ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

Διπλωματική Εργασία

Μοντελοποίηση Δεδομένων Επιβίωσης με Μηχανές Διανυσμάτων Υποστήριξης

Καστρησίου Άννα-Μαρία

Επιβλέπων Καθηγητής:

Χρήστος Κουκουβίνος, Καθηγητής Ε.Μ.Π.

Αθήνα, 2018

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

Μοντελοποίηση Δεδομένων Επιβίωσης με Μηχανές Διανυσμάτων Υποστήριξης

ΚΑΣΤΡΗΣΙΟΥ ΑΝΝΑ-ΜΑΡΙΑ

Αθήνα, 2018

Ευχαριστίες

Σε αυτό το σημείο θα ήθελα να ευχαριστήσω πολύ κάποιους ανθρώπους που με βοήθησαν και με στήριξαν για την ολοκλήρωση αυτής της εργασίας.

Αρχικά, θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου κ. Χρήστο Κουκουβίνο, καθηγητή στο Εθνικό Μετσόβιο Πολυτεχνείο, αφού με την εργασία που μου ανέθεσε, μου έδωσε τη δυνατότητα να ασχοληθώ και να μελετήσω συγκεκριμένα ερευνητικά πεδία.

Θα ήθελα επίσης να ευχαριστήσω ιδιαίτερα την υποψήφια διδάκτορα Κρυσταλλένια Δρόσου, για την καθοδήγηση, την επίβλεψη και τη συνεχή υποστήριξή της καθ' όλη τη διάρκεια εκπόνησης της εργασίας.

Τέλος, θα ήθελα να ευχαριστήσω το στενό οικογενειακό και φιλικό μου περιβάλλον για την αμέριστη στήριξή τους σε όλη την πορεία των σπουδών μου.

Άννα-Μαρία Καστρησίου

Εθνικό Μετσόβιο Πολυτεχνείο

Σχολή Εφαρμοσμένων Μαθηματικών

και Φυσικών Επιστημών

Αθήνα, 2018

Πρόλογος

Η ανάλυση επιβίωσης αποτελεί έναν κλάδο της στατιστικής για την ανάλυση της αναμενόμενης διάρκειας του χρόνου έως ότου ένα ή περισσότερα γεγονότα συμβούν, όπως είναι ο θάνατος σε βιολογικούς οργανισμούς ή η αποτυχία σε μηχανικά συστήματα. Αυτή η περιοχή ονομάζεται θεωρία αξιοπιστίας ή ανάλυση επιβίωσης και είναι ευρέως γνωστή στους τομείς της μηχανικής, της οικονομίας και της κοινωνιολογίας. Η ανάλυση επιβίωσης αποτελεί ένα όλο και πιο ενεργό και πολύ σημαντικό τομέα έρευνας. Αυτό είναι προφανές από το μεγάλο όγκο βιβλιογραφικών αναφορών που έχουν αναπτυχθεί είτε με τη μορφή βιβλίων, είτε με τη μορφή ερευνητικών εργασιών. Ιδιαίτερα στις μέρες μας όπου καλούμαστε να χειριστούμε τεράστιους όγκους δεδομένων, η επίλυση προβλημάτων για δεδομένα διάρκειας ζωής, είτε γενικότερα για αποκομμένα δεδομένα, αποτελεί πρόκληση και χρήζει ιδιαίτερης μεταχείρισης. Για να χειριστούμε τέτοια δεδομένα είναι απαραίτητα ισχυρά εργαλεία που αναφέρονται είτε σε αλγορίθμους είτε σε μεθόδους που συνήθως ανήκουν στον τομέα της μηχανικής μάθησης. Ένα από αυτά είναι και οι μηχανές διανυσμάτων υποστήριξης γνωστές ευρέως ως Support Vector Machines (SVMs).

ΠΕΡΙΛΗΨΗ

Στη συγκεκριμένη εργασία θα ασχοληθούμε με τη δημιουργία μοντέλων για δεδομένα με αποκομμένες παρατηρήσεις και ειδικότερα σε δεδομένα επιβίωσης. Αυτό επιτυγχάνεται με τη βοήθεια των μηχανών διανυσμάτων υποστήριξης (SVM) για δύο κλάσεις, αλλά και το μοντέλο του Cox.

Αρχικά, στο πρώτο κεφάλαιο αναφέρεται η ανάγκη εξόρυξης δεδομένων (Data Mining), ώστε ο τεράστιος αριθμός δεδομένων να μετατραπεί σε χρήσιμες πληροφορίες. Επίσης ορίζονται κάποιες βασικές έννοιες, μεταξύ των οποίων και οι ταξινομητές (γραμμικοί και τετραγωνικοί), τα σφάλματα, οι πίνακες σύγχυσης και απώλειας και οι καμπύλες ROC.

Στη συνέχεια, στο δεύτερο κεφάλαιο αναφέρονται μοντέλα παλινδρόμησης για αποκομμένες παρατηρήσεις, που προκύπτουν από το θάνατο ή την αποτυχία μιας μονάδας. Επίσης, ορίζονται κάποιες βασικές έννοιες, όπως ο χρόνος επιβίωσης και αποκοπής, ο εκτιμητής Nelson-Aalen, η συνάρτηση κινδύνου και το μοντέλο του Cox, ενώ αναφέρουμε και ένα παράδειγμα για τα δεδομένα του καρκίνου του μαστού. Επιπλέον, αναφερόμαστε στη μοντελοποίηση των δεδομένων επιβίωσης, καθώς εξηγούμε πώς μπορεί να χρησιμοποιηθεί και για την πρόβλεψη των δεδομένων.

Στο τρίτο κεφάλαιο ασχολούμαστε με τις μηχανές διανυσμάτων υποστήριξης (SVM), αναφερόμαστε στον ταξινομητή μέγιστου περιθωρίου και στον ταξινομητή διανυσμάτων υποστήριξης. Γίνεται περιγραφή των SVM με δύο κλάσεις καθώς επίσης γίνεται και μια αναφορά στην περίπτωση δεδομένων με περισσότερες από δύο κλάσεις, ενώ παρουσιάζεται και μια εφαρμογή της SVM σε δεδομένα καρδιακών παθήσεων. Επιπλέον περιγράφονται δύο μέθοδοι για την ανάλυση δεδομένων επιβίωσης που βασίζονται στις μηχανές διανυσμάτων υποστήριξης.

Τέλος, στο τέταρτο κεφάλαιο αναλύονται τα πειραματικά αποτελέσματα για συνθετικά (προσομοιωμένα) και για πραγματικά δεδομένα, ενώ στο πέμπτο και τελευταίο κεφάλαιο παρουσιάζονται τα συμπεράσματα της ανάλυσης SVM και του μοντέλου του Cox.

Abstract

In this assignment we will deal with the creation of models for data with censoring observations and specifically for survival data. This is achieved by using support vector machines (SVMs) for two classes, as well as the Cox model.

To begin with, the first chapter considers the need of data mining, so as to convert the large amount of data into useful information. In addition, some basic concepts are defined, including classifiers (linear or quadratic), errors, confusion and loss matrices and ROC curves.

Thereafter, chapter two focuses on regression models for censoring observations that have resulted from the death or the failure of a unit. Moreover, some key concepts are defined, such as the survival time and censoring time, the Nelson-Aalen classifier, the hazard function and the Cox model, as well as an example for breast cancer data is also provided. Additionally we refer to the modeling of survival data, while we explain how it can be used for data prediction.

In chapter three we refer to support vector machines (SVMs), the maximal margin classifier, and to the support vector classifier. We provide descriptions of SVMs with two classes, as well as a reference to instances of data with more than two classes is also made, while an application of SVMs to the heart disease data is presented. Two methods are also outlined for the analysis of survival data which are based on support vector machines.

In conclusion, in chapter four we analyse the experimental results for synthetic and real life data, whereas in chapter five, the final chapter, we present the conclusions of the SVM and Cox model analysis.

Περιεχόμενα

ΚΕΦΑΛΑΙΟ 1: ΤΑΞΙΝΟΜΗΣΗ – ΒΑΣΙΚΕΣ ΑΡΧΕΣ ΑΝΑΓΝΩΡΙΣΗΣ ΠΡΟΤΥΠΩΝ

1.1 Εξόρυξη Δεδομένων – Εισαγωγή

Τα τελευταία χρόνια η ανάγκη για παραγωγή και συλλογή δεδομένων έχει αυξηθεί ραγδαία. Η ευρεία χρήση των barcodes στα περισσότερα προϊόντα του εμπορίου καθώς και η μηχανογράφηση πολλών επιχειρηματικών και δημόσιων συναλλαγών, αλλά και η πρόοδος που έχει γίνει στα εργαλεία συλλογής δεδομένων μας έχουν δώσει τη δυνατότητα να συλλέγουμε ένα τεράστιο πλήθος δεδομένων.

Εκατομμύρια βάσεις δεδομένων χρησιμοποιούνται για τη διαχείριση των επιχειρήσεων, τη δημόσια διοίκηση, στο επιστημονικό πεδίο, στην εφαρμοσμένη μηχανική καθώς και σε πολλά άλλα πεδία. Αξίζει να σημειωθεί ότι ο αριθμός αυτός των βάσεων δεδομένων συνεχίζει να αυξάνεται με ταχείς ρυθμούς, λόγω της διαθεσιμότητας ισχυρών συστημάτων βάσεων δεδομένων. Αυτή η τεράστια αύξηση των δεδομένων, αλλά και των βάσεων δεδομένων δημιούργησε την ανάγκη για νέες τεχνικές και εργαλεία τα οποία θα μπορούν έξυπνα και αυτόματα με κατάλληλη επεξεργασία να μετατρέψουν τα δεδομένα σε χρήσιμες πληροφορίες και γνώσεις. Συνεπώς, η εξόρυξη δεδομένων (Data Mining) εξελίσσεται σε έναν ερευνητικό τομέα με αυξανόμενη σημασία (Fayyad et al., 1996, Piatetsky-Shapiro and Frawley, 1991, Silberschatz et al. 1995).

Ως εξόρυξη δεδομένων εννοούμε και τη γνώση που προκύπτει από τις βάσεις δεδομένων μέσω μιας διαδικασίας για την μη τετριμμένη εξαγωγή πεπλεγμένων, άγνωστων και δυνητικά χρήσιμων πληροφοριών (όπως περιορισμούς, κανονικότητες) από δεδομένα σε βάσεις δεδομένων (Piatetsky-Shapiro and Frawley, 1991). Σε μερικά άρθρα και έγγραφα ο όρος εξόρυξη δεδομένων (Data Mining) εμφανίζεται ως εξόρυξη γνώσης από βάσεις δεδομένων (Knowledge Mining From Data Bases), εξόρυξη γνώσης (Knowledge Extraction), ανάλυση δεδομένων (Data Analysis) κλπ. Με τις γνώσεις που προκύπτουν από τις βάσεις δεδομένων συλλέγονται ενδιαφέρουσες γνώσεις και πληροφορίες υψηλού επιπέδου, οι οποίες μπορούν να εξαχθούν από σχετικά σύνολα στις βάσεις δεδομένων και να διερευνηθούν από διαφορετικές γωνίες, έτσι ώστε με αυτόν τον τρόπο μεγάλες βάσεις δεδομένων να χρησιμεύουν ως πλούσιες και αξιόπιστες πηγές για την παραγωγή γνώσης. Η εξόρυξη πληροφορίας και γνώσης από μεγάλες βάσεις δεδομένων έχει αναγνωριστεί από πολλούς ερευνητές ως ένα βασικό θέμα έρευνας σε συστήματα βάσεων δεδομένων και έχει αποτελέσει ένα σημαντικό τομέα σε πολλές βιομηχανικές επιχειρήσεις, στις οποίες έχει αποφέρει σημαντικά έσοδα. Η εξόρυξη δεδομένων έχει κερδίσει το ενδιαφέρον πολλών διαφορετικών πεδίων, συμπεριλαμβανομένων των συστημάτων βάσεων δεδομένων της τεχνητής νοημοσύνης της στατιστικής και πολλών άλλων. Επιπλέον πολλές αναδυόμενες εφαρμογές στον τομέα της παροχής πληροφοριών

όπως είναι οι υπηρεσίες online και World Wide Web κάνουν έκκληση για διάφορες τεχνικές εξόρυξης δεδομένων για την καλύτερη κατανόηση της συμπεριφοράς των χρηστών, για τη βελτίωση παρεχόμενων υπηρεσιών, αλλά και για την αύξηση των επιχειρηματικών ευκαιριών.

Για την κατασκευή αποτελεσματικής εξόρυξης δεδομένων, κάποιος πρέπει πρώτα να εξετάσει τι είδους χαρακτηριστικά αναμένει να έχει το σύστημα και τι προκλήσεις έχει να αντιμετωπίσει στην ανάπτυξη των τεχνικών εξόρυξης δεδομένων. Παρακάτω θα δούμε κάποια βασικά χαρακτηριστικά που θα πρέπει να εξετάσει κάποιος, που θα ασχοληθεί με την εξόρυξη δεδομένων:

Χειρισμός διαφορετικών τύπων δεδομένων. Λόγω της ύπαρξης πολλών ειδών δεδομένων και βάσεων δεδομένων, τα οποία χρησιμοποιούνται σε διάφορες εφαρμογές, πρέπει κάποιος να έχει τη γνώση να εκτελέσει αποτελεσματικά την εξόρυξη δεδομένων σε διαφορετικά είδη δεδομένων.

Αποδοτικότητα και επεκτασιμότητα των αλγορίθμων εξόρυξης δεδομένων. Για την εξαγωγή αποτελεσματικών πληροφοριών από μεγάλες βάσεις δεδομένων απαιτούνται αλγόριθμοι εξόρυξης δεδομένων, οι οποίοι είναι προβλέψιμοι και αποδεκτοί σε μεγάλες βάσεις δεδομένων.

Χρησιμότητα, ασφάλεια και εκφραστικότητα των αποτελεσμάτων της εξόρυξης δεδομένων. Οι γνώσεις που παίρνουμε από την βάση δεδομένων πρέπει να απεικονίζονται με ακρίβεια, έτσι ώστε να είναι χρήσιμες για ορισμένες εφαρμογές. Η ανακρίβεια πρέπει να εκφράζεται με μέτρα αβεβαιότητας υπό τη μορφή προσέγγισης κανόνων ή από ποσοτικούς κανόνες.

Έκφραση των διαφόρων ειδών των αποτελεσμάτων της εξόρυξης δεδομένων. Από μια μεγάλη ποσότητα δεδομένων μπορούμε να πάρουμε πολλά διαφορετικά είδη γνώσεων ή ακόμα και να χρησιμοποιήσουμε αυτές τις γνώσεις σε διαφορετικές μορφές, ανάλογα με το πού θέλουμε να τις εφαρμόσουμε.

Διαδραστική εξόρυξη γνώσης σε πολλαπλά επίπεδα άντλησης πληροφοριών. Δεδομένου ότι είναι δύσκολο να προβλέψουμε τι πληροφορίες ακριβώς θα μπορούσαμε να πάρουμε από μια βάση, ένα υψηλό επίπεδο εξόρυξης δεδομένων θα πρέπει να αντιμετωπίζεται ως ένας ανιχνευτής, ο οποίος να έχει τη δυνατότητα να αποκαλύπτει μερικές πληροφορίες, οι οποίες θα βοηθούν για περαιτέρω διερεύνηση.

Εξόρυξη πληροφορίας από διαφορετικές πηγές δεδομένων. Το ευρέως διαθέσιμο δίκτυο υπολογιστών συμπεριλαμβανομένου και του διαδικτύου συνδέει πολλές πηγές δεδομένων και τις διανέμει σε ετερογενείς βάσεις δεδομένων. Η εξόρυξη γνώσης από διαφορετικές πηγές σχηματοποιημένων ή μη δεδομένων με διαφορετική σημασιολογία δημιουργεί προκλήσεις για νέα εξόρυξη δεδομένων.

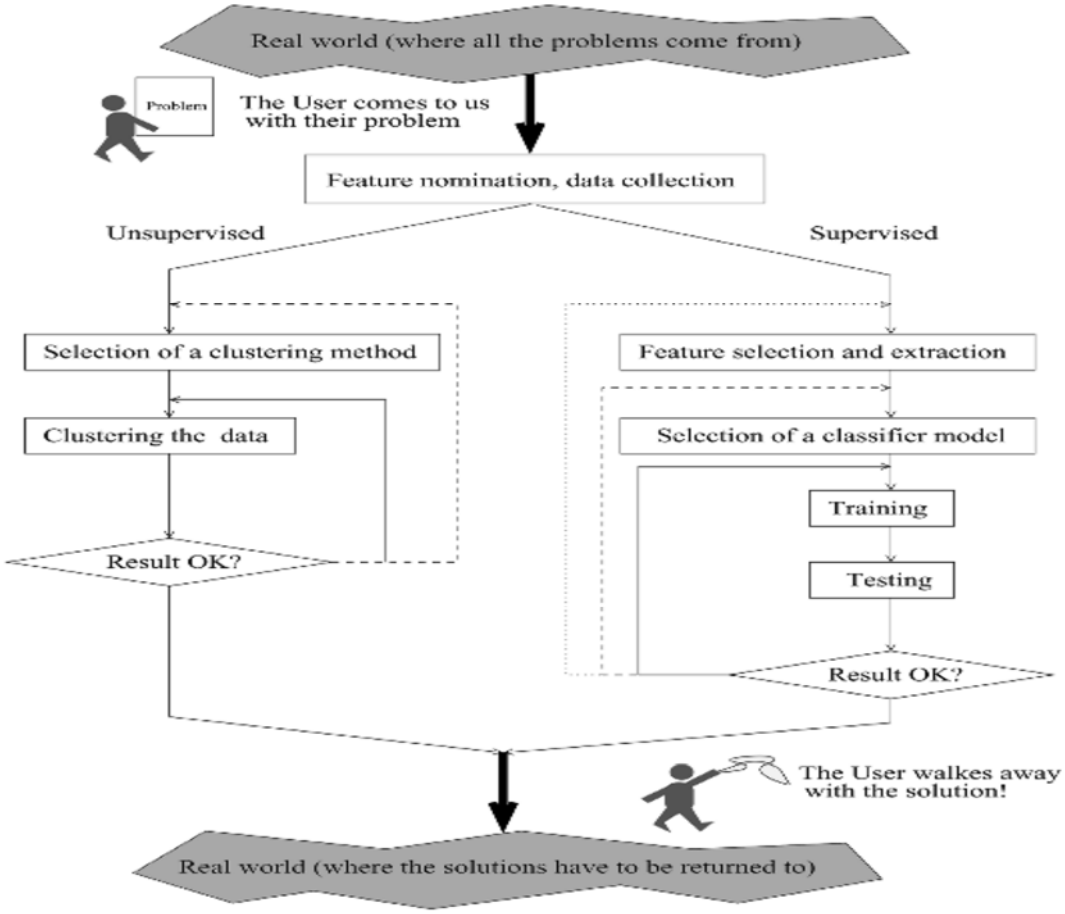
Προστασία της ιδιωτικής ζωής και ασφάλειας των δεδομένων. Όταν τα δεδομένα μπορούν να προβληθούν από πολλές διαφορετικές γωνίες και σε διαφορετικά επίπεδα, απειλείται η ασφάλεια προστασίας δεδομένων. Είναι σημαντικό να

επισημάνουμε ότι πρέπει να αναπτυχθούν μέτρα ασφαλείας για την πρόληψη παραβίασης της ιδιωτικής ζωής, δηλαδή την πρόληψη αποκάλυψης ευαίσθητων πληροφοριών (Fu, 1982).

1.2 Βασικές Έννοιες: Τάξη, Χαρακτηριστικά και Σύνολα Δεδομένων

1.2.1 Μοτίβο Κύκλου Αναγνώρισης

Η αναγνώριση προτύπων σχετίζεται με την αντιστοίχιση ετικετών σε αντικείμενα. Τα αντικείμενα περιγράφονται από ένα σύνολο μετρήσεων, τα οποία ονομάζονται χαρακτηριστικά ή λειτουργίες. Επειδή η αναγνώριση προτύπων αντιμετωπίζει διάφορες προκλήσεις για την επίλυση προβλημάτων της πραγματικής ζωής και παρά το γεγονός ότι γίνονται έρευνες σχετικά με αυτό το αντικείμενο εδώ και δεκαετίες, σύγχρονες θεωρίες εξακολουθούν να συνυπάρχουν με πιο παλιές ιδέες. Αυτό αντικατοπτρίζεται στην ποικιλία των μεθόδων και των τεχνικών που είναι διαθέσιμες για κάθε ερευνητή.



Σχήμα 1.2.1: Μοτίβο Κύκλου Αναγνώρισης

Το παραπάνω σχήμα μας δείχνει τα βασικά καθήκοντα και τα στάδια αναγνώρισης προτύπων. Ας υποθέσουμε ότι υπάρχει ένας χρήστης που μας φέρνει αντιμέτωπους με ένα πρόβλημα και ένα σύνολο δεδομένων. Καθήκον μας είναι να αποσαφηνίσουμε το πρόβλημα, να φτιάξουμε κατάλληλη ορολογία, να αναγνωρίσουμε τα πρότυπα, να το επιλύσουμε και τέλος να γνωστοποιήσουμε τη λύση στο χρήστη.

Σε περίπτωση που το σύνολο των δεδομένων δε μας είναι γνωστό, πραγματοποιείται ένα πείραμα, μέσα από το οποίο συλλέγεται το σύνολο δεδομένων. Τα σχετικά χαρακτηριστικά θα πρέπει να ορισθούν και να μετρηθούν. Το σύνολο των χαρακτηριστικών θα πρέπει να είναι όσο το δυνατόν μεγαλύτερο, συμπεριλαμβανομένων και των χαρακτηριστικών, των οποίων η σημαντικότητα να μην είναι ακόμα εμφανής μέχρι εκείνη τη στιγμή. Θα μπορούσε η σημασία τους να συνδυάζεται με άλλα χαρακτηριστικά. Συνήθως, τα όρια για τη συλλογή των δεδομένων προέρχονται από την οικονομική πλευρά του project. Ένας άλλος πιθανός λόγος για τέτοιου είδους περιορισμούς θα μπορούσε να είναι ότι μερικά χαρακτηριστικά δε μπορούν εύκολα να μετρηθούν, όπως για παράδειγμα χαρακτηριστικά που απαιτούν βλάβη ή καταστροφή του αντικειμένου, ιατρικές εξετάσεις που απαιτούν διεισδυτική εξέταση όταν υπάρχουν αντενδείξεις κλπ.

Υπάρχουν δύο κύριοι τύποι προβλημάτων αναγνώρισης προτύπων, χωρίς επίβλεψη και με επίβλεψη. Στην κατηγορία χωρίς επίβλεψη ή αλλιώς μη επιβλεπόμενη μάθηση (unsupervised learning), το πρόβλημα είναι να ανακαλυφθεί η δομή του συνόλου δεδομένων, αν υπάρχει. Αυτό συνήθως σημαίνει ότι ο χρήστης θέλει να γνωρίζει αν υπάρχουν ομάδες στα δεδομένα και ποια χαρακτηριστικά καθιστούν τα αντικείμενα παρόμοια μέσα στην ομάδα και διαφορετικά μεταξύ των ομάδων. Πολλοί αλγόριθμοι ομαδοποίησης υπάρχουν και αναπτύσσονται με μη επιβλεπόμενη μάθηση. Η επιλογή του αλγορίθμου έχει να κάνει με τις προτιμήσεις του σχεδιαστή. Διαφορετικοί αλγόριθμοι θα μπορούσαν να οδηγήσουν σε διαφορετικές δομές για το ίδιο σύνολο δεδομένων. Το αρνητικό και το θετικό αυτού του κλάδου της αναγνώρισης προτύπων είναι ότι δεν υπάρχει κανένα αληθινό έδαφος σε αντίθεση, ώστε να συγκριθούν τα αποτελέσματα. Η μόνη ένδειξη για το πόσο καλά είναι τα αποτελέσματα, είναι πιθανόν η υποκειμενική εκτίμηση του χρήστη.

Στην μάθηση με επίβλεψη (supervised learning) κάθε αντικείμενο στο σύνολο δεδομένων συνοδεύεται από μια προκαθορισμένη ετικέτα για την κάθε κατηγορία. Το καθήκον μας είναι να εκπαιδεύσουμε έναν ταξινομητή για να κάνει την ετικέτα αυτή «λογική». Πιο συχνά η διαδικασία αυτή κατηγοριοποίησης δεν μπορεί να περιγραφεί με μια αλγοριθμική μορφή. Έτσι παρέχουμε στη μηχανή/ αλγόριθμο δεξιότητες μάθησης και παρουσιάζουμε το κάθε στοιχείο του συνόλου δεδομένων συνοδευόμενο με μια ετικέτα. Η ταξινομημένη γνώση βάσει της οποίας εκπαιδεύτηκε το μηχάνημα μπορεί να είναι ασαφής, αλλά η ακρίβεια της αναγνώρισης του ταξινομητή θα είναι η ένδειξη για την επάρκειά του.

Τα χαρακτηριστικά δεν είναι όλα το ίδιο σημαντικά. Μερικά από αυτά είναι σημαντικά μόνο σε σχέση με άλλα χαρακτηριστικά και μερικά από αυτά μπορεί να προκύπτουν μη σημαντικά για το πρόβλημα που εξετάζουμε.

Η επιλογή, η εκπαίδευση και η δοκιμή ενός μοντέλου ταξινόμησης αποτελούν τον πυρήνα των εποπτευόμενων μοτίβων αναγνώρισης. Καθώς οι διακεκομμένες γραμμές και οι γραμμές με κουκίδες στο Σχήμα 1.2.1 δείχνουν το βρόγχο ρύθμισης, το μοντέλο μπορεί να κλείσει σε διαφορετικούς τόπους. Εμείς μπορεί να αποφασίσουμε να χρησιμοποιήσουμε το ίδιο μοντέλο ταξινόμησης και την εκ νέου εκπαίδευση μόνο με διαφορετικές παραμέτρους, ή και να αλλάξουμε το μοντέλο ταξινόμησης. Μερικές φορές η επιλογή και εξαγωγή χαρακτηριστικών περιλαμβάνονται στο βρόγχο.

Όταν επιτευχθεί μια ικανοποιητική λύση, τότε μπορούμε να την προσφέρουμε στο χρήστη τη δυνατότητα για επιπλέον δοκιμές και εφαρμογή.

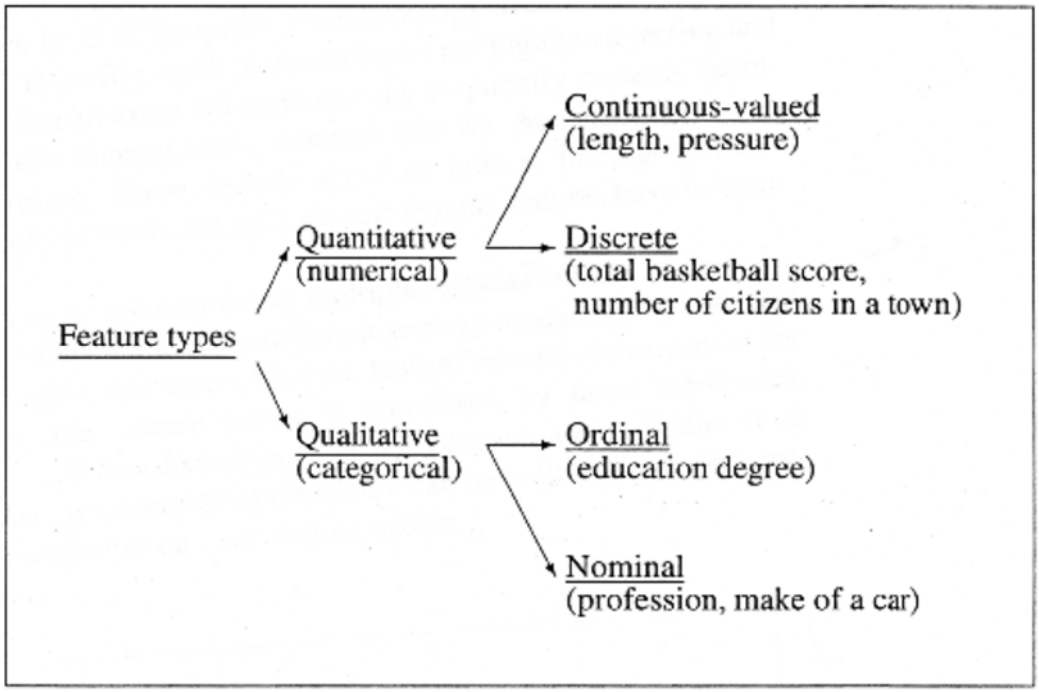
1.2.2 Τάξεις και Ετικέτες

Διαισθητικά, μια τάξη περιέχει παρόμοια αντικείμενα, ενώ αντικείμενα από διαφορετικές τάξεις είναι ανόμοια. Μερικές τάξεις έχουν μια σαφή ερμηνεία και στην απλούστερη περίπτωση αλληλοαναιρούνται. Για παράδειγμα, στην επαλήθευση της υπογραφής. Η υπογραφή είναι είτε γνήσια, είτε πλαστή. Η αληθινή κλάση είναι η μια από τις δύο, άσχετα με το αν εμείς θα μπορέσουμε να μαντέψουμε σωστά από την παρατήρηση της συγκεκριμένης υπογραφής. Σε άλλα προβλήματα, οι κλάσεις ίσως να είναι δύσκολο να προσδιοριστούν, για παράδειγμα, οι αριστερόχειρες από τους δεξιόχειρες. Η ιατρική έρευνα προκαλεί μια τεράστια ποσότητα από δυσκολίες στην ερμηνεία των δεδομένων, εξαιτίας της φυσικής μεταβλητότητας του αντικειμένου μελέτης. Για παράδειγμα, είναι συχνά επιθυμητό να γίνεται διάκριση μεταξύ χαμηλού, μέσου και υψηλού κινδύνου, αλλά μπορούμε μετά βίας να ορίσουμε απότομα κριτήρια διάκρισης μεταξύ αυτών των ετικετών τάξεων.

Θα πρέπει να υποθέσουμε ότι υπάρχουν c πιθανές τάξεις στο πρόβλημα, κατηγοριοποιημένες από ω_1 έως ω_c, οργανωμένες ως ένα σύνολο ετικετών $\Omega = \{\omega_1, \omega_2, \dots, \omega_c\}$ και ότι κάθε αντικείμενο ανήκει σε μια και μοναδική τάξη.

1.2.3 Χαρακτηριστικά

Όπως έχουμε ήδη αναφέρει, τα αντικείμενα περιγράφονται από ιδιότητες που ονομάζονται χαρακτηριστικά. Τα χαρακτηριστικά αυτά μπορεί να είναι είτε ποιοτικά, είτε ποσοτικά, όπως φαίνεται παρακάτω, στο Σχήμα 1.2.2.



Σχήμα 1.2.2: Είδη Χαρακτηριστικών

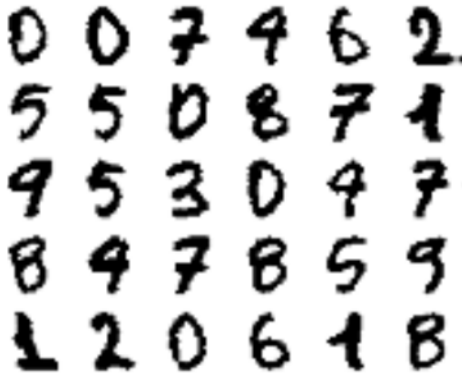
Τα διακριτά χαρακτηριστικά με ένα μεγάλο αριθμό από πιθανές τιμές αντιμετωπίζονται ως ποσοτικά. Ποιοτικά (κατηγορικά) χαρακτηριστικά είναι αυτά με μικρό αριθμό πιθανών τιμών, με ή χωρίς διαβαθμίσεις. Ένας κλάδος της αναγνώρισης προτύπων, που ονομάζεται συντακτική αναγνώρισης προτύπων (σε αντίθεση με τη στατιστική αναγνώρισης προτύπων) ασχολείται αποκλειστικά με ποιοτικά χαρακτηριστικά (Fu, 1982).

Η στατιστική αναγνώρισης προτύπων λειτουργεί με αριθμητικά χαρακτηριστικά. Αυτά περιλαμβάνουν, για παράδειγμα, τη συστολική αρτηριακή πίεση, την ταχύτητα του ανέμου, τα καθαρά κέρδη μιας εταιρείας τους τελευταίους δώδεκα μήνες, την ένταση του επιπέδου του γκρι ενός pixel. Οι τιμές των χαρακτηριστικών ενός δεδομένου αντικειμένου είναι τοποθετημένες ως ένα n-διάστατο διάνυσμα $x = [x_1, \dots, x_n]^T \in R^n$. Ο πραγματικός χώρος R^n ονομάζεται και χαρακτηριστικός χώρος, όπου κάθε άξονας αντιστοιχεί σε ένα φυσικό χαρακτηριστικό. Η απεικόνιση του πραγματικού αριθμού ($x \in R^n$) απαιτεί μια μεθοδολογία για τη μετατροπή ποιοτικών χαρακτηριστικών σε ποσοτικά. Τυπικά, τέτοιες μεθοδολογίες είναι αρκετά υποκειμενικές και απαιτούν εύρεση. Για παράδειγμα, η διεξαγωγή εξετάσεων είναι μια μεθοδολογία για την ποσοτικοποίηση της προόδου μάθησης των μαθητών. Επίσης, υπάρχουν πάρα πολλά χαρακτηριστικά, τα οποία εμείς οι άνθρωποι μπορούμε να εκτιμήσουμε διαισθητικά, αλλά δύσκολα να τα εξηγήσουμε. Αυτά περιλαμβάνουν την αίσθηση του χιούμορ, της ευφυΐας και της ομορφιάς, χαρακτηριστικά, τα οποία θεωρούμε ότι έχουν αριθμητικές εκφράσεις.

