ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ
ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

Διπλωματική Εργασία

# ΕΠΙΛΟΓΗ ΜΕΤΑΒΛΗΤΩΝ ΜΕ ΜΗΧΑΝΕΣ ΔΙΑΝΥΣΜΑΤΙΚΗΣΥΠΟΣΤΗΡΙΞΗΣ

# (VARIABLE SELECTION USING SUPPORT VECTOR MACHINES)

Γιολάντα Π. Εγγλέζου

Επιβλέπων: Χρήστος Κουκουβίνος
Καθηγητής ΕΜΠ

**Αθήνα, Ιούνιος 2014**

"The science of statistics is the chief instrumentality through which the progress of civilization is now measured, and by which its development hereafter will be largely controlled."

S. N. D. North

# Περίληψη

Η στατιστική ανάλυση και η αναγνώριση των σημαντικών μεταβλητών σε δεδομένα υψηλών διαστάσεων είναι ένα σημαντικό πρόβλημα στις μέρες μας. Η αποτυχία συνηθισμένων μεθόδων σε μεγάλα δεδομένων προβάλουν την ανάγκη να μελετηθούν καινούργιες μέθοδοι. Η εργασία αυτή ασχολείται με τεχνικές επιλογής χαρακτηριστικών, με σκοπό να βρεθεί ένα υποσύνολο χαρακτηριστικών που είναι το πιο σημαντικό για ταξινόμηση. Εκτός από τις συνηθισμένες τεχνικές επιλογής χαρακτηριστικών έχουν προταθεί πολλοί αλγόριθμοι που ασχολούνται με αυτό το πρόβλημα. Οι μηχανές διανυσματικής υποστήριξης είναι μια καινούργια μέθοδος για την άντληση πληροφοριών από ένα σύνολο δεδομένων και έχουν προταθεί πρόσφατα από πολλούς ερευνητές για σκοπούς επιλογής χαρακτηριστικών. Σε αυτή την εργασία παρουσιάζονται κάποιες μέθοδοι επιλογής χαρακτηριστικών καθώς και οι μηχανές διανυσματικής υποστήριξης, αλλά και μια πρακτική εφαρμογή αυτών των τεχνικών σε πραγματικά δεδομένα.

Το πρώτο κεφάλαιο ασχολείται γενικά με την εξόρυξη γνώσης από δεδομένα υψηλής διάστασης. Γίνεται μια εισαγωγή για τα δεδομένα υψηλής διάστασης αλλά και τα προβλήματα που παρουσιάζονται και στη συνέχεια παρουσιάζεται η ιδέα της εξόρυξης δεδομένων (data mining).

Στο δεύτερο κεφάλαιο παρουσιάζονται τεχνικές εξόρυξης γνώσης και μέθοδοι ταξινόμησης. Παρουσιάζουμε πιο αναλυτικά το πρόβλημα ταξινόμησης αλλά και τεχνικές όπως είναι τα δέντρα αποφάσεων, η λογιστική παλινδρόμηση και οι μηχανές διανυσματικής υποστήριξης.

Το τρίτο κεφάλαιο ασχολείται με την επιλογή χαρακτηριστικών. Παρουσιάζουμε τη μέθοδο, τα προβλήματα που έχουμε να αντιμετωπίσουμε και πώς μπορούμε να τα αντιμετωπίσουμε. Ακόμα παρουσιάζουμε κάποιες μεθόδους και αλγόριθμους επιλογής χαρακτηριστικών. Τέλος, παρουσιάζουμε τους αλγόριθμους SVM-RFE και Lasso που χρησιμοποιούνται για την επιλογή χαρακτηριστικών και που θα χρησιμοποιήσουμε στο πέμπτο κεφάλαιο για την εφαρμογή σε πραγματικά δεδομένα.

Το τέταρτο κεφάλαιο 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 ακρίβειας, ευαισθησίας και ειδικότητας που είναι σημαντικοί για να αποφασίσουμε για την απόδοση του μοντέλου.

Στο πέμπτο και τελευταίο κεφάλαιο γίνεται εφαρμογή στην R σε δεδομένα που πήραμε από το UCI Machine Learning Repository. Το set δεδομένων που χρησιμοποιούμε είναι το Breast Cancer Wisconsin (Diagnostic) Data Set. Εφαρμόζουμε ταξινόμηση με χρήση των μηχανών διανυσματικής υποστήριξης και κάνουμε επιλογή χαρακτηριστικών με τον αλγόριθμο SVM-RFE. Επίσης, γίνεται ταξινόμηση με λογιστική παλινδρόμηση και κάνουμε επιλογή χαρακτηριστικών με τον αλγόριθμο Lasso. Συγκρίνουμε αυτές τις δύο μεθόδους αναφέρουμε μελλοντική δουλεία που θα μπορούσε να γίνει σε αυτό το πεδίο.

# Abstract

Nowadays, the statistical analysis and the identification of important variables in high dimensional data is an important problem. The failure of conventional methods to large data sets, highlight the need to study new methods. This paper deals with feature selection techniques in order to find a subset of features that are most important for classification. Apart from the conventional feature selection techniques there have been proposed many algorithms dealing with this problem. Support vector machines is a new method for extracting information from a data set and have recently been proposed by many researchers for the purpose of feature selection. In this paper we present some methods of feature selection and support vector machines, and a practical application of these techniques to real data.

The first chapter deals generally with the mining of high-dimension data. An introduction to the high-dimensional data and problems encountered and then we present the concept of data mining.

In the second chapter we present mining techniques and classification methods. We present in detail the problem of classification and techniques such as decision trees, logistic regression and support vector machines.

The third chapter deals with feature selection. We present the method, the problems we face and how we can deal with them. Additional we present some methods and feature selection algorithms. Finally, we present two algorithms, SVM-RFE and Lasso used for feature selection and in fifth chapter are used for an application in real data.

The fourth chapter discusses the evaluation of a model using methods such as cross validation and performance of the classifiers mentioned above. Moreover, we discuss the terms of accuracy, sensitivity and specificity that are important to decide on the performance of the model.

In the fifth and final chapter we present an application in R with data we got from the UCI Machine Learning Repository. The data set we use is the Breast Cancer Wisconsin (Diagnostic) Data Set. We implement classification using support vector machines and apply feature selection with SVM-RFE algorithm. Also, we use logistic regression and apply feature selection with Lasso algorithm. We compare these two methods and finally we mention future work that could be done in this field.

# Ευχαριστίες

Πρωτίστως θα ήθελα να ευχαριστήσω εκ βάθους καρδίας τον Καθηγητή του Ε.Μ.Π. κ. Χρήστο Κουκουβίνο, όχι μόνο για τη δυνατότητα που μου προσέφερε να ασχοληθώ με αυτό το πολύ ενδιαφέρον θέμα το οποίο ανήκει στα ερευνητικά μου ενδιαφέροντα, αλλά και για τη συνεχή του προσφορά καθ' όλη τη διάρκεια των σπουδών μου που με οδήγησε τελικά στο να αγαπήσω τη Στατιστική και να επιλέξω τον τομέα αυτό για τη συνέχεια των σπουδών μου.

Ιδιαίτερες ευχαριστίες θα ήθελα να εκφράσω στην υποψήφια διδάκτορα Κρυσταλλένια Δρόσου, για την πολύτιμη βοήθεια της και το συνεχές ενδιαφέρον κατά τη διάρκεια εκπόνησης της διπλωματικής μου εργασίας.

Ακόμα θα ήθελα να ευχαριστήσω τους γονείς μου Πανίκο και Χαραλαμπία Εγγλέζου και την αδερφή μου Χριστίνα, για την αμέριστη υπομονή τους και άμεση υποστήριξή τους καθ' όλη τη διάρκεια των σπουδών μου. Επίσης αισθάνομαι την ανάγκη να ευχαριστήσω τους φίλους μου, αυτούς που βρίσκονταν κοντά μου στην Αθήνα για την καθημερινή υποστήριξη τους τα πέντε αυτά χρόνια των σπουδών μου, αλλά και αυτούς που από μακριά μου προσέφεραν τη βοήθεια και συμπαράσταση τους κατά τη διάρκεια αυτής της εργασίας. Τέλος, ένα μεγάλο ευχαριστώ στους συμφοιτητές μου για τα πέντε υπέροχα χρόνια που πέρασα μαζί τους ως φοιτήτρια του Ε.Μ.Π.

Χωρίς τη συμβολή των πιο πάνω ανθρώπων θα ήταν αδύνατη η ολοκλήρωση των σπουδών και η εκπόνηση της παρούσας διπλωματικής εργασίας.

Γιολάντα Εγγλέζου

Αθήνα, 2014

# Περιεχόμενα

# ΚΕΦΑΛΑΙΟ 1

## Εξόρυξη γνώσης από δεδομένα υψηλής διάστασης

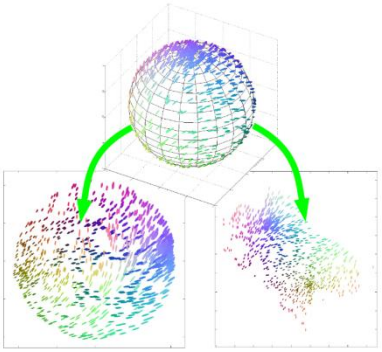
### 1.1 Δεδομένα Υψηλής Διάστασης

### (High Dimensional Data)

Στατιστική υψηλής διάστασης, είναι ένας τομέας που βασίζεται στη θεωρία των τυχαίων διανυσμάτων. Σε πολλές εφαρμογές, η διάσταση των διανυσμάτων δεδομένων μπορεί να έχει μεγαλύτερο μέγεθος από το μέγεθος του δείγματος. Ο όρος δεδομένα υψηλής διάστασης (high dimensional data) 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 σε ένα προγνωστικό μοντέλο (predictive model), όπου οι μεταβλητές χρησιμοποιούνται στην πρόβλεψη ενός συγκεκριμένου γεγονότος με βάση τα δεδομένα αυτά.

Το χαμηλό κόστος της ψηφιακής αποθήκευσης, επιτρέπει σε τεράστιες ποσότητες πληροφοριών να συλλέγονται σχετικά φτηνά και στη συνέχεια, τα χαρακτηριστικά αυτής της βάσης δεδομένων εξάγονται και αναλύονται στατιστικά. Βασικές πηγές των μεγάλων βάσεων δεδομένων είναι ο τομέας της βιομηχανίας (π.χ. πρόβλεψη της συμπεριφοράς πελατών, προσδιορισμός προϊόντος με τη μεγαλύτερη ζήτηση κλπ) και της βιολογίας (π.χ. πειράματα γενετικής με μικροδιανύσματα). 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 των κύριων επιδράσεων ή μεταβλητών.

Αξίζει να σημειωθεί 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.



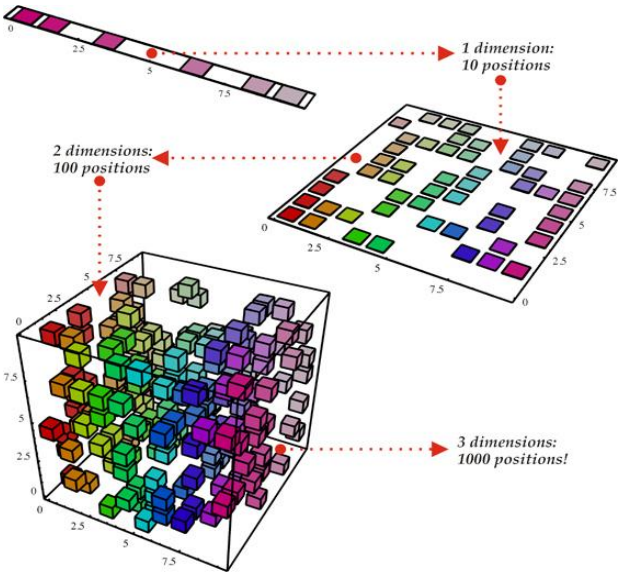
**Σχήμα 1: Οπτικοποίηση των αμοιβαίων ομοιοτήτων των οντοτήτων σε πολυδιάστατα σύνολα δεδομένων υψηλής διάστασης είναι ένα κεντρικό πρόβλημα στη διερευνητική ανάλυση των δεδομένων και την ανακάλυψη της γνώσης.**

### 1.1.1 Κατάρα της Διάστασης

Κατά την ανάλυση δεδομένων υψηλής διάστασης αντιμετωπίζουμε το πρόβλημα της κατάρας της διάστασης (curse of dimensionality). Ο όρος αυτός επινοήθηκε από τον Bellman, για να δείξει ότι ο αριθμός των δειγμάτων που απαιτούνται για να εκτιμηθεί η αυθαίρετη λειτουργία με ένα δεδομένο επίπεδο ακρίβειας αυξάνεται εκθετικά με τον αριθμό των μεταβλητών (δηλαδή, διαστάσεις) από το ότι περιλαμβάνει. Αυτό σημαίνει ότι ο αριθμός των αντικειμένων (δηλαδή, μονάδες) στο σύνολο δεδομένων που πρέπει να εξεταστούν κατά τον υπολογισμό της εκτίμησης αυξάνεται εκθετικά με την υποκείμενη διάσταση.

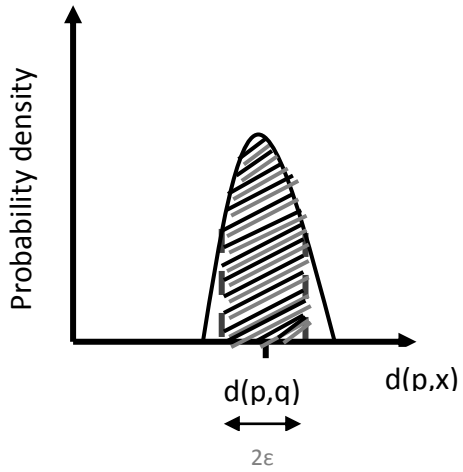
Η κατάρα της διάστασης αναφέρεται σε διάφορα φαινόμενα που προκύπτουν κατά την ανάλυση και την οργάνωση των δεδομένων σε μεγάλων διαστάσεων χώρους (συχνά με εκατοντάδες ή χιλιάδες διαστάσεις) που δεν ανταποκρίνονται σε ρυθμίσεις συνθηκών χαμηλών διαστάσεων. Υπάρχουν πολλαπλά φαινόμενα που αναφέρονται με αυτό το όνομα σε τομείς όπως η αριθμητική ανάλυση, η δειγματοληψία, Συνδυαστική, μηχανική μάθηση, εξόρυξη δεδομένων και βάσεις δεδομένων. Το κοινό θέμα των προβλημάτων αυτών είναι ότι, όταν αυξάνεται η διάσταση, ο όγκος του χώρου αυξάνεται τόσο γρήγορα έτσι ώστε τα διαθέσιμα δεδομένα σπανίζουν. Αυτές οι ελάχιστες αναφορές αποτελούν πρόβλημα για οποιαδήποτε μέθοδο που απαιτεί στατιστική σημαντικότητα. Για να αποκτηθεί

ένα στατιστικά ορθό και αξιόπιστο αποτέλεσμα, η ποσότητα των δεδομένων που απαιτούνται για να υποστηρίξουν το αποτέλεσμα συχνά αυξάνεται εκθετικά με τη διάσταση. Επίσης, η οργάνωση και η αναζήτηση των δεδομένων συχνά βασίζεται στον εντοπισμό των τομέων όπου τα αντικείμενα σχηματίζουν ομάδες με παρόμοιες ιδιότητες. Σε δεδομένα υψηλής διάστασης ωστόσο, όλα τα αντικείμενα φαίνεται να είναι αραιά και ανόμοια με πολλούς τρόπους με αποτέλεσμα η οργάνωση των δεδομένων να μην είναι αποτελεσματική.




**Σχήμα 2: Σε υψηλές διαστάσεις, τα γειτονικά σημεία γίνονται εκθετικά μεγάλα, και το καθένα απαιτεί έναν εκθετικό αριθμό παραδειγμάτων εκπαίδευσης για να το καλύψει.**

Υποθέτοντας ότι d είναι μια μετρική απόστασης και ότι η τριγωνική ανισότητα ισχύει, για να δούμε την κατάρα της διάστασης μπορούμε να παρατηρήσουμε ότι, όταν πρόκειται για δεδομένα υψηλής διάστασης, η συνάρτηση πυκνότητας πιθανότητας (ανάλογη με ένα ιστόγραμμα) των αποστάσεων των διαφόρων στοιχείων είναι πιο πυκνή και έχει μεγαλύτερη μέση τιμή. Αυτό σημαίνει ότι οι αλγόριθμοι αναζήτησης ομοιότητας θα πρέπει να εκτελέσουν περισσότερη δουλειά. Ένας τρόπος για να καταλάβει κανείς γιατί η πιο συγκεντρωμένη πυκνότητα πιθανότητας μπορεί να οδηγήσει σε πιο σύνθετη αναζήτηση ομοιότητας, είναι να παρατηρήσουμε ότι η τριγωνική ανισότητα δεν μπορεί να χρησιμοποιείται τόσο συχνά για την εξάλειψη των αντικειμένων υπό εξέταση. Ειδικότερα, η τριγωνική ανισότητα σημαίνει ότι κάθε στοιχείο x τέτοιο ώστε $| d(p,q) - d(p,x) | > \varepsilon$, δεν μπορεί να έχει απόσταση ε ή μικρότερη από το q.




**Σχήμα 3: Η πιθανότητα της εξάλειψης ενός στοιχείου από την εξέταση μέσω της χρήσης της τριγωνικής ανισότητας είναι η υπόλοιπη περιοχή κάτω από την καμπύλη.**

Έτσι, αν εξετάσουμε την συνάρτηση πυκνότητας πιθανότητας του $d\,(p,x)$ (δηλαδή, επί του οριζοντίου άξονα), βρίσκουμε ότι όταν ε είναι μικρή, ενώ η συνάρτηση πυκνότητας πιθανότητας είναι μεγάλη στα $d(p,q)$, τότε η πιθανότητα της εξάλειψης ενός στοιχείου από την εξέταση μέσω της χρήσης της τριγωνικής ανισότητας είναι η υπόλοιπη περιοχή κάτω από την καμπύλη, η οποία είναι αρκετά μικρή.

Ψάχνοντας σε υψηλής διάστασης χώρους είναι χρονοβόρο. Τα παραστατικά σημεία και η εξέταση εύρους σε υψηλές διαστάσεις είναι πολύ πιο εύκολα στην ανάλυση, από την εξέταση ομοιοτήτων γιατί δεν απαιτούν τον υπολογισμό της απόστασης.

Τα πλεονεκτήματα της διάστασης είναι λιγότερο ευρέως γνωστά, αλλά περιλαμβάνουν το φαινόμενο της συγκέντρωσης του μέτρου (concentration of measure phenomenon). Ο όρος αυτός εισήχθηκε αρχικά από τον V. Milman. Ορισμένες τυχαίες διακυμάνσεις είναι πολύ καλά ελεγχόμενες σε υψηλές διαστάσεις και παρατηρείται η επιτυχία των ασυμπτωτικών μεθόδων, που χρησιμοποιούνται ευρέως στην Στατιστική, στην ανάλυση δεδομένων υψηλής διάστασης.

Η ανάλυση δεδομένων υψηλής διάστασης παρουσιάζει μεγάλο ενδιαφέρον, τόσο για τις πολλές και διαφορετικές εφαρμογές που έχει, αλλά και την ανάγκη που προβάλει να επανεξεταστούν οι μέχρι τώρα στατιστικές μέθοδοι που χρησιμοποιούνται ώστε να εξάγονται πιο αληθή αποτελέσματα.

Σε αυτό το σημείο, θα συμβολίζουμε με p, το πλήθος των παραμέτρων πρόβλεψης και με n το πλήθος των παρατηρήσεων. Στα δεδομένα υψηλής διάστασης παρατηρούμε ότι $p >> n$.

### 1.1.2 Ψευδώς Θετικά Ποσά

Σε ένα σύνολο δεδομένων υψηλών διαστάσεων, με συνεχείς μεταβλητές, μπορούμε να παρατηρήσουμε ότι σε κάθε εκτέλεση στατιστικού τεστ, πολλές περιττές παρατηρήσεις θα εμφανίζονταν σημαντικές, εκτός από όσες είναι πραγματικά σημαντικές. Έτσι, εμφανίζεται το πρόβλημα των ψευδώς θετικών ποσοστών (false positives) εμποδίζουν τον εντοπισμό αληθών χαρακτηριστικών.

### 1.1.3 Προβλήματα Υπερποσαρμογής

Το πρόβλημα υπερποσαρμογής (overfitting) εμφανίζεται όταν ένα στατιστικό μοντέλο περιγράφει το τυχαίο σφάλμα ή τον θόρυβο αντί της επικείμενης σχέσης μεταξύ των παρατηρήσεων. Το πρόβλημα αυτό έχει ως αποτέλεσμα κακή προγνωστική απόδοση, αφού μπορεί να διογκωθούν μικρές διακυμάνσεις στα δεδομένα. Σε δεδομένα υψηλής διάστασης, όπου ο αριθμός των παρατηρήσεων είναι μικρός και ο αριθμός των μεταβλητών πρόβλεψης είναι μεγάλος, είναι πολύ εύκολο να κατασκευαστεί ένα μοντέλο που εμφανίζεται ισχυρό, αλλά έχει απογοητευτικές επιδόσεις.

### 1.1.4 Παλινδρόμηση Υψηλής Διάστασης

- ***Γραμμικό Μοντέλο***

Ένα απλό αλλά χρήσιμο μοντέλο για τα δεδομένα υψηλής διάστασης είναι το απλό γραμμικό μοντέλο. Το απλό γραμμικό μοντέλο περιλαμβάνει δύο μεταβλητές, την ανεξάρτητη ή αλλιώς επεξηγηματική μεταβλητή p-διάστασης $X_i \subset \boldsymbol{\mathcal{R}}^p$ και την εξαρτημένη ή διαφορετικά τη μεταβλητή απόκρισης $Y_i \subset \boldsymbol{\mathcal{R}}$, οι οποίες συνδέονται μεταξύ τους με τη γραμμική συνάρτηση παλινδρόμησης. Οι δύο μεταβλητές συνδέονται με την σχέση

$$Y_i = \mu + \sum_{j=1}^{p} \beta_j X_i^{(j)} + \varepsilon_i \, , i = 1, \dots, n$$

με $E[\varepsilon_i] = 0$.

Για απλότητα και χωρίς βλάβη της γενικότητας, υποθέτουμε συνήθως ότι ο σταθερός όρος είναι μηδέν και όλοι οι παράγοντες είναι στο κέντρο και μετρούνται στην ίδια κλίμακα, $Y_i = \sum_{j=1}^{p} \beta_j X_i^{(j)} + \varepsilon_i$. Έτσι ισχύουν οι σχέσεις:

$$\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n} = 0$$

και

$$\widehat{\sigma_j^2} = \frac{\sum_{i=1}^{n} (X_i^{(j)} - \overline{X_i^{(j)}})^2}{n} = 1 \,,$$

για όλα τα j.

Το μόνο ασυνήθιστο σημείο σε αυτό το γραμμικό μοντέλο είναι ότι $p \gg n$.

Επίσης, συχνά χρησιμοποιείται η αναπαράσταση $\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon$, με διάνυσμα απόκρισης $Y_{n\times 1}$, πίνακα σχεδίασης $X_{n\times p}$, διάνυσμα παραμέτρων $\beta_{n\times 1}$ και διάνυσμα σφάλματος $\varepsilon_{n\times 1}$.

Για την εκτίμηση, μπορούμε να χρησιμοποιήσουμε μια μέθοδο κανονικοποίησης, αλλιώς γνωστή ως ποινικοποιημένα ελάχιστα τετράγωνα ή ποινικοποιημένη μέγιστη πιθανοφάνεια. Ένας εύκολος τρόπος, είναι να προσπαθήσουμε να εκμεταλλευτούμε την αραιότητα, δηλαδή, να αναζητήσουμε μια λύση για το β στο οποίο πολλά από τα στοιχεία είναι μηδέν.

Η αραιότητα, μπορεί να προσδιοριστεί ποσοτικά με την $l_q$-νόρμα, όπου $1 \leq q \leq \infty$, ανάλογα με $0 < q < 1$ ή την $l_0$ κανονικοποίηση, όπου

$$\|\beta\|_0^0 = \left|\{j;\ \beta_j \neq 0\}\right|$$

που μετρά το πλήθος των μη μηδενικών καταχωρήσεων της παραμέτρου.

Ο συμβολισμός $\|\beta\|_0^0 = \sum_{j=1}^{p} |\beta_j|^0$ είναι ανάλογος με το $\|\beta\|_q^q = \sum_{j=1}^{p} |\beta_j|^q$ για $0 < q < \infty$. Σε αντίθεση με την $l_0$ κανονικοποίηση, η $l_1$-νόρμα

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

μετρά την αραιότητα με διαφορετικό τρόπο και έχει υπολογιστικό πλεονέκτημα αφού είναι κυρτή συνάρτηση στο β.

Κατά την $l_0$ κανονικοποίηση αλλάζουμε την συνάρτηση ποινής και προκύπτει $\widehat{\beta_0} = argmin_\beta(n^{-1} \| \boldsymbol{Y} - \boldsymbol{X}\beta \|^2 + \lambda\|\beta\|_0^0)$ , η $l_0$ του β. Αυτό όμως μπορεί να οδηγήσει σε μη αποδεκτή υπολογιστική πολυπλοκότητα και μια πιθανή εξάπλωση των τοπικών ακρότατων. Γι' αυτό, εναλλακτικά, μπορούμε να χρησιμοποιήσουμε την $l_1$-νόρμα για τη συνάρτηση ποινής από την οποία προκύπτει

$$\widehat{\widehat{\beta(\lambda)}} = argmin_\beta(n^{-1} \| \boldsymbol{Y} - \boldsymbol{X}\beta \|_2^2 + \lambda \| \beta \|_1).$$

Η μέθοδος αυτή ονομάζεται μέθοδος Lasso (Tibshirani, 1996). Θα τη δούμε πιο αναλυτικά σε παρακάτω κεφάλαιο.

## 1.2 Εξόρυξη Δεδομένων

## (Data Mining)

Η ραγδαία ανάπτυξη της πληροφορικής και της τεχνολογίας κατέστησε εύκολη τη συλλογή και αποθήκευση δεδομένων και συγχρόνως παρουσιάστηκε η εμφανής αδυναμία των κλασικών μεθόδων στατιστικής να αναλύσουν αυτά τα μεγάλα σετ δεδομένων. Λόγω αυτής της αδυναμίας άρχισε να αναπτύσσεται η εξόρυξη δεδομένων (data mining) γύρω στο 1990.

Το data mining έχει ως στόχο την εξαγωγή χρήσιμων μοτίβων και συμπερασμάτων από πολύ μεγάλα σετ δεδομένων. Αυτό το επιτυγχάνει συνδυάζοντας την επιστήμη της στατιστικής και της πληροφορικής, δηλαδή με τη χρήση μεθόδων που βασίζονται στην τεχνητή νοημοσύνη (artificial intelligence) και την μηχανική μάθηση (machine learning).

Γενικά υπάρχουν αντικρουόμενες απόψεις γύρω από το ποιος θα μπορούσε να είναι ένας σαφής και περιεκτικός ορισμός για την Εξόρυξη Δεδομένων. Ωστόσο, αποδεχόμαστε σαν ορισμό του data mining τον εξής :

*«Εξόρυξη Δεδομένων είναι η ανάλυση , συνήθως τεράστιων , παρατηρούμενων συνόλων δεδομένων , έτσι ώστε να βρεθούν μη παρατηρηθείσες σχέσεις και να συνοψιστούν τα δεδομένα με καινοφανείς τρόπους οι οποίοι να είναι κατανοητοί και χρήσιμοι στον κάτοχο των δεδομένων».*

Το data mining τυπικά ασχολείται με δεδομένα που έχουν συλλεχθεί για κάποιον άλλο σκοπό εκτός της data mining ανάλυσης. 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. Επίσης, οι κλασικές στατιστικές αναλύσεις χρησιμοποιούν μικρά σύνολα δεδομένων, σε αντίθεση με το data mining που χρησιμοποιεί μεγάλα σύνολα. Έτσι όμως προκύπτουν νέα προβλήματα, όπως η διαχείριση τους, η αποθήκευση τους, η πρόσβαση τους. 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.

Για τη διεξαγωγή αποτελεσματικής εξόρυξης δεδομένων, πρέπει πρώτα να εξεταστεί τι είδους χαρακτηριστικά αναμένεται να έχει ένα εφαρμοσμένο σύστημα ανακάλυψης γνώσης και τι είδους προκλήσεις μπορεί να αντιμετωπιστούν στην ανάπτυξη των τεχνικών εξόρυξης δεδομένων.

Χειρισμός διαφορετικών τύπων δεδομένων.
Επειδή υπάρχουν πολλά είδη δεδομένων και βάσεων δεδομένων που χρησιμοποιούνται στις διαφορετικές εφαρμογές, μπορεί κανείς να αναμένει ότι ένα σύστημα ανακάλυψης γνώσης θα πρέπει να είναι σε θέση να εκτελέσει αποτελεσματική εξόρυξη δεδομένων σε διαφορετικά είδη δεδομένων. Δεδομένου ότι οι περισσότερες διαθέσιμες βάσεις δεδομένων είναι σχεσιακές, είναι σημαντικό ένα σύστημα εξόρυξης δεδομένων να εκτελεί αποτελεσματική και αποδοτική ανακάλυψη γνώσης σε σχεσιακά δεδομένα. Επιπλέον, πολλές βάσεις δεδομένων περιέχουν σύνθετους τύπους δεδομένων, όπως δομημένα δεδομένα και πολύπλοκα αντικείμενα δεδομένων, δεδομένων πολυμέσων, χωρικά και χρονικά δεδομένα , δεδομένα συναλλαγών κλπ. Ένα ισχυρό σύστημα θα πρέπει να είναι σε θέση να εκτελέσει αποτελεσματική εξόρυξη δεδομένων για τέτοιου είδους πολύπλοκων τύπων δεδομένων. Ωστόσο , η ποικιλία των τύπων δεδομένων και διαφορετικών στόχων της εξόρυξης δεδομένων καθιστούν ρεαλιστικό να αναμένουμε ένα σύστημα εξόρυξης δεδομένων για να χειριστεί όλα τα είδη των δεδομένων .

Αποδοτικότητα και την επεκτασιμότητα αλγορίθμων εξόρυξης δεδομένων.
Για να είναι αποδοτική η συλλογή πληροφοριών από ένα τεράστιο ποσό δεδομένων σε βάσεις δεδομένων, οι αλγόριθμοι ανακάλυψης γνώσης πρέπει να είναι αποτελεσματικοί και επεκτάσιμοι σε μεγάλες βάσεις δεδομένων. Δηλαδή, ο χρόνος λειτουργίας ενός αλγορίθμου εξόρυξης δεδομένων πρέπει να είναι προβλέψιμος και αποδεκτός σε μεγάλες βάσεις δεδομένων.

Χρησιμότητα, βεβαιότητα και εκφραστικότητα των αποτελεσμάτων της εξόρυξης δεδομένων.
Η ανακάλυψη γνώσης θα πρέπει να απεικονίζει με ακρίβεια το περιεχόμενο της βάσης δεδομένων και είναι χρήσιμη για ορισμένες εφαρμογές. Η ατέλεια θα πρέπει να εκφράζεται με τα μέτρα της αβεβαιότητας, με τη μορφή των κατά προσέγγιση κανόνων ή ποσοτικών κανόνων. Θόρυβος και εξαιρούμενα δεδομένα θα πρέπει να αντιμετωπίζονται κομψά σε συστήματα εξόρυξης δεδομένων. Αυτό παρακινεί επίσης μια συστηματική μελέτη για τη μέτρηση της ποιότητας στην ανακάλυψη γνώσης, συμπεριλαμβανομένης της αξιοπιστίας, με τη δημιουργία στατιστικών, αναλυτικών μοντέλων και εργαλείων.

Έκφραση των διαφόρων ειδών των αποτελεσμάτων της εξόρυξης δεδομένων.
Διαφορετικά είδη της γνώσης μπορεί να ανακαλυφθούν από ένα μεγάλο όγκο δεδομένων. Επίσης, μπορεί κανείς να θέλει να εξετάσει την ανακάλυψη γνώσης από διαφορετική άποψη και να την παρουσιάσει σε διαφορετικές μορφές. Αυτό απαιτεί να εκφραστούν τόσο τα αιτήματα εξόρυξης δεδομένων και η ανακάλυψη γνώσης σε γλώσσες υψηλού επιπέδου, έτσι ώστε το έργο εξόρυξης δεδομένων να μπορεί να καθορίζεται από μη ειδικούς και η ανακάλυψη γνώσης να είναι κατανοητή και άμεσα χρησιμοποιήσιμη από τους χρήστες. Αυτό απαιτεί επίσης τα συστήματα ανακάλυψης γνώσης να υιοθετήσουν εκφραστικές τεχνικές αναπαράστασης γνώσης.

Διαδραστική εξόρυξη γνώσης σε πολλαπλά επίπεδα άντλησης.
Δεδομένου ότι είναι δύσκολο να προβλέψουμε τι ακριβώς θα μπορούσε να ανακαλυφθεί από μια βάση δεδομένων, ένα υψηλού επιπέδου ερώτημα εξόρυξης δεδομένων θα πρέπει να αντιμετωπίζεται ως ένας ανιχνευτής που μπορεί να αποκαλύψει μερικά ενδιαφέροντα ίχνη για την περαιτέρω εξερεύνηση. Η διαδραστική ανακάλυψη θα πρέπει να ενθαρρύνεται, γιατί επιτρέπει στον χρήστη να τη διαδραστική βελτίωση ενός αιτήματος εξόρυξης δεδομένων, τη σταδιακή εμβάθυνση της διαδικασίας εξόρυξης δεδομένων, και την ευελιξία να αντιμετωπιστούν τα δεδομένα και τα αποτελέσματα εξόρυξης δεδομένων σε πολλαπλά επίπεδα άντλησης και από διαφορετικές γωνίες.

Εξόρυξη πληροφοριών από διαφορετικές πηγές δεδομένων.
Το ευρέως διαθέσιμο τοπικό και ευρύ δίκτυο υπολογιστών, συμπεριλαμβανομένου του διαδικτύου, συνδέει πολλές πηγές δεδομένων από ετερογενείς βάσεις δεδομένων. Εξόρυξης γνώσης από διαφορετικές πηγές μορφοποιημένων ή μη μορφοποιημένων δεδομένων με ποικίλη σημασιολογία, δημιουργεί νέες προκλήσεις για την εξόρυξη δεδομένων. Από την άλλη πλευρά, η εξόρυξη δεδομένων μπορεί να βοηθήσει στην ανακάλυψη κανονικότητας υψηλού επιπέδου δεδομένων σε ετερογενείς βάσεις δεδομένων που δύσκολα μπορεί να ανακαλυφθεί από ένα απλό σύστημα. Επιπλέον, το τεράστιο μέγεθος της βάσης δεδομένων, η ευρεία κατανομή των δεδομένων και η υπολογιστική πολυπλοκότητα ορισμένων μεθόδων εξόρυξης δεδομένων παρακινήσει την ανάπτυξη παράλληλων και κατανεμημένων αλγορίθμων εξόρυξης δεδομένων.

Προστασία της ιδιοτικότητας και ασφάλειας των δεδομένων.
Όταν τα δεδομένα μπορούν να προβληθούν από πολλές διαφορετικές γωνίες και σε διαφορετικά επίπεδα άντλησης, απειλείται ο στόχος της προστασίας της ασφάλειας των δεδομένων και την προστασία από την εισβολή στην ιδιοτικότητα τους. Είναι σημαντικό να αναπτυχθούν μέτρα για την πρόληψη της αποκάλυψης ευαίσθητων πληροφοριών.

Κάποιοι από αυτούς τους στόχους μπορεί να προκαλέσουν αντικρουόμενα προβλήματα στην ανάπτυξη του data mining. Είναι όμως σημαντικό να παρουσιάσουμε τη γενική εικόνα των απαιτήσεων για αποτελεσματική εξόρυξη δεδομένων.

Ο κύριος στόχος του data mining, είναι η εξαγωγή νέων πληροφοριών από τα δεδομένα και η κατασκευή προγραμμάτων που χρησιμοποιούν στατιστικά αποτελέσματα. Η ανακάλυψη της γνώσης γίνεται με τεχνικές οι οποίες διακρίνονται σε δύο βασικές κατηγορίες:

- Μέθοδοι με επίβλεψη (Supervised Methods): Εδώ τα δεδομένα αποτελούνται από ένα διάνυσμα επεξηγηματικών μεταβλητών (x) και μία τιμή απόκρισης (y). Το σκεπτικό πίσω από τις μεθόδους με επίβλεψη είναι με την ανάλυση των δεδομένων να δημιουργήσει μία συνάρτηση ταξινόμησης (ταξινομητής) για να χαρτογραφήσει νέα δεδομένα που δεν συμπεριλαμβάνονται στο αρχικό σετ δεδομένων. Μερικά παραδείγματα αυτών των μοντέλων είναι τα νευρωνικά δίχτυα, δέντρα αποφάσεων, λογιστική παλινδρόμηση, μηχανές διανυσματικής υποστήριξης (SVM), boosting και bagging μέθοδοι.

- Μέθοδοι χωρίς επίβλεψη (Unsupervised Methods): Αντίθετα με τη μέθοδο εκμάθησης με επίβλεψη αυτή η μέθοδος προσπαθεί να βρει κρυμμένες δομές σε δεδομένα που δεν έχουν μεταβλητή απόκρισης. Άρα δεν έχουμε πρόβλεψη μελλοντικών αποτελεσμάτων αλλά εξερεύνηση των δομών και των σχέσεων των δεδομένων. Παραδείγματα αυτών των δομών είναι τα Kohonen Networks, blind signal separation, clustering και K-means.

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

Η εξόρυξη δεδομένων, η οποία είναι γνωστή επίσης και ως ανακάλυψη γνώσης (knowledge discovery) σε βάσεις δεδομένων, είναι μια μη τετριμμένη διαδικασία εξόρυξης, προηγουμένως άγνωστων και ενδεχομένως χρήσιμων πληροφοριών από δεδομένα σε βάσεις δεδομένων. Υπάρχουν επίσης πολλοί άλλοι όροι που αναφέρονται για να περιγράψουν αυτή την διαδικασία, που φέρουν παρόμοια ή ελαφρώς διαφορετική έννοια, όπως η εξόρυξη γνώσης από βάσεις δεδομένων

(knowledge mining from databases), εξαγωγή γνώσης (knowledge extraction), αρχαιολογία δεδομένων (archaeology data), ανάλυση δεδομένων (data analysis) κλπ.

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 το Data Mining, το οποίο 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:

- 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.

- Καθορισμός και επεξεργασία δεδομένων (data cleaning). Το βήμα αυτό περιλαμβάνει βασικές λειτουργίες 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.

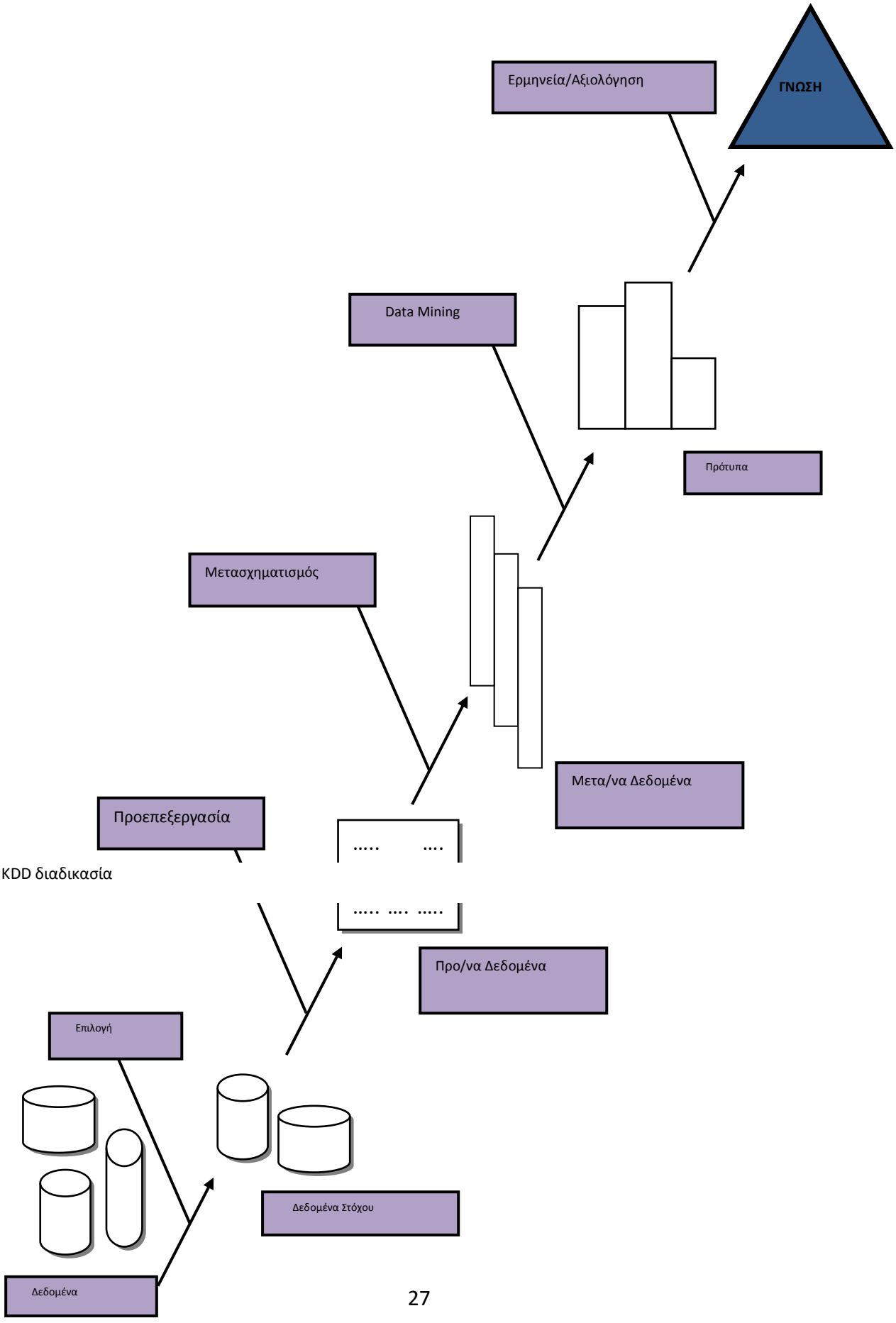
- Επιλογή των εργασιών εξόρυξης γνώσης (data mining) που θα χρησιμοποιηθούν για τις ανάγκες του προβλήματος π.χ. ταξινόμηση, πρόβλεψη, ομαδοποίηση κ.ά.

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

- 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 στον χρήστη.



