ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

Εφαρμογές των Μηχανών Διανυσματικής Υποστήριξης σε Προβλήματα Ταξινόμησης και Παλινδρόμησης

(Support Vector Machines in Classification and Regression Problems)

Διπλωματική Εργασία

της

Δανάης Π. Γιαννούλη

Επιβλέπων: Χρήστος Κουκουβίνος
Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2014

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

Εφαρμογές των Μηχανών Διανυσματικής Υποστήριξης σε Προβλήματα Ταξινόμησης και Παλινδρόμησης

(Support Vector Machines in Classification and Regression Problems)

Διπλωματική Εργασία

της

Δανάης Π. Γιαννούλη

Επιβλέπων: Χρήστος Κουκουβίνος
Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2014

Πίνακας περιεχομένων

Περίληψη

Η μηχανική μάθηση έχει ως στόχο τη δημιουργία αλγορίθμων ικανών να βελτιώνουν την απόδοση τους, αξιοποιώντας προγενέστερη γνώση και εμπειρία, με σκοπό την εξαγωγή χρήσιμων συμπερασμάτων και την περιγραφή φαινομένων, μέσω της επεξεργασίας δεδομένων τεράστιου, πολλές φορές, όγκου. Το ζητούμενο στην περίπτωση της επιβλεπόμενης μάθησης είναι η κατασκευή ενός μοντέλου που αναπαριστά τη γνώστη που αποκτήθηκε μέσω της εμπειρίας και το οποίο στη συνέχεια χρησιμοποιείται για την αξιολόγηση νέων παρατηρήσεων. Μία από τις πιο οικείες μεθόδους περιγραφής φαινομένων είναι η ταξινόμηση, η ένταξη δηλαδή κάθε παρατήρησης σε μία ομάδα, από ένα πεπερασμένο πλήθος υποψήφιων ομάδων. Η παρούσα εργασία επικεντρώνεται στην παρουσίαση ενός πολύ διαδεδομένου αλγόριθμου ταξινόμησης, προερχόμενου από τον τομέα της μηχανικής μάθησης, με όνομα «Μηχανή Διανυσματικής Υποστήριξης» (Support Vector Machine - SVM).

Η ανάπτυξη του θεωρητικού υπόβαθρου του αλγόριθμου παρουσιάζεται σταδιακά, ώστε να γίνει κατανοητή από τον αναγνώστη όλη η διαδρομή. Πιο συγκεκριμένα, το πρώτο κεφάλαιο αποτελεί μια εισαγωγή στους αλγόριθμους εξόρυξης δεδομένων (Data Mining) και σε σχετικές εφαρμογές αυτών. Στο δεύτερο κεφάλαιο παρουσιάζονται οι θεμελιώδεις έννοιες που απαιτούνται για την κατανόηση των SVMs. Στο τρίτο και στο τέταρτο κεφάλαιο γίνεται μία λεπτομερής αναφορά στις Μηχανές Διανυσματικής Υποστήριξης και στην Παλινδρόμηση με SVM, αντίστοιχα. Στη συνέχεια, στο πέμπτο κεφάλαιο παρουσιάζουμε τις μεθόδους αξιολόγησης του μοντέλου ενώ στο έκτο κεφάλαιο κάνουμε μία μικρή αναφορά στην επιλογή χαρακτηριστικών με SVM. Στο έβδομο και τελευταίο κεφάλαιο παρουσιάζουμε τρεις εφαρμογές καθώς και την ερμηνεία των αντίστοιχων αποτελεσμάτων, με σκοπό να αξιολογήσουμε τη γνώση που αποκτήσαμε.

Abstract

The aim of machine learning is to develop algorithms capable of improving their own performance, exploiting existing data, stored in huge databases, in order to discover knowledge and interpret several phenomena. Supervised learning aims in creating a model that takes into account the knowledge adapted by experience, and then uses it for evaluating new observations. One of the most common methods for describing phenomena is through classification. Where a particular object is classified to one of several available classes of objects. The presentation thesis focuses on one of the most promising classification algorithms in the field of machine learning, the «The Support Vector Machine» (SVM).

The presentation of the theoretical foundation advances gradually, starting from the most intuitive classification algorithm and reaching up to the optimized approach of SVM, so that it΄s easier for the reader to follow. More specifically, the first chapter is an introduction to data mining algorithms and some related applications. The second chapter presents the fundamental concepts required for an understanding of SVMs. In the third and fourth chapter, there is a detailed report on Support Vector Machines and Regression with SVM, respectively. Then, the fifth chapter presents the evaluation methods of the model while in the sixth chapter a short reference to the feature selection with SVM in made. In the seventh and final chapter three applications and the interpretation of the corresponding results are presented, thus we are able to evaluate the knowledge gained.

Ευχαριστίες

Αρχικά, θέλω να ευχαριστήσω θερμά τον Καθηγητή του Εθνικού Μετσόβιου Πολυτεχνείου, κ. Χρήστο Κουκουβίνο, όχι μόνο για την εμπιστοσύνη που επέδειξε στο πρόσωπό μου αναθέτοντας μου την εκπόνηση αυτής της διπλωματικής εργασίας αλλά και για το γεγονός ότι με έκανε να αγαπήσω τη στατιστική.

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

Επιπλέον, θα ήθελα να ευχαριστήσω τον συμφοιτητή μου και διπλωματούχο του Εθνικού Μετσόβιου Πολυτεχνείου, Ιωάννη Γεωργαρά, για την ιδιαίτερα σημαντική παροχή της βοήθειάς του όποτε του εζητήθη.

Κλείνοντας τον κύκλο των ευχαριστιών, θα ήθελα να ευχαριστήσω μέσα από την καρδιά μου την οικογένειά μου, τους φίλους μου καθώς επίσης και τον κ. Αριστοτέλη Παυλίδη για τη διαρκή συμπαράσταση και υποστήριξη τους καθ΄ όλη τη διάρκεια των σπουδών μου.

Γιαννούλη Δανάη

Εθνικό Μετσόβιο Πολυτεχνείο,
Σχολή Εφαρμοσμένων Μαθηματικών
και Φυσικών Επιστημών
Αθήνα, 2014

ΚΕΦΑΛΑΙΟ 1

Data Mining

1.1 Εισαγωγή

Από πολύ παλιά διαφάνηκε η ανάγκη εξόρυξης μοτίβων και πληροφοριών από δεδομένα. Οι προσπάθειες εξόρυξης πληροφοριών ξεκίνησαν γύρω στο 1700 με το γνωστό θεώρημα του Bayes και συνεχίστηκαν το 1800 με την ανάπτυξη της ανάλυσης παλινδρόμησης. Αυτές οι μέθοδοι ήταν αρκετές για την ανάλυση δεδομένων μέχρι τα μέσα του 1900 οπότε και άρχισε η ραγδαία ανάπτυξη της πληροφορικής και της τεχνολογίας γενικότερα.

Αυτή η ραγδαία ανάπτυξη της πληροφορικής κατέστησε εύκολη τη συλλογή και αποθήκευση δεδομένων και συγχρόνως παρουσιάστηκε η εμφανής αδυναμία των μέχρι τότε γνωστών 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. Το επιτυγχάνει αυτό συνδυάζοντας την επιστήμη της στατιστικής και της πληροφορικής, δηλαδή με τη χρησιμοποίηση μεθόδων που βασίζονται στην τεχνητή νοημοσύνη και εκμάθηση μηχανής.

Η εξόρυξη δεδομένων (data mining) συχνά αναφέρεται ως η γνώση που ανακαλύπτεται από βάσεις δεδομένων (Knowledge discovery in database-KDD).

Στις μέρες μας η σημαντικότητα των εφαρμογών του data mining είναι πλέον φανερή σε πολλούς τομείς όπως η βιοστατική, μετεωρολογία, επιχειρήσεις, μηχανική και φυσικά τα χρηματοοικονομικά.

1.1.1 Εξόρυξη δεδομένων και στατιστική

Η *εξόρυξη δεδομένων* είναι η ανάλυση των (συχνά μεγάλων) παρατηρούμενων συνόλων δεδομένων, για να βρούμε ανυποψίαστες σχέσεις και για να συνοψίσουμε τα δεδομένα με νέους τρόπους που να είναι τόσο κατανοητά όσο και χρήσιμα στον ιδιοκτήτη των δεδομένων.

Οι σχέσεις και οι περιλήψεις που προέρχονται μέσα από μία διαδικασία εξόρυξης δεδομένων συχνά αναφέρονται ως μοντέλα ή μοτίβα. Ο ανωτέρω ορισμός αναφέρεται σε "παρατηρούμενα δεδομένα" σε αντιδιαστολή με "πειραματικά δεδομένα". Η εξόρυξη δεδομένων συνήθως ασχολείται με τα στοιχεία που έχουν ήδη συλλεχθεί για κάποιο σκοπό εκτός από την ανάλυση εξόρυξης δεδομένων (για παράδειγμα, μπορεί να έχουν συλλεχθεί, προκειμένου να διατηρήσουμε μια έως σήμερα καταγραφή όλων των συναλλαγών σε μια τράπεζα). Αυτό σημαίνει ότι ο στόχος της εξόρυξης δεδομένων δεν παίζει κανένα ρόλο στη στρατηγική της συλλογής δεδομένων. Αυτός είναι ένας τρόπος με τον οποίο η εξόρυξη δεδομένων διαφέρει κατά πολύ από τις στατιστικές, στις οποίες τα δεδομένα συχνά συλλέγονται με τη χρήση αποτελεσματικών στρατηγικών για να απαντήσουν σε συγκεκριμένες ερωτήσεις. Για το λόγο αυτό, η εξόρυξη δεδομένων συχνά αναφέρεται ως «δευτερεύουσα» ανάλυση δεδομένων.

Ο ορισμός αναφέρει, επίσης, ότι τα σύνολα δεδομένων που εξετάζονται στην εξόρυξη δεδομένων είναι συχνά μεγάλα. Αν μόνο συμμετείχαν μικρά σύνολα δεδομένων, εμείς απλώς θα χρησιμοποιούσαμε τις κλασικές διερευνητικές μεθόδους ανάλυσης δεδομένων, όπως κάνουν οι στατιστικολόγοι. Όταν βρισκόμαστε αντιμέτωποι με μεγάλα σύνολα δεδομένων, νέα προβλήματα προκύπτουν. Μερικά από αυτά αφορούν στην καθαριότητα για το πώς να αποθηκεύσουμε ή να έχουμε πρόσβαση στα δεδομένα, και άλλα αφορούν σε περισσότερο θεμελιώδη ζητήματα, όπως το πώς θα καθορίσουμε την αντιπροσωπευτικότητα των δεδομένων, πώς θα αναλύσουμε τα δεδομένα σε ένα εύλογο χρονικό διάστημα και πώς θα αποφασίσουμε κατά πόσον μια φαινομενική σχέση είναι απλώς ένα τυχαίο περιστατικό και δεν αντανακλά καμία υποκείμενη πραγματικότητα. Συχνά, τα διαθέσιμα στοιχεία περιλαμβάνουν μόνο ένα δείγμα από τον πληθυσμό και ο στόχος μπορεί να είναι κάποια πιθανή γενίκευση από το δείγμα στον πληθυσμό.

1.1.2 Το πρόβλημα του καιρού

Το πρόβλημα του καιρού [12] αφορά σε ένα μικρό σύνολο δεδομένων που θα χρησιμοποιήσουμε για να απεικονίσουμε μεθόδους μηχανικής μάθησης. Είναι εντελώς εικονικό και υποθέτουμε ότι αφορά στις συνθήκες που είναι κατάλληλες για να παίξουμε κάποιο παιχνίδι. Σε γενικές γραμμές, οι περιπτώσεις (instances) σε ένα σύνολο δεδομένων χαρακτηρίζονται από τις τιμές των χαρακτηριστικών ή τις ιδιότητες, που μετρούν τις διαφορετικές πτυχές της παρουσίας της. Στην περίπτωση μας υπάρχουν τέσσερα χαρακτηριστικά: πρόβλεψη, θερμοκρασία, υγρασία και άνεμος. Το αποτέλεσμα είναι το αν θα παίξουμε ή όχι. Στην απλούστερη μορφή του, που δείχνεται στον Πίνακα 1.2, και τα τέσσερα χαρακτηριστικά που έχουν οι τιμές είναι συμβολικές κατηγορίες αντί για αριθμούς. Η πρόβλεψη μπορεί να πάρει τις τιμές: ήλιος, συννεφιά ή βροχή, η θερμοκρασία: ζέστη, ήπια ή δροσερή, η υγρασία μπορεί να είναι υψηλή ή κανονική και ο άνεμος μπορεί να είναι αληθής ή ψευδής. Αυτό δημιουργεί 36 πιθανούς συνδυασμούς (3x3x2x2=36), εκ των οποίων 14 εμφανίζονται στον παρακάτω πίνακα.

Ένα σύνολο κανόνων που αντλήθηκαν από αυτές τις πληροφορίες (δεν είναι απαραίτητα μια πολύ καλό) μπορεί να μοιάζει ως εξής:

Πρόβλεψη	**Θερμοκρασία**	**Υγρασία**	**Άνεμος**	**Παιχνίδι**
Ήλιος	Ζέστη	Υψηλή	Ψευδής	Όχι
Ήλιος	Ζέστη	Υψηλή	Αληθής	Όχι
Συννεφιά	Ζέστη	Υψηλή	Ψευδής	Ναι
Βροχή	Ήπια	Υψηλή	Ψευδής	Ναι
Βροχή	Δροσερή	Κανονική	Ψευδής	Ναι
Βροχή	Δροσερή	Κανονική	Αληθής	Όχι
Συννεφιά	Δροσερή	Κανονική	Αληθής	Ναι
Ήλιος	Ήπια	Υψηλή	Ψευδής	Όχι
ήλιος	δροσερή	Κανονική	Ψευδής	Ναι
Βροχή	Ήπια	Κανονική	Ψευδής	Ναι
Ήλιος	Ήπια	Κανονική	Αληθής	Ναι
Συννεφιά	Ήπια	Υψηλή	Αληθής	Ναι
Συννεφιά	Ζέστη	Κανονική	Ψευδής	Ναι
βροχή	Ήπια	Υψηλή	Αληθής	Όχι

Εάν πρόβλεψη = ήλιος και υγρασία = υψηλή *τότε παιχνίδι = όχι*
Εάν πρόβλεψη = βροχή και άνεμος = αληθής *τότε παιχνίδι = όχι*
Εάν πρόβλεψη = συννεφιά *τότε παιχνίδι = ναι*
Εάν υγρασία = κανονική *τότε παιχνίδι = ναι*
Εάν τίποτα από τα παραπάνω *τότε παιχνίδι = ναι*

Οι κανόνες αυτοί προορίζονται να ερμηνευτούν με τη σειρά: εφαρμόζεται ο πρώτος κανόνας, στη συνέχεια, αν αυτός δεν ισχύει εφαρμόζεται ο δεύτερος, και ούτω καθεξής. Ένα σύνολο κανόνων που θα πρέπει να ερμηνευθεί στη σειρά καλείται κατάλογος απόφασης. Από ένα κατάλογο απόφασης, οι κανόνες ταξινομούν σωστά όλα τα παραδείγματα του πίνακα, ενώ λαμβάνοντας μεμονωμένα κάποιον, ορισμένοι από τους κανόνες είναι εσφαλμένοι. Για παράδειγμα, ο κανόνας

εάν υγρασία = κανονική *τότε παιχνίδι = ναι*

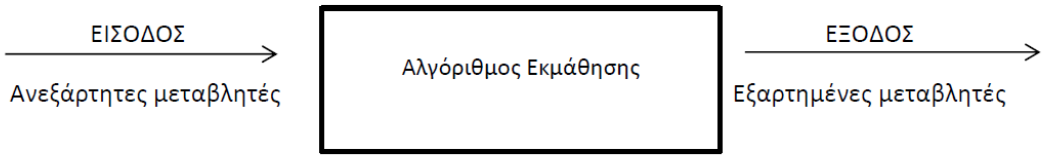
δεν ισχύει μεμονωμένα.

1.1.3 Τεχνικές εξόρυξης γνώσης

Στο data mining έχουμε 2 βασικές τεχνικές εξόρυξης γνώσης:

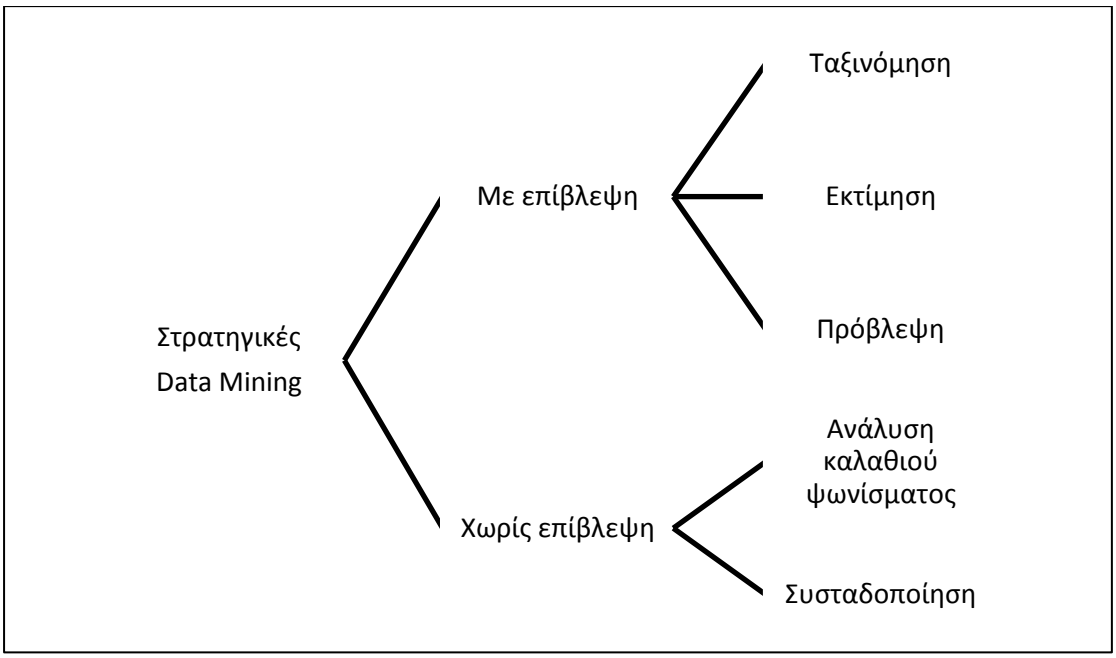
- ✓ **Μέθοδοι με επίβλεψη (Supervised Methods)**: Εδώ τα δεδομένα μου αποτελούνται από ένα διάνυσμα επεξηγηματικών μεταβλητών (*x*) και μία τιμή απόκρισης (y). Το σκεπτικό πίσω από τις μεθόδους με επίβλεψη είναι με την ανάλυση των δεδομένων

μου να δημιουργήσει η μέθοδος μία συνάρτηση ταξινόμησης (ταξινομητής) για να χαρτογραφήσει νέα δεδομένα που δεν συμπεριλαμβάνονται στο αρχικό μου σετ δεδομένων. Μερικά παραδείγματα αυτών των μοντέλων είναι τα νευρωνικά δίχτυα, δέντρα αποφάσεων, λογιστική παλινδρόμηση, μηχανές διανυσματικής υποστήριξης (SVM), boosting και bagging μέθοδοι.



Σχήμα 1.1: Data Mining με επίβλεψη

- ✓ **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 μου. Παραδείγματα αυτών των δομών είναι τα Kohonen Networks, blind signal separation, clustering και K-means.



Σχήμα 1.2: Data Mining χωρίς επίβλεψη

Μέθοδοι με επίβλεψη:

Ταξινόμηση: 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 $D = \{\overrightarrow{x_1}, \overrightarrow{x_2}, \dots . \overrightarrow{x_n}\}$ και ένα σετ κλάσεων $C = \{C_1, C_2, \dots, C_m\}$ προσπαθεί να βρει την βέλτιστη συνάρτηση $f: D \to C$.

Παραδείγματα: εύρεση περιοχών επιρρεπών σε σεισμό, να ταξινομήσεις κάποιο ως πετυχημένο, να δώσεις τιμές 0 ή 1 στα στοιχεία σου.

Εκτίμηση: 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.

Παραδείγματα: Πιθανότητα κάποιος να πάθει καρδιακή προσβολή, πιθανότητα κάποια μετοχή να ανεβεί, πόσα ξοδεύει κάποιος γνωρίζοντας τον μισθό του.

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

Παραδείγματα: Να διευκρινίσει αν κάποιος συνδρομητής μια τηλεφωνικής εταιρείας θα αλλάξει παροχέα στον επόμενο μήνα.

Μέθοδοι χωρίς επίβλεψη

- **Συσταδοποίηση:** Η μέθοδος διαχωρισμού ενός σετ δεδομένων σε διάφορα μέρη (συστάδες). Η δομή γνώσης του προγράμματος αποκτάται με αναφορά στις μετρικές ποιότητας της συστάδας. Στην αρχή της διαδικασίας ο αριθμός των συστάδων πρέπει να είναι γνωστός.

- **Ανάλυση καλαθιού ψωνίσματος (market basket analysis):** Εδώ ανακαλύπτουμε μη διαισθητικές σχέσεις μεταξύ των προϊόντων που πωλούνται. Μετά την ανάλυση ο πάροχος μπορεί να πάρει καλύτερες αποφάσεις για τη διαφήμιση, την παρουσίαση των προϊόντων και τη σειρά με την οποία θα βάλει τα προϊόντα του στα ράφια.

1.1.4 Η διαδικασία του Data Mining

Η Ανακάλυψη Γνώσης στα Δεδομένα (Knowledge discovery in databases -KDD) είναι μια αυτοματοποιημένη διαδικασία ανάλυσης και μοντελοποίησης τεράστιων αποθηκών δεδομένων. Αν και υπάρχουν και άλλες διαδικασίες data mining, η KDD είναι η μέθοδος που χρησιμοποιούν κατά κόρον οι data miners. Η KDD χωρίζεται σε 5 στάδια.

1. **Επιλογή δεδομένων:** Σε αυτό το στάδιο παίρνουμε ολόκληρο το δείγμα των δεδομένων και προσπαθούμε να αποκτήσουμε μία βασική ιδέα για το τι μας λένε, αναλύουμε τα δεδομένα μας. Επίσης, στο παρόν στάδιο θέτουμε τους στόχους και τις προσδοκίες μας.

2. **Προεπεξεργασία:** Η προεπεξεργασία είναι ένα πολύ σημαντικό στάδιο γα τη σωστή εξόρυξη δεδομένων. Εδώ προετοιμάζουμε το στοχευμένο σύνολο δεδομένων (σετ εκπαίδευσης) από το οποίο και τελικά θα εξάγουμε τη ζητούμενη συνάρτηση. Το σετ εκπαίδευσης πρέπει να είναι αρκετά μεγάλο έτσι ώστε να συμπεριλαμβάνει τις σχέσεις και τους συσχετισμούς του συνόλου των δεδομένων μας αλλά και αρκετά μικρό έτσι ώστε να εξάγει τα δεδομένα σε ένα λογικό πλαίσιο.

3. **Μετασχηματισμός δεδομένων**: Εδώ τα δεδομένα μου "καθαρίζονται", δηλαδή απομακρύνεται από αυτά ο θόρυβος, αντιμετωπίζονται οι ελλείπουσες τιμές και τέλος μειώνεται το πλήθος τους αν αυτό είναι δυνατό.

4. **Εξόρυξη δεδομένων.** Η Εξόρυξη δεδομένων χωρίζεται σε 6 υποκατηγορίες:

a. **Εύρεση ανωμαλιών:** Ο προκαθορισμός και η χρήση παράξενων δεδομένων που είτε έχουν κάποιο ιδιαίτερο ενδιαφέρον ή είναι λάθη που πρέπει να διερευνηθούν περαιτέρω.

b. **Σχέση κανόνων εκμάθησης:** Εδώ ανακαλύπτουμε ενδιαφέρουσες σχέσεις μεταξύ των μεταβλητών μας.

c. **Ομαδοποίηση:** Ομαδοποιούμε τα δεδομένα σε κατηγορίες με κοινά στοιχεία.

d. **Ταξινόμηση:** Προσδιορίζουμε σε ποια κατηγορία ανήκουν οι παρατηρήσεις μας.

e. **Παλινδρόμηση:** Προσπαθούμε να βρούμε τη ζητούμενη συνάρτηση που θα κατηγοριοποιεί τα νέα δεδομένα που θα μας παρουσιαστούν.

f. **Σύνοψη:** Παίρνουμε μια πιο συμπαγή σύνοψη των δεδομένων μας.

5. **Επεξήγηση/ Εκτίμηση αποτελεσμάτων:** Σε αυτό το στάδιο επαληθεύουμε ότι τα μοτίβα που εξάγαμε από το σετ εκπαίδευσης μας είναι σωστά. Δεν είναι απαραίτητο ότι όλα τα μοτίβα που έχουμε εξάγει είναι ορθά. Ένα από τα προβλήματα που μπορούν να παρουσιαστούν είναι το overfitting (το μοντέλο που θα επεξηγεί το τυχαίο λάθος (θόρυβο) αντί τη ζητούμενη σχέση). Για να ελέγξουμε την ορθότητα της συνάρτησης

που έχουμε εξάγει την δοκιμάζουμε σε ένα σετ επικύρωσης δεδομένων (validation set) που είναι διαφορετικό από το σετ εκπαίδευσης (training set).

1.1.5 Εφαρμογές

Όπως αναφέραμε και πριν το data mining έχει μια μεγάλη ποικιλία εφαρμογών καθώς είναι ιδανικό εργαλείο για την αντιμετώπιση μεγάλων σετ δεδομένων και τη λήψη αποφάσεων.

Στις μέρες μας το data mining χρησιμοποιείται κυρίως στην ιατρική, τηλεπικοινωνίες μετεωρολογία, διαφήμιση και φυσικά στα χρηματοοικονομικά.

Εφαρμογές στα Χρηματοοικονομικά.

Σε αντίθεση με τους άλλους κλάδους που εφαρμόζεται το data mining οι εφαρμογές του στα χρηματοοικονομικά υπερτερούν λόγω της ευκολίας συσσώρευσης και καταγραφής δεδομένων σε αντίθεση με άλλους κλάδους όπως η ιατρική και η διαφήμιση, όπου η συλλογή δεδομένων δεν είναι μόνο χρονοβόρα αλλά και πολυέξοδη (μερικές φορές ακόμα και της τάξης των εκατομμυρίων). Οι σημαντικότεροι κλάδοι των χρηματοοικονομικών στους οποίους εφαρμόζεται εκτενώς το data mining είναι η διαχείριση ρίσκου, διαχείριση χαρτοφύλακα, trading, η ταυτοποίηση πελατών και η διαχείριση σχέσεων μεταξύ των πελατών.

- **Διαχείριση χαρτοφύλακα:** Η διαχείριση ρίσκου στους χαρτοφύλακες προσεγγίζει συγκεντρωτικά την ποσοτικοποίηση του ρίσκου ενός συνόλου επενδύσεων. Ο προφανής στόχος σε αυτή την περίπτωση είναι η μεγιστοποίηση των κερδών του χαρτοφύλακα μας καθώς και η ελαχιστοποίηση του ρίσκου. Δηλαδή η μεγιστοποίηση του παράγοντα επιστροφή χρημάτων / ρίσκο. Με τη χρήση του data mining είμαστε σε θέση να αναλύσουμε και να εξάγουμε πολλές επενδυτικές στρατηγικές, μέσω των οποίων τα κεφάλαια μας θα μετακινηθούν σε ένα συνδυασμό επενδύσεων που ανάλογα με την τάση της αγοράς θα μας επιφέρει το μεγαλύτερο κέρδος με το μικρότερο δυνατό ρίσκο. Αυτό πετυχαίνεται με τη ανάλυση των τάσεων της αγοράς, ανάλυση κερδών και απωλειών, συναλλαγματικά έξοδα και επιτόκια.

- **Διαχείριση ρίσκου:** Η διαχείριση και μέτρηση ρίσκου βρίσκονται στον πυρήνα κάθε χρηματοπιστωτικού ιδρύματος. Η διαχείριση ρίσκου χωρίζεται σε 2 μεγάλες υποκατηγορίες (ρίσκο χρηματοπιστωτικής αγοράς και πιστωτικό ρίσκο) οι οποίες υπέστησαν μεγάλες αλλαγές βασισμένες στις εξειδικευμένες μεθόδους data mining.

- ✓ **Ρίσκο χρηματοπιστωτικής αγοράς:** Για τα όργανα του χρηματοπιστωτικού ιδρύματος όπως οι χρηματιστηριακοί δείκτες, επιτόκια και συνάλλαγμα η μέτρηση του ρίσκου βασίζεται σε κάποιους παράγοντες ρίσκου όπως τόκοι και οικονομική ανάπτυξη. Το κύριο ενδιαφέρον έρευνας και εφαρμογών βρίσκεται στην συναρτησιακή μορφή μεταξύ των οργάνων του χρηματοπιστωτικού συστήματος και στους παράγοντες ρίσκου. Υπάρχουν πολλές διαφορετικές προσεγγίσεις της μέτρησης του ρίσκου, όλες εξ αυτών βασισμένες σε μοντέλα που αντιπροσωπεύουν την αλληλεξάρτηση των παραγόντων με την συνολική αγορά. Οι περισσότερες από αυτές τις μεθόδους μπορούν να κατασκευαστούν μόνο με τη χρησιμοποίηση τεχνικών data mining σε ιδιωτικά δεδομένα που χρειάζονται συνεχή επίβλεψη.

- ✓ **Πιστωτικό Ρίσκο:** Η ανάλυση του πιστωτικού ρίσκου είναι το κύριο συστατικό στη διαδικασία των εμπορικών δανείων. Χωρίς αυτό ο εκάστοτε δανειστής δεν θα είχε κανένα στοιχείο για το ρίσκο του δανεισμού και άρα θα ήταν ανίκανος να αποφασίσει για το αν θα δανείσει τα χρήματά του και με τη όρους θα τα δανείσει. Εδώ βασικά το ρίσκο αντιπροσωπεύει την πιθανότητα μη επιστροφής των χρημάτων στον δανειστή. Σε αυτή την περίπτωση το μεγαλύτερο μέρος του συστήματος διαχείρισης ρίσκου θα είναι ένα τυπικό πρόβλημα data mining με επεξηγηματικές μεταβλητές την «αξία» του πιστωτή, ρυθμό επιστροφής χρημάτων, προηγούμενα δάνεια κτλ.

- **Trading:** Ένας από του περισσότερο αναπτυσσόμενους τομείς του data mining είναι η προσπάθεια για κτίσιμο εργαλείων trading, που βασισμένα στα ιστορικά δεδομένα της αγοράς θα είναι σε θέση να προβλέψει τη βραχυπρόθεσμη κίνηση των επιτοκίων, συναλλαγμάτων και μετοχών. Ο στόχος μας είναι να καταφέρουμε να εντοπίσουμε πότε η αγορά είναι “φτηνή” και πότε είναι “ακριβή”. Δηλαδή να εξάγουμε συμβουλές για το πότε η μια μετοχή/συνάλλαγμα είναι υπερεκτιμημένη ή υποτιμημένη με στόχο την πώληση/αγορά του ανάλογου αγαθού τη συγκεκριμένη χρονική στιγμή.

Αν και παραδοσιακά οι αγοραπωλησίες βασίζονται στο ένστικτο και πείρα των trader ακόμα και ο πιο έμπειρος trader μπορεί να συνυπολογίσει ένα μικρό αριθμό παραγόντων καθιστώντας την απόφασή του αμφιλεγόμενη. Λόγω των πολλών δεδομένων που είναι διαθέσιμα για την κίνηση των μετοχών/συναλλάγματος κτλ είναι προφανές πως το data mining είναι ιδανική επιλογή, αν όχι για πλήρη λήψη αποφάσεων, για βοηθητικό εργαλείο για κάθε trader έτσι ώστε να βρίσκει μοτίβα και δομές που ο ίδιος θα ήταν αδύνατο να ανακαλύψει.

- **Ξέπλυμα βρώμικου χρήματος:** Η αστυνομική λογιστική είναι ένας τομέας υπεύθυνος για την ανακάλυψη παράνομων οικονομικών συναλλαγών με κυριότερο σκοπό την εύρεση επιχορηγήσεων μηχανισμών τρομοκρατίας, όπου “καθαρά ” και “βρώμικα” κεφάλαια χρησιμοποιούνται για αγορά και κατασκευή όπλων. Ο στόχος του data mining σε αυτόν τον τομέα είναι ο εντοπισμός

ύποπτων συναλλαγών καθώς και η μείωση των «λανθασμένα θετικών» ύποπτων συναλλαγών. Αυτό μπορεί να επιτευχθεί με το data mining ψάχνοντας για ύποπτους πολύπλοκους συνδυασμούς συναλλαγών με χρησιμοποίηση κυρίως μεθόδων χωρίς επίβλεψη.

Υπάρχουν αμέτρητες άλλες εφαρμογές της μηχανής μάθησης [12]. Εδώ αναφέρουμε εν συντομία λίγες περισσότερες περιοχές για να τονίσουμε το εύρος αυτής:

- **Στον τομέα του μάρκετινγκ και των πωλήσεων:** Πρόκειται για τομείς στους οποίους οι εταιρείες έχουν τεράστιους όγκους με ακριβή καταγεγραμμένα δεδομένα, στοιχεία τα οποία είναι εξαιρετικά πολύτιμα. Σε αυτές τις εφαρμογές, οι προβλέψεις είναι το κύριο ενδιαφέρον, ο τρόπος με τον οποίο λαμβάνονται οι αποφάσεις δεν μας ενδιαφέρει.

- **Στον τομέα της υποστήριξης και εξυπηρέτησης πελατών:** Όταν ένας πελάτης αναφέρει ένα τηλεφωνικό πρόβλημα και η εταιρεία πρέπει να αποφασίσει σε τι είδους τεχνικό να αναθέσει τη δουλειά. Ένα έμπειρο σύστημα (που αναπτύχθηκε από την Bell Atlantic το 1991 και αντικαταστάθηκε το 1999) από ένα σύνολο κανόνων βασισμένων στη χρήση μηχανικής μάθησης, απέφερε κέρδος πάνω από 10 εκατομμύρια δολάρια ανά έτος εξαιτίας της λήψης λιγότερων εσφαλμένων αποφάσεων.

- Όταν κάνουμε αίτηση για ένα δάνειο, θα πρέπει να συμπληρώσουμε ένα ερωτηματολόγιο που ζητά οικονομικές και προσωπικές πληροφορίες. Αυτές οι πληροφορίες χρησιμοποιούνται από το σύστημα και η εταιρεία δανείου αποφασίζει αν θα μας χορηγήσει το δάνειο ή όχι.

- **Στη βιολογία**: Η μηχανική μάθηση χρησιμοποιείται για τον εντοπισμό των χιλιάδων γονιδίων σε κάθε νέο γονιδίωμα.

- **Στη βιοϊατρική:** Χρησιμοποιείται για να προβλέψει τη δραστικότητα του φαρμάκων όχι μόνο με την ανάλυση των χημικών ιδιοτήτων τους αλλά και των τρισδιάστατων δομών τους. Αυτό επιταχύνει την ανακάλυψη φαρμάκων και μειώνει το κόστος τους.

- **Στην αστρονομία:** Η μηχανική μάθηση έχει χρησιμοποιηθεί για την ανάπτυξη ενός πλήρως αυτόματου συστήματος καταγραφής ουράνιων αντικειμένων που είναι πάρα πολύ εξασθενημένα για να τα διακρίνουμε εύκολα.

1.2 Εξόρυξη δεδομένων και ηθική

Η χρήση των δεδομένων και ιδίως των ανθρώπινων δεδομένων, για την εξόρυξη πληροφοριών, μπορεί έχει σοβαρές ηθικές επιπτώσεις και οι επαγγελματίες των τεχνικών εξόρυξης δεδομένων πρέπει να δρουν υπεύθυνα γνωρίζοντας ηθικά ζητήματα που ενδέχεται να προκύψουν [12]. Όταν εφαρμόζεται στους ανθρώπους, εξόρυξη δεδομένων χρησιμοποιείται συχνά για να διαφόρων ειδών διακρίσεις όπως ποιός θα πάρει ένα δάνειο, ποιος θα πάρει μία ειδική προσφορά, και ούτω καθεξής. Ορισμένα είδη διακρίσεων όπως φυλετικές, σεξουαλικές, θρησκευτικές δεν είναι μόνο ανήθικες αλλά και παράνομες.

Όσοι ασχολούνται με δεδομένα, θα πρέπει να γνωρίζουν ποιος επιτρέπεται να έχει πρόσβαση σε αυτά, για ποιο σκοπό έχουν συγκεντρωθεί, και τι είδους συμπεράσματα είναι νόμιμο να αντλήσουμε από αυτά. Από ηθικής απόψεως είναι απαραίτητο να εξετάζουμε τους κανόνες της κοινότητας που χρησιμοποιούμε για να αποφύγουμε τυχόν προβλήματα.

Εκπληκτικά πράγματα προκύπτουν από την εξόρυξη δεδομένων. Για παράδειγμα, έχει αναφερθεί ότι μία από τις κορυφαίες ομάδες καταναλωτών στη Γαλλία, διαπίστωσε ότι οι άνθρωποι με κόκκινα αυτοκίνητα είναι πιο πιθανό να μην μπορούν να πληρώσουν τα δάνεια των αυτοκινήτων τους. Ποια είναι η κατάσταση μιας τέτοιας «ανακάλυψη»; Σε ποιες πληροφορίες βασίζεται; Κάτω από ποιες συνθήκες συλλέχθηκαν οι πληροφορίες; Με ποιους τρόπους είναι ηθικό να χρησιμοποιηθούν; Σαφώς, οι ασφαλιστικές εταιρείες βασίζονται σε στερεότυπα-νέοι άνδρες πληρώνουν ακριβά για την ασφάλιση αυτοκινήτων-αλλά αυτά τα στερεότυπα δεν βασίζονται αποκλειστικά σε στατιστικά συμπεράσματα, περιλαμβάνουν επίσης την κοινή λογική.