Μερικές φορές ένα αντικείμενο μπορεί να εκπροσωπείται από πολλαπλά υποσύνολα των χαρακτηριστικών. Για παράδειγμα, στην εξακρίβωση της ταυτότητας, μπορούν να χρησιμοποιηθούν τρεις διαφορετικοί τρόποι ανίχνευσης (Kittler et al., 1998): η εμπρόσθια όψη, το προφίλ και η φωνή. Υποσύνολα ειδικών χαρακτηριστικών μετρώνται για κάθε τροποποίηση και στη συνέχεια το χαρακτηριστικό διάνυσμα συντίθεται από τρία υποδιανύσματα, $x = [x^{(1)}, x^{(2)}, x^{(3)}]^T$. Αυτό καλείται διακριτή γραφική αναπαράσταση (distinct pattern representation) σύμφωνα με τους Kittler et al. (Kittler et al., 1998). Όπως θα δούμε και στη συνέχεια, ένα σύνολο ταξινομητών μπορεί να κατασκευαστεί και χρησιμοποιώντας ένα διακριτό πρότυπο εκπροσώπησης, ένα ταξινομητή για κάθε υποσύνολο χαρακτηριστικών.

1.2.4 Σύνολα Δεδομένων

Οι πληροφορίες για το σχεδιασμό ενός ταξινομητή είναι συνήθως υπό τη μορφή ενός επισημασμένου συνόλου δεδομένων $Z = \{z_1, \dots, z_N\}, z_j \in R^n$. Η ετικέτα της τάξης z_j συμβολίζεται ως $l(z_j) \in \Omega, j = 1, 2, \dots, N$.



Σχήμα 1.2.3: Παράδειγμα Χειρόγραφων Ψηφίων

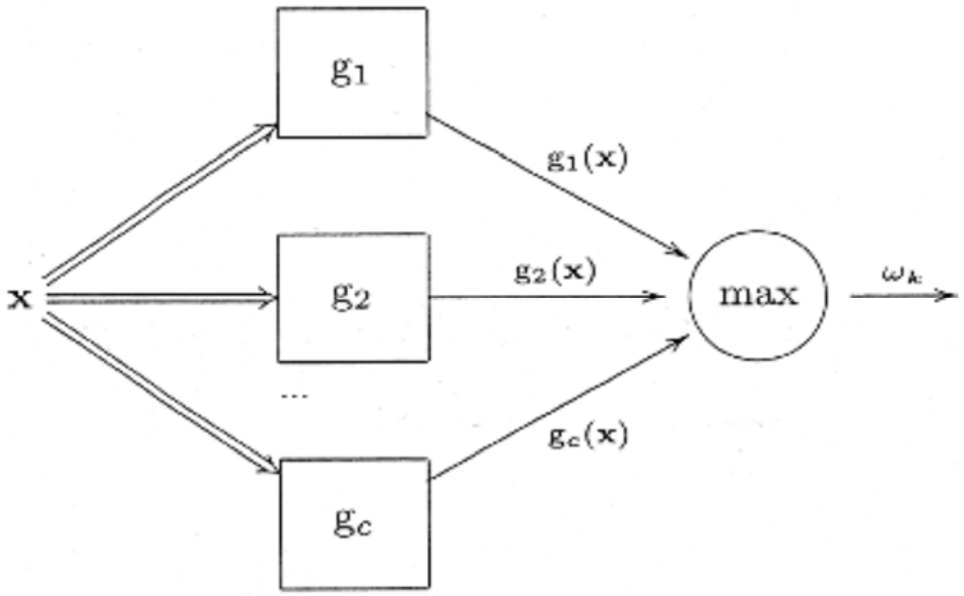
Το Σχήμα 1.2.3 παραπάνω απεικονίζει μια σειρά από παραδείγματα χειρόγραφων ψηφίων, τα οποία πρέπει να επισημανθούν από μηχάνημα σε δέκα κατηγορίες. Για να κατασκευαστεί ένα σύνολο δεδομένων, οι μαύρες και άσπρες εικόνες θα πρέπει να μετατραπούν σε χαρακτηριστικά διανύσματα. Δεν είναι πάντα εύκολο να διαμορφωθούν τα n χαρακτηριστικά, ώστε να χρησιμοποιηθούν στο πρόβλημα. Στο παράδειγμα του τρίτου σχήματος, τα ποικίλα διαφοροποιημένα χαρακτηριστικά, μπορούν να ψηφιστούν, χρησιμοποιώντας επίσης διάφορες μεταμορφώσεις στην εικόνα. Δύο πιθανά χαρακτηριστικά είναι, για παράδειγμα, τα ακαριαία εγκεφαλικά επεισόδια και ο αριθμός των κύκλων στην εικόνα του ψηφίου. Ο προσδιορισμός ενός καλού συνόλου χαρακτηριστικών προκαθορίζει σε μεγάλο βαθμό την επιτυχία ενός

συστήματος αναγνώρισης προτύπων. Θεωρούμε ότι τα χαρακτηριστικά έχουν ήδη οριστεί και έχουμε ένα έτοιμο προς χρήση σύνολο δεδομένων Z.

1.3 Ταξινομητές, Διακρίνουσες Συναρτήσεις και Περιοχές Ταξινόμησης

Ένας ταξινομητής είναι οποιαδήποτε συνάρτηση:

$$D: R^n \rightarrow \Omega \tag{1.3.1}$$



Σχήμα 1.3.1: Κανονικό μοντέλο ενός ταξινομητή. Τα διπλά βέλη δηλώνουν την n-διαστάσεων είσοδο του διανύσματος x, η έξοδος των θέσεων είναι η διακρίνουσα συνάρτηση τιμών $g_i(x)$ (βαθμωτά) και η έξοδος της μέγιστης επιλογής είναι η τάξη που έχει επισημανθεί ως $\omega_k \in \Omega$ και έχει εκχωρηθεί σύμφωνα με τον κανόνα μέγιστης ένταξης.

Στο κανονικό μοντέλο ενός ταξινομητή (Duda and Hart, 1973), το οποίο φαίνεται στο Σχήμα 1.3.1, θεωρούμε ένα σύνολο από c διακριτικές συναρτήσεις (discriminant functions) $G = \{g_1(x), \dots, g_c(x)\}$,

$$g_i: R^n \rightarrow R, i = 1, \dots, c \tag{1.3.2}$$

καθεμία από τις οποίες αποδίδει μια βαθμολογία για την αντίστοιχη τάξη. Τυπικά (και πιο φυσικά) το x επισημαίνεται στην τάξη με τη μεγαλύτερη βαθμολογία. Η επιλογή αυτή επισήμανσης ονομάζεται κανόνας μέγιστης ένταξης και είναι

$$D(x) = \omega_{i^*} \in \Omega \leftrightarrow g_{i^*}(x) = \max_{i=1,\dots,c} \{g_i(x)\} \quad (1.3.3)$$

Οι δεσμοί σπάνε τυχαία, δηλαδή, το x είναι τυχαία δεμένο σε μια από τις τάξεις. Οι διακρίνουσες συναρτήσεις κατανέμουν το χώρο των χαρακτηριστικών R^n σε c (όχι απαραίτητα συμπαγή) διαστήματα απόφασης ή περιοχές ταξινόμησης που συμβολίζονται με $R_1', \dots, R_c'$

$$R_i' = \{x | x \in R^n, g_i(x) = \max_{k=1,\dots,c} g_k(x)\}, i = 1, \dots, c \quad (1.3.4)$$

Το διάστημα απόφασης για την τάξη ω_i είναι ένα σύνολο σημείων για τα οποία οι i-οστές διακρίνουσες συναρτήσεις έχουν την υψηλότερη βαθμολογία. Σύμφωνα με τον κανόνα μέγιστης ένταξης (1.3.3), όλα τα σημεία στην περιοχή R_i' αποδίδονται στην τάξη ω_i. Τα διαστήματα απόφασης καθορίζονται από τον ταξινομητή D, ή ισοδύναμα, από τις διακρίνουσες συναρτήσεις G. Τα όρια των διαστημάτων απόφασης, που ονομάζονται όρια ταξινόμησης, και περιλαμβάνουν τα σημεία για τα οποία εφαρμόζονται οι υψηλότερες διακρίνουσες συναρτήσεις. Ένα σημείο πάνω στο όριο μπορεί να αποδοθεί σε κάποια από τις οριακές τάξεις. Αν ένα διάστημα απόφασης R_i' περιέχει δεδομένα σημεία από το σύνολο Z, με αληθινή τάξη $\omega_j, j \neq i$, οι τάξεις ω_i και ω_j καλούνται επικαλύψεις. Πρέπει να σημειωθεί ότι οι επικαλυπτόμενες τάξεις για ένα συγκεκριμένο διάστημα του χώρου των χαρακτηριστικών (που ορίζεται από έναν ορισμένο ταξινομητή D) μπορεί να είναι μη-επικαλυπτόμενες, εάν ο χώρος των χαρακτηριστικών κατανεμηθεί με άλλο τρόπο. Αν στο Z δεν υπάρχουν πανομοιότυπα σημεία με διαφορετικές επισημασμένες τάξεις, μπορούμε να κατανείμουμε το χώρο των χαρακτηριστικών σε περιοχές ταξινόμησης, έτσι ώστε οι τάξεις να είναι μη επικαλυπτόμενες. Γενικά, όσο μικρότερη είναι η επικάλυψη, τόσο καλύτερος είναι ο ταξινομητής.

Παράδειγμα:

Περιοχές Ταξινόμησης. Ένα πρόβλημα 15 σημείων και δύο τάξεων απεικονίζεται στο Σχήμα 1.3.2. Ο χώρος των χαρακτηριστικών R^2 χωρίζεται σε δύο περιοχές ταξινόμησης: ο R_1' είναι σκιασμένος (τάξη ω_1: τετράγωνα) και ο R_2' είναι μη σκιασμένος (τάξη ω_2: κουκίδες). Για δύο τάξεις μπορούμε να χρησιμοποιήσουμε μόνο μια διακρίνουσα συνάρτηση αντί για δύο:

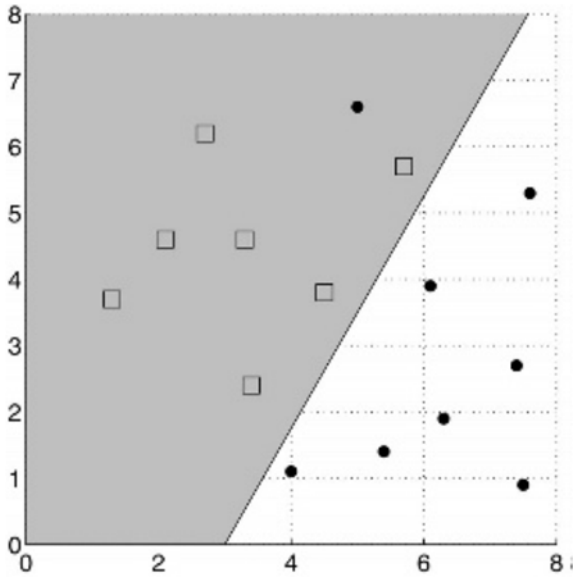
$$g(x) = g_1(x) - g_2(x) \quad (1.3.5)$$

και να εκχωρήσει την τάξη ω_1 εάν η $g(x)$ είναι θετική και την τάξη ω_2 εάν είναι αρνητική. Για το συγκεκριμένο παράδειγμα, έχουμε σχεδιάσει τα όρια ταξινόμησης που παράγονται από τη γραμμική διακρίνουσα συνάρτηση:

$$g(x) = -7x_1 + 4x_2 + 21 = 0 \quad (1.3.6)$$

Πρέπει να σημειωθεί ότι κάθε γραμμή στον R^2 είναι μια γραμμική διακρίνουσα συνάρτηση για οποιοδήποτε πρόβλημα δύο τάξεων στον R^2. Γενικά, οποιοδήποτε σύνολο συναρτήσεων $\{g_1(x), \dots, g_c(x)\}$ (γραμμικό ή μη-γραμμικό) είναι ένα σύνολο από διακρίνουσες συναρτήσεις. Είναι ένας άλλος τρόπος για το πώς αυτές οι διακρίνουσες συναρτήσεις θα διαχωρίσουν με επιτυχία τις κλάσεις.

Έστω $G^* = \{g_1^*(x), \dots, g_c^*(x)\}$ ένα σύνολο από βέλτιστες (κατά κάποιο τρόπο) διακρίνουσες συναρτήσεις. Μπορούμε να αποκτήσουμε απείρως πολλά σύνολα βέλτιστων διακρινουσών συναρτήσεων G^* με την εφαρμογή ενός μετασχηματισμού $f(g_i^*(x))$ που να διατηρεί τις τιμές της συνάρτησης για κάθε $x \in R^n$. Για παράδειγμα, το $f(\zeta)$ μπορεί να είναι $\log(\zeta)$, $\sqrt{\zeta}$ για θετικά ορισμένα $g^*(x)$, $\alpha^{\zeta}, \gamma\iota\alpha\ \alpha > 1$ κ.λ.π. Εφαρμόζοντας την ίδια f σε όλες τις διακρίνουσες συναρτήσεις στο G^*, παίρνουμε ένα ισοδύναμο σύνολο από διακρίνουσες συναρτήσεις. Χρησιμοποιώντας τον κανόνα μέγιστης ένταξης (1.3.3), το x θα επισημαίνεται στην ίδια τάξη με οποιοδήποτε από τα ισοδύναμα σύνολα διακρινουσών συναρτήσεων.



Σχήμα 1.3.2: Ένα δύο τάξεων παράδειγμα μιας γραμμικής διακρίνουσας συνάρτησης.

Εάν οι κλάσεις στο Z μπορούν να διαχωριστούν πλήρως από κάθε άλλη με ένα υπερεπίπεδο (ένα σημείο στον R, μια γραμμή στον R^2 και ένα επίπεδο στον R^3), καλούνται γραμμικώς διαχωρίσιμα. Οι δύο κλάσεις στο Σχήμα 1.3.2 δεν είναι γραμμικώς διαχωρίσιμες εξαιτίας της κουκίδας στο σημείο (5, 6.6), η οποία είναι στη λάθος πλευρά της διακρίνουσας συνάρτησης.

1.4 Σφάλμα και Ακρίβεια Ταξινόμησης

Είναι πολύ σημαντικό να γνωρίζουμε πόσο καλά αποδίδει ο ταξινομητής μας. Η επίδοση του ταξινομητή μας είναι μια ένωση χαρακτηριστικών, της οποίας η πιο σημαντική συνιστώσα είναι η ακρίβεια. Αν ήμασταν σε θέση να δοκιμάσουμε τον ταξινομητή μας σε κάθε πιθανή είσοδο αντικειμένων, θα ξέραμε ακριβώς πόσο ακριβής είναι. Δυστυχώς αυτό δεν είναι σχεδόν καθόλου πιθανό σενάριο, έτσι μια εκτίμηση της ακρίβειας πρέπει να χρησιμοποιείται αντ'αυτού.

1.4.1 Υπολογισμός Σφάλματος

Έστω ότι έχουμε διαθέσιμο ένα σύνολο δεδομένων με ετικέτα Z_{ts} μεγέθους $N_{ts} \times n$ για τη δοκιμή της ακρίβειας του ταξινομητή μας, D. Ο πιο φυσικός τρόπος για να υπολογίσουμε την εκτίμηση του σφάλματος είναι να τρέξουμε τον ταξινομητή D σε όλα τα αντικείμενα στο Z_{ts} και να βρούμε το ποσοστό των λανθασμένων αντικειμένων

$$Error(D) = \frac{N_{error}}{N_{ts}} \tag{1.4.1}$$

όπου N_{error} είναι ο αριθμός των εσφαλμένων ταξινομήσεων που διαπράχθηκαν από τον ταξινομητή D. Αυτό ονομάζεται καταμέτρηση εκτιμητή για το ποσοστό σφάλματος, επειδή βασίζεται στην καταμέτρηση των εσφαλμένων ταξινομήσεων. Έστω $s_j \in \Omega$ μια τάξη με ετικέτα, που αποδίδεται από τον εκτιμητή D στο αντικείμενο z_j. Η καταμέτρηση του εκτιμητή μπορεί να γραφεί ως

$$Error(D) = \frac{1}{N_{ts}} \sum_{j=1}^{N_{ts}} \{1 - I(l(z_j), s_j)\}, z_j \in Z_{ts} \tag{1.4.2}$$

όπου $I(a, b)$ είναι ένας δείκτης συνάρτησης που παίρνει την τιμή 1, όταν $a = b$ και 0, όταν $a \neq b$. Επίσης, σφάλμα, $Error(D)$, καλείται και το φαινομενικό ποσοστό σφάλματος. Διπλή σε αυτό το χαρακτηριστικό είναι και η φαινομενική ακρίβεια ταξινόμησης, η οποία υπολογίζεται με $1 - Error(D)$.

Για να δούμε το σφάλμα από άποψη πιθανοτήτων, μπορούμε να υιοθετήσουμε το ακόλουθο μοντέλο. Ο ταξινομητής διαπράττει σφάλμα με πιθανότητα P_D σε κάθε αντικείμενο $x \in R^n$ (μια λάθος, αλλά χρήσιμη υπόθεση). Στη συνέχεια, ο αριθμός των σφαλμάτων ακολουθεί τη διωνυμική κατανομή με παραμέτρους (P_D, N_{ts}). Μια εκτίμηση του P_D είναι

$$\widehat{P_D} = \frac{N_{error}}{N_{ts}} \tag{1.4.3}$$

το οποίο στην πραγματικότητα είναι η καταμέτρηση του σφάλματος, $Error(D)$ που ορίστηκε προηγουμένως. Αν οι N_{ts} και P_D ικανοποιούν τον εμπειρικό κανόνα:

$$N_{ts} > 30, \widehat{P_D} \times N_{ts} > 5 \; και \; \left(1 - \widehat{P_D}\right) \times N_{ts} > 5$$

τότε η διωνυμική κατανομή μπορεί να προσεγγιστεί από την κανονική κατανομή. Το 95% διάστημα εμπιστοσύνης για το σφάλμα είναι

$$[\widehat{P_D} - 1.96\sqrt{\frac{\widehat{P_D}\left(1 - \widehat{P_D}\right)}{N_{ts}}}, \widehat{P_D} + 1.96\sqrt{\frac{\widehat{P_D}(1 - \widehat{P_D})}{N_{ts}}}] \tag{1.4.4}$$

Υπάρχουν οι λεγόμενες τροποποιήσεις εξομάλυνσης του εκτιμητή καταμέτρησης (Glick, 1978), των οποίων ο σκοπός είναι η μείωση της διακύμανσης του εκτιμητή του P_D. Η δυαδική ένδειξη της συνάρτησης $I(a, b)$ στη σχέση (1.4.2) μπορεί να αντικατασταθεί από μια συνάρτηση εξομάλυνσης λαμβάνοντας τιμές στο διάστημα $[0, 1] \subset R$.

1.4.2 Κατάρτιση και Έλεγχος Συνόλων Δεδομένων

Έστω ότι έχουμε ένα σύνολο δεδομένων Z μεγέθους $N \times n$, που περιέχει $n - διαστάσεων$ χαρακτηριστικά που περιγράφουν N αντικείμενα. Θα θέλαμε να χρησιμοποιήσουμε όσο το δυνατόν περισσότερο τα στοιχεία για την κατασκευή του ταξινομητή (εκπαίδευση), αλλά και όσο το δυνατόν μη ορατά δεδομένα για τον έλεγχο της απόδοσης εκτενέστερα (έλεγχος). Όμως, αν χρησιμοποιήσουμε όλα τα στοιχεία για την εκπαίδευση και τα ίδια για τον έλεγχο, θα μπορούσαμε να επανεκπαιδεύσουμε τον ταξινομητή, έτσι ώστε να μαθαίνει τέλεια τα διαθέσιμα δεδομένα και να παραλείπει τα μη ορατά. Για το λόγο αυτό είναι σημαντικό να έχουμε ένα ξεχωριστό σύνολο δεδομένων, στο οποίο θα εξετάζουμε το τελικό προϊόν. Οι κυριότερες εναλλακτικές για την καλύτερη χρήση των Z συνοψίζονται ως εξής:

- Επανα-αντικατάσταση (Resubstitution) (Μέθοδος R). Σχεδιασμός ενός ταξινομητή D στο Z και έλεγχός του στο Z. Το $\widehat{P_D}$ είναι θετικά μεροληπτικό.
- Αντοχή (Μέθοδος H). Παραδοσιακά, η διάσπαση του Z στη μέση, χρησιμοποιεί το ένα μισό για την εκπαίδευση και το άλλο μισό για τον υπολογισμό του $\widehat{P_D}$. Το $\widehat{P_D}$ είναι αρνητικά μεροληπτικό. Χωρίζεται σε άλλες αναλογίες που επίσης χρησιμοποιούνται. Μπορούμε να ανταλλάξουμε τα δύο υποσύνολα, να πάρουμε μια διαφορετική εκτίμηση για το $\widehat{P_D}$ και να βγάλουμε το μέσο όρο των δύο. Μια παραλλαγή αυτής της μεθόδου είναι το ανακάτεμα των δεδομένων, όπου πραγματοποιούμε L τυχαίες διασπάσεις του Z στην εκπαίδευση και τον έλεγχο των τμημάτων και να βγάλουμε το μέσο όρο όλων των εκτιμήσεων L του $\widehat{P_D}$ που υπολογίζεται στα αντίστοιχα τμήματα δοκιμών.
- Διασταυρωμένη επικύρωση (ονομάζεται επίσης και μέθοδος περιστροφής, orp-method). Επιλέγουμε έναν ακέραιο K (κατά προτίμηση ένα παράγοντα του N) και τυχαία χωρίζουμε το Z σε K υποσύνολα μεγέθους N/K. Στη συνέχεια, χρησιμοποιούμε ένα υποσύνολο για να ελέγξουμε την απόδοση των D εκπαιδεύοντας την ένωση των $K-1$ υποσυνόλων που έχουν απομείνει. Αυτή η διαδικασία επαναλαμβάνεται K φορές, επιλέγοντας ένα διαφορετικό μέρος για έλεγχο κάθε φορά. Για να πάρουμε την τελική τιμή του $\widehat{P_D}$ παίρνουμε το μέσο όρο των εκτιμήσεων K. Όταν $K = N$, η μέθοδος ονομάζεται leave-one-out (ή μέθοδος U).
- Bootstrap. Αυτή η μέθοδος έχει σχεδιαστεί για να διορθώσει την θετική μεροληψία της μεθόδου R. Αυτό γίνεται από τη δημιουργία τυχαίων L συνόλων πληθικότητας N από το αρχικό σύνολο Z, με επανατοποθέτηση. Τότε εκτιμάμε και βρίσκουμε το μέσο όρο του ποσοστού του σφάλματος των ταξινομητών που έχει χτιστεί σε αυτά τα σύνολα.

Το ερώτημα σχετικά με τον καλύτερο τρόπο για να οργανώσουμε το πείραμα εκπαίδευσης και ελέγχου έχει απασχολήσει εδώ και πολύ καιρό (Toussaint, 1974). Η αναγνώριση προτύπων έχει πλέον ξεπεράσει το στάδιο όπου ο υπολογισμός των πηγών ήταν ο καθοριστικός παράγοντας για το ποια μέθοδο θα χρησιμοποιήσουμε. Ωστόσο, ακόμα και με τη σύγχρονη τεχνολογία των υπολογιστών, το πρόβλημα δεν έχει εξαφανιστεί. Τα ολοένα και αυξανόμενα μεγέθη των συνόλων δεδομένων που συλλέγονται σε διάφορους τομείς της επιστήμης και της πρακτικής συνιστούν μια νέα πρόκληση. Είμαστε πίσω στη χρήση της παλιάς μεθόδου hold-out, πρώτον επειδή οι άλλες μπορεί να είναι πολύ χρονοβόρες, και δεύτερον, επειδή η ποσότητα των δεδομένων μπορεί να είναι τόσο υπερβολική ώστε μικρά τμήματα αρκούν για την εκπαίδευση και τον έλεγχο. Για παράδειγμα, ας θεωρήσουμε ένα σύνολο δεδομένων που λαμβάνεται από την ανάλυση λιανικής πώλησης, το οποίο περιλαμβάνει εκατοντάδες χιλιάδες συναλλαγές. Χρησιμοποιώντας μια εκτίμηση του σφάλματος πάνω από, ας πούμε, 10.000 σημεία δεδομένων, μπορεί εύκολα να

συρρικνωθεί το διάστημα εμπιστοσύνης και να κάνει την εκτίμηση αρκετά αξιόπιστη.

Καθίσταται πλέον μια κοινή πρακτική να χρησιμοποιούνται τρία αντί για δύο σύνολα δεδομένων: ένα για την εκπαίδευση, ένα για την επικύρωση και ένα για τη δοκιμή. Όπως και προηγουμένως, το σύνολο της δοκιμής παραμένει αόρατο κατά τη διάρκεια της διαδικασίας εκπαίδευσης. Το σύνολο δεδομένων της επικύρωσης δρα ως ψευδο-δοκιμή. Συνεχίζουμε τη διαδικασία εκπαίδευσης μέχρι η βελτίωση των συνόλων εκπαίδευσης να μην συνδυάζεται πλέον με τη βελτίωση επιδόσεων στο σύνολο επικύρωσης. Στο σημείο αυτό η εκπαίδευση θα πρέπει να σταματήσει έτσι ώστε να αποφευχθεί η υπερβολική εκπαίδευση. Δεν είναι όλα τα σύνολα δεδομένων αρκετά μεγάλα ώστε να επιτρέψουν σε ένα μέρος επικύρωσης να κοπεί. Πολλά από τα σύνολα δεδομένων από τη βάση δεδομένων UCI Machine Learning Repository (στη διεύθυνση http://www.ics.uci.edu/~mlearn/MLRepository.html), συχνά χρησιμοποιούνται ως σημεία αναφοράς στο σημείο αναγνώρισης και μηχανικής μάθησης, αλλά μπορεί να είναι ακατάλληλα για ένα τριπλό χωρισμό σε εκπαίδευση/ επικύρωση/ έλεγχο. Ο λόγος είναι ότι τα υποσύνολα δεδομένων θα είναι υπερβολικά μικρά και οι εκτιμήσεις του σφάλματος στις υποομάδες αυτές θα είναι αναξιόπιστες. Στη συνέχεια, το σταμάτημα της εκπαίδευσης στο σημείο που προτείνεται από το σύνολο επικύρωσης μπορεί να είναι ανεπαρκές, η εκτίμηση της ακρίβειας ελέγχου μπορεί να είναι ανακριβής και ο ταξινομητής μπορεί να είναι φτωχός εξαιτίας της ανεπάρκειας των δεδομένων εκπαίδευσης.

Όταν οι πολλαπλές εκπαιδεύσεις και έλεγχοι πραγματοποιήθηκαν, υπάρχει το ζήτημα σχετικά με το ποιος από τους ταξινομητές που χτίστηκε κατά τη διάρκεια αυτής της διαδικασίας θα πρέπει να χρησιμοποιηθεί στο τέλος. Για παράδειγμα σε μια δεκαπλάσια διασταυρωμένη επικύρωση, χτίζουμε δέκα διαφορετικούς ταξινομητές χρησιμοποιώντας διαφορετικά σύνολα δεδομένων. Οι παραπάνω μέθοδοι έχουν σημασία μόνο για να μας δώσουν μια εκτίμηση της ακρίβειας ενός συγκεκριμένου μοντέλου που κατασκευάστηκε για το πρόβλημα στο χέρι. Βασιζόμαστε στην υπόθεση ότι η ακρίβεια ταξινόμησης θα αλλάξει ομαλά με τις αλλαγές στο μέγεθος των δεδομένων εκπαίδευσης (Dietterich, 1978). Ως εκ τούτου, αν είμαστε ικανοποιημένοι με την ακρίβεια και τη μεταβλητότητά της σε διαφορετικές υποομάδες εκπαίδευσης, μπορεί να αποφασίσουμε τελικά να εκπαιδεύσουμε ένα ενιαίο ταξινομητή για όλο το σύνολο δεδομένων. Εναλλακτικά, μπορούμε να κρατήσουμε τους ταξινομητές κατασκευασμένους σε όλη την εκπαίδευση και να εξετάσουμε τη χρήση τους μαζί σε ένα σύνολο.

1.4.3 Πίνακες Σύγχυσης και Πίνακες Απώλειας

Για να μάθουμε πώς τα λάθη διανέμονται σε όλες τις κλάσεις κατασκευάζουμε ένα πίνακα σύγχυσης χρησιμοποιώντας το σύνολο δεδομένων ελέγχου, Z_{ts}. Η είσοδος α_{ij} ενός τέτοιου πίνακα υποδηλώνει τον αριθμό των στοιχείων τουZ_{ts}, των οποίων η πραγματική κλάση είναι ω_i, και τα οποία αντιστοιχίζονται από το D στην κλάση ω_j.

Ο πίνακας σύγχυσης για το γραμμικό ταξινομητή για τα δεδομένα 15 σημείων απεικονίζεται στον Πίνακα 1.4.1.

Πίνακας 1.4.1: Πίνακας σύγχυσης

True Class	$D(x)$	
	ω_1	ω_2
ω_1	7	0
ω_2	1	7

Η εκτίμηση της ακρίβειας του ταξινομητή μπορεί να υπολογιστεί ως το ίχνος του πίνακα διαιρούμενο με το συνολικό άθροισμα των εισόδων, $(7+7)/15$ στη συγκεκριμένη περίπτωση. Οι πρόσθετες πληροφορίες που παρέχει ο πίνακας σύγχυσης είναι όπου έχουν προκληθεί λανθασμένες ταξινομήσεις. Αυτό είναι σημαντικό για προβλήματα με μεγάλο αριθμό κλάσεων επειδή μια μεγάλη μη διαγώνια είσοδος ενός πίνακα μπορεί να υποδηλώνει ένα δύσκολο δύο τάξεων πρόβλημα, το οποίο χρειάζεται να αντιμετωπιστεί ξεχωριστά.

Παράδειγμα: Πίνακας Σύγχυσης για Δεδομένα Επιστολής. Ένα σύνολο δεδομένων επιστολής που είναι διαθέσιμο από το UCI Machine Learning Repository Database περιέχει δεδομένα που εξάγονται από 20000 ασπρόμαυρες εικόνες από κεφάλαια του English Letters. Δεκαέξι αριθμητικά χαρακτηριστικά περιγράφουν κάθε εικόνα $(N = 20000, c = 26, n = 16)$. Για το σκοπό αυτής της εικόνας θα χρησιμοποιηθεί η μέθοδος hold-out. Το σύνολο δεδομένων διασπάται τυχαία στη μέση. Το ένα μισό χρησιμοποιήθηκε για την εκπαίδευση ενός γραμμικού ταξινομητή και το άλλο μισό για τον έλεγχο. Οι ετικέτες των δεδομένων ελέγχου αντιστοιχήθηκαν με τις ετικέτες που λήφθηκαν από τον ταξινομητή και ο 26×26 πίνακας σύγχυσης κατασκευάστηκε. Εάν ο ταξινομητής ήταν ιδανικός και όλες οι ετικέτες ταίριαζαν, ο πίνακας σύγχυσης θα ήταν διαγώνιος.

Ο Πίνακας 1.4.1 παρουσιάζει τη γραμμή του πίνακα σύγχυσης που αντιστοιχεί στην κλάση “H”. Οι είσοδοι δείχνουν πόσες φορές η αληθής κλάση “H” είναι λάθος για το έγγραφο στην αντίστοιχη στήλη. Ο έντονα γραμμένος αριθμός είναι η διαγώνια

είσοδος που δείχνει πόσες φορές η κλάση “H” έχει αναγνωριστεί σωστά. Έτσι, από το σύνολο των 379 παραδειγμάτων του “H” στο σύνολο ελέγχου, μόνο οι 165 έχουν επισημανθεί σωστά από τον ταξινομητή. Περιέργως, ο μεγαλύτερος αριθμός λαθών, 37, είναι για το γράμμα “O”.

Τα σφάλματα στην ταξινόμηση δεν είναι εξίσου δαπανηρά. Για να ληφθούν υπόψη τα διαφορετικά κόστη των λαθών εισάγουμε τον πίνακα ζημιών. Ορίζουμε ένα πίνακα ζημιών (Πίνακας 1.4.2) με εισόδους λ_{ij} που δηλώνουν τη ζημιά που υπέστη από την αντίστοιχη ετικέτα ω_i, δεδομένου ότι η πραγματική ετικέτα του αντικειμένου είναι ω_j.

Πίνακας 1.4.2: Η “H”-γραμμή στον πίνακα σύγχυσης για το σύνολο δεδομένων γραμμάτων που λαμβάνονται από ένα γραμμικό ταξινομητή και έχουν πραγματοποιηθεί σε 10000 σημεία.

“H” mistaken for:	A	B	C	D	E	F	G	H	I	J	K	L	M
No of times:	2	12	0	27	0	2	1	165	0	0	26	0	1
“H” mistaken for:	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
No of times:	31	37	4	8	17	1	1	13	3	1	27	0	0

Αν ο ταξινομητής δεν είναι σίγουρος για την ετικέτα, μπορεί να αρνηθεί να πάρει μια απόφαση. Μια επιπλέον κατηγορία (που ονομάζεται refuse-to-decide) συμβολίζεται με ω_{c+1} μπορεί να προστεθεί στο Ω. Επιλέγοντας την ω_{c+1} θα πρέπει να είναι λιγότερο δαπανηρή από την επιλογή μιας λανθασμένης κλάσης. Για ένα πρόβλημα με c αρχικές κλάσεις και μια επιλογή άρνησης, ο πίνακας ζημιών θα είναι μεγέθους $(c+1) \times c$. Οι πίνακες ζημιών καθορίζονται συνήθως από το χρήστη. Ένας μηδέν-ένα $(0-1)$ πίνακας ζημιών ορίζεται ως $\lambda_{ij} = 0\ \gamma\iota\alpha\ i = j\ \kappa\alpha\iota\ \lambda_{ij} = 1\ \gamma\iota\alpha\ i \neq j$, το οποίο σημαίνει ότι όλα τα σφάλματα είναι εξίσου δαπανηρά.

