ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ
ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

Διπλωματική Εργασία

Στατιστικές Μέθοδοι για την Ανάλυση Δεδομένων Υψηλής Διάστασης (Statistical Methods for the Analysis of High Dimensional Data)

ΔΡΟΣΟΥ Π. ΚΡΥΣΤΑΛΛΕΝΙΑ

Επιβλέπων: Χρήστος Κουκουβίνος
Καθηγητής Ε.Μ.Π

Αθήνα, 2013

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

Στατιστικές Μέθοδοι για την Ανάλυση Δεδομένων Υψηλής Διάστασης

ΔΡΟΣΟΥ Π. ΚΡΥΣΤΑΛΛΕΝΙΑ

Αθήνα, 2013

Πρόλογος

Τα δεδομένα υψηλής διάστασης (High Dimensional Data) είναι σήμερα ο κανόνας και όχι η εξαίρεση σε τομείς όπως η τεχνολογία της πληροφορίας (information technology), η βιοπληροφορική (bio informatics) ή η αστρονομία (astronomy). Η λέξη "υψηλής-διάστασης" αναφέρεται στην κατάσταση όπου ο αριθμός των άγνωστων παραμέτρων που πρόκειται να εκτιμηθεί είναι μία ή αρκετές τάξεις μεγέθους μεγαλύτερος από τον αριθμό των δειγμάτων στα δεδομένα. Η κλασική στατιστική συμπερασματολογία δεν μπορεί να χρησιμοποιηθεί για τέτοιου είδους προβλήματα. Για παράδειγμα, η μέθοδος ελαχίστων τετραγώνων ενός γραμμικού μοντέλου έχει πολλές περισσότερες άγνωστες παραμέτρους από τις παρατηρήσεις και έτσι ο προσδιορισμός των αντίστοιχων τυπικών σφαλμάτων και των μέτρων σημαντικότητας είναι αδύνατος. Είναι, μάλλον, προφανές ότι χωρίς επιπλέον παραδοχές – υποθέσεις, ή χωρίς περιορισμούς σε συγκεκριμένη κατηγορία μοντέλων, η στατιστική επαγωγή για μεγάλες βάσεις δεδομένων είναι αδύνατη. Ένα καθιερωμένο πλαίσιο για την προσαρμογή πολλών παραμέτρων με βάση την παραδοχή μιας ομαλής δομής επιτρέπει την εκτίμηση ομαλών συναρτήσεων. Τα τελευταία χρόνια έχουμε γίνει μάρτυρες μιας επανάστασης των μεθοδολογικών, υπολογιστικών και μαθηματικών επιτευγμάτων, που επιτρέπουν την στατιστική συμπερασματολογία για υψηλής διάστασης δεδομένα βασιζόμενη στην παραδοχή κάποιων εννοιών της σποραδικότητας (sparsity). Η μετατόπιση του κέντρου βάρους από την ομαλότητα των περιορισμών στους περιορισμούς σποραδικότητας (sparsity constraints), ή συνδυασμό των δύο, ανοίγει το δρόμο για πολλές άλλες εφαρμογές που περιλαμβάνουν σύνθετα δεδομένα. Για παράδειγμα, η παραδοχή της σποραδικότητας (sparsity assumption) ότι η κατάσταση της υγείας ενός ατόμου εξαρτάται μόνο από λίγους μεταξύ πολλών χιλιάδων βιοδεικτών φαίνεται πολύ πιο ρεαλιστική από την εξέταση ενός μοντέλου όπου χιλιάδες μεταβλητές θα συνεισφέρουν με ομαλό τρόπο στην κατάσταση της υγείας.

Περιεχόμενα

Περιεχόμενα Σχημάτων

Περιεχόμενα Πινάκων

Περίληψη

Το πρόβλημα της στατιστικής μοντελοποίησης και του εντοπισμού των σημαντικών μεταβλητών σε μεγάλα σύνολα δεδομένων είναι ένα συνηθισμένο ζήτημα στις μέρες μας. Η εργασία αυτή ασχολείται με την στατιστική ανάλυση ενός μεγάλου διαστάσεων συνόλου δεδομένων. Διεξάγουμε μία σεισμική ανάλυση ευαισθησίας κινδύνου χρησιμοποιώντας σεισμικά δεδομένα που αποκτήθηκαν στην Ελλάδα κατά τη διάρκεια των ετών 1962-2003. Ο κύριος σκοπός της ανάλυσης είναι η εξαγωγή γνώσης υψηλού επιπέδου για τη χρήση ή τη λήψη αποφάσεων σ΄αυτό τον τομέα. Οκτώ μη παραμετρικοί ταξινομητές που προέρχονται από μεθόδους εξόρυξης δεδομένων (πολυστρωματικά Perceptrons (MLP) Νευρωνικά Δίκτυα, Radial Basis Function Νευρωνικά (RBFN) δίκτυα, δίκτυα Bayes, διανυσματικές μηχανές υποστήριξης (SVMs), δέντρα ταξινόμησης και παλινδρόμησης (C & RT, CHAID, C5.0 αλγόριθμο, QUEST) μας απασχολούν σε αυτή την εργασία, σε σύγκριση με τη Λογιστική Παλινδρόμηση και L_1-νόρμα SVM όσον αφορά τη συνολική ακρίβεια ταξινόμησης, την ευαισθησία, την ειδικότητα και την περιοχή κάτω από την καμπύλη ROC (AUROC). Ο στόχος αυτής της εργασίας είναι διπλός. Αφενός να αξιολογήσει τη σημασία των διαφόρων μεταβλητών εισόδου, προκειμένου να εντοπίσει τους πιθανούς παράγοντες κινδύνου των μεγάλων σεισμών και αφετέρου να εξετάσει ποιοι ταξινομητές είναι οι πλέον κατάλληλοι για μία μεγάλων διαστάσεων ανάλυση δεδομένων, ανιχνεύοντας αποτελεσματικά τις σύνθετες μη γραμμικές σχέσεις και ενδεχομένως οδηγώντας σε πιο ακριβείς προβλέψεις.

Συγκεκριμένα, το πρώτο κεφάλαιο ασχολείται με την βασική ιδέα των δεδομένων υψηλής διάστασης και το δεύτερο με τις τεχνικές εξόρυξης δεδομένων και μηχανικής μάθησης. Το κεφάλαιο 3 παρουσιάζει βασικές μεθόδους για τη μείωση των δεδομένων, όπως η παραγοντική ανάλυση και ανάλυση των κυρίων συνιστωσών. Στο τέταρτο κεφάλαιο αναφερόμαστε σε μεθόδους ταξινόμησης και παρουσιάζουμε, όπως έχουμε ήδη αναφέρει, οκτώ μη-παραμετρικούς ταξινομητές (Τεχνητά Νευρωνικά Δίκτυα (MLP, RBFN), Bayesian δίκτυα, Μηχανές Διανυσματικής Υποστήριξης (SVM), Δέντρα Ταξινόμησης και Παλινδρόμησης (C&RT, CHAID, QUEST , C5.0), Λογιστική Παλινδρόμηση και l1-νόρμα svm). Το κεφάλαιο 5 αναφέρεται στην αξιολόγηση ενός μοντέλου με τη χρήση μεθόδων, όπως η πολλαπλή επικύρωση και στην απόδοση των ταξινομητών που αναφέρονται παραπάνω. Επιπλέον, συζητούνται οι όροι της ευαισθησίας και ειδικότητας και γίνεται μια σύντομη αναφορά στην περιοχή κάτω από την καμπύλη (AUC). Στο τελευταίο - κεφάλαιο 6 παρουσιάζουμε το λογισμικό Clementine το οποίο θα εφαρμοστεί σε σεισμικά δεδομένα και προχωράμε στην συνέχεια με την εφαρμογή και την ερμηνεία των αποτελεσμάτων.

Abstract

The problem of statistical modelling and identifying the significant variables in large data sets is common nowadays. This paper deals with the statistical analysis of a large dimensional data set; we conduct with a seismic hazard sensitivity analysis using seismic data from Greece acquired during the years 1962 - 2003. The main purpose of the analysis is to extract high-level knowledge for the domain user or decision-maker. Eight non parametric classifiers derived from data mining methods (Multilayer Perceptrons (MLP) Neural Networks, Radial Basis Function Neural (RBFN) Networks, Bayesian Networks, Support Vector Machines (SVMs), Classification and Regression Tree (C&RT), Chi-square Automatic Interaction Detection (CHAID), C5.0 algorithm and Quick, Unbiased, Efficient Statistical Tree (QUEST)) are employed in this work, and are compared to Logistic Regression and l_1-norm SVM in terms of overall classification accuracy, sensitivity, specificity, and Area under the ROC curve (AUROC). The goal of this paper is twofold; assess the importance of several input variables in order to detect the possible risk factors of large earthquakes and examine which classifiers are most suited for a large dimensional data analysis, detecting effectively complex nonlinear relationships and potentially lead to more accurate predictions.

Specifically, the first Chapter deals with the main concept of high dimensional data and the second one with data mining and machine learning techniques. Chapter 3 present basic methods for data reduction such as factor analysis and principal component analysis. In the fourth chapter we refer to Classification methods and present, as we have already mentioned, eight non-parametric classifiers (Artificial Neural Networks (MLP, RBFN), Bayesian Networks, Support vector Machines, Classification and Regression Tree (C&RT, CHAID, QUEST, C5.0), Logistic Regression and l_1-norm support vector machine). Chapter 5 refers to the evaluation of a model using methods like cross validation and to the performance of the classifiers mentioned above. In addition the terms of sensitivity and specificity are discussed and a brief reference to the the area under the curve (AUC) is presented as well. In the final-chapter 6 we present the Clementine software which we applied to seismic data and proceed with the implementation and interpretation of the results.

Ευχαριστίες

Η εκπόνηση της παρούσας διπλωματικής εργασίας πραγματοποιήθηκε υπό την επίβλεψη του καθηγητή του Εθνικού Μετσόβιου Πολυτεχνείου, κ. Χρήστο Κουκουβίνο, τον οποίο θα ήθελα να ευχαριστήσω θερμά για την ανάθεση της συγκεκριμένης εργασίας, δίνοντας μου με αυτό τον τρόπο την δυνατότητα να ασχοληθώ με ένα θέμα το οποίο ανήκει στα ερευνητικά μου ενδιαφέροντα.

Παράλληλα ιδιαίτερες ευχαριστίες θα ήθελα να εκφράσω στους υποψήφιους διδάκτορες Ανδρουλάκη Μάνο και Παρπούλα Χριστίνα, για την πολύτιμη βοήθεια τους και το συνεχές ενδιαφέρον κατά τη διάρκεια εκπόνησης της διπλωματικής μου εργασίας.

Θα ήθελα επίσης να εκφράσω την ευγνωμοσύνη μου στους γονείς μου για την διαρκή τους υποστήριξη, που επέτρεψε την επιτυχή διεκπαιρέωση των σπουδών μου.

Χωρίς τη συμβολή των παραπάνω ανθρώπων θα ήταν αδύνατη η δημιουργία και παρουσίαση αυτής της εργασίας.

Δρόσου Κρυσταλλένια

Εθνικό Μετσόβιο Πολυτεχνείο,
Σχολή Εφαρμοσμένων Μαθηματικών
και Φυσικών Επιστημών
Αθήνα, 2013

Κεφάλαιο 1:

Δεδομένα υψηλής διάστασης

1.1 Τι είναι τα υψηλής διάστασης δεδομένα;

Πολλές τάσεις στον τομέα της στατιστικής, καθοδηγούνται από τους τύπους των συνόλων δεδομένων που παράγονται από τη βιομηχανία και την επιστήμη, ιδιαίτερα όταν αυτοί παρουσιάζουν νέα και γεμάτα προκλήσεις προβλήματα. Αναμφίβολα, από τέτοιου είδους προβλήματα προέκυψε, αυτό που αναφέρεται ως «μεγάλων διαστάσεων» δεδομένα (high dimensional data). Συγκεκριμένα, ο όρος αναφέρεται σε καταστάσεις όπου υπάρχουν πολλές μεταβλητές ή στοιχεία, που είναι διαθέσιμα για χρήση σε οποιοδήποτε στατιστικό μοντέλο ή ανάλυση. Ένα μοντέλο εδώ είναι κάθε στατιστικό πλαίσιο που κατασκευάστηκε χρησιμοποιώντας τέτοια δεδομένα, και πιο συχνά αναφέρεται σ'ένα προγνωστικό μοντέλο (predictive model), όπου οι μεταβλητές χρησιμοποιούνται στην πρόβλεψη ενός συγκεκριμένου γεγονότος με βάση τα δεδομένα αυτά.

Κύρια πηγή των δεδομένων αυτών είναι η *βιομηχανία*, ιδιαίτερα οργανώσεις με μεγάλες βάσεις δεδομένων των πελατών. Το χαμηλό κόστος της ψηφιακής αποθήκευσης, επιτρέπει τεράστιες ποσότητες πληροφοριών που αφορούν κάθε πελάτη να συλλέγονται σχετικά φτηνά και στη συνέχεια τα χαρακτηριστικά αυτής της βάσης δεδομένων εξάγονται και αναλύονται στατιστικά. Ένας οργανισμός μπορεί να ζητήσει να κάνει χρήση μιας βάσης δεδομένων είτε για την καλύτερη πρόβλεψη της συμπεριφοράς των πελατών, είτε για την καλύτερη κατανόηση των αλλαγών στη βάση των πελατών τους, ή ακόμη και για τον προσδιορισμό του προϊόντος για το οποίο υπάρχει μεγαλύτερη ζήτηση. Ας πάρουμε, για παράδειγμα, το σύνολο δεδομένων που χρησιμοποιείται στην 2009 KDD CUP1, η οποία

αποτελεί μέρος του ετήσιου διαγωνισμού εξόρυξης δεδομένων. Το σύνολο δεδομένων εκπαίδευσης (training set) (που χρησιμοποιείται για την οικοδόμηση στατιστικών μοντέλων) αποτελείται από 50.000 πελάτες μιας γαλλικής τηλεφωνικής εταιρείας. Για καθένα από τους πελάτες υπάρχουν 15.000 μεταβλητές, ή χωριστά κομμάτια πληροφοριών που είναι διαθέσιμα. Οι μεταβλητές αυτές κωδικοποιούν σχεδόν τα πάντα που η εταιρεία γνωρίζει για τον πελάτη, από δημογραφικά στοιχεία μέχρι τα τρέχοντα και παρελθοντικά προϊόντα που έχουν χρησιμοποιηθεί. Σ'αυτό το συγκεκριμένο πρόβλημα, ο σκοπός ήταν να χρησιμοποιηθούν αυτές οι πληροφορίες για να προβλεφθεί η πιθανότητα ο πελάτης να επιλέξει έναν ανταγωνιστή, καθώς και το εάν ο πελάτης θα αναβαθμίσει το συμβόλαιο κινητής τηλεφωνίας ή εάν θα ανταποκριθεί σε κάποιο υλικό μάρκετινγκ. Η αφθονία των πληροφοριών, επέτρεψε αυτές οι προβλέψεις να εκτελούνται με εκπληκτική ακρίβεια.

Μια άλλη βασική πηγή των μεγάλων βάσεων δεδομένων είναι ο τομέας της *βιολογίας* και των συναφών επιστημών. Ο σύγχρονος εξοπλισμός επιτρέπει την ταυτόχρονη μέτρηση πολλών διαφορετικών συστατικών, τα οποία συλλέγονται σε μία ενιαία βάση. Για παράδειγμα, τα πειράματα γενετικής με μικροδιανύσματα (microarray) μετρούν τη σχετική δραστηριότητα χιλιάδων γονιδίων από ένα μόνο δείγμα ιστού. Αυτό επιτρέπει σε έναν ερευνητή να ψάξει για τα γονίδια που είναι ιδιαίτερα δραστήρια σε μία ομάδα πειραμάτων συγκριτικά με κάποια άλλη. Για παράδειγμα, οι ασθενείς με ένα συγκεκριμένο καρκίνο μπορεί να έχουν ένα γονίδιο το οποίο εμφανίζει πολύ μεγαλύτερη δραστηριότητα από εκείνη που παρατηρήθηκε σε ασθενείς χωρίς καρκίνο. Αυτό θα οδηγήσει τους ερευνητές να εστιάσουν σε αυτό το συγκεκριμένο γονίδιο με την ελπίδα της καλύτερης κατανόησης του καρκίνου. Μία βασική διαφορά μεταξύ των δύο παραπάνω παραδειγμάτων είναι ο αριθμός των παρατηρήσεων. Ενώ η τάξη μεγέθους των μεταβλητών ή συστατικών είναι παρόμοια για τα δύο παραδείγματα που αναφέραμε προηγουμένως (σε χιλιάδες), η βάση δεδομένων των πελατών έχει 50.000 εγγραφές από τις οποίες μπορεί να αντλήσει πληροφορίες, ενώ το δεύτερο πείραμα (μικροσυστοιχιών) θα περιλαμβάνει σπάνια περισσότερα από εκατό δείγματα. Αυτό επηρεάζει δραματικά το τι μπορεί να επιτευχθεί για κάθε πρόβλημα. Εάν υπάρχουν πολλές παρατηρήσεις, ένα εξαιρετικά πολύπλοκο μοντέλο ενσωματώνει εκατοντάδες παράγοντες και αλληλεπιδράσεις που θα μπορούσαν να κατασκευαστούν με τη λογική ότι οι περισσότεροι από αυτούς είναι πραγματικά σημαντικοί. Αντιστρόφως, όταν υπάρχει ένας μικρός αριθμός παρατηρήσεων η ανάλυση είναι συνήθως, αναγκαστικά, απλούστερη, με έμφαση στην αξιόπιστη ανίχνευση μόνο το κύριων επιδράσεων ή μεταβλητών.

Αξίζει να σημειωθεί εδώ ότι τα προβλήματα που αντιμετωπίζονται στο πλαίσιο μεγάλων βάσεων δεδομένων (high dimensional statistics) είναι σχετικά πρόσφατα και η προσπάθεια επίλυσής τους γίνεται με παραδοσιακές τεχνικές σε νέου τύπου βάσεις δεδομένων. Ωστόσο, αυτές οι βάσεις συχνά υπονομεύουν μία παραδοσιακή τεχνική. Πολλές παραδοσιακές τεχνικές υπάρχουν για την επίλυση προβλημάτων χαμηλής διάστασης (low dimensional data) αλλά χρειάζονται επανεξέταση για να λειτουργήσουν σε ένα καινούργιο πλαίσιο

δεδομένων. Μια σύμβαση στη στατιστική είναι ο αριθμός των παρατηρήσεων να καλείται n, και ο αριθμός των μεταβλητών πρόβλεψης p.

1.2 Τι κάνει ιδιαίτερη τη στατιστική ανάλυση των δεδομένων υψηλής διάστασης

Η στατιστική συμπερασματολογία για τα δεδομένα υψηλής διάστασης είναι ιδιαίτερα ενδιαφέρουσα όχι μόνο επειδή περιλαμβάνει ενδιαφέρουσες εφαρμογές, αλλά επειδή ένα μεγάλο μέρος της παραδοσιακής στατιστικής ανάλυσης θα πρέπει να επανεξεταστεί. Στη συνέχεια ακολουθούν μερικά παραδείγματα.

1.2.1 Σημαντικότητα και ψευδώς θετικά ποσοστά

Ας υποθέσουμε ότι έχουμε ένα μεγάλων διαστάσεων σύνολο δεδομένων στο οποίο κάθε παρατήρηση ανήκει σε μία από τις δύο πιθανές τάξεις, και καθεμία από τις p μεταβλητές είναι συνεχής. Ένα παραδοσιακό t-test μπορεί να εκτελεστεί για κάθε μεταβλητή μεμονωμένα, αξιολογώντας έτσι πόσο σημαντική είναι η παρατηρούμενη διαφορά στο μέσο μεταξύ των δύο κλάσεων. Για ένα δεδομένο επίπεδο σημαντικότητας, αναμένουμε ότι αυτή η αναλογία των μεταβλητών, η οποία δεν έχει καμία πραγματική σχέση με τις κλάσεις, θα παραβιάσει αυτό το επίπεδο σημαντικότητας; και αυτό το επίπεδο που αφορά στις περιττές μεταβλητές είναι αυτό που εμφανίζεται ως ψευδώς θετικό αποτελέσματα. Σε επίπεδο σημαντικότητας 5% και σε σύνολο μεταβλητών p = 10.000, αυτό θα σήμαινε ότι έως και 500 περιττές μεταβλητές θα εμφανίζονταν σημαντικές, εκτός από κάποιες που είναι πραγματικά σημαντικές. Έτσι, υπάρχει μεγάλη πιθανότητα πολλών ψευδώς θετικών (false positives) ποσοστών που εμποδίζουν την ικανότητα να εντοπιστούν τυχόν αληθή χαρακτηριστικά. Υπάρχουν αρκετές αναφορές σχετικά με το πώς μπορεί να επιλεγεί το επίπεδο σημαντικότητας σε τέτοιες περιπτώσεις. Μία δημοφιλής προσέγγιση είναι να γίνει μία προσπάθεια ελέγχου του ψεύτικου ποσοστού που ανακαλύφθηκε (ο αριθμός των ψευδώς θετικών διαιρεμένος με τον αριθμό των απορρίψεων, βλέπε Benjamini και Hochberg (1995) και Farcomeni (2008)). Ακόμη και αν αυτό δίνει καλύτερη αίσθηση για το επίπεδο των σφαλμάτων που σχετίζονται με ένα πρόβλημα, το θεμελιώδες πρόβλημα παραμένει; η πραγματική επίδραση πρέπει να εμφανίζεται πολύ έντονα για να διακρίνεται σαφώς ο θόρυβος.

1.2.2 Τα προβλήματα της υπερπροσαρμογής (overfitting[1])

Η ενότητα αυτή ξεκινάει με ένα μικρό πείραμα. Παίρνουμε ένα διάνυσμα δεδομένων (π.χ. leukemia microarray (βλ. Miller, H. R., (2010)). Αυτό το σύνολο δεδομένων περιλαμβάνει n = 72 παρατηρήσεις, p = 7129 μεταβλητές και κάθε παρατήρηση ανήκει σε μία από τις δύο κλάσεις (τύπος οξείας Λευχαιμίας). Ας υποθέσουμε ότι επιλέγουμε τυχαία τα 2/3 των δεδομένων για την εκπαίδευση (training set) και εφαρμόζουμε λογιστική παλινδρόμηση (Hosmer και Lemeshow,2000), χρησιμοποιώντας 20 τυχαία επιλεγμένες μεταβλητές. Στη συνέχεια, γίνεται έλεγχος για το πόσο καλά το μοντέλο αυτό προβλέπει για τα δύο τρίτα των δεδομένων, σε σύγκριση με το υπόλοιπο ένα τρίτο. Αυτό το πείραμα εφαρμόζεται 50 φορές και τα αποτελέσματα υπολογίζονται κατά μέσο όρο. Τα μοντέλα προβλέφθηκαν σωστά με ακρίβεια 99% στο σύνολο εκπαίδευσης (training set), ενώ η ακρίβεια του συνόλου ελέγχου (test set) ήταν κάτω από το 70%, έτσι δεν είναι πολύ μακριά από το 50% που θα περίμενε κανείς από ένα καθαρά τυχαίο μοντέλο.

Αυτό δείχνει μια γενική αρχή, η οποία είναι ότι όταν υπάρχουν πολλές μεταβλητές, είναι πολύ εύκολο να χτιστεί ένα μοντέλο που προσαρμόζεται άρτια στα δεδομένα. Με αυτό εννοούμε ότι η φαινομενική επίδοση του μοντέλου για το σύνολο δεδομένων εκπαίδευσης είναι υπερβολικά αισιόδοξη, και η απόδοση του μοντέλου σε νέα, «αόρατα» δεδομένα θα είναι μη αποδοτική. Αυτό αποτελεί μία ιδιαίτερη πρόκληση σε καταστάσεις όπου το n είναι μικρό και το p μεγάλο, διότι είναι πολύ εύκολο να κατασκευαστεί ένα μοντέλο που εμφανίζεται ισχυρό μεν, αλλά έχει απογοητευτικές επιδόσεις στο μέλλον δε.

1.2.3 Υπολογιστική πολυπλοκότητα

Παρά την αναπόφευκτη αλήθεια ότι κάθε σχόλιο στην υπολογιστική επιβάρυνση θα ήταν παρωχημένο, θα επισημανθούν ορισμένες παρατηρήσεις σχετικά με πρακτικούς περιορισμούς στην ανάλυση δεδομένων υψηλής διάστασης. Η κύρια παρατήρηση είναι ότι αν το p είναι μεγάλο, κάθε μέθοδος που παίρνει χρόνο $0(p^a)$ για κάποιο χρονικό διάστημα $\alpha > 1$ είναι πιθανό να είναι ανέφικτη. Αυτό περιορίζει ορισμένα είδη προσεγγίσεων. Για παράδειγμα, αν υπάρχει η υπόνοια ότι κάποια απόκριση εξαρτάται από το πολύ k μεταβλητές, τότε ο πιο φυσικός (και βέλτιστος) τρόπος ανεύρεσης αυτών θα ήταν να

[1] Το overfitting συμβαίνει όταν ένα στατιστικό μοντέλο περιγράφει το τυχαίο σφάλμα ή το θόρυβο αντί της υποκείμενης σχέσης. Overfitting εμφανίζεται γενικά όταν ένα μοντέλο είναι υπερβολικά πολύπλοκο, όπως όταν έχει πάρα πολλές παραμέτρους σε σχέση με τον αριθμό των παρατηρήσεων. Ένα μοντέλο το οποίο έχει overfit θα έχει γενικά κακή προγνωστική απόδοση, καθώς μπορεί να διογκωθούν μικρές διακυμάνσεις στα δεδομένα.

δοκιμαστούν όλα τα δυνατά υποσύνολα των k μεταβλητών και να επιλεχθεί αυτό που αποδίδει καλύτερα, σύμφωνα με κάποια μέτρα. Ωστόσο, υπάρχουν $\binom{p}{k} = 0(p^k)$ τέτοια υποσύνολα, τα οποία μεγαλώνουν ραγδαία ανάλογα τα p και k. Για παράδειγμα, αν $p = 10.000$ τότε υπάρχουν περίπου 50 εκατομμύρια δυνατότητες, όταν το k είναι μόνο δύο. Σε αυτές τις συνθήκες, είναι απαραίτητη η ανεύρεση τρόπων για την αποφυγή τέτοιων υπολογισμών.

1.3 Χρυσοί κανόνες

Μετά από την προηγούμενη ενότητα, αυτές είναι κάποιες κατευθυντήριες γραμμές που έχουν προκύψει από έρευνες σε υψηλής διάστασης δεδομένα. Είναι ένα μείγμα, κοινής λογικής και ενός σημαντικού ποσοστού πειραματισμού.

Δεν υπάρχει ενιαία προσέγγιση με καλή επίδοση σε όλα τα προβλήματα μεγάλων διαστάσεων.

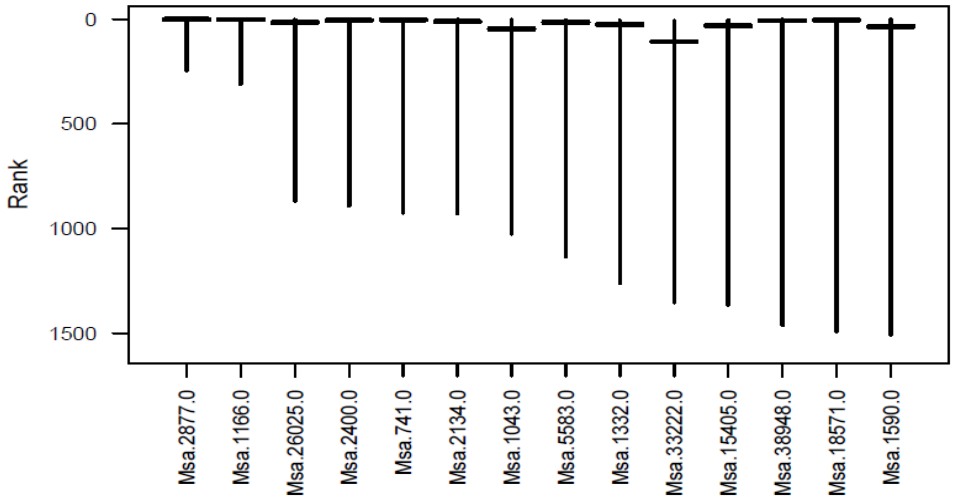
Ακριβώς όπως και στα παραδείγματα που προηγήθηκαν παρουσιάστηκε ένα ευρύ φάσμα των τύπων προβλημάτων, ο καλύτερος τρόπος για την προσέγγιση ενός προβλήματος, είτε πρόκειται για ένα πρόβλημα πρόβλεψης είτε για κάτι άλλο, θα διαφέρει σημαντικά. Αν και αυτό, προφανώς, αποτελεί ένα όφελος για τους εν ενεργεία ερευνητές, οι οποίοι μπορεί να συνεχίσει να ψάξουν αποδοτικές μεθόδους για ποικίλα σενάρια, αυτό σημαίνει ότι η ανάθεση της «καλύτερης» ή «χειρότερης» μεθόδου έχει χάσει τη σημασία της σε τέτοιου είδους προβλήματα.

Ως ένα παράδειγμα, ο Dettling (2004) συγκρίνει την ακρίβεια των επτά ταξινομητών σε έξι (microarray) συνόλων δεδομένων. Τέσσερις από τις μεθόδους αποδίδουν καλύτερα σε τουλάχιστον ένα από αυτά, αποδεικνύοντας ότι ακόμη και όταν υψηλής διαστασιολογικά δεδομένα περιορίζονται σε έναν απλό τύπο, η ανεύρεση της καλύτερης μεθόδου δεν είναι δυνατή.

Το ποσοστό ανίχνευσης των πραγματικών επιδράσεων με ακρίβεια είναι συχνά τρομακτικά μικρό

Αυτό σχετίζεται με το θέμα των ψευδώς θετικών που περιγράφονται παραπάνω. Μεγάλο μέρος της διατριβής του Miller (2010) ασχολείται με την εκτίμηση της ακρίβειας της κατάταξης, όπου οι μεταβλητές ενός συνόλου δεδομένων διατάσσονται σύμφωνα με κάποιο

κριτήριο σημαντικότητας. Το Σχήμα 1 δίνει ένα παράδειγμα μιας τέτοιας κατάταξης στην οποία χρησιμοποιούνται 90% διαστήματα εμπιστοσύνης για την κατάταξη καθεμιάς από τις πιο σημαντικές μεταβλητές (14 μεταβλητές) που συμπεριλαμβάνονται (υπολογίζεται με τη βοήθεια του bootstrap), από ένα σύνολο δεδομένων μικροσυστοιχιών με n = 30 και ρ = 6319. Μπορούμε να παρατηρήσουμε τα ευρεία διαστήματα για κάθε μεταβλητή, συμπεριλαμβανομένων εκείνων που κρίνονται ότι είναι τα πιο σημαντικά. Έτσι, όταν το πείραμα πραγματοποιήθηκε μία δεύτερη φορά, η πλέον σημαντική μεταβλητή από την τρέχουσα κατάταξη θα μπορούσε εύλογα να αναμένεται να καταταχθεί οπουδήποτε στις 200 πιο σημαντικές, που σημαίνει ότι είναι πολύ απίθανο ότι θα ανιχνευθεί και πάλι ως ένας ιδιαίτερα σημαντικός παράγοντας.



Σχήμα 1: Σημαντικές μεταβλητές και κατάταξη διαστημάτων εμπιστοσύνης για το παράδειγμα (Ro131) του κεφαλαίου 2(Miller (2010)).

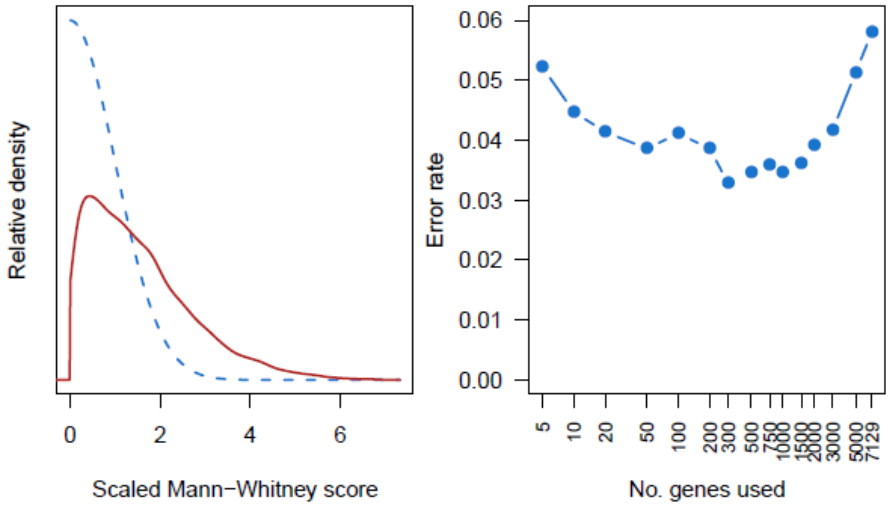
Η υπόθεση της αραιότητας (sparsity) είναι σχεδόν πάντα μη ρεαλιστική, αλλά σχεδόν πάντα χρήσιμη

Ένα αραιό μοντέλο (sparse) είναι εκείνο που χρησιμοποιεί σχετικά λίγες από τις διαθέσιμες μεταβλητές. Οι ποινικοποιημένες μέθοδοι, δίνουν συχνά αραιές λύσης και γίνονται όλο και πιο δημοφιλείς σήμερα. Η αρχή που διέπει ο παραπάνω κανόνας είναι ότι, ακόμη και αν η πραγματική κατάσταση δεν είναι αραιή, υπάρχει μικρή πιθανότητα της σωστής ενσωμάτωσης όλων αυτών των επιδράσεων, και έτσι ένα αραιό μοντέλο που ενσωματώνει μόνο τις ισχυρότερες μεταβλητές θα έχει γενικά καλύτερες επιδόσεις. Για παράδειγμα, το 2009 η ομάδα του Πανεπιστημίου της Μελβούρνης (Melbourne KDD Cup 2009 team) έκτισε μοντέλα πρόβλεψης για το σύνολο των δεδομένων, που χρησιμοποιούν

μόνο 200 μεταβλητές, λιγότερο από το 2% των διαθέσιμων μεταβλητών. Παρά το γεγονός της αραιότητας, αυτά τα μοντέλα ήταν αρκετά ισχυρά για να κερδίσουν μέρος του διαγωνισμού (βλ. Miller et al, 2009) έτσι η αραιότητα, παρά τη μείωση της ακρίβειας ενός συγκεκριμένου μοντέλου, αφαιρεί το σχετικό θόρυβο με την εκτίμηση των αδύναμων στοιχείων. Ως δεύτερο, κάπως πιο περίπλοκο, παράδειγμα αυτής της αρχής, εξετάζονται τα γραφήματα της συνάρτησης πυκνότητας πιθανότητας στον αριστερό πίνακα του Σχήματος 2. Αυτό δείχνει την κατανομή των κλιμακωτών του Mann-Whitney αποτελεσμάτων δοκιμής για καθένα από τα 7129 γονίδια στο σύνολο δεδομένων των μικροσυστοιχιών λευχαιμίας, το οποίο περιγράφεται στον Miller (2010). Το γράφημα περιέχει μια πιο λεπτομερή περιγραφή της μεθοδολογίας. Η διακεκομμένη γραμμή παριστάνει την πυκνότητα των αποτελεσμάτων του τεστ υποθέτοντας ότι δεν υπήρχε καμία σχέση μεταξύ των επιπέδων και των δύο κατηγοριών της απόκρισης. Η μεγάλη παρέκκλιση της πραγματικής πυκνότητας από αυτό υποδηλώνει ότι ένα καλό ποσοστό των γονιδίων, ίσως τουλάχιστον το 30%, έχουν κάποια σχέση με την απόκριση. Ας υποθέσουμε ότι θέλουμε να οικοδομήσουμε ένα τυχαίο μοντέλο πρόβλεψης (Breiman, 2001a) χρησιμοποιώντας τα d κορυφαία γονίδια κατάταξης που βασίζονται στο παρόν Mann-Whitney στατιστικό αποτέλεσμα της δοκιμής.

Με βάση τις παραπάνω παρατηρήσεις, κάποιος μπορεί να σκεφτεί ότι ένα καλό μοντέλο μπορεί να χρειάζεται τουλάχιστον $d = 2000$. Ωστόσο, τα αποτελέσματα στον δεξί πίνακα του Σχήματος 2 (όπου έχουμε διασπάσει επανειλημμένα τα δεδομένα σε δύο τρίτα εκπαίδευση / ένα τρίτο δοκιμή, επιλέχθηκαν γονίδια και χτίστηκαν μοντέλα χρησιμοποιώντας τα δεδομένα εκπαίδευσης, και μετρήθηκε η απόδοση για το σύνολο ελέγχου), δείχνουν ότι, προτιμάται το μοντέλο μεγέθους 300, ή μια τάξη μεγέθους μικρότερη. Έτσι, υποπίπτοντας στην πλευρά της αραιότητας μπορεί να βελτιώσουμε συχνά την προγνωστική ακρίβεια.

Ως μια μικρή προειδοποίηση για το παραπάνω επιχείρημα, πιστεύεται ότι πολλά γονιδιακά προβλήματα μπορεί στην πραγματικότητα να είναι λιγότερο αραιά απ' ότι πιστεύονταν αρχικά (βλέπε για παράδειγμα, Goldstein, 2009, Hirschhorn, 2009, και η Kraft και Hunter, 2009). Βλέπε Hall et αl. (2010) για τις πρόσφατες προσπάθειες για την κατασκευή αποτελεσματικών μοντέλων με χαμηλότερους βαθμούς αραιότητας.



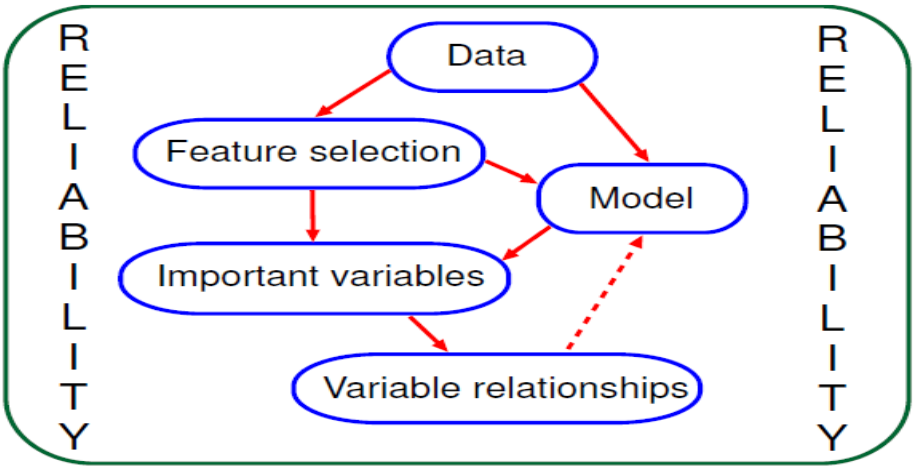
Σχήμα 2: Επιλογή μοντέλου για το σύνολο δεδομένων Λευχαιμία. Ο αριστερός πίνακας δείχνει την κατανομή του στατιστικού τεστ Mann-Whitney για τα 7129 γονίδια στο σύνολο δεδομένων, σε σύγκριση με την μηδενική (διακεκομμένη). Ο δεξιός πίνακας παρουσιάζει τα μη ταξινομημένα ποσοστά για τα τυχαία μοντέλα όπου χρησιμοποιούμε διαφορετικούς αριθμούς από τα γονίδια που κατατάχθηκαν στις κορυφαίες θέσεις, σύμφωνα με το τεστ. Περίπου 300 γονίδια φαίνονται βέλτιστα.

Επικύρωση, και σωστή επικύρωση

Στην ενότητα 1.2.2 έχει ήδη διερευνηθεί πόσο εύκολο είναι να παραχθεί ένα overfitted μοντέλο. Στο συγκεκριμένο παράδειγμα που δίνεται, τα μοντέλα ήταν σαφές ότι ήταν φτωχά, επειδή είχαν επικυρωθεί. Ο πιο συνηθισμένος τρόπος για να γίνει αυτό είναι με την εκπαίδευση-δοκιμή επικύρωση (train-test validation) ή τη διασταυρωμένη επικύρωση (cross-validation) (βλέπε Hastie et al., 2001), όπου μέρος των δεδομένων προορίζεται να αξιολογήσει την προσαρμογή του μοντέλο στο τέλος της διαδικασίας. Αν και απλό, αυτό παραμένει ένας από τους λίγους αποδοτικούς τρόπους για να εξακριβώσουμε σωστά το πόσο καλά εκτελείται μια τεχνική. Είναι επίσης ιδιαίτερα σημαντική η σωστή πραγματοποίηση της επικύρωσης. Ένα παράδειγμα που το απεικονίζει είναι αυτό όπου πραγματοποιείται αρχικά ένα βήμα επιλογής μεταβλητών (variable selection), και ακολουθείται από ένα βήμα επιλογής μοντέλου. Αν η επιλογή μεταβλητών γίνει με το συνολικό σύνολο δεδομένων, τότε ακόμη και αν το βήμα επιλογής του μοντέλου επικυρώνεται το τελικό μοντέλο θα είναι υπερβολικά αισιόδοξο. Για το λόγο αυτό, η ιδέα της χρησιμοποίησης δύο στρωμάτων της διασταυρωμένης επικύρωσης, το οποίο εισήχθη από τον Stone(1974), χρησιμοποιείται συνήθως σε πολλά σενάρια (και συνεπώς έχει ενσωματωθεί εντός του βιοστατιστικού πακέτου της R Rmagpie2 για την ανάλυση των

δεδομένων microarray). Η εργασία στο κεφάλαιο 8 του Miller(2010) παρέχει ένα άλλο ρητό σενάριο όπου αυτή η μεθοδολογία είναι κατάλληλη.

Αξίζει να επισημάνουμε, κλείνοντας, ότι η εικόνα που ακολουθεί είναι ιδιαίτερα χαρακτηριστική στην περίπτωση που εργαζόμαστε με δεδομένα υψηλής διάστασης, εφόσον πολλοί είναι οι συγγραφείς που εργάζονται με βάση αυτό το πλάνο. Για παράδειγμα ο Miller (2010) στο σύνολο της διδακτορικής του διατριβής εργάζεται αρκετά συχνά με μειωμένο αριθμό μεταβλητών.



Σχήμα 3: Ένα πιθανό πλαίσιο για τη στατιστική ανάλυση δεδομένων υψηλής διάστασης

1.4 Από την απλή γραμμική στην υψηλής διάστασης παλινδρόμηση (From simple linear regression to high dimension)

Σε αυτή την ενότητα, υποδεικνύεται η φύση του γενικού θέματος της ανάλυσης δεδομένων υψηλής διάστασης ξεκινώντας με ένα πολύ στοιχειώδη στατιστικό μοντέλο δείχνοντας πώς μια απλή διαδικασία της εξέλιξης μας οδηγεί γρήγορα σε σφαίρες, όπου οι δυσκολίες της μεγάλης διάστασης καθίστανται σαφείς.

(α) Το παραδοσιακό σενάριο

Το απλό γραμμικό μοντέλο περιλαμβάνει δύο μεταβλητές, την ανεξάρτητη ή αλλιώς επεξηγηματική μεταβλητή x και την εξαρτημένη ή διαφορετικά τη μεταβλητή απόκρισης y, οι οποίες συνδέονται μεταξύ τους με τη γραμμική συνάρτηση παλινδρόμησης. Έχουμε,

λοιπόν, τα δεδομένα στη μορφή ενός ζεύγους των μετρήσεων $\{x_i\,, y_i\,; i = 1, 2, \dots, n\}$. Οι δύο μεταβλητές, υποθέτουμε ότι συνδέονται μέσω της σχέσης :

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$$

για κάθε i.

Υποθέτουμε επίσης ότι, ανεξάρτητα για κάθε i, $\varepsilon_i \sim N(0\,, \sigma^2)$ δηλαδή το ε_i, που ονομάζεται τυχαίο σφάλμα, ακολουθεί την κανονική κατανομή με μέσο μηδέν και διασπορά σ^2. Έτσι, η «μέση» σχέση μεταξύ της μεταβλητής πρόβλεψης και της μεταβλητής απόκρισης ακολουθεί μια ευθεία γραμμή με σταθερά β_1 και κλίση β_2 . Μια τέτοια ευθεία θα έχει τη μορφή :

$$E(y_i \mid x_i) = \beta_1 + \beta_2\, x_i$$

και αναφέρεται ως απλό γραμμικό μοντέλο παλινδρόμησης του y επί του x, και χαρακτηρίζεται «απλό» επειδή υπάρχει μόνο μία επεξηγηματική μεταβλητή. Οι δύο «παράμετροι», β_1 και β_2, είναι άγνωστες σταθερές που πρέπει να εκτιμηθούν. Σε μία πιθανή εφαρμογή, η μεταβλητή πρόβλεψης θα μπορούσε να είναι η «ηλικία» και η μεταβλητή απόκρισης μπορεί να είναι «συστολική αρτηριακή πίεση. (Συνήθως, σ^2 είναι επίσης άγνωστη σταθερά, αλλά για λόγους απλότητας εδώ θα το πάρουμε ως γνωστό).

Υπάρχει ένας βολικός συμβολισμός με τη χρήση διανύσματος-πίνακα, για το μοντέλο για το πλήρες σύνολο των n ζευγών των δεδομένων ορίζεται,

$$y = X\beta + \varepsilon$$

όπου το $n \times 1$ διάνυσμα y περιέχει τις αποκρίσεις, το διάνυσμα β περιέχει τις δύο (γενικά p) παραμέτρους εκτός του σ^2, το $n \times 1$ διάνυσμα ε περιέχει το «θόρυβο» και το $n \times \rho$ είναι ο λεγόμενος πίνακας σχεδιασμού X ολοκληρώνει το μοντέλο. Στην απλή γραμμική παλινδρόμηση, κάθε στοιχείο στην πρώτη στήλη του X είναι 1.

Ο συνήθης τρόπος εκτίμησης της άγνωστης κλίσης και του σημείου τομής στο β είναι να χρησιμοποιήσουμε την προσέγγιση ελαχίστων τετραγώνων του Gauss και να βρούμε

$$\hat{\beta} = arg\min_{\beta} \sum_i (y_i - \beta_1 - \beta_2 x_i)^2,$$

πράγμα που σημαίνει ότι το β είναι ο ελαχιστοποιητής του αθροίσματος των τετραγώνων της συνάρτησης στη δεξιά πλευρά. Στο γενικό συμβολισμό διανύσματος-πίνακα, αυτό μπορεί να γραφεί από την άποψη της Ευκλείδειας ή της l_2 νόρμας, ως εξής

$$\hat{\beta} = arg\min_{\beta} \|y - X\beta\|_2^{\,2},$$

Υπάρχει και μια άλλη σημαντική ερμηνεία της $\hat{\beta}$. Η υπόθεσή μας για την κατανομή του ε συνεπάγεται ότι $y \sim N_n(X\beta, \sigma^2 I)$, στην οποία το N_n υποδηλώνει τώρα μια n-περιγραφική πολυ-περιγραφική Gaussian κατανομή, με $X\beta$ ως το διάνυσμα των μέσων και $\sigma^2 I$ τον πίνακα συνδιακύμανσης, και όπου I είναι ο $n \times n$ ταυτοτικός πίνακας. Η συνάρτηση πυκνότητας πιθανότητας για το y είναι τότε

$$p(y|X,\beta) = \left\{\sqrt{2\pi\sigma^2}\right\}^{-n/2} exp\left\{-\frac{\|y - X\beta\|_2{}^2}{2\sigma^2}\right\}.$$

Τα διαθέσιμα δεδομένα παρέχουν το y και το X. Όταν αντιμετωπίζονται ως συνάρτηση των παραμέτρων, ονομάζονται πλέον συνάρτηση πιθανοφάνειας και, σαφώς,

$$\hat{\beta} = arg\max_{\beta} p(y|X,\beta),$$

Έτσι, το $\hat{\beta}$ είναι ο λεγόμενος «εκτιμητής μέγιστης πιθανοφάνειας» του β. Η χρήση εκτιμητών μέγιστης πιθανοφάνειας είναι ένα πολύ κοινό παράδειγμα για στατιστικά συμπεράσματα.

Στο πρόβλημά μας,το $\hat{\beta}$ ικανοποιεί

$$X^T X\hat{\beta} = X^T y$$

και

$$\hat{\beta} = (X^T X)^{-1} X^T y,$$

ο ακριβής τύπος στη δεύτερη εξίσωση είναι διαθέσιμος υπό την προϋπόθεση ότι ο $X^T X$ μπορεί να αναστραφεί. (Εδώ X^T είναι ο ανάστροφος πίνακας του X.)

Επιπλέον, αν το μοντέλο είναι σωστό,

$$\hat{\beta} \sim N_p(\beta, \sigma^2 (X^T X)^{-1}),$$

από την οποία το διάστημα εμπιστοσύνης για το β μπορεί να ληφθεί, με ελαφρά τροποποίηση εάν, όπως είναι η συνήθης περίπτωση, το σ^2 πρέπει να εκτιμηθεί.

Σε γενικές γραμμές, για πολλά σενάρια μέγιστης πιθανοφάνειας σχετικά με τα παραμετρικά μοντέλα τα οποία εμπλέκουν ένα σταθερό αριθμό παραμέτρων β, για μεγάλο n, περίπου αν και σπάνια ακριβώς,

$$\hat{\beta} \sim N_p\left(\beta, \Sigma_{\hat{\beta}}\right),$$

για ένα συγκεκριμένο πίνακα $\Sigma_{\hat{\beta}}$. Έτσι, η $\hat{\beta}$ είναι ασυμπτωτικά αμερόληπτη, από το ότι κατά μέσο όρο δεν υπερεκτιμά ή υποεκτιμά το β, και κατανέμεται κανονικά.

(β) Μία μεγάλων διαστάσεων πραγματικότητα και τι πρέπει να κάνουμε γι 'αυτό

Για την παραπάνω κομψή και απλή ανάλυση, πρέπει να έχουμε $p \leq n$, διαφορετικά $(X^T X)$ είναι μη αντιστρέψιμος και οι παράμετροι στο μοντέλο παλινδρόμησης δεν μπορεί να εκτιμηθούν μοναδικά. Επιπρόσθετα, γενικά στο περιβάλλον της μέγιστης πιθανοφάνειας, ιδιαίτερα αν το p δεν έχει καθοριστεί, η ασυμπτωτική θεωρία καταρρέει. Τι θα συμβεί αν $p > n$ ή ακόμα και $\rho \gg n$ στο πρόβλημα παλινδρόμησης, για παράδειγμα το p είναι πολύ μεγαλύτερο από το n; μία προσέγγιση να αποφευχθεί η μη αντιστρεψιμότητα του $(X^T X)$ είναι να χρησιμοποιήσουμε μια μέθοδο της νομιμοποίησης (ή κανονικοποίησης), αλλιώς γνωστή ως *ποινικοποιημένα ελάχιστα τετράγωνα* ή *ποινικοποιημένη μέγιστη πιθανοφάνεια*. Ένα πρώτο παράδειγμα αυτής της μεθόδου είναι η παλινδρόμηση κορυφογραμμής (Hoerl και Kennard 1970), στην οποία εκτιμούμε το β με

$$\hat{\beta}_R = S_{\lambda_2} X^T y$$

Όπου $S_{\lambda_2} = (X^T X + \lambda_2 I)^{-1}$, και το θετικό βαθμωτό μέγεθος λ_2 ονομάζεται παράμετρος κορυφογραμμής ή κανονικοποιημένη σταθερά. Η κατανομή συχνοτήτων του β_R είναι

$$\hat{\beta}_R \sim N_p\left(S_{\lambda_2} X^T X\beta, \sigma^2 S_{\lambda_2}(X^T X) S_{\lambda_2}\right).$$

Έτσι ο εκτιμητής β_R είναι τώρα μεροληπτικός, αλλά μπορεί να υπολογιστεί. Καθώς το λ_2 αυξάνει, αυξάνει η μεροληψία, αλλά η «διακύμανση» μειώνεται, για να την αντισταθμίσει.

Υπάρχουν διάφορες ερμηνείες για το $\hat{\beta}_R$:

(i) $\widehat{\beta_R} = arg\min_\beta\{\|y - X\beta\|_2^{\,2} + \lambda_2\|\beta\|_2^{\,2}\}$

(ii) Το $\widehat{\beta_R}$ ελαχιστοποιεί $\|y - X\beta\|_2^{\,2}$ και υπόκειται στον περιορισμό $\|\beta\|_2^{\,2} \leq c_2(\lambda_2)$, για κάποια $c_2(\lambda_2)$ που εξαρτώνται από το λ_2, και

(iii)Το $\widehat{\beta_R}$ ελαχιστοποιεί $\|\beta\|_2^{\,2}$ και υπόκειται στον περιορισμό $\|y - X\beta\|_2^{\,2} \leq b_2(\lambda_2)$, για κάποια $b_2(\lambda_2)$ που εξαρτώνται από το λ_2.

Η πρώτη ερμηνεία δείχνει ότι το $\widehat{\beta_R}$ αντιστοιχεί σε αυτό που ονομάζεται l_2 κανονικοποίηση, διότι η συνάρτηση ποινής δίνεται από την l_2 ή από την τετραγωνική νόρμα. Στις άλλες ερμηνείες, το λ_2 ή το αντίστροφό του είναι ένας Lagrange πολλαπλασιαστής.

Ωστόσο, αν και αντιστρέψιμος, ο $X^T X + \lambda_2 I$ είναι $p \times p$ και ακόμα ένας ενδεχομένως πολύ μεγάλος πίνακας. Η κυρίαρχη στρατηγική σε τρέχουσες προσεγγίσεις σε αυτό το είδος της δυσκολίας είναι να προσπαθήσουμε να εκμεταλλευτούμε την αραιότητα, με άλλα λόγια, να αναζητήσουμε μια λύση για το β στο οποίο πολλά από τα στοιχεία είναι μηδέν. Μετά από όλα, εάν $n < \rho$, είναι εύλογο ότι διαισθητικά μόνο αραιές λύσεις μπορούν να ληφθούν «αξιόπιστα». Επιπλέον, στην πράξη, αν υπάρχει τεράστιος αριθμός εκτιμητών, είναι συχνά επιστημονικά ευλογοφανές ότι μόνο ένα μικρό ποσοστό είναι πιθανό να επιρεάζουν ως εκτιμητικοί παράγοντες. Με αυτό κατά νου, θα προσπαθήσουμε αλλάζοντας τη συνάρτηση ποινής, να εξετάσουμε τι καλείται l_0 κανονικοποίηση. Οι τρεις μας ισοδύναμοι τύποι είναι τώρα ως εξής:

(i) $\widehat{\beta_0} = arg\min_\beta\{\|y - X\beta\|_2^2 + \lambda_0\|\beta\|_0\}$, όπου $\|\beta\|_0$, ο αριθμός των μη-μηδενικών στοιχείων στο β, είναι η l_0 του β

(ii) Το $\widehat{\beta_0}$ ελαχιστοποιεί $\|y - X\beta\|_2^2$ και υπόκειται στον περιορισμό $\|\beta\|_0 \leq c_0(\lambda_0)$

(iii)Το $\widehat{\beta_0}$ ελαχιστοποιεί $\|\beta\|_0$ και υπόκειται στον περιορισμό $\|y - X\beta\|_2^2 \leq b_0(\lambda_0)$

Το πρόβλημα σχετικά με την εφαρμογή αυτής της προσέγγισης είναι ο συνδυαστικός χαρακτήρας του: δεν υπάρχει εναλλακτική λύση για την εξέταση, ξεχωριστά, κάθε ρύθμισης των μηδενικών και μη μηδενικών τιμών στο β, και αυτό οδηγεί σε μη αποδεκτή υπολογιστική πολυπλοκότητα και μια πιθανή εξάπλωση των τοπικών ακροτάτων.

Μια ενδιάμεση στρατηγική είναι να βασιστεί η συνάρτηση ποινής για τη νόρμα του l_1

$$\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|,$$

που οδηγεί στους ακόλουθους ισοδύναμους τύπους:

(i) $\widehat{\beta_L} = arg\min_\beta\{\|y - X\beta\|_2^2 + \lambda_1\|\beta\|_1\}$, *για κάποια* λ_1

(ii) Το $\widehat{\beta_L}$ ελαχιστοποιεί $\|y - X\beta\|_2^2$ και υπόκειται στον περιορισμό $\|\beta\|_1 \leq c_1(\lambda_1)$

(iii)Το $\widehat{\beta_L}$ ελαχιστοποιεί $\|\beta\|_1$ και υπόκειται στον περιορισμό $\|y - X\beta\|_2^2 \leq b_1(\lambda_1)$

Ο δείκτης L επιλέγεται διότι αυτή η μέθοδος ονομάζεται *Lasso* (Tibshirani, 1996). Η μέθοδος έχει τα διπλά πλεονεκτήματα ώστε να είναι υπολογιστικά εφικτή για την προσαρμογή του μοντέλου του τετραγωνικού προγραμματισμού, και γενικά οδηγεί σε αραιές λύσεις. Η τελευταία μπορεί να εξηγηθεί ανεπίσημα στο πλαίσιο του δεύτερου τύπου. Η λύση $\widehat{\beta_L}$ θα είναι στο σημείο της επαφής μιας «λείας» συνάρτησης σφάλματος αθροισμάτων τετραγώνων και μία κυρτή, τμηματικά-επίπεδη επιφάνεια περιορισμού. Το σημείο επαφής είναι πολύ πιθανό να είναι σε μία κορυφή της επιφάνειας περιορισμού και επομένως σε ένα σημείο όπου τα στοιχεία του β είναι μηδέν. Όπως και με την κανονικοποίηση της l_2, η Lasso έχει και μία Bayesian ερμηνεία.

Μερική ισχυρή υποστήριξη για την στρατηγική Lasso προέρχεται από την εξέταση χωρίς-θόρυβο εκδόσεων του προβλήματος: ελαχιστοποιώντας τη $\|\beta\|_0$ ή τη $\|\beta\|_1$ υπό τον περιορισμό $y = X\beta$. Αν η λύση στο πρόβλημα του l_0 είναι αρκετά αραιή − έχοντας το πολύ ένα αρκετά μικρό αριθμό k μη-μηδενικών καταχωρήσεων, ας πούμε − τότε οι λύσεις στο πρόβλημα l_0 και στο προβλημα l_1 συμπίπτουν (βλ. για παράδειγμα, Candes και Tao (2005) και Donoho (2006).

Οι κύριοι στόχοι οποιασδήποτε έρευνας της αραιότητας στο γραμμικό μοντέλο μπορεί να περιγραφούν ανεπίσημα ως εξής:

(i) Για τον ακριβή προσδιορισμό των συστατικών του β που δεν είναι μηδέν, δηλ., την υποστήριξη του β
(ii) Για να εκτιμηθούν αξιόπιστα οι πραγματικές τιμές των μη μηδενικών στοιχείων, υποθέτοντας ότι το μοντέλο είναι σωστό, και
(iii) Να κάνει αυτό «τόσο καλά» ώστε κάποιος θα μπορούσε να αναμένεται να κάνει αν η ταυτότητα των μη μηδενικών στοιχείων ήταν γνωστά από την αρχή, η λεγόμενη oracle ιδιότητα.

Κεφάλαιο 2:

Εξόρυξη δεδομένων και μηχανική μάθηση (Data mining and machine learning)

2.1 Εισαγωγή

Η ραγδαία ανάπτυξη των ηλεκτρονικών υπολογιστών και η εφαρμογή τους σε όλες τις ανθρώπινες δραστηριότητες, σε συνδυασμό με την ψηφιακή επεξεργασία και αποθήκευση δεδομένων, οδήγησε στη δημιουργία βάσεων δεδομένων πολύ μεγάλων διαστάσεων. Για την εξόρυξη των πληροφοριών που υποβόσκουν σε αυτές τις βάσεις απαιτείται η επεξεργασία τους, όμως οι έως τώρα αναλύσεις των δεδομένων, αφήνουν ακόμη μεγάλα ποσά πληροφορίας «κρυμμένα». Υπό αυτές τις συνθήκες γεννήθηκε η ανάγκη για νέες τεχνικές και εργαλεία που θα μετατρέπουν έξυπνα και αυτόματα τα δεδομένα σε χρήσιμες πληροφορίες και γνώση. Ο επιστημονικός κλάδος που ασχολείται με το αντικείμενο αυτό είναι γνωστός με τον όρο Data Mining (εξόρυξης πληροφορίας από δεδομένα). Παρόλο που είναι δύσκολο να καθοριστούν με ακρίβεια το εύρος και τα όρια μελέτης αυτού του κλάδου, παραβλέπουμε τις λεπτομέρειες και δεχόμαστε ως ορισμό του data mining τον παρακάτω:

«Data mining είναι η ανάλυση συχνά μεγάλων παρατηρούμενων συνόλων δεδομένων με σκοπό να βρούμε σχέσεις που δεν υποψιαζόμαστε και να συνοψίσουμε τα δεδομένα με καινοτόμους τρόπους, κατανοητούς και χρήσιμους για τον κάτοχο των δεδομένων».

Ο παραπάνω ορισμός αναφέρεται σε "παρατηρούμενα " αντί για "πειραματικά δεδομένα". Ο λόγος είναι ότι το data mining τυπικά ασχολείται με δεδομένα που έχουν συλλεχθεί για κάποιον άλλο σκοπό εκτός της data mining ανάλυσης, δηλαδή οι αντικειμενικοί στόχοι του data mining δεν παίζουν κανένα ρόλο στην στρατηγική που ακολουθείται για τη συλλογή

των δεδομένων. Αυτό είναι ένα σημείο στο οποίο το data mining διαφοροποιείται από τις συνηθισμένες στατιστικές αναλύσεις οι οποίες συχνά συλλέγουν δεδομένα χρησιμοποιώντας αποτελεσματικές στρατηγικές για να απαντήσουν σε συγκεκριμένα ερωτήματα, γι' αυτό και συχνά το data mining αναφέρεται ως δευτερεύουσα ανάλυση δεδομένων. Επίσης οι κλασικές στατιστικές αναλύσεις χρησιμοποιούν μικρά σύνολα δεδομένων, σε αντίθεση με το data mining που χρησιμοποιεί μεγάλα σύνολα από τα οποία όμως προκύπτουν νέα προβλήματα, όπως η διαχείριση τους, η αποθήκευση τους, η πρόσβαση τους και άλλα. Οι σχέσεις που προκύπτουν από το data mining συχνά αναφέρονται ως μοντέλα ή πρότυπα (*patterns*) και περιλαμβάνουν γραμμικές εξισώσεις, κανόνες, συστάδες (*clusters*), γραφήματα, δέντρα και επαναλαμβανόμενα πρότυπα σε χρονοσειρές.