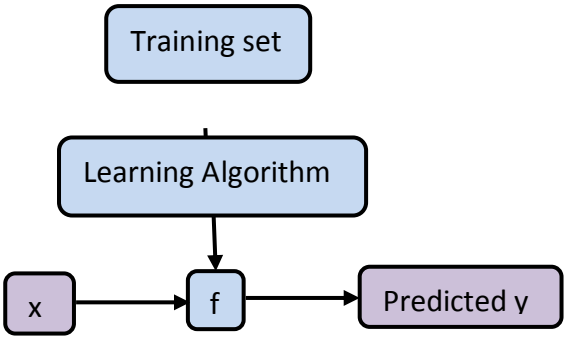

**Σχήμα 4:** KDD διαδικασία

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)**: Η ταξινόμηση είναι η δημοφιλέστερη και πιο κατανοητή στρατηγική του data mining και στηρίζεται σε 4 βασικά συστατικά:

- Κλάσεις : Είναι η εξαρτημένη κατηγορική μεταβλητή του μοντέλου.
- Ανεξάρτητες μεταβλητές (predictors): Δίνουν τα χαρακτηριστικά των δεδομένων που συμμετέχουν στη διαδικασία ταξινόμησης.
- Σετ δεδομένων εκμάθησης (training set): Το σετ δεδομένων εκμάθησης περιλαμβάνει τα 2 πρώτα συστατικά (κλάσεις και ανεξάρτητες μεταβλητές). Το μοντέλο εκπαιδεύεται σε αυτό το σετ για να προβλέψει τα μελλοντικά σημεία.
- Σετ δοκιμής: Νέα δεδομένα στα οποία ελέγχεται η ακρίβεια του μοντέλου μας.

Πιο συγκεκριμένα η ταξινόμηση έχοντας ένα σετ εκμάθησης $D = \{x_1, x_2, \dots, x_n\}$ και ένα σετ κλάσεων $C = \{y_1, y_2, \dots, y_m\}$ προσπαθεί να βρει την βέλτιστη συνάρτηση $f : D \to C$.




**Σχήμα 5: Παραστατική παρουσίαση της διαδικασίας ταξινόμησης. Έχοντας το σετ εκμάθησης προσπαθεί να βρει τη βέλτιστη συνάρτηση $f$ για την εκτίμηση της τιμής εξόδου μιας άγνωστης μεταβλητής.**

**Εκτίμηση (estimation):** Ο στόχος εδώ είναι να εκτιμήσουμε την τιμή εξόδου μιας άγνωστης μεταβλητής. Η διαφορά αυτής της στρατηγικής με την ταξινόμηση είναι ότι ενώ στην ταξινόμηση η μεταβλητή μας είναι κατηγορική εδώ είναι αριθμητική.

**Πρόβλεψη (prediction):** Ενώ οι δύο πρώτες στρατηγικές χρησιμοποιούνταν για να εκτιμήσουν συμπεριφορά (παρούσα) η πρόβλεψη χρησιμοποιείται για να προβλέψει τη μελλοντική συμπεριφορά.

**Ομαδοποίηση (grouping):** Καθορισμός των αντικειμένων που ανήκουν σε συγκεκριμένη ομάδα.

**Συσταδοποίηση (clustering):** Κατάτμηση ενός πληθυσμού σε ένα αριθμό υποομάδων ή συστάδων.

**Περιγραφή και Οπτικοποίηση (description and visualization):** Διερευνητικό ή οπτικό data mining.

Υπάρχουν πολλές τεχνικές εξόρυξης δεδομένων. Πιο συγκεκριμένα παρακάτω θα ασχοληθούμε με την τεχνική επιλογής χαρακτηριστικών (feature selection).
Οι αλγόριθμοι για την ανάλυση δεδομένων έχουν μελετηθεί από στατιστικούς και έχουν χρησιμοποιηθεί σε ποικίλους κλάδους. Ωστόσο, νέοι αλγόριθμοι χρειάζεται να σχεδιαστούν για να αντιμετωπίσουν τους περιορισμούς των υπαρχουσών τεχνικών που προκύπτουν από τους νέους τύπους δεδομένων που συλλέγονται. Η ραγδαία ανάπτυξη στην τεχνολογία πληροφοριών έκανε ικανή τη συγκέντρωση τεράστιων ποσών δεδομένων. 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 αναπτυσσόμενο πεδίο του Data Mining διορθώνει τους περιορισμούς υπαρχουσών τεχνικών ανάλυσης δεδομένων απευθυνόμενο σε αυτούς τους νέους τύπους δεδομένων. Τα τελευταία χρόνια το πεδίο αυτό εξελίχθηκε ραγδαία και συνεχίζει να παράγει μεγάλο αριθμό αλγορίθμων που απευθύνονται σε περαιτέρω περιορισμούς.

Υπάρχουν πολλές απαιτήσεις που χρειάζεται να αντιμετωπίσει ένας αναλυτής για την αποτελεσματική ανάπτυξη και δημιουργία αλγορίθμων. 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 υποθέσεις που παράγονται από αυτούς τους αλγορίθμους πρέπει να εκτιμηθούν από γερές στατιστικές μεθοδολογίες για να χρησιμοποιηθούν στην πράξη.

Σε πολλές περιπτώσεις η ανάλυση δεδομένων γίνεται καίριο κομμάτι της καθημερινής μας ζωής και γι’ αυτό προβάλλεται επιτακτική ανάγκη να επιλυθούν τα προβλήματα που προκύπτουν ώστε να βελτιώσουν αισθητά την καθημερινή μας ζωή.

# ΚΕΦΑΛΑΙΟ 2

## Τεχνικές Εξόρυξης Γνώσης

**(Μέθοδοι ταξινόμησης)**

Για την επιτυχή διεκπεραίωση των διαφόρων εργασιών data mining έχουν αναπτυχθεί πολλές τεχνικές. Κάποιες από τις πιο σημαντικές τεχνικές είναι οι ακόλουθες:

- Τα δέντρα αποφάσεων (Decision Trees)
- Τα νευρωνικά δίκτυα (Neural Networks)
- Λογιστική παλινδρόμηση (Logistic Regression)
- Ταξινόμηση (Classification)
- Μηχανές Διανυσματικής Υποστήριξης (Support Vector Machines)

Οι παραπάνω τεχνικές διαφέρουν ως προς την ακρίβεια και τη δυνατότητα κατανόησης τους.

## 2.1 Ταξινόμηση

## (Classification)

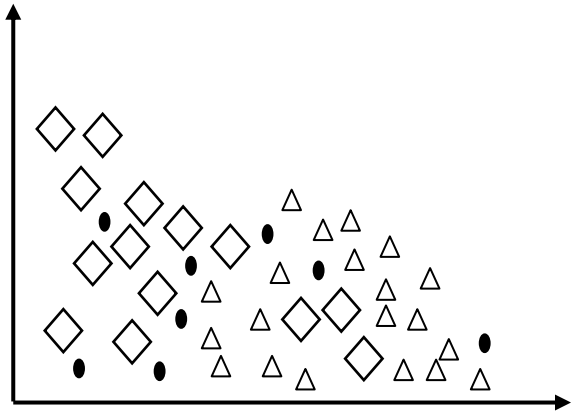
Στην Στατιστική και στην εκμάθηση μηχανών (machine learning), η ταξινόμηση (classification) είναι μια απαραίτητη διαδικασία ώστε να αναγνωρίσουμε σε ποιο σετ κατηγοριών (sub-populations) ανήκει μια νέα παρατήρηση, με βάση ένα σύνολο δεδομένων κατάρτισης που περιέχει τις παρατηρήσεις. Οι επιμέρους παρατηρήσεις αναλύονται σε ένα σύνολο μετρήσιμων ιδιοτήτων γνωστά ως χαρακτηριστικά (features). Αυτές οι παρατηρήσεις μπορεί να είναι κατηγορικές, αριθμητικές, ακέραιες ή πραγματικές. Ένας αλγόριθμος που υλοποιεί την ταξινόμηση, ειδικά σε μια συγκεκριμένη εφαρμογή, είναι γνωστός ως ένα

ταξινομητής (classifier). Ο όρος “ταξινομητής” μερικές φορές αναφέρεται επίσης στην μαθηματική συνάρτηση, που υλοποιείται από έναν αλγόριθμο ταξινόμησης, η οποία αντιστοιχίζει τα δεδομένα εισόδου σε μια κατηγορία.

Στην στατιστική, όπου η ταξινόμηση γίνεται συχνά με λογιστική παλινδρόμηση (logistic regression), δέντρα αποφάσεων, νευρωνικά δίκτυα, ή κάποια παρόμοια διαδικασία, οι παρατηρήσεις ονομάζονται επεξηγηματικές μεταβλητές (ή ανεξάρτητες μεταβλητές) και οι κατηγορίες που πρέπει να προβλεφθούν θεωρούνται τα πιθανά αποτελέσματα της εξαρτημένης μεταβλητής (predictors). Στην μηχανική μάθηση, οι παρατηρήσεις συχνά γνωστές ως περιπτώσεις(instances), οι ερμηνευτικές μεταβλητές ονομάζονται χαρακτηριστικά (features), τα οποία ομαδοποιούνται σε ένα διάνυσμα χαρακτηριστικών γνωρισμάτων, καθώς και οι πιθανές κατηγορίες που πρέπει να προβλεφθούν ονομάζονται τάξεις (classes).

Τι είναι ταξινόμηση;
Η ταξινόμηση αποτελεί μία από τις βασικές τεχνικές εξόρυξης δεδομένων. Βασίζεται στην εξέταση των χαρακτηριστικών ενός νέου αντικειμένου (μη κατηγοριοποιημένου), το οποίο με βάση τα χαρακτηριστικά αυτά, αντιστοιχίζεται σε ένα προκαθορισμένο σύνολο κλάσεων. Η διαδικασία της κατηγοριοποίησης χαρακτηρίζεται από ένα σαφή καθορισμό των κατηγοριών και το σύνολο που χρησιμοποιείται για την εκπαίδευση του μοντέλου αποτελείται από προκαθορισμένα παραδείγματα. Η ταξινόμηση δεδομένων είναι μια διαδικασία η οποία βρίσκει τις κοινές ιδιότητες μεταξύ ενός συνόλου αντικειμένων σε μια βάση δεδομένων και ταξινομεί τα αντικείμενα αυτά σε διαφορετικές κλάσεις (τάξεις) σύμφωνα με ένα μοντέλο ταξινόμησης.

Πιο συγκεκριμένα η διαδικασία ταξινόμησης, παίρνει ένα σύνολο από δείγματα δεδομένων, το καθένα αποτελούμενο από μετρήσεις σε ένα σύνολο μεταβλητών, με τις σχετικές ετικέτες, που ονομάζονται τύποι κλάσεων, και τα χρησιμοποιεί για να μάθουν ένα συγκεκριμένο μοντέλο. Χρησιμοποιώντας αυτό το μοντέλο, οι ετικέτες των νέων δειγμάτων μπορεί να υπολογιστούν.



**Σχήμα 6: Σημεία δύο διαστάσεων που ανήκουν σε διαφορετικές κλάσεις (ρόμβοι και τρίγωνα).*Στο πιο πάνω σχεδιάγραμμα παρατηρούμε σημεία δύο διαστάσεων που ανήκουν σε διαφορετικές κλάσεις (ρόμβοι και τρίγωνα). Ένας ταξινομητής θα μάθει ένα μοντέλο χρησιμοποιώντας αυτά τα σημεία και στην συνέχεια θα χρησιμοποιήσει αυτό το μοντέλο για να ταξινομήσει με ακρίβεια τα νέα σημεία που συμβολίζονται με μαύρη κουκίδα.***

Η ταξινόμηση μπορεί να εφαρμοστεί σε πολλά διαφορετικά πεδία όπως:

Α)Βιολογία. Κατά τα τελευταία χρόνια η μελέτη των μικροσυστοιχιών γονιδιακής έκφρασης έγινε πολύ δημοφιλής. Κάθε μικροσυστοιχία είναι ένα δείγμα, η οποία δίνει το επίπεδο έκφρασης πολλών γονιδίων 13 σε ένα άτομο, και τα δείγματα από διαφορετικές τάξεις (π.χ. υπό ανάρρωση άτομα και υγιή άτομα) είναι δεδομένη. Δεδομένα γονιδιακής έκφρασης χρησιμοποιήθηκαν επιτυχώς για την ταξινόμηση των ασθενών σε διάφορες κλινικές ομάδες, με αποτέλεσμα τον εντοπισμό νέων ομάδων της νόσου και των σχετικών γονιδίων για αυτό το κλινικό φαινόμενο.

Β) Ακτινοδιαγνωστική. Είναι μια σχετικά πρόσφατη ιατρική ειδικότητα που προέκυψε από το χωρισμό της παλιότερης ειδικότητας της ακτινολογίας. Οι ταξινομητές ασχολούνται με τη μελέτη διαγνωστικών εικόνων που παράγονται με διάφορες μεθόδους και την εξαγωγή διαγνωστικών συμπερασμάτων.

Γ) Οπτική αναγνώριση χαρακτήρων (Optical character recognition-OCR). Εδώ χρησιμοποιείται η ταξινόμηση για να μεταφραστούν εικόνες χειρόγραφες, δακτυλογραφημένες και από τυπωμένο κείμενο σε επεξεργάσιμο κείμενο σε μηχανές.

Δ)Ταξινόμηση Εγγράφων. Ταξινομούνται τα έγγραφα σε μία ή περισσότερες κλάσεις ή κατηγορίες.

Αυτά είναι μερικά από τα πολλά παραδείγματα σε διάφορους τομείς που υπάρχουν.

### 2.1.1 Μαθηματικό Μοντέλο

Το πρόβλημα της ταξινόμησης μπορεί να οριστεί ως εξής:
Δείγμα είναι ένα ζευγάρι $(x_i, y_i)$, όπου $x_i$ είναι ένα διάνυσμα μετρήσεων p-διάστασης, $x_i \epsilon R^p$ και $y_i$ είναι η ετικέτα του δείγματος που υποδεικνύει σε ποια κλάση ανήκει. Έστω μια βάση δεδομένων $D = \{d_1, d_2, \dots, d_n\}$, όπου $d_i$ είναι πλειάδες της μορφής $\langle d_{i1}, d_{i2}, \dots, d_{ip} \rangle$, που καλούνται στοιχεία ή εγγραφές ή παραδείγματα. Έστω επίσης, Ένα σύνολο κλάσεων $C = \{c_1, c_2, \dots, c_m\}$.

Το πρόβλημα της κατηγοριοποίησης συνιστάται στον προσδιορισμό της απεικόνισης $f : D \to C$, όπου κάθε $d_i$ αντιστοιχεί σε μια κλάση $c_j$. Η απεικόνιση αυτή ονομάζεται μοντέλο.

Έτσι μια κλάση $c_j$ ορίζεται ως το σύνολο των παραδειγμάτων που κατατάσσονται σ' αυτήν:

$$c_j = \{\frac{d_i}{f(d_i)} = c_j, 1 \le i \le n, d_i \epsilon D\},$$

όπου κάθε διάνυσμα $d_i$ θεωρείται ως ένα διάνυσμα. Τα $d_{ik}, k = 1, \dots, p$ είναι τιμές (διακριτές ή αριθμητικές), που αναφέρονται στα αντίστοιχα φυσικά χαρακτηριστικά (features) $x_1, x_2, \dots, x_p$. Γι' αυτό και ένα τέτοια διάνυσμα ονομάζεται διάνυσμα χαρακτηριστικών (feature vector). Κάθε χαρακτηριστικό $x_k$, μπορεί να πάρει τιμές $D_{x_k} = \{x_{k_i}, i = 1, \dots, r\}$. Επομένως σε ένα παράδειγμα $d_{ik}$ είναι μια από τις $x_{k_i}$, άρα $d_{ik} \epsilon D_{x_k}$.

Οι κλάσεις αναφέρονται και αυτές σε ένα χαρακτηριστικό $x_f$, που ονομάζεται χαρακτηριστικό στόχου (target feature). Πιο συγκεκριμένα, οι κλάσεις αντιστοιχούν στις διαφορετικές τιμές που μπορεί να πάρει το χαρακτηριστικό στόχου.

Ο πολύ μεγάλος αριθμός των διαθέσιμων χαρακτηριστικών, μπορεί να αποτελέσει ένα πραγματικό πρόβλημα για τους ταξινομητές, ειδικά όταν έχουμε σχετικά λίγα δείγματα. Έτσι, σε πολλές περιπτώσεις είναι ανάγκη να επιλέξουμε μόνο ένα μικρό υποσύνολο από τα διαθέσιμα χαρακτηριστικά που θα συμβάλουν στο μέγιστο στην ταξινόμηση. Αυτό ονομάζεται επιλογή χαρακτηριστικών (feature selection).

## 2.2 Δέντρα Αποφάσεων

Τα δέντρα απόφασης είναι πολύ ισχυρά εργαλεία που χρησιμοποιούνται ευρέως για τις περιπτώσεις της ταξινόμησης και της πρόβλεψης. Ένα δέντρο απόφασης αντιπροσωπεύει μια σειρά από IF THEN κανόνες ξεκινώντας από τη ρίζα του δέντρου και καταλήγοντας στα φύλλα του.

Οι εσωτερικοί κόμβοι ενός δέντρου απόφασης περιέχουν τα γνωρίσματα του προβλήματος, οι ακμές περιέχουν τις δυνατές τιμές των γνωρισμάτων και τα φύλλα περιέχουν τις πιθανές κλάσεις του προβλήματος. Απαραίτητο για την κατασκευή ενός δέντρου απόφασης είναι ένα σύνολο από στιγμιότυπα εκπαίδευσης, κάθε στιγμιότυπο του οποίου περιγράφεται από κάποια γνωρίσματα και την κλάση του προβλήματος στην οποία ανήκει.

Ξεκινώντας από τη ρίζα του δέντρου ο αλγόριθμος διασπά το σύνολο των στιγμιότυπων εκπαίδευσης σε υποσύνολα με βάση τη βέλτιστη ιδιότητα (best attribute) του κόμβου (η βέλτιστη ιδιότητα ενός κόμβου καθορίζεται από κάποιο κριτήριο όπως το information gain, το gain ratio, δείκτη Gini (Index Gini). Επομένως, μπορούμε να πούμε συμπερασματικά ότι ως ρίζα επιλέγουμε εκείνο το χαρακτηριστικό που δίνει το μέγιστο κέρδος πληροφορίας και για να το ποσοτικοποιήσουμε θα χρησιμοποιήσουμε την έννοια της εντροπίας. Έτσι προκύπτει ένα πλήθος υποσυνόλων που το καθένα περιέχει λιγότερα παραδείγματα από το αρχικό σύνολο. Για καθένα από αυτά τα επιμέρους υποσύνολα εφαρμόζεται επαναληπτικά η παραπάνω διαδικασία χρησιμοποιώντας τα εναπομείναντα γνωρίσματα, οπότε η διάσπαση των στιγμιότυπων προχωρά και σταματά όταν όλα τα στιγμιότυπα του υποσυνόλου ανήκουν στην ίδια κλάση ή έχουν εξαντληθεί όλα τα γνωρίσματα.

Εκτός από το σύνολο των στιγμιότυπων εκπαίδευσης, υπάρχει και το σύνολο ελέγχου με βάση το οποίο ελέγχεται η απόδοση του δέντρου, δηλαδή η ακρίβεια με την οποία το κατασκευασμένο δέντρο απαντά στο πρόβλημα της ταξινόμησης. Στην περίπτωση αυτή δίνουμε ως είσοδο στο δέντρο τις τιμές των γνωρισμάτων του στιγμιότυπου ελέγχου και περιμένουμε ως απάντηση την τάξη του στιγμιότυπου. Το πλήθος των λανθασμένων απαντήσεων (δηλαδή τα στιγμιότυπα στα οποία το δέντρο απάντησε διαφορετική κλάση από την πραγματική) καθορίζει την ακρίβεια του δέντρου.

Έτσι, η διαδικασία κατασκευής δέντρου απόφασης είναι η εξής:

- Επιλογή χαρακτηριστικού για τη θέση του αρχικού κόμβου (ρίζας) και δημιουργία κλάδων για κάθε πιθανή τιμή του χαρακτηριστικού.

- Διάσπαση υποδειγμάτων σε υποσύνολα ένα για κάθε κλάδο που εκτείνεται από τη ρίζα.
- Επανάληψη των παραπάνω για κάθε κλάδο με χρήση μόνο του υποσυνόλου των υποδειγμάτων κάθε κλάδου.
- Ολοκλήρωση της διαδικασίας όταν όλα τα υποδείγματα σε ένα κόμβο ανήκουν στην ίδια τάξη.

Όταν έχει ολοκληρωθεί η διαδικασία ανακάλυψης γνώσης με χρήση του αλγορίθμου, τότε το δέντρο μπορεί να αναπαρασταθεί ως σύνολο κανόνων της μορφής:

« If <ΣΥΝΟΛΟ ΣΥΝΘΗΚΩΝ> then <ΣΥΜΡΕΡΑΣΜΑ>» .

Η ανακάλυψη γνώσης με χρήση αλγορίθμων δέντρων απόφασης αποτελεί μια από τις πλέον δημοφιλείς τεχνικές επαγωγικής εκμάθησης και έχει μεγάλη εφαρμογή στη διάγνωση ιατρικών περιπτώσεων, στην εκτίμηση πιθανού ρίσκου από πιστοληπτικές τραπεζικές εργασίες κ.α.

### 2.2.1 Κανόνες Διακοπής

Οι κανόνες διακοπής-ολοκλήρωσης της διαδικασίας ελέγχουν αν η διαδικασία κατασκευής δέντρου πρέπει να σταματήσει ή όχι. Χρησιμοποιούνται οι εξής κανόνες διακοπής:

- Αν ο κόμβος γίνει καθαρός: δηλαδή αν όλες οι περιπτώσεις μέσα σε ένα κόμβο έχουν πανομοιότυπες τιμές της εξαρτημένης μεταβλητής τότε ο κόμβος δε θα διαχωριστεί.
- Αν όλες οι περιπτώσεις μέσα σε ένα κόμβο έχουν πανομοιότυπες τιμές για κάθε μεταβλητή πρόβλεψης τότε ο κόμβος δε κα διαχωριστεί.
- Αν το βάθος του πρόσφατου δέντρου πλησιάζει την τιμή του μέγιστου ορίου βάθους το οποίο καθορίζεται από το χρήστη, η διαδικασία κατασκευής δέντρου θα σταματήσει.
- Αν το μέγεθος ενός κόμβου είναι μικρότερο από την ελάχιστη τιμή μεγέθους του κόμβου που ορίζεται από τον χρήστη τότε ο κόμβος δεν θα διαχωριστεί.
- Αν ο διαχωρισμός ενός κόμβου έχει σαν αποτέλεσμα ένα θυγατρικό κόμβο του οποίου το μέγεθος είναι μικρότερο από την ελάχιστη τιμή μεγέθους του κόμβου που ορίζεται από το χρήστη τότε ο κόμβος δε θα διαχωριστεί.
- Ο καλύτερος διαχωρισμός για ένα κόμβο αποδίδει μια μείωση στη μη καθαρότητα η οποία είναι μικρότερη από την ελάχιστη αλλαγή στη μη καθαρότητα που ορίζεται από το χρήστη.

### 2.2.2 Διαδικασία Κλαδέματος

Το κλάδεμα (pruning) αναφέρεται στη διαδικασία του ελέγχου ενός πλήρους αναπτυσσόμενου δέντρου και της αφαίρεσης των διαχωρισμών των κάτω επιπέδων που δεν έχουν σημαντική συνεισφορά στην ακρίβεια του δέντρου. Το λογισμικό στο κλάδεμα του δέντρου προσπαθεί να δημιουργήσει το μικρότερο δέντρο του οποίου το ρίσκο λανθασμένης ταξινόμησης δεν είναι πολύ μεγαλύτερο από το ρίσκο λανθασμένης ταξινόμησης του μεγαλύτερου πιθανού δέντρου. Η διαδικασία αφαιρεί ένα κλαδί δέντρου, αν το κόστος το οποίο σχετίζεται με τη μεγαλύτερη πολυπλοκότητα του δέντρου είναι μεγαλύτερο από το κέρδος το οποίο σχετίζεται με το εάν έχουμε ένα άλλο επίπεδο κόμβων (κλαδί). Χρησιμοποιεί ένα δείκτη ο οποίος μετρά το ρίσκο λανθασμένης ταξινόμησης και την πολυπλοκότητα του δέντρου αφού στόχος μας είναι να ελαχιστοποιήσουμε και τα δύο.

Το μέτρο κόστους πολυπλοκότητας (cost complexity) ορίζεται ως εξής:

$$R_a(T) = R(T) + a|\tilde{T}|$$

όπου
$R(T)$ είναι το ρίσκο λανθασμένης ταξινόμησης του δέντρου T
$|\tilde{T}|$ είναι το πλήθος των τερματικών κόμβων για το δέντρο T
$\alpha$ είναι το κόστος πολυπλοκότητας ανά τερματικό κόμβο για το δέντρο

Η τιμή α υπολογίζεται από τον αλγόριθμο κατά τη διάρκεια του κλαδέματος.

Κάθε δέντρο που μπορούμε να παράγουμε έχει ένα μέγιστο μέγεθος ($T_{max}$), όπου σε κάθε τερματικό κόμβο περιέχεται μόνο μια καταχώρηση. Στην περίπτωση που το κόστος πολυπλοκότητας είναι μηδενικό ($\alpha = 0$) , το μέγιστο δέντρο έχει το χαμηλότερο ρίσκο, αφού κάθε εγγραφή προβλέπεται τέλεια. Επομένως, όσο μεγαλύτερη είναι η τιμή του α, τόσο μικρότερος είναι ο αριθμός των τερματικών κόμβων στο $T(\alpha)$, δηλαδή το δέντρο με το μικρότερο κόστος πολυπλοκότητας για το δοσμένο α. Όταν το α αυξάνεται από το 0 τότε παράγει μια πεπερασμένη ακολουθία από υποδέντρα ($T_1, T_2, T_3$ ), το καθένα με λιγότερους τερματικούς κόμβους από το προηγούμενο. Το κλάδεμα κόστους πολυπλοκότητας δουλεύει αφαιρώντας τον πιο αδύναμο διαχωρισμό. Οι εξισώσεις που ακολουθούν εκφράζουν το κόστος πολυπλοκότητας για τον κόμβο {t}, που είναι ένας οποιοσδήποτε ξεχωριστός-μόνος κόμβος και για $T_t$, τον υπο-κλάδο του {t}:

$$R_a(\{t\}) = R(t) + a$$

και

$$R_a(T_t) = R(T_t) + a|\tilde{T}_t|$$

Στην περίπτωση που το $R_a(T_t)$ είναι μικρότερο από το $R_a(\{t\})$, το κλαδί $T_t$ έχει μικρότερο κόστος πολυπλοκότητας από αυτό του ξεχωριστού κόμβου {t}.

Η διαδικασία ανάπτυξης του δέντρου εξασφαλίζει ότι για $\alpha = 0$ ισχύει

$$R_a(\{t\}) \geq R_a(T_t) \quad (1)$$

Καθώς το α αυξάνεται από το 0, τα $R_a(\{t\})$ και $R_a(T_t)$ αυξάνονται γραμμικά με το $R_a(T_t)$ να αυξάνεται με ταχύτερο ρυθμό. Τελικά, βρίσκουμε ένα κάτω όριο α' τέτοιο ώστε $R_a(\{t\}) < R_a(T_t)$ για όλα τα α<α'. Συμπεραίνουμε ότι όταν το α γίνεται μεγαλύτερο από το α', το κόστος πολυπλοκότητας του δέντρου μειώνεται αν κόψουμε το υποκλάδι (sub branch) $T_t$ κάτω από το $\{t\}$.

Μπορούμε εύκολα να υπολογίσουμε το κάτω όριο λύνοντας την (1) ούτως ώστε να βρούμε τη μεγαλύτερή τιμή του α για την οποία ισχύει η ανισότητα, η οποία συμβολίζεται και ως $g(t)$. Συνεπώς, προκύπτει:

$$\alpha \leq g(t) = \frac{R(t) - R(T_t)}{|\tilde{T}_t| - 1}$$

Μπορούμε να ορίσουμε σαν το πιο αδύναμο (link) σύνδεσμο (t) στο δέντρο T τον κόμβο ο οποίος παίρνει τη μικρότερη τιμή του $g(t)$:

$$g(\dot{t}) = \min_{t \in T} g(t)$$

Κατά συνέπεια, καθώς το α αυξάνεται, ο $\dot{t}$ είναι ο πρώτος κόμβος για τον οποίο ισχύει $R_a(\{t\}) = R_a(T_t)$. Στο σημείο αυτό, το $\{t\}$ προτιμάται από το $T_{\dot{t}}$ και το υποκλάδι κλαδεύεται.

Συνοπτικά ο αλγόριθμος κλαδέματος βασίζεται στα εξής βήματα:

- Ορίζουμε το $\alpha_1 = 0$ και ξεκινάμε με το δέντρο για το οποίο $T_1 = T(0)$ δηλαδή το πλήρως αναπτυσσόμενο δέντρο.
- Αυξάνουμε το α μέχρι το κλάδεμα ενός κλαδιού. Έπειτα, κλαδεύουμε το κλαδί από το δέντρο και υπολογίζουμε την εκτίμηση του ρίσκου του δέντρου το οποίο έχουμε κλαδέψει.
- Επαναλαμβάνουμε το προηγούμενο βήμα μέχρι να απομείνει μόνο ο αρχικός κόμβος ρίζα, αποδίδοντας μια σειρά από υποδέντρα $T_1, T_2, \ldots, T_k$.
- Στην περίπτωση που επιλέξουμε τον κανόνα του τυπικού σφάλματος, τότε διαλέγουμε το μικρότερο δέντρο $T_{opt}$ για το οποίο

$$R(T_{opt}) \leq \min_k R(T_k) + m \times SE(R(T))$$

- Στην περίπτωση που δεν επιλέγουμε τον κανόνα τυπικού σφάλματος τότε διαλέγουμε το δέντρο με τη μικρότερη τιμή ρίσκου $R(T)$.

## 2.3 Λογιστική Παλινδρόμηση

Το μοντέλο της Λογιστικής παλινδρόμησης (logistic regression) αποτελεί ειδική περίπτωση των γενικευμένων γραμμικών μοντέλων. Άρχισε να χρησιμοποιείται ευρέως κατά την δεκαετία του '50 , κυρίως με εφαρμογές στη βιοστατιστική. Είναι μια μέθοδος στατιστικής ανάλυσης που χρησιμοποιεί ένα σύνολο ανεξάρτητων μεταβλητών για τη διερεύνηση της κίνησης μιας κατηγορικής εξαρτημένης μεταβλητής.

Η Λογιστική Παλινδρόμηση είναι χρήσιμη σε καταστάσεις στις οποίες επιθυμούμε την πρόβλεψη ύπαρξης ή της απουσίας ενός χαρακτηριστικού ή συμβάντος. Η πρόβλεψη αυτή βασίζεται στην κατασκευή ενός γραμμικού μοντέλου και συγκεκριμένα στον προσδιορισμό των τιμών που παίρνουν οι συντελεστές ενός συνόλου ανεξάρτητων μεταβλητών που χρησιμοποιούνται ως μεταβλητές πρόβλεψης.

Σε πολλές εφαρμογές η εξαρτημένη μεταβλητή παίρνει δυο μόνο τιμές, οι οποίες αντιστοιχούν σε δύο ενδεχόμενα. Για παράδειγμα , το αν ο ασθενής ζει ή απεβίωσε, το αν ο άνεργος βρίσκει εργασία ή όχι, το αν ραγίζει ή αντέχει ένα δοκάρι. Οι τιμές της μεταβλητής αποτελούν μια αυθαίρετη κωδικοποίηση των δύο ενδεχομένων, συνήθως 0 και 1.

Εάν ορίσουμε την τιμή $y = 1$ σαν «επιτυχία» και την τιμή $y = 0$ σαν «αποτυχία», τότε η $y$ είναι τυχαία μεταβλητή της κατανομής Bernoulli, δηλαδή $y \sim B(p)$, με μέση τιμή $E(y) = p$ και διασπορά $V(y) = p(1-p)$.

Γενικεύοντας σε μια σειρά από n επαναλήψεις ορίζουμε

$$y = αριθμός\ επιτυχιών\ σε\ n\ δοκιμές.$$

Υπό την υπόθεση ότι η πιθανότητα επιτυχίας είναι ίδια σε κάθε δοκιμή και οι δοκιμές είναι ανεξάρτητες μεταξύ τους, ισχύει η Διωνυμική Κατανομή (Binomial):

$$y \sim b(n, p)$$

με συνάρτηση πυκνότητας

$$f(y) = \binom{n}{p} p^y (1-p)^{n-y}, όπου\ y = 1,2, \dots, n$$

όπου p η πιθανότητα επιτυχίας η οποία είναι παράμετρος της κατανομής.

Η Διωνυμική κατανομή αποτελεί τη βασική κατανομή για την περιγραφή και ανάλυση μιας μεταβλητής αυτής της φύσης. Η μέση τιμή της y είναι ίση με $E(y) = np$ και η διασπορά με $V(y) = np(1-p)$. Στην ειδική περίπτωση που $n = 1$ μιλάμε για δυαδικά δεδομένα, αλλιώς για διωνυμικά δεδομένα.

Σε πολλές περιπτώσεις η τ. μ. ενδέχεται να εξαρτάται από κάποιες επεξηγηματικές μεταβλητές. Η εξάρτηση της από τις επεξηγηματικές μεταβλητές (ανεξάρτητες μεταβλητές ή συμμεταβλητές) εισάγεται μέσω της εξάρτησης της πιθανότητας επιτυχίας p από τις x (π. χ. η πιθανότητα να μείνει κάποιος άνεργος εξαρτάται από το φύλο, την ηλικία, το μορφωτικό επίπεδο κ. ά.). Πιο συγκεκριμένα, κατασκευάζεται το αποκαλούμενο μοντέλο λογιστικής παλινδρόμησης, το οποίο είναι ένα γενικευμένο γραμμικό μοντέλο και εκφράζεται μέσω της σχέσης:

$$n_x = g\big(E(y_x)\big) = g(\mu_x) = x'\beta$$

με την ακόλουθη δομή:

1. $y_x \sim b(n_x, \mu_x)$ $(n_x > 1, \delta\iota\omega\nu\upsilon\mu\iota\kappa\acute{\alpha}\ \delta\epsilon\delta o\mu\acute{\epsilon}\nu\alpha)$
   ή $y_x \sim B(n_x, \mu_x)$ $(n_x = 1, \delta\upsilon\alpha\delta\iota\kappa\acute{\alpha}\ \delta\epsilon\delta o\mu\acute{\epsilon}\nu\alpha)$

2. $n_x = g(\mu_x) = ln\frac{\mu_x}{n_x - \mu_x} = ln\frac{p_x}{1-p_x} = logit(p_x) = x'\beta$

3. Ανεξαρτησία παρατηρήσεων $y_x$,
   όπου:
   $n_x$ είναι ο αριθμός των επαναλήψεων της τιμής του διανύσματος $x$ των επεξηγηματικών μεταβλητών.

Αντιστρέφοντας τη συνάρτηση σύνδεσης προκύπτει:

$$p_x = \frac{e^{n_x}}{1 + e^{n_x}}$$

για την οποία ισχύει ο περιορισμός $0 < p_x < 1$.

Για κάθε παρατήρηση i το μοντέλο γράφεται ως:

$$\ln\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}, i = 1, \ldots, n$$

όπου

$$p_i = p_{xi} = \frac{e^{\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}}} = \frac{1}{1 + e^{-x_i'\beta}}$$

η πιθανότητα επιτυχίας

$$x_i'\beta = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$$

είναι ο linear predictor.

$$E(y_i) = n_i p_i = n_i \frac{e^{x_i'\beta}}{1 + e^{x_i'\beta}}$$

### 2.3.1 Εκτίμηση παραμέτρων με τη μέθοδο μεγίστης πιθανοφάνειας (Maximum likelihood)

Ας υποθέσουμε ότι τα δεδομένα μας είναι χωρισμένα σε κατηγορίες. Δηλαδή, έχουμε $n_i$ στο πλήθος πειραματικές μονάδες στο i-οστό σημείο δεδομένων (για παράδειγμα μπορούμε να θεωρήσουμε ότι το $n_i$ είναι το πλήθος των πειραματόζωων στα οποία έχουμε δώσει μια συγκεκριμένη δοσολογία φαρμάκου). Το μοντέλο μας βάση της εξίσωσης (1), γράφεται στη μορφή:

$$E(y_i) = n_i P(x_i) = n_i \frac{1}{1 + e^{-x_i'\beta}}, i = 1, \dots, m$$

Με $y_1, y_2, \dots, y_m$ να είναι οι παρατηρούμενες τιμές των ανεξάρτητων διωνυμικών τυχαίων μεταβλητών. Σε αυτή την περίπτωση ισχύει:

$$var(y_i) = n_i P(x_i)[1 - P(x_i)]$$

και το άθροισμα

$$\sum_{i=1}^{m} n_i = n$$

είναι το συνολικό πλήθος του δείγματος.

Η συνάρτηση πιθανότητας μιας απλής διωνυμικής τυχαίας μεταβλητής y, με παραμέτρους n, P δίνεται από τον τύπο:

$$\binom{n}{y} P^y (1 - P)^{n-y}$$

Ωστόσο, ο όρος $\binom{n}{y}$ δεν περιλαμβάνει το β, οπότε δε μπορεί να χρησιμοποιηθεί. Επομένως, η log πιθανοφάνεια για το λογιστικό μοντέλο παλινδρόμησης δίνεται από τον τύπο:

$$\ln[\mathcal{L}(P; y)] = \sum_{i=1}^{m} \left\{ y_i \ln\left[\frac{P(x_i)}{1 - P(x_i)}\right] + n_i \ln[1 - P(x_i)] \right\} \quad (2)$$

Είναι εφικτό τώρα να εισάγουμε τη μορφή του λογιστικού μοντέλου στην εξίσωση (1).

Ο όρος $\ln\left[\frac{P(x_i)}{1-P(x_i)}\right]$ ονομάζεται logit και γράφεται ως:

$$\ln\left[\frac{P(x_i)}{1-P(x_i)}\right] = x_i'\beta = \beta_0 + \sum_{j=1}^{k} x_{ij}\beta_j, \qquad i = 1,2,\dots,m, \qquad m \geq k+1$$

Σαν αποτέλεσμα η log πιθανοφάνεια της εξίσωσης (2) γράφεται ως:

$$\ln[\mathcal{L}(P;y)] = \sum_{i=1}^{m}\sum_{j=1}^{k} y_i x_{ij}\beta_j - \sum_{i=1}^{m} n_i \ln\left(1 + e^{\sum_{j=1}^{k} x_{ij}\beta_j}\right) \quad (3)$$

Στη συνέχεια η εξίσωση (3) πρέπει να μεγιστοποιηθεί ως προς τον όρο $\beta_j$. Σε μορφή πινάκων η εξίσωση γράφεται ως:

$$\ln[\mathcal{L}(\beta;y)] = \beta' X y - \sum_{i=1}^{m} n_i \ln\left(1 + e^{x_i'\beta}\right) \quad (4)$$

όπου:
Χ είναι ο κλασικός πίνακας του μοντέλου που συναντάμε και στη γραμμική παλινδρόμηση και y το διάνυσμα απόκρισης.

Παραγωγίζουμε τώρα την εξίσωση (4) ως προς β:

$$\frac{\partial ln[\mathcal{L}(\beta_i;y)]}{\partial\beta} = X'y - \sum_{i=1}^{m}\left[\frac{n_i}{1+e^{x_i'\beta}}\right] e^{x_i'\beta} x_i$$

Γνωρίζοντας ότι ισχύει:

$$\frac{e^{x_i'\beta}}{1+e^{x_i'\beta}} = \frac{1}{1+e^{-x_i'\beta}} = P(x_i)$$

Προκύπτει ότι:

$$\frac{\partial ln[\mathcal{L}(\beta_i;y)]}{\partial\beta} = X'y - \sum_{i=1}^{m} n_i P(x_i) x_i$$

Εφόσον, ο όρος $n_i P(x_i)$ αποτελεί τον μέσο της διωνυμικής τυχαίας μεταβλητής το δεξί μέλος, της παραπάνω σχέσης γράφεται σε μορφή πινάκων ως $X'(y-\mu)$ όπου:

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_m \end{bmatrix}$$

και $\mu_i = n_i P(x_i)$. Σαν αποτέλεσμα ο εκτιμητής μεγίστης πιθανοφάνειας (Maximum Likelihood Estimator-MLE) είναι η λύση της εξίσωσης (score equation):

$$X'(y - \mu) = 0 \quad (5)$$

Για τη λύση της εξίσωσης (5) μπορούμε να χρησιμοποιήσουμε μια επαναληπτική διαδικασία για να παράγουμε τις εκτιμήσεις $b_0, b_1, \dots, b_k$ των όρων $\beta_0, \beta_1, \dots, \beta_k$ για τις $p = k + 1$ παραμέτρους του μοντέλου. Μια τέτοια αριθμητική μέθοδος είναι αυτή των σταθμισμένων ελάχιστων τετραγώνων (weighted least squares).

Ο τύπος για το σταθμισμένο άθροισμα τετραγώνων των υπολοίπων είναι:

$$S = \sum_{i=1}^{m} \left[ \frac{(y_i - \mu_i)^2}{\sigma_i^2} \right]$$

όπου
$\mu_i = n_i P(x_i)$ και $\sigma_i^2$ είναι η διωνυμική διακύμανση στο i-οστό σημείο δεδομένων με:

$$\sigma_i^2 = n_i P(x_i)[1 - P(x_i)] = n_i \frac{e^{-x_i'\beta}}{(1 + e^{-x_i'\beta})^2}$$

Ελαχιστοποιούμε το S:

$$minS = \min_{\beta} \sum_{i=1}^{m} \left[ \frac{(y_i - \mu_i)^2}{\sigma_i^2} \right]$$

Η διακύμανση $\sigma_i^2$ είναι σταθερή, επομένως παραγωγίζουμε μόνο τον αριθμητή του S και παίρνουμε:

$$2 \sum_{i=1}^{m} \left[ \frac{(y_i - \mu_i)^2}{\sigma_i^2} \right] \left( \frac{\partial \mu_i}{\partial \beta} \right)$$

Ισχύει ότι:

$$\frac{\partial \mu_i}{\partial \beta} = n_i P(x_i)[1 - P(x_i)] x_i = \sigma_i^2 x_i$$

Επομένως, η λύση που παίρνουμε από την ελαχιστοποίηση του σταθμισμένου αθροίσματος τετραγώνων των υπολοίπων με σταθερό $\sigma_i^2$ είναι:

$$\sum_{i=1}^{m} (y_i - \mu_i) x_i = 0$$

η οποία είναι παρόμοια με την εξίσωση $X'(y-\mu)=0$. Άρα, μια επαναληπτική μέθοδος όπως η παραπάνω μπορεί να χρησιμοποιηθεί για να προσδιοριστούν οι αριθμητικές τιμές των $b_0, b_1, \dots, b_k$ δηλαδή των εκτιμητών μεγίστης πιθανοφάνειας.