1.4.4 Καμπύλες ROC (Receiver Operating Characteristic)

Στην στατιστική μια καμπύλη ROC είναι ένα διάγραμμα, το οποίο απεικονίζει, οργανώνει και επιλέγει ταξινομητές με βάση την απόδοσή τους. Οι καμπύλες αυτές αποτελούν μια τυποποιημένη μέθοδο, η οποία συνοψίζει την απόδοση ενός ταξινομητή συσχετίζοντας το μεταξύ εναλλαγών αληθώς θετικών (TP) και ψευδώς θετικών (FP) ποσοστό σφαλμάτων. Στην ουσία μια καμπύλη ROC είναι ένα δυαδικό σύστημα $[0,1] \times [0,1]$, ξεκινώντας από το $(0,0)$, το οποίο αποτελεί το σημείο απόφασης που είναι μεγαλύτερο από όλες τις μετρήσεις σήματος και θορύβου και

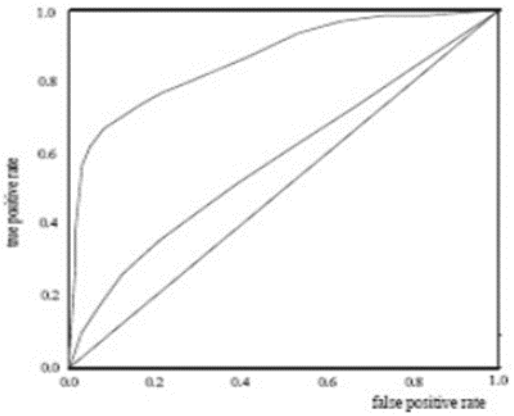
καταλήγοντας στο $(1,1)$, το οποίο αποτελεί το σημείο απόφασης που είναι μικρότερο από όλες τις μετρήσεις.

Στο σημείο αυτό είναι απαραίτητο να οριστεί μια περιοχή, η οποία βρίσκεται κάτω από την καμπύλη ROC, η περιοχή AUC, η οποία και αποτελεί ένα αποδεκτό σύστημα μέτρησης για μια καμπύλη ROC. Η μικρότερη τιμή που μπορεί να πάρει η (AUC) είναι 0.5 καθώς είμαστε στο δυαδικό $[0,1] \times [0,1]$ και μια τυχαία εικασία μπορεί να παραχθεί μεταξύ του $(0,0)$ και του $(1,1)$. Επιπλέον, η (AUC) συνδέεται άμεσα με τον δείκτη Gini για τον οποίο ισχύει, σύμφωνα με τους Hang και Till (2001):

$$GINI = 2 \times AUC - 1 \tag{1.4.5}$$

Οι καμπύλες ROC αποτελούν την καλύτερη επιλογή για αποφάσεις σχετικά με τις (TP) και (FP). Αυτό συμβαίνει καθώς το εμβαδόν που σχηματίζεται κάτω από την καμπύλη αποτελεί ένα μέτρο ποιότητας για το διαχωρισμό σήματος-θορύβου και το οποίο αποτελεί βασικό εργαλείο για τη στατιστική συμπερασματολογία.

Παρακάτω παρουσιάζεται ένα διάγραμμα ROC. Η επιφάνεια κάτω από την καμπύλη ROC αποτελεί όπως ήδη αναφέρθηκε παραπάνω την περιοχή (AUC). Όσο πιο κοντά βρισκόμαστε στην διαγώνιο, τόσο λιγότερο ακριβές είναι το μοντέλο. Ο κάθετος άξονας αποτελεί το επίπεδο αληθώς θετικών και ο οριζόντιος άξονας το επίπεδο των ψευδώς θετικών. Το μοντέλο είναι πιο ακριβές όταν το εμβαδόν προσεγγίζει την τιμή 1.0.



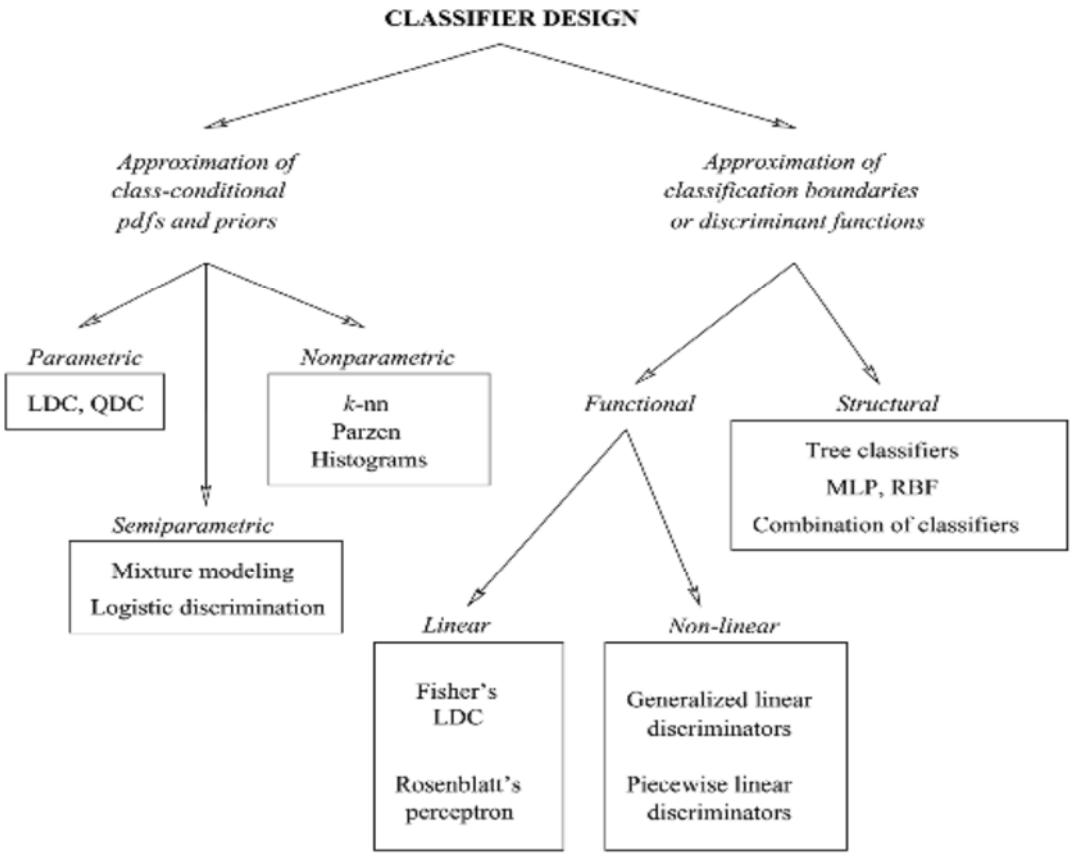
Σχήμα 1.4.1: Μορφή καμπύλης ROC.

1.5 Ταξινόμηση των Μεθόδων Σχεδιασμού ενός Ταξινομητή

Με δεδομένο ότι στα προβλήματα της πραγματικής ζωής συνήθως δεν γνωρίζουμε τις πραγματικές εκ των προτέρων πιθανότητες, ούτε τις συνθήκες των συναρτήσεων πυκνότητας πιθανότητας (σ.π.π./pdf) που μελετάμε, μπορούμε να σχεδιάσουμε

μόνο λανθασμένες εκδοχές του ταξινομητή Bayes. Η στατιστική αναγνώριση προτύπων παρέχει μια ποικιλία των μοντέλων του ταξινομητή (Devijver and Kittler, 1978, Duda and Hart, 1973, Fukunaga, 1972, Patrick, 1972, Tou and Gonzalez, 1974, Bishop, 1995, Looney, 1997, Ripley, 1996). Το Σχήμα 1.5.1 δείχνει μια πιθανή ταξινόμηση των μεθόδων σχεδιασμού ενός ταξινομητή. Τα κουτιά περιέχουν αντιπροσωπευτικά μοντέλα ταξινόμησης από τις αντίστοιχες κατηγορίες.

Μια λύση είναι να προσπαθήσουμε να εκτιμήσουμε το $P(\omega_i)$ και το $p(x|\omega_i), i = 1, \ldots, c$ από το Z και αντικαθιστά τις εκτιμήσεις $\hat{P}(\omega_i)$ *και* $\hat{p}(x|\omega_i)$ στις διακριτικές συναρτήσεις (discriminant functions) $g_i(x) = P(\omega_i)p(x|\omega_i), i = 1, \ldots, c.$ Αυτό ονομάζεται προσέγγιση plug-in για το σχεδιασμό του ταξινομητή. Η προσέγγιση $p(x|\omega_i)$ ως συνάρτηση του x χωρίζει τις μεθόδους σχεδιασμού του ταξινομητή σε δύο μεγάλες ομάδες: τις παραμετρικές και τις μη παραμετρικές. Στο γράφημα που ακολουθεί (Σχήμα 1.5.1) παρουσιάζονται οι μέθοδοι σχεδιασμού του ταξινομητή που προέρχονται τόσο μέσω της προσέγγισης των σ.π.π., αλλά επίσης και εμπειρικά μέσω της προσέγγισης των ορίων απόφασης ταξινομητή ή των διακριτικών συναρτήσεων.



Σχήμα 1.5.1: Μια ταξινόμηση των μεθόδων σχεδιασμού του ταξινομητή.

Η διάκριση μεταξύ των ομάδων δεν είναι σαφής. Για παράδειγμα, τα δίκτυα λειτουργίας ακτινικής βάσης (radial basis function networks ή RBF) από την ομάδα των δομικών προσεγγίσεων των διακριτικών συναρτήσεων μπορούν να μετακινηθούν προς την ομάδα των λειτουργικών προσεγγίσεων ή ακόμα και στην ομάδα των ημιπαραμετρικών μοντέλων σ.π.π (Traven, 1991). Ομοίως, η πιο κοντινή γειτονική k (k-nn) μέθοδος, παρόλο που θεωρητικά συνδέεται με την μη παραμετρική εκτίμηση σ.π.π, παράγει μια άμεση εκτίμηση των διακριτικών συναρτήσεων και μπορεί να τεθεί υπό την κλάση των διαρθρωτικών σχεδίων για την προσέγγιση της διακριτικής συνάρτησης.

Δεν υπάρχει συναίνεση για μια ενιαία ταξινόμηση, ή ακόμα και για τον ορισμό των παραμετρικών και μη παραμετρικών ταξινομητών. Ο Lippmann (1991) απαριθμεί πέντε είδη ταξινομητών:

- Πιθανολογικός (LDC, QDC, Parzen)
- Παγκόσμιος (πολυστρωματικό νευρωνικό δίκτυο, MLP)
- Τοπικός (ακτινική λειτουργία βάσης νευρωνικών δικτύων, RBF)
- Τύπος Πλησιέστερου Γείτονα (k-nn, μαθαίνοντας νευρωνικά δίκτυα διανυσματικής κβάντωσης, LVQ)
- Αποκλεισμός Μορφοποίησης (δέντρα δυαδικής απόφασης, συστήματα βάσει κανόνων).

Ο Holmström et al. (Holmström et al., 1997) θεωρεί μια άλλη ομαδοποίηση:

- Ταξινομητές που βασίζονται στην Εκτίμηση της Πυκνότητας:
  - Παραμετρικοί (LDC, QDC)
  - Μη Παραμετρικοί (k-nn, μέθοδος πυρήνα, πεπερασμένα μείγματα, RBF).
- Ταξινομητές που βασίζονται στην Παλινδρόμηση:
  - Παραμετρικοί (γραμμική παλινδρόμηση, λογιστική παλινδρόμηση, MLP)
  - Μη Παραμετρικοί (επιδίωξη προβολής, πρόσθετα μοντέλα).
- Άλλους Ταξινομητές (για παράδειγμα, το πρωτότυπο που βασίζεται: LVQ, k-nn για μικρά k).

Μερικοί συγγραφείς διακρίνουν μεταξύ νευρωνικών και μη νευρωνικών ταξινομητών, τους τοπικούς και παγκόσμιους ταξινομητές και ούτω καθεξής.

1.6 Γραμμικοί και Τετραγωνικοί Ταξινομητές

Οι γραμμικοί και τετραγωνικοί ταξινομητές πήραν το όνομά τους από τον τύπο της διακριτικής συνάρτησης που χρησιμοποιούσαν. Έτσι, κάθε σύνολο γραμμικών συναρτήσεων $g_i: R^n \to R, i = 1, \dots, c$,

$$g_i(x) = w_{i0} + w_i^T x, \qquad x, w_i \in R^n, w_{i0} \in R \tag{1.6.1}$$

μπορεί να θεωρηθεί ως ένας γραμμικός ταξινομητής.

1.6.1 Γραμμικός Διακριτικός Ταξινομητής

Η εκπαίδευση των γραμμικών ταξινομητών έχει μελετηθεί αυστηρά στη βιβλιογραφία της αναγνώρισης προτύπων (Duda and Hart, 1973), που χρονολογείται από τη γραμμική διακριτική του Fisher, το 1936 (Fisher, 1936). Παρακάτω παράγουμε το γραμμικό ταξινομητή ως ταξινομητή ελάχιστου σφάλματος (Bayes) για κανονικά κατανεμημένες τάξεις με ίσους πίνακες συνδιακύμανσης. Θα ονομάζουμε αυτό το μοντέλο γραμμικά διακριτικό ταξινομητή (LDC). Ο LDC είναι απλό να υπολογιστεί από τα δεδομένα και είναι αρκετά ισχυρός, τα αποτελέσματα μπορεί να είναι εκπληκτικά καλά ακόμα και όταν οι κλάσεις δεν έχουν κανονικές κατανομές.

Οποιοδήποτε σύνολο διακριτικών συναρτήσεων που λαμβάνεται από ένα μονότονο μετασχηματισμό από τις εκ των προτέρων πιθανότητες $P(\omega_i|x)$ αποτελεί ένα σύνολο με ελάχιστο σφάλμα. Έχουμε σχηματίσει ένα τέτοιο σύνολο λαμβάνοντας

$$g_i(x) = \log[P(\omega_i)p(x|\omega_i)], \quad i = 1, \dots, c \tag{1.6.2}$$

όπου $P(\omega_i)$ είναι η εκ των προτέρων πιθανότητα για την κλάση ω_i και $p(x|\omega_i)$ είναι η υπό συνθήκη κλάση συνάρτηση πυκνότητας πιθανότητας (pdf). Ας υποθέσουμε ότι όλες οι κλάσεις είναι κανονικά κατανεμημένες με μέση τιμή μ_i και πίνακες συνδιακύμανσης Σ_i, δηλαδή, $p(x|\omega_i) \sim N(\mu_i, \Sigma_i)$, $i = 1, \dots, c$. Στη συνέχεια, η εξίσωση 1.6.2 λαμβάνει τη μορφή

$$\begin{aligned} g_i(x) &= \log[P(\omega_i)] + \log\left\{\frac{1}{(2\pi)^{n/2}\sqrt{|\Sigma_i|}}\exp[-\frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)]\right\} \\ &= \log[P(\omega_i)] - \frac{n}{2}\log(2\pi) - \frac{1}{2}\log(|\Sigma_i|) - \frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i) \end{aligned} \tag{1.6.3}$$

Ας υποθέσουμε ότι όλες οι κλάσεις πινάκων συνδιασποράς είναι ίδιες, δηλαδή $\Sigma_i = \Sigma$ και $p(x|\omega_i) \sim N(\mu_i, \Sigma)$. Ανοίγοντας τις παρενθέσεις στον τελευταίο όρο της εξίσωσης (1.6.3) και απορρίπτοντας όλους τους όρους που δεν εξαρτώνται από το ω_i, παίρνουμε ένα νέο σύνολο ξεχωριστών συναρτήσεων

$$g_i(x) = \log[P(\omega_i)] - \frac{1}{2}\mu_i^T \Sigma^{-1}\mu_i + \mu_i^T \Sigma^{-1}x = w_{i0} + w_i^T x \quad (1.6.4)$$

όπου $w_{i0} \in R \; και \; w_i \in R^n$ είναι οι συντελεστές της γραμμικής διακριτικής συνάρτησης g_i.

Στην πραγματικότητα, οι κλάσεις δεν είναι ούτε στην κανονική κατανομή, ούτε είναι γνωστές οι πραγματικές τιμές των μ_i και Σ_i. Ακόμα, μπορούμε να υπολογίσουμε τους συντελεστές w_{i0} και w_i από τα δεδομένα χρησιμοποιώντας τις εκτιμήσεις από τις μέσες τιμές και τους πίνακες συνδιασποράς, αλλά ο ταξινομητής που προκύπτει δεν θα είναι ισοδύναμος με τον ταξινομητή ελάχιστου σφάλματος (Bayes).

Η γραμμική διακριτική ανάλυση (Linear Discriminant Analysis, LDA) είναι μια γενίκευση της γραμμικής διακριτικής του Fisher και είναι μια μέθοδος που χρησιμοποιείται στη στατιστική, την αναγνώριση προτύπων και την μηχανική μάθηση για την εύρεση γραμμικών συνδυασμών των επεξηγηματικών μεταβλητών που χαρακτηρίζει ή διαχωρίζει δύο ή περισσότερες κλάσεις αντικειμένων ή γεγονότων. Ο προκύπτων συνδυασμός μπορεί να χρησιμοποιηθεί ως γραμμικός ταξινομητής ή συνηθέστερα για την μείωση των διαστάσεων πριν από την ταξινόμηση.

1.6.2 Τετραγωνικός Διακριτικός Ταξινομητής

Υποθέτουμε ότι οι κλάσεις είναι κανονικά κατανεμημένες, αλλά τώρα με πίνακες συνδιακύμανσης συγκεκριμένης κλάσης, δηλαδή με, $p(x|\omega_i) \sim N(\mu_i, \Sigma_i)$. Το σύνολο των βέλτιστων διακριτικών συναρτήσεων λαμβάνεται από την Εξίσωση (1.6.3) μέσω της απόρριψης όλων των όρων που δεν εξαρτώνται από την ετικέτα ω_i,

$$g_i(x) = w_{i0} + w_i^T x + W_i x \quad (1.6.5)$$

Όπου

$$w_{i0} = \log[P(\omega_i)] - \frac{1}{2}\mu_i^T \Sigma_i^{-1}\mu_i - \frac{1}{2}\log(|\Sigma_i|) \quad (1.6.6)$$

$$w_i = \Sigma_i^{-1}\mu_i \quad (1.6.7)$$

Και

$$W_i = -\frac{1}{2}\Sigma_i^{-1} \quad (1.6.8)$$

Οι εκτιμήσεις των παραμέτρων για τις LDC και τον τετραγωνικό διακριτικό ταξινομητή (QDC) υπολογίζονται από τα δεδομένα. Έστω N_i ο αριθμός των αντικειμένων στο σύνολο δεδομένων Z από την κλάση $\omega_i, i = 1, \dots, c$ $και$ $l(z_j) \in \Omega$ η ετικέτα της κλάσης του $z_j \in Z$. Οι μέσες τιμές λαμβάνονται από

$$\widehat{\mu_\iota} = \frac{1}{N_i} \sum_{l(z_j)=\omega_i} z_j \tag{1.6.9}$$

και οι πίνακες συνδιακύμανσης[1] από

$$\widehat{\Sigma_\iota} = \frac{1}{N_i} \sum_{l(z_j)=\omega_i} (z_j - \widehat{\mu_\iota})(z_j - \widehat{\mu_\iota})^T \tag{1.6.10}$$

Ο κοινός πίνακας συνδιασποράς για τις LDC λαμβάνεται ως ο σταθμισμένος μέσος όρος των ξεχωριστά εκτιμώμενων κλάσεων των υπό συνθηκών πινάκων συνδιακύμανσης.

$$\hat{\Sigma} = \frac{1}{N} \sum_{i=1}^{c} N_i \Sigma_i \tag{1.6.11}$$

1.6.3 Χρησιμοποιώντας Στάθμες Δεδομένων με ένα Γραμμικό και Τετραγωνικό Διακριτικό Ταξινομητή

Για τους σκοπούς του σχεδιασμού τα σύνολα των ταξινομητών είναι σημαντικό να έχουν ένα μηχανισμό που να ενσωματώνει τις στάθμες δεδομένων σε LDC και QDC. Ο πιο φυσικός τρόπος γι' αυτό είναι να υπολογίσουμε τα σταθμισμένα $\widehat{\mu_\iota}$ και $\widehat{\Sigma_\iota}$. Έστω $W(j)$ το βάρος του αντικειμένου $z_j \in Z$, W^i είναι το άθροισμα από όλα τα βάρη των αντικειμένων στο Z από το ω_i, και $W = \sum_{i=1}^{c} W^i$ να είναι το συνολικό άθροισμα από όλα τα βάρη του Z. Τότε

$$\widehat{\mu_\iota}^{(w)} = \frac{1}{W^i} \sum_{l(z_j)=\omega_i} W(j) z_j \tag{1.6.12}$$

και

[1] Χρησιμοποιούμε τον εκτιμητή μέγιστης πιθανοφάνειας των πινάκων συνδιακύμανσης και σημειώνουμε ότι ο ταξινομητής είναι μεροληπτικός. Για τον αμερόληπτο ταξινομητή παίρνουμε $\hat{\Sigma}_i = 1/(N_i - 1) \sum_{l(z_j)=\omega_i} (z_j - \hat{\mu}_i)(z_j - \hat{\mu}_i)^T$.

$$\widehat{\Sigma_i}^{(w)} = \frac{1}{W^i} \sum_{i(z_j)=\omega_i} W(j)(z_j - \widehat{\mu_i})(z_j - \widehat{\mu_i})^T \quad (1.6.13)$$

Για τον κοινό πίνακα συνδιακύμανσης για την LDC,

$$\hat{\Sigma}^{(w)} = \frac{\sum_{i=1}^{c} W^i \Sigma_i}{W} \quad (1.6.14)$$

1.6.4 Κανονικοποιημένη Διακριτική Ανάλυση

Από τότε που ο $\hat{\Sigma}$ (για LDC) ή ο $\widehat{\Sigma_i}$ (για QDC) έπρεπε να αναστρέφονται, είναι σημαντικό ότι αυτοί οι πίνακες δεν είναι μοναδικοί ή κοντά στο να είναι μοναδικοί. Αυτό συχνά δημιουργεί ένα πρόβλημα, ειδικά για μικρά μεγέθη δεδομένων (μικρό N) και υψηλών διαστάσεων δεδομένα (μεγάλο N). Όταν το N είναι μικρότερο από τον αριθμό των παραμέτρων που πρέπει να εκτιμηθούν, μερικές από τις παραμέτρους δεν μπορούν να αναγνωριστούν από τα δεδομένα, τότε λέμε ότι το πρόβλημα είναι κακώς «τοποθετημένο» (ill-posed). Όταν το N υπερβαίνει μόνο οριακά τον αριθμό των παραμέτρων που πρέπει να εκτιμηθούν, το πρόβλημα αυτό καλείται ανεπαρκώς δημιουργημένο. Για να ξεπεραστεί αυτό, μπορούμε να χρησιμοποιήσουμε την κανονικοποίηση.

Η κανονικοποιημένη διακριτική ανάλυση (RDA – Regularized Discriminant Analysis) προτείνεται από τον Friedman (Friedman, 1989) για να αντιμετωπίσει τα «άρρωστα» και τα ανεπαρκώς δημιουργημένα προβλήματα. Θα πάρουμε τον ορισμό του Friedman για την κανονικοποίηση ως «μια προσπάθεια για τη βελτίωση των εκτιμήσεων από την επίδρασή τους μακριά από το δείγμα που βασίζεται στις τιμές μπροστά σε αυτές που θεωρούνται πιο φυσικά αληθοφανείς».

Θεωρούμε δύο τρόπους για τη σταθεροποίηση των εκτιμήσεων $\widehat{\Sigma_i}$. Αρχικά, αν θεωρήσουμε ότι ο μέσος όρος των $\widehat{\Sigma_i}$ είναι σταθμισμένος από τον αριθμό των παρατηρήσεων, έστω Σ, τότε η Εξίσωση (1.6.11), είναι ήδη μια κίνηση καλής κανονικοποίησης. Όμως, χρησιμοποιώντας ένα Σ για όλες τις κλάσεις θα μειώσουμε το QDC σε LDC. Μπορούμε να υπολογίσουμε το ύψος της μείωσης αυτής με την εισαγωγή μιας παραμέτρου, $\lambda \in [0, 1]$, και να χρησιμοποιήσουμε

$$\Sigma_i(\lambda) = \widehat{(1-\lambda)}\widehat{\Sigma_i} + \lambda\hat{\Sigma} \quad (1.6.15)$$

Ο Friedman (Friedman, 1989) χρησιμοποιεί τις σταθμισμένες εκτιμήσεις

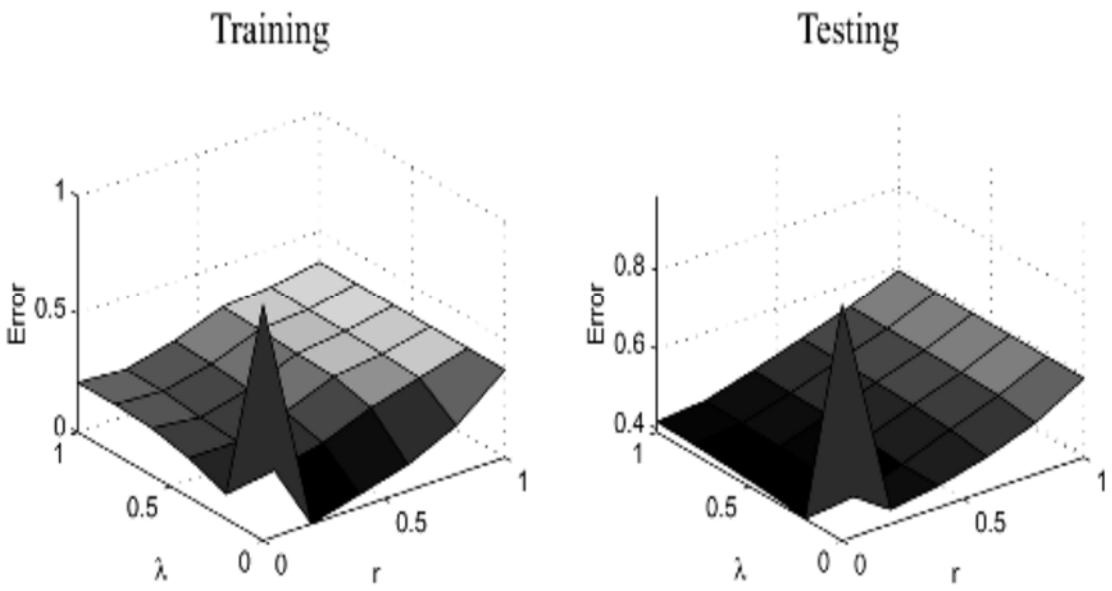
$$\widehat{\Sigma_i}^{(w)}(\lambda) = \frac{(1-\lambda)W^i\widehat{\Sigma_i}^{(w)} + \lambda W\widehat{\Sigma}^{(w)}}{(1-\lambda)W^i + \lambda W} \tag{1.6.16}$$

Και στους δύο τύπους, για $\lambda = 1$ σημαίνει ότι το QDC γίνεται LDC, επειδή όλες οι κλάσεις μοιράζονται τον ίδιο πίνακα συνδιακύμανσης $\widehat{\Sigma}$, και για $\lambda = 0$ σημαίνει ότι καμία κανονικοποίηση δεν μπορεί να πραγματοποιηθεί. Έτσι, το λ εκτείνεται στο διάστημα μεταξύ 0 και 1 και παράγει μια οικογένεια ταξινομητών μεταξύ LDC ΚΑΙ QDC.

Το πρόβλημα θα μπορούσε να είναι η «άρρωστη» ή ελλιπής δημιουργία ακόμα και για το LDC. Ο δεύτερος τρόπος κανονικοποίησης των εκτιμήσεων είναι να προσθέσουμε έναν όρο στο $\widehat{\Sigma_i}$, ο οποίος θα μειωθεί μπροστά σε ένα πολλαπλό πίνακα ταύτισης

$$\widehat{\Sigma_i}(r) = (1-r)\widehat{\Sigma_i} + \frac{r}{n}tr\left(\widehat{\Sigma_i}\right)I \tag{1.6.17}$$

όπου το tr δηλώνει το ίχνος του πίνακα και το I δηλώνει τον πίνακα ταύτισης μεγέθους $n \times n$. Η εκτίμηση αυτή έχει ως αποτέλεσμα την «εξισορρόπηση» των ιδιοτιμών του $\widehat{\Sigma_i}$, του οποίου οι μετρητές μεροληψίας είναι συνυφασμένοι με τις εκτιμήσεις που βασίζονται στο δείγμα των ιδιοτιμών (Friedman, 1989). Η παράμετρος $r \in [0,1]$ καθορίζει σε ποιο βαθμό θέλουμε να εξισώσουμε τις ιδιοτιμές. Για $r = 0$, δεν υπάρχει κανονικοποίηση και για $r = 1$, ο $\widehat{\Sigma_i}$ είναι ένας διαγώνιος πίνακας με ιδιοτιμές ίσες με το μέσο όρο των ιδιοτιμών του δείγματος που βασίζονται στην εκτίμηση του πίνακα συνδιακύμανσης.



Σχήμα 1.6.1: Τα ποσοστά σφάλματος εκπαίδευσης και ελέγχου για ζεύγη τιμών (λ, r) για το RDA με το σύνολο δεδομένων γραμμάτων.

Ανακαλούμε το παράδειγμα από το Σχήμα 1.6.1, όπου μια παράμετρος κανονικοποίησης μεταβλήθηκε. Σε αυτό το παράδειγμα κανονικοποιούμε τον πίνακα συνδιακύμανσης χρησιμοποιώντας την Εξίσωση (1.6.17) για 20 διαφορετικές τιμές του r.

Ο Friedman ορίζει τον RDA να είναι ένας συνδυασμός των δύο τρόπων κανονικοποίησης έτσι ώστε

$$\hat{\Sigma}_i(\lambda, r) = (1 - r)\hat{\Sigma}_i(\lambda) + \frac{r}{n} tr\left[\hat{\Sigma}_i(\lambda)\right] I \tag{1.6.18}$$

Οι τιμές των δύο παραμέτρων, λ και r, πρέπει να καθορίζονται από τα δεδομένα. Ανάλογα με το πρόβλημα, ένα διαφορετικό ζεύγος τιμών μπορεί να προτιμάται. Μια διαδικασία διασταυρωμένης επικύρωσης έχει προταθεί από τους (Friedman, 1989) για την επιλογή ενός πλέγματος από ζεύγη (λ, r).

Παράδειγμα: Επιλέγοντας τις τιμές των παραμέτρων για την Κανονικοποιημένη Διακριτική Ανάλυση. Για την προσομοίωση ενός μικρού συνόλου δεδομένων, χρησιμοποιούμε τα πρώτα 200 αντικείμενα από τα δεδομένα της επιστολής για την εκπαίδευση και τα υπόλοιπα 19800 αντικείμενα για τον έλεγχο. Με δεδομένο ότι υπάρχουν 26 κλάσεις και 16 χαρακτηριστικά, είναι πιθανό ότι οι εκτιμήσεις του δείγματος των πινάκων της συνδιακύμανσης θα επωφεληθούν από την κανονικοποίηση. Το Σχήμα 1.6.1 δείχνει τις επιφάνειες εκπαίδευσης και ελέγχου πάνω από την τετράγωνη μονάδα $[0,1]^2$ που εκτείνεται στο (λ, r).

Υπήρχαν απλοί πίνακες συνδιασποράς για $\lambda = 0 \; και \; r = 0$, οι οποίοι είναι η αιτία της υψηλής κορυφής του σφάλματος γύρω από την περιοχή. Μια πολύ μικρή

κανονικοποίηση εμφανίζεται να είναι επαρκής για να κάνει τα περισσότερα από το γραμμικό-τετραγωνικό συμβιβαστικό μοντέλο. Το σφάλμα εκπαίδευσης για $\lambda = 0, r = 0.2$ εμφανίζεται να είναι 0 και το αντίστοιχο σφάλμα ελέγχου ήταν 42.8%. Το καλύτερο σφάλμα ελέγχου, 39.2% βρέθηκε για $\lambda = 0.2, r = 0$, το οποίο προτείνει ένα μοντέλο, όχι πολύ μακριά από το QDC. Όπως φαίνεται στο Σχήμα, η περαιτέρω κανονικοποίηση θα αυξήσει το σφάλμα, που οδηγεί τελικά σε 37% σφάλμα εκπαίδευσης και 59.1% σφάλμα ελέγχου του LDA που έχει πραγματοποιηθεί σε αυτό το σύνολο δεδομένων.

ΚΕΦΑΛΑΙΟ 2: ΑΝΑΛΥΣΗ ΕΠΙΒΙΩΣΗΣ – ΜΟΝΤΕΛΟ COX

2.1 Μοντέλα Παλινδρόμησης για Δεξιά Αποκοπή

Έχουμε παρουσιάσει στατιστικές μεθόδους για τα δεξιά και αριστερά αποκομμένα δεδομένα επιβίωσης. Γι' αυτόν τον τύπο δεδομένων υποθέτουμε ότι υπάρχει ένα μοναδικό γεγονός, το οποίο προκαλεί το θάνατο ή την αποτυχία μιας μεμονωμένης μονάδας δειγματοληψίας. Αυτές οι μονάδες θα μπορούσαν να είναι δείγμα είτε ανθρώπου, είτε ζώου, που υποβάλλονται σε κάποιο είδος θεραπείας ή ηλεκτρονικής μονάδας. Εδώ, εάν ο χρόνος της αποτυχίας είναι T, τότε ενδιαφερόμαστε να βγάλουμε ένα συμπέρασμα σχετικά με τη συνάρτηση επιβίωσης $S(t) = \Pr[T > t]$. Σε εφαρμογές της μηχανικής η συνάρτηση αυτή ονομάζεται συνάρτηση αξιοπιστίας. Στη συνέχεια θα εστιάσουμε σε βιολογικές εφαρμογές των μεθόδων.

Για πολλές εφαρμογές ο χρόνος μέχρι την αποτυχία δεν είναι παρατηρήσιμος για όλα τα άτομα της μελέτης, αλλά αποσπασματικές πληροφορίες, ότι ο χρόνος ενός γεγονότος είναι μεγαλύτερος από κάποιο χρονικό διάστημα αποκοπής, είναι όλες διαθέσιμες. Τέτοιες παρατηρήσεις ονομάζονται δεξιά αποκομμένες παρατηρήσεις και οι πληροφορίες που μας δίνουν είναι απλά ότι ο χρόνος αποτυχίας για ένα άτομο είναι πέρα από το χρόνο αποκοπής του.