Το θέμα είναι ότι η εξόρυξη δεδομένων είναι απλώς ένα εργαλείο στην όλη διαδικασία: οι άνθρωποι είναι αυτοί που παίρνουν τα αποτελέσματα και μαζί με άλλες γνώσεις αποφασίζουν τι μέτρα θα εφαρμόσουν.

1.3 Machine Learning

Η μηχανική μάθηση (machine learning) είναι μια περιοχή της τεχνητής νοημοσύνης η οποία αφορά αλγόριθμους και μεθόδους που επιτρέπουν στους υπολογιστές να «μαθαίνουν». Με τη μηχανική μάθηση καθίσταται εφικτή η κατασκευή προσαρμόσιμων (adaptable) προγραμμάτων υπολογιστών τα οποία λειτουργούν με βάση την αυτοματοποιημένη ανάλυση συνόλων δεδομένων και όχι τη διαίσθηση των μηχανικών που τα προγραμμάτισαν. Η μηχανική μάθηση επικαλύπτεται σημαντικά με τη στατιστική, αφού και τα δύο πεδία μελετούν την ανάλυση δεδομένων.

Το 1959, ο πρωτοπόρος σχεδιαστής παιχνιδιών Άρθρουρ Σάμουελ όρισε ως μηχανική μάθηση: "Το πεδίο μελέτης όπου δίνει στους υπολογιστές την δυνατότητα να μαθαίνουν χωρίς να έχουν προγραμματιστεί".

Το 1997 ο Τομ Μ. Μιτσέλ έδωσε ένα πιο επίσημο ορισμό ο οποίος χρησιμοποιείται ευρέως: *"Ένα πρόγραμμα υπολογιστή λέγεται ότι μαθαίνει από μια εμπειρία Ε σε σχέση μια σειρά από έργα Τ και μια μέτρηση της απόδοσης Ρ ή οποία βελτιώνεται με την εμπειρία Ε"*
("A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E").

Αρκετοί άνθρωποι θεωρούν την εξόρυξη δεδομένων συνώνυμη με την μηχανική μάθηση. Δεν υπάρχει καμία αμφιβολία ότι κάποιες μέθοδοι εξόρυξης δεδομένων χρησιμοποιούν κατάλληλους αλγόριθμους μηχανικής μάθησης [9]. Επαγγελματίες της μηχανικής μάθησης χρησιμοποιούν τα δεδομένα ως ένα σύνολο εκπαίδευσης, για να εκπαιδεύσουν έναν αλγόριθμο από τους πολλούς τύπους που χρησιμοποιούνται από τους επαγγελματίες μηχανικής μάθησης, όπως τα δίκτυα Bayes, οι μηχανές διανυσματικής υποστήριξης, τα δέντρα απόφασης και πολλά άλλα.

Υπάρχουν περιπτώσεις όπου η χρήση των δεδομένων με τον τρόπο αυτό έχει νόημα. Η τυπική περίπτωση μηχανικής μάθησης είναι μια καλή προσέγγιση είναι όταν έχουμε μικρή ιδέα για το τι ψάχνουμε για τα δεδομένα. Για παράδειγμα, είναι μάλλον ασαφές τι είναι αυτό που κάνει μία ταινία αρεστή ή όχι στους θεατές. Έτσι, σε απάντηση της "Netflix πρόκληση" για την επινόηση ενός αλγόριθμου που προβλέπει τις αξιολογήσεις των ταινιών από τους χρήστες, με βάση ένα δείγμα των απαντήσεων τους, οι αλγόριθμοι μηχανικής μάθησης έχουν αποδειχθεί αρκετά επιτυχείς.

Από την άλλη πλευρά, η μηχανική μάθηση δεν έχει αποδειχθεί επιτυχής σε καταστάσεις όπου μπορούμε να περιγράψουμε τους στόχους της εξόρυξης (mining) πιο άμεσα. Μια ενδιαφέρουσα περίπτωση είναι η προσπάθεια από το Whiz Bang! Labs1 να χρησιμοποιήσει τη μηχανική μάθηση για να εντοπίσει βιογραφικά ανθρώπων στο Web. Δεν ήταν σε θέση να το κάνει καλύτερα από αλγόριθμους σχεδιασμένους στο χέρι που ψάχνουν για κάποιες από τις προφανείς λέξεις και φράσεις που εμφανίζονται στο τυπικό βιογραφικό. Δεδομένου ότι ο καθένας ο οποίος έχει εξετάσει ή γράψει ένα βιογραφικό έχει μια αρκετά καλή ιδέα για το τι περιέχει το βιογραφικό, δεν υπήρχε κανένα μυστήριο σχετικά με το τι κάνει μια ιστοσελίδα για ένα βιογραφικό να διαφέρει (what makes a Web page a resume). Έτσι, δεν υπήρχε κανένα πλεονέκτημα της μηχανικής μάθησης σε σύγκριση με την άμεση σχεδίαση ενός αλγορίθμου που ανακαλύπτει βιογραφικά.

ΚΕΦΑΛΑΙΟ 2

ΘΕΜΕΛΙΩΔΕΙΣ ΕΝΝΟΙΕΣ

2.1 Εισαγωγή

Στο κεφάλαιο που ακολουθεί θα εισάγουμε ορισμένες βασικές έννοιες που αφορούν στις μηχανές διανυσματικής υποστήριξης [5]. Αρχικά, θα δώσουμε τον ορισμό ενός κυρτού προβλήματος και μίας κυρτής συνάρτησης και θα αναπτύξουμε τη σπουδαιότητα του να ασχολούμαστε με τέτοιου είδους προβλήματα βελτιστοποίησης. Στη συνέχεια, θα αναφερθούμε σε ταξινομητές (SVM) μέγιστου περιθωρίου. Η ιδιότητα αυτή είναι ιδιαίτερα χρήσιμη καθώς μας βοηθάει να αποφύγουμε τη λάθος ταξινόμηση των δεδομένων. Έπειτα, θα δώσουμε ιδιαίτερη έμφαση στη θεωρία Lagrange, την οποία και θα εφαρμόσουμε για να παράγουμε τις μηχανές διανυσματικής υποστήριξης. Ακολούθως, θα ασχοληθούμε με τη VC διάσταση. Τέλος, θα συνοψίσουμε τα βασικότερα πλεονεκτήματα των μηχανών διανυσματικής υποστήριξης.

2.1.1 Μηχανές Διανυσματικής Υποστήριξης και Κυρτή Βελτιστοποίηση

Βασική έννοια για τις μηχανές διανυσματικής υποστήριξης είναι η έννοια της κυρτότητας. Αυτό γίνεται ξεκάθαρο αν αναλογιστεί κανείς ότι ουσιαστικά έχουμε να αντιμετωπίσουμε ένα πρόβλημα βελτιστοποίησης. Η έννοια της κυρτότητας και κατ' επέκταση των κυρτών συναρτήσεων αποτελεί βασικό κομμάτι στην επίλυση προβλημάτων βελτιστοποίησης. Ένα κυρτό σύνολο είναι μια περιοχή όπως η παρακάτω. Είναι σαφές ότι, οποιοδήποτε ευθύγραμμο τμήμα συνδέει δύο σημεία της περιοχής παραμένει μέσα στο σύνολο:

Σχήμα 2.1: Κυρτό σύνολο

Αντιθέτως, το ακόλουθο σύνολο είναι μη κυρτό καθώς μέρος ενός ευθύγραμμου τμήματος δεν βρίσκεται μέσα στο σύνολο:

Σχήμα 2.2: Μη κυρτό σύνολο

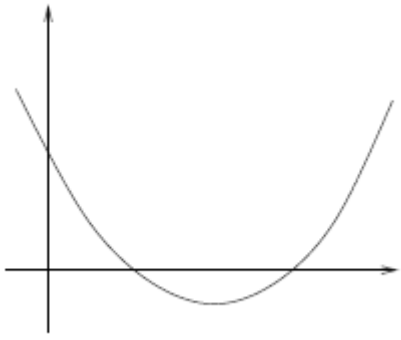
Ορισμός (Κυρτή Περιοχή)

Ένα σύνολο Α είναι κυρτό αν για κάθε $x_1, x_2 \in A$, ισχύει

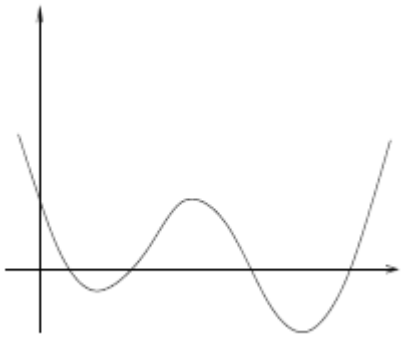
$$\lambda x_1 + (1-\lambda)x_2 \in A, \qquad \forall 0 \leq \lambda \leq 1.$$

Σημειώνουμε ότι το $\lambda x_1 + (1-\lambda)x_2 \in A,\ \forall 0 \leq \lambda \leq 1$, περιλαμβάνει όλα τα σημεία της περιοχής που συνδέουν τα x_1 και x_2.

Μία έννοια που σχετίζεται στενά με τα παραπάνω είναι η κυρτή συνάρτηση. Στο ακόλουθο σχήμα φαίνεται η διαφοροποίηση μεταξύ μίας κυρτής και μιας μη κυρτής συνάρτησης.



Κυρτή συνάρτηση



Μη κυρτή συνάρτηση

Η κύρια διαφορά μεταξύ των δύο παραπάνω σχημάτων είναι ότι η κυρτή συνάρτηση έχει ένα μοναδικό "τοπικό ελάχιστο", ενώ η μη κυρτή συνάρτηση έχει δύο.

Ορισμός (Τοπικό Ελάχιστο)

Μία συνάρτηση f, με πεδίο ορισμού A, θα λέμε ότι παρουσιάζει στο $x_0 \in A$ τοπικό ελάχιστο, όταν υπάρχει θετικός δ, τέτοιος ώστε να ισχύει:

$$f(x_0) \leq f(x) \text{ για κάθε } x \in (x_0 - \delta, x_0 + \delta) \cap A.$$

Το x_0 λέγεται θέση ή σημείο τοπικού ελαχίστου.

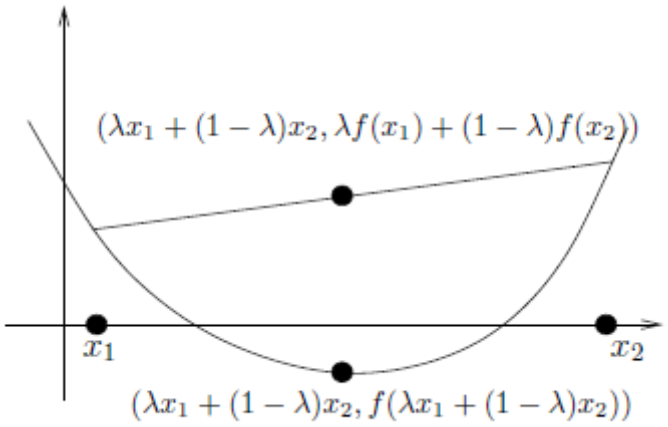
Το $f(x_0)$ λέγεται τοπικό ελάχιστο της f.

Ορισμός (Κυρτή Συνάρτηση)

Μια συνάρτηση $f(x)$ είναι κυρτή αν για κάθε x_1, x_2, ισχύει

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2), \qquad \forall 0 \leq \lambda \leq 1.$$

Δηλαδή, όπως φαίνεται στο ακόλουθο Σχήμα 2.3, κάθε ευθύγραμμο τμήμα που συνδέει δύο σημεία $(x_1, f(x_1))$ και $(x_2, f(x_2))$ είναι πάνω από την $f(x)$, όπου το x είναι μεταξύ των x_1 και x_2.



Σχήμα 2.3

Εφαρμογή στο πρόβλημα των μηχανών διανυσματικής υποστήριξης

Η βασική συνάρτηση για τη λύση του προβλήματος βελτιστοποίησης στα SVM είναι, προφανώς, η αντικειμενική συνάρτηση. Θα δείξουμε ότι η αντικειμενική συνάρτηση SVM

$$f(x) = \frac{1}{2} w^T w + C \sum_{i=1}^{l} {\xi_i}^2$$

είναι κυρτή. Για να το δείξουμε αυτό θα θέλαμε να έχουμε

$$\frac{1}{2}(\lambda w^1 + (1-\lambda)w^2)^T(\lambda w^1 + (1-\lambda)w^2) + C \sum_{i=1}^{l}(\lambda \xi_i^1 + (1-\lambda)\xi_i^2)$$

$$\leq \lambda(\frac{1}{2} {w^1}^T w^1 + C \sum_{I=1}^{l} \xi_i^2) + (1-\lambda)\,(\frac{1}{2} {w^2}^T w^2 + C \sum_{I=1}^{l} \xi_i^2)$$

Είναι εύκολο να απαλείψουμε τους όρους που περιλαμβάνουν το ξ_i. Στη συνέχεια, η διαφορά μεταξύ του δεξιού και του αριστερού μέρους της ανισότητας είναι:

$$\lambda\left(\frac{1}{2} {w^1}^T w^1\right) + (1-\lambda)\left(\frac{1}{2} {w^2}^T w^2\right) - \frac{1}{2}(\lambda w^1 + (1-\lambda)w^2)^T(\lambda w^1 + (1-\lambda)w^2)$$

$$= \frac{1}{2}\lambda(1-\lambda)(w^1 - w^2)^T(w^1 - w^2) \geq 0 \qquad (2.1)$$

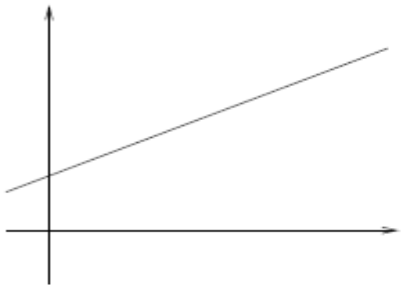
Από τα παραπάνω είναι σαφές ότι οι γραμμικές συναρτήσεις στον R^2 είναι κυρτές συναρτήσεις. Αναλογιζόμενοι αυτό, μπορούμε να δώσουμε τον ορισμό της «αυστηρά κυρτής» συνάρτησης:

Ορισμός (Αυστηρά Κυρτή Συνάρτηση)

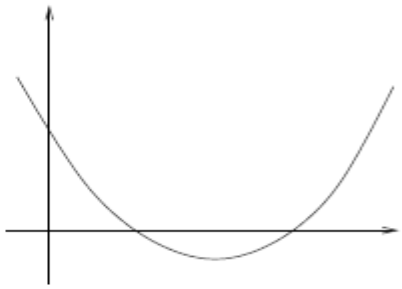
Μια συνάρτηση $f\ (x)$ είναι αυστηρά κυρτή αν για κάθε $x_1 \neq x_2$, ισχύει

$$f(\lambda x_1 + (1-\lambda)\, x_2) < \lambda f(x_1) + (1-\lambda)\, f(x_2), \qquad \forall\ 0 < \lambda < 1$$

Σαφώς, η $f(w) = \frac{1}{2} w^T w$ είναι μια αυστηρά κυρτή συνάρτηση καθώς στην σχέση (2.1), εάν 0 <λ <1, το "≥" γίνεται ">."



Κυρτή



Αυστηρά κυρτή

Στο παραπάνω σχήμα φαίνεται ο διαχωρισμός μεταξύ μίας κυρτής και μίας αυστηρά κυρτής συνάρτησης.

Για μία αυστηρά κυρτή συνάρτηση, εάν είναι τετραγωνική, μπορούμε εύκολα να βρούμε το ολικό ελάχιστο της:

Θεώρημα

Θεωρούμε μια αυστηρά κυρτή τετραγωνική συνάρτηση που έχει τη μόρφη

$$f(x) = \frac{1}{2}x^T Q x + p^T x$$

Τότε, $\bar{x}$ είναι το μοναδικό ολικό ελάχιστο $\Leftrightarrow Q\bar{x} + p = 0$.

Απόδειξη

Αρχικά, ισχυριζόμαστε ότι για την αυστηρά κυρτή τετραγωνική συνάρτηση ισχύει:

$$d^T Q d > 0, \forall d \neq 0$$

Για οποιαδήποτε δύο διανύσματα x και d,

$$f(x + \lambda d) = \frac{1}{2}(x + \lambda d)^T Q(x + \lambda d) + p^T(x + \lambda d) =$$

$$= f(x) + \lambda(Qx + p)^T d + \frac{1}{2}\lambda^2 d^T Q d.$$

Από την άλλη μεριά έχουμε, για κάθε $0 < \lambda < 1$,

$$f(x + \lambda d) = f\big((1 - \lambda)x + \lambda(x + d)\big) \leq (1 - \lambda)f(x) + \lambda f(x + d) =$$
$$= (1 - \lambda)f(x) + \lambda f(x) + \lambda(Qx + p)^T d + \frac{\lambda}{2} d^T Q d$$

Έτσι,

$$\frac{\lambda^2}{2} d^T Q d \leq \frac{\lambda}{2} d^T Q d$$

και επομένως:

$$d^T Q d \geq 0.$$

Τώρα μπορούμε να αποδείξουμε το θεώρημα:

"⇐" Για κάθε $\hat{x} \neq \bar{x}$. Αν $d \equiv \hat{x} - \bar{x}$, τότε:

$$f(\hat{x}) = \frac{1}{2}(\bar{x}+d)^T Q(\bar{x}+d) + p^T(\bar{x}+d) =$$

$$= f(\bar{x}) + (Q\bar{x}+p)^T d + \frac{1}{2}d^T Qd > f(\bar{x}).$$

Επομένως, το $\bar{x}$ είναι το μοναδικό ολικό ελάχιστο.

"$\Longrightarrow$" Εάν $Q\bar{x}+p \neq 0$, θεωρούμε $d = -t(Q\bar{x}+p),$ με $t > 0$.

$$f(\bar{x}+d) = f(\bar{x}) + (Q\bar{x}+p)^T d + \frac{1}{2}d^T Qd$$

$$= f(\bar{x}) - t(Q\bar{x}+p)^T(Q\bar{x}+p) + \frac{1}{2}t^2(Q\bar{x}+p)^T Q(Q\bar{x}+p).$$

Αν $(Q\bar{x}+p)^T Q(Q\bar{x}+p) > 0$ και

$$t < \frac{(Q\bar{x}+p)^T(Q\bar{x}+p)}{(Q\bar{x}+p)^T Q(Q\bar{x}+p)},$$

Τότε

$$f(\bar{x}+d) < f(\bar{x})$$

και το $\bar{x}$ δεν είναι ολικό βέλτιστο. Αυτό προκαλεί μια αντίφαση. Από την άλλη πλευρά, αν

$$(Q\bar{x}+p)^T Q(Q\bar{x}+p) = 0,$$

για κάθε $t > 0$, $f(\bar{x}) < f(\bar{x}+d)$ προκαλεί επίσης μια αντίφαση.

Σημειώνουμε ότι αν ο Q είναι θετικά ορισμένος,

$$Qx + p = 0$$

έχει μια μοναδική λύση την $-Q^{-1}p$.

Έτσι, η $f\ (x)$ έχει ένα μοναδικό ολικό ελάχιστο.

2.1.2 Πρόβλημα Βελτιστοποίησης

Προβλήματα βελτιστοποίησης είναι τα προβλήματα στα οποία θέλουμε να επιλέξουμε την καλύτερη λύση από έναν αριθμό πιθανών ή εφικτών λύσεων [4]. Τυπικά, οι εφικτές λύσεις ταξινομούνται από μια αντικειμενική συνάρτηση και ο στόχος είναι να βρεθεί η εφικτή λύση που ελαχιστοποιεί (ή μεγιστοποιεί) την τιμή αυτής της συνάρτησης. Στα περισσότερα προβλήματα βελτιστοποίησης έχουμε επίσης ένα σύνολο περιορισμών που περιορίζουν το χώρο λύσεων, δηλαδή, οι περιορισμοί θέτουν τα όρια ως προς το αν μία λύση είναι εφικτή ή όχι. Επίσημα, μπορούμε να εκφράσουμε ένα πρόβλημα βελτιστοποίησης ως εξής:

$$min_{\bar{x}}\varphi(\bar{x}),$$

έτσι ώστε

$$h_i(\bar{x}) \geq c_i,$$

με $i = 1, \dots l$ και για κάθε $\bar{x} \in R^n$. Εδώ η συνάρτηση $\varphi: R^n \to R$ είναι η αντικειμενική συνάρτηση, και κάθε συνάρτηση $h_i: R^n \to R$ καλείται περιορισμός με όριο το c_i. Κάθε τιμή $\bar{x} \in R^n$ που ικανοποιεί τους περιορισμούς ονομάζεται εφικτή λύση. Η βελτιστοποίηση έχει ως στόχο την εύρεση της εφικτής λύσης $\overline{x^*}$,που ελαχιστοποιεί την αντικειμενική συνάρτηση έτσι ώστε για κάθε άλλη εφικτή λύση, έχουμε

$$\varphi(\overline{x^*}) \leq \varphi(\bar{q})$$

Έχουμε ορίσει προβλήματα βελτιστοποίησης που έχουν να κάνουν με ελαχιστοποίηση. Αυτό δεν μας περιορίζει από τη στιγμή που μπορούμε να μετατρέψουμε ένα οποιοδήποτε πρόβλημα μεγιστοποίησης σε ένα πρόβλημα ελαχιστοποίησης χρησιμοποιώντας έναν από τους παρακάτω μετασχηματισμούς:

$$\max \varphi(\bar{x}) = \min\big(-\varphi(\bar{x})\big)\,, \qquad \max \varphi(\bar{x}) = min\,\frac{1}{\varphi(\bar{x})}$$

εφόσον η $1/(\varphi(\bar{x}))$ είναι καλά ορισμένη.

Τα προβλήματα βελτιστοποίησης ταξινομούνται σύμφωνα με οι ιδιότητες των αντίστοιχων αντικειμενικών συναρτήσεων και των περιορισμών τους. Για παράδειγμα, ένα γραμμικό πρόβλημα βελτιστοποίησης έχει μια γραμμική αντικειμενική συνάρτηση και γραμμικούς περιορισμούς. Όταν η αντικειμενική συνάρτηση ή οι περιορισμοί δεν είναι γραμμικοί, το πρόβλημα βελτιστοποίησης θεωρείται μη γραμμικό.

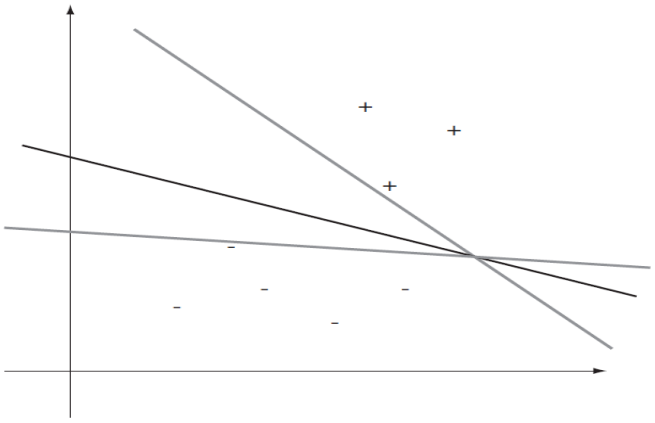
Εμείς θα ασχοληθούμε με κυρτά προβλήματα βελτιστοποίησης. Ένα κυρτό πρόβλημα βελτιστοποίησης έχει μια

- ✓ κυρτή αντικειμενική συνάρτηση και
- ✓ γραμμικούς περιορισμούς.

Αξίζει να σημειώσουμε ότι υπάρχουν αποδοτικοί αλγόριθμοι, που επωφελούνται από την κυρτότητα μιας συνάρτησης, προκειμένου να λύσουν ένα κυρτό πρόβλημα βελτιστοποίησης. Μια τέτοια τεχνική είναι ο τετραγωνικός προγραμματισμός.

2.1.3 Μέγιστο Περιθώριο

Αν έχουμε ένα γραμμικά διαχωρίσιμο σύνολο δεδομένων για ένα δυαδικό πρόβλημα ταξινόμησης, ίσως διαισθητικά βγάλουμε το συμπέρασμα ότι η βέλτιστη επιφάνεια απόφασης είναι αυτή που ισαπέχει από τα όρια των κλάσεων. Ανεπίσημα, μπορούμε να το δικαιολογήσουμε αυτό με το επιχείρημα ότι το σύνολο εκπαίδευσης είναι μόνο μια προσεγγιστική αναπαράσταση του συνόλου των δεδομένων μας και άρα, τοποθετώντας την επιφάνεια απόφασης σε ίση απόσταση από τα αντίστοιχα όρια των κλάσεων θα αυξήσουμε την πιθανότητα της σωστής ταξινόμησης των σημείων που δεν ανήκουν στο σύνολο εκπαίδευσης. Εάν, επιπλέον, μεγιστοποιήσουμε τις αποστάσεις ανάμεσα στην επιφάνεια απόφασης και στα όρια των κλάσεων θα αυξήσουμε την πιθανότητα εύρεσης της βέλτιστης επιφάνειας απόφασης. Στο Σχήμα 2.4 απεικονίζεται ένας R^2 χώρος, όπου η έντονη γραμμή θεωρείται μια καλύτερη επιφάνεια απόφασης από τις γκρίζες γραμμές.

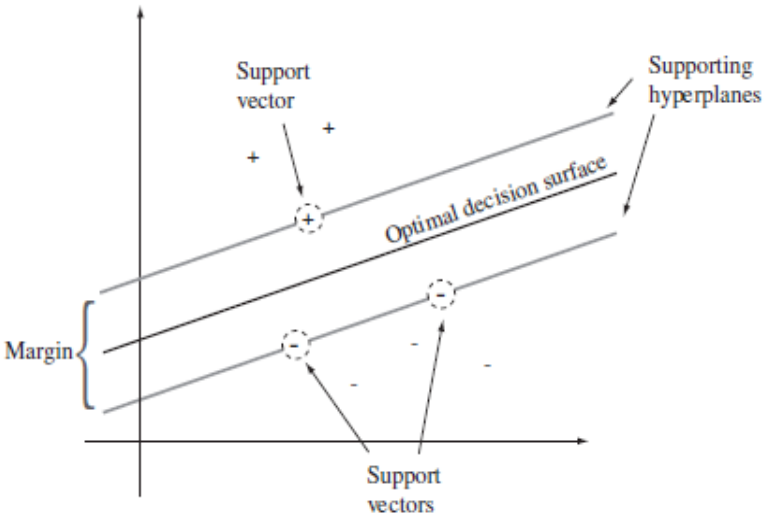


Σχήμα 2.4

Στο Σχήμα 2.5 βλέπουμε τα εξής:

Υπάρχουν δύο υπερεπίπεδα υποστήριξης τα οποία «ελαφρώς» αγγίζουν τα όρια των κλάσεων. Η απόσταση ανάμεσα στα υπερεπίπεδα είναι το περιθώριο και η βέλτιστη επιφάνεια απόφασης βρίσκεται στο κέντρο του περιθωρίου. Παρατηρούμε ότι το μέγεθος του περιθωρίου περιορίζεται από τα σημεία της κάθε κλάσης που είναι σε κύκλο, τα οποία ονομάζονται ***διανύσματα υποστήριξης***.

Εάν ένα υπερεπίπεδο υποστήριξης μετακινηθεί προς την κλάση του, τότε θα πάψει να είναι υπερεπίπεδο υποστήριξης διότι στοιχεία της ίδιας κλάσης θα εμφανίζονται και στις δύο πλευρές του υπερεπιπέδου. Ακόμα, εύκολα μπορούμε να αντιληφθούμε ότι το περιθώριο είναι μέγιστο. Κάθε περιστροφή ή μεταφορά της επιφάνειας απόφασης θα οδηγήσει σε ένα μικρότερο περιθώριο.



Σχήμα 2.5

2.2 Μεθοδολογία Επίλυσης του Μοντέλου

Στη θεωρία βελτιστοποίησης, όταν από ένα αρχικό πρόβλημα βελτιστοποίησης υπολογίσουμε το δυϊκό του, είμαστε συχνά σε θέση να εξάγουμε νέες γνώσεις που απορρέουν από το πρόβλημα. Αυτές οι νέες γνώσεις μπορούν να οδηγήσουν σε καινούργιες τεχνικές για την επίλυση του προβλήματος βελτιστοποίησης ή, όπως θα δούμε στην περίπτωση των μηχανών διανυσματικής υποστήριξης, μπορεί να οδηγήσουν σε εντελώς νέες κατηγορίες αλγορίθμων βελτιστοποίησης. Θα αναπτύξουμε μια ιδιαίτερα κατάλληλη τεχνική για την εύρεση του δυϊκού προβλήματος βελτιστοποίησης γνωστή ως *Lagrangian dual*. Ας υποθέσουμε ότι έχουμε ένα πρόβλημα βελτιστοποίησης της μορφής

$$min_{\bar{x}}\varphi(\bar{x}), \qquad (2.2)$$

έτσι ώστε

$$g_i(\bar{x}) \geq 0, \qquad (2.3)$$

με $i = 1, \dots l$ και για κάθε $\bar{x} \in R^n$. Εδώ υποθέτουμε ότι η φ είναι μια κυρτή αντικειμενική συνάρτηση και ότι οι περιορισμοί g_i είναι γραμμικοί. Όπως και πριν, οι γραμμικοί περιορισμοί είναι οι περιορισμοί που σχηματίζουν γραμμές, επίπεδα ή υπερεπίπεδα στον R^n. Για ένα πρόβλημα βελτιστοποίησης, η μορφή αυτή είναι ταυτόσημη με την αρχική μορφή του προβλήματος, αν πάρουμε τους περιορισμούς να είναι $g_i(\bar{x}) = h_i(\bar{x}) - c_i$. Συχνά αναφέρουμε αυτή τη μορφή ως το *αρχικό πρόβλημα βελτιστοποίησης*.
Τώρα, μπορούμε να κατασκευάσουμε ένα νέο πρόβλημα βελτιστοποίησης, το οποίο ονομάζεται πρόβλημα βελτιστοποίησης Lagrange, με βάση το αρχικό μας πρόβλημα:

$$max_{\bar{a}} min_{\bar{x}} L(\bar{a}, \bar{x}) = max_{\bar{a}} min_{\bar{x}} \left(\varphi(\bar{x}) - \sum_{i=1}^{l} a_i g_i(\bar{x})\right), \quad (2.4)$$

έτσι ώστε

$$a_i \geq 0, \quad (2.5)$$

με $i = 1, \dots l$ και για κάθε $\bar{x} \in R^n$. Η νέα αντικειμενική συνάρτηση $L(\bar{a}, \bar{x})$ ονομάζεται Λαγκρανζιανή και ενσωματώνει την αρχική αντικειμενική συνάρτηση φ μαζί με ένα γραμμικό συνδυασμό των περιορισμών g_i. Οι τιμές $a_1, \dots, a_l$ ονομάζονται *πολλαπλασιαστές Lagrange,* και όταν είναι βολικό τους γράφουμε με τη μορφή διανύσματος ως εξής:

$$\bar{a} = (a_1, \dots, a_l). \quad (2.6)$$

Έχουμε ακριβώς ένα πολλαπλασιαστή Lagrange a_i για κάθε περιορισμό g_i. Καλούμε το $\bar{x}$ αρχική μεταβλητή και το $\bar{a}$ δυική μεταβλητή.

Αυτό το νέο πρόβλημα βελτιστοποίησης που υπολογίζουμε έχει το ασυνήθιστο χαρακτηριστικό γνώρισμα των δύο εμφωλευμένων συναρτήσεων βελτιστοποίησης με αντίθετους στόχους βελτιστοποίησης. Ας υποθέσουμε ότι έχουμε ορίσει το διάνυσμα $\bar{x}$ να παίρνει την τιμή $\bar{x}^*$. Τότε το πρόβλημα βελτιστοποίησης ανάγεται σε πρόβλημα μεγιστοποίησης

$$max_{\bar{a}} L(\bar{a}, \bar{x}^*) = max_{\bar{a}} \left(\varphi(\bar{x}^*) - \sum_{i=1}^{l} a_i g_i(\bar{x}^*)\right) \quad (2.7)$$

Αντίθετα, αν ορίσουμε το $\bar{a}$ να πάρει την τιμή $\bar{a}^*$, παίρνουμε το πρόβλημα ελαχιστοποίησης

$$min_{\bar{x}} L(\bar{a}^*, \bar{x}) = min_{\bar{x}} \left(\varphi(\bar{x}) - \sum_{i=1}^{l} a_i{}^* g_i(\bar{x})\right) \quad (2.8)$$

Οι λύσεις στο πρόβλημα βελτιστοποίησης Lagrange είναι τα σημεία τα οποία μεγιστοποιούν τη συνάρτηση $L(\bar{a}, \bar{x})$ ως προς τη δυική μεταβλητή $\bar{a}$ και ταυτόχρονα ελαχιστοποιούν τη συνάρτησης ως προς την αρχική μεταβλητή $\bar{x}$. Αυτό σημαίνει ότι οι λύσεις είναι στάσιμα σημεία στο γράφημα της συνάρτησης $L(\bar{a}, \bar{x})$. Υποθέτουμε, όμως, ότι η αρχική αντικειμενική συνάρτηση $\varphi(x)$ είναι κυρτή και οι περιορισμοί $g_i(\bar{x})$ είναι

γραμμικοί, άρα θα έχουμε ένα μοναδικό στάσιμο σημείο. Επειδή το στάσιμο σημείο αντιπροσωπεύει μία λύση, το L θα ελαχιστοποιείται ως προς το $\bar{x}$ και η μερική παράγωγος της L ως προς $\bar{x}$ σε αυτό το σημείο θα πρέπει να είναι μηδέν, δηλαδή:

$$\frac{\partial L}{\partial \bar{x}} = \bar{0} \tag{2.9}$$

Έστω $\bar{x}^*$ η τιμή του $\bar{x}$ στο στάσιμο σημείο της L. Τότε η μερική παράγωγος της L ως προς $\bar{x}$ στο στάσιμο σημείο μας δίνει:

$$\frac{\partial L}{\partial \bar{x}}(\bar{a}, \bar{x}^*) = \bar{0} \tag{2.10}$$

Εδώ το σημείο $\bar{x}^*$ αντιπροσωπεύει το βέλτιστο της L ως προς το $\bar{x}$.

Μία από τις ενδιαφέρουσες, και για εμάς ιδιαίτερης σημασίας, ιδιότητες της βελτιστοποίησης Lagrange είναι ότι, υπό ορισμένες συνθήκες, μια λύση για το πρόβλημα Lagrange είναι επίσης μια λύση και για το αρχικό πρόβλημα βελτιστοποίησης. Για να το δούμε αυτό, ας υποθέσουμε ότι $\bar{a}^*$ και $\bar{x}^*$ είναι μια λύση για το πρόβλημα Lagrange έτσι ώστε:

$$max_{\bar{a}} min_{\bar{x}} L(\bar{a}, \bar{x}) = L(\bar{a}^*, \bar{x}^*) = \varphi(\bar{x}^*) - \sum_{i=1}^{l} \bar{a}^* g_i(\bar{x}^*). \tag{2.11}$$

Τότε η $\bar{x}^*$ είναι λύση για την αρχική αντικειμενική συνάρτηση αν και μόνο ισχύουν οι ακόλουθες προϋποθέσεις:

$$\frac{\partial L}{\partial \bar{x}}(\bar{a}^*, \bar{x}^*) = \bar{0}, \tag{2.12}$$

$$a_i{}^* g_i(\bar{x}^*) = 0, \tag{2.13}$$

$$g_i(\bar{x}^*) \geq 0, \tag{2.14}$$

$$\bar{a}^* \geq 0 \tag{2.15}$$

για κάθε $i = 1, \dots l$. Η πιο ενδιαφέρουσα από αυτές τις συνθήκες είναι ίσως η δεύτερη (2.13), που δείχνει ότι κάθε περιορισμός g_i υπολογιζόμενος στο $\bar{x}^*$ και πολλαπλασιαζόμενος με τον αντίστοιχο πολλαπλασιαστή Lagrange $\bar{a}^*$ δίνει την τιμή μηδέν.

Αυτό είναι αναγκαίο και μπορούμε να το δούμε από την προηγούμενη σχέση (2.11), όπου ο όρος $\sum_{i=1}^{l} \bar{a}^* g_i(\bar{x}^*)$ πρέπει να εξαφανιστεί έτσι ώστε $L(\bar{a}^*, \bar{x}^*) = \varphi(\bar{x}^*)$. Οι υπόλοιπες συνθήκες είναι απλές. Η εξίσωση (2.12) εξασφαλίζει ότι η τιμή $\bar{x}^*$ αποτελεί στάσιμο σημείο, και οι εξισώσεις (2.14) και (2.15) είναι οι αρχικοί περιορισμοί του αρχικού και του προβλήματος βελτιστοποίησης Lagrange, έτσι ώστε να διασφαλιστεί ότι τα σημεία $\bar{a}^*$ και $\bar{x}^*$ βρίσκονται στις αντίστοιχες εφικτές περιοχές.

Οι προϋποθέσεις αυτές αναφέρονται ως τις συνθήκες ***Karush - Kuhn - Tucker*** (KKTconditions). Επιπλέον, εξαιτίας της σημαντικότητας της, η σχέση (2.13) αναφέρεται ως συμπληρωματική συνθήκη ΚΚΤ [4].