### 2.3.2 Άλλες μορφές στατιστικής συμπερασματολογίας για τις οποίες γίνεται χρήση της λογιστικής παλινδρόμησης

Όπως έχουμε δει η λογιστική παλινδρόμηση χρησιμοποιείται σε πολλές διαφορετικές περιπτώσεις για την εξαγωγή συμπερασμάτων, όπως για παράδειγμα σε κλινικές δοκιμές όπου πρέπει να συγκρίνουμε τα αποτελέσματα διαφορετικών θεραπειών των οποίων το αποτέλεσμα έχει δυαδική μορφή. Για την βελτίωση του μοντέλου εξετάζεται η σημασία της κάθε μεταβλητής. Σε πολλές περιπτώσεις τα δεδομένα που έχουμε δεν είναι ομαδοποιημένα, δηλαδή $n_i = 1$. Όταν όμως οι πειραματικές μονάδες του δείγματος είναι σχετικά ομοιογενείς, τότε η λογιστική παλινδρόμηση μπορεί να πάρει τη μορφή μιας καμπύλης «δόσης-απόκρισης», όπου μετράει την ανταπόκριση ενός ασθενή ανάλογα με την δοσολογία που του χορηγείται. Σε τέτοια περίπτωση ισχύει ότι $k = 1$ *και* $p = 2$ και το μοντέλο παίρνει τη μορφή:

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

Σε αρκετές περιπτώσεις τα όρια εμπιστοσύνης για τα $\beta_0$ και $\beta_1$ καθώς επίσης και τα όρια εμπιστοσύνης για τους συντελεστές της $P(x_i)$ είναι σημαντικά για τους ερευνητές. Το ενδιαφέρον για την μελέτη του κάθε συντελεστή παλινδρόμησης ξεχωριστά πηγάζει από την ανάγκη να προσδιορίσουμε τους λόγους πιθανοτήτων (odd ratios).

Η ιδέα προσδιορισμού του λόγου πιθανοτήτων είναι αποτέλεσμα της χρήσης του logit(P) που δίνεται από τον τύπο:

$$Log[\frac{P}{1-P}]$$

Στη γενική σχέση (1) του λογιστικού μοντέλου παλινδρόμησης, το $logit[P(x_i)]$ δίνεται από τη σχέση:

$$\ln\left[\frac{P(x_i)}{1-P(x_i)}\right] = x_i'\beta$$

και μέσω του μετασχηματισμού P, γραμμικοποιείται η λογιστική συνάρτηση.

### 2.3.3 Ιδιότητες της διασποράς των εκτιμητών μέγιστης πιθανοφάνειας στη λογιστική παλινδρόμηση

Είναι ευρέως γνωστό ότι οι εκτιμητές μεγίστης πιθανοφάνειας παρουσιάζουν ασυμπτωτικές ιδιότητες στη διακύμανση και τη συνδιακύμανση, στον πίνακα πληροφορίας. Στην περίπτωση ενός γραμμικού μοντέλου με κανονικά, ανεξάρτητα και ομοιόμορφα κατανεμημένα σφάλματα, ο πίνακας πληροφορίας για τους εκτιμώμενους συντελεστές παλινδρόμησης εκφράζεται από τον τύπο:

$$I(b) = \frac{X'X}{\sigma^2}$$

όπου $\sigma^2$ είναι η διακύμανση του σφάλματος. Σαν αποτέλεσμα σε αυτή την περίπτωση ο πίνακας διακύμανσης-συνδιακύμανσης (variance-covariance matrix) των εκτιμώμενων συντελεστών είναι:

$$I^{-1}(b) = (X'X)^{-1}\sigma^2$$

Ο πίνακας πληροφορίας παρουσιάζει, κατά μια έννοια, την ποιότητα των πληροφοριών των παραμέτρων που διατίθενται από τα δεδομένα μας. Ένας σχετικά μεγάλος πίνακας πληροφορίας σημαίνει μικρότερες διακυμάνσεις στους εκτιμώμενους συντελεστές του μοντέλου. Μπορούμε να υπολογίσουμε τον πίνακα πληροφορίας με διάφορες μεθόδους, όπως με τη βοήθεια της εξίσωσης (5):

$$\begin{aligned} I(b) &= var(score) \\ &= var[X'(y-\mu)] \end{aligned}$$

όπου var συμβολίζουμε τον πίνακα διακύμανσης-συνδιακύμανσης.

Χρησιμοποιώντας τον τελεστή τυπικής διακύμανσης, η παραπάνω εξίσωση αποκτά τη μορφή:

$$var[X'(y-\mu)] = X'var[(y-\mu)]X$$

Για το μοντέλο λογιστικής παλινδρόμησης έχουμε υποθέσει για τις $y_1, y_2, \dots, y_m$ ανεξάρτητες παρατηρήσεις, ότι κάθε $y_i$ παρατήρηση είναι μια διωνυμική τυχαία μεταβλητή με μέσο όρο $n_i P(x_i)$ και διασπορά $\sigma_i^2 = n_i[P(x_i)][1-P(x_i)]$. Επομένως, ισχύει

$$V = diag\{\sigma_1^2, \sigma_2^2, \dots, \sigma_m^2\}$$

και

$$I(b) = X'VX$$

Ο ασυμπτωτικός πίνακας διακύμανσης-συνδιακύμανσης του b, δίνεται τελικά από τον τύπο:

$$var(b) = (X'VX)^{-1}$$

Συνεπώς τα εκτιμώμενα τυπικά σφάλματα βρίσκονται στα διαγώνια στοιχεία του V το οποίο αντικαθιστά το $\hat{V}$ από τη στιγμή που τα β της $P(x_i)$ έχουν αντικατασταθεί από τα εκτιμώμενα b.

### 2.3.4 Συμπερασματολογία με χρήση της μεθόδου Wald στη λογιστική παλινδρόμηση

Η πρώτη εφαρμογή της μεθόδου Wald πραγματεύεται με έλεγχο υποθέσεων για κάθε ξεχωριστό συντελεστή του μοντέλου της λογιστικής παλινδρόμησης. Πιο συγκεκριμένα, θέλουμε να ελέγξουμε:

$$H_0: \beta_j = 0 \text{ έναντι } H_1: \beta_j \neq 0$$

με το $\beta_j$ να εμφανίζεται στον linear predictor $x_i'\beta$ του λογιστικού μοντέλου στην εξίσωση (1).

Για την εκτίμηση της μεγίστης πιθανοφάνειας $b_j$ ισχύει ότι:

$$z_j = \frac{b_j - \beta_j}{\sigma b_j}$$

ο οποίος ακολουθεί τυπική κανονική κατανομή N(0,1) και έτσι ισχύει ότι το

$$z_j^2 = (\frac{b_j}{\sigma b_j})^2$$

ακολουθεί ασυμπτωτικά την $\mathcal{X}_1^2$ κατανομή υπό την $H_0$ υπόθεση, όπου $\sigma b_j$ είναι το κατάλληλο διαγώνιο στοιχείο του ασυμπτωτικού πίνακα variance-covariance των b.

Στην πράξη αντικαθιστούμε τα $\sigma b_j$ με τα $\hat{\sigma} b_j$. . Ο έλεγχος που διεξάγουμε, είναι ο

συνηθισμένος μονομερής ή διμερής έλεγχος (one or two sided test). Για τον υπολογισμό των τιμών της $\mathcal{X}_1^2$ και της p-τιμής για κάθε συντελεστή του απαιτούμενου μοντέλου γίνεται χρήση διαφόρων στατιστικών πακέτων.

Μια δεύτερη μορφή της Wald συμπερασματολογίας έχει να κάνει με τον υπολογισμό του διαστήματος εμπιστοσύνης της διωνυμικής πιθανότητας για κάποια δοσμένα ή αυθαίρετα δεδομένα. Θα μπορούσε να χρησιμοποιηθεί η μέθοδος Δέλτα για το σκοπό αυτό αλλά λόγω ύπαρξης του linear predictor $x_i'\beta$ στο λογιστικό μοντέλο ακολουθείται μια εναλλακτική διαδικασία υπολογισμού των διαστημάτων εμπιστοσύνης.

Στη λογιστική παλινδρόμηση πρέπει να έχουμε υπόψη ότι θ μέση απόκριση στο $x = x_i$ δίνεται από τον τύπο $\frac{1}{1+e^{-x_i'\beta}}$ και άρα είναι πιθανότητα. Για παράδειγμα, ένας μηχανικός πιθανόν να απαιτεί ένα 95% διάστημα εμπιστοσύνης για την πιθανότητα «ελαττωματικού» προϊόντος σε μια βιομηχανία όπου οι συνθήκες παραγωγής ορίζονται ως $x = x_i$.

Η σημειακή εκτίμηση της πιθανότητας δίνεται από το $\widehat{y_i} = \hat{P}(x_i)$.

Στο λογιστικό μοντέλο $P = \frac{1}{1+e^{-x_i'\beta}}$ το P είναι μια μονότονη εξίσωση του x'. Μπορούμε να ορίσουμε ένα 100(1-α)% διάστημα εμπιστοσύνης στο P, χρησιμοποιώντας ένα διάστημα εμπιστοσύνης στο $x'\beta$. Προφανώς, ο linear predictor περιέχει όρους που είναι γραμμικοί στο β και μπορούμε να εκμεταλλευτούμε το γεγονός ότι ο b (εκτιμητής μεγίστης πιθανοφάνειας του β) είναι ασυμπτωτικά κανονικός. Συνεπώς, ένα άνω διάστημα εμπιστοσύνης για το $x'\beta$, παράγει ένα άνω διάστημα για το P.

Ασυμπτωτικά ισχύει ότι:

$$x'b \sim N(x'\beta, x'(X'VX)^{-1}x)$$

Έτσι το διάστημα εμπιστοσύνης για το $x'\beta$ δίνεται από τον τύπο:

$$x'b \pm z_{\frac{a}{2}}\sqrt{x'(X'VX)^{-1}x}$$

Μια πολύ σημαντική ιδιότητα της λογιστικής παλινδρόμησης θεωρείται η παρακάτω:

$$\frac{\partial P(x_i)}{\partial \beta} = n_i[P(x_i)][1 - P(x_i)]x_i$$

ή γενικότερα

$$\frac{\partial \mu_i}{\partial \beta} = [var(y_i)]x_i$$

Από την παραπάνω σχέση προκύπτει:

$$var\left[\hat{P}(x_i)\right] = [var(y_i)]^2 x_i'(X'VX)^{-1}x_i$$

Έτσι, το διάστημα πρόβλεψης μπορεί να βρεθεί όπως και σε όλα τα γραμμικά μοντέλα.

Καταρχήν

$$\frac{y_i - \hat{P}(x_i)}{n_i[P(x_i)][1 - P(x_i)]\sqrt{1 + x_i'(X'VX)^{-1}x_i}} \sim N(0,1) \alpha\sigma\upsilon\mu\pi\tau\omega\tau\iota\kappa\acute{\alpha}.$$

Επομένως, ένα 100(1-α)% διάστημα εμπιστοσύνης για το $y_i$ μπορεί να βρεθεί από τη σχέση:

$$\hat{P}(x_i) \pm z_{\frac{a}{2}}\{n_i[P(x_i)][1 - P(x_i)]\}\sqrt{1 + x_i'(X'VX)^{-1}x_i}\,, i = 1,2, \dots, m$$

Στην πράξη πρέπει να αντικαταστήσουμε το $\hat{P}(x_i)$ στον πίνακα V.

### 2.3.5 Συμπερασματολογία με χρήση πιθανοφάνειας στη λογιστική παλινδρόμηση

Με τη συμπερασματολογία πιθανοφάνειας, μπορούμε να ενισχύσουμε τον έλεγχο υποθέσεων, χρησιμοποιώντας τη log likelihood. Η χρήση της μοιάζει αρκετά με τη χρήση της αρχής του επιπλέον αθροίσματος τετραγώνων (extra sum of squares principles) των γραμμικών μοντέλων. Για παράδειγμα, στα γραμμικά μοντέλα μπορούμε να χρησιμοποιήσουμε κάτω από τη μηδενική υπόθεση ένα μοντέλο ελαττωμένο (reduced model), δηλαδή η μηδενική υπόθεση θέτει σε ένα υποσύνολο συντελεστών παλινδρόμησης την τιμή μηδέν. Ο έλεγχος χρησιμοποιεί τη διαφορά στο άθροισμα τετραγώνων του σφάλματος:

$$SS_E(reduced) - SS_E(full)$$

Η διαφορά στο άθροισμα τετραγώνων του σφάλματος αντικαθίσταται, στη λογιστική παλινδρόμηση, από τη διαφορά της log πιθανοφάνειας. Ασυμπτωτικά ισχύει:

$$-2\ln\left[\frac{\mathcal{L}(reduced)}{\mathcal{L}(full)}\right]\sim\mathcal{X}^2_{\Delta}$$

όπου

$\mathcal{L}(\ )$ είναι η πιθανοφάνεια και στη περίπτωσή μας ζητάμε την πιθανοφάνεια για το πλήρες και το ελαττωμένο μοντέλο.

η παράμετρος $\Delta$ είναι η διαφορά στον αριθμό των παραμέτρων ανάμεσα στο πλήρες και το ελαττωμένο μοντέλο.

## 2.4 Μηχανές Διανυσματικής Υποστήριξης

Στη μηχανική μάθηση, μηχανές διανυσματικής υποστήριξης (Support Vector Machines-SVMs), είναι μοντέλα μάθησης με επίβλεψη, με τους σχετικούς αλγόριθμους εκμάθησης που αναλύουν τα δεδομένα για να αναγνωρίσουν τα πρότυπα και τα οποία χρησιμοποιούνται για την ταξινόμηση και την ανάλυση παλινδρόμησης. Ένα μοντέλο SVM είναι μια αναπαράσταση του δοκιμαστικού συνόλου (training set) ως σημεία στο χώρο, τα οποία χαρτογραφούνται έτσι ώστε τα δεδομένα των επιμέρους κατηγοριών να χωρίζονται από ένα σαφές κενό που είναι όσο το δυνατόν ευρύτερο. Τα νέα δεδομένα στη συνέχεια αντιστοιχίζονται με το ίδιο διάστημα και προβλέπεται αν ανήκουν σε μια κατηγορία με βάση σε ποια πλευρά του χάσματος θα πέσουν.

Εκτός από την εκτέλεση γραμμικής ταξινόμησης, τα SVMs μπορούν να εκτελέσουν αποτελεσματικά και μια μη γραμμική κατάταξη σύμφωνα με το τέχνασμα του πυρήνα (kernel trick) όπως ονομάζεται, που γίνεται με την έμμεση χαρτογράφηση των εισόδων σε χώρους μεγάλων διαστάσεων.

Πιο συγκεκριμένα, μια μηχανή διανυσματικής υποστήριξης κατασκευάζει ένα υπερεπίπεδο ή σύνολο από υπερεπίπεδα σε έναν υψηλής ή άπειρης διάστασης χώρο, που μπορεί να χρησιμοποιηθεί για την ταξινόμηση, παλινδρόμηση, ή άλλες εργασίες. Διαισθητικά, ένας καλός διαχωρισμός επιτυγχάνεται από το υπερεπίπεδο που έχει τη μεγαλύτερη απόσταση από το πλησιέστερο σημείο των δεδομένων κατάρτισης οποιασδήποτε κατηγορίας (γνωστό ως περιθώριο λειτουργιών-functional margin), δεδομένου ότι σε γενικές γραμμές όσο μεγαλύτερο είναι το περιθώριο τόσο χαμηλότερη είναι η λάθος γενίκευση του ταξινομητή.

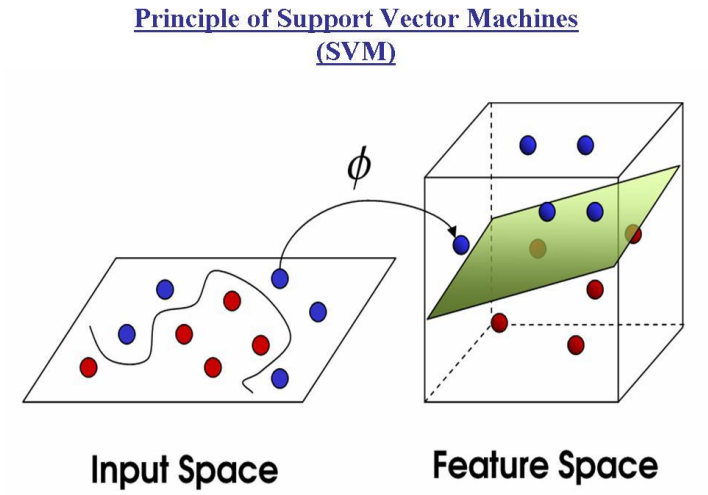
Οι μηχανές διανυσματικής υποστήριξης δημιουργήθηκαν σαν ιδέα από τους Cortes και Vapnik το 2000. Εκπαιδεύονται από την επίλυση ενός περιορισμένου προβλήματος ταξινόμησης και υλοποιούν τη χαρτογράφηση των συντελεστών παραγωγής σε ένα υψηλό τρισδιάστατο χώρο χρησιμοποιώντας ένα σύνολο μη γραμμικών βασικών συναρτήσεων. Η ανάπτυξη της SVM μεθόδου είναι διαφορετική από τους συνήθεις αλγόριθμους που χρησιμοποιούνται για τη μάθηση και παρέχει μια νέα άποψη μάθησης. Τα πιο σημαντικά χαρακτηριστικά της τεχνικής είναι η δυαδικότητα, οι πυρήνες, η κυρτότητα και η σποραδικότητα.

Η τεχνική SVM μπορεί να χρησιμοποιηθεί για την επίλυση διαφόρων προβλημάτων του πραγματικού κόσμου:

- Είναι χρήσιμη για την κατηγοριοποίηση κειμένου αφού η εφαρμογή τους μπορεί να μειώσει σημαντικά την ανάγκη για επισημασμένες

περιπτώσεις κατάρτισης τόσο στο επαγωγικό πρότυπο αλλά και τις μεταβιβαστικές ρυθμίσεις.

- Ταξινόμηση των εικόνων μπορεί επίσης να πραγματοποιηθεί χρησιμοποιώντας SVMs. Τα πειραματικά αποτελέσματα δείχνουν ότι οι SVMs επιτυγχάνουν σημαντικά μεγαλύτερη ακρίβεια αναζήτησης από τα παραδοσιακά συστήματα βελτίωσης.
- Τα SVMs είναι επίσης χρήσιμα στην ιατρική επιστήμη για την ταξινόμηση των πρωτεϊνών με μέχρι 90% από τις ενώσεις που ταξινομούνται, να ταξινομούνται σωστά.
- Επίσης, χειρόγραφοι χαρακτήρες μπορούν να αναγνωριστούν με τη χρήση SVMs.




**Σχήμα 7: Παρατηρούμε τη βασική αρχή στην οποία βασίζονται οι μηχανές διανυσματικής υποστήριξης (SVM). Υποδεικνύει το διαχωριστικό υπερεπίπεδο όταν εφαρμόζεται στο χώρο των χαρακτηριστικών (feature space).**

Η τεχνική SVM είναι μια τεχνική πλήρης επίβλεψης. Γνωστές ετικέτες ελέγχουν αν το σύστημα εκτελείται στο σωστό δρόμο ή όχι. Μια διαδικασία ταξινόμησης περιλαμβάνει τα δεδομένα εκπαίδευσης και τα δεδομένα εξέτασης που αποτελούνται από περιπτώσεις δεδομένων (στιγμιότυπα). Στόχος της τεχνικής αυτής είναι να παράγει ένα μοντέλο που να προβλέπει μία τιμή-στόχο των δεδομένων στο σύνολο δοκιμών. Ένα βήμα από την διαδικασία ταξινόμησης είναι η αναγνώριση, δηλαδή η επιλογή χαρακτηριστικών ή εξαγωγή χαρακτηριστικών.

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

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

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

Το βέλτιστο διαχωριστικό υπερεπίπεδο διαχωρίζει τις δύο κλάσεις και μεγιστοποιεί την απόσταση στο πλησιέστερο σημείο από κάθε κατηγορία (Vapnik, 1996). Όχι μόνο κάνει αυτό δηλαδή παρέχει μία μοναδική λύση στο πρόβλημα του διαχωριστικού υπερεπιπέδου, αλλά με τη μεγιστοποίηση του περιθωρίου μεταξύ των δύο κατηγοριών για τα δεδομένα εκπαίδευσης, οδηγεί σε καλύτερη απόδοση ταξινόμησης για τα δεδομένα δοκιμών.

### 2.4.1 Γραμμικά Διαχωριζόμενα Δεδομένα

Δοθέντος των δεδομένων εκπαίδευσης $D$, ένα σετ από n σημεία της μορφής

$$D = \{(x_i,y_i)|\ x_i \in R^p, y_i\ \in \{-1,1\}\}_{i=1}^{n}$$

*ό*που το $y_i$ παίρνει τις τιμές είτε 1, είτε -1 οι οποίες υποδεικνύουν την κατηγορία στην οποία ανήκει το σημείο $x_i$. Κάθε $x_i$ είναι ένα διάνυσμα p διάστασης πραγματικών αριθμών. Θέλουμε να βρούμε το υπερεπίπεδο μέγιστου περιθωρίου (max margin hyperplane) που χωρίζει τα σημεία που έχουν $y_i$= 1 από εκείνα που έχουν $y_i$= -1. Κάθε υπερεπίπεδο μπορεί να γραφτεί ως το σύνολο των σημείων x που ικανοποιούν την εξίσωση

$$w.x - b = 0.$$

Στην παραπάνω εξίσωση, $w$ είναι το κανονικό διάνυσμα του υπερεπιπέδου, δηλαδή το κάθετο διάνυσμα στο υπερεπίπεδο. Η παράμετρος $\frac{b}{\|w\|}$, καθορίζει την απόσταση του υπερεπιπέδου από την αρχή.

Αν τα δεδομένα εκπαίδευσης είναι γραμμικά διαχωριζόμενα, μπορούμε να επιλέξουμε δύο υπερεπίπεδα που διαχωρίζουν τα δεδομένα και δεν υπάρχουν σημεία ενδιάμεσά τους και στη συνέχεια προσπαθούμε να μεγιστοποιήσουμε τη μεταξύ τους απόσταση. Η περιοχή που περικλείουν ονομάζεται περιθώριο (margin). Αυτά τα υπερεπίπεδα μπορούν να περιγραφούν από τις εξισώσεις

$$w.x - b = 1$$

και

$$w.x - b = -1.$$

Χρησιμοποιώντας τη γεωμετρία, βρίσκουμε ότι η απόσταση των δύο υπερεπιπέδων είναι $\frac{2}{\|w\|}$, άρα πρέπει α ελαχιστοποιηθεί το $\|w\|$. Επίσης, για να αποφύγουμε τα σημεία μας να «πέσουν» στο περιθώριο προσθέτουμε τους ακόλουθους περιορισμούς:

$w.x - b \geq 1$, για $x_i$ στην πρώτη κλάση

ή

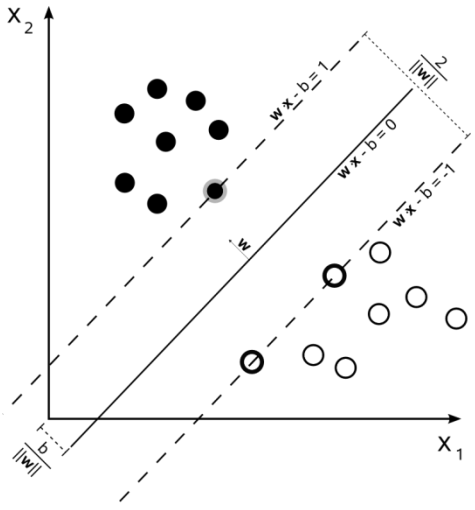
$w.x - b \leq 1$, για $x_i$ στη δεύτερη κλάση.

Αυτοί οι περιορισμοί μπορούν να ξαναγραφτούν μαζί ως

$y_i(w.x - b) \geq 1$, για $1 \leq i \leq n$.

Τέλος, συνδυάζοντας όλα τα παραπάνω παίρνουμε το παρακάτω πρόβλημα βελτιστοποίησης:

$$min\|w\|$$
$$\textit{υπό τον περιορισμό } y_i(w.x - b) \geq 1, \textit{για } 1 \leq i \leq n.$$




**Σχήμα 8: Παρατηρούμε τον διαχωρισμό δύο ομάδων με το διαχωριστικό υπερεπίπεδο με την τεχνική των (SVM). Η απόσταση των δύο υπερεπιπέδων είναι $\frac{2}{\|w\|}$ όπως φαίνεται στο σχήμα και έχουμε τους δύο περιορισμούς $w.x - b \geq 1$ και $w.x - b \leq 1$ έτσι ώστε να αποφύγουμε τα δεδομένα μας να πέσουν στο περιθώριο.**

Ελαχιστοποιώντας το $\|w\|$, είναι ισοδύναμο με την ελαχιστοποίηση του $\frac{1}{2}||w||^2$ και η χρήση αυτής κάνει εφικτή την εκτέλεση βελτιστοποίησης του τετραγωνικού προγραμματισμού (Quadratic Programming Optimization). Τελικά εμείς πρέπει να υπολογίσουμε:

$$min\frac{1}{2}||w||^2$$
$$\textit{υπό τον περιορισμό } y_i(w.x-b)-1 \geq 0, \textit{για } 1 \leq i \leq n.$$

Σε αυτό το πρόβλημα πρέπει να κατανέμουμε στους περιορισμούς τους πολλαπλασιαστές Lagrange α, όπου $\alpha_i \geq 0, \forall i$.

$$\begin{aligned} L_p &= \frac{1}{2}||w||^2 - \alpha[y_i(w.x-b)-1], \forall i \\ &= \frac{1}{2}||w||^2 - \sum_{i=1}^{n}\alpha_i[\,y_i(w.x-b)-1] \\ &= \frac{1}{2}||w||^2 - \sum_{i=1}^{n}\alpha_i[\,y_i(w.x-b) + \sum_{i=1}^{n}a_i \end{aligned}$$

Θέλουμε να βρούμε το w και το b τα οποία ελαχιστοποιούν, και το α, το οποίο μεγιστοποιεί την εξίσωση. Μπορούμε να το κάνουμε αυτό διαφορίζοντας την ως προς το w και το b, και θέτοντας τις παραγώγους ίσες με το μηδέν:

$$\frac{\partial L_p}{\partial w} = 0 \Rightarrow w = \sum_{i=1}^{n}\alpha_i y_i x_i$$
$$\frac{\partial L_p}{\partial b} = 0 \Rightarrow \sum_{i=1}^{n}\alpha_i y_i = 0$$

Αντικαθιστώντας τις δύο τελευταίες στην προηγούμενη εξίσωση, παίρνουμε μία άλλη μορφή η οποία εξαρτάται από το α, και τότε πρέπει να μεγιστοποιήσουμε:

$L_D = \sum_{i=1}^{n} a_\iota - \frac{1}{2}\sum_{i,j} a_i\alpha_j y_i y_j x_i x_j$ , υπό $\alpha_i \geq 0$ *και* $\sum_{i=1}^{n}\alpha_i y_i = 0$

$= \sum_{i=1}^{n} a_\iota - \frac{1}{2}\sum_{i,j} a_i H_{ij}\alpha_j$, όπου $H_{ij} = y_i y_j x_i x_j$

$= \sum_{i=1}^{n} a_\iota - \frac{1}{2}\alpha^T H\alpha$ , υπό $\alpha_i \geq 0$ *και* $\sum_{i=1}^{n}\alpha_i y_i = 0$

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 του πυρήνα, και περιγράφεται παρακάτω. Αφού το πρόβλημα έχει μετατοπιστεί από την ελαχιστοποίηση $L_p$ στη μεγιστοποίηση της $L_D$, θα πρέπει να βρεθεί:

$$max[\sum_{i=1}^{n} a_\iota - \frac{1}{2}\alpha^T H\alpha]$$

$$\text{Υπό } \alpha_i \geq 0 \; και \; \sum_{i=1}^{n} \alpha_i y_i = 0$$

### 2.4.2 Μη Γραμμικά Διαχωριζόμενα Δεδομένα

Για δεδομένα που δεν είναι πλήρως γραμμικά διαχωρίσιμα θα χαλαρώσουμε τους περιορισμούς για να επιτρέπουν ελαφρώς μη ταξινομημένα σημεία. Αυτό γίνεται με την εισαγωγή μιας μεταβλητής χαλάρωσης $\xi_i, i = 1, .., L$ που μετρά το βαθμό της εσφαλμένης ταξινόμησης των δεδομένων $x_i$,

$$x_i . w + b \geq 1 - \xi_i, για \; y_i = 1$$
$$x_i . w + b \geq 1 - \xi_i, για \; y_i = -1$$
$$\xi_i \geq 0 \; \forall i.$$

Αφού συνδυάσουμε όλες τις παραπάνω εξισώσεις:

$y_i(x_i . w + b) - 1 + \xi_i \geq 0, όπου \; \xi_i \geq 0 \; \forall i.$

Σε αυτό το «μαλακό» SVM περιθώριο (soft margin), 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, είναι να βρούμε:

$$min \frac{1}{2}||w||^2 + C \sum_{i=1}^{n} \xi_i, \text{ υπό τον περιορισμό } y_i(x_i . w + b) - 1 + \xi_i \geq 0, \forall i$$

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 $\xi_i$ και να μεγιστοποιηθεί ως προς α:

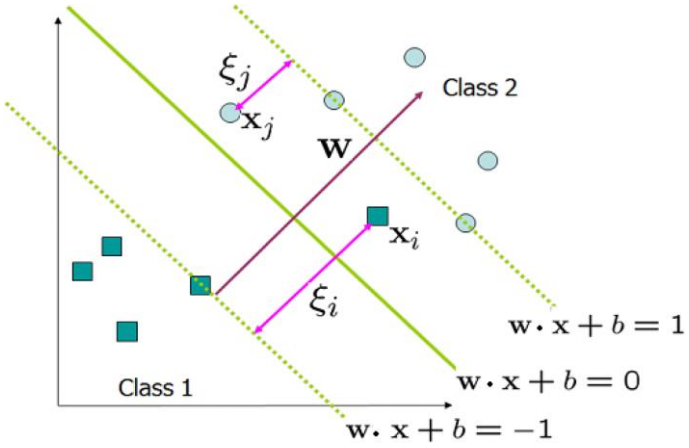
$$L_p = \frac{1}{2}||w||^2 \left| + C \sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \alpha_i [\, y_i(x_i . w + b) - 1 + \xi_i \right] - \sum_{\iota=1}^{n} \mu_i \, \xi_i$$

Διαφορίζοντας την προηγούμενη εξίσωση $L_p$ ως προς το w, το b και το $\xi_i$και θέτοντας τις μερικές παραγώγους ίσες με το μηδέν:

$$\frac{\partial L_p}{\partial w} = 0 \Rightarrow w = \sum_{i=1}^{n} a_\iota y_i x_i$$

$$\frac{\partial L_p}{\partial b} = 0 \Rightarrow \sum_{i=1}^{n} \alpha_i y_i = 0$$

$$\frac{\partial L_p}{\partial \xi_i} = 0 \Rightarrow C = \alpha_i + \mu_i$$




**Σχήμα 9: Γραφικό παράδειγμα μη γραμμικά διαχωρισμένων δεδομένων με την εισαγωγή της μεταβλητής χαλάρωσης $\xi_i, i = 1,..,L$ που μετρά το βαθμό της εσφαλμένης ταξινόμησης των δεδομένων $x_i$.**

Τέλος, αντικαθιστώντας όπως προηγουμένως και συνδυάζοντας τις προηγούμενες εξισώσεις πρέπει να βρούμε:

$$max \sum_{i=1}^{n} \alpha_i - \frac{1}{2} a^T H a$$

υπό τους περιορισμούς $0 \leq \alpha_i \leq C$ και $\sum_{i=1}^{n} \alpha_i y_i = 0$.

### 2.4.3 Μηχανές Διανυσματικής Υποστήριξης για Παλινδρόμηση

Μια άλλη εκδοχή της μεθόδου SVM για παλινδρόμηση προτάθηκε και είναι γνωστή ως διανυσματική υποστήριξη παλινδρόμησης (support vector regression-SVR). Το μοντέλο που δημιουργείται για την ταξινόμηση εξαρτάται μόνο από ένα υποσύνολο του συνόλου δεδομένων, επειδή η συνάρτηση ποινής για τη δημιουργία του μοντέλου δεν ενδιαφέρεται για τα σημεία εκπαίδευσης

που βρίσκονται πέρα από το περιθώριο. Ανάλογα, το μοντέλο που δημιουργείται από το SVR εξαρτάται μόνο από ένα υποσύνολο του συνόλου δεδομένων, επειδή η συνάρτηση ποινής για τη δημιουργία του μοντέλου αγνοεί τα δεδομένα εκπαίδευσης που βρίσκονται κοντά στο μοντέλο πρόβλεψης, εντός ορίου ε.

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:

$$\{x_i, y_i\} \text{ όπου } i = 1, \dots, n, y_i \in R, x_i \in R^D.$$
$$y_i = x_i \,.\, w + b$$

Πιο αναλυτικά, 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 από την πραγματική τιμή $t_i$, δηλαδή αν $|t_i - y_i| < \varepsilon$. Η περιοχή $y_i \pm \varepsilon$ ονομάζεται ε-insensitive σωλήνας (tube). Επίσης, οι μεταβλητές εξόδου που είναι εκτός του σωλήνα, δίνουν μία από τις δύο χαλαρές μεταβλητές ανάλογα αν βρίσκονται πάνω ($\xi^+$) ή κάτω ($\xi^-$) από τον σωλήνα:

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

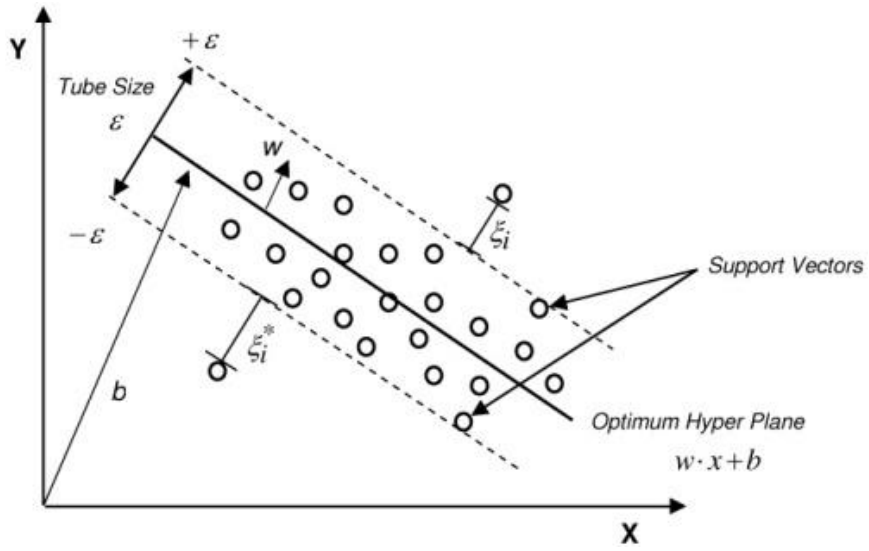
Η συνάρτηση σφάλματος για την SVM παλινδρόμηση μπορεί να γραφεί ως εξής:

$$C \sum_{i=1}^{n} (\xi^+ + \xi^-) + \frac{1}{2} ||w||^2$$

Αυτή η συνάρτηση πρέπει να ελαχιστοποιηθεί υπό τους περιορισμούς $\xi^+ \geq 0$, $\xi^- \geq 0$ και τις εξισώσεις $t_i \leq y_i + \varepsilon + \xi^+$, $t_i \geq y_i - \varepsilon - \xi^+$. Για το λόγο αυτό εισάγουμε και πάλι τους πολλαπλασιαστές Lagrange

$a_i^+ \geq 0, a_i^- \geq 0, \mu_i^+ \geq 0, \mu_i^- \geq 0 \; \forall i.$

Με τη διαδικασία που ακολουθήσαμε στις προηγούμενες υποενότητες δημιουργούμε την $L_p$, διαφορίζουμε ως προς w, b, $\xi^+, \xi^-$ και θέτουμε τις παραγώγους ίσες με μηδέν. Στη συνέχεια βρίσκουμε το $L_D$ και μεγιστοποιούμε ως προς $a_i^+$ και $a_i^-$. Τέλος βρίσκουμε τις παραμέτρους που χρειαζόμαστε.




**Σχήμα 10: Μηχανές Διανυσματικής Υποστήριξης με τις μεταβλητές χαλάρωσης. Ελαχιστοποίησης της $C\sum_{i=1}^{n}(\xi^{+}+\xi^{-})+\frac{1}{2}||w||^{2}$ υπό τους περιορισμούς $\xi^{+}\geq 0$, $\xi^{-}\geq 0$ και τις εξισώσεις $t_i \leq y_i+\varepsilon+\xi^{+}$, $t_i \geq y_i-\varepsilon-\xi^{+}$.**

### 2.4.4 Η μέθοδος των πυρήνων

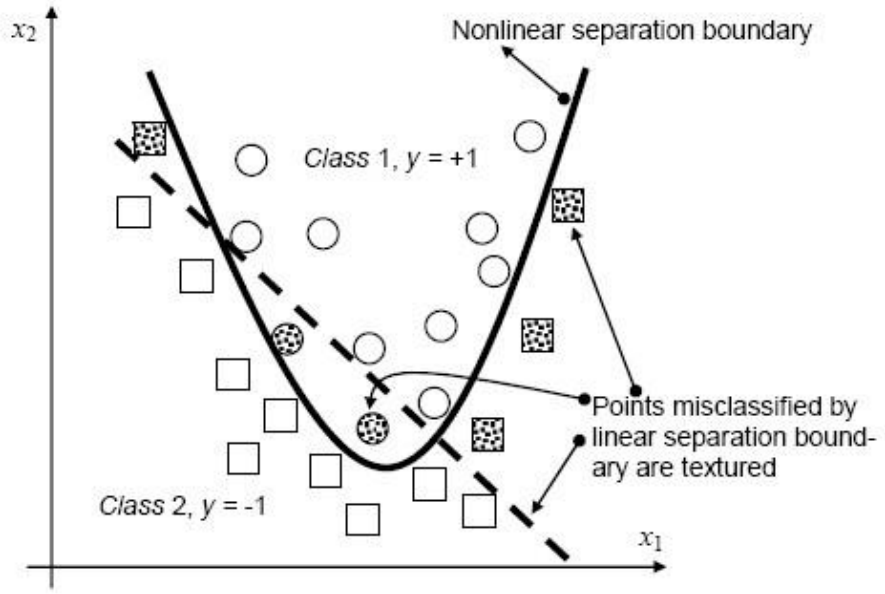
Οι μέθοδοι των πυρήνων είναι μία πολύ δημοφιλής και επιτυχημένη περιοχή της μηχανικής μάθησης. Η κοινή βάση τους είναι το αποκαλούμενο τέχνασμα του πυρήνα (kernel trick), το οποίο μπορεί να εφαρμοστεί σε οποιοδήποτε γραμμικό αλγόριθμο ο οποίος βασίζεται μόνο στα δεδομένα από την άποψη των εσωτερικών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

Για τους αλγόριθμους μηχανικής μάθησης, το τέχνασμα του πυρήνα είναι ένας τρόπος χαρτογράφησης παρατηρήσεων από ένα γενικό σύνολο S σε ένα εσωτερικό προϊόν χώρου V (εξοπλισμένη με το φυσικό πρότυπο της), χωρίς να χρειάζεται να υπολογίσει τη χαρτογράφηση ρητά, επειδή οι παρατηρήσεις θα αποκτήσουν γραμμική δομή στο χώρο V. Οι γραμμικές ταξινομήσεις στο χώρο V είναι ισοδύναμες με γενικές ταξινομήσεις στο χώρο S. Το τέχνασμα ή η μέθοδος που χρησιμοποιείται για να αποφευχθεί η ρητή χαρτογράφηση είναι η χρήση αλγορίθμων μάθησης που απαιτούν μόνο γινόμενα μεταξύ των φορέων στο χώρο V, για να επιλεγεί η χαρτογράφηση, έτσι ώστε τα δεδομένα υψηλών διαστάσεων να μπορεί να υπολογιστούν εντός του αρχικού χώρου, με τη βοήθεια μιας συνάρτησης πυρήνα.

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

$$H_{ij} = y_i y_j K(x_i, x_j) = x_i \,.\, x_j = x_i^T x_j$$

Η $K(x_i, x_j)$ είναι ένα παράδειγμα μιας οικογένειας συναρτήσεων που ονομάζονται συναρτήσεις πυρήνα (kernel functions) (ο $K(x_i, x_j)$ είναι γνωστός ως γραμμικός πυρήνας). Στο σύνολο των συναρτήσεων του πυρήνα όλα βασίζονται στον υπολογισμό εσωτερικών γινομένων των δύο διανυσμάτων.




**Σχήμα 4: Μη γραμμικό διαχωριστικό σύνορο με τη βοήθεια της μεθόδου των πυρήνων. Χρησιμοποιήθηκε ο πολυωνιμικός πυρήνας $K(x_i, x_j) = (x_i . x_j + a)^b$.**

Μερικοί δημοφιλείς πυρήνες για την ταξινόμηση και παλινδρόμηση είναι:

*Ο γραμμικός πυρήνας (Linear Kernel)*

$$K(x_i, x_j) = x_i . x_j$$

*Ο πολυωννυμικός πυρήνας (Polynomian Kernel)*

$$K(x_i, x_j) = (x_i . x_j + a)^b$$

Είναι μια συνάρτηση πυρήνα που χρησιμοποιείται πιο συχνά με μηχανές διανυσματικής υποστήριξης που αναπαριστά την ομοιότητα διανυσμάτων σε ένα χώρο χαρακτηριστικών από τα πολυώνυμα των αρχικών μεταβλητών, επιτρέποντας τη μάθηση των μη γραμμικών μοντέλων.

*Ο σιγμοειδής πυρήνας (Neural Kernel)*

$$K(x_i, x_j) = tanh(ax_i . x_j - b)$$

*Ακτινική Βάση Πυρήνα (Gaussian Radial Basis Kernel)*

$$K(x_i, x_j) = e^{\frac{||x_i - x_j||^2}{2\sigma^2}}$$

Ένα σύνολο δεδομένων το οποία δεν είναι γραμμικά διαχωρίσιμο σε δισδιάστατο χώρο δεδομένων x μπορεί να διαχωριστεί στο μη γραμμικό χώρο των χαρακτηριστικών έμμεσα από αυτή τη μη- γραμμική συνάρτηση πυρήνα.

Γενικά υπάρχουν πολλές συναρτήσεις πυρήνα. Υπάρχουν όμως και συγκεκριμένες απαιτήσεις για μία συνάρτηση έτσι ώστε να μπορεί να χρησιμοποιηθεί ως συνάρτηση του πυρήνα.

### 2.4.5 Επιλογή μοντέλου- Επιλογή 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.

Η αναζήτηση πλέγματος (grid search) αναφέρεται σε μία εξαντλητική αναζήτηση μέσω ενός υποσυνόλου του παραμετρικού (hyperparameter) χώρου του αλγορίθμου μάθησης για να λύσει το πρόβλημα της επιλογής μοντέλου (model selection) ή της βελτιστοποίησης των παραμέτρων (hyperparameter). Ένας αλγόριθμος αναζήτησης του δικτύου θα πρέπει να καθοδηγείται από κάποια μετρική απόδοση, μετρούμενη με την διασταυρωμένη επικύρωση (cross validation) σε ένα σύνολο εκπαίδευσης.

Εφόσον η έρευνα του δικτύου είναι μια εξαντλητική μέθοδος, και ως εκ τούτου είναι ενδεχομένως μια δαπανηρή μέθοδος, έχουν προταθεί πολλές εναλλακτικές λύσεις.