Υπάρχουν πολλοί τύποι των δεξιά αποκομμένων παρατηρήσεων. Υπάρχει ο τύπος I, όπου σε κάθε αντικείμενο έχει ανατεθεί ένας σταθερός χρόνος αποκοπής, μετά από τον οποίο κάθε παρατήρηση στο αντικείμενο σταματά. Αυτός ο τρόπος αποκοπής χρησιμοποιείται σε αξιόπιστες εφαρμογές για να μειωθεί ο χρόνος μελέτης ή σε βιολογικές μελέτες όταν το «παράθυρο» παρατηρήσεων είναι σταθερό. Για τον τύπο II διαγραφόμενων παρατηρήσεων ο αριθμός των αποτυχιών είναι σταθερός και ο χρόνος αποκοπής είναι τυχαίος. Ο τρόπος αυτός χρησιμοποιείται πιο συχνά σε εφαρμογές μηχανικής για να μειωθεί ο χρόνος μελέτης. Τέλος, υπάρχει και η προοδευτική αποκοπή, όπου αυτά που παρατηρούνται είναι ο μικρότερος χρόνος του συμβάντος T και ένας τυχαίος χρόνος αποκοπής C. Ο χρόνος αποκοπής C, συνήθως ο χαμένος χρόνος παρακολούθησης (lost-to-follow-up time), αντανακλάται όταν το άτομο αποχωρήσει από τη μελέτη ή σταματήσει να ακολουθείται. Οι περισσότερες αναλύσεις υποθέτουν ότι ο χρόνος αποκοπής είναι μη πληροφοριακός ή ανεξάρτητος από το χρόνο επιβίωσης. Η αποκοπή είναι μη κατατοπιστική αν δεν υπάρχουν πληροφορίες σχετικά με το χρόνο επιβίωσης, T, που διατίθεται από το μέγεθος του χρόνου αποκοπής.

Σε ορισμένες περιπτώσεις, εκτός του ότι είναι δεξιά αποκομμένα τα δεδομένα, είναι και αριστερά αποκομμένα. Τα δεδομένα λέγονται αριστερά αποκομμένα μόνο αν τα πιθανά αντικείμενα που είχαν κάποιο γεγονός αποκοπής σε κάποιο χρόνο τ, βρίσκονται σε κίνδυνο σε οποιαδήποτε χρονική στιγμή μετά το τ. Το κλασικό παράδειγμα είναι η μελέτη του «Channing house», η οποία εξέτασε τις πιθανότητες επιβίωσης για τους ηλικιωμένους σε έναν οίκο ευγηρίας. Σε αυτή τη μελέτη σε μια δεδομένη χρονική στιγμή, t_0, μόνο τα άτομα που εισέρχονται στον οίκο ευγηρίας σε μια ηλικία πριν από την t_0, θεωρούνται ότι διατρέχουν κίνδυνο θανάτου σε αυτή την ηλικία, και κάθε ασθενής με την ηλικία εισόδου μεγαλύτερη από την t_0 δεν περιλαμβάνεται στο σύνολο κινδύνου. Αφού οι αριστερά αποκομμένες απλά μοντελοποιούν το σύνολο κινδύνου, οι μέθοδοι που αναπτύχθηκαν για τα δεξιά αποκομμένα δεδομένα είναι συνήθως οι ίδιες με εκείνες για τα αριστερά-δεξιά αποκομμένα δεδομένα.

Επαγωγικές τεχνικές για τα δεδομένα επιβίωσης μπορούν να εντοπίσουν τις ρίζες τους σε τρεις βασικές δημοσιεύσεις. Η πρώτη είναι η δημοσίευση των Kaplan και Meier (1958), οι οποίοι ανέπτυξαν μια μη παραμετρική εκτίμηση της συνάρτησης επιβίωσης. Αυτός ο «Kaplan-Meier» εκτιμητής ήταν μια σύγχρονη εκδοχή των κλασικών πινάκων ζωής που χρησιμοποιούνται σε αναλογιστικές επιστήμες και αναπτύχθηκε από τον Edmond Halley το 1693. Στη συνέχεια ακολούθησε η δημοσίευση του David Cox το 1972.

Η τρίτη και ίσως η σημαντικότερη ερευνητική εργασία είναι η ανάπτυξη της θεωρίας των διαδικασιών μέτρησης (counting processes) και η χρήση τους στην ανάλυση επιβίωσης από τον Odd Aalen. Η πρωτοποριακή του εργασία σχετικά με τη διαδικασία μέτρησης και τα martingales, ξεκίνησε το 1975 με την διδακτορική του διατριβή και είχε βαθιά επιρροή στις τεχνικές ανάλυσης επιβίωσης. Τα συμπεράσματα των θεμελιωδών ποσοτήτων που σχετίζονται με αθροιστικά ποσοστά κινδύνου στην ανάλυση επιβίωσης και στα μοντέλα για την ανάλυση ιστορικών γεγονότων βασίζονται κατά κανόνα στο έργο του Aalen.

Το μοντέλο του Cox ή μοντέλο αναλογικών κινδύνων είναι ίσως η πιο συνηθισμένη μέθοδος στην ανάλυση επιβίωσης. Αυτό το μοντέλο βασίζεται στη μοντελοποίηση της συχνότητας του κινδύνου $\lambda(t|Z)$. Εδώ το ποσοστό κινδύνου είναι ο ρυθμός με τον οποίο τα άτομα αντιμετωπίζουν το συμβάν, δηλαδή

$$\lambda(t|Z) = -\frac{d\ln\big(S(t|Z)\big)}{dt} = \frac{f(t|Z)}{S(t|Z)} \tag{2.1.1}$$

Εδώ η $S(T|Z)$ $(f(T|Z))$ είναι η πιθανότητα επιβίωσης (πυκνότητα) με δεδομένο ένα διάνυσμα Z των συμμεταβλητών. Το μοντέλο του Cox στην πιο ευρέως

χρησιμοποιούμενη διατύπωσή του θεωρεί ότι μπορούμε να γράψουμε το $\lambda(t|Z)$ ως:

$$\lambda(t|Z) = \lambda_0(t)\exp\{\beta^t Z\} \tag{2.1.2}$$

όπου β είναι ένα διάνυσμα παραμέτρων και $\lambda_0(t)$ η αρχική τιμή για τον κίνδυνο. Να σημειωθεί ότι για το μοντέλο αυτό, αν έχουμε δύο άτομα με δύο σύνολα συνδιακύμανσης:

$$\frac{\lambda(t|Z_1)}{\lambda(t|Z_2)} = \frac{\lambda_0(t)\exp\{\beta^t Z_1\}}{\lambda_0(t)\exp\{\beta^t Z_2\}} = \exp\{\beta^t(Z_1 - Z_2)\} \tag{2.1.3}$$

τα οποία είναι ανεξάρτητα του t. Έτσι, αυτό ονομάζεται «μοντέλο αναλογικού κινδύνου».

Το μοντέλο του Cox είναι το πιο δημοφιλές μοντέλο παλινδρόμησης για τα δεδομένα επιβίωσης. Οι ιδιότητες του μπορούν να προκύψουν χρησιμοποιώντας τις τεχνικές της διαδικασίας μέτρησης του Aalen. Υπάρχουν επίσης πακέτα συμπερασματολογίας γι' αυτό σχεδόν κάθε στατιστικό πακέτο. Μπορεί αρκετά εύκολα να επεκταθεί σε συμμεταβλητές εξαρτώμενες από το χρόνο και σε μοντέλα πολλαπλών καταστάσεων (models for multistate models) ή μοντέλα με τυχαία αποτελέσματα. Περισσότερες πληροφορίες σχετικά με την ανάλυση επιβίωσης και το μοντέλο του Cox μπορεί κάποιος να αναζητήσει στο βιβλίο της Καρώνη-Ρίτσαρντσον (2009).

2.2 Βασικές Στατιστικές Έννοιες

2.2.1 Χρόνος Επιβίωσης και Χρόνος Αποκοπής

Για να ρυθμίσουμε το πλαίσιο των δεδομένων επιβίωσης με δεξιά αποκοπή, δύο τυχαίες μεταβλητές χρειάζεται να οριστούν

$T_{surv.}$: *ο χρόνος επιβίωσης*, $T_{cens.}$: *ο χρόνος αποκοπής*.

Ο χρόνος αποκοπής $T_{cens.}$ συχνά συμβολίζεται με C. Διαφορετικοί μηχανισμοί αποκοπής μπορούν να διακριθούν. Ένας κοινός μηχανισμός για τα κλινικά δεδομένα είναι η διαχειριζόμενη αποκοπή (administrative censoring), όπου ο χρόνος αποκοπής καθορίζεται από τη διακοπή της μελέτης. Για τους περισσότερους σκοπούς επαρκεί να θεωρήσουμε ότι η αποκοπή είναι τυχαία. Η κρίσιμη συνθήκη για τη στατιστική ανάλυση είναι ότι ο χρόνος επιβίωσης $T_{surv.}$ και ο χρόνος

αποκοπής $T_{cens.}$ είναι ανεξάρτητοι. Με την παρουσία των επεξηγηματικών μεταβλητών η συνθήκη αυτή μπορεί να εξασθενήσει την ανεξαρτησία της $T_{surv.}$ και $T_{cens.}$, εξαρτώμενων από τις επεξηγηματικές μεταβλητές.

Και για τις δύο τυχαίες μεταβλητές οι αθροιστικές συναρτήσεις κατανομής (cumulative distribution functions) $F_{surv}(t) = P(T_{surv} \leq t)$ και $F_{cens}(t) = P(T_{cens} \leq t)$ μπορούν να οριστούν. Η συνάρτηση κατανομής του χρόνου επιβίωσης καλείται «συνάρτηση αποτυχίας». Στην ανάλυση επιβίωσης είναι συχνά πιο βολικό να χρησιμοποιηθούν οι συμπληρωματικές συναρτήσεις, η συνάρτηση επιβίωσης $S(t)$ καθώς και η συνάρτηση αποκοπής $G(t)$ που ορίζονται με

$$S(t) = 1 - F_{surv}(t) = P(T_{surv} > t) \quad (2.2.1)$$
$$G(t) = 1 - F_{cens}(t) = P(T_{cens} > t) \quad (2.2.2)$$

Εκτιμάται ότι ο T_{surv} έχει μια συνεχή κατανομή, πράγμα που σημαίνει ότι η συνάρτηση επιβίωσης $S(t)$ είναι συνεχής και διαφορίσιμη.

Συνοψίζοντας, σε ένα σύνολο δεδομένων επιβίωσης η πιο σημαντική πληροφορία μας δίνεται από (μια εκτίμηση από) τη συνάρτηση επιβίωσης, αλλά είναι επίσης σημαντικό να δείξουμε (μια εκτίμηση από) τη συνάρτηση αποκοπής. Η συνάρτηση αποκοπής περιγράφει την κατανομή των χρόνων παρακολούθησης αν κανένα άτομο δεν είχε πεθάνει.

Στην πράξη είναι ως επί το πλείστον αδύνατο να παρατηρήσουμε τόσο το T_{surv} όσο και το T_{cens}. Ο παρατηρούμενος «χρόνος επιβίωσης» T είναι ο μικρότερος μεταξύ των δύο,

$$T = \min(T_{surv}, T_{cens}) \quad (2.2.3)$$

Επιπλέον, είναι γνωστό αν οι T_{surv} ή T_{cens} έχουν παρατηρηθεί. Αυτό αποδεικνύεται από το δείκτη συμβάντος D. Ο συνήθης ορισμός είναι

$$D = \begin{cases} 0, \text{εάν } T = T_{cens}; \\ 1, \text{εάν } T = T_{surv}. \end{cases} \quad (2.2.4)$$

Έτσι, η πληροφορία σχετικά με την κατάσταση επιβίωσης συνοψίζεται στο ζεύγος (T, D).

2.2.2 Ο Εκτιμητής Kaplan-Meier

Το σημείο εκκίνησης για τη στατιστική ανάλυση είναι ένα δείγμα από n ανεξάρτητες παρατηρήσεις

$$(t_1, d_1), (t_2, d_2), \dots, (t_n, d_n)$$

από το (T, D). Από τους παρατηρούμενους χρόνους επιβίωσης $t_1, \dots, t_n$ εκείνοι με $d_i = 1$ καλούνται χρόνοι γεγονότος (event times). Οι παρατηρούμενοι χρόνοι επιβίωσης με $d_i = 0$ ονομάζονται χρόνοι αποκοπής.

Είναι βολικό να χρησιμοποιούμε το συμβολισμό από τον τομέα των διαδικασιών μέτρησης

$$Y_i(t) = 1\{t_i \geq t\}$$
$$\bar{Y}(t) = \sum_{i=1}^{n} Y_i(t)$$
$$N_i(t) = 1\{t_i \leq t, d_i = 1\}$$
$$\bar{N}(t) = \sum_{i=1}^{n} N_i(t)$$

Το σύνολο κινδύνου $R(t)$ ορίζεται ως $R(t) = \{i,\ t_i \geq t\}$. Το μέγεθός του δίνεται από το $\bar{Y}(t)$. Υποθέτοντας ότι τα T_{surv} και T_{cens} είναι ανεξάρτητα, τόσο το $S(t)$ όσο και το $G(t)$ μπορούν να εκτιμηθούν από την εκδοχή του εκτιμητή Kaplan-Meier (Kaplan και Meier, 1958). Ο εκτιμητής της συνάρτησης επιβίωσης δίνεται

$$\hat{S}_{KM}(t) = \prod_{s \leq t} (1 - \frac{\Delta \bar{N}(s)}{\bar{Y}(s)}) \tag{2.2.5}$$

όπου $\Delta \bar{N}(t)$ είναι ο αριθμός των γεγονότων σε χρόνο t. Ο εκτιμητής $\hat{G}_{KM}(t)$ ορίζεται ομοίως.

Η φόρμουλα επίσης καλύπτει την περίπτωση αλληλένδετων χρόνων ενός γεγονότος. Το τυπικό σφάλμα της εκτίμησης δίνεται από τον τύπο του Greenwood (Greenwood, 1926)

$$se^2\left(\hat{S}_{KM}(t)\right) = \hat{S}_{KM}^2(t) \cdot \sum_{s \leq t} \frac{\Delta \bar{N}(s)}{\bar{Y}(s)(\bar{Y}(s) - \Delta \bar{N}(s))} \tag{2.2.6}$$

Η εκτίμηση $\hat{S}_{KM}(t)$ μαζί με το τυπικό της σφάλμα $se(\hat{S}_{KM}(t))$ μπορούν να χρησιμοποιηθούν για την κατασκευή ενός $(1 - \alpha) \cdot 100\%$ κατά σημείο διάστημα εμπιστοσύνης για τη συνάρτηση επιβίωσης $S(t)$. Το πιο απλό διάστημα

εμπιστοσύνης είναι το $\hat{S}_{KM}(t) \pm z_{1-\alpha/2} se(\hat{S}_{KM}(t))$, με $z_{1-\alpha/2}$ το $1-\alpha/2$ εκατοστημόριο της τυπικής κανονικής κατανομής. Θα μπορούσε ωστόσο να βρεθεί και εκτός του διαστήματος $[0,1]$. Αυτό μπορεί να διορθωθεί με τη χρήση μετασχηματισμών όπως $\ln(S(t))$ ή $\ln(-\ln(S(t)))$. Τέτοια μετασχηματισμένα διαστήματα εμπιστοσύνης απαιτούνται για μικρότερα μεγέθη δειγμάτων. Για περισσότερες λεπτομέρειες παραπέμπουμε στους Borgan και Liestol (1990).

2.2.3 Η Συνάρτηση Κινδύνου

Σύμφωνα με την υπόθεση της συνεχούς κατανομής με διαφορετική συνάρτηση επιβίωσης, η συνάρτηση κινδύνου, επίσης γνωστή και ως «δύναμη της θνησιμότητας» (force of morality) ορίζεται ως

$$\lambda(t)dt = P(T < t + dt | T \geq t) \tag{2.2.7}$$

Από το $P(T > t + dt | T \geq t) = S(t + dt)/S(t)$ ο ακόλουθος εναλλακτικός ορισμός μπορεί να ληφθεί ως

$$\lambda(t) = -\frac{S'(t)}{S(t)} = -\frac{d\ln(S(t))}{dt} \tag{2.2.8}$$

Μια σχετική έννοια είναι η αθροιστική συνάρτηση κινδύνου, συμβολίζεται με $\Lambda(t)$ και ορίζεται ως

$$\Lambda(t) = \int_0^t \lambda(s)ds \tag{2.2.9}$$

Προφανώς, οι $\Lambda(t)$ και $S(t)$ είναι στενά συνδεδεμένες: $\Lambda(t) = -\ln\big(S(t)\big), S(t) = \exp(-\Lambda(t))$.

Δεδομένου ότι η συνάρτηση κινδύνου $\lambda(t)$ είναι καλά ορισμένη εάν η συνάρτηση επιβίωσης $S(t)$ είναι διαφορίσιμη, είναι δύσκολο να εκτιμήσουμε σωστά τη συνάρτηση κινδύνου επειδή η εκτίμηση Kaplan-Meier της συνάρτησης επιβίωσης είναι μη διαφορίσιμη συνάρτηση και κάποια εξομάλυνση είναι απαραίτητη πριν ληφθεί μια σωστή εκτίμηση της συνάρτησης κινδύνου.

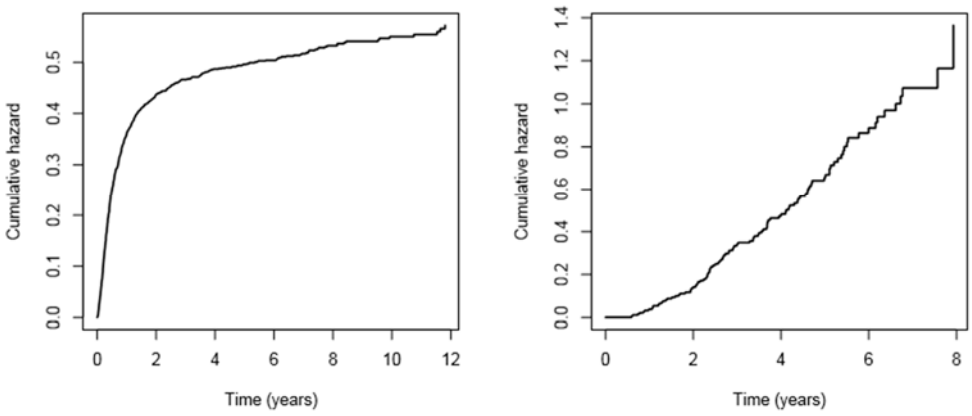
Είναι πολύ πιο εύκολο να εκτιμήσουμε την αθροιστική συνάρτηση κινδύνου $\Lambda(t)$. Ένας τρόπος να γίνει αυτό είναι να χρησιμοποιήσουμε τη σύνδεση μεταξύ του $\Lambda(t)$ και του $S(t)$ και να ορίσουμε

$$\hat{\Lambda}_{KM}(t) = -\ln\left(\hat{S}_{KM}(t)\right) = \sum_{s \le t} \ln(1 - \frac{\Delta \bar{N}(s)}{\bar{Y}(s)}) \tag{2.2.10}$$

Μια εναλλακτική είναι ο εκτιμητής Nelson-Aalen (Nelson, 1969; Aalen, 1975). Ο εκτιμητής και το τυπικό του σφάλμα δίνονται από

$$\hat{\Lambda}_{NA}(t) = \int_0^t \frac{d\bar{N}(s)}{\bar{Y}(s)}, \qquad se^2\left(\hat{\Lambda}_{NA}(t)\right) = \int_0^t \frac{d\bar{N}(s)}{\bar{Y}(s)^2} \tag{2.2.11}$$

Εάν το μέγεθος του δείγματος είναι μεγάλο, υπάρχει πολύ μικρή διαφορά μεταξύ των $\Lambda_{KM}(t)$ και $\Lambda_{NA}(t)$ ή αντίστοιχα μεταξύ των $\hat{S}_{KM}(t)$ και $\hat{S}_{NA}(t) = \exp(-\hat{\Lambda}_{NA}(t))$. Τα άλματα στο $\hat{\Lambda}_{NA}(t)$ ορίζουν μια διακριτή εκτίμηση του κινδύνου συγκεντρωμένα στους χρόνους συμβάντος: $\hat{\lambda}_{NA}(t) = \Delta \bar{N}(t)/\bar{Y}(t)$. Αυτός ο ορισμός ισχύει επίσης και με την παρουσία συνδέσεων.



Σχήμα 2.2.1: Εκτιμητές Nelson-Aalen του αθροιστικού κινδύνου για όλους τους ασθενείς (αριστερά) και των ασθενών με χρόνια μυελογενή λευχαιμία (δεξιά).

Παρά το γεγονός ότι η συνάρτηση κινδύνου είναι δύσκολο να εκτιμηθεί, παίζει ουσιαστικό εννοιολογικό ρόλο στον τρόπο σκέψης σχετικά με τη διαδικασία της επιβίωσης. Στο κλινικό περιβάλλον, το σχήμα της συνάρτησης κινδύνου καθορίζει την μακροπρόθεσμη προοπτική για ένα ασθενή. Μια φθίνουσα συνάρτηση κινδύνου σημαίνει ότι η πρόγνωση γίνεται καλύτερη όσο ζεις περισσότερο. Μια αύξουσα συνάρτηση κινδύνου σημαίνει ότι η πρόγνωση χειροτερεύει όσο ζεις περισσότερο. Η γραφική παράσταση της (εκτιμώμενης) αθροιστικής συνάρτησης κινδύνου για ένα σύνολο δεδομένων μπορεί να είναι ένας βολικός τρόπος για την ανίχνευση μιας αύξουσας ή φθίνουσας συνάρτησης κινδύνου. Μια κυρτή

αθροιστική συνάρτηση κινδύνου δείχνει προς μια αύξηση του κινδύνου, ενώ μια κοίλη αθροιστική συνάρτηση κινδύνου πηγαίνει σε μείωση του κινδύνου. Ένα παράδειγμα δανεισμένο από τους Van Houwelingen και Putter (2012) δίνεται στο Σχήμα (2.2.1) δείχνοντας τα διαγράμματα των εκτιμήσεων Nelson-Aalen για το θάνατο ή την υποτροπή σε όλους τους ασθενείς (αριστερά) και για το θάνατο στους ασθενείς με χρόνια μυελογενή λευχαιμία (δεξιά). Σε περίπτωση θεραπείας, η αθροιστική συνάρτηση κινδύνου θα φτάσει το ανώτατο όριο. Για περισσότερη συζήτηση της ερμηνείας των καμπυλών κινδύνου βλέπουμε Klein και Moeschberger (2003).

2.3 Μοντέλο του Cox

2.3.1 Το Αναλογικών Κινδύνων (Cox) Μοντέλο

Για την ανάπτυξη μοντέλων για δεδομένα επιβίωσης σε ένα πληθυσμό ατόμων, κάποιος χρειάζεται έναν απλό τρόπο περιγραφής της μεταβολής στην επιβίωση μεταξύ των ατόμων. Ένα δημοφιλές μοντέλο είναι να εξεταστεί η ειδική ατομική συνάρτηση κινδύνου $\lambda_i(t)$ και να κάνει την υπόθεση αναλογικών κινδύνων, η οποία θα είναι

$$\lambda_i(t) = c_i \lambda_0(t) \tag{2.3.1}$$

όπου c_i είναι μια σταθερά και $\lambda_0(t)$ είναι μια συνάρτηση κινδύνου, η οποία είναι αριστερά ακαθόριστη.

Το αποτέλεσμα των συμμεταβλητών σχετικά με την επικινδυνότητα μπορεί εύκολα να μοντελοποιηθεί λαμβάνοντας $c_i = \exp(X_i^T \beta)$ που οδηγεί στο μοντέλο παλινδρόμησης αναλογικών κινδύνων, περισσότερο γνωστό ως το μοντέλο παλινδρόμησης του Cox, το οποίο εισήχθη στο Cox (1972),

$$\lambda(t|X) = \lambda_0(t)\exp(X^T \beta) \tag{2.3.2}$$

Εδώ, το $\lambda_0(t)$ είναι η βασική γραμμή κινδύνου που καθορίζει το σχήμα της συνάρτησης επιβίωσης, το X είναι το διάνυσμα στήλη των συμμεταβλητών ενός ατόμου και β είναι ένα διάνυσμα στήλης από συντελεστές παλινδρόμησης. Είναι κοινή πρακτική να μην ορίζουμε ένα παραμετρικό μοντέλο για τη βασική γραμμή κινδύνου. Αυτό είναι σύμφωνο με την πρακτική για να δείξει την Kaplan-Meier εκτίμηση της συνάρτησης επιβίωσης, ως σύνοψη των δεδομένων. Όπως και σε μοντέλα παλινδρόμησης για άλλους τύπους δεδομένων, το διάνυσμα της

συμμεταβλητής X μπορεί να περιέχει μετασχηματισμούς και αλληλεπιδράσεις των παραγόντων κινδύνου. Θα πρέπει να σημειωθεί ότι δεν υπάρχει σταθερός όρος στον φορέα παλινδρόμησης. Η σταθερά απορροφάται στη βασική γραμμή κινδύνου: $\ln(\lambda_0(t))$ μπορεί να θεωρηθεί ως μια χρονο-εξαρτώμενη τομή στο γραμμικό μοντέλο για $\ln(\lambda(t|X))$. Το συμπέρασμα είναι ότι το κεντράρισμα των συμμεταβλητών, αντικαθιστώντας το X με $X - E[X]$, θα αλλάξει τη βασική γραμμή, αλλά όχι τους συντελεστές παλινδρόμησης. Σε κάποιο λογισμικό το κεντράρισμα αυτό εφαρμόζεται και δεν είναι πάντα εύκολο να κατανοήσουμε σε τι αναφέρεται η αναφερόμενη βασική γραμμή κινδύνου. Η συνάρτηση επιβίωσης που εξάγεται από το μοντέλο δίνεται από

$$S(t|X) = \exp\big(-\exp(X^T\beta)\,\Lambda_0(t)\big) = S_0(t)^{\exp(X^T\beta)}. \tag{2.3.3}$$

Εδώ, το $\Lambda_0(t) = \int_0^t \lambda_0(s)ds$ είναι η αθροιστική βασική γραμμή κινδύνου και το $S_0(t) = \exp(-\Lambda_0(t))$ η βασική συνάρτηση επιβίωσης. Η γραμμική πρόβλεψη (linear predictor) $X^T\beta$ είναι γνωστή ως προγνωστικός δείκτης και συμβολίζεται με PI.

Η οριακή συνάρτηση επιβίωσης επιτυγχάνεται με τη λήψη της αναμενόμενης συνάρτησης επιβίωσης $E[S(t|X)]$ στον πληθυσμό. Δεδομένου ότι το X εμφανίζεται στον εκθέτη, το $E[S(t|X)]$ δεν είναι το ίδιο με την εκτιμώμενη επιβίωση για το μέσο άτομο: $S(t) = E[S(t|X)] \neq S(t|\,E[X])$. Η διαφορά μεταξύ των $E[S(t|\,X)]$ και $S(t|\,E[X])$ μπορεί να αναμένεται να είναι μικρή εάν η διακύμανση του προγνωστικού δείκτη $PI = X^T\beta$ είναι μικρή και $S(t|X)$ δεν απέχει πολύ από τη μονάδα.

2.3.2 Τοποθέτηση του Μοντέλου του Cox

Είναι πιο ενδιαφέρον να διαβάσετε το πρωτότυπο έγγραφο από τον Cox (1972) και τη γραπτή συζήτηση που ακολουθεί. Η έμφαση του Cox είναι στην εκτίμηση των συντελεστών παλινδρόμησης χρησιμοποιώντας αυτό που ονομάζεται «μερική πιθανότητα». Η εκτίμηση της βασικής γραμμής κινδύνου έχει από καιρό παραμεληθεί. Ωστόσο το αποτέλεσμα του δείκτη κινδύνου μπορεί μόνο να κατανοηθεί πλήρως, αν ο βασικός κίνδυνος είναι γνωστός. (Ο καλύτερος τρόπος για την κατανόηση του μοντέλου είναι η απεικόνιση των εκτιμώμενων καμπύλων επιβίωσης για τις αντιπροσωπευτικές τιμές του διανύσματος συμμεταβλητής X).

Για να τονιστεί η σημασία και των δύο πλευρών (βασική γραμμή κινδύνου και συντελεστές παλινδρόμησης), η πλήρης πιθανότητα των δεδομένων θα πρέπει να ληφθεί ως το σημείο εκκίνησης για την τοποθέτηση του μοντέλου.

Τα διαθέσιμα δεδομένα είναι ένα δείγμα από n ανεξάρτητες παρατηρήσεις από την τριπλέτα (T, D, X), που είναι

$$(t_1, d_1, x_1), (t_2, d_2, x_2), \dots, (t_n, d_n, x_n)\ . \tag{2.3.4}$$

Η λογαριθμική πιθανότητα των δεδομένων δοθέντων των συμμεταβλητών δίνεται από

$$l(\lambda_0, \beta) = \sum_{i=1}^{n}(-\Lambda_0(t_i)\exp\left(x_i^T\beta\right) + d_i(\ln\left(\lambda_0(t_i)\right) + x_i^T\beta)). \tag{2.3.5}$$

Αυτή η σχέση θα μεγιστοποιηθεί με τη συγκέντρωση όλων των κινδύνων στους χρόνους εκδήλωσης. Αυτό οδηγεί σε μια διακριτή εκδοχή του κινδύνου, όπως περιγράφεται στην παράγραφο 2.2.3, στην οποία ο αθροιστικός κίνδυνος ορίζεται ως

$$\Lambda_0(t) = \sum_{s \le t} \lambda_0(s). \tag{2.3.6}$$

Συνδέοντας τη σχέση αυτή με τη σχέση για τη λογαριθμική πιθανοφάνεια και αναδιατάσσοντας μερικούς όρους έχουμε

$$l(\lambda_0, \beta) = \sum_t (-\lambda_0(t) \sum_i Y_i(t)\exp\left(x_i^T\beta\right) + \ln\left(\lambda_0(t)\right)\Delta\bar{N}(t) + \sum_i \Delta N_i(t) x_i^T\beta). \tag{2.3.7}$$

Ο τύπος αυτός επιτρέπει δεσμούς. Για συγκεκριμένη τιμή του β η σχέση αυτή είναι μέγιστη, ονομάζεται «εκτιμητής Breslow» (Breslow, 1974) και δίνεται από

$$\hat{\lambda}_0(t|\beta) = \frac{\Delta\bar{N}(t)}{\sum_i Y_i(t)\exp(x_j^T\beta)}. \tag{2.3.8}$$

Η προκύπτουσα μέγιστη λογαριθμική πιθανοφάνεια είναι

$$l(\lambda_0(\beta), \beta) = pl(\beta) + \sum_t (-\Delta\bar{N}(t) + \ln(\Delta\bar{N}(t))). \tag{2.3.9}$$

Εδώ το $pl(\beta)$ είναι η μερική λογαριθμική πιθανοφάνεια του Cox και ορίζεται ως

$$pl(\beta) = \sum_{i=1}^{n} \int_0^{\infty} \ln\left(\frac{\exp\left(x_i^T \beta\right)}{\sum_j Y_j(t) \exp\left(x_j^T \beta\right)}\right) dN_i(t). \quad (2.3.10)$$

Ο Cox δεν έλαβε αυτή τη σχέση ως μια κατανομή πιθανοφάνειας, αλλά χρησιμοποίησε μια προϋπόθεση. Ο όρος $\exp(x_i^T \beta)/\sum_j Y_j(t)\exp(x_i^T \beta)$ $\gamma\iota\alpha\ t = t_i$ μπορεί να ερμηνευτεί ως η πιθανότητα το ατόμου i να είναι αυτό που πέθανε τη χρονική στιγμή t_i με δεδομένο σύνολο κινδύνου $R(t_i)$ τους ανθρώπους που είναι ακόμα ζωντανοί και στη συνέχεια τη χρονική στιγμή ακριβώς πριν από την t_i. Η συνθήκη περιπλέκεται με την παρουσία των δεσμών. Ο παραπάνω ορισμός επιτρέπει την παρουσία των δεσμών και προτάθηκε από τον Breslow (Breslow, 1972). Είναι απόλυτα έγκυρος εφόσον οι δεσμοί είναι τυχαίοι και μόνο εξαιτίας της στρογγυλοποίησης των παρατηρούμενων χρόνων.

Έτσι, η υπολογιστική διαδικασία είναι να εκτιμηθεί το β μεγιστοποιώντας τη μερική λογαριθμική πιθανοφάνεια και να εκτιμηθεί η βασική γραμμή κινδύνου από τον εκτιμητή Breslow με $\beta = \hat{\beta}$.

Η συνάρτηση επιβίωσης δοθείσας της συμμεταβλητής x μπορεί να εκτιμηθεί είτε από τον αναλογικό εκτιμητή Nelson-Aalen

$$\hat{S}_{NA}(t|x, \beta) = \exp\left(-\hat{\Lambda}_0(t) \exp\left(x^T \hat{\beta}\right)\right), \quad (2.3.11)$$

είτε από τον αναλογικό εκτιμητή Kaplan-Meier

$$\hat{S}_{PL}(t|x, \hat{\beta}) = \prod_{s \le t} (1 - \exp(x^T \hat{\beta}) \hat{\lambda}_0(s)). \quad (2.3.12)$$

Τα περισσότερα πακέτα λογισμικού παρέχουν το $\hat{S}_{NA}$. Στην R, και το $\hat{S}_{NA}$ και το $\hat{S}_{PL}$ μπορούν να υπολογιστούν, μέσω του τύπου της συνάρτησης **survfit()** στο πακέτο επιβίωσης. Να σημειωθεί ότι το $\hat{S}_{NA}$ πάντα αποφέρει μια σωστή συνάρτηση επιβίωσης, ενώ το $\hat{S}_{PL}$ θα αποφέρει περίεργα αποτελέσματα εάν $\exp\left(x^T \hat{\beta}\right) \lambda_0(t_i) > 1\ \gamma\iota\alpha\ \kappa\acute{\alpha}\pi o\iota\alpha\ t_i$. Στην πράξη, όμως, υπάρχει πολύ μικρή διαφορά μεταξύ των δύο μεθόδων αν το μέγεθος του δείγματος είναι σχετικά μεγάλο. Η συμπεριφορά του μικρού δείγματος αυτών των εκτιμητών και επιπλέον παραλλαγών συζητείται στον Andersen (1996).