Η ιδέα στην οποία στηρίζεται το data mining είναι η κατασκευή υπολογιστικών προγραμμάτων που χρησιμοποιούν στατιστικά αποτελέσματα για το κρισάρισμα των βάσεων δεδομένων και την εξαγωγή προτύπων και άλλων πληροφοριών. Φυσικά σε μια τέτοια διαδικασία επιλογής προτύπων, υπάρχουν και πολλά προβλήματα, όπως το να προκύψουν πρότυπα χωρίς ενδιαφέρον, ξεπερασμένα, πολύπλοκα ή να είναι αποτέλεσμα συμπτώσεων της συγκεκριμένης βάσης δεδομένων. Επιπλέον τα πραγματικά δεδομένα μπορεί να είναι διαστρεβλωμένα ή ελλιπή με αποτέλεσμα να μην προκύπτουν ακριβή συμπεράσματα. Μπορεί δηλαδή να προκύπτουν εξαιρέσεις σε κάθε κανόνα καθώς και περιπτώσεις οι οποίες δεν καλύπτονται από κανένα κανόνα. Για αυτό οι αλγόριθμοι θα πρέπει να είναι αρκετά ανθεκτικοί για να ανταποκρίνονται σε μη τέλεια δεδομένα και να μπορούν να εξάγουν και κανόνες μη ακριβείς μεν, χρήσιμους δε. Η εύρεση ισχυρών προτύπων, αν υπάρχουν, είναι ένα πολύ χρήσιμο εργαλείο για την ακριβή πρόβλεψη μελλοντικών δεδομένων, για τη γενίκευση από ένα δείγμα του συνόλου στο πλήρες σύνολο και για τη συμπίεση μεγάλων δεδομένων σε μικρότερα, με σκοπό να γίνουν πιο κατανοητά και πιο χρήσιμα. Στη βιβλιογραφία το data mining συνίσταται και με τον όρο knowledge discovery in databases (KDD – ανακάλυψη γνώσης από βάσεις δεδομένων), όρο δανεισμένο από την τεχνητή νοημοσύνη, χωρίς όμως να διαφοροποιείται το πεδίο μελέτης του.

Περισσότερες από τις τεχνικές για την εύρεση και την περιγραφή δομικών σχεδίων στα δεδομένα έχουν αναπτυχθεί διαμέσου ενός πεδίου γνωστό ως μηχανική μάθηση (machine learning). Επομένως θα μπορούσαμε να πούμε ότι η μηχανική μάθηση είναι η τεχνική βάση του data mining και χρησιμοποιείται για την εξόρυξη πληροφορίας από ακατέργαστα δεδομένα σε μεγάλες βάσεις δεδομένων.

Διαδικασία KDD

Επεξεργαζόμενοι μια τεράστια βάση δεδομένων είναι πιθανό να ανακαλύψουμε την ύπαρξη «κρυμμένης γνώσης». Δηλαδή, μπορεί να εντοπίσουμε συσχετίσεις, αλληλεξαρτήσεις ή

ομαδοποιήσεις μεταξύ των δεδομένων, πράγματα τα οποία να μην είναι άμεσα εμφανή. Το είδος αυτής της «γνώσης» θεωρείται ότι δεν είναι εκ των προτέρων διαθέσιμο αλλά μπορεί να αποδειχθεί πολύ χρήσιμο.

Την ανάγκη αυτή ανάκτησης γνώσης έρχεται να καλύψει η εξόρυξη δεδομένων (data mining), η οποία αποτελεί τον πυρήνα της γενικότερης μεθοδολογίας της ανακάλυψης της γνώσης από βάσεις δεδομένων (Knowledge Discovery in Databases - KDD).

Η KDD είναι μια αυτοματοποιημένη διαδικασία ανάλυσης και μοντελοποίησης τεράστιων αποθηκών δεδομένων. Πρόκειται για μια συγκροτημένη μεθοδολογία αναγνώρισης έγκυρων και πρωτότυπων προτύπων μέσα από πολύ μεγάλους και περίπλοκους πίνακες δεδομένων, με στόχο τα πρότυπα που θα προκύψουν να είναι χρήσιμα και κατανοητά.

Τα βασικά βήματα της KDD διαδικασίας είναι τα ακόλουθα (Σχήμα 4):

- **Ανάπτυξη και κατανόηση του πεδίου της εφαρμογής** συμπεριλαμβανόμενης οποιασδήποτε σχετικής προηγούμενης γνώσης για το πρόβλημα καθώς επίσης και των στόχων / προσδοκιών των τελικών χρηστών.

- **Δημιουργία του στοχευόμενου συνόλου δεδομένων (target data)**, το οποίο θα περιλαμβάνει τα δεδομένα από τα οποία πρόκειται να εξαχθεί η γνώση. Το βήμα αυτό είναι εξαιρετικά κρίσιμο καθώς η ποιότητα των δεδομένων επηρεάζει την απόδοση του συστήματος αποκάλυψης γνώσης.

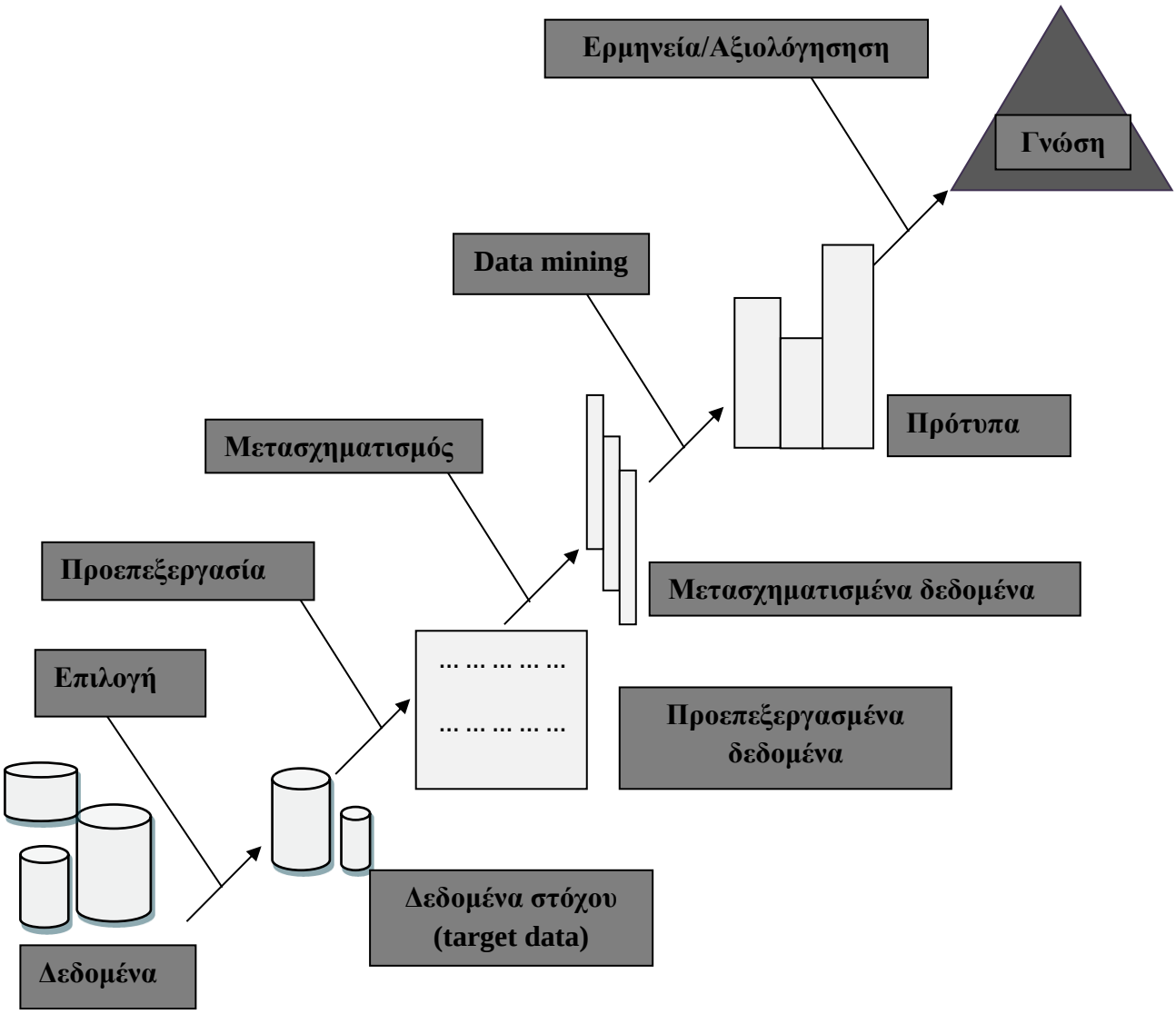
- **Καθαρισμός και επεξεργασία δεδομένων (data cleaning).** Το βήμα αυτό περιλαμβάνει βασικές λειτουγίες όπως η απομάκρυνση του θορύβου, η αντιμετώπιση του προβλήματος των δεδομένων με ελλιπείς τιμές κ.ά.

- **Μείωση της ποσότητας των δεδομένων (data reduction).** Το βήμα αυτό περιλαμβάνει την εύρεση χρήσιμων χαρακτηριστικών για την αναπαράσταση των δεδομένων του προβλήματος ανάλογα με τους στόχους της ανακάλυψης γνώσης , τη μείωση του πλήθους αυτών των χαρακτηριστικών κ.ά.

- **Επιλογή των εργασιών εξόρυξης γνώσης (data mining)** που θα χρησιμοποιηθούν για τις ανάγκες του προβλήματος π.χ ταξινόμηση , πρόβλεψη , ομαδοποίηση κ.α.

- **Επιλογή των αλγορίθμων εξόρυξης γνώσης (data mining)** που θα χρησιμοποιηθούν για την αναζήτηση προτύπων στα δεδομένα. Το βήμα αυτό περιλαμβάνει την επιλογή του κατάλληλου μοντέλου, την επιλογή των κατάλληλων παραμέτρων του μοντέλου κ.ά.

- **Data Mining**: αναζήτηση στα δεδομένα των προτύπων που μας ενδιαφέρουν.
- **Ερμηνεία των προτύπων** που ανακαλύφθηκαν από την KDD διαδικασία – πιθανόν να χρειαστεί να επιστρέψουμε και πάλι σε κάποια από τα παραπάνω βήματα.
- **Ενοποίηση της γνώσης που έχει εξαχθεί**: Σε αυτό το βήμα, η εξορυγμένη γνώση ενσωματώνεται στο σύστημα και χρησιμοποιούνται κάποιες τεχνικές αντιπροσώπευσης αυτής προκειμένου να παρουσιαστεί ευκρινώς στο χρήστη.



Σχήμα 4: Διαδικασία KDD

2.2 Μηχανική Μάθηση και Στατιστική

Πολλοί αντιμετωπίζουν στρεβλά την έκρηξη του εμπορικού ενδιαφέροντος (και των διαφημιστικών εκστρατειών) σε αυτόν τον τομέα, εξισώνοντας την εξόρυξη δεδομένων με τη στατιστική καθώς και την εμπορία. Στην πραγματικότητα, δεν θα πρέπει να κοιτάξουμε για μια διαχωριστική γραμμή μεταξύ της μηχανικής μάθησης και της στατιστικής γιατί υπάρχει μια συνέχεια - και μάλιστα πολυδιάστατη – των τεχνικών ανάλυσης δεδομένων. Μερικές τεχνικές προέρχονται από τις δεξιότητες που διδάσκονται στα βασικά μαθήματα στατιστικής και άλλες είναι πιο στενά συνδεδεμένες με το είδος της μηχανικής μάθησης που έχει προκύψει από την επιστήμη των υπολογιστών. Ιστορικά, οι δύο πλευρές είχαν μάλλον διαφορετικές παραδόσεις. Αν αναγκαστούμε να επισημάνουμε μία απλή διαφορά όπου δίνουν έμφαση, θα μπορούσε να είναι ότι, η στατιστική έχει μεγαλύτερη σχέση με τον έλεγχο υποθέσεων, ενώ η μηχανική μάθηση ασχολείται περισσότερο με τη διαμόρφωση της διαδικασίας της γενίκευσης ως αναζήτησης μέσω μιας πιθανής υπόθεσης. Αλλά αυτό είναι μια πρόχειρη υπεραπλούστευση: η στατιστική είναι πολλά περισσότερα από έναν έλεγχο υποθέσεων, καθώς και πολλές τεχνικές μηχανικής μάθησης δεν συνεπάγονται την αναζήτηση σε όλα.

Στο παρελθόν, πολλές κοινές μέθοδοι έχουν αναπτυχθεί παράλληλα στη στατιστική και τη μηχανική μάθηση. Μία από αυτές είναι τα δέντρα αποφάσεων. Τέσσερις στατιστικολόγοι (Breiman et al., 1984) δημοσίευσαν ένα βιβλίο σχετικά με την ταξινόμηση και τα δέντρα παλινδρόμησης στα μέσα της δεκαετίας του 1980. Παράλληλα, καθ' όλη τη δεκαετία του 1970 και στις αρχές του 1980 ένας εξέχων ερευνητής της μηχανικής μάθησης, J.Ross Quinlan, ανέπτυξε ένα σύστημα για την συναγωγή δέντρων ταξινόμησης από παραδείγματα. Είναι αξιοσημείωτο ότι αυτά τα δύο ανεξάρτητα έργα παρήγαγαν αρκετά παρόμοιες μεθόδους για την δημιουργία δέντρων από παραδείγματα, και οι ερευνητές έλαβαν γνώση ο ένας για την εργασία του άλλου πολύ αργότερα. Ένας άλλος τομέας στον οποίο παρόμοιες μέθοδοι έχουν προκύψει περιλαμβάνει τη χρήση της μεθόδου των πλησιέστερων - γειτόνων για την ταξινόμηση. Αυτές είναι συνήθεις στατιστικές τεχνικές που έχουν προσαρμοστεί σε μεγάλο βαθμό από τους ερευνητές μηχανικής μάθησης, τόσο για τη βελτίωση των επιδόσεων της ταξινόμησης όσο και για να καταστεί η διαδικασία πιο αποτελεσματική υπολογιστικά.

Τώρα, όμως, οι δύο προοπτικές συγκλίνουν. Πολλές τεχνικές της μηχανικής μάθησης ενσωματώνουν μεγάλη στατιστική σκέψη. Από το ξεκίνημα της διαδικασίας, κατά την κατασκευή και τελειοποίηση του αρχικού σετ του παραδείγματος, εφαρμόζονται οι συνήθεις στατιστικές μέθοδοι: οπτικοποίηση δεδομένων, επιλογή χαρακτηριστικών, απόρριψη των ακραίων τιμών, και ούτω καθεξής. Οι περισσότεροι αλγόριθμοι μάθησης χρησιμοποιούν στατιστικές δοκιμές κατά την κατασκευή κανόνων ή δέντρων και για τη διόρθωση μοντέλων που είναι "overfitted" υπό την έννοια ότι εξαρτώνται σε πολύ μεγάλο βαθμό από τις

λεπτομέρειες των συγκεκριμένων παραδειγμάτων που χρησιμοποιούνται για την παραγωγή τους. Οι στατιστικές δοκιμές χρησιμοποιούνται για την επικύρωση των μοντέλων μηχανικής μάθησης και για την αξιολόγηση των αλγορίθμων μηχανικής μάθησης.

Για να είναι τα αποτελέσματα της εφαρμογής του data mining σε πρακτικά προβλήματα ασφαλή, δε θα πρέπει να στηρίζονται μόνο στην εφαρμογή των αλγορίθμων του σε υποδείγματα, δηλαδή στη μηχανική μάθηση (*machine learning*) μέσω υποδειγμάτων, αλλά να συνδυάζονται και με τη στατιστική ανάλυση. Μπορούμε επομένως να πούμε ότι η στατιστική ανάλυση και οι αλγόριθμοι εξόρυξης πληροφορίας από βάσεις δεδομένων αποτελούν τα δύο βασικά συστατικά του data mining για την ανάλυση δεδομένων πρακτικών προβλημάτων.

Η στατιστική ανάλυση έχει χρησιμοποιηθεί και στο παρελθόν, από το 19^{ο} αιώνα, για την ερμηνεία πραγματικών δεδομένων. Μεγάλη ανάπτυξη έλαβε μετά το 2^{ο} Παγκόσμιο πόλεμο και σήμερα κατέχει κυρίαρχη θέση. Το πλεονέκτημα της στατιστικής ανάλυσης είναι η ικανότητα της να ερμηνεύει τα δεδομένα με μαθηματικό τρόπο ενώ το μειονέκτημά της εντοπίζεται στη δυσκολία της να γενικεύσει σε πολύ μεγάλα σύνολα δεδομένων.

Αλγόριθμοι και data mining

Οι αλγόριθμοι για την ανάλυση δεδομένων έχουν μελετηθεί από στατιστικούς και έχουν χρησιμοποιηθεί σε ποικίλους κλάδους εδώ και πολλούς αιώνες, όμως νέοι αλγόριθμοι χρειάζεται να σχεδιαστούν για να διευθετήσουν τους περιορισμούς των υπαρχουσών τεχνικών που προκύπτουν από τους νέους τύπους δεδομένων που συλλέγονται. Η πρόσφατη πρόοδος στην τεχνολογία πληροφοριών έκανε ικανή τη συγκέντρωση τεράστιων ποσών δεδομένων στο εμπόριο και σε διάφορους επιστημονικούς κλάδους. Πολλά από αυτά τα σύνολα δεδομένων είναι υψηλών διαστάσεων, ετερογενή, διασκορπισμένα ή χώρου– χρόνου και οι παραδοσιακές τεχνικές δε μπορούν να εφαρμοστούν σε αυτά.

Το αναδυόμενο πεδίο του data mining διορθώνει τους περιορισμούς υπαρχουσών τεχνικών ανάλυσης δεδομένων απευθυνόμενο σε αυτούς τους νέους τύπους δεδομένων. Μέσα σε 10 χρόνια το πεδίο του data mining εξελίχθηκε ραγδαία και συνεχίζει να παράγει μεγάλο αριθμό αλγορίθμων που απευθύνονται σε περαιτέρω περιορισμούς.

Υπάρχουν πολλές προκλήσεις και απαιτήσεις που θα πρέπει να αντιμετωπίσει ο αναλυτής για την ανάπτυξη αποτελεσματικών αλγορίθμων. Μία από αυτές τις απαιτήσεις είναι η τεράστια ποικιλία των τύπων των δεδομένων και των στόχων του data mining, η οποία καθιστά αδύνατη την ύπαρξη ενός μοναδικού και ταυτόχρονα αποτελεσματικού συστήματος data mining και η οποία επιβάλλει την κατασκευή ιδιαίτερων αλγορίθμων για συγκεκριμένους τύπους δεδομένων. Μία άλλη απαίτηση που θα πρέπει να ικανοποιούν οι

αλγόριθμοι είναι η αποτελεσματικότητα τους και η δυνατότητά τους να κλιμακώνονται σε μεγάλες βάσεις δεδομένων σε αποδεκτό και αναμενόμενο χρόνο. Επιπλέον, οι αλγόριθμοι θα πρέπει να διαχειρίζονται σωστά το θόρυβο και τα δεδομένα που αποτελούν εξαιρέσεις, να καταλήγουν σε αποτελέσματα που απεικονίζουν με ακρίβεια τα περιεχόμενα της βάσης δεδομένων και να παρέχουν τη δυνατότητα στον αναλυτή να εξετάσει την ανακαλυφθείσα γνώση από διάφορες οπτικές γωνίες και διαφορετικές μορφές, ενθαρρύνοντας την αλληλεπίδρασή τους. Τέλος, βασική απαίτηση που πρέπει να ικανοποιούν οι αλγόριθμοι του data mining είναι ότι η μη τελειότητα που παρουσιάζουν τα αποτελέσματά τους θα πρέπει να μπορεί να εκφραστεί με μέτρα αβεβαιότητας σε μορφή προσεγγιστικών ή ποσοτικών κανόνων, έτσι οδηγούμαστε στη μελέτη της μέτρησης της ποσότητας της ανακαλυφθείσας γνώσης καθώς και του ενδιαφέροντος που παρουσιάζει και της αξιοπιστίας της, κατασκευάζοντας στατιστικά μοντέλα και εργαλεία.

Στις περισσότερες εφαρμογές του data mining σε πρακτικά προβλήματα, οι αλγόριθμοι ανάλυσης προτύπων συσχέτισης που εφαρμόζονται παράγουν ένα μεγάλο σύνολο προτύπων και ο καθορισμός των χρήσιμων προτύπων απαιτεί στενή αλληλεπίδραση ανάμεσα στους σχεδιαστές των αλγορίθμων και στους ειδικούς των εφαρμογών καθώς και η βαθιά γνώση του χώρου αποτελεί το κλειδί για να αναγνωρίσουμε και να εκτιμήσουμε τα χρήσιμα χαρακτηριστικά και τα πρότυπα που έχουν νόημα. Οι αλγόριθμοι του data mining από τη φύση τους, οδηγούνται από τα δεδομένα. Οι υποθέσεις που παράγονται από αυτούς τους αλγορίθμους πρέπει να εκτιμηθούν από γερές στατιστικές μεθοδολογίες για να χρησιμοποιηθούν στην πράξη.

Η κύρια κινητήρια δύναμη πίσω από την ανάπτυξη των στατιστικών τεχνικών και των αλγορίθμων του data mining είναι τα προβλήματα ανάλυσης δεδομένων που χρειάζεται να επιλυθούν. Οι χώροι από τους οποίους αναδύονται τέτοια προβλήματα είναι τόσο ποικίλοι, όσες και οι προσεγγίσεις των λύσεων. Τα προβλήματα που διευθετούνται μπορεί να έχουν πλατιά επίδραση και να επηρεάζουν την καθημερινή ζωή πολλών ανθρώπων, όπως για παράδειγμα στην περίπτωση της σωστής αναγνώρισης σε μία δόλια συναλλαγή μίας πιστωτικής κάρτας ή στη επιστροφή σχετικών αποτελεσμάτων σε μία μηχανή αναζήτησης. Αυτά είναι μερικά παραδείγματα όπου η ανάλυση δεδομένων γίνεται ακέραιο κομμάτι της ζωής μας.

2.3 Στατιστική μάθηση (Statistical learning) τα είδη μάθησης

Ο όρος αντίληψη αναφέρεται στο αντικείμενο της μάθησης και στόχος του data mining είναι η εύρεση κατανοητής και λειτουργικής περιγραφής μιας αντίληψης με εφαρμογή κατάλληλων και αποτελεσματικών αλγορίθμων. Η περιγραφή της ζητούμενης αντίληψης,

δηλαδή η απεικόνιση, μπορεί να γίνει με διάφορους τρόπους, ανάλογα με τις ανάγκες του προβλήματός μας.

Η στατιστική μάθηση διαδραματίζει σημαντικό ρόλο σε πολλούς τομείς της επιστήμης, των οικονομικών και της βιομηχανίας. Ακολουθούν μερικά παραδείγματα των μαθησιακών προβλημάτων:

- Πρόβλεψη εάν ένας ασθενής, στο νοσοκομείο λόγω καρδιακής προσβολής, θα έχει μια δεύτερη καρδιακή προσβολή. Η πρόβλεψη πρόκειται να βασιστεί σε δημογραφικά χαρακτηριστικά, διατροφικά χαρακτηριστικά και κλινικές μετρήσεις για αυτόν τον ασθενή.

- Πρόβλεψη της τιμής μιας μετοχής σε 6 μήνες από τώρα, βάσει των μέτρων απόδοσης της εταιρείας και των οικονομικών δεδομένων.

- Προσδιορισμός των αριθμών σ' ένα χειρόγραφο κώδικα, από μια ψηφιακή εικόνα.

- Υπολογισμός της ποσότητας της γλυκόζης στο αίμα ενός διαβητικού ατόμου, από το υπέρυθρο φάσμα απορρόφησης του αίματος του ατόμου.

- Προσδιορισμός των παραγόντων κινδύνου για καρκίνο του προστάτη, με βάση κλινικές και δημογραφικές μεταβλητές.

Η επιστήμη της μάθησης διαδραματίζει καίριο ρόλο στους τομείς των στατιστικών δεδομένων, της εξόρυξης και της τεχνητής νοημοσύνης και διασταυρώνεται με τους τομείς της μηχανικής και άλλων κλάδων.

Σ' ένα τυπικό σενάριο, έχουμε ένα αποτέλεσμα μέτρησης (outcome), συνήθως ποσοτικό (όπως τιμή της μετοχής) ή κατηγορηματικό (όπως καρδιακή προσβολή / χωρίς καρδιακή προσβολή), που επιθυμούμε να προβλέψουμε βασιζόμενοι σε ένα σύνολο χαρακτηριστικών (όπως η δίαιτα και κλινικές μετρήσεις). Έχουμε ένα σύνολο δεδομένων εκπαίδευσης (training set), στα οποία παρατηρούμε τις μετρήσεις - αποτέλεσμα και χαρακτηριστικά - για ένα σύνολο αντικειμένων (όπως οι άνθρωποι). Χρησιμοποιώντας αυτά τα δεδομένα χτίζουμε ένα μοντέλο πρόβλεψης (ή learner) το οποίο θα μας δώσει τη δυνατότητα να προβλέψουμε το αποτέλεσμα για νέα αντικείμενα. Ένας καλός "μαθητής" (learner) είναι αυτός που προβλέπει με ακρίβεια, ένα τέτοιο αποτέλεσμα.

Σε ένα τυπικό πρόβλημα του data mining έχουμε, λοιπόν, ένα σύνολο δεδομένων εκπαίδευσης *(training set)* στο οποίο γνωρίζουμε την τιμή του αποτελέσματος και τις τιμές των χαρακτηριστικών που μας ενδιαφέρουν, και προσπαθούμε με βάση αυτά τα δεδομένα να κατασκευάσουμε ένα μοντέλο πρόβλεψης. Το μοντέλο αυτό θα το χρησιμοποιήσουμε στη συνέχεια για να προβλέψουμε το αποτέλεσμα νέων συνόλων δεδομένων εξέτασης *(test set),* στα οποία σύνολα είναι γνωστές οι τιμές των χαρακτηριστικών αλλά δεν είναι γνωστή η τιμή του αποτελέσματος, δηλαδή η τιμή της τάξης.

Τα παραπάνω παραδείγματα περιγράφουν αυτό που ονομάζεται πρόβλημα εποπτευόμενης μάθησης.

Στο σημείο αυτό μπορούμε, λοιπόν, να διακρίνουμε δύο μορφές μάθησης:

- Τη *μάθηση με επίβλεψη* ή εποπτευόμενη μάθηση (supervised learning). Στη μάθηση με επίβλεψη τα δεδομένα εκπαίδευσης συνοδεύονται από ετικέτες για την κλάση στην οποία ανήκει το καθένα, δηλαδή μπορούμε να πούμε ότι η διαδικασία μάθησης οδηγείται από την παρουσία των αποτελεσμάτων της κλάσης. Ονομάζεται "εποπτευόμενη" λόγω της παρουσίας της μεταβλητής έκβασης που καθοδηγεί τη διαδικασία μάθησης. Μερικές από αυτές τις supervised τεχνικές είναι και τα νευρωνικά δίκτυα (neural networks), τα δέντρα αποφάσεων (decision trees), η λογιστική παλινδρόμηση (logistic regression), οι μηχανές διανυσματικής υποστήριξης (support vector machines).

- Τη *μάθηση χωρίς επίβλεψη* ή μη εποπτευόμενη μάθηση (unsupervised learning). Στη μάθηση χωρίς επίβλεψη δεν είναι γνωστό σε ποια κλάση ανήκουν τα δεδομένα εκπαίδευσης, δηλαδή γνωρίζουμε μόνο τις τιμές των χαρακτηριστικών και όχι την τιμή του αποτελέσματος. Στην μάθηση χωρίς επίβλεψη, παρατηρούμε μόνο τα χαρακτηριστικά και δεν έχουμε μετρήσεις από το αποτέλεσμα. Το καθήκον μας είναι περισσότερο να περιγράψουμε πώς τα δεδομένα οργανώνονται ή ομαδοποιούνται.

Τα περισσότερα προβλήματα που συναντάμε στην πράξη ανήκουν στην κατηγορία της μάθησης με επίβλεψη.

Η γνώση που προκύπτει από μία διαδικασία εξόρυξης πληροφοριών από δεδομένα, μπορεί να κατηγοριοποιηθεί με διάφορους τρόπους, ανάλογα με το στόχο του προβλήματος που εξετάζουμε. Οι τρόποι αυτοί θα πρέπει να είναι κατανοητοί, συνοπτικοί και εύχρηστοι. Έτσι, τα κύρια είδη μάθησης που διακρίνουμε είναι τα εξής:

- *Ταξινόμηση* (Classification): είναι η ταξινόμηση των υποδειγμάτων σε μία προκαθορισμένη τάξη (class).
- *Συσχέτιση* (Association): είναι η ανακάλυψη συσχετίσεων μεταξύ των διαφόρων χαρακτηριστικών του συνόλου δεδομένων.
- *Συστηματοποίηση ή Ομαδοποίηση δεδομένων* (Clustering): είναι η εύρεση ομάδων αντικειμένων με υψηλό βαθμό ομοιότητας και εκχώρηση υποδειγμάτων στις ομάδες αυτές.
- *Αριθμητική πρόβλεψη (Prediction)*: είναι η πρόβλεψη μίας αριθμητικής ποσότητας και είναι όμοια με την ταξινόμηση μόνο που εδώ η τάξη είναι αριθμητική.

Άλλα δύο είδη μάθησης που συναντάμε συχνά είναι:

- Γενικεύσεις δεδομένων και συνόψεις (data generalization and summarization tools) : παρουσιάζουν τα γενικά χαρακτηριστικά ή μια υψηλού επιπέδου περιληπτική άποψη ενός συνόλου δεδομένων ειδικής χρήσης από μία βάση δεδομένων.
- Εύρεση προτύπων βασισμένων στην ομοιότητα (Pattern based similarity search) : εύρεση όμοιων προτύπων σε χρονικά ή χωρο – χρονικά σύνολα δεδομένων.

Στην παρούσα διπλωματική εργασία θα ασχοληθούμε εκτενέστερα με την ταξινόμηση των υποδειγμάτων σε μία προκαθορισμένη τάξη (class).

2.4 Οι τύποι των δεδομένων

Ένα σύνολο δεδομένων (data set) ή μία βάση δεδομένων, όπως διαφορετικά λέγεται, είναι ένα σύνολο μετρήσεων που συλλέγουμε κατά την παρατήρηση ενός περιβάλλοντος ή μιας διαδικασίας. Στην πιο απλή περίπτωση, έχουμε μία συλλογή από n αντικείμενα και για κάθε αντικείμενο έχουμε ένα σύνολο των ίδιων p μετρήσεων. Σε αυτή την περίπτωση μπορούμε να απεικονίσουμε τις μετρήσεις σε ένα πίνακα $n \times p$. Για παράδειγμα τα αντικείμενα μπορεί να είναι ασθενείς, πελάτες πιστωτικών καρτών ή άλλα μεμονωμένα αντικείμενα όπως αστέρια και γαλαξίες.

Οι γραμμές του πίνακα αποτελούν την είσοδο (*input*) των αλγορίθμων που εφαρμόζουμε, ονομάζονται υποδείγματα (examples, instances) και είναι ανεξάρτητες μεταξύ τους. Στη βιβλιογραφία συναντάμε και εναλλακτικούς ορισμούς των υποδειγμάτων, όπως εγγραφές (records), αντικείμενα (objects), περιπτώσεις (cases), οντότητες (entities), ή άτομα (individuals), ανάλογα με την ορολογία του προβλήματος που έχουμε να εξετάσουμε.

Η άλλη διάσταση του πίνακα αποτελεί τις p μετρήσεις που καταγράφουμε για κάθε αντικείμενο και θεωρούμε ότι οι μετρήσεις αυτές είναι οι ίδιες για κάθε αντικείμενο, παρόλο

που μπορεί να μην συμβαίνει αυτό, όπως για παράδειγμα στην περίπτωση όπου διαφορετικοί ιατρικοί έλεγχοι εφαρμόζονται σε διαφορετικούς ασθενείς. Οι p στήλες του πίνακα δεδομένων αναφέρονται ως μεταβλητές (variables), χαρακτηριστικά (features, attributes) ή πεδία (fields), ανάλογα και εδώ με το πεδίο έρευνας και δεν είναι πάντα ανεξάρτητες μεταξύ τους καθώς υπάρχουν περιπτώσεις όπου η τιμή ενός χαρακτηριστικού εξαρτάται από την τιμή ενός άλλου χαρακτηριστικού.

Η τιμή ενός χαρακτηριστικού είναι η μέτρηση της ποσότητας στην οποία το χαρακτηριστικό αναφέρεται και μπορεί να είναι ονομαστική (ή ποιοτική) ή αριθμητική (ή ποσοτική). Τα αριθμητικά χαρακτηριστικά ονομάζονται και συνεχή και μπορεί να είναι πραγματικός ή ακέραιος αριθμός. Να σημειώσουμε ότι ο όρος συνεχής χρησιμοποιείται καταχρηστικά αφού τα χαρακτηριστικά με ακέραια τιμή δεν είναι «συνεχή» με την αυστηρά μαθηματική έννοια. Τα ονομαστικά χαρακτηριστικά είναι διακριτά σύμβολα που αποτελούνται από ένα περιγραφικό όνομα και οι τιμές που παίρνουν είναι από ένα προκαθορισμένο σύνολο πιθανών τιμών. Μεταξύ των τιμών αυτών δεν συνεπάγεται καμία σχέση διάταξης ή απόστασης και κατά συνέπεια, δεν έχει νόημα καμία μαθηματική πράξη μεταξύ αυτών παρά μόνο η σύγκριση ως προς την ισότητα της τιμής του χαρακτηριστικού μεταξύ των υποδειγμάτων. Στη βιβλιογραφία συναντάμε και άλλη ορολογία για τα ονομαστικά χαρακτηριστικά, όπως ρητά ή κατηγορικά (categorical), ή απαριθμημένα (enumerated) ή διακριτά (discrete). Ο όρος απαριθμημένα χρησιμοποιείται κατά κύριο λόγο στην πληροφορική για να δηλώσει ένα ρητό τύπο δεδομένων όμως ο ακριβής ορισμός προϋποθέτει διάταξη.

Άλλοι τύποι χαρακτηριστικών είναι τα τακτικά (ordinal), τα περιοδικά (interval) και τα αναλογικά (ratio). Στα τακτικά χαρακτηριστικά ορίζεται η έννοια της διάταξης μεταξύ των διαφόρων τιμών, όμως δεν ορίζεται η έννοια της απόστασης μεταξύ τους και άρα δε μπορούν να εκτελεστούν αριθμητικές πράξεις. Τα τακτικά χαρακτηριστικά συχνά αναφέρονται και ως αριθμητικά ή συνεχή χωρίς όμως να υπαινίσσεται η έννοια της συνέχειας με μαθηματικό τρόπο. Για παράδειγμα όταν ένα παρατηρούμενο χαρακτηριστικό ενός συνόλου δεδομένων είναι η «θερμοκρασία», τότε μπορεί να έχουμε τις παρακάτω πιθανές τιμές « ζέστη – ήπιο - κρύο». Σε αυτές τις τιμές είναι προφανής η διάταξη «ζέστη» > «ήπιο» > «κρύο», όμως δεν έχει νόημα η πρόσθεση ή η αφαίρεσή τους. Η διάκριση ανάμεσα στα ονομαστικά και στα τακτικά χαρακτηριστικά δεν είναι πάντα ευκρινής. Τα περιοδικά χαρακτηριστικά έχουν διατεταγμένες αλλά και μετρήσιμες σε σταθερές και ισαπέχουσες μονάδες. Ως παράδειγμα αναφέρουμε το χαρακτηριστικό «θερμοκρασία» όπου οι τιμές του εκφράζονται τώρα σε βαθμούς Fahrenheit και όχι με ονομαστικό τρόπο. Στις περιπτώσεις αυτές έχει νόημα να υπολογίσουμε τη διαφορά μεταξύ δύο τιμών και να τη συγκρίνουμε με τη διαφορά άλλων τιμών, όμως δεν έχει νόημα το άθροισμα ή το γινόμενο τιμών, καθώς δεν ορίζεται το σημείο μηδέν, δηλαδή το σημείο αναφοράς. Στα αναλογικά χαρακτηριστικά ορίζεται το σημείο μηδέν και τα χρησιμοποιούμε ως πραγματικούς αριθμούς όπου όλες οι

μαθηματικές πράξεις έχουν νόημα. Ωστόσο, ο ορισμός του σημείου μηδέν είναι συνήθως σχετικός και όχι απόλυτος.

Μια ειδική περίπτωση ονομαστικών χαρακτηριστικών είναι τα δυαδικά (Boolean) όπου έχουμε μόνο δύο πιθανές τιμές, συνήθως της μορφής ναι / όχι ή σωστό / λάθος.

Στην πράξη, οι μέθοδοι εξόρυξης πληροφορίας από δεδομένα χρησιμοποιούν τα ονομαστικά και τα τακτικά χαρακτηριστικά. Όλοι οι αλγόριθμοι που χρησιμοποιούνται δε δέχονται και τις δύο μορφές χαρακτηριστικών και γι' αυτό συχνά η εφαρμογή κάποιου συγκεκριμένου αλγορίθμου προϋποθέτει τη μετατροπή ενός ή περισσότερων χαρακτηριστικών από τον ένα τύπο στον άλλο.

Το χαρακτηριστικό το οποίο θεωρούμε ως αποτέλεσμα (*output*) των παρατηρήσεων μας, αυτό δηλαδή το χαρακτηριστικό που κατά κύριο λόγο θέλουμε να μελετήσουμε, ονομάζεται τάξη ή κλάση (class).

Για τους δύο τύπους των αποτελεσμάτων (output), είναι λογικό να σκεφτούμε τη χρήση των εισροών (inputs) για να προβλέψουμε την έξοδο. Για παράδειγμα, λαμβάνοντας υπόψη κάποιες συγκεκριμένες ατμοσφαιρικές μετρήσεις σήμερα και χθες, θέλουμε να προβλέψουμε το επίπεδο του όζοντος αύριο.

Αυτή η διάκριση σε τύπο εξόδου έχει οδηγήσει σε μια σύμβαση στην ονομασία για το αντικείμενο της πρόβλεψης: παλινδρόμηση (regression) όταν η πρόβλεψη αφορά σε ποσοτικά αποτελέσματα και την ταξινόμηση (classification) όταν η πρόβλεψη αφορά σε ποιοτικά αποτελέσματα.

2.5 Μέθοδοι επιλογής χαρακτηριστικών (filter, wrapper)

Στην μηχανική μάθηση και στατιστική, η επιλογή χαρακτηριστικών είναι η διαδικασία της επιλογής ενός υποσυνόλου των σχετικών χαρακτηριστικών για χρήση στην κατασκευή του μοντέλου. Η κεντρική υπόθεση, όταν χρησιμοποιούμε μια τεχνική επιλογής χαρακτηριστικών είναι ότι τα δεδομένα περιέχουν πολλά περιττά ή άσχετα χαρακτηριστικά. Τα περιττά χαρακτηριστικά είναι εκείνα που δεν παρέχουν περισσότερες πληροφορίες από ότι τα ήδη επιλεγμένα χαρακτηριστικά, και τα άσχετα χαρακτηριστικά δεν παρέχουν χρήσιμες πληροφορίες σε οποιοδήποτε πλαίσιο. Οι τεχνικές επιλογής χαρακτηριστικών είναι ένα υποσύνολο του ευρύτερου τομέα της εξαγωγής χαρακτηριστικών. Η εξαγωγή χαρακτηριστικών δημιουργεί νέα χαρακτηριστικά από συναρτήσεις των αρχικών χαρακτηριστικών, ενώ η επιλογή χαρακτηριστικών επιστρέφει ένα υποσύνολο τους. Οι τεχνικές επιλογής χαρακτηριστικών χρησιμοποιούνται συχνά σε περιοχές όπου υπάρχουν πολλά χαρακτηριστικά και συγκριτικά λίγα δείγματα (ή πειραματικά σημεία (datapoints)). Η αρχέτυπη

περίπτωση είναι η χρήση της επιλογής χαρακτηριστικών στην ανάλυση μικροσυστοιχιών DNA, όπου υπάρχουν πολλές χιλιάδες των χαρακτηριστικών, και μερικές δεκάδες έως εκατοντάδες δείγματα. Οι τεχνικές επιλογής χαρακτηριτικών παρέχουν τρία κύρια οφέλη κατά την κατασκευή μοντέλων πρόβλεψης:

- Βελτίωση επεξηγηματικότητας του μοντέλου,
- Μικρότερους χρόνους εκπαίδευσης
- Ενισχυμένη γενίκευση με τη μείωση του overfitting.

Η επιλογή χαρακτηριστικών είναι επίσης χρήσιμη ως μέρος της διαδικασίας ανάλυσης των δεδομένων, καθώς δείχνει ποια χαρακτηριστικά είναι σημαντικά για την πρόβλεψη, και πως σχετίζονται μεταξύ τους.

Ένας αλγόριθμος επιλογής χαρακτηριστικών μπορεί να θεωρηθεί ως ο συνδυασμός μιας τεχνικής αναζήτησης για την πρόταση νέων υποσυνόλων χαρακτηριστικών, μαζί με ένα μέτρο αξιολόγησης που πετυχαίνει τα διαφορετικά υποσύνολα χαρακτηριστικών. Ο απλούστερος αλγόριθμος είναι να δοκιμάσουμε κάθε δυνατό υποσύνολο των χαρακτηριστικών για την εύρεση εκείνου που ελαχιστοποιεί το ποσοστό σφάλματος. Αυτή είναι μια εξαντλητική αναζήτηση του χώρου, και είναι υπολογιστικά δυσεπίλυτο για όλα τα υποσύνολα, αλλά το μικρότερο από τα σύνολα των χαρακτηριστικών. Η επιλογή των μετρικών αξιολόγησης επηρεάζει σε μεγάλο βαθμό τον αλγόριθμο, και αυτές οι μετρήσεις αξιολόγησης είναι εκείνες οι οποίες διακρίνουν τρεις κύριες κατηγορίες αλγορίθμων επιλογής χαρακτηριστικών: τα περιτυλίγματα (wrapper), τα φίλτρα (filters) και τις ενσωματωμένες μεθόδους (embedded methods).

Οι Wrapper μέθοδοι χρησιμοποιούν ένα μοντέλο πρόβλεψης για να σκοράρουν υποσύνολα χαρακτηριστικών. Κάθε νέο υποσύνολο χρησιμοποιείται για να εκπαιδεύσει ένα μοντέλο, το οποίο έχει δοκιμαστεί σε ένα σύνολο hold-out. Μετρώντας τον αριθμό των λαθών που έγιναν σε αυτό το σύνολο hold-out (το ποσοστό σφάλματος του μοντέλου), δίνει τη βαθμολογία για αυτό το υποσύνολο. Όσον αφορά τις μεθόδους περιτυλίγματος που εκπαιδεύουν ένα νέο μοντέλο για κάθε υποσύνολο, είναι πολύ υψηλής έντασης υπολογισμών, αλλά συνήθως παρέχουν το σύνολο χαρακτηριστικών με τις καλύτερες επιδόσεις για το συγκεκριμένο τύπο του μοντέλου.

Οι Filter μέθοδοι χρησιμοποιούν ένα εντολοδοχικό (proxy) μέτρο αντί του ποσοστού σφάλματος για να σκοράροουν ένα υποσύνολο χαρακτηριστικό. Το μέτρο αυτό έχει επιλεγεί να είναι γρήγορο στον υπολογισμό, ενώ ακόμη συλλαμβάνει τη χρησιμότητα των

γνωρισμάτων του συνόλου χαρακτηριστικών. Κοινά μέτρα περιλαμβάνουν την αμοιβαία πληροφορία, το συντελεστή συσχέτισης (product-moment) του Pearson, καθώς και την ενδοαπόσταση. Τα φίλτρα (filters) είναι συνήθως λιγότερο υπολογιστικά εντατικά από τα περιτυλίγματα (wrappers), αλλά παράγουν ένα σύνολο χαρακτηριστικών το οποίο δεν συντονίζεται σε ένα συγκεκριμένο τύπο του μοντέλου πρόβλεψης. Πολλά φίλτρα παρέχουν μια κατάταξη χαρακτηριστικών παρά ένα ρητά καλύτερο υποσύνολο χαρακτηριστικών, και το σημείο αποκοπής στην κατάταξη επιλέγεται μέσω της διασταυρωμένης επικύρωσης.

Οι ενσωματωμένες μέθοδοι είναι μία catch-all ομάδα τεχνικών που εκτελούν επιλογή χαρακτηριστικών, ως μέρος της διαδικασίας κατασκευής του μοντέλου. Το πρότυπο αυτής της προσέγγισης είναι η μέθοδος LASSO για την κατασκευή ενός γραμμικού μοντέλου, το οποίο τιμωρεί τους συντελεστές παλινδρόμησης, συρρικνώνοντας πολλούς από αυτούς στο μηδέν. Τυχόν χαρακτηριστικά που έχουν μη-μηδενικούς συντελεστές παλινδρόμησης «επιλέγονται» από τον αλγόριθμο LASSO. Μια άλλη δημοφιλής προσέγγιση είναι ο αναδρομικός αλγόριθμος εξάλειψης χαρακτηριστικού (Recursive Feature Elimination algorithm), που χρησιμοποιείται συνήθως με τις μηχανές διανυσματικής υποστήριξης για να κατασκευάσει επανειλημμένα ένα μοντέλο και να αφαιρέσει χαρακτηριστικά με χαμηλό βάρος. Οι προσεγγίσεις αυτές τείνουν να είναι μεταξύ των φίλτρων και των περιτυλιγμάτων από πλευράς υπολογιστικής πολυπλοκότητας.

Όσον αφορά στη στατιστική, η πιο δημοφιλής μορφή της επιλογής χαρακτηριστικών είναι η σταδιακή υποχώρηση (stepwise regression). Είναι ένας άπληστος αλγόριθμος που προσθέτει το καλύτερο χαρακτηριστικό (ή διαγράφει το χειρότερο χαρακτηριστικό) σε κάθε επανάληψη. Το κύριο θέμα του ελέγχου είναι να αποφασιστεί πότε θα σταματήσει τον αλγόριθμο. Στη μηχανική μάθηση, αυτό γίνεται συνήθως με διασταυρωμένη επικύρωση. Όσον αφορά στη στατιστική, κάποια κριτήρια είναι βελτιστοποιημένα. Αυτό οδηγεί στο εγγενές πρόβλημα της ένθεσης (nesting). Περισσότερες εύρωστες μέθοδοι έχουν διερευνηθεί, όπως το branch and bound και τμηματικά το γραμμικό δίκτυο.

Κριτήρια βελτιστοποίησης

Υπάρχει μια ποικιλία από κριτήρια βελτιστοποίησης που μπορεί να χρησιμοποιηθεί για τον έλεγχο επιλογής χαρακτηριστικών. Τα παλαιότερα είναι το στατιστικό Cp-Mallows και το κριτήριο πληροφορίας Akaike (AIC). Αυτά προσθέτουν μεταβλητές εάν το t-στατιστικό είναι μεγαλύτερο από $\sqrt{2}$.

Άλλα κριτήρια είναι το Bayesian κριτήριο πληροφορίας (BIC), το οποίο χρησιμοποιεί το $\sqrt{logn}$, το ελάχιστο μήκος περιγραφής (MDL), το οποίο χρησιμοποιεί ασυμπτωτικά το $\sqrt{logn}$, το Bonnferroni / RIC που χρησιμοποιεί το $\sqrt{2logp}$, μέγιστη εξάρτηση επιλογή χαρακτηριστικού, καθώς και μια ποικιλία από νέα κριτήρια που υποκινούνται από την ανακάλυψη του ψευδούς ποσοστού (FDR), τα οποία χρησιμοποιούν κάτι κοντά στο $\sqrt{2\log\frac{p}{q}}$.

Κεφάλαιο 3:

Μέθοδοι μείωσης διαστάσεων

3.1 Εισαγωγή

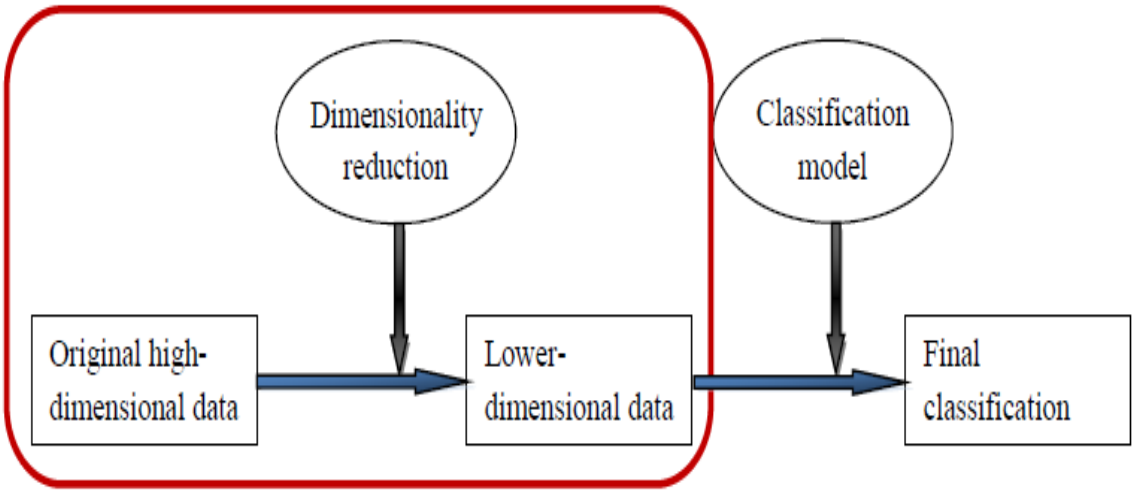
Οι βάσεις δεδομένων που χρησιμοποιούνται στα περισσότερα πρακτικά προβλήματα εξόρυξης γνώσης από δεδομένα είναι πολύ μεγάλων διαστάσεων (High Dimensional) καθώς αποτελούνται από εκατομμύρια εγγραφές και εκατοντάδες μεταβλητές (χαρακτηριστικά). Η ταξινόμηση αυτών των δεδομένων είναι ένα δύσκολο πρόβλημα, επειδή ο τεράστιος αριθμός των μεταβλητών δημιουργεί προκλήσεις με τις συμβατικές μεθόδους ταξινόμησης και καθιστά πολλές κλασικές τεχνικές ανέφικτες. Μια φυσική λύση είναι να προηγηθεί ένα βήμα μείωσης διαστάσεων της βάσης δεδομένων πριν εφαρμοστεί η τεχνική της ταξινόμησης.

Θεωρητικά, η ύπαρξη περισσότερων μεταβλητών κάνει περισσότερο αποδοτική τη διαδικασία μάθησης όμως στην πράξη η προσθήκη μη σχετικών μεταβλητών «συγχύζει» τους αλγορίθμους που εφαρμόζουμε. Η χρήση πολλών μεταβλητών για τη μοντελοποίηση μίας σχέσης με μία μεταβλητή απόκρισης μπορεί να περιπλέξει την ερμηνεία της ανάλυσης και να παραβεί την αρχή της φειδωλότητας (principle of parsimony), σύμφωνα με την οποία πρέπει να κρατήσουμε τον αριθμό των μεταβλητών σε ένα μέγεθος το οποίο να μπορούμε εύκολα να εξηγήσουμε. Επιπλέον, η παραμονή πάρα πολλών μεταβλητών μπορεί να οδηγήσει σε υπερπροσαρμογή όπου η γενικότητα των αποτελεσμάτων που βρίσκουμε παρεμποδίζεται επειδή νέα δεδομένα δε συμπεριφέρονται το ίδιο με τα δεδομένα εκπαίδευσης για όλες τις μεταβλητές.

Οι μέθοδοι μείωσης των διαστάσεων της βάσης δεδομένων χρησιμοποιούν τις συσχετίσεις μεταξύ των μεταβλητών για να μειώσουν τον αριθμό τους, να επιβεβαιώσουν ότι αυτές οι μεταβλητές είναι ανεξάρτητες και να ερμηνεύσουν τα αποτελέσματα.

Δύο από τις μεθόδους μείωσης των διαστάσεων της βάσης δεδομένων είναι

- Η ανάλυση κυρίων συνιστωσών (principal components analysis)
- Παραγοντική ανάλυση (Factor analysis)



Σχήμα 5: Αναπαράσταση της διαδικασίας που ακολουθείται, συνήθως, σε δεδομένα υψηλής διάστασης

3.2 Ανάλυση κυρίων συνιστωσών

Η ανάλυση κύριων συνιστωσών (principal component analysis, PCA) είναι μία μέθοδος μείωσης των διαστάσεων μίας βάσης δεδομένων η οποία αναπτύχθηκε το 1901 από τον Karl Pearson. Η μέθοδος έχει σκοπό να δημιουργήσει γραμμικούς συνδυασμούς των αρχικών μεταβλητών, τους οποίους ονομάζουμε συνιστώσες, έτσι ώστε οι γραμμικοί αυτοί συνδυασμοί να είναι ασυσχέτιστοι μεταξύ τους αλλά να περιέχουν όσο γίνεται μεγαλύτερο μέρος της διακύμανσης των αρχικών μεταβλητών. Το κέρδος από μία τέτοια διαδικασία είναι πως από ένα σύνολο συσχετισμένων μεταβλητών καταλήγουμε σε ένα μικρότερο σύνολο ασυσχέτιστων μεταβλητών κάτι το οποίο είναι πολύ χρήσιμο για ορισμένες στατιστικές μεθόδους και για την εξόρυξη γνώσης από δεδομένα. Για παράδειγμα υπενθυμίζουμε το πρόβλημα της πολυσυγγραμμικότητας στην παλινδρόμηση όπου αν χρησιμοποιήσουμε τις συσχετισμένες μεταβλητές οι εκτιμήσεις που θα πάρουμε δε θα είναι συνεπείς, ενώ αν χρησιμοποιήσουμε ασυσχέτιστες μεταβλητές το πρόβλημα παύει να υπάρχει. Επίσης, αν οι κύριες συνιστώσες που θα προκύψουν μπορούν να ερμηνεύσουν ένα μεγάλο ποσοστό της διακύμανσης τότε αυτό σημαίνει πως αντί να έχουμε m μεταβλητές, όπως είχαμε αρχικά, έχουμε λιγότερες, με κόστος βέβαια ότι χάνουμε κάποιο, ελπίζουμε

μικρό, ποσοστό της συνολικής μεταβλητότητας. Σε βάσεις δεδομένων πολύ μεγάλων διαστάσεων μπορούμε δηλαδή αντί να αποθηκεύσουμε όλες τις μεταβλητές να αποθηκεύσουμε μόνο κάποιο αριθμό κύριων συνιστωσών, κάτι το οποίο μπορεί να αποδειχθεί ζωτικής σημασίας. Επιπλέον, η μέθοδος κύριων συνιστωσών είναι πολύ χρήσιμη για τις περιπτώσεις που έχουμε λίγες παρατηρήσεις και πολλές μεταβλητές, για τις περιπτώσεις που θέλουμε να εξετάσουμε τις συσχετίσεις ανάμεσα στις μεταβλητές και να εξετάσουμε πόσο μοιάζουν ή όχι, καθώς και τις περιπτώσεις που θέλουμε να αναγνωρίσουμε δίνοντας ονόματα στις καινούργιες μεταβλητές (τις συνιστώσες) παρατηρώντας ποιες αρχικές μεταβλητές έχουν μεγάλη επίδραση σε αυτές. Αυτό είναι πολύ χρήσιμο σε κάποιες επιστήμες καθώς μας επιτρέπει να ποσοτικοποιήσουμε μη μετρήσιμες ποσότητες.

Σύμφωνα με την ανάλυση των κύριων συνιστωσών έχουμε ότι m μεταβλητές περιέχουν σχεδόν τόση πληροφορία όση k συνιστώσες τους, δηλαδή k γραμμικοί συνδυασμοί τους, όπου $k < m$ συνιστώσες έτσι ώστε το σύνολο δεδομένων με το οποίο δουλεύουμε να περιέχει τώρα n υποδείγματα στις k συνιστώσες αντί για n υποδείγματα στις m μεταβλητές. Η μέθοδος των κύριων συνιστωσών στηρίζεται στη φασματική ανάλυση ενός τετραγωνικού πίνακα, σύμφωνα με την οποία αν ξεκινήσουμε από ένα *τετραγωνικό πίνακα Α* μπορούμε να καταλήξουμε σε ένα *διαγώνιο πίνακα Λ*, πολλαπλασιάζοντας κατάλληλα με ένα *ορθογώνιο πίνακα P* ο οποίος αποτελείται από τα κανονικοποιημένα διανύσματα. Περιγράψαμε, λοιπόν, την γενική ιδέα με βάση την οποία γίνεται η εύρεση των κύριων συνιστωσών. Περαιτέρω ανάλυση παρεκκλίνει από τους στόχους της παρούσας εργασίας.

Το πιο σημαντικό κομμάτι της ανάλυσης κυρίων συνιστωσών είναι η απόφαση για τον αριθμό των συνιστωσών που θα κρατήσουμε. Επιλέγοντας λιγότερες κύριες συνιστώσες από όσες μεταβλητές είχαμε αρχικά, αναγκαστικά χάνουμε κάποια πληροφορία, αυτή η απώλεια αποτελεί και το κόστος μας για το κέρδος μας να μειώσουμε τις διαστάσεις του προβλήματος. Αυτό που επιδιώκουμε είναι να κρατήσουμε ένα μικρό αριθμό συνιστωσών, με το μικρότερο δυνατό κόστος. Για να αποφασίσουμε τον αριθμό των συνιστωσών που θα κρατήσουμε έχουν αναπτυχθεί πολλά κριτήρια, μερικά από τα οποία είναι το ποσοστό συνολικής διακύμανσης που εξηγούν οι συνιστώσες (proportion of variance explained criterion), Scree plot, κριτήριο των ιδιοτιμών (eigenvalue criterion) ή κριτήριο του Kaiser, το ποσοστό της διακύμανσης των αρχικών μεταβλητών που ερμηνεύεται, την κανονική προσέγγιση, το bootstrap και το cross validation.

3.3 Παραγοντική ανάλυση

Η παραγοντική ανάλυση (factor analysis) είναι μία μέθοδος μείωσης των διαστάσεων της βάσης δεδομένων η οποία όμως έχει διαφορετικό στόχο από την μέθοδο ανάλυσης κύριων συνιστωσών που εξετάσαμε παραπάνω. Η μέθοδος ανάλυσης κύριων συνιστωσών

δημιουργεί ορθογώνιους γραμμικούς συνδυασμούς των μεταβλητών για να περιγράψει τις μεταβλητές και τη μεταβλητότητά τους καθώς και για να αντικαταστήσει το αρχικό σύνολο μεταβλητών με ένα μικρότερο σύνολο ασυσχέτιστων συνιστωσών. Αντιθέτως, η παραγοντική ανάλυση δημιουργεί ένα μοντέλο για τα δεδομένα κάτω από κάποιες υποθέσεις και άρα είναι πιο λεπτομερής και πιο στατιστική μέθοδος. Επιπλέον, η παραγοντική ανάλυση προσπαθεί να ερμηνεύσει περισσότερο τη δομή και την συνδιακύμανση των μεταβλητών, παρά τη διακύμανση τους, όπως συμβαίνει στην ανάλυση κύριων συνιστωσών.

Τα κυριότερα προβλήματα που παρουσιάζονται στην παραγοντική ανάλυση είναι ότι στηρίζεται σε ένα πλήθος υποθέσεων οι οποίες δεν είναι απαραίτητα ρεαλιστικές για πραγματικά προβλήματα και συνήθως ο ερευνητής δεν μπορεί να τις ελέγξει εύκολα. Ένα δεύτερο πρόβλημα της μεθόδου είναι ότι δεν έχει μοναδική λύση. Όπως θα δούμε και στη συνέχεια, μπορούμε να χρησιμοποιήσουμε διάφορες μεθόδους εκτίμησης και ακόμα και για την ίδια μέθοδο εκτίμησης μπορούμε να πάρουμε ένα μεγάλο αριθμό ισοδύναμων εκτιμήσεων. Έτσι βασιζόμενοι στα ίδια δεδομένα μπορούμε να πάρουμε διαφορετικά αποτελέσματα. Επίσης, οι παράγοντες που προκύπτουν μπορούν να δεχτούν διαφορετικές ερμηνείες οι οποίες μπορεί και να έρχονται σε αντιπαράθεση. Ένα τελευταίο πρόβλημα της μεθόδου είναι ότι ο αριθμός των παραγόντων που χρειάζεται να εξάγουμε ώστε τα αποτελέσματα να είναι χρήσιμα, δεν είναι προφανής και εξαρτάται και από τη μέθοδο εκτίμησης που θα χρησιμοποιηθεί. Αυτό επιτρέπει στον αναλυτή να δουλεύει σε μία μεροληπτική βάση έτσι ώστε να εμφανίζει τα αποτελέσματα όπως τον συμφέρουν.

Παρά τα προβλήματα της, η μέθοδος αποτελεί πολύτιμο εργαλείο σε πολλές επιστήμες κυρίως λόγω του ότι δίνει τη δυνατότητα στον αναλυτή να ποσοτικοποιήσει μη παρατηρήσιμες ποσότητες.