Οι Zhu et al. (2004), κάνουν εισαγωγή ενός ενιαίου σχεδιασμού (UD, uniform design) και της μεθόδου παλινδρόμησης των μηχανών διανυσματικής υποστήριξης (SVR) για τη μείωση του κόστους υπολογισμού της παραδοσιακής μεθόδου GS. Έτσι το όρια του σφάλματος (error bounds) της SVM υπολογίζονται μόνο σε ορισμένους κόμβους που επιλέγονται από τη μέθοδο UD. Τότε η μέθοδος παλινδρόμησης SVM (SVR), εκπαιδεύεται από τα αποτελέσματα που υπολογίστηκαν. Στη συνέχεια, οι τιμές του ορίου σφάλματος της SVM εκτιμώνται σε άλλους κόμβους από τη συνάρτηση SVR και οι βελτιστοποιημένες παράμετροι μπορούν να επιλεγούν με βάση τα εκτιμώμενα αποτελέσματα.

Οι Chapelle et al. (2002) ανέπτυξαν την μέθοδο των κλίσεων (gradient descent method) που βασίζεται σε ορισμένα όρια σφάλματος των SVM, όπως το όριο RM (RM bound). Ωστόσο, το όριο RM είναι ανακριβές σε ορισμένες περιπτώσεις και οι αρχικές τιμές των παραμέτρων έχουν ισχυρές επιδράσεις σχετικά με τις προκύπτουσες παραμέτρους και την αναζήτηση της αποτελεσματικότητας.

Διάφορες πρόσφατες μελέτες έχουν αναφέρει ότι η SVM (μηχανή διανυσματικής υποστήριξης) τεχνική είναι γενικά ικανή να παρέχει υψηλότερες επιδόσεις όσον αφορά την ακρίβεια ταξινόμησης, απ' ότι οι άλλοι αλγόριθμοι ταξινόμησης δεδομένων. Ωστόσο, για ορισμένα σύνολα δεδομένων, η απόδοση της SVM είναι πολύ ευαίσθητη στο πώς επιλέγονται, η παράμετρος του κόστους και οι παράμετροι του πυρήνα. Ως εκ τούτου, ο χρήστης πρέπει κανονικά να διεξάγει εκτεταμένες διαδικασίες cross validation, προκειμένου να υπολογίσει την βέλτιστη ρύθμιση παραμέτρων. Η διαδικασία αυτή αναφέρεται συνήθως ως επιλογή μοντέλου. Ένα πρακτικό πρόβλημα με την επιλογή του μοντέλου είναι ότι αυτή η διαδικασία είναι πολύ χρονοβόρα. Για παράδειγμα, εάν η διαδικασία επιλογής μοντέλου που εγκρίθηκε να χρησιμοποιηθεί για την κατασκευή μιας SVM για ένα σύνολο δεδομένων, είναι το πλήρες πλέγμα αναζήτησης διαδικασίας επιλογής μοντέλου (the complete grid-search model selection process) θα λάβει πολλές ώρες σε έναν υπολογιστή. Δεδομένου ότι το σύνολο δεδομένων δεν θεωρείται μεγάλο, το πώς να επιταχυνθεί η διαδικασία επιλογής του μοντέλου για τις SVM γίνεται ένα κρίσιμο ζήτημα και έχουν διεξαχθεί διάφορες μελέτες για την αντιμετώπιση αυτού του ζητήματος κατά τα τελευταία έτη. Οι μελέτες αυτές μοιράζονται ένα κοινό έδαφος με στόχο τη μείωση του χώρου αναζήτησης στους συνδυασμούς παραμέτρων.

Στην εργασία των G.Lebrun et. al (2006) προτείνεται μια νέα μέθοδος μάθησης για την κατασκευή μιας δίτιμης συνάρτησης αποφάσεων (Binary Decision function (BDF)) στις Διανυσματικές μηχανές υποστήριξης (SVMs) μειώνοντας την πολυπλοκότητα και καθιστώντας αποτελεσματική τη γενίκευση. Στόχος είναι η κατασκευή ενός γρήγορου και αποτελεσματικού SVM ταξινομητή. Ορίζεται ένα κριτήριο για την αξιολόγηση της ποιότητας της συνάρτησης αποφάσεων (DFQ, Decision function Quality), η οποία λαμβάνει υπ' όψιν το ποσοστό αναγνώρισης και την πολυπλοκότητα της BDF. Για την απλοποίηση του συνόλου εκπαίδευσης χρησιμοποιείται Vector Quantization (VQ). Η επιλογή μοντέλου γίνεται με βάση την επιλογή του απλούστερου επιπέδου, ενός υποσυνόλου χαρακτηριστικών και των παραμέτρων του SVM (hyperparameters) και εκτελείται για την βελτιστοποίηση της DFQ. Ο χώρος όπου γίνεται η αναζήτηση για την επιλογή του καλύτερου μοντέλου είναι τεράστιος, έτσι χρησιμοποιείται ο Tabu Search (TS) για να βρεθεί ένα καλό υπο-βέλτιστο μοντέλο σε ευάγωγες περιπτώσεις.

Γενικά, η κακή επιλογή των ρυθμιστικών παραμέτρων μπορεί να μειώσει δραματικά την απόδοση των SVMs. Το πρόβλημα της επιλογής μιας καλής ρυθμιστικής παραμέτρου για μια καλύτερη ικανότητα στη γενίκευση είναι το λεγόμενο πρόβλημα επιλογής μοντέλου (model selection). Θα ήταν επιθυμητό να έχουμε ένα αποτελεσματικό και αυτόματο σύστημα επιλογής μοντέλου κάνοντας έτσι τα SVMs πρακτικά σε εφαρμογές της πραγματικής ζωής, ιδιαίτερα, για τους ανθρώπους που δεν είναι εξοικειωμένοι με τις παραμέτρους (parameters tuning) στα SVMs. Είναι ένας από τους πιο ελπιδοφόρους αλγορίθμους μάθησης για την ταξινόμηση καθώς και για την παλινδρόμηση και πρέπει να γίνουν ακόμη σημαντικές έρευνες έτσι ώστε να είναι πάντα αποτελεσματικοί.

# ΚΕΦΑΛΑΙΟ 3

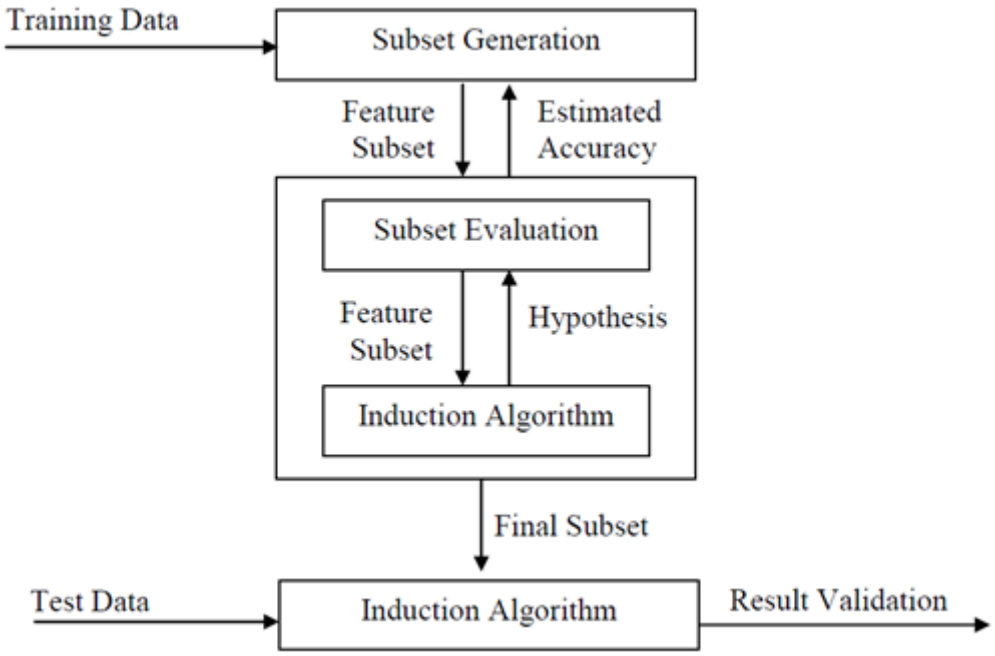
## Feature Selection

### (Επιλογή Χαρακτηριστικών)

Επιλογή χαρακτηριστικών (Feature Selection) είναι ένας όρος που χρησιμοποιείται συνήθως στον τομέα της εξόρυξης δεδομένων (Data Mining) για να περιγράψει τα εργαλεία και τις τεχνικές που διατίθενται για τη μείωση των εισόδων σε ένα μέγεθος για επεξεργασία και ανάλυση.

Η επιλογή χαρακτηριστικών έχει γίνει το επίκεντρο πολλών ερευνών σε τομείς εφαρμογής στους οποίους τα σύνολα δεδομένων περιλαμβάνουν δεκάδες, εκατοντάδες, χιλιάδες μεταβλητές διαθέσιμες μεταβλητές.

Στην εκμάθηση μηχανών (machine learning) και στην Στατιστική, η επιλογή χαρακτηριστικών (feature selection), γνωστή και ως μεταβλητή επιλογή(variable selection, attribute selection, variable subset selection) είναι η διαδικασία της επιλογής ενός υποσυνόλου των σχετικών χαρακτηριστικών για την κατασκευή του μοντέλου. Η κεντρική ιδέα της επιλογής χαρακτηριστικών είναι η επιλογή ενός υποσυνόλου από τις εισαγόμενες μεταβλητές εξαλείφοντας χαρακτηριστικά με μικρές ή καθόλου προγνωστικές πληροφορίες.




### 3.1 Η μέθοδος Επιλογής Χαρακτηριστικών

Οι τεχνικές επιλογής χαρακτηριστικών περιλαμβάνουν τρία βασικά πλεονεκτήματα, όσο αφορά την κατασκευή ενός μοντέλου:

- Βελτιώνουν την απόδοση πρόβλεψης των παραγόντων πρόβλεψης.
- Παρέχει ταχύτερα πιο αποδοτικούς προγνωστικούς παράγοντες.
- Παρέχουν καλύτερη κατανόηση της υποκείμενης διαδικασίας που παρήγαγε τα δεδομένα.

Από τις τεχνικές επιλογής χαρακτηριστικών προκύπτουν πολλά προνόμια, όπως η διευκόλυνση στην οπτικοποίηση δεδομένων και στην κατανόησή τους, μειώνουν τις απαιτήσεις μέτρησης και αποθήκευσης δεδομένων, μειώνουν τους χρόνους κατάρτισης και αξιοποίησης, αψηφούν το πρόβλημα των μεγάλων διαστάσεων για τη βελτίωση της απόδοσης πρόβλεψης.

Ένας αλγόριθμος επιλογής χαρακτηριστικών μπορεί να θεωρηθεί ως ο συνδυασμός της τεχνικής αναζήτησης για να προτείνει νέα υποσύνολα χαρακτηριστικών, μαζί με ένα μέτρο αξιολόγησης που βάζει στα διαφορετικά υποσύνολα. Ο απλούστερος αλγόριθμος είναι να δοκιμαστεί κάθε δυνατό υποσύνολο των χαρακτηριστικών βρίσκοντας το ένα το οποίο ελαχιστοποιεί το ποσοστό σφάλματος. Αυτή είναι μια εξαντλητική αναζήτηση του χώρου, και είναι υπολογιστικά δυσεπίλυτο για όλους, εκτός από το μικρότερο χαρακτηριστικό σετ. Η επιλογή των μετρικών αξιολόγησης επηρεάζει σε μεγάλο βαθμό τον αλγόριθμο. Οι μετρήσεις αξιολόγησης, οι οποίες διακρίνουν τις τρεις

βασικές κατηγορίες χαρακτηριστικών των αλγορίθμων επιλογής: τα περιτυλίγματα, τα φίλτρα και τις ενσωματωμένες μεθόδους.

Όσο αφορά την Στατιστική, ο καλύτερος αλγόριθμος επιλογής χαρακτηριστικών είναι stepwise regression. Ο αλγόριθμος αυτός προσθέτει το καλύτερο χαρακτηριστικό (ή διαγράφει το χειρότερο) σε κάθε εκτέλεση. Το κύριο θέμα του αλγορίθμου είναι να οριστεί πότε να σταματήσει. Στην μηχανική μάθηση αυτό γίνεται με cross validation. Στην Στατιστική ορισμένα κριτήρια έχουν βελτιωθεί.

Η επιλογή υποομάδας, αξιολογεί ένα υποσύνολο των χαρακτηριστικών ως ομάδα για την καταλληλότητα του. Οι αλγόριθμοι επιλογής υποσυνόλου μπορούν να διασπαστούν σε Συσκευαστές (Wrappers), Φίλτρα (Filters) και Ενσωματοτές (Embedded).

- Οι συσκευαστές χρησιμοποιούν έναν αλγόριθμο αναζήτησης, ο οποίος αναζητεί μέσα από το χώρο των πιθανών χαρακτηριστικών και ταυτόχρονα αξιολογεί κάθε υποσύνολο εκτελώντας ένα μοντέλο για το υποσύνολο.
- Τα περιτυλίγματα μπορεί να είναι υπολογιστικά ακριβά και έχουν κίνδυνο πάνω τοποθέτηση στο μοντέλο.
- Φίλτρα είναι παρόμοια με τους Συσκευαστές στην προσέγγιση αναζήτησης, αλλά αντί να αξιολογηθεί έναντι ενός μοντέλου, ένα απλούστερο φίλτρο αξιολογείται.

Γενικά, αξιολογούν επαναληπτικά ένα υποψήφιο υποσύνολο των χαρακτηριστικών, στη συνέχεια τροποποιεί το υποσύνολο και αξιολογεί αν το νέο υποσύνολο είναι βελτιωμένο σε σχέση με το παλιό. Η αξιολόγηση των υποσυνόλων απαιτεί βαθμολόγηση των τάξεων ενός υποσυνόλου των χαρακτηριστικών. Εξαντλητική έρευνα είναι γενικά ανέφικτη, οπότε κάποιο κριτήριο διακοπής (implementor) ορίζεται το σημείο ώστε να σταματήσει ο αλγόριθμος και το υποσύνολο των χαρακτηριστικών με την υψηλότερη βαθμολογία που ανακαλύφθηκε μέχρι εκείνο το σημείο έχει επιλεγεί ως το ικανοποιητικό υποσύνολο . Το κριτήριο διακοπής ποικίλλει ανάλογα με τον αλγόριθμο.

Υπάρχουν πολλά κριτήρια βελτιστοποίησης που χρησιμοποιούνται για να διαχειριστούν την επιλογή χαρακτηριστικών σε ένα σύνολο δεδομένων. Από τα πιο παλιά είναι το στατιστικό κριτήριο Cp-Mallows's και το Akaike information criterion (AIC). Αυτά τα κριτήρια προσθέτουν μεταβλητές όσο η p-value της κατανομής student (t) είναι μεγαλύτερη από $\sqrt{2}$. Άλλα κριτήρια είναι τα Bayesian Information Criterion (BIC), Minimum description length (MDL), Bonnferoni/RIC, maximum dependency feature selection και πολλά άλλα.

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. 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.

## 3.2 Κατάταξη Μεταβλητής

Πολλοί αλγόριθμοι επιλογής χαρακτηριστικών/μεταβλητών περιλαμβάνουν κατάταξη μεταβλητών ως κύριο ή βοηθητικό μηχανισμό επιλογής λόγω της απλότητας, επεκτασιμότητα και καλής εμπειρικής επιτυχίας. Αρκετές εργασίες σε αυτό ζήτημα χρησιμοποιούν κατάταξης μεταβλητής ως βασική μέθοδο (e.g., Bekkerman et al., 2003, Caruana and de Sa, 2003, Forman, 2003, Weston et al., 2003).

Κατάταξη μεταβλητής δεν χρησιμοποιείται απαραίτητα για την κατασκευή προγνωστικών παραγόντων. Μία από τις κοινές χρήσεις της στον τομέα ανάλυσης μικροσυστοιχιών είναι να ανακαλύψει πού οδηγεί μια σειρά από θεραπείες: Ένα κριτήριο κατάταξης χρησιμοποιείται για να βρουν γονίδια που εισάγουν διακρίσεις μεταξύ των υγιών και ασθενών της νόσου. Τέτοια γονίδια μπορεί για παράδειγμα να κωδικοποιούν πρωτεΐνες που μπορούν οι ίδιες να χρησιμοποιηθούν ως φάρμακα. Θεωρούμε ότι σε αυτά τα κριτήρια, το τμήμα κατάταξης ορίζεται για τις μεμονωμένες μεταβλητές, ανεξάρτητα από το πλαίσιο των άλλων.

## 3.3 Αρχές της Μεθόδου και Συμβολισμοί

Θεωρούμε μια σειρά από m παραδείγματα $\{x_k, y_k\}\ k = 1, \dots, m$ που αποτελείται από n μεταβλητές εισόδου $x_k$ και μια μεταβλητή εξόδου $y_k$. Η κατάταξη μεταβλητής κάνει χρήση μιας συνάρτησης βαθμολόγησης S(i) η οποία υπολογίζεται από τα $x_k$ και $y_k$. Κατά συνθήκη, υποθέτουμε ότι μια υψηλή βαθμολογία είναι ενδεικτική μιας πολύτιμης μεταβλητής και ότι οι μεταβλητές ταξινομούνται κατά φθίνουσα σειρά S(i). Για να χρησιμοποιηθεί η κατάταξη μεταβλητής για να χτίσει παράγοντες πρόβλεψης, ένθετα υποσύνολα ενσωματώνουν προοδευτικά όλο και περισσότερες μεταβλητές φθίνοντος ενδιαφέροντος ορίζονται.

### 3.3.1 Κριτήρια Συσχέτισης

Ας εξετάσουμε πρώτα την πρόβλεψη ενός συνεχούς αποτελέσματος y. Ο συντελεστής συσχέτισης του Pearson ορίζεται ως εξής:

$$\mathcal{R}(i) = \frac{cov(X_i, Y)}{\sqrt{var(X_i)var(Y)}}$$

Η εκτίμηση του R(i) δίνεται από τον τύπο:

$$R(i) = \frac{\sum_{k=1}^{m}(x_{k,i} - \bar{x}_i)(y_k - \bar{y})}{\sqrt{\sum_{k=1}^{m}(x_{k,i} - \bar{x}_i)^2 \sum_{k=1}^{m}(y_k - \bar{y})^2}}$$

Κριτήρια συσχέτισης όπως είναι το R(i) μπορούν να ανιχνεύσουν μόνο γραμμικές εξαρτήσεις μεταξύ μεταβλητών και στόχου. Ένας απλός τρόπος για να εξαλείψουμε αυτόν τον περιορισμό είναι να κάνει μια μη γραμμική προσαρμογή του στόχου με ενιαίες μεταβλητές και την κατάταξη σύμφωνα με την προσαρμογή που ταιριάζει. Λόγω του ρίσκου της υπερπροσαρμογής κάνουμε μια μη γραμμική προεπεξεργασία και στη συνέχεια χρησιμοποιούμε έναν απλό συντελεστή συσχέτισης.

### 3.3.2 Θεωρητική Πληροφορία Κριτηρίων Κατάταξης

Αρκετές προσεγγίσεις στο πρόβλημα επιλογής μεταβλητής χρησιμοποιώντας θεωρητική πληροφορία κριτηρίων έχουν προταθεί. Πολλοί βασίζονται σε εμπειρικές εκτιμήσεις της αμοιβαίας πληροφορίας μεταξύ κάθε μεταβλητής και του στόχου:

$$I(i) = \int_{x_i} \int_{y} p(x_i, y) log \frac{p(x_i, y)}{p(x_i)p(y)} dxdy$$

όπου $p(x_i)$ και $p(y)$ είναι οι πυκνότητες πιθανότητας των $x_i$ και $y$ αντίστοιχα. Το κριτήριο I(i) είναι ένα μέτρο εξάρτησης μεταξύ της πυκνότητας της μεταβλητής $x_i$ και του στόχου $y$.

Η δυσκολία είναι ότι οι πυκνότητες $p(x_i)$, $p(y)$ και $p(x_i, y)$ είναι άγνωστες και είναι δύσκολο να υπολογιστούν από τα δεδομένα. Στην περίπτωση διακριτών μεταβλητών είναι ευκολότερο γιατί το ολοκλήρωμα γίνεται άθροισμα:

$$I(i) = \sum_{x_i} \sum_{y} P(X = x_i, Y = y)\, log \frac{P(X = x_i, Y = y)}{P(X = x_i)P(Y = y)}$$

## 3.4 Feature construction και Feature Selection/Reduction

Η ουσία και η χρησιμότητα της επιλογής χαρακτηριστικών μπορεί να συνοψισθεί στην επόμενη φράση:

*«Δεδομένου ενός αριθμού χαρακτηριστικών, με ποίον τρόπο μπορεί κάποιος να επιλέξει τα πιο σημαντικά από αυτά, ώστε να μειώσει τον αριθμό τους και ταυτόχρονα να διατηρεί όσο το δυνατό περισσότερη από την πληροφορία που αυτά φέρουν για τη διάκριση μεταξύ των κλάσεων;»*

Feature Selection/Reduction

Η διαδικασία αυτή είναι γνωστή ως επιλογή ή μείωση χαρακτηριστικών (feature selection/reduction) και είναι εξαιρετικά σημαντική και κρίσιμη για ένα σύστημα ταξινόμησης. Εάν επιλέξουμε χαρακτηριστικά με μικρή διακριτική ισχύ/ικανότητα (discrimination power), η σχεδίαση θα οδηγήσει σε ένα σύστημα με πολύ χαμηλή απόδοση, ενώ αν επιλεγούν χαρακτηριστικά που φέρουν μεγάλη ποσότητα πληροφορίας η σχεδίαση απλοποιείται σε πολύ μεγάλο βαθμό. Τελικά σε μια πιο ποιοτική περιγραφή του προβλήματος, θα μπορούσαμε να πούμε πως επιθυμούμε την επιλογή χαρακτηριστικών που οδηγούν σε μεγάλη απόσταση μεταξύ των κλάσεων και μικρή διακύμανση εντός των κλάσεων.

Feature construction

Η τέχνη της μηχανικής μάθησης ξεκινά με το σχεδιασμό των κατάλληλων αναπαραστάσεων δεδομένων. Καλύτερη απόδοση επιτυγχάνεται συχνά με τη χρήση χαρακτηριστικών που προέρχονται από την αρχική είσοδο. Χτίζοντας μια αναπαράσταση χαρακτηριστικών είναι μια ευκαιρία για να ενσωματωθεί η γνώση των δεδομένων και μπορεί να έχει πολύ συγκεκριμένη εφαρμογή.

Δύο διαφορετικοί στόχοι μπορούν να επιδιωχθούν για τη δόμηση χαρακτηριστικών: η επίτευξη της καλύτερης ανακατασκευής των δεδομένων ή να είναι πιο αποτελεσματική για την πραγματοποίηση προβλέψεων. Το πρώτο πρόβλημα είναι το πρόβλημα μάθησης χωρίς επίβλεψη. Είναι στενά συνδεδεμένο με εκείνο της συμπίεσης των δεδομένων και πολλοί αλγόριθμοι χρησιμοποιούνται και στα δύο αυτά πεδία. Το δεύτερο είναι εκείνο της μάθησης με επίβλεψη.

Έχουν προταθεί πολλοί τρόποι για την αντιμετώπιση του προβλήματος της δόμησης χαρακτηριστικών (feature construction).

### 3.4.1 Ομαδοποίηση (Clustering)

Η ομαδοποίηση έχει χρησιμοποιηθεί από καιρό για τη δόμηση χαρακτηριστικών. Η βασική ιδέα είναι να αντικαταστήσει μια ομάδα από «όμοιες» μεταβλητές με ένα σύμπλεγμα κέντρου βάρους, που γίνεται ένα χαρακτηριστικό. Οι πιο δημοφιλείς αλγόριθμοι περιλαμβάνουν K-means και ιεραρχική ομαδοποίηση.

Η ομαδοποίηση συνδέεται συνήθως με την ιδέα της εκμάθησης χωρίς επίβλεψη. Μπορεί να είναι χρήσιμο να εισάγουν κάποια εποπτεία στη διαδικασία ομαδοποίησης για να λάβει περισσότερα διακριθέντα χαρακτηριστικά. Αυτή είναι η ιδέα της αναδιανεμητικής ομαδοποίησης (Pereira et al., 1993).

Εφαρμογές επεξεργασίας κειμένου είναι οι συνήθεις στόχοι για αυτές τις τεχνικές. Τα πρότυπα είναι πλήρη έγγραφα και οι μεταβλητές προέρχονται από μια αναπαράσταση «bag-of-words»: Κάθε μεταβλητή συνδέεται σε μια λέξη και είναι ανάλογη προς το κλάσμα των εγγράφων στα οποία εμφανίζεται η λέξη. Στην εφαρμογή της δόμησης χαρακτηριστικών, οι μέθοδοι ομαδοποίησης ομαδοποιούν λέξεις, όχι έγγραφα. Στις εργασίες κατηγοριοποίησης κειμένου, η επίβλεψη προέρχεται από τη γνώση των κατηγοριών εγγράφου. Εισάγεται με την αντικατάσταση διανυσμάτων μεταβλητών που περιέχουν την συχνότητα έγγραφων με το μικρότερο διάνυσμα μεταβλητών που περιέχει τη συχνότητα κατηγορίας εγγράφων, δηλαδή, οι λέξεις που εκπροσωπήθηκαν ως διανομές πάνω στις κατηγορίες εγγράφων.

### 3.4.2 Παραγοντοποίηση Πίνακα (Matrix Factorization)

Μια άλλη ευρέως χρησιμοποιούμενη μέθοδος δόμησης χαρακτηριστικών είναι η μοναδική αξία αποσύνθεσης (Singular Value Decomposition SVD). Ο στόχος της SVD είναι να σχηματίσει μια σειρά από χαρακτηριστικά που είναι γραμμικοί συνδυασμοί των αρχικών μεταβλητών, οι οποίες παρέχουν την καλύτερη δυνατή αποκατάσταση των αρχικών δεδομένων υπό την έννοια των ελαχίστων τετραγώνων (Dudaet al., 2001). Είναι μια μέθοδος χωρίς επίβλεψη δόμησης χαρακτηριστικών. Τα πιο κατατοπιστικά χαρακτηριστικά γνωρίσματα που εξάγονται από επίλυση ενός προβλήματος βελτιστοποίησης που παρακολουθεί την ανταλλαγή μεταξύ ανασυγκρότησης δεδομένων και συμπίεσης. Τα χαρακτηριστικά που βρέθηκαν ως πολλαπλασιαστές Lagrange του στόχου βελτιστοποιούνται. Μη αρνητικοί πίνακες P διαστάσεων $m \times n$ εκπροσωπούν την από κοινού διανομή των δύο τυχαίων μεταβλητών. Τα χαρακτηριστικά γνωρίσματα που εξάγονται από τις θεωρητικές πληροφορίες I-προβολές, αποδίδουν έναν ανακατασκευασμένο πίνακα στην ειδική εκθετική μορφή

$\check{P} = \frac{1}{Z} e^{\Phi Y}$. Για ένα σετ από d χαρακτηριστικά, Φ είναι ένας $m \times (d+2)$ πίνακας του οποίου η (d+1)-στήλη αποτελείται από 1, Y είναι ένας $(d+2) \times n$ πίνακας του οποίου η (d+2)-στήλη αποτελείται από 1 και Z είναι η σταθερά κανονικοποίησης. Παρόμοια με τα SVD, η λύση δείχνει τη συμμετρία του προβλήματος σε σχέση με τα πρότυπα και τις μεταβλητές.

### 3.4.3 Επιλογή Χαρακτηριστικών με Επίβλεψη (Supervised Feature Selection)

Εξετάζουμε τρεις προσεγγίσεις για την επιλογή των χαρακτηριστικών στις περιπτώσεις κατά τις οποίες πρέπει να διακρίνονται χαρακτηριστικά από μεταβλητές επειδή και οι δύο εμφανίζονται ταυτόχρονα στο ίδιο σύστημα:

- Nested-μέθοδος Επιλογής Υποσυνόλου
- Μέθοδοι Φίλτρου
- Άμεση Αντικειμενική Βελτιστοποίηση

## 3.5 Χώρος Χαρακτηριστικών

Πριν προχωρήσουμε στην ανάλυση μεθόδων επιλογής χαρακτηριστικών θα πρέπει να δοθεί ένας σύντομος ορισμός για το τι είναι ο χώρος χαρακτηριστικών. Ο χώρος χαρακτηριστικών (feature space) στην αναγνώριση προτύπων είναι ο χώρος όπου κάθε δείγμα προτύπου αναπαρίσταται σαν ένα σημείο ενός N-διάστατου χώρου. Η διάσταση N συνδέεται με τον αριθμό των χαρακτηριστικών που χρησιμοποιούνται για την επιλογή του μοντέλου μας. Παρόμοια δείγματα συγκεντρώνονται μεταξύ τους, πράγμα το οποίο επιτρέπει τη χρήση της εκτίμησης πυκνότητας για την αναγνώριση των μοντέλων. Η επιλογή χαρακτηριστικών για τη δημιουργία εύρωστων μοντέλων εκμάθησης, τα οποία βρίσκουν εφαρμογή κυρίως στην εκμάθηση μηχανών. Με την απομάκρυνση άσχετων και πλεοναζόντων χαρακτηριστικών από τα δεδομένα μας επιτυγχάνουμε τη βελτίωση της απόδοσης του συστήματός μας και βοηθούμαστε να κατανοήσουμε καλύτερα τα δεδομένα, καταλαβαίνοντας ποια είναι τα πιο σημαντικά χαρακτηριστικά και πώς αυτά συσχετίζονται μεταξύ τους.

## 3.6 Επιλογή Υποσυνόλου Χαρακτηριστικών

Παρακάτω αναφέρουμε μερικούς τρόπους επιλογής χαρακτηριστικών:

Βαθμωτή Επιλογή Χαρακτηριστικών:
Σε αυτή την περίπτωση τα χαρακτηριστικά αντιμετωπίζονται ξεχωριστά. Μπορούν να υιοθετηθούν οποιαδήποτε από τα μέτρα διαχωρισιμότητας κλάσεων. Υπολογίζουμε την τιμή ενός κριτηρίου C(k) για κάθε ένα χαρακτηριστικό k=1,2, …, m. Στη συνέχεια κατατάσσονται σε φθίνουσα σειρά. Τα l χαρακτηριστικά που επιλέγονται αποτελούν τις l καλύτερες τιμές του κριτηρίου C(k) και έπειτα προχωρούμε στο σχεδιασμό ενός διανύσματος χαρακτηριστικών με αυτά.

Το κυριότερο πλεονέκτημα της ξεχωριστής αντιμετώπισης του καθενός χαρακτηριστικού είναι η υπολογιστική απλότητα. Παρόλα αυτά όμως δεν λαμβάνονται υπόψη οι υπάρχουσες συσχετίσεις (correlations) μεταξύ των χαρακτηριστικών.

Επιλογή Διανύσματος Χαρακτηριστικών
Η ξεχωριστή αντιμετώπιση των χαρακτηριστικών έχει, όπως αναφέρθηκε, το πλεονέκτημα υπολογιστικής απλότητας. Σε πολύπλοκα προβλήματα και σε περιπτώσεις όπου τα χαρακτηριστικά εμφανίζουν μεγάλες συσχετίσεις μεταξύ τους η τεχνική αυτή δε θα ήταν αποδοτική. Για το λόγο αυτό θα εστιάσουμε σε τεχνικές οι οποίες βαθμονομούν τις ταξινομικές ικανότητες διανυσμάτων χαρακτηριστικών. Γίνεται εύκολα αντιληπτό πως στην περίπτωση αυτή η υπολογιστική πολυπλοκότητα είναι ένας σοβαρός περιοριστικός παράγοντας. Εάν θέλουμε να είμαστε πιστοί στην έννοια της βελτιστοποίησης θα έπρεπε να σχηματίζαμε όλους τους δυνατούς συνδυασμούς διανυσμάτων των l χαρακτηριστικών. Ανάλογα με τον κανόνα βελτιστοποίησης με τον οποίο εργαζόμαστε μπορούμε να χωρίσουμε την επιλογή διανύσματος χαρακτηριστικών σε δύο κατηγορίες-προσεγγίσεις.

## 3.7 Αλγόριθμοι Επιλογής με βάση το ποσοστό αναγνώρισης/Συσκευαστές (Wrappers)

Η πρώτη προσέγγιση είναι η προσέγγιση με βάση το ποσοστό αναγνώρισης που πετυχαίνει το σύστημα ταξινόμησης χρησιμοποιώντας το (wrapper approach). Στην προσέγγιση αυτή, ο κανόνας για την επιλογή χαρακτηριστικών είναι ανεξάρτητος του τύπου του ταξινομητή που θα χρησιμοποιήσουμε στη σχεδίαση του συστήματός μας. Για κάθε έναν συνδυασμό πρέπει να χρησιμοποιήσουμε ένα από τα μέτρα διαχωρισιμότητας κλάσεων και να διαλέξουμε με βάση αυτό τον καλύτερο από τους συνδυασμούς. Συνολικά ο αριθμός των διαφορετικών

συνδυασμών με l χαρακτηριστικά σε ένα σύνολο m χαρακτηριστικών $\binom{m}{l} = \frac{m!}{l!(m-l))!}$. Για ένα μεγάλο σύνολο χαρακτηριστικών γίνεται αντιληπτό πως ο αριθμός αυτός είναι πρακτικά πολύ μεγάλος. Ακόμα στις περισσότερες περιπτώσεις δε γνωρίζουμε ποιος είναι ο βέλτιστος αριθμός χαρακτηριστικών.

Οι σημαντικότερες τεχνικές προσέγγισης είναι η προς τα πίσω επιλογή (Sequential Backward Selection) και η προς τα εμπρός επιλογή (Sequential Forward Selection).

### 3.7.1 Sequential Backward Selection

Η παρουσίαση της μεθόδου θα γίνει με ένα παράδειγμα. Θεωρούμε ένα σύνολο με $m = 4$ χαρακτηριστικά $[x_1, x_2, x_3, x_4]$ εκ των οποίων ζητούμε να γίνει επιλογή του βέλτιστου διανύσματος με δύο χαρακτηριστικά. Η διαδικασία αποτελείται από τα ακόλουθα βήματα:

1. Υιοθετούμε ένα κριτήριο διαχωρισιμότητας C και υπολογίζουμε τις επιμέρους τιμές για κάθε χαρακτηριστικό του διανύσματος $[x_1, x_2, x_3, x_4]^{\mathrm{T}}$.
2. Απομακρύνουμε ένα χαρακτηριστικό και για κάθε έναν από τους συνδυασμούς που θα προκύψουν $[x_1, x_2, x_3]^{\mathrm{T}}$, $[x_1, x_2, x_4]^{\mathrm{T}}$, $[x_1, x_3, x_4]^{\mathrm{T}}$, $[x_2, x_3, x_4]^{\mathrm{T}}$ υπολογίζουμε την αντίστοιχη τιμή του κριτηρίου μας. Επιλέγουμε το συνδυασμό με την καλύτερη τιμή, π.χ. το $[x_1, x_2, x_3]^{\mathrm{T}}$.
3. Από το 3-διάστατο διάνυσμα που προέκυψε απομακρύνουμε ένα χαρακτηριστικό για κάθε έναν συνδυασμό $[x_1, x_2]^{\mathrm{T}}$, $[x_1, x_3]^{\mathrm{T}}$, $[x_2, x_3]^{\mathrm{T}}$ και υπολογίζουμε πάλι την τιμή του κριτηρίου για κάθε ένα από αυτά και επιλέγουμε αυτό με την καλύτερη τιμή του κριτηρίου.

Έτσι, ξεκινώντας από έναν αριθμό m χαρακτηριστικών, σε κάθε βήμα απομακρύνουμε ένα χαρακτηριστικό μέχρι να καταλήξουμε σε ένα διάνυσμα l χαρακτηριστικών.

Ο αλγόριθμος αυτός λειτουργεί καλύτερα όταν το μέγεθος m του ζητούμενου υποσυνόλου είναι μεγάλο καθώς αρχικά το εύρος αναζήτησης του είναι μεγαλύτερο.

Το κύριο μειονέκτημα της μεθόδου αυτής είναι ότι δε είναι δυνατή η επαναξιολόγηση ενός χαρακτηριστικού αφότου αυτό έχει απομακρυνθεί.

### 3.7.2 Sequential Forward Selection

Στην ίδια λογική είναι και η τεχνική Sequential Forward Selection, μόνο που η διαδικασία για την επιλογή γίνεται ανάποδα.

1. Για κάθε χαρακτηριστικό υπολογίσουμε την τιμή του κριτηρίου. Θεωρούμε ότι το χαρακτηριστικό με την καλύτερη τιμή είναι το $x_1$.
2. Σχηματίζουμε όλα τα διδιάστατα διανύσματα που χρησιμοποιούν το βέλτιστο χαρακτηριστικό του πρώτου βήματος $[x_1, x_2]^{\mathrm{T}}$, $[x_1, x_3]^{\mathrm{T}}$, $[x_1, x_4]^{\mathrm{T}}$. Υπολογίζουμε και πάλι την τιμή του κριτηρίου μας και συνεχίζουμε στο ίδιο μοτίβο μέχρι τον ζητούμενο αριθμό χαρακτηριστικών l.

Ο αλγόριθμος Forward Selection λειτουργεί καλύτερα όταν το μέγεθος m του ζητούμενου υποσυνόλου είναι μικρό καθώς αρχικά το εύρος αναζήτησης του είναι μεγαλύτερο.

Το κύριο μειονέκτημα της μεθόδου αυτής είναι ότι δεν είναι δυνατή η επαναξιολόγηση και απομάκρυνση ενός χαρακτηριστικού αφότου αυτό έχει προστεθεί στο σύνολο. Συχνά π.χ. κάποια χαρακτηριστικά καθίστανται περιττά έπειτα από την πρόσθεση νέων.

Πρέπει να αναφέρουμε ότι ο αλγόριθμος αυτός αποτελεί μια στρατηγική αναζήτησης στην οποία εκτός από το τελικό ποσοστό αναγνώρισης μπορεί να εφαρμοστεί και από οποιαδήποτε άλλη αντικειμενική συνάρτηση (π.χ. συσχέτιση μεταξύ των χαρακτηριστικών). Ωστόσο, αποτελεί έναν από τους κλασικούς αλγορίθμους τύπου wrapper, γι' αυτό και τον αναφέρουμε σε αυτή την κατηγορία.

Και οι δύο τεχνικές αυτές προφανώς είναι υποβέλτιστες μιας και δεν μπορεί να δειχθεί ότι το παραχθέν υποσύνολο είναι το βέλτιστο. Συνήθως κατηγορούνται ότι είναι μέθοδοι ωμής βίας που απαιτούν μεγάλα ποσά υπολογισμού.

### 3.7.3 Επιλογή με ταυτόχρονη προσθήκη και αφαίρεση χαρακτηριστικών σε κάθε βήμα

Η μέθοδος αυτή προσπαθεί να καλύψει τις αδυναμίες των δύο προηγούμενων προσθέτοντας τη δυνατότητα επαναξιολόγησης χαρακτηριστικών που προστέθηκαν ή απομακρύνθηκαν από το σύνολο προηγουμένως. Συγκεκριμένα σε κάθε επανάληψη προσθέτει έναν αριθμό από L χαρακτηριστικά και στη συνέχεια αφαιρεί R.

Το κύριο μειονέκτημα αυτής της μεθόδου είναι ότι υπάρχει έλλειψη γνώσης για τους κατάλληλους αριθμούς L και R που πρέπει να χρησιμοποιηθούν ώστε να υπάρξει καλύτερο αποτέλεσμα.

### 3.8 Αλγόριθμοι Επιλογής με εφαρμογή ειδικού φίλτρου (Filters)

Η δεύτερη προσέγγιση ονομάζεται προσέγγιση φίλτρου (filter approach). Η διαφορά σε σχέση με την προηγούμενη προσέγγιση έγκειται στο γεγονός ότι αυτή τη φορά αντί να θεωρούμε σαν κριτήριο κάποιο από τα μέτρα διαχωρισιμότητας κλάσεων, θεωρούμε σαν κριτήριο την απόδοση αυτού καθεαυτού του ταξινομητή μας. Ο συνδυασμός χαρακτηριστικών που τελικά επιλέγεται είναι αυτός που παρουσιάζει το μικρότερο σφάλμα ταξινόμησης. Η υπολογιστική πολυπλοκότητα ανεβαίνει ανάλογα με το είδος του ταξινομητή που χρησιμοποιείται.

Οι αλγόριθμοι αυτής της κατηγορίας χρησιμοποιούν ως κριτήριο για την αξιολόγηση των υποσυνόλων συναρτήσεις ανεξάρτητες με το ποσοστό αναγνώρισης. Συνήθως οι συναρτήσεις αυτές μετράνε μεγέθη σχετικά με τη θεωρία πληροφορίας, όπως είναι η αμοιβαία πληροφορία μεταξύ των χαρακτηριστικών ή η πληροφορία που δίνουν τα χαρακτηριστικά για το πρόβλημα. Κύριος σκοπός των μεθόδων αυτών είναι να επιλεγούν χαρακτηριστικά τα οποία να είναι όσο το δυνατόν περισσότερο ασυσχέτιστα μεταξύ τους και παράλληλα σχετικά με το πρόβλημα.

#### 3.8.1 Αλγόριθμος επιλογής με βάση το F-score

Ο αλγόριθμος F-score αποτελεί μια απλή μέθοδο για τη μέτρηση της διαφορετικότητας δύο συνόλων πραγματικών αριθμών. Δεδομένων των διανυσμάτων χαρακτηριστικών των δειγμάτων εκπαίδευσης $x_k, k = 1, .., m$ εάν $n_+$ και $n_-$ είναι ο αριθμός των δειγμάτων των δύο διαφορετικών κλάσεων τότε το F-score για το i-οστό χαρακτηριστικό ορίζεται ως:

$$F(i) = \frac{(\bar{x}_i^{(+)} - \bar{x}_i)^2 + (\bar{x}_i^{(-)} - \bar{x}_i)^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n_+}(x_{k,i}^{(+)} - \bar{x}_i^{(+)})^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n_+}(x_{k,i}^{(-)} - \bar{x}_i^{(-)})^2}$$

όπου τα $\bar{x}_i$, $\bar{x}_i^{(+)}, \bar{x}_i^{(-)}$, είναι αντίστοιχα οι μέσες τιμές του i-οστού χαρακτηριστικού συνολικά στα δείγματα, στα θετικά μόνο δείγματα και στα αρνητικά δείγματα. Με $x_{k,i}^{(+)}$ συμβολίζουμε το i-οστό χαρακτηριστικό του k-οστού θετικού δείγματος (αντίστοιχα για τα αρνητικά). Ο αριθμητής είναι ένα

μέτρο για το πόσο διαφέρουν τα θετικά από τα αρνητικά δείγματα, ενώ ο παρανομαστής μετράει τη διαφορετικότητα στο εσωτερικό του κάθε συνόλου.

Όπως λογικά προκύπτει, όσο μεγαλύτερο είναι το F-score ενός χαρακτηριστικού, τόσο περισσότερο κατάλληλο κρίνεται για να διαχωρίσει τις δύο αυτές κλάσεις.

Ένα μειονέκτημα αυτής της μεθόδου είναι ότι δε λαμβάνει υπόψη τη συμπληρωματική πληροφορία μεταξύ των χαρακτηριστικών που επιλέγει. Κάθε χαρακτηριστικό αξιολογείται ξεχωριστά, ανεξάρτητα με την πιθανή συνδυαστική δύναμή του με τα υπόλοιπα.