Έχει δειχθεί από τους Tsiatis (1981) και Andersen και Gill (1982) ότι η μερική πιθανότητα μπορεί να θεωρηθεί ως κανονική πιθανότητα, με την έννοια ότι η εκτίμηση $\hat{\beta}$ έχει μια ασυμπτωτική κανονική κατανομή με μέση τιμή β και πίνακα διασποράς που δίνεται από τον αντίστροφα παρατηρούμενο πίνακα πληροφορίας

Fisher. Η πρώτη παράγωγος, γνωστή και ως συνάρτηση βαθμολογίας ή εξίσωση εκτίμησης για το β δίνεται από

$$\frac{\partial pl(\beta)}{\partial \beta} = \sum_i \int_0^\infty Y_i(t)\big(x_i - \bar{x}(\beta, t)\big) dN_i(t), \tag{2.3.13}$$

με $\bar{x}(\beta, t)$ το σταθμισμένο μέσο όρο των x_j στο σύνολο κινδύνου $R(t)$, το οποίο είναι

$$\bar{x}(\beta, t) = \frac{\sum_j Y_j(t) x_j \exp(x_j^T \beta)}{\sum_j Y_j(t) \exp(x_j^T \beta)}. \tag{2.3.14}$$

Η πληροφορία κατά Fisher της μερικής πιθανότητας δίνεται από

$$I_{pl}(\beta) = -\frac{\partial^2 pl(\beta)}{\partial \beta^2} = \int_0^\infty var(x|\beta, t) d\bar{N}(t), \tag{2.3.15}$$

με

$$var(x|\,\beta, t) = \frac{\sum_j Y_j (x_j - \bar{x}(\beta, s)(x_j - \bar{x}(\beta, t)))^T \exp(x_j^T \beta)}{\sum_j Y_j(t) \exp(x_j^T \beta)} \tag{2.3.16}$$

ο σταθμισμένος πίνακας συνδιακύμανσης στο $R(t)$.

Ομοίως, φαίνεται στα ίδια paper (Andersen και Gill, 1982; Tsiatis, 1981) ότι οι εκτιμήσεις των επιμέρους πιθανοτήτων επιβίωσης $\hat{S}(t|x)$ είναι ασυμπτωτικά κανονικές με μέσες τιμές $S(t|x)$ και πίνακα διασποράς που μπορεί να λαμβάνεται από την παρατηρούμενη πληροφορία Fisher της πλήρους πιθανότητας $l(\lambda_0, \beta)$. Είναι αδιάφορο ποια μέθοδος (NA ή PL) χρησιμοποιείται για να εκτιμήσουμε τις πιθανότητες, επειδή είναι ασυμπτωτικά ισοδύναμες. Η ασυμπτωτική διασπορά του $\hat{S}(t|x) = \hat{S}(t|x, \beta)$ περιπλέκεται από το γεγονός ότι αυτό εξαρτάται από το $\hat{\beta}$ τόσο άμεσα όσο και έμμεσα μέσω της εξάρτησης του $\hat{\Lambda}_0(t)$ στο $\hat{\beta}$. Η ασυμπτωτική διασπορά του $-\ln(\hat{S}(t|x)) = \hat{\Lambda}_0(t)\exp(x^T\hat{\beta})$ μπορεί να εκτιμηθεί με συνέπεια από

$$\int_0^t \left(\frac{\exp\big(x^T\hat{\beta}\big)}{\sum_j Y_j(s)\exp\big(x_j^T\hat{\beta}\big)}\right)^2 d\bar{N}(s) + \hat{q}(t|x)^T I_{pl}^{-1}\big(\hat{\beta}\big)\hat{q}(t|x), \tag{2.3.17}$$

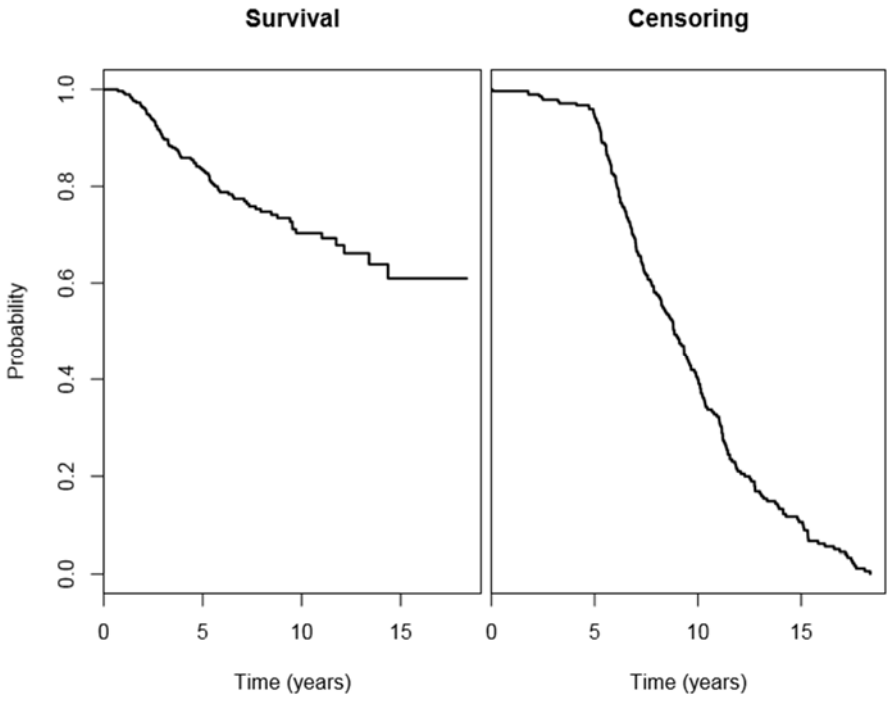
με

$$\hat{q}(t|x) = \int_0^t (x - \bar{x}(\hat{\beta}, s)) \frac{\exp\big(x^T\hat{\beta}\big)}{\sum_j Y_j(s)\exp\big(x_j^T\hat{\beta}\big)} d\bar{N}(s). \tag{2.3.18}$$

Ο τύπος βασίζεται στο εύρημα ότι τα $\hat{\beta}$ και $\hat{\Lambda}_0(t)$ είναι ασυμπτωτικά ανεξάρτητα αν το X επικεντρώνεται δυναμικά στο $\bar{x}(\hat{\beta}, t)$ και το $\hat{\Lambda}_0(t)$ αντικαθίσταται από το $\hat{\Lambda}_0(t)\exp(\bar{x}(\hat{\beta}, t)^T\hat{\beta})$. Ο τύπος μπορεί να χρησιμοποιηθεί για την κατασκευή διαστημάτων εμπιστοσύνης για το $\hat{S}(t|x)$ στην κλίμακα του ln, ή στην κλίμακα πιθανοτήτων. Ωστόσο, δεν έχουν όλα τα πακέτα λογισμικού τη δυνατότητα υπολογισμού τέτοιων διαστημάτων εμπιστοσύνης.

2.4 Παράδειγμα: Δεδομένα Καρκίνου του Μαστού στο NKI

Το παράδειγμα που χρησιμοποιείται σε όλο αυτό το κεφάλαιο είναι για το ολλανδικό σύνολο δεδομένων που αναφέρονται στον καρκίνο του μαστού και αφορούν δεδομένα σχετικά με τη συνολική επιβίωση των ασθενών με καρκίνο του μαστού, όπως συλλέγονται από το Ολλανδικό Ινστιτούτο Καρκίνου (Dutch Cancer Institute - NKI) στο Άμστερνταμ. Αυτό το σύνολο δεδομένων έγινε πολύ γνωστό επειδή χρησιμοποιήθηκε σε μια από τις πρώτες επιτυχημένες μελέτες που σχετίζονται με την επιβίωση του καρκίνου του μαστού σε δεδομένα γονιδιακής έκφρασης. Τα ευρήματα αυτής της μελέτης αναφέρθηκαν σε δύο σημαντικές ερευνητικές εργασίες , των van’t Veer et al. (2002) και van de Vijver et al. (2002) αντίστοιχα.



Σχήμα 2.4.1: Συναρτήσεις επιβίωσης και αποκοπής για το σύνολο δεδομένων του καρκίνου του μαστού.

Το σύνολο δεδομένων περιέχει τα κλινικά και γονιδιακά δεδομένα σε 295 ασθενείς με 79 συμβάντα, όπως αναφέρεται από τους van de Vijver et al. (2002). Τα δεδομένα αναλύθηκαν από τον van Houwelingen σε συνεργασία με τους στατιστικολόγους του NKI και δημοσιεύθηκαν από τους van Houwelingen et al. (2006). Οι συναρτήσεις επιβίωσης και αποκοπής αυτού του συνόλου δεδομένων φαίνονται στο Σχήμα 2.4.1. Η καμπύλη επιβίωσης φαίνεται να σταθεροποιείται μακροπρόθεσμα, με ποσοστό επιβίωσης περίπου 60%. Η καμπύλη αποκοπής δείχνει ότι η μέση παρακολούθηση (median follow-up) στο σύνολο δεδομένων είναι περίπου 9 χρόνια. Τα δεδομένα είναι διαθέσιμα στην ιστοσελίδα των van Houwelingen και Putter (2012) www.msbi.nl/DynamicPrediction. Το σύνολο δεδομένων περιέχει κλινικές και γονιδιακές πληροφορίες για τους ασθενείς. Σε αυτό το κεφάλαιο χρησιμοποιείται μόνο η κλινική πληροφορία. Η χρήση της υψηλών διαστάσεων γονιδιακής πληροφορίας είναι πέραν του αντικειμένου της παρούσας εργασίας. Οι πληροφορίες σχετικά με τους κλινικούς παράγοντες κινδύνου που διατίθενται μετά από χειρουργική επέμβαση, δίνονται στον Πίνακα 2.4.1.

Από τις κατηγορικές συμμεταβλητές, η ιστολογική διαβάθμιση (Histological Grade) και η αγγειακή εμβολή (Vascular Invasion) φαίνεται να έχουν σημαντική μονοδιάστατη επίδραση. Για τις συνεχείς συμμεταβλητές, οι μονοδιάστατοι συντελεστές παλινδρόμησης του Cox δίνονται στον Πίνακα 2.4.1. Καμία προσπάθεια δεν γίνεται σε αυτό το στάδιο για να βελτιστοποιηθεί η συναρτησιακή μορφή αυτών των συμμεταβλητών. Ένα απλό μοντέλο εφαρμόζεται με μια γραμμική επίδραση σε κάθε συνεχή συμμεταβλητή. Προφανώς, ο όγκος της διαμέτρου, η ηλικία του ασθενούς και το επίπεδο οιστρογόνων έχουν σημαντική μονοδιάστατη επίδραση. Ο Πίνακας 2.4.1 δίνει επίσης τους συντελεστές παλινδρόμησης για το μοντέλο του Cox, συμπεριλαμβανομένων όλων των συμμεταβλητών. Το επίπεδο οιστρογόνων και η ιστολογική διαβάθμιση φαίνονται να είναι οι πιο σημαντικοί δείκτες για την πρόβλεψη.

Πίνακας 2.4.1: Οι κλινικοί παράγοντες κινδύνου και οι επιδράσεις τους για επιβίωση στο σύνολο δεδομένων του καρκίνου του μαστού. Παρουσιάζονται οι εκτιμώμενοι συντελεστές παλινδρόμησης (B) και τα τυπικά τους σφάλματα (SE) σε ξεχωριστά για κάθε («μονοδιάστατο») παράγοντα κινδύνου μοντέλα του Cox και σε μοντέλα του Cox συμπεριλαμβανομένων όλων των συμμεταβλητών («πολυμεταβλητά»).

			univariate		multivariate	
Covariate	Category	Freq.	B	SE	B	SE
Chemotherapy	No	185	0	0	0	0
	Yes	110	-0.235	0.240	-0.423	0.298
Hormonal surgery	No	255	0	0	0	0
	Yes	40	-0.502	0.426	-0.172	0.442
Type of surgery	Excision	161	0	0	0	0

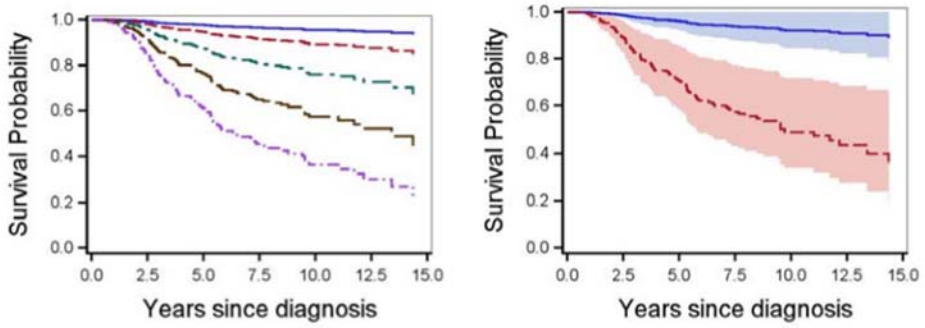
	Mastectomy	134	0.185	0.225	0.154	0.249	
Histological grade	Intermediate	101	0	0	0	0	
	Poorly differentiated	119	0.789	0.248	0.266	0.281	
	Well differentiated	75	-1.536	0.540	-1.308	0.547	
Vascular invasion	-	185	0	0	0	0	
	+	80	0.682	0.234	0.603	0.253	
	+/-	30	-0.398	0.474	-0.146	0.491	

Covariate	Min	Max	Mean	SD	B	SE	B	SE
Diameter (mm)	2	50	22.54	8.86	0.037	0.011	0.020	0.013
# positive nodes	0	13	1.38	2.19	0.064	0.046	0.074	0.052
Age at diagnosis	26	53	43.98	5.48	-0.058	0.020	-0.039	0.020
Estrogen level	-1.591	0.596	-0.260	0.567	-1.000	0.183	-0.750	0.211

Σημείωση: Το SD σημαίνει σταθερή απόκλιση

Για τις κατηγορικές συμμεταβλητές, η επίδραση συνήθως εκφράζεται ως αναλογία κινδύνου σε σχέση με τη βασική γραμμή κατηγορίας $HR = \exp(B)$ και το αντίστοιχο 95% διάστημα εμπιστοσύνης $(\exp(B - 1.96 \cdot SE), \exp(B + 1.96 \cdot SE))$. Για παράδειγμα, η μονοδιάστατη αναλογία κινδύνου για την μαστεκτομή: Η εκτομή (excision) ισούται με 1.203 με 95% διάστημα εμπιστοσύνης $(0.774, 1.870)$. Για συνεχείς συμμεταβλητές, το $\exp(B)$ θα δώσει την αναλογία του κινδύνου ανά μονάδα που αυξάνεται. Αυτό εξαρτάται από την κλιμάκωση της συμμεταβλητής. Για παράδειγμα, η ηλικία διάγνωσης στον Πίνακα 2.4.1 μετριέται σε χρόνια. Η αναλογία κινδύνου ανά έτος είναι πολύ κοντά στο ένα (0.944). Είναι πιο λογικό να εξετάσουμε την αναλογία κινδύνου κάθε δέκα χρόνια. Η μονοδιάστατη επίδραση αναλογίας κινδύνου ανά δέκα χρόνια δίνεται από το $\exp\big(10 \cdot (-0.058)\big) = 0.560$ με 95% διάστημα εμπιστοσύνης $(0.378, 0.829)$.

Η διακύμανση στην επιβίωση είναι άμεσα συνδεδεμένη με την τυπική απόκλιση του προγνωστικού δείκτη $PI = X^T\hat{\beta}$. Σε αυτά τα δεδομένα ισούται με 1.125. Η διακύμανση στην επιβίωση φαίνεται στο αριστερά πλαίσιο του Σχήματος 2.4.2 χρησιμοποιώντας εκατοστημόρια του δείκτη πρόβλεψης και η αβεβαιότητα των εκτιμώμενων καμπυλών επιβίωσης φαίνεται στο δεξιά πλαίσιο του Σχήματος 2.4.2 αντιπαραθέτοντας δύο ασθενείς. Ο ασθενής 1 έχει αρνητική αγγειακή εμβολή και καλά-διαφοροποιημένη ιστολογία και ο ασθενής 2 έχει θετική αγγειακή εμβολή και κακώς-διαφοροποιημένη ιστολογία. Και οι δύο έχουν συνεχείς συμμεταβλητές στη μέση τιμή και οι άλλες κατηγορικές συμμεταβλητές στη βασική τιμή αναφοράς.



Σχήμα 2.4.2: Καμπύλες πρόβλεψης επιβίωσης για εκατοστιαίες τιμές (10 (κορυφή), 25, 50, 75, 90 (κάτω)) του δείκτη πρόβλεψης (αριστερά) και καμπύλες πρόβλεψης επιβίωσης με 95% κατά σημείο διάστημα εμπιστοσύνης για δύο ασθενείς (δεξιά).

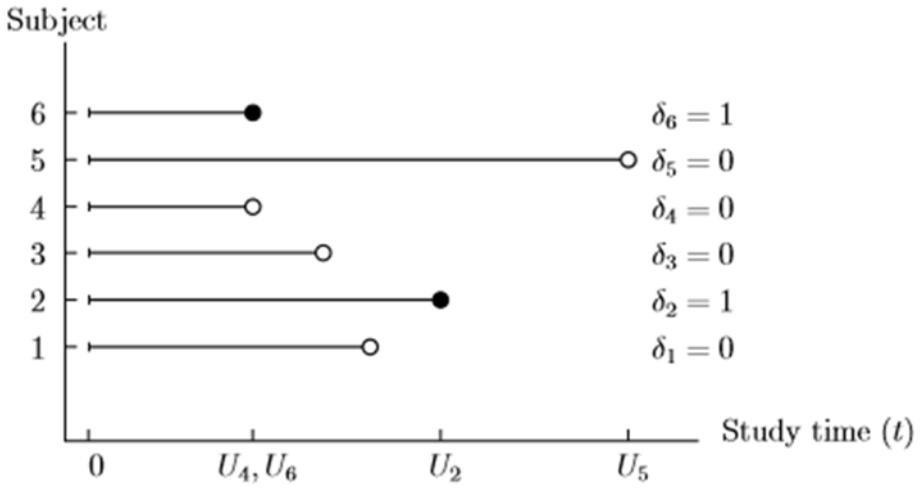
2.5 Μοντελοποίηση της Πρόβλεψης Δεδομένων Επιβίωσης

Ας υποθέσουμε ότι ο T δηλώνει την ώρα του συμβάντος, όπως ο θάνατος ή η ζωή, το C δηλώνει τον αποκομμένο χρόνο, για παράδειγμα, το τέλος μιας μελέτης ή το χρόνο ενός ατόμου που αποχωρεί από τη μελέτη. Οι T υποτίθεται ότι είναι ανεξάρτητοι και πανομοιότυπα κατανεμημένοι με τη συνάρτηση πυκνότητας πιθανότητας $\varphi(t)$ και τη συνάρτηση επιβίωσης $S(t)$. Για το δεξιά αποκομμένο τμήμα, γνωρίζουμε μόνο ότι $T_i > C_i$ με παρατηρούμενες τις C_i. Στη συνέχεια, τα δεδομένα επιβίωσης μπορούν να παρασταθούν από ζεύγη τυχαίων μεταβλητών $(U_i, \delta_i), i = 1, \dots, n$. Τα δ_i δείχνουν αν ο παρατηρούμενος χρόνος επιβίωσης U_i αντιστοιχεί σε ένα συμβάν ($\delta_i = 1$) ή αποκόπτεται ($\delta_i = 0$). Το U_i είναι ίσο με το T_i αν τηρείται η διάρκεια ζωής ή το συμβάν και ίσο με C_i αν αποκόπτεται. Μαθηματικά τα U_i και δ_i ορίζονται ως

$$U_i = \min(T_i, C_i) \tag{2.5.1}$$

και

$$\delta_i = I(T_i \leq C_i) = \begin{cases} 0, & \textit{για αποκομμένες παρατηρήσεις}, \\ 1, & \textit{για συμβάν που προκλήθηκε}. \end{cases} \tag{2.5.2}$$



Σχήμα 2.5.1: Παράδειγμα δεδομένων επιβίωσης σε μια κλίμακα μελέτης-χρόνου. Οι ακριβείς παρατηρήσεις υποδεικνύονται από μαύρες τελείες και οι αποκομμένες παρατηρήσεις από άδειες τελείες.

Στο Σχήμα 2.5.1, τα αντικείμενα 4 και 6 έχουν τον ίδιο παρατηρούμενο χρόνο επιβίωσης ($U_4 = U_6$), αλλά οι δείκτες διαγραφής τους είναι διαφορετικοί ($\delta_4 = 0, \delta_6 = 1$). Ως εκ τούτου, στην ανάλυση επιβίωσης, μας δίνεται ένα σύνολο δεδομένων $(x_i, U_i, \delta_i), i = 1, \dots, n$, όπου $x_i \in R^d, U_i \in R_+$ *και* $\delta_i \in \{0, 1\}$. Αντίθετα, στην εποπτευόμενη μάθηση, μας δίνεται ένα σύνολο δεδομένων εκπαίδευσης, $(x_i, y_i), i = 1, \dots, n$, όπου $x_i \in R^d$ και $y_i \in R$. Οι τιμές-στόχοι y_i μπορεί να είναι πραγματικές τιμές, όπως στο πρότυπο παλινδρόμησης, ή ετικέτες δυαδικής κλάσης όπως στην ταξινόμηση.

Η κλασική στατιστική προσέγγιση για την μοντελοποίηση των δεδομένων επιβίωσης αποσκοπεί στην εκτίμηση της συνάρτησης επιβίωσης $S(t)$, η οποία είναι η πιθανότητα ο χρόνος θανάτου να είναι μεγαλύτερος από ένα ορισμένο χρονικό διάστημα t. Γενικότερα, ο στόχος είναι να εκτιμηθεί η $S(t|x)$, ή η συνάρτηση επιβίωσης να εξαρτάται από τα χαρακτηριστικά του ασθενούς, που υποδηλώνονται ως χαρακτηριστικό διάνυσμα x. Υποθέτοντας ότι το πιθανοθεωρητικό μοντέλο $S(t|x)$ είναι γνωστό ή μπορεί να εκτιμηθεί με ακρίβεια από τα διαθέσιμα δεδομένα, το μοντέλο αυτό παρέχει πλήρη στατιστικό χαρακτηρισμό των δεδομένων. Ειδικότερα, μπορεί να χρησιμοποιηθεί για την πρόβλεψη και την επεξήγηση (δηλαδή, προσδιορίζοντας τα χαρακτηριστικά εισόδου που συνδέονται στενά με ένα αποτέλεσμα, όπως ο θάνατος).

Σε πολλές εφαρμογές, ο στόχος είναι να εκτιμηθεί (προβλεφθεί) η επιβίωση σε ένα προκαθορισμένο σημείο του χρόνου r, για παράδειγμα, η επιβίωση των ασθενών με καρκίνο δύο χρόνια μετά την αρχική διάγνωση, ή η κατάσταση της επιβίωσης των ασθενών ένα χρόνο μετά τη διαδικασία μεταμόσχευσης μυελού των οστών. Σε

γενικές γραμμές ο r μπορεί να είναι περίπου το μισό του μέγιστου παρατηρούμενου χρόνου επιβίωσης. Στη συνέχεια θα περιγράψουμε τη δυνατή μορφοποίηση (formalization) του προβλήματος αυτού υπό την προληπτική διευθέτηση, οδηγώντας στο σχηματισμό της δυαδικής ταξινόμησης.

Πρόβλημα ταξινόμησης (Classification problem): Γνωρίζοντας τα δεδομένα εκπαίδευσης επιβίωσης, $(x_i, U_i, \delta_i, y_i), i = 1, \dots, n,$ όπου $x_i \in R^d, U_i \in R_+, \delta_\iota \in \{0, 1\}$ $\kappa\alpha\iota$ $y_i \in \{-1, +1\}$, εκτιμάμε ένα μοντέλο ταξινόμησης $f(x)$ που προβλέπει την κατάσταση ενός αντικειμένου σε ένα προκαθορισμένο χρόνο r που βασίζεται στην είσοδο ή στις συμμεταβλητές x.

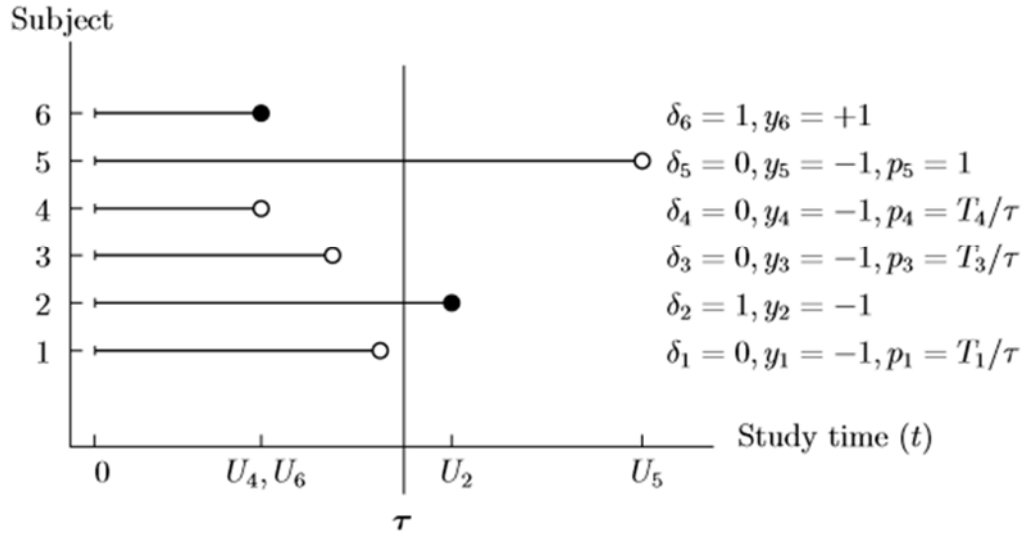
Η κατάσταση του αντικειμένου i σε χρόνο r είναι μια δυαδική ετικέτα κλάσης μέσω της παρακάτω κωδικοποίησης

$$y_i = \begin{cases} +1, & \alpha\nu\ U_i < r, \\ -1, & \alpha\nu\ U_i \geq r. \end{cases} \tag{2.5.3}$$

Να σημειωθεί ότι τα U_i και δ_i είναι διαθέσιμα μόνο για εκπαίδευση και όχι για πρόβλεψη (ή επίπεδο ελέγχου). Έτσι η πρόκληση για μοντελοποίηση της πρόβλεψης είναι να αναπτυχθούν νέες φόρμουλες ταξινόμησης που ενσωματώνουν αβέβαιης φύσης αποκομμένα δεδομένα.

Σε μια υποθετική μελέτη όπως φαίνεται στο Σχήμα 2.5.2, ας υποθέσουμε ότι η κατάσταση του αντικειμένου δίνεται από τη σχέση (2.5.3), τότε δεν υπάρχει καμία αμφιβολία στις καταστάσεις των αντικειμένων 2 και 6. Ομοίως, η κατάσταση επιβίωσης του αντικειμένου 5 είναι γνωστή, έστω και αν η παρατήρηση έχει αποκοπεί. Ωστόσο, οι καταστάσεις επιβίωσης για τα αντικείμενα 1, 3 και 4 είναι άγνωστες αφού οι παρατηρούμενοι χρόνοι επιβίωσης είναι μικρότεροι από r.

Υπάρχουν δύο απλοϊκοί τρόποι για να ενσωματώσουμε αποκομμένα δεδομένα σε τυπική διαμόρφωση ταξινόμησης:



Σχήμα 2.5.2: Παράδειγμα των δεδομένων επιβίωσης υπό τη ρύθμιση του προβλήματος πρόβλεψης. Ο στόχος είναι να βρούμε ένα μοντέλο που θα προβλέπει τις καταστάσεις των αντικειμένων σε χρόνο τ.

- Αντιμετωπίζουμε τον αποκομμένο χρόνο ως τον πραγματικό χρόνο του συμβάντος, δηλαδή αντικαθιστούμε το T_i με C_i. Αυτή η προσέγγιση υποτιμά τον πραγματικό χρόνο του συμβάντος επειδή $T_i > C_i$.
- Απλά αγνοούμε τα αποκομμένα δεδομένα και εκτιμάμε ένα δυαδικό ταξινομητή χρησιμοποιώντας μόνο τις ακριβείς παρατηρήσεις. Αυτή η προσέγγιση δίνει αναντίστοιχα μοντέλα, όπως αγνοούμε τις πληροφορίες που διατίθενται στα αποκομμένα δεδομένα.

Έχουν διερευνηθεί δύο διαφορετικές στρατηγικές για την ενσωμάτωση των αποκομμένων δεδομένων σε ταξινομητές που βασίζονται στην SVM:

1. Να σημειωθεί ότι οι αποκομμένες πληροφορίες είναι διαθέσιμες/ γνωστές για τα δεδομένα εκπαίδευσης, αλλά δεν είναι γνωστές κατά τη διάρκεια της πρόβλεψης. Τα αποκομμένα δεδομένα μπορούν να θεωρηθούν ως επιπλέον πληροφορία στο πλαίσιο του παραδείγματος της λεγόμενη μάθησης με τη χρήση πρόσθετης πληροφορίας (LUPI- Learning Using Privileged Information) (Vapnik, 2006, Vapnik and Vashist, 2009).
2. Μπορούμε να αναθέσουμε στις πιθανότητες να αντανακλούν την αβέβαιη κατάσταση των δειγμάτων των αποκομμένων δεδομένων. Ένας απλός κανόνας είναι να καθοριστεί η πιθανότητα ένα αντικείμενο να είναι ζωντανό σε χρόνο τ ανάλογα με τον (γνωστό) χρόνο επιβίωσης, όπως φαίνεται στο Σχήμα 2.5.2. Δηλαδή, $\Pr(y_i = -1|x_i) = U_i/\tau$ ή $\Pr(y_i = +1|x_i) = 1 - U_i/\tau$. Η ιδέα είναι ότι αν το U_i είναι μικρό, είναι πιο πιθανό το αντικείμενο i να μην επιβιώσει στο χρόνο τ. Από την άλλη, αν το U_i είναι πολύ κοντά

στο r, το αντικείμενο i θα είναι ζωντανό σε χρόνο r με μεγάλη πιθανότητα. Ως εκ τούτου τα δεδομένα επιβίωσης $(x_i, U_i, \delta_i), i = 1, \dots, n$, μπορούν να μεταφραστούν σε $(x_i, U_i, l_i), i = 1, \dots, n$. Για ακριβείς παρατηρήσεις, $l_i = y_i \in \{-1, +1\}, i = 1, \dots, m$. Για αποκομμένες παρατηρήσεις, $l_i = p_i \in [0, 1], i = m + 1, \dots, n$, όπου

$$p_i = \Pr(y_i = -1 | x_i) = U_i / r \tag{2.5.4}$$

θεωρεί την αβεβαιότητα σχετικά με την ένταξη κατηγορίας των x_i. Η έννοια της ανάθεσης πιθανότητας με την αβέβαιη κατάσταση μπορεί να επεκταθεί σε ακριβείς παρατηρήσεις. Για μια ακριβή παρατήρηση, έχουμε την κατάστασή της y_i με πιθανότητα $p_i = 1$. Στη συνέχεια τα δεδομένα επιβίωσης αναπαρίστανται ως $(x_i, U_i, p_i, y_i), i = 1, \dots, n$. Αυτή η επισημοποίηση των αποκομμένων δεδομένων οδηγεί στη λεγόμενη προσέγγιση μοντελοποίησης SVM με αβέβαιες ετικέτες (Niaf et al., 2011).

Και οι δυο προσεγγίσεις μοντελοποίησης παρουσιάζονται παρακάτω στην παράγραφο 3.5.

Τέλος, περιγράφουμε την εφαρμογή της κλασσικής ανάλυσης επιβίωσης υπό τον έξυπνο καθορισμό. Κλασικά μοντέλα ανάλυσης επιβίωσης περιγράφουν την εμφάνιση του συμβάντος μέσω των καμπυλών επιβίωσης και τα ποσοστά επικινδυνότητας και αναλύουν την εξάρτιση (αυτού του γεγονότος) για συμμεταβλητές μέσω των μοντέλων παλινδρόμησης (Aalen et al., 2008). Μια από τις πιο δημοφιλείς εκτιμήσεις της καμπύλης επιβίωσης είναι η προσέγγιση του μοντέλου του Cox που βασίζεται στο μοντέλο αναλογικών κινδύνων. Μόλις μια συνάρτηση επιβίωσης είναι γνωστή ή εκτιμώμενη (από τα δεδομένα εκπαίδευσης) μπορεί να χρησιμοποιηθεί για πρόβλεψη. Συγκεκριμένα, για τους νέους ελέγχους εισόδου x η πρόβλεψη επιτυγχάνεται με ένα απλό κανόνα κατωφλίου (simple thresholding rule)

$$y_i = \begin{cases} +1, & \alpha\nu\ S(t|x_i) < r, \\ -1, & \alpha\nu\ S(t|x_i) \geq r, \end{cases} \tag{2.5.5}$$

όπου η τιμή r του κατώτατου ορίου θα πρέπει να αντανακλά τα κόστη εσφαλμένης ταξινόμησης δεδομένων εκ των προτέρων. Εμείς θα υποθέσουμε ίσα κόστη εσφαλμένης ταξινόμησης. Επομένως, το κατώτατο όριο έχει οριστεί σε $r = 0{,}5$. Η προσέγγιση αυτή θα χρησιμοποιηθεί για την εκτίμηση της ακριβούς πρόβλεψης (σφάλμα ελέγχου) του μοντέλου του Cox από εμπειρικές συγκρίσεις που παρουσιάζονται στα Κεφάλαια 4 και 5.