3.5 Αλγόριθμος επιλογής μεταβλητών

Τα προβλήματα εξόρυξης δεδομένων περιλαμβάνουν συχνά εκατοντάδες ή ακόμα και χιλιάδες, των μεταβλητών. Ως αποτέλεσμα, το μεγαλύτερο μέρος του χρόνου και της προσπάθειας που δαπανάται στη διαδικασία οικοδόμησης του μοντέλο περιλαμβάνει την εξέταση των μεταβλητών που πρέπει θα συμπεριληφθούν στο μοντέλο. Η προσαρμογή ενός νευρωνικού δικτύου ή ενός δέντρου απόφασης σε ένα μεγάλο σύνολο μεταβλητών μπορεί να απαιτεί περισσότερο χρόνο από ότι είναι πρακτικό. Η επιλογή χαρακτηριστικών επιτρέπει στο σύνολο μεταβλητών να μειωθεί σε μέγεθος, δημιουργώντας ένα πιο εύχρηστο σύνολο χαρακτηριστικών για την μοντελοποίηση. Η προσθήκη της επιλογής χαρακτηριστικών στην αναλυτική διαδικασία έχει πολλά πλεονεκτήματα:

- Απλοποιεί και περιορίζει το πεδίο εφαρμογής της στα χαρακτηριστικά που είναι απαραίτητα για την ανάπτυξη ενός μοντέλου πρόβλεψης.
- Ελαχιστοποιεί τον υπολογιστικό χρόνο και τις απαιτήσεις μνήμης για την οικοδόμηση ενός μοντέλου πρόβλεψης επειδή η εστίαση μπορεί να κατευθύνεται προς το υποσύνολο των προβλέψεων που είναι πιο σημαντικό.
- Οδηγεί σε πιο ακριβή και / ή πιο φειδωλά μοντέλα.
- Μειώνει το χρόνο για τη δημιουργία βαθμολογιών επειδή το προγνωστικό μοντέλο βασίζεται σε ένα μόνο υποσύνολο της πρόβλεψης.

Για να μειώσουμε, λοιπόν, στο ελάχιστο τις πιθανές επιλογές μεταβλητών, ο αλγόριθμος της επιλογής των χαρακτηριστικών (Feature Selection Algorithm) μπορεί να χρησιμοποιηθεί για να προσδιορίσει τα πεδία εκείνα τα οποία είναι πιο σημαντικά για τη δεδομένη ανάλυση.

Η επιλογή χαρακτηριστικών αποτελείται από τρία βήματα :

- Screening (κρισάρισμα) :

  Σε αυτό το βήμα απομακρύνονται οι μη σημαντικές και προβληματικές μεταβλητές πρόβλεψης καθώς και εγγραφές, όπως στην περίπτωση που έχουμε μεταβλητές με πολλές ελλειπούσες τιμές ή μεταβλητές με πολύ μεγάλη ή πολύ μικρή διακύμανση για να τις καθιστά χρήσιμες.

- Ranking (Στοίχιση) :

  Σε αυτό το βήμα ξεχωρίζονται οι εναπομείναντες μεταβλητές πρόβλεψης και καθορίζονται ranks βασισμένα στη σημαντικότητα.

- Selection (Επιλογή) :

  Σε αυτό το βήμα αναγνωρίζεται το υποσύνολο των χαρακτηριστικών που θα χρησιμοποιηθεί στα μοντέλα που ακολουθούν κρατώντας μόνο τις πιο σημαντικές μεταβλητές πρόβλεψης και φιλτράροντας ή αποκλείοντας όλες τις υπόλοιπες.

Εν κατακλείδι, τα πλεονεκτήματα από την επιλογή χαρακτηριστικών είναι ότι η διαδικασία της μοντελοποίησης απλοποιείται και φυσικά γίνεται ταχύτερη. Μειώνοντας τον αριθμό των πεδίων που χρησιμοποιούνται στο μοντέλο μειώνεται ο χρόνος αξιολόγησης του μοντέλου και επιπρόσθετα αποκτούμε απλούστερα και ακριβέστερα μοντέλα τα οποία μπορούν πολύ πιο εύκολα να εξηγηθούν .

Κεφάλαιο 4:

Μέθοδοι Ταξινόμησης

4.1 Εισαγωγή

Η ταξινόμηση δεδομένων είναι μία διαδικασία η οποία βρίσκει τις κοινές ιδιότητες μεταξύ ενός συνόλου αντικειμένων σε μία βάση δεδομένων και ταξινομεί τα αντικείμενα αυτά σε διαφορετικές κλάσεις (τάξεις) σύμφωνα με ένα μοντέλο ταξινόμησης. Για να κατασκευάσουμε ένα τέτοιο μοντέλο ταξινόμησης, μία δειγματική βάση δεδομένων $E = \{t_1,\ t_2, \dots, t_n,\}$ θεωρείται ως το σύνολο εκπαίδευσης (training set) στο οποίο κάθε εγγραφή αποτελείται από το ίδιο σύνολο πολλαπλών χαρακτηριστικών όπως οι εγγραφές σε μία μεγάλη βάση δεδομένων W και επισπρόσθετα κάθε εγγραφή έχει μία γνωστή ετικέτα (label) κλάσης. Το σύνολο των κλάσεων το συμβολίζουμε με $C = \{c_1,\ c_2, \dots, c_n,\}$.

Ο αντικειμενικός σκοπός της ταξινόμησης είναι πρώτον να αναλύσει τα δεδομένα του συνόλου εκπαίδευσης και να αναπτύξει μία ακριβή περιγραφή ή ένα μοντέλο για κάθε κλάση χρησιμοποιώντας τα χαρακτηριστικά που είναι διαθέσιμα στα δεδομένα. Με άλλα λόγια το πρόβλημα της κατηγοριοποίησης έγκειται στον ορισμό μίας απεικόνισης $f : E \longrightarrow C$ όπου κάθε εγγραφή t_i ανατίθεται σε μία κλάση c_i. Οι περιγραφές κλάσεων που προκύπτουν χρησιμοποιούνται στη συνέχεια για να ταξινομήσουν μελλοντικά δεδομένα (test set) στη βάση δεδομένων W ή για να αναπτύξουν μια καλύτερη περιγραφή την οποία ονομάζουμε «κανόνες ταξινόμησης» για κάθε κλάση στη βάση δεδομένων. Επομένως, μπορούμε να θεωρήσουμε ότι με την ταξινόμηση διαμερίζουμε το σύνολο Ε σε κλάσεις ισοδυναμίας και επιπλέον ότι το πρόβλημα της πρόβλεψης είναι ένα πρόβλημα ταξινόμησης όπου έχουμε άπειρο αριθμό κλάσεων.

Η ταξινόμηση βρίσκει πολλές εφαρμογές σε ιατρικές διαγνώσεις, στο marketing και αλλού και αποτελεί αντικείμενο μελέτης για τη στατιστική, τη μηχανική μάθηση και βέβαια το data

mining. Πρόκειται για μάθηση με επίβλεψη (supervised learning) καθώς οι ομάδες ταξινόμησης είναι εκ των προτέρων γνωστές και το πραγματικό αποτέλεσμα κάθε υποδείγματος είναι επίσης γνωστό. Επομένως, είναι δυνατό να μετράμε το βαθμό αξιοπιστίας σε μη χρησιμοποιημένα για τη διαμόρφωση της αντίληψης δεδομένα ή υποκείμενα, ανάλογα με το βαθμό αποδοχής της περιγραφής.

Η τυπική προσέγγιση που χρησιμοποιούν οι τεχνικές ταξινόμησης είναι η δημιουργία ενός μοντέλου μέσω της αξιολόγησης του συνόλου δεδομένων εκπαίδευσης και η εφαρμογή του μοντέλου σε νέα δεδομένα. Οι πιο κοινές τεχνικές είναι τα δέντρα αποφάσεων (decision trees), τα νευρωνικά δίκτυα (Neural Networks), οι μηχανές διανυσματικής υποστήριξης (Support Vector Machines), η λογιστική παλινδρόμηση (logistic regression) και τα Bayesian Network Models με τα οποία θα ασχοληθούμε στη συνέχεια.

4.2 Λογιστική Παλινδρόμηση (Logistic Regression)

Μία ευρέως χρησιμοποιούμενη τακτική για την ταξινόμηση υποδειγμάτων σε κλάσεις είναι η παλινδρόμηση. Με τη χρήση μοντέλων παλινδρόμησης επιτυγχάνουμε τη διαμέριση του χώρου σε περιοχές, δηλαδή σε κλάσεις ισοδυναμίας, και επιπλέον καθιστούμε εύκολη την πρόβλεψη της τάξης μελλοντικών υποδειγμάτων. Έτσι λοιπόν, μπορούμε να εφαρμόσουμε οποιαδήποτε τεχνική παλινδρόμησης, όπως απλή ή πολλαπλή γραμμική παλινδρόμηση, μη γραμμική ή λογαριθμική παλινδρόμηση, για την ταξινόμηση υποδειγμάτων σε κλάσεις.

4.2.1 Ορισμός

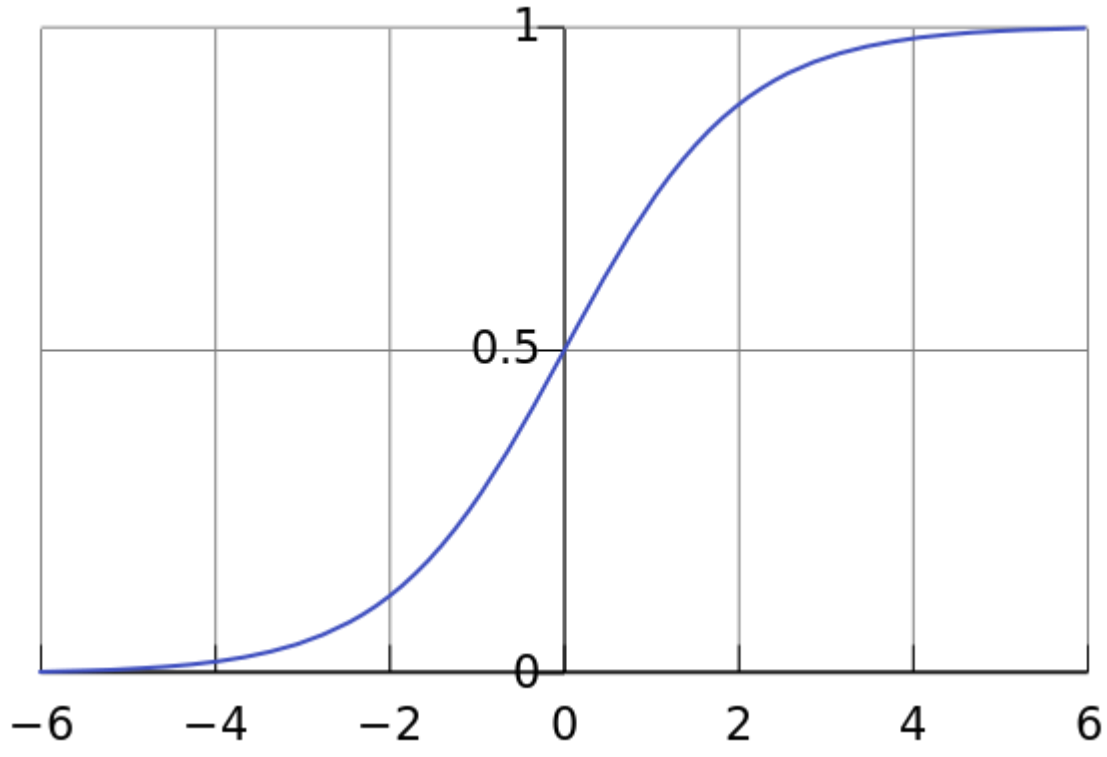
Η γραμμική παλινδρόμηση χρησιμοποιείται για να προσεγγίσει τη σχέση μεταξύ μίας συνεχούς μεταβλητής απόκρισης και ενός συνόλου ανεξάρτητων επεξηγηματικών (predictor) μεταβλητών. Υπάρχουν όμως περιπτώσεις όπου η μεταβλητή απόκρισης δεν είναι συνεχής αλλά κατηγορική (categorical). Για τις περιπτώσεις αυτές η γραμμική παλινδρόμηση δεν είναι κατάλληλη, όμως ο αναλυτής μπορεί να χρησιμοποιήσει μία ανάλογη μέθοδο, τη λογιστική παλινδρόμηση, η οποία έχει αρκετά κοινά σημεία με τη γραμμική παλινδρόμηση. Η λογιστική παλινδρόμηση αναφέρεται σε μεθόδους που περιγράφουν τη σχέση μεταξύ μίας κατηγορικής μεταβλητής απόκρισης και ενός συνόλου επεξηγηματικών μεταβλητών. Στην παρούσα διπλωματική εργασία θα ασχοληθούμε μόνο

με τη λογιστική παλινδρόμηση για δυαδικές μεταβλητές, δηλαδή με την περίπτωση όπου η μεταβλητή απόκρισης έχει μόνο δύο κατηγορίες. Η περίπτωση όπου η μεταβλητή απόκρισης έχει περισσότερες από δύο κατηγορίες είναι εκτός των σκοπών της παρούσας εργασίας και δε θα την εξετάσουμε.

Ας δούμε τώρα πως δημιουργείται η γραμμή της λογιστικής παλινδρόμησης. Θεωρούμε τη δεσμευμένη μέση τιμή της Y δεδομένου ότι X = x, την οποία συμβολίζουμε με E (Y | x). Αυτή η τιμή εκφράζει την αναμενόμενη τιμή της μεταβλητής απόκρισης για δοθείσα τιμή της επεξηγηματικής μεταβλητής. Υπενθυμίζουμε ότι στη γραμμική παλινδρόμηση θεωρούμε τη μεταβλητή απόκρισης ως τυχαία μεταβλητή που ορίζεται από τη σχέση:

$$Y = \beta_0 + \beta_1 * x + \varepsilon$$

Επομένως, καθώς το σφάλμα έχει μηδενική μέση τιμή, έχουμε $E\ (Y \mid x) = \beta_0 + \beta_1 * x$ για τη γραμμική παλινδρόμηση και τις πιθανές τιμές να βρίσκονται πάνω σε ολόκληρο τον άξονα των πραγματικών αριθμών.



Σχήμα 6: Η συνάρτηση λογιστικής παλινδρόμησης, με $\beta_0 + \beta_1 * x$ στον οριζόντιο άξονα και π(x) στον κατακόρυφο άξονα.

Η δεσμευμένη μέση τιμή για τη λογιστική παλινδρόμηση, η οποία συμβολίζεται με π(x) αντί για $E(Y \mid x)$, έχει διαφορετική μορφή από ότι στη γραμμική παλινδρόμηση και συγκεκριμένα ορίζεται από τη σχέση

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 * x}}{1 + e^{\beta_0 + \beta_1 * x}} \qquad (1)$$

Οι καμπύλες που προκύπτουν από την (1) ονομάζονται σιγμοειδείς καθώς έχουν το σχήμα ενός S και επομένως είναι μη γραμμικές. Η λογιστική παλινδρόμηση έχει επιλεγεί για να μοντελοποιήσει διχοτομικά δεδομένα διότι είναι ευέλικτη και εύκολη στην ερμηνεία. Η ελάχιστη τιμή για το π(x) λαμβάνεται στο $\lim_{\alpha \to -\infty} \frac{e^{\alpha}}{(1+ e^{\alpha})} = 0$ ενώ η μέγιστη στο $\lim_{\alpha \to \infty} \frac{e^{\alpha}}{(1+ e^{\alpha})} = 1$.

Επομένως το π(x) μπορεί να θεωρηθεί ως η πιθανότητα να εμφανιστεί ένα θετικό αποτέλεσμα (για παράδειγμα ασθένεια) για τις εγγραφές με $X = x$, με $0 \leq \pi(x) \leq 1$, και $1 - \pi(x)$ μπορεί να θεωρηθεί ως η πιθανότητα να μην εμφανιστεί ένα θετικό αποτέλεσμα για αυτές τις εγγραφές.

Σφάλματα

Τα μοντέλα παλινδρόμησης υποθέτουν ότι $Y = \beta_0 + \beta_1 * x + \varepsilon$, όπου ο όρος του σφάλματος ε είναι κανονικά κατανεμημένος με μέσο μηδέν και σταθερή διακύμανση.

Η υπόθεση που γίνεται για το μοντέλο στη λογιστική παλινδρόμηση είναι διαφορετική. Καθώς η απόκριση είναι διχοτομική, τα σφάλματα μπορούν να έχουν μόνο δύο πιθανές μορφές :

- αν $Y = 1$ (δηλαδή εμφάνιση της ασθένειας), το οποίο συμβαίνει με πιθανότητα π(x) (η πιθανότητα η απόκριση να είναι θετική), τότε

  $$\varepsilon = 1 - \pi(x)$$

  είναι η κάθετη απόσταση μεταξύ του δεδομένου σημείου $Y = 1$ και της καμπύλης $\pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{1+ e^{\beta_0 + \beta_1 x}}$ ακριβώς κάτω από αυτό για $X = x$.

Από την άλλη,

- αν $Y = 0$ (δηλαδή απουσία της ασθένειας), το οποίο συμβαίνει με πιθανότητα 1 – π(x) (η πιθανότητα η απόκριση να είναι αρνητική), τότε

$$\varepsilon = 0 - \pi(x)$$

είναι η κάθετη απόσταση μεταξύ του δεδομένου σημείου Y = 0 και της καμπύλης π(x) ακριβώς από πάνω του για $X = x$.

Επομένως η διακύμανση του ε είναι $\pi(x) * [\, 1 - \pi(x)\,]$, η οποία είναι η διακύμανση για τη διωνυμική κατανομή, και η μεταβλητή απόκρισης στη λογιστική παλινδρόμηση

$$Y = \pi(x) + \varepsilon$$

υποθέτουμε ότι ακολουθεί τη διωνυμική κατανομή με πιθανότητα επιτυχίας π(x).

Ένας χρήσιμος μετασχηματισμός που χρησιμοποιείται για τη λογιστική παλινδρόμηση είναι ο $\log it$ μετασχηματισμός ο οποίος ορίζεται ως εξής:

$$g(x) = \ln \frac{\pi(x)}{1 - \pi(x)} = \beta_0 + \beta_1 x$$

Ο μετασχηματισμός αυτός εμφανίζει πολλές ενδιαφέρουσες ιδιότητες του μοντέλου γραμμικής παλινδρόμησης, όπως τη γραμμικότητα, τη συνέχεια και το εύρος από (- ∞, +∞).

4.2.2 Προσαρμογή του μοντέλου

Μέθοδος μέγιστης πιθανοφάνειας

Μία από τις πιο ενδιαφέρουσες ιδιότητες της γραμμικής παλινδρόμησης είναι ότι με την εφαρμογή της μεθόδου ελαχίστων τετραγώνων μπορούν να προκύψουν λύσεις κλειστής μορφής για τις βέλτιστες τιμές των συντελεστών παλινδρόμησης. Κάτι τέτοιο δυστυχώς δεν ισχύει και για τους συντελεστές της λογιστικής παλινδρόμησης και άρα, θα πρέπει να στραφούμε στη μέθοδο μέγιστης πιθανοφάνειας η οποία βρίσκει εκτιμήσεις των παραμέτρων για τις οποίες η πιθανοφάνεια των δεδομένων είναι μέγιστη.

Η συνάρτηση πιθανοφάνειας $I(\boldsymbol{\beta} \mid x)$ είναι μία συνάρτηση των παραμέτρων $\boldsymbol{\beta} = \beta_0, \beta_1, \ldots, \beta_m$ η οποία εκφράζει την πιθανότητα του παρατηρούμενου δεδομένου x. Βρίσκοντας τις τιμές $\boldsymbol{\beta} = \beta_0, \beta_1, \ldots, \beta_m$ που μεγιστοποιούν την $I(\boldsymbol{\beta} \mid x)$ βρίσκουμε τις εκτιμήσεις της μέγιστης πιθανοφάνειας. Η πιθανότητα θετικής απόκρισης δοθέντος του

δεδομένου x είναι $\pi(x) = P(Y = 1 \mid x)$ και η πιθανότητα αρνητικής απόκρισης είναι $1 - \pi(x) = P(Y = 0 \mid x)$. Τότε, παρατηρήσεις των οποίων η απόκριση είναι θετική $(X_i = x_i, Y_i = 1)$ θα συμβάλουν με πιθανότητα π(x) στην πιθανοφάνεια ενώ παρατηρήσεις των οποίων η απόκριση είναι αρνητική $(X_i = x_i, Y_i = 0)$ θα συμβάλουν με πιθανότητα $1 - \pi(x)$ στην πιθανοφάνεια. Επομένως, καθώς $Y_i = 0$ ή 1, η κατανομή στην πιθανοφάνεια της i – οστής παρατήρησης μπορεί να εκφραστεί ως

$$[\pi(x_i)]^{y_i}[1 - \pi(x_i)]^{1-y_i}.$$

Η υπόθεση ότι οι παρατηρήσεις είναι ανεξάρτητες μας επιτρέπει να εκφράσουμε τη συνάρτηση πιθανοφάνειας ως γινόμενο ξεχωριστών όρων:

$$I(\boldsymbol{\beta} \mid x) = \prod_{i=1}^{n} [\pi(x_i)]^{y_i}[1 - \pi(x_i)]^{1-y_i}$$

Ο λογάριθμος της πιθανοφάνειας $L(\boldsymbol{\beta} \mid x) = \ln[I(\boldsymbol{\beta} \mid x)]$ δίνεται από τη σχέση

$$L(\boldsymbol{\beta} \mid x) = \ln[I(\boldsymbol{\beta} \mid x)] = \sum_{i=1}^{n} \{y_i \ln[\pi(x_i)] + (1 - y_i) \ln[1 - \pi(x_i)]\} \quad (2)$$

Οι εκτιμήσεις της μέγιστης πιθανοφάνειας μπορούν να βρεθούν παραγωγίζοντας το λογάριθμο πιθανοφάνειας $L(\boldsymbol{\beta} \mid x)$ ως προς τις παραμέτρους και θέτοντας τις παραγώγους αυτές ίσες με μηδέν. Δυστυχώς, αντίθετα με ό,τι συμβαίνει στη γραμμική παλινδρόμηση, δεν προκύπτουν λύσεις κλειστής μορφής για αυτές τις παραγώγους. Επομένως πρέπει να εφαρμοστούν άλλες μέθοδοι όπως επαναληπτικά σταθμισμένα ελάχιστα τετράγωνα (iterative weighted least squares).

4.2.3 Εφαρμογή της Λογιστικης παλινδρόμησης

Σκοπός του παραδείγματος είναι να μελετήσουμε τη χρήση της λογιστικής παλινδρόμησης για την ανάλυση της επίδρασης μιας ουσίας σε ένα πείραμα τοξικότητας.

Ο παρακάτω πίνακας δείχνει την επίδραση διαφορετικών δόσεων νικοτίνης στην κοινή μύγα των φρούτων.

Συγκέντρωση x $(g/100cc)$	**Αριθμός εντόμων** N	**Αριθμός εντόμων που απεβίωσαν** y	**Ποσοστό**
0.10	47	8	17.0
0.15	53	14	26.4
0.20	55	24	43.6
0.30	52	32	61.5
0.50	46	38	82.6
0.70	54	50	92.6
0.95	52	50	96.2

Πίνακας 1: Δεδομένα Προβλήματος-πείραμα τοξικότητας

Με χρήση της λογιστικής παλινδρόμησης θα καταλήξουμε σε ένα κατάλληλο μοντέλο και θα εκτιμήσουμε τις αποτελεσματικές δόσεις (ED), τις τιμές δηλαδή της νικοτίνης που οδηγούν σε μια συγκεκριμένη τιμή πιθανότητας P. Τέτοιες ποσότητες χρησιμοποιούνται συχνά για να χαρακτηρίσουν τα αποτελέσματα μιας πειραματικής διαδικασίας. Θα εκτιμήσουμε την ED_{50}, όπου ED_P είναι η *τιμή του* x για την οποία η πιθανότητα του θανάτου μιας μύγας των φρούτων παίρνει την τιμή P.

Το στατιστικό πακέτο (PROC LOGIST from SAS) δίνει τα ακόλουθα αποτελέσματα:

Ανάλυση των εκτιμητών μέγιστης Πιθανοφάνειας

Μεταβλητή	B.E.	Εκτίμηση Παραμέτρου	Τυπικό σφάλμα	Wald chi-square	Pr>Chi-square	Κανονικοποιημένη εκτίμηση
INTERCEPT	1	-1.7361	0.2420	51.4482	0.0001	
X	1	6.2954	0.7422	71.9399	0.0001	1.024917
INTERCEPT	1	3.1236	0.3349	86.9818	0.0001	
LOGX	1	2.1279	0.2214	92.3628	0.0001	0.898802

Χρησιμοποιήθηκαν δυο λογιστικά μοντέλα με διαφορετική μορφή το καθένα για τον γραμμικό εκτιμητή. Αρχικά χρησιμοποιήθηκε το τυπικό μοντέλο της εξίσωσης (1) με το τυπικό linear predictor $\beta_0 + \beta_1 x$. Ακόμη , χρησιμοποιήθηκε το μοντέλο

$$P = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \ln x)}}$$

Σε τέτοιου είδους πειράματα συχνά αντικαθιστούμε το x με το $\ln x$. Αυτό είναι ιδιαίτερα χρήσιμο όταν το x έχει μεγάλο εύρος τιμών. Οι $P - τιμές$ των παραμέτρων που παράχθηκαν από τα στατιστικά X^2 του Wald ελέγχου είναι αρκετά σημαντικές και για τα δύο μοντέλα, έτσι έχουμε δύο υποψήφια μοντέλα. Μια μέθοδος για να συγκρίνουμε τα δύο μοντέλα είναι να συγκρίνουμε το εύρος των διαστημάτων εμπιστοσύνης γύρω από το $\hat{y}$ (Lewis, Montgomery and Myers 2001). Μια άλλη σχετική μέθοδος είναι να παρατηρήσουμε το τυπικό σφάλμα του εκτιμώμενου predictor $x'b$ για τα δύο μοντέλα. Στον παρακάτω πίνακα παρουσιάζονται τα τυπικά σφάλματα των γραμμικών εκτιμητών (linear predictors).

$b_0 + b_1 x$	$b_0' + b_i' \ln x$
0.1844	0.2440
0.1607	0.1763
0.1428	0.1439
0.1336	0.1408
0.2139	0.2041
0.3432	0.2646
0.5194	0.3246

Στην περίπτωσή μας είναι δύσκολο να επιλέξουμε ανάμεσα στα δύο μοντέλα χρησιμοποιώντας τις παραπάνω πληροφορίες, παρόλο που τα τυπικά σφάλματα είναι αρκετά μικρότερα για το log μοντέλο στις υψηλές δόσεις. Η χρήση των υπολοίπων (residuals) για την εξέταση αυτών των μοντέλων με τον ίδιο τρόπο που χρησιμοποιούνται στα συνηθισμένα γραμμικά μοντέλα θέλει προσοχή, καθώς τα υπόλοιπα δεν έχουν κοινή διακύμανση (όπως συζητήθηκε εκτενέστερα προηγουμένως).

Υπολογίζουμε το ED_{50} χρησιμοποιώντας και τα δυο μοντέλα για το Linear predictor:

- Για το μοντελο $b_0 + b_1 x$, ο $\widehat{ED}_{50}$ δίνεται από την εξίσωση:

$$\widehat{ED}_{50} = \frac{b_0}{b_1} = 0.277\ γραμμάρια\ ανά\ 100cc$$

- Για το μοντέλο $b_0' + b_1' \ln x$, το $\widehat{ED}_{50}$ δίνεται από την εξίσωση

$$\widehat{ED}_{50} = e^{-1.42} = 0.242\ \text{γραμμάρια ανά}\ 100cc$$

Δύναμη της εφαρμογής της

Τα μοντέλα της λογιστικής παλινδρόμησης είναι συχνά αρκετά ακριβή. Μπορούν να χειριστούν συμβολικά και αριθμητικά πεδία εισόδου. Μπορούν επίσης να δώσουν προβλεπόμενες πιθανότητες για όλες τις κατηγορίες-στόχους, έτσι ώστε μία δεύτερη καλύτερη εικασία μπορεί εύκολα να εντοπιστεί. Τα λογιστικά μοντέλα είναι πιο αποτελεσματικά όταν τα μέλη της ομάδας είναι πραγματικά κατηγορηματικά πεδία. Αν τα μέλη της ομάδας βασίζονται στις τιμές ενός πεδίου συνεχούς κλίμακας (για παράδειγμα, υψηλό δείκτη νοημοσύνης (IQ) σε σχέση με χαμηλό IQ), θα πρέπει να κάνουμε χρήση της γραμμικής παλινδρόμησης ώστε να επωφεληθούμε από την υψηλότερη πληροφορία που προσφέρει όλο το φάσμα των τιμών. Τα λογιστικά μοντέλα μπορούν επίσης να εκτελέσουν αυτόματη επιλογή πεδίων, αν και άλλες προσεγγίσεις, όπως τα μοντέλα των δέντρων ή η επιλογή μεταβλητών (Feature selection) μπορεί να το κάνει αυτό πιο γρήγορα για μεγάλα σύνολα δεδομένων. Τέλος, δεδομένου ότι τα λογιστικά μοντέλα είναι καλύτερα κατανοητά από πολλούς αναλυτές και οι επιστήμονες στον τομέα του data mining, μπορούν να χρησιμοποιηθούν από κάποιους ως βασική γραμμή έναντι όποιων άλλων τεχνικών μοντελοποίησης μπορούν να συγκριθούν.

4.3 Δέντρα αποφάσεων (Decision Trees)

Ένα δέντρο απόφασης είναι ένα εργαλείο υποστήριξης αποφάσεων που χρησιμοποιεί ένα δέντρο που μοιάζει με γραφική παράσταση ή το μοντέλο των αποφάσεων και των πιθανών συνεπειών τους, συμπεριλαμβανομένων των αποτελεσμάτων ενός τυχαίου γεγονότος, το κόστος των πόρων και τη χρησιμότητα. Είναι ένας τρόπος για να παρουσιάσουμε έναν αλγόριθμο.

Τα δέντρα αποφάσεων, που χρησιμοποιούνται στη στατιστική, στην εξόρυξη δεδομένων και στη μηχανική μάθηση, χρησιμοποιούν ένα δέντρο απόφασης ως προγνωστικό μοντέλο το οποίο χαρτογραφεί τις παρατηρήσεις σχετικά με ένα αντικείμενο σε συμπεράσματα σχετικά με την τιμή στόχο του αντικειμένου. Περισσότερα περιγραφικά ονόματα για τέτοια μοντέλα δέντρων είναι: δέντρα ταξινόμησης ή δέντρα παλινδρόμησης. Σε αυτές τις δομές δέντρων, τα «φύλλα» αντιπροσωπεύουν ετικέτες και τα «κλαδιά» αποτελούν συνδέσμους χαρακτηριστικών (features) που οδηγούν σε αυτές τις ετικέτες κατηγορίας.

4.3.1 Θεωρητικό Υπόβαθρο

Οι μέθοδοι που βασίζονται στα δέντρα διαμερίζουν το χώρο των χαρακτηριστικών σε ένα σύνολο ορθογωνίων, και στη συνέχεια τοποθετούν ένα απλό μοντέλο (όπως ένα σταθερό) σε κάθε ένα από αυτά. Αν και εννοιολογικά απλό, είναι ωστόσο και πολύ ισχυρό. Αρχικά θα περιγράψουμε πρώτα μια δημοφιλή μέθοδο δέντρων: τα δέντρα παλινδρόμησης και ταξινόμησης που ονομάζεται CART, και αργότερα τη συγκρίνουμε με την C4.5, ένα σημαντικό ανταγωνιστή της.

Ας εξετάσουμε ένα πρόβλημα παλινδρόμησης με τη συνεχή μεταβλητή απόκρισης Y και επεξηγηματικές μεταβλητές (inputs) X_1 και X_2, όπου καθεμία λαμβάνει τιμές στο μοναδιαίο διάστημα. Το επάνω αριστερό γράφημα στο Σχήμα 7 δείχνει ένα διαχωρισμό του χώρου χαρακτηριστικών με γραμμές που είναι παράλληλες στους άξονες συντεταγμένων. Σε κάθε διαχωρισμό μπορούμε να μοντελοποιήσουμε το Y με μία διαφορετική σταθερά. Ωστόσο, υπάρχει ένα πρόβλημα: αν και κάθε διαχωριστική γραμμή έχει μια απλή περιγραφή όπως $X_1 = c$, μερικές από τις προκύπτουσες περιοχές είναι περίπλοκο για να περιγραφούν.

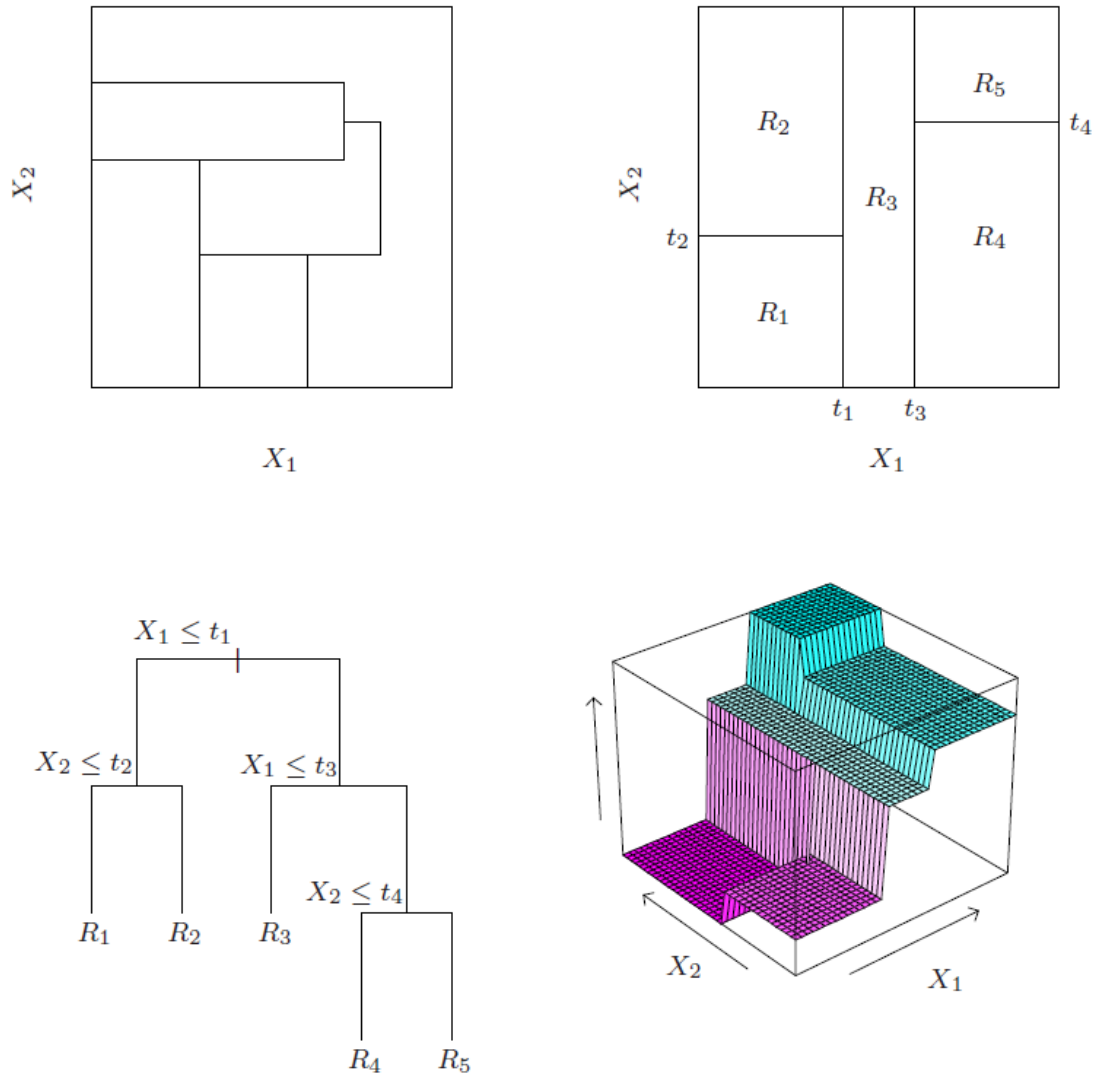
Για λόγους απλούστευσης, θα περιορίσουμε την προσοχή σε αναδρομικά δυαδικά χωρίσματα όπως εκείνα στον άνω δεξί πίνακα του Σχήματος 7. Έχουμε χωρίσει πρώτα το χώρο σε δύο περιοχές, και μοντελοποιήσαμε τη μεταβλητή από τη μέση τιμή του Y σε κάθε μία περιοχή. Επιλέγουμε τη μεταβλητή και το διαχωριστικό-σημείο για να επιτευχθεί η καλύτερη προσαρμογή. Στη συνέχεια, μία ή και οι δύο από αυτές τις περιοχές είναι χωρισμένες σε δύο περισσότερες περιοχές, και αυτή η διαδικασία συνεχίζεται, μέχρις ότου εφαρμοστεί κάποιος κανόνας διακοπής.

Για παράδειγμα, στο πάνω δεξί τμήμα του Σχήματος 7, πρώτα χωρίζουμε σε $X_1 = t_1$. Στη συνέχεια, η περιοχή $X_1 \leq t_1$ χωρίζεται σε $X_2 = t_2$ και η περιοχή $X_1 > t_1$ είναι χωρισμένη σε $X_1 = t_3$. Τέλος, η περιοχή $X_1 > t_3$ χωρίζεται σε $X_2 \leq t_4$.

Το αποτέλεσμα αυτής της διαδικασίας είναι μία διαμέριση στις πέντε περιοχές $R_1, R_2, \ldots, R_5$ όπως φαίνεται στο ακόλουθο σχήμα. Το αντίστοιχο μοντέλο παλινδρόμησης προβλέπει την Y με μία σταθερά c_m στην περιοχή R_m δηλαδή,

$$\hat{f}(X) = \sum_{m=1}^{5} c_m \, I\{(X_1, X_2) \in R_m\}$$

Αυτό το ίδιο μοντέλο μπορεί να παρασταθεί με το δυαδικό δέντρο στο κάτω αριστερά γράφημα του Σχήματος 7.



Σχήμα 7: Χωρίσματα και CART. Το πάνω δεξιά γράφημα δείχνει μία διαμέριση ενός δισδιάστατου χώρου χαρακτηριστικών με αναδρομική δυαδική διάσπαση, όπως χρησιμοποιείται στο CART, που εφαρμόζεται σε ορισμένα ψευδή στοιχεία. Ο επάνω αριστερά πίνακας δείχνει μια γενική διαμέριση που δεν μπορεί να ληφθεί από αναδρομική δυαδική διάσπαση. Στον κάτω αριστερό πίνακα φαίνεται το αντίστοιχο δέντρο της διαμέρισης στο πάνω δεξιά πλαίσιο, και στον κάτω δεξιά πίνακα εμφανίζεται ένα γράφημα με προοπτική της προβλεπόμενης επιφάνειας.

Το σύνολο των δεδομένων βρίσκεται στην κορυφή του δέντρου. Οι παρατηρήσεις που ικανοποιούν τη συνθήκη σε κάθε κόμβο έχουν εκχωρηθεί στον αριστερό κλάδο, και οι άλλες προς τη δεξιά διακλάδωση. Οι τερματικοί κόμβοι ή τα φύλλα του δένδρου αντιστοιχούν στις περιοχές $R_1, R_2, \ldots, R_5$. Το κάτω δεξιά γράφημα του Σχήματος 7 είναι ένα διάγραμμα προοπτικής της επιφάνειας παλινδρόμησης από αυτό το μοντέλο. Για παράδειγμα, εμείς επιλέξαμε τους κόμβους $c_1 = -5$, $c_2 = -7$, $c_3 = 0$, $c_4 = 2$, $c_5 = 4$ για να δημιουργήσουμε αυτό το γράφημα.

Ένα βασικό πλεονέκτημα του επαναληπτικού δυαδικού δέντρου είναι η επεξηγηματικότητά του. Ο διαχωρισμός του χώρου των χαρακτηριστικών περιγράφεται πλήρως από ένα δέντρο. Με περισσότερες από δύο εισόδους, χωρίσματα, όπως ότι στην πάνω δεξιά πλευρά του Σχήματος 7 είναι δύσκολο να σχεδιαστούν, αλλά η δυαδική αναπαράσταση δέντρων λειτουργεί με τον ίδιο τρόπο. Αυτή η παράσταση είναι επίσης δημοφιλής μεταξύ των ιατρικών επιστημόνων, ίσως επειδή μιμείται τον τρόπο με τον οποίο σκέφτεται ένας γιατρός. Το δέντρο στρωματοποιεί τον πληθυσμό σε στρώματα υψηλής και χαμηλής έκβασης (outcome), στη βάση των χαρακτηριστικών των ασθενών.

4.3.2 Δέντρα παλινδρόμησης

Ας στραφούμε τώρα στο ζήτημα του πώς θα δημιουργηθεί ένα δέντρο παλινδρόμησης. Τα δεδομένα μας αποτελούνται από p εισόδους και μια απόκριση, για καθεμία από τις N παρατηρήσεις: δηλαδή, (x_i, y_i) για $i = 1, 2, \ldots, N$, με $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$. Ο αλγόριθμος θα πρέπει να αποφασίσει αυτόματα σχετικά με το διαχωρισμό των μεταβλητών και τα διαχωριστικά σημεία, και επίσης την τοπολογία (σχήμα) που θα πρέπει να έχει το δέντρο. Ας υποθέσουμε πρώτα ότι έχουν ένα διαχωρισμό σε M περιοχές $R_1, R_2, \ldots, R_M$ και μοντελοποιούμε την απόκριση ως μια σταθερά c_m σε κάθε περιοχή:

$$f(x) = \sum_{m=1}^{M} c_m \, I\{x \in R_m\}$$

Εάν λάβουμε ως κριτήριο την ελαχιστοποίηση του αθροίσματος τετραγώνων

$$\sum (y_i - f(x_i))^2$$

είναι εύκολο να δούμε ότι η καλύτερη $\widehat{c_m}$ είναι ακριβώς ο μέσος όρος των y_i στην περιοχή R_m:

$$\widehat{c_m} = ave(y_i | \, x_i \in R_m)$$

Τώρα, η εύρεση της καλύτερης δυαδικής διαμέρισης, όσον αφορά το άθροισμα των ελαχίστων τετραγώνων, είναι γενικά υπολογιστικά ανέφικτη. Ως εκ τούτου προχωράμε με ένα άπληστο αλγόριθμο. Ξεκινώντας με όλα τα δεδομένα, εξετάζουμε μια *διασπασμένη μεταβλητή*[2] j και το σημείο διαχωρισμού s, και καθορίζουμε το ζεύγος των ημι-επιπέδων

[2] Η διάσπαση μιας μεταβλητής είναι μια μέθοδος αποσύνθεσης που χαλαρώνει ένα σύνολο περιορισμών. Όταν η μεταβλητή x εμφανίζεται σε δύο σύνολα των περιορισμών, είναι δυνατό να υποκαταστήσει τις νέες μεταβλητές x_1 στον πρώτο περιορισμό και x_2 στο δεύτερο, και στη συνέχεια να ενωθούν οι δύο μεταβλητές σε μια με ένα "συνδετικό" περιορισμό, ο οποίος απαιτεί $x_1 = x_2$. Αυτή η νέα σύνδεση-περιορισμός μπορούν να

$$R_1(j,s) = \{X|X_j \leq s\} \text{ και } R_2(j,s) = \{X|X_j > s\}$$

Στη συνέχεια αναζητούμε τη διασπασμένη μεταβλητή (splitting variable) j και το σημείο διαχωρισμού (split point) s που επιλύουν:

$$\min_{j,s}\left[\min_{c_1}\sum_{x_i\in R_1(j,s)}(y_i - c_1)^2 + \min_{c_2}\sum_{x_i\in R_2(j,s)}(y_i - c_2)^2\right]$$

Για κάθε επιλογή j και s, το εσωτερικό ελαχιστοποίησης επιλύεται με

$$\widehat{c_1} = ave(y_i|\, x_i \in R_1(j,s)) \text{ και } \widehat{c_2} = ave(y_i|\, x_i \in R_2(j,s))$$

Για κάθε διασπασμένη μεταβλητή, ο προσδιορισμός του σημείου διαχωρισμού μπορεί να γίνει πολύ γρήγορα και ως εκ τούτου σαρώνοντας μέσω όλων των εισόδων, είναι εφικτός ο καθορισμός του βέλτιστου ζεύγους (j,s).

Έχοντας βρει τον καλύτερο διαχωρισμό, διαμερίζουμε τα δεδομένα στις δύο περιοχές του αποτελέσματος και επαναλαμβάνουμε τη διαδικασία διαχωρισμού για κάθε μία από τις δύο περιοχές. Τότε αυτή η διαδικασία επαναλαμβάνεται σε όλες τις περιοχές που προκύπτουν.

Πόσο μεγάλο θα έπρεπε να κάνουμε το δέντρο; Είναι σαφές ότι ένα πολύ μεγάλο δέντρο μπορεί να κάνει υπερεκτίμηση των δεδομένων (overfit), ενώ ένα μικρό δέντρο μπορεί να μην συλλάβει τη σημαντική δομή. Το μέγεθος του δέντρου είναι μία ρυθμιστική παράμετρος που διέπει την πολυπλοκότητα του μοντέλου, και το βέλτιστο μέγεθος του δέντρου πρέπει να επιλέγεται προσαρμοστικά από τα δεδομένα. Μία προσέγγιση θα μπορούσε να ήταν η διαίρεση των κόμβων του δέντρου μόνο εάν η μείωση του αθροίσματος των τετραγώνων λόγω της διάσπασης υπερβαίνει κάποιο όριο (threshold). Η στρατηγική αυτή είναι κοντόφθαλμη, ωστόσο, μια φαινομενικά άχρηστη διάσπαση μπορεί να οδηγήσει σε πολύ καλή διάσπαση στη συνέχεια.

Η προτιμώμενη στρατηγική είναι να αναπτυχθεί ένα μεγάλο δέντρο έστω T_0, σταματώντας την διαδικασία διάσπασης μόνο όταν επιτευχθεί κάποιο ελάχιστο μέγεθος κόμβων (ας πούμε 5). Στη συνέχεια, αυτό το μεγάλο δέντρο «κλαδεύεται» χρησιμοποιώντας κλάδεμα του κόστους-πολυπλοκότητας (cost-complexity prunining), τα οποία θα περιγράψουμε στη συνέχεια.

χαλαρώσουν με ένα πολλαπλασιαστή Lagrange. Για πολλά προβλήματα, όταν η ισότητα των μεταβλητών διάσπασης είναι χαλαρή, τότε το σύστημα αποσυντίθεται, και κάθε υποσύστημα μπορεί να λυθεί ανεξάρτητα, με σημαντική μείωση του χρόνου υπολογισμού και αποθήκευσης μνήμης.

Ορίζουμε ένα υπόδεντρο $T \subset T_0$ να είναι οποιοδήποτε δέντρο που μπορεί να ληφθεί με κλάδεμα του T_0, δηλαδή, συμπτύσσοντας οποιοδήποτε αριθμό των εσωτερικών του (μη-τερματικών) κόμβων. Συμβολίζουμε τους τερματικούς κόμβους με το δείκτη m, με τον κόμβο m να αντιπροσωπεύει την περιοχή R_m. Έστω $|T|$ χαρακτηρίζει τον αριθμό των τερματικών κόμβων στο T. Θέτουμε:

$$\widehat{c_m} = \frac{1}{N_m} \sum_{x_i \in R_m} y_i$$

$$Q_m(T) = \frac{1}{N_m} \sum_{x_i \in R_m} (y_i - \widehat{c_m})^2$$

Ορίζουμε το κριτήριο του κόστους περιπλοκότητας (cost complexity criterion)

$$C_a(T) = \sum_{m=1}^{|T|} N_m \, Q_m(T) \; + \; a|T|$$

Η ιδέα είναι να βρούμε, για κάθε α, το υπόδεντρο $T_\alpha \subseteq T_0$ για την ελαχιστοποίηση του $C_a(T)$. Η ρυθμιστική παράμετρος $\alpha \geq 0$ διέπει το δίλημμα (tradeoff) μεταξύ του μεγέθους του δέντρου και της καλής προσαρμογής του στα δεδομένα. Μεγάλες τιμές του α έχουν ως αποτέλεσμα μικρότερα δέντρα T_α, και αντίστροφα, για μικρότερες τιμές του α. Όπως υποδηλώνει ο συμβολισμός, με $\alpha = 0$ η λύση είναι το πλήρες δέντρο T_0. Θα συζητήσουμε στη συνέχεια για το πώς θα επιλέξουμε προσαρμοστικά το α.

Για κάθε α μπορεί κάποιος να δείξει ότι υπάρχει ένα μοναδικό μικρότερο υπόδεντρο T_α που ελαχιστοποιεί το $C_a(T)$. Για να βρούμε το T_α χρησιμοποιούμε το κλάδεμα του πιο αδύναμου κρίκου (weakest link pruning): έχουμε διαδοχικά συμπτύξει τον εσωτερικό κόμβο που παράγει τη μικρότερη ανά κόμβο αύξηση στο άθροισμα $\sum_{m=1}^{|T|} N_m \, Q_m(T)$, και συνεχίζουμε μέχρι να παράξουμε το δέντρο με έναν κόμβο (ρίζα). Αυτό δίνει μία (πεπερασμένη) ακολουθία από υπόδενδρα, και μπορεί κανείς να δείξει αυτή η ακολουθία πρέπει να περιέχει το T_α. Δείτε Breiman et al. (1984) ή Ripley (1996) για λεπτομέρειες. Η εκτίμηση του α επιτυγχάνεται με πέντε- ή δέκα- φορές διασταυρωμένη επικύρωση: επιλέγουμε την τιμή $\hat{\alpha}$ που ελαχιστοποιεί το άθροισμα των τετραγώνων της διασταυρωμένης επικύρωσης. Το τελικό μας δέντρο είναι το $T_{\hat{\alpha}}$.

4.3.3 Δέντρα ταξινόμησης

Εάν ο στόχος είναι ένα αποτέλεσμα ταξινόμησης που λαμβάνει τιμές $1, 2, \ldots, K$, οι μόνες αλλαγές που απαιτούνται στον αλγόριθμο του δέντρου αφορούν στα κριτήρια για το διαχωρισμό των κόμβων και στο κλάδεμα του δέντρου. Για την παλινδρόμηση χρησιμοποιήθηκε το τετραγωνικό-σφάλμα του κόμβου που μετράει (impurity measure $Q_m(T)$) το $Q_m(T)$ που ορίστηκε προηγουμένως, αλλά αυτό δεν είναι κατάλληλο για την ταξινόμηση. Σε έναν κόμβο m που αντιπροσωπεύει την περιοχή R_m με N_m παρατηρήσεις, έστω

$$\hat{p}_{mk} = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i = k) ,$$

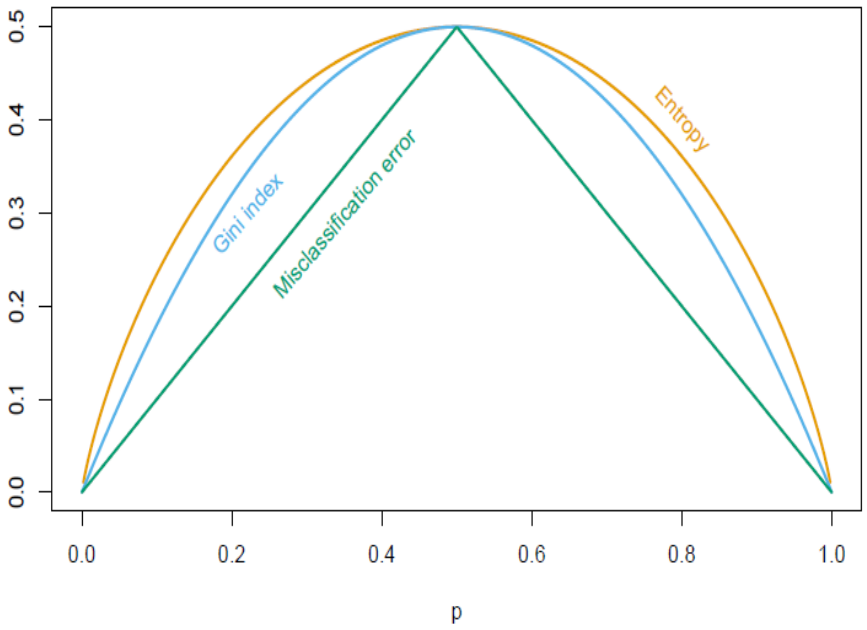
το ποσοστό της τάξης, με k παρατηρήσεις στον κόμβο m. Ταξινομούμε τις παρατηρήσεις στον κόμβο m στην κατηγορία $k(m) = argmax_k \hat{p}_{mk}$, την επικρατέστερη κλάση στον κόμβο m. Διαφορετικά μέτρα $Q_m(T)$ του (impurity) κόμβου περιλαμβάνουν τα ακόλουθα:

Σφάλμα εσφαλμένης ταξινόμησης: (*Misclassification error*)	$\frac{1}{N_m}\sum_{x_i \in R_m} I(y_i \neq k(m)) = 1 - \hat{p}_{mk}(m).$
Δείκτης Gini: (*Gini index*)	$\sum_{k \neq k'} \hat{p}_{mk}\hat{p}_{mk'} = \sum_{k=1}^{K} \hat{p}_{mk}\,(1 - \hat{p}_{mk}).$
Διασταυρωμένη εντροπία ή απόκλιση: (*Cross-entropy or deviance*)	$-\sum_{k=1}^{K} \hat{p}_{mk}\, log\, \hat{p}_{mk}.$

Για δύο τάξεις, εάν το p είναι η αναλογία στη δεύτερη κατηγορία, αυτά τα τρία μέτρα είναι $1 - max(p, 1 - p)$, $2p(1 - p)$ και $-plog\, p - (1 - p)\, log\, (1 - p)$, αντιστοίχως. Αυτά φαίνονται στο Σχήμα 8. Και τα τρία είναι παρόμοια, αλλά η διασταυρωμένη εντροπία και ο δείκτης Gini είναι διαφορίσιμα, και ως εκ τούτου, πιο δεκτικά σε αριθμητική βελτιστοποίηση.

Επιπλέον, η διασταυρωμένη εντροπία και ο δείκτης Gini είναι πιο ευαίσθητα στις μεταβολές πιθανότητας του κόμβου από ότι στο ποσοστό ταξινόμησης. Για παράδειγμα, σ' ένα πρόβλημα δύο-κλάσεων με 400 παρατηρήσεις σε κάθε κατηγορία (αυτό δηλώνεται με (400, 400)), ας υποθέσουμε ότι η μία διάσπαση δημιούργησε κόμβους (300, 100) και (100, 300), ενώ η άλλη δημιούργησε κόμβους (200, 400) και (200, 0). Και οι δύο διασπάσεις παράγουν ένα ποσοστό μη ταξινόμησης της τάξεως του 0.25, αλλά η δεύτερη διάσπαση παράγει ένα

σκέτο κόμβο (pure node) και είναι πιθανώς προτιμότερη. Και ο δείκτης Gini αλλά και η διασταυρωμένη-εντροπία είναι χαμηλότερα για τη δεύτερη διάσπαση.



Σχήμα 8: Τα μέτρα προσμίξεων του κόμβου (node impurity measures) για την ταξινόμηση δύο τάξεων, ως συνάρτηση του ποσοστού p στην τάξη 2. Η διασταυρωμένη-εντροπία έχει κλιμακωθεί για να περάσει από το (0.5, 0.5).

Για το λόγο αυτό, όταν αυξάνεται το δέντρο πρέπει να χρησιμοποιούνται είτε ο δείκτης Gini είτε η διασταυρωμένη εντροπία. Για να καθοδηγήσουν το κλάδεμα του κόστους-πολυπλοκότητας (cost-complexity pruning), οποιαδήποτε από τα τρία μέτρα μπορούν να χρησιμοποιηθούν, αλλά τυπικά θα είναι το ποσοστό μη ταξινόμησης.

Ο δείκτης Gini μπορεί να ερμηνευθεί με δύο ενδιαφέροντες τρόπους. Αντί να ταξινομούμε τις παρατηρήσεις στην επικρατέστερη τάξη (τάξη πλειοψηφίας) στον κόμβο, θα μπορούσαμε να τις κατατάξουμε στην τάξη k με πιθανότητα $\hat{p}_{mk}$. Στη συνέχεια, το ποσοστό σφάλματος εκπαίδευσης αυτού του κανόνα στον κόμβο είναι $\sum_{k \neq k'} \hat{p}_{mk}\hat{p}_{mk'}$ (δείκτης Gini). Ομοίως, αν κωδικοποιήσουμε κάθε παρατήρηση ως 1 για την κλάση k και μηδέν διαφορετικά, η διακύμανση πάνω στον κόμβο αυτής της 0-1 απόκρισης είναι

$$\hat{p}_{mk}\,(1 - \hat{p}_{mk})$$

Συνοψίζοντας πάνω στις k κλάσεις πάλι παίρνουμε το δείκτη Gini.

Στην παρούσα διπλωματική εργασία θα εστιάσουμε σε 4 ευρέως διαδεδομένους αλγορίθμους δέντρων απόφασης: τα δέντρα ταξινόμησης και παλινδρόμησης (**C&RT**) (βλ. Breiman et al. (1984)), **C5.0** (βλ. Quinlan(1993)) , Chi-square Automatic Interaction Detection (**CHAID**) (βλ. Kass(1980)), και Quick, Unbiased, Efficient Statistical Tree (**QUEST**) (βλ. Loh και Shih(1997)).

4.3.4 Άλλα θέματα

Κατηγορική Πρόβλεψη

Όταν διαχωρίζουμε μία πρόβλεψη που έχει q δυνατές μη ταξινομημένες τιμές, υπάρχουν $2^{q-1}-1$ δυνατά χωρίσματα των q τιμών σε δύο ομάδες, και οι υπολογισμοί γίνονται απαγορευτικοί για μεγάλο q. Ωστόσο, με μία έκβαση (outcome) $0-1$, ο υπολογισμός αυτός απλοποιείται. Έτσι, διατάσσουμε τις τάξεις πρόβλεψης σύμφωνα με το ποσοστό που υπάγεται στην κλάση έκβασης 1. Στη συνέχεια, χωρίζουμε το ποσοστό σαν να ήταν μια διατεταγμένη πρόβλεψη. Κάποιος μπορεί να δείξει ότι αυτό δίνει τη βέλτιστη διάσπαση, από την άποψη της διασταυρωμένης-εντροπίας ή του δείκτη Gini, μεταξύ όλων δυνατών $2^{q-1}-1$ διασπάσεων. Αυτό το αποτέλεσμα ισχύει και για ένα ποσοτικό αποτέλεσμα και την απώλεια του τετραγωνικού σφάλματος- οι κατηγορίες ταξινομούνται με την αύξηση του μέσου του αποτελέσματος. Αν και διαισθητικές, οι αποδείξεις αυτών των ισχυρισμών δεν είναι ασήμαντες. Η απόδειξη για τη δυαδική έκβαση δίνεται στο Breiman et al. (1984) και Ripley (1996). Η απόδειξη για την ποσοτική περίπτωση μπορεί να βρεθεί στον Fisher (1958). Για τα αποτελέσματα για διάφορες κατηγορίες (multicategory), δεν είναι δυνατές τέτοιες απλουστεύσεις, αν και έχουν προταθεί διάφορες προσεγγίσεις (Loh και Vanichsetakul, 1988).

Ο αλγόριθμος κατάτμησης τείνει να ευνοεί την κατηγορηματική πρόβλεψη με πολλά επίπεδα q, ο αριθμός των χωρισμάτων αυξάνεται εκθετικά στο q, και όσο περισσότερες επιλογές έχουμε, τόσο περισσότερες πιθανότητες έχουμε να βρούμε ένα καλό αποτέλεσμα για τα δεδομένα με το χέρι. Αυτό μπορεί να οδηγήσει σε σοβαρή υπερπροσαρμογή (overfitting) εάν το q είναι μεγάλο, και τέτοιες μεταβλητές θα πρέπει να αποφεύγονται.

Άλλες Διαδικασίες δημιουργίας δέντρων

Η προηγούμενη συζήτηση επικεντρώνεται κυρίως στην εφαρμογή των δέντρων CART (Δέντρο ταξινόμησης και παλινδρόμησης). Η άλλη δημοφιλής μέθοδος είναι η ID3 και οι νεότερες εκδόσεις της, C4.5 και C5.0 (Quinlan, 1993). Οι πρώτες εκδόσεις του προγράμματος περιορίζονταν σε κατηγορηματική πρόβλεψη, και χρησιμοποιούσαν έναν

top-down κανόνα χωρίς κλάδεμα. Με τις πιο πρόσφατες εξελίξεις, ο C5.0 έχει γίνει αρκετά παρόμοιος με τον CART. Το πιο σημαντικό χαρακτηριστικό μοναδικό στον C5.0 είναι ένα σύστημα για την εξαγωγή συνόλων κανόνων. Μετά από την ανάπτυξη ενός δέντρου, οι κανόνες διάσπασης που καθορίζουν τους τερματικούς κόμβους μπορεί μερικές φορές να απλοποιηθούν: δηλαδή, μία ή περισσότερες συνθήκες μπορεί να απαλειφθούν χωρίς να αλλάζει το υποσύνολο των παρατηρήσεων που εμπίπτουν στον κόμβο. Θα καταλήξουμε σε ένα απλοποιημένο σύνολο κανόνων που καθορίζουν κάθε τερματικό κόμβο. Αυτά δεν ακολουθούν πλέον μια δομή δέντρου, αλλά η απλότητα τους μπορεί να καταστεί πιο ελκυστική για τον χρήστη.

Η αστάθεια των Δένδρων

Ένα σημαντικό πρόβλημα με τα δέντρα είναι η υψηλή διασπορά τους. Συχνά, μια μικρή αλλαγή στα δεδομένα μπορεί να οδηγήσει σε μια πολύ διαφορετική σειρά από διασπάσεις, κάνοντας την ερμηνεία κάπως επισφαλή. Ο κύριος λόγος για αυτή την αστάθεια είναι η ιεραρχική φύση της διαδικασίας: το αποτέλεσμα ενός σφάλματος στην κορυφαία-πρώτη διάσπαση διαδίδεται σε όλες τις επόμενες διασπάσεις κάτω από αυτή. Κάποιος μπορεί να το κατευνάσει αυτό σε κάποιο βαθμό, προσπαθώντας να χρησιμοποιήσει ένα πιο σταθερό κριτήριο διάσπασης, αλλά η εγγενής αστάθεια δεν αφαιρείται. Αυτό είναι το τίμημα που πρέπει να καταβληθεί για την εκτίμηση μιας απλής δομής δέντρου από τα δεδομένα. Το Bagging υπολογίζει κατά μέσο όρο πολλά δέντρα για να μειώσει αυτή τη διασπορά.