Η επίλυση ενός προβλήματος βελτιστοποίησης με τη μέθοδο των πολλαπλασιαστών Lagrange όπου η αρχική αντικειμενική συνάρτηση είναι κυρτή μπορεί να απλοποιηθεί με την αξιοποίηση του γεγονότος ότι η βέλτιστη λύση $\bar{x}^*$ πρέπει να αποτελεί το μοναδικό στάσιμο σημείο. Επομένως, η επίλυση της εξίσωσης για $\bar{x}^*$ μας επιτρέπει να κατασκευάσουμε μια σχέση που θα μας επιτρέψει να αναδιατυπώσουμε το αρχικό πρόβλημα βελτιστοποίησης όσον αφορά τη δυική μεταβλητή μόνο, $L(\bar{a}, \bar{x}^*) = \varphi'(\bar{a})$, και μπορούμε να βρούμε το βέλτιστο ως προς την δυική μεταβλητή καθώς το πρόβλημα βελτιστοποίησης Lagrange

$$max_{\bar{a}}\varphi'(\bar{a}), \tag{2.16}$$

υπόκειται στους περιορισμούς

$$a_i \geq 0 \tag{2.17}$$

για κάθε $i = 1, \dots l$.

Καλούμε την συνάρτηση φ' Lagrangian dual (μερικές φορές ονομάζεται επίσης Wolfe dual). Αυτό σημαίνει ότι μπορούμε να λύσουμε το αρχικό πρόβλημα βελτιστοποίησης χρησιμοποιώντας το δυϊκό του Lagrangian,

$$max_{\bar{a}}\varphi'(\bar{a}) = \varphi'(\bar{a}^*) = L(\bar{a}^*, \bar{x}^*) = \varphi'(\bar{x}^*) \tag{2.18}$$

όπου τα $\bar{a}^*$ και $\bar{x}^*$ πρέπει να ικανοποιούν τις συνθήκες ΚΚΤ.

Πριν προχωρήσουμε, θα εξηγήσουμε τις παραπάνω έννοιες μέσα από ένα παράδειγμα όπου θα ξεκινήσουμε με ένα αρχικό πρόβλημα βελτιστοποίησης,στη συνέχεια θα το μετατρέψουμε σε λαγκραζιανό και θα το λύσουμε με το δυϊκό του Lagrange.

<u>Παράδειγμα</u>

Θεωρούμε το κυρτό πρόβλημα βελτιστοποίησης

$$min\ \varphi(x) = min\frac{1}{2}x^2 \tag{2.19}$$

με τους γραμμικούς περιορισμούς

$$g(x) = x - 2 \geq 0, \tag{2.20}$$

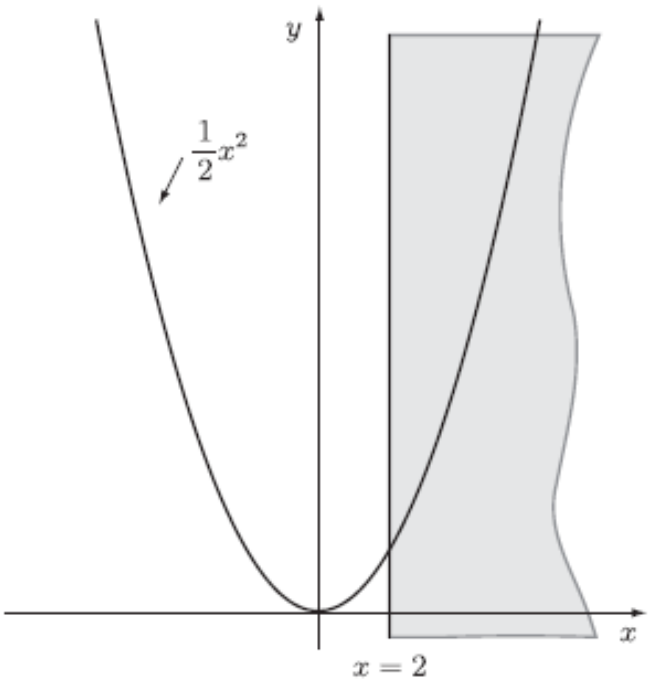
με $x \in R$.

Εδώ η κλασική μέθοδος εύρεσης του ελαχίστου με τη λήψη της παραγώγου της φ ως προς x και ο καθορισμός της τιμής της στο μηδέν, δηλαδή

$$\frac{d\varphi}{dx} = 0 \tag{2.21}$$

αποτυγχάνει, επειδή η τιμή $x = 0$ δεν αποτελεί μέρος της εφικτής περιοχής, δεδομένου ότι δεν πληροί τον περιορισμό $x - 2 \geq 0$. Ως εκ τούτου, πρέπει να βρούμε

μια τιμή του x που να βρίσκεται στην εφικτή περιοχή και να ελαχιστοποιεί την αντικειμενική συνάρτηση. Σε αυτό το απλό πρόβλημα βελτιστοποίησης είναι εύκολο να δούμε ότι η τιμή του x που ικανοποιεί τον περιορισμό και ελαχιστοποιεί την αντικειμενική συνάρτηση είναι $x = 2$, όπως απεικονίζεται στο Σχήμα 3. Εδώ, η γκρίζα περιοχή αντιπροσωπεύει όλα τα σημεία (x, y) που ικανοποιούν τον περιορισμό $x - 2 \geq 0$. Επομένως, μόνο το τμήμα της αντικειμενικής συνάρτηση φ που εμπίπτει σε αυτήν την περιοχή μπορεί να χρησιμοποιηθεί για τη βελτιστοποίηση.



Σχήμα 2.6: Μέρος –Κομμάτι της αντικειμενικής συνάρτησης $\boldsymbol{\varphi(x) = \frac{1}{2}x^2}$. Η γκρίζα περιοχή παριστάνει όλα τα σημεία $(\boldsymbol{x}, \boldsymbol{y})$ που ικανοποιούν τον περιορισμό $\boldsymbol{x - 2 \geq 0}$.

Για να λυθεί το πρόβλημα βελτιστοποίησης με τη χρήση του δυϊκού Lagrange πρέπει πρώτα να κατασκευάσουμε τη λαγκραζιανή εξίσωση :

$$L(a, x) = \frac{1}{2}x^2 - a(x - 2) \tag{2.22}$$

Όπως αναμένεται για μία κυρτή αντικειμενική συνάρτηση, έχουμε ένα μοναδικό στάσιμο σημείο στο γράφημα, όπως φαίνεται στο Σχήμα 4. Επιπλέον, γνωρίζουμε ότι αυτό το στάσιμο σημείο υπάρχει όταν η κλίση ως προς τη μεταβλητή x είναι μηδέν.

Πιο συγκεκριμένα,

$$\frac{\partial L}{\partial x}(a, x^*) = x^* - a = 0 \tag{2.23}$$

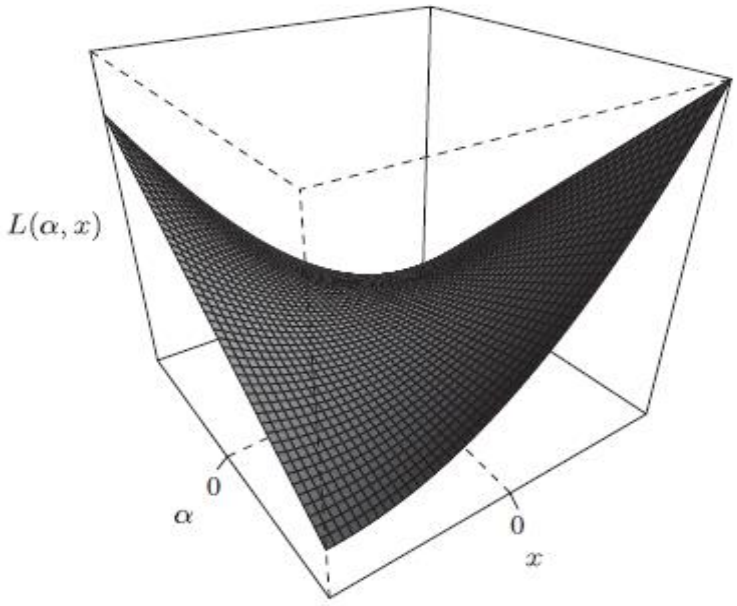
Εδώ το x^* αντιπροσωπεύει την τιμή που ελαχιστοποιεί την συνάρτηση Lagrangian ως προς το x στο στάσιμο σημείο.

Λύνοντας ως προς x^* έχουμε:

$$x^* = a \tag{2.24}$$

Επομένως, συνδέοντας τις (2.24) και (2.25) παίρνουμε:

$$L(a, x^*) = \frac{1}{2}a^2 - a^2 + 2a = 2a - \frac{1}{2}a^2 \tag{2.25}$$



Σχήμα 2.7: Γράφημα της συνάρτησης Lagrange $L(a, x) = \frac{1}{2}x^2 - a(x - 2)$.

Αυτή η συνάρτηση δεν εξαρτάται από τη μεταβλητή x, και μπορούμε επομένως να την ξαναγράψουμε με $\varphi'(a) = L(a, x^*)$,

$$max_a \varphi'(a) = max_a \left(2a - \frac{1}{2}a^2\right), \tag{2.26}$$

υπό τους περιορισμούς

$$\alpha \geq 0. \tag{2.27}$$

Τώρα, γνωρίζουμε ότι η $L(a, x)$ έχει ένα μοναδικό στάσιμο σημείο, και αυτό συνεπάγεται ότι η συνάρτηση $\varphi'(a) = L(a, x^*)$ έχει ένα μοναδικό μέγιστο. Αυτό το μοναδικό μέγιστο συμβαίνει όταν η κλίση είναι ίση με το μηδέν. Μπορούμε να υπολογίσουμε την τιμή της a^* σε αυτό το μοναδικό μέγιστο ως εξής:

$$\frac{d\varphi'}{da}(a^*) = 2 - a^* = 0. \quad (2.28)$$

Λύνοντας ως προς a παίρνουμε τη λύση για το δυϊκό πρόβλημα Lagrange $a = 2$.

Στη συνέχεια, σύμφωνα με την σχέση, έχουμε $x^* = a^* = 2$. Η λύση αυτή συμπίπτει με τη λύση που φαίνεται από το Σχήμα.

Μπορούμε λοιπόν, να δείξουμε ότι η λύση στο αρχικό πρόβλημα βελτιστοποίησης και στο δυϊκό πρόβλημα Lagrange συμπίπτουν, αποδεικνύοντας ότι ικανοποιούνται οι συμπληρωματικές συνθήκες ΚΚΤ:

$$a^* g(x^*) = a^*(x^* - 2) = 2(2 - 2) = 0.$$

2.3 Βασικές Έννοιες για τις Μηχανές Αναγνώρισης Προτύπων

Η βασική διαφορά που έχουν τα SVM δίκτυα από τους περισσότερους τύπους Νευρωνικών Δικτύων είναι ότι αποτελούν μια ακριβή υλοποίηση της μεθόδου ελαχιστοποίησης του δομικού ρίσκου. Αυτή η αρχή βασίζεται στο γεγονός ότι ο ρυθμός λάθους μιας μηχανής μάθησης σε δεδομένα ελέγχου (ρυθμός λάθους γενίκευσης) είναι φραγμένος από το άθροισμα του ρυθμού λάθους εκπαίδευσης και ενός όρου που εξαρτάται από την VC (Vapnik – Chervonenkis) διάσταση. Στην περίπτωση των γραμμικά διαχωρίσιμων προτύπων, το SVM δίκτυο μηδενίζει τον πρώτο όρο και ελαχιστοποιεί τον δεύτερο όρο. Για το λόγο αυτό ένα SVM δίκτυο μπορεί να εμφανίζει μεγάλη ικανότητα γενίκευσης.

Η μάθηση στα SVM δίκτυα είναι επιβλεπόμενη και γίνεται αφού δοθεί στο δίκτυο ολόκληρο το σύνολο εκμάθησης. Όπως και τα περισσότερα συστήματα με επιβλεπόμενη μάθηση, ένα SVM δίκτυο μπορεί να εκτελέσει εργασίες όπως ο διαχωρισμός προτύπων και η προσέγγιση συναρτήσεων. Η μάθηση γίνεται όπως και στα υπόλοιπα δίκτυα με την ελαχιστοποίηση μιας συνάρτησης κόστους. Ο τρόπος που γίνεται η μάθηση στο SVM δίκτυο αποτελεί ένα από τα μειονεκτήματα του. Αυτό, γιατί αφού τελειώσει η εκπαίδευση του δικτύου, αν υποτεθεί ότι βρίσκεται ακόμη ένα σύνολο εκμάθησης, δεν είναι δυνατόν να προστεθεί η νέα αυτή γνώση στο δίκτυο. Πρέπει να γίνει η εκπαίδευση από την αρχή, διαδικασία πολλές φορές χρονοβόρα. Αυτό το μειονέκτημα, όμως, αντισταθμίζεται από τα πλεονεκτήματα που έχουν τα SVM δίκτυα έναντι στους υπόλοιπους τύπους Νευρωνικών Δικτύων που θα φανούν στην συνέχεια.

2.3.1 Ένα φράγμα στην ικανότητα Γενίκευσης – Εμπειρικό Ρίσκο

Έστω ένα σύνολο από N παρατηρήσεις, το οποίο και ονομάζεται σύνολο εκμάθησης. Κάθε παρατήρηση αποτελείται από ένα ζεύγος: ένα διάνυσμα $x_i \in R^m$ και την απόκρισή του d_i. Θεωρείται ότι υπάρχει μια άγνωστη κατανομή πιθανότητας $P(\mathbf{x}, d)$ με βάση την οποία προκύπτουν τα δεδομένα αυτά. Έστω μια μηχανή που έχει ως σκοπό να μάθει την αντιστοίχιση $\mathbf{x}_i \mapsto d_i$. Η μηχανή έχει οριστεί να κάνει ένα σύνολο από απεικονίσεις $\mathbf{x} \mapsto f(\mathbf{x}, a)$ όπου a μια παράμετρος και είναι ντετερμινιστική, δηλαδή για κάποια είσοδο $\mathbf{x}$ και μια συγκεκριμένη επιλογή του a δίνει πάντα την ίδια έξοδο $f(\mathbf{x}, a)$. Η παράμετρος a επιλέγεται από την εκπαίδευση του δικτύου. Άρα για ένα Νευρωνικό Δίκτυο η παράμετρος αυτή μπορεί να θεωρηθεί ότι σχετίζεται με την επιλογή των βαρών του. Η αναμενόμενη τιμή του ρυθμού λαθών για μια εκπαιδευμένη μηχανή είναι:

$$R(a) = \int \frac{1}{2} \left| d - f(\mathbf{x}, a) \right| dP(\mathbf{x}, y) \tag{2.29}$$

Η ποσότητα $R(a)$ ονομάζεται αναμενόμενο ρίσκο. Το εμπειρικό ρίσκο $R_{emp}(a)$ είναι η ποσότητα:

$$R_{emp}(a) = \frac{1}{2N} \sum_{i=1}^{N} \left| d_i - f(\mathbf{x}_i, a) \right| \tag{2.30}$$

και όπως φαίνεται, είναι ο μέσος ρυθμός λαθών που προκύπτει στο σύνολο εκμάθησης και είναι μια σταθερή τιμή για δεδομένο σύνολο εκμάθησης $\{\mathbf{x}_i, d_i\}_{i=1}^{N}$ και δεδομένη επιλογή του a. Η ποσότητα $\frac{1}{2}\left| d_i - f(\mathbf{x}_i, a) \right|$ καλείται απώλεια και μπορεί να πάρει μόνο τις τιμές 0 και 1 όταν $d_i = \pm 1$. Έστω η τέτοιο ώστε $0 \le \eta \le 1$. Για απώλειες που παίρνουν τις τιμές αυτές με πιθανότητα $1-\eta$, ισχύει η ακόλουθη ανισότητα που θέτει ένα άνω φράγμα για το αναμενόμενο ρίσκο:

$$R(a) \le R_{emp}(a) + \sqrt{\frac{h(\log(2l/h) + 1 - \log(\eta/4)}{l}} \tag{2.31}$$

όπου h είναι ένας μη αρνητικός ακέραιος που είναι η VC διάσταση που αναφέρθηκε προηγουμένως. Αυτό που αντιπροσωπεύει το δεξί μέρος της σχέσης (2.31) είναι, δηλαδή, το άνω φράγμα για το ρίσκο με πιθανότητα h. Η σχέση (2.31) είναι χρήσιμη γιατί συνήθως το πρώτο της μέλος δεν μπορεί να υπολογιστεί. Το δεύτερο, αντίθετα, μπορεί να υπολογιστεί γνωρίζοντας την VC διάσταση.

2.3.2 VC διάσταση

Η VC διάσταση είναι μια ιδιότητα του συνόλου των συναρτήσεων $\{f(a)\}$. Όταν ένα σύνολο συναρτήσεων έχει VC διάσταση h, τότε υπάρχει τουλάχιστον ένα υποσύνολο του συνόλου εκμάθησης που περιέχει h σημεία που μπορούν να διαχωριστούν. Η VC διάσταση αυξάνεται όσο αυξάνεται το h. Έτσι, προτιμώνται μηχανές που έχουν όσο το δυνατόν μικρότερη VC διάσταση, θεωρώντας ότι έχουν ίδιο εμπειρικό ρίσκο. Παρά το ότι όπως θα περίμενε κανείς θα ήταν επιθυμητή η μεγάλη τιμή της, αυτό δεν γίνεται αφού όπως μπορεί να δειχθεί, ακόμα και άπειρη VC διάσταση δεν εγγυάται ότι θα μπορούν να διαχωρίζονται πολύ μικρά σύνολα διανυσμάτων.

2.4 Τα πλεονεκτήματα και τα μειονεκτήματα των Μηχανών Διανυσματικής Υποστήριξης

Όλες οι τεχνικές ταξινόμησης έχουν πλεονεκτήματα και μειονεκτήματα, τα οποία είναι περισσότερο ή λιγότερο σημαντικά ανάλογα με τα δεδομένα τα οποία αναλύονται. Οι Μηχανές Διανυσματικής Υποστήριξης μπορεί να είναι ένα χρήσιμο εργαλείο για ανάλυσης σε περιπτώσεις μη κανονικότητας των δεδομένων, για παράδειγμα όταν τα δεδομένα δεν είναι κανονικά κατανεμημένα ή ακολουθούν μια άγνωστη κατανομή [1]. Τα τέσσερα πιο σημαντικά χαρακτηριστικά της SVM είναι η **δυαδικότητα, οι πυρήνες, η κυρτότητα και η σποραδικότητα**.

Τα πλεονεκτήματα των SVM μπορούν να συνοψιστούν στα εξής:

- Βασίζονται σε πολύ απλές και ξεκάθαρες ιδέες από τη θεωρία στατιστικής μάθησης (Vapnik, 1995) και μπορούν να χρησιμοποιηθούν για την πρόβλεψη μελλοντικών δεδομένων.

- Έχουν ισχυρό θεωρητικό υπόβαθρο και έτσι μπορεί να ακολουθηθεί η διαδικασία βήμα προς βήμα.

- Η εκπαίδευση των Μηχανών Διανυσματικής Υποστήριξης είναι σχετικά εύκολη. Κλιμακώνεται σε σχετικά καλές υψηλές διαστάσεις των δεδομένων και η εξισορρόπηση μεταξύ της ταξινόμησης της πολυπλοκότητας και του λάθους μπορεί να ελεγχθεί ρητά. Το μόνο που απαιτείται είναι η καλή λειτουργία του πυρήνα.

- Με την εισαγωγή του πυρήνα, οι SVM αποκτούν ευελιξία στην επιλογή της μορφής του διαχωριστικού ορίου που διαχωρίζει τις κλάσεις, οι οποίες δεν χρειάζεται να είναι γραμμικά διαχωρίσιμες και ακόμη δεν χρειάζεται να έχουν την ίδια συνάρτηση για όλα τα δεδομένα, δεδομένου ότι η συνάρτηση του είναι μη παραμετρική και λειτουργεί τοπικά.

- Δεδομένου ότι ο πυρήνας περιέχει σιωπηρά ένα μη γραμμικό μετασχηματισμό, καμία υπόθεση σχετικά με τη λειτουργική μορφή του μετασχηματισμού, η οποία καθιστά τα δεδομένα γραμμικά διαχωρίσιμα, δεν είναι απαραίτητη. Ο μετασχηματισμός λαμβάνει χώρα εμμέσως σε μια ισχυρή θεωρητική βάση και η ανθρώπινη κρίση/τεχνογνωσία των ειδικών εκ των προτέρων δεν είναι απαραίτητη.

- Η μέθοδος δεν απαιτεί τη γνώση της στατιστικής κατανομής των δεδομένων.

- Η SVM εκπαιδεύεται από την επίλυση ενός περιορισμένου τετραγωνικού προβλήματος βελτιστοποίησης. Η SVM υλοποιεί τη χαρτογράφηση των συντελεστών παραγωγής σε ένα υψηλό τρισδιάστατο χώρο χρησιμοποιώντας ένα σύνολο μη γραμμικών βασικών συναρτήσεων.

- Οι SVM προσφέρουν μια μοναδική, βέλτιστη και ολική λύση, αφού η εκπαίδευσή τους γίνεται με την επίλυση ενός προβλήματος τετραγωνικού κυρτού προγραμματισμού. Αυτό είναι ένα πλεονέκτημα σε σύγκριση σε Νευρωνικά Δίκτυα, τα οποία έχουν πολλαπλές λύσεις που σχετίζονται με τα τοπικά ελάχιστα και για το λόγο αυτό μπορεί να μην είναι ισχυρά πάνω από διαφορετικά δείγματα.

- Η SVM μπορεί να χρησιμοποιηθεί για μια ποικιλία από αναπαραστάσεις, όπως τα νευρωνικά δίκτυα, splines, πολυωνυμικούς εκτιμητές, κ.λπ., αλλά υπάρχει μια μοναδική βέλτιστη λύση για κάθε επιλογή των SVM παραμέτρων. Αυτό είναι διαφορετικό σε άλλες μηχανές μάθησης, όπως τα τυποποιημένα Νευρωνικά Δίκτυα που χρησιμοποιούν την προς τα πίσω διάδοση. Με λίγα λόγια η ανάπτυξη των SVM είναι εντελώς διαφορετική από τους συνήθεις αλγόριθμους που χρησιμοποιούνται για τη μάθηση και η SVM παρέχει μια νέα άποψη μάθησης. Σε πληθώρα πραγματικών εφαρμογών έχουν επιδείξει ισάξια ή καλύτερη επίδοση συγκριτικά με άλλες ανταγωνιστικές μεθόδους.

- Οι SVM παρέχουν μια καλή γενίκευση και εκτός δείγματος, εάν οι παράμετροι C και R (στην περίπτωση ενός Gaussian πυρήνα) είναι κατάλληλα επιλεγμένοι.

- Ξεπερνούν σε σημαντικό βαθμό το πρόβλημα υπερπροσαρμογής στα δεδομένα (overfitting).

- Οι SVM μπορεί να είναι ισχυρή, ακόμη και όταν το δείγμα εκπαίδευσης έχει κάποια διαστρέβλωση.

- Οι SVM λειτουργούν ως μια από τις καλύτερες προσεγγίσεις για τη μοντελοποίηση δεδομένων. Συνδυάζουν τον γενικευμένο έλεγχο ως μια τεχνική για τον έλεγχο των διαστάσεων.

- Μπορούν να παράγουν περίπλοκα μη γραμμικά μοντέλα, που έχουν συγκεκριμένη συναρτησιακή διατύπωση.

- Το πλήθος των παραμέτρων που απαιτούν ρύθμιση στις μηχανές διανυσμάτων υποστήριξης είναι σημαντικά μικρότερο από το αντίστοιχο άλλων μεθοδολογιών.

Ένα κοινό μειονέκτημα των μη-παραμετρικών τεχνικών, όπως η SVM είναι η έλλειψη διαφάνειας των αποτελεσμάτων.Η ερμηνεία των αποτελεσμάτων είναι όμως εφικτή και μπορεί να βασίζεται σε γραφική απεικόνιση.

ΚΕΦΑΛΑΙΟ 3

ΜΗΧΑΝΕΣ ΔΙΑΝΥΣΜΑΤΙΚΗΣ ΥΠΟΣΤΗΡΙΞΗΣ

3.1 Εισαγωγή

Οι Μηχανές Διανυσματικής Υποστήριξης (SVMs) είναι μια νέα μέθοδος που αναπτύχθηκε για ταξινόμηση από τον Vapnik και την ομάδα του στο AT&T Bell Labs. Οι Μηχανές Διανυσματικής Υποστήριξης έχουν εφαρμογές σε προβλήματα ταξινόμησης καθώς αντικατέστησαν τα πολυστρωματικά δίκτυα. Η ταξινόμηση επιτυγχάνεται είτε με μία γραμμική είτε με μία μη γραμμική επιφάνεια διαχωρισμού (διαχωριστικό υπερεπίπεδο) στο χώρο εισόδου του συνόλου των δεδομένων. Στόχος των Μηχανών Διανυσματικής Υποστήριξης, όπως και των υπολοίπων ταξινομητών είναι να ελαχιστοποιηθεί το αναμενόμενο σφάλμα εξόδου του δείγματος. Βασικό χαρακτηριστικό των SVMs είναι η χαρτογράφηση του συνόλου δυαδικών δεδομένων εκπαίδευσης σε έναν χώρο υψηλότερων διαστάσεων και σε δεύτερη φάση ο διαχωρισμός των δύο κλάσεων με ένα υπερεπιπέδο μέγιστου περιθωρίου. Για να κατανοήσουμε την προσέγγιση των SVM, πρέπει να κατανοήσουμε δύο βασικές ένοοιες που συνδέονται άμεσα με αυτά και σε μεγάλο βαθμό τα χαρακτηρίζουν: τη δυαδικότητα και τους πυρήνες. Στο παρόν κεφάλαιο, αρχικά θα εξετάσουμε αυτές τις έννοιες για την απλή περίπτωση και στη συνέχεια θα δείξουμε πως μπορεί να επεκταθεί σε πιο σύνθετες εφαρμογές.

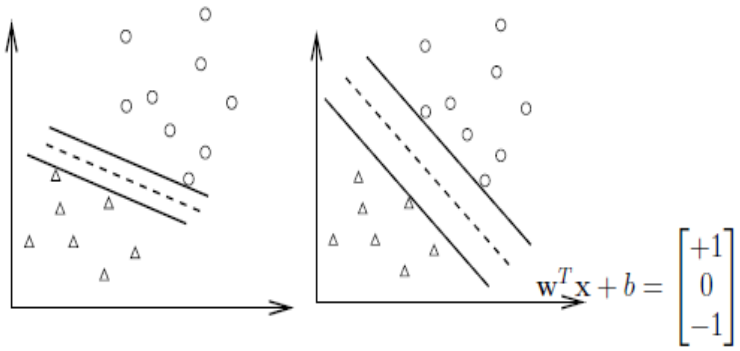
3.1.1 Γραμμικές Μηχανές Διανυσματικής Υποστήριξης

Η αρχική ιδέα των γραμμικών SVM ταξινομητών είναι να χρησιμοποιήσει ένα γραμμικό διαχωριστικό υπερεπίπεδο για να δημιουργήσει ένα ταξινομητή που θα διαχωρίζει τις δύο κλάσεις [5]. Παίρνοντας ως δεδομένο ένα σύνολο διανυσμάτων x_i, $i = 1, \dots l$ μήκους n, και ένα διάνυσμα y που ορίζεται ως εξής:

$$y_i = \begin{cases} 1, & \text{αν το } x_i \text{ ανήκει στη κλάση 1} \\ -1, & \text{αν το } x_i \text{ ανήκει στην κλάση 2} \end{cases}$$

ο SVM ταξινομητής προσπαθεί να βρει το διαχωριστικό υπερεπίπεδο με το μεγαλύτερο περιθώριο μεταξύ των δύο κλάσεων, που μετράται κατά μήκος μιας γραμμής κάθετης προς το υπερεπίπεδο. Για παράδειγμα, στο Σχήμα 3.1, οι δύο κλάσεις θα μπορούσαν να διαχωριστούν πλήρως από μια διακεκομμένη γραμμή $w^T x + b = 0$. Θα θέλαμε να

βρούμε τη γραμμή με το μεγαλύτερο περιθώριο. Με άλλα λόγια, αυτό που επιθυμούμε είναι η απόσταση μεταξύ των δύο κλάσεων δεδομένων εκπαίδευσης να είναι όσο το δυνατόν μεγαλύτερη. Αυτό σημαίνει ότι θα βρούμε ένα υπερεπίπεδο με παραμέτρους w και b έτσι ώστε η απόσταση μεταξύ των $w^T x + b = \pm 1$ να μεγιστοποιείται.

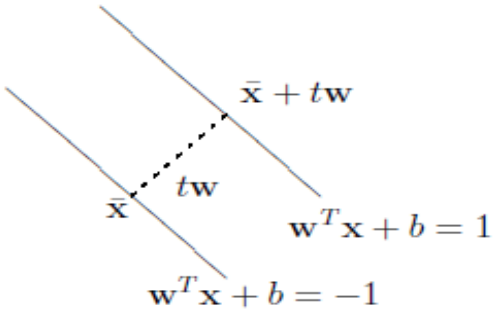


Σχήμα 3.1: Διαχωριστικό υπερεπίπεδο

Η απόσταση μεταξύ $w^T x + b = +1$ και $w^T x + b = -1$ μπορεί να υπολογιστεί με τον ακόλουθο τρόπο:

(i) Θεωρούμε ένα σημείο $\bar{x}$ στο $w^T x + b = -1$

(ii) Έστω w το κάθετο διάνυσμα στο υπερεπίπεδο $w^T x + b = -1$, το w και το υπερεπίπεδο είναι κάθετα μεταξύ τους.



(iii) Ξεκινώντας από το $\bar{x}$ και κινούμενοι κατά μήκος της κατεύθυνσης w, υποθέτουμε ότι το $\bar{x} + tw$ αγγίζει το επίπεδο $w^T x + b = 1$. Έτσι

$$\left.\begin{matrix} w^T(\bar{x} + tw) + b = 1 \\ w^T\bar{x} + b = -1 \end{matrix}\right\} \rightarrow tw^T w = 2$$

Οπότε η απόσταση (δηλαδή το μήκος του tw) είναι $\|tw\| = \frac{2\|w\|}{w^T w} = \frac{2}{\|w\|}$. Σημειώνουμε ότι $\|w\| = \sqrt{w_1^2 + \cdots + w_n^2}$. Η ελαχιστοποίηση του $\frac{2}{\|w\|}$ είναι ισοδύναμη με τη μεγιστοποίηση $\frac{w^T w}{2}$, επομένως έχουμε το ακόλουθο πρόβλημα:

$$min_{w,b}\ \frac{1}{2} w^T w$$

s.t. $y_i(w^T x_i + b) \geq 1, i = 1, \ldots l$ (3.1)

Ο περιορισμός $y_i(w^T x_i + b) \geq 1$ σημαίνει ότι

$$w^T x_i + b \geq 1 \text{ αν } y_i = 1, \quad w^T x_i + b \leq -1 \text{ αν } y_i = -1$$

Επομένως, τα δεδομένα της κλάσης 1 πρέπει να είναι στην δεξιά πλευρά του $w^T x + b = 0$ ενώ τα δεδομένα της άλλης κλάσης 2 πρέπει να βρίσκονται στην αριστερή πλευρά. Σημειώνουμε ότι ο λόγος της μεγιστοποίησης της απόστασης μεταξύ $w^T x + b = \pm 1$ βασίζεται στον Vapnik (Vapnik's Structural Risk Minimization -Vapnik, 1998).

Το ακόλουθο παράδειγμα δίνει μια απλή απεικόνιση του διαχωριστικού υπερεπιπέδου μέγιστου περιθωρίου.

<u>Παράδειγμα 3.1.1</u>

Δοθέντων δύο δεδομένων εκπαίδευσης στον R^1 όπως απεικονίζονται στο ακόλουθο σχήμα:

Ποίο είναι το διαχωριστικό υπερεπίπεδο;

Τώρα τα δύο δεδομένα είναι $x_1 = 1, x_2 = 0$ με $y = [+1, -1]^T$. Επιπλέον, $w \in R^1$, έτσι η (3.1) γίνεται

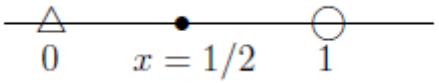
$$min_{w,b}\ \frac{1}{2} w^2$$

s.t. $w \cdot 1 + b \geq 1,$ (3.2)

$-1(w \cdot 0 + b) \geq 1.$ (3.3)

Από την (3.3), $-b \geq 1$ και θέτοντας αυτό στη (2.2) $w \geq 2$. Με άλλα λόγια, για κάθε (w, b) που ικανοποιεί τις (3.2), (3.3) $w \geq 2$. Καθώς ελαχιστοποιούμε το $\frac{1}{2}w^2$,η μικρότερη

δυνατότητα είναι $w = 2$. Έτσι, η (w, β) = (2, -1) είναι η βέλτιστη λύση. Το διαχωριστικό υπερεπίπεδο είναι $2x - 1 = 0$, στη μέση των δύο δεδομένων εκπαίδευσης:

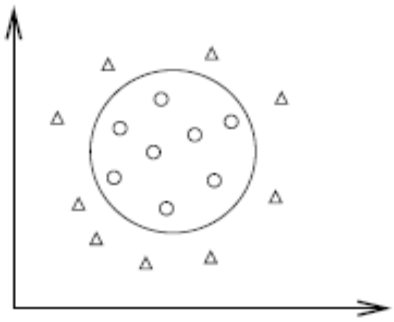


3.1.2 Χαρτογράφηση δεδομένων σε χώρους υψηλότερων διαστάσεων

Ωστόσο, πρακτικά προβλήματα μπορεί να μην είναι γραμμικά διαχωρίσιμα όπως περιγράψαμε στην προηγούμενη παράγραφο. Στο Σχήμα 2.2 διακρίνουμε μία τέτοια περίπτωση όπου δεν είναι εφικτός ο διαχωρισμός με ένα υπερεπίπεδο. Δηλαδή, δεν υπάρχει (w,b) ώστε να ικανοποιείται ο περιορισμούς της (2.1). Τότε λέμε ότι η (3.1) είναι «ανέφικτη». Έτσι οι Cortes and Vapnik (1995) εισήγαγαν τις χαλαρές μεταβλητές $\xi_i,\ i = 1, \dots, l$ στο πρόβλημα ελαχιστοποίησης:

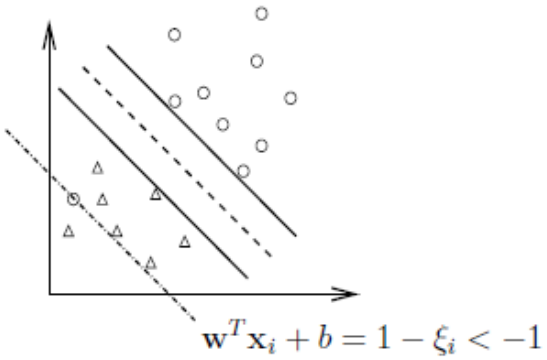
$$min_{w,b,\xi} \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i$$

s.t. $y_i(w^T x_i + b) \geq 1 - \xi_i,\ \xi_i \geq 0,\ i = 1, \dots l$ (3.4)



Σχήμα 3.2: Ένα παράδειγμα όπου τα δεδομένα δεν είναι γραμμικώς διαχωρίσιμα

Δηλαδή, με τους περιορισμούς (3.4) «επιτρέπουμε» στα δεδομένα εκπαίδευσης να μην είναι στη σωστή πλευρά του διαχωριστικού υπερεπίπεδου $w^T x + b = 0$. Αυτό συμβαίνει όταν $\xi_i > 1$ και ένα παράδειγμα είναι στο παρακάτω σχήμα:



Έχουμε ότι $\xi \geq 0$ καθώς αν $\xi < 0$, $y_i(w^T x_i + b) \geq 1 - \xi_i \geq 1$ άρα τα δεδομένα εκπαίδευσης είναι ήδη στη σωστή πλευρά. Το νέο πρόβλημα είναι πάντα εφικτό αφού για κάθε (w,b)

$$\xi_i \equiv \max\big(0,1 - y_i(w^T x + b)\big), i = 1, \dots, l,$$

οδηγούμαστε στο (w,b,ξ) που είναι μία εφικτή λύση.

Χρησιμοποιώντας αυτή τη ρύθμιση, μπορεί να ανησυχούμε ότι για γραμμικά διαχωρίσιμα δεδομένα, κάποια $\xi_i > 1$ και ως εκ τούτου τα αντίστοιχα δεδομένα είναι λάθος ταξινομημένα. Στην περίπτωση που τα περισσότερα δεδομένα εκτός από κάποια θορυβώδη μπορούν να διαχωριστούν από μια γραμμική συνάρτηση, θα θέλαμε το $w * x + b = 0$ να ταξινομεί σωστά την πλειοψηφία των σημείων. Έτσι, στην αντικειμενική συνάρτηση προσθέτουμε έναν όρος ποινής $C \sum_{i=1}^{l} \xi_i$ όπου $C > 0$ είναι μία παράμετρος ποινής. Για να έχουμε την αντικειμενική αξία όσο το δυνατόν μικρότερη, τα περισσότερα ξ_i πρέπει να είναι μηδέν, έτσι ώστε ο περιορισμός να παίρνει την αρχική του μορφή. Θεωρητικά μπορούμε να αποδείξουμε ότι αν τα δεδομένα είναι γραμμικά διαχωρίσιμα και το C είναι μεγαλύτερο από ένα συγκεκριμένο αριθμό, το πρόβλημα (3.4) ανάγεται στο (3.1) και όλα τα ξ_i είναι μηδέν [5].

Επισημάνουμε ότι η σταθερά C, που καλείται κόστος (cost) μας επιτρέπει να ελέγχουμε το ‘’trade-off’’ μεταξύ του μεγέθους του περιθωρίου και του σφάλματος. Τονίζουμε ότι το C, είναι θετικό και δε μπορεί να πάρει την τιμή μηδέν (αν θέσουμε $C = 0$ θα σημαίνει ότι αγνοούμε τις χαλαρές μεταβλητές).