Ο αλγόριθμος F-score αποτελεί μια πολύ απλή και γρήγορη μέθοδο που δίνει συχνά ικανοποιητικά αποτελέσματα. Συνήθως επιλέγεται ένα σύνολο χαρακτηριστικών των οποίων το F-score είναι μεγαλύτερο από κάποιο κατώφλι που ορίζουμε.

### 3.8.2 Αλγόριθμος Μέγιστης Σχετικότητας και Ελάχιστου Πλεονασμού (MRMR)

Ο αλγόριθμος MRMR στοχεύει στην επιλογή ενός υποσυνόλου του οποίου τα χαρακτηριστικά έχουν μέγιστη συσχέτιση με το πρόβλημα αναγνώρισης που εξετάζουμε, και ταυτόχρονα, ελάχιστη αμοιβαία πληροφορία μεταξύ τους. Συγκεκριμένα τα αρχικά του αλγόριθμου υποδηλώνουν δύο επιδιώξεις, την μέγιστη σχετικότητα και τον ελάχιστο πλεονασμό. Στην συνέχεια παρουσιάζουμε με περισσότερη λεπτομέρεια τα βασικά σημεία του αλγορίθμου.

Ορίζουμε αρχικά την συνάρτηση αμοιβαίας πληροφορίας (mutual information) δύο διακριτών μεταβλητών ως:

$$I(x,y) = \sum_{i,j} p(x_i, y_i) log \frac{p(x_i, y_j)}{p(x_i)p(y_j)}$$

Χρησιμοποιούμε την συνάρτηση αμοιβαίας πληροφορίας ως ένα μέτρο ομοιότητας μεταξύ των δύο μεταβλητών. Η ιδέα πίσω από το κριτήριο του ελάχιστου πλεονασμού είναι να διαλέξουμε χαρακτηριστικά τα οποία έχουν μεταξύ τους μικρή αμοιβαία πληροφορία.

Αν με S συμβολίσουμε το σύνολο των επιλεγμένων χαρακτηριστικών από το αρχικό σύνολο Ω, τότε το μέτρο του πλεονασμού γι' αυτό το σύνολο δίνεται από τον τύπο:

$$W_S = \frac{1}{|S|^2} \sum_{i,j \in S} I(i,j)$$

όπου με |S| συμβολίζουμε το πλήθος των χαρακτηριστικών του S. Το κριτήριο του ελάχιστου πλεονασμού αναζητά το υποσύνολο S το οποίο ελαχιστοποιεί το $W_S$.

Έστω ότι με C συμβολίζουμε το σύνολο των κλάσεων που εμφανίζονται στο πρόβλημα μας. Χρησιμοποιώντας και πάλι τη συνάρτηση αμοιβαίας πληροφορίας, μπορούμε να μετρήσουμε τη σχέση των χαρακτηριστικών ενός συνόλου με τη μεταβλητή του συνόλου των κλάσεων $C = \{c_1, c_2, \dots, c_n\}$ του προβλήματος μας:

$$V_{C,S} = \frac{1}{|S|} \sum_{j \in S} I(C,j)$$

Το κριτήριο της μέγιστης σχετικότητας αναζητεί ένα σύνολο S που μεγιστοποιεί το $V_{C,S}$ για το συγκεκριμένο πρόβλημα αναγνώρισης με κλάσεις που ανήκουν στο C.

Τελικά ο αλγόριθμος MRMR προσπαθεί να ικανοποιήσει και τα δύο πιο πάνω κριτήρια μεγιστοποιώντας το λόγο:

$$\max_{S \subset \Omega} \frac{\sum_{i \in S} I(C,i)}{\frac{1}{|S|} \sum_{i,j \in S} I(i,j)}$$

Για την εύρεση της βέλτιστης λύσης, η πολυπλοκότητα του αλγορίθμου είναι $O(|\Omega|^{|S|})$ όπου |Ω| ο συνολικός αριθμός των χαρακτηριστικών στο πρόβλημά μας, καθώς χρειάζεται εξαντλητική αναζήτηση όλων των υποσυνόλων μεγέθους |S|. Ωστόσο στην πράξη κυρίως για λόγους πολυπλοκότητας υπολογίζουμε μια λύση κοντά στη βέλτιστη. Ο τρόπος είναι ο παρακάτω:

Αρχικά επιλέγουμε το 1ο χαρακτηριστικό του συνόλου, το οποίο είναι αυτό που μεγιστοποιεί το κριτήριο της μέγιστης συχνότητας. Στη συνέχεια τα υπόλοιπα χαρακτηριστικά επιλέγονται με ακολουθητικό τρόπο όμοια με τη στρατηγική Frequential Forward Selection που περιγράψαμε προηγουμένως. Πιο συγκεκριμένα, αν S είναι το τρέχον σύνολό μας, τότε κάθε φορά προσθέτουμε το χαρακτηριστικό i για το μεγιστοποιείται ο λόγος

$$\max_{S \subset \Omega} \frac{\sum_{i \in S} I(C,i)}{\frac{1}{|S|} \sum_{i,j \in S} I(i,j)}$$

Η διαδικασία τερματίζει μόλις φτάσουμε στο επιθυμητό μέγεθος συνόλου. Χρειάζεται να πούμε ότι προκειμένου να χρησιμοποιήσουμε την συνάρτηση αμοιβαίας πληροφορίας για διακριτές μεταβλητές όπως ορίστηκε παραπάνω, πρέπει αρχικά να διαφοροποιήσουμε όλα τα χαρακτηριστικά του προβλήματος.

### 3.8.3 Συνδυαστικός Αλγόριθμος Επιλογής Χαρακτηριστικών

Όπως αναφέραμε και προηγουμένως, στους αλγόριθμους τύπου wrapper, η χρησιμότητα των χαρακτηριστικών αξιολογείται σύμφωνα με την ακρίβεια αναγνώρισης που πετυχαίνει το σύστημά μας χρησιμοποιώντας τα. Η διαδικασία αυτή είναι συνήθως χρονοβόρα, επειδή περιλαμβάνει εκπαίδευση του συστήματος για κάθε υποσύνολο χαρακτηριστικών, ενώ επιπλέον αγνοείται η συσχέτιση και η αλληλεπίδραση μεταξύ των χαρακτηριστικών που επιλέγονται. Από την άλλη, οι μέθοδοι τύπου filter λαμβάνουν υπόψη τους κριτήρια αμοιβαίας πληροφορίας και στατιστικής συσχέτισης μεταξύ των χαρακτηριστικών διατηρώντας παράλληλα χαμηλή πολυπλοκότητα που τους επιτρέπει να είναι απλές και γρήγορες στην υλοποίηση. Ωστόσο οι αλγόριθμοι φίλτρου αγνοούν τελείως τον τύπο του ταξινομητή που χρησιμοποιούμε κάτι που συχνά παίζει σημαντικό ρόλο στο τελικό αποτέλεσμα.

Προκύπτει λοιπόν με λογικό τρόπο η ιδέα ανάπτυξης ενός συνδυαστικού συστήματος επιλογής χαρακτηριστικών που θα χρησιμοποιεί και τις δύο παραπάνω κατηγορίες. Μια πιθανή λύση είναι ο αλγόριθμος δύο σταδίων, όπου στο πρώτο στάδιο προηγείται κάποια μέθοδος φίλτρου ενώ στο δεύτερο στάδιο εφαρμόζεται μέθοδος wrapper στο υποσύνολο που προέκυψε. Με αυτό τον τρόπο συνδυάζονται τα θετικά σημεία και των δύο μεθόδων, ενώ παράλληλα μειώνεται σημαντικά η πολυπλοκότητα καθώς ο wrapper έχει πλέον ως είσοδο ένα αρκετά μικρότερο σύνολο από ασυσχέτιστα χαρακτηριστικά.

Για τις δύο προσεγγίσεις που παρατέθηκαν παραπάνω έχουν προταθεί μια σειρά από αποδοτικές τεχνικές. Μερικές από αυτές είναι υποβέλτιστες και μερικές βέλτιστες.

## 3.9 Nested-Μέθοδος Επιλογής Υποσυνόλου

Ορισμένες ενσωματωμένες μέθοδοι καθοδηγούν την αναζήτησή τους με την εκτίμηση των μεταβολών στην αντικειμενική αξία της συνάρτησης που προκύπτουν κάνοντας τις κινήσεις στο μεταβλητό χώρο υποσυνόλων. Σε συνδυασμό με άπληστες στρατηγικές αναζήτησης (προς τα πίσω ή προς τα εμπρός επιλογή χαρακτηριστικών) δίνουν ένθετα υποσύνολα μεταβλητών.

Καλούμε s τον αριθμό των μεταβλητών που επιλέγονται στο συγκεκριμένο βήμα του αλγορίθμου και J(s) η τιμή της αντικειμενικής συνάρτησης της εκπαιδευόμενης μηχανής μάθησης χρησιμοποιώντας μια τέτοια μεταβλητή υποσύνολο. Η πρόβλεψη στην αλλαγή της αντικειμενικής συνάρτησης επιτυγχάνεται με:

1. Πεπερασμένος υπολογισμός διαφορών: Η διαφορά μεταξύ J(s) και J(s+1) ή J(s-1) υπολογίζεται για τις μεταβλητές που είναι υποψήφιες για πρόσθεση ή αφαίρεση.
2. Τετραγωνική Προσέγγιση της Συνάρτησης Κόστους: Η μέθοδος αυτή προτάθηκε αρχικά για εφαρμογή στα νευρωνικά δίκτυα (LeCun et al., 1990). Μπορεί α χρησιμοποιηθεί για προς τα πίσω αφαίρεση μεταβλητών μέσω του κλαδέματος των μεταβλητών εισόδου σύμφωνα με το βάρος $w_i$. Μια δεύτερη προσέγγιση είναι η επέκταση Taylor της συνάρτησης J. Κατά τη βέλτιστη τιμή των J, ο όρος πρώτης τάξης μπορεί να αγνοηθεί, αποδίδοντας για τη μεταβλητή i για τη διακύμανση $DJ_i = \frac{1}{2}\frac{\partial^2 J}{\partial w_i^2}(Dw_i)^2$. Η μεταβολή στο βάρος $Dw_i$ αντιστοιχεί στην αφαιρούμενη μεταβλητή i.
3. Ευαισθησία Υπολογισμού της Αντικειμενικής Συνάρτησης: Η απόλυτη τιμή των τετραγώνων της παραγώγου του J σε σχέση με το $x_i$ (ή σε σχέση με το $w_i$).

Μερικοί αλγόριθμοι εκπαίδευσης χρησιμοποιούν τη μέθοδο πεπερασμένων διαφορών, επειδή μπορούν να υπολογιστούν αποτελεσματικά ακριβείς διαφορές, χωρίς την επανεκπαίδευση νέου μοντέλου για κάθε υποψήφια μεταβλητή. Μια τέτοια περίπτωση είναι το γραμμικό μοντέλο ελαχίστων τετραγώνων. Η ορθοκανονικοποίηση Gram-Schmidt επιτρέπει την προς τα εμπρός επιλογή χαρακτηριστικών προσθέτοντας σε κάθε βήμα την μεταβλητή που μειώνει το μέσο τετραγωνικό σφάλμα. Άλλοι αλγόριθμοι, όπως η μέθοδος των πυρήνων, υπολογίζουν αποτελεσματικά προσεγγίσεις των διαφορών. Η μέθοδος των πυρήνων είναι μια μέθοδος της μορφής $f(x) = \sum_{k=1}^{m} a_k K(x, x_k)$ όπου K είναι η συνάρτηση πυρήνα που μετρά τη συσχέτιση μεταξύ $x$ και $x_k$. Η μεταβολή στην J(s) υπολογίζεται κρατώντας τις $a_k$ τιμές σταθερές. Αυτή η διαδικασία προτάθηκε αρχικά για τα SVMs (Guyon et al., 2002).

## 3.10 Άμεση Αντικειμενική Βελτιστοποίηση

Σε γενικές γραμμές, η αντικειμενική συνάρτηση αποτελείται από δύο όρους που συγκρίνονται μεταξύ τους: (α) καλή προσαρμογή (goodness-of-fit) που πρέπει να μεγιστοποιηθεί και (β) ο αριθμός των μεταβλητών που πρέπει να ελαχιστοποιηθεί. Η προσέγγιση αυτή φέρει ομοιότητες με αντικειμενικές συναρτήσεις με δύο μέρη, τον όρο καλής προσαρμογής και τον όρο

κανονικοποιήσης, ιδιαίτερα όταν η επίδραση του όρου κανονικοποίησης είναι να συρρικνώσει τον παραμετρικό χώρο. Η αντιστοιχία αυτή έχει καθιερωθεί επίσημα (Weston et al. 2003) για τη συγκεκριμένη περίπτωση της ταξινόμησης με γραμμικούς προγνωστικούς παράγοντες $f(x) = w \cdot x + b$, στα SVMs (Boser et al., 1992, Vapnik, 1998). Χρησιμοποιούνται συντελεστές συρρίκνωσης του τύπου $||w||_p^p = (\sum_{i=1}^{n} w_i^p)^{\frac{1}{p}}$ $(l_p - νόρμα)$. Στο όριο καθώς το $p \to 0$ η $l_p - νόρμα$ είναι απλώς ο αριθμός των βαρών, για παράδειγμα ο αριθμός των μεταβλητών. Weston et al. προχωρούν δείχνοντας ότι η φόρμουλα της $l_0 - νόρμας$ των SVMs μπορεί να λυθεί κατά προσέγγιση:

1. Εκπαίδευση ενός κανονικού γραμμικού SVM (χρησιμοποιώντας $l_1 - νόρμα$ ή $l_2 - νόρμα$).
2. Επανακλιμακοποίηση των μεταβλητών εισόδου από τον πολλαπλασιασμό τους με τις απόλυτες τιμές των συνιστωσών του διανύσματος βαρών w λαμβάνεται.
3. Επαναλάβει τα πρώτα 2 βήματα μέχρι τη σύγκλιση.

Η μέθοδος θυμίζει την προς τα πίσω αφαίρεση μεταβλητών με βάση τη μικρότερη τιμή |wi|. Η κανονικοποίηση μεταβλητών είναι σημαντική για μια τέτοια μέθοδο έτσι ώστε να λειτουργήσει σωστά.

### 3.11 Μέθοδοι Επικύρωσης

Εδώ παρουσιάζονται όλα τα ζητήματα που αφορούν την εκτός δείγματος πρόβλεψης απόδοσης (πρόβλεψη γενίκευσης) και την επιλογή του μοντέλου. Αυτά είναι που εμπλέκονται σε διάφορες πτυχές της επιλογής μεταβλητής ή χαρακτηριστικού: να προσδιοριστεί ο αριθμός των μεταβλητών που είναι "σημαντικές", να καθοδηγήσει και να σταματήσει την αναζήτηση για τα καλά υποσύνολα μεταβλητών, να επιλέξει υπερπαραμέτρους (hyperparameters), και να αξιολογήσουν την τελική απόδοση του συστήματος.

Θα πρέπει πρώτα να διακριθεί το πρόβλημα της επιλογής μοντέλου από εκείνο της αξιολόγησης της τελικής απόδοσης της προγνώσεως. Για το σκοπό αυτό, είναι σημαντικό να αναιρέσει ένα ανεξάρτητο σύνολο ελέγχου. Το υπόλοιπο των δεδομένων χρησιμοποιείται τόσο για την κατάρτιση αλλά και την εκτέλεση επιλογής μοντέλου. Επιπλέον πειραματική εκλέπτυνση μπορεί να προστεθεί με την επανάληψη ολόκληρου του πειράματος για αρκετές εναλλαγές του συνόλου ελέγχου.

Για να εκτελεστεί η επιλογή του μοντέλου (συμπεριλαμβανομένων των μεταβλητών / χαρακτηριστικών επιλογής και βελτιστοποίηση των

υπερπαραμέτρων) στα δεδομένα που δεν χρησιμοποιούνται για τις δοκιμές μπορεί να γίνει περαιτέρω διαχωρισμός μεταξύ σταθερού κατάρτισης και συνόλων επικύρωσης, ή μπορούν να χρησιμοποιηθούν διάφορες μέθοδοι διασταυρωμένης επικύρωσης (cross-validation). Το πρόβλημα κατόπιν επαναφέρεται πίσω στην σημασία εκτίμησης των διαφορών σε σφάλματα επικύρωσης. Για ένα σταθερό σύνολο επικύρωσης, στατιστικές δοκιμές μπορούν να χρησιμοποιηθούν, αλλά η ισχύς τους είναι αμφίβολη για διασταυρωμένη επικύρωση, επειδή παραβιάζονται παραδοχές της ανεξαρτησίας. Επίσης, αν υπάρχουν πολλά παραδείγματα, μπορεί να μην είναι αναγκαίο να χωρίσει τα δεδομένα εκπαίδευσης και μπορούν να χρησιμοποιηθούν συγκρίσεις των σφαλμάτων εκπαίδευσης με τις στατιστικές δοκιμές.

Επιλέγοντας ποιο τμήμα των δεδομένων πρέπει να χρησιμοποιηθεί για την εκπαίδευση και για την επικύρωση είναι ένα ανοικτό πρόβλημα. Πολλοί συγγραφείς καταφεύγουν στη χρήση της «leave-one-out» διαδικασίας διασταυρωμένης επικύρωσης, μολονότι είναι γνωστό ότι εκτιμάται ένα υψηλό σφάλμα γενίκευσης στη διακύμανση (Vapnik, 1982) για να δώσει υπερβολικά αισιόδοξα αποτελέσματα, ιδιαίτερα όταν δεν είναι σωστά, ανεξάρτητα και πανομοιότυπα δείγματα δεδομένων από την «αληθή» διανομή. Η διαδικασία «leave-one-out» συνίσταται σε αφαίρεση ενός παραδείγματος από το σύνολο εκπαίδευσης, την κατασκευή του προγνωστικός παράγοντα βάσει μόνο των υπόλοιπων στοιχείων της κατάρτισης, τότε οι δοκιμές γίνονται στο αφαιρεμένο παράδειγμα. Με αυτό τον τρόπο κανείς δοκιμάζει όλα τα παραδείγματα των δεδομένων εκπαίδευσης και παίρνει τους μέσους όρους των αποτελεσμάτων. Όπως αναφέρθηκε προηγουμένως, υπάρχουν ακριβή ή κατά προσέγγιση φόρμουλες του «leave-one-out» σφάλματος για έναν αριθμό μηχανών μάθησης.

Για κατάταξη μεταβλητής ή άλλες μεθόδους κατάταξης υποσυνόλων, μπορεί να ληφθεί μια άλλη στατιστική προσέγγιση. Η ιδέα είναι να εισαγάγει έναν ανιχνευτή στα δεδομένα που είναι μια τυχαία μεταβλητή. Χονδρικά, οι μεταβλητές που έχουν μια συνάφεια μικρότερη ή ίση με εκείνη του ανιχνευτή θα πρέπει να απορρίπτονται. Οι Bi et al. (2003) θεωρούν ότι μια πολύ απλή εφαρμογή αυτής της ιδέας είναι να εισάγουν στα δεδομένα τους τρεις επιπλέον τυχαίες ψευδομεταβλητές από μια κατανομή Gauss και να τις υποβάλουν στη διαδικασία επιλογής μεταβλητών με τις «αληθείς μεταβλητές». Στη συνέχεια, διαγράφονται όλες οι μεταβλητές που είναι λιγότερο σημαντικές από μία από τις τρεις ψευδομεταβλητές (σύμφωνα με το κριτήριο βάρους). Οι Stoppiglia et al. (2003) προτείνουν μια πιο εξελιγμένη μέθοδο για την προς τα εμπρός μέθοδο επιλογής Gram-Schmidt . Για έναν ανιχνευτή κατανομής Gauss, παρέχουν ένα αναλυτικό τύπο για τον υπολογισμό της κατάταξης του ανιχνευτή που σχετίζεται με ένα συγκεκριμένο κίνδυνο για την αποδοχή άσχετων μεταβλητών. Μια μη παραμετρική παραλλαγή της μεθόδου ανιχνευτή συνίσταται στη δημιουργία ψευδομεταβλητών από τυχαίο ανακάτεμα πραγματικών

διανυσμάτων μεταβλητών. Σε μια προς τα εμπρός διαδικασία επιλογής, η εισαγωγή των ψευδομεταβλητών δεν διαταράσσει την επιλογή, επειδή οι ψευδομεταβλητές μπορεί να απορρίπτονται όταν επιλέγονται.

## 3.12 Μέθοδος Επιλογής Χαρακτηριστικών για τις Μηχανές Διανυσματικής Υποστήριξης

Η επιλογή σχετικών χαρακτηριστικών για τους ταξινομητές μηχανών διανυσματικής υποστήριξης (SVM) είναι σημαντική για διάφορους λόγους, όπως η απόδοση, γενίκευση, υπολογιστική αποδοτικότητα και επεξηγηματικότητα. Παραδοσιακές SVM προσεγγίσεις για την επιλογή χαρακτηριστικών συνήθως εξαγάγουν τα χαρακτηριστικά γνωρίσματα και μαθαίνουν τις παραμέτρους των SVM ανεξάρτητα. Η ανεξάρτητη εκτέλεση αυτών των δύο βημάτων μπορεί να οδηγήσει σε απώλεια πληροφοριών που σχετίζονται με τη διαδικασία ταξινόμησης. Πειράματα σε διάφορες βάσεις δεδομένων παρουσιάζουν σημαντική μείωση των χαρακτηριστικών που χρησιμοποιούνται και διατηρούν ταυτόχρονα την απόδοση ταξινόμησης.

Κατά την τελευταία δεκαετία, τα SVMs έχουν γίνει το σημείο αναφοράς για πολλά προβλήματα ταξινόμησης λόγω της ευελιξίας τους, της υπολογιστικής αποδοτικότητας και της ικανότητα να διαχειριστούν δεδομένα υψηλής διάστασης. Όπως και άλλοι ταξινομητές, σε πολλές μεθόδους μάθησης με επίβλεψη, η αποτυχία να απορρίψει άσχετα χαρακτηριστικά (π.χ. θόρυβος , ακραίες τιμές) θα επηρεάσει την απόδοση του συστήματος το οποίο περιλαμβάνει την ακρίβεια ταξινόμησης, υπολογιστική αποδοτικότητα και τη σύγκλιση της μάθησης. Πρώτον, η έμμεση νομιμοποίηση που επιτυγχάνεται με το χαρακτηριστικό κλάδεμα αυξάνει συνήθως την ικανότητα γενίκευσης ταξινομητών και αυτό οδηγεί γενικά σε υψηλότερη ακρίβεια ταξινόμησης. Δεύτερον, χρησιμοποιώντας άσχετα χαρακτηριστικά, επίσης, αυξάνει σημαντικά τον χρόνο υπολογισμού. Τρίτον, πάρα πολλές δυνατότητες μπορεί να καταστήσει αδύνατη τη σύγκλιση, που οδηγούν σε αποφάσεις τυχαίας κατάταξης. Εκτός από την απόδοση του συστήματος, ο προσδιορισμός των σημαντικών μεταβλητών που έχουν διαισθητική φυσική ερμηνεία είναι μια άλλη κρίσιμη απαίτηση σε πολλές εφαρμογές. Λόγω των παραπάνω λόγων, η επιλογή χαρακτηριστικών έχει κεντρικό ρόλο σε διάφορους τομείς, συμπεριλαμβανομένης της επεξεργασίας σήματος, Στατιστική, νευρωνικά δίκτυα, αναγνώρισης προτύπων και μηχανικής μάθησης.

Παραδοσιακά, η επιλογή χαρακτηριστικών γίνεται ανεξάρτητα από την εκμάθηση των παραμέτρων ταξινόμησης. Ωστόσο, η χωριστή εκτέλεση αυτών δύο σταδίων θα μπορούσε να οδηγήσει σε απώλεια των πληροφοριών που σχετίζονται με τα καθήκοντα κατάταξης. Πρόσφατα διάφορες προσεγγίσεις για

την από κοινού επιλογή χαρακτηριστικών και SVM κατασκευής έχουν προταθεί. Ωστόσο, τα προβλήματα βελτιστοποίησης αυτών των μεθόδων δεν είναι κυρτά και η διαδικασία της κατάρτισης ταξινόμησης συγκλίνει συχνά σε ένα τοπικό ελάχιστο.

Υπάρχουν πολλές τεχνικές για την επιλογή χαρακτηριστικών για τα SVMs. Μια δημοφιλής τεχνική για την επιλογή των χαρακτηριστικών είναι η RELIEF, η οποία εκχωρεί ένα βάρος με ένα ιδιαίτερο χαρακτηριστικό που βασίζεται στις διαφορές μεταξύ των τιμών από τον πλησιέστερο γείτονα ζεύγη. Οι Cao et al αναπτύσσουν περαιτέρω αυτή τη μέθοδο με την εκμάθηση χαρακτηριστικών με τα βάρη στους χώρους του πυρήνα. Αυτή η μέθοδος γίνεται συχνά ως επεξεργασία δεδομένων με βήμα, ανεξάρτητα από την κατασκευή ταξινομητή. Οι De la Torre και Vinyals εισάγουν μια παραμετροποιημένη σειρά Taylor για την επέκταση του πυρήνα που ουσιαστικά είναι κάτω σταθμισμένη για την ταξινόμηση με SVMs. Πρόσφατα, έχουν υπάρξει επίσης αρκετές εργασίες που στοχεύουν στην εκμάθηση πινάκων πυρήνα για την ταξινόμηση. Μια δημοφιλής προσέγγιση είναι να ορίσουμε μια παραμετροποιημένη οικογένεια πινάκων του πυρήνα και να γίνει βελτιστοποίηση των παραμέτρων του πυρήνα για να ευθυγραμμιστούν με ένα ιδανικό πυρήνα. Μια άλλη δημοφιλής προσέγγιση είναι να καθοριστεί μια επιθυμητή ιδιότητα και να γίνει εκμάθηση ενός πυρήνα που παρουσιάζει αυτή η ιδιότητα. Σε αυτές τις προσεγγίσεις , ο πυρήνας μαθαίνεται ανεξάρτητα από τις παραμέτρους των SVM.

Για να αντιμετωπιστεί το πρόβλημα της από κοινού μάθησης παραμέτρων SVM και πυρήνων, οι Chapelle et al και Weston et al προτείνουν μια μέθοδο για την επιλογή των παραμέτρων SVM συμπεριλαμβανομένων και των παραμέτρων των πυρήνων με την ελαχιστοποίηση της «leave-one-out» διασταυρωμένης επικύρωσης (LOOCV σφάλμα). Ωστόσο, δεδομένου ότι το σφάλμα LOOCV δεν μπορούσε να εκφραστεί αναλυτικά, αντί αυτού πρότειναν να ελαχιστοποιήσουν κάποιες διαφορίσιμες συναρτήσεις που ήταν άνω φράγματα του σφάλματος LOOCV. Οι Mangasarian &Wild θέσπισαν την τροποποίηση της αντικειμενικής συνάρτησης των SVMs και εκτελείται επιλογή χαρακτηριστικών κατ' επανάληψη σαρώνοντας όλα τα χαρακτηριστικά για να αποφασίσει εάν θα επιλεγεί ή θα καταργηθεί ένα χαρακτηριστικό γνώρισμα ανάλογα με την μείωση της τιμής της αντικειμενικής συνάρτησης.

Ένας τρόπος για να επιλεγεί ένα υποσύνολο των «καλών» χαρακτηριστικών είναι να καθαρίζουμε τα περιττά. Οι Hermes και Buhmann που ξεκίνησαν με την κατασκευή ενός ταξινομητή SVM χρησιμοποιώντας όλα τα διαθέσιμα χαρακτηριστικά και αφαιρώντας αναδρομικά αυτών που είχαν τις λιγότερες επιπτώσεις σχετικά με την λειτουργία απόφασης. Ομοίως, ο Avidan χρησιμοποίησε μια άπληστη διαδοχική προς τα εμπρός μέθοδο επιλογής για την εξεύρεση ενός υποσυνόλου χαρακτηριστικών και φορέων υποστήριξης που

προσεγγίζουν τη λύση που λαμβάνεται από τα SVM χρησιμοποιώντας όλα τα διαθέσιμα χαρακτηριστικά.

Για περαιτέρω περιορισμό των παραμέτρων των SVMs, ορισμένοι συγγραφείς πρότειναν την τροποποίηση της αντικειμενικής συνάρτησης των SVMs, συμπεριλαμβάνοντας όρους νομιμοποίησης ή περιορισμούς σχετικά με την παράμετρο w των SVMs. Για παράδειγμα, οι Chan et al περιελάμβανε δύο πρόσθετους περιορισμούς σχετικά με τις $l_1$ και $l_2$ νόρμες του w στη διαμόρφωση των SVMs για να επιτευχθεί μια αραιότητα στο διάνυσμα βαρών w. Οι Stoeckel & Fung πρόσθεσαν έναν περιορισμό στο w. Αυτός ο περιορισμός ήταν να έχουν το βάρος για κάθε εικονοστοιχείο το οποίο δεν εξαρτιόταν μόνο από το ίδιο το εικονοστοιχείο, αλλά και από τους γείτονές του. Οι Dundar et al προσέθεσαν έναν όρο νομιμοποίησης για το w στην αντικειμενική συνάρτηση για την ενθάρρυνση της λειτουργίας απόφασης να παράγουν παρόμοια αποτελέσματα για τα γειτονικά εικονοστοιχεία.

## 3.13 Αλγόριθμος SVM-RFE (Recursive Feature Elimination)

Ο αλγόριθμος SVM-RFE είναι μια μέθοδος συσκευαστών (wrapper method) για την επιλογή χαρακτηριστικών χρησιμοποιώντας μηχανές διανυσματικής υποστήριξης. Η μέθοδος SVM-RFE ταξινομεί όλα τα χαρακτηριστικά σύμφωνα με κάποια συνάρτηση σκορ (score function) και εξαλείφει ένα ή περισσότερα χαρακτηριστικά με τα χαμηλότερα σκορ. Αυτή η διαδικασία επαναλαμβάνεται μέχρι να επιτευχθεί η ταξινόμηση με μεγαλύτερη ακρίβεια. Παράγει την κατάταξη των χαρακτηριστικών χρησιμοποιώντας την προς τα πίσω αφαίρεση χαρακτηριστικών.

Λόγω της επιτυχίας του αλγορίθμου στην επιλογή γονιδίων για την ταξινόμηση του καρκίνου, ο SVM-RFE απέκτησε μεγάλη δημοτικότητα και είναι πολύ γνωστή μέθοδος ως μία από τις πιο αποτελεσματικές μεθόδους επιλογής χαρακτηριστικών. Παρόλα αυτά, η SVM-RFE είναι μια άπληστη μέθοδος που έχει μόνο ως στόχο να βρει τον καλύτερο δυνατό συνδυασμό για την ταξινόμηση.

Βασική ιδέα της μεθόδου, είναι να εξαλείψει περιττά γονίδια και να αποδώσει καλύτερα και πιο συμπαγή υποσύνολα του γονιδίου. Τα χαρακτηριστικά εξαλείφονται σύμφωνα με ένα κριτήριο σχετικό με την υποστήριξή τους στη συνάρτηση διάκρισης και οι SVM εκπαιδεύονται εκ νέου σε κάθε βήμα. Η SVM-RFE είναι μια μέθοδος με βάση το βάρος. Σε κάθε στάδιο, οι συντελεστές του διανύσματος βάρους ενός γραμμικού SVM χρησιμοποιούνται ως κριτήριο κατάταξης των χαρακτηριστικών.

Ο αλγόριθμος SVM-RFE μπορεί να διασπαστεί σε τέσσερα βήματα:

1) Εκπαιδεύει τα SVM με βάση το σύνολο εκπαίδευσης (training set).
2) Υπολογίζει το κριτήριο ταξινόμησης (ranking criterion) για όλα τα χαρακτηριστικά.
3) Αφαιρεί τα χαρακτηριστικά με το μικρότερο αποτέλεσμα στο κριτήριο ταξινόμησης.
4) Επαναλαμβάνει την πιο πάνω διαδικασία με το σύνολο εκπαίδευσης περιορισμένο στα εναπομείναντα χαρακτηριστικά.

Για την κατάρτιση του SVM χρειάζεται η λύση ενός προβλήματος τετραγωνικού προγραμματισμού. Δοθέντος ενός συνόλου εκπαίδευσης με ζεύγη $(x_i, y_i)$ $για\ i = 1, \dots, n$ το πρόβλημα γράφεται στη μορφή:

$$minimize\ W(a) = -\sum_{i=1}^{n} a_i + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j a_i a_j K\left(x_i, x_j\right)$$

$$s.t. \sum_{i=1}^{n} a_i y_i = 0\ \ \forall i, 0 \le a_i \le C$$

Ο αλγόριθμος SVM-RFE που προτάθηκε από τους Guyon, Weston, Barnhill, Vapnik (2002) επιστρέφει μια κατάταξη των χαρακτηριστικών ενός προβλήματος ταξινόμησης με την κατάρτιση ενός SVM με ένα γραμμικό πυρήνα και την αφαίρεση των χαρακτηριστικών με τη μικρότερη τιμή του κριτηρίου κατάταξης. Αυτό το κριτήριο είναι η τιμή του w που είναι η επιλογή του υπερεπιπέδου που δίδεται από τον SVM.

<u>Το πρόβλημα ταξινόμησης:</u>
Στόχος είναι να αντιμετωπίσουμε προβλήματα ταξινόμησης. Είσοδος είναι ένα διάνυσμα που καλούμε «πρότυπο» n συνιστωσών που καλούμε χαρακτηριστικά. Καλούμε F τον n-διάστατο χώρο χαρακτηριστικών. Περιοριζόμαστε σε χαρακτηριστικά που είναι συντελεστές γονιδιακής έκφρασης και μοτίβα που αντιστοιχούν σε ασθενείς καθώς επίσης και σε προβλήματα ταξινόμησης δύο κατηγοριών. Οι δύο κατηγορίες διαχωρίζονται με τα σύμβολα + και -. Δίνεται ένα σετ εκπαίδευσης $\{x_1, x_2, \dots, x_k, \dots, x_l\}$ με γνωστές ετικέτες κλάσης $\{y_1, y_2, \dots, y_k, \dots, y_l\}$ όπου $y_k \in \{-1,1\}$. Τα πρότυπα εκπαίδευσης χρησιμοποιούνται για να δημιουργήσουν μια συνάρτηση απόφασης (discriminant function) $D(x)$ που είναι μια βαθμωτή συνάρτηση προτύπων ειδόδου $x$. Νέα πρότυπα ταξινομούνται σύμφωνα με το πρόσημο της συνάρτησης απόφασης:

$$D(x) > 0, x \in στην\ κλάση +$$
$$D(x) < 0, x \in στην\ κλάση -$$
$$D(x) = 0, όριο\ απόφασης$$

Όπου $D(x) = w \cdot x + b$.

Μείωση του χώρου διαστάσεων και επιλογής χαρακτηριστικών:
Support Vector Machines (SVMs), μέθοδος που χρησιμοποιεί τεχνικές κανονικοποίησης, χρησιμοποιείται για το σκοπό αυτό, τη μείωση δηλαδή του χώρου επιλογής χαρακτηριστικών. Νέα χαρακτηριστικά λαμβάνονται που είναι γραμμικοί συνδυασμοί των αρχικών χαρακτηριστικών.

Κατάταξη χαρακτηριστικών με συντελεστές συσχέτισης:
Αξιολογώντας το πόσο καλά ένα μεμονωμένο χαρακτηριστικό συμβάλλει στο διαχωρισμό (π.χ. καρκίνος έναντι κανονικού) μπορεί να παράγει μια απλή χαρακτηριστική (γονίδιο) κατάταξη. Διάφοροι συντελεστές συσχέτισης χρησιμοποιούνται ως κριτήρια κατάταξης. Ο συντελεστής που χρησιμοποιείται στην Golub (1999) ορίζεται ως:

$$w_i = \frac{\mu_i(+) - \mu_i(-)}{\sigma_i(+) + \sigma_i(-)}$$

Όπου
$\mu_i(+)$: μέσος κλάσης +
$\mu_i(-)$: μέσος κλάσης -
$\sigma_i(+)$: τυπική απόκλιση κλάσης +
$\sigma_i(-)$: τυπική απόκλιση κλάσης -

Μεγάλες θετικές τιμές του $w_i$ δείχνουν ισχυρή συσχέτιση με την τάξη (+), ενώ οι μεγάλες αρνητικές τιμές $w_i$ δείχνουν ισχυρή συσχέτιση με την τάξη (-). Η αρχική μέθοδος της Golub (1999), είναι να επιλεγεί ίσος αριθμός γονιδίων με θετικό και αρνητικό συντελεστή συσχέτισης. Άλλοι (Furey, 2000), έχουν χρησιμοποιήσει την απόλυτη τιμή του $w_i$ ως κριτήριο κατάταξης. Επίσης στην εργασία Pavlidis (2000), οι συγγραφείς έχουν χρησιμοποιήσει ένα σχετικό συντελεστή $\frac{(\mu_i(+)-\mu_i(-))^2}{(\sigma_i(+))^2+(\sigma_i(-))^2}$ παρόμοιο με το κριτήριο διαχωρισμού του Fisher.

Αυτό που χαρακτηρίζει την κατάταξη χαρακτηριστικών με τις μεθόδους συσχέτισης είναι οι παραδοχές που γίνονται για έμμεση ορθογωνιότητα. Κάθε $w_i$ συντελεστής υπολογίζεται με πληροφορίες σχετικά με ένα ενιαίο χαρακτηριστικό (γονίδιο) και δεν λαμβάνει υπόψη την αμοιβαία ενημέρωση μεταξύ των χαρακτηριστικών.

Κριτήριο Κατάταξης και Ταξινόμηση
Μια πιθανή χρήση της κατάταξης χαρακτηριστικών είναι ο σχεδιασμός της προγνωστικής κλάσης (ή ταξινομητή) με βάση ένα προεπιλεγμένο υποσύνολο χαρακτηριστικών. Κάθε χαρακτηριστικό που συσχετίζεται (ή δεν συσχετίζονται) με το διαχωρισμό που μας ενδιαφέρει είναι από μόνο του μια τέτοια κατηγορία πρόβλεψης, έστω και ατελής. Αυτό υποδηλώνει μια απλή μέθοδος ταξινόμησης βάσει της σταθμισμένης ψήφου: τα χαρακτηριστικά ψηφίζουν αναλογικά με τον συντελεστή συσχέτισης τους. Τέτοια είναι η μέθοδος που χρησιμοποιείται σε

Golub (1999). Το σταθμισμένο σύστημα ψηφοφορίας δίνει ένα συγκεκριμένο γραμμικό ταξινομητή διάκρισης:

$$D(x) = w \cdot (x - \mu), \qquad \text{όπου } \mu = \frac{\mu(+) + \mu(-)}{2}$$

Οι συντελεστές χαρακτηριστικών κατάταξης μπορούν να χρησιμοποιηθούν ως βάρη ταξινομητών (classifier weights). Αμοιβαίως, ως συντελεστές χαρακτηριστικών κατάταξης μπορούν να χρησιμοποιηθούν τα βάρη πολλαπλασιάζοντας τις εισόδους ενός δεδομένου ταξινομητή. Οι είσοδοι που σταθμίζονται με την μεγαλύτερη τιμή, επηρεάζουν περισσότερο την απόφαση κατάταξης. Ως εκ τούτου, εάν ο ταξινομητής εκτελείται καλά, αυτές οι είσοδοι με τις μεγαλύτερες τιμές βαρών αντιστοιχούν στα πιο κατατοπιστική χαρακτηριστικά. Ειδικότερα , υπάρχουν πολλοί αλγόριθμοι για την εκπαίδευση γραμμικών συναρτήσεων διάκρισης που μπορεί να παρέχουν μια καλύτερη λειτουργία κατάταξης από τους συντελεστές συσχέτισης. Αυτοί οι αλγόριθμοι περιλαμβάνουν τη γραμμική διακρίνουσα του Fisher, και SVMs. Και οι δύο μέθοδοι είναι γνωστές στη Στατιστική ως «πολυμεταβλητοί» ταξινομητές, που σημαίνει ότι είναι βελτιστοποιημένα κατά τη διάρκεια της εκπαίδευσης για το χειρισμό πολλών μεταβλητών ( ή χαρακτηριστικών) ταυτόχρονα. Η μέθοδος Golub (1999), σε αντίθεση , είναι ένας συνδυασμός των πολλαπλών «μονομεταβλητών» ταξινομητών .

Ο αλγόριθμος SVM-RFE (Support Vector Machines Recursive Feature Elimination):

Ο συνδυασμένος αλγόριθμος SVM-RFE είναι μία εφαρμογή του RFE που χρησιμοποιεί το πλάτος των βαρών ως κριτήριο ταξινόμησης. Παρακάτω παρατίθεται μία γενική εφαρμογή του γραμμικού αλγόριθμου, με χρήση της εκπαίδευσης SVM.

***Ο Αλγόριθμος SVM-RFE:***

Κριτήριο Ταξινόμησης

$$w_i = \frac{\mu_i(+) - \mu_i(-)}{\sigma_i(+) + \sigma_i(-)}$$

Εισόδος:

$$X = [x_1, x_2, \dots, x_{k,} \dots, x_l]$$
$$Y = [y_1, y_2, \dots, y_k, \dots, y_l]$$

Έξοδος:

Ταξινομημένη λίστα χαρακτηριστικών r

Εναπομείναντα χαρακτηριστικά

$$s = [\,]$$

Λίστα Κατάταξης

$$r = [\,]$$

Στην περίπτωση γραμμικού πυρήνα, ορίζουμε μια συνάρτηση ποινής $\frac{1}{2}||w||^2$. Τότε, το λιγότερο ευαίσθητο χαρακτηριστικό που έχει την ελάχιστη τιμή του κριτηρίου κατάταξης εξαλείφεται πρώτα. Σε αυτό το χαρακτηριστικό γίνεται κατάταξη n. Γίνεται επανεκπαίδευση χωρίς το προηγούμενο χαρακτηριστικό και αφαιρείται το χαρακτηριστικό με την μικρότερη τιμή κριτηρίου κατάταξης. Αφού αυτό το χαρακτηριστικό αφαιρεθεί γίνεται κατάταξη n-1. Με τον τρόπο αυτό επαναλαμβάνεται αυτή η διαδικασία μέχρι να μην υπάρχει η δυνατότητα ένα χαρακτηριστικό να βρίσκεται στα αριστερά, και άρα μπορούμε να ταξινομήσουμε όλα τα χαρακτηριστικά. Ο αλγόριθμος έχει ως εξής:

$$Repeat\ until\ \boldsymbol{s} = [1,2,\dots,n]$$

Περιορισμός παραδειγμάτων εκπαίδευσης για την καλή λειτουργία των δεικτών

$$X_0 = X(:,\boldsymbol{s})$$

Εκπαίδευση ταξινομητή

$$a = SVM - train(X,\boldsymbol{y})$$

Υπολογισμός του διανύσματος βάρους διάστασης μήκους ($\boldsymbol{s}$)

$$\boldsymbol{w} = \sum_{k=1}^{n} a_k y_k x_k$$

Υπολογισμός του κριτηρίου κατάταξης

$$c_i = (w_i)^2\,, for\ all\ i$$

Βρίσκει το χαρακτηριστικό με τη μικρότερη τιμή κριτηρίου κατάταξης

$$f = argmin(\boldsymbol{c})$$

Ενημέρωση λίστας κατάταξης

$$\boldsymbol{r} = [\boldsymbol{s}(f),\boldsymbol{r}]$$

Εξάλειψη χαρακτηριστικού με τη μικρότερη τιμή του κριτηρίου κατάταξης

$$\boldsymbol{s} = \boldsymbol{s}(1{:}\,f-1, f+1{:}\,length(\boldsymbol{s}))$$

$$End.$$

***Αλγόριθμος SVM-train:***

Είσοδος: χαρακτηριστικά εκπαίδευσης $\{x_1,\dots,x_k,\dots,x_l\}$
και ετικέτες κλάσης $\{y_1,\dots,y_k,\dots,y_l\}$

$$minimize\ a_k$$

$$J = \frac{1}{2}\sum_{hk} y_h y_k a_h a_k (x_h x_k + \lambda\delta_{hk}) - \sum_k a_k$$

$$s.t.\ 0 \le a_k \le C\ \kappa\alpha\iota\ \sum_k a_k y_k = 0$$

Έξοδος: παράμετροι $a_k$

1. Let n be the initial number of features.
2. While $m \geq 0$
3. Estimate the direction vector w of the separating hyperplane using linear SVM.
4. Rank features according to the components of |w|.
5. Remove the feature with the smallest weight in absolute value
6. $m = m - 1$. More than one features can be removed in each iteration.
7. Estimate classification accuracy of the m surviving features using a linear SVM classifier.
8. End While.
9. Output as marker genes the set of surviving features achieving maximum accuracy performance.