ΚΕΦΑΛΑΙΟ 3: ΜΗΧΑΝΕΣ ΔΙΑΝΥΣΜΑΤΩΝ ΥΠΟΣΤΗΡΙΞΗΣ

Σε αυτό το κεφάλαιο θα συζητήσουμε για τη μηχανή διανυσμάτων υποστήριξης (SVM), μια προσέγγιση για την ταξινόμηση, που αναπτύχθηκε στη επιστημονική κοινότητα των υπολογιστών τη δεκαετία του 1990 και έχει αυξηθεί σε δημοτικότητα από τότε. Οι SVM έχουν αποδειχθεί ότι αποδίδουν καλά σε μια σειρά από ρυθμίσεις και συχνά θεωρούνται ένας από τους καλύτερους «έξω από το κουτί» ταξινομητές.

Η μηχανή διανυσμάτων υποστήριξης είναι μια γενίκευση ενός απλού και διαισθητικού ταξινομητή που ονομάζεται ταξινομητής μέγιστου περιθωρίου, τον οποίο εισάγουμε στην παράγραφο 3.1. Παρόλο που είναι κομψός και απλός, θα δούμε ότι αυτός ο ταξινομητής, δυστυχώς, δεν μπορεί να εφαρμοστεί στα περισσότερα σύνολα δεδομένων, αφού απαιτεί οι κλάσεις να μπορούν να διαχωριστούν από ένα γραμμικό όριο. Στην παράγραφο 3.2 έχουμε εισαγάγει τον ταξινομητή διανυσμάτων υποστήριξης, μια επέκταση του ταξινομητή μέγιστου περιθωρίου που μπορεί να εφαρμοστεί σε ένα ευρύτερο φάσμα περιπτώσεων. Στην παράγραφο 3.3 εισάγουμε τη μηχανή διανυσματικής υποστήριξης, η οποία είναι μια περαιτέρω επέκταση του ταξινομητή διανυσμάτων υποστήριξης για να δεχθεί μη γραμμικά όρια κλάσης. Οι μηχανές διανυσμάτων υποστήριξης προορίζονται για ρύθμιση δυαδικής ταξινόμησης, στην οποία υπάρχουν δύο κλάσεις. Στην παράγραφο 3.4 θα συζητήσουμε επεκτάσεις των μηχανών διανυσμάτων υποστήριξης στην περίπτωση των περισσότερων από δύο κλάσεις. Σχετικά με το πρόβλημα ταξινόμησης γενικότερα, και πιο συγκεκριμένα μέσω των μηχανών διανυσμάτων υποστήριξης περισσότερες πληροφορίες μπορείτε να βρείτε στην εργασία της Drosou (2013). Ιδιαίτερο ενδιαφέρον παρουσιάζουν αρκετές εφαρμογές των μηχανών διανυσμάτων υποστήριξης, όπως η εφαρμογή σε δεδομένα υψηλής διάστασης (Parpoula et al. , 2013), η χρησιμότητα τους στην ανάλυση ιατρικών δεδομένων (Drosou and Koukouvinos, 2017) καθώς επίσης και η αποδοτικότητα τους στην ανάλυση μη ισορροπημένων δεδομένων (Drosou et al., 2013, Drosou, 2015).

Οι άνθρωποι συχνά αναφέρονται αόριστα στον ταξινομητή μέγιστου περιθωρίου και τον ταξινομητή διανυσμάτων υποστήριξης. Για να αποφευχθεί αυτή η σύγχυση, θα εξετάσουμε στη συνέχεια προσεκτικά αυτές τις έννοιες.

3.1 Μέγιστος Ταξινομητής Περιθωρίου

Σε αυτή την ενότητα θα ορίσουμε ένα υπερεπίπεδο και θα εισάγουμε την έννοια ενός βέλτιστου ταξινομητή.

3.1.1 Τι Είναι Υπερεπίπεδο;

Σε ένα χώρο p-διαστάσεων, ένα υπερεπίπεδο είναι ένας επίπεδα συσχετισμένος υπόχωρος με (p-1)-διαστάσεις. Για παράδειγμα, σε δύο διαστάσεις, ένα υπερεπίπεδο είναι ένας επίπεδος μιας διάστασης υπόχωρος, με άλλα λόγια, μια γραμμή. Σε τρεις διαστάσεις, ένα υπερεπίπεδο είναι ένας επίπεδος υπόχωρος δύο διαστάσεων, δηλαδή ένα επίπεδο. Σε p>3 διαστάσεις, μπορεί να είναι δύσκολο να απεικονίσουμε ένα υπερεπίπεδο, αλλά η έννοια ενός (p-1) διαστάσεων επίπεδου υπόχωρου εξακολουθεί να εφαρμόζεται.

Ο μαθηματικός ορισμός για το υπερεπίπεδο είναι αρκετά απλός. Σε δύο διαστάσεις, ένα υπερεπίπεδο ορίζεται από την εξίσωση

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 = 0 \tag{3.1.1}$$

με παραμέτρους β_0, β_1 και β_2. Όταν λέμε ότι η σχέση (3.1.1) «ορίζει» το υπερεπίπεδο, εννοούμε ότι κάθε $X = (X_1, X_2)^T$ το οποίο κρατάει η (3.1.1) είναι ένα σημείο στο υπερεπίπεδο. Να σημειωθεί ότι η (3.1.1) είναι απλά η εξίσωση της ευθείας, αφού όντως στις δύο διαστάσεις το υπερεπίπεδο είναι μια γραμμή.

Η εξίσωση (3.1.1) μπορεί εύκολα να επεκταθεί για p-διαστάσεις:

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p = 0 \tag{3.1.2}$$

που ορίζει ένα p-διαστάσεων υπερεπίπεδο, πάλι με την έννοια ότι αν ένα σημείο $X = (X_1, X_2, \dots, X_n)^T$ σε χώρο p-διαστάσεων (δηλαδή ένα διάνυσμα μήκους p) ικανοποιεί τη σχέση (3.1.2), τότε το X βρίσκεται στο υπερεπίπεδο.

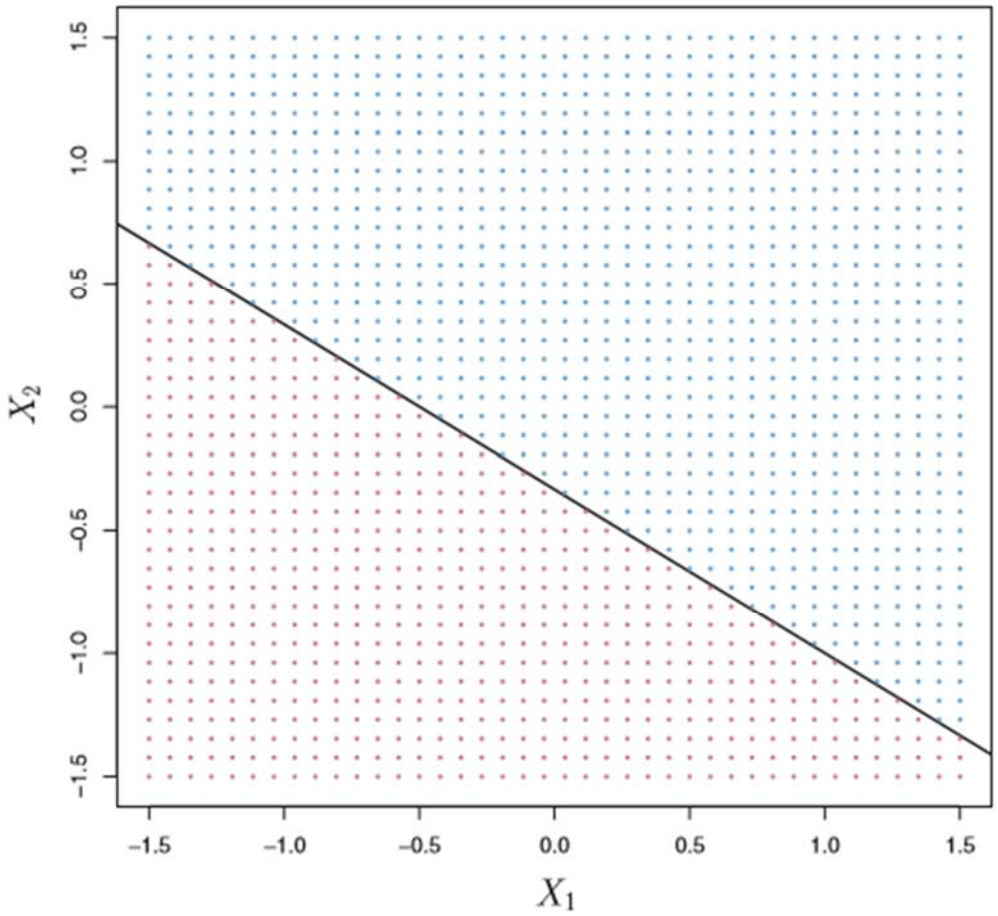
Τώρα, ας υποθέσουμε ότι το X δεν ικανοποιεί την (3.1.2). Καλύτερα,

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p > 0. \tag{3.1.3}$$

Τότε, αυτό μας λέει ότι το X βρίσκεται στη μια πλευρά του υπερεπιπέδου. Από την άλλη πλευρά, εάν

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p < 0, \tag{3.1.4}$$

τότε το X βρίσκεται στην άλλη πλευρά του υπερεπιπέδου. Έτσι μπορούμε να σκεφτούμε το υπερεπίπεδο ως διαιρούμενους χώρους p-διαστάσεων σε δύο ίσα μέρη. Κάποιος μπορεί εύκολα να προσδιορίσει σε ποια πλευρά του υπερεπιπέδου έγκειται ένα σημείο απλά από τον υπολογισμό του σημείου στην αριστερή πλευρά από την (3.1.2). Ένα υπερεπίπεδο σε χώρο δύο διαστάσεων παρουσιάζεται στο Σχήμα 3.1.1.



Σχήμα 3.1.1: Απεικονίζεται το υπερεπίπεδο $1 + 2X_1 + 3X_2 = 0$. Η μπλε περιοχή είναι το σύνολο των σημείων για τα οποία $1 + 2X_1 + 3X_2 > 0$ και η μωβ περιοχή είναι το σύνολο των σημείων για τα οποία $1 + 2X_1 + 3X_2 < 0$.

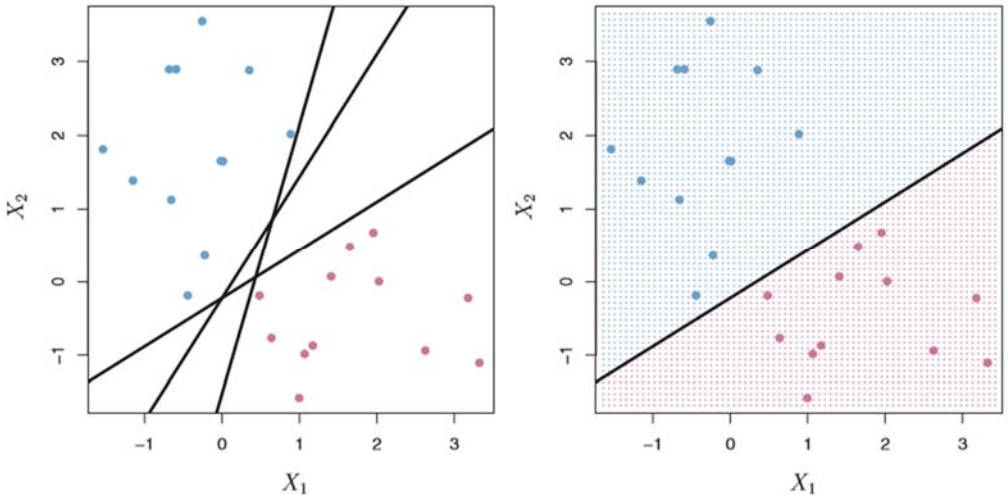
3.1.2 Ταξινόμηση Με Διαχωριστικό Υπερεπίπεδο

Τώρα υποθέτουμε ότι έχουμε έναν $n \times p$ πίνακα δεδομένων $\boldsymbol{X}$ ο οποίος αποτελείται από n παρατηρήσεις εκπαίδευσης σε ένα χώρο p-διαστάσεων,

$$x_1 = \begin{pmatrix} x_{11} \\ \vdots \\ x_{1p} \end{pmatrix}, \dots, x_n = \begin{pmatrix} x_{n1} \\ \vdots \\ x_{np} \end{pmatrix}, \tag{3.1.5}$$

και ότι αυτές οι παρατηρήσεις διαιρούνται σε δύο κλάσεις, που είναι $y_1, \dots, y_n \in \{-1, 1\}$ όπου το -1 αντιπροσωπεύει τη μια κλάση και το 1 την άλλη κλάση. Επίσης έχουμε παρατηρήσεις ελέγχου, ένα p-διάνυσμα από παρατηρούμενα χαρακτηριστικά $x^* = (x_1^* \dots x_p^*)^T$. Ο στόχος μας είναι να αναπτύξουμε ένα ταξινομητή που βασίζεται σε δεδομένα εκπαίδευσης, που θα ταξινομήσουν σωστά την παρατήρηση ελέγχου χρησιμοποιώντας τις χαρακτηριστικές μετρήσεις. Τώρα θα δούμε μια νέα προσέγγιση που βασίζεται στην ιδέα διαχωρισμού του υπερεπιπέδου.

Ας υποθέσουμε ότι είναι δυνατόν να κατασκευαστεί ένα υπερεπίπεδο που χωρίζει τις παρατηρήσεις εκπαίδευσης απόλυτα σύμφωνα με τις ετικέτες κατηγορίας τους (class labels). Παραδείγματα τριών τέτοιων διαχωρισμένων υπερεπιπέδων φαίνονται στο αριστερό πλαίσιο του Σχήματος 3.1.2. Μπορούμε να ονομάσουμε τις παρατηρήσεις από την μπλε κατηγορία ως $y_i = 1$ και αυτές από την μωβ κατηγορία ως $y_i = -1$.



Σχήμα 3.1.2: Αριστερά: Υπάρχουν δύο κλάσεις παρατηρήσεων, που φαίνονται με μπλε και μωβ, η καθεμία από τις οποίες έχει μετρήσεις σε δύο μεταβλητές. Τρία διαχωριστικά υπερεπίπεδα, από πολλά πιθανά, φαίνονται με μαύρο. Δεξιά: Ένα διαχωριστικό υπερεπίπεδο φαίνεται με μαύρο. Το μπλε και το μωβ πλέγμα δείχνουν τον κανόνα απόφασης που ελήφθη από τον ταξινομητή που βασίζεται στο διαχωριστικό υπερεπίπεδο: Η παρατήρηση ελέγχου που πέφτει στο μπλε τμήμα του πλέγματος θα ανατεθεί στην μπλε τάξη και η παρατήρηση ελέγχου που πέφτει στο μωβ τμήμα του πλέγματος θα ανατεθεί στην μωβ τάξη.

Στη συνέχεια ένα διαχωρισμένο υπερεπίπεδο έχει την ιδιότητα ότι

$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} > 0 \text{ , } \alpha\nu \; y_i = 1, \tag{3.1.6}$$

και

$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} < 0 \text{ , } \alpha\nu \; y_i = -1. \tag{3.1.7}$$

Ισοδύναμα ένα διαχωρισμένο υπερεπίπεδο έχει την ιδιότητα ότι

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}) > 0 \tag{3.1.8}$$

για όλα τα $i = 1, \dots, n$.

Αν υπάρχει ένα διαχωριστικό υπερεπίπεδο, μπορούμε να το χρησιμοποιήσουμε για να κατασκευάσουμε έναν ταξινομητή: σε μια παρατήρηση ελέγχου έχει εκχωρηθεί μια κλάση ανάλογα με το σε ποια πλευρά του υπερεπιπέδου βρίσκεται. Η δεξιά πλευρά του πλαισίου του Σχήματος 3.1.2 δείχνει ένα παράδειγμα ενός τέτοιου ταξινομητή. Δηλαδή, ταξινομούμε την παρατήρηση ελέγχου x^* που βασίζεται στο πρόσημο της $f(x^*) = \beta_0 + \beta_1 x_1^* + \beta_2 x_2^* + \cdots + \beta_p x_p^*$. Εάν η $f(x^*)$ είναι θετική, τότε αναθέτουμε την παρατήρηση ελέγχου στην κλάση 1 και αν η $f(x^*)$ είναι αρνητική, τότε την αναθέτουμε στην κλάση -1. Μπορούμε επίσης να κάνουμε χρήση του μεγέθους της $f(x^*)$. Αν η $f(x^*)$ απέχει πολύ από το μηδέν, τότε αυτό σημαίνει ότι το x^* βρίσκεται μακριά από το υπερεπίπεδο και έτσι μπορούμε να είμαστε σίγουροι για την ανάθεση της τάξης μας για το x^*. Από την άλλη πλευρά, αν η $f(x^*)$ είναι κοντά στο μηδέν, τότε το x^* βρίσκεται κοντά στο υπερεπίπεδο και έτσι είμαστε λιγότερο βέβαιοι για την κλάση που πρέπει να εκχωρηθεί στο x^*. Δεν αποτελεί έκπληξη και όπως βλέπουμε και από το Σχήμα 3.1.2, ένας ταξινομητής που βασίζεται σε ένα διαχωριστικό υπερεπίπεδο οδηγεί σε ένα γραμμικό όριο απόφασης.

3.1.3 Ταξινομητής Μέγιστου Περιθωρίου

Σε γενικές γραμμές, εάν τα δεδομένα μας μπορούν τέλεια να διαχωριστούν χρησιμοποιώντας ένα υπερεπίπεδο, τότε θα υπάρχει στην πραγματικότητα ένας άπειρος αριθμός από τέτοια υπερεπίπεδα. Αυτό οφείλεται στο γεγονός ότι ένα δεδομένο διαχωριστικό υπερεπίπεδο μπορεί συνήθως να μετατοπίσει ένα μικροσκοπικό κομμάτι προς τα πάνω ή προς τα κάτω, ή να περιστραφεί, χωρίς να έρχεται σε επαφή με καμία από τις παρατηρήσεις. Τρία πιθανά διαχωριστικά υπερεπίπεδα φαίνονται στο αριστερό μέρος του πλαισίου του Σχήματος 3.1.2. Για να κατασκευαστεί ένας ταξινομητής που βασίζεται σε ένα διαχωριστικό υπερεπίπεδο, πρέπει να έχουμε ένα λογικό τρόπο να αποφασίσουμε ποια από τα άπειρα δυνατά διαχωριστικά υπερεπίπεδα να χρησιμοποιήσουμε.

Μια φυσική επιλογή είναι το υπερεπίπεδο μέγιστου περιθωρίου (επίσης γνωστό και ως υπερεπίπεδο βέλτιστου ταξινομητή), το οποίο είναι ένα διαχωριστικό υπερεπίπεδο, είναι το πλέον απομακρυσμένο υπερεπίπεδο από τις παρατηρήσεις εκπαίδευσης. Δηλαδή, μπορούμε να υπολογίσουμε την (κάθετη) απόσταση από

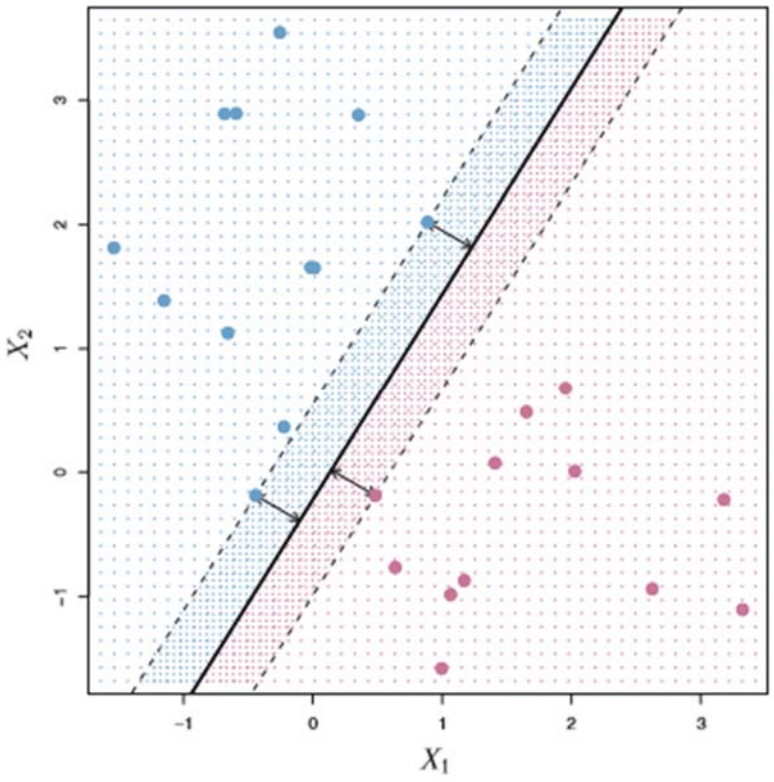
κάθε παρατήρηση εκπαίδευσης σε ένα δεδομένο διαχωριστικό υπερεπίπεδο. Η μικρότερη από αυτές τις αποστάσεις είναι η ελάχιστη απόσταση από τις παρατηρήσεις στο υπερεπίπεδο και είναι γνωστή ως περιθώριο. Το υπερεπίπεδο μέγιστου περιθωρίου είναι ένα διαχωριστικό υπερεπίπεδο για το οποίο το περιθώριο είναι μεγαλύτερο. Δηλαδή είναι το υπερεπίπεδο που έχει την πιο μακρινή ελάχιστη απόσταση στις παρατηρήσεις εκπαίδευσης. Στη συνέχεια μπορούμε να ταξινομήσουμε μια παρατήρηση ελέγχου με βάση την πλευρά του υπερεπιπέδου μέγιστου περιθωρίου που αυτή βρίσκεται. Αυτό είναι γνωστό ως ταξινομητής μέγιστου περιθωρίου. Ελπίζουμε ότι ο ταξινομητής που έχει μεγάλο περιθώριο στα δεδομένα εκπαίδευσης θα έχει επίσης ένα μεγάλο περιθώριο στα δεδομένα ελέγχου και ως εκ τούτου θα ταξινομήσει σωστά τις παρατηρήσεις ελέγχου. Αν και ο ταξινομητής μέγιστου περιθωρίου είναι συχνά επιτυχής, μπορεί επίσης να οδηγήσει σε υπερπροσαρμογή όταν το p είναι μεγάλο.

Αν $\beta_0, \beta_1, \dots, \beta_p$ είναι οι συντελεστές του υπερεπιπέδου μέγιστου περιθωρίου, τότε ο ταξινομητής μέγιστου περιθωρίου ταξινομεί την παρατήρηση ελέγχου x^* που βασίζεται στο πρόσημο της $f(x^*) = \beta_0 + \beta_1 x_1^* + \beta_2 x_2^* + \cdots + \beta_p x_p^*$.

Το Σχήμα 3.1.3 δείχνει το υπερεπίπεδο μέγιστου περιθωρίου στο σύνολο δεδομένων του Σχήματος 3.1.2. Συγκρίνοντας το δεξιά πλαίσιο του Σχήματος 3.1.2 με το Σχήμα 3.1.3, βλέπουμε ότι το υπερεπίπεδο μέγιστου περιθωρίου που φαίνεται στο Σχήμα 3.1.3 οδηγεί πράγματι σε μια μεγαλύτερη ελάχιστη απόσταση μεταξύ των παρατηρήσεων και του διαχωριστικού υπερεπιπέδου, που είναι ένα μεγαλύτερο περιθώριο. Κατά μια έννοια, το υπερεπίπεδο μέγιστου περιθωρίου αντιπροσωπεύει τη μεσαία γραμμή του ευρύτερου «σωλήνα» (”slab”) που μπορούμε να τοποθετήσουμε μεταξύ των δύο κατηγοριών.

Εξετάζοντας το Σχήμα 3.1.3 βλέπουμε ότι τρεις παρατηρήσεις εκπαίδευσης είναι σε ίση απόσταση από το υπερεπίπεδο μέγιστου περιθωρίου και βρίσκονται κατά μήκος των διακεκομμένων γραμμών δείχνοντας το πλάτος του περιθωρίου. Αυτές οι τρεις παρατηρήσεις είναι γνωστές ως διανύσματα υποστήριξης (support vectors), αφού είναι διανύσματα σε χώρο p-διαστάσεων (στο Σχήμα 3.1.3, $p = 2$) που «υποστηρίζουν» το υπερεπίπεδο μέγιστου περιθωρίου με την έννοια ότι αν αυτά τα σημεία μεταφέρθηκαν ελαφρώς τότε το υπερεπίπεδο μέγιστου περιθωρίου θα κινηθεί επίσης. Είναι ενδιαφέρον ότι το υπερεπίπεδο μέγιστου περιθωρίου εξαρτάται άμεσα από τα διανύσματα υποστήριξης, αλλά όχι από τις άλλες παρατηρήσεις: μια κίνηση σε οποιαδήποτε από τις άλλες παρατηρήσεις δεν θα επηρέαζε το διαχωριστικό επίπεδο, υπό την προϋπόθεση ότι η κίνηση των παρατηρήσεων δεν το προκαλεί να διασχίσει το όριο που θέτει το περιθώριο. Το γεγονός ότι το υπερεπίπεδο μέγιστου περιθωρίου εξαρτάται άμεσα μόνο από ένα μικρό υποσύνολο των παρατηρήσεων είναι μια σημαντική ιδιότητα που θα

προκύψει αργότερα σε αυτό το κεφάλαιο όταν συζητάμε για τον ταξινομητή διανυσμάτων υποστήριξης και μηχανές διανυσμάτων υποστήριξης.



Σχήμα 3.1.3: Υπάρχουν δύο κατηγορίες παρατηρήσεων, οι οποίες εμφανίζονται με μπλε και με μωβ. Το υπερεπίπεδο μέγιστου περιθωρίου παρουσιάζεται ως συμπαγής γραμμή. Το περιθώριο είναι η απόσταση από τη συνεχή γραμμή σε μια από τις διακεκομμένες γραμμές. Τα δύο μπλε σημεία και το μωβ σημείο που βρίσκονται στις διακεκομμένες γραμμές είναι διανύσματα υποστήριξης και η απόσταση από εκείνα τα σημεία στο περιθώριο υποδεικνύεται με βέλη. Το μωβ και το μπλε πλέγμα δείχνουν τον κανόνα απόφασης που ελήφθη από τον ταξινομητή και βασίζεται σε αυτό το διαχωριστικό επίπεδο.

3.1.4 Κατασκευή Του Μέγιστου Ταξινομητή Περιθωρίου

Θεωρούμε τώρα το καθήκον κατασκευής του υπερεπιπέδου μέγιστου περιθωρίου που βασίζεται σ' ένα σύνολο από n παρατηρήσεις εκπαίδευσης $x_1, \dots, x_n \in R^p$ και συσχετισμένες ετικέτες κλάσης $y_1, \dots, y_n \in \{-1, 1\}$. Συνοπτικά, το υπερεπίπεδο μέγιστου περιθωρίου είναι η λύση του προβλήματος βελτιστοποίησης

$$\max_{\beta_0, \beta_1, \dots, \beta_p} M \tag{3.1.9}$$

που ικανοποιεί τις

$$\sum_{j=1}^{p} \beta_j^2 = 1, \tag{3.1.10}$$

$$y_i\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}\right) \geq M, \quad \forall\, i = 1, \dots, n. \tag{3.1.11}$$

Αυτό το πρόβλημα βελτιστοποίησης (3.1.9)-(3.1.11) είναι στην πραγματικότητα πιο απλό απ' ότι φαίνεται. Αρχικά, ο περιορισμός στην (3.1.11) ότι

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} \geq M, \qquad \forall\, i = 1, \ldots, n$$

εγγυάται ότι κάθε παρατήρηση θα είναι στη σωστή πλευρά του υπερεπιπέδου, υπό τη συνθήκη ότι το M θα είναι θετικό. (Στην πραγματικότητα, για να είναι η κάθε παρατήρηση στη σωστή πλευρά του υπερεπιπέδου θα πρέπει απλά $y_i\big(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}\big) > 0$, έτσι ο περιορισμός στην (3.1.11) στην πραγματικότητα απαιτεί η κάθε παρατήρηση να είναι στη σωστή πλευρά του υπερεπιπέδου, με λίγο χώρο, με την προϋπόθεση ότι το M είναι θετικό.)

Επίσης, να σημειωθεί ότι η σχέση (3.1.10) δεν είναι πραγματικά ένας περιορισμός στο υπερεπίπεδο, αφού εάν $\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} = 0$ ορίζει ένα υπερεπίπεδο, τότε το ίδιο κάνει και η $k\big(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}\big) = 0, \gamma\iota\alpha\ \kappa\acute{\alpha}\theta\varepsilon\ k \neq 0$. Ωστόσο η (3.1.10) προσθέτει νόημα στην (3.1.11), η μια μπορεί να δείξει ότι με αυτό τον περιορισμό η κάθετη απόσταση από την i-οστή παρατήρηση στο υπερεπίπεδο δίνεται από

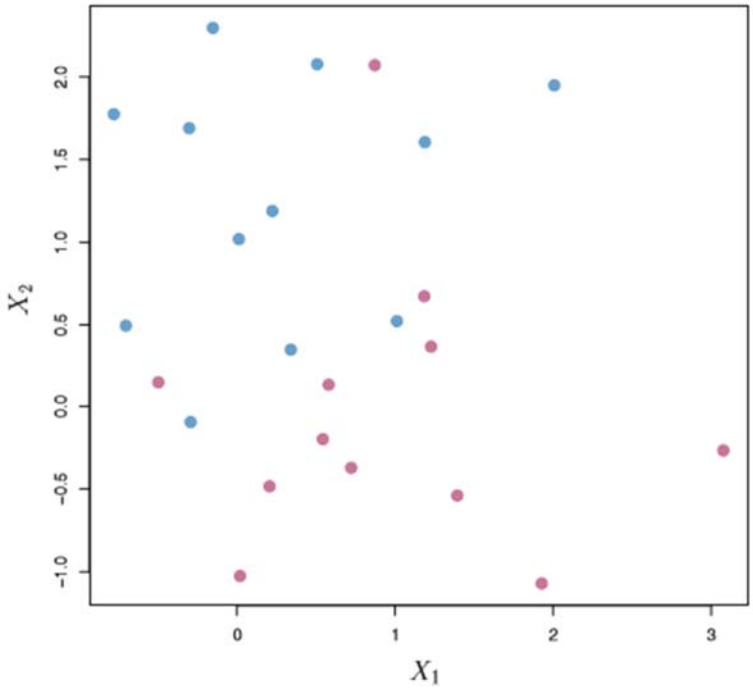
$$y_i\big(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}\big).$$

Επομένως, οι περιορισμοί (3.1.10) και (3.1.11) διασφαλίζουν ότι κάθε παρατήρηση είναι στη σωστή πλευρά από το υπερεπίπεδο και τουλάχιστον απόσταση M από το υπερεπίπεδο. Όμως το M αντιπροσωπεύει το περιθώριο του υπερεπιπέδου και το πρόβλημα βελτιστοποίησης επιλέγει $\beta_0, \beta_1, \ldots, \beta_p$ για να μεγιστοποιήσει το M. Αυτός είναι ακριβώς ο ορισμός του υπερεπιπέδου μέγιστου περιθωρίου. Το πρόβλημα (3.1.9)-(3.1.11) μπορεί να λυθεί αποτελεσματικά, αλλά δε θα ασχοληθούμε με τις λεπτομέρειες αυτής της βελτιστοποίησης.

3.1.5 Η Μη-Διαχωριστική Περίπτωση

Ο ταξινομητής μέγιστου περιθωρίου είναι ένας πολύ φυσικός τρόπος για να πραγματοποιηθεί μια ταξινόμηση, αν υπάρχει διαχωριστικό υπερεπίπεδο. Όμως, όπως έχουμε αφήσει να εννοηθεί, σε πολλές περιπτώσεις δεν υπάρχει διαχωριστικό υπερεπίπεδο και έτσι δεν υπάρχει ταξινομητής μέγιστου περιθωρίου. Σε αυτή την περίπτωση το πρόβλημα βελτιστοποίησης (3.1.9)-(3.1.11) δεν έχει λύση με $M > 0$. Ένα παράδειγμα φαίνεται στο Σχήμα 3.1.4. Σε αυτή την περίπτωση δεν μπορούμε να διαχωρίσουμε ακριβώς τις δύο τάξεις. Ωστόσο, όπως θα δούμε στη συνέχεια,

μπορούμε να επεκτείνουμε την έννοια του διαχωριστικού υπερεπιπέδου προκειμένου να αναπτυχθεί ένα υπερεπίπεδο που να διαχωρίζει τις κλάσεις χρησιμοποιώντας το λεγόμενο «μαλακό περιθώριο» (soft margin). Η γενίκευση του ταξινομητή μέγιστου περιθωρίου στην μη διαχωριστική περίπτωση είναι γνωστή ως ταξινομητής διανυσμάτων υποστήριξης.



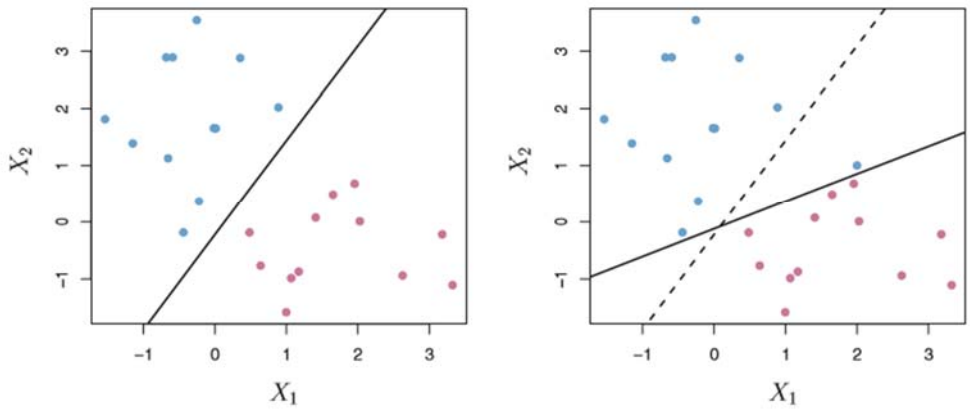
Σχήμα 3.1.4: Υπάρχουν δύο κλάσεις από παρατηρήσεις, που εμφανίζονται με μπλε και μωβ χρώμα. Σε αυτή την περίπτωση, οι δύο κλάσεις δεν είναι διαχωρίσιμες από ένα υπερεπίπεδο και έτσι ο ταξινομητής μέγιστου περιθωρίου δεν μπορεί να χρησιμοποιηθεί.