Συνοψίζοντας:

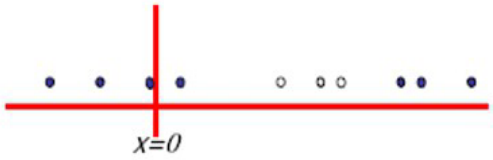
μεγάλο C ~ μικρό περιθώριο ~ μικρό λάθος

μικρό C ~ μεγάλο περιθώριο ~ μεγάλο λάθος

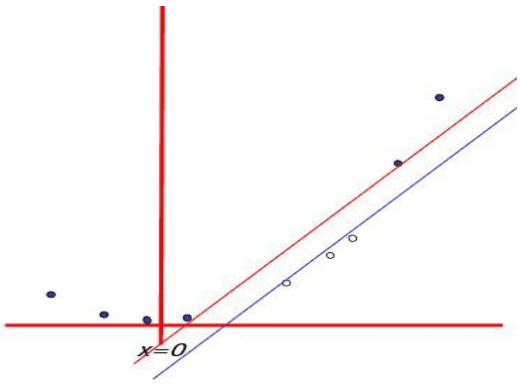
Δυστυχώς, όμως, μια τέτοια ρύθμιση δεν είναι αρκετή για πρακτική χρήση. Εάν τα δεδομένα κατανέμονται με ένα μη γραμμικό τρόπο, χρησιμοποιώντας μόνο μια γραμμική συνάρτηση πολλά δεδομένα εκπαίδευσης είναι στη λάθος πλευρά του

υπερεπιπέδου. Έτσι εμφανίζεται λάθος τοποθέτηση και η συνάρτηση απόφασης δεν αποδίδει καλά.

Για να τοποθετήσουμε τα δεδομένα κατάρτισης καλύτερα, μπορούμε να σκεφτούμε τη χρήση μιας μη γραμμικής καμπύλης, όπως στο σχήμα 3.3. Το πρόβλημα είναι ότι είναι πολύ δύσκολο να μοντελοποιήσουμε μη γραμμικές καμπύλες. Είμαστε εξοικειωμένοι με ελλειπτικές, υπερβολικές ή παραβολικές καμπύλες, οι οποίες απέχουν πολύ στην πράξη. Αντί να χρησιμοποιήσουμε πιο εξελιγμένες καμπύλες, μια άλλη προσέγγιση είναι να χαρτογραφήσουμε τα δεδομένα σε ένα χώρο υψηλότερων διαστάσεων, όπως φαίνεται στο επόμενο σχήμα:



Σχήμα 3.3: Δύο κλάσεις που δεν είναι γραμμικώς διαχωρίσιμες στον R^1



Σχήμα 3.4: Οι εικόνες των δύο κλάσεων στο χώρο R^2 είναι γραμμικώς διαχωρίσιμες.

Έτσι οι μη γραμμικές μηχανές διανυσματικής υποστήριξης μετατρέπουν τον αρχικό χώρο εισόδου σε έναν υψηλότερων διαστάσεων χώρο των χαρακτηριστικών. Πιο συγκεκριμένα, το δεδομένο εκπαίδευσης x χαρτογραφείται σε ένα διάνυσμα σε ένα χώρο υψηλότερων διαστάσεων:

$$\varphi(x) = [\varphi(x_1), \varphi(x_2), \dots]$$

Σε αυτόν τον υψηλότερων διαστάσεων χώρο, είναι πιο πιθανό ότι τα δεδομένα μπορούν να διαχωριστούν γραμμικά.

Ένα ακραίο παράδειγμα είναι να χαρτογραφήσουμε ένα παράδειγμα δεδομένων $x \in R^1$ σε ένα άπειρων διαστάσεων χώρο:

$$\varphi(x) = \left[1, \frac{x}{1!}, \frac{x^2}{2!}, \frac{x^3}{3!}, \dots\right]^T$$

Στη συνέχεια, προσπαθούμε να βρούμε ένα γραμμικά διαχωρίσιμο υπερεπίπεδο σε ένα χώρο υψηλότερων διαστάσεων, έτσι η (2.4) γίνεται:

$$min_{w,b,\xi} \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i$$

s.b. $y_i(w^T \varphi(x_i) + b) \geq 1 - \xi_i,\ \ \xi_i \geq 0,\ \ i = 1, \dots l$ (3.5)

3.1.3 Το δυϊκό πρόβλημα

Το πρόβλημα που παραμένει είναι ο τρόπος που θα λύσουμε αποτελεσματικά το πρόβλημα (3.5). Ειδικά όταν τα δεδομένα χαρτογραφούνται σε έναν χώρο υψηλότερων διαστάσεων, ο αριθμός των μεταβλητών (w,b) γίνεται πολύ μεγάλος ή ακόμα και άπειρος. Για να μπορέσουμε να χειριστούμε αυτή τη δυσκολία προχωράμε στην επίλυση του δυϊκού προβλήματος της εξίσωσης (3.5):

$$min_{\boldsymbol{a}} \frac{1}{2} \sum_{i,j=1}^{l} a_i a_j y_i y_j \varphi(x_i)^T \varphi(x_j) - \sum_{i=1}^{l} a_i$$

s.t. $0 \leq a_i \leq C \quad i = 1, \dots l,$ (3.6)

$$\sum_{i=1}^{l} y_i \boldsymbol{a_i} = 0$$

Αυτό το νέο πρόβλημα βέβαια έχει κάποια σχέση με το αρχικό πρόβλημα (3.5), και ελπίζουμε ότι μπορεί να επιλυθεί ευκολότερα. Ορισμένες φορές είναι προτιμότερο να γράφουμε την εξίσωση (3.6) σε μορφή πινάκων για λόγους ευκολίας επίλυσης του προβλήματος:

$$min_a \frac{1}{2} a^T Q a - e^T a$$

s.t. $0 \leq a_i \leq C \quad i = 1, \dots l\ ,$ (3.7)

$$y^T a = 0$$

Στην (3.7), e είναι το μοναδιαίο διάνυσμα, C είναι το άνω όριο, Q είναι ένας *l*l* θετικά ημιορισμένος πίνακας, $Q_{I,J} \equiv y_i y_j K(x_i, x_j)$ και $K(x_i, x_j) \equiv \varphi(x_i)^T \varphi(x_j)$ είναι ο πυρήνας, στον οποίο θα αναφερθούμε στην παράγραφο 3.1.4.

Αν (3.7) ονομάσουμε το δυϊκό του προβλήματος (2.5), τότε το πρόβλημα (3.5) θα αναφέρεται ως το αρχικό πρόβλημα. Ας υποθέσουμε $(\overline{w}, \overline{b}, \overline{\xi})$ και $\overline{\alpha}$ είναι οι βέλτιστες

λύσεις του αρχικού και του δυϊκού προβλήματος, αντίστοιχα, τότε ισχύουν οι ακόλουθες δύο ιδιότητες:

$$\bar{w} = \sum_{i=1}^{l} \bar{a}_i y_i \varphi(x_i) \tag{3.8}$$

$$\frac{1}{2}\bar{w}^T\bar{w} + C\sum_{i=1}^{l}\bar{\xi}_i = e^T\bar{a} - \frac{1}{2}\bar{a}^T Q\bar{a} \tag{3.9}$$

Με άλλα λόγια, αν $\bar{a}$ είναι η λύση του δυϊκού προβλήματος, η βέλτιστη λύση $\bar{w}$ του αρχικού προβλήματος λαμβάνεται εύκολα από την εξίσωση (3.8). Ας υποθέσουμε ότι το βέλτιστο $\bar{b}$ είναι επίσης εύκολο να βρεθεί, τότε η συνάρτηση απόφασης μπορεί επίσης εύκολα να προσδιοριστεί.

Έτσι, το κρίσιμο σημείο είναι αν το δυϊκό πρόβλημα είναι πιο εύκολο να λυθεί από το αρχικό. Ο αριθμός των μεταβλητών στο δυϊκό, ο οποίος ισούται με το μέγεθος του συνόλου εκπαίδευσης l, είναι ένας σταθερός αριθμός. Αντίθετα, ο αριθμός των μεταβλητών στο αρχικό προβλήματος ποικίλλει ανάλογα με το πώς τα δεδομένα χαρτογραφούνται σε έναν χώρο υψηλότερων διαστάσεων. Ως εκ τούτου, η μετάβαση από το αρχικό πρόβλημα στο δυϊκό συνεπάγεται τη λύση ενός προβλήματος βελτιστοποίησης πεπερασμένων διαστάσεων αντί ενός, ενδεχομένως, απειροδιάστατου προβλήματος.

Εμείς καταδεικνύουμε αυτή την σχέση αρχικού-δυϊκού προβλήματος με τη χρήση των δεδομένων του παραδείγματος 3.1.1, χωρίς να γίνει η χαρτογράφηση τους σε έναν χώρο υψηλότερων διαστάσεων. Δεδομένου ότι το πρόβλημα είναι γραμμικά διαχωρίσιμο, είναι καλό να εξετάσουμε την (3.1) χωρίς τις χαλαρές μεταβλητές $\xi_i, i = 1, \dots, l$.

Τότε το δυϊκό πρόβλημα είναι το ακόλουθο:

$$min_{\boldsymbol{a}} \ \frac{1}{2}\sum_{i,j=1}^{l} a_i a_j y_i y_j {x_i}^T x_j - \sum_{i=1}^{l} a_i$$

s.t.

$$0 \le a_i \quad i = 1, \dots l, \ \sum_{i=1}^{l} y_i \boldsymbol{a_i} = 0$$

Χρησιμοποιώντας τα δεδομένα του παραδείγματος 3.1.1, η αντικειμενική συνάρτηση είναι:

$$\frac{1}{2}{a_1}^2 - (a_1 + a_2) = \frac{1}{2}[a_1\ a_2]\begin{bmatrix}1 & 0\\ 0 & 0\end{bmatrix}\begin{bmatrix}a_1\\ a_2\end{bmatrix} - [1\ 1]\begin{bmatrix}a_1\\ a_2\end{bmatrix}$$

Οι περιορισμοί είναι

$$a_1 - a_2 = 0, 0 \le a_1, 0 \le a_2$$

Αντικαθιστώντας $a_1 = a_2$ στην αντικειμενική συνάρτηση

$$\frac{1}{2}{a_1}^2 - 2a_1$$

Παίρνει τη μικρότερη τιμή για $a_1 = 2$. Καθώς η $[2\ 2]^T$ ικανοποιεί τους περιορισμούς $0 \leq a_1,\ 0 \leq a_2$, αυτή είναι και η βέλτιστη λύση. Χρησιμοποιώντας την αρχική-δυική σχέση (3.8)

$$w = y_1 a_1 x_1 + y_2 a_2 x_2 = 1 \cdot 2 \cdot 1 + (-1) \cdot 2 \cdot 0 = 2$$

Την ίδια που βρήκαμε λύνοντας το αρχικό πρόβλημα (!).

Ο υπολογισμός του b είναι εύκολος, αλλά δε θα γίνει στην παράγραφο αυτή. Το υπόλοιπο θέμα της χρήσης του δυϊκού προβλήματος αφορά στο εσωτερικό γινόμενο $\varphi(x_i)^T \varphi(x_j)$. Αν $\varphi(x)$ είναι ένα άπειρο διάνυσμα, δεν υπάρχει τρόπος να το γράψουμε και στη συνέχεια να υπολογίσουμε το εσωτερικό γινόμενο. Έτσι, ακόμη και αν το δυϊκό πρόβλημα έχει το πλεονέκτημα του πεπερασμένου αριθμού των μεταβλητών, δε μπορούμε να γράψουμε το πρόβλημα πριν από την επίλυσή του. Αυτό επιλύεται χρησιμοποιώντας ειδικές συναρτήσεις χαρτογράφησης φ έτσι ώστε το $\varphi(x_i)^T \varphi(x_j)$ να μπορεί να υπολογιστεί αποτελεσματικά. Λεπτομέρειες για την λειτουργία της χαρτογράφηση δίνονται στην επόμενη παράγραφο.

3.1.4 Πυρήνες και συναρτήσεις απόφασης

Ας ξεκινήσουμε με ένα παράδειγμα.

Παράδειγμα 3.4.1

Έστω $\varphi(x)\colon R^2 \to R^3$ με $\varphi(\bar{x}) = ({x_1}^2, {x_2}^2, \sqrt{2}x_1 x_2)$. Το εσωτερικό γινόμενο $\varphi(\bar{x})\varphi(\bar{y})$ με $\bar{x}, \bar{y} \in R$ μπορεί να υπολογιστεί ως εξής:

$$\varphi(\bar{x}) \cdot \varphi(\bar{y}) = \left({x_1}^2, {x_2}^2, \sqrt{2}x_1 x_2\right) \cdot \left({y_1}^2, {y_2}^2, \sqrt{2}y_1 y_2\right) = {x_1}^2 {y_1}^2 + {x_2}^2 {y_2}^2 + 2x_1 x_2 y_1 y_2$$
$$= (x_1 y_1 + x_2 y_2)(x_1 y_1 + x_2 y_2) = (\bar{x} \cdot \bar{y})(\bar{x} \cdot \bar{y}) = (\bar{x} \cdot \bar{y})^2$$

Ας ρίξουμε μία πιο προσεκτική ματιά σε αυτό το μετασχηματισμό.

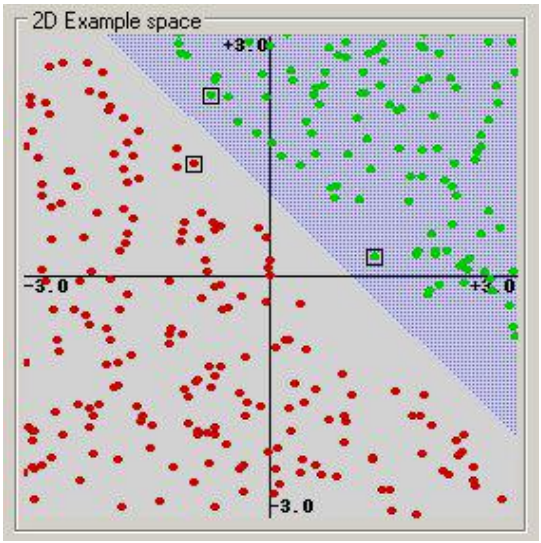
Με μία κατάλληλη χαρτογράφηση

$$\varphi\colon R^n \to R^m \text{ με } m \geq n$$

οι συναρτήσεις της μορφής: $K(\bar{x}, \bar{y}) = \varphi(\bar{x}) \cdot \varphi(\bar{y})$ με $\bar{x}, \bar{y} \in R^n$

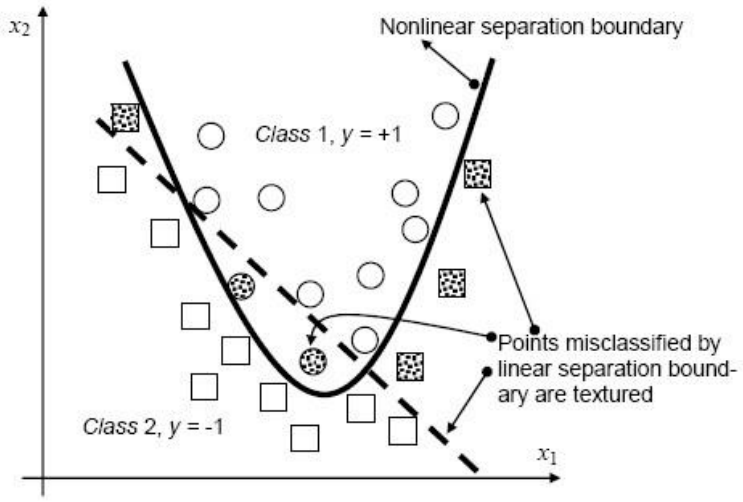
καλούνται *πυρήνες* ή *συναρτήσεις πυρήνων.* Ορισμένοι δημοφιλείς πυρήνες είναι:

1. **Ο γραμμικός** (Linear): $K(\vec{x}, \vec{x}) = \vec{x} \cdot \vec{x} + \gamma$,
   όπου η παράμετρος γ καθορίζεται από τον χρήστη. Αυτός ο τύπος πυρήνα έχει εφαρμογή μόνο σε γραμμικώς διαχωρίσιμα προβλήματα. Παράδειγμα ενός τέτοιου πυρήνα φαίνεται στο παρακάτω Σχήμα.



Σχήμα 3.5:. Παράδειγμα γραμμικού πυρήνα. Τα σημεία που περικλείονται από τετραγωνικό πλαίσιο αποτελούν διανύσματα υποστήριξης.

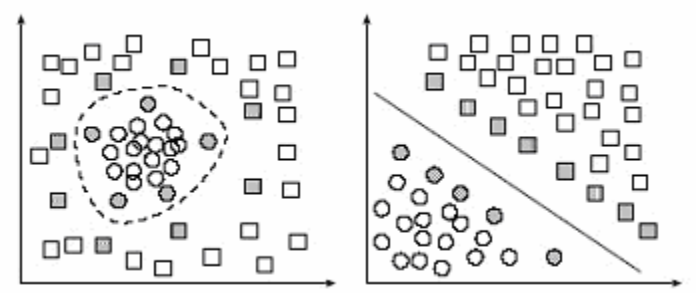
2. **Ο πολυωνυμικός** (Polynomial): $K(\vec{x}, \vec{x}) = (\vec{x}, \vec{x}' + 1)^p$,
   όπου η παράμετρος p καθορίζεται από τον χρήστη. Στο Σχήμα 3.6 παρουσιάζεται η περίπτωση εφαρμογής SVM γραμμικού πυρήνα και πολυωνυμικού πυρήνα σε μη-γραμμικό πρόβλημα. Είναι σαφής η μεγαλύτερη διαχωριστική ικανότητα που παρουσιάζει το SVM με πολυωνυμικό πυρήνα.



***Σχήμα*.3.6**: Παράδειγμα διαχωρισμού μη-γραμμικού προβλήματος με χρήση SVMs γραμμικού (διακεκομμένη γραμμή) και πολυωνυμικού πυρήνα (συνεχής γραμμή). Είναι σαφής η μεγαλύτερη διαχωριστική ικανότητα του SVM πολυωνυμικού πυρήνα.

3. **Ο Ακτινωτός** (Radial Basis Function (RBF) ή Gaussian kernel):
   $K(\vec{x},\vec{x}) = e^{-\frac{\|\vec{x}-\vec{x}''\|}{2\sigma^2}}$,

   όπου η παράμετρος σ καθορίζεται από τον χρήστη. Παράδειγμα ενός τέτοιου πυρήνα φαίνεται στο Σχήμα 3.7



***Σχήμα* 3.7:** Στην αριστερή υπο-εικόνα παρουσιάζεται ο τρόπος που θα φαινόταν ο διαχωρισμός με radial basis πυρήνα στον αρχικό χώρο (χώρο εισόδου), ενώ στη δεξιά έχει προηγηθεί ο μετασχηματισμός σε χώρο υψηλότερης διάστασης με τη βοήθεια radial basis πυρήνα όπου ο διαχωρισμός είναι πια γραμμικός. Τα σημεία που χρωματίζονται σε κλίμακα του γκρι αποτελούν διανύσματα υποστήριξης

4. **Ο Σιγμοειδής** (Sigmoid): $K(\vec{x},\vec{x}) = \tanh(\kappa \cdot \vec{x} \cdot \vec{x}' + \theta)$,

   όπου κ και θ είναι παράμετροι των πυρήνων.

Αφού λυθεί το πρόβλημα (3.7) και βρούμε το a, το διάνυσμα για το οποίο $\alpha_\iota > 0$ ονομάζεται *διάνυσμα υποστήριξης.* Τότε η συνάρτηση απόφασης μορφοποιείται ως εξής:

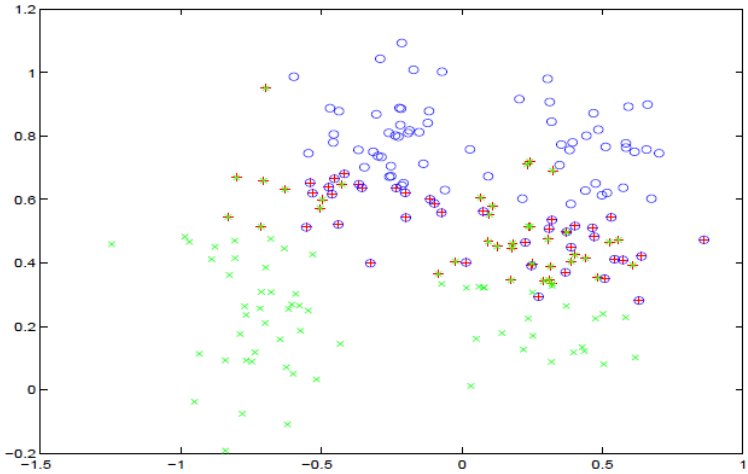
$$f(x) = sign(w^T\varphi(x) + b) = sign(\sum_{i=1}^{l} y_i a_i \varphi(x_i)^T \varphi(x) + b) \quad (3.10)$$

Με άλλα λόγια, αν θεωρήσουμε ένα τυχαίο διάνυσμα x, εάν παίρνοντας την εξίσωση $\sum_{i=1}^{l} y_i a_i \varphi(x_i)^T \varphi(\boldsymbol{x}) + b > 0$, δηλαδή προκύπτει αποτέλεσματα μεγαλύτερο του μηδενός τότε μπορούμε να το κατατάξουμε στην κλάση 1. Σε αντίθετη περίπτωση, πιστεύουμε ότι είναι στη δεύτερη κλάση. Είναι σημαντικό να επισημάνουμε ότι μόνο τα διανύσματα υποστήριξης θα επηρεάσουν τα αποτελέσματα στο στάδιο πρόβλεψης. Σε γενικές γραμμές, ο αριθμός των διανυσμάτων υποστήριξης δεν είναι μεγάλος. Ως εκ τούτου, μπορούμε να πούμε ότι οι μηχανές διανυσματικής υποστήριξης χρησιμοποιούνται για την εύρεση σημαντικών δεδομένων (διανύσματα υποστήριξης) από δεδομένα εκπαίδευσης.

Χρησιμοποιούμε το Σχήμα 3.3 ως παράδειγμα όπου οι δύο κλάσεις των δεδομένων εκπαίδευσης δεν είναι γραμμικά διαχωρίσιμες. Χρησιμοποιώντας τον πυρήνα RBF, παίρνουμε ένα υπερεπίπεδο $w^T\varphi(x) + b = 0$ το οποίο στον αρχικό χώρο, είναι όντως μια μη γραμμική καμπύλη.

$$\sum_{i=1}^{l} y_i a_i \varphi(x_i)^T \varphi(\boldsymbol{x}) + b = 0 \qquad (3.11)$$

Στο παρακάτω σχήμα, όλα τα σημεία με κόκκινο χρώμα είναι διανύσματα υποστήριξης και επιλέγονται και από τις δύο κατηγορίες των δεδομένων εκπαίδευσης. Σαφώς τα διανύσματα υποστήριξης που είναι κοντά στην μη γραμμική καμπύλη (3.11) είναι τα πιο σημαντικά σημεία.



Σχήμα 3.8: Τα διανύσματα υποστήριξης (που είναι σημειωμένα με κόκκινο) είναι τα σημαντικά δεδομένα από τα δεδομένα εκπαίδευσης.

3.2 Το θεώρημα του Mercer

Το θεώρημα του Mercer (Mercer, 1908, Courant and Hilbert, 1970) μπορεί να διατυπωθεί ως εξής:

Έστω $K(\mathbf{x},\mathbf{x}_i)$ ένας συμμετρικός ($K(\mathbf{x},\mathbf{x}_i)\ K(\mathbf{x}_i,\mathbf{x})$) πυρήνας, ο οποίος είναι ορισμένος στο $\mathbf{a} \le \mathbf{x} \le \mathbf{b}$ και αντίστοιχα και για το x_i. Ο πυρήνας μπορεί να αναπτυχθεί στην σειρά:

$$K(\mathbf{x},\mathbf{x}_i) = \sum_{i=1}^{\infty} \lambda_i \phi_i(\mathbf{x}) \phi_i(\mathbf{x}_i) \qquad (3.12)$$

με θετικούς συντελεστές λ_i για κάθε i. Για να συγκλίνει αυτή η σειρά απόλυτα και ομοιόμορφα, πρέπει η συνθήκη:

$$\int_{\mathbf{b}}^{\mathbf{a}} \int_{\mathbf{b}}^{\mathbf{a}} K(\mathbf{x},\mathbf{x}_i)\psi(\mathbf{x})\psi(\mathbf{x}_i)d\mathbf{x}d\mathbf{x}_i \ge 0$$

να ικανοποιείται για όλες τις $\psi(\bullet)$ για τις οποίες $\int_{\mathbf{b}}^{\mathbf{a}} \psi^2(\mathbf{x})d\mathbf{x} < \infty$.

Το θεώρημα του Mercer προσδιορίζει αν ένας πυρήνας είναι εσωτερικό γινόμενο σε κάποιον χώρο και συνεπώς μπορεί να χρησιμοποιηθεί σε ένα SVM δίκτυο, αλλά δεν λέει τίποτα για τον χώρο αυτό, ούτε για τον τρόπο κατασκευής των ϕ_i.

3.2.1 Η VC διάσταση των SVM δικτύων

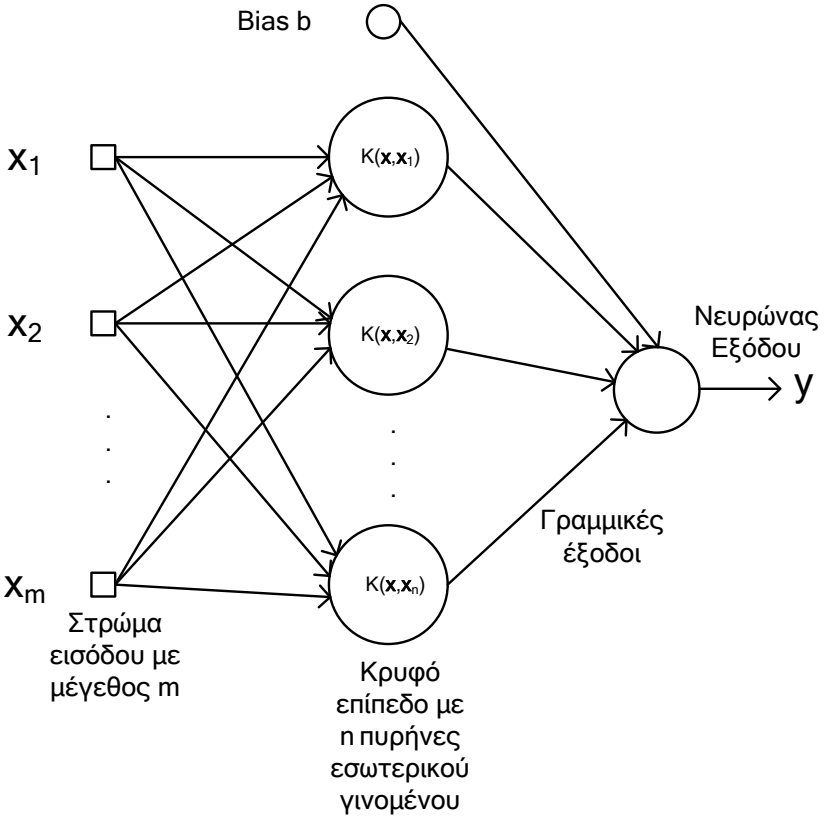
Αν και η VC διάσταση των SVM δικτύων είναι πολύ μεγάλη, εώς και άπειρη, όπως θα δειχθεί παρακάτω, και θα περίμενε κανείς μικρή ικανότητα για γενίκευση, στην πράξη αυτό δεν συμβαίνει και τελικά τα SVM δίκτυα έχουν μεγάλη ικανότητα για γενίκευση. Η VC διάσταση ενός SVM δικτύου δίνεται από το παρακάτω θεώρημα:

«Αν K ένας πυρήνας που ικανοποιεί το θεώρημα του Mercer και H ο αντίστοιχος χώρος χαρακτηριστικών, τότε αν η διάσταση του H είναι d_H, η VC διάσταση του SVM δικτύου θα είναι $d_H + 1$»

3.2.2 Η αρχιτεκτονική ενός SVM δικτύου

Η αρχιτεκτονική ενός SVM δικτύου φαίνεται στο σχήμα 3.4. Αποτελείται από δύο επίπεδα, ένα κρυφό και ένα φανερό. Το δίκτυο έχει m εισόδους, όση και η διάσταση του χώρου εισόδου. Το κρυφό επίπεδο αποτελείται από n νευρώνες, όση και η διάσταση του χώρου χαρακτηριστικών. Κάθε νευρώνας δέχεται το διάνυσμα εισόδου $\mathbf{x}$ και πραγματοποιεί τον μη γραμμικό μετασχηματισμό $K(\mathbf{x}, \mathbf{x}_i)$. Στο φανερό επίπεδο καταλήγουν όλοι οι νευρώνες του κρυφού επιπέδου, πολλαπλασιασμένοι με τα κατάλληλα βάρη που έχουν υπολογισθεί από τον αλγόριθμο βελτιστοποίησης, αλλά και το bias b. Το επίπεδο αυτό αποτελείται από έναν νευρώνα, τον νευρώνα εξόδου, ο οποίος απλά στέλνει στην έξοδο το άθροισμα των μη γραμμικών μετασχηματισμών πολλαπλασιασμένων με τα βάρη και του b. Είναι φανερή η απλότητα που έχει το δίκτυο σε σχέση με άλλους τύπους Νευρωνικών Δικτύων, αφού ουσιαστικά μόνο το κρυφό επίπεδο εκτελεί όλους τους μετασχηματισμούς.

Επίσης το δεύτερο (φανερό) επίπεδο μπορεί να παρομοιαστεί με ένα perceptron και έτσι λειτουργεί στον χώρο χαρακτηριστικών



Σχήμα 3.9: Αρχιτεκτονική ενός SVM δικτύου

3.3 Μηχανές Διανυσματικής Υποστήριξης για προβλήματα ταξινόμησης πολλών κλάσεων

Μέχρι τώρα υποθέσαμε ότι τα δεδομένα ανήκουν μόνο σε δύο κλάσεις. Ωστόσο πολλές πρακτικές εφαρμογές περιλαμβάνουν περισσότερες από δύο κλάσεις με αποτέλεσμα την επιτακτική ανάγκη γενίκευσης του προβλήματος δυαδικής ταξινόμησης [5]. Για παράδειγμα, η χειρόγραφη ψηφιακή αναγνώριση θεωρεί δεδομένα 10 κλάσεων: ψηφία από 0 έως 9. Υπάρχουν πολλοί τρόποι για να επεκτείνουμε τις μηχανές διανυσματικής υποστήριξης για τέτοιες περιπτώσεις. Εδώ, θα συζητήσουμε δύο απλές μεθόδους:

3.3.1 One-against-all

Αυτή η γνωστή μέθοδος θα έπρεπε να ονομάζεται "one-against-the-rest". Κατασκευάζει δυαδικά μοντέλα (SVM) με μια κατηγορία ως θετική και τις υπόλοιπες ως αρνητικές. Επεξηγούμε αυτή τη μέθοδο με μια απλή περίπτωση προβλήματος με τέσσερις τάξεις. Οι τέσσερις μηχανές διανυσματικής υποστήριξης δύο κλάσεων είναι:

$y_i = 1$	$y_i = -1$	**Συνάρτηση Απόφασης**
Κλάση 1	Κλάσεις 2,3,4	$f^1(x) = (w^1)^T x + b^1$
Κλάση 2	Κλάσεις 1,3,4	$f^2(x) = (w^2)^T x + b^2$
Κλάση 3	Κλάσεις 1,2,4	$f^3(x) = (w^3)^T x + b^3$
Κλάση 4	Κλάσεις 1,2,3	$f^4(x) = (w^4)^T x + b^4$

Για οποιοδήποτε δεδομένο x, αν ανήκει στην κλάση *i*, θα περίμενε κανείς ότι:

$$f^i(x) \geq 1 \text{ και } f^j(x) \leq -1, i \neq j$$

Αυτή η "προσδοκία" απορρέει άμεσα από τη ρύθμιση/σύνθεση των τεσσάρων προβλημάτων των δύο κλάσεων και από την παραδοχή ότι τα δεδομένα είναι σωστά διαχωρισμένα. Ως εκ τούτου, η $f^i(x)$ έχει τις μεγαλύτερες τιμές μεταξύ των $f^1(x), \dots, f^4(x)$ και επομένως, ο κανόνας απόφασης είναι:

$$Αναμενόμενη\ κλάση\ =\ arg\,max_{i=1,\dots 4} f^i(x)$$

3.3.2 One-against-one

Και αυτή η μέθοδος κατασκευάζει αρκετές μηχανές διανυσματικής υποστήριξης δύο κατηγοριών, αλλά η καθεμιά έχει δεδομένα εκπαίδευσης μόνο από δύο διαφορετικές κατηγορίες. Έτσι, η μέθοδος αυτή μερικές φορές ονομάζεται "σοφό ζεύγος" προσέγγισης. Για το ίδιο παράδειγμα των τεσσάρων κλάσεων, έξι προβλήματα δύο κλάσεων κατασκευάστηκαν:

$y_i = 1$	$y_i = -1$	**Συνάρτηση Απόφασης**
Κλάση 1	Κλάση 2	$f^{12}(x) = (w^{12})^T x + b^{12}$
Κλάση 1	Κλάση 3	$f^{13}(x) = (w^{13})^T x + b^{13}$
Κλάση 1	Κλάση 4	$f^{14}(x) = (w^{14})^T x + b^{14}$
Κλάση 2	Κλάση 3	$f^{23}(x) = (w^{23})^T x + b^{23}$
Κλάση 2	Κλάση 4	$f^{24}(x) = (w^{24})^T x + b^{24}$
Κλάση 3	Κλάση 4	$f^{34}(x) = (w^{34})^T x + b^{34}$

Οποιοδήποτε δεδομένο x, το βάζουμε στις έξι συναρτήσεις. Αν το πρόβλημα των κλάσεων i και j δείξει ότι το x δεδομένο θα πρέπει να είναι στην i, η κλάση i παίρνει μία ψήφο. Για παράδειγμα, ας υποθέσουμε ότι:

Κλάσεις	Νικητής
1 2	1
1 3	1
1 4	1
2 3	2
2 4	4
3 4	3

Τότε έχουμε:

Κλάση	1	2	3	4
Πλήθος ψήφων	3	1	1	1

Έτσι, το x προβλέπεται ότι θα είναι στην κλάση 1.

Για ένα σύνολο δεδομένων, με k διαφορετικές κλάσεις, η μέθοδος αυτή κατασκευάζει $\binom{k}{2} = \frac{k(k-1)}{2}$ μηχανές διανυσματικής υποστήριξης δύο κλάσεων. Μπορεί κάποιος να ανησυχεί ότι μερικές φορές περισσότερες από μία κλάσεις λαμβάνουν τις περισσότερες ψήφους. Πρακτικά, αυτή η κατάσταση δεν συμβαίνει τόσο συχνά και υπάρχουν κάποιες άλλες στρατηγικές για να το αντιμετωπίσουμε.

ΚΕΦΑΛΑΙΟ 4

ΠΑΛΙΝΔΡΟΜΗΣΗ ΜΕ ΜΗΧΑΝΕΣ ΔΙΑΝΥΣΜΑΤΙΚΗΣ ΥΠΟΣΤΗΡΙΞΗΣ

4.1 Εισαγωγή

Ιστορικά, οι μηχανές διανυσματικής υποστήριξης αναπτύχθηκαν στο πλαίσιο της ταξινόμησης προβλημάτων. Ωστόσο, υπάρχει μια άλλη σημαντική κατηγορία προβλημάτων: τα προβλήματα παλινδρόμησης. Τα προβλήματα παλινδρόμησης διαφέρουν από τα προβλήματα ταξινόμησης στο ότι οι παρατηρήσεις ενός συνόλου εκπαίδευσης δεν έχουν μια διακριτή ετικέτα (± 1), αλλά, αντ 'αυτού, συνδέονται με έναν οποιοδήποτε αριθμό. Τυπικά, ο αριθμός αυτός ανήκει στο σύνολο των πραγματικών αριθμών. Σε αυτό το κεφάλαιο θα δείξουμε πώς μπορούν οι μηχανές διανυσματικής υποστήριξης να προσαρμοστούν ώστε να χειριστούν προβλήματα παλινδρόμησης.