### 3.14 Ο εκτιμητής Lasso (Least Absolute Shrinkage and Selection Operator)

LASSO είναι μια καινοτόμος μέθοδος επιλογής μεταβλητών για την παλινδρόμηση. Η επιλογή μεταβλητών για την παλινδρόμηση είναι εξαιρετικά σημαντικό όταν έχουμε στη διάθεσή μας μια μεγάλη συλλογή των πιθανών συμμεταβλητών από την οποία ελπίζουμε να επιλεγεί ένα φειδωλή σύνολο για την αποτελεσματική πρόβλεψη μιας μεταβλητής απόκρισης. Η LASSO ελαχιστοποιεί το υπολειπόμενο άθροισμα των τετραγώνων που υπόκεινται προς το άθροισμα της απόλυτης τιμής των συντελεστών είναι μικρότερη από μια σταθερά. Βοηθά όχι μόνο να βελτιώσει την ακρίβεια πρόβλεψης όταν ασχολείται με τη συγγραμμικότητα των δεδομένων, αλλά και μεταφέρει ιδιότητες όπως επεξηγηματικότητα και την αριθμητική σταθερότητα. Το έργο αυτό, επίσης, περιλαμβάνει ένα ζευγάρι των αριθμητικών προσεγγίσεων για την επίλυση LASSO.

Η μέθοδος αυτή (Tibshirani 1996) είναι ένα χρήσιμο εργαλείο για την επίτευξη συρρίκνωσης και επιλογής μεταβλητών ταυτόχρονα. Αφού η LASSO χρησιμοποιεί την ποινή $l_1$, η βελτιστοποίηση πρέπει να βασίζεται στον τετραγωνικό προγραμματισμό ή γενικά σε ένα μη-γραμμικό πρόγραμμα το οποίο είναι γνωστό ότι είναι υπολογιστικά εντατική. Επιτυγχάνει καλύτερη ακρίβεια πρόβλεψης από συρρίκνωση ως παλινδρόμηση κορυφογραμμής, αλλά την ίδια στιγμή, δίνει ένα αραιό διάλυμα, που σημαίνει ότι ορισμένοι

συντελεστές είναι ακριβώς μηδέν. Ως εκ τούτου, πιστεύεται ότι η LASSO για να επιτευχθεί η συρρίκνωση και η επιλογή μεταβλητής πρέπει να γίνει ταυτόχρονα.

Οι Knight και Fu (2000) απόδειξαν μερικά ασυμπτωτικού τύπου αποτελέσματα εκτιμητών για την LASSO. Οι Chen et al. (1999) και Bakin (1999) εφάρμοσαν την ιδέα του LASSO στη θεωρία κυματιδίων και ανέπτυξαν μια μέθοδο που ονομάζεται «basis pursuit». Οι Gunn και Kandola (2002) που εφάρμοσαν τη LASSO στη μηχανή πυρήνα, και Zhang et al. (2003) την εφάρμοσαν στην εξομάλυνση μοντέλων σφήνας.

Ένα πρόβλημα στην LASSO είναι ότι η αντικειμενική συνάρτηση δεν είναι διαφορίσιμη, και ως εκ τούτου ειδικές τεχνικές βελτιστοποίηση είναι απαραίτητες. Ο Tibshirani (1996) χρησιμοποιεί τον τετραγωνικό προγραμματισμό (QP) για τετραγωνική παλινδρόμηση και η επαναληπτική διαδικασία σταθμισμένων ελαχίστων τετραγώνων με QP για γενικευμένα γραμμικά μοντέλα. Οι Osborne et al. (2000) πρότεινε ένα ταχύτερο αλγόριθμο QP για LASSO, ο οποίος τέθηκε σε εφαρμογή από Lokhorst et al. (1999). Οι αλγόριθμοι που βασίζονται στον Τετραγωνικό Προγραμματισμό, ωστόσο, δεν μπορεί να εφαρμόζεται εύκολα σε μεγάλα σύνολα δεδομένων, όταν η διάσταση των εισροών είναι πολύ μεγάλη. Επιπλέον, οι αλγόριθμοι μπορούν να μην συγκλίνουν στη βέλτιστη λύση όταν η συνάρτηση απώλειας δεν είναι άλλη από την απώλεια τετραγώνων. Εκτός από QP, οι Grandvalet και Canu (1999) εφάρμοσαν ένα σημειακό αλγόριθμο χρησιμοποιώντας την ισοτιμία μεταξύ της προσαρμοστική παλινδρόμησης κορυφογραμμής και LASSO και οι Perkins et al. (2003) ανέπτυξαν ένα «stagewise» κάθοδο κλίσης αλγόριθμο που ονομάζεται «graftying». Αυτοί οι αλγόριθμοι, ωστόσο, δεν μπορούν να οδηγήσουν σε ολική σύγκλιση.

Η μέθοδος Lasso έχει διπλά πλεονεκτήματα ώστε να είναι υπολογιστικά εφικτή για την προσαρμογή του μοντέλου του τετραγωνικού προγραμματισμού, και γενικά οδηγεί σε αραιές λύσεις.

Θεωρούμε το απλό γραμμικό μοντέλο παλινδρόμησης με δεδομένα $(\boldsymbol{x}^i, y_i), i = 1, .., n$, όπου $\boldsymbol{x}^i = (x_{i1}, \dots, x_{ip})^{\mathrm{T}}$ και $y_i$ η μεταβλητή απόκρισης.

Η συνάρτηση απώλειας της παλινδρόμησης Lasso ορίζεται ως:

$$L = \sum_i (y_i - \sum_p \beta_p x_{ip})^2 + \lambda \sum_p ||\beta_p||_1$$

όπου $x_{ip}$ είναι p-οστός προγνωστικός παράγοντας (χαρακτηριστικό) του i-οστού δεδομένου.

Οι παράμετροι στο γραμμικό μοντέλο εκτιμούνται με τη μέθοδο Lasso.

$$\widehat{\widehat{\beta(\lambda)}} = argmin_{\beta}(n^{-1} \parallel \boldsymbol{Y} - \boldsymbol{X}\beta \parallel_2^2 + \lambda \parallel \beta \parallel_1),$$

όπου $\parallel \boldsymbol{Y} - \boldsymbol{X}\beta \parallel_2^2 = \sum_{i=1}^{n}(Y_i - (X\beta)_i)^2$, $\parallel \beta \parallel_1 = \sum_{j=1}^{p} |\beta_j|$ και $\lambda \geq 0$ είναι μια συνάρτηση ποινής. Ο εκτιμητής έχει την ιδιότητα να κάνει επιλογή μεταβλητών (variable selection) αφού $\widehat{\widehat{\beta_j(\lambda)}}=0$ για κάποια j (εξαρτάται από την επιλογή του λ) και το $\widehat{\widehat{\beta_j(\lambda)}}$ μπορεί να θεωρηθεί ως ένας συρρικνωμένος εκτιμητής ελαχίστων τετραγώνων (LASSO=Least Absolute Shrinkage and Selection Operator).

Η βελτιστοποίηση της $\widehat{\widehat{\beta(\lambda)}} = argmin_{\beta}(n^{-1} \parallel \boldsymbol{Y} - \boldsymbol{X}\beta \parallel_2^2 + \lambda \parallel \beta \parallel_1)$ είναι κυρτή και καθιστά δυνατό τον αποτελεσματικότερο υπολογισμό του εκτιμητή. Επιπρόσθετα, αυτό το πρόβλημα βελτιστοποίησης είναι ισοδύναμο με το

$$\widehat{\widehat{\beta_{primal}}}(R) = argmin_{\beta}(n^{-1} \parallel \boldsymbol{Y} - \boldsymbol{X}\beta \parallel_2^2)$$

με ένα προς ένα αντιστοιχία μεταξύ του λ και του R .

Λόγω της $l_1$-γεωμετρίας, η μέθοδος Lasso, εκτελεί επιλογή μεταβλητών λόγω του ότι μια εκτιμώμενη συνιστώσα μπορεί να είναι ακριβώς μηδέν. Το άθροισμα τετραγώνων των υπολοίπων φτάνει σε μια ελάχιστη τιμή, αν η περιμετρική γραμμή χτυπήσει την $l_1$-μπάλα στη γωνιά που αντιστοιχεί στην πρώτη συνιστώσα $\widehat{\widehat{\beta_{primal,1}}}$ και είναι ίση με μηδέν. Αυτό το φαινόμενο δε συμβαίνει με την παλινδρόμηση Ridge,

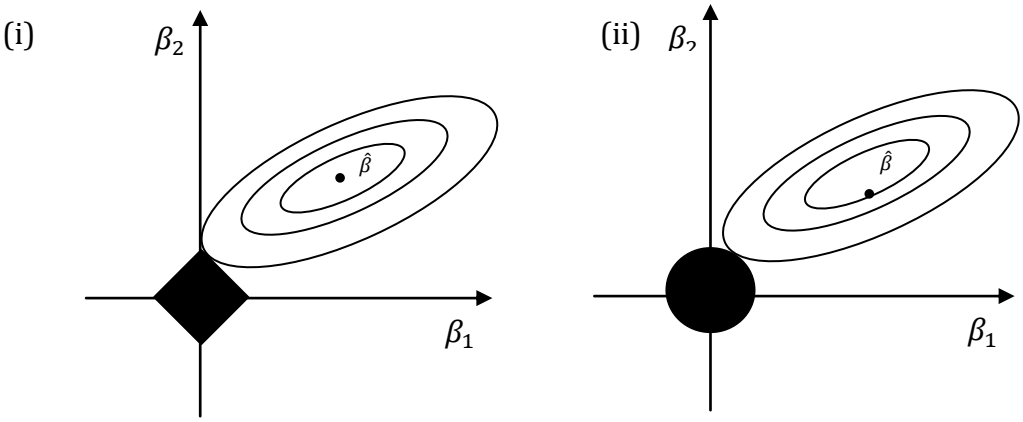
$$\widehat{\widehat{\beta_{Ridge}}}(\lambda) = argmin_{\beta}(n^{-1} \parallel \boldsymbol{Y} - \boldsymbol{X}\beta \parallel_2^2 + \lambda \parallel \beta \parallel_2^2)$$

με την ισοδύναμη αρχική λύση

$$\widehat{\widehat{\beta_{Ridge;primal}}}(R) = argmin_{\beta}(n^{-1} \parallel \boldsymbol{Y} - \boldsymbol{X}\beta \parallel_2^2)$$

με ένα προς ένα αντιστοιχία μεταξύ του λ και του R.

Γεωμετρική Ερμηνεία




**Σχήμα 12:** ***(i): περιμετρικές γραμμές του αθροίσματος των τετραγώνων τω υπολοίπων, με*** $\widehat{\boldsymbol{\beta}}$ ***να είναι ο εκτιμητής ελαχίστων τετραγώνων και η*** $\boldsymbol{l_1}$**-μπάλα που αντιστοιχεί στο πρόβλημα Lasso.** ***(ii):ανάλογα με το σχήμα (i), αλλά με την*** $\boldsymbol{l_2}$**-μπάλα που αντιστοιχεί στην παλινδρόμηση Ridge**.

Γενικά, η στατιστική συμπερασματολογία σε υψηλές διαστάσεις είναι εφικτή, όταν οδηγεί σε λογική ακρίβεια ή ασυμπτωτική συνοχή αν

$$\log(p) \cdot sparsity(\beta) \ll n$$

ανάλογα με το πώς ορίζουμε την σποραδικότητα και την υπό εξέταση ρύθμιση.

Η λύση $\beta_L$ 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 $l_2$ κανονικοποίηση της, 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 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.

Ο αλγόριθμος Lasso

Καθορίζουμε $\lambda \geq 0$. Το πρόβλημα μπορεί να εκφραστεί ως πρόβλημα ελαχίστων τετραγώνων με $2^p$ περιορισμούς ανισότητας. Έχουμε $L(\beta) = \sum_i (y_i - \sum_p \beta_p x_{ip})^2$ και τον περιορισμό $\sum_j |\beta_j| \leq \lambda$ που μπορεί να γραφτεί

επίσης ως $\delta_i^T \beta \leq \lambda$ για όλα τα i. Για ένα δοσμένο β, $E = \{i: \delta_i^T \beta = \lambda\}$ και $S = \{i: \delta_i^T \beta < \lambda\}$. Το σύνολο Ε είναι το σύνολο ισότητας που αντιστοιχεί στους περιορισμούς για τους οποίους ισχύει η ισότητα και το S είναι το χαλαρό σύνολο για τους περιορισμούς που δεν ισχύει ακριβώς η ισότητα. Συμβολίζουμε με $G_E$ τον πίνακα του οποίου οι γραμμές είναι τα $\delta_i, i \in E$, και $\mathbf{1}$ είναι ένα διάνυσμα που αποτελείται μόνο από τους αριθμούς 1 με μήκος ίσο με τον αριθμό των γραμμών του $G_E$.

Ο αλγόριθμος ξεκινά με $E = \{i_0\}$, όπου $\delta_{i_0} = sign(\hat{\beta})$. Λύνει το πρόβλημα ελαχίστων τετραγώνων $\delta_{i_0}^T \beta \leq \lambda$ και ελέγχει αν ισχύει $\sum_j |\beta_j| \leq \lambda$. Αν ισχύει τότε οι υπολογισμοί έχουν ολοκληρωθεί, διαφορετικά συνεχίζεται η διαδικασία μέχρι να ισχύει $\sum_j |\beta_j| \leq \lambda$.

Παρακάτω παρουσιάζεται μια περίληψη του αλγορίθμου:

1. Ξεκινώντας με $E = \{i_0\}$ και $\delta_{i_0} = sign(\hat{\beta}^0)$, όπου $\hat{\beta}^0$ είναι η συνολική εκτίμηση ελαχίστων τετραγώνων.
2. Βρίσκω το $\hat{\beta}$ που ελαχιστοποιεί το $L(\beta)$ υπό τον περιορισμό $G_E \leq \lambda \mathbf{1}$.
3. Καθώς ισχύει $\{\sum_j |\beta_j| > \lambda\}$
4. Προσθέτω το i στο σύνολο Ε όπου $\delta_i = sign(\hat{\beta})$. Βρίσκω το $\hat{\beta}$ που ελαχιστοποιεί το $L(\beta)$ υπό τον περιορισμό $G_E \leq \lambda \mathbf{1}$.

Αυτή η διαδικασία πρέπει α συγκλίνει σε ένα πεπερασμένο αριθμό βημάτων, αφού σε κάθε βήμα προσθέτεται το στοιχείο i στο σύνολο Ε και υπάρχουν συνολικά $2^p$ στοιχεία. Η τελευταία επανάληψη είναι η λύση στο αρχικό πρόβλημα και οι περιορισμοί ικανοποιούνται για το σύνολα Ε και S στη σύγκλιση.

## 3.15 Γεωμετρική ισοτιμία μεταξύ της λύσης υπερεπιπέδων Lasso παλινδρόμησης και SVM

Έστω $X = [x_1, x_2, \dots, x_N]$ πίνακας δείγματος όπου κάθε στήλη $x_i = (x_1, \dots, x_K)^{\mathrm{T}}$ αντιπροσωπεύει ένα διάνυσμα δείγματος από Κ χαρακτηριστικά. Διάνυσμα χαρακτηριστικών μπορεί να προκύψει από τη μεταφορά μιας γραμμής στον πίνακα του δείγματος, π.χ. $f_q = (x_{1q}, x_{2q}, \dots, x_{Nq})^{\mathrm{T}}$. Έστω επίσης $y = (y_1, y_2, .., y_N)^{\mathrm{T}}$ το διάνυσμα απόκρισης που περιλαμβάνει τις αποκρίσεις που αντιστοιχούν σε όλα τα δείγματα.

Τώρα έστω ένα δειγματικό χώρο του οποίου η κάθε βάση αντιπροσωπεύεται από ένα $x_i$ στον πίνακα δείγματος. Στο δειγματικό χώρο, τόσο τα χαρακτηριστικά $f_q$ καθώς και το διάνυσμα απόκρισης y μπορεί να θεωρηθεί ως

ένα σημείο σε αυτό το χώρο. Μπορεί να δειχθεί ότι η λύση της παλινδρόμησης Lasso έχει μια πολύ διαισθητική έννοια g στο δειγματικό χώρο: οι συντελεστές παλινδρόμησης μπορεί να θεωρηθεί ως το βάρος των χαρακτηριστικών διανυσμάτων στο δειγματικό χώρο. Επιπλέον, όλα τα μη μηδενικά σταθμισμένα διανύσματα χαρακτηριστικών είναι σε δύο παράλληλα υπερεπίπεδα στο δειγματικό χώρο. Αυτά διαθέτουν φορείς, μαζί με την μεταβλητή απόκρισης, καθορίζουν τις κατευθύνσεις των δύο αυτών υπερεπιπέδων. Αυτή η γεωμετρική άποψη μπορεί να προέρχονται από την ακόλουθη αναδιατύπωση της παλινδρόμησης Lasso (Perkins et al., 2003):

$$|\sum_i (y_i - \sum_p \beta_p x_{ip}) x_{iq}| \leq \frac{\lambda}{2}, \quad \forall q$$
$$\Rightarrow \left| f_q (y - [f_1, \dots, f_K]\beta) \right| \leq \frac{\lambda}{2}, \forall q$$

Είναι προφανές από την παραπάνω εξίσωση ότι το $f_q(y - [f_1, \dots, f_K]\beta)$ καθορίζει τον προσανατολισμό ενός διαχωρισμού υπερεπιπέδου. Μπορεί να δειχθεί ότι η ισότητα ισχύει μόνο για μη μηδενικά σταθμισμένο χαρακτηριστικά, και όλα τα μηδενικά σταθμισμένα διανύσματα χαρακτηριστικών είναι μεταξύ των υπερεπιπέδων με περιθώριο $\frac{\lambda}{2}$.

Τα διαχωριστικά υπερεπίπεδα λόγω (γραμμικό) SVM έχουν παρόμοιες ιδιότητες με εκείνες της παλινδρόμησης υπερεπιπέδων που περιγράφονται παραπάνω, αν και αυτά των SVM ορίζονται στον χαρακτηριστικό χώρο (στον οποία κάθε άξονας αντιπροσωπεύει ένα χαρακτηριστικό και κάθε σημείο εκπροσωπεί ένα δείγμα) αντί του δειγματικού χώρου. Σε ένα SVM, όλα τα μη μηδενικά σταθμισμένα δείγματα είναι επίσης σχετικά τα δύο περιθώρια $\frac{\lambda}{2}$ διαχωρισμού των υπερεπιπέδων (όπως συμβαίνει στη Lasso παλινδρόμηση), ενώ όλα τα μηδενικά σταθμισμένα δείγματα είναι τώρα έξω από το ζεύγος των υπερεπιπέδων. Είναι επίσης γνωστό ότι οι ταξινομητές υπερεπιπέδων σε SVM μπορεί να επεκταθούν και σε υπερεπιφάνειες εισάγοντας πυρήνες που ορίζονται, για διανύσματα δειγμάτων. Με αυτόν τον τρόπο, μπορεί να μοντελοποιήσει SVM μη γραμμικών εξαρτήσεων μεταξύ των δειγμάτων και του ορίου ταξινόμησης. Δεδομένης της ομοιότητας των γεωμετρικών δομών της Lasso παλινδρόμησης και SVM, μπορεί κανείς να επιδιώξει παράλληλα να εφαρμόσει παρόμοια «τεχνάσματα πυρήνα» των διανυσμάτων χαρακτηριστικών σε Lasso παλινδρόμηση, ώστε η δυνατότητα επιλογής χαρακτηριστικών μπορεί να επεκταθεί και σε μη γραμμικά μοντέλα.

# ΚΕΦΑΛΑΙΟ 4

## Κριτήρια Αξιολόγησης Μοντέλου

Η αξιολόγηση ενός μοντέλου είναι πολύ σημαντικό κεφάλαιο της στατιστικής ανάλυσης, γιατί κατευθύνει την επιλογή της μεθόδου εκμάθησης αλλά και την επιλογή του μοντέλου και μας δίνει ένα μέτρο ποιότητας του τελικώς επιλεγμένου μοντέλου.

Παρακάτω αναφέρουμε μεθόδους αξιολόγησης της απόδοσης του επιλεγμένου μοντέλου.

### 4.1 Μεροληψία, Διασπορά, Περιπλοκότητα

Θεωρούμε την περίπτωση μιας ποσοτικής μεταβλητής απόκρισης. Έχουμε μια μεταβλητή $Y$-στόχο, ένα διάνυσμα εισόδων $X$ και μια πρόβλεψη του μοντέλου $\hat{f}(X)$ που έχει υπολογιστεί από σύνολο εκπαίδευσης $T$. Η λειτουργία απώλειας για την μέτρηση των σφαλμάτων μεταξύ $Y$ και $\hat{f}(X)$συμβολίζεται με $L(Y, \hat{f}(X))$ και

$$L\left(Y, \hat{f}(X)\right) = \begin{cases} (Y - \hat{f}(X))^2 \text{ , } \tau\varepsilon\tau\rho\alpha\gamma\omega\nu\iota\kappa\acute{o}\ \sigma\phi\acute{\alpha}\lambda\mu\alpha\ (squared\ error) \\ \left|Y - \hat{f}(X)\right|, \quad \alpha\pi\acute{o}\lambda\upsilon\tau o\ \sigma\phi\acute{\alpha}\lambda\mu\alpha\ (absolute\ error) \end{cases}$$

Το σφάλμα γενίκευσης (generalization error), γνωστό και ως σφάλμα δοκιμών (test error) είναι μια λειτουργία που μετρά πόσο καλά μια μηχανή μάθησης γενικεύεται σε πρωτοεμφανιζόμενα δεδομένα. Μετράται ως η απόσταση μεταξύ του σφάλματος σχετικά με το σύνολο εκπαίδευσης και το σύνολο των δοκιμών και είναι κατά μέσο όρο για το σύνολο των πιθανών στοιχείων κατάρτισης που μπορούν να δημιουργηθούν μετά από κάθε επανάληψη της διαδικασίας μάθησης. Έχει αυτό το όνομα επειδή η λειτουργία αυτή δείχνει την ικανότητα μιας μηχανής που μαθαίνει με τον προκαθορισμένο αλγόριθμο να συναγάγει έναν κανόνα (ή να γενικεύσει) που χρησιμοποιείται από το μηχάνημα να παράγει δεδομένα που βασίζονται μόνο σε μερικά παραδείγματα.

Το θεωρητικό μοντέλο υποθέτει μια κατανομή πιθανοτήτων των παραδειγμάτων και μια συνάρτηση δίνοντας τον ακριβή στόχο. Το μοντέλο μπορεί επίσης να περιλαμβάνει το θόρυβο στο παράδειγμα (στην είσοδο και/ή έξοδο). Το γενικευμένο σφάλμα συνήθως ορίζεται ως η αναμενόμενη τιμή του τετραγώνου της διαφοράς μεταξύ της συνάρτησης και του ακριβές στόχου (μέσο τετραγωνικό σφάλμα-mean squared error). Σε συγκεκριμένες περιπτώσεις , η κατανομή και ο στόχος είναι άγνωστα και χρησιμοποιούνται στατιστικές εκτιμήσεις.

Η απόδοση ενός αλγορίθμου μηχανικής μάθησης μετράται με γραφήματα των γενικευμένων τιμών σφάλματος μέσα από τη διαδικασία της μάθησης και καλούνται καμπύλες μάθησης.

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

Μία σχετική ποσότητα είναι το αναμενόμενο σφάλμα πρόβλεψης (expected prediction error):

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

Το σφάλμα εκπαίδευσης είναι η μέση απώλεια πάνω στο δείγμα εκπαίδευσης

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

Καθώς το μοντέλο γίνεται πιο πολύπλοκο, χρησιμοποιεί τα δεδομένα εκπαίδευσης περισσότερο και είναι σε θέση να προσαρμοστούν σε περισσότερο πολύπλοκες υποκείμενες δομές. Ως εκ τούτου, υπάρχει μείωση στην μεροληψία αλλά αύξηση στη διασπορά διακύμανσης. Υπάρχει κάποια 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. Το σφάλμα εκπαίδευσης μειώνεται σταθερά με την πολυπλοκότητα του μοντέλου, συνήθως πέφτει στο μηδέν εάν αυξηθεί αρκετά η πολυπλοκότητα του μοντέλου. Ωστόσο, σε ένα μοντέλο με μηδενικό σφάλμα εκπαίδευσης παρατηρούμε το πρόβλημα υπερπροσαρμογής στην κατάρτιση δεδομένων και τυπικά δεν θα γενικευθεί καλά. Το ίδιο συμβαίνει για μια ποιοτική ή κατηγορηματική μεταβλητή απόκρισης.

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.

Κατά την αξιολόγηση του μοντέλου είναι σημαντικό να έχουμε υπόψη μας δύο σημαντικούς στόχους:

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

Εκτίμηση Μοντέλου (Model Assessment): έχοντας επιλέξει ένα τελικό μοντέλο, εκτιμάται το σφάλμα 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 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.

Η προσέγγιση στο στάδιο της επικύρωσης γίνεται είτε αναλυτικά (AIC, BIC, MDL, SRM) ή από την αποτελεσματική επαναχρησιμοποίηση του δείγματος (cross-validation και 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.

## 4.2 Διασταυρωμένη Επικύρωση
## Cross-Validation

Διασταυρωμένη επικύρωση, μερικές φορές ονομάζεται εκτίμηση περιστροφής, είναι ένα μοντέλο τεχνικής επικύρωσης για την αποτίμηση πώς τα αποτελέσματα της στατιστικής ανάλυσης θα γενικευθούν σε ένα ανεξάρτητο σύνολο δεδομένων. Χρησιμοποιείται κυρίως σε χώρους όπου ο στόχος είναι η πρόβλεψη, δηλαδή θέλει να εκτιμήσει με ακρίβεια το πώς ένα προγνωστικό μοντέλο θα εκτελεστεί στην πράξη. Αξίζει να επισημανθεί ότι σε ένα πρόβλημα πρόβλεψης, στο μοντέλο δίνεται συνήθως ένα γνωστό σύνολο δεδομένων στα οποία θα εκτελεστεί εκπαίδευση (σύνολο δεδομένων κατάρτισης/εκπαίδευσης), καθώς και ένα σύνολο άγνωστων δεδομένων (ή πρωτοεμφανιζόμενα στοιχεία) έναντι του οποίου το μοντέλο ελέγχεται (σύνολο δοκιμής). Ο στόχος της διασταυρωμένης επικύρωσης είναι να καθορίσει ένα σύνολο δεδομένων για να ελέγξει το μοντέλο στη φάση της κατάρτισης (δηλαδή, το σύνολο δεδομένων επικύρωσης), προκειμένου να περιοριστούν τα προβλήματα, όπως είναι το πρόβλημα της υπερπροσαρμογή, δίνοντας μια εικόνα για το πώς το μοντέλο θα γενικευτεί σε ένα ανεξάρτητο σύνολο δεδομένων (δηλαδή , ένα άγνωστο σύνολο δεδομένων, για παράδειγμα από ένα πραγματικό πρόβλημα ).

Ένας γύρος της διασταυρωμένης επικύρωσης περιλαμβάνει διαχωρισμό ενός δείγματος των δεδομένων σε συμπληρωματικά υποσύνολα, που εκτελεί την ανάλυση σε ένα υποσύνολο (ονομάζεται σύνολο εκπαίδευσης) και την επικύρωση της ανάλυσης από την άλλη υποομάδα (ονομάζεται σύνολο επικύρωσης ή ομάδα ελέγχου). Για να μειωθεί η μεταβλητότητα, εκτελούνται πολλαπλές επαναλήψεις διασταυρωμένης επικύρωσης χρησιμοποιώντας διαφορετικές κατατμήσεις, και παίρνουμε το μέσο όρο των αποτελεσμάτων επικύρωσης από τις διαφορετικές επαναλήψεις.

Η διασταυρωμένη επικύρωση είναι σημαντική για τις υποθέσεις που προτείνονται από τα δεδομένα, κυρίως όταν τα δείγματα είναι επικίνδυνα, δαπανηρά ή αδύνατα στη συλλογή.

Η μέθοδος αυτή υπολογίζει άμεσα το αναμενόμενο σφάλμα $Err = E\left[L\left(Y, \hat{f}(X)\right)\right]$, το μέσο σφάλμα γενίκευσης όταν η μέθοδος $\hat{f}(X)$ εφαρμόζεται σε ένα ανεξάρτητο δείγμα δοκιμής από την κοινή κατανομή των X και Y.

Ας υποθέσουμε ότι έχουμε ένα μοντέλο με μια ή περισσότερες άγνωστες παραμέτρους, καθώς και ένα σύνολο δεδομένων στο οποίο το μοντέλο μπορεί να προσαρμοστεί (το σύνολο δεδομένων εκπαίδευσης). Η διαδικασία προσαρμογής

βελτιστοποιεί τις παραμέτρους του μοντέλου για να κάνει το μοντέλο να προσαρμοστεί στα δεδομένα εκπαίδευσης όσο το δυνατόν καλύτερα. Αν στη συνέχεια να λάβει ένα ανεξάρτητη δείγμα της επικύρωσης δεδομένων από τον ίδιο πληθυσμό, όπως τα δεδομένα εκπαίδευσης, θα αποδειχθεί γενικά ότι το μοντέλο δεν ταιριάζει με τα δεδομένα επικύρωσης, καθώς και ότι ταιριάζει με τα δεδομένα εκπαίδευσης. Αυτό ονομάζεται υπερπροσαρμογή, και είναι ιδιαίτερα πιθανό να συμβεί όταν το μέγεθος του συνόλου δεδομένων εκπαίδευσης είναι μικρό, ή όταν ο αριθμός των παραμέτρων του μοντέλου είναι μεγάλος. Διασταυρωμένη επικύρωση είναι ένας τρόπος για να προβλέψουμε την προσαρμογή του μοντέλου σε ένα υποθετικό σύνολο επικύρωσης, όταν δεν είναι διαθέσιμο ένα σαφές σύνολο επικύρωσης.

### 4.2.1 k- φορές Διασταυρωμένη Επικύρωση

Στην k-φορές διασταυρωμένη επικύρωση (k-fold cross validation) το αρχικό δείγμα χωρίζεται τυχαία σε k επιμέρους δείγματα ίσου μεγέθους. Από τα επιμέρους δείγματα k, ένα ενιαίο υποσύνολο από αυτά διατηρείται για τη δοκιμή του μοντέλου ως τα δεδομένα επικύρωσης, και τα υπόλοιπα k - 1 δείγματα χρησιμοποιούνται ως δεδομένα εκπαίδευσης. Η διαδικασία διασταυρωμένης επικύρωσης επαναλαμβάνεται k φορές (folds), με το καθένα από τα επιμέρους δείγματα k να χρησιμοποιείται ακριβώς μία φορά ως τα δεδομένα επικύρωσης. Από τα k αποτελέσματα από τις επαναλήψεις, γίνεται στη συνέχεια ο μέσος όρος (ή αλλιώς συνδυάζονται) για να παραχθεί μια ενιαία εκτίμηση. Το πλεονέκτημα αυτής της μεθόδου είναι ότι όλες οι παρατηρήσεις που χρησιμοποιούνται τόσο για την εκπαίδευση αλλά και για την επικύρωση, και κάθε παρατήρηση χρησιμοποιείται για την επικύρωση ακριβώς μια φορά. Η 10–φορές διασταυρωμένη επικύρωση χρησιμοποιείται συνήθως, αλλά γενικά το k παραμένει μια αόριστη παράμετρος.

Σε στρωματοποιημένη k- φορές διασταυρωμένη επικύρωση, οι επαναλήψεις επιλέγονται έτσι ώστε η μέση τιμή απόκρισης να είναι περίπου ίση σε όλες τις φορές. Στην περίπτωση μιας διχοτόμου ταξινόμησης, αυτό σημαίνει ότι κάθε επανάληψη περιέχει περίπου τις ίδιες αναλογίες των δύο τύπων κατηγορίας.

Η μέθοδος της διασταυρωμένης επικύρωσης μπορεί να χρησιμοποιηθεί για τη σύγκριση των επιδόσεων των διαφόρων διαδικασιών μοντελοποίησης πρόβλεψης. Για παράδειγμα, ας υποθέσουμε ότι ενδιαφέρομαστε για την οπτική αναγνώριση χαρακτήρων, και εξετάζουμε τα δεδομένα χρησιμοποιώντας είτε τις μηχανές διανυσματικής υποστήριξης (SVMs) ή τη μέθοδο των k πλησιέστερων γειτόνων (kNN) για να προβλέψει τον αληθινό χαρακτήρα από μια εικόνα ενός χειρογράφου χαρακτήρα. Χρησιμοποιώντας cross-validation, θα μπορούσαμε να

συγκρίνουμε αντικειμενικά αυτές τις δύο μεθόδους στο πλαίσιο των αντίστοιχων κλασμάτων των εσφαλμένα ταξινομημένων χαρακτήρων.

### 4.2.2 Εφαρμογές

Cross validation, μπορεί επίσης να χρησιμοποιηθεί στην επιλογή χαρακτηριστικών (feature selection). Ας υποθέσουμε ότι χρησιμοποιούν τα επίπεδα έκφρασης 20 πρωτεϊνών για να προβλέψει εάν ένας ασθενής με καρκίνο θα ανταποκριθεί σε ένα φάρμακο. Μια πρακτική μέθοδος θα είναι να καθοριστεί ποιο υποσύνολο των 20 χαρακτηριστικών θα πρέπει να χρησιμοποιούνται για να παράγουν το καλύτερο μοντέλο πρόβλεψης. Για τις περισσότερες διαδικασίες μοντελοποίησης, αν συγκρίνουμε το χαρακτηριστικό υποσύνολο με τη χρήση των ποσοστών σφάλματος στο δείγμα, η καλύτερη απόδοση θα συμβεί όταν όλα τα 20 χαρακτηριστικά χρησιμοποιούνται. Ωστόσο, στο πλαίσιο της διασταυρωμένης επικύρωσης, το μοντέλο με την καλύτερη προσαρμογή, θα περιλαμβάνει μόνο ένα υποσύνολο από τα χαρακτηριστικά που θεωρούνται πραγματικά σημαντικά.

## 4.3 Accuracy και Precision

Στους τομείς της Επιστήμης, της Μηχανικής, της Βιομηχανίας, και στη Στατιστική, η ακρίβεια (accuracy) του συστήματος μέτρησης είναι ο βαθμός της εγγύτητας των μετρήσεων της ποσότητας με την πραγματική (αληθής) τιμή της ποσότητας του. Η ακρίβεια (precision) του συστήματος μέτρησης, που σχετίζεται με την επαναληψιμότητα και την αναπαραγωγικότητα, είναι ο βαθμός στον οποίο οι επανειλημμένες μετρήσεις υπό αμετάβλητες συνθήκες δείχνουν τα ίδια αποτελέσματα. Αν και οι δύο αυτοί ορισμοί accuracy και precision μπορεί να είναι συνώνυμες στην καθομιλουμένη χρήση, έρχονται σκόπιμα σε αντίθεση στο πλαίσιο της επιστημονικής μεθόδου.

Ένα σύστημα μέτρησης μπορεί να είναι accurate, αλλά όχι precise, και το αντίθετο ή και τα δύο μαζί. Για παράδειγμα, αν ένα πείραμα περιέχει ένα συστηματικό σφάλμα, τότε αυξάνοντας το μέγεθος του δείγματος γενικά αυξάνει την precision, αλλά δεν βελτιώνει την accuracy. Το αποτέλεσμα θα είναι μια συνεπής και ανακριβή σειρά των αποτελεσμάτων από το ελαττωματικό πείραμα. Η εξάλειψη του συστηματικού σφάλματος βελτιώνει την accuracy, αλλά δεν αλλάζει την precision.

Ένα σύστημα μέτρησης θεωρείται έγκυρο εάν είναι ακριβές και σαφές. Σχετικοί όροι περιλαμβάνουν την μεροληψία (bias-μη τυχαία ή κατευθυνόμενη επίδραση που προκαλείται από έναν παράγοντα ή παράγοντες που δεν σχετίζονται με την ανεξάρτητη μεταβλητή) και το σφάλμα (τυχαία μεταβλητότητα). Εκτός από την ορθότητα (trueness) και την ακρίβεια (precision), οι μετρήσεις μπορεί επίσης να έχουν μια ανάλυση μέτρησης (measurement resolution), η οποία είναι η μικρότερη αλλαγή στην υποκείμενη φυσική ποσότητα που παράγει μια απόκριση στη μέτρηση.

### 4.3.1 Ποσοτικοποίηση

Στην ιδανική περίπτωση μια συσκευή μέτρησης είναι τόσο ακριβή και σαφή, με τις μετρήσεις κοντά και σφιχτά συγκεντρωμένες γύρω από τη γνωστή τιμή. Η ορθότητα και η ακρίβεια της διαδικασίας μέτρησης που καθορίζονται συνήθως με κάποια επανειλημμένη μέτρηση από ένα ανιχνεύσιμο πρότυπο αναφοράς . Τέτοια πρότυπα που ορίζονται στο Διεθνές Σύστημα Μονάδων (SI) και διατηρήθηκε από τους εθνικούς οργανισμούς τυποποίησης.

Το ίδιο ισχύει και όταν οι μετρήσεις επαναλαμβάνονται και κατά μέσο όρο. Στην περίπτωση αυτή, ο όρος τυπικό σφάλμα εφαρμόζεται σωστά: η ακρίβεια του μέσου είναι ίση με τη γνωστή τυπική απόκλιση της μεθόδου διαιρεμένη με την τετραγωνική ρίζα του μέσου όρου του αριθμού των μετρήσεων. Περαιτέρω, το κεντρικό οριακό θεώρημα δείχνει ότι η κατανομή πιθανότητας των μετρήσεων κατά μέσο όρο θα είναι πιο κοντά σε μια κανονική κατανομή από εκείνη των μεμονωμένων μετρήσεων.

Όσον αφορά την ακρίβεια (accuracy) μπορούμε να διακρίνουμε:

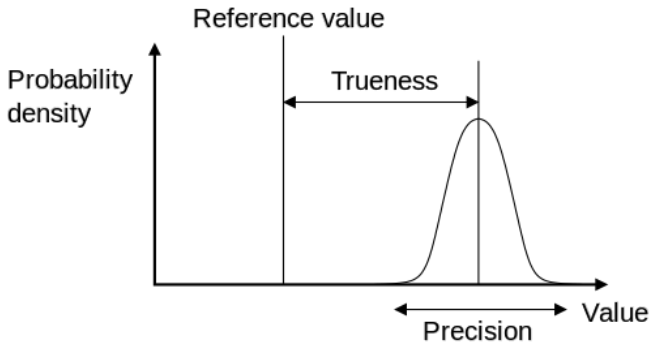
- τη διαφορά μεταξύ του μέσου όρου των μετρήσεων και της τιμής αναφοράς, την μεροληψία (bias). Ο καθορισμός και η διόρθωση της μεροληψίας είναι απαραίτητη για την ταξινόμηση.
- το συνδυασμένο αποτέλεσμα και την ακρίβεια (precision).

Η ακρίβεια (precision) είναι μερικές φορές στρωματοποιημένη σε:

- Επαναληψιμότητα: η μεταβολή που προκύπτει όταν όλες οι προσπάθειες που καταβάλλονται για να κρατηθούν σταθερές οι συνθήκες, χρησιμοποιώντας το ίδιο όργανο και το χειριστή, και επαναλαμβάνοντας σε σύντομο χρονικό διάστημα.
- Αναπαραγωγιμότητα: η μεταβολή που προκύπτει χρησιμοποιώντας την ίδια διαδικασία μέτρησης μεταξύ των διαφόρων μέσων και φορέων, καθώς και σε μεγαλύτερες χρονικές περιόδους.

### 4.3.2 Ορολογία του ISO 5725

Σύμφωνα με το πρότυπο ISO 5725-1, οι όροι ορθότητα (trueness) και ακρίβεια (precision) χρησιμοποιούνται για να περιγράψουν την ακρίβεια (accuracy) της μέτρησης. Η ορθότητα αναφέρεται στην εγγύτητα του μέσου όρου των αποτελεσμάτων των μετρήσεων με την πραγματική (αληθή) αξία και ακρίβεια και αναφέρεται στην εγγύτητα της συμφωνίας στο πλαίσιο των επιμέρους αποτελεσμάτων.




Ως εκ τούτου, σύμφωνα με το πρότυπο ISO, ο όρος accuracy αναφέρεται στους όρους ορθότητα (trueness) και ακρίβεια (precision).

### 4.3.3 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, προσμετρείται ως ψευδώς θετική.

Δεδομένου ενός ταξινομητή και μια σειράς από περιπτώσεις (στο σύνολο δοκιμής), μπορεί να κατασκευαστεί ένας 2X2 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.

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

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

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

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

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

ΑΚΡΙΒΕΙΑ (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.

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

Όταν έχουμε ακρίβεια 100% σημαίνει ότι οι μετρούμενες τιμές είναι ακριβώς οι ίδιες με τις αληθείς τιμές.

ΑΚΡΙΒΕΙΑ (PRECISION) ή αλλιώς θετική προγνωστική αξία είναι το ποσοστό των αληθώς θετικών έναντι όλων των θετικών αποτελεσμάτων (αληθώς θετικά και ψευδώς θετικά).

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

## 4.4 Ευαισθησία και Ειδικότητα (Sensitivity and Specificity)

Η ευαισθησία (sensitivity) και η ειδικότητα (specificity) είναι στατιστικές μετρήσεις των επιδόσεων μιας δοκιμής δυαδικής ταξινόμησης, γνωστή στο πεδίο της Στατιστικής ως συνάρτηση ταξινόμησης. Η ευαισθησία (ονομάζεται επίσης το πραγματικό θετικό ποσοστό ή το ποσοστό ανάκλησης σε ορισμένους τομείς) μετρά την αναλογία των πραγματικά θετικών που έχουν αναγνωριστεί σωστά (π.χ. το ποσοστό των ασθενών που έχουν αναγνωριστεί σωστά ως έχει η κατάσταση). Ειδικότητα (μερικές φορές ονομάζεται το πραγματικό αρνητικό ποσοστό) μετρά το ποσοστό των αρνητικών που έχουν αναγνωριστεί σωστά (π.χ. το ποσοστό των υγιή ανθρώπων που έχουν αναγνωριστεί σωστά). Τα δύο αυτά μέτρα συνδέονται στενά με τις έννοιες του τύπου λάθους I και τύπου λάθους II. Ένα τέλειο μοντέλο πρόβλεψης θα περιγραφεί με 100% ευαισθησία και με 100 % ειδικότητα. Ωστόσο, θεωρητικά κάθε μοντέλο πρόβλεψης θα διαθέτει ένα ελάχιστο όριο σφάλματος γνωστό ως το ποσοστό σφάλματος Bayes.