3.2 Ταξινομητής Διανυσμάτων Υποστήριξης

3.2.1 Επισκόπηση Του Ταξινομητή Διανυσμάτων Υποστήριξης

Στο Σχήμα 3.1.4 βλέπουμε ότι οι παρατηρήσεις που ανήκουν σε δύο κατηγορίες δεν είναι απαραιτήτως διαχωρίσιμες από ένα υπερεπίπεδο. Στην πραγματικότητα, ακόμα και αν υπάρχει ένα διαχωριστικό υπερεπίπεδο, τότε υπάρχουν περιπτώσεις στις οποίες ένας ταξινομητής που βασίζεται σ' ένα διαχωριστικό υπερεπίπεδο μπορεί να μην είναι επιθυμητός. Ένας ταξινομητής που βασίζεται σε ένα διαχωριστικό υπερεπίπεδο θα ταξινομήσει απαραιτήτως τέλεια όλες τις παρατηρήσεις εκπαίδευσης. Αυτό μπορεί να οδηγήσει σε ευαισθησία στις μεμονωμένες παρατηρήσεις. Ένα παράδειγμα φαίνεται στο Σχήμα 3.2.1. Η προσθήκη μιας και μόνο παρατήρησης στο δεξιά πλαίσιο του Σχήματος 3.2.1 οδηγεί

σε μια δραματική αλλαγή στο υπερεπίπεδο μέγιστου περιθωρίου. Το υπερεπίπεδο μέγιστου περιθωρίου που προκύπτει δεν είναι ικανοποιητικό, διότι έχει μόνο ένα μικρό περιθώριο. Αυτό είναι προβληματικό, διότι όπως αναφέρθηκε προηγουμένως, η απόσταση από μια παρατήρηση από το υπερεπίπεδο μπορεί να θεωρηθεί ως μια ένδειξη ότι η παρατήρησή μας έχει ταξινομηθεί σωστά. Επιπλέον, το γεγονός ότι το υπερεπίπεδο μέγιστου περιθωρίου είναι εξαιρετικά ευαίσθητο σε μια αλλαγή μιας και μόνο παρατήρησης υποδεικνύει ότι μπορεί να έχει υπερπροσαρμοστεί στα δεδομένα εκπαίδευσης.



Σχήμα 3.2.1: Αριστερά: Δύο κλάσεις παρατηρήσεων που φαίνονται με μπλε και μωβ, μαζί με το υπερεπίπεδο μέγιστου περιθωρίου. Δεξιά: Μια επιπλέον μπλε παρατήρηση έχει προστεθεί και οδηγεί σε δραματική αλλαγή του υπερεπιπέδου μέγιστου περιθωρίου και φαίνεται ως συμπαγής γραμμή. Η διακεκομμένη γραμμή δείχνει το υπερεπίπεδο μέγιστου περιθωρίου λήφθηκε από την απουσία του επιπρόσθετου σημείου.

Σε αυτή την περίπτωση, μπορεί να είμαστε πρόθυμοι να εξετάσουμε ένα ταξινομητή που βασίζεται σε ένα υπερεπίπεδο το οποίο δεν μπορεί απόλυτα να διαχωρίσει τις δύο τάξεις προς το συμφέρον

- της μεγαλύτερης ανθεκτικότητας σε μεμονωμένες παρατηρήσεις και
- της καλύτερης ταξινόμησης των περισσότερων παρατηρήσεων εκπαίδευσης.

Δηλαδή, θα μπορούσε να αξίζει τον κόπο να ταξινομήσεις εσφαλμένα μερικές παρατηρήσεις εκπαίδευσης ώστε να γίνει καλύτερη δουλειά στην ταξινόμηση των υπολοίπων παρατηρήσεων.

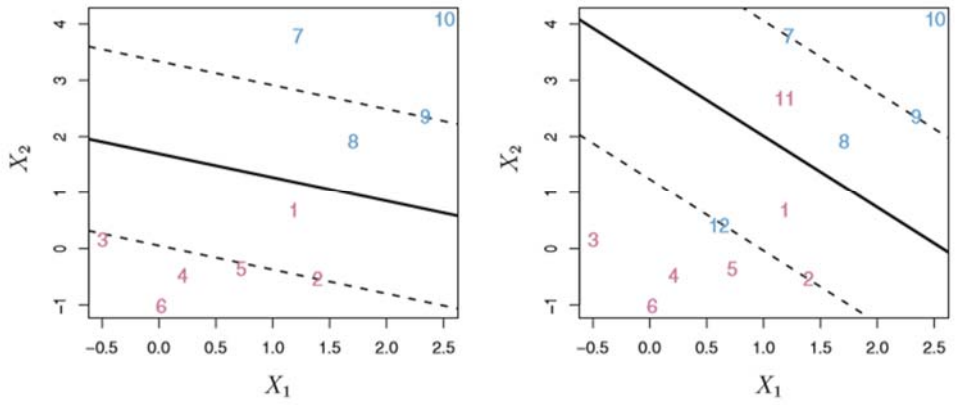
Ο ταξινομητής διανυσμάτων υποστήριξης, ο οποίος μερικές φορές ονομάζεται και ταξινομητής μαλακού περιθωρίου (soft margin classifier), κάνει αυτό ακριβώς. Επιδιώκει το μεγαλύτερο δυνατό περιθώριο, έτσι ώστε η κάθε παρατήρηση να μην είναι μόνο στη σωστή πλευρά του υπερεπιπέδου, αλλά και στη σωστή πλευρά του περιθωρίου, ενώ εμείς αντίθετα επιτρέπουμε μερικές παρατηρήσεις να μην είναι στη σωστή πλευρά του περιθωρίου ή ακόμα και στην εσφαλμένη πλευρά του

υπερεπιπέδου. (Το περιθώριο είναι μαλακό επειδή μπορεί να παραβιάζεται από μερικές παρατηρήσεις εκπαίδευσης.) Ένα παράδειγμα φαίνεται στο αριστερό πλαίσιο του Σχήματος 3.2.2. Οι περισσότερες από τις παρατηρήσεις βρίσκονται στη σωστή πλευρά του περιθωρίου. Ωστόσο, ένα μικρό υποσύνολο των παρατηρήσεων βρίσκεται στη λάθος πλευρά του περιθωρίου.

Μια παρατήρηση μπορεί να μην είναι μόνο από τη λάθος πλευρά του περιθωρίου, αλλά και από τη λάθος πλευρά του υπερεπιπέδου. Οι παρατηρήσεις που βρίσκονται στη λάθος πλευρά του υπερεπιπέδου αντιστοιχούν σε παρατηρήσεις εκπαίδευσης που έχουν ταξινομηθεί εσφαλμένα από τον ταξινομητή διανυσμάτων υποστήριξης. Το δεξιά πλαίσιο του Σχήματος 3.2.2 απεικονίζει αυτή την υπόθεση.

3.2.2 Λεπτομέρειες Του Ταξινομητή Διανυσμάτων Υποστήριξης

Ο ταξινομητής διανυσμάτων υποστήριξης ταξινομεί μια παρατήρηση ελέγχου ανάλογα με το σε ποια πλευρά του υπερεπιπέδου βρίσκεται. Το υπερεπίπεδο έχει επιλέξει να διαχωρίσει σωστά τις περισσότερες από τις παρατηρήσεις εκπαίδευσης στις δύο τάξεις, αλλά μερικές μπορεί να ταξινομήσει εσφαλμένα.



Σχήμα 3.2.2: Αριστερά: Ένας ταξινομητής διανυσμάτων υποστήριξης που ταίριαζε σ' ένα μικρό σύνολο δεδομένων. Το υπερεπίπεδο παρουσιάζεται ως μια συμπαγής γραμμή και τα περιθώρια εμφανίζονται ως διακεκομμένες γραμμές. Μωβ παρατηρήσεις: Οι παρατηρήσεις 3, 4, 5 και 6 είναι στη σωστή πλευρά του περιθωρίου, η παρατήρηση 2 είναι στο περιθώριο και η παρατήρηση 1 είναι στη λάθος πλευρά του περιθωρίου. Μπλε παρατηρήσεις: Οι παρατηρήσεις 7 και 10 είναι στη σωστή πλευρά του περιθωρίου, η παρατήρηση 9 είναι στο περιθώριο και η παρατήρηση 8 είναι στη λάθος πλευρά του περιθωρίου. Καμία παρατήρηση δεν είναι στη λάθος πλευρά του υπερεπιπέδου. Δεξιά: Το ίδιο με το αριστερά πλαίσιο με δύο επιπλέον σημεία, το 11 και το 12. Αυτές οι δύο παρατηρήσεις είναι στη λάθος πλευρά του υπερεπιπέδου και στη λάθος πλευρά του περιθωρίου.

Είναι η λύση στο πρόβλημα βελτιστοποίησης

$$\max_{\beta_0,\beta_1,\dots,\beta_p,\epsilon_1,\dots,\epsilon_n} M \tag{3.2.1}$$

που ικανοποιεί τις

$$\sum_{j=1}^{p} \beta_j^2 = 1, \tag{3.2.2}$$

$$y_i\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}\right) \geq M(1 - \epsilon_i), \tag{3.2.3}$$

$$\epsilon_i \geq 0, \sum_{i=1}^{n} \epsilon_i \leq C, \tag{3.2.4}$$

όπου C είναι μια μη αρνητική παράμετρος ρύθμισης. Όπως στην (3.1.11) το M είναι το πλάτος του περιθωρίου, επιδιώκουμε να κάνουμε αυτή την ποσότητα όσο το δυνατόν μεγαλύτερη. Στην 3.2.3 τα $\epsilon_1, \dots, \epsilon_n$ είναι χαλαρές μεταβλητές (slack variables) που επιτρέπουν στις μεμονωμένες παρατηρήσεις να είναι στη εσφαλμένη πλευρά του περιθωρίου ή του υπερεπιπέδου. Θα τα εξηγήσουμε όλα λεπτομερώς στη συνέχεια. Μόλις έχουμε λύσει τις (3.2.1)-(3.2.4), ταξινομούμε την παρατήρηση ελέγχου x^* όπως και πριν, απλά καθορίζοντας σε ποια πλευρά του υπερεπιπέδου θα βρίσκεται. Δηλαδή, έχουμε ταξινομήσει την παρατήρηση ελέγχου με βάση το πρόσημο της $f(x^*) = \beta_0 + \beta_1 x_1^* + \cdots + \beta_p x_p^*$.

Το πρόβλημα (3.2.1)-(3.2.4) φαίνεται περίπλοκο, αλλά η εικόνα για τη συμπεριφορά του μπορεί να γίνει μέσα από μια σειρά από απλές παρατηρήσεις που παρουσιάζονται παρακάτω. Πρώτα απ' όλα, η χαλαρή μεταβλητή ϵ_i μας λέει που βρίσκεται η i-οστή παρατήρηση, σε σχέση με το υπερεπίπεδο και το περιθώριο. Αν $\epsilon_i = 0$ τότε η i-οστή παρατήρηση είναι στη σωστή πλευρά του περιθωρίου, όπως είδαμε στην παράγραφο 3.1.4. Αν $\epsilon_i > 0$ τότε η i-οστή παρατήρηση είναι στη λάθος πλευρά του περιθωρίου και μπορούμε να πούμε ότι η i-οστή παρατήρηση έχει παραβιάσει το περιθώριο. Αν $\epsilon_i > 1$ τότε είναι στη λάθος πλευρά του υπερεπιπέδου.

Μπορούμε τώρα να εξετάσουμε το ρόλο της ρυθμιστικής παραμέτρου (tuning parameter) C. Στη σχέση (3.2.3), το C οριοθετεί το άθροισμα των ϵ_i και έτσι καθορίζει τον αριθμό και τη σοβαρότητα των παραβιάσεων στο περιθώριο (και στο υπερεπίπεδο) που θα ανεχθούμε. Μπορούμε να σκεφτούμε το C ως τον περιορισμό για την ποσότητα που το περιθώριο μπορεί να παραβιαστεί από n παρατηρήσεις. Αν $C = 0$ τότε δεν υπάρχει περιορισμός για τις παραβιάσεις στο περιθώριο και τότε έχουμε την περίπτωση όπου $\epsilon_1 = \cdots = \epsilon_n = 0$, στην οποία περίπτωση οι (3.2.1)-(3.2.4) απλά ισοδυναμούν με το πρόβλημα βελτιστοποίησης του υπερεπιπέδου μέγιστου περιθωρίου (3.1.9)-(3.1.11). (Φυσικά ένα υπερεπίπεδο μέγιστου περιθωρίου υπάρχει μόνο εάν οι δύο κλάσεις είναι διαχωρίσιμες.) Για $C > 0$ τότε δεν μπορούν περισσότερες από C παρατηρήσεις να είναι στη λάθος πλευρά του υπερεπιπέδου, επειδή αν μια παρατήρηση είναι στη λάθος πλευρά του

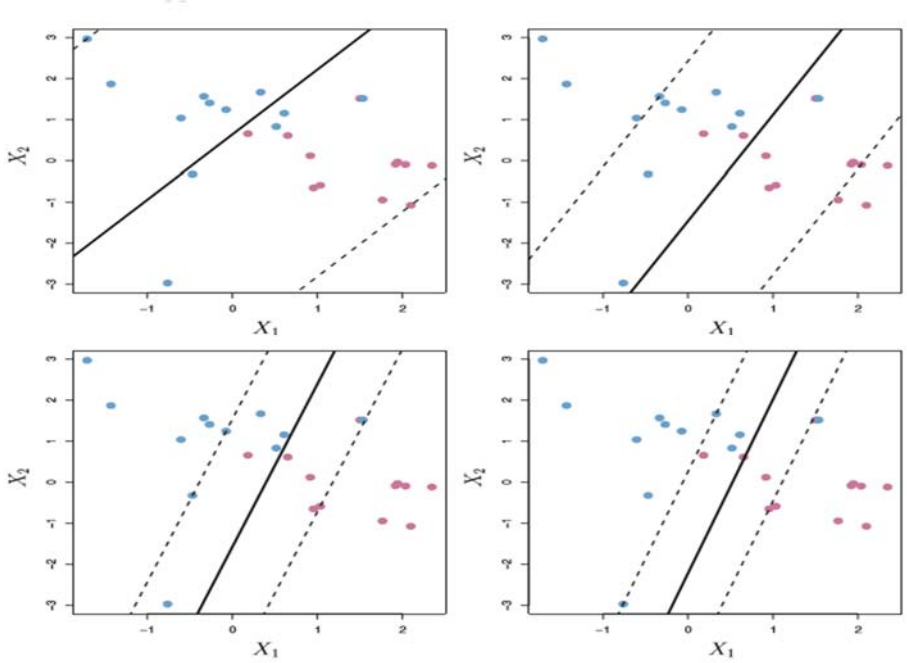
υπερεπιπέδου τότε $\epsilon_i > 1$ και η (3.2.4) απαιτεί ότι $\sum_{i=1}^{n} \epsilon_i \leq C$. Δεδομένου ότι ο προϋπολογισμός του C αυξάνεται, γινόμαστε πιο ανεκτοί στις παραβιάσεις του περιθωρίου και έτσι το περιθώριο θα διευρυνθεί. Αντίθετα, όσο το C μειώνεται, θα γινόμαστε λιγότερο ανεκτικοί στις παραβιάσεις του περιθωρίου και έτσι το περιθώριο θα στενεύει. Ένα παράδειγμα φαίνεται στο Σχήμα 3.2.3.

Πρακτικά, το C αντιμετωπίζεται ως μια παράμετρος ρύθμισης που γενικά επιλέγεται μέσω μιας διασταυρωμένης επικύρωσης (cross validation). Όπως και με τις παραμέτρους ρύθμισης που έχουμε δει, το C ελέγχει τη στάθμιση μεροληψίας - διακύμανσης της τεχνικής στατιστικής μάθησης. Όταν το C είναι μικρό, αναζητούμε στενά περιθώρια που σπάνια παραβιάζονται, αυτό ισοδυναμεί με ένα ταξινομητή που μπορεί ιδιαίτερα να προσαρμοστεί στα δεδομένα, τα οποία μπορεί να έχουν χαμηλή μεροληψία και υψηλή διακύμανση. Από την άλλη πλευρά όταν το C είναι μεγάλο το περιθώριο είναι μεγαλύτερο και επιτρέπουμε περισσότερες παραβιάσεις σε αυτό, αυτό ισοδυναμεί με λιγότερο σκληρή προσαρμοστικότητα των δεδομένων και εξασφαλίζουν έναν ταξινομητή που έχει δυνητικά μεγαλύτερη μεροληψία αλλά ίσως να έχει μικρότερη διακύμανση.

Το πρόβλημα βελτιστοποίησης (3.2.1)-(3.2.4) έχει μια πολύ ενδιαφέρουσα ιδιότητα: αποδεικνύεται ότι μόνο οι παρατηρήσεις που είτε βρίσκονται στο περιθώριο είτε παραβιάζουν το περιθώριο θα επιδράσουν στο υπερεπίπεδο και ως εκ τούτου και στον ταξινομητή που λαμβάνεται. Με άλλα λόγια, μια παρατήρηση που βρίσκεται αυστηρά στη σωστή πλευρά του περιθωρίου δεν επηρεάζει τον ταξινομητή διανυσμάτων υποστήριξης. Αλλάζοντας τη θέση της παρατήρησης δε θα άλλαζε εντελώς ο ταξινομητής, προϋποθέτοντας ότι η θέση του παραμένει στη σωστή πλευρά του περιθωρίου. Οι παρατηρήσεις που βρίσκονται ακριβώς στο περιθώριο ή στη λάθος πλευρά του περιθωρίου για την κλάση τους, είναι γνωστές ως διανύσματα υποστήριξης. Αυτές οι παρατηρήσεις επηρεάζουν τον ταξινομητή διανυσμάτων υποστήριξης.

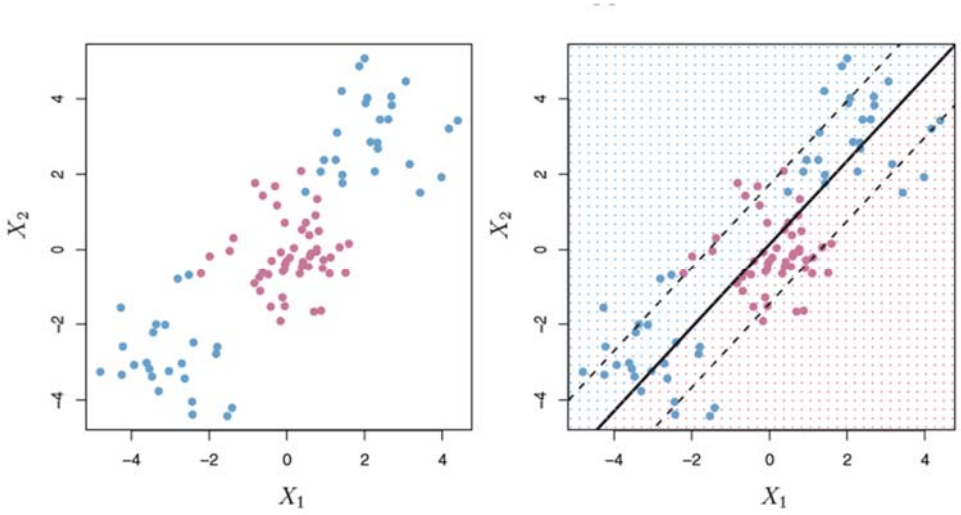
Το γεγονός ότι μόνο τα διανύσματα υποστήριξης επηρεάζουν τον ταξινομητή είναι σύμφωνα με τους προηγούμενους ισχυρισμούς μας ότι το C ελέγχει τη στάθμιση μεροληψίας - διακύμανσης του ταξινομητή διανυσμάτων υποστήριξης. Όταν η παράμετρος ρύθμισης C είναι μεγάλη τότε το περιθώριο είναι μεγάλο, πολλές παρατηρήσεις παραβιάζουν το περιθώριο και έτσι υπάρχουν πολλά διανύσματα υποστήριξης. Σε αυτή την περίπτωση, πολλές παρατηρήσεις εμπλέκονται στον καθορισμό του υπερεπιπέδου. Το πάνω αριστερά πλαίσιο του Σχήματος 3.2.3 απεικονίζει τη ρύθμιση αυτή: αυτός ο ταξινομητής έχει χαμηλή διακύμανση (αφού πολλές παρατηρήσεις είναι διανύσματα υποστήριξης) αλλά δυνητικά υψηλή μεροληψία. Αντίθετα, αν το C είναι μικρό, τότε θα υπάρχουν λιγότερα διανύσματα υποστήριξης και συνεπώς, ο ταξινομητής που προκύπτει θα έχει χαμηλή

μεροληψία, αλλά υψηλή διακύμανση. Το κάτω δεξιά πλαίσιο του Σχήματος 3.2.3 απεικονίζει αυτή τη ρύθμιση, με μόνο οκτώ διανύσματα υποστήριξης.



Σχήμα 3.2.3: Ένας ταξινομητής διανυσμάτων υποστήριξης προσαρμόστηκε χρησιμοποιώντας τέσσερις τιμές από την παράμετρο ρύθμισης C στις (3.2.1)-(3.2.4). Η μεγαλύτερη τιμή της C χρησιμοποιήθηκε στο πάνω αριστερά πλαίσιο και η μικρότερες τιμές στα πάνω δεξιά, κάτω αριστερά και κάτω δεξιά πλαίσια. Όταν το C είναι μεγάλο, τότε υπάρχει μεγάλη ανοχή για τις παρατηρήσεις που βρίσκονται στη λάθος πλευρά του περιθωρίου και έτσι το περιθώριο θα είναι μεγάλο. Όσο το C μειώνεται, η ανοχή για τις παρατηρήσεις που βρίσκονται στη λάθος πλευρά του περιθωρίου μειώνεται και το περιθώριο στενεύει.

Το γεγονός ότι ο κανόνας απόφασης του ταξινομητή διανυσμάτων υποστήριξης βασίζεται μόνο σε ένα δυνητικά μικρό υποσύνολο παρατηρήσεων εκπαίδευσης (τα διανύσματα υποστήριξης) σημαίνει ότι είναι αρκετά ισχυρό για τη συμπεριφορά των παρατηρήσεων που είναι πολύ μακριά από το υπερεπίπεδο. Αυτή η ιδιότητα είναι διαφορετική από μερικές άλλες μεθόδους ταξινόμησης, όπως η γραμμική διακριτική ανάλυση (Linear Discriminant Analysis). Υπενθυμίζουμε ότι ο κανόνας ταξινόμησης LDA εξαρτάται από τη μέση τιμή των παρατηρήσεων σε κάθε κατηγορία, καθώς και ο πίνακας συνδιακύμανσης σε κάθε κατηγορία υπολογίζεται χρησιμοποιώντας όλες τις παρατηρήσεις. Η λογιστική παλινδρόμηση, αντίθετα με την LDA, έχει πολύ χαμηλή ευαισθησία στις παρατηρήσεις μακριά από το όριο απόφασης.



Σχήμα 3.2.4: Αριστερά: Οι παρατηρήσεις που εμπίπτουν σε δύο κατηγορίες, με μη γραμμικό σύνορο μεταξύ τους. Δεξιά: Ο ταξινομητής διανυσμάτων υποστήριξης επιδιώκει ένα γραμμικό όριο και κατά συνέπεια το παρουσιάζει πολύ μειωμένο.

3.3 Μηχανές Διανυσμάτων Υποστήριξης

Αρχικά συζητάμε ένα γενικό μηχανισμό για τη μετατροπή ενός γραμμικού ταξινομητή σε έναν που να παράγει μη γραμμικά όρια απόφασης. Στη συνέχεια εισάγουμε τη μηχανή διανυσμάτων υποστήριξης, η οποία το κάνει αυτό με αυτόματο τρόπο.

3.3.1 Ταξινόμηση Με Μη-Γραμμικά Όρια Απόφασης (Decision Boundaries)

Ο ταξινομητής διανυσμάτων υποστήριξης είναι μια φυσική προσέγγιση για την ταξινόμηση στη ρύθμιση δύο κλάσεων, αν το όριο μεταξύ των δύο κατηγοριών είναι γραμμικό. Όμως, στην πράξη είμαστε μερικές φορές αντιμέτωποι με μη γραμμικά όρια κλάσης. Για παράδειγμα, ας εξετάσουμε τα δεδομένα στο αριστερά πλαίσιο του Σχήματος 3.2.4. Είναι σαφές ότι ο ταξινομητής διανυσμάτων υποστήριξης ή οποιοσδήποτε γραμμικός ταξινομητής θα εκτελεστεί ελάχιστα εδώ. Πράγματι, ο ταξινομητής διανυσμάτων υποστήριξης που φαίνεται στο δεξιά πλαίσιο του Σχήματος 3.2.4 είναι άχρηστος εδώ.

Υπάρχουν περιπτώσεις που βρισκόμαστε αντιμέτωποι με ανάλογες καταστάσεις. Η απόδοση της γραμμικής παλινδρόμησης μπορεί να πάσχει από μια μη γραμμική

σχέση μεταξύ των προγνωστικών και του αποτελέσματος. Στην περίπτωση αυτή, θα εξετάσουμε τη διεύρυνση του χώρου χαρακτηριστικών χρησιμοποιώντας συναρτήσεις πρόβλεψης, όπως τετραγωνικούς και κυβικούς όρους, προκειμένου να αντιμετωπιστεί αυτή η μη γραμμικότητα. Στην περίπτωση του ταξινομητή διανυσμάτων υποστήριξης θα μπορούσαμε να αντιμετωπίσουμε το πρόβλημα των ενδεχόμενων μη γραμμικών ορίων μεταξύ των κλάσεων με παρόμοιο τρόπο, με τη διεύρυνση του χαρακτηριστικού χώρου χρησιμοποιώντας τετραγωνικό, κυβικό και ακόμα και υψηλότερης τάξης πολυωνυμική συνάρτηση πρόγνωσης. Για παράδειγμα, αντί να προσαρμόσουμε έναν ταξινομητή διανυσμάτων υποστήριξης χρησιμοποιώντας p χαρακτηριστικά

$$X_1, X_2, \dots, X_p,$$

θα μπορούσαμε αντίθετα να προσαρμόσουμε ένα ταξινομητή διανυσμάτων υποστήριξης χρησιμοποιώντας $2p$ χαρακτηριστικά

$$X_1, X_1^2, X_2, X_2^2, \dots, X_p, X_p^2.$$

Τότε οι (3.2.1)-(3.2.4) θα γινόντουσαν

$$\max_{\beta_0,\beta_{11},\beta_{12},\dots,\beta_{p1},\beta_{p2},\epsilon_1,\dots,\epsilon_n} M \tag{3.3.1}$$

που ικανοποιεί τις

$$y_i\left(\beta_0 + \sum_{j=1}^{p}\beta_{j1}x_{ij} + \sum_{j=1}^{p}\beta_{j2}x_{ij}^2\right) \geq M(1-\epsilon_i),$$

$$\sum_{i=1}^{n}\epsilon_i \leq C, \qquad \epsilon_i \geq 0, \qquad \sum_{j=1}^{p}\sum_{k=1}^{2}\beta_{jk}^2 = 1.$$

Γιατί αυτή οδηγεί σε ένα μη γραμμικό όριο απόφασης. Στον διευρυμένο χαρακτηριστικό χώρο το όριο απόφασης που προκύπτει από τη σχέση (3.3.1) είναι στην πραγματικότητα γραμμικό. Αλλά στον αρχικό χαρακτηριστικό χώρο το όριο απόφασης είναι της μορφής $q(x) = 0$, όπου το q είναι ένα τετραγωνικό πολυώνυμο και οι λύσεις του είναι γενικά μη γραμμικές. Θα μπορούσε κάποιος επιπλέον να θέλει να διευρύνει το χαρακτηριστικό χώρο με ανώτερης τάξης όρους πολυωνύμου ή με όρους αλληλεπίδρασης της μορφής $X_j X_{j'}$ $για\, j \neq j'$. Εναλλακτικά, άλλες συναρτήσεις των προγνωστικών θα μπορούσαν να θεωρηθούν τα πολυώνυμα. Δεν είναι δύσκολο να δούμε ότι υπάρχουν πολλοί πιθανοί τρόποι για να διευρύνουμε το χαρακτηριστικό χώρο και ότι αν δεν είμαστε προσεκτικοί, θα μπορούσαμε να καταλήξουμε με ένα τεράστιο αριθμό χαρακτηριστικών. Τότε οι υπολογισμοί θα γινόντουσαν ανεξέλεγκτοι. Η μηχανή διανυσμάτων υποστήριξης, την οποία θα παρουσιάσουμε στη συνέχεια, μας επιτρέπει να διευρύνουμε το χαρακτηριστικό

χώρο που χρησιμοποιείται από τον ταξινομητή διανυσμάτων υποστήριξης με τέτοιο τρόπο που να οδηγεί σε αποτελεσματικούς υπολογισμούς.

3.3.2 Μηχανή Διανυσμάτων Υποστήριξης

Η μηχανή διανυσμάτων υποστήριξης SVM είναι μια επέκταση του ταξινομητή διανυσμάτων υποστήριξης που προκύπτει από τη διεύρυνση του χαρακτηριστικού χώρου με ένα συγκεκριμένο τρόπο, χρησιμοποιώντας πυρήνες. Θα συζητήσουμε τώρα την επέκταση αυτή, οι λεπτομέρειες της οποίας είναι οι συναρτήσεις πυρήνα που αποτελούν περίπλοκο εργαλείο και βρίσκονται πέρα από το πεδίο της συγκεκριμένης εργασίας. Όμως, η κύρια ιδέα περιγράφεται στην παράγραφο 3.3.1: μπορεί να θέλουμε να διευρύνουμε το χαρακτηριστικό μας χώρο, προκειμένου να υποδεχτούμε ένα μη γραμμικό όριο ανάμεσα στις κλάσεις. Η προσέγγιση του πυρήνα που περιγράφουμε εδώ είναι απλά μια αποτελεσματική υπολογιστική προσέγγιση για τη θέσπιση αυτής της ιδέας.

Εμείς δεν έχουμε συζητήσει ακριβώς πώς ο ταξινομητής διανυσμάτων υποστήριξης υπολογίζεται επειδή οι λεπτομέρειες γίνονται κάπως τεχνικά. Όμως, αποδεικνύεται ότι η λύση στο πρόβλημα του ταξινομητή διανυσμάτων υποστήριξης (3.2.1)-(3.2.4) περιλαμβάνει μόνο τα εσωτερικά γινόμενα των παρατηρήσεων (σε αντίθεση με τις ίδιες τις παρατηρήσεις). Το εσωτερικό γινόμενο δύο r-διανυσμάτων a και b ορίζεται ως $\langle a|b\rangle = \sum_{i=1}^{r} a_i b_i$. Έτσι το εσωτερικό γινόμενο των δύο παρατηρήσεων x_i, x_i' δίνεται από

$$\langle x_i|x_i'\rangle = \sum_{j=1}^{p} x_{ij}x_{i'j}. \tag{3.3.2}$$

Μπορεί να αποδειχθεί ότι

- Ο γραμμικός ταξινομητής διανυσμάτων υποστήριξης μπορεί να παρουσιαστεί ως

$$f(x) = \beta_0 + \sum_{i=1}^{n} a_i\langle x|x_i\rangle, \tag{3.3.3}$$

  όπου υπάρχουν n παράμετροι a_i, $i = 1, \dots, n$, μια σε κάθε παρατήρηση εκπαίδευσης.
- Για να εκτιμήσουμε τις παραμέτρους $a_1, \dots, a_n$ και β_0, τα μόνα που χρειαζόμαστε είναι τα $\binom{n}{2}$ εσωτερικά γινόμενα $\langle x_i|x_{i'}\rangle$ μεταξύ όλων των ζευγών των παρατηρήσεων εκπαίδευσης. (Η σημείωση $\binom{n}{2}$ σημαίνει

$n(n-1)/2$, και δίνει τον αριθμό των ζευγών μεταξύ ενός συνόλου n στοιχείων.)

Παρατηρούμε ότι στη σχέση (3.3.3), προκειμένου να αξιολογηθεί η συνάρτηση $f(x)$, πρέπει να υπολογίσουμε το εσωτερικό γινόμενο μεταξύ του σημείου x και του σημείου εκπαίδευσης x_i. Όμως, αποδεικνύεται ότι το a_i είναι μη μηδενικό μόνο για τα διανύσματα εμπιστοσύνης στη λύση - δηλαδή, αν μια παρατήρηση εκπαίδευσης δεν είναι διάνυσμα υποστήριξης, τότε το a_i του ισούται με μηδέν. Έτσι, αν το S είναι η συλλογή των δεικτών αυτών των σημείων υποστήριξης, μπορούμε να ξαναγράψουμε οποιαδήποτε συνάρτηση λύσης της μορφής (3.3.3) ως

$$f(x) = \beta_0 + \sum_{i \in S} a_i \langle x | x_i \rangle, \tag{3.3.4}$$

η οποία συνήθως περιλαμβάνει πολύ λιγότερα στοιχεία απ' ότι στη σχέση (3.3.3).

Για να συνοψίσουμε, στην εκπροσώπηση του γραμμικού ταξινομητή $f(x)$ και στον υπολογισμό των συντελεστών, το μόνο που χρειαζόμαστε είναι τα εσωτερικά γινόμενα.