4.1.1 Απλή και πολλαπλή γραμμική ταξινόμηση

Από στατιστικής άποψης μπορούμε να σκεφτούμε τη γραμμική παλινδρόμηση ως την τοποθέτηση ενός υπερεπιπέδου μέσω ενός συνόλου σημείων εκπαίδευσης με ελάχιστο σφάλμα. Αυτό το σφάλμα παλινδρόμησης ονομάζεται *κατάλοιπο* ή *υπόλοιπο*, και ορίζεται ως η διαφορά μεταξύ της αναμενόμενης και της πραγματικής τιμής των δεδομένων εκπαίδευσης του μοντέλου. Στόχος της γραμμικής παλινδρόμησης είναι να ελαχιστοποιήσει τα κατάλοιπα. Για να γίνει αυτό πιο σαφές, ας υποθέσουμε ένα σύνολο εκπαίδευσης παλινδρόμησης της μορφής:

$$D = \{(\overline{x_1}, y_1), (\overline{x_2}, y_2), \dots, (\overline{x_l}, y_l)\} \subset R^n \times R \tag{4.1}$$

Ας υποθέσουμε ότι η $\hat{f}(x)$ είναι ένα μοντέλο παλινδρόμησης στο D, τότε η ποσότητα:

$$\rho_i = y_i - \hat{f}(\overline{x_i}) \tag{4.2}$$

για $(x_i, y_i) \in D$, ονομάζεται *υπόλοιπο*, και μετρά τη διαφορά μεταξύ προβλεπόμενης και της πραγματική παρατήρησης. Για ένα τέλειο μοντέλο τα κατάλοιπα είναι όλα

μηδέν, δηλαδή, οι προβλεπόμενες τιμές ταιριάζουν ακριβώς με τις τιμές που παρατηρούνται στα δεδομένα εκπαίδευσης. Ωστόσο, είναι υπερβολικά αισιόδοξο να υποθέσουμε ότι μπορούμε να κατασκευάσουμε τέτοια «τέλεια» μοντέλα για τα δεδομένα εκπαίδευσης στον πραγματικό κόσμο. Ως εκ τούτου, θα πρέπει να κατασκευάσουμε μοντέλα, όπου τα υπόλοιπα ελαχιστοποιούνται. Στην γραμμική παλινδρόμηση αυτό επιτυγχάνεται υπολογίζοντας το ελάχιστο άθροισμα των τετραγώνων των σφαλμάτων (γνωστή ως μέθοδος ελαχίστων τετραγώνων):

$$min \sum_{i=1}^{l} {\rho_i}^2 = min_{\hat{f}} \sum_{i=1}^{l} (y_i - \hat{f}(\bar{x_\iota}))^2 \quad (4.3)$$

με $(x_i, y_i) \in D$. Παρατηρούμε ότι το σφάλμα εξαρτάται από τον τύπο $\hat{f}$ που επιλέγουμε να υπολογίσουμε τα κατάλοιπα. Αυτό μας δίνει ένα πρόβλημα βελτιστοποίησης που μας επιτρέπει να υπολογίσουμε το βέλτιστο γραμμικό μοντέλο παλινδρόμησης $\widehat{f^\star}$:

$$\hat{f}^* = argmin_{\hat{f}} \sum_{i=1}^{l} (y_i - \hat{f}(\bar{x_\iota}))^2 \quad (4.4)$$

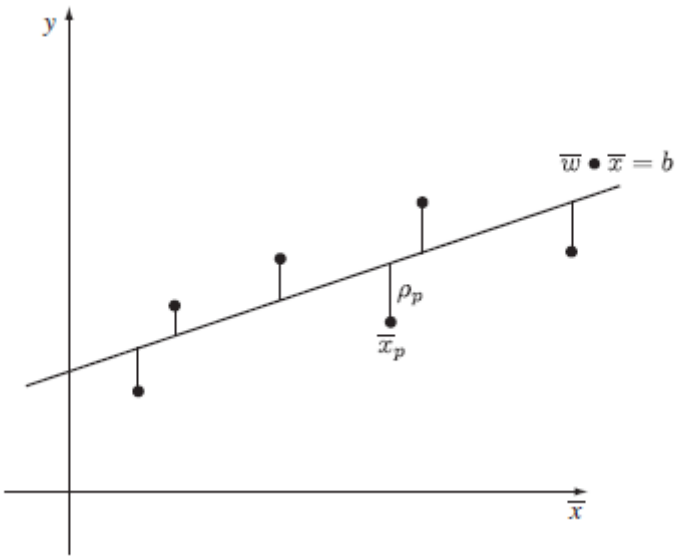
Εκμεταλλευόμενοι το γεγονός ότι κατασκευάζουμε γραμμικά μοντέλα, μπορούμε να ξαναγράψουμε την προηγούμενη εξίσωση ως:

$$(\bar{w}^*, b^*) = argmin_{\bar{w},b} \sum_{i=1}^{l} (y_i - \bar{w} \cdot \bar{x}_i + b)^2 \quad (4.5)$$

Όπου το βέλτιστο μοντέλο παλινδρόμησης είναι:

$$\hat{f}^*(\bar{x}) = \bar{w}^* \bar{x} - b^* \quad (4.6)$$

Στην απλή γραμμική παλινδρόμηση, δηλαδή, τα προβλήματα παλινδρόμησης, όπου τα δεδομένα εκπαίδευσης έχουν τη μορφή $(x, y) \in R \times R$, η εξίσωση (4.5) μπορεί να λυθεί αναλυτικά. Το Σχήμα 4.1 απεικονίζει ένα μοντέλο γραμμικής παλινδρόμησης για την απλή γραμμική παλινδρόμηση. Κάθε σημείο στο γράφημα αναπαριστά μια παρατήρηση $(x, y) \in R \times R$ και οι κάθετες γραμμές αντιπροσωπεύουν τα κατάλοιπα. Το βέλτιστο μοντέλο κατασκευάζεται έτσι ώστε να ελαχιστοποιούνται τα κατάλοιπα. Στην πολλαπλή γραμμική παλινδρόμηση, δηλαδή τα προβλήματα παλινδρόμησης, όπου τα δεδομένα εκπαίδευσης έχουν τη μορφή $(\bar{x}, y) \in R_n \times R$, η εξίσωση (4.5) μπορεί να μην μπορεί να επιλυθεί αναλυτικά εκτός αν τα δεδομένα πληρούν ορισμένες προϋποθέσεις συγγραμμικότητας.



Σχήμα 4.1: Γραμμική παλινδρόμηση με υπόλοιπα. Εδώ το σημείο $\overline{x_p}$ είναι μια παρατήρηση και ρ_p είναι το υπόλοιπο αυτής της παρατήρησης δεδομένου του μοντέλου $\bar{w} \bullet \bar{x} = b$.

4.2 Παλινδρόμηση με μηχανές μέγιστου περιθωρίου

Ένα ισχυρό κίνητρο για την ανάπτυξη των διανυσμάτων υποστήριξης στα μοντέλα παλινδρόμησης είναι η απλή επέκταση από τη γραμμική στη μη γραμμική παλινδρόμηση χρησιμοποιώντας το τέχνασμα του πυρήνα. Για την ανάπτυξη των μηχανών διανυσματικής υποστήριξης στο πλαίσιο της παλινδρόμησης, ξεκινάμε με την πρωταρχική προσαρμογή των μηχανών μέγιστου περιθωρίου. Στην περίπτωση αυτή οι βασικές ιδέες είναι ουσιαστικά οι ίδιες όπως στην περίπτωση της ταξινόμησης μηχανών μέγιστου περιθωρίου: δοθέντος ενός υπερεπίπεδου, θέλουμε να μεγιστοποιήσουμε τις αποστάσεις των παρατηρήσεων από αυτό το υπερεπίπεδο.

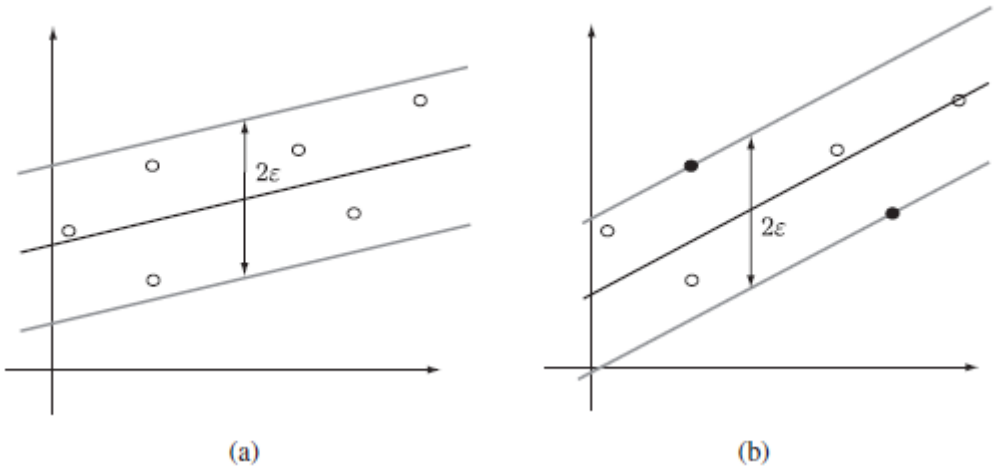
Ο Hamel (2009) ξεκινάει υποθέτοντας ότι έχουμε ένα πρόβλημα παλινδρόμησης, όπου όλες οι παρατηρήσεις των δεδομένων εκπαίδευσης της παλινδρόμησης :

$$D = \{(\overline{x_1}, y_1), (\overline{x_2}, y_2), \dots, (\overline{x_l}, y_l)\} \subseteq R^n \times R$$

προσαρμόζονται σε έναν υπερ-σωλήνα πλάτους 2ε με $\varepsilon > 0$ (βλ. Σχήμα 4.2α). Μπορούμε να ερμηνεύσουμε αυτόν τον υπερ-σωλήνα ως ένα μοντέλο παλινδρόμησης, με το να φανταστούμε ότι υπάρχει υπερεπίπεδο τοποθετημένο ακριβώς στο κέντρο του σωλήνα που μοντελοποιεί τις παρατηρήσεις. Τώρα, τυπικά υπάρχουν πολλοί διαφορετικοί τρόποι για να τοποθετηθεί ο υπερ-σωλήνας πλάτους 2ε και τα δεδομένα εκπαίδευσης να βρίσκονται όλα εντός του σωλήνα. Ωστόσο, υπάρχει μια βέλτιστη ευθυγράμμιση του υπερ-σωλήνα έτσι ώστε όσο το δυνατόν περισσότερες παρατηρήσεις να ωθούνται πιο κοντά στα εξωτερικά όρια του υπερ-σωλήνα. Με άλλα λόγια, η βέλτιστη ευθυγράμμιση του υπερ-σωλήνα επιτυγχάνεται όταν οι αποστάσεις

των παρατηρήσεων από το κέντρο του υπερεπίπεδου μεγιστοποιούνται. Αυτό απεικονίζεται στο Σχήμα 4.2b, όπου οι μαύροι κύκλοι αντιπροσωπεύουν τις παρατηρήσεις που λειτουργούν ως περιορισμοί στη βελτιστοποίηση.

Αυτό είναι παρόμοιο με το πρόβλημα της μεγιστοποίησης του περιθωρίου της συνάρτησης απόφασης και αποδεικνύεται ότι μπορούμε να χρησιμοποιήσουμε το ίδιο πρόβλημα βελτιστοποίησης για την εξεύρεση της βέλτιστης ευθυγράμμισης του υπερ-σωλήνα που χρησιμοποιήσαμε για την εύρεση του μέγιστου περιθωρίου της συνάρτησης απόφασης στο κεφάλαιο 2, ρυθμίζοντας κατάλληλα τους περιορισμούς.



Σχήμα 4.2: Επίλυση προβλημάτων παλινδρόμησης με γραμμικά μοντέλα χρησιμοποιώντας ε (υπερ-)σωλήνα: (α) μοντέλο παλινδρόμησης όπου όλες οι παρατηρήσεις είναι εντός του υπερ-σωλήνα και απεικονίζεται με γκρι γραμμές, (β) βέλτιστο μοντέλο παλινδρόμησης με μέγιστο περιθώριο.

Βελτιστοποιούμε την αρχική αντικειμενική συνάρτηση,

$$\min \varphi(\overline{w}, b) = min_{\overline{w},b} \frac{1}{2} \overline{w} \cdot \overline{w} \tag{4.7}$$

ώστε οι περιορισμοί

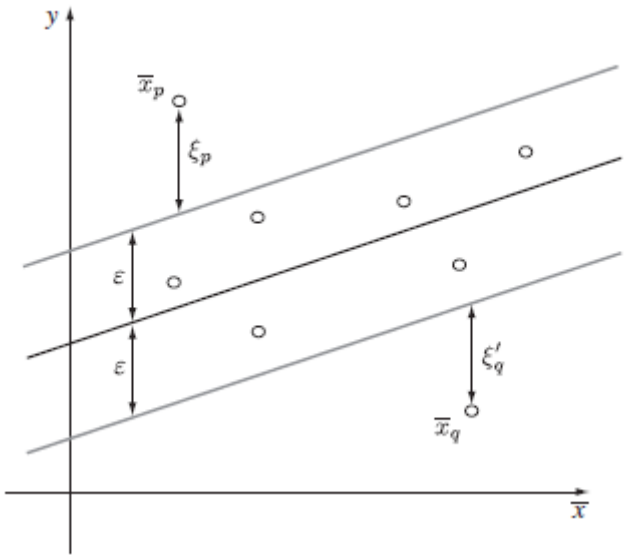
$$y_i - \hat{f}(x_i) \leq \varepsilon \tag{4.8}$$

$$\hat{f}(x_i) - y_i \leq \varepsilon \tag{4.9}$$

να ικανοποιούνται για $i = 1, \ldots, l$ και όπου $\hat{f}(\bar{x}) = \overline{w} \cdot \bar{x} - b$. Οι δύο ανισωτικοί περιορισμοί μπορούν επίσης να γραφτούν με την ενιαία ανισότητα, $\left| y_i - \hat{f}(x_i) \right| \leq \varepsilon$. Είναι ίσως αυτή η μορφή στην οποία γίνεται πιο προφανές ότι οι περιορισμοί εξασφαλίζουν ότι όλες οι παρατηρήσεις πρέπει να είναι εντός του (υπερ-)σωλήνα. Αυτό μας επιτρέπει να ερμηνεύσουμε τη βελτιστοποίηση ως υπολογισμό όπου θα προσαρμόζουμε την περιστροφή και την αντιστάθμιση του μοντέλου μέχρι να

μεγιστοποιηθούν οι αποστάσεις των παρατηρήσεων από το κέντρο του υπερεπίπεδου με τον περιορισμό να διατηρούνται όλες οι παρατηρήσεις εντός του (υπερ-)σωλήνα.

Προηγουμένως κάναμε την υπόθεση ότι είναι δυνατόν να χωρέσουν όλες οι παρατηρήσεις σε έναν υπερ-σωλήνα πλάτους 2ε. Φυσικά, στον πραγματικό κόσμο αυτό θα είναι δύσκολο να συμβεί. Για τις παρατηρήσεις που δεν βρίσκονται στον υπερ-σωλήνα με μια σταθερή τιμή του ε, εισάγουμε τις *χαλαρές μεταβλητές* που μας λένε πόσο πολύ πρέπει να διορθώσουμε αυτές τις παρατηρήσεις για να μετακινηθούν στον υπερ-σωλήνα. Αυτό απεικονίζεται στο Σχήμα 4.3. Για οποιαδήποτε παρατήρηση $(\overline{x_i}, y_i)$, χρησιμοποιούμε ένα ζευγάρι χαλαρών μεταβλητών, ξ_i και ξ_i' που επισημαίνουν την απαραίτητη διόρθωση για αυτή την παρατήρηση.



Σχήμα 4.3 Γραμμική παλινδρόμηση μέγιστου περιθωρίου με *χαλαρές μεταβλητές*.

Εάν καμία διόρθωση για την παρατήρηση δεν είναι απαραίτητη, και οι δύο χαλαρές μεταβλητές μηδενίζονται. Εάν η παρατήρηση είναι παραπάνω από τον υπερ-σωλήνα, θέτουμε τη χαλαρή μεταβλητή ξ_i ίση με την απόλυτη τιμή που είναι αναγκαία για να μετακινηθεί η παρατήρηση στον (υπερ-)σωλήνα, και ορίζουμε την άλλη χαλαρή μεταβλητή ξ_i' ίση με μηδέν. Αντίθετα, αν η παρατήρηση είναι κάτω από τον υπερ-σωλήνα, θέτουμε τη χαλαρή μεταβλητή ξ_i' ίση με την απόλυτη τιμή που είναι απαραίτητο ώστε να μετακινηθεί η παρατήρηση στον υπερ-σωλήνα και αφήνουμε την άλλη χαλαρή μεταβλητή ξ_i να είναι μηδέν. Πιο συγκεκριμένα, ορίζουμε τις χαλαρές μεταβλητές,

$$\xi_i = \begin{cases} 0, & \alpha\nu\ y_i - \hat{f}(\overline{x_i}) \le \varepsilon \\ \left| y_i - \hat{f}(\overline{x_i}) \right| - \varepsilon, & \alpha\lambda\lambda\iota\acute{\omega}\varsigma \end{cases} \tag{4.10}$$

$$\xi_i' = \begin{cases} 0, & \alpha\nu\ \hat{f}(\overline{x_\iota}) - y_i \leq \varepsilon \\ \left| y_i - \hat{f}(\overline{x_\iota}) \right| - \varepsilon, & \alpha\lambda\lambda\iota\acute{\omega}\varsigma \end{cases} \quad (4.11)$$

για κάθε $i = 1, \dots, l$ με $(\overline{x_\iota}, y_i) \in D$.

Εδώ οι χαλαρές μεταβλητές ξi είναι μηδέν, εκτός από τις παρατηρήσεις που βρίσκονται πάνω από τον υπερ-σωλήνα. Αντίθετα, οι χαλαρές μεταβλητές ξ_i' είναι μηδέν εκτός από τις παρατηρήσεις που βρίσκονται κάτω από τον υπερ-σωλήνα. Λαμβάνοντας υπόψη αυτό, η εξεύρεση του βέλτιστου υπερεπίπεδου τότε γίνεται ένα ”trade-off“ μεταξύ της μεγιστοποίησης του περιθωρίου στον υπερ-σωλήνα και της ελαχιστοποίησης της τιμής των χαλαρών μεταβλητών. Για να εκφραστεί αυτό ως πρόβλημα βελτιστοποίησης, προσθέτουμε τις χαλαρές μεταβλητές ως όρο ποινής στην αντικειμενική συνάρτηση στην (4.7). Μπορούμε τώρα να αναφερθούμε στην παλινδρόμηση με μηχανές μέγιστου περιθωρίου, ως ακολούθως:

Πρόταση 4.1 Δοθέντος ενός συνόλου εκπαίδευσης παλινδρόμησης

$$D = \{(\overline{x_1}, y_1), (\overline{x_2}, y_2), \dots, (\overline{x_l}, y_l)\} \subseteq R^n \times R,$$

μπορούμε να υπολογίσουμε το βέλτιστο μοντέλο παλινδρόμησης $\hat{f}^*(\bar{x}) = \overline{w}^* \overline{x} - b^*$

ως τη βελτιστοποίηση

$$min\ \ \varphi\left(\overline{w}, b, \bar{\xi}, \bar{\xi}'\right) = min_{\overline{w}, b, \bar{\xi}, \bar{\xi}'} \frac{1}{2} \overline{w} \cdot \overline{w} + C \sum_{i=1}^{l} (\xi_i + \xi_i') \quad (4.12)$$

υπό τους περιορισμούς

$$y_i - \hat{f}(x_i) \leq \xi_i + \varepsilon, \quad (4.13)$$
$$\hat{f}(x_i) - y_i \leq \xi_i' + \varepsilon, \quad (4.14)$$
$$0 \leq \xi_i, \xi_i' \quad (4.15)$$

για κάθε $i = 1, \dots, l$ έτσι ώστε $\hat{f}(\bar{x}) = \overline{w} \cdot \overline{x} - b$.

Στην Πρόταση 4.1 η σταθερά ποινής C διαμορφώνει το ‘’trade-off’’ μεταξύ της μεγιστοποίησης του περιθωρίου και της ελαχιστοποίησης των μεταβλητών χαλάρωσης.

4.3 Παλινδρόμηση με μηχανές διανυσματικής υποστήριξης

Μπορούμε να εξάγουμε το δυϊκό του προβλήματος για τη βελτιστοποίηση μέγιστου περιθωρίου με την κατασκευή της βελτιστοποίησης Lagrangian

$$max_{\bar{\alpha}} min_{\bar{x}} L(\bar{a}, \bar{x}) = max_{\bar{\alpha}} min_{\bar{x}} (\varphi(\bar{x}) - \sum_{i=1}^{l} a_i g_i(\bar{x})) \quad (4.16)$$

υπό τους περιορισμούς

$$a_i \geq 0 \qquad \text{για κάθε } i = 1, \dots l. \quad (4.17)$$

Εδώ $g_i(\bar{x}) \geq 0$ είναι οι περιορισμοί ανισότητας, ενώ οι μεταβλητές $\bar{\alpha}$ και $\bar{x}$ ονομάζονται δυικές και πρωταρχικές μεταβλητές του προβλήματος βελτιστοποίησης, αντίστοιχα.
Ως πρώτο βήμα για την κατασκευή της βελτιστοποίησης Lagrange δημιουργούμε τους περιορισμούς ανισότητας. Αυτό γίνεται εύκολα αναδιατυπώνοντας τους περιορισμούς που περιλαμβάνονται στο αρχικό πρόβλημα βελτιστοποίησης της Πρότασης 4.1:

$$\xi_i + \varepsilon - y_i + \hat{f}(\bar{x}_\iota) \geq 0 \quad (4.18)$$
$$\xi_i' + \varepsilon - \hat{f}(\bar{x}_\iota) + y_i \geq 0 \quad (4.19)$$
$$\xi_i \geq 0 \quad (4.20)$$
$$\xi_i' \geq 0 \quad (4.21)$$

Το γεγονός ότι έχουμε τέσσερις ομάδες με ανισοτικούς περιορισμούς σημαίνει ότι θα πρέπει να εισάγουμε τέσσερις ομάδες δυικών μεταβλητών στην βελτιστοποίηση Lagrange. Αντικαθιστώντας την αντικειμενική συνάρτηση και τους περιορισμούς ανισότητας στην βελτιστοποίηση Lagrangian έχουμε τα ακόλουθα:

$$max_{\bar{\alpha},\bar{\alpha}',\bar{\beta},\bar{\beta}'} min_{\bar{w},b,\bar{\xi},\bar{\xi}'} L\left(\bar{\alpha}, \bar{\alpha}', \bar{\beta}, \bar{\beta}', \bar{w}, b, \bar{\xi}, \bar{\xi}'\right) =$$

$$= max_{\bar{\alpha},\bar{\alpha}',\bar{\beta},\bar{\beta}'} min_{\bar{w},b,\bar{\xi},\bar{\xi}'} \left(\frac{1}{2}\bar{w} \cdot \bar{w}\right) + C \sum_{i=1}^{l} (\xi_i + \xi_i') - \sum_{i=1}^{l} a_i \left(\xi_i + \varepsilon - y_i + \hat{f}(\bar{x}_\iota)\right) - \sum_{i=1}^{l} a_i' \left(\xi_i + \varepsilon - \hat{f}(\bar{x}_\iota) + y_i\right) - \sum_{i=1}^{l} \beta_i \xi_i - \sum_{i=1}^{l} \beta_i' \xi_i'$$

υπό τους περιορισμούς;

$$a_i, a_i', \beta_i, \beta_i' \geq 0 \quad (4.22)$$

για κάθε $i = 1, \dots, l$ και όπου $\hat{f}(\bar{x}) = \bar{w} \cdot \bar{x} - b$.

Με δεδομένη μία λύση της βελτιστοποίησης Lagrange

$$max_{\bar{\alpha},\bar{\alpha}',\bar{\beta},\bar{\beta}'} min_{\bar{w},b,\bar{\xi},\bar{\xi}'} L\left(\bar{\alpha},\bar{\alpha}',\bar{\beta},\bar{\beta}',\bar{w},b,\bar{\xi},\bar{\xi}'\right) = L(\bar{\alpha}^*,\bar{\alpha}'^*,\bar{\beta}^*,\bar{\beta}'^*,\bar{w}^*,b^*,\bar{\xi}^*,\bar{\xi}^*) \tag{4.23}$$

Γνωρίζουμε ότι θα ικανοποιούνται οι συνθήκες ΚΚΤ:

$$\frac{\partial L(\bar{\alpha}^*,\bar{\alpha}'^*,\bar{\beta}^*,\bar{\beta}'^*,\bar{w}^*,b^*,\bar{\xi}^*,\bar{\xi}^*)}{\partial \bar{w}} = \bar{0} \tag{4.24}$$

$$\frac{\partial L(\bar{\alpha}^*,\bar{\alpha}'^*,\bar{\beta}^*,\bar{\beta}'^*,\bar{w}^*,b^*,\bar{\xi}^*,\bar{\xi}^*)}{\partial b} = 0 \tag{4.25}$$

$$\frac{\partial L(\bar{\alpha}^*,\bar{\alpha}'^*,\bar{\beta}^*,\bar{\beta}'^*,\bar{w}^*,b^*,\bar{\xi}^*,\bar{\xi}^*)}{\partial \xi_i} = 0 \tag{4.26}$$

$$\frac{\partial L(\bar{\alpha}^*,\bar{\alpha}'^*,\bar{\beta}^*,\bar{\beta}'^*,\bar{w}^*,b^*,\bar{\xi}^*,\bar{\xi}^*)}{\partial \xi_i'} = 0 \tag{4.27}$$

$$\alpha_\iota{}^*\left(\xi_i{}^* + \varepsilon - y_i + \hat{f}^*(\bar{x_\iota})\right) = 0 \tag{4.28}$$
$$\alpha_\iota{}'^*\left(\xi_i{}'^* + \varepsilon - \hat{f}^*(\bar{x_\iota}) + y_i\right) = 0 \tag{4.29}$$
$$\beta_i{}^*\xi_i{}^* = 0 \tag{4.30}$$
$$\beta_i{}'^*\xi_i{}'^* = 0 \tag{4.31}$$

$$\xi_i{}^* + \varepsilon - y_i + \hat{f}^*(\bar{x_\iota}) \geq 0 \tag{4.32}$$
$$\xi_i{}'^* + \varepsilon - \hat{f}^*(\bar{x_\iota}) + y_i \geq 0 \tag{4.33}$$
$$\xi_i{}^*, \xi_i{}'^* \geq 0 \tag{4.34}$$
$$\alpha_\iota, \alpha_\iota{}' \geq 0 \tag{4.35}$$
$$\beta_i{}^*, \beta_i{}'^* \geq 0 \tag{4.36}$$

για κάθε $i = 1, \dots, l$ και $\hat{f}^*(\bar{x}) = \bar{w}^*\,\bar{x} - b^*$ είναι η βέλτιστη συνάρτηση παλινδρόμησης. Χρησιμοποιώντας στις συνθήκες ΚΚΤ δεν είναι δύσκολο να αποδειχθεί ότι ισχύει η παρακάτω πρόταση.

Πρόταση 4.2 Δοθέντος ενός συνόλου εκπαίδευσης παλινδρόμησης

$$D = \{(\bar{x_1}, y_1), (\bar{x_2}, y_2), \dots, (\bar{x_l}, y_l)\} \subseteq R^n \times R,$$

μπορούμε να υπολογίσουμε το βέλτιστο μοντέλο διανυσματικής υποστήριξης για την παλινδρόμηση $\hat{f}^*(\bar{x}) = \bar{w}^*\,\bar{x} - b^*$ με το δυϊκό πρόβλημα

$$max_{\bar{a},\bar{a}'}\varphi'(\bar{a},\bar{a}') = \\ = max_{\bar{a},\bar{a}'}\left(-\frac{1}{2}\sum_{i=1}^{l}\sum_{i=1}^{l}(a_i - a_i\,') + \sum_{i=1}^{l} y_i\,(a_i - a_i')\right) - \varepsilon\sum_{i=1}^{l}(a_i + a_i')$$

(4.37)

υπό τους περιορισμούς

$$\sum_{i=1}^{l}(a_i + a_i') = 0 \quad (4.38)$$

$$C \geq a_i, a_i' \geq 0 \quad (4.39)$$

για κάθε $i = 1, \dots, l$ όπου

$$\bar{w}^* = \sum_{i=1}^{l}(a_i^* - a_i'^*)\bar{x}_i \quad (4.40)$$

$$b^* = \frac{1}{l}\sum_{i=1}^{l}\bar{w}^* \cdot \bar{x}_i - y_i \quad (4.41)$$

Σε μοντέλα παλινδρόμησης με διανύσματα υποστήριξης μπορούμε να ερμηνεύσουμε μια παρατήρηση $(\bar{x}_i, y_i)$ για την οποία ο συντελεστής $(a_i - a_i')$ είναι μη μηδενικός ως διάνυσμα υποστήριξης. Παρατηρούμε ότι η λύση στο βέλτιστο μοντέλο παλινδρόμησης

$$\hat{f}^*(\bar{x}) = \bar{w}^* \cdot \bar{x} - b^* = \sum_{i=1}^{l}(a_i^* - a_i'^*)\bar{x}_i \cdot \bar{x} - \frac{1}{l}\sum_{i=1}^{l}\sum_{j=1}^{l}(a_i^* - a_i'^*)\,\bar{x}_i \cdot \bar{x}_j - y_j \quad (4.42)$$

εξαρτάται μόνο από τα διανύσματα υποστήριξης. Ως εκ τούτου, μπορούμε να αναφερόμαστε σε αυτό το μοντέλο ως παλινδρόμηση με μηχανές διανυσματικής υποστήριξης. Ως τελευταία παρατήρηση, είναι ιδιαίτερα σημαντικό το γεγονός ότι η γραμμική παλινδρόμηση με μηχανές διανυσματικής υποστήριξης μπορεί να επεκταθεί σε μη γραμμική παλινδρόμηση με την εφαρμογή του τεχνάσματος του πυρήνα για τη βελτιστοποίηση και το μοντέλο. Δηλαδή, μπορούμε να αντικαταστήσουμε το εσωτερικό γινόμενο στη βελτιστοποίηση και στο μοντέλο με μια κατάλληλη συνάρτηση πυρήνα για να επεκτείνουμε τα διανύσματα υποστήριξης παλινδρόμησης για μη γραμμικά σύνολα.

4.4 Εκτίμηση μοντέλου

Παρόμοια με την ταξινόμηση μοντέλων διανυσματικής υποστήριξης, τα μοντέλα παλινδρόμησης με διανύσματα υποστήριξης έχουν έναν αριθμό ελεύθερων παραμέτρων που πρέπει να είναι συντονισμένες. Αυτές είναι οι πυρήνες K με τις αντίστοιχες παραμέτρους λ, η παράμετρος ε, και η σταθερά κόστους C. Οι τεχνικές hold-out και cross-validation, καθώς και bootstrapping μεταφέρονται στα μοντέλα παλινδρόμησης με διανύσματα υποστήριξης, με τη διαφορά ότι η συνάρτηση απώλειας 0-1 αντικαθίσταται από μια συνάρτηση απώλειας που υπολογίζει πόσο καλά το μοντέλο ταιριάζει με τις παρατηρήσεις. Δηλαδή, αντί να μετράμε πόσο εσφαλμένη είναι η ταξινόμηση, μετράμε πόσο διαφορετική είναι η προβλεπόμενη τιμή από την παρατηρούμενη τιμή.

Η πιο κοινή εκτίμηση σφάλματος για τη συνάρτηση παλινδρόμησης είναι το μέσο τετραγωνικό σφάλμα (MSE). Ορίζουμε μια συνάρτηση απώλειας L_2 που υπολογίζει το τετράγωνο του υπολοίπου για μία παρατήρηση $(\bar{x}, y)$ δοθέντος ενός μοντέλου $\hat{f}$,

$$L_2\left(y, \hat{f}(\bar{x})\right) = (y - \hat{f}(\bar{x}))^2 \tag{4.43}$$

Τώρα, έστω ένα σύνολο εκπαίδευσης για την παλινδρόμηση:

$$D = \{(\overline{x_1}, y_1), (\overline{x_2}, y_2), \dots, (\overline{x_l}, y_l)\} \subset R^n \times R$$

ορίζουμε το μέσο τετραγωνικό σφάλμα υπολογισμένο στο D,

$$MSE_D\left[\widehat{f_D}[k, \lambda, \varepsilon, C]\right] = \frac{1}{l}\sum_{i=1}^{l} L_2(y_i, \widehat{f_D}[k, \lambda, \varepsilon, C](\overline{x_i})) \tag{4.44}$$

με $(\overline{x_i}, y_i) \in D$. Εδώ χρησιμοποιούμε το μοντέλο $\widehat{f_D}$, όπου ο δείκτης υποδεικνύει ότι κατασκευάστηκε χρησιμοποιώντας το σετ D. Σε αυτή την περίπτωση μπορούμε να ερμηνεύσουμε το μέσο τετραγωνικό σφάλμα ως τη μέση απώλεια L_2 του μοντέλου $\widehat{f_D}$, από τα δεδομένα του συνόλου D.

Μπορούμε να γενικεύσουμε αυτή την τεχνική (hold-out testing technique) διασπώντας το σύνολο D σε δυο μη επικαλυπτόμενα σύνολα τα P και Q, έτσι ώστε:

$$D = P \cup Q, \tag{4.45}$$

όπου χρησιμοποιούμε το P ως σύνολο εκπαίδευσης και το Q ως σύνολο δοκιμής. Το σφάλμα αναμονής (hold-out error) μπορεί τότε να υπολογιστεί ως εξής:

$$MSE_Q\left[\widehat{f_P}[k, \lambda, \varepsilon, C]\right] = \frac{1}{|Q|}\sum_{(\overline{x_i}, y_i) \in Q} L2(y_i, \widehat{f_P}[k, \lambda, \varepsilon, C](\overline{x_i})) \tag{4.46}$$

Για το σφάλμα αναμονής υπολογίζουμε τη μέση απώλεια στο σύνολο Q ενός μοντέλου εκπαιδευμένου στο P. Ένα βέλτιστο μοντέλο θα ελαχιστοποιήσει το μέσο τετραγωνικό σφάλμα στο σύνολο Q.

Δεδομένης της συνάρτησης απώλειας L_2, είναι εύκολο να εξάγουμε μία αντίστοιχη έκφραση για το σφάλμα διασταύρωσης (cross-validated error). Είναι επίσης εύκολο να γενικεύσει κανείς τον υπολογισμό του διαστήματος εμπιστοσύνης με βάση το bootstrap από την ταξινόμηση στην παλινδρόμηση. Τέλος θα πρέπει να αναφέρουμε ότι μια άλλη δημοφιλής μέθοδος προσδιορισμού του σφάλματος ενός μοντέλου παλινδρόμησης είναι η ρίζα του μέσου τετραγωνικού σφάλματος (root-mean-squared error) που προέρχεται από το μέσο τετραγωνικό σφάλμα απλά με τη λήψη της τετραγωνικής του ρίζας.

ΚΕΦΑΛΑΙΟ 5

ΑΞΙΟΛΟΓΗΣΗ ΜΟΝΤΕΛΟΥ

5.1 Εισαγωγή

Αρκετά παγκόσμια δεδομένα που σχετίζονται με περισσότερα σχέδια γνώσης είναι πολύ μεγάλα για να θεωρηθούν ως σύνολα κατάρτισης για μηχανές διανυσματικής υποστήριξης. Ως εκ τούτου, είναι αναγκαίο να χρησιμοποιούμε υποσύνολα των δεδομένων αυτών ως σύνολα εκπαίδευσης. Από τη στιγμή που χρησιμοποιούμε μόνο ένα υποσύνολο των παγκόσμιων δεδομένων ως σύνολο εκπαίδευσης, βρισκόμαστε αντιμέτωποι με το ερώτημα: Πόσο καλά αποδίδει το μοντέλο, στις περιπτώσεις όπου τα δεδομένα δεν είναι μέρος του συνόλου της κατάρτισης; Σε αυτό το κεφάλαιο θα παρουσιάσουμε τεχνικές που θα μας επιτρέψουν να ποσοτικοποιήσουμε την απόδοση των μοντέλων μας, στο πλαίσιο αυτής της αβεβαιότητας. Ιδιαίτερη έμφαση θα δοθεί στις ROC καμπύλες (Receiver Operating Characteristics). Ένα γράφημα ROC είναι μια τεχνική για την οπτικοποίηση, οργάνωση και ταξινόμηση με βάση την απόδοση των δεδομένων. Τα ROC γραφήματα εδώ και καιρό έχουν χρησιμοποιηθεί στη θεωρία ανίχνευσης σημάτων για να περιγράψουν την σχέση μεταξύ των συντελεστών και ποσοστών εσφαλμένων συναγερμών ή ταξινομητές. Η ROC ανάλυση έχει εκτεταμένη χρήση στην απεικόνιση και στην ανάλυση της συμπεριφοράς των διαγνωστικών συστημάτων. Τέλος, θα αναφέρουμε στη μέθοδο της διασταυρωμένης επικύρωσης (cross validation).

5.1.1 Στατιστικός έλεγχος υποθέσεων

Σε πολλά προβλήματα δεν ενδιαφερόμαστε να εκτιμήσουμε με κάποια ακρίβεια την τιμή μίας παραμέτρου αλλά να διαπιστώσουμε αν η παράμετρος είναι μικρότερη ή μεγαλύτερη από μία δεδομένη τιμή που έχει σημασία για το πρόβλημά μας. Δηλαδή επιθυμούμε να ελέγξουμε αν μία ή περισσότερες παράμετροι ενός πληθυσμού ικανοποιούν μία βασική υπόθεση έναντι μίας εναλλακτικής υπόθεσης. Σχεδόν πάντα δεν είμαστε σε θέση να καταγράψουμε όλο τον πληθυσμό οπότε αρκούμαστε σε ένα τυχαίο δείγμα από αυτόν. Με βάση αυτό το τυχαίο δείγμα θέλουμε να πάρουμε μία απόφαση: να απορρίψουμε ή όχι αν ισχύει η βασική υπόθεση.

Ο έλεγχος υπόθεσης (hypothesis testing) επεξεργάζεται στατιστικά εργαλεία (τον εκτιμητή και την κατανομή του) σε μία διαδικασία λήψης απόφασης. Για τη διαδικασία ελέγχου μίας στατιστικής υπόθεσης πρώτα ορίζουμε τη στατιστική υπόθεση, και μετά

υπολογίζουμε το στατιστικό ελέγχου και την περιοχή απόρριψης και τέλος αποφασίζουμε για την υπόθεση με βάση την ένδειξη που έχουμε από το δείγμα.

Η στατιστική υπόθεση (statistical hypothesis) μπορεί να είναι μία οποιαδήποτε στατιστική δήλωση ή πρόταση που θέτουμε υπό έλεγχο με βάση τις παρατηρήσεις. Η μηδενική υπόθεση (null hypothesis) την οποία θέτουμε υπό έλεγχο συμβολίζεται με H_0 ενώ η εναλλακτική υπόθεση (alternative hypothesis) την οποία δεχόμαστε αν απορρίψουμε την H_0 συμβολίζεται με H_1. Οι δυνατές αποφάσεις του ελέγχου είναι:

1. Σωστή απόφαση: Αποδεχόμαστε την H_0 όταν η H_0 είναι σωστή. Η πιθανότητα αυτής της απόφασης είναι P(αποδοχή της H_0/ H_0 σωστή)=1-α.
2. **Σφάλμα τύπου II** (type II error): Αποδεχόμαστε την H_0 όταν η H_0 είναι λανθασμένη. Η πιθανότητα αυτού του σφάλματος είναι P(αποδοχή της H_0/ H_0 λανθασμένη) =β.
3. **Σφάλμα τύπου I** (type I error): Απορρίπτουμε την H_0 όταν η H_0 είναι σωστή. Η πιθανότητα αυτού του σφάλματος είναι το επίπεδο σημαντικότητας P(απόρριψη της H_0/ H_0 σωστή) =α.
4. Σωστή απόφαση: Απορρίπτουμε την H_0 και η H_0 είναι λανθασμένη. Η πιθανότητα αυτής της απόφασης είναι P(απόρριψη της H_0/ H_0 λανθασμένη) =1-β, και δηλώνει την ισχύ του ελέγχου (power of the test).