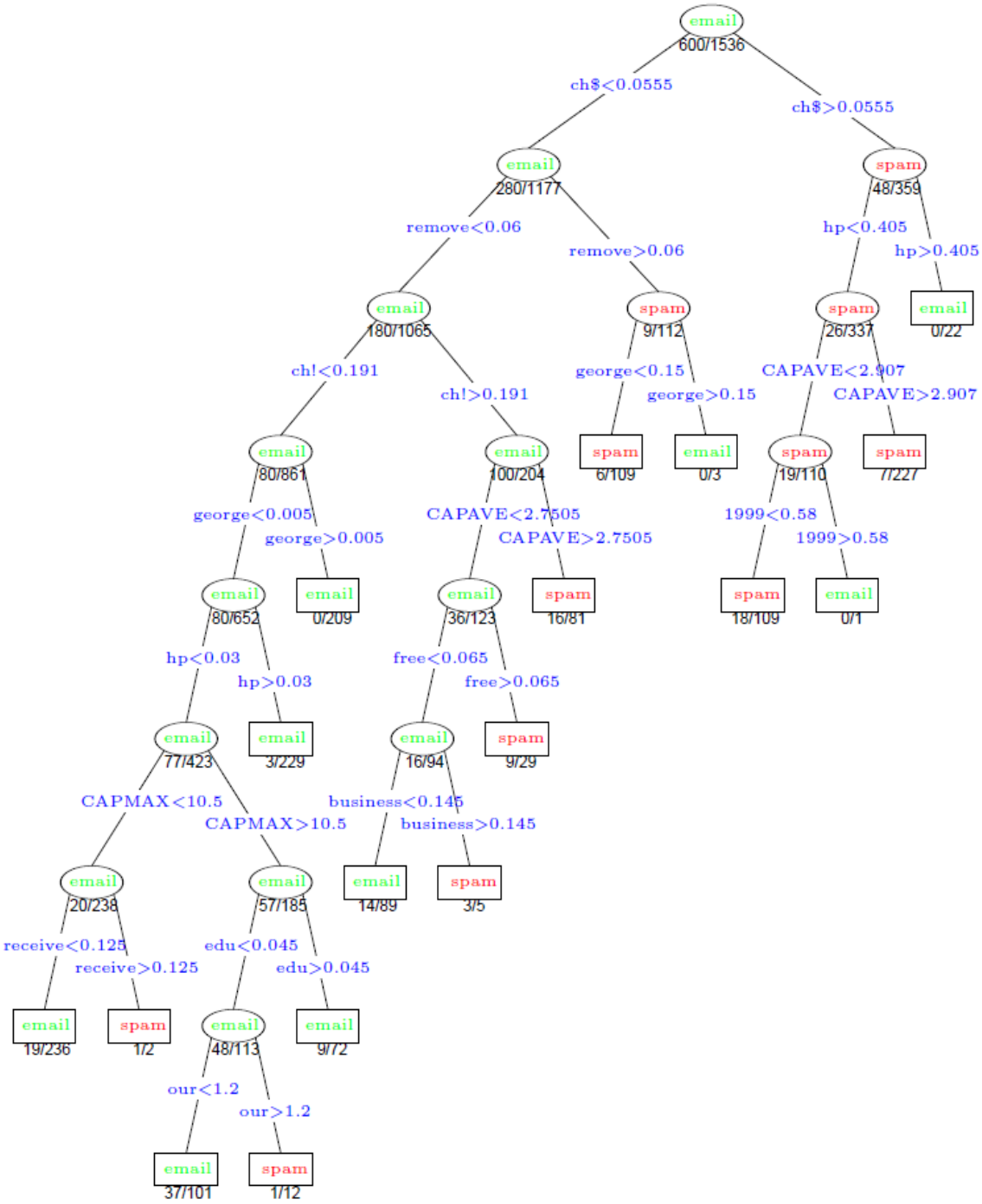
Ακολουθεί ένα γράφημα που δείχνει τη μορφή που έχει μία δομή δέντρου. Το παράδειγμα (spam example) που ακολουθεί εξετάστηκε από τους Hastie et al. (2001).

4.3.5 Αλγόριθμοι Δέντρων αποφάσεων

<u>Αλγόριθμος C&RT</u>

Τα αρχικά C&RT σημαίνουν δέντρα ταξινόμησης και παλινδρόμησης (Classification and Regression Trees), βασικά περιγράφονται στο βιβλίο του με το ίδιο όνομα (Breiman, Friedman, Olshen, και Stone, 1984). Ο C&RT χωρίζει τα δεδομένα σε δύο υποσύνολα έτσι ώστε οι εγγραφές εντός του κάθε υποσυνόλου να είναι περισσότερο ομοιογενείς από ό,τι στην προηγούμενη υποομάδα. Είναι μια επαναληπτική διαδικασία - καθένα από αυτά τα δύο υποσύνολα κατόπιν διαιρείται (split) και πάλι, και η διαδικασία επαναλαμβάνεται μέχρι το κριτήριο της ομοιογένειας ή μέχρι κάποιο άλλο κριτήριο διακοπής να ικανοποιείται (όπως κάνουν όλες οι μέθοδοι των δέντρων). Το ίδιο πεδίο πρόβλεψης μπορεί να χρησιμοποιηθεί πολλές φορές σε διαφορετικά επίπεδα στο δέντρο. Χρησιμοποιεί

διαχωρισμό (ή διάσπαση/split) με υποκατάστατα για να γίνει η καλύτερη δυνατή χρήση των δεδομένων με τις τιμές που λείπουν.



Σχήμα 9: Το «κλαδεμένο» δέντρο για το παράδειγμα spam. Η διασπασμένη μεταβλητή φαίνεται με μπλε χρώμα στα κλαδιά του δέντρου, και η ταξινόμηση φαίνεται σε κάθε κόμβο. Οι αριθμοί κάτω από τους τερματικούς κόμβους καταδεικνύουν το ποσοστό των μη ταξινομημένων δεδομένων δοκιμών.

Ο C&RT είναι αρκετά ευέλικτος. Επιτρέπει στη διαδικασία δημιουργίας του δένδρου να θεωρήσει άνισο κόστος εσφαλμένης ταξινόμησης. Επίσης, επιτρέπει να οριστεί η μία εκ των προτέρων κατανομή πιθανότητας σε ένα πρόβλημα ταξινόμησης. Μπορούμε να εφαρμόσουμε αυτόματο κλάδεμα του κόστους-πολυπλοκότητας σ'ένα C&RT για να αποκτήσουμε ένα πιο γενικεύσιμο δέντρο.

Παράμετροι του μοντέλου

Ο C&RT λειτουργεί επιλέγοντας μία διάσπαση (split) σε κάθε κόμβο, έτσι ώστε κάθε θυγατρικός κόμβος που δημιουργήθηκε από τη διάσπαση να είναι πιο καθαρός από τον κόμβο γονέα του. Εδώ η καθαρότητα αναφέρεται σε ομοιότητα των τιμών του πεδίου στόχου. Σε έναν εντελώς καθαρό κόμβο, όλες οι εγγραφές έχουν την ίδια τιμή για το πεδίο στόχου. Ο C&RT μετρά την ακαθαρσία-μη καθαρότητα μιας διάσπασης (split) σε έναν κόμβο ορίζοντας ένα μέτρο μη καθαρότητας. Τα ακόλουθα βήματα χρησιμοποιούνται για την κατασκευή ενός C&RT δέντρου (ξεκινώντας από τον κόμβο ρίζα που περιέχει όλες τις εγγραφές) βρίσκει την καλύτερη διάσπαση για κάθε εκτιμητή.

Για κάθε πεδίο πρόβλεψης, βρίσκει την καλύτερη δυνατή διάσπαση, ως ακολούθως:

- ✓ **Πεδία κλίμακας (αριθμητικά)**. Ταξινομεί τις τιμές των πεδίων για τις εγγραφές στον κόμβο από το μικρότερο στο μεγαλύτερο. Επιλέγει κάθε σημείο με τη σειρά ως ένα σημείο διαχωρισμού (split), και υπολογίzει τη στατιστική μη καθαρότητα για τον προκύπτον θυγατρικό κόμβο της διάσπασης. Επιλέγει το καλύτερο σημείο διαχωρισμού για το πεδίο, ως αυτό που παράγει τη μεγαλύτερη μείωση στη σχέση μη καθαρότητας στην μη καθαρότητα του διασπασμένου κόμβου.

- ✓ **Συμβολικά (κατηγορηματικά) πεδία**. Εξετάζει κάθε δυνατό συνδυασμό των τιμών ως δύο υποσύνολα. Για κάθε συνδυασμό, υπολογίζει την μη καθαρότητα των θυγατρικών κόμβων για τη διάσπαση με βάση αυτό το συνδυασμό. Επιλέγει το καλύτερο σημείο διαχωρισμού για το πεδίο, όπως αυτό που παράγει τη μεγαλύτερη μείωση στη σχέση της μη καθαρότητας με την μη καθαρότητα του κόμβου διάσπασης.

Κανόνες Διακοπής

Κανόνες ελέγχου που καθορίζουν τον τρόπο με τον οποίο ο αλγόριθμος αποφασίζει πότε να σταματήσει τον διαχωρισμό των κόμβων στο δέντρο. Η ανάπτυξη του δέντρου προχωρά μέχρι κάθε κόμβος-φύλλο στο δέντρο να προκαλέσει τουλάχιστον έναν κανόνα διακοπής.

Οποιαδήποτε από τις ακόλουθες προϋποθέσεις θα αποτρέψει έναν κόμβο από την κατάτμηση:

- Ο κόμβος είναι καθαρός (όλες οι εγγραφές έχουν την ίδια τιμή για το πεδίο-στόχο)
- Όλες οι εγγραφές στον κόμβο έχουν την ίδια τιμή για όλα τα πεδία πρόβλεψης που χρησιμοποιούνται από το μοντέλο.
- Το βάθος του δέντρο για τον τρέχοντα κόμβο (ο αριθμός των αναδρομικών διασπάσεων του κόμβου που ορίζουν τον τρέχοντα κόμβο) είναι το μέγιστο βάθος του δέντρου (προεπιλογή ή καθορίζεται από το χρήστη).
- Ο αριθμός των εγγραφών στον κόμβο είναι μικρότερος από το ελάχιστο μέγεθος του γονεακού κόμβου (προεπιλογή ή καθορίζεται από το χρήστη).
- Ο αριθμός των εγγραφών σε οποιονδήποτε από τους θυγατρικούς κόμβους που προκύπτουν από καλύτερη διάσπαση (split) του κόμβου είναι μικρότερη από το ελάχιστο μέγεθος του θυγατρικού κόμβου (προεπιλογή ή καθορίζεται από το χρήστη).
- Η καλύτερη διάσπαση (split) για τον κόμβο αποδίδει μία μείωση στην μη καθαρότητα η οποία είναι μικρότερη από την ελάχιστη μεταβολή στην μη καθαρότητα (προεπιλογή ή καθορίζεται από το χρήστη).

Αλγόριθμος CHAID

Τα αρχικά CHAID σημαίνουν Χ-τετράγωνο Αυτόματος ανιχνευτής αλληλεπίδρασης (Chi-squared Automatic Interaction Detector). Είναι μία πολύ αποδοτική στατιστική τεχνική για κατάτμηση, ή δημιουργία ενός δέντρου, που αναπτύχθηκε από τον Kass το 1980. Χρησιμοποιώντας ως κριτήριο της σημαντικότητα ενός στατιστικού τεστ, ο CHAID αξιολογεί όλες τις τιμές ενός δυναμικού πεδίου πρόβλεψης. Συγχωνεύει τις τιμές που κρίνονται να είναι στατιστικά ομογενείς (παρόμοιες) σε σχέση με την μεταβλητή-στόχο και διατηρεί όλες τις άλλες αξίες που είναι ετερογενείς (ανόμοιες). Αυτό στη συνέχεια επιλέγει τον καλύτερο εκτιμητή για να σχηματίσει το πρώτο κλαδί του δέντρου απόφασης, έτσι ώστε κάθε θυγατρικός κόμβος να είναι κατασκευασμένος από μια ομάδα με ομοιογενείς τιμές του επιλεγμένου πεδίου. Αυτή η διαδικασία συνεχίζεται κατ 'επανάληψη μέχρι το δέντρο να έχει αναπτυχθεί πλήρως. Το στατιστικό τεστ που χρησιμοποιήθηκε εξαρτάται από τη μέτρηση

του επιπέδου του πεδίου στόχου. Εάν το πεδίο στόχου είναι συνεχές, χρησιμοποιείται ένα F τεστ. Εάν το πεδίο στόχου είναι κατηγορηματικό, χρησιμοποιείται ένα X^2-τεστ.

Ο CHAID δεν είναι μια δυαδική μέθοδος δέντρου. Δηλαδή, μπορεί να παράγει περισσότερες από δύο κατηγορίες σε οποιοδήποτε συγκεκριμένο επίπεδο στο δέντρο. Ως εκ τούτου, τείνει να δημιουργήσει ένα ευρύτερο δέντρο από ό,τι κάνουν οι δυαδικές μέθοδοι. Λειτουργεί για όλους τους τύπους των μεταβλητών, και δέχεται και τα βάρη και τις μεταβλητές συχνότητας. Χειρίζεται τις τιμές που λείπουν χειρίζοντας τες όλες ως μία έγκυρη κατηγορία.

Εξαντλητικός CHAID

Ο εξαντλητικός CHAID είναι μια τροποποίηση του CHAID που αναπτύχθηκε για να αντιμετωπίσει ορισμένες από τις αδυναμίες της μεθόδου CHAID (Biggs, de Ville, και Suen, 1991). Ειδικότερα, μερικές φορές ο CHAID δεν μπορεί να βρεί τη βέλτιστη διάσπαση (split) για μια μεταβλητή, δεδομένου ότι σταματά τη συγχώνευση των κατηγοριών, μόλις διαπιστώσει ότι όλες οι υπόλοιπες κατηγορίες είναι στατιστικά διαφορετικές. Ο εξαντλητικός CHAID το θεραπεύει αυτό, συνεχίζοντας να συγχωνεύσει τις κατηγορίες της μεταβλητής πρόβλεψης μέχρι να μείνουν μόνο δύο υπερκατηγορίες. Στη συνέχεια εξετάζει τη σειρά των συγχωνεύσεων για τον εκτιμητή και βρίσκει το σύνολο των κατηγοριών που δίνει την ισχυρότερη σχέση με τη μεταβλητή στόχο, και υπολογίζει μια προσαρμοσμένη p-τιμή για την αυτή τη σχέση. Ετσι, ο εξαντλητικός CHAID μπορεί να βρει την καλύτερη διάσπαση για κάθε εκτιμητή, και στη συνέχεια να επιλέξει ποιος εκτιμητής θα διασπαστεί συγκρίνοντας με τις προσαρμοσμένες p-τιμές. Ο εξαντλητικός CHAID είναι ταυτόσημος με τον CHAID στις στατιστικές δοκιμές που χρησιμοποιεί και τον τρόπο με τον οποίο μεταχειρίζεται τις ελλειπούσες τιμές. Επειδή η μέθοδος του συνδυασμού κατηγοριών των μεταβλητών είναι πιο εμπεριστατωμένη από ότι στον CHAID, χρειάζεται περισσότερος χρόνος για να υπολογιστεί. Ωστόσο, αν έχετε να διαθέσετε τον αντίστοιχο χρόνο, ο εξαντλητικός CHAID είναι γενικά ασφαλέστερος στη χρήση από τον CHAID. Βρίσκει συχνά πιο χρήσιμες διασπάσεις, αν και εξαρτάται από τα δεδομένα. Αξίζει να παρατηρήσουμε ότι μπορεί να μην βρούμε καμία διαφορά μεταξύ των αποτελεσμάτων του εξαντλητικού αλγορίθμου CHAID και του CHAID.

Παράμετροι του μοντέλου

Ο CHAID δουλεύει με όλους τους τύπους των συνεχών ή κατηγορηματική πεδίων. Ωστόσο, τα συνεχή προγνωστικά πεδία κατηγοριοποιούνται αυτόματα για το σκοπό της ανάλυσης.

Αλγόριθμος QUEST

Τα αρχικά QUEST σημαίνουν γρήγορο, αμερόληπτο, αποτελεσματικό στατιστικό δέντρο (Quick, Unbiased, Efficient Statistical Tree). Είναι ένας σχετικά νέος δυαδικός αλγόριθμος δέντρου (Loh και Shih, 1997). Ασχολείται χωριστά με την επιλογή του πεδίου διάσπασης και την επιλογή του split-σημείου. Η μονοπαραγοντική διάσπαση στον QUEST εκτελεί περίπου αμερόληπτα επιλογή πεδίων. Δηλαδή, αν όλα τα πεδία πρόβλεψης είναι εξίσου κατατοπιστικά με βάση το πεδίο στόχου, ο QUEST επιλέγει οποιοδήποτε από τα πεδία πρόβλεψης με ίσες πιθανότητες. Ο QUEST προσφέρει πολλά από τα πλεονεκτήματα του C&RT, αλλά, όπως και ο C&RT, τα δέντρα μπορεί να προκύψουν δυσκίνητα. Μπορούμε να εφαρμόσουμε αυτόματο κλάδεμα του κόστους-πολυπλοκότητας σε ένα δέντρο QUEST για να μειώσουμε το μέγεθός του. Ο QUEST χρησιμοποιεί υποκατάστατη διάσπαση για να χειριστεί τις ελλειπούσες τιμές.

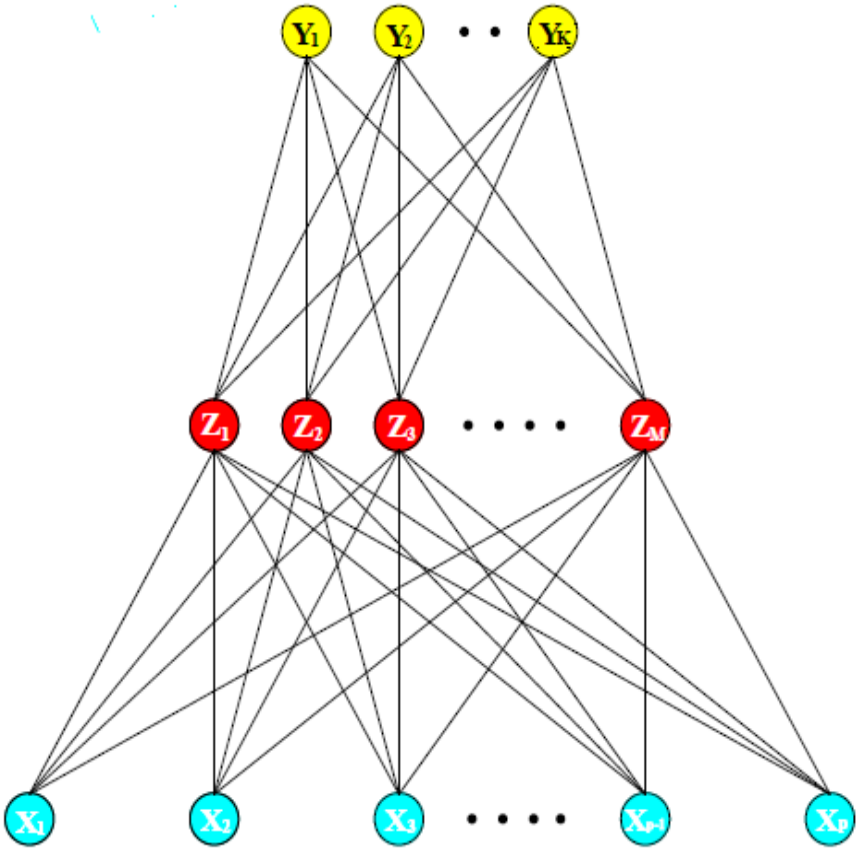
4.4 Τεχνητά Νευρωνικά Δίκτυα (Neural Networks)

4.4.1 Εισαγωγή

Τα νευρωνικά δίκτυα, όπως και άλλες μέθοδοι που έχουμε αναπτύξει, ανήκουν στην κατηγορία μεθόδων μάθησης που αναπτύχθηκε χωριστά σε διάφορους τομείς -στατιστική και τεχνητή νοημοσύνη- και είναι, βασικά, βασισμένη σε ταυτόσημα μοντέλα. Η κεντρική ιδέα είναι να εξάγουμε γραμμικούς συνδυασμούς των εισροών, όπως προκύπτουν τα χαρακτηριστικά, και στη συνέχεια να μοντελοποιήσουμε το στόχο ως μια μη γραμμική συνάρτηση αυτών των χαρακτηριστικών. Το αποτέλεσμα είναι μία ισχυρή μέθοδος μάθησης, με διαδεδομένες εφαρμογές σε πολλούς τομείς.

Ο όρος νευρωνικό δίκτυο έχει εξελιχθεί ώστε να περιλαμβάνει μια μεγάλη κλάση των μοντέλων και μεθόδων μάθησης. Εδώ περιγράφουμε το πιο ευρέως χρησιμοποιούμενο "vanilla" νευρικό δίκτυο, που ονομάζεται μερικές φορές «δίκτυο μονού κρυφού στρώματος πίσω-διάδοσης», ή «μονό perceptron στρώμα». Υπήρξε μια μεγάλη δημοσιότητα γύρω από τα νευρωνικά δίκτυα, κάνοντας τα να φαίνονται μαγικά και μυστηριώδη. Όπως έχουμε κάνει σαφές σε αυτή την παράγραφο, είναι απλά μη γραμμικά στατιστικά μοντέλα, ακριβώς όπως το μοντέλο παλινδρόμησης projection pursuit (βλ. Hastie (2001)).

Ένα νευρωνικό δίκτυο είναι μια δύο-σταδίων διαδικασία παλινδρόμησης ή ένα μοντέλο ταξινόμησης, που συνήθως αντιπροσωπεύεται από ένα *διάγραμμα δικτύου* όπως στο ακόλουθο σχήμα.



Σχήμα 10: Σχηματική αναπαράσταση ενός ενιαίου κρυμμένο στρώμα, feedforward νευρωνικό δίκτυο.

Αυτό το δίκτυο εφαρμόζεται τόσο για την παλινδρόμηση όσο και για την ταξινόμηση.

Για την παλινδρόμηση, τυπικά το $K = 1$ και υπάρχει μόνο μία μονάδα εξόδου Y_1 στην κορυφή. Ωστόσο, αυτά τα δίκτυα μπορούν να χειριστούν πολλαπλές ποσοτικές αποκρίσεις σε μία ενιαία τάση, γι 'αυτό θα ασχοληθούμε με την γενική περίπτωση.

Για την ταξινόμηση της K-τάξης υπάρχουν K μονάδες στην κορυφή, με την k-οστή μονάδα να μοντελοποιεί την πιθανότητα της τάξης k. Υπάρχουν K μετρήσεις στόχου Y_κ, $k = 1,\ldots,K$, καθεμία από τις οποίες κωδικοποιούνται ως μία 0 - 1 μεταβλητή για την k-οστή κλάση.

Τα χαρακτηριστικά Z_m που προκύπτουν, δημιουργούνται από γραμμικούς συνδυασμούς των εισροών, και στη συνέχεια ο στόχος Y_κ μοντελοποιείται ως μια συνάρτηση γραμμικών συνδυασμών των Z_m,

$$Z_m = \sigma(\alpha_{0m} + \alpha^T{}_m X), \;\; m = 1, \dots, M$$

$$T_k = \beta_{0k} + \beta^T{}_k Z, \;\; k = 1, \dots, K \qquad (4.4.1)$$

$$f_k(X) = g_k(T), \;\; k = 1, \dots, K$$

Όπου $Z = (Z_1, Z_2, \dots, Z_M)$, και $T = (T_1, T_2, \dots, T_K)$

Η συνάρτηση ενεργοποίησης $\sigma(v)$ επιλέγεται συνήθως να είναι η σιγμοειδής

$$\sigma(v) = \frac{1}{(1 + e^{-v})}$$

Στο σχήμα 11 φαίνεται το γράφημα του της συνάρτησης $1/(1 + e^{-v})$.

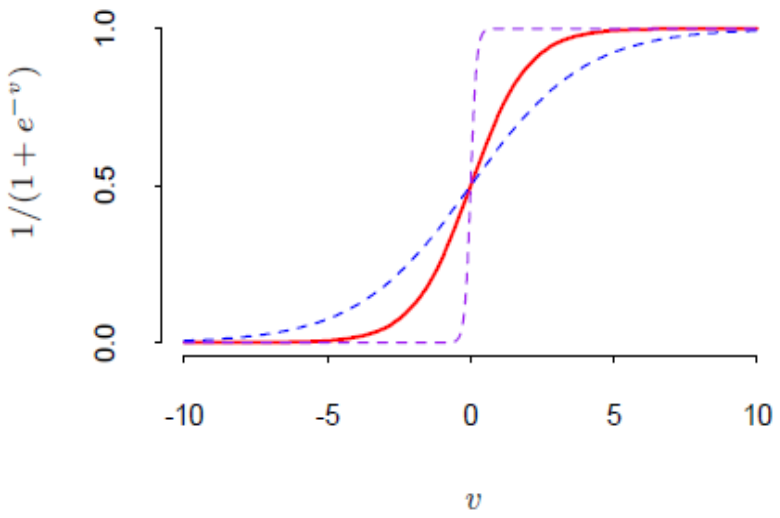
Μερικές φορές χρησιμοποιείται η συνάρτηση Gaussian radial basis (βλ. Hastie et al.(2001)) για την $\sigma(v)$, που παράγει αυτό που είναι γνωστό ως *radial basis function network*. Τα διαγράμματα των νευρωνικών δικτύων, όπως σχήμα 10 είναι μερικές φορές σχεδιασμένα με μία επιπλέον μονάδα *μεροληψίας* που βρίσκεται μέσα σε κάθε μονάδα των κρυφών στρωμάτων και των στρωμάτων εξόδου.

Σκεφτείτε τη σταθερά "1" ως ένα πρόσθετο χαρακτηριστικό εισόδου, αυτή η μεροληπτική μονάδα αιχμαλωτίζει τους όρους α_{0m} και β_{0k} στο μοντέλο (4.4.1).

Η συνάρτηση εξόδου $g_k(T)$ επιτρέπει μια τελική μετατροπή του διανύσματος T των εξόδων. Για την παλινδρόμηση επιλέγουμε συνήθως την ταυτοτική συνάρτηση $g_k(T) = T_k$. Η πρόωρη εργασία στην K-κλάση ταξινόμησης χρησιμοποιεί και πάλι αυτή τη συνάρτηση, αλλά αυτό αργότερα εγκαταλείφθηκε υπέρ της συνάρτησης Softmax

$$g_k(T) = \frac{e^{T_k}}{\sum_{l=1}^{K} e^{T_l}}$$

Αυτός είναι φυσικά ακριβώς ο μετασχηματισμός που χρησιμοποιείται στο μοντέλο multilogit (βλ. Hastie et al. (2001)), και παράγει θετικές εκτιμήσεις που αθροίζονται σ' ένα ποσό.



Σχήμα 11: Γράφημα της σιγμοειδούς συνάρτησης σ(v)=1/((1+exp(-v)) **(**κόκκινη-συμπαγής καμπύλη), που χρησιμοποιείται συνήθως στο κρυφό στρώμα ενός νευρωνικού δικτύου. Περιλαμβάνονται οι σ (sv) για s = ½ (μπλε-διακεκομμένη καμπύλη) και s = 10 (μωβ- διακεκομμένη καμπύλη). Η παράμετρος κλίμακας s ελέγχει το ποσοστό ενεργοποίησης, και μπορούμε να δούμε ότι μεγάλες ποσότητες s σε μία δύσκολη ενεργοποίηση σε v = 0. Σημειώνουμε ότι το σ(s (v - v_0)) μετατοπίζει το όριο ενεργοποίησης από 0 έως v_0.

Οι μονάδες στο μέσο του δικτύου, που υπολογίζουν τα παραγόμενα χαρακτηριστικά Z_m, καλούνται *κρυφές μονάδες*, διότι οι τιμές Z_m δεν μπορούν να παρατηρηθούν άμεσα. Σε γενικές γραμμές μπορεί να υπάρχουν περισσότερα από ένα κρυφά στρώματα. Μπορούμε να σκεφτούμε το Z_m ως βασική επέκταση (basis-expansion) των αρχικών εισόδων X. Το νευρωνικό δίκτυο είναι τότε ένα απλό γραμμικό μοντέλο, ή γραμμικό μοντέλο multilogit, που χρησιμοποιεί αυτούς τους μετασχηματισμούς ως εισροές. Υπάρχει, ωστόσο, μια σημαντική βελτίωση επί των τεχνικών της βασικής επέκτασης (basis-expansion). Εδώ οι παράμετροι της βασικής συνάρτησης «μαθαίνουν» από τα δεδομένα. (Περισσότερα για τις βασικές συναρτήσεις βλ. Hastie et al.(2001)).

Σημειώνουμε ότι εάν το σ είναι η ταυτοτική συνάρτηση (identity function), τότε ολόκληρο το μοντέλο συρρικνώνεται σε ένα γραμμικό μοντέλο στις εισόδους. Ως εκ τούτου ένα νευρωνικό δίκτυο μπορεί να θεωρηθεί ως μια μη γραμμική γενίκευση του γραμμικού μοντέλου, τόσο για την παλινδρόμηση όσο και για την ταξινόμηση. Με την εισαγωγή του μη γραμμικού μετασχηματισμού σ, μεγεθύνεται σε μεγάλο βαθμό η τάξη των γραμμικών μοντέλων. Στο σχήμα 11 βλέπουμε ότι το ποσοστό ενεργοποίησης της σιγμοειδούς εξαρτάται από την νόρμα του α_m, και αν το $\|\alpha_m\|$ είναι πολύ μικρό, η μονάδα θα είναι πράγματι λειτουργική στο γραμμικό μέρος της συνάρτησης ενεργοποίησης της.

Σημειώνουμε επίσης ότι το μοντέλο του νευρωνικού δικτύου με ένα κρυφό στρώμα έχει ακριβώς την ίδια μορφή όπως το projection pursuit model (βλ. Hastie et al.(2001)). Η διαφορά είναι ότι το μοντέλο PPR χρησιμοποιεί απαραμετρικές συναρτήσεις $g_m(v)$, ενώ τα νευρωνικά δίκτυα χρησιμοποιούν μια πολύ απλούστερη συνάρτηση που βασίζεται στη $\sigma(v)$, με τρεις ελεύθερες παραμέτρους στα ορίσματά του. Αναλυτικότερα, βλέποντας το μοντέλο του νευρωνικού δικτύου ως ένα μοντέλο PPR, ορίζουμε

$$\begin{aligned} g_m(\omega_m{}^T X) &= \beta_m \sigma(\alpha_{0m} + \alpha_m{}^T X) \\ &= \beta_m \sigma\big(\alpha_{0m} + \|\alpha_m\|(\omega_m{}^T X)\big) \end{aligned}$$

όπου $\omega_m = \alpha_m/\|\alpha_m\|$ είναι η m-οστή μονάδα-διάνυσμα. Επειδή το $\sigma_{\beta,\alpha_0,s}(v) = \beta\sigma(\alpha_0 + sv)$ έχει μικρότερη πολυπλοκότητα από ένα περίσσοτερο γενικό απαραμετρικό $g(v)$, δεν είναι έκπληξη ότι ένα νευρωνικό δίκτυο μπορεί να χρησιμοποιήσει 20 ή 100 τέτοιες συναρτήσεις, ενώ το μοντέλο PPR συνήθως χρησιμοποιεί λιγότερους όρους (M = 5 ή 10, για παράδειγμα).

Τέλος, σημειώνουμε ότι το όνομα "νευρωνικά δίκτυα" προέρχεται από το γεγονός ότι αναπτύχθηκαν για πρώτη φορά ως μοντέλα για τον ανθρώπινο εγκέφαλο. Κάθε μονάδα αντιπροσωπεύει ένα νευρώνα, καθώς και οι συνδέσεις (links στο σχ. 10) αντιπροσωπεύουν συνάψεις. Στα αρχικά μοντέλα, οι νευρώνες καίγονταν όταν το συνολικό σήμα που περνούσε στη μονάδα υπερέβαινε ένα συγκεκριμένο όριο-κατώφλι. Στο παραπάνω μοντέλο, αυτό αντιστοιχεί με τη χρήση μιας βηματικής συνάρτησης για το $\sigma(Z)$ και το $g_k(T)$. Αργότερα, το νευρωνικό δίκτυο αναγνωρίστηκε ως ένα χρήσιμο εργαλείο για μη γραμμική στατιστική μοντελοποίηση, και γι 'αυτό το σκοπό η βηματική συνάρτηση δεν είναι αρκετά λεία για βελτιστοποίηση. Ως εκ τούτου, η βηματική συνάρτηση αντικαταστάθηκε από μια ομαλότερη συνάρτηση κατωφλίου, τη σιγμοειδή στο Σχήμα 11.

4.4.2 Προσαρμογή των νευρωνικών δικτύων

Το νευρωνικό δίκτυο έχει άγνωστες παραμέτρους, που συχνά αποκαλούνται βάρη (weights), και αναζητούμε τιμές για αυτά που κάνουν την καλύτερη προσαρμογή του μοντέλου στα δεδομένα εκπαίδευσης. Ορίζουμε το πλήρες σύνολο των βαρών με θ, το οποίο αποτελείται από

$$\{\alpha_{0m}, \alpha_m;\ m = 1,2,\dots,M\} \quad M(p+1)\beta\acute{\alpha}\rho\eta,$$

$$\{\beta_{0m}, \beta_m;\ k = 1,2, \dots, K\} \quad K(M+1)\beta\acute{\alpha}\rho\eta.$$

Για παλινδρόμηση, χρησιμοποιούμε ως *μέτρο για την προσαρμογή* τα σφάλματα των αθροισμάτων τετραγώνων (συνάρτηση σφάλματος):

$$R(\theta) = \sum_{\kappa=1}^{K} \sum_{i=1}^{N} (y_{ik} - f_k(x_i))^2$$

Για την ταξινόμηση χρησιμοποιούμε είτε τετραγωνικό σφάλμα ή διασταυρωμένη-εντροπία (απόκλιση):

$$R(\theta) = -\sum_{i=1}^{N} \sum_{\kappa=1}^{K} y_{ik} log f_k(x_i)$$

και ο αντίστοιχος ταξινομητής είναι $G(x) = argmax_k f_k(x)$.

Με τη λειτουργία ενεργοποίησης Softmax και την συνάρτηση σφάλματος της διασταυρωμένης-εντροπίας, το μοντέλο του νευρωνικού δικτύου είναι ακριβώς ένα γραμμικό μοντέλο λογιστικής παλινδρόμησης στις κρυφές μονάδες, και όλες οι παράμετροι εκτιμώνται με τη μέθοδο της μέγιστης πιθανοφάνειας.

Συνήθως δεν θέλουμε την συνολική ελαχιστοποιητή του $R(\theta)$, δεδομένου ότι είναι πιθανό να προκύψει μια overfit λύση. Αντ' αυτού απαιτείται κάποια νομιμοποίηση: αυτό επιτυγχάνεται άμεσα μέσω ενός όρου ποινής, ή έμμεσα, με την πρόωρη διακοπή.

Η γενική προσέγγιση για την ελαχιστοποίηση της $R(\theta)$ είναι η κλίση καθόδου (gradient descent), που ονομάζεται πίσω-διάδοση (back-propagation) σε αυτή την περίπτωση. Λόγω της σύνθετης μορφής του μοντέλου, η κλίση μπορεί εύκολα να παραχθεί χρησιμοποιώντας τον κανόνα της αλυσίδας για τη διαφοροποίηση. Αυτό μπορεί να υπολογίζεται από ένα προς τα εμπρός και προς τα πίσω σάρωμα του δικτύου, παρακολουθώντας μόνο τις ποσότητες που είναι τοπικές σε κάθε μονάδα.

Εδώ παραθέτουμε την πίσω-διάδοση με κάθε λεπτομέρεια για την απώλεια του τετραγωνικού σφάλματος. Έστω $z_{mi} = \sigma(\alpha_{0m} + {\alpha_m}^T x_i)$, από την (4.4.1) και θέτοντας $z_i = (z_{1i}, z_{2i}, \dots, z_{Mi})$.

Μετά έχουμε

$$R(\theta) = \sum_{i=1}^{N} R_i = \sum_{i=1}^{N}\sum_{\kappa=1}^{K}\left(y_{ik} - f_k(x_i)\right)^2$$

Με παραγώγους

$$\frac{\partial R_i}{\partial \beta_{\kappa m}} = -2\left(y_{ik} - f_k(x_i)\right) g'_k\left(\beta_k{}^T z_i\right) z_{mi},$$
$$\frac{\partial R_i}{\partial a_{ml}} = -\sum_{\kappa=1}^{K} 2\left(y_{ik} - f_k(x_i)\right) g'_k\left(\beta_k{}^T z_i\right)\beta_{km}\sigma'(\alpha_m{}^T x_i)x_{il}. \quad (4.4.2)$$

Λαμβάνοντας υπόψη αυτές τις παραγώγους, μια ενημερωμένη έκδοση της κλίσης καθόδου στην $(r+1)$ επανάληψη έχει τη μορφή:

$$\beta_{\kappa m}{}^{(r+1)} = \beta_{\kappa m}{}^{(r)} - \gamma_r \sum_{i=1}^{N} \frac{\partial R_i}{\partial \beta_{\kappa m}{}^{(r)}},$$
$$a_{ml}{}^{(r+1)} = a_{ml}{}^{(r)} - \gamma_r \sum_{i=1}^{N} \frac{\partial R_i}{\partial a_{ml}{}^{(r)}}, \quad (4.4.3)$$

Όπου γ_r είναι το ποσοστό μάθησης, που συζητήσαμε πιο πάνω.

Τώρα γράφουμε την (4.4.2) ως εξής

$$\frac{\partial R_i}{\partial \beta_{\kappa m}} = \delta_{\kappa i} z_{mi},$$
$$\frac{\partial R_i}{\partial a_{ml}} = s_{mi} x_{il}. \quad (4.4.4)$$

Οι ποσότητες $\delta_{\kappa i}$ και s_{mi} είναι «σφάλματα» από το τρέχον μοντέλο στην έξοδο και στις μονάδες του κρυφού στρώματος, αντίστοιχα. Από τους ορισμούς τους, αυτά τα σφάλματα ικανοποιούν

$$s_{mi} = \sigma'(\alpha_m{}^T x_i) \sum_{\kappa=1}^{K} \beta_{km} \delta_{\kappa i} \tag{4.4.5}$$

γνωστές ως εξισώσεις πίσω-διάδοσης. Χρησιμοποιώντας αυτό, η ενημερωμένη έκδοση στην σχέση, που αναφέραμε προηγουμένως, μπορεί να υλοποιηθεί με έναν αλγόριθμο δύο-περασμάτων (two-pass algorithm). Στο *προς τα εμπρός πέρασμα*, τα τρέχοντα βάρη καθορίζονται και οι προβλεπόμενες τιμές $\widehat{f_\kappa}(x_i)$ υπολογίζονται από τον τύπο (4.4.5). Στο *προς τα πίσω πέρασμα*, υπολογίζονται τα σφάλματα $\delta_{\kappa i}$, και στη συνέχεια πίσω-πολλαπλασιάζονται μέσω της (4.4.5) για να δώσει τα σφάλματα s_{mi}. Και τα δύο σύνολα σφαλμάτων χρησιμοποιούνται στη συνέχεια για τον υπολογισμό των κλίσεων για τις ενημερώσεις στην (4.4.3), μέσω της (4.4.4).

Αυτή η διαδικασία των δύο περασμάτων είναι αυτό που είναι γνωστό ως πίσω-διάδοση. Έχει επίσης κληθεί ο κανόνας δέλτα (delta rule) (Widrow και Hoff, 1960). Τα υπολογιστικά στοιχεία για την διασταυρωμένη-εντροπία έχουν την ίδια μορφή με εκείνα της συνάρτησης σφάλματος του αθροίσματος τετραγώνων.

Τα πλεονεκτήματα της πίσω διάδοσης είναι η απλή, τοπική της φύση. Στον αλγόριθμο της πίσω διάδοσης, κάθε κρυφή μονάδα περνά και λαμβάνει πληροφορίες μόνο από και προς τις μονάδες που μοιράζονται μια σύνδεση. Ως εκ τούτου, μπορεί να εφαρμοστεί αποτελεσματικά σε μια παράλληλη αρχιτεκτονική υπολογιστών.

Οι ενημερώσεις στην (4.4.3) είναι ένα είδος μάθησης συνόλου (batch learning), με τις ενημερώσεις της παραμέτρου να είναι ένα άθροισμα πάνω σε όλες τις περιπτώσεις εκπαίδευσης. Μάθηση μπορεί επίσης να πραγματοποιείται σε απευθείας σύνδεση-επεξεργά-ζοντας κάθε παρατήρηση, μία κάθε φορά, ενημερώνοντας την κλίση μετά από κάθε περί-πτωση εκπαίδευσης, και περνώντας μέσα από τις περιπτώσεις εκπαίδευσης πολλές φορές. Στην περίπτωση αυτή, τα αθροίσματα στις εξισώσεις (4.4.3) αντικαθίστανται από ένα ενιαίο άθροισμα. Το training epoch αναφέρεται σε ένα σκούπισμα ολόκληρου του συνόλου εκπαίδευσης. Η online εκπαίδευση δίνει τη δυνατότητα στο δίκτυο να χειριστεί πολύ μεγάλα σύνολα εκπαίδευσης, καθώς επίσης και να ενημερώσει τα βάρη καθώς εισέρχονται νέες παρατηρήσεις.

Το ποσοστο μάθησης γ_r για τη μάθηση του συνόλου (batch learning), συνήθως θεωρείται ότι είναι μια σταθερά, και μπορεί επίσης να βελτιστοποιηθεί με την αναζήτηση γραμμής (line search) που ελαχιστοποιεί τη συνάρτηση σφάλματος σε κάθε ενημέρωση. Με την online μάθηση το γ_r θα πρέπει να μειωθεί στο μηδέν καθώς η επανάληψη $r \to \infty$. Αυτή η

μάθηση είναι μια μορφή της *στοχαστικής προσέγγισης* (Robbins και Munro, 1951). Αποτελέσματα σε αυτόν τον τομέα, εξασφαλίζουν τη σύγκλιση αν $\gamma_r \to 0, \sum_r \gamma_r = \infty,$ και $\sum_r {\gamma_r}^2 < \infty$(πληρούνται, για παράδειγμα, με γ_r = 1 / r). 6

Η πίσω-διάδοση μπορεί να είναι πολύ αργή, και για το λόγο αυτό δεν είναι συνήθως η μέθοδος επιλογής. Δεύτερης-τάξης τεχνικές όπως η μέθοδος του Νεύτωνα εδώ δεν φαίνονται να είναι ελκυστικές, επειδή ο πίνακας των δεύτερων παραγώγων της R (ο εσσιανός πίνακας) μπορεί να είναι πολύ μεγάλος. Καλύτερες προσεγγίσεις για την προσαρμογή περιλαμβάνουν συζευγμένες κλίσεις και μεθόδους μετρικών των μεταβλητών. Αυτά αποφεύγουν ρητά τον υπολογισμό του πίνακα των δεύτερων παραγώγων, ενώ εξακολουθούν να παρέχουν ταχύτερη σύγκλιση.

4.4.3 Μερικά θέματα στην εκπαίδευση των νευρωνικών δικτύων

Υπάρχει μια «τέχνη» στην εκπαίδευση νευρωνικών δικτύων. Το μοντέλο είναι γενικά υπερπαραμετρικό, και το πρόβλημα βελτιστοποίησης είναι ασταθές και μη κυρτό εκτός και αν ακολουθούνται ορισμένες κατευθυντήριες γραμμές. Σε αυτή την ενότητα θα συνοψίσουμε μερικά από τα σημαντικά ζητήματα.

Τιμές εκκίνησης

Σημειώνουμε ότι εάν τα βάρη είναι κοντά στο μηδέν, τότε το ενεργό μέρος της σιγμοειδούς (Σχήμα 10), είναι περίπου γραμμικό, και επομένως το νευρωνικό δίκτυο καταρρέει περίπου σε ένα γραμμικό μοντέλο. Συνήθως οι τιμές εκκίνησης-αρχικές τιμές για τα βάρη επιλέγονται να είναι τυχαίες τιμές κοντά στο μηδέν. Ως εκ τούτου, το μοντέλο ξεκινά σχεδόν γραμμικό, και γίνεται μη γραμμικό καθώς τα βάρη αυξάνονται. Μεμονωμένες μονάδες εντοπίζονται με τις οδηγίες και την εισαγωγή μη γραμμικότητας, όπου χρειάζεται. Χρήση ακριβώς μηδενικών βαρών οδηγεί σε μηδενικές παραγώγους και τέλεια συμμετρία, και ο αλγόριθμος δεν κινείται ποτέ. Αντ΄αυτού ξεκινώντας με μεγάλα βάρη συχνά οδηγούμαστε σε κακές λύσεις.

Υπερπροσαρμογή (overfitting)

Συχνά τα νευρωνικά δίκτυα έχουν πάρα πολλά βάρη και θα υπερπροσαρμόσουν (overfit) τα δεδομένα αν κάνουμε ολική ελαχιστοποίηση του R. Στις πρώτες εξελίξεις των νευρωνικών δικτύων, είτε από το σχεδιασμό είτε τυχαία, χρησιμοποιήθηκε ένας πρόωρος κανόνας διακοπής για να αποφευχθεί το overfitting. Εδώ έχουμε εκπαιδεύσει το μοντέλο μόνο για

μια στιγμή, και να σταματήσει πολύ πριν εμείς προσεγγίσουμε το ολικό ελάχιστο. Δεδομένου ότι τα βάρη ξεκινούν από μία πολύ συστηματοποιημένη (γραμμική) λύση, αυτό έχει ως αποτέλεσμα τη συρρίκνωση του τελικού μοντέλου σε ένα γραμμικό μοντέλο. Ένα σύνολο δεδομένων επικύρωσης είναι χρήσιμο για τον προσδιορισμό του πότε να σταματήσει, αφού περιμένουμε το σφάλμα επικύρωσης να αρχίσει να αυξάνεται.

Μια πιο σαφής μέθοδος για την συστηματοποίηση είναι η φθορά του βάρους (weight decay), το οποίο είναι ανάλογο στην παλινδρόμηση κορυφογραμμής και χρησιμοποιείται για γραμμικά μοντέλα. Προσθέτουμε μία ποινή στη συνάρτηση σφάλματος $R(\theta) + \lambda J(\theta)$, όπου

$$J(\theta) = \sum_{km} {\beta_{km}}^2 + \sum_{ml} {a_{ml}}^2 \qquad (4.4.6)$$

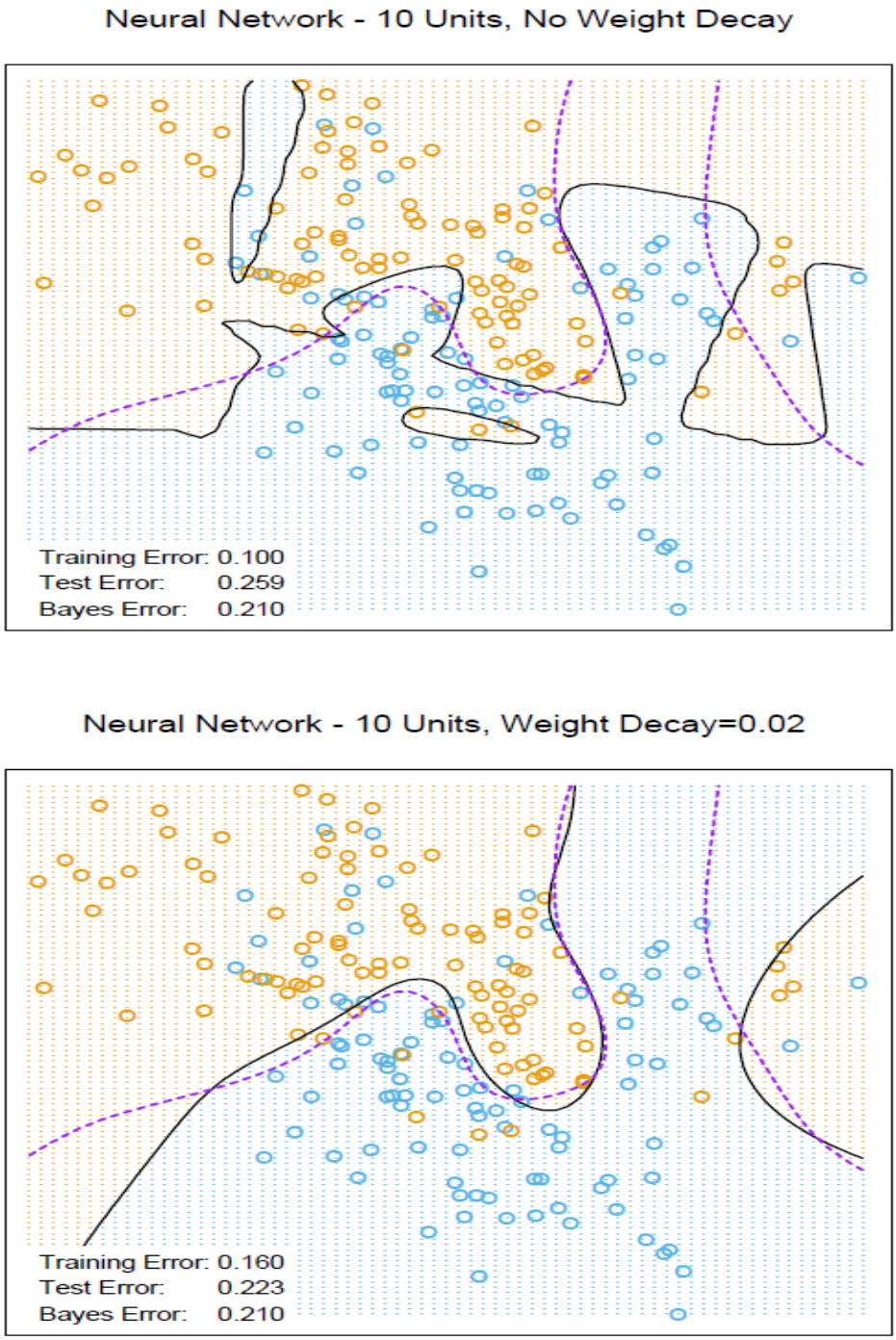
και $\lambda \geq 0$ είναι μια παράμετρος συντονισμού (tuning). Οι μεγαλύτερες τιμές του λ θα τείνουν να συρρικνώσουν τα βάρη προς το μηδέν: τυπικά η διασταυρωμένη-επικύρωση χρησιμοποιείται για να εκτιμηθεί το λ. Η επίδραση της ποινής είναι απλά για να προσθέσει τους όρους $2\beta_{km}$ και $2\alpha_{ml}$ για τις αντίστοιχες εκφράσεις παραγώγων (4.4.3).

Άλλες μορφές για την ποινή έχουν προταθεί, για παράδειγμα,

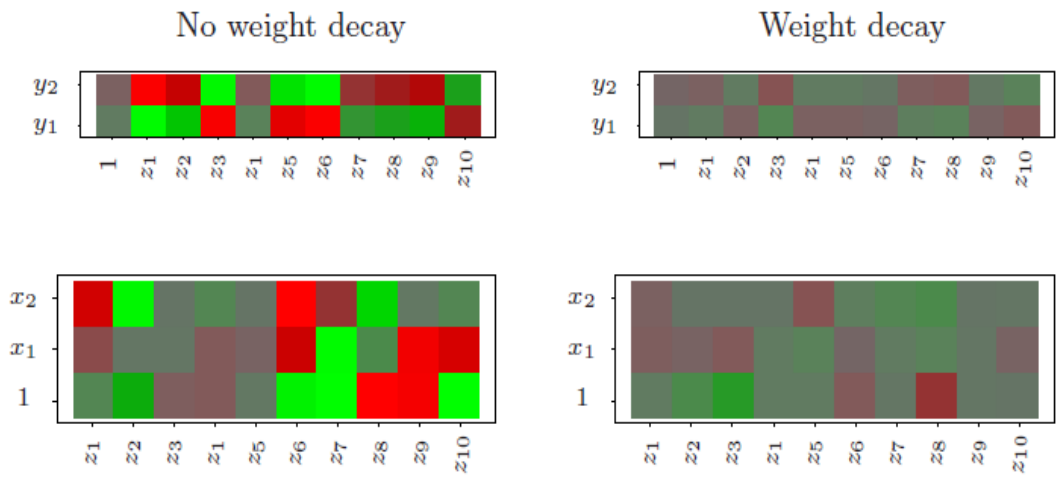
$$J(\theta) = \sum_{km} \frac{{\beta_{km}}^2}{1 + {\beta_{km}}^2} + \sum_{ml} \frac{{a_{ml}}^2}{1 + {a_{ml}}^2} \qquad (4.4.7)$$

γνωστή ως ποινή αποβολής βάρους (weight elimination penalty). Αυτό έχει το αποτέλεσμα της συρρίκνωσης μικρότερων βαρών περισσότερο απ'ότι κάνει η (4.4.6).

Το Σχήμα 12 δείχνει το αποτέλεσμα της εκπαίδευση ενός νευρικού δικτύου με δέκα κρυφές μονάδες, χωρίς τα decay βάροι (άνω πίνακας) και με το decay βάρος (κάτω πίνακας), στο παράδειγμα μίγμα του κεφαλαίου 2 στο Hastie et al. (2001). Τα decay βάροι έχουν σαφώς βελτιώσει την πρόβλεψη.



Σχήμα 12: Ένα νευρωνικό δίκτυο στο παράδειγμα του κεφαλαίου 2 (mixture example) στο Hastie et al. (2001). Ο άνω πίνακας δεν χρησιμοποιεί decay βάροι, και γίνεται υπερπροσαρμογή των δεδομένων εκπαίδευσης. Ο κάτω πίνακας χρησιμοποιεί decay βάρη, και επιτυγχάνει κοντά το ποσοστό σφάλματος Bayes (διακεκομμένο μωβ όριο). Και οι δύο χρησιμοποιούν τη Softmax συνάρτηση ενεργοποίησης και το σφάλμα διασταυρωμένης-εντρόπιας.



Σχήμα 13: Χάρτες θερμότητας των εκτιμώμενων βαρών από την εκπαίδευση των νευρωνικών δικτύων από το Σχήμα 12. Η παρουσίαση ποικίλει από το φωτεινό πράσινο (αρνητικές) στο έντονο κόκκινο (θετικό).

Το σχήμα 13 δείχνει τους χάρτες θερμότητας των εκτιμώμενων βαρών από την εκπαίδευση (οι εκδόσεις σε κλίμακα του γκρι ονομάζονται Hinton διαγράμματα). Βλέπουμε ότι τα decay βάροι, έβλαψαν τα βάρη και στα δύο στρώματα: τα προκύπτοντα βάρη κατανέμονται αρκετά ομοιόμορφα πάνω στις δέκα κρυμμένες μονάδες.

Κλιμάκωση των εισόδων

Επειδή η κλιμάκωση των εισόδων προσδιορίζει την αποτελεσματική κλιμάκωση των βαρών στο κάτω στρώμα, μπορεί να έχει μια μεγάλη επίδραση στην ποιότητα του τελικού αποτελέσματος. Κατ' αρχάς είναι καλύτερο να τυποποιηθούν όλες οι είσοδοι για να έχουν μέση τιμή μηδέν και τυπική απόκλιση ένα. Αυτό εξασφαλίζει ότι όλες οι είσοδοι έχουν ίση μεταχείριση στην η διαδικασία συστηματοποίησης (regularization process), και επιτρέπει σε κάποιον να επιλέξει μια σημαντική περιοχή για την εκκίνηση των τυχαίων βαρών. Με τυποποιημένες εισόδους, είναι τυπικό να ληφθούν τυχαία ομοιόμορφα βάρη πάνω από στο εύρος [-0.7, 0.7].

Αριθμός των κρυφών μονάδων και στρωμάτων

Σε γενικές γραμμές είναι καλύτερο να έχουμε πάρα πολλές κρυφές μονάδες απ'ότι πολύ λίγες. Με πολύ λίγες κρυφές μονάδες, το μοντέλο μπορεί να μην έχει αρκετή ευελιξία για να συλλάβει τις μη γραμμικότητες στα δεδομένα. Με πάρα πολλές κρυφές μονάδες, τα

επιπλέον βάρη μπορεί να συρρικνωθούν προς το μηδέν αν χρησιμοποιείται κατάλληλη συστηματοποίηση. Τυπικά, ο αριθμός των κρυμμένων μονάδων είναι κάπου στην κλίμακα 5 έως 100, με τον αριθμό να αυξάνει με τον αριθμό των εισόδων και τον αριθμό των περιπτώσεων εκαπίδευσης. Είναι πιο κοινό να καταθέσουμε ένα αρκετά μεγάλο αριθμό μονάδων και την εκπαίδευσή τους με συστηματοποίηση. Μερικοί ερευνητές χρησιμοποιούν διασταυρωμένη-επικύρωση για να εκτιμήσουν το βέλτιστο αριθμό, αλλά αυτό φαίνεται περιττό εάν η διασταυρωμένη-επικύρωση χρησιμοποιείται για την εκτίμηση της συστηματοποίησης της παραμέτρου (regularization parameter). Η επιλογή του αριθμού των κρυμμένων στρωμάτων οδηγείται από το υπόβαθρο της γνώσης και τον πειραματισμό. Κάθε στρώμα εξάγει τα χαρακτηριστικά της εισόδου για την παλινδρόμηση ή ταξινόμηση. Χρήση πολλαπλών κρυμμένων στρωμάτων επιτρέπει την κατασκευή ιεραρχικών χαρακτηριστικών σε διαφορετικά επίπεδα ανάλυσης.

Πολλαπλά ελάχιστα

Η συνάρτηση σφάλματος $R(\theta)$ είναι μη κυρτή, έχοντας πολλά τοπικά ελάχιστα. Ως αποτέλεσμα, η τελική λύση που λαμβάνεται εξαρτάται αρκετά από την επιλογή των αρχικών βαρών. Κάποιος πρέπει να προσπαθήσει τουλάχιστον μια σειρά από τυχαίες διαμορφώσεις εκκίνησης, και να επιλέξει τη λύση που δίνει χαμηλότερο (ποινικοποιημένο) σφάλμα. Πιθανώς μία καλύτερη προσέγγιση είναι να χρησιμοποιήσουμε τις μέσες προβλέψεις πάνω στη συλλογή των δικτύων ως την τελική πρόβλεψη (Ripley, 1996). Αυτό είναι προτιμότερο να παίρνεις το μέσο όρο των βαρών, δεδομένου ότι η μη γραμμικότητα του μοντέλου προϋποθέτει ότι αυτή η κατά μέσο όρο λύση θα μπορούσε να είναι αρκετά φτωχή. Μία άλλη προσέγγιση είναι μέσω bagging, η οποία παίρνει το μέσο όρο των προβλέψεων της εκπαίδευσης του δικτύου από τυχαία διαταρασσόμμενες εκδόσεις των δεδομένων εκπαίδευσης.

4.5 Μοντέλα Μπεϋζιανών δικτύων (Bayesian networks models)

Ένα Μπεϋζιανό δίκτυο (Bayes network), δίκτυο πεποιθήσεων, Μπεϋζιανό μοντέλο ή πιθανολογικά κατευθυνόμενο άκυκλο γραφικό μοντέλο είναι ένα πιθανοτικό γραφικό μοντέλο (ένα είδος στατιστικού μοντέλου), που αντιπροσωπεύει ένα σύνολο τυχαίων μεταβλητών και τις εξαρτήσεις τους, μέσω ενός κατευθυνόμενου άκυκλου γραφήματος (DAG). Για παράδειγμα, ένα Μπεϋζιανό δίκτυο θα μπορούσε να αντιπροσωπεύει τις σχέσεις μεταξύ των πιθανολογικών ασθενειών και των συμπτωμάτων. Λαμβάνοντας υπόψη τα συμπτώματα, το δίκτυο μπορεί να χρησιμοποιηθεί για να υπολογίσει τις πιθανότητες της παρουσίας διαφόρων ασθενειών.

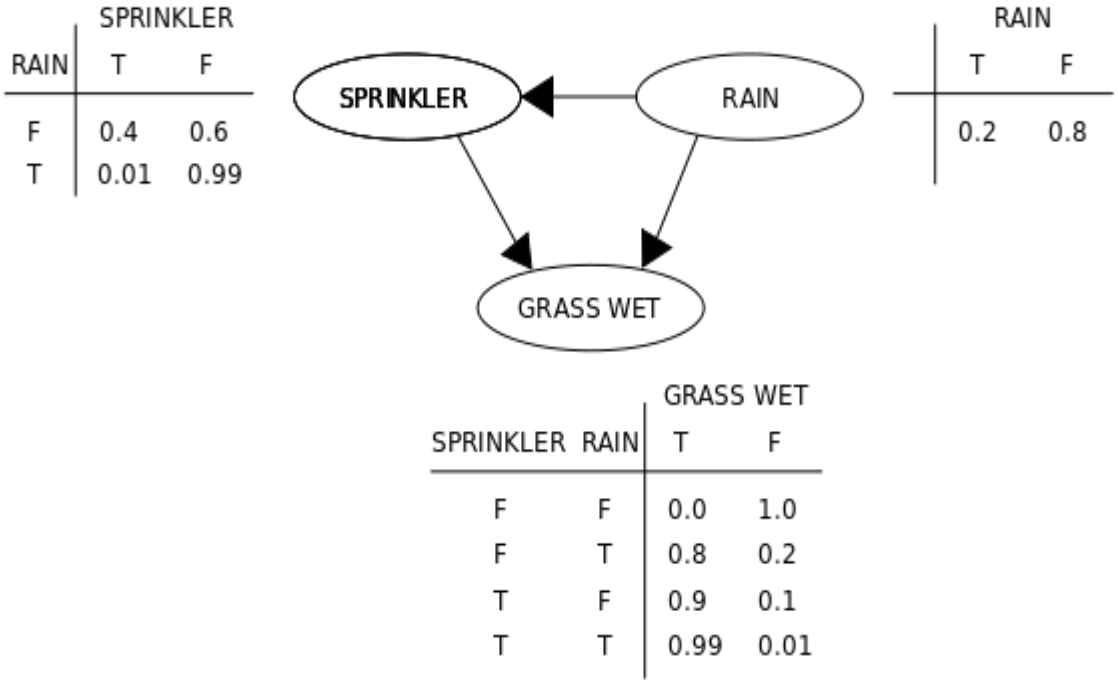
Επισήμως, τα Μπεϋζιανά δίκτυα είναι κατευθυνόμενα άκυκλα γραφήματα των οποίων οι κόμβοι αποτελούν τυχαίες μεταβλητές στην Μπεϋζιανή λογική: μπορεί να είναι παρατηρήσιμες ποσότητες, λανθάνουσες μεταβλητές, άγνωστες παράμετροι και υποθέσεις. Οι ακμές αντιπροσωπεύουν τις εξαρτήσεις, οι κόμβοι που δεν είναι συνδεδεμένοι αντιπροσωπεύουν μεταβλητές που είναι δυνητικά ανεξάρτητες μεταξύ τους. Κάθε κόμβος συνδέεται με μια συνάρτηση πιθανότητας που λαμβάνει ως είσοδο ένα συγκεκριμένο σύνολο τιμών για τις γονεϊκέςμεταβλητές του κόμβου και δίνει την πιθανότητα της μεταβλητής που αντιπροσωπεύεται από τον κόμβο. Για παράδειγμα, αν οι γονείς είναι m Boolean μεταβλητές, τότε η συνάρτηση πιθανότητας θα μπορούσε να αντιπροσωπεύεται από έναν πίνακα 2^m καταχωρήσεων, μια καταχώρηση για κάθε ένα από τους 2^m πιθανούς συνδυασμούς των γονέων του είναι αληθείς ή ψευδείς. Παρόμοιες ιδέες μπορούν να εφαρμοστούν σε μη κατευθυνόμενα, και ενδεχομένως κυκλικά, γραφήματα, τέτοια ονομάζονται ως Μαρκοβιανά δίκτυα.