ΕΥΑΙΣΘΗΣΙΑ (SENSITIVITY) ή αλλιώς το ποσοστό των αληθές θετικών αποτελεσμάτων, δηλαδή των θετικών ενδείξεων στον πληθυσμό, δηλαδή η πιθανότητα το τεστ να είναι θετικό δεδομένου ότι κάποιος έχει το χαρακτηριστικό που εξετάζουμε και δίνεται από τον τύπο:

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

ΕΙΔΙΚΟΤΗΤΑ (SPECIFICITY) ή αλλιώς το ποσοστό των αληθώς αρνητικών αποτελεσμάτων, δηλαδή των αρνητικών ενδείξεων στον πληθυσμό, δηλαδή η πιθανότητα το τεστ να είναι αρνητικό δεδομένου ότι κάποιος δεν έχει το χαρακτηριστικό που εξετάζουμε και υπολογίζεται από τον τύπο:

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

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.
Ισχύει

$$TPR = 1 - FNR$$

Όπου

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

## 4.5 Επιπολασμός (Prevalence)

Ως επιπολασμό, ορίζουμε το σύνολο των θετικών προς το σύνολο του πληθυσμού και υπολογίζεται ως εξής:

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

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

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

Συνοψίζοντας όλα τα παραπάνω μέτρα παίρνουμε τον παρακάτω πίνακα:

| Αποτελέσματα του τεστ | Πραγματική Τιμή | | | |
|---|---|---|---|---|
| | Θετικό (P) | TP | FP | Θετική Προγνωστική Αξία (PPV/PRECITION) |
| | Αρνητικό (N) | FP | TN | Αρνητική Προγνωστική Αξία (NPV) |
| | | Ευαισθησία Sensitivity | Ειδικότητα Specificity | ΑΚΡΙΒΕΙΑ ACCURACY |

## 4.6 ROC Καμπύλες

Η πραγματοποίηση προβλέψεων αποτελεί ένα από τα σημαντικότερα μελήματα κάθε επιχείρησης και επιστημονικού πεδίου σχετικά με την αναζήτηση πληροφορίας. Το γεγονός αυτό καθιστά την εξασφάλιση προγνωστικής ακρίβειας απαραίτητη στον σχεδιασμό και την σύγκριση μοντέλων, αλγορίθμων και τεχνολογιών που παράγουν προβλέψεις. Οι καμπύλες ROC (Receiver Operating Characteristic-Λειτουργικό Χαρακτηριστικό Δέκτη) συμβάλλουν στην εξασφάλιση της επιθυμητής ακρίβειας στις προβλέψεις. Έτσι, αποτελούν χρήσιμη τεχνική για την οργάνωση, επιλογή και απεικόνιση ταξινομητών με βάση τη γραφική τους παράσταση.

Ιστορικά οι καμπύλες ROC χρονολογούνται στις αρχές τις δεκαετίας του ‘50, όταν οι φοιτητές του τμήματος ηλεκτρολόγων μηχανικών του πανεπιστημίου του Michigan -WW Peterson και TG Birdsall - εφάρμοσαν τη στατιστική θεωρία αποφάσεων σε προβλήματα της θεωρίας λήψης σημάτων (signal detection theory). Αρχικά, προτάθηκαν ως γραφική μέθοδος μέτρησης της ποιότητας λήψης σήματος από ένα δείκτη σε ατελή διαγνωστικά συστήματα. 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 ROC καμπύλη ορίζεται ως το μοναδιαίο τετράγωνο, [0,1]X[0,1], το οποίο ξεκινά από το σημείο (0, 0) - όταν το σημείο απόφασης είναι μεγαλύτερο από όλες τις μετρήσεις θορύβου και σήματος – για να καταλήξει στο σημείο (1, 1) – στην 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 και σήματος και χρησιμοποιείται συχνά στην Στατιστική Συμπερασματολογία των ROC καμπυλών.

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.

### 4.6.1 Βασικές Έννοιες

Συχνά στην ιατρική έρευνα αναπτύσσονται διαγνωστικοί έλεγχοι σε συνεχή ή διακριτή κλίμακα για το διαχωρισμό των πληθυσμών υγιών και ασθενών. Η κλινική χρησιμότητα μιας διαγνωστικής δοκιμασίας (εργαστηριακό αποτέλεσμα ή κλινικό εύρημα) προσδιορίζεται κατά κύριο λόγο από τη διακριτική της ικανότητα, δηλαδή από την ακρίβεια με την οποία διακρίνει αρρώστους με ή χωρίς το υπό διερεύνηση νόσημα, για το οποίο αυτή επιτελείται.

Στην θεωρία ανίχνευσης σημάτων, μια ROC καμπύλη είναι μια γραφική παράσταση της ευαισθησίας (sensitivity) ή των αληθών θετικών, έναντι του 1-ειδικότητα (specificity) ή ψευδώς θετικών, για ένα σύστημα δυαδικής ταξινόμησης, καθώς το όριο ταξινόμησης ποικίλει. Ο δείκτης λειτουργικού χαρακτηριστικού μπορεί ισοδύναμα να εκπροσωπηθεί με τη γραφική παράσταση του ποσοστού των αληθώς θετικών (TPR = True Positive Rate) έναντι του ποσοστού των ψευδώς θετικών (FPR =False Positive Rate). Είναι ακόμη γνωστή ως καμπύλη σχετικού λειτουργικού χαρακτηριστικού, αφού

αποτελεί τη σύγκριση δύο λειτουργικών χαρακτηριστικών (TPR, FPR ) καθώς το κριτήριο αλλάζει.

Τώρα, ας εξετάσουμε ένα πρόβλημα πρόβλεψης διπλής κλάσης (δυαδική ταξινόμηση), στο οποίο το αποτέλεσμα χαρακτηρίζεται ως θετική (P) ή αρνητική (N) κλάση. Υπάρχουν τέσσερις πιθανές εκβάσεις για ένα δυαδικό ταξινομητή. Αν το αποτέλεσμα της πρόβλεψης είναι P και η πραγματική τιμή είναι επίσης P, αυτό ονομάζεται αληθώς θετικό. Ωστόσο, εάν η πραγματική τιμή είναι N, λέγεται ψευδώς θετικό. Αντίθετα, ένα αληθώς αρνητικό έχει προκύψει όταν τόσο το αποτέλεσμα της πρόβλεψης όσο και η πραγματική τιμή είναι N ενώ η πραγματική τιμή είναι P.

### 4.6.2 Βασική Ορολογία

Ευαισθησία ή Ποσοστό Θετικών: $TPR = \frac{TP}{P} = \frac{TP}{TP+FN}$

Ποσοστό Ψευδώς Θετικών: $FPR = \frac{FP}{P} = \frac{FP}{FP+TN}$

Ειδικότητα ή Ποσοστό Αληθώς Αρνητικών: $SPC = \frac{TN}{N} = \frac{TP}{FP+TN} = 1 - FPR$

Ακρίβεια (Accuracy): $accuracy = \frac{TP+TN}{P+N}$

Θετική Προβλεπόμενη Τιμή: $PPV = \frac{TP}{TP+FP}$

Αρνητική Προβλεπόμενη Τιμή: $NPV = \frac{TN}{TN+FN}$

Αρνητικός Λόγος Πιθανοφανειών: $LR = \lambda = \frac{FNR}{TNR}$

Επιπολασμός (Prevalence): $PRV = \frac{TP+FN}{P+N}$

Συντελεστής Συσχέτισης του Matthews: $MCC = \frac{TP \times TN - FP \times FN}{\sqrt{PNP'N'}}$

F1-score: $F_1 = \frac{2TP}{P+P'}$

Από ένα πίνακα συνάφειας μπορούμε να βρούμε πολλές μετρικές:

- Η ευαισθησία (TPR) καθορίζει ένα διαγνωστικό test που ταξινομεί σωστά τα θετικά περιστατικά μεταξύ όλων των διαθέσιμων θετικών δειγμάτων

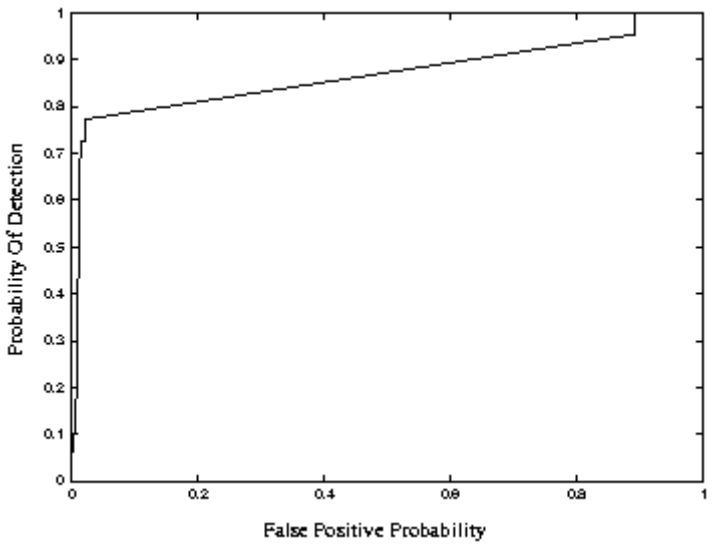
κατά τη διάρκεια δοκιμής. Ο όρος ευαισθησία χρησιμοποιείται για να ορίσει τη μικρότερη συγκέντρωση μιας ουσίας που μπορεί η μέθοδος να ανιχνεύσει και να μετρήσει. Από την άλλη πλευρά, ο όρος 1- ειδικότητα (FPR) ορίζει πόσα λανθασμένα θετικά αποτελέσματα εμφανίζονται μεταξύ όλων των διαθέσιμων αρνητικών δειγμάτων κατά τη διάρκεια της δοκιμής. Ο όρος ειδικότητα χρησιμοποιείται για να ορίσει την ιδιότητα της μεθόδου να ανιχνεύσει και να μετράει μόνο την ουσία που θέλει να μετρήσει.

- Η θετική προβλεπόμενη τιμή (PPV) ερμηνεύεται ως η πιθανότητα εμφάνισης θετικού περιστατικού μεταξύ όλων των θετικών προβλέψεων και η αρνητική προβλεπόμενη τιμή (NPV) ως η πιθανότητα εμφάνισης αρνητικού περιστατικού μεταξύ όλων των αρνητικών προβλέψεων.
- Ο θετικός λόγος πιθανοφανειών (L) εκφράζει πόσες φορές πιο συχνά εμφανίζεται το θετικό αποτέλεσμα στους πάσχοντες σε σχέση με τους μη πάσχοντες από το νόσημα που διερευνάται. Ο αρνητικός λόγος πιθανοφανειών (λ) εκφράζει πόσες φορές πιο συχνά εμφανίζεται το αρνητικό αποτέλεσμα στους μη πάσχοντες σε σχέση με τους πάσχοντες από το νόσημα που διερευνάται.

### 4.6.3 Σχεδιασμός Καμπύλης ROC

Η καμπύλη ROC εκφράζει τη σχέση του ποσοστού των αληθώς θετικών (%ΑΘ) και ψευδώς θετικών (%ΨΘ=1-%ΑΘ) αποτελεσμάτων της διαγνωστικής δοκιμασίας, καθώς μεταβάλλεται προοδευτικά προς μια κατεύθυνση το διαχωριστικό όριο (ΔΟ). Ένας χώρος ROC ορίζεται από το %ΨΘ στον x και το %ΑΘ στον y άξονα αντίστοιχα και κάθε σημείο της εμπειρικής καμπύλης ROC προσδιορίζεται από ένα ορισμένο ζεύγος (%ΑΘ, %ΨΘ ).

Η ROC καμπύλη όπως είδαμε και πιο πάνω ορίζεται ως το μοναδιαίο τετράγωνο [0,1]X[0,1], το οποίο ξεκινά από το σημείο (0, 0) για να καταλήξει στο σημείο (1,1). Η καλύτερη δυνατή μέθοδος πρόβλεψης θα απέφερε ένα σημείο στην επάνω αριστερή γωνιά ή τη συντεταγμένη (0,1) του χώρου ROC, που αντιπροσωπεύει το 100% ευαισθησία (μηδέν ψευδώς αρνητικά) και 100% ειδικότητα (μηδέν ψευδώς θετικά). Μια εντελώς τυχαία εικασία θα έδινε ένα σημείο κατά μήκος μιας διαγώνιας γραμμής (γραμμή της μη διάκρισης), από την κάτω αριστερή προς την πάνω δεξιά γωνιά.




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).

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.

### 4.6.4 Περιοχές και σημεία που έχουν προβλεπτικές ικανότητες

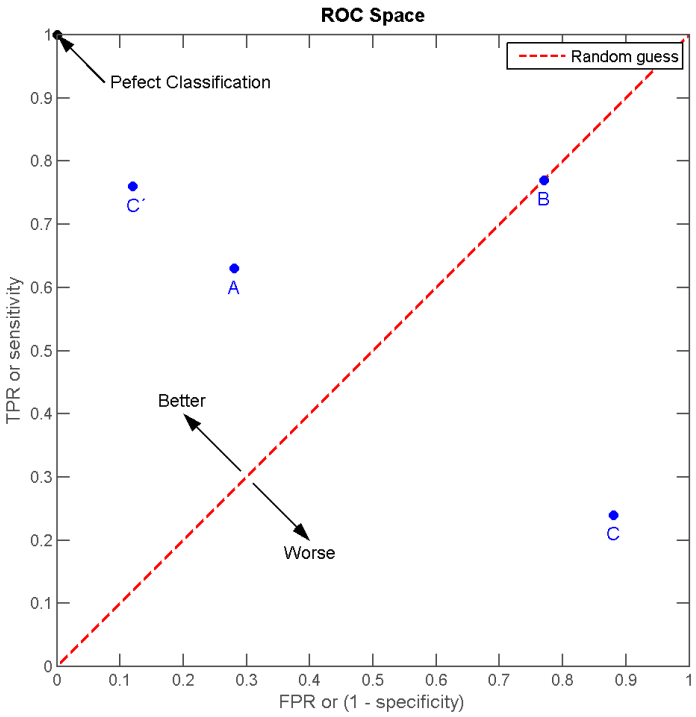
Σε ένα επίπεδο όπου απεικονίζεται μια ROC καμπύλη, ορίζονται επιμέρους σημεία και περιοχές με ιδιαίτερη προβλεπτική ικανότητα. Ορισμένες εκδοχές θεωρούνται οι εξής:

- Το σημείο τομής της ROC καμπύλης με την κάθετη διαγώνιο.
- Η περιοχή μεταξύ της καμπύλης ROC και της διαγωνίου.
- Η περιοχή κάτω από την ROC καμπύλη (AUC).

- d', η απόσταση μεταξύ του μέσου της κατανομής στο σύστημα υπό θόρυβο μείον το μέσο της κατανομής στο σύστημα υπό σήματα, δια την τυπική απόκλιση με την προϋπόθεση ότι οι δύο αυτές κατανομές είναι κανονικές με την ίδια τυπική απόκλιση. Βάση των υποθέσεων αυτών, μπορεί να αποδειχθεί ότι το σχήμα της ROC καμπύλης εξαρτάται μόνο από την d'.

### 4.6.5 Μέθοδος Αντικατοπτρισμού σημείου

Η διαγώνιος χωρίζει το χώρο ROC. Τα σημεία πάνω από την διαγώνιο αντιπροσωπεύουν τα αποτελέσματα της καλής ταξινόμησης. Τα σημεία κάτω από τη διαγώνιο δίνουν πενιχρά αποτελέσματα. Παρατηρούμε, ωστόσο ότι ένα φτωχό μέσο πρόβλεψης μπορεί απλά να αποκτήσει σημεία πάνω από τη διαγώνιο, αν τα αποτελέσματα του πίνακα συνάφειας αντιστραφούν.




Τα αποτελέσματα της μεθόδου A έχει την καλύτερη προβλεπτική ικανότητα μεταξύ των A, B, C. Το αποτέλεσμα της B ακουμπάει στη διαγώνιο, άρα η ακρίβεια της B είναι 50%. Ωστόσο, όταν πάρουμε το συμμετρικό του σημείου C ως προς το κεντρικό σημείο (0.5, 0.5), η προκύπτουσα μέθοδος C' είναι ακόμη καλύτερη από την A. Αυτή η μέθοδος αντικατοπτρισμού αντιστρέφει τις προβλέψεις οποιασδήποτε μεθόδου ή test παράγει ο πίνακας συνάφειας C. Παρόλο που η αρχική μέθοδος έχει αρνητική προβλεπτική ικανότητα, απλά

αντιστρέφοντας τις αποφάσεις της οδηγούμαστε σε μια νέα διαγνωστικέ μέθοδο C' με θετική προβλεπτική ικανότητα. Στην περίπτωση αυτή η απόσταση σημείου από την διαγώνιο είναι ο καλύτερος δείκτης για την προβλεπτική δύναμη της μεθόδου. Αν το αποτέλεσμα είναι κάτω από τη συγκεκριμένη γραμμή, όλες οι προβλέψεις της μεθόδου πρέπει να αντιστραφούν, ώστε το αποτέλεσμα να βρεθεί πάνω από τη γραμμή και να αξιοποιηθεί η ισχύς της μεθόδου.

### 4.6.6 Έννοια και χρήση του Εμβαδού κάτω από την καμπύλη ROC

Η περιοχή κάτω από την καμπύλη ROC (AUC) ισούται με την πιθανότητα ένας ταξινομητής να κατατάξει ένα τυχαία επιλεγμένο θετικό παράδειγμα υψηλότερα από ένα τυχαία επιλεγμένο αρνητικό. Η AUC ερμηνεύεται αλλιώς ως η πιθανότητα, η τιμή του test για έναν ασθενή (N+) να είναι η μεγαλύτερη από την τιμή του test για ένα άτομο που δεν έχει την ασθένεια (N-). Δηλαδή,
$AUC = P(N+ > N-)$ και εκτιμάται από τον τύπο:

$$w = \frac{1}{n_+ n_-} \sum_{i=1}^{n_+} \sum_{j=1}^{n_-} I\left(N_i^+, N_j^-\right)$$

όπου
$n_+$ και $n_-$ το πλήθος των ατόμων με ή χωρίς τη νόσο αντίστοιχα
$N_i^+$ η τιμή του διαγνωστικού test για το i άτομο της ομάδας ασθενών
$N_j^-$ η τιμή του διαγνωστικού test για το j άτομο της ομάδας μη ασθενών

$$I\left(N_i^+, N_j^-\right) = \begin{cases} 1, & N_i^+ > N_j^- \\ \frac{1}{2}, & N_i^+ = N_j^- \\ 0, & N_i^+ < N_j^- \end{cases}$$

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

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

# Κεφάλαιο 5

# Εφαρμογή σε Πραγματικά Δεδομένα

## 5.1 Καρκίνος Μαστού

Ο καρκίνος του μαστού αποτελεί την πιο συχνή νεοπλασία στις γυναίκες στον ανεπτυγμένο κόσμο και είναι η κυριότερη αιτία θανάτου για τις γυναίκες μεταξύ 40 και 49 ετών. Είναι δε, η δεύτερη σημαντικότερη αιτία θανάτου μετά τον καρκίνο του πνεύμονα για το γυναικείο φύλο. Μια στις τρεις διαγνώσεις δείχνει καρκίνο του μαστού.

Σύμφωνα με τον παγκόσμιο οργανισμό υγείας, ο καρκίνος του στήθους διαγιγνώσκεται ετησίως σε πάνω από 1,1 εκατομμύρια ανθρώπους, σε παγκόσμια κλίμακα. Στην Ευρώπη, ο καρκίνος του στήθους είναι πολύ διαδεδομένος. Το 2004 διαγνώστηκαν περίπου 350.000 νέες περιπτώσεις καρκίνου του στήθους, ενώ οι θάνατοι που είχαν αιτιώδη σχέση με την εν λόγω ασθένεια ανήλθαν σε 130.000, ποσοστό που αντιστοιχεί στο 17.5% του συνόλου των θανάτων από καρκίνο.

Η έγκαιρη ανίχνευση του καρκίνου του μαστού, όπως άλλωστε και οποιασδήποτε μορφής καρκίνου, παίζει σημαντικό ρόλο στην επιβίωση. Ο αποτελεσματικότερος τρόπος παρακολούθησης του καρκίνου του μαστού είναι η μαστογραφία, η οποία μπορεί να ανιχνεύσει μια κακοήθεια ήδη 2 χρόνια πριν από την ψηλάφηση του όγκου. Γενικά, η μαστογραφία ανιχνεύει σε διαφόρους βαθμούς (σε διαφορετικά επίπεδα) τις παρακάτω μορφές καρκίνου του στήθους: ομαδοποιημένες αποτιτανώσεις, ακτινοειδείς ανωμαλίες, περιγραμμένους όγκους και αρχιτεκτονικές δυσμορφίες.

Για την ανάλυση των μαστογραφιών έχουν εξετασθεί πολλές μέθοδοι με διαφορετικό επίπεδο επιτυχίας η καθεμιά. Χαρακτηριστικά αναφέρουμε ότι

έχουν χρησιμοποιηθεί κλασικές τεχνικές, προερχόμενες από το πεδίο της επεξεργασίας εικόνας, με σκοπό τον ταχύτερο και ακριβέστερο εντοπισμό του Καρκίνου. Η μέχρι τώρα έρευνα έχει δείξει ότι οι εφαρμοσθείσες υπολογιστικές τεχνικές έχουν καλά ποσοστά επιτυχίας στην ανάλυση ψηφιακών μαστογραφιών. Το γεγονός αυτό έχει δημιουργήσει την πεποίθηση ότι τα υπολογιστικά συστήματα τα οποία επικεντρώνουν την προσοχή τους σε περιοχές ενδιαφέροντος, οι οποίες δεν χαρακτηρίζονται εύκολα από τους ακτινολόγους, έχουν υψηλές πιθανότητες να βοηθήσουν στην έγκαιρη ανίχνευση και διάγνωση.

## 5.2 Περιγραφή Δεδομένων

Από το UCI Machine Learning Repository, χρησιμοποιήσαμε το σετ δεδομένων Breast Cancer Wisconsin (Diagnostic) Data Set. Τα χαρακτηριστικά υπολογίζονται από μια ψηφιακή εικόνα αναρρόφησης της μάζας του μαστού με μια λεπτή βελόνα (FNA). Περιγράφουν χαρακτηριστικά των κυτταρικών πυρήνων που περιέχονται στην εικόνα. Μερικές τέτοιες εικόνες μπορούν να βρεθούν στο http://www.cs.wisc.edu/~street/images/.

Το σετ δεδομένων περιλαμβάνει 569 παρατηρήσεις και 32 χαρακτηριστικά. Τα 32 αυτά χαρακτηριστικά είναι:

1. Αριθμός Ταυτότητας
2. Διάγνωση με M=malignant (κακοήθη) και B=benign (καλοήθη)
3. Δέκα χαρακτηριστικά πραγματικών τιμών υπολογίζονται για κάθε πυρήνα κυττάρου (3-32). Αυτά τα χαρακτηριστικά είναι:
    i. ακτίνα (μέσος όρος των αποστάσεων από το κέντρο προς τα σημεία στην περίμετρο)
    ii. υφή (τυπική απόκλιση των τιμών της κλίμακας του γκρι)
    iii. περίμετρος
    iv. έκταση
    v. ομαλότητα (τοπική παραλλαγή στο μήκος ακτίνας)
    vi. συμπάγεια $\frac{περίμετρος^2}{έκταση - 1.0}$
    vii. κοιλότητα (βαρύτητα των κοίλων τμημάτων του περιγράμματος)
    viii. κοίλα σημεία (αριθμός κοίλων τμημάτων του περιγράμματος)
    ix. συμμετρία
    x. fractal διάσταση

Η διανομή κλάσεων είναι 357 καλοήθη και 212 κακοήθη.

## 5.3 Χειρισμός Δεδομένων στην R

Σε αυτή την τελευταία ενότητα, συγκρίνουμε μερικούς ταξινομητές από το πεδίο της μηχανικής μάθησης και στη συνέχεια παρουσιάζουμε τα αποτελέσματα της διαδικασίας μοντελοποίησης.

Όπως αναφέραμε θα χρησιμοποιήσουμε δεδομένα που αφορούν τον καρκίνο του μαστού. Κατεβάζουμε από το UCI Machine Learning Repository τα δεδομένα Breast Cancer Wisconsin (Diagnostic) Data Set. Για να μπορέσουμε να κάνουμε σωστά την ταξινόμηση των δεδομένων αντικαθιστούμε τη μεταβλητή διάγνωσης τύπου χαρακτήρα (M και B) με 1 και -1 αντίστοιχα η οποία είναι η μεταβλητή απόκρισης y.

Γενικά η μορφή των παρατηρήσεων είναι της μορφής $x_1, x_2, \dots, x_{30}$ όπου κάθε $x_i$ ανήκει στο $R^{569}$. Η μεταβλητή απόκρισης $y$ παίρνει τις τιμές 1 και -1 και έχει μήκος επίσης 569. Κάθε παρατήρηση συνοδεύεται από τον αριθμό ταυτότητας του ασθενή.

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. Ο διαχωρισμός γίνεται 75%-25% αντίστοιχα και περιλαμβάνει 427 από τις περιπτώσεις στο σύνολο εκπαίδευσης και 142 από τις περιπτώσεις στο σύνολο δοκιμής.

### 5.3.1 Επιλογή χαρακτηριστικών με τον αλγόριθμο RFE και ταξινόμηση χρησιμοποιώντας SVM

Αφού κατεβάσουμε και αποθηκεύσουμε τα δεδομένα μας σε ένα αρχείο txt, τα εισάγουμε στην R σε ένα πίνακα. Στη συνέχεια χωρίζουμε στα διάφορα διανύσματα τα δεδομένα μας $x_1, x_2, \dots, x_{30}$. Για να μπορέσουμε να εκτελέσουμε την ταξινόμηση του μοντέλου με το SVM δημιουργούμε ένα πλαίσιο δεδομένων (mydataframe), μετατρέπουμε την μεταβλητή απόκρισης σε διάνυσμα παραγόντων και κατεβάζουμε το πακέτο "e1071". Στη συνέχεια κατεβάζουμε το πακέτο "caret" και με την εντολή "*confusionMatrix"* παίρνουμε τις τιμές Accuracy, Sensitivity, Specificity οι οποίες πρέπει να είναι ψηλές.

Στη συνέχεια, αφού δημιουργήσουμε το μοντέλο με τη βοήθεια του SVM συγκρίνουμε τις διαφορετικές τιμές των accuracy, sensitivity και specificity για τα δύο σετ αλλά και για τους διάφορους πυρήνες των SVM (Linear, Radial, Polynomial, Sigmoid).

Παρακάτω παρουσιάζουμε συγκεντρωτικά τον πίνακα αποτελεσμάτων:

| | ACCURACY | | SENSITIVITY | | SPECIFICITY | |
|---|---|---|---|---|---|---|
| | TRAIN | TEST | TRAIN | TEST | TRAIN | TEST |
| Linear | 0,9883 | 0.9718 | 0.9748 | 0.9623 | 0.9963 | 0.9775 |
| Radial | 0.9859 | 0,9718 | 0,9686 | 0,9623 | 0.9963 | 0,9775 |
| Polynomial | 0,9040 | 0,9366 | 0,7421 | 0,8302 | 1,0000 | 1,0000 |
| Sigmoid | 0,9485 | 0,9718 | 0,9057 | 0,9434 | 0,9739 | 0,9888 |

Ο γραμμικός πυρήνας έχει μεγαλύτερη ακρίβεια και ευαισθησία στο σύνολο εκπαίδευσης και από τις πιο ψηλές τιμές ακρίβειας στο σύνολο δοκιμής. Η ειδικότητα παίρνει την πιο ψηλή τιμή (100%) με τον πολυωνυμικό πυρήνα. Το μοντέλο μας έχει τον καλύτερο συνδυασμό τιμών ακρίβειας, ευαισθησίας και ειδικότητας χρησιμοποιώντας τον γραμμικό πυρήνα. Επίσης, παρατηρούμε υψηλές τιμές και στο σύνολο δοκιμής, άρα το μοντέλο μας μπορεί να κάνει μια καλή πρόβλεψη.

Τα αποτελέσματα όπως τα παίρνουμε από την R για την περίπτωση του γραμμικού πυρήνα για το σύνολο εκπαίδευσης, φαίνονται παρακάτω.

Confusion Matrix and Statistics

| | Reference | |
|---|---|---|
| Prediction | -1 | 1 |
| -1 | 267 | 4 |
| 1 | 1 | 155 |

Από το πιο πάνω συμπεραίνουμε ότι το μοντέλο μας προέβλεψε 267 τιμές αρνητικές σωστά (TP=267), 4 τιμές προέβλεψε αρνητικές αλλά λάθος (FP=4), μία τιμή προέβλεψε θετική αλλά λάθος (FN=1) και 155 τιμές προέβλεψε θετικές σωστά (TN=155). Έτσι, μπορούμε να βρούμε τις τιμές accuracy, sensitivity και specificity από τους τύπους που αναφέραμε στο προηγούμενο κεφάλαιο.

$$accuracy = \frac{TN + TP}{total} = \frac{267 + 155}{427} = 0{,}9883$$

$$sensitivity = \frac{TN}{TN + FP} = \frac{155}{159} = 0{,}9748$$

$$specificity = \frac{TP}{TP + FN} = \frac{267}{268} = 0{,}9963$$

Και παρακάτω βλέπουμε αναλυτικά τα αποτελέσματα που μας δίνει η R.

Accuracy : 0.9883
95% CI : (0.9729, 0.9962)
No Information Rate : 0.6276
P-Value [Acc > NIR] : <2e-16

Kappa : 0.9749
Mcnemar's Test P-Value : 0.3711

Sensitivity : 0.9748
Specificity : 0.9963
Pos Pred Value : 0.9936
Neg Pred Value : 0.9852
Prevalence : 0.3724
Detection Rate : 0.3630
Detection Prevalence : 0.3653
Balanced Accuracy : 0.9856

'Positive' Class : 1

Με τον ίδιο τρόπο εκτελούμε για το σύνολο εκπαίδευσης και το σύνολο δοκιμής για όλους του πυρήνες, παίρνουμε τα αποτελέσματα στην R όπως προηγουμένως και συμπληρώνουμε τον συγκεντρωτικό πίνακα αποτελεσμάτων που παραθέσαμε πιο πάνω.

Εφαρμογή SVM-RFE:

Στη συνέχεια εφαρμόζουμε τον αλγόριθμο RFE για να γίνει κατάταξη των μεταβλητών μας (κατά φθίνουσα σειρά) και εκτελούμε την ίδια διαδικασία με πριν αλλά αυτή τη φορά θα πάρουμε διαφορετικό αριθμό μεταβλητών κάθε φορά. Αρχικά τις 20 καλύτερες από τα αποτελέσματα του RFE, μετά τις 15 καλύτερες και στη συνέχεια τις 10 καλύτερες. Συγκρίνουμε και πάλι τις τιμές accuracy, sensitivity και specificity κάθε φορά.

Παρακάτω παραθέτουμε συγκεντρωτικά τα αποτελέσματά μας:

Η ταξινομημένη λίστα μεταβλητών από τον αλγόριθμο RFE είναι:

| |
|---|
| 28 27 29 26 25 12 21 1 7 8 9 6 5 13 22 3 30 17 16 18 2 14 15 23 11 20 10 19 24 4 |

Από τον πιο πάνω πίνακα παρατηρούμε τη σημαντικότητα των 30 μεταβλητών κατά φθίνουσα σειρά. Δηλαδή έχουμε $x_{28}, x_{27}, x_{29}, x_{26}, x_{25}, x_{12}$ κλπ. Εκτελούμε

την ταξινόμηση επιλέγοντας τις καλύτερες μεταβλητές. Εισάγουμε σε ένα πλαίσιο δεδομένων τις 20 πρώτες, μετά τις 15 πρώτες και τέλος τις 10 πρώτες.

Για τις 20 καλύτερες μεταβλητές:

| | ACCURACY | | SENSITIVITY | | SPECIFICITY | |
|---|---|---|---|---|---|---|
| *20 μεταβλητές* | TRAIN | TEST | TRAIN | TEST | TRAIN | TEST |
| Linear | 0,9859 | 0,9859 | 0,9752 | 0,9804 | 0,9925 | 0,9890 |
| Radial | 0,9836 | 0,9789 | 0,9627 | 0,9608 | 0,9962 | 0,9890 |
| Polynomial | 0,9180 | 0,9155 | 0,7826 | 0,7647 | 1,0000 | 1,0000 |
| Sigmoid | 0,8993 | 0,9155 | 0,8509 | 0,8824 | 0,9286 | 0,9341 |

Παρατηρούμε και πάλι στο σύνολο εκπαίδευσης αλλά και στο σύνολο δοκιμής παίρνουμε υψηλότερες τιμές ακρίβειας και ευαισθησίας χρησιμοποιώντας τον γραμμικό πυρήνα, ενώ η ειδικότητα παίρνει τη μεγαλύτερη τιμή χρησιμοποιώντας τον πολυωνυμικό πυρήνα. Οι τιμές του συνόλου εκπαίδευσης είναι εξίσου υψηλές άρα το μοντέλο μας δίνει μια καλή πρόβλεψη. Τέλος οι τιμές, αφού κρατήσαμε τις 20 καλύτερες μεταβλητές αφαιρώντας 10, δεν παρουσιάζουν σημαντική αλλαγή εως καθόλου.

Για τις 15 καλύτερες μεταβλητές:

| | ACCURACY | | SENSITIVITY | | SPECIFICITY | |
|---|---|---|---|---|---|---|
| *15 μεταβλητές* | TRAIN | TEST | TRAIN | TEST | TRAIN | TEST |
| Linear | 0,9859 | 0,9718 | 0,9701 | 0,9556 | 0,9962 | 0,9794 |
| Radial | 0,9813 | 0,9648 | 0,9641 | 0,9556 | 0,9923 | 0,9691 |
| Polynomial | 0,9087 | 0,9155 | 0,7665 | 0,7778 | 1,0000 | 0,9794 |
| Sigmoid | 0,9157 | 0,9014 | 0,8862 | 0,8889 | 0,9346 | 0,9072 |

Στο σύνολο εκπαίδευσης αλλά και στο σύνολο δοκιμής παίρνουμε υψηλότερες τιμές ακρίβειας και ευαισθησίας χρησιμοποιώντας τον γραμμικό πυρήνα, ενώ η ειδικότητα παίρνει τη μεγαλύτερη τιμή χρησιμοποιώντας τον πολυωνυμικό πυρήνα. Οι τιμές του συνόλου εκπαίδευσης είναι εξίσου υψηλές άρα το μοντέλο μας δίνει μια καλή πρόβλεψη. Παρατηρούμε, αφού αφαιρέσαμε τις επιπλέον πέντε μεταβλητές ότι υπάρχει μια μικρή πτώση των τιμών ακρίβειας, ευαισθησίας και ειδικότητας.

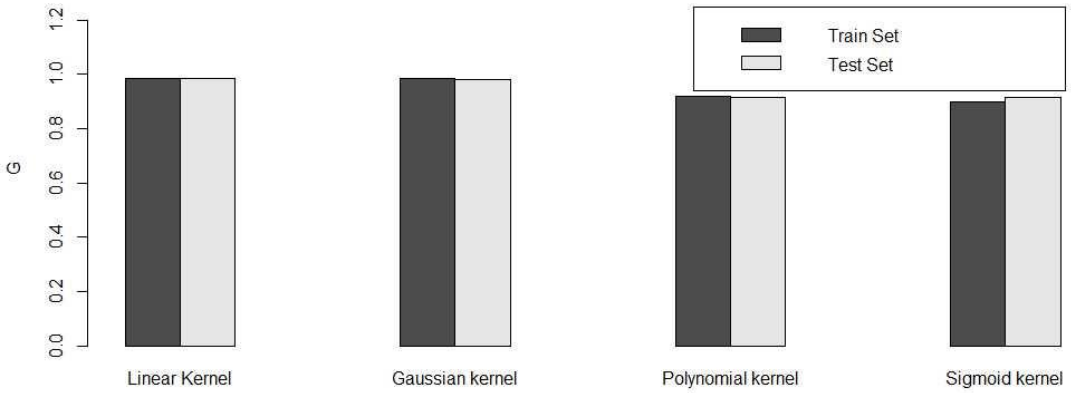
Για τις 10 καλύτερες μεταβλητές:

| | ACCURACY | | SENSITIVITY | | SPECIFICITY | |
|---|---|---|---|---|---|---|
| *10 μεταβλητές* | TRAIN | TEST | TRAIN | TEST | TRAIN | TEST |
| Linear | 0,9696 | 0,9648 | 0,9375 | 0,9038 | 0,9888 | 1,0000 |
| Radial | 0,9649 | 0,9648 | 0,9187 | 0,9038 | 0,9925 | 1,0000 |
| Polynomial | 0,9087 | 0,9014 | 0,7625 | 0,7308 | 0,9963 | 1,0000 |
| Sigmoid | 0,8970 | 0,9296 | 0,8750 | 0,8846 | 0,9101 | 0,9556 |

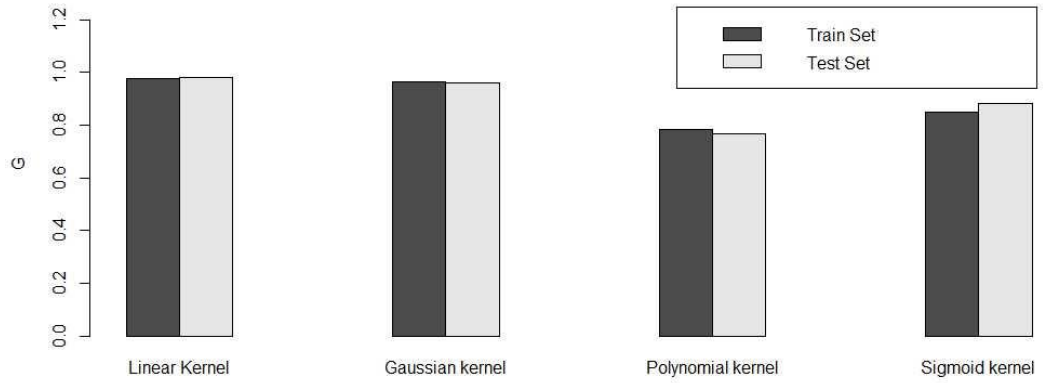
Οι τιμές είναι και πάλι υψηλότερες χρησιμοποιώντας τον γραμμικό πυρήνα, σε αυτή την περίπτωση και η τιμή τις ειδικότητας. Το μοντέλο μας δίνει μια αρκετά καλή πρόβλεψη. Παρατηρούμε όμως, αφού αφαιρέσαμε επιπλέον πέντε μεταβλητές, κρατώντας μόνο τις 10 από τις 30 οι τιμές πέφτουν αισθητά σε σχέση με τα προηγούμενα μοντέλα.

Γενικά, συγκρίνοντας τις τέσσερις αυτές διαφορετικές περιπτώσεις, (30 μεταβλητές, 20 μεταβλητές, 15 μεταβλητές και 10 μεταβλητές) μπορούμε να πούμε ότι το καλύτερο μοντέλο που δίνει τις καλύτερες τιμές ακρίβειας, ευαισθησίας και ειδικότητας είναι το μοντέλο με τις 20 μεταβλητές, γιατί μας δίνει αρκετά ψηλές τιμές με πολύ καλή πρόβλεψη όπως και το αρχικό με όλες τις μεταβλητές. Μπορούμε να δούμε και γραφικά με barplots πως κυμαίνονται οι τιμές της ακρίβειας, ευαισθησίας και ειδικότητας.

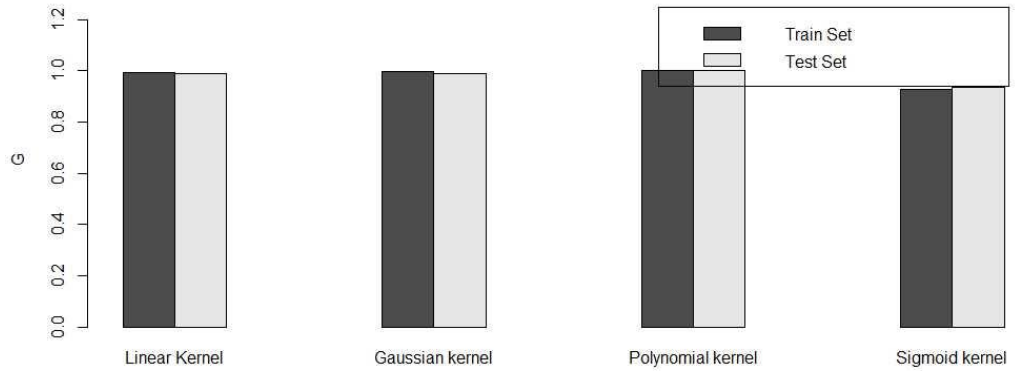
Ακρίβεια (accuracy):



Ευαισθησία (sensitivity):




Ειδικότητα (specificity):




Η υψηλή τιμή ακρίβειας μας αποδεικνύει ότι η δυαδική ταξινόμηση του μοντέλου γίνεται σωστά. Η ευαισθησία, δηλαδή το ποσοστό των αληθώς θετικών αποτελεσμάτων στην ταξινόμηση μας δείχνει ότι το μοντέλο μας είναι σε θέση να προσδιορίσει θετικά αποτελέσματα. Τέλος, η ειδικότητα, δηλαδή το ποσοστό των αληθώς αρνητικών αποτελεσμάτων στην ταξινόμηση, δείχνει την ικανότητα του μοντέλου μας να εντοπίζει αρνητικά αποτελέσματα. Αυτά μπορούμε να τα παρατηρήσουμε από το σύνολο δοκιμής που δίνει τόσο ψηλές τιμές.

### 5.3.2 Επιλογή χαρακτηριστικών με τον αλγόριθμο Lasso και ταξινόμηση χρησιμοποιώντας Λογιστική Παλινδρόμηση

Η Λογιστική Παλινδρόμηση προσαρμόζεται στα δεδομένα σε μια λογιστική (σιγμοειδές) συνάρτηση και κάνει προβλέψεις της πιθανότητας εμφάνισης ενός

γεγονότος. Μία λογιστική συνάρτηση χρησιμοποιείται επειδή μπορεί να πάρει οποιεσδήποτε τιμές (θετική ή αρνητική) και παράγει μια τιμή μεταξύ 0 και 1. Η λογιστική συνάρτηση επηρεάζεται από μια συνάρτηση logit που λαμβάνει μία μεταβλητή που προέρχεται από ένα άθροισμα των σταθμισμένων χαρακτηριστικών. Η λειτουργία logit είναι ο φυσικός λογάριθμος της πιθανότητας η εξαρτημένη μεταβλητή να ισούται με ένα.