Οι 4 δυνατές περιπτώσεις στην απόφαση του ελέγχου δίνονται στον πίνακα.

	Αποδοχή της H_0	Απόρριψη της H_0
H_0 σωστή	ορθή απόφαση $(1-a)$	σφάλμα τύπου I (a)
H_0 λανθασμένη	σφάλμα τύπου II (β)	ορθή απόφαση $(1-\beta)$

Πίνακας 5.1: Οι 4 περιπτώσεις στην απόφαση ελέγχου με την αντίστοιχη πιθανότητα σε παρένθεση.

Για να είναι ένας έλεγχος ακριβής θα πρέπει το πραγματικό σφάλμα τύπου I να είναι στο επίπεδο σημαντικότητας α στο οποίο γίνεται ο έλεγχος. Στην πράξη αυτό δεν είναι βέβαια εφικτό.

Αλλά μπορούμε να το διαπιστώσουμε αν έχουμε τη δυνατότητα να κάνουμε προσομοιώσεις. Για να υπολογίσουμε το πραγματικό σφάλμα τύπου I θα πρέπει να γνωρίζουμε ότι η H_0 είναι σωστή και να επαναλάβουμε τον έλεγχο σε M διαφορετικά δείγματα ίδιου τύπου και στο ίδιο επίπεδο σημαντικότητας α. Αν η H_0 απορρίπτεται m φορές σε επίπεδο σημαντικότητας α θα πρέπει για να είναι ο έλεγχος ακριβής (να έχει σωστή σημαντικότητα) η αναλογία m/M να είναι κοντά στο α. Επίσης μας ενδιαφέρει ο έλεγχος να έχει μεγάλη ισχύ. Την ισχύ αυτού του ελέγχου μπορούμε υπολογιστικά να την μετρήσουμε και πάλι με προσομοιώσεις όπου τώρα θα πρέπει να γνωρίζουμε ότι η H_0 δεν είναι σωστή και επιπλέον ότι ο έλεγχος έχει σωστή σημαντικότητα. Οι παραπάνω προσομοιώσεις συνήθως ακολουθούνται για να αξιολογήσουμε το στατιστικό που χρησιμοποιείται στον έλεγχο.

p-value (ή significance value)

Αν η περιοχή απόρριψης της H_0 είναι της μορφής T(x)>c (όπου T κατάλληλη στατιστική συνάρτηση $T(X)=T(X_1,X_2,\dots X_n)$, του τυχαίου δείγματος $(X_1,X_2,\dots X_n)$ τότε το p-value των τιμών x του δείγματος είναι η πιθανότητα

$$p-value = P(T(X) > T(x)|\, H_0) = 1 - F_{T|H_0}(T(x))$$

η οποία μπορεί να θεωρηθεί ότι εκφράζει την πιθανότητα να εμφανιστεί ένα τόσο ή ακόμα πιο ακραίο δείγμα από αυτό που εμφανίστηκε, δεδομένου ότι ισχύει η H_0. Διαισθητικά αν το p-value είναι κοντά στο μηδέν τότε συμπεραίνουμε ότι είναι απίθανο, δεδομένης, της H_0 να εμφανιστεί αυτό το δείγμα και όπως είναι φυσικό φτάνουμε στο συμπέρασμα ότι μάλλον δεν ισχύει η H_0. Πράγματι, αν p-value<α απορρίπτουμε την H_0. Επομένως αντί να εξετάζουμε αν T(x)>c, ισοδύναμα εξετάζουμε αν:

- αν $p-value < \alpha$: απορρίπτουμε την H_0, ενώ
- αν $p-value \geq \alpha$: δεν απορρίπτουμε την H_0

Αν το p-value είναι πάρα πολύ μικρό (π.χ. 0,0001) τότε απορρίπτουμε την H_0 χωρίς δεύτερη σκέψη ενώ αν το p-value είναι σχετικά μικρό (π.χ. κοντά στο 0,05) τότε μπορεί μεν να απορρίψουμε την H_0 αλλά με κάποια επιφυλακτικότητα.

Το πλεονέκτημα από τη χρήση του p-value είναι ότι δεν απορρίπτουμε ή δεχόμαστε απλώς την H_0, αλλά μπορούμε να δούμε και πόσο πιθανή ήταν η εμφάνιση του δείγματος x που πήραμε (υπό την H_0) ενώ επίσης μπορούμε να την συγκρίνουμε άμεσα με όποιο α και αν επιλέξουμε.

5.2. Κριτήρια Απόδοσης του Μοντέλου - Confusion Matrix

Όταν έχουμε να κάνουμε με ένα δυαδικό πρόβλημα ταξινόμησης, υπάρχουν τέσσερις πιθανές εκβάσεις, όταν ένα μοντέλο εφαρμόζεται σε μια παρατήρηση. Ας είναι $(\bar{x}, y) \in R^n \times \{+1, -1\}$ μια παρατήρηση και έστω $f: R^n \times \{+1, -1\}$ να είναι ένα μοντέλο. Στη συνέχεια, έχουμε τις ακόλουθα τέσσερις δυνατότητες όταν το μοντέλο εφαρμόζεται στην παρατήρηση:

$$\hat{f}(\bar{x}) = \begin{cases} +1, & \alpha\nu\ y{=}+1, \ \ true\ positive \\ -1, & \alpha\nu\ y{=}+1, \ \ false\ negative \\ +1, & \alpha\nu\ y{=}-1, \ \ false\ positive \\ -1, & \alpha\nu\ y{=}-1, \ \ true\ negative \end{cases} \tag{5.1}$$

Εάν η τιμή εξόδου $\hat{f}(\bar{x})$ του μοντέλου ταιριάζει με την ετικέτα y της παρατήρησης, έχουμε είτε μια αληθώς θετική ή μια αληθώς αρνητική έκβαση. Εάν η τιμή εξόδου του μοντέλου δεν ταιριάζει με την παρατηρούμενη ετικέτα, έχουμε είτε ένα ψευδώς θετικό

ή ψευδώς αρνητικό αποτελέσματα, και τα δύο εκ των οποίων είναι λάθος αποτελέσματα.

Σε πολλές περιπτώσεις, είναι σημαντικό να γίνεται διάκριση μεταξύ των δύο αυτών λάθος αποτελεσμάτων, όταν γίνεται αξιολόγηση για την επίδοση του μοντέλου. Ας εξετάσουμε το ακόλουθο κλινικό παράδειγμα. Υποθέτουμε ότι θα αναπτύξουμε ένα μοντέλο που, δεδομένων των παραμέτρων της ιστολογικής βιοψίας, θα προβλέψει αν αυτός ο ιστός είναι καρκινικός ή όχι. Σύμφωνα με την παραπάνω συζήτηση, το μοντέλο μπορεί να δεσμευτεί με τα δύο είδη των λαθών. Μπορεί να διαπράξει ένα ψευδώς θετικό λάθος, δηλαδή, να προβλέψει ότι το δείγμα ιστού είναι καρκινικό όταν δεν είναι. Μπορεί επίσης να δεσμευθεί με ένα ψευδώς αρνητικό σφάλμα. Εδώ το μοντέλο προβλέπει ότι το δείγμα του ιστού δεν είναι καρκινικό όταν στην πραγματικότητα είναι. Σε κλινικό περιβάλλον το τελευταίο είναι ένα πολύ πιο σοβαρό σφάλμα από ότι το πρώτο, δεδομένου ότι ένα ψευδώς αρνητικό λάθος συνεπάγεται ότι ο ασθενής θα παραμείνει χωρίς θεραπεία, ενώ ένα ψευδώς θετικό συνήθως οδηγεί σε περισσότερες δοκιμές μέχρις ότου ανιχνευθεί το ψευδώς θετικό σφάλμα και ο ασθενής απορριφθεί κατάλληλα.

Κατά την ανάλυση επίδοσης του μοντέλου σε αυτούς τους τύπους καταστάσεων πρέπει να κατανοήσουμε τους διαφορετικούς τύπους των σφαλμάτων που το μοντέλο μας διαπράττει. Μια αναπαράσταση της απόδοσης του μοντέλου που ονομάζεται πίνακας σύγχυσης (Confusion Matrix) διακρίνει τους δύο τύπους λαθών και ως εκ τούτου είναι το εργαλείο της επιλογής κατά την ανάλυση της επίδοσης του μοντέλου, όπου το ένα ή το άλλο είδος σφάλματος μπορεί να έχει σοβαρές συνέπειες.

Ένας πίνακας σύγχυσης για ένα δυαδικό μοντέλο ταξινόμησης είναι ένας 2 × 2 πίνακας, στον οποίο εμφανίζονται οι παρατηρούμενες ετικέτες έναντι των προβλεπόμενων ετικετών για ένα σύνολο δεδομένων. Ένας τρόπος για να απεικονίσουμε έναν πίνακα σύγχυσης είναι να θεωρήσουμε ότι η εφαρμογή ενός μοντέλου $\hat{f}$ σε μια παρατήρηση $(\bar{x}, y)$ θα μας δώσει δύο ετικέτες. Η πρώτη ετικέτα, y, οφείλεται στην παρατήρηση, και η δεύτερη ετικέτα, $\hat{y} = \hat{f}(\bar{x})$, οφείλεται στην πρόβλεψη του μοντέλου. Δηλαδή, για κάθε παρατήρηση έχουμε ένα ζεύγος ετικετών $(y, \hat{y})$. Αυτό το ζεύγος των ετικετών προσδιορίζει τις συντεταγμένες κάθε παρατήρησης μέσα στον πίνακα σύγχυσης: Η πρώτη ετικέτα διευκρινίζει τη γραμμή του πίνακα και η δεύτερη ετικέτα διευκρινίζει τη στήλη του πίνακα. Ως εκ τούτου, μια παρατήρηση με το ζεύγος ετικέτας $(y, \hat{y})$ θα χαρτογραφηθεί επάνω σε έναν πίνακα, όπως στον παρακάτω πίνακα:

	Predicted ($\hat{y}$)	
Observed (y)	$+1$	-1
$+1$	True positive (TP)	False negative (FN)
-1	False positive (FP)	True negative (TP)

Πίνακας 5.2: Διάταξη ενός πίνακα σύγχυσης (Confusion Matrix)

Οι αληθώς θετικές προβλέψεις χαρτογραφούνται στην επάνω αριστερή γωνία του πίνακα σύγχυσης, και οι αληθώς αρνητικές προβλέψεις χαρτογραφούνται στην κάτω δεξιά γωνία του πίνακα. Οι ψευδώς θετικές και ψευδώς αρνητικές αντιστοιχίζονται στην κάτω αριστερά και πάνω δεξιά γωνία του πίνακα, αντίστοιχα. Για ένα μοντέλο το οποίο δεν διαπράττει τυχόν λάθη, όλες οι προβλέψεις θα πρέπει να χαρτογραφηθούν στα πάνω αριστερά και κάτω δεξιά πεδία. Για μοντέλα που διαπράττουν λάθη, βλέπουμε ότι τα λάθη θα αντιστοιχηθούν στον πίνακα σύμφωνα με τον τύπο του σφάλματος που το μοντέλο διαπράττει.

Ο Πίνακας 5.2 δείχνει έναν πίνακα σύγχυσης ενός μοντέλου που εφαρμόζεται σε ένα σύνολο 200 παρατηρήσεων. Σε αυτό το σύνολο των παρατηρήσεων το μοντέλο διαπράττει 7 ψευδώς αρνητικά σφάλματα και 4 ψευδώς θετικά σφάλματα επιπροσθέτως των 95 αληθώς θετικών και 94 αληθώς αρνητικών προβλέψεων.

	Predicted	
Observed	$+1$	-1
$+1$	95	7
-1	4	94

Πίνακας 5.3: Τυπικός Πίνακας Σύγχυσης(Confusion Matrix) για μοντέλο που εφαρμόζεται σε ένα συγκεκριμένο σύνολο παρατηρήσεων.

Αν αυτό ήταν ένα μοντέλο για το παραπάνω παράδειγμα, το γεγονός ότι το μοντέλο διαπράττει σχεδόν διπλάσια ψευδώς αρνητικά λάθη από ό, τι ψευδώς θετικά λάθη θα είναι αιτία ανησυχίας. Εδώ θα ήταν σκόπιμο να οικοδομήσουμε ένα νέο μοντέλο με πιο ισορροπημένα λάθη. Μόνο ο πίνακας σύγχυσης είναι σε θέση να παρέχει αυτού του είδους τη διορατικότητα. Το συνολικό σφάλμα του μοντέλου είναι:

$$err = \frac{1}{200}(4 + 7) = 0{,}055$$

Ένα μοντέλο που διαπράττει ένα σφάλμα πρόβλεψης 5,5% φαίνεται σαν ένα λογικό μοντέλο. Ωστόσο, στο πλαίσιο όπου τα σφάλματα του μοντέλου μπορεί να έχουν σοβαρές συνέπειες, θα πρέπει να εξετάσουμε πιο προσεκτικά τα είδη των λαθών που διαπράττει ένα μοντέλο.

Δεδομένου ενός πίνακα σύγχυσης για ένα μοντέλο όπως στον Πίνακα 4.1, μπορούμε να υπολογίσουμε το σφάλμα του μοντέλου μας και την ακρίβεια (accuracy και precision) απευθείας από τον πίνακα,

ως εξής:

$$err = \frac{FP+FN}{TP+TN+FP+FN} \quad (5.2)$$

$$accurracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (5.3)$$

και

$$precision = \frac{TP}{TP+FP} \quad (5.4)$$

αντιστοίχως. Εκτός από αυτές τις μετρικές έχουμε δύο άλλες μετρήσεις που συνήθως χρησιμοποιούνται για να χαρακτηρίσουμε την επίδοση του μοντέλου: ευαισθησία (sensitivity) και ειδικότητα (specificity). Η *ευαισθησία* του μοντέλου ορίζεται ως το πηλίκο των αληθώς θετικών προβλέψεων δια του αθροίσματος όλων των θετικών παρατηρήσεων,

$$sensitivity = TPR = \frac{TP}{P} = \frac{TP}{TP+FN} \quad (5.5)$$

Η *ειδικότητα* ενός μοντέλου ορίζεται ως οι αληθώς αρνητικές προβλέψεις διαιρούμενες με το άθροισμα όλων των αρνητικών παρατηρήσεων,

$$specificity = TNR = \frac{TN}{N} = \frac{TN}{TN+FP} \quad (5.6)$$

Μια ευαισθησία του 1,0 για ένα μοντέλο σημαίνει ότι το μοντέλο προβλέπει όλες τις θετικές παρατηρήσεις σωστά: με άλλα λόγια, το μοντέλο δεν διαπράττει καμία ψευδώς αρνητική πρόβλεψη.

Μία εξειδίκευσή του 1,0 για ένα μοντέλο σημαίνει ότι το μοντέλο προβλέπει όλες τις αρνητικές παρατηρήσεις σωστά, με άλλα λόγια, το μοντέλο δεν διαπράττει καμία ψευδώς θετική πρόβλεψη.
Πηγαίνοντας πίσω στον πίνακα σύγχυσης του μοντέλου μας που δίνεται στον Πίνακα 5.3, μπορούμε να υπολογίζουν τις μετρικές ως εξής:

$$err = \frac{4+7}{95+94+4+7} = 0{,}055$$

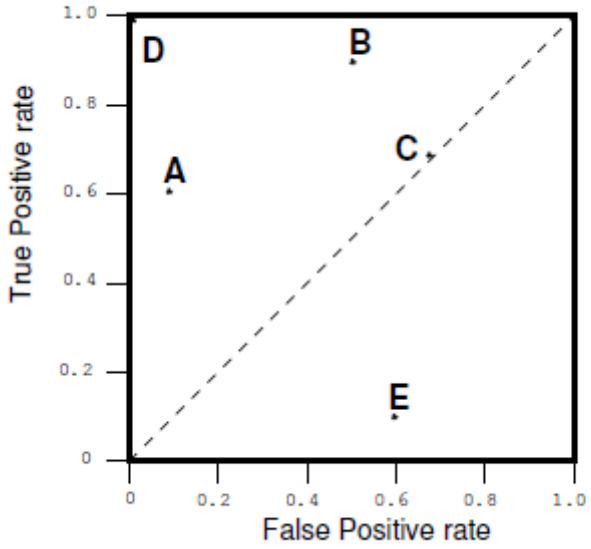
$$acc = \frac{95+94}{95+94+4+7} = 0{,}945$$

$$sensitivity = \frac{95}{95+7} = 0{,}93$$

$$specificity = \frac{94}{94+4} = 0{,}96.$$

5.3 ROC γραφήματα και ερμηνεία

Τα ROC γραφήματα (Receiver Operating Characteristics) είναι δισδιάστατα διαγράμματα στα οποία το TP ποσοστό σχεδιάζεται στον Y άξονα και το FP ποσοστό σχεδιάζεται στον X άξονα. Ένα γράφημα ROC απεικονίζει τη σχετική μεταβολή μεταξύ του κέρδους (αληθώς θετικά) και του κόστους (ψευδώς θετικά). Το Σχήμα 5.1 δείχνει ένα γράφημα ROC με πέντε ταξινομητές χαρακτηρισμένους με A έως E. Ένας διακριτός ταξινομητής είναι αυτός που έχει ως αποτέλεσμα μόνο μια τιμή κλάσης. Κάθε επιμέρους ταξινομητής παράγει ένα (FP rate, TP rate) ζευγάρι, το οποίο αντιστοιχεί σε ένα μόνο σημείο στο χώρο ROC. Οι ταξινομητές στο σχήμα 5.1 είναι όλοι διακριτοί ταξινομητές [2].



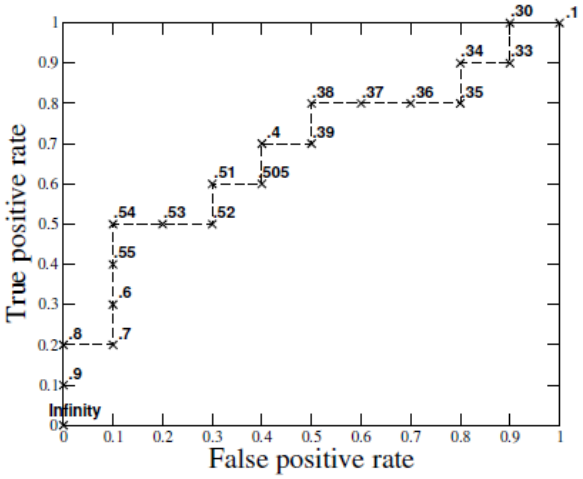
Σχήμα 5.1: Ένα βασικό ROC γράφημα με πέντε διακριτούς ταξινομητές.

Αρκετά σημεία στο χώρο ROC είναι σημαντικό να επισημανθούν. Το κάτω αριστερά σημείο (0,0) αντιπροσωπεύει μια ταξινόμηση η οποία δε δεσμεύει κανένα ψευδώς θετικό αποτέλεσμα αλλά και κανένα αληθώς θετικό. Η αντίθετη στρατηγική, της άνευ όρων έκδοσης θετικών ταξινομήσεων, αντιπροσωπεύεται από το ανώτερο δεξιό σημείο (1,1). Το σημείο (0,1) αντιπροσωπεύει την τέλεια ταξινόμηση. Η απόδοση του σημείου D είναι τέλεια, όπως φαίνεται στο Σχήμα 5.1.

Άτυπα, ένα σημείο στο ROC χώρο είναι καλύτερο από ένα άλλο αν είναι στα βορειοδυτικά (όπου το TP ποσοστό είναι υψηλότερο, το FP ποσοστό είναι χαμηλότερο, ή και τα δύο) του πρώτου. Ταξινομητές που εμφανίζονται στην αριστερή πλευρά του

γραφήματος ROC, κοντά στον άξονα Χ, μπορούν να θεωρηθούν ως «συντηρητικοί»: κάνουν θετικές ταξινομήσεις μόνο με ισχυρές ενδείξεις, ώστε να κάνουν ελάχιστα ψευδώς θετικά λάθη, αλλά συχνά έχουν χαμηλά αληθώς θετικά ποσοστά, επίσης. Ταξινομητές στην επάνω δεξιά πλευρά του ROC γραφήματος μπορεί να θεωρηθούν ως "φιλελεύθεροι": κάνουν θετικές ταξινομήσεις με ασθενή στοιχεία, έτσι ώστε να ταξινομούν σχεδόν όλα τα θετικά σωστά, αλλά συχνά έχουν υψηλά ποσοστά ψευδώς θετικών. Στο Σχήμα 5.1, το Α είναι πιο συντηρητικό από το Β.

Η διαγώνιος y = x αντιπροσωπεύει τη στρατηγική της τυχαίας πρόβλεψης μιας κλάσης. Για παράδειγμα, αν ένας ταξινομητής τυχαία μαντεύει τη θετική κλάση τις μισές φορές, μπορεί να αναμένεται ότι θα ταξινομήσει το ήμισυ των θετικών και το ήμισυ των αρνητικών σωστά. Από αυτά προκύπτει το σημείο (0.5,0.5) στην καμπύλη ROC. Αν αυτός βρίσκει τη θετική κλάση στο 90% του χρόνου, τότε μπορεί να προβλεφθεί ότι θα πάρει το 90% των αληθώς θετικών, αλλά και τα ψευδώς θετικά αποτελέσματα της θα αυξηθούν στο 90%, αποδίδοντας το (0.9,0.9) στο ROC χώρο. Έτσι, ένας τυχαίος ταξινομητής θα παράγει ένα σημείο ROC το οποίο θα "γλιστρά" πότε εμπρός και πότε πίσω στη διαγώνιο με βάση την συχνότητα με την οποία βρήκε τη θετική κλάση. Για να φύγουμε από τη διαγώνιο, μέσα στην πάνω τριγωνική περιοχή, ο ταξινομητής πρέπει να εκμεταλλευτεί κάποιες πληροφορίες όσον αφορά τα δεδομένα. Στο Σχήμα 5.1, η απόδοση του σημείου C είναι σχεδόν τυχαία. Στο (0.7,0.7), το C μπορεί να θεωρηθεί σαν μια πρόβλεψη της θετικής κλάσης στο 70% των περιπτώσεων. Κάθε ταξινομητής που εμφανίζεται στο κατώτερο δεξί τρίγωνο έχει χειρότερη απόδοση από μια τυχαία εικασία. Αυτό το τρίγωνο είναι ως εκ τούτου συνήθως άδειο στα ROC γραφήματα. Ωστόσο, αξίζει να σημειωθεί ότι ο χώρος απόφασης είναι συμμετρικός ως προς τη διαγώνιο που χωρίζει τα δύο τρίγωνα. Αν αναιρέσουμε έναν ταξινομητή- δηλαδή, αν αντιστρέψουμε τις αποφάσεις ταξινόμησης σε κάθε περίπτωση-οι αληθώς θετικές ταξινομήσεις γίνονται ψευδώς θετικές, και τα ψευδώς θετικά αποτελέσματα της γίνονται αληθώς θετικά. Ως εκ τούτου, κάθε ταξινομητής που παράγει ένα σημείο στο κάτω δεξιά τρίγωνο μπορεί να εξαλειφθεί για να παράγει ένα σημείο στο επάνω αριστερό τρίγωνο. Στο Σχήμα 5.1, το σημείο Ε έχει πολύ χειρότερη απόδοση από ό, τι ένα τυχαίο, και στην πραγματικότητα είναι η άρνηση του Α. Δεδομένου ενός γραφήματος ROC στο οποίο η απόδοση ενός ταξινομητή φαίνεται να είναι ελαφρώς καλύτερη από έναν τυχαίο, είναι φυσικό να αναρωτηθεί κανείς αν αυτή η απόδοση του ταξινομητή είναι πραγματικά σημαντική ή μήπως είναι μόνο καλύτερη από την τυχαία. Δεν υπάρχει αποτελεσματικό τεστ για αυτή, αλλά ο Forman (2002) έχει δείξει μια μεθοδολογία η οποία αντιμετωπίζει το ζήτημα αυτό με ROC καμπύλες.



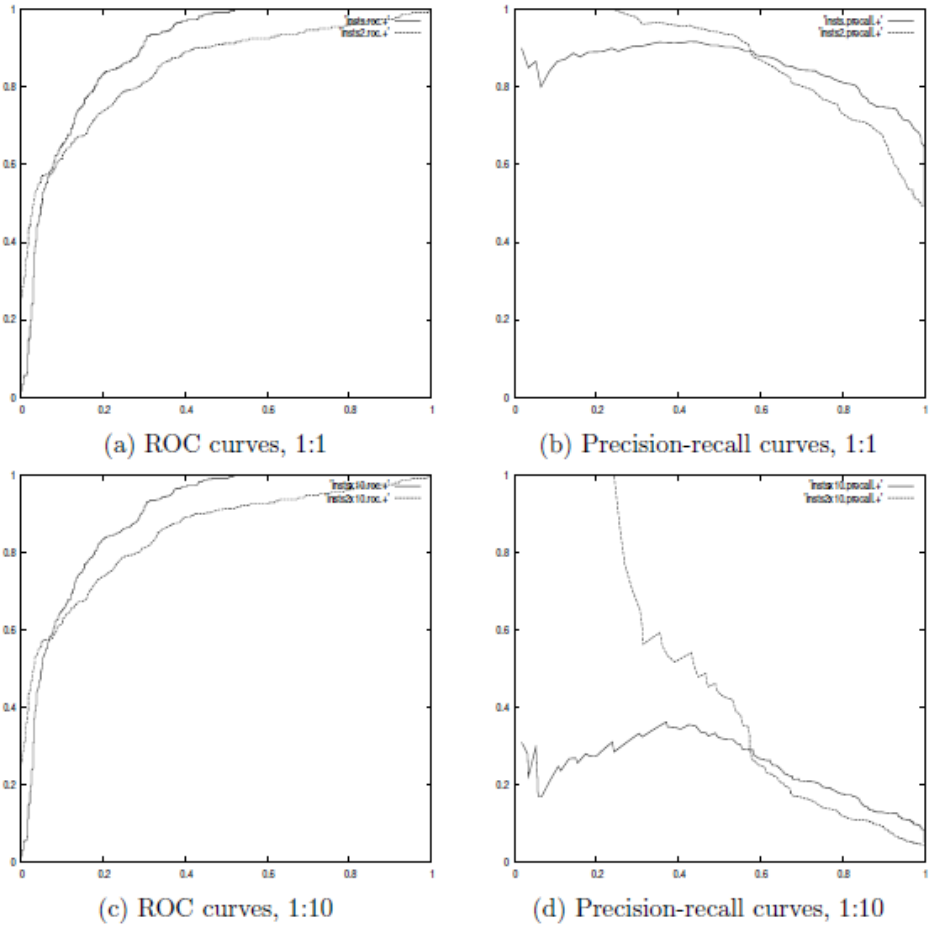
Inst#	Class	Score	Inst#	Class	Score
1	p	.9	11	p	.4
2	p	.8	12	n	.39
3	n	.7	13	p	.38
4	p	.6	14	n	.37
5	p	.55	15	n	.36
6	p	.54	16	n	.35
7	n	.53	17	p	.34
8	n	.52	18	n	.33
9	p	.51	19	p	.30
10	n	.505	20	n	.1

Σχήμα 5.2: Η ROC "καμπύλη" που δημιουργήθηκε από το κατώτατο όριο ενός συνόλου δοκιμών. Ο πίνακας στα δεξιά δείχνει είκοσι δεδομένα και την τιμή που αποδίδεται σε κάθε ένα ταξινομητή βαθμολόγησης. Το γράφημα στα αριστερά δείχνει την αντίστοιχη καμπύλη ROC με κάθε σημείο χαρακτηρισμένη από το όριο που το παράγει.

Το Σχήμα 5.2 δείχνει ένα παράδειγμα μίας ROC καμπύλης σε ένα σύνολο δοκιμής από είκοσι δεδομένα. Τα δεδομένα, δέκα θετικά και δέκα αρνητικά, φαίνονται στον πίνακα δίπλα από το γράφημα. Κάθε καμπύλη ROC που παράγεται από ένα πεπερασμένο σύνολο των περιπτώσεων είναι στην πραγματικότητα μια συνάρτηση ’’ βήμα”(step function), η οποία προσεγγίζει μια αληθινή καμπύλη καθώς ο αριθμός των περιπτώσεων προσεγγίζει το άπειρο. Η συνάρτηση βήμα στο Σχήμα 5.2 λαμβάνεται από ένα πολύ μικρό σύνολο δεδομένων έτσι ώστε η παραγωγή κάθε σημείου να μπορεί να γίνει κατανοητή. Στον πίνακας του Σχήματος 5.2, τα δεδομένα ταξινομούνται βάση τα αποτελέσματα τους και κάθε σημείο στο γράφημα ROC είναι χαρακτηρισμένο από το κατώτερο όριο που το παράγει. Το κατώτερο όριο στο $+\infty$ παράγει το σημείο (0,0). Καθώς χαμηλώνουμε το όριο στο 0.9 το πρώτο θετικό δεδομένο είναι ταξινομημένο θετικά, αποδίδοντας το σημείο (0,0.1). Καθώς το όριο μειώνεται περαιτέρω, η καμπύλη ανεβαίνει πάνω και προς τα δεξιά, καταλήγοντας στο (1,1) με κατώτερο όριο το 0.1. Σημειώνουμε ότι η μείωση του ορίου αυτού αντιστοιχεί στην μετακίνηση από τις «συντηρητικές» στις «φιλελεύθερες» περιοχές του γραφήματος. Αν και το σύνολο της δοκιμής είναι πολύ μικρό, μπορούμε να κάνουμε κάποιες δοκιμαστικές παρατηρήσεις σχετικά με τον ταξινομητή. Ο ταξινομητής φαίνεται να έχει καλύτερες επιδόσεις στην πιο συντηρητική περιοχή του γραφήματος καθώς το ROC σημείο (0.1,0.5) παράγει την υψηλότερη ακρίβεια (70%). Αυτό είναι ισοδύναμο με το να πούμε ότι ο ταξινομητής είναι καλύτερος στον εντοπισμό των πιθανώς θετικών από ό,τι στον εντοπισμό των πιθανώς αρνητικών δεδομένων. Σημειώστε επίσης ότι η καλύτερη ακρίβεια του ταξινομητή εμφανίζεται στο όριο του .54, και όχι στο 0.5, όπως θα μπορούσαμε να αναμένουμε (που αποδίδει 60%).

Συνοψίζοντας, η καμπύλη ROC που αντιστοιχεί είναι *το συνεχές γράφημα που ορίζουν τα σημεία (FP, TP) για όλα τα δυνατά σημεία απόφασης στο μοναδιαίο τετράγωνο [0,1]×[0,1] και ξεκινά από το σημείο (0,0) για να καταλήξει στο σημείο (1,1*). Για μια επαρκή κλίμακα σημείων απόφασης (αρκετά μεγάλη ώστε να περιλαμβάνει όλες μετρήσεις που έχουν καταγραφεί) ορίζονται αντίστοιχοι πίνακες και τα αντίστοιχα σημεία (FP, TP) τα οποία ενώνονται με ευθύγραμμα τμήματα και ορίζουν την καμπύλη ROC.

Σημεία στην άνω αριστερή γωνία του διαγράμματος αποτελούν τιμές κατάλληλες ως σημεία διαχωρισμού, μιας και έχουν χαμηλό αριθμό ψευδώς θετικών περιπτώσεων και υψηλή ευαισθησία. Συνεπώς οι καμπύλες ROC είναι η διαγραμματική απεικόνιση των χαρακτηριστικών ενός ποσοτικού τεστ και μας βοηθούν στην εξέταση της απόδοσης του τεστ για διαφορετικά σημεία προγνωστικού ελέγχου καθώς επίσης και στην επιλογή του σημείου απόφασης όσον αφορά αν ένας έλεγχος θεωρείται θετικός ή αρνητικός.



Καμπύλες ROC και καμπύλες ακρίβειας-ανάκλησης

Οι ROC καμπύλες έχουν μια ελκυστική ιδιότητα: είναι «αναίσθητες» σε αλλαγές στην κατανομή κλάσης. Εάν αλλάξει το ποσοστό από θετικά σε αρνητικά δεδομένα σε ένα σύνολο δοκιμών, οι καμπύλες ROC δεν θα αλλάξουν.

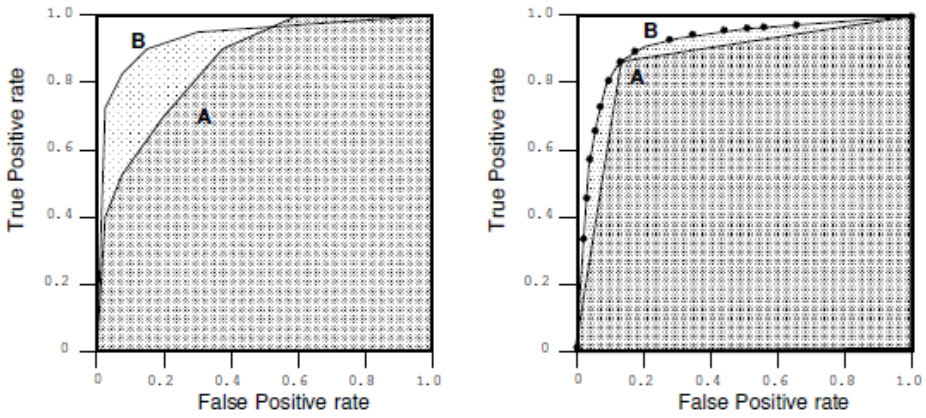
Για να δούμε το αποτέλεσμα της παραποίησης στην κλάση, θεωρούμε τις καμπύλες του παρακάτω Σχήματος, που δείχνουν δύο ταξινομητές αξιολογημένους από τις καμπύλες ROC και τις καμπύλες ακρίβειας. Στα (a) και (b), το σύνολο δοκιμής έχει μία ισορροπημένη 1:1 κλάση κατανομής. Τα γραφήματα (c) και (d) δείχνουν τους ίδιους ταξινομητές στον ίδιο τομέα, αλλά ο αριθμός των αρνητικών δεδομένων έχει αυξηθεί 10 φορές. Σημειώνουμε ότι μόνο η κατανομή κλάσης διαφέρει. Παρατηρούμε ότι τα ROC γραφήματα στα (a) και (c) είναι πανομοιότυπα, ενώ τα γραφήματα ακρίβειας στα (b), (d) διαφέρουν δραματικά.

5.3.1 Η περιοχή κάτω από την ROC καμπύλη (AUC)

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

Η AUC έχει μια σημαντική στατιστική ιδιότητα: η AUC ενός ταξινομητή είναι ισοδύναμη με την πιθανότητα που ο ταξινομητής θα ταξινομήσει ένα τυχαίο επιλεγμένο θετικό παράδειγμα υψηλότερα από ένα τυχαία επιλεγμένο αρνητικό παράδειγμα. Αυτό είναι ισοδύναμο με τη δοκιμή Wilcoxon των βαθμίδων (Hanley & McNeil, 1982). 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 + 1 = 2 x AUC.

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



Σχήμα 5.3: Δύο ROC γραφήματα**.** *Το γράφημα στα αριστερά δείχνει την περιοχή κάτω από δύο καμπύλες ROC. Το γράφημα στα δεξιά δείχνει την περιοχή κάτω από τις καμπύλες του διακριτού ταξινομητή A και του πιθανού ταξινομητή B.*

Είναι δυνατόν για ένα υψηλής-AUC ταξινομητή να έχει χειρότερες επιδόσεις σε μια περιοχή της ROC από ένα χαμηλό-AUC ταξινομητής. Το AUC μπορεί να υπολογιστεί εύκολα χρησιμοποιώντας κατάλληλο αλγόριθμο (ROC Graphs: Notes and Practical Considerations for Data Mining Researchers). Αντί της συλλογής ROC σημείων, ο αλγόριθμος προσθέτει διαδοχικά περιοχές τραπεζοειδών στην Περιοχή.

5.3.2 Επιλογή Βέλτιστου Σημείου Απόφασης με Βάση την Καμπύλη

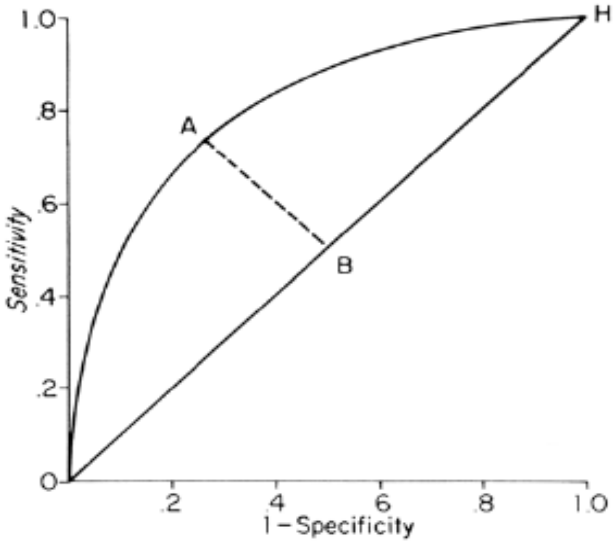
Η συνεχής καμπύλη ROC είναι το γράφημα της συμμεταβολής των ζευγών εσφαλμένα θετικών και ορθώς θετικών αποφάσεων (FP, TP) για όλα τα σημεία απόφασης. Συχνά όμως ο ερευνητής πρέπει να επιλέξει ένα συγκεκριμένο σημείο απόφασης - χρησιμοποιώντας κάποιο κριτήριο για αυτήν την επιλογή- ως λειτουργικό σημείο του διαγνωστικού ελέγχου, σύμφωνα με το οποίο θα κατατάσσει τα δεδομένα με το μικρότερο κόστος (ή με το μικρότερο σφάλμα). Η καμπύλη ROC μπορεί να χρησιμοποιηθεί για την επιλογή του βέλτιστου σημείου απόφασης.