Υπάρχουν αποτελεσματικοί αλγόριθμοι που εκτελούν συμπεράσματα και μάθηση σε Μπεϋζιανά δίκτυα. Μπεϋζιανά δίκτυα που μοντελοποιούν ακολουθίες μοντέλων των μεταβλητών (π.χ. σήματα ομιλίας ή αλληλουχιών πρωτεϊνών) ονομάζονται δυναμικά Μπεϋζιανά δίκτυα. Οι γενικεύσεις των Μπεϋζιανών δικτύων που μπορεί να αντιπροσωπεύσουν και να λύσουν τα προβλήματα απόφασης υπό αβεβαιότητα ονομάζονται διαγράμματα επιρροής.

Ένα Μπεϋζιανό δίκτυο παρέχει ένα σύντομο τρόπο περιγραφής της κοινής κατανομής πιθανότητας για ένα δεδομένο σύνολο τυχαίων μεταβλητών. Έστω ότι V είναι ένα σύνολο τυχαίων κατηγορικών μεταβλητών και $G = (V, E)$ είναι ένα κατευθυνόμενο άκυκλο γράφημα κόμβων V και ένα σύνολο κατευθυνόμενων ακμών Ε. Ένα μοντέλο Μπεϋζιανού δικτύου (Bayesian Network Models) αποτελείται από το γράφημα G μαζί με έναν πίνακα υπό συνθήκη πιθανοτήτων για κάθε δεδομένο κόμβο από τους μητρικούς του κόμβους. Με δεδομένη την τιμή των μητρικών του, κάθε κόμβος θεωρείται ότι είναι ανεξάρτητος από όλους τους κόμβους που δεν είναι απόγονοι του. Η κοινή κατανομή πιθανότητας για τις μεταβλητές V μπορεί στη συνέχεια να υπολογιστεί ως γινόμενο των υπό συνθήκη πιθανοτήτων για όλους τους κόμβους, δεδομένων των γονιών του κάθε κόμβου. Δεδομένου ενός συνόλου μεταβλητών V και ένος αντίστοιχου συνόλου δεδομένων δείγματος, παρουσιάζουμε το έργο της τοποθέτησης κατάλληλου μοντέλου Μπεϋζιανού δικτύου.

Το θέμα του καθορισμού των κατάλληλων ακμών στο γράφημα G ονομάζεται δομή μάθησης, ενώ το έργο της εκτίμησης του πίνακα δεσμευμένης πιθανότητας δοθέντων των μητρικών κόμβων ονομάζεται παράμετρος μάθησης.

Ένα απλό παράδειγμα ενός Μπεϋζιανού δικτύου είναι το ακόλουθο



Σχήμα 14: Ένα απλό Μπεϋζιανό δίκτυο

Εκμάθηση παραμέτρων

Για να προσδιοριστεί πλήρως το Μπεϋζιανό δίκτυο και ως εκ τούτου η πλήρης παρουσίαση της από κοινού κατανομής πιθανότητας, είναι απαραίτητο να διευκρινιστεί, για κάθε κόμβο X η κατανομή πιθανότητας του X εξαρτάται από τους γονείς του X. Η κατανομή των X που εξαρτάται από τους γονείς του μπορεί να έχει οποιαδήποτε μορφή. Είναι σύνηθες να συνεργαστεί με διακριτή ή Gaussian κατανομή δεδομένου ότι απλοποιεί τους υπολογισμούς. Μερικές φορές μόνο οι περιορισμοί συνάρτηση κατανομής είναι γνωστοί, και κάποιος μπορεί να χρησιμοποιήσει στη συνέχεια την αρχή της μέγιστης εντροπίας για να καθορίσει μία ενιαία συνάρτηση κατανομής, εκείνη με τη μεγαλύτερη εντροπία δεδομένων των περιορισμών. (Κατ 'ανάλογο τρόπο, στο συγκεκριμένο πλαίσιο ενός δυναμικού Μπεϋζιανού δικτύου, κάποιος μπορεί να καθορίσει τη δεσμευμένη κατανομή για την χρονική εξέλιξη του κρυφού για να μεγιστοποιήσει το ποσοστό της εντροπία της σιωπηρής στοχαστικής διαδικασίας.)

Συχνά, αυτές οι δεσμευμένες κατανομές περιλαμβάνουν παραμέτρους που είναι άγνωστες και πρέπει να εκτιμηθούν με βάση τα δεδομένα, μερικές φορές χρησιμοποιώντας την προσέγγιση μέγιστης πιθανοφάνειας. Άμεση μεγιστοποίηση της πιθανότητας (ή εκ των υστέρων-posterior πιθανότητας) είναι συχνά πολύπλοκη όταν υπάρχουν απαρατήρητες μεταβλητές. Μια κλασική προσέγγιση σε αυτό το πρόβλημα είναι ο αλγόριθμος EM (expectation-maximization) που εναλλάσσει τις υπολογιστικές αναμενόμενες τιμές των μη παρατηρούμενων μεταβλητών υπό την προϋπόθεση των παρατηρούμενων δεδομένων, με τη μεγιστοποίηση της πλήρους πιθανότητα (ή posterior), θεωρώντας ότι προηγουμένως οι αναμενόμενες τιμές που υπολογίστηκαν είναι σωστές. Κάτω από ήπιες συνθήκες κανονικότητας αυτή η διαδικασία συγκλίνει στις τιμές της μεγιστοποιημένης πιθανοφάνειας (ή το μέγιστο posterior) για τις παραμέτρους.

Μια πιο πλήρης Μπεϋζιανή προσέγγιση στις παραμέτρους είναι η μεταχείρηση τους ως πρόσθετες απαρατήρητες μεταβλητές και ο υπολογισμός μιας πλήρους posterior κατανομής σε όλους τους κόμβους εξαρτάται από τις υπο συνθήκη πάνω στα δεδομένα, και στη συνέχεια, να ενσωματώσει τις παραμέτρους. Αυτή η προσέγγιση μπορεί να είναι ακριβή και να οδηγήσει σε μεγάλων διαστάσεων μοντέλα, έτσι ώστε στην πράξη η κλασική παραμετρικές προσεγγίσεις είναι πιο συχνές.

Εφαρμογή

Τα Μπεϋζιανά δίκτυα χρησιμοποιούνται για τη γνώση μοντέλων στην υπολογιστική βιολογία και στη βιοπληροφορική (ρυθμιστικά δίκτυα γονιδίου, πρωτεϊνική δομή, στη γονιδιακή ανάλυση της έκφρασης (Friedman et al. (2000)), τη μάθηση epistasis από τα σύνολα δεδομένων GWAS (Jiang et al. (2011))) στην ιατρική, (J. Uebersax (2004)) στη βιοπαρακολούθηση, (Jiang και Cooper (2010)) στην κατάταξη εγγράφων, στην ανάκτηση πληροφοριών, (Luis et al. (2004)) στη σημασιολογική αναζήτηση (Koumenides και Shadbolt (2012)), στην επεξεργασία εικόνας, στη συγχώνευση δεδομένων, στα συστήματα υποστήριξης αποφάσεων (Díez et al. (1997)), στη μηχανική, στα τυχερά παιχνίδια και στο νόμο.

4.6 Μηχανές Διανυσματικής υποστήριξης (Support Vector Machines)

4.6.1 Εισαγωγικά στοιχεία

Οι Μηχανές Διανυσματικής Υποστήριξης σαν ιδέα δημιουργήθηκαν από τον Cortes και τον Vapnik (2000). Στην παρούσα εργασία θα γίνει μία παρουσίαση των βασικών εννοιών που απαρτίζουν τις μηχανές διανυσματικής υποστήριξης.

Η SVM τεχνική βασίζεται στην στατιστική θεωρία της μάθησης[3] και μπορεί να χρησιμοποιηθεί για την πρόβλεψη μελλοντικών δεδομένων. Εκπαιδεύεται από την επίλυση ενός περιορισμένου προβλήματος ταξινόμησης και υλοποιεί τη χαρτογράφηση των συντελεστών παραγωγής σε ένα υψηλό τρισδιάστατο χώρο χρησιμοποιώντας ένα σύνολο μη γραμμικών βασικών συναρτήσεων. Η SVM τεχνική μπορεί να χρησιμοποιηθεί για μια ποικιλία από αναπαραστάσεις όπως τα νευρωνικά δίκτυα, τα splines, τους πολυωνυμικούς εκτιμητές κ.λπ., αλλά υπάρχει μια μοναδική βέλτιστη λύση για κάθε επιλογή των SVM παραμέτρων. Αυτό είναι διαφορετικό σε άλλες μηχανές μάθησης όπως τα τυποποιημένα Νευρωνικά Δίκτυα που χρησιμοποιούν την προς τα πίσω διάδοση. Με λίγα λόγια η ανάπτυξη της SVM μεθόδου είναι εντελώς διαφορετική από τους συνήθεις αλγόριθμους που χρησιμοποιούνται για τη μάθηση και έτσι η SVM τεχνική παρέχει μία νέα άποψη μάθησης. Τα τέσσερα πιο σημαντικά χαρακτηριστικά της SVM τεχνικής είναι η δυαδικότητα**,** οι πυρήνες, η κυρτότητα και η σποραδικότητα.

Οι μηχανές διανυσματικής υποστήριξης λειτουργούν ως μία από τις καλύτερες προσεγγίσεις για τη μοντελοποίηση δεδομένων. Συνδυάζουν τον γενικευμένο έλεγχο ως μία τεχνική για τον έλεγχο των διαστάσεων. Η χαρτογράφηση του πυρήνα παρέχει μια κοινή βάση για τα περισσότερα από τα συνηθισμένα απασχολούμενα αρχιτεκτονικά μοντέλα, που επιτρέπει τις συγκρίσεις που πρέπει να εκτελεστούν.

Στα προβλήματα ταξινόμησης επιτυγχάνεται γενικευμένος έλεγχος με τη μεγιστοποίηση του περιθωρίου κέρδους, το οποίο αντιστοιχεί στην ελαχιστοποίηση του διανύσματος σε ένα κανονικό πλαίσιο. Η ελαχιστοποίηση του διανύσματος βάρους μπορεί να χρησιμοποιηθεί ως

[3] Οι αλγόριθμοι μηχανικής μάθησης έχουν στόχο τις αναπαραστάσεις απλών λειτουργιών. Ως εκ τούτου, στόχος της εκπαίδευσης είναι το αποτέλεσμα μιας υπόθεσης που πραγματοποιεί σωστή ταξινόμηση των δεδομένων εκπαίδευσης και οι αρχικοί αλγόριθμοι εκμάθησης έχουν σχεδιαστεί για να βρίσκουνε μία τέτοια λύση που να ταιριάζει με τα δεδομένα. Η SVM τεχνική αποδίδει καλύτερα σε όρους που δεν είναι πάνω στη γενίκευση, σε αντίθεση με τα νευρωνικά δίκτυα τα οποία καταλήγουν πιο εύκολα σε γενίκευση.

κριτήριο και σε προβλήματα παλινδρόμησης με μία τροποποιημένη λειτουργία απώλειας. Οι μελλοντικές κατευθύνσεις περιλαμβάνουν μία τεχνική για την επιλογή της λειτουργίας και έναν επιπλέον έλεγχο της ικανότητας. Τέλος, οι νέες κατευθύνσεις που αναφέρονται στη νέα προσέγγιση της SVM τεχνικής σχετίζονται με σκευάσματα μάθησης που προτάθηκαν πρόσφατα από τον Vapnik .

4.6.2 Η SVM μέθοδος για την δυαδική ταξινόμηση

Η SVM είναι μία χρήσιμη τεχνική για την ταξινόμηση των δεδομένων. Ακόμη και αν τα Νευρωνικά Δίκτυα θεωρούνται ότι είναι ευκολότερα στη χρήση, μερικές φορές λαμβάνονται μη ικανοποιητικά αποτελέσματα. Μια διαδικασία ταξινόμησης περιλαμβάνει συνήθως τα δεδομένα εκπαίδευσης και εξέτασης που αποτελούνται από κάποιες περιπτώσεις (στιγμιότυπα) δεδομένων. Κάθε στιγμιότυπο, στο σύνολο της κατάρτισης, περιέχει μία τιμή-στόχο και διάφορα χαρακτηριστικά. Ο στόχος της SVM τεχνικής είναι να παράγει ένα μοντέλο το οποίο προβλέπει την τιμή-στόχο των δεδομένων στο σύνολο των δοκιμών.

Η SVM είναι ένα παράδειγμα μάθησης με πλήρη επίβλεψη[4]. Γνωστές ετικέτες βοηθάνε στην αναφορά εάν το σύστημα εκτελείται σε σωστό δρόμο ή όχι. Αυτή η πληροφορία παραπέμπει σε μια επιθυμητή απάντηση, είτε αφορά στην επικύρωση της ακρίβειας του συστήματος, ή στη χρησιμοποίηση για να μάθει το σύστημα να ενεργεί σωστά. Ένα βήμα για την SVM ταξινόμηση περιλαμβάνει την αναγνώριση, κάτι το οποίο είναι άρρηκτα συνδεδεμένο με τις γνωστές κατηγορίες. Αυτό ονομάζεται επιλογή χαρακτηριστικών ή εξαγωγή χαρακτηριστικών. Η δυνατότητα επιλογής και της SVM ταξινόμησης έχει από κοινού χρήση, ακόμη και όταν η πρόβλεψη των άγνωστων δειγμάτων δεν είναι απαραίτητη. Μπορεί να χρησιμοποιηθεί για να προσδιορίσει τα βασικά σύνολα που εμπλέκονται στις διεργασίες για διάκριση μεταξύ των τάξεων.

Η απλούστερη μορφή επίλυσης ενός προβλήματος πρόβλεψης είναι η δυαδική κατηγοριοποίηση (binary classification), όπου πρέπει να γίνει ένας διαχωρισμός σε αντικείμενα που ανήκουν σε μία από δύο κατηγορίες οι οποίες συμβολίζονται με θετικό (+1)

[4]Μέθοδοι με επίβλεψη **(supervised methods)**: Οι αλγόριθμοι εκμάθησης με επίβλεψη είναι εκείνοι που χρησιμοποιούνται στην ταξινόμηση και στην πρόβλεψη. Ουσιαστικά μοντελοποιούν μια μεταβλητή απόκρισης βασιζόμενοι σε μια ή περισσότερες επεξηγηματικές μεταβλητές (input variables). Μερικές από αυτές τις supervised τεχνικές είναι και τα νευρωνικά δίκτυα (neural networks), τα δέντρα αποφάσεων (decision trees) και η λογιστική παλινδρόμηση (logistic regression).

ή αρνητικό (-1) πρόσημο. Οι SVMs χρησιμοποιούν για την επίλυση αυτού του προβλήματος:

α) διαχωρισμό δεδομένων με μεγάλο περιθώριο (large margin separation) και

β) πράξεις στο επίπεδο των kernels (πυρήνων) (kernel functions).

<u>Γραμμικά διαχωρίσιμα δεδομένα</u>

<u>Θεωρία</u>

Έχουμε L σημεία εκπαίδευσης, όπου κάθε είσοδος x_i έχει D χαρακτηριστικά (δηλαδή είναι διάστασης D) και είναι σε μία από τις δύο κατηγορίες y_i = -1 ή +1, δηλαδή τα δεδομένα εκπαίδευσης είναι της μορφής:

$$\{x_i, y_i\}, \qquad \text{όπου} \qquad i = 1, \dots, L, \quad y_i \in \{-1,1\}, \quad \mathbf{x} \in \Re^D$$

Εδώ υποθέτουμε ότι τα δεδομένα είναι γραμμικά διαχωρίσιμα, πράγμα που σημαίνει ότι μπορούμε να δημιουργήσουμε μια γραμμή επί του γραφήματος των x_1 vs x_2 που χωρίζει τις δύο κλάσεις όταν $D = 2$, και ένα υπερεπίπεδο σε γραφήματα $x_1, x_2, \dots, x_D$ όταν $D > 2$. Αυτό το υπερεπίπεδο μπορεί να περιγραφεί από την εξίσωση $\boldsymbol{w} \cdot \boldsymbol{x} + b = 0$ όπου:

- Το **w** είναι κάθετο προς το υπερεπίπεδο.
- $\frac{b}{\|\boldsymbol{w}\|}$ είναι η κάθετη απόσταση από το υπερεπίπεδο προς την αρχή.

Τα διανύσματα υποστήριξης είναι τα παραδείγματα που βρίσκονται πλησιέστερα προς το διαχωριστικό υπερεπίπεδο και ο στόχος της μηχανής διανυσματικής υποστήριξης (SVM) είναι να προσανατολίσει το υπερεπίπεδο κατά τέτοιο τρόπο ώστε να είναι όσο το δυνατόν μακρύτερα από τα πλησιέστερα μέλη των δύο τάξεων.

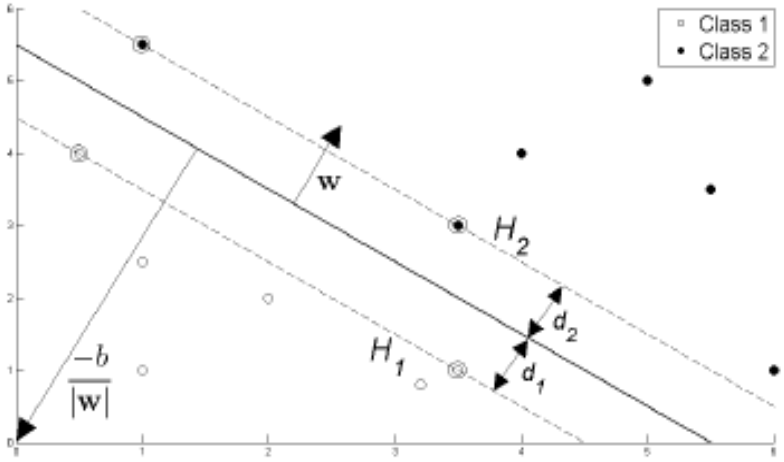
Αναφερόμενοι στο Σχήμα 15, η υλοποίηση SVM στηρίζεται στην επιλογή των μεταβλητών **w** και b, έτσι ώστε τα δεδομένα εκπαίδευσης να μπορούν να περιγραφούν με:

$$\boldsymbol{x}_i \cdot \boldsymbol{w} + b \geq +1 \qquad \gamma\iota\alpha \;\; y_i = +1$$

$$\boldsymbol{x}_i \cdot \boldsymbol{w} + b \leq -1 \qquad \gamma\iota\alpha \;\; y_i = -1$$

Αυτές οι εξισώσεις μπορούν να συνδυαστούν ως εξής :

$$y_i\,(\boldsymbol{x}_i \cdot \boldsymbol{w} + b)\; - 1 \;\geq\; 0 \qquad \forall i \tag{4.6.1}$$



Σχήμα 15 : Υπερεπίπεδο ανάμεσα σε δύο γραμμικά διαχωρισμένες κλάσεις

Αν εξετάσουμε τώρα μόνο τα σημεία που κείνται πλησιέστερα προς το διαχωριστικό υπερ-επίπεδο, δηλαδή τα διανύσματα υποστήριξης (που φαίνονται σε κύκλους στο διάγραμμα), τότε τα δύο επίπεδα H_1 και H_2 όπου αυτά τα σημεία κείνται μπορούν να περιγραφούν με:

$$\boldsymbol{x}_i \cdot \boldsymbol{w} + b \;=\; +1 \qquad \gamma\iota\alpha\;\; H_1$$

$$\boldsymbol{x}_i \cdot \boldsymbol{w} + b \;=\; -1 \qquad \gamma\iota\alpha\;\; H_2$$

Αναφερόμενοι στο Σχήμα 1, έχουμε ορίσει d_1 ως την απόσταση από το H_1 προς το υπερεπίπεδο και d_2 ως την απόσταση από το H_2 προς αυτό. Η ίση απόσταση του υπερεπιπέδου από το H_1 και το H_2 σημαίνει ότι $d_1 = d_2$, και είναι μια ποσότητα γνωστή ως περιθώριο του SVM. Για να προσανατολίσει το υπερεπίπεδο να είναι όσο το δυνατόν πιο μακριά γίνεται από τα διανύσματα υποστήριξης, θα πρέπει να μεγιστοποιήσει αυτό το περιθώριο.

Η απλή γεωμετρία διανυσμάτων δείχνει ότι το περιθώριο είναι ίσο με $\frac{1}{\|\boldsymbol{w}\|}$ και η μεγιστοποίηση αυτού υπό τους περιορισμούς (4.6.1) είναι ισοδύναμη με την εύρεση:

$$min\;\|\boldsymbol{w}\| \qquad \text{s.t.} \qquad y_i\,(\boldsymbol{x}_i \cdot \boldsymbol{w} + b)\; - 1 \;\geq\; 0 \qquad \forall i$$

Ελαχιστοποιώντας το $\|\boldsymbol{w}\|$ είναι ισοδύναμο με την ελαχιστοποίηση του $\frac{1}{2}\|\boldsymbol{w}\|^2$ και η χρήση αυτού κάνει δυνατή την εκτέλεση της βελτιστοποίησης του Τετραγωνικού

προγραμματισμού (Quadratic programming optimization). Εμείς χρειάζεται να υπολογίσουμε :

$$\min \frac{1}{2} \|\boldsymbol{w}\|^2 \quad \text{s.t.} \quad y_i \, (\boldsymbol{x}_i \cdot \boldsymbol{w} + b) \, - 1 \, \geq \, 0 \qquad \forall i \tag{4.6.2}$$

Προκειμένου να ληφθεί μέριμνα για τους περιορισμούς σε αυτή την ελαχιστοποίηση, θα πρέπει να κατανέμουμε σε αυτούς, τους πολλαπλασιαστές Lagrange $\boldsymbol{a}$, όπου $a_i \geq 0, \forall i$:

$$\begin{aligned} L_P \; &\equiv \frac{1}{2}\|\boldsymbol{w}\|^2 - \boldsymbol{a} \, [\, y_i \, (\boldsymbol{x}_i \cdot \boldsymbol{w} + b) \, - 1 \, \forall i \,] \\ &\equiv \frac{1}{2}\|\boldsymbol{w}\|^2 - \, \sum_{\boldsymbol{i=1}}^{\boldsymbol{L}} \boldsymbol{a_i} \, [\, y_i \, (\boldsymbol{x}_i \cdot \boldsymbol{w} + b) \, - 1 \,] \\ &\equiv \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{\boldsymbol{i=1}}^{\boldsymbol{L}} \boldsymbol{a_i} \, y_i \, (\boldsymbol{x}_i \cdot \boldsymbol{w} + b) \, + \sum_{\boldsymbol{i=1}}^{\boldsymbol{L}} \boldsymbol{a_i} \end{aligned} \tag{4.6.3}$$

Θέλουμε να βρούμε το **w** και το b τα οποία ελαχιστοποιούν, και το $\boldsymbol{a}$ το οποίο μεγιστοποιεί την (4.6.2) (κρατώντας τα $a_i \geq 0, \forall i$). Μπορούμε να το κάνουμε αυτό διαφορίζοντας την L_P ως προς το **w** και το b, και θέτοντας τις παραγώγους ίσες με το μηδέν :

$$\frac{\partial L_P}{\partial \mathbf{w}} = 0 \; \Rightarrow \; \mathbf{w} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{L}} \boldsymbol{a_i} \, y_i \, \boldsymbol{x}_i \tag{4.6.4}$$

$$\frac{\partial L_P}{\partial \mathrm{b}} = 0 \; \Rightarrow \; \sum_{\boldsymbol{i=1}}^{\boldsymbol{L}} \boldsymbol{a_i} \, y_i = 0 \tag{4.6.5}$$

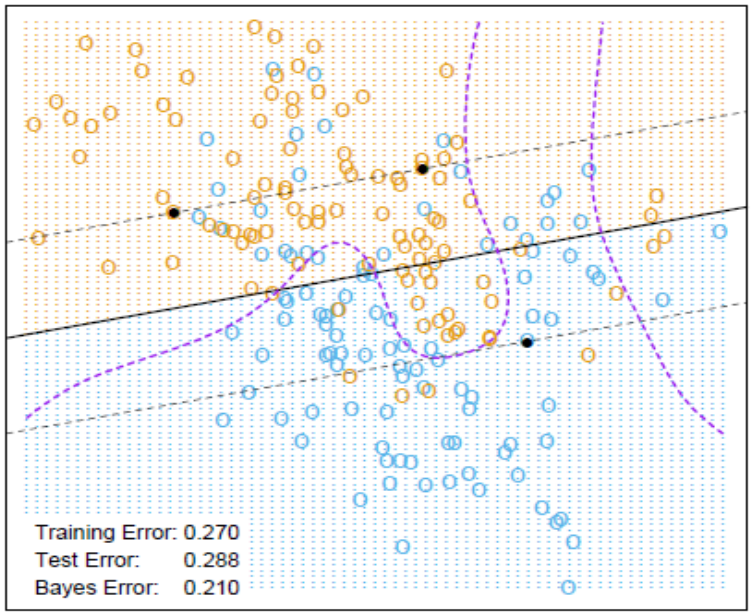
Αντικαθιστώντας την (4.6.4) και (4.6.5) στην (4.6.3) παίρνουμε μία άλλη μορφή η οποία εξαρτάται από το $\boldsymbol{a}$, και τότε πρέπει να μεγιστοποιήσουμε:

$$\begin{aligned} L_D \; &\equiv \sum_{\boldsymbol{i=1}}^{\boldsymbol{L}} \boldsymbol{a_i} \; - \frac{\boldsymbol{1}}{\boldsymbol{2}} \sum_{\boldsymbol{i,j}} a_i a_j \, y_i y_j \, \boldsymbol{x}_i \boldsymbol{x}_j \qquad \text{s.t.} \qquad a_i \geq 0 \;\; \forall i, \sum_{i=1}^{L} a_i \, y_i = \, 0 \\ &\equiv \sum_{\boldsymbol{i=1}}^{\boldsymbol{L}} \boldsymbol{a_i} \; - \frac{\boldsymbol{1}}{\boldsymbol{2}} \sum_{\boldsymbol{i,j}} a_i H_{ij} a_j \qquad \text{όπου} \qquad H_{ij} \equiv \; y_i y_j \, \boldsymbol{x}_i \boldsymbol{x}_j \\ &\equiv \sum_{\boldsymbol{i=1}}^{\boldsymbol{L}} \boldsymbol{a_i} \; - \frac{\boldsymbol{1}}{\boldsymbol{2}} \, \boldsymbol{\alpha^T H \alpha} \qquad \text{s.t.} \qquad a_i \geq 0 \;\; \forall i \, , \;\; \sum_{i=1}^{L} a_i \, y_i = \, 0 \end{aligned} \tag{4.6.6}$$

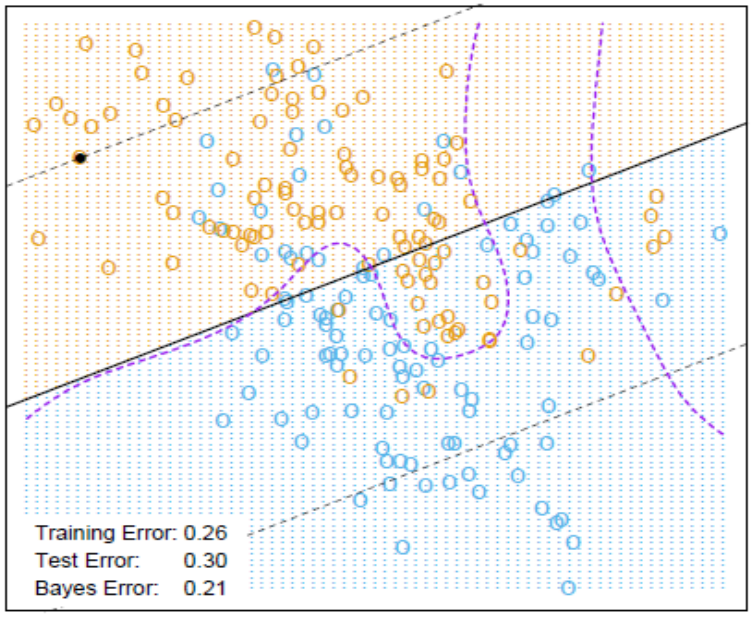
Αυτή η νέα σύνθεση L_D αναφέρεται ως η διπλή μορφή της πρωτοβάθμιας L_P. Αξίζει να σημειωθεί ότι η διπλή μορφή απαιτεί μόνο να υπολογιστεί το γινόμενο όλων των διανυσμάτων εισόδου x_i. Αυτό είναι σημαντικό για το τέχνασμα πυρήνα, και περιγράφεται παρακάτω.

Αφού το πρόβλημα έχει μετατοπιστεί από την ελαχιστοποίηση L_P στη μεγιστοποίηση της L_D, θα πρέπει να βρεθεί:

$$max \left[\sum_{\boldsymbol{i=1}}^{\boldsymbol{L}} \boldsymbol{a_i} \; - \frac{\boldsymbol{1}}{\boldsymbol{2}} \, \boldsymbol{\alpha^T H \alpha} \right] \qquad s.t. \qquad a_i \geq 0 \;\; \forall i \, , \qquad \sum_{i=1}^{L} a_i \, y_i = \, 0$$



$C = 10000$



$C = 0.01$

Σχήμα 16: Το γραμμικό όριο του διανύσματος υποστήριξης για τα δεδομένα του παραδείγματος του κεφαλαίου 2 (mixture example) στο Hastie et al. (2001), με δύο επικαλυπτόμενες κλάσεις, για δύο διαφορετικές τιμές του C. Οι διακεκομμένες γραμμές καταδεικνύουν τα περιθώρια, όπου f(x)=±1. Τα σημεία υποστήριξης (α_i>0) είναι όλα τα σημεία στη λάθος πλευρά του περιθωρίου τους. Οι μαύρες τελείες είναι εκείνα τα σημεία υποστήριξης που είναι ακριβώς στο περιθώριο (ξ_i=0, α_i>0). Στο άνω σχήμα το 62% των παρατηρήσεων είναι σημεία υποστήριξης, ενώ στο κάτω σχήμα είναι το 85%. Η διακκεκομμένη μοβ καμπύλη στο πίσω μέρος είναι το όριο απόφασης του Bayes.

Αυτό είναι ένα κυρτό τετραγωνικό πρόβλημα βελτιστοποίησης και διατρέχουμε μια QP επίλυση η οποία θα επιστρέψει το $\boldsymbol{\alpha}$ και από την (4.6.4) θα μας δώσει το **w**. Αυτό που απομένει είναι να υπολογιστεί το b.

Αντικαθιστώντας στην (4.6.4) την (4.6.5) και χρησιμοποιώντας την (4.6.1) και τέλος παίρνοντας το μέσο όρο όλων των x_s βρίσκουμε το b.

Έχουμε τώρα τις μεταβλητές w και b που ορίζουν το βέλτιστο διαχωριστικό προσ-ανατολισμό του υπερεπιπέδου, και ως εκ τούτου τη μηχανή διανυσματικής υποστήριξης.

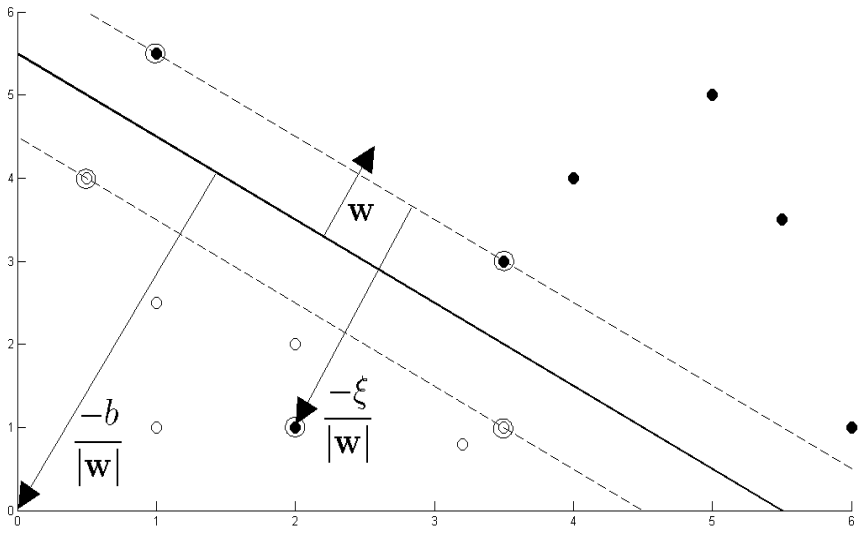
<u>Μη γραμμικά διαχωρίσιμα δεδομένα</u>

Προκειμένου να επεκταθεί η μεθοδολογία SVM για να διαχειριστεί τα δεδομένα που δεν είναι πλήρως γραμμικά διαχωρίσιμα, θα χαλαρώσουμε τους περιορισμούς (4.6.1) για να επιτρέπουν τα ελαφρώς μη ταξινομημένα σημεία. Αυτό γίνεται με την εισαγωγή μιας θετικής χαλαρής μεταβλητής $\xi_i, i = 1, \dots, L$

$$\boldsymbol{x}_i \cdot \boldsymbol{w} + b \geq +1 - \xi_i \quad \gamma\iota\alpha \; y_i = +1 \tag{4.6.7}$$

$$\boldsymbol{x}_i \cdot \boldsymbol{w} + b \leq -1 + \xi_i \quad \gamma\iota\alpha \; y_i = -1$$

$$\xi_i \geq 0 \; \forall i$$



Σχήμα 17 : Υπερεπίπεδο διαμέσου δύο μη γραμμικά διαχωρίσιμων κλάσεων

Αυτές οι εξισώσεις μπορούν να συνδυαστούν ως εξής :

$$y_i\,(\boldsymbol{x}_i \cdot \boldsymbol{w} + b)\; - 1 + \;\xi_i \geq\; 0 \quad \text{όπου} \quad \xi_i \;\geq 0\; \forall i$$

Σε αυτό το «μαλακό» SVM περιθώριο (soft margin[5]), τα σημεία δεδομένων για την εσφαλμένη πλευρά του ορίου περιθωρίου έχουν μια ποινή που αυξάνει με την απόσταση από αυτό. Καθώς προσπαθούμε να μειώσουμε τον αριθμό των μη ταξινομημένων σημείων, ένας λογικός τρόπος για να προσαρμόσουμε την αντικειμενική μας συνάρτηση μας (4.6.2), είναι να βρούμε:

$$\min \tfrac{1}{2}\;\|\boldsymbol{w}\|^2 + C\;\textstyle\sum_{i=1}^{L} \xi_i \qquad s.t. \qquad y_i\,(\boldsymbol{x}_i \cdot \boldsymbol{w} + b)\; - 1 + \;\xi_i \geq\; 0\;\; \forall i$$

όπου η παράμετρος C ελέγχει το trade-off μεταξύ της ποινής της χαλαρής μεταβλητής και του μεγέθους του περιθωρίου. Η αναδιατύπωση ως Lagrangian, η οποία όπως και πριν θα πρέπει να ελαχιστοποιηθεί σε σχέση με τα w, b και ξ_i και να μεγιστοποιηθεί ως προς $\boldsymbol{a}$ (όπου $a_i \;\geq 0,\; \mu_i \;\forall i$):

$$L_P \;\equiv\; \tfrac{1}{2}\|\boldsymbol{w}\|^2 + \;C\;\textstyle\sum_{i=1}^{L} \xi_i - \sum_{\boldsymbol{i=1}}^{\boldsymbol{L}} a_i\,[\,y_i\,(\boldsymbol{x}_i \cdot \boldsymbol{w} + b) - 1 + \;\xi_i\;] - \sum_{i=1}^{L} \mu_i\,\xi_i$$

Διαφορίζοντας την L_P ως προς το **w,** το b και το ξ_i και θέτοντας τις παραγώγους ίσες με το μηδέν:

$$\frac{\partial L_P}{\partial \mathbf{w}} = 0 \;\Rightarrow\; \mathbf{w} = \;\sum_{\boldsymbol{i=1}}^{\boldsymbol{L}} \boldsymbol{a_i}\; y_i\; \boldsymbol{x}_i$$

$$\frac{\partial L_P}{\partial \mathrm{b}} = 0 \;\Rightarrow\; \sum_{\boldsymbol{i=1}}^{\boldsymbol{L}} \boldsymbol{a_i}\; y_i = 0$$

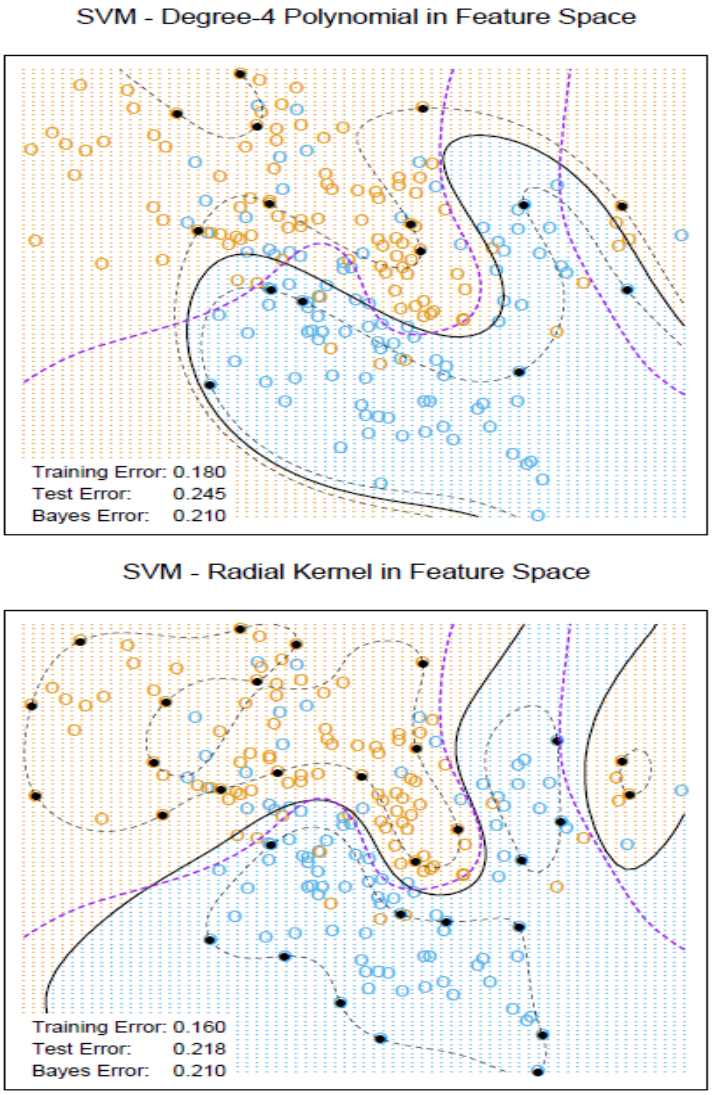
$$\frac{\partial L_P}{\partial \xi_i} = 0 \;\Rightarrow\; C = \;a_i + \;\mu_i \qquad (4.6.8)$$

Αντικαθιστώντας αυτά, η L_P έχει την ίδια μορφή όπως η σχέση (4.6.6) προηγουμένως. Ωστόσο η (4.6.8) μαζί με τα $\mu_i \geq 0$, συνεπάγεται ότι α ≤ C. Επομένως χρειάζεται να βρούμε:

[5] Για να αποτραπεί η υπερβολική χρήση των λάθος τοποθετημένων σημείων, ορίζεται η σταθερά C η οποία θέτει τους όρους για τη μεγιστοποίηση του περιθωρίου και την ελαχιστοποίηση των λάθος κατηγοριοποιήσεων. Η μέθοδος αυτή ονομάζεται soft-margin SVM.

$$max\left[\sum_{i=1}^{L} \boldsymbol{a_i} \; - \frac{1}{2}\, \boldsymbol{\alpha^T H \alpha}\right] \qquad s.t. \qquad 0 \leq a_i \leq C \;\; \forall i, \qquad \sum_{i=1}^{L} a_i\, y_i = 0$$

Στη συνέχεια το b υπολογίζεται με τον ίδιο τρόπο όπως προηγουμένως στην (4.6.2), αν και σε αυτή την περίπτωση το σύνολο των διανυσμάτων υποστήριξης χρησιμοποιείται για τον υπολογισμό του b που προσδιορίζεται με την εύρεση των δεικτών i, όπου $0 \leq a_i \leq C$.



Σχήμα 18: Δύο μη γραμμικά SVMs για τα δεδομένα του παραδείγματος του κεφαλαίου 2 (mixture example) στο Hastie et al. (2001). Το άνω γράφημα χρησιμοποιεί 4^{ου} βαθμού πολυωνιμικό πυρήνα, το κάτω ένα radial basis πυρήνα (με γ = 1). Σε κάθε περίπτωση C ήταν συντονισμένοι για την επίτευξη περίπου της καλύτερη δυνατή απόδοση σφάλματος δοκιμής, και το C=1 λειτούργησε καλά και στις δύο περιπτώσεις. Ο radial basis πυρήνας αποδίδει το καλύτερο (κοντά στο βέλτιστο Bayes), όπως θα ήταν αναμενόμενο δοθέντος των δεδομένων που προκύπτουν από το μείγμα των Gaussians. Η διακκεκομμένη μοβ καμπύλη στο πίσω μέρος είναι το όριο απόφασης του Bayes.

4.6.3 Η SVM μέθοδος για την παλινδρόμηση

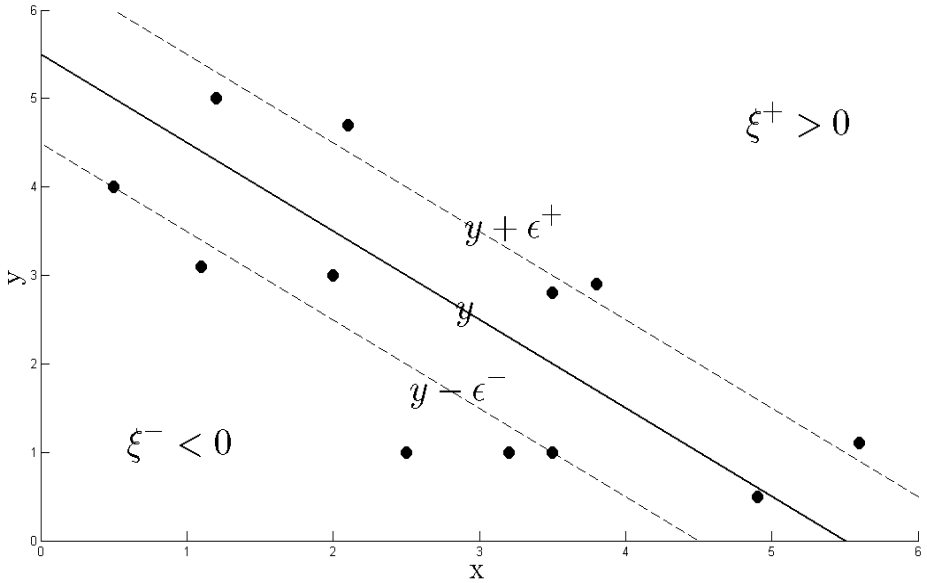
Οι SVMs μπορούν επίσης να εφαρμοστούν σε προβλήματα παλινδρόμησης με την εισαγωγή μιας εναλλακτικής λειτουργίας απώλειας . Η λειτουργία απώλειας πρέπει να τροποποιηθεί ώστε να συμπεριλάβει το μέτρο της απόστασης. Η παλινδρόμηση μπορεί να είναι γραμμική και μη γραμμική. Τα γραμμικά μοντέλα αποτελούνται κυρίως από τις ακόλουθες λειτουργίες απώλειας, εντατικές λειτουργίες απώλειας, τετραγωνική και Huber λειτουργίες απώλειας.

Ομοίως με τα προβλήματα ταξινόμησης, ένα μη γραμμικό μοντέλο συνήθως απαιτεί επαρκή δεδομένα. Με τον ίδιο τρόπο, μια μη γραμμική χαρτογράφηση μπορεί να χρησιμοποιηθεί για να χαρτογραφήσει τα δεδομένα σε ένα υψηλό διαστάσεων χώρο χαρακτηριστικών, όπου η γραμμική παλινδρόμηση εκτελείται. Η προσέγγιση του πυρήνα και πάλι χρησιμοποιείται για την αντιμετώπιση της διάστασης. Στη μέθοδο παλινδρόμησης υπάρχουν εκτιμήσεις που βασίζονται σε προγενέστερη γνώση του προβλήματος και τη διανομή του θορύβου.

Αντί να προσπαθούμε να κατατάξουμε νέες άγνωστες μεταβλητές x΄ σε μία από τις δύο κατηγορίες $y' = \pm 1$, τώρα επιθυμούμε να προβλέψουμε μια πραγματική τιμή εξόδου για το y' και έτσι τα δεδομένα εκπαίδευσης μας είναι της μορφής:

$$\{x_i, y_i\}, \qquad \text{όπου} \qquad i = 1, \dots, L, \quad y_i \in \Re, \quad \mathbf{x} \in \Re^D$$

$$y_i = \boldsymbol{x}_i \cdot \boldsymbol{w} + b \tag{4.6.9}$$



Σχήμα 19 : Παλινδρόμηση με ε- insensitive σωλήνα (tube)

Η SVM παλινδρόμηση θα χρησιμοποιήσει μια πιο εξελιγμένη λειτουργία ποινής από πριν, μη χορηγώντας ποινή εάν η προβλεπόμενη τιμή y_i είναι μικρότερη από απόσταση ε μακριά από την πραγματική τιμή t_i, δηλαδή αν $|t_i - y_i| < \varepsilon$. Αναφερόμενοι στο Σχήμα 19, η περιοχή που οριοθετείται από $y_i \pm \varepsilon$ λέγεται ε - insensitive σωλήνας (tube). Μια άλλη μορφοποίηση στη λειτουργία ποινής είναι ότι οι μεταβλητές εξόδου που είναι εκτός του σωλήνα δίνουν μία από τις δύο χαλαρές μεταβλητές, ποινές ανάλογα με το αν βρίσκονται πάνω (ξ^+) ή κάτω από το σωλήνα (ξ^-) (όπου $\xi^+ > 0,\ \xi^- > 0,\ \forall i$):

$$t_i \leq y_i + \varepsilon + \xi^+$$

$$t_i \geq y_i - \varepsilon - \xi^+ \qquad (4.6.10)$$

Η συνάρτηση σφάλματος για την SVM παλινδρόμηση (SVR) μπορεί να γραφεί ως εξής:

$$C \sum_{i=1}^{L} (\xi_i^+ + \xi_i^-) + \frac{1}{2}\|\boldsymbol{w}\|^2$$

Αυτή χρειάζεται να ελαχιστοποιηθεί υπό τους περιορισμούς $\xi^+ \geq 0$, $\xi^- \geq 0$, $\forall i$ και την (4.6.9) και (4.6.10). Για να το κάνουμε αυτό εισάγουμε πολλαπλασιαστές Lagrange

$\alpha_i^+ \geq 0, \alpha_i^- \geq 0, \mu_i^+ \geq 0, \mu_i^- \geq 0\ ,\ \forall i$:

Με την ίδια διαδικασία δημιουργούμε την L_P , διαφορίζουμε ως προς w, b, ξ^+ και ξ^- και θέτουμε τις παραγώγους ίσες με το μηδέν. Με ανάλογο τρόπο αντικαθιστούμε και βρίσκουμε το L_D το οποίο θέλουμε να μεγιστοποιήσουμε ως προς α_i^+ και α_i^- ($\alpha_i^+ \geq 0$, $\alpha_i^- \geq 0,\ \forall i$).

Με τα βήματα λοιπόν που περιγράφηκαν στα προηγούμενα κεφάλαια βρίσκουμε τις παραμέτρους που χρειαζόμαστε.

4.6.4 Το SVM ως ποινικοποιημένη μέθοδος

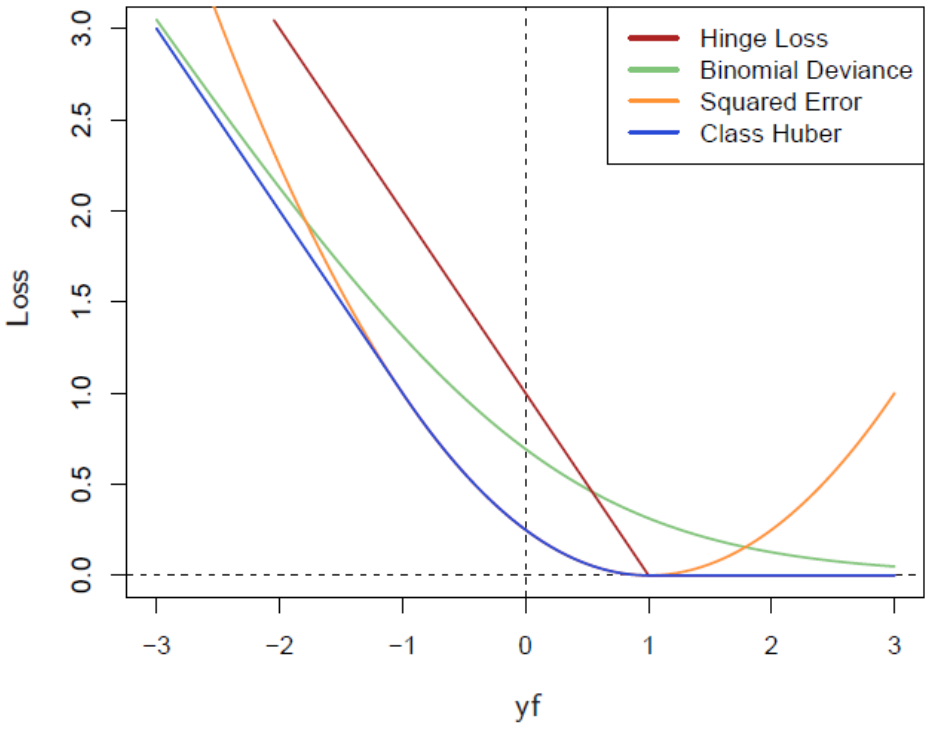
Με το $f(x) = h(x)^T\beta + \beta_0$, θεωρούμε το πρόβλημα βελτιστοποίησης

$$\min_{\beta_0,\beta} \sum_{i=1}^{N} [1 - y_i f(x_i)]_+ + \lambda\|\beta\|^2 \qquad (4.6.11)$$

όπου ο δείκτης "+" υποδηλώνει θετικό τμήμα. Αυτό έχει τη μορφή *απώλεια* + *ποινή*, το οποίο είναι ένα γνώριμο παράδειγμα στην εκτίμηση συναρτήσεων. Είναι εύκολο να δούμε ότι η λύση για την (4.6.10) με $\lambda = \frac{1}{C}$, είναι ίδια με την (4.6.7).

Η εξέταση της «hinge» απώλειας $L(y, f) = [1 - yf]_+$ δείχνει ότι είναι λογική για την ταξινόμηση δύο τάξεων, όταν συγκρίνεται με άλλες πιο παραδοσιακές συναρτήσεις απώλειας.

Το Σχήμα 20 συγκρίνει την απώλεια λογαριθμοπιθανοφάνειας για τη λογιστική παλινδρόμηση, καθώς και την απώλεια του τετραγωνικού σφάλματος και μια παραλλαγή αυτών. Η (αρνητική) λογαριθμοπιθανοφάνεια ή διωνυμική απόκλιση έχει παρόμοιες ουρές, όπως η SVM απώλεια (hinge loss[6]), δίνοντας μηδενική ποινή σε σημεία που είναι καλά μέσα στο περιθώριο τους, και μια γραμμική ποινή στα σημεία που είναι στη λάθος πλευρά και πολύ μακριά. Το σφάλμα τετραγώνων, από την άλλη πλευρά δίνει μια τετραγωνική ποινή, καθώς και τα σημεία μέσα στο ίδιο τους το περιθώριο έχουν επίσης μια ισχυρή επιρροή στο μοντέλο.



Σχήμα 20: Η συνάρτηση απώλειας των διανυσμάτων υποστήριξης (hinge loss), σε σύγκριση με την αρνητική λαγοριθμοπιθανοφάνεια απώλεια (διωνυμική απόκλιση) για τη λογιστική παλινδρόμηση, η απώλεια τετραγωνικού σφάλματος, και μία "Huberized" εκδοχή της τετραγωνικής hinge loss. Όλα εμφανίζονται σαν συνάρτηση του yf αντί του f, λόγω της συμμετρίας μεταξύ της περίπτωσης y=1 και της y= -1. Η απόκλιση και η Huber έχουν τις ίδιες

[6] Στη μηχανική μάθηση, η «hinge loss» είναι μια συνάρτηση που χρησιμοποιείται για την εκπαίδευση των ταξινομητών . Η «hinge loss» χρησιμοποιείται για «μεγιστοποίηση του περιθωρίου" ταξινόμησης, κυρίως για τις μηχανές διανυσματικής υποστήριξης (SVMs).

ασύμπτωτες με την απώλεια του SVM, αλλά έχουν στρογγυλοποιηθεί στο εσωτερικό. Όλα κλιμακώνονται να έχουν τον περιορισμό της κλίσης της αριστερής-ουράς του -1.

Η τετραγωνική hinge απώλεια $L(y,f) = [1 - yf]^2_+$ είναι σαν την τετραγωνική, εκτός του ότι είναι μηδέν για τα σημεία μέσα στο περιθώριο τους. Αυξάνεται ακόμα τετραγωνικά στην αριστερή ουρά, και θα είναι λιγότερο εύρωστη από ότι η hinge ή η απόκλιση στο να ταξινομηθεί εσφαλμένα παρατηρήσεις. Πρόσφατα οι Rosset και Zhu (2007) πρότειναν μια "Huberized" έκδοση της τετραγωνικής hinge απώλειας, η οποία μετατρέπει ομαλά σε μία γραμμική απώλεια στην $yf = -1$. Μπορούμε να χαρακτηρίσουμε αυτές τις λειτουργίες απώλειας από την άποψη του τι εκτιμούν σε επίπεδο πληθυσμού. Θεωρούμε την ελαχιστοποίηση $EL(Y, f\ (X))$.

Ο Πίνακας 12.1 συνοψίζει τα αποτελέσματα. Ενώ η hinge απώλεια εκτιμά τον ταξινομητή $G\ (x)$ η ίδια, όλες οι άλλες εκτιμούν ένα μετασχηματισμό της τάξης εκ των υστέρων πιθανότητας. Η "Huberized" τετραγωνική hinge απώλεια δίνει ελκυστικές ιδιότητες της λογιστικής παλινδρόμησης (ομαλή συνάρτηση απώλειας, εκτιμήσεις πιθανοτήτων), όπως και η SVM hinge απώλεια (σημεία υποστήριξης).

Loss Function	$L[y, f(x)]$	Minimizing Function
Binomial Deviance	$\log[1 + e^{-yf(x)}]$	$f(x) = \log \frac{\Pr(Y = +1\|x)}{\Pr(Y = -1\|x)}$
SVM Hinge Loss	$[1 - yf(x)]_+$	$f(x) = \text{sign}[\Pr(Y = +1\|x) - \frac{1}{2}]$
Squared Error	$[y - f(x)]^2 = [1 - yf(x)]^2$	$f(x) = 2\Pr(Y = +1\|x) - 1$
"Huberised" Square Hinge Loss	$-4yf(x)$, $yf(x) < -1$ $[1 - yf(x)]^2_+$ otherwise	$f(x) = 2\Pr(Y = +1\|x) - 1$

Σχήμα 21: Οι minimizers του πληθυσμού για τις διαφορετικές συναρτήσεις απώλειας στο σχήμα 20. Η λογιστική παλινδρόμηση χρησιμοποιεί τη διωνυμική λογαριθμοπιθανοφάνεια ή απόκλιση. Η γραμμική διακριτή ανάλυση χρησιμοποιεί την απώλεια του τετραγωνικού-σφάλματος. Η hinge απώλεια του SVM εκτιμά τη λειτουργία των εκ των υστέρων πιθανοτήτων, ενώ οι άλλες εκτιμούν ένα γραμμικό μετασχηματισμό αυτών των πιθανοτήτων.

Ο τύπος (4.6.10) ρίχνει το SVM ως τακτοποιημένο πρόβλημα εκτίμησης συνάρτησης, όπου οι συντελεστές της γραμμικής επέκτασης $f\ (x)\ =\ \beta_0\ +\ h\ (x)^T\ \beta$ έχουν συρρικνωθεί προς το μηδέν (εκτός από τη σταθερά). Αν το h (x) παριστάνει μια ιεραρχική βάση έχοντας κάποια διατεταγμένη δομή (όπως η διάταξη στην τραχύτητα), τότε η ομοιόμορφη συρρίκνωση γίνεται πιο λογική αν το σκληρότερο h_j στο διάνυσμα h έχει μικρότερη νόρμα. Όλες οι συναρτήσεις απώλειες του σχήματος 12.1 τετραγωνικό-σφάλμα είναι οι λεγόμενες "μεγιστοποιημένες συναρτήσεις απώλειας περιθωρίου" (Rosset et al., 2004b).

Αυτό σημαίνει ότι εάν τα δεδομένα είναι διαχωρίσιμα, τότε το όριο του β_λ στην (4.6.10), καθώς $\lambda \to 0$ καθορίζει το βέλτιστο διαχωριστικό υπερεπίπεδο.

Γενική L1-νόρμα στις μηχανές διανυσματικής υποστήριξης για την επιλογή χαρακτηριστικών

Έχει αποδειχθεί από τους Nguyen et al.(2011) ότι η παραδοσιακή L1-νόρμα SVM που προτάθηκε από τους Bradley και Mangasarian (1998) μπορεί να γενικευθεί σε μια γενική L1-νόρμα SVM (GL1-SVM).

Επιπλέον, έχει αποδειχθεί ότι η επίλυση του νέου προτεινόμενου προβλήματος βελτιστοποίησης (GL1-SVM) δίνει μικρότερη ποινή σφάλματος και διευρύνει το περιθώριο μεταξύ των δύο υπερεπιπέδων διανυσμάτων υποστήριξης, έτσι δίνει ίσως την καλύτερη ικανότητα γενίκευσης των SVM από την επίλυση της παραδοσιακής -L1 νόρμας SVM.

GL1-SVM μπορεί επίσης να αντιμετωπιστεί ως μια ειδική περίπτωση ορισμένων γενικών επιλογέων χαρακτηριστικών (Nguyen et al. (2010)).

4.6.5 Πυρήνες

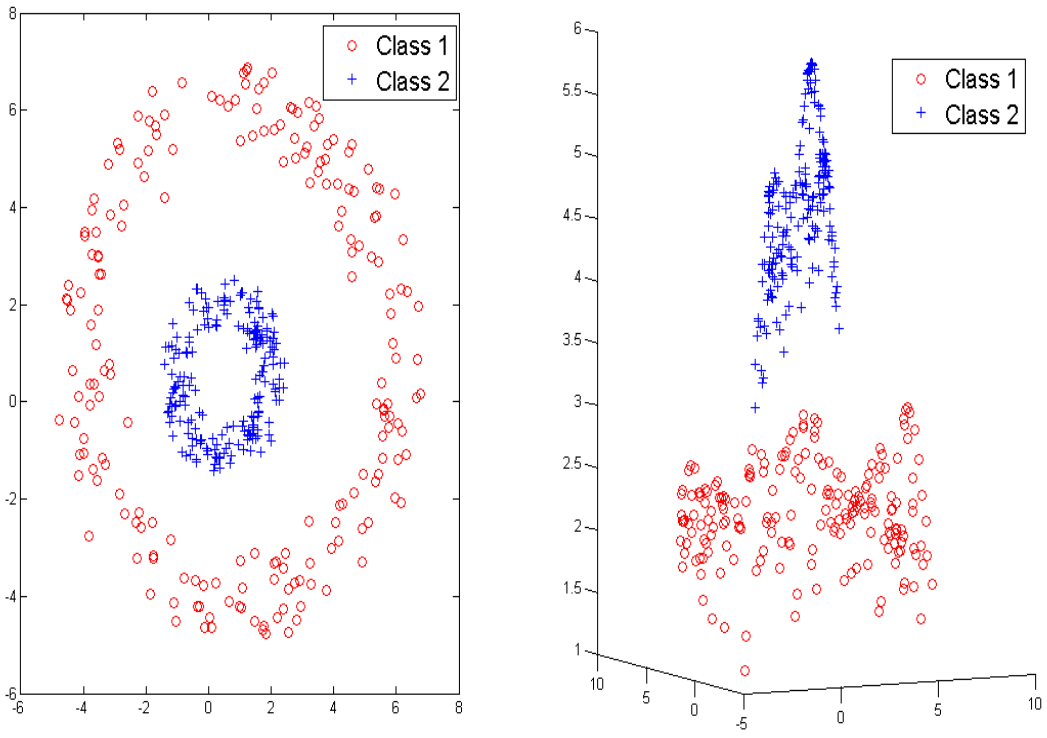
Οι μέθοδοι των πυρήνων είναι μία πολύ δημοφιλής και επιτυχημένη περιοχή της μηχανικής μάθησης. Η κοινή βάση τους είναι το αποκαλούμενο κόλπο του πυρήνα (kernel trick), το οποίο μπορεί να εφαρμοστεί σε οποιονδήποτε γραμμικό αλγόριθμο ο οποίος βασίζεται μόνο στα δεδομένα από την άποψη των εσωτερικών γινομένων μεταξύ δύο παραδειγμάτων.