Όπως και πριν εισάγουμε με τον ίδιο τρόπο τα δεδομένα μας στην R. Χωρίζουμε και πάλι σε σύνολο εκπαίδευσης και σύνολο δοκιμής με τον ίδιο τρόπο. Αυτή τη φορά προσαρμόζουμε το μοντέλο μας χρησιμοποιώντας λογιστική παλινδρόμηση. Δημιουργούμε το μοντέλο με τη βοήθεια της συνάρτησης glm και συγκρίνουμε τις διαφορετικές τιμές των accuracy, sensitivity και specificity για τα δύο σετ.

| | ACCURACY | | SENSITIVITY | | SPECIFICITY | |
|---|---|---|---|---|---|---|
| | TRAIN | TEST | TRAIN | TEST | TRAIN | TEST |
| | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 |

Τα πιο πάνω αποτελέσματα που φαίνονται στον πίνακα προκύπτουν όπως δείξαμε στο προηγούμενο κεφάλαιο από τα αποτελέσματα που παίρνουμε από τον confusion matrix του μοντέλου.

Παρατηρούμε 100% ακρίβεια, ευαισθησία και ειδικότητα στο σύνολο εκπαίδευσης αλλά και στο σύνολο δοκιμής, άρα το μοντέλο μας μπορεί να κάνει μια πολύ καλή πρόβλεψη.

Εφαρμογή Lasso:

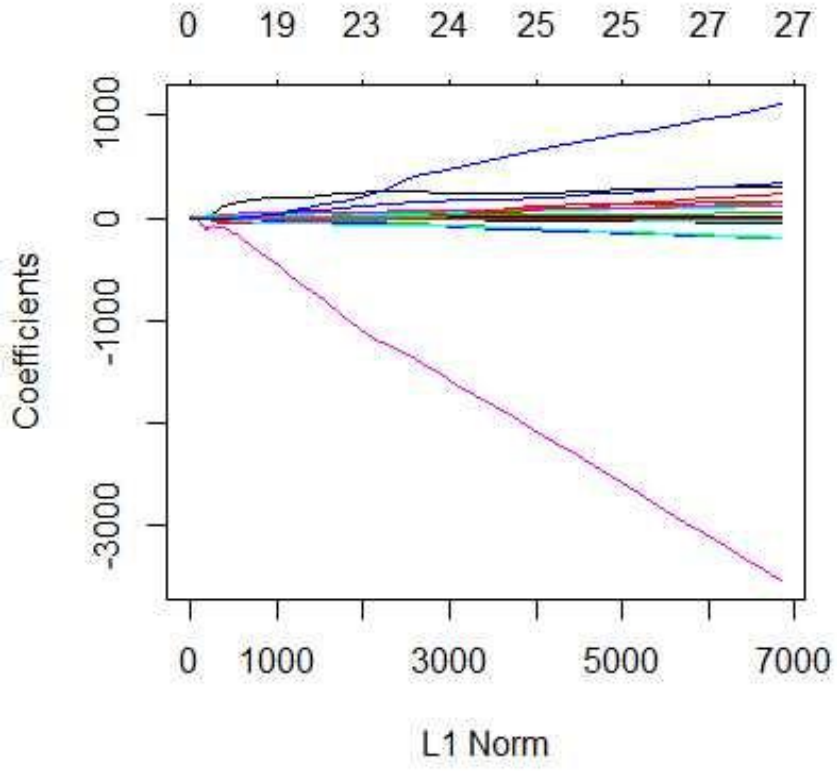
Αφού είδαμε τις τιμές που παίρνει η ακρίβεια, ειδικότητα και ευαισθησία εφαρμόζοντας λογιστική παλινδρόμηση στο μοντέλο χρησιμοποιώντας και τις 30 μεταβλητές, θα εφαρμόσουμε τη μέθοδο Lasso για να πάρουμε τις καλύτερες μεταβλητές και στη συνέχεια θα βρούμε και πάλι τις τιμές ακρίβειας, ειδικότητας και ευαισθησίας στο νέο μειωμένο μοντέλο.

Για να εφαρμόσουμε τη μέθοδο Lasso κατεβάζουμε στην R το πακέτο "glmnet" που περιέχει τη συνάρτηση glmnet. Δημιουργούμε ένα πίνακα x που περιέχει τις 30 μεταβλητές $x_1, x_2, x_3, \dots, x_{30}$ και καλούμε τη συνάρτηση. Ένα μέρος από τα αποτελέσματα που παίρνουμε παρουσιάζονται παρακάτω:

```
Call:  glmnet(x = x, y = Y, family = "binomial", alpha = 1)
      Df       %Dev    Lambda
 [1,]  0 -3.641e-15 3.837e-01
 [2,]  2  8.186e-02 3.496e-01
 ……
```

Στα πιο πάνω αποτελέσματα, Df είναι ο αριθμός των μη μηδενικών μεταβλητών, %dev το ποσοστό της τυπικής απόκλισης και lambda η τιμή του λ. Εξ' ορισμού η συνάρτηση glmnet καλεί 100 τιμές του λ αλλά μπορεί να σταματήσει πιο γρήγορα σε περίπτωση που η τιμή %dev δεν αλλάξει αρκετά από το ένα λ στο επόμενο.

Μπορούμε επίσης να απεικονίσουμε τα δεδομένα χρησιμοποιώντας την εντολή plot(fit) (fit ονομάσαμε το μοντέλο που παίρνουμε με την εφαρμογή της συνάρτησης glmnet).




Κάθε γραμμή αντιστοιχεί σε μια μεταβλητή. Δείχνει την πορεία των συντελεστών σε σχέση με την $l_1 - νόρμα$ όλου του διανύσματος συντελεστών καθώς το λ παίρνει τις διάφορες τιμές. Ο άξονας δηλώνει τον αριθμό των μη μηδενικών συντελεστών για το κάθε λ, που είναι οι αποτελεσματικοί βαθμοί ελευθερίας (df) για τη Lasso.

Χρησιμοποιώντας διασταυρωμένη επικύρωση μπορούμε να βρούμε το καλύτερο λ και τους συντελεστές με το μικρότερο σφάλμα. Η τιμή που παίρνουμε είναι 0.00239857 και από τα προηγούμενα αποτελέσματα μπορούμε να επιλέξουμε το καλύτερο μοντέλο. Οι πιο κοντινές τιμές στο λ αυτό είναι:

[55,] 17  9.109e-01 2.524e-03
 [56,] 17  9.133e-01 2.300e-03
 [57,] 17  9.154e-01 2.096e-03

Έτσι το μοντέλο που θα χρησιμοποιήσουμε θα περιέχει 17 μη μηδενικές μεταβλητές. Με την εντολή coef(fit, s = 0.002) μπορούμε να δούμε ποιες είναι αυτές οι μεταβλητές αλλά και τις τιμές των συντελεστών τους.

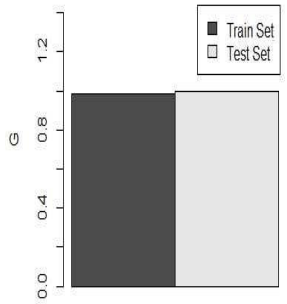
Από τα πιο πάνω προκύπτει ότι οι μεταβλητές που επιλέγουμε με τη μέθοδο Lasso είναι οι $x_2, x_7, x_8, x_{10}, x_{11}, x_{12}, x_{15}, x_{16}, x_{20}, x_{21}, x_{22}, x_{23}, x_{24}, x_{25}, x_{27}, x_{28}, x_{29}$.

Στη συνέχεια, σε ένα νέο πλαίσιο δεδομένων βάζουμε αυτές τις μεταβλητές με τη μεταβλητή απόκρισης και κάνουμε και πάλι με τον ίδιο τρόπο λογιστική παλινδρόμηση για να ελέγξουμε και πάλι τις τιμές της ακρίβειας, ευαισθησίας και ειδικότητας.

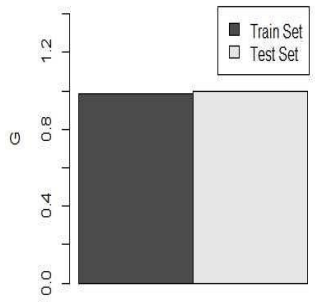
| | ACCURACY | | SENSITIVITY | | SPECIFICITY | |
|---|---|---|---|---|---|---|
| *17 μεταβλητές* | TRAIN | TEST | TRAIN | TEST | TRAIN | TEST |
| | 0,9859 | 1,0000 | 0,9810 | 1,0000 | 0,9888 | 1,0000 |

Παρατηρούμε ότι οι τιμές διατηρούνται ακόμα ψηλές για το σύνολο εκπαίδευσης ενώ είναι 100% για το σύνολο δοκιμής. Έτσι, το μειωμένο μοντέλο κάνει πολύ καλή πρόβλεψη για τα δεδομένα μας. Παρακάτω παρουσιάζουμε και τα barplots για τις τιμές ακρίβειας, ευαισθησίας και ειδικότητας.

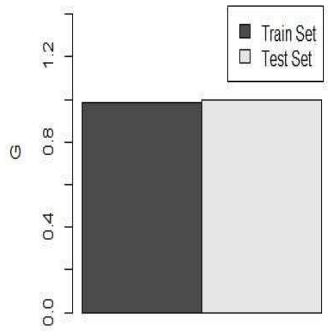
<u>Ακρίβεια (accuracy):</u>



Ευαισθησία (Sensitivity):




Ειδικότητα (Specificity):




Θα εκτελέσουμε την ίδια διαδικασία κρατώντας 20 μεταβλητές, 15 μεταβλητές και 10 μεταβλητές όπως ακριβώς κάναμε και χρησιμοποιώντας τη μέθοδο SVM στο προηγούμενο υποκεφάλαιο. Βρίσκουμε ποιες μεταβλητές θα κρατήσουμε κάθε φορά με την προηγούμενη διαδικασία της μεθόδου Lasso.

| | ACCURACY | | SENSITIVITY | | SPECIFICITY | |
|---|---|---|---|---|---|---|
| *20 μεταβλητές* | TRAIN | TEST | TRAIN | TEST | TRAIN | TEST |
| | 0,9906 | 1,0000 | 0,9809 | 1,0000 | 0,9963 | 1,0000 |

| | ACCURACY | | SENSITIVITY | | SPECIFICITY | |
|---|---|---|---|---|---|---|
| *15 μεταβλητές* | TRAIN | TEST | TRAIN | TEST | TRAIN | TEST |
| | 0,9906 | 1,0000 | 0,9805 | 1,0000 | 0,9963 | 1,0000 |

| | ACCURACY | | SENSITIVITY | | SPECIFICITY | |
|---|---|---|---|---|---|---|
| *10 μεταβλητές* | TRAIN | TEST | TRAIN | TEST | TRAIN | TEST |
| | 0,9906 | 1,0000 | 0,9871 | 1,0000 | 0,9926 | 1,0000 |

Γενικά, χρησιμοποιώντας λογιστική παλινδρόμηση στο σύνολο εκπαίδευσης έχουμε πολύ ψηλές τιμές ακρίβειας, ευαισθησίας και ειδικότητας ενώ στο σύνολο δοκιμής όσες μεταβλητές και να κρατήσουμε έχουμε πάντα τιμές 100%. Μπορεί να γίνει πολύ καλή πρόβλεψη στο μοντέλο. Η τιμή ακρίβειας μας αποδεικνύει ότι η δυαδική ταξινόμηση του μοντέλου γίνεται σωστά. Η ευαισθησία, δηλαδή το ποσοστό των αληθώς θετικών αποτελεσμάτων στην ταξινόμηση μας δείχνει ότι το μοντέλο μας είναι σε θέση να προσδιορίσει θετικά αποτελέσματα. Τέλος, η ειδικότητα, δηλαδή το ποσοστό των αληθώς αρνητικών αποτελεσμάτων στην ταξινόμηση, δείχνει την ικανότητα του μοντέλου μας να εντοπίζει αρνητικά αποτελέσματα.

### 5.3.3 Σύγκριση Μεθόδων – Συμπεράσματα

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 στα δεδομένα καρκίνου του μαστού για πρόβλεψη καλοήθη ή κακοήθη όγκου.

Όταν είναι ενδιαφέρον να προβλέψουμε την ομάδα στην οποία μια νέα παρατήρηση ανήκει, τα μοντέλα SVM είναι μια εφικτή εναλλακτική λύση για την Λογιστική Παλινδρόμηση.

Τα SVMs έχουν χρησιμοποιηθεί ευρέως στη μηχανική μάθηση. Οι μηχανές διανυσματικής υποστήριξης (SVMs) είναι γρήγορες στην εκπαίδευση, αλλά απαιτούν μια κατάλληλη επιλογή της συνάρτησης του πυρήνα. Από την ανάλυση που κάναμε, βρήκαμε ότι ένα ιδανικό μοντέλο θα ήταν κρατώντας τις 20 από τις 30 μεταβλητές. Η επιλογή χαρακτηριστικών με SVMs είναι μία από τις πολλές εφαρμογές SVM. Έχουμε παρουσιάσει μεθόδους επιλογής χαρακτηριστικών με SVMs. Μια γενική εισαγωγή για τα SVMs δόθηκε πριν την εισαγωγή στις μεθόδους επιλογής χαρακτηριστικών από το SVM μιας και ήταν το κύριο εργαλείο μηχανικής μάθησης. Παρουσιάσαμε τον αλγόριθμο SVM – RFE και στη συνέχεια παρουσιάζονται τα πειραματικά αποτελέσματα. Τα πειραματικά αποτελέσματα έδειξαν ότι ο αλγόριθμος είναι εξίσου καλός όσο άλλες

υπάρχουσες μέθοδοι στην ακρίβεια και επίσης είναι πολύ ταχύς. Επίσης, τα πειραματικά αποτελέσματα υποστηρίζουν ότι ένα υποσύνολο των χαρακτηριστικών είναι αρκετά σημαντικό για την κατάρτιση του μηχανήματος και την πρόβλεψη του στόχου. Στο δικό μας πείραμα καρκίνου του μαστού, μόνο 20 γονίδια δίνουν την καλύτερη ακρίβεια.

Οι δύο μέθοδοι ταξινόμησης (SVM και Λογιστική Παλινδρόμηση) δίνουν εξίσου καλά αποτελέσματα και πολύ ψηλές τιμές ακρίβειας, ευαισθησίας και ειδικότητας. Η Λογιστική Παλινδρόμηση όσο αφορά το σύνολο δοκιμής μας δίνει 100% ακρίβεια, ευαισθησία και ειδικότητα χωρίς όμως η μέθοδος SVM να υστερεί σε μεγάλο βαθμό με τιμές να κυμαίνονται από 98,5%-99,5%.

Οι μέθοδοι επιλογής χαρακτηριστικών (SVM-RFE και Lasso) κάνουν επιλογή μεταβλητών με τη διαφορά η μέθοδος Lasso να μας δίνει το καλύτερο μοντέλο, πόσες και ποιες μεταβλητές πρέπει να κρατήσουμε σε αντίθεση με τον αλγόριθμο SVM-RFE που πρέπει να κάνουμε κατ' επανάληψη διαδικασίες για να βρούμε τον καλύτερο αριθμό μεταβλητών. Οι μεταβλητές που επιλέγονται τελικά και από τους δύο αλγόριθμους είναι οι ίδιες, άρα μπορούμε να πούμε ότι είναι εξίσου αξιόπιστες μέθοδοι.

Για τη σύγκριση των αποτελεσμάτων, έπειτα από την επιλογή των σημαντικών μεταβλητών με τις δύο μεθόδους, κάνουμε γραφική παρουσίαση των τιμών ακρίβειας, ευαισθησίας και ειδικότητας για κάθε μέθοδο, SVM-RFE και Lasso. Με αυτά τα γραφήματα αυτά παρατηρούμε τις ψηλές τιμές του συνόλου εκπαίδευσης και συνόλου δοκιμής και με τις δύο μεθόδους. Η μέθοδος SVM, όπως αναφέραμε είναι μία πολύ υποσχόμενη μέθοδος για την ταξινόμηση και μπορεί να συγκριθεί επαρκώς με την πολυχρησιμοποιημένη Λογιστική Παλινδρόμηση και μέθοδο Lasso.

## 5.4 Μελλοντική Δουλειά

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 ιατρικά δεδομένα. Ωστόσο, υπάρχουν μερικά ενδιαφέροντα θέματα που είναι ανοιχτά και θα πρέπει να διερευνηθούν στο μέλλον.

Αυτό που επανεξετάσαμε είναι οι μέθοδοι επιλογής χαρακτηριστικών, που δίνουν μια κατάταξη για όλες τις μεταβλητές. Με αυτές τις μεθόδους, δεν ξέρουμε ποια από τα πολλά χαρακτηριστικά γνωρίσματα είναι καλύτερα χωρίς διεξαγωγή του υπολογισμού για την ακρίβεια για τα κορυφαία χαρακτηριστικά k, όπου $k = 1, \ldots, n$. Δηλαδή, δεν ξέρουμε τον βέλτιστο αριθμός των χαρακτηριστικών. Έρευνα για το σκοπό αυτό παραμένει ανεξερεύνητη.

Επίσης, η έρευνα σε SVMs είναι υπό εξερεύνηση, ιδιαίτερα για να βρεθούν αναλυτικά οι παράμετροι για το πείραμα. Στο πείραμα μας κάναμε επανάληψη για τις διάφορους πυρήνες κατ' επανάληψη για ένα δυνατό σύνολο παραμέτρων. Αυτό απαιτεί πολλά προκαταρκτικά πειράματα πριν γίνει η επιλογή χαρακτηριστικών.

Γενικά, τα δεδομένα εκπαίδευσης πρέπει προεπεξεργαστούν πριν από την εκπαίδευση ενός SVM. Υπάρχει σημαντική διαφορά στη διαδικασία εκπαίδευσης στην ταχύτητα κατάρτισης μεταξύ προεπεξεργασμένων δεδομένων και μη προεπεξεργασμένων δεδομένων. Μερικές φορές η προεπεξεργασία επηρεάζει τις τιμές της ακρίβειας. Η προεπεξεργασία περνά μέσα από πολλά στάδια: τη λήψη του λογάριθμου και στη συνέχεια ομαλοποίηση. Η ομαλοποίηση περιλαμβάνει αφαίρεση της μέσης τιμής και διαιρώντας το αποτέλεσμα με την τυπική απόκλιση. Αλλά και για άλλα σύνολα δεδομένων, μπορεί να υπάρχουν διαφορετικά βήματα προεπεξεργασίας. Το πώς η προεπεξεργασία πρέπει να γίνει και πώς επηρεάζει την ταχύτητα κατάρτισης και την ακρίβεια είναι επίσης ένα θέμα που πρέπει να διερευνηθεί.

Τέλος, περαιτέρω εργασία περιλαμβάνει η αξιολόγηση των μεθόδων SVM και Λογιστικής Παλινδρόμησης για άλλες διανομές πιθανότητας, διαφορετικές μήτρες διακύμανσης-συνδιακύμανσης μεταξύ των ομάδων, και υψηλής διάστασης δεδομένα, με λιγότερα δεδομένα από τις παρατηρήσεις, π.χ., τα γενετικά δεδομένα με έως 5 εκατομμύρια γονοτύπων και 1000 περιπτώσεις.

# Βιβλίογραφία

## (References )


1. Asparoukhova, K. & Krzanowskib, J. (2001), 'A comparison of discriminant procedures for binary variables', Computational Statistics & Data Analysis 38, 139–160.

2. C. Koukouvinos, C. Parpoula, K. Drosou, K. Mylona. (2014). A new variable Selection Approach Inspired by Supersaturated Designs Given a Large-Dimensional Dataset. *Journal of Data Science*. 12, 35-52.

3. C. Koukouvinos, K. Mylona, C. Parpoula. (2013). A combination of a model of variable selection and data mining techniques for high-dimensional statistical modeling. *International Journal of Information and Decision Sciences*. 5 (2), 154-168.

4. Chapelle, O., Vapnik, V.N., Bousquet, O., Mukherjee, S., (2002). Choosing Multiple Parametersfor Support Vector Machines. Machine Learning, 46, 131-159.

5. D. L. Donoho. (2000). High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality. *AMS Math Challenges Lecture*.

6. D. Morariu, L. N. Vintan, V. Tresp. (2006). Evaluating some Feature Selection Methods for an Improved SVM Classifier. *International Journal of Electrical and Computer Engineering*. 8 (1), 575-585.

7. D. Zeg (2010). *STATISTICAL LEARNING AND HIGH-DIMENSIONAL DATA*. North Carolina at Chapel Hill: Department of Biostatistics, University of North Carolina at Chapel Hill. 7-118.

8. D.T., Larose, (2005). Discovering Knowledge in Data. An Introduction to Data Mining, John Wiley and Sons, Hoboken, New Jersey.

9. D.W., Hosmer, S., Lemeshow, (2000). Applied Logistic Regression, John Wiley and Sons, New York, second edition.

10. De Martino F, Valente G, Staeren N, Ashburner J, Goebel R., Formisano E. (2008). Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns. *NeuroImage*,43, 44-48.

11. Diego Alejandro Salazar, Jorge Iván Vélez2, Juan Carlos Salazar. (2012). Comparison between SVM and Logistic Regression: Which One is Better to Discriminate?. *Revista Colombiana de Estadística Número especial en Bioestadística*. 35 (2), 223-237.

12. Drucker, Harris; Burges, Christopher J. C.; Kaufman, Linda; Smola, Alexander J.; and Vapnik, Vladimir N. (1997); "Support Vector Regression Machines", in *Advances in Neural Information Processing Systems 9, NIPS 1996*, 155–161, MIT Press.

13. Eun Seog Youn (2002). Feature Selection in Support Vector Machines, Master Thesis to the graduate school of the University of Florida in Partial Fulfillment of the requirements for the degree of Master of Science.

14. F. M. Zaman, H. Hirose. (2009). Double SVMBagging: A New Double Bagging with Support Vector Machine. *Engineering Letter*. 17 (2).

15. Fan Li, Yiming Yang, Eric P. Xing. (2006). From Lasso regression to Feature vector machine. *School of Computer Science, Carnegie Mellon University*.

16. Guyon I, Weston J, Barnhill SMD, Vapnik V(2002): Gene Selection for Cancer Classification using Support Vector Machines.*Machine Learning*, 46(1–3):389-422.

17. Hastie, T., Tibshirani, R., Friedman J., (2009). The Elements of Statistical Learning: Data Mining, Inference and Prediction, Springer, Berlin, 2nd edition.

18. Hsu, Chih-Wei; Chang, Chih-Chung; and Lin, Chih-Jen (2003). A Practical Guide to Support Vector Classification (Technical report). Department of Computer Science and Information Engineering, National Taiwan University.

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

20. John Robert Taylor (1999). *An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements*. University Science Books. pp. 128–129.

21. L. Parsons, E. Haque, H. Liu. (2004). Subspace clustering for high dimensional data: A Review. *Newsletter ACM SIGKDD Explorations Newsletter - Special issue on learning from imbalanced datasets*. 6 (1), 90-105.

22. L. Yu, H. Liu. (2003). Feature Selection for High Dimensional Data: A fast correlation-based filter solution. *Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC*.

23. Lee, Y.; Lin, Y.; and Wahba, G. (2001). "Multicategory Support Vector Machines". *Computing Science and Statistics 33*.

24. Lukas Meier, Sara van de Geer, Peter Bühlmann. (2008). The group lasso for logistic regression. *J. R. Statis.*. 70 (1), 53-71.

25. M. L. Samb , F. Camara, S. Ndiaye, Y. Slimani, M. A. Esseghir. (2012). A Novel RFE-SVM-based Feature Selection Approach for Classification .*International Journal of Advanced Science and Technology*. 43, 27-36.

26. M.-S. Chen, J. Han, P. S. Yu. (1996). Data Mining: An Overview from Database Perspective. *IEEE Transactions on Knowledge and Data Enjineering*. 8 (6), 866-883.

27. Minh Hoai Nguyen, Fernando De la Torre. (2009). Optimal Feature Selection for Support Vector Machines. *Pattern Recognition*.

28. N. Shigei, H. Miyajima. (2009). Bagging and Boosting Algorithms for Support Vector Machine Classifiers. *Proceedings of the 8th WSEAS Int. Conf. on ARTIFICIAL INTELLIGENCE, KNOWLEDGE ENGINEERING & DATA BASES*. 1790-5109.

29. P. Bühlmann, P. Rütimann, M. Kalisch. (2013). Controlling false positive selections in high-dimensional regression and causal inference.*Statistical Methods in Medical Research*. 22 (5), 466-492.

30. P. Bühlmann, S. Van De Geer (2011). *Statistics for High Dimensional Data, Methods, Theory and Applications*. Zurich, Switzerland: Springer Series in Statistics.

31. P. Ravikumar (2001). *High Dimensional Statistics*. University of Texas.

32. Picard, Richard; Cook, Dennis (1984). "Cross-Validation of Regression Models". *Journal of the American Statistical Association* 79 (387): 575–583.

33. Powers, David M W (2007–2011). "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation". *Journal of Machine Learning Technologies* **2** (1): 37–63.

34. R. Agrawal, J. Gehrke, D. Gunopulos, P. Raghavan. (1998). Automatic subspace clustering of high dimensional data for data mining applications. *Newsletter, ACM SIGMOD Record*. 27 (2), 94-105.

35. R.-E. Fan; K.-W. Chang; C.-J. Hsieh; X.-R. Wang; C.-J. Lin (2008). "LIBLINEAR: A library for large linear classification*". Journal of Machine Learning Research 9*: 1871–1874.

36. Ricardo Ramos Guerra, Jörg Stork (2013). *Building and analyzing SVM ensembles with Bagging and AdaBoost on big data sets*. Cologne University of Applied Sciences, Germany.

37. Robert Tibshirani. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*. 58 (1), 267-288.

38. Roman Zakharov, Pierre Dupont. (2013). Stable LASSO for High-Dimensional Feature Selection through Proximal Optimization. *ROKS, Machine Learning Group, ICTEAM Institute, Universite catholique de Louvain, B-1348 Louvain-la-Neuve, Belgium*.

39. S.-J. Wanga, A. Mathew, Y. Chen, L.-F. Xi, L. Mab, J. Lee. (2009). Empirical analysis of support vector machine ensemble classifiers. *Expert Systems with Applications*. 36, 6466-6476.

40. Smola, A.J., Schölkopf, B.,(2004). A tutorial on support vector regression, Statistics and Computing, Volume 14, Issue 3, pp 199-222.

41. T. Hastie, R. Tibshirani, J. Friedman (2008). *The Elements of Statistical Learning - Data Mining, Inference, and Prediction*. 2nd ed. Stanford, California: Springer Series in Statistics. 20-764.

42. T. Siva Tian, (2009). Dimensionality Reduction for Classification with high-dimensional data, University of Southern California.

43. W.W., Hauck and A., Donner, (1997). Wald's test applied to hypotheses in logit analysis. Journal of the American Statistical Association, Vol. 72, pp. 851-853.

44. Y. S. Kim, W. N. Street, F. Menczer. *Feature Selection in Data Mining.*Available: http://www.ime.unicamp.br/~wanderson/Artigos/feature_selection_in_mining.pdf. Last accessed 28th Jan 2014.

45. Ying Yu. (2012). SVM-RFE Algorithm for Gene Feature Selection.*Department of Electrical and Computer Engineering, University of Delaware, Newark*.

46. Yuanyuan Ding, Dawn Wilkins. (2006). Improving the Performance of SVM-RFE to Select Genes in Microarray Data. *BMC Bioinformatics*. 7(S12).

47. Yuchun Tang, Yan-Qing Zhang, Zhen Huang, Xiaohua Hu. (2005). Granular SVM-RFE gene selection algorithm for reliable prostate cancer classification on microarray expression data. *Bioinformatics and Bioengineering. BIBE 2005. Fifth IEEE Symposium*, 290-293.

48. Yuchun Tang, Yan-Qing Zhang, Zhen Huang. (2007). Development of Two-Stage SVM-RFE Gene Selection Strategy for Microarray Expression Data Analysis. *Computational Biology and Bioinformatics*. 4 (3), 365-381.

49. Zhu, Y., Li, C., and Zhang, Y., (2004). A Practical Parameters Selection Method for SVM ISNN 2004, LNCS 3173, pp. 518–523.

50. Zong-Xia Xie, Qing-Hua Hu, Da-Ren Yu. (2006). Improved Feature Selection Algorithm Based on SVM and Correlation. *Harbin Institute of Technology, Harbin, P.R. China*. ISNN 2006, LNCS 3971, 1373–1380.

# Παράρτημα

## (Κώδικες στην R)

1. Εισαγωγή Δεδομένων και δημιουργία Πλαισίου Δεδομένων

```
mydata=matrix(scan("C:\\Users\\yiolanda\\Desktop\\diplomatiki\\cancer_data.
txt",sep=",",na.strings="?"),ncol=32,byrow=T)
y=mydata[,2]
x1=mydata[,3]
x2=mydata[,4]
x3=mydata[,5]
x4=mydata[,6]
x5=mydata[,7]
x6=mydata[,8]
x7=mydata[,9]
x8=mydata[,10]
x9=mydata[,11]
x10=mydata[,12]
x11=mydata[,13]
x12=mydata[,14]
x13=mydata[,15]
x14=mydata[,16]
x15=mydata[,17]
x16=mydata[,18]
x17=mydata[,19]
x18=mydata[,20]
x19=mydata[,21]
x20=mydata[,22]
x21=mydata[,23]
x22=mydata[,24]
x23=mydata[,25]
x24=mydata[,26]
```

```
x25=mydata[,27]
x26=mydata[,28]
x27=mydata[,29]
x28=mydata[,30]
x29=mydata[,31]
x30=mydata[,32]
Y=as.factor(y)
mydataframe=data.frame(Y,x1,x2,x3,x4,x5,x6,x7,x8,x9,x10,x11,x12,x13,x14,x15,x16,
x17,x18,x19,x20,x21,x22,x23,x24,x25,x26,x27,x28,x29,x30)
load("e1071")
load("caret")
X=mydataframe[,2:31]
```

2. Διαχωρισμός Δεδομένων σε Train Set και Test Set (75%-25%)

```
index<-1:nrow(mydataframe)
testindex<-sample(index, trunc(length(index)/4))
testset<-mydataframe[testindex,]
trainset<-mydataframe[-testindex,]
xtrainset <-trainset[,2:31]
ytrainset <- trainset[,1]
xtestset <-testset[,2:31]
ytestset <- testset[,1]
ytrainset<-as.factor(ytrainset)
ytestset<-as.factor(ytestset)
```

3. Δημιουργία μοντέλου και σύγκριση accuracy&sensitivity για τα δύο σετ και τους διάφορους πυρήνες SVM (linear, radial, polynomial, sigmoid)

*Linear Kernel*

```
model_linear<-svm(Y~.,data=trainset, type='C-classification', kernel='linear')

predtrain<-predict(model_linear, xtrainset)
resultstrain<-confusionMatrix(predtrain, ytrainset,  positive="1")
```

*predtest<-predict(model_linear, xtestset)*
*resultstest<-confusionMatrix(predtest, ytestset,  positive="1")*

*<u>Radial Kernel</u>*
*model_radial<-svm(Y~.,data=trainset, type='C-classification', kernel='radial')*

*predtrain<-predict(model_radial, xtrainset)*
*resultstrain<-confusionMatrix(predtrain, ytrainset,  positive="1")*

*predtest<-predict(model_radial, xtestset)*
*resultstest<-confusionMatrix(predtest, ytestset,  positive="1")*

*<u>Polynomial Kernel</u>*
*model_linear<-svm(Y~.,data=trainset, type='C-classification', kernel='polynomial')*

*predtrain<-predict(model_linear, xtrainset)*
*resultstrain<-confusionMatrix(predtrain, ytrainset,  positive="1")*

*predtest<-predict(model_linear, xtestset)*
*resultstest<-confusionMatrix(predtest, ytestset,  positive="1")*

*<u>Sigmoid Kernel</u>*
*model_linear<-svm(Y~.,data=trainset, type='C-classification', kernel='sigmoid')*

*predtrain<-predict(model_linear, xtrainset)*
*resultstrain<-confusionMatrix(predtrain, ytrainset,  positive="1")*

*predtest<-predict(model_linear, xtestset)*
*resultstest<-confusionMatrix(predtest, ytestset,  positive="1")*

4. <u>Αλγόριθμος SVM_RFE:</u>

*svmrfeFeatureRanking = function(x,y){*
*n = ncol(x)*
*survivingFeaturesIndexes = seq(1:n)*

```
featureRankedList = vector(length=n)
rankedFeatureIndex = n
while(length(survivingFeaturesIndexes)>0){

#train the support vector machine
svmModel = svm(x[, survivingFeaturesIndexes], y, scale=FALSE, type="C-classification", kernel="linear" )

#compute the weight vector
w = t(svmModel$coefs)%*%svmModel$SV

#compute ranking criteria
rankingCriteria = w * w

#rank the features
ranking = sort(rankingCriteria, index.return = TRUE)$ix

#update feature ranked list
featureRankedList[rankedFeatureIndex] = survivingFeaturesIndexes[ranking[1]]
rankedFeatureIndex = rankedFeatureIndex - 1

#eliminate the feature with smallest ranking criterion
(survivingFeaturesIndexes = survivingFeaturesIndexes[-ranking[1]])
}
return (featureRankedList)
}
```

5. Δημιουργία νέου πλαισίου δεδομένων με βάση τα αποτελέσματα του αλγορίθμου SVM-RFE:

   20 μεταβλητές:

```
mydataframe20=data.frame(Y,x28,x27,x29,x26,x25,x12,x21,x1,x7,x8,x9,x6,x5,x13,x22,x3,x30,x17,x16,x18)
index<-1:nrow(mydataframe20)
testindex<-sample(index, trunc(length(index)/4))
testset<-mydataframe20[testindex,]
```

```
trainset<-mydataframe20[-testindex,]
xtrainset <-trainset[,2:21]
ytrainset <- trainset[,1]
xtestset <-testset[,2:21]
ytestset <- testset[,1]
ytrainset<-as.factor(ytrainset)
ytestset<-as.factor(ytestset)
```

Όμοια για 15 μεταβλητές και για 10 μεταβλητές.

6. Λογιστική Παλινδρόμηση:
   Αφού εισάγουμε τα δεδομένα μας στην R χωρίσουμε σε train set και test set, αρχικά εκτελούμε για το train set:

```
logit=glm(Y~x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11+x12+x13+x14+x15+x16+x17+x18+x19+x20+x21+x22+x23+x24+x25+x26+x27+x28+x29+x30,binomial(link="logit"),trainset, control=glm.control(maxit=100))
```

Με τις πιο κάτω εντολές δημιουργούμε το confusion Matrix για τη λογιστική παλινδρόμηση:

```
confusion.glm <- function(data, model) {
prediction <- ifelse(predict(model, data, type='response') > 0.5, TRUE, FALSE)
 confusion  <- table(prediction, as.logical(model$y))
confusion<-cbind(confusion,c(1 -confusion[1,1]/(confusion[1,1]+confusion[2,1]), 1 -confusion[2,2]/(confusion[2,2]+confusion[1,2])))
confusion  <- as.data.frame(confusion)
names(confusion) <- c('FALSE', 'TRUE', 'class.error')
confusion
}

confusion.glm(xtrainset,logit)
```

Όμοια για το test set:

*logit=glm(Y~x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11+x12+x13+x14+x15+x16+x17+x18+x19+x20+x21+x22+x23+x24+x25+x26+x27+x28+x29+x30,binomial(link="logit"),testset, control=glm.control(maxit=100))*

*confusion.glm(xtestset,logit)*

7. Αλγόριθμος Lasso:

*x <-*
*as.matrix(data.frame(x1,x2,x3,x4,x5,x6,x7,x8,x9,x10,x11,x12,x13,x14,x15,x16,x17,x18,x19,x20,x21,x22,x23,x24,x25,x26,x27,x28,x29,x30))*

*fit<-glmnet(x,Y,alpha=1,family='binomial')*
*plot(fit)*

Τα αποτελέσματα που παίρνουμε από τη χρήση της συνάρτησης *glmnet* παρουσιάζονται παρακάτω:

```
Call:  glmnet(x = x, y = Y, family = "binomial", alpha = 1)
      Df       %Dev    Lambda
 [1,]  0 -3.641e-15 3.837e-01
 [2,]  2  8.186e-02 3.496e-01
 [3,]  2  1.554e-01 3.185e-01
 [4,]  2  2.180e-01 2.902e-01
 [5,]  2  2.724e-01 2.645e-01
 [6,]  3  3.205e-01 2.410e-01
 [7,]  3  3.636e-01 2.196e-01
 [8,]  3  4.027e-01 2.001e-01
 [9,]  3  4.380e-01 1.823e-01
[10,]  2  4.700e-01 1.661e-01
[11,]  2  4.989e-01 1.513e-01
[12,]  2  5.253e-01 1.379e-01
[13,]  2  5.495e-01 1.256e-01
[14,]  3  5.718e-01 1.145e-01
[15,]  3  5.922e-01 1.043e-01
[16,]  4  6.138e-01 9.504e-02
[17,]  4  6.343e-01 8.660e-02
[18,]  4  6.531e-01 7.890e-02
[19,]  4  6.706e-01 7.190e-02
[20,]  4  6.867e-01 6.551e-02
[21,]  4  7.017e-01 5.969e-02
[22,]  4  7.155e-01 5.439e-02
[23,]  4  7.284e-01 4.955e-02
[24,]  4  7.403e-01 4.515e-02
[25,]  5  7.515e-01 4.114e-02
```

```
[26,]  5  7.622e-01 3.749e-02
[27,]  6  7.722e-01 3.416e-02
[28,]  6  7.818e-01 3.112e-02
[29,]  7  7.911e-01 2.836e-02
[30,]  7  8.000e-01 2.584e-02
[31,]  7  8.083e-01 2.354e-02
[32,]  7  8.160e-01 2.145e-02
[33,]  8  8.232e-01 1.955e-02
[34,]  8  8.299e-01 1.781e-02
[35,]  8  8.362e-01 1.623e-02
[36,]  8  8.421e-01 1.479e-02
[37,]  8  8.476e-01 1.347e-02
[38,]  9  8.527e-01 1.228e-02
[39,]  9  8.575e-01 1.118e-02
[40,]  9  8.619e-01 1.019e-02
[41,] 10  8.667e-01 9.286e-03
[42,] 10  8.713e-01 8.461e-03
[43,] 10  8.756e-01 7.709e-03
[44,] 10  8.796e-01 7.024e-03
[45,] 10  8.832e-01 6.400e-03
[46,] 10  8.865e-01 5.832e-03
[47,] 10  8.895e-01 5.314e-03
[48,] 11  8.925e-01 4.842e-03
[49,] 12  8.955e-01 4.411e-03
[50,] 13  8.984e-01 4.020e-03
[51,] 13  9.011e-01 3.662e-03
[52,] 13  9.037e-01 3.337e-03
[53,] 15  9.061e-01 3.041e-03
[54,] 15  9.086e-01 2.771e-03
[55,] 17  9.109e-01 2.524e-03
[56,] 17  9.133e-01 2.300e-03
[57,] 17  9.154e-01 2.096e-03
[58,] 16  9.172e-01 1.910e-03
[59,] 16  9.188e-01 1.740e-03
[60,] 17  9.204e-01 1.585e-03
[61,] 17  9.219e-01 1.445e-03
[62,] 16  9.232e-01 1.316e-03
[63,] 17  9.245e-01 1.199e-03
[64,] 16  9.258e-01 1.093e-03
[65,] 15  9.268e-01 9.957e-04
[66,] 16  9.277e-01 9.072e-04
[67,] 16  9.288e-01 8.266e-04
[68,] 17  9.298e-01 7.532e-04
[69,] 18  9.308e-01 6.863e-04
[70,] 19  9.319e-01 6.253e-04
[71,] 19  9.330e-01 5.698e-04
[72,] 19  9.341e-01 5.191e-04
[73,] 20  9.352e-01 4.730e-04
[74,] 21  9.365e-01 4.310e-04
[75,] 21  9.377e-01 3.927e-04
[76,] 22  9.394e-01 3.578e-04
[77,] 23  9.409e-01 3.260e-04
```

```
 [78,] 23  9.421e-01 2.971e-04
 [79,] 23  9.433e-01 2.707e-04
 [80,] 23  9.444e-01 2.466e-04
 [81,] 23  9.454e-01 2.247e-04
 [82,] 23  9.465e-01 2.048e-04
 [83,] 23  9.476e-01 1.866e-04
 [84,] 24  9.487e-01 1.700e-04
 [85,] 24  9.497e-01 1.549e-04
 [86,] 23  9.507e-01 1.411e-04
 [87,] 23  9.515e-01 1.286e-04
 [88,] 24  9.524e-01 1.172e-04
 [89,] 25  9.531e-01 1.068e-04
 [90,] 25  9.539e-01 9.728e-05
 [91,] 25  9.546e-01 8.864e-05
 [92,] 25  9.553e-01 8.076e-05
 [93,] 26  9.558e-01 7.359e-05
 [94,] 26  9.564e-01 6.705e-05
 [95,] 26  9.568e-01 6.109e-05
 [96,] 26  9.572e-01 5.567e-05
 [97,] 27  9.576e-01 5.072e-05
 [98,] 27  9.580e-01 4.621e-05
 [99,] 27  9.583e-01 4.211e-05
[100,] 27  9.587e-01 3.837e-05
```

```
cv.fit <- cv.glmnet(x,y,alpha=1)
best_lambda <- cv.fit$lambda.min
coef(fit, s = 0.002)
```

```
x1           .
x2           0.002779505
x3           .
x4           .
x5           .
x6           .
x7           1.543348141
x8          25.721677018
x9           .
x10        -19.261480644
x11          9.241793789
x12         -0.592775885
x13          .
x14          .
x15         93.776388364
x16        -43.752042161
x17          .
x18          .
x19          .
x20        -87.399592049
x21          0.543030712
```

| | |
|---|---|
| x22 | 0.274757393 |
| x23 | 0.003807088 |
| x24 | 0.002584771 |
| x25 | 23.765848649 |
| x26 | . |
| x27 | 4.983585982 |
| x28 | 18.922600415 |
| x29 | 8.423391295 |
| x30 | . |

8. Λογιστική Παλινδρόμηση για το μειωμένο μοντέλο:

*mydataframe17=data.frame(Y,x2,x7,x8,x10,x11,x12,x15,x16,x20,x21,x22,x23,x24,x25,x27,x28,x29)*

*index<-1:nrow(mydataframe)*
*testindex<-sample(index, trunc(length(index)/4))*
*testset<-mydataframe[testindex,]*
*trainset<-mydataframe[-testindex,]*
*xtrainset <-trainset[,2:31]*
*ytrainset <- trainset[,1]*
*xtestset <-testset[,2:31]*
*ytestset <- testset[,1]*
*ytrainset<-as.factor(ytrainset)*
*ytestset<-as.factor(ytestset)*

*logit=glm(Y~x2+x7+x8+x10+x11+x12+x15+x16+x20+x21+x22+x23+x24+x25+x27+x28+x29,binomial(link="logit"),trainset, control=glm.control(maxit=100))*

*confusion.glm(xtrainset, logit)*

*logit=glm(Y~x2+x7+x8x10+x11+x12+x15+x16+x20+x21+x22+x23+x24+x25+x27+x28+x29,binomial(link="logit"),testset, control=glm.control(maxit=100))*

*confusion.glm(xtestset, logit)*

Όμοια για 20 μεταβλητές, 15 μεταβλητές και 10 μεταβλητές.