Ανάλογα με τη φύση του διαγνωστικού ελέγχου υπάρχουν τρία κριτήρια επιλογής του σημείου απόφασης.

- Το αυστηρό κριτήριο είναι αυτό που χρησιμοποιείται στον βιομηχανικό ποιοτικό έλεγχο και με βάση αυτό επιλέγεται το σημείο απόφασης που αντιστοιχεί σε FP=0.01 ή 0.05 (δηλαδή επιλογή του σφάλματος τύπου I σύμφωνα με τη μεθοδολογία των στατιστικών ελέγχων υποθέσεων). Σύμφωνα με το αυστηρό κριτήριο μόνο μετρήσεις στα άκρα της κατανομής των φυσιολογικών θα χαρακτηρίζονται ως μη-αποδεκτές.

- Με βάση το επιεικές κριτήριο το σημείο απόφασης επιλέγεται ώστε το κλάσμα TP να είναι κοντά στη μονάδα. Οι συσκευές εντόπισης βλαβών στα αεροσκάφη λειτουργούν με αυτό το κριτήριο.

- Το δημοφιλέστερο κριτήριο σε ότι αφορά την επιλογή του σημείου απόφασης με χρήση της καμπύλης ROC βρίσκεται μεταξύ των δύο προηγούμενων και στηρίζεται στην μεγιστοποίηση των ορθών αποφάσεων (θετικών και αρνητικών). Το σημείο της καμπύλης ROC στο οποίο μεγιστοποιούνται οι ορθές αποφάσεις δηλαδή όπου TN+TP=max (αντίστοιχα ελαχιστοποιούνται οι εσφαλμένες αποφάσεις με FN+FP=min) είναι αυτό με τη μέγιστη απόσταση από την κύρια διαγώνιο. Αυτό αντιστοιχεί στο σημείο απόφασης που ορίζεται από το σημείο τομής των συναρτήσεων πιθανότητας των κατανομών.



Σχήμα 5.4: Βέλτιστο σημείο απόφασης με βάση την καμπύλη ROC.

Συγκεκριμένα, έστω ότι οι συντεταγμένες του σημείου Α στο Σχήμα 4.3 είναι (FP, TP) = (1-TN, TP). Το ευθύγραμμο τμήμα ΟΗ έχει κλίση 1 και ικανοποιεί την y=x. Το ΑΒ είναι κάθετο στο ΟΗ και θα ικανοποιεί την TP=-(1-TN)+λ, όπου λ ο σταθερός όρος. Λύνοντας ως προς λ έχουμε, λ=TP-TN+1. Το Β ανήκει και στις δύο ευθείες οπότε οι συντεταγμένες του θα είναι:

$$\left(\frac{TP+1-TN}{2}, \frac{FP+1-TN}{2}\right)$$

Άρα το μήκος του ΑΒ:

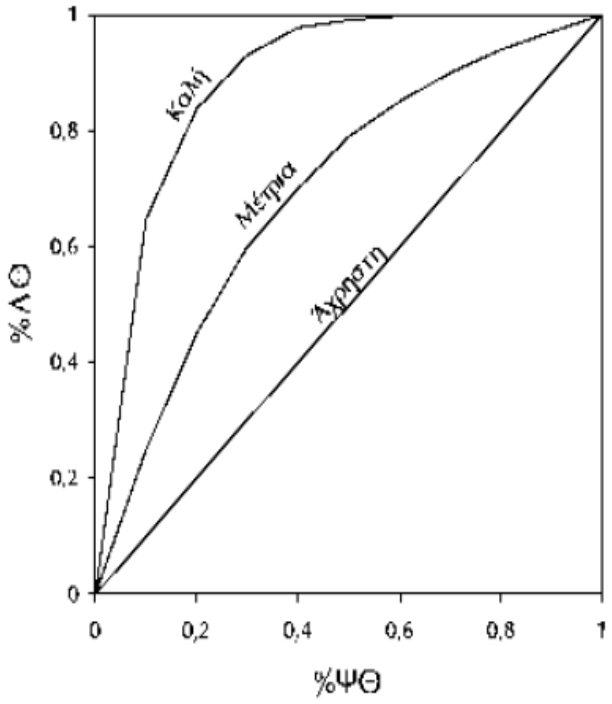
$$\sqrt{\left[(1-TN)-\frac{TP+1-TN}{2}\right]^2+\left[TP-\frac{TP+1-TN}{2}\right]^2}=\frac{\sqrt{2}}{2}(TP+TN-1)$$

γίνεται μέγιστο όταν η ποσότητα TN+TP είναι μέγιστη.

5.3.3 Σύγκριση Διαγνωστικών Δοκιμασιών

Με τη βοήθεια των καμπυλών ROC είναι δυνατή η οπτική και ποσοτική σύγκριση, τόσο της συνολικής (ανεξάρτητα από το επιλεγμένο όριο) διακριτικής ικανότητας δύο ή περισσοτέρων δοκιμασιών που χρησιμοποιούνται για τη διάγνωση του ίδιου αποτελέσματος, όσο και της διαγνωστικής ποιότητας που αντιστοιχεί σε ένα συγκεκριμένο –στο επιλεγμένο- διαχωριστικό όριο κάθε δοκιμασίας.

Η δοκιμασία της οποίας η καμπύλη ROC εμφανίζει μεγαλύτερη κυρτότητα προς την άνω αριστερή γωνία (οπτική εκτίμηση) και επομένως η περιοχή (εμβαδόν) κάτω από αυτή είναι μεγαλύτερη (ποσοτική εκτίμηση), εμφανίζει και τη μεγαλύτερη συνολική διακριτική ικανότητα (Σχήμα 5.4).



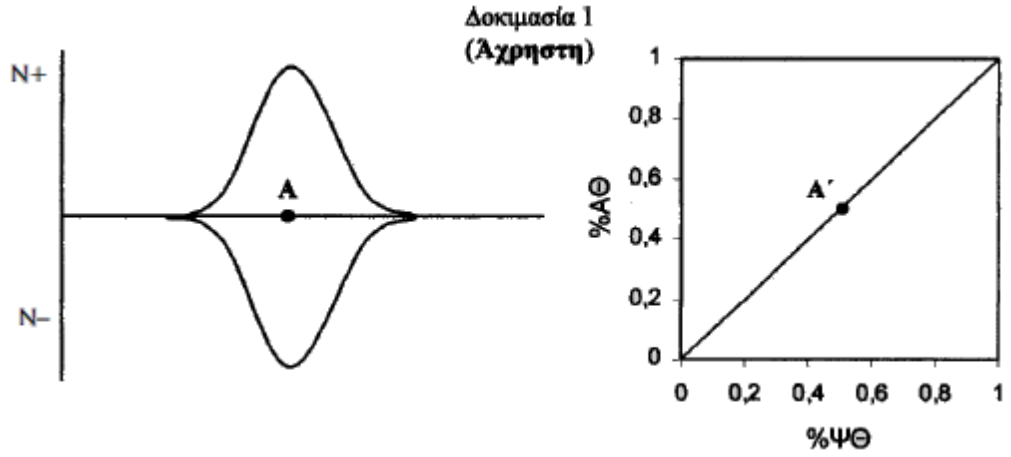
Σχήμα 5.5: Συγκριτική αξιολόγηση των δοκιμασιών βάσει των καμπύλων ROC

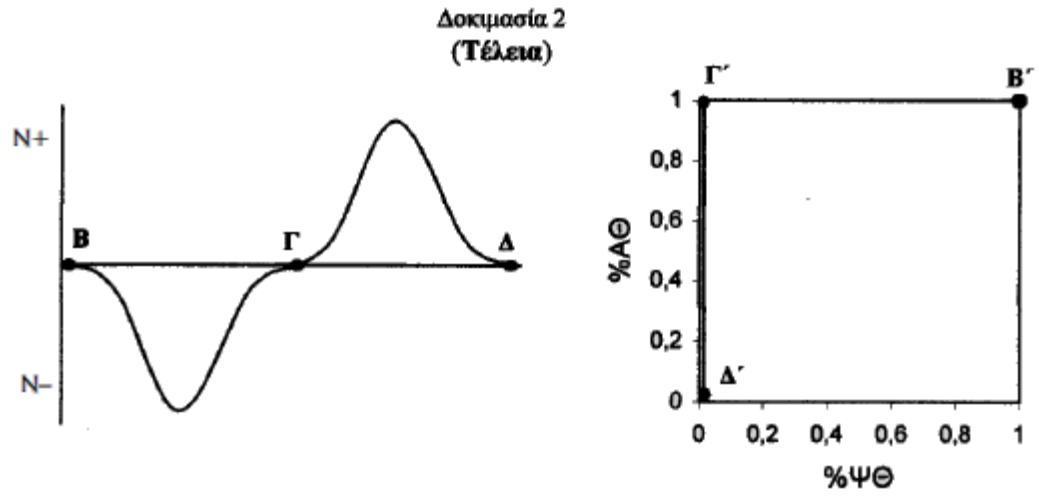
5.3.4 Εκτίμηση της Διακριτικής Ικανότητας μίας Δοκιμασίας

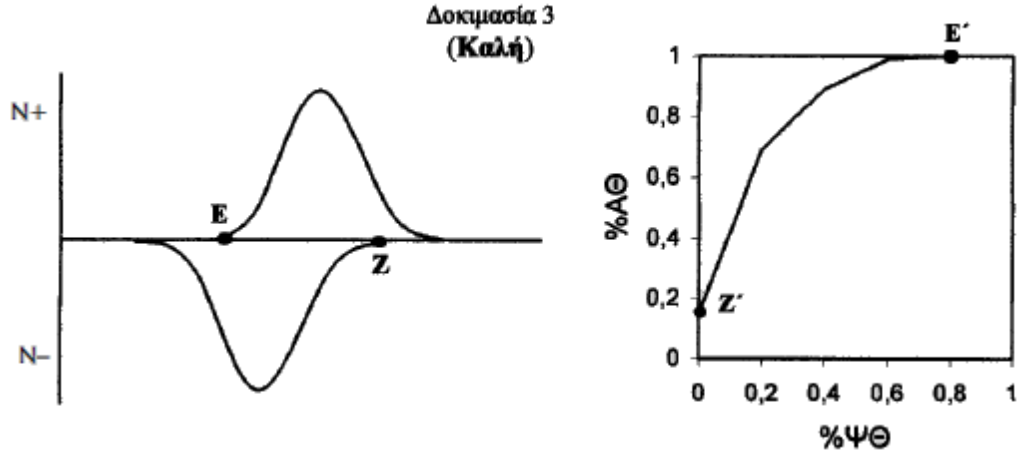
Το σχήμα και η θέση της καμπύλης ROC παρέχουν πληροφορίες σχετικά με τη διακριτική ικανότητα μίας δοκιμασίας (οπτική εκτίμηση). Όσο μεγαλύτερη είναι η κυρτότητα της καμπύλης ROC προς την αριστερή άνω γωνία του τετραγώνου, τόσο μεγαλύτερη είναι η διακριτική ικανότητα. Όταν η καμπύλη βρίσκεται κοντά στη διαγώνιο, η δοκιμασία έχει μικρή ή καμιά διακριτική ισχύ, ενώ η καμπύλη ROC που

βρίσκεται κάτω από τη διαγώνιο αφορά σε ευρήματα, των οποίων η απουσία και όχι η παρουσία συνδέεται προς το δεδομένο που εξετάζουμε.

Το εμβαδόν της περιοχής κάτω από την καμπύλη ROC αποτελεί ένα ενδεικτικό μέτρο του συνολικού πληροφοριακού περιεχομένου της διαγνωστικής δοκιμασίας ανεξάρτητα από το επιλεγμένο διαχωριστικό όριο (ποσοτική εκτίμηση). Το εμβαδόν της περιοχής κάτω από την καμπύλη ROC εκφράζει την πιθανότητα που υφίσταται να ταξινομήσει η δοκιμασία ορθά ένα τυχαίο ζεύγος ενός δεδομένου και λαμβάνει τιμές από 0-1. Μια περιοχή ίση με 1 (εμβαδόν του τετραγώνου) αντιπροσωπεύει μια «τέλεια» δοκιμασία και, αντίθετα μα περιοχή ίση με 0,5 μια «άχρηστη» δοκιμασία. Η συνολική διακριτική ικανότητα της δοκιμασίας είναι τόσο μεγαλύτερη, όσο η περιοχή κάτω από την καμπύλη είναι μεγαλύτερη από 0,5 και πλησιάζει το 1.







Σχήμα 5.6: Κατανομές συχνότητας των αποτελεσμάτων τριών δοκιμασιών με διαφορετική διακριτική ικανότητα και οι αντίστοιχες καμπύλες

5.4 Προβλήματα με περισσότερες από δύο κλάσεις

Μέχρι αυτό το σημείο έχουμε ασχοληθεί με προβλήματα στα οποία υπάρχουν μόνο δύο κλάσεις. Οι δύο άξονες σε μία καμπύλη ROC αντιπροσωπεύουν τα "tradeoffs" μεταξύ των σφαλμάτων (ψευδώς θετικά) και των σωστών (αληθώς θετικά) που ένας ταξινομητής διαπράττει ανάμεσα σε δύο τάξεις. Το μεγαλύτερο μέρος της ανάλυσης είναι απλό, λόγω της συμμετρίας που υπάρχει σε ένα πρόβλημα δύο κλάσεων. Η προκύπτουσα απόδοση μπορεί εύκολα να απεικονιστεί σε ένα δισδιάστατο χώρο.

5.4.1 ROC Γραφήματα και AUC Πολλών Κλάσεων

Με περισσότερες από δύο κλάσεις, η κατάσταση γίνεται πολύ πιο περίπλοκη. Με n κλάσεις ο πίνακας σύγχυσης γίνεται ένας nxn πίνακας που περιέχει τις n σωστές ταξινομήσεις (the major diagonal entries) και $n^2 - n$ πιθανά σφάλματα (the off-diagonal entries). Αντί να διαχειριζόμαστε τα trade-offs μεταξύ των TP και FP, έχουμε n οφέλη και $n^2 - n$ σφάλματα. Με τρεις μόνο κλάσεις, η επιφάνεια γίνεται $3^2 - 3 = 6$-διαστάσεων πολύτοπο. Ο Lane (2000) έγραψε ένα σύντομο έγγραφο στο οποίο δίνει έμφαση σε σχετικά ζητήματα και τις προοπτικές για την αντιμετώπισή τους.

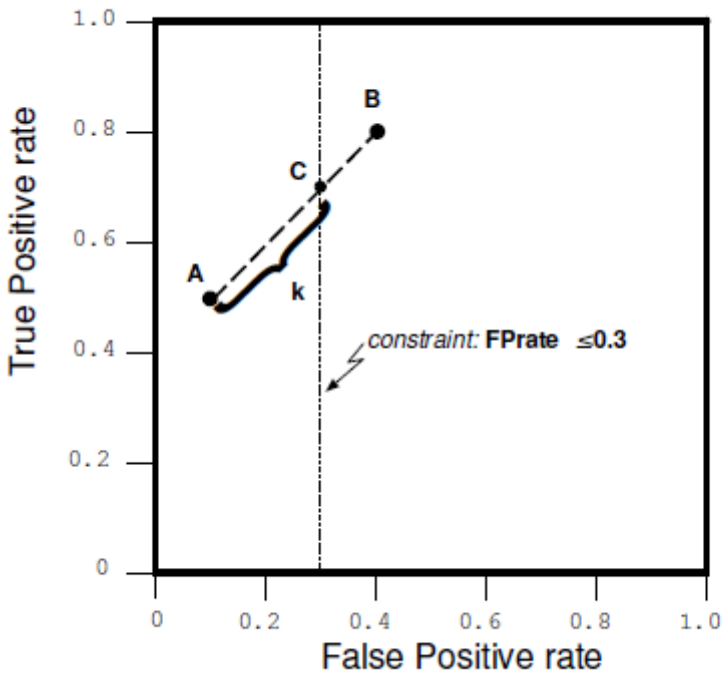
Μία μέθοδος για να διαχειριστούμε n κλάσεις είναι να παράγουμε n διαφορετικά ROC διαγράμματα, ένα για κάθε κλάση. Αυτό ονομάζεται η διατύπωση *κλάσης αναφοράς* (the class reference formulation). Ειδικότερα, αν C είναι το σύνολο όλων των κλάσεων, το ROC γράφημα i σχεδιάζεται έτσι ώστε η ταξινόμηση να παριστάνει την κλάση c_i θετική και όλες τις άλλες κλάσεις αρνητικές, δηλαδή,

$$P_i = c_i$$

$$N_i = \bigcup_{j \neq i} c_j \in C$$

Ενώ αυτό είναι μια βολική διαμόρφωση, θέτει σε κίνδυνο ένα από τα πλεονεκτήματα των ROC γραφημάτων, το ότι παραμένουν ανεπηρέαστα σε αλλαγές στις κατανομές. Επειδή κάθε Ni περιλαμβάνει την ένωση των $n-1$ κλάσεων, οι αλλαγές στον επιπολασμό εντός αυτών των κλάσεων μπορούν να τροποποιήσουν ROC γράφημα c_i. Για παράδειγμα, ας υποθέσουμε ότι κάποια κλάση $c_k \in N$ είναι ιδιαίτερα εύκολο να εντοπιστεί. Ένας ταξινομητής για την κλάση $c_i, i \neq k$ μπορεί να "εκμεταλλευτεί" κάποια χαρακτηριστικά της c_k, ώστε να παράγει χαμηλά όρια για τα δεδομένα της c_k. Η αύξηση του επιπολασμού της c_k μπορεί να αλλάξει την απόδοση του ταξινομητή, και αυτό θα ισοδυναμούσε στην αλλαγή του στόχου. Με αυτόν τον τρόπο θα τροποποιηθεί-μεταβληθεί η ROC καμπύλη. Ωστόσο, λαμβάνοντας υπόψη αυτή την προειδοποίηση, η μέθοδος αυτή μπορεί να λειτουργήσει καλά στην πράξη και να παρέχει ευελιξία στην αξιολόγηση.

Η AUC είναι ένα μέτρο της διάκρισης ενός ζεύγους κλάσεων. Σε ένα πρόβλημα δύο κλάσεων, η AUC έχει μια ενιαία βαθμωτή τιμή, αλλά σε ένα πρόβλημα πολλών κλάσεων δημιουργείται το πρόβλημα του συνδυασμού τιμών πολλαπλών διακεκριμένων ζευγών(Hand and Till's-2001).



Παρεμβολή ταξινομητών

Οι Hand and Till's ασχολήθηκαν με ένα μέτρο που δεν επηρεάζεται από την κατανομή της κλάσης και το σφάλμα του κόστους. Η διαδικασία είναι αρκετά λεπτομερής για να τη συνοψίσουμε εδώ, αλλά βασίζεται στο γεγονός ότι η AUC είναι ισοδύναμη με την πιθανότητα ότι ο ταξινομητής θα ταξινομήσει ένα τυχαία επιλεγμένο θετικό παράδειγμα υψηλότερα από ένα τυχαία επιλεγμένο αρνητικό παράδειγμα. Από αυτή την πιθανολογική μορφή, παράγεται ένας τύπος που μετρά τις διακεκριμένες κλάσεις ενός μη σταθμισμένου απλού ζεύγους. Το μέτρο τους, το οποίο καλείται M, είναι ίσο με:

$$AUC_{total} = \frac{2}{|C|(|C|-1)} \sum_{\{c_i, c_j\} \in C} AUC(c_i, c_j)$$

όπου n είναι ο αριθμός των κλάσεων και $AUC(c_i, c_j)$ είναι η περιοχή κάτω από την ROC καμπύλη των δύο κλάσεων c_i, c_j.

Το άθροισμα υπολογίζεται για όλα τα ζεύγη των διακεκριμένων κλάσεων, χωρίς να λαμβάνουμε υπόψη τη σειρά αυτών. Υπάρχουν $|C|(|C|-1)/2$ τέτοια ζευγάρια. Παρόλο που το μέτρο αυτό είναι αποδοτικό, δεν υπάρχει εύκολος τρόπος για να παραστήσουμε την επιφάνεια που υπολογίζουμε.

5.5 Διασταυρωμένη Επικύρωση(Cross Validation)

Η Διασταυρωμένη Επικύρωση είναι μια μέθοδος αξιολόγησης του μοντέλου που είναι καλύτερη από τα υπόλοιπα (residuals). Η μέθοδος εκτιμάει το σφάλμα πρόβλεψης (prediction error). Το πρόβλημα με την αξιολόγηση μέσω υπολοίπων είναι ότι δεν δίνουν μια ένδειξη του πόσο καλή απόδοση θα έχει ο «εκπαιδευόμενος» (ταξινομητής) όταν θα κληθεί να προβεί σε νέες προβλέψεις για δεδομένα που δεν έχει ήδη δει. Ένας τρόπος για να ξεπεραστεί αυτό το πρόβλημα είναι να μη χρησιμοποιήσει ο ταξινομητής όλο το σύνολο των διαθέσιμων δεδομένων. Μερικά από τα στοιχεία που περιέχει το σύνολο θα αφαιρεθούν πριν από την έναρξη της εκπαίδευσης. Στη συνέχεια, όταν η εκπαίδευση ολοκληρωθεί τα δεδομένα που αφαιρέθηκαν θα χρησιμοποιηθούν για τον έλεγχο της απόδοσης του μοντέλου. Αυτή είναι η βασική ιδέα για μια ολόκληρη κατηγορία των μεθόδων αξιολόγησης μοντέλων που ονομάζεται διασταυρωμένη επικύρωση (*cross validation*).

Πιο συγκεκριμένα, αρχικά χωρίζουμε το σύνολο των δεδομένων σε N τμήματα(folds) με N<<l έτσι ώστε:

$$D = Q_1 \cup Q_2 \cup \ldots \cup Q_{N-1} \cup Q_N$$

και

$$Q_i \cap Q_j = \emptyset$$

$$|Q_i| = |Q_j| = \frac{l}{N},$$

για $i, j = 1, \dots N$ και $i \neq j$.

Θα χρησιμοποιήσουμε κάθε τμήμα (fold) για δοκιμή ακριβώς μια φορά, και τα υπόλοιπα θα τα χρησιμοποιήσουμε για να εκπαιδεύσουμε το μοντέλο. Ας είναι Q_i ένα τμήμα του συνόλου δεδομένων D, τότε μπορούμε να κατασκευάσουμε το αντίστοιχο σύνολο εκπαίδευσης, ως εξής:

$$P_i = D - Q_i$$

για $i, j = 1, \dots N$. Μπορούμε να υπολογίσουμε το σφάλμα κάποιου τυχαίου τμήματος Q_i ως εξής:

$$err_{Q_i}\left[\widehat{f_{P_i}}[k, \lambda, C]\right] = \frac{1}{|Q_i|} \sum_{(x_j, y_j)} L\left[y_i, \widehat{f_{P_i}}[k, \lambda, C](\overline{x_j})\right]$$

όπου $\widehat{f_{P_i}}[k, \lambda, C]$ είναι το μοντέλο εκπαίδευσης του συνόλου δεδομένων Pi με παραμέτρους k, λ, και C. Στη διασταυρωμένη επικύρωση κάθε τμήμα έχει περισσότερα από ένα στοιχεία, ως εκ τούτου, το σφάλμα υπολογίζεται ως η μέση απώλεια πάνω σε αυτό το τμήμα.
Μπορούμε να υπολογίσουμε το σφάλμα της διασταυρωμένης επικύρωσης (CVE) με τη βοήθεια του παρακάτω τύπου:

$$CVE_D[k, \lambda, C] = \frac{1}{N} \sum_{i=1}^{N} \widehat{f_{P_i}}[k, \lambda, C]$$

Επομένως, το CVE υπολογίζεται πάνω στο σύνολο των παραμέτρων k,λ,C και τα ίδια τα μοντέλα αποτελούν μέρος του σφάλματος υπολογισμού. Για την εύρεση του βέλτιστου συνόλου των παραμέτρων πρέπει να ελαχιστοποιήσουμε το CVE:

$$(k^*, \lambda^*, C^*) = argmin_{k,\lambda,C} CVE_D[k, \lambda, C]$$

και το βέλτιστο μοντέλο $\widehat{f_{P_i}}[k^*, \lambda^*, C^*]$ μπορεί να κατασκευαστεί χρησιμοποιώντας όλα τα δεδομένα του συνόλου D.

Έχει δειχθεί ότι η διασταυρωμένη επικύρωση με τιμές για το N:3, 5 και 10 είναι πολύ αποτελεσματική.

ΚΕΦΑΛΑΙΟ 6

ΕΠΙΛΟΓΗ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ ΜΕ ΜΗΧΑΝΕΣ ΔΙΑΝΥΣΜΑΤΙΚΗΣ ΥΠΟΣΤΗΡΙΞΗΣ

6.1 Εισαγωγή

Η επιλογή χαρακτηριστικών είναι ένα σημαντικό πρόβλημα της εξόρυξης δεδομένων. Σε αυτό το κεφάλαιο, αφού παρουσιάσουμε το πρόβλημα, θα αναφερθούμε στην αναγκαιότητα επίλυσης του. Στη συνέχεια, θα περιγράψουμε εν συντομία, δύο μεθόδους επιλογής χαρακτηριστικών (filter, wrapper), οι οποίες μας εξασφαλίζουν σπουδαία οφέλη κατά την κατασκευή μοντέλων πρόβλεψης.

6.1.1 Το πρόβλημα της επιλογής χαρακτηριστικών

Με τον όρο επιλογή χαρακτηριστικών αναφερόμαστε στην εξεύρεση ενός υποσυνόλου των χαρακτηριστικών που είναι πιο σημαντικά στην ταξινόμηση. Οι Μηχανές Διανυσματικής Υποστήριξης (SVMs) είναι μια νέα μέθοδος για την εξαγωγή πληροφοριών από ένα σύνολο δεδομένων. Οι Μηχανές Διανυσματικής Υποστήριξης έχουν αρκετές εφαρμογές όπως στη βιοπληροφορική, την αναγνώριση προσώπου, την κατηγοριοποίηση κειμένων και ούτω καθεξής. Οι ερευνητές έχουν αρχίσει πρόσφατα να χρησιμοποιούν τις SVMs για την επιλογή χαρακτηριστικών.

Με δεδομένο ένα παράδειγμα (instance) $x = (x_1, \dots, x_n)$ θεωρούμε ένα πρόβλημα ταξινόμησης. Συνήθως, μόνο ένας μικρός αριθμός από τα χαρακτηριστικά του x δίνει σαφείς πληροφορίες για την ταξινόμηση. Το πρόβλημα επιλογής χαρακτηριστικών έγκειται στην αναγνώριση του μικρού υποσυνόλου των χαρακτηριστικών που είναι σχετικά με τον στόχο μας (απόκριση). Ένα μικρό υποσύνολο σχετικών χαρακτηριστικών έχει περισσότερη ισχύ από τη χρήση περισσότερων χαρακτηριστικών. Αυτό είναι αντι-διαισθητικό, δεδομένου ότι περισσότερα χαρακτηριστικά θα έπρεπε να δώσουν περισσότερες πληροφορίες και ως εκ τούτου θα έπρεπε να δίνουν μεγαλύτερη διακριτική ισχύ. Αλλά αν ένα χαρακτηριστικό είναι άσχετο (δε σχετίζεται με τα υπόλοιπα), αυτό το χαρακτηριστικό δεν έχει επίδραση στο στόχο μας (απόκριση). Επίσης, εάν ένα χαρακτηριστικό είναι περιττό, τότε αυτό δεν προσθέτει τίποτα νέο. Τα οφέλη της επιλογής χαρακτηριστικών είναι:

- Η μείωση του χρόνου υπολογισμού
- Η εύρεση ενός καλύτερου υπερεπίπεδου και

- Η καλύτερη κατανόηση των δεδομένων.

Σε ένα πείραμα, για παράδειγμα, που παρουσιάζεται στην εργασία «Feature Seletion in Support Vector Machines», Eun Seog Youn(2002),έχουμε ένα σύνολο δεδομένων για τον καρκίνο του παχέος εντέρου. Κάθε παράδειγμα αποτελείται από 2.000 συστατικά (επίπεδα έκφρασης γονιδίων). Οι ερευνητές πιστεύουν ακράδαντα ότι μόνο ένα μικρό υποσύνολο των γονιδίων είναι υπεύθυνα για τον καρκίνο του παχέος εντέρου. Πειραματικά αποτελέσματα και άλλες ερευνητικές μελέτες το υποστηρίζουν αυτό.

6.2 Μέθοδοι επιλογής χαρακτηριστικών

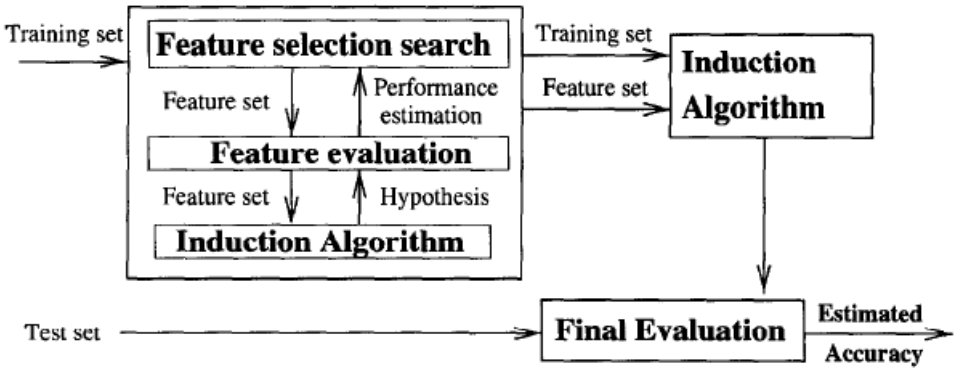
Ένας αλγόριθμος επιλογής χαρακτηριστικών μπορεί να θεωρηθεί ως ο συνδυασμός μιας τεχνικής αναζήτησης για την πρόταση νέων υποσυνόλων χαρακτηριστικών, μαζί με ένα μέτρο αξιολόγησης που πετυχαίνει τα διαφορετικά υποσύνολα χαρακτηριστικών. Ο απλούστερος αλγόριθμος είναι να δοκιμάσουμε κάθε δυνατό υποσύνολο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

1. **Wrappers**

Οι προσεγγίσεις περιτυλιγμάτων (wrappers) είναι διαδικασίες ευρείας αναζήτησης που αξιολογούν την ποιότητα της λειτουργίας του υποσύνολο χρησιμοποιώντας την ακρίβεια πρόβλεψης του συστήματος μάθησης. Περιλαμβάνουν τεχνικές όπως τη διαδοχική προς τα εμπρός και προς τα πίσω επιλογή χαρακτηριστικών, τις παραλλαγές των «ορειβατών του λόφου» (hill climbers), αναζήτηση «πρώτα το καλύτερο» (best-first), την ακτινική αναζήτηση και τους τυχαίους αλγορίθμους όπως τη προσομοίωση ανόπτησης (Simulated Annealing) και γενετικούς αλγόριθμους (Genetic).

Οι wrappers συχνά δίνουν τα καλύτερα αποτελέσματα (όσον αφορά την τελική διαγνωστική ακρίβεια του αλγόριθμου) σε σχέση με τα filters, διότι η επιλογή χαρακτηριστικού είναι βελτιστοποιημένη για το συγκεκριμένο αλγόριθμο. Ωστόσο, εφόσον ένας μαθησιακός αλγόριθμος χρησιμοποιήθηκε για την αξιολόγηση κάθε μίας του συνόλου των χαρακτηριστικών, τα wrappers είναι απαγορευτικά δαπανηρά, και μπορεί να είναι δυσεπίλυτο για μεγάλες βάσεις δεδομένων να περιέχουν πολλά χαρακτηριστικά. Επιπλέον, δεδομένου ότι η επιλογή χαρακτηριστικών είναι μια διαδικασία στενά συνδεδεμένη με τον αλγόριθμο μάθησης, τα wrappers είναι λιγότερο

γενικά από τα filters και η όλη διαδικασία πρέπει να επαναληφθεί στην εναλλαγή από τον ένα αλγόριθμο μάθησης στο άλλο.



Σχήμα 6.1: Η μέθοδος περιτυλίγματος (wrapper) για την επιλογή υποσυνόλου χαρακτηριστικών. Ο αλγόριθμος επαγωγής χρησιμοποιείται ως ένα «μαύρο κουτί» (καμία γνώση του αλγόριθμου δεν είναι απαραίτητη).

2. Filters

Το filter προσέγγισης αξιολογεί τα ανεξάρτητα χαρακτηριστικά των ταξινομητών και προσπαθεί να αφαιρέσει τα άσχετα χαρακτηριστικά από το σύνολο προτού χρησιμοποιηθούν από τον αλγόριθμο μάθησης. Τα παραδείγματα των χαρακτηριστικών μέτρων αξιολόγησης είναι εγγενείς ιδιότητες των στοιχείων- πιθανά μέτρα απόστασης, πιθανά μέτρα εξάρτησης, εσωτερικά μέτρα απόστασης, μέτρα θεωρητικών πληροφοριών όπως εντροπία κλπ. Συστήματα όπως το FOCUS, η σταυρωτή εντροπία filter και η RELIEF και οι παραλλαγές της, το δέντρο αποφάσεων filter είναι μερικά από τα γνωστά filter συστήματα.

Τα μέτρα αυτά αντιλαμβάνονται τη σχέση του χαρακτηριστικού με τον στόχο. Οι προσεγγίσεις των filters είναι υπολογιστικά λιγότερο δαπανηρές και πιο γενικές, αλλά επιστρέφουν ένα μεγάλο υποσύνολο χαρακτηριστικών. Επίσης, μερικοί filters αλγόριθμοι που περιγράφηκαν προηγουμένως δεν χειρίζονται το θόρυβο στα δεδομένα (Focus), ενώ άλλοι απαιτούν το επίπεδο του θορύβου να καθορίζεται κατά προσέγγιση από τον χρήστη εκ των προτέρων.

Σχήμα 6.2: Η μέθοδος του φίλτρου (filter). Τα χαρακτηριστικά φιλτράρονται ανεξαρτήτως του αλγορίθμου επαγωγής.

Μια άλλη αξιοσημείωτη παρατήρηση για αυτά τα έργα είναι ότι δεν υπάρχει αλγόριθμος που λειτουργεί ιδανικά σε όλους τους τομείς, όπως φαίνεται από την διακύμανση των πειραματικών αποτελεσμάτων. Αυτό είναι κατανοητό διότι η επιλογή των χαρακτηριστικών είναι μια εξαιρετικά συγκεκριμένη εργασία. Η εξεύρεση του βέλτιστου συνόλου χαρακτηριστικών είναι συνήθως δυσεπίλυτη, καθώς και πολλά προβλήματα που σχετίζονται με την επιλογή χαρακτηριστικών έχουν αποδειχθεί ότι είναι δύσκολα. Για περισσότερο πρακτικά προβλήματα, η βέλτιστη λύση μπορεί να διασφαλιστεί μόνο αν ένα κριτήριο για την αξιολόγηση των χαρακτηριστικών βρεθεί, αλλά αυτή η υπόθεση, είναι σπάνια στον πραγματικό κόσμο. Ως εκ τούτου, είμαστε αναγκασμένοι να βρούμε λύσεις που είναι μεταξύ της ποιότητας (γενίκευση WRT, προγνωστική ακρίβεια) και του χρόνου.

ΚΕΦΑΛΑΙΟ 7

ΕΦΑΡΜΟΓΕΣ

7.1 Εισαγωγή

Στο παρόν κεφάλαιο θα αξιολογήσουμε πειραματικά την εφαρμογή των μεθόδων Διανυσματικής Υποστήριξης σε προβλήματα ταξινόμησης δύο και τριών κλάσεων καθώς επίσης και σε ένα πρόβλημα παλινδρόμησης. 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.

Για τα προβλήματα με Μηχανές Διανυσματικής Ταξινόμησης (γραμμικά ή μη), ένας ταξινομητής θα πρέπει να παρέχει υψηλές τιμές των ACC, sensitivity, specificity και της AUROC, και η γενικευμένη απόδοση συχνά εκτιμάται με holdout επικύρωση (εκπαίδευση/ δοκιμή). Όσο αφορά στα προβλήματα παλινδρόμησης με Μηχανές Διανυσματικής Ταξινόμησης, ο ταξινομητής θα πρέπει να παρέχει χαμηλές τιμές του μέσου τετραγωνικού σφάλματος (MSE) ή της τετραγωνικής ρίζας αυτού (RMSE).

7.2 Εισαγωγή στην R

<u>Το πακέτο R</u>

- Το R είναι ένα *υπολογιστικό πακέτο* που προσφέρει δυνατότητες διαχείρισης και *στατιστικής ανάλυσης δεδομένων* καθώς και δυνατότητες κατασκευής γραφημάτων
- Βασίζεται στην *γλώσσα προγραμματισμού* S (που χρησιμοποιεί και το στατιστικό πακέτο S plus) και πρόκειται για *λογισμικό ανοικτού κώδικα* (open source) που διατίθεται ελεύθερα.
- Μπορεί να χρησιμοποιηθεί είτε με κατευθείαν εντολές που υπάρχουν είτε με προγράμματα που ο χρήστης μπορεί να προγραμματίσει για επίλυση πιο πολύπλοκων στατιστικών προβλημάτων.
- Στις συγκεκριμένες σημειώσεις χρησιμοποιούμε την έκδοση 3.1.0
- Για μια πιο αναλυτική εισαγωγή παραπέμπουμε στο αρχείο R-intro.pdf (An introduction to R, W.N. Venables, D.M. Smith and the R development core team) που παρέχεται στα help files του R.