Κατά την εφαρμογή της SVM τεχνικής για γραμμικά διαχωρίσιμα δεδομένα είχαμε ξεκινήσει δημιουργώντας ένα πίνακα Η από το γινόμενο των μεταβλητών εισόδου:

$$H_{ij}\ \equiv\ y_i y_j\ k\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) =\ \boldsymbol{x}_i\ \cdot\ \boldsymbol{x}_j =\ \boldsymbol{x}_i^{\boldsymbol{T}}\ \boldsymbol{x}_j \qquad (4.6.12)$$

Η $k\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right)$ είναι ένα παράδειγμα μιας οικογένειας συναρτήσεων που ονομάζονται συναρτήσεις πυρήνα (kernel functions) (ο $k\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) = \boldsymbol{x}_i^{\boldsymbol{T}}\ \boldsymbol{x}_j$ είναι γνωστός ως γραμμικός πυρήνας). Το σύνολο των συναρτήσεων του πυρήνα αποτελείται από παραλλαγές του (4.6.13) με την έννοια ότι όλα βασίζονται στον υπολογισμό εσωτερικών γινομένων των δύο διανυσμάτων.

Αυτό σημαίνει ότι αν οι λειτουργίες μπορούν να αναδιατυπωθούν σε ένα χώρο υψηλότερης διάστασης από κάποια πιθανά μη-γραμμικά χαρακτηριστικά χαρτογράφησης της συνάρτησης $x \rightarrow \varphi(x)$, μόνο τα εσωτερικά γινόμενα της αντίστοιχης εισόδου στο χώρο των χαρακτηριστικών χρειάζεται να καθοριστούν, χωρίς να χρειάζεται να υπολογιστεί ρητά η φ. Ο λόγος που αυτό το τέχνασμα του πυρήνα είναι χρήσιμο είναι ότι υπάρχουν πολλά προβλήματα ταξινόμησης/ παλινδρόμησης που δεν είναι γραμμικά διαχωρίσιμα στο χώρο των εισόδων x, τα οποία μπορεί να είναι σε ένα υψηλότερης διάστασης χώρο χαρακτηριστικών δεδομένης μιας κατάλληλης χαρτογράφησης $x \rightarrow \varphi(x)$. Για περισσότερες λεπτομέρειες σχετικά με τις συναρτήσεις πυρήνων στην ταξινόμηση παραπέμπουμε στον Herbrich (2002).



Σχήμα 22: Διχοτόμηση δεδομένων, ανασχηματισμός με τη χρήση του πυρήνα RBF.

Αναφερόμενοι στο Σχήμα 22, αν ορίσουμε τον πυρήνα μας να είναι :

$$k\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) = e^{-\left(\frac{\left\|x_i - x_j\right\|^2}{2\sigma^2}\right)} \quad (4.6.13)$$

τότε ένα σύνολο δεδομένων το οποία δεν είναι γραμμικά διαχωρίσιμο σε δισδιάστατο χώρο δεδομένων **x** (όπως στην αριστερή πλευρά του Σχήματος 22) μπορεί να διαχωριστεί στο μη γραμμικό χώρο των χαρακτηριστικών (δεξιά πλευρά του Σχήματος 22) έμμεσα από αυτή τη μη- γραμμική συνάρτηση πυρήνα - γνωστή ως Ακτινική Βάση Πυρήνα (Radial Basis Kernel).

Άλλοι δημοφιλείς πυρήνες για ταξινόμηση και παλινδρόμηση είναι ο πολυωνυμικός πυρήνας (Polynomial Kernel)

$$k\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) = \left(\boldsymbol{x}_i \cdot \boldsymbol{x}_j + a\right)^{\boldsymbol{b}}$$

και ο σιγμοειδής πυρήνας

$$k\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) = \tanh(a\boldsymbol{x}_i \cdot \boldsymbol{x}_j - b)$$

όπου a και b είναι οι παράμετροι που καθορίζουν τη συμπεριφορά του πυρήνα.

Υπάρχουν πολλές συναρτήσεις πυρήνα, συμπεριλαμβανομένων και εκείνων που ενεργούν πάνω στα σύνολα, στις χορδές και ακόμα και στη μουσική. Υπάρχουν συγκεκριμένες απαιτήσεις για μία συνάρτηση έτσι ώστε να μπορεί να χρησιμοποιηθεί ως συνάρτηση του πυρήνα, που βρίσκονται πέρα από το πεδίο εφαρμογής της παρούσας πολύ σύντομης εισαγωγής στην περιοχή.

4.6.6 Μέθοδοι επιλογής μοντέλου/παραμέτρων για τις μηχανές διανυσματικής υποστήριξης

Η απόδοση των μηχανών διανυσματικής υποστήριξης (SVM) επηρεάζεται σημαντικά από τις παραμέτρους του μοντέλου. Μία κοινώς χρησιμοποιούμενη μέθοδος επιλογής παραμέτρων SVM, είναι το πλέγμα αναζήτησης (GS), η οποία είναι πολύ χρονοβόρα. Έχουν γίνει αρκετές προσπάθειες από διάφορους ερευνητές για την μείωση του υπολογιστικού κόστους. Οι Ou et al. (2003) πρότειναν ένα μηχανισμό για τη μείωση των δεδομένων με σκοπό την επίσπευση της διαδικασίας επιλογής μοντέλου στην SVM μέθοδο. Τα πειραματικά αποτελέσματα δείχνουν ότι ο προτεινόμενος μηχανισμός είναι σε θέση να μειώσει σημαντικά το χρόνο για να πραγματοποιηθεί η επιλογή μοντέλου με το ελάχιστο κόστος. Τον επόμενο χρόνο, οι Zhu et al. (2004), μέσω της εισαγωγής ενός ενιαίου

σχεδιασμού (UD, uniform design) αλλά και με τη χρήση της μεθόδου παλινδρόμησης των μηχανών διανυσματικής υποστήριξης (SVR) κατάφεραν να μειώσουν το κόστος υπολογισμού της παραδοσιακής μεθόδου GS. Μία άλλη προσέγγιση του ίδιου προβλήματος έγινε από τους G.Lebrun et. al (2006) όπου προτείνεται μια νέα μέθοδος μάθησης για την κατασκευή μιας δίτιμης συνάρτησης αποφάσεων (Binary Decision function (BDF)) στις μηχανές διανυσματικής υποστήριξης (SVMs) μειώνοντας την πολυπλοκότητα και καθιστώντας αποτελεσματική τη γενίκευση. Στόχος είναι η κατασκευή ενός γρήγορου και αποτελεσματικού SVM ταξινομητή. Οι Hwang, et al. (2007) πρότειναν έναν ένθετο σχεδιασμό ενιαίων μεθοδολογιών (UD, uniform design) για την αποτελεσματική, ισχυρή (robust) και αυτόματη επιλογή μοντέλου για τις μηχανές διανυσματικής υποστήριξης (SVMs). Η προτεινόμενη μέθοδος εφαρμόζεται για να επιλεγεί το σύνολο των υποψήφιων συνδυασμών παραμέτρων και εκτελείται ένα k-fold cross-validation για να αξιολογηθεί η γενικευμένη απόδοση του κάθε συνδυασμού παραμέτρων.

Κεφάλαιο 5 :

Αξιολόγηση μοντέλου

5.1 Εισαγωγή

Η γενικευμένη απόδοση της μεθόδου εκμάθησης σχετίζεται με την ικανότητα της πρόβλεψης σε ανεξάρτητα δεδομένα δοκιμών (test data). Η αξιολόγηση της απόδοσης είναι εξαιρετικά σημαντικό στην πράξη, δεδομένου ότι κατευθύνει την επιλογή της μεθόδου μάθησης ή το μοντέλο, και μας δίνει ένα μέτρο της ποιότητας του τελικώς επιλεγμένου μοντέλου. Σε αυτό το κεφάλαιο θα περιγράψουμε κάποιες μεθόδους για την αξιολόγηση της απόδοσης.

5.2 Μεροληψία, Διασπορά και περιπλοκότητα μοντέλου

Το Σχήμα 23 απεικονίζει το σημαντικό ζήτημα στην εκτίμηση της δυνατότητας της μεθόδου μάθησης για να γενικευθεί. Θεωρούμε την περίπτωση μιας ποσοτικής μεταβλητής απόκρισης. Έχουμε μια μεταβλητή Y-στόχο, ένα διάνυσμα από εισόδους X, και μία πρόβλεψη του μοντέλου $\hat{f}(X)$ που έχει υπολογιστεί από ένα σετ εκπαίδευσης T. Η λειτουργία απώλειας για τη μέτρηση των σφαλμάτων μεταξύ Y και $\hat{f}(X)$ συμβολίζεται με $L\,(Y, \hat{f}(X))$. Τυπικές επιλογές είναι:

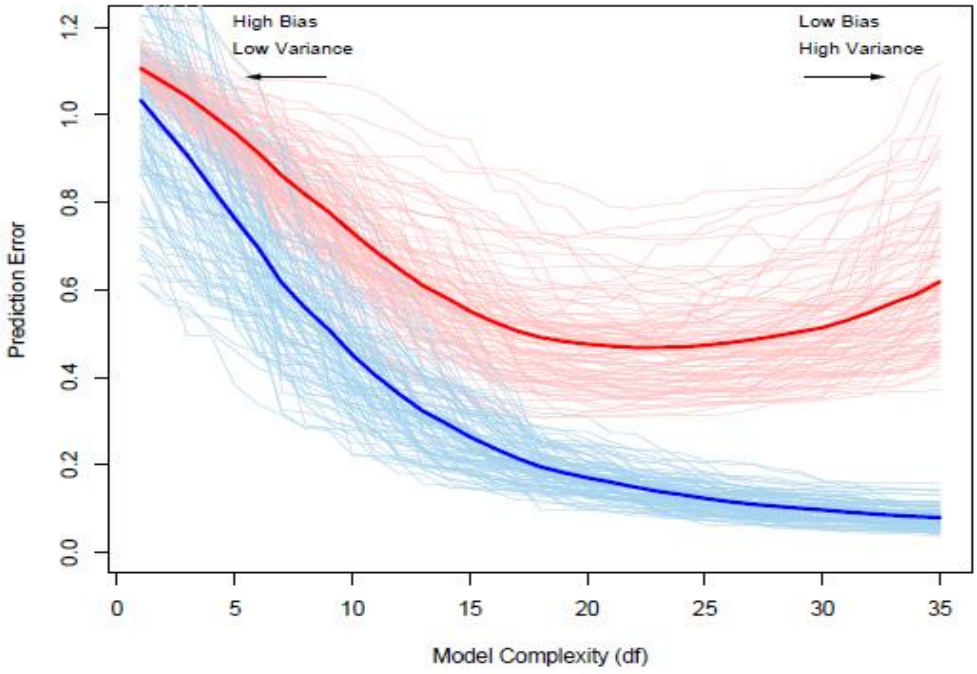
$$L\left(Y, \hat{f}(X)\right) = \begin{cases} (Y - \hat{f}(X))^2 & squared\ error \\ \left|Y - \hat{f}(X)\right| & absolute\ error \end{cases}$$

Το σφάλμα δοκιμών (test error), που αναφέρεται επίσης ως σφάλμα γενίκευσης (generalization error), είναι το σφάλμα πρόβλεψης πάνω σ' ένα ανεξάρτητο σύνολο δοκιμών (test sample):

$$Err_T = E[\, L\, \left(Y, \hat{f}(X)\right) \Big|\, T\,]$$

όπου και οι δύο X και Y επιλέγονται τυχαία από την κοινή τους κατανομή (πληθυσμό). Εδώ το T σύνολο εκπαίδευσης είναι σταθερό, και το σφάλμα δοκιμής αναφέρεται στο σφάλμα για αυτό το συγκεκριμένο σύνολο εκπαίδευσης. Μία σχετική ποσότητα είναι το αναμενόμενο σφάλμα πρόβλεψης ή αλλιώς το αναμενόμενο σφάλμα της δοκιμής (expected prediction error / expected test error):

$$Err = E[L\, \left(Y, \hat{f}(X)\right)] = E[Err_T]$$



Σχήμα 23: Συμπεριφορά του σφάλματος του συνόλου δοκιμών και του συνόλου εκπαίδευσης καθώς ποικίλει η περιπλοκότητα του μοντέλου. Οι ανοιχτόχρωμες μπλε καμπύλες δείχνουν το σφάλμα της εκπαίδευσης err, ενώ οι ανοιχτόχρωμες κόκκινες καμπύλες δείχνουν το υποθετικό σφάλμα δοκιμών Err_T για 100 σετ εκπαίδευσης μεγέθους 50 το καθένα, καθώς η πολυπλοκότητα του μοντέλου μεγαλώνει. Οι συμπαγείς καμπύλες δείχνουν το αναμενόμενο σφάλμα δοκιμών Err και το αναμενόμενο σφάλμα εκπαίδευσης E(err)

Σημειώνουμε ότι αυτός ο αναμενόμενος μέσος όρος είναι πάνω σε ότι είναι τυχαίο, συμπεριλαμβανομένης και της τυχαιότητας στο σύνολο εκπαίδευσης που παρήγαγε την $\hat{f}(X)$. Το σχήμα 23 παρουσιάζει το σφάλμα πρόβλεψης (ανοιχτόχρωμες κόκκινες καμπύλες)

Err_T για 100 προσομοιωμένα σετ εκπαίδευσης το κάθε ένα μεγέθους 50. Η μέθοδος Lasso[7] είναι εκείνη που χρησιμοποιήθηκε για να παράγει την ακολουθία που ταιριάζει. Η συμπαγής κόκκινη καμπύλη είναι ο μέσος όρος, και ως εκ τούτου, η εκτίμηση του Err. Η εκτίμηση του Err_T θα είναι ο στόχος μας, αν και θα δούμε ότι το Err είναι πιο επιδεκτικό σε στατιστική ανάλυση, και οι περισσότερες μέθοδοι εκτιμούν αποτελεσματικά το αναμενόμενο σφάλμα. Επίσης, δεν φαίνεται να είναι δυνατό να εκτιμηθεί αποτελεσματικά το υποθετικό σφάλμα (conditional error), αφού δίνεται μόνο η πληροφορία για το ίδιο σύνολο εκπαίδευσης.

Το σφάλμα εκπαίδευσης είναι η μέση απώλεια πάνω στο δείγμα εκπαίδευσης

$$\overline{err} = \frac{1}{N}\sum_{i=1}^{N} L\left(y_i, \hat{f}(x_i)\right)$$

Θα θέλαμε να γνωρίζουμε το αναμενόμενο σφάλμα (test error) της μέτρησης του εκτιμώμενου μοντέλου μας $\hat{f}$. Καθώς το μοντέλο γίνεται όλο και πιο πολύπλοκο, χρησιμοποιεί τα δεδομένα εκπαίδευσης περισσότερο και είναι σε θέση να προσαρμοστούν σε περισσότερο πολύπλοκες υποκείμενες δομές. Ως εκ τούτου, υπάρχει μια μείωση στην μεροληψία αλλά αύξηση στη διασπορά διακύμανση. Υπάρχει κάποιο ενδιάμεση πολυπλοκότητα ενός μοντέλου που δίνει το ελάχιστο αναμενόμενο σφάλμα δοκιμών (test error). Δυστυχώς το σφάλμα εκπαίδευσης δεν είναι μια καλή εκτίμηση του σφάλματος δοκιμής, όπως φαίνεται στο Σχήμα 23. Το σφάλμα εκπαίδευσης μειώνεται σταθερά με την πολυπλοκότητα του μοντέλου, συνήθως πέφτει στο μηδέν εάν αυξηθεί αρκετά η πολυπλοκότητα του μοντέλου. Ωστόσο, ένα μοντέλο με μηδενικό σφάλμα εκπαίδευσης είναι overfit στην εκπαίδευση δεδομένων και τυπικά δεν θα γενικευθεί καλά. Η ιστορία είναι παρόμοια για μια ποιοτική ή κατηγορηματική μεταβλητή απόκρισης.

Στο κεφάλαιο αυτό περιγράφουμε ένα αριθμό μεθόδων για την εκτίμηση του αναμενόμενου σφάλματος δοκιμών (test error) για ένα μοντέλο. Συνήθως το μοντέλο μας θα έχει μια ρυθμιστική παράμετρο ή παραμέτρους και έτσι μπορούμε να γράψουμε τις προβλέψεις μας ως $f_a(x)$. Η παράμετρος ρύθμισης διαφέρει ανάλογα με την πολυπλοκότητα του μοντέλου μας, και θέλουμε να βρούμε την τιμή αυτού που ελαχιστοποιεί σφάλμα, δηλαδή αυτού που παράγει την ελάχιστη καμπύλη του σφάλματος δοκιμών (test error) στο σχήμα 23.

Είναι σημαντικό να σημειωθεί ότι υπάρχουν στην πραγματικότητα δύο ξεχωριστοί στόχοι που θα μπορούσαμε να έχουμε κατά νου:

[7] Η Lasso είναι μια μέθοδος συρρίκνωσης όπως η μέθοδος κορυφογραμμής, με λεπτές αλλά σημαντικές διαφορές.

Επιλογή Μοντέλου (Model selection): εκτίμηση της απόδοσης των διαφορετικών μοντέλων για να επιλέξουμε το καλύτερο.

Εκτίμηση Μοντέλου (Model assessment): έχοντας επιλέξει ένα τελικό μοντέλο, εκτίμηση του σφάλματος πρόβλεψης του μοντέλου αυτού (σφάλμα γενίκευσης) για τα νέα δεδομένα.

Αν έχουμε ένα αρκετά μεγάλο σύνολο δεδομένων, η καλύτερη προσέγγιση για τα δύο προβλήματα είναι να διαιρέσουμε το σύνολο δεδομένων τυχαία σε τρία μέρη: ένα σύνολο εκπαίδευσης (*training set*), ένα σύνολο επικύρωσης (*validation set*), και ένα σύνολο δοκιμών (test set). Το σύνολο εκπαίδευσης χρησιμοποιείται για να προσαρμόσει τα μοντέλα; το σύνολο επικύρωσης χρησιμοποιείται για την εκτίμηση του σφάλματος πρόβλεψης (prediction error) για την επιλογή μοντέλου και τέλος, το σύνολο δοκιμής χρησιμοποιείται για την εκτίμηση του σφάλματος γενίκευσης (generalization error) του τελικού επιλεχθέντος μοντέλου. Ιδανικά, το σετ δοκιμής θα πρέπει να κρατείται σε απομονωμένο και να το φέρνουμε στην επιφάνεια μόνο στο τέλος της ανάλυσης των δεδομένων. Ας υποθέσουμε ότι αντί να χρησιμοποιούμε το test-set επανειλημμένα, επιλέγουμε το μοντέλο με το μικρότερο σφάλμα δοκιμών (test error). Τότε το σφάλμα δοκιμής του τελικά επιλεγμένου μοντέλου θα υποτιμήσει το πραγματικό σφάλμα της δοκιμής (test error), μερικές φορές σε σημαντικό βαθμό. Είναι δύσκολο να δοθεί ένας γενικός κανόνας για το πώς να επιλέξουμε τον αριθμό των παρατηρήσεων σε κάθε ένα από τα τρία μέρη, καθώς αυτό εξαρτάται από την αναλογία signal-tonoise στα δεδομένα και το μέγεθος του δείγματος εκπαίδευσης. Μια τυπική διάσπαση θα μπορούσε να είναι 50% για την εκπαίδευση, και 25% για κάθε ένα από τα σύνολα δοκιμής και επικύρωσης:

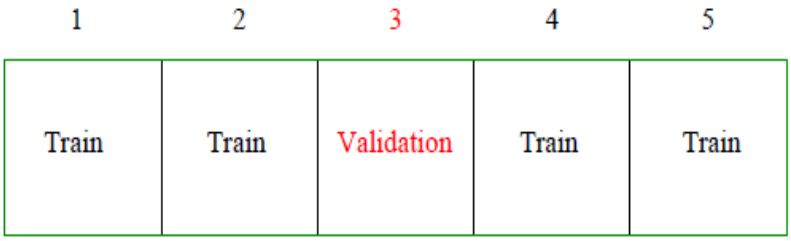
Συνήθως έχουμε περιπτώσεις όπου υπάρχουν ανεπαρκή στοιχεία για να χωριστούν τα δεδομένα σε τρία μέρη. Και πάλι είναι πάρα πολύ δύσκολο να δοθεί ένας γενικός κανόνας σχετικά με κατά πόσο ο όγκος των δεδομένων εκπαίδευσης είναι αρκετός; μεταξύ άλλων, αυτό εξαρτάται από την αναλογία signal-to-noise ratio σήματος προς θόρυβο της υποκείμενης λειτουργίας, και η πολυπλοκότητα των μοντέλων που ταιριάζουν με τα δεδομένα.

Η προσέγγιση στο στάδιο της επικύρωσης γίνεται είτε αναλυτικά (AIC, BIC, MDL, SRM) ή από την αποτελεσματική επαναχρησιμοποίηση του δείγματος (cross-validation και η bootstrap). Εκτός από τη χρήση αυτών των μεθόδων στην επιλογή μοντέλου, μπορούμε επίσης να εξετάσουμε σε ποιο βαθμό κάθε μέθοδος παρέχει μια αξιόπιστη εκτίμηση του σφάλματος δοκιμής του τελικά επιλεχθέντος μοντέλου.

5.3 Διασταυρωμένη επικύρωση (Cross-validation)

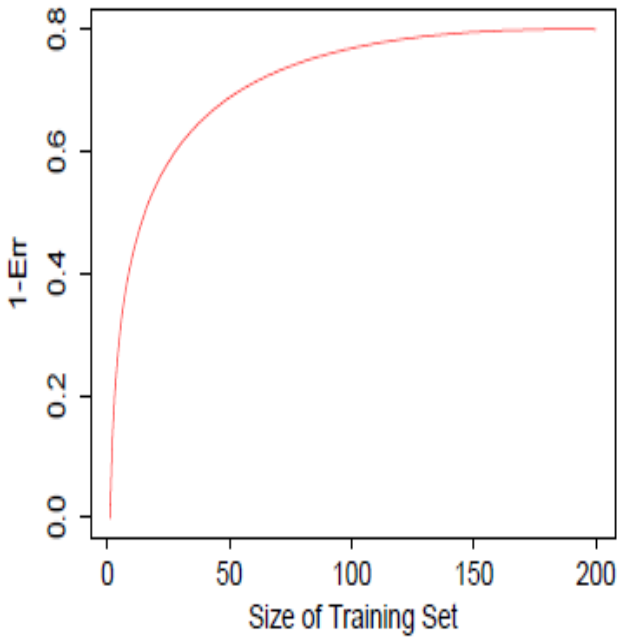
Πιθανώς η απλούστερη και πιο ευρέως χρησιμοποιούμενη μέθοδος για την εκτίμηση του σφάλματος πρόβλεψης (prediction error) είναι η διασταυρωμένη επικύρωση. Αυτή η μέθοδος υπολογίζει άμεσα το αναμενόμενοο εξω-δείγματος σφάλμα $Err = E\left[L\left(Y, \hat{f}(X)\right)\right]$ το μέσο σφάλμα γενίκευσης (generalization error) όταν η μέθοδος $\hat{f}(X)$ εφαρμόζεται σε ένα ανεξάρτητο δείγμα δοκιμής από την κοινή κατανομή των X και Y. Όπως αναφέρθηκε προηγουμένως, μπορούμε να ελπίζουμε ότι η διασταυρωμένη επικύρωση εκτιμά το υπό όρους σφάλμα (conditional error), με το σύνολο εκπαίδευσης T που έχει καθοριστεί. Αλλά, η διασταυρωμένη επικύρωση συνήθως είναι καλή υπολογιστικά μόνο για το αναμενόμενο σφάλμα πρόβλεψης (prediction error).

Στην ιδανική περίπτωση, αν είχαμε αρκετά δεδομένα, θα αναιρέσουμε το σύνολο επικύρωσης και θα το χρησιμοποιήσουμε για να αξιολογήσουμε την απόδοση του μοντέλου που προβλέψαμε. Δεδομένου ότι τα δεδομένα συχνά σπανίζουν, αυτό δεν είναι συνήθως δυνατό. Για την φινέτσα του προβλήματος, η K-φορές διασταυρωμένη επικύρωση χρησιμοποιεί μέρος των διαθέσιμων δεδομένων για να προσαρμόσει το μοντέλο, και ένα διαφορετικό μέρος για να το δοκιμάσει. Έχουμε χωρίσει τα δεδομένα σε K τμήματα, περίπου ίσου μεγέθους, για παράδειγμα, όταν K = 5, το σενάριο μοιάζει με το ακόλουθο:



Για το k-οστό τμήμα (τρίτο στο παραπάνω σχήμα), προσαρμόζουμε το μοντέλο με τα άλλα K-1 μέρη των δεδομένων, και υπολογίζουμε το σφάλμα πρόβλεψης (prediction error) του προσαρμοσμένου μοντέλου όταν προβλέπουμε το κ-οστό τμήμα των δεδομένων. Το κάνουμε αυτό για k = 1, 2,. . . , K και συνδυάζουμε τις K εκτιμήσεις του σφάλματος πρόβλεψης (prediction error).

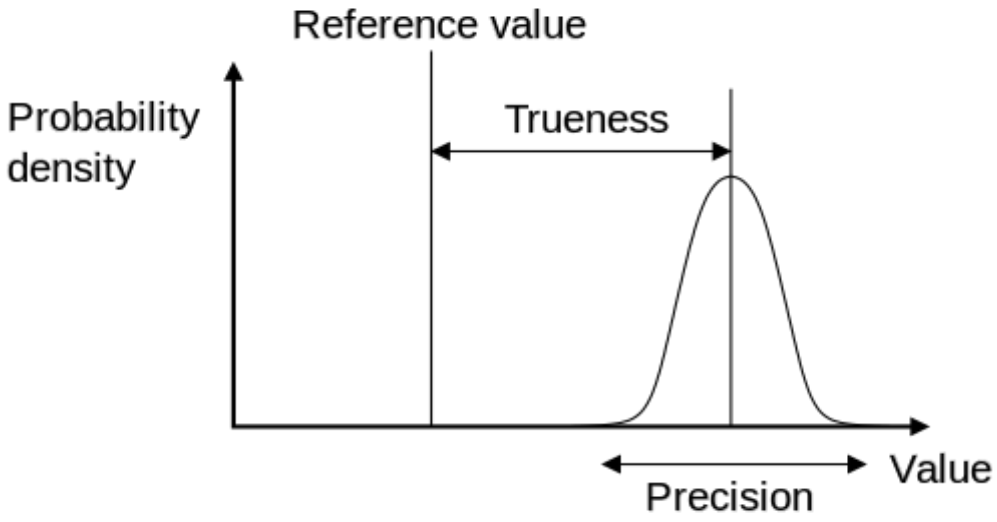
Τι τιμή θα πρέπει να επιλέξουμε για το K; Με K = N, ο εκτιμητής της διασταυρωμένης επικύρωσης είναι περίπου αμερόληπτος για το αληθές (αναμενόμενο) σφάλμα πρόβλεψης (prediction error), αλλά μπορεί να έχει μεγάλη διασπορά, επειδή τα N "σύνολα εκπαίδευσης" είναι τόσο όμοια το ένα στο άλλο. Η υπολογιστική επιβάρυνση είναι επίσης σημαντική, απαιτώντας N εφαρμογές της μεθόδου εκμάθησης.



Σχήμα 24: Υποθετική καμπύλη μάθησης για έναν ταξινομητή σε ένα συγκεκριμένο έργο: ένα γράφημα 1 - Err σε σχέση με το μέγεθος του συνόλου εκπαίδευσης N. Με ένα σύνολο δεδομένων από 200 παρατηρήσεις, μία 5-fold διασταυρωμένη επικύρωση θα χρησιμοποιεί σύνολα εκπαίδευσης μεγέθους 160, τα οποία θα συμπεριφέρονται σαν το πλήρες σύνολο. Ωστόσο, με ένα σύνολο δεδομένων των 50 παρατηρήσεων η 5-fold διασταυρωμένη επικύρωση θα χρησιμοποιεί σύνολα εκπαίδευσης μεγέθους 40, και αυτό θα είχε ως αποτέλεσμα μία σημαντική υπερεκτίμηση του σφάλματος πρόβλεψης.

5.4 Κριτήρια αποδόσης του μοντέλου−Αξιολόγηση ταξινομητών

Η ακρίβεια (accuracy), ενός συστήματος μέτρησης είναι ο βαθμός της εγγύτητας των μετρήσεων μιας ποσότητας με την πραγματική (πραγματική) τιμή της ποσότητας αυτής. Η ακρίβεια (precision), ενός συστήματος μέτρησης, που ονομάζεται επίσης αναπαραγωγιμότητα ή επαναληψιμότητα, είναι ο βαθμός στον οποίο επαναλαμβανόμενες μετρήσεις υπό αμετάβλητες συνθήκες δείχνουν τα ίδια αποτελέσματα. Ένα σύστημα μέτρησης μπορεί να είναι ακριβές (accurate), αλλά όχι precise, και το αντίθετο ή και τα δύο μαζί. Για παράδειγμα, αν ένα πείραμα περιέχει ένα συστηματικό σφάλμα, αυξάνοντας στη συνέχεια το μέγεθος του δείγματος γενικά αυξάνει η presicion, αλλά δεν βελτιώνεται η accuracy. Ένα σύστημα μέτρησης ορίζεται ως έγκυρο εάν είναι τόσο ακριβές όσο και σαφές (accurate και precise). Σχετικοί όροι περιλαμβάνουν τη μεροληψία (bias) (μη-τυχαίες ή κατευθυνόμενες επιδράσεις που προκαλούνται από έναν παράγοντα ή παράγοντες που δεν σχετίζονται με την ανεξάρτητη μεταβλητή) και το σφάλμα (τυχαία μεταβλητότητα).



Σχήμα 25: Η ακρίβεια αποτελείται από την Ορθότητα/trueness (εγγύτητα των αποτελεσμάτων της μέτρησης με την αληθή τιμή) και την precision (επαναληψιμότητα/αναπαραγωγιμότητα των μετρήσεων)

Σύμφωνα με το πρότυπο ISO 5725-1, οι όροι της ορθότητας (trueness) και της ακρίβειας (precision) χρησιμοποιούνται για να περιγράψουν την ακρίβεια (accuracy) της μέτρησης. Η ορθότητα (trueness) αφορά την εγγύτητα του μέσου όρου των αποτελεσμάτων των μετρήσεων με την «σωστή» τιμή και η ακρίβεια (precision) αναφέρεται στην εγγύτητα της συμφωνίας στο πλαίσιο των επιμέρους αποτελεσμάτων. Ως εκ τούτου, σύμφωνα με το πρότυπο ISO, ο όρος «accuracy» αναφέρεται τόσο στην ορθότητα (trueness) όσο και στην ακρίβεια (precision).

Η ακρίβεια (accuracy) χρησιμοποιείται επίσης ως ένα στατιστικό μέτρο του πόσο καλά ένα τεστ σε μία δυαδική ταξινόμηση προσδιορίζει σωστά ή αποκλείει μια κατάσταση.

Ακολουθεί ο πίνακας συνάφειας (contingency):

		Πραγματική τιμή	
		Αληθές (T)	Ψευδές (F)
Προβλεπόμενη Τιμή (αποτέλεσμα του τεστ)	Θετικό (P)	Αληθώς Θετικά (TP)	Ψευδώς Θετικά (FP)
	Αρνητικό (N)	Ψευδώς Αρνητικά (FN)	Αληθώς Αρνητικά (TN)

Πίνακας 2: Πίνακας Συνάφειας

Δεδομένου ενός ταξινομητή και ενός παραδείγματος, υπάρχουν τέσσερα πιθανά αποτελέσματα.

TP: Αν η περίπτωση είναι *θετική* και είναι ταξινομημένη ως *θετική*, υπολογίζεται ως μια αληθώς θετική.

FN: Αν η περίπτωση είναι *θετική* και έχει ταξινομηθεί ως *αρνητική*, αυτό υπολογίζεται ως ψευδώς αρνητική.

TN: Αν η περίπτωση είναι *αρνητική* και έχει ταξινομηθεί ως *αρνητική*, αυτή υπολογίζεται ως μια αληθώς αρνητική.

FP: Αν η περίπτωση είναι *αρνητική* και έχει ταξινομηθεί ως *θετική*, προσμετράται ως ψευδώς θετική.

Δεδομένου ενός ταξινομητή και μια σειράς από περιπτώσεις (στο σύνολο δοκιμής), μπορεί να κατασκευαστεί ένας 2×2 πίνακας συνάφειας (ονομάζεται επίσης πίνακας έκτακτης ανάγκης) όπου αντιπροσωπεύονται οι διατάξεις του συνόλου των περιπτώσεων. Αυτός ο πίνακας αποτελεί τη βάση για πολλές μετρήσεις.

Οι αριθμοί κατά μήκος των κυρίων διαγωνίων αντιπροσωπεύουν τις σωστές αποφάσεις, και οι αριθμοί εκτός της διαγωνίου αντιπροσωπεύουν τα λάθη − τη σύγχυση − μεταξύ των διαφόρων κατηγοριών.

Το *Αληθώς Θετικό ποσοστό* (True Positive rate/TPR) (ονομάζεται επίσης ποσοστό επιτυχίας και ανάκληση) ενός ταξινομητή υπολογίζεται ως εξής:

$$TP\ rate \ \cong \ \frac{\alpha\lambda\eta\theta\acute{\omega}\varsigma\ \theta\varepsilon\tau\iota\kappa\acute{\alpha}}{\sigma\acute{\upsilon}\nu o\lambda o\ \theta\varepsilon\tau\iota\kappa\acute{\omega}\nu}$$

Το *Ψευδώς Θετικό ποσοστό* (False Positive rate/TPR) ενός ταξινομητή είναι:

$$FP\ rate \ \cong \ \frac{\psi\varepsilon\upsilon\delta\acute{\omega}\varsigma\ \theta\varepsilon\tau\iota\kappa\acute{\alpha}}{\sigma\acute{\upsilon}\nu o\lambda o\ \alpha\rho\nu\eta\tau\iota\kappa\acute{\omega}\nu}$$

ACCURACY ΚΑΙ PRECISION

Ακρίβεια (***accuracy***) είναι η αναλογία των πραγματικών αποτελεσμάτων (και τα δύο, αληθώς θετικά (TP) και αληθώς αρνητικά (TN)) στον πληθυσμό. Είναι μια παράμετρος της δοκιμής/τεστ.

$$accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

Accuracy 100% σημαίνει ότι οι μετρούμενες τιμές είναι ακριβώς ίδιες με τις τιμές που δίνονται.

Από την άλλη πλευρά, η ακρίβεια ή *θετική προγνωστική αξία* ορίζεται ως το ποσοστό των αληθώς θετικών έναντι όλων των θετικών αποτελεσμάτων (τόσο αληθώς θετικά όσο και ψευδώς θετικά).

$$precision = \frac{TP}{TP + FP}$$

ΕΥΑΙΣΘΗΣΙΑ ΚΑΙ ΕΙΔΙΚΟΤΗΤΑ

Η ευαισθησία και η ειδικότητα είναι στατιστικά μέτρα της απόδοσης ενός τεστ δυαδικής ταξινόμησης, γνωστές στη στατιστική ως συναρτήσεις ταξινόμησης.

Αυτά τα δύο μέτρα συνδέονται στενά με τις έννοιες των σφαλμάτων τύπου I και τύπου II. Ένας τέλειος εκτιμητής θα πρέπει να περιγράφεται με 100% ευαίσθησία και 100% ειδικότητα.

Οι συχνότερα, λοιπόν, χρησιμοποιούμενες και αναφερόμενες συνιστώσες της διαγνωστικής ποιότητας μιας δοκιμασίας, που καθορίζουν τη διακριτική της ικανότητα είναι

- το *ποσοστό των αληθώς θετικών αποτελεσμάτων,* το ποσοστό δηλαδή των θετικών ενδείξεων στον πληθυσμό (true positive rate, **TPR**) ή η **ευαισθησία** (sensitivity) της δοκιμασίας, δηλαδή η πιθανότητα το τεστ να είναι θετικό δεδομένου ότι κάποιος έχει το χαρακτηριστικό που εξετάζουμε και δίνεται από τον τύπο:

  $$SE = TPR = \frac{\alpha\lambda\eta\theta\acute{\omega}\varsigma\ \theta\varepsilon\tau\iota\kappa\acute{\alpha}}{\sigma\acute{\upsilon}\nu o\lambda o\ \theta\varepsilon\tau\iota\kappa\acute{\omega}\nu} = \frac{TP}{P} = \frac{TP}{TP + FN}$$

  Η ευαισθησία σχετίζεται με την ικανότητα του τεστ να προσδιορίσει θετικά αποτελέσματα. Μια δοκιμή με υψηλή ευαισθησία έχει χαμηλό ποσοστό σφάλματος τύπου II.

και

- το *ποσοστό των αληθώς αρνητικών αποτελεσμάτων*, δηλαδή το ποσοστό των αρνητικών ενδείξεων στον πληθυσμό (true negative rate, **TNR**) ή η **ειδικότητα** (specificity) της δοκιμασίας, δηλαδή η πιθανότητα το τεστ να ειναι αρνητικό δεδομένου ότι κάποιος δεν έχει το χαρακτηριστικό που εξετάζουμε και υπολογίζεται ως εξής:

$$SPC = TNR = \frac{\alpha\lambda\eta\theta\acute{\omega}\varsigma\ \alpha\rho\nu\eta\tau\iota\kappa\acute{\omega}\nu}{\sigma\acute{\upsilon}\nu o\lambda o\ \alpha\rho\nu\eta\tau\iota\kappa\acute{\omega}\nu} = \frac{TN}{N} = \frac{TN}{FP + TN}$$

  Η ειδικότητα σχετίζεται με την ικανότητα του τεστ να εντοπίσει αρνητικά αποτελέσματα. Μια δοκιμή με υψηλή εξειδίκευση έχει χαμηλό ποσοστό σφάλματος τύπου Ι.

Τα ποσοστά αυτά, καθώς και τα συμπληρωματικά τους (ποσοστό ψευδώς αρνητικών (**FNR**) και ψευδώς θετικών αποτελεσμάτων (**FPR**), αντίστοιχα) ονομάζονται *πιθανοφάνειες* (likelihood) ή, αλλιώς, λειτουργικά χαρακτηριστικά (operating characteristics) της διαγνωστικής δοκιμασίας.

Προφανώς ισχύει:

$$TPR = 1 - FNR$$

όπου $FNR = \frac{FN}{P} = \frac{FN}{TP+FN}$.

ΘΕΤΙΚΗ ΚΑΙ ΑΡΝΗΤΙΚΗ ΠΡΟΓΝΩΣΤΙΚΗ ΑΞΙΑ

Άλλες δύο χρήσιμες έννοιες που αφορούν στους διαγνωστικούς ελέγχους θα πρέπει να εισάγουμε σε αυτό το σημείο. Η *θετική προγνωστική* (ή διαγνωστική ή προβλεπόμενη) *αξία* (*positive predictive value*) που συμβολίζεται με **PPV** :

$$PPV = \frac{TP}{TP + FP}$$

Παρόμοια ορίζεται και η αρνητική προγνωστική (ή διαγνωστική ή προβλεπόμενη) αξία (negative predictive value) που συμβολίζεται με **NPV** :

$$NPV = \frac{TN}{TN + FN}$$

Τα PPV και NPV είναι αντίστοιχα με τα σφάλματα τύπου I και II στους αντίστοιχους ελέγχους υποθέσεων.

ΕΠΙΠΟΛΑΣΜΟΣ

Ως *επιπολασμό* (*prevalence*) ορίζουμε το σύνολο των θετικών προς το σύνολο του πληθυσμού και υπολογίζεται ως εξής:

$$PRV = \frac{TP + FN}{P + N}$$

Θα μπορούσαμε να πούμε ότι οι διαγνωστικές έννοιες PPV και NPV λειτουργούν συμπληρωματικά με τον επιπολασμό που εκφράζει την πιθανότητα προ δοκιμασίας(Pretest Probability).

Η ακρίβεια μπορεί να προσδιοριστεί από την ευαισθησία και ειδικότητα, εφόσον είναι γνωστός ο επιπολασμός, χρησιμοποιώντας την εξίσωση:

$$accuracy = (sensitivity)(prevalence) + (specificity)(1 - prevalence)$$

Στον ακόλουθο πίνακα δίνουμε μία πιο συγκεντρωτική εικόνα για τα παραπάνω μέτρα:

		Πραγματική τιμή: Αληθές (T)	Πραγματική τιμή: Ψευδές (F)	
Αποτέλεσμα του τεστ	Θετικό (P)	**TP**	**FP**	→ Θετική προγνωστική αξία ($PPV/precision$)
	Αρνητικό (N)	**FN**	**TN**	→ Αρνητική προγνωστική αξία (NPV)
		↓ Ευαισθησία Sensitivity	↓ Ειδικότητα Specificity	Ακρίβεια Accuracy

Πίνακας 3: Συγκεντρωτικός πίνακας-πίνακας συνάφειας και μέτρα

5.5 ROC καμπύλες

5.5.1 Εισαγωγή

Η πραγματοποίηση προβλέψεων αποτελεί μέλημα ζωτικής σημασίας κάθε επιχείρησης και επιστημονικού πεδίου όσον αφορά την αναζήτηση πληροφορίας και την περαιτέρω έρευνα. Είναι λοιπόν αναγκαία η πραγματοποίηση προβλέψεων και η εξασφάλιση προγνωστικής ακρίβειας στον σχεδιασμό και την σύγκριση μοντέλων, αλγορίθμων και τεχνολογιών που παράγουν προβλέψεις. Οι ROC (Receiver Operating Characteristic: Λειτουργικό Χαρακτηριστικό Δέκτη) καμπύλες συμβάλλουν στην εξασφάλιση της επιθυμητής ακρίβειας στις προβλέψεις. Αποτελούν χρήσιμη τεχνική για την απεικόνιση, την οργάνωση και την επιλογή ταξινομητών με βάση την απόδοση τους. Η ROC είναι επίσης γνωστή ως μια καμπύλη σχετικής χαρακτηριστικής λειτουργίας (relative operating characteristic), γιατί είναι μια σύγκριση των δύο χαρακτηριστικών λειτουργίας (TPR και FPR) καθώς αλλάζει το κριτήριο.

Η καμπύλη ROC ορίζεται ως το μοναδιαίο τετράγωνο $[0,1] \times [0,1]$ και ξεκινά από το σημείο (0,0) (όταν το σημείο απόφασης είναι μεγαλύτερο από όλες τις μετρήσεις θορύβου και σήματος) για να καταλήξει στο (1,1)) (για την περίπτωση που το σημείο απόφασης είναι μικρότερο από όλες τις μετρήσεις). Το εμβαδόν που ορίζεται κάτω από την καμπύλη αποτελεί ένα μέτρο της ποιότητας διαχωρισμού θορύβου – σήματος και χρησιμοποιείται συχνά στη στατιστική συμπερασματολογία των καμπυλών ROC.

Τα τελευταία χρόνια έχουν συντελέσει σημαντικά στην ανάπτυξη της Στατιστικής Θεωρίας με εφαρμογές κυρίως στη μη-παραμετρική στατιστική, τα λεγόμενα U-statistics, τους ελέγχους καλής προσαρμογής, τα γενικευμένα γραμμικά μοντέλα, τους τυχαίους περίπατους και την ανάλυση επιβίωσης. Η χρήση τους αφορά κυρίως την εκτίμηση της ποιότητας διαγνωστικών εργαλείων, τη σύγκριση μεταξύ τους, τη βέλτιστη επιλογή μοντέλων, τον ποιοτικό έλεγχο, τη σύγκριση αλγορίθμων μηχανικής μάθησης, την επιλογή βέλτιστων σημείων απόφασης και τη θεωρία αποφάσεων.

Τα ROC γραφήματα είναι εννοιολογικά απλά, αλλά υπάρχουν κάποιες μη προφανείς περιπλοκές που προκύπτουν όταν χρησιμοποιούνται στην έρευνα. Υπάρχουν επίσης κοινές παρερμηνείες και παγίδες κατά τη χρήση τους στην πράξη.

5.5.2 ROC Γραφήματα και Ερμηνεία

Η σχέση του ποσοστού των αληθώς θετικών (***TPR***) και ψευδώς θετικών (***FPR***) αποτελεσμάτων της διαγνωστικής δοκιμασίας, καθώς μεταβάλλεται προοδευτικά προς μια κατεύθυνση το διαχωριστικό όριο αυτής, παριστάνεται γραφικά με την **καμπύλη ROC** (*Receiver Operating Characteristic Curve*) ή καμπύλη λειτουργικών χαρακτηριστικών.

Ας θεωρήσουμε ένα πρόβλημα πρόβλεψης δύο κατηγοριών (δυαδική ταξινόμηση), όπου τα αποτελέσματα επισημαίνονται είτε ως θετικά (P) ή αρνητικά (N). Υπάρχουν τέσσερις πιθανές εκβάσεις από ένα δυαδικό ταξινομητή, όπως ακριβώς παρουσιάζονται προηγουμένως στον πίνακα συνάφειας.

Για τον σχεδιασμό μιας καμπύλης ROC αρκούν μόνο τα TPR (πόσα σωστά θετικά αποτελέσματα συμβαίνουν ανάμεσα σε όλα τα θετικά κατά τη διάρκεια του τεστ TP/TP+FN) και FPR (πόσα λάθος θετικά αποτελέσματα συμβαίνουν ανάμεσα σε όλα τα αρνητικά κατα τη διάρκεια του τεστ FP/FP+TN)

Τα ROC γραφήματα είναι δισδιάστατα διαγράμματα στα οποία το TP ποσοστό σχεδιάζεται στον $y-$ άξονα και το FP ποσοστό σχεδιάζεται στον $x-$ άξονα και απεικονίζουν σχετικά trade-offs ανάμεσα στο αληθώς θετικά (TP / οφέλη) και ψευδώς θετικά (FP / κόστοι). Επομένως θα μπορούσαμε να πούμε ότι ένα γράφημα ROC απεικονίζει τη σχετική μεταβολή μεταξύ του κέρδους (αληθώς θετικά) και του κόστους (ψευδώς θετικά). Η καμπύλη αυτή εγγράφεται μέσα σε ένα τετράγωνο, στις τέσσερις γωνίες του οποίου αντιστοιχούν οι ακραίες τιμές (0 και 1) του %ΑΘ και του %ΨΘ αποτελεσμάτων, καθώς και των συμπληρωματικών αυτών ποσοστών (%ΨΑ και %ΑΑ).

Λόγω του ότι το $TPR = sensitivity$ και το $FPR = 1 - specificity$ το ROC γράφημα καλείται κάποιες φορές και $(sensitivity)vs(1 - specificity)$ διάγραμμα. Κάθε πρόβλεψη ή περίπτωση του πίνακα συνάφειας αντιπροσωπεύει ένα σημείο στο χώρο ROC.

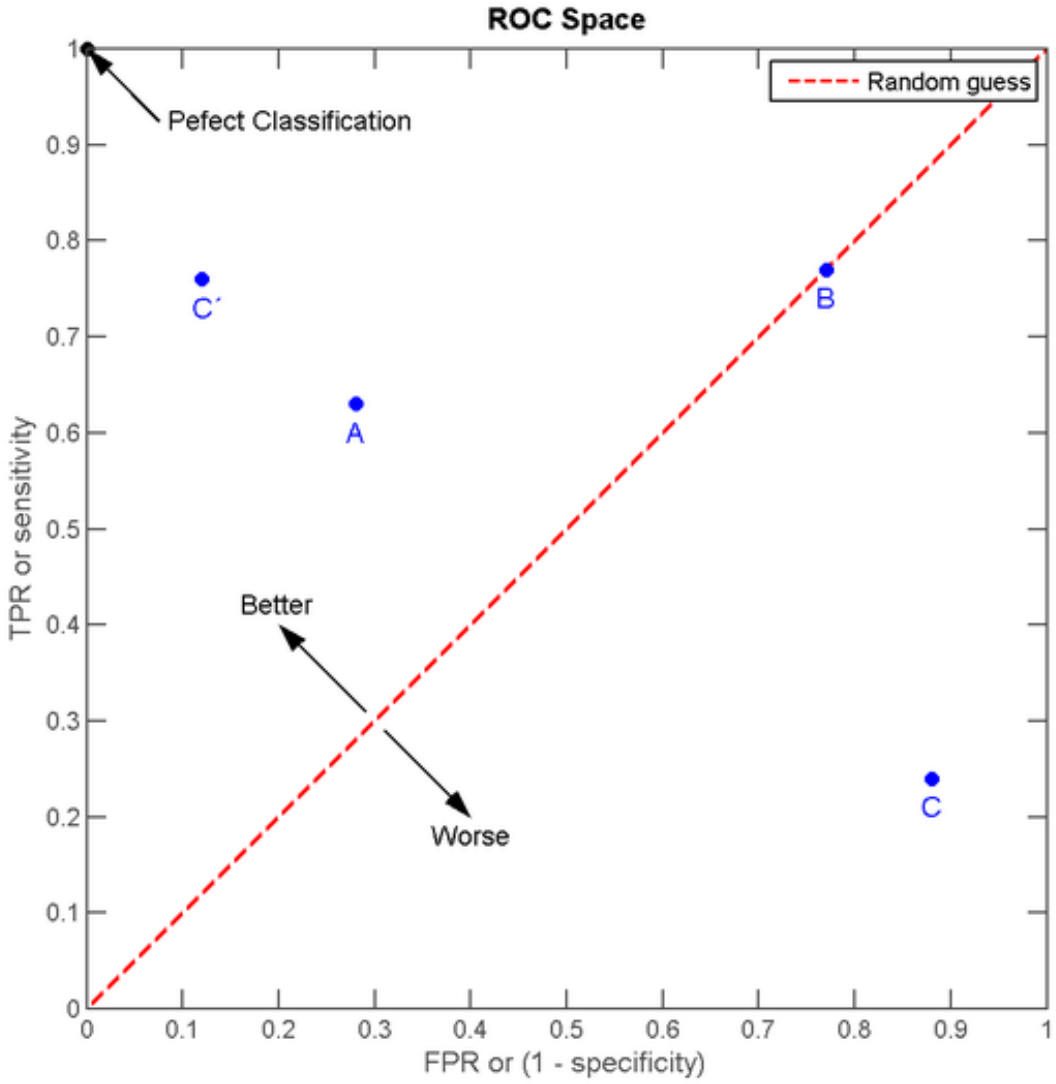
Η καλύτερη δυνατή μέθοδος πρόβλεψης θα αποδώσει ένα σημείο στην επάνω αριστερή γωνία ή στη συντεταγμένη (0,1) του χώρου ROC, που αντιπροσωπεύει την περίπτωση που έχουμε 100% ευαισθησία (όχι ψευδώς αρνητικά) και την 100% ειδικότητα (δεν υπάρχουν ψευδώς θετικά). Το σημείο (0,1) ονομάζεται επίσης μια τέλεια ταξινόμηση (perfect classification). Μια εντελώς τυχαία εικασία θα δώσει ένα σημείο κατά μήκος μιας διαγώνιας γραμμής (η λεγόμενη γραμμή της μη-διάκρισης) από κάτω αριστερά προς τα πάνω δεξιά γωνία (ανεξάρτητα από τις θετικές και αρνητικές τιμές βάσης). Ένα διαισθητικό παράδειγμα μιας τυχαίας εικασίας είναι μια απόφαση με το νόμισμα (κορώνα ή γράμματα). Δεδομένου ότι το μέγεθος του δείγματος αυξάνει, ένα τυχαίο σημείο του ROC ταξινομητή «μεταναστεύει» προς το σημείο (0.5,0.5).

Η διαγώνιος (y = x) διαιρεί το χώρο ROC. Έτσι, μια τυχαία κατάταξη θα παράγει ένα σημείο ROC το οποίο μετακινείται εμπρός και πίσω στη διαγώνιο με βάση τη συχνότητα με την οποία εικάζει τη θετική τάξη. Για να ξεφύγουμε από τη διαγώνιο στην άνω τριγωνική περιοχή, η ταξινόμηση πρέπει να εκμεταλλευτεί κάποιες πληροφορίες όσον αφορά τα δεδομένα. Τα σημεία πάνω από τη διαγώνιο αντιπροσωπεύουν καλά αποτελέσματα ταξινόμησης (καλύτερα από τυχαία), σημεία κάτω από το όριο όχι καλά αποτελέσματα (χειρότερα από τυχαία).

Τα αποτελέσματα από τα τέσσερα παραπάνω αποτελέσματα στο χώρο ROC δίδεται στο σχήμα 27. Το αποτέλεσμα της μεθόδου A δείχνει σαφώς την καλύτερη προβλεπτική δύναμη μεταξύ των A, B, και C. Το αποτέλεσμα της B βρίσκεται στη γραμμή της τυχαίας εικασίας (διαγώνια γραμμή), (η ακρίβεια (accuracy) του B είναι 50%). Ωστόσο, όταν η C αντικατοπτρίζεται απέναντι του κεντρικού σημείου (0.5,0.5), η προκύπτουσα μέθοδος C΄ είναι ακόμα καλύτερη από την A. Αυτή η μέθοδος που αντικατοπτρίζεται, απλώς αντιστρέφει τις προβλέψεις της οποιασδήποτε μεθόδου ή των δοκιμών που παράγονται από τον πίνακα συνάφειας που παρήγαγε τη C.

Παρά το γεγονός ότι η αρχική μέθοδος **C** έχει αρνητική προγνωστική δύναμη, απλά αντιστρέφοντας τις αποφάσεις της οδηγούμαστε σε μια νέα μέθοδο πρόβλεψης **C**΄, η οποία έχει θετική προγνωστική δύναμη. Όταν η μέθοδος **C** προβλέπει για παράδειγμα p ή n, η μέθοδος **C**΄ θα προβλέψει n ή p, αντιστοίχως. Με τον τρόπο αυτό, το κριτήριο της **C**΄ θα εκτελέσει το καλύτερο. Όσο πιο κοντά είναι ένα αποτέλεσμα του πίνακα συνάφειας στην επάνω αριστερή γωνία, τόσο καλύτερη είναι η πρόβλεψη, αλλά η απόσταση από την γραμμή της τυχαίας εικασίας προς οποιαδήποτε κατεύθυνση είναι ο καλύτερος δείκτης για την προβλεπτική ικανότητα μια μεθόδου. Εάν το αποτέλεσμα είναι κάτω από τη γραμμή (δηλαδή η μέθοδος είναι χειρότερη από ότι μια τυχαία εικασία), το σύνολο των προβλέψεων της μεθόδου πρέπει να αναστραφεί, προκειμένου να χρησιμοποιήσει την ισχύ της, μετακινώντας έτσι το αποτέλεσμα πάνω από τη γραμμή της τυχαίας εικασίας.

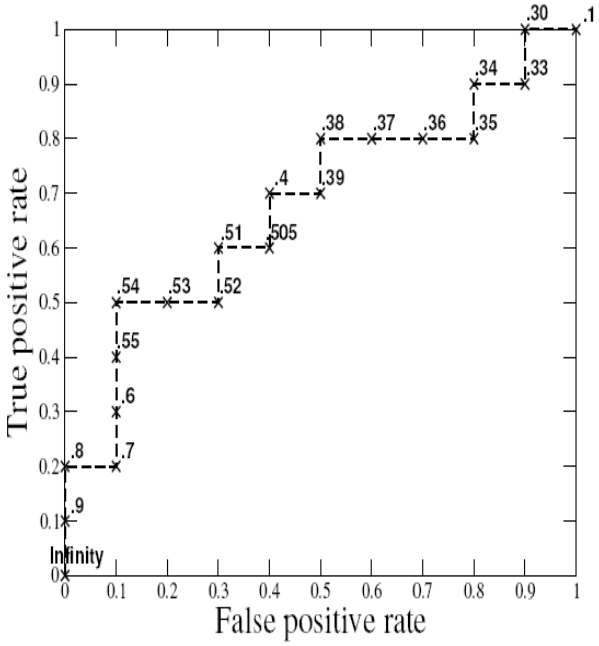
Κάθε ταξινομητής, λοιπόν, που εμφανίζεται στο κάτω δεξιό τρίγωνο εκτελεί δυσμενέστερα από ότι η τυχαία εικασία. Αυτό το τρίγωνο είναι ως εκ τούτου συνήθως άδειο στα ROC γραφήματα. Ωστόσο, σημειώνουμε ότι ο χώρος απόφασης είναι συμμετρικός σχετικά με τη διαγώνιο που χωρίζει τα δύο τρίγωνα. Αν αντιστρέψετε μια ταξινόμηση δηλαδή, η αντίστροφη της ταξινόμησης των αποφάσεων σε κάθε περίπτωση, οι αληθώς θετικά ταξινομήσεις γίνονται ψευδώς θετικά λάθη και τα ψευδώς θετικά γίνονται αληθώς θετικά.



Σχήμα 26: Ο χώρος ROC και το γράφημα 4 παραδειγμάτων πρόβλεψης

Ως εκ τούτου, οποιαδήποτε ταξινόμηση που παράγει ένα σημείο στο κάτω δεξιά τρίγωνο μπορεί να εξαλειφθεί για να παραχθεί ένα σημείο στο επάνω αριστερό τρίγωνο.

Ακολουθεί άλλο ένα κατατοπιστικό σχήμα:



Inst#	Class	Score	Inst#	Class	Score
1	p	.9	11	p	.4
2	p	.8	12	n	.39
3	n	.7	13	p	.38
4	p	.6	14	n	.37
5	p	.55	15	n	.36
6	p	.54	16	n	.35
7	n	.53	17	p	.34
8	n	.52	18	n	.33
9	p	.51	19	p	.30
10	n	.505	20	n	.1

Σχήμα 27: Η ROC καμπύλη δημιουργήθηκε από ένα σύνολο ορίων ελέγχου. Ο πίνακας στα δεξιά δείχνει είκοσι δεδομένα και το σκορ ανατεθεί σε κάθε ένα τη βαθμολόγηση. Το γράφημα στα αριστερά δείχνει την αντίστοιχη καμπύλη ROC με κάθε σημείο χαρακτηρισμένο από το όριο που το παράγει**.**

5.5.3 Η περιοχή κάτω από την ROC καμπύλη (AUC)

Η καμπύλη ROC είναι μια δισδιάστατη απεικόνιση της απόδοσης της ταξινόμησης. Για να συγκρίνουμε τους ταξινομητές μπορεί να χρειαστεί να μειώσουμε την απόδοση ROC σε μία ενιαία βαθμωτή τιμή που αντιπροσωπεύει την αναμενόμενη απόδοση. Μια κοινή μέθοδος είναι να υπολογίσουμε το εμβαδόν κάτω από την καμπύλη ROC, η συντομογραφία της είναι AUC (Bradley, 1997, Hanley & McNeil, 1982). Εφόσον η AUC είναι μέρος της περιοχής της μονάδας, η τιμή της θα είναι πάντα μεταξύ 0 και 1. Ωστόσο, επειδή μια τυχαία εικασία παράγει τη διαγώνια γραμμή μεταξύ $(0,0)$ και $(1,1)$, η οποία έχει έκταση 0,5, κανένας ρεαλιστικός ταξινομητής δεν θα πρέπει να έχει AUC λιγότερο από 0,5.

Η AUC ενός ταξινομητή , όταν χρησιμοποιούμε κανονικοποιημένες μονάδες είναι ισοδύναμη με την πιθανότητα ο ταξινομητής να ταξινομήσει ένα τυχαία επιλεγμένο θετικό παράδειγμα υψηλότερα από ένα τυχαία επιλεγμένο αρνητικό παράδειγμα. Αυτό είναι ισοδύναμο με το Mann-Whitney U (Hanley και McNeil (1982), Mason και Graham (2002)) οι δοκιμές των οποίων οποιαδήποτε θετικά ταξινομούνται υψηλότερα από από τα αρνητικά. Αυτό είναι επίσης ισοδύναμο με τη δοκιμή Wilcoxon των βαθμίδων (Mason και Graham (2002)). Η AUC είναι επίσης στενά συνδεδεμένη με το δείκτη Gini (Breiman, Friedman, Olshen, & Stone, 1984), ο οποίος είναι διπλάσιος από το χώρο ανάμεσα στην διαγώνιο και

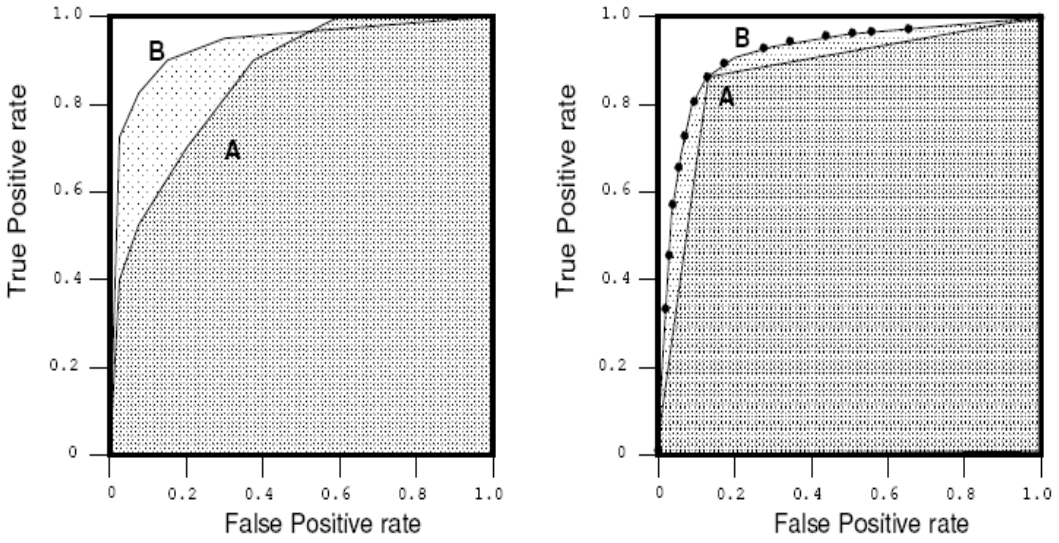
στην καμπύλη ROC. Ο Hand και Till (2001) επισημαίνουν ότι

$$Gini = 2 \times AUC - 1$$

Το παρακάτω σχήμα δείχνει τις περιοχές κάτω από τις δύο ROC καμπύλες, Α και Β. Ο ταξινομητής Β έχει μεγαλύτερη έκταση και συνεπώς καλύτερη μέση απόδοση. Επίσης δείχνει την περιοχή κάτω από την καμπύλη ενός δυαδικού ταξινομητή Α και ενός αθροιστικού ταξινομητή Β. Ο ταξινομητής Α αντιπροσωπεύει την απόδοση του Β όταν ο Β χρησιμοποιείται με ένα συγκεκριμένο όριο. Αν και η απόδοση και των δύο είναι ίση σε συγκεκριμένο σημείο (Β όριο), οι επιδόσεις του Β είναι κατώτερες του Α σε περαιτέρω από αυτό το σημείο.