7.3 Βασικά Βήματα

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

I. Αρχικά εισάγουμε τα δεδομένα στο πρόγραμμα μέσω ενός txt αρχείου.

II. Καθορίζουμε τον τύπο των δεδομένων (factor, data frame).

III. Χωρίζουμε τα δεδομένα σε:

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

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

Στα προβλήματα που ακλουθούν, χωρίσαμε, όπως τα δεδομένα σε *σύνολο εκπαίδευσης* που αποτελείται από το 75% των περιπτώσεων και *σύνολο δοκιμής* που αποτελείται από το 25% των περιπτώσεων.

7.3.1 Πρώτη εφαρμογή

Τα δεδομένα που κατεβάσαμε (Bank Marketing Data Set) προέρχονται από την τοποθεσία:

"[Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014"

Προτείνεται μια προσέγγιση εξόρυξης δεδομένων (Data Mining) για να προβλέψει την επιτυχία των κλήσεων τηλεμάρκετινγκ για την πώληση τραπεζικών μακροπρόθεσμων καταθέσεων. Τα δεδομένα έχουν συλλεχθεί από μία Πορτογαλική τράπεζα λιανικής, από το Μάιο του 2008 έως το Νοέμβριο του 2010. Συχνά, χρειάστηκαν περισσότερες από μία επαφές με τον πελάτη. Υπάρχουν 4 σύνολα δεδομένων για το σκοπό αυτό. Εμείς θα χρησιμοποιήσουμε το μικρότερο σύνολο, που ενδείκνυται για δοκιμές απαιτητικών αλγόριθμων μηχανικής μάθησης, όπως SVM. **Σκοπός** μας είναι να προβλέψουμε αν ο πελάτης θα εγγραφεί σε μία προθεσμιακή κατάθεση. Προφανώς, λοιπόν, έχουμε ένα πρόβλημα ταξινόμησης δύο κλάσεων.

Το σύνολο δεδομένων που χρησιμοποιήσαμε έχει 4521 καταγραφές (instances), με 16 επεξηγηματικές μεταβλητές και 1 μεταβλητή απόκρισης (response). Έχουμε, λοιπόν, τον παρακάτω πίνακα δεδομένων:

Πίνακας Δεδομένων

ΠΡΟΣΩΠΙΚΑ ΔΕΔΟΜΕΝΑ	
X1	Ηλικία
X2	Επάγγελμα
X3	Οικογενειακή κατάσταση
X4	Μόρφωση
X5	Εάν έχει άλλη πίστωση που εκκρεμεί
X6	Μέσο ετήσιο ποσό όλων των τρεχούμενων λογαριασμών του πελάτη
X7	Στεγαστικό δάνειο
X8	Προσωπικό δάνειο
ΕΠΑΦΗ ΜΕ ΤΟΝ ΠΕΛΑΤΗ	
X9	Τρόπος επικοινωνίας
X10	Ημέρα που έγινε η τελευταία επαφή
X11	Μήνας που έγινε η τελευταία επαφή
X12	Διάρκεια της τελευταίας επαφής
ΙΣΤΟΡΙΚΟ	
X13	Ο αριθμός των επαφών που πραγματοποιήθηκαν για κάθε πελάτη για αυτή την εκστρατεία

X14	Ο αριθμός των ημερών που πέρασαν από την τελευταία επαφή για οποιαδήποτε άλλη εκστρατεία
X15	Συνολικός αριθμός προηγούμενων επαφών πριν από αυτή την εκστρατεία
X16	Αποτέλεσμα της προηγούμενης εκστρατείας
ΜΕΤΑΒΛΗΤΗ ΑΠΟΚΡΙΣΗΣ	
Y	Εάν ο πελάτης θα εγγραφεί στην προθεσμιακή κατάθεση που του προτείνεται

Όπως αναφέραμε προηγουμένως, πρόκειται για ένα σύνολο δεδομένων που αποτελείται από 4521 εγγραφές, που χωρίζονται με τυχαίο τρόπο σε σύνολο εκπαίδευσης που αποτελείται από το 75% των περιπτώσεων (3391) και σύνολο δοκιμής που αποτελείται από το 25% των περιπτώσεων (1130) για να εκτιμήσουμε την απόδοση των ταξινομητών σε νέα δεδομένα.

Κάνοντας χρήση των πακέτων e1072, caret, pROC, θα χρησιμοποιήσουμε τους έτοιμους αλγόριθμούς που υπάρχουν για ταξινόμηση με Μηχανές Διανυσματικής Υποστήριξης. Πρώτα, θα εκπαιδεύσουμε τον ταξινομητή μας και έπειτα θα τον αξιολογήσουμε. Για το σκοπό αυτό, θα εξετάσουμε τη συμπεριφορά του ταξινομητή μας με καθένα από τους τέσσερις πυρήνες; γραμμικό, ακτινωτό, πολυωνυμικό και σιγμοειδή τόσο στο σύνολο εκπαίδευσης όσο και στο σύνολο δοκιμής. Στη συνέχεια, με κατάλληλα μέτρα (Confusion Matrix, ROC Curves, AUC) θα εκτιμήσουμε την απόδοση του.

Διατηρώντας την παράμετρο του κόστους σταθερή και ίση με τη μονάδα, COST=1 (by default cost=1) παίρνουμε τα παρακάτω αποτελέσματα:

	Accuracy		**Sensitivity**		**Specificity**	
	Train	***Test***	***Train***	***Test***	***Train***	***Test***
Linear	0.8947	0.9044	0.17602	0.24806	0.98866	0.98901
Radial	0.9047	0.8903	0.23214	0.054264	0.99266	0.998002
Polynomial	0.8859	0.892	0.020408	0.069767	0.999000	0.998002
Sigmoid	0.8826	0.8823	0.21173	0.007752	0.97032	0.995005

Πίνακας 7.1: Πίνακας αξιολόγησης

Από τον Πίνακα 7.1 προκύπτει ότι:

- Σε κάθε περίπτωση, η ακρίβεια (accuracy) του μοντέλου είναι ικανοποιητική. Ο ακτινωτός(radial) πυρήνας, στο σύνολο εκπαίδευσης, μας δίνει το καλύτερο αποτέλεσμα (0.9047) και ακολουθεί ο γραμμικός (linear), στο σύνολο δοκιμής (0.9044).

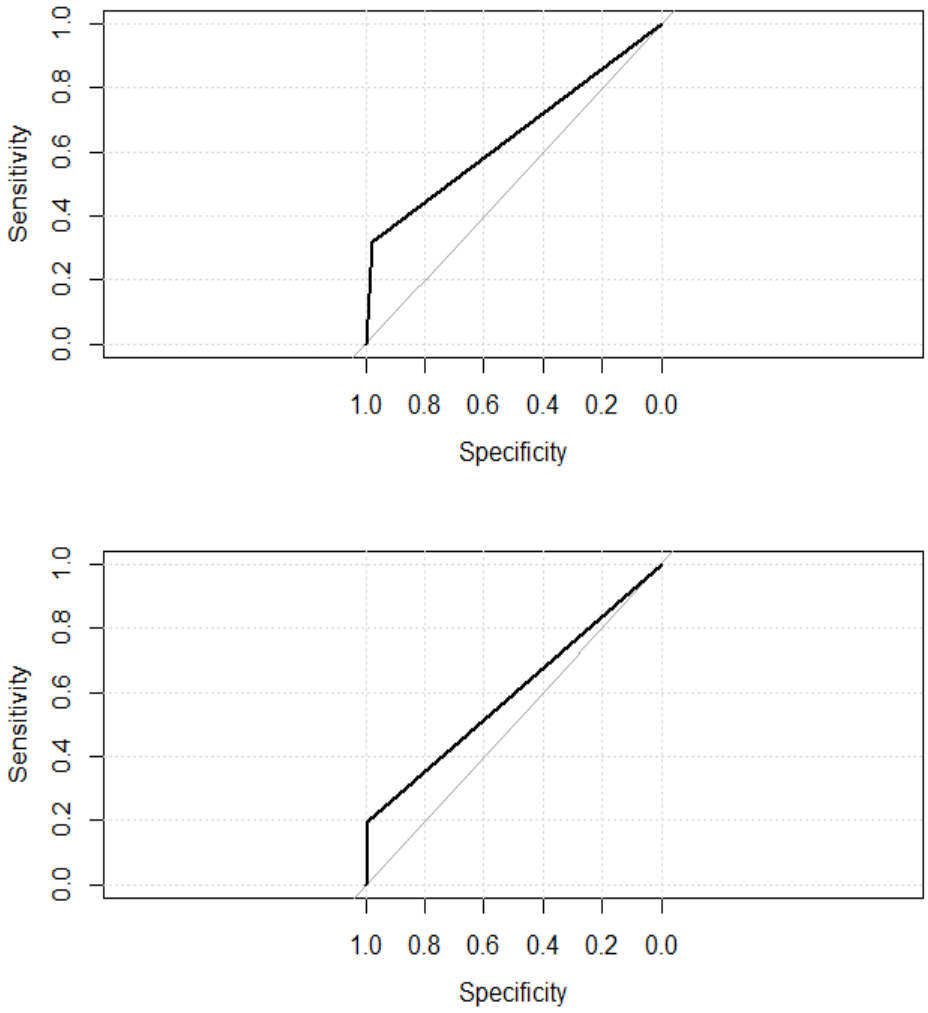
- Η εξειδίκευση (specificity) επίσης του μοντέλου, σε κάθε περίπτωση, είναι αξιοσημείωτη, καθώς μία εξειδίκευση πάνω από 0.9 για ένα μοντέλο σημαίνει ότι το μοντέλο διαπράττει ελάχιστες ψευδώς θετικές προβλέψεις.
- Ο δείκτης ευαισθησίας (sensitivity) του μοντέλου θα μπορούσε να μας προβληματίσει, καθώς οι τιμές που παρατηρούμε είναι ιδιαίτερα χαμηλές (η υψηλότερη τιμή είναι 0.23214). Με άλλα λόγια θα λέγαμε ότι το μοντέλο μας διαπράττει αρκετές ψευδώς αρνητικές προβλέψεις.
  Με μία πιο προσεκτική ματιά, μπορούμε να συμπεράνουμε ότι οι χαμηλές τιμές οφείλονται στην ανισορροπία του συνόλου των δεδομένων. Πιο συγκεκριμένα, η δίτιμη μεταβλητή απόκρισης y (με τιμές ναι/όχι) χωρίζεται σε δύο κλάσεις όπου η μία είναι πολύ μεγαλύτερη από την άλλη (521 ναι, 4000 όχι).

Στο σημείο αυτό θα μπορούσαμε να πούμε ότι ο ακτινωτός και ο γραμμικός πυρήνας είναι οι καλύτεροι για την ταξινόμηση των δεδομένων μας – παρουσιάζουν αυξημένη ακρίβεια και εξειδίκευση και την μεγαλύτερη ευαισθησία μεταξύ των υπολοίπων. Θα προσπαθήσουμε να ενισχύσουμε αυτό το συμπέρασμα με τη βοήθεια των ROC καμπυλών και του εμβαδού κάτω από αυτές. Όπως έχουμε αναφέρει στο κεφάλαιο 5, η καμπύλη ROC είναι μια δισδιάστατη απεικόνιση της απόδοσης της ταξινόμησης. Επιπλέον, το εμβαδόν κάτω από την καμπύλη ROC (AUC) μας βοηθάει να συγκρίνουμε ταξινομητές και να αποφανθούμε ποιος έχει μεγαλύτερη μέση απόδοση.

AUC				
	Linear	**Radial**	**Polynomial**	**Sigmoid**
Train	0.5823	0.6124	0.5097	0.591
Test	0.6185	0.5261	0.5339	0.5014

Πίνακας 7.2: Εμβαδόν κάτω από την καμπύλη ROC για κάθε ταξινομητή

Από τον Πίνακα 7.2 είναι εύκολο να επαληθεύσουμε τις αρχικές μας προβλέψεις. Το εμβαδό κάτω από την καμπύλη ROC είναι μεγαλύτερο στην περίπτωση του γραμμικού πυρήνα (0.6185) και ακολουθεί ο ακτινωτός πυρήνας με μικρή διαφορά (0.6124). Το Σχήμα 7.1 δείχνει τις περιοχές κάτω από τις δύο ROC καμπύλες για τους δύο αυτούς πυρήνες. Ο ταξινομητής Α έχει ελαφρώς μεγαλύτερη έκταση και συνεπώς καλύτερη μέση απόδοση.



Σχήμα 7.2: Δύο ROC γραφήματα**.** *Το πρώτο γράφημα (επάνω) δείχνει την περιοχή κάτω από την καμπύλη ROC. Για τον ταξινομητή Α με το γραμμικό πυρήνα. Το δεύτερο γράφημα (κάτω) δείχνει την περιοχή κάτω από την καμπύλη του διακριτού ταξινομητή Β με τον ακτινωτό πυρήνα.*

7.3.2 Δεύτερη εφαρμογή

Τα δεδομένα που κατεβάσαμε (Iris Data Set) προέρχονται από την τοποθεσία:

"Bache, K. & Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science."

Είναι ίσως η πιο γνωστή βάση δεδομένων που μπορεί να βρεθεί στη βιβλιογραφία αναγνώρισης προτύπων. Τα δεδομένα περιλαμβάνουν 3 κλάσεις των 50 περιπτώσεων η κάθε μία, όπου κάθε κλάση αναφέρεται σε ένα είδος του φυτού ίρις. Μία κλάση είναι γραμμικά διαχωρίσιμα από τις άλλες 2, οι άλλες, όμως, δεν είναι γραμμικά διαχωρίσιμα μεταξύ τους. **Σκοπός** μας είναι να προβλέψουμε την κλάση (κατηγορία) του φυτού ίρις.

Το σύνολο δεδομένων που χρησιμοποιήσαμε έχει 150 εγγραφές (instances), με 4 επεξηγηματικές μεταβλητές και 1 μεταβλητή απόκρισης (response). Έχουμε, λοιπόν, τον παρακάτω πίνακα δεδομένων:

Πίνακας Δεδομένων

ΕΠΕΞΗΓΗΜΑΤΙΚΕΣ ΜΕΤΑΒΛΗΤΕΣ	
X1	Μήκος σέπαλου*
X2	Πλάτος σέπαλου*
X3	Μήκος πέταλου
X4	Πλάτος πέταλου
ΜΕΤΑΒΛΗΤΗ ΑΠΟΚΡΙΣΗΣ	
Y	Κατηγορία φυτού (Setosa, Versicolour, Virginica)

**Το σέπαλο είναι μέρος του φυτού*

Όπως αναφέραμε προηγουμένως, πρόκειται για ένα σύνολο δεδομένων που αποτελείται από 150 εγγραφές, που χωρίζονται με τυχαίο τρόπο σε σύνολο εκπαίδευσης που αποτελείται από το 75% των περιπτώσεων (113) και σύνολο δοκιμής που αποτελείται από το 25% των περιπτώσεων (37) για να εκτιμήσουμε την απόδοση των ταξινομητών σε νέα δεδομένα.

Στο σημείο αυτό, με τη βοήθεια των Μηχανών Διανυσματικής Υποστήριξης θα εκπαιδεύσουμε και θα ελέγξουμε τον ταξινομητή μας. Διατηρώντας την παράμετρο του κόστους σταθερή (cost=1), θα ελέγξουμε την ακρίβεια του ταξινομητή με τους τέσσερις γνωστούς πυρήνες: γραμμικό, ακτινωτό, πολυωνυμικό, σιγμοειδή, στο σύνολο εκπαίδευσης και στο σύνολο δοκιμής.

	Accuracy	
	Train	***Test***
Linear	0.9912	0.973
Radial	0.9823	0.9459
Polynomial	0.9292	0.8378
Sigmoid	0.9292	0.973

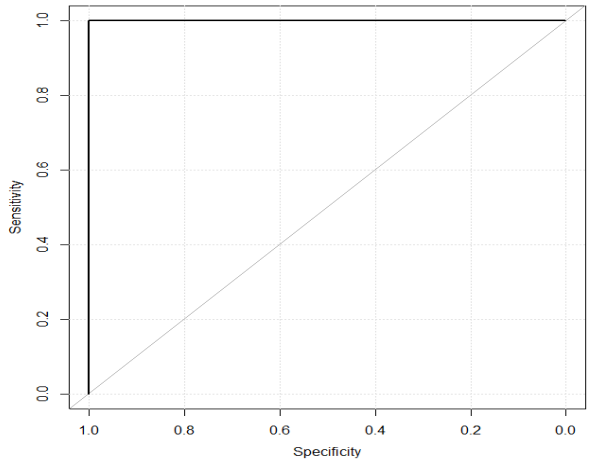
Πίνακας 7.3: Πίνακας Αξιολόγησης

Από τον Πίνακα 7.3 μπορούμε να παρατηρήσουμε ότι η ακρίβεια του ταξινομητή είναι αρκετά υψηλή. Παρόλο αυτά, μεγαλύτερη ακρίβεια έχουμε στην περίπτωση του γραμμικού πυρήνα (0.9912) και ακολουθεί ο ακτινωτός (0.9823), στο σύνολο εκπαίδευσης και στις δύο περιπτώσεις. Ένας επιπλέον έλεγχος της απόδοσης του ταξινομητή (βλέπε Πίνακα 7.4) καθιστά σαφή την υπεροχή του γραμμικού πυρήνα έναντι του ακτινωτού. Όπως βλέπουμε στον πίνακα 7.4, ο γραμμικός πυρήνας παρουσιάζει μεγαλύτερη ευαισθησία και εξειδίκευση συνολικά και στις τρεις κλάσεις:

	Κλάση 1		**Κλάση 2**		**Κλάση 3**	
	Γραμμικός	**Ακτινωτός**	**Γραμμικός**	**Ακτινωτός**	**Γραμμικός**	**Ακτινωτός**
Sensitivity	1.0000	1.0000	0.9722	0.9722	1.0000	0.9744
Specificity	1.0000	1.0000	1.0000	0.9870	0.9865	0.9865

Πίνακας 7.4: Πίνακας Αξιολόγησης

Τέλος, σημειώνουμε ότι το εμβαδόν κάτω από την καμπύλη ROC σε κάθε περίπτωση είναι ίσο με τη μονάδα AUC=1, και παραθέτουμε ενδεικτικά το γράφημα της καμπύλης ROC στην περίπτωση του γραμμικού πυρήνα:



Σχήμα 7.3: ROC γραφήματα που *δείχνει την περιοχή κάτω από την καμπύλη ROC για τον ταξινομητή με το γραμμικό πυρήνα.*

7.3.3 Τρίτη εφαρμογή

Τα δεδομένα που κατεβάσαμε (Boston Housing Data Set) προέρχονται από την τοποθεσία:

"Bache, K. & Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science."

Πρόκειται για μία βάση δεδομένων (housing.data) που αναφέρεται στις αξίες των σπιτιών στα προάστια της Βοστώνης. Έχουν γίνει 506 καταγραφές (instances), με 13 συνεχείς επεξηγηματικές μεταβλητές και 1 μεταβλητή απόκρισης (response). Σημειώνουμε ότι στο πείραμα αυτό εμείς δε χρησιμοποιήσαμε την τέταρτη μεταβλητή (4. 4. CHAS: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise). Έχουμε, λοιπόν, τον παρακάτω πίνακα δεδομένων:

Πίνακας Δεδομένων

ΕΠΕΞΗΓΗΜΑΤΙΚΕΣ ΜΕΤΑΒΛΗΤΕΣ	
X1	Ποσοστό εγκληματικότητας ανά πόλη
X2	Ποσοστό των κατοικημένων περιοχών για τμήματα άνω των 25.000τ.μ.
X3	Αναλογία των στρεμμάτων μη-λιανικού εμπορίου ανά πόλη
X4	Συγκέντρωση των οξειδίων του αζώτου (μέρη ανά 10 εκατομμύρια)
X5	Μέσος αριθμός των δωματίων ανά κατοικία
X6	Ποσοστό ιδιοκατοικημένων μονάδων που έχουν κατασκευαστεί πριν από το 1940
X7	Σταθμισμένες αποστάσεις για πέντε κέντρα απασχόλησης στη Βοστώνη
X8	Δείκτης της πρόσβασης σε ακτινικούς αυτοκινητόδρομους
X9	Ποσοστό φόρου ιδιοκτησίας πλήρους αξίας ανά 25.000$
X10	Αναλογία μαθητή-δασκάλου ανά πόλη
X11	1000 (Bk - 0,63) ^ 2 όπου Bk είναι το ποσοστό των Μαύρων ανά πόλη
X12	% χαμηλότερη κατάσταση του πληθυσμού
ΜΕΤΑΒΛΗΤΗ ΑΠΟΚΡΙΣΗΣ	
Y	Μέση τιμή των ιδιοκατοικημένων κατοικιών σε $1000

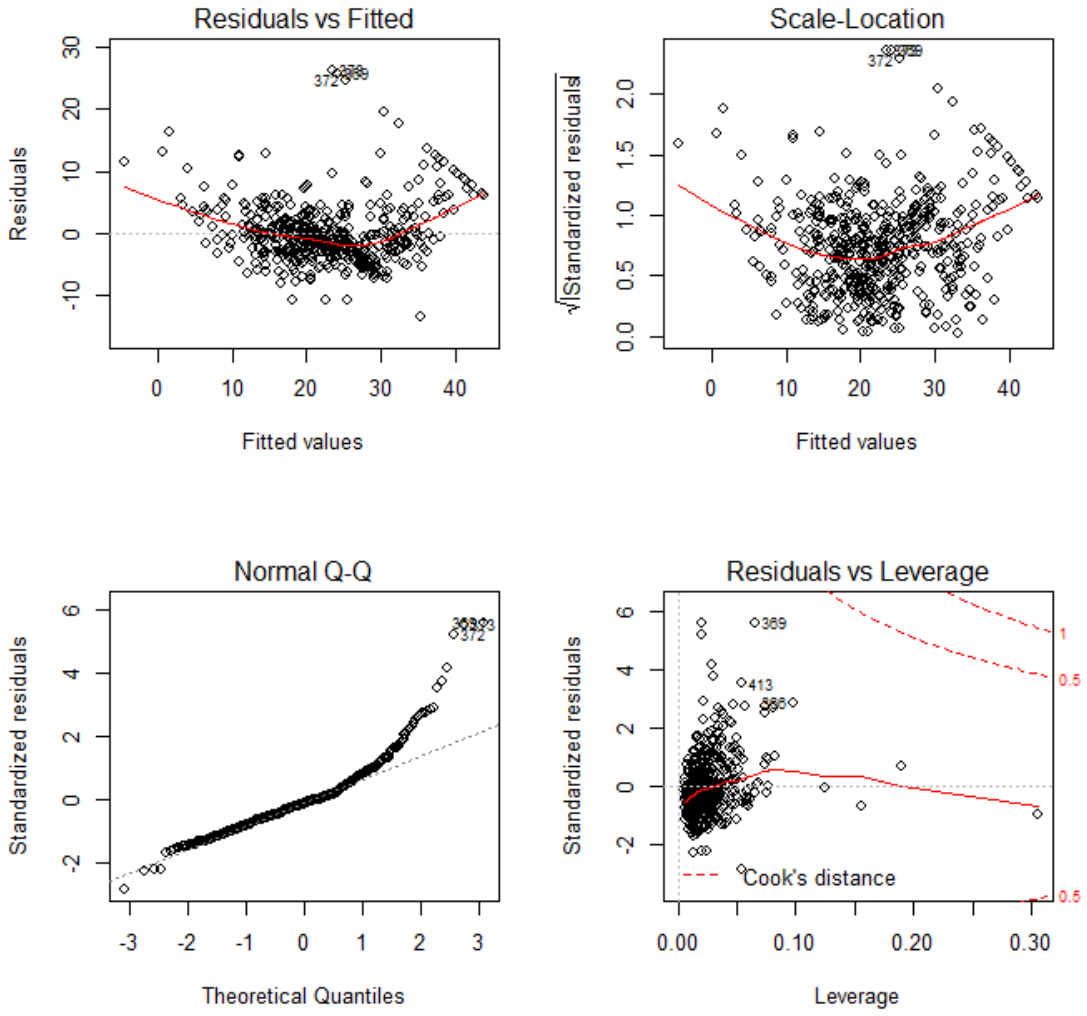
Όπως αναφέραμε προηγουμένως, πρόκειται για ένα σύνολο δεδομένων που αποτελείται από 506 εγγραφές, που χωρίζονται με τυχαίο τρόπο σε σύνολο εκπαίδευσης που αποτελείται από το 75% των περιπτώσεων (380) και σύνολο δοκιμής

που αποτελείται από το 25% των περιπτώσεων (126) για να εκτιμήσουμε την απόδοση των ταξινομητών σε νέα δεδομένα.

Βήμα 1ο

Πρώτα, θα εφαρμόσουμε το γραμμικό μοντέλο παλινδρόμησης στα δεδομένα μας. Η ιδέα μας είναι να συγκρίνουμε τη συμπεριφορά της Παλινδρόμησης με Μηχανές Διανυσματικής Υποστήριξης (SVR: Support Vector Regression) με τη μέθοδο της γραμμικής παλινδρόμησης. *Είναι η SVR καλύτερη για το πρόβλημα μας;*

Ξεκινώντας, θα κάνουμε ορισμένους διαγνωστικούς ελέγχους για την ετεροσκεδαστικότητα, την ομαλότητα, και την επιρροή των υπολοίπων του μοντέλου:



Σχήμα 7.3: Διαγράμματα διαγνωστικών ελέγχων

Τα διαγράμματα του Σχήματος 7.3 είναι αρκετά ικανοποιητικά επομένως τώρα είμαστε σε θέση να προσαρμόσουμε τα δεδομένα μας σε ένα μοντέλο γραμμικής παλινδρόμησης. Τα αποτελέσματα που προκύπτουν είναι τα ακόλουθα:

```
> summary(fit)

Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 +
    x10 + x11 + x12, data = mydata)

Residuals:
     Min       1Q   Median       3Q      Max
-13.3968  -2.8103  -0.6455   1.9141  26.3755

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  36.891960   5.146516   7.168 2.79e-12 ***
x1           -0.113139   0.033113  -3.417 0.000686 ***
x2            0.047052   0.013847   3.398 0.000734 ***
x3            0.040311   0.061707   0.653 0.513889
x4          -17.366999   3.851224  -4.509 8.13e-06 ***
x5            3.850492   0.421402   9.137  < 2e-16 ***
x6            0.002784   0.013309   0.209 0.834407
x7           -1.485374   0.201187  -7.383 6.64e-13 ***
x8            0.328311   0.066542   4.934 1.10e-06 ***
x9           -0.013756   0.003766  -3.653 0.000287 ***
x10          -0.990958   0.131399  -7.542 2.25e-13 ***
x11           0.009741   0.002706   3.600 0.000351 ***
x12          -0.534158   0.051072 -10.459  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.787 on 493 degrees of freedom
Multiple R-squared:  0.7355,    Adjusted R-squared:  0.7291
F-statistic: 114.3 on 12 and 493 DF,  p-value: < 2.2e-16
```

Το μοντέλο φαίνεται ικανοποιητικό, ο συντελεστής προσδιορισμού είναι $R^2 = 73.55\%$, δηλαδή 73.55% της διασποράς της εξαρτημένης μεταβλητής εξηγείται από το μοντέλο.

Βήμα 2ο

Θα εξετάσω εάν η παλινδρόμηση με Μηχανές Διανυσματικής Υποστήριξης δίνει καλύτερα αποτελέσματα.

Θα επιλέξουμε την πιο αποδοτική τεχνική για τα δεδομένα κάνοντας χρήση της SVR. Διατηρώντας την παράμετρο του κόστους C σταθερή και ίση με τη μονάδα (C=1), θα εξετάσουμε ποιος πυρήνας όταν εφαρμοστεί στα δεδομένα μας δίνει την καλύτερη απόδοση (γραμμικός, πολυωνυμικός, ακτινωτός, σιγμοειδής).

Εκτίμηση μου μοντέλου

Ο δείκτης απόδοσης που θα χρησιμοποιήσουμε για να συγκρίνουμε τους ταξινομητές είναι το μέσο τετραγωνικό σφάλμα (RMSE), το οποίο έχουμε αναλύσει στο κεφάλαιο 4. Όσο μικρότερη είναι η τιμή του RMSE τόσο καλύτερα αποτελέσματα παίρνουμε.

Έχουμε, λοιπόν, τον παρακάτω πίνακα:

	Γραμμικός	**Πολυωνυμικός**	**Ακτινωτός**	**Σιγμοειδής**
RMSE	4.9555761	3.2601716	2.8825632	41.253405023
R-squared	0.7091373	0.8809268	0.9033784	0.003291498

Παρατηρούμε ότι ο ακτινωτός πυρήνας μας δίνει τη μικρότερη ρίζα του μέσου τετραγωνικού σφάλματος (RMSE=2.8825632). Επιπλέον, ο ακτινωτός πυρήνας έχει το μεγαλύτερο R-squared=90.34%, δηλαδή ο συντελεστής προσδιορισμού είναι πού καλός και σαφώς καλύτερος από αυτόν που βρήκαμε στο 1ο Βήμα.

7.4 Περίληψη – Συμπεράσματα

Αυτή η διπλωματική επικεντρώθηκε στις Μηχανές Διανυσματικής Υποστήριξης (SVMs).

Περιγράψαμε λεπτομερώς τις Μηχανές Διανυσματικής Υποστήριξης με σκοπό να αναπτύξουμε μια γραμμική μεθοδολογία για την εκτέλεση προβλημάτων ταξινόμησης και παλινδρόμησης. Δίνοντας ιδιαίτερη έμφαση στο τέχνασμα του πυρήνα προσπαθήσαμε να επεκταθούμε σε μη γραμμικά προβλήματα ταξινόμησης και παλινδρόμησης, όσο είναι δυνατόν.

Η θεωρητική βάση των Μηχανών Διανυσματικής Υποστήριξης έχει ερευνηθεί εντατικά κατά τα τελευταία λίγα χρόνια. Οι πρόοδοι, επιπλέον, στη θεωρία βελτιστοποίησης έχουν οδηγήσει σε ταχύτερες μεθόδους εκπαίδευσης.

Πολλές νέες εφαρμογές των SVMs έχουν προκύψει, συμπεριλαμβανομένης της πρόγνωσης του καιρού, του ελέγχου του ομιλητή, της ανίχνευσης προσώπου και πολλών άλλων.

Παρά τα ορισμένα μειονεκτήματα των SVMs, όπως η κατάλληλη επιλογή της συνάρτησης του πυρήνα και η σωστή επιλογή των παραμέτρων αυτού, οι SVMs αποτελούν ένα εύκολο εργαλείο για ευέλικτη μοντελοποίηση.

Οι SVMs μπορούν να επιφέρουν καλύτερα αποτελέσματα σε σύνολα δεδομένων που αποτελούνται από μεγάλο πλήθος δεδομένων και χαρακτηριστικών. Σε αυτή τη διπλωματική εργασία, οι εφαρμογές που έγιναν επιβεβαιώνουν ότι τα ποσοστά ακρίβειας και η σωστή κατηγοριοποίηση ήταν ιδιαίτερα ικανοποιητικά.

Ελπίζουμε ότι αυτή η εργασία θα πείσει τους αναγνώστες ότι οι Μηχανές Διανυσματικής Υποστήριξης αποτελούν ένα ισχυρό εργαλείο πρόβλεψης που έρχεται να προστεθεί στις ήδη υπάρχουσες μεθοδολογίες.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] Auria, L., & Moro, R. A. (2008). *Support vector machines (SVM) as a technique for solvency analysis* (No. 811). Discussion papers//German Institute for Economic Research.

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

[3] Fletcher, T. (2009). Support vector machines explained. *Tutorial paper., Mar*.

[4] Hamel, L. H. (2011). *Knowledge discovery with support vector machines* (Vol. 3). John Wiley & Sons.

[5] Lin, C. J. (2006). A guide to support vector machines. *Department of Computer Science, National Taiwan University*.

[6] Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial intelligence*, *97*(1), 273-324.

[7] Maldonado, S., & Weber, R. (2009). A wrapper method for feature selection using Support Vector Machines. *Information Sciences*, *179*(13), 2208-2217.

[8] Moore, A. W. (2001). VC-dimension for characterizing classifiers. *Tutorial at http://www-2. cs. cmu. edu/awm/tutorials/vcdim08. pdf*.

[9] Rajaraman, A., Leskovec, J., & Ullman, J., (2014). *Mining of Massive Datasets*. 2nd Edition.

[10] Rakotomamonjy, A., (2003). Variable Selection Using SVM-based Criteria, Journal of Machine Learning Research 01/2003, 3:1357-1370.

[11] Shah, R. S. (2007), *Support Vector Machines for Classification and Regression*, Thesis (M. Sc) – McGill University.

[12] Witten, I. H., & Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann. pp 3-39

[13] Youn, E. S. (2002). *Feature Selection in Support Vector Machines*. Thesis (M. Sc) - University of Florida.

[14] Δρόσου, Π. Κ. (2013). *Στατιστικές Μέθοδοι για την Ανάλυση Δεδομένων Υψηλής Διάστασης.* Διπλωματική εργασία στο ΕΜΠ.

[15] Καρακόλιος, Κ. (2013). *Εφαρμογή Μηχανών Διανυσμάτων Υποστήριξης (Support Vector Machines) σε προβλήματα ταξινόμησης πολλών κλάσεων*. Διπλωματική εργασία στο Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης.

[16] Κοκολάκης, Γ. και Φουσκάκης, Δ. (2009). *Στατιστική Θεωρία & Εφαρμογές.*

[17] Ολύμπιος, Σ. (2013). *Bagging και Boosting Μέθοδοι για την Κατασκευή Μοντέλων με Εφαρμογές στα Χρηματοοικονομικά*. Πτυχιακή εργασία στο ΕΜΠ.

[18] Παναγιωτοπούλου, Σ. (2012). *Συμπερασματολογία και ερμηνεία των καμπυλών λειτουργικού χαρακτηριστικού δέκτη μέσω των γενικευμένων γραμμικών μοντέλων.* Διπλωματική εργασία στο ΕΜΠ.

[19] Παπαδάκη, Μ. (2012). *Μηχανές διανυσματικής υποστήριξης (SVMs) και εφαρμογές σε πραγματικά σεισμολογικά δεδομένα*. Διπλωματική εργασία στο ΕΜΠ.

- ***R tutorial***

[20] Kuhn, M. (2013). A Short Introduction to the caret Package.

[21] Maindonald J. H., (2004), *Using R for Data Analysis and Graphics Introduction, Code and Commentary*, Centre for Bioinformation Science, Australian National University.

[22] Meyer, D., Dimitriadou, E., Hornik, K., Lin, C., & Meyer, M. D. (2013). Package 'e1071'.

[23] Karatzoglou, A., Smola, A., Hornik, K., Karatzoglou, M. A., SparseM, S., & Yes, L. (2013). The kernlab package.

[24] Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., ... & Robin, M. X. (2014). Package 'pROC'.

[25] Venables, W. N., Smith, D. M., & R Development Core Team. (2005). An introduction to R.

[26] Verzani, J. (2014). *Using R for introductory statistics*. CRC Press.

- ***Διαδικτυακοί Ιστότοποι***

[26] http://www.wikipedia.org

[27] http://mlr.cs.umass.edu/ml/index.html

ΠΑΡΑΡΤΗΜΑ

Λίγα λόγια για τον Vladimir N. Vapnik

Ο Vladimir Naumovich Vapnik (Ρωσικά: Владимир Наумович Вапник) είναι ένας από τους κύριους κατασκευαστές της θεωρίας Vapnik-Chervonenkis. Γεννήθηκε στη Σοβιετική Ένωση. Έλαβε το μεταπτυχιακό του στα Μαθηματικά από το Πανεπιστήμιο του Ουζμπεκιστάν (Uzbek State University) το 1958 και το Διδακτορικό στον τομέα της Στατιστικής από το Institute of Control Sciences, στη Μόσχα το 1964. Εργάστηκε σε αυτό το Ίδρυμα από το 1961 έως το 1990 και έγινε Επικεφαλής του Ερευνητικού Τμήματος της Επιστήμης Υπολογιστών. Στο τέλος του 1990, μετακόμισε στις ΗΠΑ και εργάστηκε στο "Adaptive Systems Research Department at AT&T Bell Labs in Holmdel", στο Νιού Τζέρσεϊ. Ο Vapnik έφυγε από το AT & T το 2002 και εντάχθηκε στο "NEC Laboratories" στο Πρίνσετον, Νιού Τζέρσεϊ όπου εργάζεται μέχρι και σήμερα στην ομάδα Μηχανικής Μάθησης. Κατέχει, επίσης, θέση Καθηγητή της Επιστήμης των Υπολογιστών και της Στατιστικής στο Royal Holloway, στο Πανεπιστήμιο του Λονδίνου από το 1995, καθώς και μια θέση καθηγητή της Επιστήμης των Υπολογιστών στο Πανεπιστήμιο Κολούμπια της Νέας Υόρκης από το 2003. Έλαβε το 2005 το βραβείο "Gabor Award", το 2008 το "Paris Kanellakis Award", το 2010 το "Neural Networks Pioneer Award", το 2012 το " IEEE Frank Rosenblatt Award", το 2012 το Μετάλλιο "Benjamin Franklin in Computer and Cognitive Science from the Franklin Institute", και το 2013 το" C & C Award" από το NEC C & C Foundation.

Στα εργαστήρια AT & T, ο Vapnik και οι συνεργάτες του ανέπτυξαν τη θεωρία των Μηχανών Διανυσματικής Υποστήριξης. Απέδειξαν την απόδοσή τους σε μια σειρά από προβλήματα που παρουσιάζουν ενδιαφέρον στην κοινότητα της μηχανικής μάθηση, συμπεριλαμβανομένης της αναγνώρισης χειρογράφου.