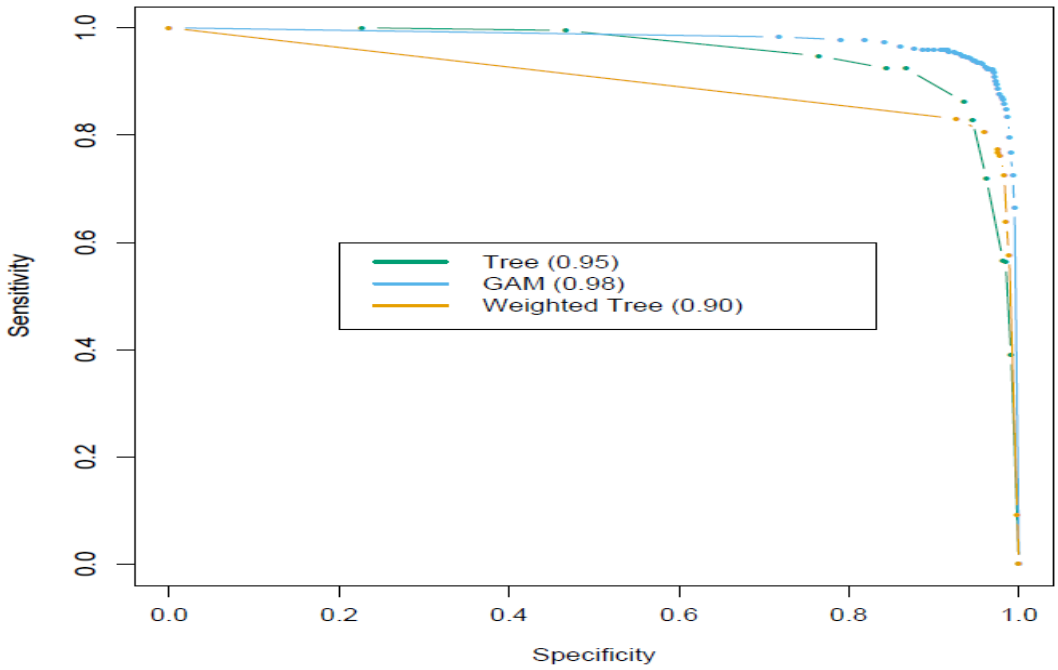
Είναι δυνατόν για ένα υψηλής−AUC ταξινομητή να έχει χειρότερες επιδόσεις σε μια περιοχή της ROC από ένα χαμηλό−AUC ταξινομητής.

Αντί της συλλογής ROC σημείων, ο αλγόριθμος προσθέτει διαδοχικά περιοχές τραπεζο-ειδών στην περιοχή.



Σχήμα 28: Δύο γραφήματα ROC**.** Το γράφημα στα αριστερά δείχνει την περιοχή κάτω από δύο καμπύλες ROC. Το γράφημα στα δεξιά δείχνει την περιοχή κάτω από τις καμπύλες του διακριτού ταξινομητή Α και του πιθανού ταξινομητή Β.

Τέλος, παραθέτουμε (για την σύγκριση με προηγούμενες μεθόδους) την καμπύλη ROC για το παράδειγμα (spam example) που συζητήσαμε προηγουμένως των Hastie et al. (2001).



Σχήμα 29: οι ROC καμπύλες για τους κανόνες ταξινόμησης στα δεδομένα του παραδείγματος spam. Καμπύλες που είναι πιο κοντά στη βορειοανατολική γωνία αντιπροσωπεύουν καλύτερους ταξινομητές. Στην περίπτωση αυτή ο ταξινομητής GAM κυριαρχεί των δέντρων. Το σταθμισμένο δέντρο επιτυγχάνει καλύτερη ευαισθησία (sensitivity) για μεγαλύτερη ειδικότητα (specificity) από ό, τι το μη σταθμισμένο δέντρο. Οι αριθμοί στην λεζάντα αντιπροσωπεύουν την περιοχή κάτω από την καμπύλη.

Κεφάλαιο 6:

Εφαρμογή σε πραγματικά δεδομένα

Στην έκτη και τελευταία ενότητα, συγκρίνουμε μερικούς ταξινομητές από το πεδίο της μηχανικής μάθησης και στη συνέχεια παρουσιάζουμε τα αποτελέσματα της διαδικασίας μοντελοποίησης. Για να παρέχουμε μία αμερόληπτη εκτίμηση για την ποιότητα ταξινόμησης του κάθε μοντέλου χρησιμοποιώντας τη μέθοδο της διάκρισης (discrimination), οι τιμές των κριτηρίων απόδοσης υπολογίζονται από ένα σύνολο δεδομένων που δεν χρησιμοποιήθηκε στη διαδικασία μοντελοποίησης. Για το σκοπό αυτό χρησιμοποιήσαμε από το πραγματικό σύνολο δεδομένων, ένα μέρος (το σύνολο δοκιμής) το οποίο αφήσαμε στην άκρη για αυτό το σκοπό.

Ένας ταξινομητής θα πρέπει να παρέχει υψηλές τιμές των ACC, sensitivity, specificity και της AUROC, και η γενικευμένη απόδοση συχνά εκτιμάται με holdout επικύρωση (εκπαίδευση/δοκιμή).

6.1 Περιγραφή των δεδομένων – Εισαγωγή στο Clementine

Clementine

Το πρόγραμμα Clementine του πακέτου SPSS προσφέρει μια ποικιλία μεθόδων μοντελοποίησης που λαμβάνονται από τη μάθηση μηχανής, την τεχνητή νοημοσύνη αλλά και τη στατιστική. Οι μέθοδοι που διατίθενται στην παλέτα του Modeling μας επιτρέπουν να αντλήσουμε νέες πληροφορίες από τα δεδομένα μας και να αναπτύξουμε μοντέλα πρόβλεψης. Κάθε μέθοδος έχει ορισμένες δυναμικές και είναι κατάλληλη για συγκεκριμένα είδη προβλημάτων.

Ως μια εφαρμογή εξόρυξης δεδομένων, το Clementine προσφέρει μια στρατηγική προσέγγιση για την εξεύρεση χρήσιμων σχέσεων σε μεγάλα σύνολα δεδομένων.

Περιγραφή δεδομένων

Πρόκειται για ένα σύνολο δεδομένων υψηλής διάστασης που αποτελείται από 10333 σεισμούς, που χωρίζονται με τυχαίο τρόπο σε σύνολο εκπαίδευσης που αποτελείται από το 75% των περιπτώσεων (7749) και σύνολο δοκιμής που αποτελείται από το 25% των περιπτώσεων (2584) για να εκτιμήσουμε την απόδοση των ταξινομητών σε νέα δεδομένα.

Πίνακας Δεδομένων

*Μεταβλητή Απόκρισης-Δίτιμη-**output***	Y	magnitude (**0**(<6.5), **1**(>=6.5))
*Συνεχείς Μεταβλητές-**inputs***	x_1	**χρόνο, χρόνια**
	x_4	**γεωγραφικό πλάτος**
	x_7	**υπερ απόσταση,** μετρημένη σε βαθμούς (°)
	x_8	**αζιμούθιο[8], μετρημένο σε βαθμούς** (°)
	x_9	**επίκεντροx,** η τεταγμένη του επικέντρου
	x_{10}	**επίκεντροy,** η τετμημένη του επικέντρου
	x_{11}	**βάθος,** το βάθος του σεισμού κυμαίνεται από 0-700 km
Κατηγορικές μεταβλητές-**inputs**	x_2	**nome** (1 έως 54, όλοι οι νόμοι στην Ελλάδα)
	x_5	**ένταση**, (1 έως 12 βαθμούς)

Πίνακας 4: Περιγραφή του συνόλου δεδομένων

[8] είναι μια από τις οριζόντιες συντεταγμένες καθώς αποτελεί και μια γωνία του τριγώνου θέσεως. Συνήθως μετριέται από τον Βορρά.

Το υπό εξέταση σύνολο δεδομένων αποτελείται από τη μεταβλητή απόκρισης y η οποία αναφέρεται στο μέγεθος του σεισμού και κωδικοποιείται με 0-1 ($μέγεθος > 6.5: 1$, $αλλιώς: 0$), όπου η σκάλα μεγέθους χρησιμοποιείται για να εκφράσει την ενέργεια που απελευθερώνεται από ένα σεισμό, προέκυψαν 9 σημαντικοί παράγοντες από την απόδοση των τεχνικών επιλογής μεταβλητών από τους Koukouvinos et al.(2012) και 10333 παραδείγματα.

Πριν πραγματοποιηθεί η ανάλυση των δεδομένων με τη χρηση των ταξινομητών που αναλύσαμε προηγουμένως, θα πρέπει να εφαρμόσουμε κάποια βασικά βήματα:

I. Αρχικά εισάγουμε τα δεδομένα στο πρόγραμμα μέσω ενός xls (excel) αρχείου.

II. Τοποθετούμε ένα **Type node** στο stream canvas του προγράμματος Clementine όπου καθορίζονται:

a. ο ***τύπος*** (type - *εύρος (range)*/ *Διακριτή (discrete)*/ *Δίτιμη(flag)*/ *Σύνολο (set)*/ *typeless*) των δεδομένων για κάθε πεδίο και

b. η ***κατεύθυνση*** (direction – *out/in/both/none*) που επιδεικνύει το ρόλο που παίζει κάθε πεδίο στη μοντελοποίηση.

III. Τοποθετούμε έναν **Partition node** στο stream canvas με σκοπό να χωριστούν τα δεδομένα σε:

a. ***δεδομένα εκπαίδευσης*** (training set) και με βάση αυτά, γνωρίζοντας την τιμή του αποτελέσματος προσπαθούμε να κατασκευάσουμε ένα μοντέλο πρόβλεψης.

b. ***δεδομένα ελέγχου-εξέτασης*** (test dataset). Το μοντέλο που δημιουργήσαμε θα το χρησιμοποιήσουμε στη συνέχεια για να προβλέψουμε το αποτέλεσμα νέων συνόλων δεδομένων εξέτασης (test set), στα οποία σύνολα είναι γνωστές οι τιμές των χαρακτηριστικών αλλά δεν είναι γνωστή η τιμή του αποτελέσματος, δηλαδή η τιμή της τάξης.

Στην περίπτωση που ο αλγόριθμος που εφαρμόζουμε στηρίζεται σε κατασκευή και εκτίμηση μοντέλου, τα δεδομένα διαχωρίζονται σε τρία υποσύνολα:

a. ***δεδομένα εκπαίδευσης*** (training set), τα οποία χρησιμοποιούνται για την προσαρμογή του μοντέλου

b. ***δεδομένα επαλήθευσης-επικύρωσης*** (quiz set-validation set) που χρησιμοποιούνται για την εκτίμηση του σφάλματος πρόβλεψης για την επιλογή του μοντέλου
c. ***δεδομένα ελέγχου-εξέτασης*** (test dataset), που χρησιμοποιούνται για τον υπολογισμό της γενικευμένης τιμής σφάλματος του τελικά επιλεγμένου μοντέλου.

Καθένα από αυτά τα σύνολα θα πρέπει να επιλεγεί ανεξάρτητα.

Στο πρόβλημά μας, χωρήσαμε, όπως επισημάναμε προηγουμένως σε *σύνολο εκπαίδευσης* που αποτελείται από το 75% των περιπτώσεων (7749) και *σύνολο δοκιμής* που αποτελείται από το 25% των περιπτώσεων (2584).

IV. Τοποθετούμε έναν **Feature selection node** στο stream canvas τον οποίο συνδέουμε στον Partition node με σκοπό να επιλεγούν , οι σημαντικές μεταβλητές. Στην εφαρμογή δεν πραγματοποιήσαμε αυτό το βήμα διότι κρατήσαμε όλες τις μεταβλητές στην μοντελοποίηση που θα ακολουθήσει.

Στόχος – Σύνοψη της διαδικασίας που θα ακολουθήσουμε

- Ο κόμβος της μοντελοποίησης (*Modeling node* π.χ. ) εκτιμάει το μοντέλο μελετώντας τις εγγραφές (records) για τις οποίες η έκβαση (outcome) είναι γνωστή, και παράγει ένα model nugget (ονομάζονται «nuggets» λόγω της αναπαράστασής τους με ένα αντικείμενο σε σχήμα διαμαντιού π.χ. ).
- Το model nugget μπορεί να προστεθεί σε οποιοδήποτε stream με τα αναμενώμενα πεδία για να σκοράρει εγγραφές. Σκοράρωντας τις εγγραφές για τις οποίες ήδη γνωρίζαμε την έκβαση (outcome), μπορούμε να αξιολογήσουμε (evaluate) πόσο καλά αυτό εκτελείται (perform).
- Μόλις είμαστε ικανοποιημένοι ότι το μοντέλο συμπερφέρεται καλά μπορούμε να κάνουμε υπολογισμούς για νέα δεδομένα για να προβλέψουμε πως θα ανταποκριθεί.

6.2 Μέτρα αξιολόγησης

Στην προηγούμενη ενότητα αναφερθήκαμε εκτενέστερα στα μέτρα αξιολόγησης. Ωστόσο ο τρόπος που παρουσιάζονται στο πρόγραμμα του Clementine είναι λίγο διαφορετικός. Συγκεκριμένα έχουμε:

		Αποτέλεσμα του τεστ		Σύνολο
		αρνητικό (0)	θετικό (1)	
Πραγματική τιμή	Αρνητικό (0)	d / TN (πραγματικά αρνητικά)	b / FP (λανθασμένα θετικά)	b + d
	Θετικό (1)	c / FN (λανθασμένα αρνητικά)	a / TP (πραγματικά θετικά)	a + c
Σύνολο		c+d	a + b	a+b+c+d

Πίνακας 5: Πίνακας συνάφειας-Clementine

Όπου:

- ευαισθησία $= \frac{a}{a+c} = \frac{TP}{TP+FN}$
- ειδικότητα $= \frac{d}{b+d} = \frac{TN}{FP+TN}$
- θετική προγνωστική αξία $(PPV) = \frac{a}{a+b} = \frac{TP}{TP+FP}$
- αρνητική προγνωστική αξία $(NPV) = \frac{d}{d+c} = \frac{TN}{TN+FN}$
- ακρίβεια $(accuracy) = \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+FP+FN+TN}$

6.3 Λογιστική παλινδρόμηση

6.3.1 Μέθοδος Forwards

Η λογιστική παλινδρόμηση, όπως έχουμε ήδη επισημάνει, είναι μια στατιστική τεχνική για την ταξινόμηση των αρχείων με βάση τις τιμές των πεδίων εισαγωγής. Είναι ανάλογο με τη γραμμική παλινδρόμηση, αλλά παίρνει μια κατηγορηματική μεταβλητή ως τομέα-στόχο αντί ενός αριθμητικού εύρους.

Για να εκτιμήσουμε το μοντέλο με τη μέθοδο της λογιστικής παλινδρόμησης ακολουθούμε την διαδικασία που περιγράφουμε στη συνέχεια.

Επιλέγουμε τη διαδρομή:

Modeling ⟶ *Classification* ⟶ *Logistic*

Δημιουργώντας με αυτό τον τρόπο έναν **Modeling Node**  μπορούμε να επιλέξουμε τις παραμέτρους του μοντέλου μας και να εκτελούμε το μοντέλο.

Τον **Logistic Node** τον συνδέουμε στον Partition node έτσι ώστε να ληφθεί υπόψην στη μοντελοποίηση ο διαχωρισμός σε σύνολο εκπαίδευσης και σύνολο εξέτασης.

Modeling Node:

Model name: (costum) Forwards.

Use partitioned data: το σημειώνουμε διότι αυτή η επιλογή μας εξασφαλίζει ότι χρησιμοποιούνται τα δεδομένα μόνο από το το training set για την κατασκευή του μοντέλου.

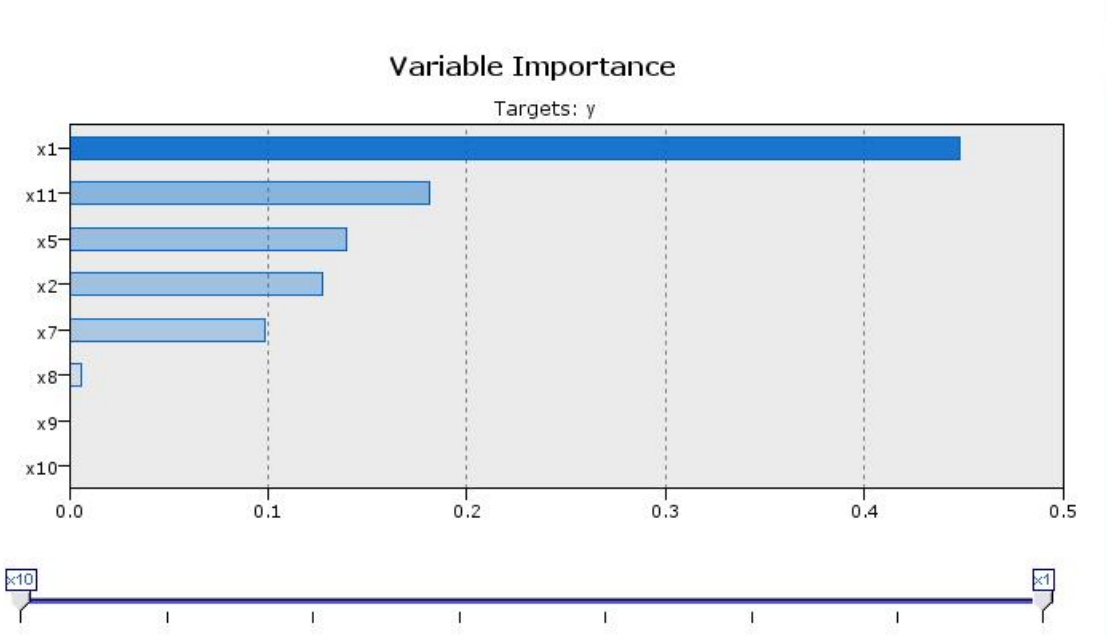
Procedure: Binomial

Method: **Forwards**

Εκτελούμε το μοντέλο και παράγουμε ένα **model nugget** που είναι απαραίτητο στη συνέχεια για την αξιολόγηση του μοντέλου.

Τα αποτελέσματα που προέκυψαν από την προσαρμογή του μοντέλου είναι τα ακόλουθα:

Η σειρά των μεταβλητών ανάλογα με την σημαντικότητα τους είναι η ακόλουθη



Σχήμα 30: κατάταξη των μεταβλητών ανάλογα με τη σημαντικότητά τους (LR)

Τέλος συνδέουμε έναν **Analysis node** για να δούμε τα μέτρα αξιολόγησης του μοντέλου καθώς και ένα **evaluation chart** για να σχεδιάσουμε την roc καμπύλη.

Στη συνέχεια ακολουθεί η αξιολόγηση του μοντέλου μέσω του

Coincidence matrices

Results for output field y

Comparing $L-y with y

'Partition'	1 Training		2 Testing	
Correct	6,655	85.85%	2,211	85.66%
Wrong	1,097	14.15%	370	14.34%
Total	7,752		2,581	

Coincidence Matrix for $L-y (rows show actuals)

'Partition' = 1 Training	0.000000	1.000000
0.000000	1,622	588
1.000000	509	5,033
'Partition' = 2 Testing	0.000000	1.000000
0.000000	585	230
1.000000	140	1,626

Performance Evaluation

'Partition' = 1 Training	
0.000000	0.982
1.000000	0.225
'Partition' = 2 Testing	
0.000000	0.938
1.000000	0.247

Αποτέλεσμα του τεστ

Πραγματική τιμή	Σύνολο εκπαίδευσης		Σύνολο Ελέγχου		
	0	**1**		**0**	**1**
0	TN = 1622	FP = 588	**0**	TN = 585	FP = 230
1	FN = 509	TP=5033	**1**	FN = 140	TP = 1626

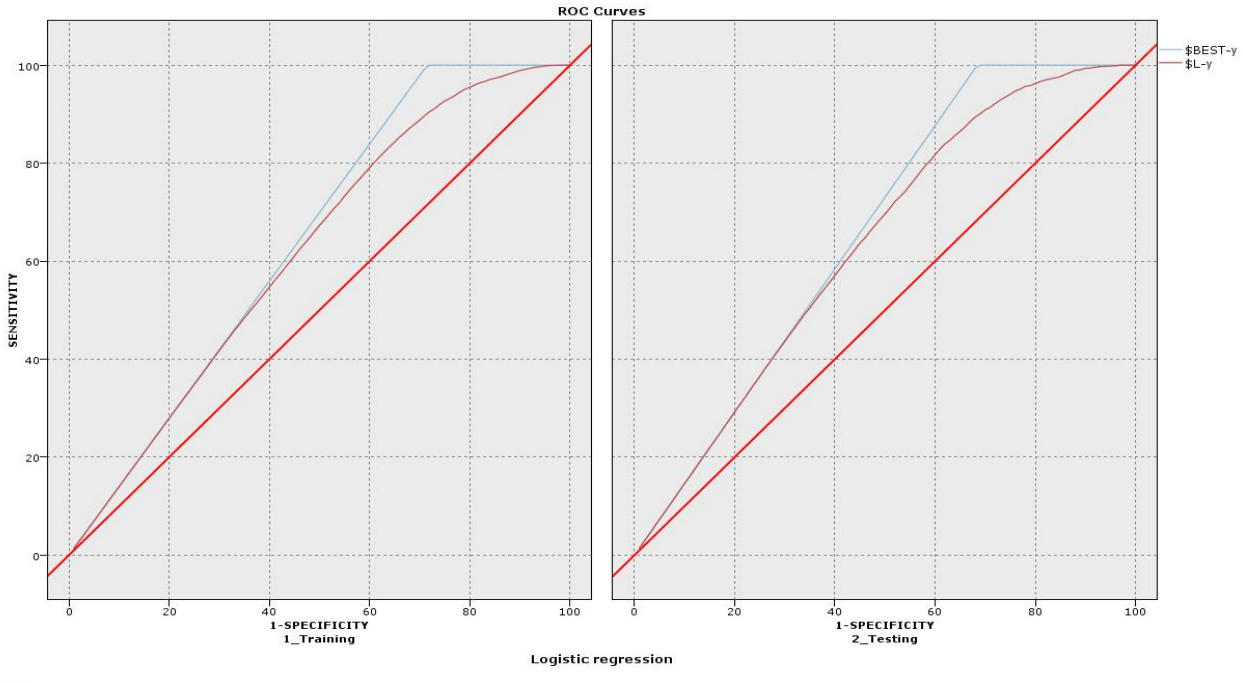
Πίνακας 6: Πίνακας συνάφειας(LR)

Επομένως έχουμε

	Σύνολο εκπαίδευσης	Σύνολο ελέγχου
Ευαισθησία (sensitivity)	90,81%	92,07%
Ειδικότητα (specificity)	73,39%	71,77%
Θετική προγνωστική αξία	89,53%	87,60%
Αρνητική προγνωστική αξία	76,11%	80,60%
Ακρίβεια (accuracy)	85,84%	85,66%

Πίνακας 7: Μέτρα αξιολόγησης(LR)

Αξιολόγηση *της λογιστικής παλινδρόμησης*



Σχήμα 31: ROC καμπύλη για το εκτιμώμενο μοντέλο($L-y) της λογιστικής παλινδρόμησης

6.3.2 Μέθοδος Backwards

Model name: (costum) Backwards.

Use partitioned data: το σημειώνουμε διότι αυτή η επιλογή μας εξασφαλίζει ότι χρησιμοποιούνται τα δεδομένα μόνο από το το training set για την κατασκευή του μοντέλου.

Procedure: Binomial

Method: **Backwards**

Στη συνέχεια ακολουθούμε την ίδια διαδικασία με προηγουμένως και πέρνουμε τα ακόλουθα αποτελέσματα

Coincidence matrices

Results for output field y

Comparing $L-y with y

'Partition'	1 Training		2 Testing	
Correct	6,655	85.85%	2,211	85.66%
Wrong	1,097	14.15%	370	14.34%
Total	7,752		2,581	

Coincidence Matrix for $L-y (rows show actuals)

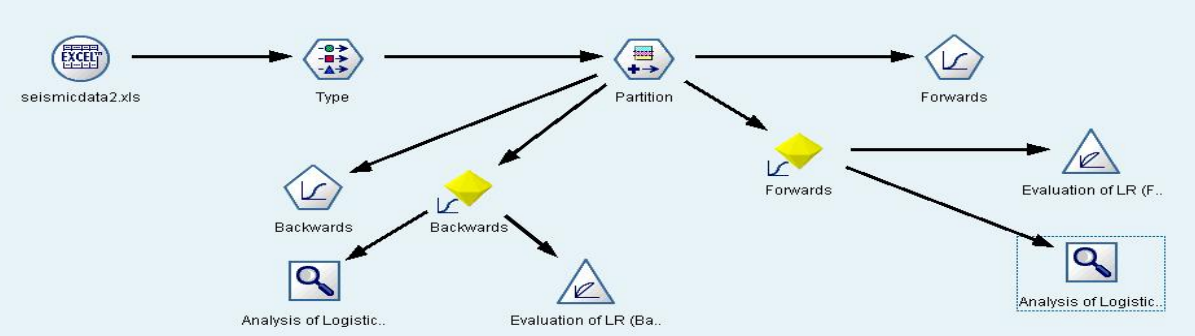
'Partition' = 1 Training	0.000000	1.000000
0.000000	1,622	588
1.000000	509	5,033
'Partition' = 2 Testing	0.000000	1.000000
0.000000	585	230
1.000000	140	1,626

Performance Evaluation

'Partition' = 1 Training	
0.000000	0.982
1.000000	0.225
'Partition' = 2 Testing	
0.000000	0.938
1.000000	0.247

Παρατηρούμε ότι τα αποτελέσματα που πήραμε είναι ακριβώς τα ίδια με την μέθοδο forwards. Επομένως οι δύο μέθοδοι είναι ισοδύναμες, αφού δίνουν την ίδια ακρίβεια.

Συνοπτικά το stream canvas που προέκυψε από την λογιστική παλινδρόμηση είναι το ακόλουθο:



6.3.3 Συμπεράσματα ανάλυσης

Στόχος μας εξ αρχής ήταν να δημιουργήσουμε ένα φειδωλό μοντέλο που θα εξηγεί με τον καλύτερο δυνατό τρόπο το σημαντικότατο πρόβλημα που έχουμε στην πραγματική ζωή και αφορά στην πρόβλεψη των σεισμών.

Ο πιο βασικός στόχος, λοιπόν, είναι αφενός να ανιχνεύσουμε τους εκείνους τους παράγοντες που έχουν συνεισφορά σ'έναν σεισμό και αφετέρου να εξερευνήσουμε τον τρόπο με τον οποίο αυτό πετυχαίνεται.

Όπως επισημάναμε και στην αρχή του κεφαλαίου το σύνολο δεδομένων μας αποτελείται από 10333 σεισμούς, από τη μεταβλητή απόκρισης y η οποία αναφέρεται στο μέγεθος του σεισμού και κωδικοποιείται με 0-1 ($μέγεθος > 6.5: 1,\ αλλιώς: 0$), όπου η σκάλα μεγέθους χρησιμοποιείται για να εκφράσει την ενέργεια που απελευθερώνεται από ένα σεισμό, και 9 παράγοντες.

Το εκτιμώμενο μοντέλο που προέκυψε από την εφαρμογή της λογιστικής παλινδρόμησης είναι το ακόλουθο

$$1.8814 - 1.5934 * x_1 + 0.4791 * x_2 + 0.4844 * x_4 + 0.9662 * x_5 + 1.2827 * x_7 - 0.1536 * x_8 \\ -2.3002 * x_9 - 1.3152 * x_{10} + 1.3955\, x_{11}$$

Παρατηρούμε ότι ο ποιος σημαντικός παράγοντας για τα σεισμολογικά δεδομένα βρέθηκε να είναι η μεταβλητή «χρόνια» (x_1). Ειδικότερα, σύμφωνα με τον αρνητικό συντελεστή του στο μοντέλο καταλήγουμε στο συμπέρασμα ότι τα τελευταία χρόνια παρατηρούνται σεισμοί μικρότερου μεγέθους, δηλαδή καθώς περνούν τα χρόνια φθίνει και η ένταση των σεισμικών δονήσεων. Αυτός ο συσχετισμός μεταξύ του έτους και του μεγέθους, μπορεί να υποδεικνύει μια περιοδικότητα της σεισμικής δραστηριότητας και χρειάζεται περαιτέρω έρευνα από τους σεισμολόγους. Το θέμα αυτό είναι πολύ κρίσιμο για την πρόγνωση των σεισμών αλλά και για την εκτίμηση των επιδράσεων τους στις κτιριακές εγκαταστάσεις, και γι'αυτό το λόγο πολλοί ερευνητές έχουν επικεντρωθεί στην μελέτη αυτών τις τελευταίες δεκαετίες. Ο δεύτερος στατιστικά σημαντικός παράγοντας είναι η υπεραπόσταση, η οποία είναι η Ευκλείδεια απόσταση, που σημαίνει ότι

$$υπεραπόσταση = \sqrt{(epicenterx)^2 + (epicentery)^2}$$

Αυτά τα δύο γεωγραφικά χαρακτηριστικά του επικέντρου μπορεί να είναι χρήσιμα για τον προσδιορισμό περιοχών που παρουσιάζουν υψηλό κίνδυνο για ισχυρούς σεισμούς. Οι σεισμολόγοι ενδιαφέρονται πολύ για μελέτες σχετικά με αυτές τις περιοχές, δεδομένου ότι μπορούν να τις συνδυάσουν με τα συμπεράσματα του σχετικά με τις τεκτονικές πλάκες.

Λαμβάνοντας υπόψη αυτά τα χαρακτηριστικά και τις χρονικές στιγμές που απελευθερώνεται ενέργεια μέσω των σεισμών, οι ερευνητές έχουν χρήσιμες πληροφορίες για την πρόγνωση τους. Ένα ακόμα σημαντικό στοιχείο είναι το «βάθος». Το αποτέλεσμα είναι πιθανό, δεδομένου ότι είναι γνωστό ότι η ισχυρότερη σεισμική δραστηριότητα παρατηρείται στην εξωτερική επιφάνεια της Γης. Μεταξύ των 9 παραγόντων, που συμπεριλαμβάνονται στο σύνολο δεδομένων, υπάρχουν παράγοντες που είναι εγγενείς ενός σεισμού, όπως το βάθος (x_{11}), κάποια γεωλογικά χαρακτηριστικά, όπως η υπεραπόσταση (X_7) και γεωγραφικά χαρακτηριστικά, όπως το γεωγραφικό πλάτος (x_4), το αζιμούθιο (x_8), η τεταγμένη του επίκεντρου (X_9) και η τετμημένη του επικέντρου (x_{10}).

6.4 Δέντρα αποφάσεων

6.4.1 C5.0

Ο κόμβος C5.0 κατασκευάζει είτε ένα δέντρο απόφασης ή ένα σύνολο κανόνων. Το μοντέλο λειτουργεί διασπώντας το δείγμα με βάση το πεδίο που παρέχουν το μέγιστο κέρδος πληροφοριών σε καθένα επίπεδο. Το πεδίο στόχος πρέπει να είναι κατηγορικό. Επιτρέπονται πολλαπλές διασπάσεις σε περισσότερες από δύο υποομάδες.

Για να εκτιμήσουμε το μοντέλο με τη μέθοδο του δέντρου C5.0 ακολουθούμε την διαδικασία που περιγράφουμε στη συνέχεια.

Επιλέγουμε τη διαδρομή:

Modeling ⟶ *Classification* ⟶ *C5.0*

Δημιουργώντας με αυτό τον τρόπο έναν **Modeling Node**  μπορούμε να επιλέξουμε τις παραμέτρους του μοντέλου μας και να εκτελούμε το μοντέλο.

Τον **C5.0 Node** τον συνδέουμε στον Partition node έτσι ώστε να ληφθεί υπόψην στη μοντελοποίηση ο διαχωρισμός σε σύνολο εκπαίδευσης και σύνολο εξέτασης.

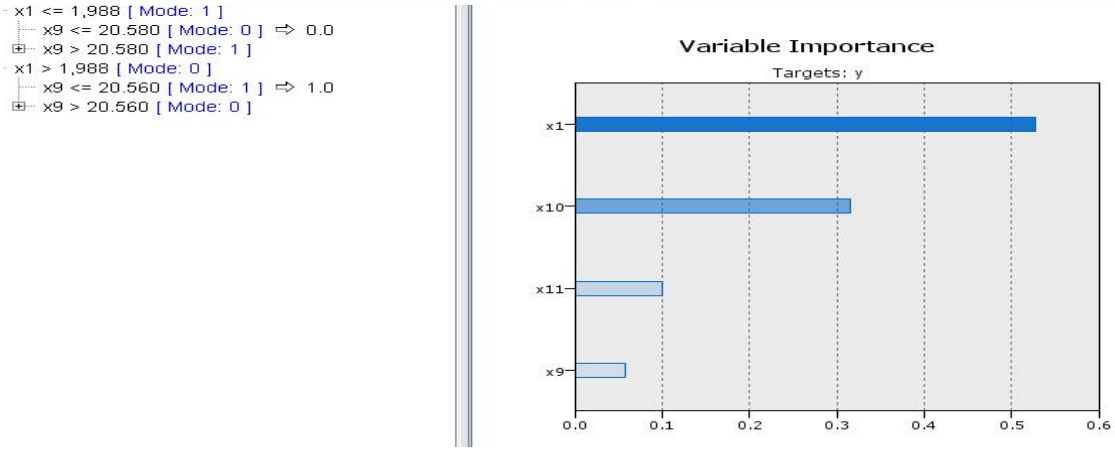
Modeling Node:

Model name: (costum) C5.0.

Use partitioned data: το σημειώνουμε διότι αυτή η επιλογή μας εξασφαλίζει ότι χρησιμοποιούνται τα δεδομένα μόνο από το το training set για την κατασκευή του μοντέλου.

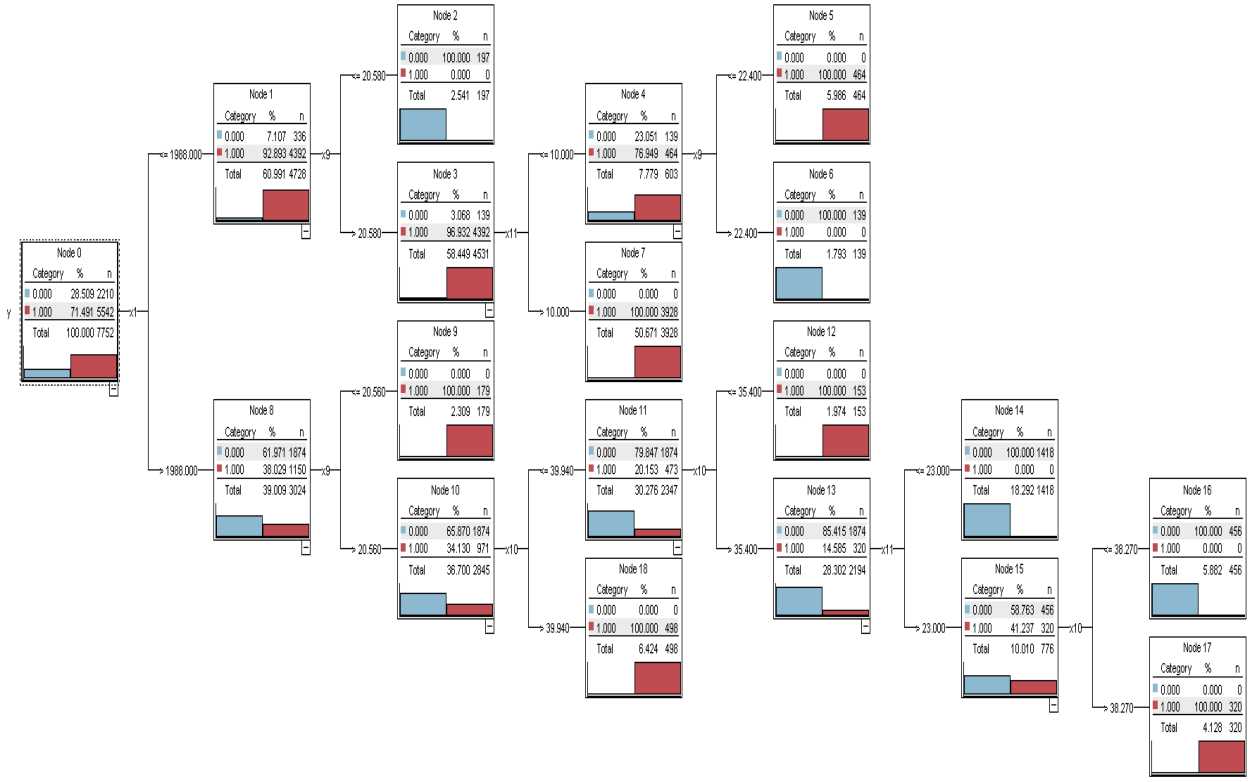
Εκτελούμε το μοντέλο και παράγουμε ένα **model nugget** που είναι απαραίτητο στη συνέχεια για την αξιολόγηση του μοντέλου.

Τα αποτελέσματα που προέκυψαν από την προσαρμογή του μοντέλου είναι τα ακόλουθα:



Σχήμα 32: κατάταξη των μεταβλητών ανάλογα με τη σημαντικότητά τους (C5.0)

Το δέντρο που προέκυψε είναι το ακόλουθο:



Σχήμα 33: Δέντρο που προέκυψε εφαρμόζοντας τον C5.0

Τέλος συνδέουμε έναν **Analysis node** για να δούμε τα μέτρα αξιολόγησης του μοντέλου καθώς και ένα **evaluation chart** για να σχεδιάσουμε την roc καμπύλη.

Στη συνέχεια ακολουθεί η αξιολόγηση του μοντέλου μέσω του

Coincidence matrices

Results for output field y

Comparing $C-y with y

'Partition'	1 Training		2 Testing	
Correct	7,752	100%	2,581	100%
Wrong	0	0%	0	0%
Total	7,752		2,581	

Coincidence Matrix for $C-y (rows show actuals)

'Partition' = 1 Training	0.000000	1.000000
0.000000	2,210	0
1.000000	0	5,542
'Partition' = 2 Testing	0.000000	1.000000
0.000000	815	0
1.000000	0	1,766

Performance Evaluation

'Partition' = 1 Training	
0.000000	1.255
1.000000	0.336
'Partition' = 2 Testing	
0.000000	1.153
1.000000	0.379

Αποτέλεσμα του τεστ

Πραγματική τιμή	Σύνολο εκπαίδευσης			Σύνολο Ελέγχου		
	0	**1**		**0**	**1**
0	TN = 2210	FP = 0	**0**	TN = 815	FP = 0
1	FN = 0	TP=5542	**1**	FN = 0	TP = 1766

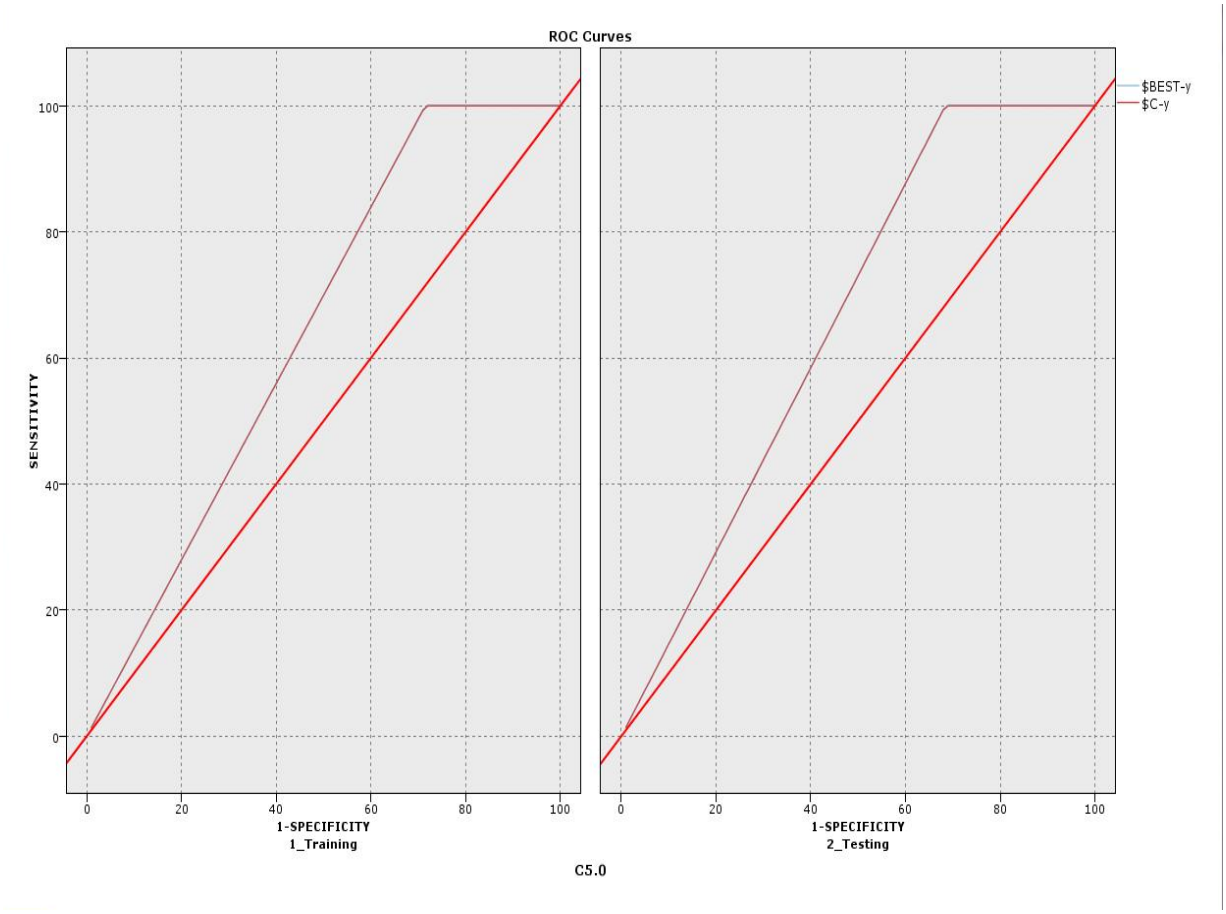
Πίνακας 8: Πίνακας συνάφειας (C5.0)

Επομένως έχουμε

	Σύνολο εκπαίδευσης	Σύνολο ελέγχου
Ευαισθησία (sensitivity)	100%	100%
Ειδικότητα (specificity)	100%	100%
Θετική προγνωστική αξία	100%	100%
Αρνητική προγνωστική αξία	100%	100%
Ακρίβεια (accuracy)	100%	100%

Πίνακας 9: Μέτρα αξιολόγησης(C5.0)

***Αξιολόγηση** του δέντρου C5.0*



Σχήμα 34: ROC καμπύλη για το εκτιμώμενο μοντέλο($L-y) του δέντρου C5.0

6.4.2 CHAID

Ο κόμβος CHAID δημιουργεί δέντρα απόφασης προσδιορίζοντας την βέλτιστη διάσπαση με την χρήση των Χ-τετράγωνο στατιστικών. Σε αντίθεση με το δέντρο C&RT και τους κόμβους QUEST, ο CHAID μπορεί να δημιουργήσει μη δυαδικά δέντρα, πράγμα που σημαίνει ότι ορισμένες διασπάσεις έχουν περισσότερους από δύο κλάδους. Τα πεδια στόχου και τα πεδία πρόβλεψης μπορεί να είναι συνεχή ή κατηγορηματικά. Η εξαντλητική CHAID είναι μια τροποποίηση του CHAID που κάνει μια πιο εμπεριστατωμένη εργασία του εξετάζοντας όλες τις πιθανές διασπάσεις, αλλά παίρνει αρκετό χρόνο για να τις υπολογίσει.

Για να εκτιμήσουμε το μοντέλο με τη μέθοδο του δέντρου CHAID ακολουθούμε την διαδικασία που περιγράφουμε στη συνέχεια.

Επιλέγουμε τη διαδρομή:

Modeling ⟶ *Classification* ⟶ *CHAID*

Δημιουργώντας με αυτό τον τρόπο έναν **Modeling Node** 

μπορούμε να επιλέξουμε τις παραμέτρους του μοντέλου μας και να εκτελούμε το μοντέλο.

Τον **CHAID Node** τον συνδέουμε στον Partition node έτσι ώστε να ληφθεί υπόψην στη μοντελοποίηση ο διαχωρισμός σε σύνολο εκπαίδευσης και σύνολο εξέτασης.

Modeling Node:

Model name: (costum) CHAID

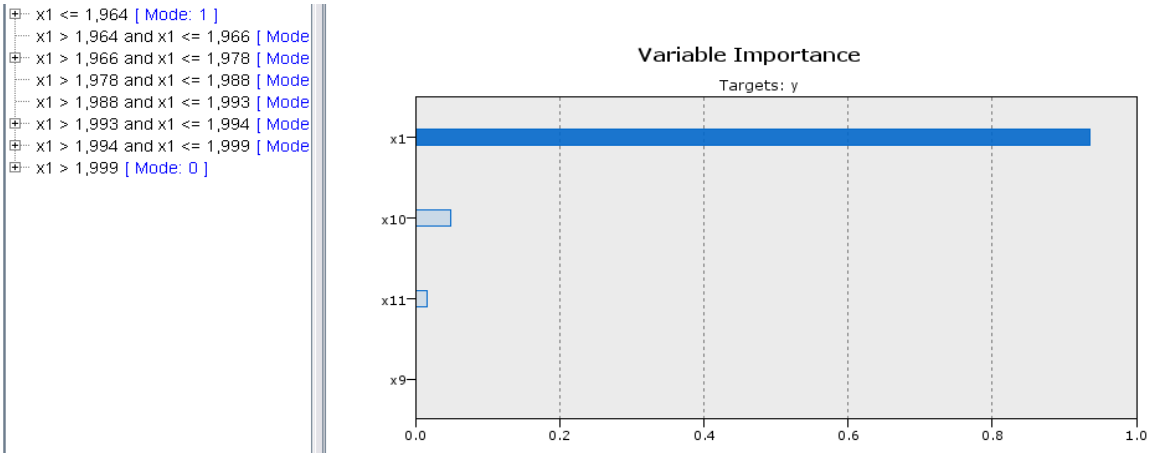
Use partitioned data: το σημειώνουμε διότι αυτή η επιλογή μας εξασφαλίζει ότι χρησιμοποιούνται τα δεδομένα μόνο από το το training set για την κατασκευή του μοντέλου.

Method: CHAID

Εκτελούμε το μοντέλο και παράγουμε ένα **model nugget** που είναι απαραίτητο στη συνέχεια για την αξιολόγηση του μοντέλου.

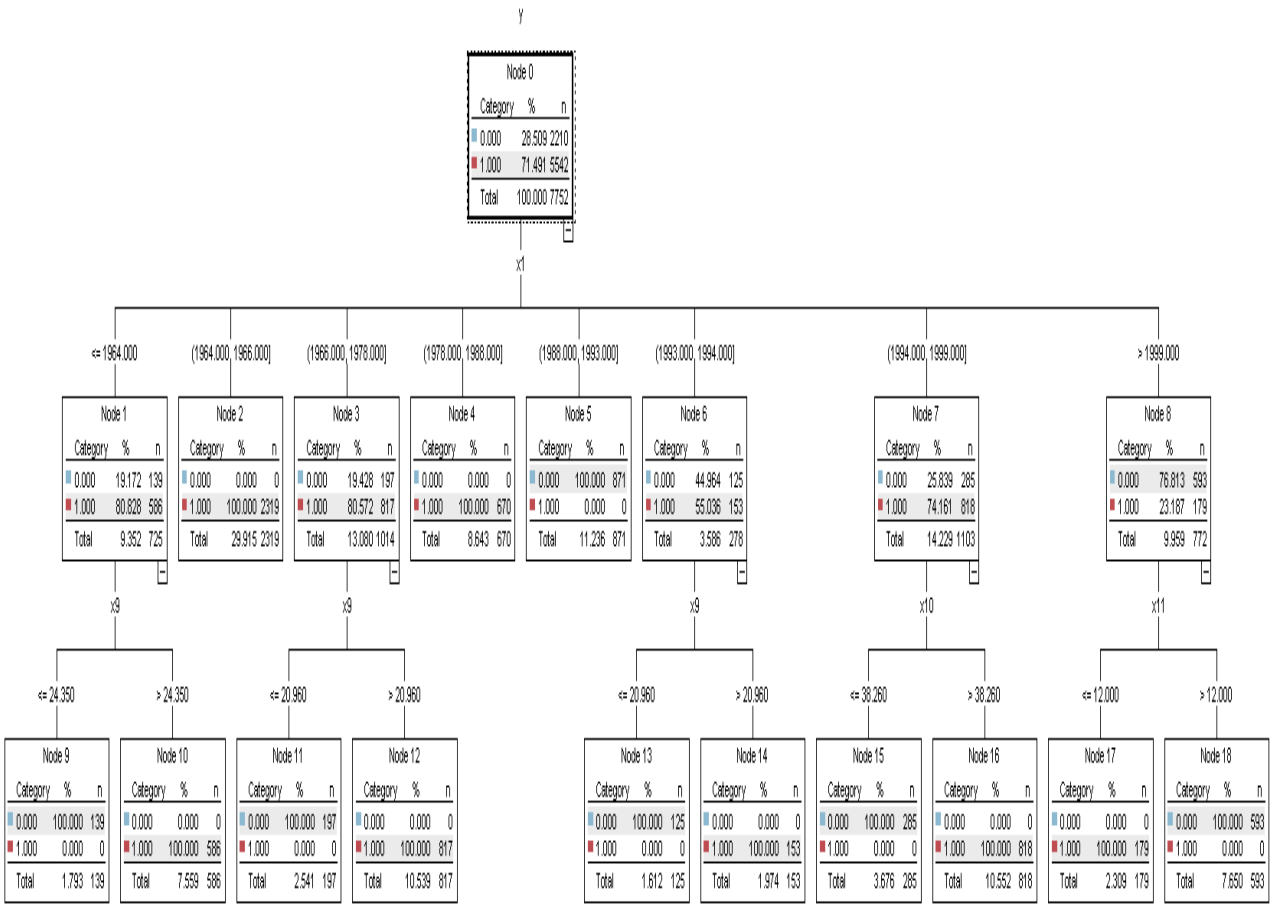
Τα αποτελέσματα που προέκυψαν από την προσαρμογή του μοντέλου είναι τα ακόλουθα:

Η σειρά των μεταβλητών ανάλογα με την σημαντικότητα τους είναι η ακόλουθη



Σχήμα 35: κατάταξη των μεταβλητών ανάλογα με τη σημαντικότητά τους (CHAID)

Το δέντρο που προέκυψε είναι το ακόλουθο:



Σχήμα 36: Δέντρο που προέκυψε εφαρμόζοντας τον CHAID

Τέλος συνδέουμε έναν **Analysis node** για να δούμε τα μέτρα αξιολόγησης του μοντέλου καθώς και ένα **evaluation chart** για να σχεδιάσουμε την roc καμπύλη.

Στην τελευταία ενότητα αυτού του κεφαλαίου παρουσιάζεται ένας συγκεντρωτικός πίνακας όπου παρουσιάζουμε τα μέτρα που μας ενδιαφέρουν για όλα τα δέντρα αποφάσεων.

6.4.3 C&RT

Ο κόμβος του δένδρου Ταξινόμησης και Παλινδρόμησης (C&R) δημιουργεί ένα δέντρο απόφασης που μας επιτρέπει να προβλέψουμε ή να ταξινομήσουμε τις μελλοντικές παρατηρήσεις. Η μέθοδος χρησιμοποιεί αναδρομικό διαχωρισμό για να χωρίσει τις εγγραφές της εκπαίδευσης σε τμήματα ελαχιστοποιώντας την ακαθαρσία σε κάθε βήμα, όπου ένας

κόμβος θεωρείται «καθαρός» εάν το 100% των περιπτώσεων στον κόμβο πέφτουν σε μια συγκεκριμένη κατηγορία του πεδίου στόχου. Τα πεδία στόχου και πρόβλεψης μπορεί να είναι συνεχή ή κατηγορηματικά αλλά όλες οι διασπάσεις είναι δυαδικές (μόνο δύο υποομάδες).

Για να εκτιμήσουμε το μοντέλο με τη μέθοδο του δέντρου C&R Tree ακολουθούμε την διαδικασία που περιγράφουμε στη συνέχεια.

Επιλέγουμε τη διαδρομή:

Modeling ⟶ *Classification* ⟶ *C&R Tree*

Δημιουργώντας με αυτό τον τρόπο έναν **Modeling Node**  μπορούμε να επιλέξουμε τις παραμέτρους του μοντέλου μας και να εκτελούμε το μοντέλο.

Τον **C&R Tree Node** τον συνδέουμε στον Partition node έτσι ώστε να ληφθεί υπόψην στη μοντελοποίηση ο διαχωρισμός σε σύνολο εκπαίδευσης και σύνολο εξέτασης.

Modeling Node:

Model name: (costum) C&RT

Use partitioned data: το σημειώνουμε διότι αυτή η επιλογή μας εξασφαλίζει ότι χρησιμοποιούνται τα δεδομένα μόνο από το το training set για την κατασκευή του μοντέλου.

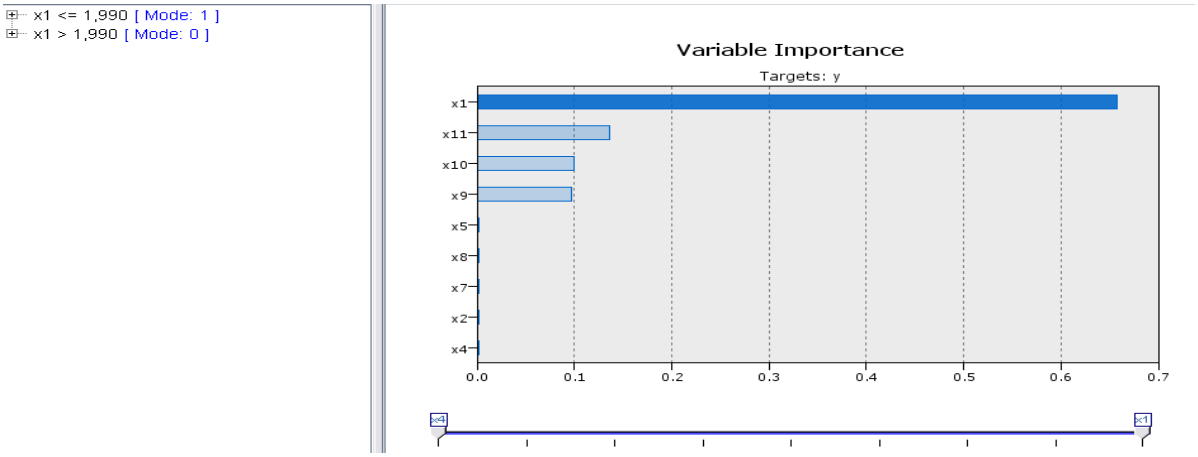
Expert: ***Gini / Twoing***

Εκτελούμε το μοντέλο και παράγουμε ένα **model nugget** που είναι απαραίτητο στη συνέχεια για την αξιολόγηση του μοντέλου.

Τέλος συνδέουμε έναν **Analysis node** για να δούμε τα μέτρα αξιολόγησης του μοντέλου καθώς και ένα **evaluation chart** για να σχεδιάσουμε την roc καμπύλη.

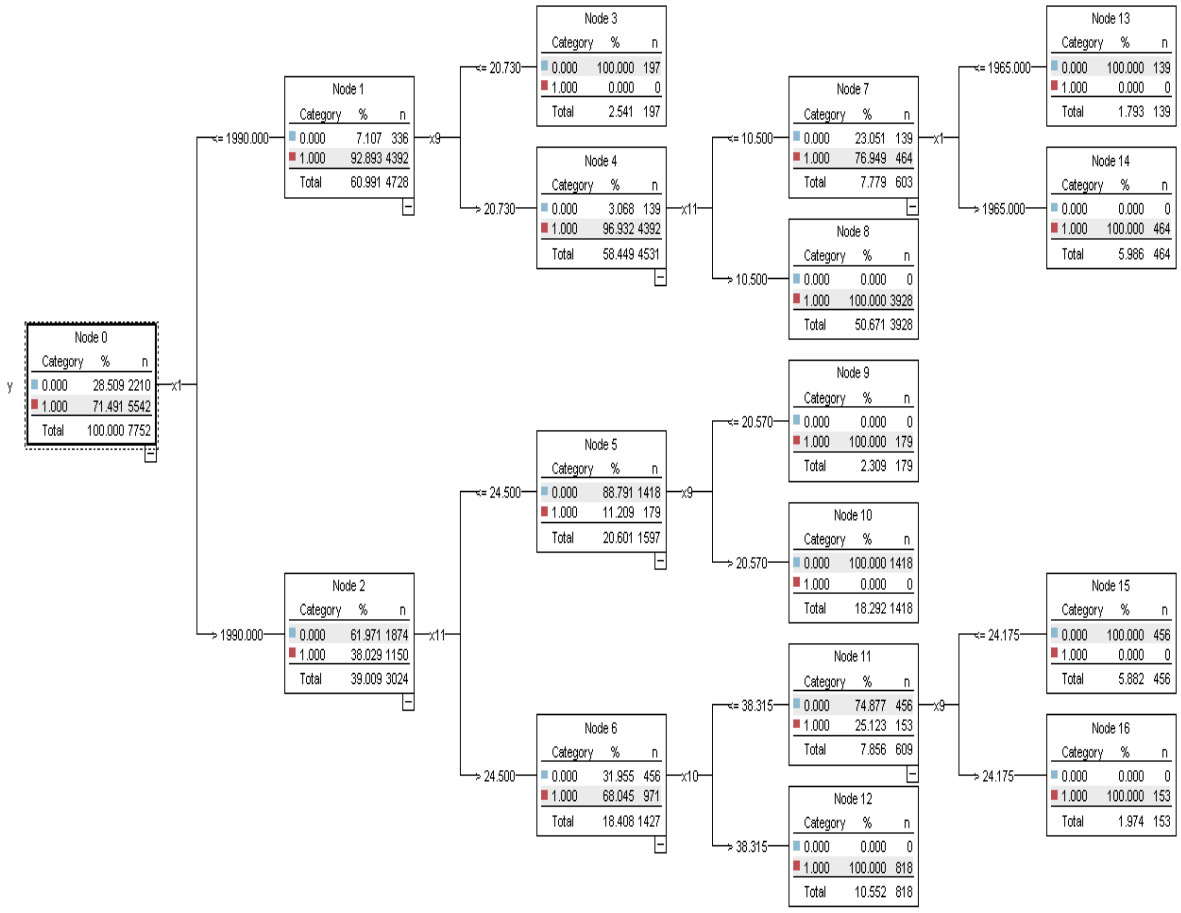
Στην τελευταία ενότητα αυτού του κεφαλαίου παρουσιάζεται ένας συγκεντρωτικός πίνακας όπου παρουσιάζουμε τα μέτρα που μας ενδιαφέρουν για όλα τα δέντρα αποφάσεων.

Τα αποτελέσματα που προέκυψαν από την προσαρμογή του μοντέλου είναι τα ακόλουθα:



Σχήμα 37: κατάταξη των μεταβλητών ανάλογα με τη σημαντικότητά τους (C&RT)

Το δέντρο που προέκυψε είναι το ακόλουθο:



Σχήμα 38: Δέντρο που προέκυψε εφαρμόζοντας τον C&RT με τη χρήση του μέτρου Gini

6.4.4 QUEST

Ο κόμβος QUEST παρέχει μια δυαδική μέθοδο ταξινόμησης για την κατασκευή δέντρων απόφασης, έχοντας σχεδιαστεί για να μειώσει το χρόνο επεξεργασίας που απαιτείται για τις μεγάλες C&R Tree αναλύσεις, ενώ επίσης, μειώνει την τάση που βρέθηκε σε μεθόδους δέντρων ταξινόμησης για να ευνοεί μεταβλητές που επιτρέπουν περισσότερες διασπάσεις.

Οι επεξηγηματικές μεταβλητές μπορεί να είναι συνεχείς, αλλά το πεδίο-στόχος πρέπει να είναι κατηγορηματική. Όλες οι διασπάσεις είναι δυαδικές.

Για να εκτιμήσουμε το μοντέλο με τη μέθοδο του δέντρου QUEST ακολουθούμε την διαδικασία που περιγράφουμε στη συνέχεια.

Επιλέγουμε τη διαδρομή:

Modeling → *Classification* → *QUEST*

Δημιουργώντας με αυτό τον τρόπο έναν **Modeling Node** 

μπορούμε να επιλέξουμε τις παραμέτρους του μοντέλου μας και να εκτελούμε το μοντέλο.

Τον **QUEST Node** τον συνδέουμε στον Partition node έτσι ώστε να ληφθεί υπόψην στη μοντελοποίηση ο διαχωρισμός σε σύνολο εκπαίδευσης και σύνολο εξέτασης.

Ακολουθόντας ακριβώς την ίδια διαδικασία με προηγουμένως λαμβάνουμε το δέντρο απόφασης καθώς και τα μέτρα αξιολόγησης που είναι απαράιτητα για την αξιολόγηση του μοντέλου που προέκυψε από την εφαρμογή της μεθόδου QUEST.

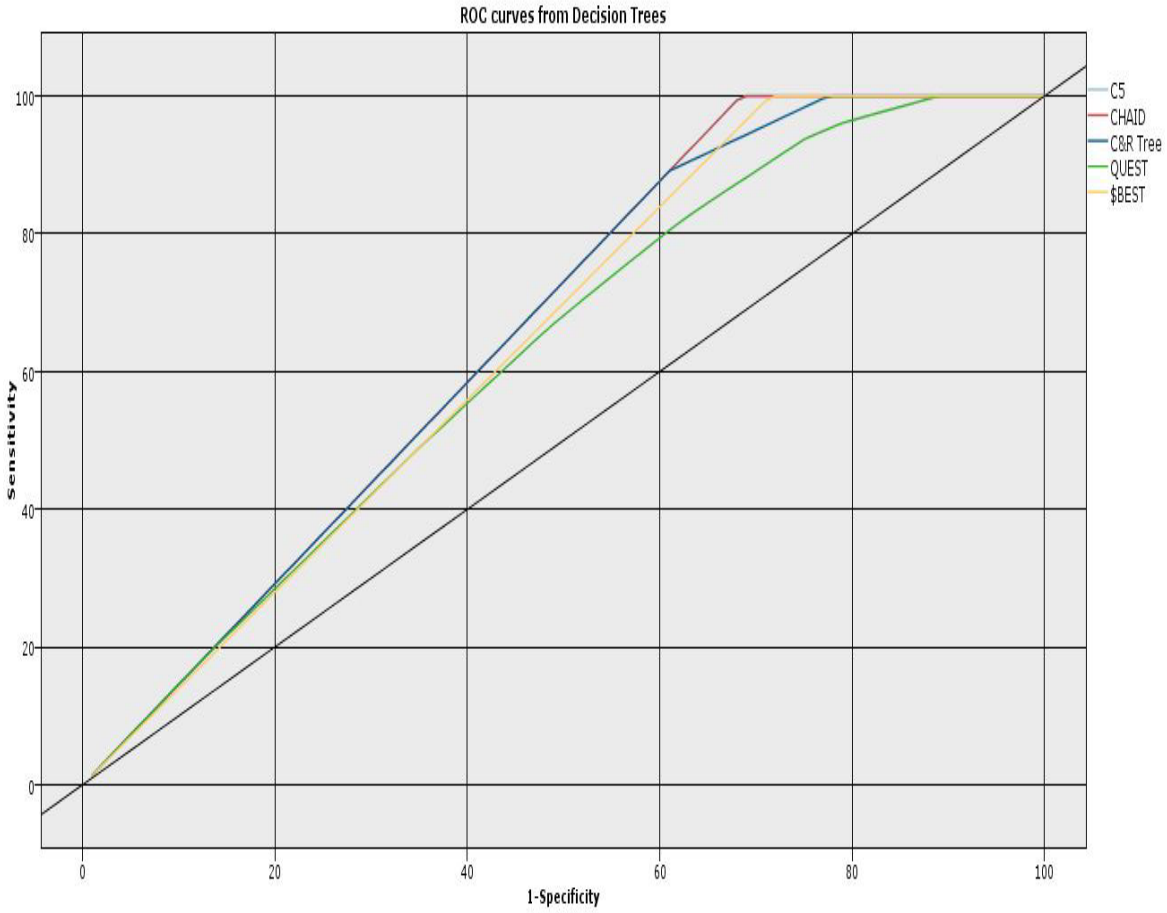
6.4.5 Σύγκριση δέντρων

: Με τον τρόπο αυτό μπορούμε να συγκρίνουμε τους ταξινομητές που επιθυμούμε. Έτσι για τα δέντρα αποφάσεων έχουμε τον ακόλουθο συγκεντρωτικό πίνακα:

	Ακρίβεια		Ευαισθησία		Ειδικότητα	
Ταξινομητής	Σύνολο Εκπαίδευσης	Σύνολο Ελέγχου	Σύνολο Εκπαίδευσης	Σύνολο Ελέγχου	Σύνολο Εκπαίδευσης	Σύνολο Ελέγχου
C5.0	100%	100%	100%	100%	100%	100%
CHAID	99,24%	99,35%	99,7%	99,72%	98,05%	98,49%
C&RT	99,24%	99,25%	98,95%	99,14%	100%	100%
QUEST	99,07%	98,85%	97,64%	97,89%	97,43%	97,95%

Πίνακας 10: Πίνακας σύγκρισης δέντρων απόφασης

Ο αλγόριθμος C5.0 έχει σαφώς καλύτερη ακρίβεια ταξινόμησης, ευαισθησία, και ειδικότητα που φτάνει το απόλυτο ποσοστό του 100% τόσο στο σύνολο εκπαίδευσης όσο και στο σύνολο δοκιμής. Συγκρίνοντας τα δέντρα CHAID και C&RT, το πρώτο έχει περισσότερες για το σύνολο εκπαίδευσης είναι τα ίδια και για τους δύο ταξινομητές. Ο CHAID έχει σωστά ταξινομημένες εγγραφές στο σύνολο ελέγχου (99,35%), ενώ τα αντίστοιχα ποσοστά επαρκή ειδικότητα ενώ το ίδιο μέτρο για τον C&RT είναι 100% πράγμα που σημαίνει ότι ο ταξινομητής αναγνωρίζει όλα τα πραγματικά αρνητικά (TN). Με άλλα λόγια αυτό σημαίνει ότι ο C&RT έχει χαμηλό Type I ποσοστό σφάλματος. Το μέτρο αυτό από μόνο του δεν μας λέει πόσο καλά ο ταξινομητής αναγνωρίζει θετικές περιπτώσεις και γι 'αυτό είναι αναγκαίο να ληφθεί υπόψη και η ευαισθησία των χρησιμοποιούμενων ταξινομητών. Όταν οι δύο αλγόριθμοι αξιολογήθηκαν σύμφωνα με την ευαισθησία, ο CHAID είχει σαφές πλεονέκτημα έχοντας υψηλότερα ποσοστά, πράγμα που σημαίνει ότι το Type II ποσοστά σφάλματος είναι χαμηλότερα από τα αντίστοιχα του C&RT αλγορίθμου. Ο QUEST έχει τη χειρότερη επίδοση σε σύγκριση με τους C5.0, CHAID και C&RT από την άποψη της ακρίβειας, της ευαισθησίας και της ειδικότητας. Το σχήμα 35 εμφανίζει τις καμπύλες ROC που προέρχονται από όλους τους αλγορίθμους των δέντρων αποφάσεων. Η AUROC είναι 1, 1, 0,98 και 0,95 για το C5.0, CHAID, C & RT και QUEST αντίστοιχα. Σε γενικές γραμμές, ο C5.0 και ο CHAID φαίνεται να έχουν υψηλότερες επιδόσεις από τον C&RT, και στη συνέχεια ακολουθεί ο QUEST με τη χειρότερη επίδοση στις τιμές των κριτηρίων.



Σχήμα 39: Καμπύλες ROC που προέκυψαν από τα δέντρα αποφάσεων

6.5 Νευρωνικά δίκτυα

Ο κόμβος του νευρωνικού δικτύου (Neural Net) χρησιμοποιεί ένα απλοποιημένο μοντέλο του τρόπου με τον οποίο γίνονται οι διεργασιες πληροφοριών στον ανθρώπινο εγκέφαλο. Λειτουργεί με την προσομοίωση ενός μεγάλου αριθμού απλών διασυνδεδεμένων μονάδων επεξεργασίας που μοιάζουν με αφηρημένες εκδόσεις των νευρώνων. Τα νευρωνικά δίκτυα είναι ισχυρές γενικές εκτιμητικές συναρτήσεις και απαιτούν ελάχιστη στατιστική ή μαθηματική γνώση για να εκπαιδευτούν ή να εφαρμοστούν.

Για να εκτιμήσουμε το μοντέλο με τη μέθοδο Neural net ακολουθούμε την διαδικασία που περιγράφουμε στη συνέχεια.

Επιλέγουμε τη διαδρομή:

Modeling ⟶ *Classification* ⟶ *Neural net*

Δημιουργώντας με αυτό τον τρόπο έναν **Modeling Node**  μπορούμε να επιλέξουμε τις παραμέτρους του μοντέλου μας και να εκτελούμε το μοντέλο.

Τον **Neural net Node** τον συνδέουμε στον Partition node έτσι ώστε να ληφθεί υπόψην στη μοντελοποίηση ο διαχωρισμός σε σύνολο εκπαίδευσης και σύνολο εξέτασης.

6.5.1 MLP

Modeling Node:

Model name: (costum) Neural Net-MLP

Use partitioned data: το σημειώνουμε διότι αυτή η επιλογή μας εξασφαλίζει ότι χρησιμοποιούνται τα δεδομένα μόνο από το το training set για την κατασκευή του μοντέλου.

Method: Multiple

Topologies: 3 / 5 / 7 / 9

Εκτελούμε το μοντέλο σε κάθε μία από τις περιπτώσεις και παράγουμε ένα **model nugget** που είναι απαραίτητο στη συνέχεια για την αξιολόγηση του μοντέλου.

6.5.2 RBFN

Modeling Node:

Model name: (costum) Neural Net-MLP

Use partitioned data: το σημειώνουμε διότι αυτή η επιλογή μας εξασφαλίζει ότι χρησιμοποιούνται τα δεδομένα μόνο από το το training set για την κατασκευή του μοντέλου.

Method: RBFN

Topologies: 3 / 5 / 7 / 9

Εκτελούμε το μοντέλο και παράγουμε ένα **model nugget** που είναι απαραίτητο στη συνέχεια για την αξιολόγηση του μοντέλου.

Τα αποτελέσματα που προέκυψαν από την προσαρμογή του μοντέλου είναι τα ακόλουθα:

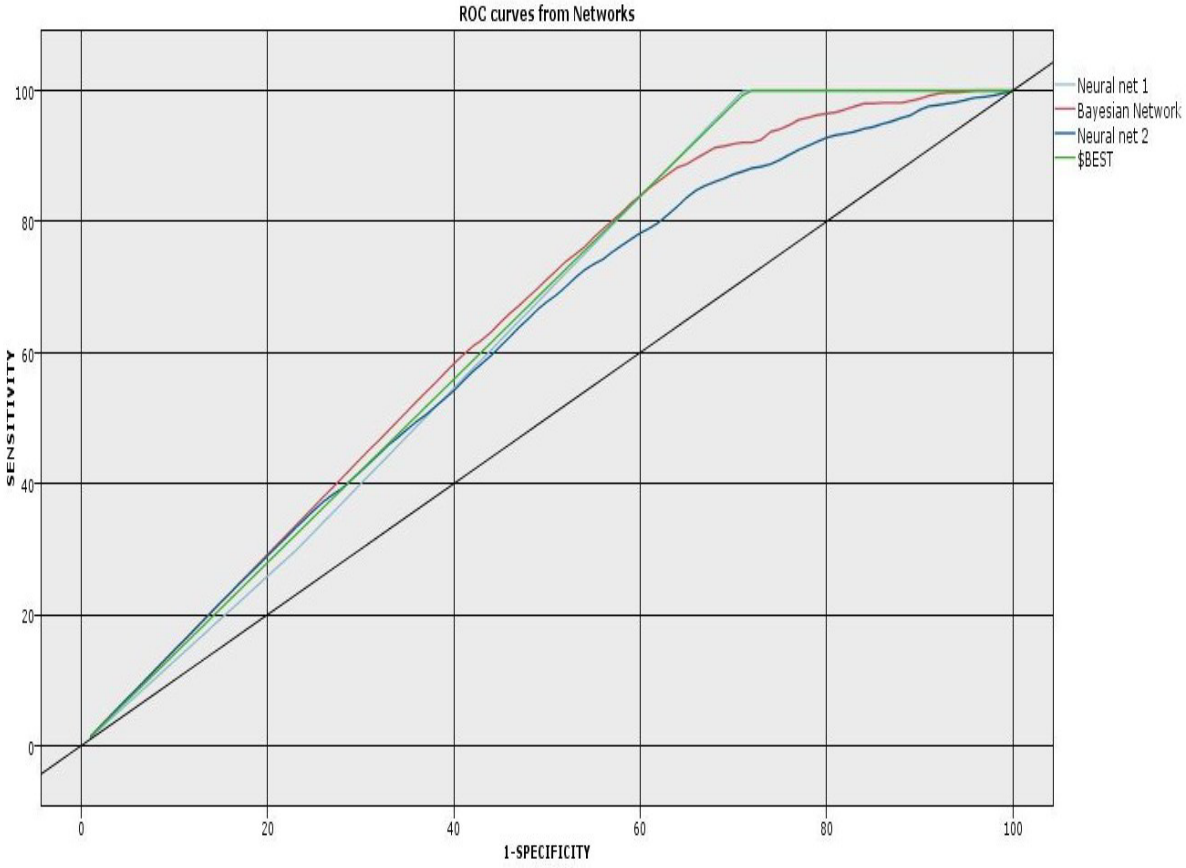
6.5.3 Σύγκριση δικτύων

Μέθοδος	Κρυμμένες μονάδες	Εκτιμώμενη ακρίβεια (%)
RBFN	3	**83.51**
RBFN	5	81.23
RBFN	7	80.53
RBFN	9	81.70
MLP	3	**97.69**
MLP	5	97.5
MLP	7	97.45
MLP	9	97.33

Πίνακας 11: Πίνακας οπου απεικονίζονται οι κρυμμένες μονάδες και η εκτιμώμενη ακρίβεια

Εξετάζουμε αρκετό αριθμό μονάδων, όπως 3, 5, 7 και 9 στο κρυφό στρώμα, ώστε να προσδιοριστεί ο βέλτιστος αριθμός των νευρώνων στο κρυμμένο στρώμα. Ο παραπάνω πίνακας δείχνει την εκτιμώμενη ακρίβεια της δυαδικής ταξινόμησης με 3, 5, 7 και 9 μονάδες στο κρυμμένο στρώμα για κάθε μία από τις μεθόδους MLP και RBFN. Ο αριθμός των νευρώνων στο κρυμμένο στρώμα επιλέγεται να είναι 3, δεδομένου ότι αυτή η τιμή βρέθηκε να είναι η βέλτιστη, δίνοντας την υψηλότερη αναμενόμενη ακρίβεια τόσο για την MLP όσο και για την RBFN.

Το δίκτυο MLP έχει σαφώς μεγαλύτερη ακρίβεια ταξινόμησης σε σύγκριση με το RBFN και το Bayesian δίκτυο τόσο για το σύνολο εκπαίδευσης όσο και για το σύνολο δοκιμής. Επιπλέον, το MLP δίκτυο επιτυγχάνει άριστα αποτελέσματα για την ειδικόττηα, και φθάνει στο απόλυτο ποσοστό του 100% για την ευαισθησία. Το σχήμα 40 απεικονίζει τις καμπύλες ROC που προέρχονται από τους MLP, RBFN και τα Bayesian δίκτυα (που θα παρουσιάσουμε στη συνέχεια). Η AUROC είναι 0.94, 0.93 και 0.86 για τα δίκτυα MLP, Bayesian και RBFN αντίστοιχα. Σε γενικές γραμμές, η τεχνική του MLP δικτύου υπερτερεί του Bayesian δικτύου, και στη συνέχεια ακολουθεί ο RBFN λαμβάνοντας υπόψη όλα τα σύνολα που προέρχονται από τη διάσπαση σε σύνολο εκπαίδευσης και σύνολο ελέγχου.



Σχήμα 40: Καμπύλες ROC που προέκυψε από τα δίκτυα

Ακολουθεί ένας πίνακας με τη σύγκριση όλων των δικτύων.

	Ακρίβεια		Ευαισθησία		Ειδικότητα	
Ταξινομητής	Σύνολο εκπαίδευσης	Σύνολο ελέγχου	Σύνολο εκπαίδευσης	Σύνολο ελέγχου	Σύνολο εκπαίδευσης	Σύνολο ελέγχου
MLP (Neural Net)	97.46%	97.37%	100 %	100 %	91.08%	91.65 %
Bayesian Network	87.65%	88.73 %	89.96 %	91.21 %	81.85 %	83.53 %
RBFN (Neural Net)	83.4%	84.11%	84.01 %	84.82 %	81.85%	82.57 %

Πίνακας 12: Σύγκριση προόδου της εκτέλεσης των δικτύων

6.6 Μηχανές διανυσματικής υποστήριξης

Ο κόμβος των μηχανών διανυσματικής υποστήριξης (SVM) μας δίνει τη δυνατότητα να ταξινομήσει τα δεδομένα σε μια από τις δύο ομάδες χωρίς να έχουμε υπερπροσαρμογή (Overfitting). Το SVM λειτουργεί καλά με ευρεία σύνολα δεδομένων, δηλαδή όταν έχουμε ένα πολύ μεγάλο αριθμό πεδίων πρόβλεψης.

Για να εκτιμήσουμε το μοντέλο με τη μέθοδο SVM ακολουθούμε την διαδικασία που περιγράφουμε στη συνέχεια.

Επιλέγουμε τη διαδρομή:

Modeling → *Classification* → *SVM*

Δημιουργώντας με αυτό τον τρόπο έναν **Modeling Node**  μπορούμε να επιλέξουμε τις παραμέτρους του μοντέλου μας και να εκτελούμε το μοντέλο.

Τον **SVM Node** τον συνδέουμε στον Partition node έτσι ώστε να ληφθεί υπόψην στη μοντελοποίηση ο διαχωρισμός σε σύνολο εκπαίδευσης και σύνολο εξέτασης.

Χρησιμοποιήσαμε του πυρήνες RBF, γραμμικό, σιγμοειδή και πολυωνυμικό και αρκετούς συνδυασμούς παραμέτρων για καθέναν από αυτούς (grid search).

Τέλος συγκρίναμε τα ωέλτιστα μοντέλα που προέκυψαν από τον καλύτεροσυνδυασμό παραμέτρων. Στην επόμενη ενότητα παρουσιάζουμε τα αποτελέσματα της μελέτης.

6.6.1 Συγκεντρωτικοί πίνακες – Grid search

Εκτός από τη λειτουργία του πυρήνα, η παράμετρος κανονικοποίησης C και η τιμή της γάμμα επιλέγονται από διάφορες υποψήφιες τιμές, και όπως επισημάναμε και προηγουμένως επιλέχθηκαν εκείνες που έχουν ως αποτέλεσμα την καλύτερη απόδοση. Αν ο τύπος του πυρήνα έχει επιλεγεί να είναι ο πολυωνυμικός ή ο σιγμοειδής η παράμετρος της μεροληψίας ορίζει την τιμή του συντελεστή στη συνάρτηση του πυρήνα και στις περισσότερες περιπτώσεις είναι κατάλληλη η προεπιλεγμένη τιμή 0. Ο βαθμός της παραμέτρου είναι ενεργοποιημένος μόνο εάν ο τύπος του πυρήνα έχει οριστεί να είναι ο πολυωνυμικός και ελέγχει την πολυπλοκότητα (διάσταση) του χώρου χαρτογράφησης (στη μελέτη μας, θέτουμε την τιμή d = 3).

Η παράμετρος κανονικοποίησης C ελέγχει το trade-off μεταξύ της μεγιστοποίησης του περιθωρίου και ελαχιστοποιώντας τον όρο του σφάλματος εκπαίδευσης. Η τιμή αυτή θα πρέπει κανονικά να είναι μεταξύ 1 και 10. Η αύξηση της τιμής βελτιώνει την ακρίβεια ταξινόμησης για τα δεδομένα εκπαίδευσης, αλλά αυτό μπορεί επίσης να οδηγήσει σε υπερπροσαρμογή. Η τιμή γάμμα πρέπει κανονικά να είναι μεταξύ

$$\frac{3}{k} \ (= \ 0.333)$$

και

$$\frac{6}{k} \ (= \ 0.666)$$

, όπου k είναι ο αριθμός των μεταβλητών εισόδου (9 στη μελέτη μας).

Η επιλογή παραμέτρων στα SVM μπορεί να θεωρηθεί ως μία διαδικασία βελτιστοποίησης, δεδομένου ότι η μέθοδος αναζήτησης πλέγματος εκτελείται με τις παραμέτρους ελέγχου C και γ, και να επιτύχει το καλύτερο δυνατό μοντέλο. Οι παρακάτω πίνακες δείχνουν τα αποτελέσματα από την αναζήτηση του πλέγματος που βασίζεται στα δεδομένα μας, χρησιμοποιώντας τους τέσσερις πυρήνες. Σε αυτή τη συγκριτική μελέτη, το καλύτερο μοντέλο με την υψηλότερη αναμενόμενη ακρίβεια επιτυγχάνεται χρησιμοποιώντας $C = 10$ και $\gamma = 0.66$ για όλους τους πυρήνες. Μετά τον εντοπισμό των βέλτιστων παραμέτρων κανονικοποίησης, έχουμε εκπαιδεύσει το τελικό μας μοντέλο και εκτιμήσαμε την προγνωστική ακρίβεια.

Το SVM με πολυωνυμικό πυρήνα έχει σαφώς μεγαλύτερη ακρίβεια ταξινόμησης σε σύγκριση με το l_1-norm SVM και τα SVMs με RBF, γραμμικό και σιγμοειδή πυρήνα για το σύνολο εκπαίδευσης και δοκιμής. Επιπλέον, η SVM με τον πολυωνυμικό πυρήνα έχει σημειώσει τις υψηλότερες τιμές για την ευαισθησία και την ειδικότητα. Το Gaussian SVM (RBF) ξεπερνά το γραμμικό SVM και το l_1-norm SVM από την άποψη της ακρίβειας, της ευαισθησίας και της ειδικότητας, και στη συνέχεια ακολουθεί το σιγμοειδές SVM. Στο σχήμα 41 εμφανίζονται οι καμπύλες ROC που προέρχονται από όλα τα SVMs με τους θεωρούμενους τέσσερις πυρήνες, και το σχήμα 42 εμφανίζει το l_1-norm SVM. Η AUROC είναι 0.98, 0.97, 0.90, 0.62 για τα SVMs με τον πολυωνυμικό, RBF, γραμμικό και σιγμοειδές πυρήνας αντίστοιχα. Η AUROC για το l_1-norm SVM παίρνει τη χαμηλότερη τιμή ίση με 0.59. Το l_1-norm SVM εκτελείται σχεδόν παρόμοια με το γραμμικό SVM από την άποψη της ακρίβειας, της ευαισθησίας και της ειδικότητας. Σημειώνουμε, ότι το l_1- norm SVM αναγνώρισε και τους 11 πιθανούς παράγοντες κινδύνου των μεγάλων σεισμών, ως στατιστικά σημαντικούς. Το l_1-norm SVM τείνει να έχει υψηλότερο σφάλματα τύπου I και χαμηλότερο σφάλμα τύπου II, με άλλα λόγια, αυτό σημαίνει ότι το l_1-norm SVM τείνει να σημειώνει σε μεγαλύτερο ποσοστό ανενεργές μεταβλητές να είναι ενεργές, και σε χαμηλότερο ποσοστό ενεργές μεταβλητές να είναι ανενεργές.

Προγνωστική ακρίβεια(%)										
Gaussian RBF										
0.34	c=1	c=2	c=3	c=4	c=5	c=6	c=7	c=8	c=9	c=10
Σύνολο εκπαίδευσης	86.24	87.33	87.81	88.05	88.31	88.45	88.62	88.69	88.83	**88.94**
Σύνολο ελέγχου	86.44	86.98	87.45	87.76	88.03	88.3	88.53	88.76	88.96	**89.23**
0.35	c=1	c=2	c=3	c=4	c=5	c=6	c=7	c=8	c=9	**c=10**
Σύνολο εκπαίδευσης	86.51	87.47	87.9	88.24	88.39	88.56	88.66	88.83	89.01	**89.13**
Σύνολο ελέγχου	86.48	87.1	87.6	87.99	88.22	88.53	88.76	89.07	89.35	**89.42**
0.4	c=1	c=2	c=3	c=4	c=5	c=6	c=7	c=8	c=9	**c=10**
Σύνολο εκπαίδευσης	86.96	87.73	88.13	88.44	88.54	88.85	89.01	89.11	89.25	**89.49**
Σύνολο ελέγχου	86.94	87.52	87.95	88.22	88.8	89.07	89.42	89.42	89.69	**89.89**
0.45	c=1	c=2	c=3	c=4	c=5	c=6	c=7	c=8	c=9	**c=10**
Σύνολο εκπαίδευσης	87.19	87.95	88.44	88.61	88.89	89.14	89.29	89.51	89.71	**89.89**
Σύνολο ελέγχου	87.18	87.87	88.18	88.88	89.27	89.46	89.73	89.85	89.85	**89.97**
0.5	c=1	c=2	c=3	c=4	c=5	c=6	c=7	c=8	c=9	**c=10**
Σύνολο εκπαίδευσης	87.45	88.13	88.54	88.91	89.18	89.4	89.64	89.95	90.14	**90.36**
Σύνολο ελέγχου	87.45	87.07	88.61	89.31	89.46	89.73	89.97	89.97	90	**90.31**
0.55	c=1	c=2	c=3	c=4	c=5	c=6	c=7	c=8	c=9	**c=10**
Σύνολο εκπαίδευσης	87.58	88.45	88.78	89.2	89.46	89.8	90.07	90.29	90.43	**90.56**
Σύνολο ελέγχου	87.64	88.26	89.07	89.46	89.73	90.04	90	90.43	90.59	**90.78**
0.6	c=1	c=2	c=3	c=4	c=5	c=6	c=7	c=8	c=9	**c=10**
Σύνολο εκπαίδευσης	87.9	88.51	89.28	89.37	89.85	90.07	90.36	90.42	90.66	**90.94**
Σύνολο ελέγχου	87.72	88.49	89.35	89.66	89.97	90.12	90.47	90.66	90.78	**91.24**
0.65	c=1	c=2	c=3	c=4	c=5	c=6	c=7	c=8	c=9	**c=10**
Σύνολο εκπαίδευσης	88.02	88.71	89.22	89.76	90.03	90.38	90.54	90.85	91.02	**91.15**
Σύνολο ελέγχου	87.87	89	89.5	89.93	90.2	90.55	90.78	91.2	91.44	**91.67**
0.66	c=1	c=2	c=3	c=4	c=5	c=6	c=7	c=8	c=9	**c=10**
Σύνολο εκπαίδευσης	87.96	88.73	89.25	89.81	90.09	90.42	90.6	90.89	91.01	**91.24**
Σύνολο ελέγχου	87.87	88.96	89.54	89.97	90.39	90.59	90.78	91.28	91.44	**91.82**

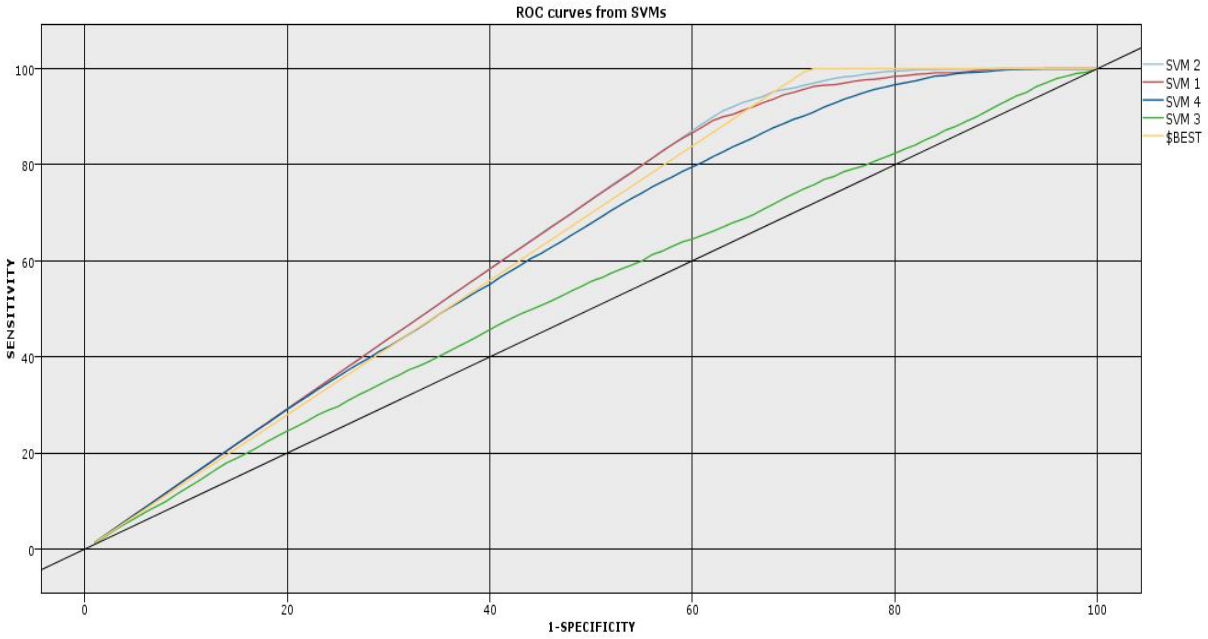
Πίνακας 13: Αποτελέσματα του grid search για τον RBF πυρήνα

	Προγνωστική ακρίβεια(%)									
Σιγμοειδής	c=1	c=2	c=3	c=4	c=5	c=6	c=7	c=8	c=9	**c=10**
Σύνολο εκπαίδευσης	71.7	71.71	71.71	71.72	71.72	71.72	71.72	71.71	71.71	71.74
Σύνολο ελέγχου	68.38	68.46	68.46	68.5	68.5	68.46	68.46	68.46	68.46	68.5
Γραμμικός	c=1	c=2	c=3	c=4	c=5	c=6	c=7	c=8	c=9	**c=10**
Σύνολο εκπαίδευσης	83.76	83.93	84	84.08	84.15	84.13	84.13	84.15	84.13	84.18
Σύνολο ελέγχου	83.92	84.23	84.35	84.54	84.54	84.62	84.62	84.66	84.7	84.73
Πολυωνυμικός	c=1	c=2	c=3	c=4	c=5	c=6	c=7	c=8	c=9	**c=10**
Σύνολο εκπαίδευσης	88.34	89.23	90.26	90.96	91.34	91.65	91.96	92.13	92.52	92.87
Σύνολο ελέγχου	88.45	89.11	90.55	91.13	91.71	91.98	92.13	92.44	92.75	93.26

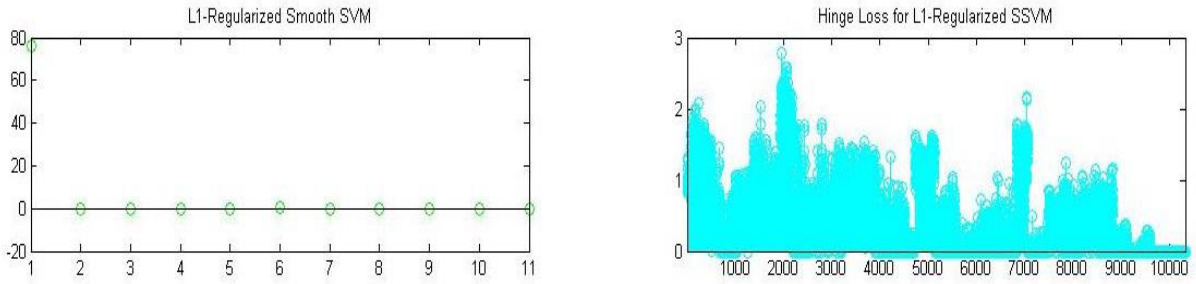
Πίνακας 14: Αποτελέσματα του grid search για σιγμοειδή, γραμμικό και πολυωνυμικό πυρήνα

	Ακρίβεια		Ευαισθησία		Ειδικότητα	
Ταξινομητής	Σύνολο εκπαίδευσης	Σύνολο ελέγχου	Σύνολο εκπαίδευσης	Σύνολο ελέγχου	Σύνολο εκπαίδευσης	Σύνολο ελέγχου
Polynomial (SVM 2)	92.86%	93.52%	95.05 %	95.97 %	87.37%	87.36 %
Gaussian RBF(SVM 1)	91.24%	91.82 %	93.82 %	94.79 %	84.75 %	85.39 %
Linear (SVM4)	84.35%	84.73%	90.76 %	92.52 %	67.51%	67.85 %
l1-norm SVM	83.96%	84%	89.9%	90%	69.62%	70%
Sigmoid (SVM3)	71.71%	68.46%	99.4 %	99.43 %	23%	14.7 %

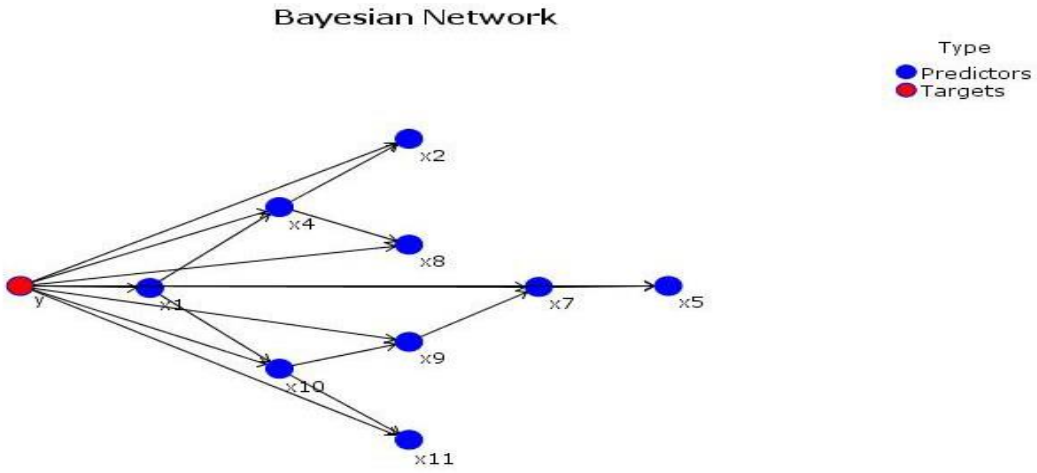
Πίνακας 15: Γενική σύγκριση της απόδοσης των SVM



Σχήμα 41: Καμπύλες ROC που προέκυψαν από τα SVM



Σχήμα 42: L1 norm SVM



Σχήμα 43: Μπεϋζιανό δίκτυο

6.7 Bayesian Network

Ο κόμβος του Bayesian Network μας δίνει τη δυνατότητα να οικοδομήσουμε ένα μοντέλο πιθανοτήτων, συνδυάζοντας παρατηρώντας και καταγράφοντας τις ενδείξεις με τις πραγματικές γνώσεις για τον καθορισμό της πιθανότητας των περιστατικών. Στην τρέχοουσα έκδοση του Clementine 12.0, ο κόμβος επικεντρώνεται στο δέντρο Augmented Naïve Bayes (TAN) και στα Markov Blanket δίκτυα που κατά κύριο λόγο χρησιμοποιούνται για την ταξινόμηση.

Για να εκτιμήσουμε το μοντέλο με τη μέθοδο Bayes Net ακολουθούμε την διαδικασία που περιγράφουμε στη συνέχεια.

Επιλέγουμε τη διαδρομή:

Modeling ⟶ *Classification* ⟶ *Bayes Net*

Δημιουργώντας με αυτό τον τρόπο έναν **Modeling Node**  μπορούμε να επιλέξουμε τις παραμέτρους του μοντέλου μας και να εκτελούμε το μοντέλο.

Τον **Bayes Net Node** τον συνδέουμε στον Partition node έτσι ώστε να ληφθεί υπόψην στη μοντελοποίηση ο διαχωρισμός σε σύνολο εκπαίδευσης και σύνολο εξέτασης.

Modeling Node:

Model name: (costum) Bayes Net

Use partitioned data: το σημειώνουμε διότι αυτή η επιλογή μας εξασφαλίζει ότι χρησιμοποιούνται τα δεδομένα μόνο από το το training set για την κατασκευή του μοντέλου.

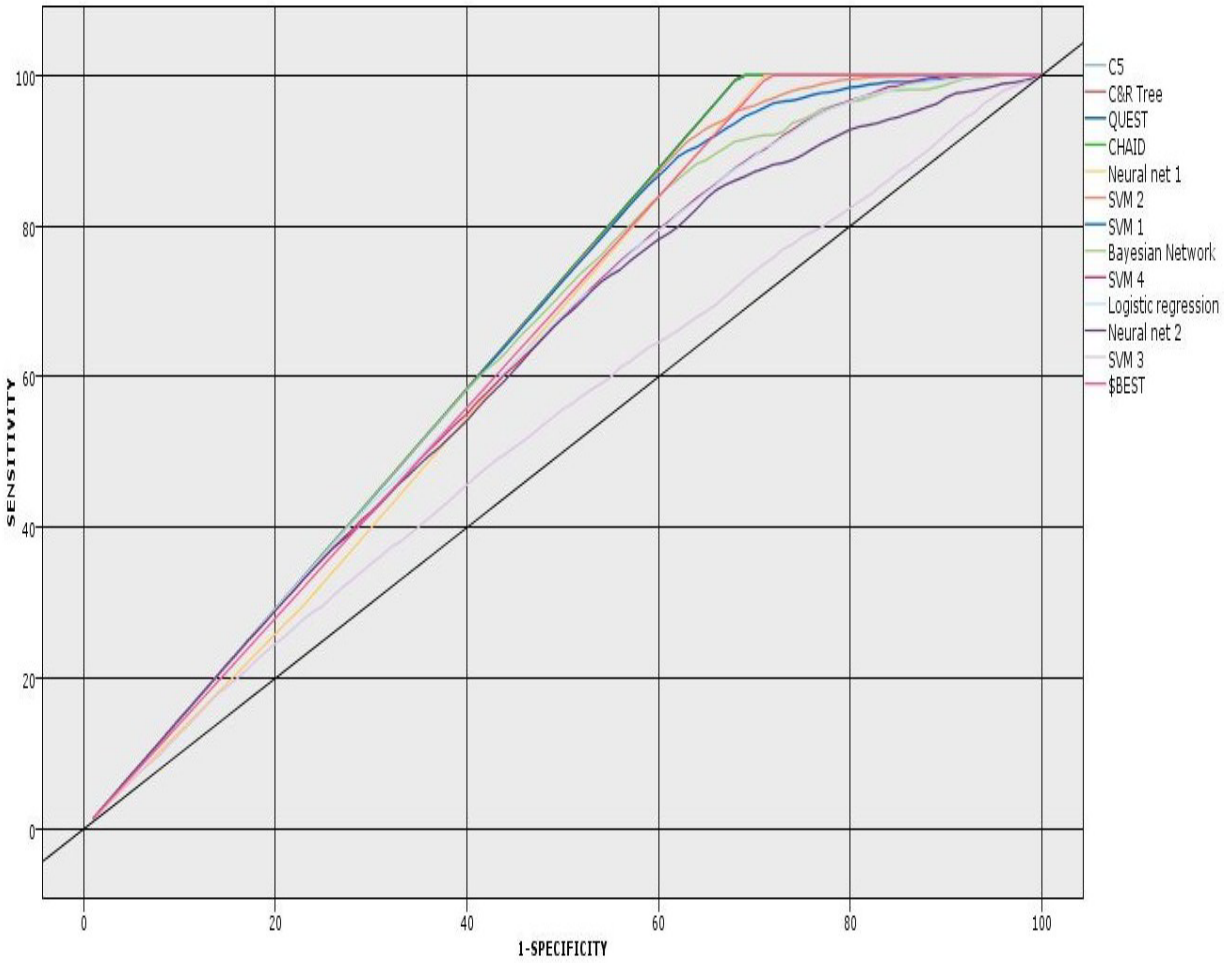
Εκτελούμε το μοντέλο σε κάθε μία από τις περιπτώσεις και παράγουμε ένα **model nugget** που είναι απαραίτητο στη συνέχεια για την αξιολόγηση του μοντέλου. Τα αποτελέσματα που προέκυψαν από την προσαρμογή του μοντέλου είναι τα ακόλουθα: (Το σχήμα 43 απεικονίζει το Bayesian δίκτυο για τις 9 στατιστικά σημαντικές μεταβλητές.)

6.8 Συνολική σύγκριση των ταξινομητών

Ο Πίνακας 12 κατατάσσει τα καλύτερα υποψήφια μοντέλα σύμφωνα με τα κριτήρια επίδοσής τους. Με τον τρόπο αυτό βοηθά τον πειραματιστή να επιλέξει την καλύτερη προσέγγιση για μια συγκεκριμένη ανάλυση. Το σχήμα 45 απεικονίζει τις ROC καμπύλες που προέρχονται από όλες τις μεθόδους που χρησιμοποιούνται σε αυτή την συγκριτική μελέτη. Όσο περισσότερο η καμπύλη βρίσκεται πάνω από τη γραμμή αναφοράς, τόσο πιο ακριβής είναι η δοκιμή.

	Ακρίβεια		Περιοχή κάτω από την καμπύλη
Ταξινομητής	Σύνολο εκπαίδευσης	Σύνολο ελέγχου	Σύνολο ελέγχου
C5.0	100%	100%	1
CHAID	99.24%	99.35%	1
C&RT	99.24%	99.25%	0.98
QUEST	99.07%	98.85%	0.95
MLP (Neural Net)	97.46%	97.37%	0.94
Polynomial (SVM 2)	92.86%	93.52%	0.98
Gaussian RBF(SVM1)	91.24%	91.82 %	0.97
Bayesian Network	87.65%	88.73 %	0.93
Linear (SVM4)	84.35%	84.73%	0.90
Λογιστική παλινδρόμηση	84.13%	84.23%	0.90
RBFN(Neural Net)	83.4%	84.11%	0.86
Sigmoid (SVM3)	71.71%	68.46%	0.62
l1-norm SVM	83.96%	84%	0.59

Πίνακας 16: Αναλυτική σύγκριση των ταξινομητών μέσω της συνολικής ακρίβειας και της AUROC (με φθίνουσα σειρά)



Σχήμα 44: ROC καμπύλες που προέκυψαν απ'ολους τους ταξινομητές

6.9 Γενική συζήτηση

Ο πρόσφατος πολλαπλασιασμός των μεγάλων διαστάσεων βάσεων δεδομένων καθιστά την επιλογή των επεξηγηματικών μεταβλητών ζωτικής σημασίας για την δημιουργία μοντέλου για μεγάλες βάσεις δεδομένων με πολύπλοκη δομή. Σ' αυτή την εργασία παρουσιάζουμε μια εκτενή συγκριτική ανάλυση των διαφόρων ταξινομητών μηχανικής μάθησης σε πραγματικά σεισμολογικά δεδομένα. Ο αλγόριθμος C5.0 είχε τα καλύτερα αποτελέσματα όσον αφορά την ακρίβεια, την ευαισθησία και την ειδικότητα. Τόσο ο αλγόριθμος CHAID όσο και ο C&RT, παρουσιάζουν επαρκή αποτελέσματα σε αυτά τα στατιστικά μέτρα που αφορούν στην απόδοση των ταξινομητών, και ξεπερνουν στην επίδοση τον αλγόριθμο QUEST.

Το MLP νευρωνικό δίκτυο ξεπέρασε σε μεγάλο βαθμό το Μπεϋζιανό (Bayesian) δίκτυο, και στη συνέχεια ακολούθησε το δίκτυο RBFN. Το SVM με πολυωνυμικό πυρήνα είχε σαφώς καλύτερη απόδοση ταξινόμησης σε σύγκριση με το l_1-νόρμα SVM και τα SVMs με RBF, γραμμικό και σιγμοειδή πυρήνα. Σε γενικές γραμμές, τα SVMs (με πολυώνυμο, Gaussian RBF και γραμμικό πυρήνα) έχουν αποδειξει την άριστη απόδοση ταξινόμησης, αφού βρέθηκαν να είναι περισσότερο αποδοτικά από την μέθοδο λογιστικής παλινδρόμησης (LR) και το δίκτυο RBFN. Το δίκτυο RBFN και το σιγμοειδές SVM παρατηρήθηκαν να έχουν την χειρότερη επίδοση ταξινόμησης. Το l_1-νόρμα SVM εκτελούνται σχεδόν όμοια με το γραμμικό SVM από την άποψη της ακρίβειας (ACC), της ευαισθησία και της ειδικότητας. Το l_1-νόρμα SVM δεν έχει την τάση να δηλώνει σε παρόμοιο ποσοστό τις ανενεργές μεταβλητές να είναι ενεργές ή τις ενεργές μεταβλητές να είναι ανενεργές, ως εκ τούτου, δεν θα μπορούσε να θεωρηθεί ως συντηρητική με αυτή την έννοια.

Η αξιολόγηση της αξιοπιστίας των αλγορίθμων ταξινόμησης είναι απαραίτητη για τη διασφάλιση της ποιότητας των δεδομένων. Στην συγκεκριμένη εργασία χρησιμοποιήσαμε τα μέτρα της ευαισθησίας και της ειδικότητας για τη σύγκριση των αλγορίθμων, ώστε να παρέχουμε χρήσιμα αποτελέσματα, αφού είναι προφανές ότι οι σεισμολόγοι προσπαθούν να κάνουν μία καλή πρόβλεψη των σεισμών, κάτι που τους υποχρεώνει να είναι πιο προσεκτικοί στην έρευνά τους. Στη σεισμική μελέτη μας, το εμπόδιο αυτό είναι ο διπλός κίνδυνος εσφαλμένης πρόβλεψης των σεισμών. Από τη μια πλευρά, αν οι σεισμολόγοι καταλήξουν στο ότι ένας σεισμός μπορεί να συμβεί σε συγκεγκριμένο χώρο και χρόνο, η κυβέρνηση θα πρέπει να ξεκινήσει μια μεγάλη εκστρατεία ετοιμότητας σεισμού που κοστίζει αλλά και φοβίζει τους πολίτες. Έτσι, εάν αυτή η πρόβλεψη δεν γίνει πραγματικότητα, αυτό θα έχει μεγάλο αντίκτυπο στην οικονομία και την κοινωνική ζωή. Από την άλλη πλευρά, εάν ένας ισχυρός σεισμός συμβεί χωρίς προηγούμενη πληροφορία, αυτό θα είναι επικίνδυνο και να οδηγήσει σε πολλούς θανάτους.

Η αξία αυτής της συγκριτικής μελέτης στέκεται όχι μόνο στην ανακάλυψη της γνώσης, αλλά επίσης και στην ικανότητα να υπολογίσει τα ποσοστά σφάλματος Τύπου Ι και Τύπου ΙΙ για κάθε χρησιμοποιούμενη μέθοδο. Σε γενικές γραμμές, παρατηρήσαμε ότι οι ταξινομητές μηχανικής μάθησης αναγνωρίζουν αποτελεσμετικά τους περισσότερους στατιστικά σημαντικούς παράγοντες για την κατασκευή του μοντέλου, δίνοντας επιπλέον σημαντικότητα στο έτος (x_1) της σεισμικής δραστηριότητας. Χάρη σε αυτό το αποτέλεσμα, οι σεισμολόγοι έχουν επιπλέον πληροφορίες σχετικά με την περιοδικότητα των σεισμών, η οποία είναι ένας από τους βασικούς παράγοντες για την πρόγνωση των σεισμών.

Ούτε τα τεχνητά νευρωνικά δίκτυα (ANNs) ούτε οι μηχανές διανυσματικής υποστήριξης (SVMs) είναι τέλειες τεχνικές. Οι μηχανές διανυσματικής υποστήριξης (SVMs) είναι γρήγορες στην εκπαίδευση, αλλά απαιτούν μια κατάλληλη επιλογή της συνάρτησης του πυρήνα. Τα τεχνητά νευρωνικά δίκτυα είναι πιο αργά στην εκπαίδευση, αλλά είναι γρήγορα

στην ταξινόμηση και εύρωστα στο θόρυβο. Τα τεχνητά νευρωνικά δίκτυα έχουν ευρέως αναπτυχθεί για την αντιμετώπιση μη-γραμμικών σεισμικών δεδομένων. Ελπίζουμε ότι αυτή η εργασία θα πείσει τους πειραματιστές να χρησιμοποιούν όχι μόνο τα τα τεχνητά νευρωνικά δίκτυα αλλά και τις μηχανές διανυσματικής υποστήριξης για την εξαγωγή μοτίβων για δεδομένα υψηλής διάστασης σε παράγοντες κινδύνου ενός σεισμού. Οι μηχανές διανυσματικής υποστήριξης θα πρέπει να θεωρηθούν ένα ισχυρό εργαλείο πρόβλεψης που έρχεται να προστεθεί στην ήδη υπάρχουσα μεθοδολογία της λογιστικής παλινδρόμησης. Έτσι, ένα από τα πλέον υποσχόμενα θέματα για περαιτέρω μελέτη είναι η χρήση των μηχανών διανυσματικής υποστήριξης ως μια εναλλακτική μέθοδο για την υποστήριξη ανακάλυψης της γνώσης για σεισμικά δεδομένα. Ωστόσο, υπάρχουν μερικά ενδιαφέροντα θέματα που είναι ανοιχτά και θα πρέπει να διερευνηθούν στο μέλλον.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] Ashida, Y., (1996). Data processing of reflection seismic data by use of neural network, Journal of Applied Geophysics 35, 89-98.

[2] Bao F., He X. and Zhao, F., (2010). Applying Data Mining to the Geosciences Data, 2010 International Conference on Computer, Mechatronics, Control and Electronic Engineering (CMCE), 5, 290-293.

[3] Benbrahim, M., Daoudi A., Benjelloun D. and Ibenbrahim, A., (2005). Discrimination of Seismic Signals Using Artificial Neural Networks, Engineering and Technology, 4, 4-7.

[4] Benjamini, Y. and Hochberg, Y., (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological) 289-300.

[5] Biggs, D., B. de Ville, and E. Suen. (1991). A method of choosing multiway partitions for classification and decision trees. *Journal of Applied Statistics*, 18, 49–62.

[6] Bishop, C., (1995). Neural Networks for Pattern Recognition, Oxford, Oxford University Press.

[7] Bradley, P. and Mangasarian, O.L., (1998). Feature selection via concave minimization and support vector machines. In Proceedings of the Fifteenth International Conference (ICML), pp. 82-90.

[8] Breiman, L., J.H.Friedman, R.A.Olshen, and C.J. Stone. (1984). Classification and regression trees. Monterey, CA: Wadsworth.

[9] BS ISO 5725-1: "Accuracy (trueness and precision) of measurement methods and reults - Part 1: General principles and definitions", pp.1 (1994).

[10] Burges, C.J.C., (1998). A tutorial on Support Vector Machines for Pattern Recognition,Data Mining and Knowledge Discovery, 2, 121-167.

[11] Candes, E. J., Tao, T. (2005) Decoding by linear programming. *IEEE Trans. Inform. Theory* **51**, 4203–4215. (doi:10.1109/TIT.2005.858979)

[12] Deighton, M. and Petrou, M., (2009). Data mining for large scale 3D seismic data analysis, Machine Vision and Applications, 20, 11-22.

[13] Dettling, M., (2004). BagBoosting for tumor classification with gene expression data. Bioinformatics, 20 3583-3593.

[14] Chen M. - S. and Han J., Yu P.,(1996). Data Mining: an overview from database perspective, IEEE Transactions on Knowledge and Data Engineering.

[15] Díez, F.J., Mira, J., Iturralde E., Zubillaga S. (1997). DIAVAL, a Bayesian expert system for echocardiography. Artificial Intelligence in Medicine (Elsevier) 10 (1): 59–73.

[16] Diersen, S., Lee, E., Spears, D., Chen, P. and Wang, L., (2011). Classification of Seismic Windows Using Artificial Neural Networks, Procedia Computer Science 00 (2011), 1-10.

[17] Donoho, D. L. (2006). For most large underdetermined systems of linear equations the minimal l_1 norm is also the sparsest solution. *Comm. Pure Appl. Math.* **59**, 797–829. (doi:10.1002/cpa.20132)

[18] Dreiseitl, S. and Ohno-Machado, L., (2002). Logistic regression and artificial neural network classification models: a methodology review, Journal of Biomedical Informatics, 35, 352-359.

[19] Farcomeni, A., (2008). A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. Statistical Methods in Medical Research, 17 347-388.

[20] Fawcett, T., (2003). ROC Graphs: Notes and Practical Considerations for Data Mining Researchers, Intelligent Enterprise Technologies Laboratory HP Laboratories Palo Alto HPL-2003-4 January 7th.

[21] Fisher, W. (1958). On grouping for maximum homogeniety, Journal of the American Statistical Association 53(284): 789–798.

[22] Friedman, N.; Geiger, D.; Goldszmidt, M. (1997). Bayesian network classifiers, Machine Learning 29 (2/3): 131.

[23] Friedman, N., Linial, M., Nachman, I., D. Pe'er (2000). Using Bayesian Networks to Analyze Expression Data. Journal of Computational Biology (Larchmont, New York: Mary Ann Liebert, Inc.) 7 (3/4): 601–620.

[24] Fung, G. and Mangasarian, O.L., (2004). A feature selection newton method for support vector machine classification, Comput. Optim. Appl. J., 28, 185-202.

[25] F.C. Garcia-Lopez, M. Garcia-Torres, B. Melian, J.A. Moreno-Perez, J.M. Moreno-Vega. Solving feature subset selection problem by a Parallel Scatter Search, *European Journal of Operational Research*, vol. 169, no. 2, pp. 477-489, 2006.

[26] Goldstein, D. (2009). Common genetic variation and human traits. New England J. Med., 360 1696-1698.

[27] Guyon, I., Elisseeff, A., (2003). An Introduction to Variable and Feature Selection, Journal of Machine Learning Research 3, 1157-1182

[28] Hall, M.,A., (1999), Correlation-based Feature Selection for Machine Learning, Phd thesis, Department of Computer Science , the University of Waikato, Hamilton, New Zealand.

[29] Hall, P., Jin, J. and Miller, H. (2010). Feature selection when there are many influential features. Manuscript.

[30] Hastie, T., Tibshirani, R., Friedman, J., (2001). The elements of statistical learning, Springer Series in Statistics, Springer-Verlag, New York. Data mining, Inference and Prediction.

[31] Hirschhorn, J. (2009). Genomewide association studies-illuminating biologic pathways. New England J. Med., 360 1699-1701

[32] Hoerl, A. E. & Kennard, R. W. 1970 Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* **12**, 55–82. (doi:10.2307/1267351)

[33] Hosmer, D. and Lemeshow, S. (2000). Applied logistic regression. Wiley-Interscience, New York.

[34] Iain M. Johnstone and D. Michael Titterington, (2012). Statistical challenges of high-dimensional data, Mathematical, Physical & Engineering Sciences, 2012.

[35] JCGM 200:2008 International vocabulary of metrology — Basic and general concepts and associated terms (VIM)

[36] Jiang X, Cooper GF. (2010). A Bayesian spatio-temporal method for disease outbreak detection. J Am Med Inform Assoc 17 (4): 462–71.

[37] Jiang, X.; Neapolitan, R.E.; Barmada, M.M.; Visweswaran, S. (2011). Learning Genetic Epistasis using Bayesian Network Scoring Criteria. BMC Bioinformatics 12, 89.

[38] Johnstone, I., M., Titterington, D.,M. (2009). Statistical challenges of high-dimensional data. *Phil. Trans. R. Soc. A,* **367**, 4237-4253.

[39] Karatzoglou, A., Meyer, D. and Hornik, K., (2006). Support Vector Machines in R, Journal of Statistical Software, 15(9), 1-28.

[40] Karayiannis N.B. and Weigun, G.M., (1997). Growing radial basis neural networks: merging supervised and unsupervised learning with network growth techniques, IEEE Trans Neural Netw, vol. 8, 1492-1505.

[41] Kass, G., (1980). An exploratory technique for investigating large quantities of categorical data, Applied Statistics, 29, 119-127.

[42] Kerh T., Huang, C. and Gunaratnam D., (2011). Neural Network Approach for Analyzing Seismic Data to Identify Potentially Hazardous Bridges, Mathematical Problems in Engineering, Vol. 2011, Article ID 464353, 15 pages, DOI:10.1155/2011/464353.

[43] Koukouvinos, C., Mylona, K. and Parpoula, C., (2013). A combination of a model of variable selection and data mining techniques for high-dimensional statistical modelling, International Journal of Information and Decision Sciences, Vol. 5, No. 2, pp. 154-168.

[44] Koumenides, C. L., Shadbolt, N. R. (2012). Combining link and content-based information in a Bayesian inference model for entity search. In Proceedings of the 1st Joint International Workshop on Entity-Oriented and Semantic Search (JIWES '12). ACM, New York, NY, USA, , Article 3 , 6 pages.

[45] Kraft, P. and Hunter, D. (2009). Genetic Risk Prediction-Are We There Yet New? England J. Med., 360 1701-1703.

[46] Lebrun, G., Lezoray, O., Charrier, C. and Cardot, H., (2006). A New Model Selection Method for SVM, E. Corchado et al. (Eds.): IDEAL 2006, LNCS 4224, pp. 99–107, Springer-Verlag Berlin Heidelberg 2006.

[47] Leon, F. and Atanasiu, G.M., (2006). Data Mining Methods for GIS Analysis of Seismic Vulnerability, Proceedings of the First International Conference on Software and Data Technologies (ICSOFT 2006), 2, 153-156, INSTICC Press, Portugal, ISBN 972-886569-4.

[48] Lewis, S.L., Montgomery, D.C., and Myers, R.H. (2001b). "Confidence Interval Coverage for Designated Experiments Analyzed with GLMs," Journal of Quality Technology, 33, pp. 279-292.

[49] Loh, W.Y., and Shih, Y.S., (1997). Split selection methods for classification trees, Statistica Sinica, 7, 815-840.

[50] Luis M. de Campos, Juan M. Fernández-Luna and Juan F. Huete (2004). "Bayesian networks and information retrieval: an introduction to the special issue". Information Processing & Management (Elsevier) 40 (5): 727–733.

[51] Mason, S.J., Graham, N.E., (2002). "Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation". *Quarterly Journal of the Royal Meteorological Society* (128): 2145–2166.

[52] Marketos, G., Theodoridis, Y. and Kalogeras, I.S., (2008) Seismological Data Warehousing and Mining: A survey, International Journal of Data Warehousing & Mining, 4 (1), 1-16.

[53] Miller, H. R., (2010). Statistical Methods for the Analysis of High-Dimensional Data, Phd Thesis Department of Mathematics and Statistics the University of Melbourne.

[54] Miller, H., Clarke, S., Lane, S., Lonie, A., Lazaridis, D., Petrovski, S. and Jones, O. (2009). Predicting customer behaviour: The university of melbourne's kdd cup report. The 2009 Knowledge Discovery in Data Competition (KDD Cup 2009), Challenges in Machine Learning, Volume 3, JMLR Workshop and Conference Proceedings pp.45-55.

[55] Mohsin, S. and Azam, F., (2011). Computational seismic algorithmic comparison for earthquake prediction, International Journal of Geology, 5(3), 53-59.

[56] Myers, R.H., Montgomery, D.C. and Vining, G.G., (2002). Generalized Linear Models: With Applications Engineering and the Sciences, John Wiley and Sons, New York.

[57] Nguyen, H., Franke, K., Petrovic, S. (2010). "Towards a Generic Feature-Selection Measure for Intrusion Detection", In Proc. International Conference on Pattern Recognition (ICPR), Istanbul, Turkey.

[58] Nguyen, H. T., Franke, K., Petrovic, S. (2011). On General Definition of L1-norm Support-Vector Machine for Feature Selection, In Proceedings of the International Journal of Machine Learning and Computing, ISSN: 2010-3700.

[59] Ou, Y.Y., Chen, C.Y., Hwang, S.C., Oyang, Y.J., (2003). Expediting model selection for support vector machines based on data reduction. In: Proc. IEEE International Conference on Systems, Man and Cybernetics (SMC2003), pp. 786–791.

[60] Pepe, M.S., (2000). Receiver operating characteristic methodology, J. Am. Statist. Assoc., 95, 308-11.

[61] Preethi, G. and Santhi, B., (2011). Study on Techniques of Earthquake Prediction, International Journal of Computer Applications, 29 (4), 55-58. Quinlan, J.(1993) C4.5: Programs for Machine Learning, Morgan Kaufmann Publisher, San Mateo, California.

[62] Ripley, B. D. (1996). Pattern Recognition and Neural Networks, Cambridge University Press.

[63] Robbins, H. and Munro, S. (1951). A stochastic approximation method, Annals of Mathematical Statistics 22: 400–407.

[64] Rosset, S. and Zhu, J. (2007). Piecewise linear regularized solution paths, Annals of Statistics 35(3): 1012–1030.

[65] Rosset, S., Zhu, J. and Hastie, T. (2004b). Margin maximizing loss functions, in S. Thrun, L. Saul and B. Sch¨olkopf (eds), Advances in Neural Information Processing Systems 16, MIT Press, Cambridge, MA.

[66] Scherbaum, F., Delavaud, E. and Riggelsen, C., (2009). Model Selection in Seismic

Hazard Analysis: An Information-Theoretic Perspective, Bulletin of the Seismological Society of America, 99(6), 3234-3247.

[67] Schmidt, M., Fung, G. and Rosales, R., (2007). Fast Optimization Methods for L1-Regularization: A Comparative Study and Two New Approaches, European Conference on Machine Learning (ECML 2007), 1-12.

[68] Smola, A.J., Scholkopf, B., (1998). A Tutorial on Support Vector Regression, NeuroCOLT2 Technical Report Series, NC2-TR-1998-030.

[69] Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. J. Roy. Statist. Soc. Ser. B, 36 111-147. With discussion by G. A. Barnard, A. C. Atkinson, L. K. Chan, A. P. Dawid, F. Downton, J. Dickey, A. G. Baker, O. Barndor_-Nielsen, D. R. Cox, S. Giesser, D. Hinkley, R. R. Hocking, and A. S. Young, and with a reply by the authors.

[70] Taylor, J. R. (1999). An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements. University Science Books. pp. 128–129. ISBN 0-935702-75-X.

[71] T. Siva Tian, (2009). Dimensionality Reduction for Classification with high-dimensional data, university of southern California.

[72] Tibshirani, R. 1996 Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B* **58**,267–288.

[73] Uebersax, J. (2004). Genetic Counseling and Cancer Risk Modeling: An Application of Bayes Nets. Marbella, Spain: Ravenpack International.

[74] Vapnik, V.N., (2000). The Nature of Statistical Learning Theory. 2nd end. Springer-Verlag, Berlin Heidelberg New York.

[75] Wan, S., Lei, T.C. and Chou, T.Y., (2010). A novel data mining technique of analysis and classification for landslide problems, Nat Hazards, 52, 211-230, DOI: 10.1007/s11069-009-9366-3.

[76] Widrow, B. and Hoff, M. (1960). Adaptive switching circuits, IRE WESCON Convention record, Vol. 4. pp 96-104; Reprinted in Andersen and Rosenfeld (1988).

[77] Witten, I. H., Frank, E., (2005). Data Mining, Practical machine learning tools and techniques with Java Implementations, 2nd edn, Morgan Kaufmann Publishers,

[78] San Francisco., Elsevier.

[79] Zhu, Y., Li, C., and Zhang, Y., (2004). A Practical Parameters Selection Method for SVM ISNN 2004, LNCS 3173, pp. 518–523.

[80] Zhu, J., Rosset, S., Hastie, T. and Tibshirani, R., (2003). 1-norm support vector machines, In Advances in Neural Information Processing Systems 16, Proceedings of the 2003 Conference.

[81] Zou, H., and Hastie, T., (2005). Regularization and variable selection via the elastic net, J. Roy. Statist. Soc. Ser. B, 67, 301-320.

[82] Ευθαλία Δ. Μάσσου, (2008). Αλγόριθμοι εξόρυξης πληροφορίας και στατιστική ανάλυση δεδομένων, Master Thesis, Εθνικό Μετσόβιο Πολυτεχνείο, Σχολή Εφαρμοσμένων Μαθηματικών Και Φυσικών Επιστημών, Τομέας Μαθηματικών.