Τώρα ας υποθέσουμε ότι κάθε φορά που το εσωτερικό γινόμενο (3.3.2) εμφανίζεται στην απεικόνιση (3.3.3), ή στον υπολογισμό της λύσης για τον ταξινομητή διανυσμάτων υποστήριξης, μπορούμε να το αντικαταστήσουμε με μια γενίκευση του εσωτερικού γινομένου της μορφής

$$K(x_i, x_{i'}), \tag{3.3.5}$$

όπου το K είναι κάποια συνάρτηση που θα αναφέρεται ως πυρήνας. Ένας πυρήνας είναι μια συνάρτηση πυρήνα που ποσοτικοποιεί την ομοιότητα δύο παρατηρήσεων. Για παράδειγμα, μπορούσαμε απλά να πάρουμε

$$K(x_i, x_{i'}) = \sum_{j=1}^{p} x_{ij} x_{i'j'}, \tag{3.3.6}$$

η οποία θα μας δώσει πίσω ακριβώς τον ταξινομητή διανυσμάτων υποστήριξης. Η εξίσωση (3.3.6) είναι γνωστή ως γραμμικός πυρήνας διότι ο ταξινομητής διανυσμάτων υποστήριξης είναι γραμμικός στα χαρακτηριστικά, ο γραμμικός πυρήνας ουσιαστικά ποσοτικοποιεί την ομοιότητα από ένα ζεύγος παρατηρήσεων χρησιμοποιώντας τη συσχέτιση του Pearson (Pearson correlation). Αλλά θα μπορούσε κανείς αντίθετα να επιλέξει μια άλλη μορφή για τη σχέση (3.3.5). Για παράδειγμα κάποιος θα μπορούσε να αντικαταστήσει κάθε δείγμα του $\sum_{j=1}^{p} x_{ij} x_{i'j}$ με την ποσότητα

$$K(x_i, x_{i'}) = (1 + \sum_{j=1}^{p} x_{ij} x_{i'j})^d. \quad (3.3.7)$$

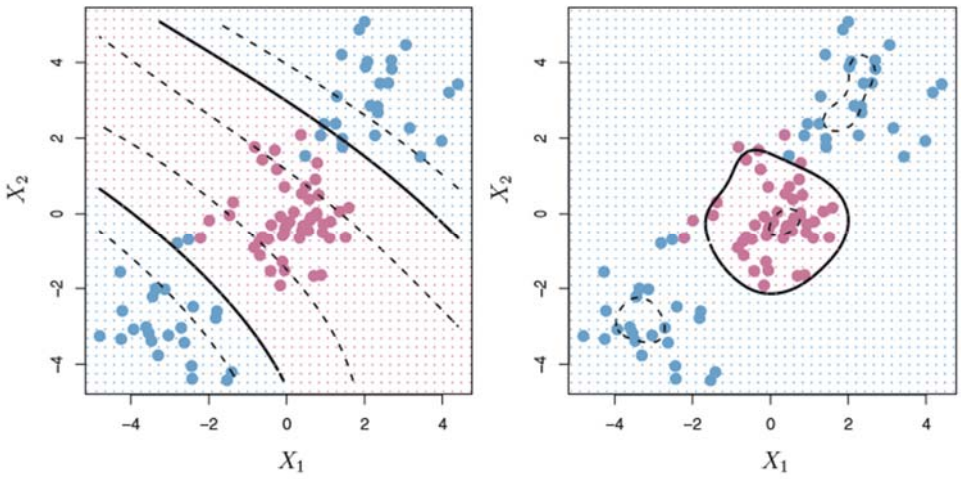
Αυτή είναι γνωστή ως ένας πυρήνας του πολυωνύμου βαθμού d, όπου το d είναι ένας θετικός ακέραιος αριθμός. Χρησιμοποιώντας ένα τέτοιο πυρήνα με $d > 1$, αντί του κανονικού γραμμικού πυρήνα (3.3.6), στον ταξινομητή διανυσμάτων υποστήριξης ο αλγόριθμος οδηγεί σε ένα πολύ πιο ευέλικτο όριο απόφασης. Ουσιαστικά ισοδυναμεί με την προσαρμογή του ταξινομητή διανυσμάτων υποστήριξης σε έναν υψηλότερων διαστάσεων χώρο που περιλαμβάνει πολυώνυμα βαθμού d, απ' ότι στον αρχικό χαρακτηριστικό χώρο. Όταν ο ταξινομητής διανυσμάτων υποστήριξης συνδυάζεται με ένα μη γραμμικό πυρήνα όπως στη σχέση (3.3.7), ο ταξινομητής που προκύπτει είναι γνωστός ως μηχανή διανυσμάτων υποστήριξης. Να σημειωθεί ότι σε αυτή την περίπτωση η (μη γραμμική) συνάρτηση έχει τη μορφή

$$f(x) = \beta_0 + \sum_{i \in S} a_i K(x, x_i). \quad (3.3.8)$$

Το αριστερό πλαίσιο του Σχήματος (3.3.1) δείχνει ένα παράδειγμα μιας SVM με ένα πυρήνα του πολυωνύμου που εφαρμόζεται σε μη γραμμικά δεδομένα από το Σχήμα (3.2.4). Η προσαρμογή είναι μια σημαντική βελτίωση σε σχέση με το γραμμικό ταξινομητή διανυσμάτων υποστήριξης. Όταν $d = 1$, τότε η SVM μειώνει τον ταξινομητή διανυσμάτων υποστήριξης που φαίνεται νωρίτερα σε αυτό το κεφάλαιο.

Ο πυρήνας του πολυωνύμου που φαίνεται στη σχέση (3.3.7) είναι ένα παράδειγμα ενός πιθανού μη γραμμικού πυρήνα, αλλά οι εναλλακτικές λύσεις αφθονούν. Μια άλλη δημοφιλής επιλογή είναι ο ακτινικός πυρήνας, που λαμβάνει τη μορφή

$$K(x_i, x_{i'}) = \exp\left(-\gamma \sum_{j=1}^{p} \left(x_{ij} - x_{i'j}\right)^2\right). \quad (3.3.9)$$



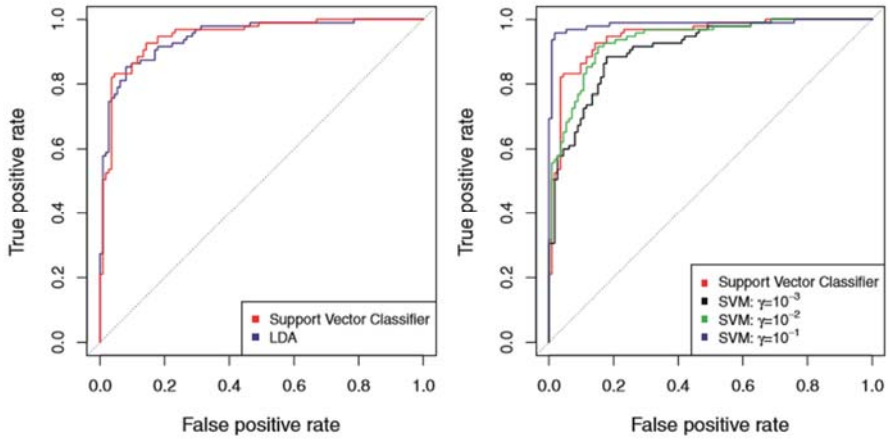
Σχήμα 3.3.1: Αριστερά: Μια SVM με ένα πυρήνα πολυωνύμου τρίτου βαθμού που εφαρμόζεται σε μη γραμμικά δεδομένα στο Σχήμα (3.2.4), με αποτέλεσμα έναν πολύ πιο κατάλληλο κανόνα απόφασης. Δεξιά: Εφαρμόζεται μια SVM με ακτινικό πυρήνα. Σε αυτό το παράδειγμα, ο πυρήνας είναι ικανός να συλλάβει το όριο απόφασης.

Στη σχέση (3.3.9) το γ είναι μια θετική σταθερά. Το δεξιά πλαίσιο του Σχήματος (3.3.1) δείχνει ένα παράδειγμα μιας SVM με ακτινικό πυρήνα σε μη γραμμικά δεδομένα και κάνει επίσης πολύ καλή δουλειά στο διαχωρισμό των δύο κλάσεων.

Πώς λειτουργεί πραγματικά ο ακτινικός πυρήνας (3.3.9); Αν μια δεδομένη παρατήρηση ελέγχου $x^* = (x_1^*, \dots, x_p^*)^T$ είναι μακριά από μια παρατήρηση εκπαίδευσης x_i όσο αφορά την Ευκλείδεια απόσταση, τότε το $\sum_{j=1}^{p}(x_j^* - x_{ij})^2$ θα είναι μεγάλο και έτσι το $K(x^*, x_i) = \exp(-\gamma \sum_{j=1}^{p}(x_j^* - x_{ij})^2)$ θα είναι πολύ μικρό. Αυτό σημαίνει ότι στη σχέση (3.3.8), το x_i δεν παίζει σχεδόν κανένα ρόλο στην $f(x^*)$. Υπενθυμίζουμε ότι η προβλεπόμενη ετικέτα κλάσης για την παρατήρηση ελέγχου x^* βασίζεται στο πρόσημο της $f(x^*)$. Με άλλα λόγια, οι παρατηρήσεις εκπαίδευσης που είναι μακριά από το x^* δεν θα παίξουν ουσιαστικά κανένα ρόλο στην προβλεπόμενη ετικέτα κατηγορίας για το x^*. Αυτό σημαίνει ότι ο ακτινικός πυρήνας έχει πολύ τοπική συμπεριφορά, με την έννοια ότι μόνο οι κοντινές παρατηρήσεις εκπαίδευσης έχουν ένα αποτέλεσμα στις ετικέτες κατηγορίας μιας παρατήρησης ελέγχου.

Ποιο είναι όμως το πλεονέκτημα της χρήσης ενός πυρήνα και όχι απλώς τη διεύρυνση του χαρακτηριστικού χώρου χρησιμοποιώντας συναρτήσεις από τα αρχικά χαρακτηριστικά, όπως στη σχέση (3.3.1); Ένα πλεονέκτημα είναι υπολογιστικό και ισοδυναμεί με το γεγονός ότι χρησιμοποιώντας τους πυρήνες, αρκεί να υπολογίσουμε μόνο το $K(x_i, x_i')$ για όλα τα $\binom{n}{2}$ διακριτά ζεύγη i, i'. Αυτό μπορεί να γίνει χωρίς σαφώς να εργάζονται στη διεύρυνση του χαρακτηριστικού χώρου. Αυτό είναι σημαντικό διότι σε πολλές εφαρμογές της SVM, ο διευρυμένος

χαρακτηριστικός χώρος είναι τόσο μεγάλος που οι υπολογισμοί είναι δυσεπίλυτοι. Για μερικούς πυρήνες, όπως τον ακτινικό πυρήνα στη σχέση (3.3.9), ο χαρακτηριστικός χώρος είναι αφανής και άπειρος σε διαστάσεις, έτσι δε θα μπορούσε ποτέ να κάνει εκεί τους υπολογισμούς.



Σχήμα 3.3.2: Καμπύλες ROC για το σύνολο δεδομένων εκπαίδευσης για την καρδιά. Αριστερά: Συγκρίνονται ο ταξινομητής διανυσμάτων υποστήριξης και ο LDA. Δεξιά: Ο ταξινομητής διανυσμάτων υποστήριξης συγκρίνεται με μια SVM χρησιμοποιώντας μια ακτινική βάση πυρήνα με $\gamma = 10^{-3}, 10^{-2}$ και 10^{-1}.

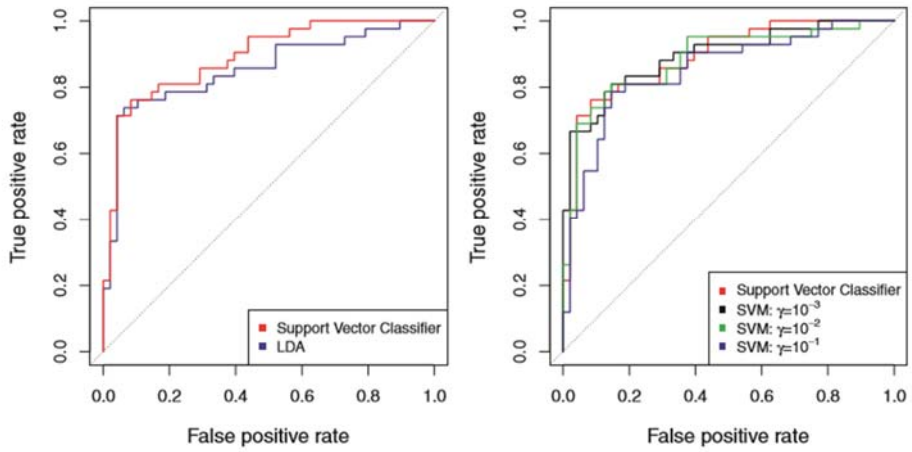
3.3.3 Εφαρμογή Σε Δεδομένα Καρδιακών Παθήσεων

Αυτή τη στιγμή είμαστε σε θέση να διερευνήσουμε πώς μια SVM συγκρίνεται με την LDA σε δεδομένα για την καρδιά. Τα δεδομένα αποτελούνται από 297 άτομα, τα οποία εμείς τυχαία χωρίζουμε σε 207 παρατηρήσεις εκπαίδευσης και 90 παρατηρήσεις ελέγχου.

Αρχικά προσαρμόζουμε την LDA και τον ταξινομητή διανυσμάτων υποστήριξης στα δεδομένα εκπαίδευσης. Να σημειωθεί ότι ο ταξινομητής διανυσμάτων υποστήριξης είναι ισοδύναμος με μια SVM χρησιμοποιώντας ένα πυρήνα πολυωνύμου βαθμού $d = 1$. Το αριστερό πλαίσιο του Σχήματος (3.3.2) παρουσιάζει τις καμπύλες ROC για τις προβλέψεις του συνόλου εκπαίδευσης τόσο για την LDA όσο και για τον ταξινομητή διανυσμάτων υποστήριξης. Και οι δύο ταξινομητές υπολογίζουν τα αποτελέσματα της μορφής $\hat{f}(X) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_p X_p$ για κάθε παρατήρηση. Για οποιαδήποτε δεδομένη αποκοπή t, ταξινομούμε τις παρατηρήσεις στην καρδιακή νόσο ή σε καμία κατηγορία καρδιοπάθειας ανάλογα με το αν $\hat{f}(X) < t$ ή $\hat{f}(X) \geq t$. Η καμπύλη ROC λαμβάνεται με σχηματισμό αυτών των προβλέψεων και τον υπολογισμό των ψευδώς θετικών και αληθώς θετικών ποσοστών για ένα εύρος τιμών του t. Ένας βέλτιστος ταξινομητής θα «αγκαλιάσει» την πάνω αριστερή γωνία του γραφήματος ROC. Σε αυτή την περίπτωση και η LDA και ο ταξινομητής

διανυσμάτων υποστήριξης έχουν καλές επιδόσεις, αν και υπάρχει μια πρόταση ότι ο ταξινομητής διανυσμάτων υποστήριξης μπορεί να είναι ελαφρώς ανώτερος.

Το δεξιά πλαίσιο του Σχήματος (3.3.2) παρουσιάζει τις καμπύλες ROC για τις SVM χρησιμοποιώντας έναν ακτινικό πυρήνα, με διάφορες τιμές του γ. Καθώς το γ αυξάνεται και η προσαρμογή γίνεται περισσότερο μη γραμμική, οι καμπύλες ROC βελτιώνονται. Χρησιμοποιώντας $\gamma = 10^{-1}$ φαίνεται να δίνουν μια σχεδόν τέλεια καμπύλη ROC. Όμως αυτές οι καμπύλες αντιπροσωπεύουν ποσοστά σφάλματος εκπαίδευσης, τα οποία μπορεί να είναι παραπλανητικά ως προς τις επιδόσεις σε νέα δεδομένα ελέγχου. Στο Σχήμα (3.3.3) εμφανίζονται καμπύλες ROC που έχουν υπολογιστεί σε 90 παρατηρήσεις ελέγχου. Παρατηρούμε μερικές διαφορές από τις καμπύλες εκπαίδευσης ROC.



Σχήμα 3.3.3: Καμπύλες ROC για το σύνολο ελέγχου σε δεδομένα καρδιάς. Αριστερά: Συγκρίνονται ο ταξινομητής διανυσμάτων υποστήριξης και η LDA. Δεξιά: Ο ταξινομητής διανυσμάτων υποστήριξης συγκρίνεται με μια SVM χρησιμοποιώντας μια ακτινική βάση πυρήνα με $\gamma = 10^{-3}, 10^{-2}$ και 10^{-1}.

Στο αριστερό πλαίσιο του Σχήματος (3.3.3), ο ταξινομητής διανυσμάτων υποστήριξης φαίνεται να έχει μικρό πλεονέκτημα σε σχέση με την LDA (αν και αυτές οι διαφορές δεν είναι στατιστικά σημαντικές). Στο δεξιά πλαίσιο, η SVM χρησιμοποιώντας $\gamma = 10^{-1}$, η οποία έδειξε τα καλύτερα αποτελέσματα για τα δεδομένα εκπαίδευσης, παράγει τις χειρότερες εκτιμήσεις στα δεδομένα ελέγχου. Αυτό είναι για ακόμα μια φορά απόδειξη ότι ενώ μια πιο ευέλικτη μέθοδος θα παράγει συχνά χαμηλότερα ποσοστά σφάλματος εκπαίδευσης, αυτό δεν οδηγεί απαραίτητα σε βελτιωμένη επίδοση σε δεδομένα ελέγχου. Οι SVM με $\gamma = 10^{-2}$ και $\gamma = 10^{-3}$ λειτουργούν συγκριτικά με τον ταξινομητή διανυσμάτων υποστήριξης και οι τρεις να ξεπεράσουν την SVM με $\gamma = 10^{-1}$.

3.4 Οι SVM Με Περισσότερες Από Δύο Κλάσεις

Μέχρι στιγμής η συζήτησή μας έχει περιοριστεί στην περίπτωση της δυαδικής ταξινόμησης (binary classification): δηλαδή ταξινόμηση σε ρύθμιση δύο κλάσεων. Πώς μπορούμε να επεκτείνουμε τις SVM στην πιο γενική περίπτωση που θα έχουμε κάποιο αυθαίρετο αριθμό κλάσεων; Αποδεικνύεται ότι η έννοια του διαχωριστικού υπερεπιπέδου, στο οποίο βασίζονται οι SVM, δεν προσφέρεται φυσικά για περισσότερες από δύο κλάσεις. Αν και μια σειρά από προτάσεις για την επέκταση των SVM στην περίπτωση K-κλάσεων έχουν γίνει, οι δύο πιο δημοφιλείς είναι οι προσεγγίσεις ένα-προς-ένα και ένα εναντίον όλων. Εμείς θα συζητήσουμε εν συντομία και τις δύο αυτές προσεγγίσεις.

3.4.1 Ταξινόμηση Ένα Προς Ένα

Ας υποθέσουμε ότι θέλουμε να εκτελέσουμε μια ταξινόμηση με χρήση των SVM και υπάρχουν $K > 2$ κλάσεις. Η ένα-προς-ένα ή όλων των ζευγών η προσέγγιση κατασκευάζει $\binom{K}{2}$ SVM, καθεμία από τις οποίες συγκρίνει ένα ζευγάρι κλάσεων. Για παράδειγμα, μια τέτοια SVM μπορεί να συγκρίνει την k-οστή κλάση, που κωδικοποιείται ως $+1$ με την k'-οστή κλάση που κωδικοποιείται ως -1. Έχουμε ταξινομήσει μια παρατήρηση ελέγχου χρησιμοποιώντας καθένα από τους $\binom{K}{2}$ ταξινομητές και αντιστοιχίζουμε τον αριθμό των φορών που η παρατήρηση ελέγχου έχει εκχωρηθεί σε καθεμία από τις K κατηγορίες. Η τελική ταξινόμηση γίνεται με την ανάθεση της παρατήρησης ελέγχου στην οποία αποδιδόταν πιο συχνά σε αυτές τις $\binom{K}{2}$ ταξινομήσεις ζευγών (pairwise classifications).

3.4.2 Ταξινόμηση Ένα Εναντίον Όλων

Η ένα εναντίον όλων είναι μια εναλλακτική διαδικασία για την εφαρμογή των SVM στην περίπτωση των $K > 2$ κλάσεων. Προσαρμόζουμε K SVM, κάθε φορά συγκρίνοντας μια από τις K κλάσεις στις υπόλοιπες $K - 1$ κλάσεις. Έστω οι $\beta_{0k}, \beta_{1k}, \dots, \beta_{pk}$ δηλώνουν τις παραμέτρους που προκύπτουν από μια σύγκριση της SVM της k-οστής κλάσης (που κωδικοποιείται ως $+1$) με τις υπόλοιπες (που κωδικοποιούνται ως -1). Έστω x^* υποδηλώνει την παρατήρηση ελέγχου. Αναθέτουμε την παρατήρηση στην κλάση για την οποία η $\beta_{0k} + \beta_{1k}x_1^* + \beta_{2k}x_2^* + \cdots + \beta_{pk}x_p^*$ είναι μεγαλύτερη, καθώς αυτή ισοδυναμεί με ένα υψηλό επίπεδο

αυτοπεποίθησης ότι η παρατήρηση ελέγχου ανήκει στην k-οστή κλάση και όχι σε κάποια από τις άλλες κατηγορίες.

3.5 Μηχανές Διανυσμάτων Υποστήριξης για Δεδομένα Επιβίωσης

Η ενότητα αυτή παρουσιάζει δύο πρόσφατα προηγμένες φόρμουλες, που βασίζονται στην SVM, κατάλληλες για την προγνωστική μοντελοποίηση των δεδομένων επιβίωσης. Η παρουσίαση ξεκινά με μια γενική περιγραφή αυτών των σκευασμάτων που βασίζονται στην SVM, ακολουθούμενη από την ειδική περιγραφή των ενσωματωμένα αποκομμένων δεδομένων σε αυτή τη φόρμουλα.

3.5.1 SVM+

Μια στρατηγική για να χειριστούμε τα δεδομένα επιβίωσης είναι το περιβάλλον γνωστό και ως «Μάθηση με Χρήση Προνομιακών Πληροφοριών» (LUPI – Learning Using Privileged Information) που αναπτύχθηκε από τον Vapnik (Vapnik, 2006, Vapnik and Vashist, 2009). Σε ένα πλούσιο κόσμο δεδομένων, υπάρχουν συχνά πρόσθετες πληροφορίες σχετικές με τα δείγματα εκπαίδευσης. Αυτές οι πρόσθετες πληροφορίες μπορούν εύκολα να αγνοηθούν από τυπικές επαγωγικές μεθόδους όπως η SVM. Η αποτελεσματική χρήση των πρόσθετων πληροφοριών κατά τη διάρκεια της εκπαίδευσης συχνά οδηγεί σε μια βελτιωμένη γενίκευση (Vapnik and Vashist, 2009).

Σύμφωνα με τη ρύθμιση LUPI, μας δίνεται ένα σύνολο από τριάδες $(x_i, x_i^*, y_i), i = 1, \dots, n$, όπου $x_i \in R^d, x_i^* \in R^k$ και $y_i \in \{-1, +1\}$. Το (x, y) είναι το «σύνηθες» δεδομένο εκπαίδευσης με ετικέτα και το (x^*) υποδηλώνει την επιπλέον προνομιακή πληροφορία που διατίθεται μόνο για τα δεδομένα εκπαίδευσης. Σημειώνεται ότι η πρόσθετη πληροφορία ορίζεται σε ένα διαφορετικό χαρακτηριστικό χώρο. Η προσέγγιση αυτή της SVM+ απεικονίζει εισόδους, x_i και x_i^*, σε δύο διαφορετικούς χώρους:

- Χώρος Απόφασης Z μέσω της απεικόνισης $\Phi(x): x \mapsto z$, ο οποίος είναι ο ίδιος χαρακτηριστικός χώρος που χρησιμοποιείται στην κλασσική SVM.
- Χώρος Διόρθωσης Z^* μέσω της απεικόνισης $\Phi^*(x): x \mapsto z^*$, ο οποίος αντικατοπτρίζει τις πρόσθετες πληροφορίες για τα δεδομένα εκπαίδευσης.

Ο στόχος της SVM+ είναι να εκτιμήσει τη συνάρτηση απόφασης $(w \cdot z) + b$ χρησιμοποιώντας τη συνάρτηση διόρθωσης $\xi(z^*) = (w^* \cdot z^*) + d \geq 0$ ως οι πρόσθετοι περιορισμοί στα σφάλματα εκπαίδευσης (ή χαλαρών μεταβλητών) στο

χώρο απόφασης. Ο ταξινομητής της SVM+ εκτιμάται από τα δεδομένα εκπαίδευσης επιλύοντας το ακόλουθο πρόβλημα βελτιστοποίησης:

Ελαχιστοποίηση της

$$\frac{1}{2}||w||^2 + \frac{\gamma}{2}||w^*||^2 + C\sum_{i=1}^{n}\xi_i$$

που υπακούει σε

$$\begin{gathered}\xi \succcurlyeq 0 \\ y_i((w \cdot z_i) + b \geq 1 - \xi_i, i = 1, \dots, n \\ \xi_i = (w^* \cdot z_i^*) + d, i = 1, \dots, n\end{gathered} \tag{3.5.1}$$

με $w \in R^d, b \in R, w^* \in R^k, d \in R$ και $\xi \in R_+^n$ ως μεταβλητές. Το σύμβολο $\succcurlyeq$ δηλώνει την κατά συνιστώσα ανισότητα και το R_+ δηλώνει τους μη αρνητικούς πραγματικούς αριθμούς.

Η προγνωστική μοντελοποίηση των δεδομένων επιβίωσης μπορεί να επισημοποιηθεί υπό την διαμόρφωση της SVM+/LUPI (3.5.1) όπως εξηγείται στη συνέχεια. Τα διαθέσιμα δεδομένα επιβίωσης (x_i, U_i, p_i, y_i) μπορούν να αναπαρασταθούν ως (x_i, x_i^*, y_i), όπου $x_i^* = (U_i, p_i)$ είναι η επιπρόσθετη πληροφόρηση. Στη συνέχεια το πρόβλημα της ανάλυσης επιβίωσης μπορεί να επισημοποιηθεί και να μοντελοποιηθεί χρησιμοποιώντας το παράδειγμα της SVM+/LUPI.

3.5.2 SVM με Αβέβαιες Ετικέτες

Αυτή η ενότητα περιγράφει τη νέα φόρμουλα που βασίζεται στην SVM (Niaf et al., 2011), η οποία εισάγει τις ετικέτες αβέβαιης κλάσης. Δηλαδή, κάποιες περιπτώσεις (δείγματα εκπαίδευσης) δεν συνδέονται με πεπερασμένης κλάσης ετικέτες. Για τέτοιες αβέβαιες ετικέτες, παρέχονται μόνο τα επίπεδα εμπιστοσύνης (ή οι πιθανότητες) όσο αφορά τις συμμετοχές της κατηγορίας. Στο πλαίσιο της ανάλυσης επιβίωσης, ακριβείς παρατηρήσεις έχουν γνωστές ετικέτες κλάσης (class labels) και οι αποκομμένες παρατηρήσεις έχουν αβέβαιες ετικέτες κλάσης.

Για τα μη διαχωριζόμενα δεδομένα επιβίωσης, έχουμε το ακόλουθο πρόβλημα βελτιστοποίησης:

Ελαχιστοποίηση της

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

που υπακούει σε

$$\begin{gathered}\xi \succcurlyeq 0 \\ y_i\big((w \cdot x_i) + b\big) \geq 1 - \xi_i, i = 1, \dots, m \\ \xi^- \succcurlyeq 0 \\ \xi^+ \succcurlyeq 0\end{gathered} \tag{3.5.2}$$

$$q_i^- - \xi_i^- \le (w \cdot x_i) + b \le q_i^+ + \xi_i^+, i = m+1, \dots, n.$$

με $w \in R^d, b \in R, \xi \in R_+^m, \xi^- \in R_+^{n-m}$ $\kappa\alpha\iota$ $\xi^+ \in R_+^{n-m}$ ως μεταβλητές. Το πρώτο τμήμα των περιορισμών είναι για τις ακριβείς παρατηρήσεις. Όσο αφορά τις αποκομμένες παρατηρήσεις, τις τιμές απόφασής τους, $(w \cdot x_i) + b$, οριοθετούνται από τα q_i^- και q_i^+. Τα όρια είναι συναρτήσεις των p_i, α και η, δηλαδή,

$$q_i^- = -\frac{1}{\alpha}\log\left(\frac{1}{p_i - \eta} - 1\right), \qquad q_i^+ = -\frac{1}{\alpha}\log\left(\frac{1}{p_i + \eta} - 1\right),$$

όπου $\alpha = \log(1/\eta - 1)$ είναι μια σταθερά και η είναι η μέγιστη τυπική απόκλιση της εκτίμησης της πιθανότητας από το p_i (Niaf et al., 2011, Platt, 1999).

Η τιμή του p_i που ορίζεται στη σχέση (2.5.4) κωδικοποιεί την πληροφορία σχετικά με το χρόνο επιβίωσης τόσο για τις αποκομμένες όσο και για τις ακριβείς παρατηρήσεις, που είναι διαθέσιμες στα δεδομένα εκπαίδευσης. Η σύνθεση αυτή μπορεί να επεκταθεί και σε μη γραμμική (πυρήνας) παραμετροποίηση χρησιμοποιώντας την τυποποιημένη μεθοδολογία της SVM. Αυτή η μέθοδος είναι γνωστή (και θα αναφέρεται) ως pSVM.

ΚΕΦΑΛΑΙΟ 4: ΠΕΙΡΑΜΑΤΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ

4.1 Εμπειρικές Συγκρίσεις για Συνθετικά Δεδομένα

Αυτή η ενότητα περιγράφει τις εμπειρικές συγκρίσεις μεταξύ των pSVM, SVM+/LUPI μεθόδων και την προσέγγιση του μοντέλου του Cox (Aalen et al., 2008). Η πρακτική εφαρμογή αυτών των μεθόδων για πεπερασμένα στοιχεία, περιλαμβάνει επιπλέον απλουστεύσεις, όπως θα συζητηθεί παρακάτω:

- Για την SVM+, η μη γραμμικότητα μοντελοποιείται μόνο στη διόρθωση χώρου (Liang et al., 2009). Δηλαδή, σε όλα τα πειράματα η απόφαση χώρου χρησιμοποιεί γραμμική παραμετροποίηση και η διόρθωση χώρου υλοποιείται μέσω μη γραμμικών (RBF) πυρήνων.
- Η pSVM χρησιμοποιεί είτε γραμμική, είτε μη γραμμική χαρτογράφηση στα πειράματα.

Κατά συνέπεια, η pSVM με RBF πυρήνα έχει τρεις παραμέτρους ρύθμισης, $C, \tilde{C}$ $και$ σ (παράμετρος πλάτους RBF), ενώ η SVM+ με RBF πυρήνα έχει τρεις παραμέτρους ρύθμισης, C, γ $και$ σ. Επιπλέον, η pSVM με γραμμικό πυρήνα έχει δύο παραμέτρους συντονισμού (C $και$ $\tilde{C}$). Αντίθετα, δεν υπάρχει καμία παράμετρος συντονισμού στην προσέγγιση του μοντέλου του Cox.

Οι εμπειρικές συγκρίσεις έχουν σχεδιαστεί για να κατανοήσουμε τα σχετικά πλεονεκτήματα και περιορισμούς των μεθόδων που βασίζονται στην SVM για την μοντελοποίηση των συνόλων δεδομένων επιβίωσης με διάφορα στατιστικά χαρακτηριστικά, όπως τον αριθμό των δειγμάτων εκπαίδευσης, τον θόρυβο στους παρατηρούμενους χρόνους επιβίωσης και την αναλογία αποκοπής. Το συνθετικό σύνολο δεδομένων δημιουργείται ως εξής (Cherkassky and Ma, 2002):

- Ρυθμίζουμε τον αριθμό των χαρακτηριστικών εισόδου d στο 30.
- Δημιουργούμε $x \in R^d$ με κάθε στοιχείο x_i να είναι ένας τυχαίος αριθμός ομοιόμορφα κατανεμημένος μέσα στο $[-1,1]$.
- Ορίζουμε το συντελεστή του διανύσματος ως $\beta = [1, 1, 2, 3, 3, 1, 1, 1, 1, 0, 2, 0, 2, 2, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]$.
- Δημιουργούμε την ώρα του συμβάντος T την ακόλουθη κατανομή $Exp((\beta x) + 2)$. Ο «θόρυβος του Gauss» (Gaussian noise) $\nu \sim N(0, 0.2)$ επίσης προστίθεται στο χρόνο συμβάντος T. Δημιουργώντας το χρόνο αποκοπής C που ακολουθεί την κατανομή $Exp(\lambda)$.
- Ο χρόνος επιβίωσης και η ένδειξη συμβάντος λαμβάνονται σύμφωνα με τις σχέσεις 2.5.1 και 2.5.2. Ο ρυθμός της εκθετικής κατανομής, λ, χρησιμοποιείται για να ελέγξει την αναλογία αποκοπής στο σύνολο εκπαίδευσης.
- Αντιστοιχίζουμε την ετικέτα της κλάσης σε κάθε διάνυσμα δεδομένων από τον κανόνα 2.5.3. Ο χρόνος που μας ενδιαφέρει, τ, βρίσκεται στην

ενδιάμεση τιμή μεταξύ των χρόνων επιβίωσης. Με αυτόν τον τρόπο, η εκ των προτέρων πιθανότητα για κάθε κλάση είναι περίπου η ίδια.

- Δημιουργούμε 400 δείγματα για εκπαίδευση, 400 για επαλήθευση και 2000 για έλεγχο.

Αυτό το σύνολο δεδομένων συνάδει με πιθανολογικές υποθέσεις (δηλαδή, εκθετική κατανομή), στις οποίες βασίζεται η κλασική προσέγγιση μοντελοποίησης. Έτσι, η προσέγγιση μοντελοποίησης του Cox αναμένεται να είναι πολύ ανταγωνιστική για το συνθετικό σύνολο δεδομένων.

Η ακόλουθη πειραματική διαδικασία χρησιμοποιήθηκε σε όλα τα πειράματα:

- Εκτιμάμε τον ταξινομητή χρησιμοποιώντας τα δεδομένα εκπαίδευσης.
- Βρίσκουμε τη βέλτιστη ρύθμιση παραμέτρων για κάθε μέθοδο, χρησιμοποιώντας τα δεδομένα επικύρωσης. Για την προσέγγιση του μοντέλου του Cox, τα δεδομένα επικύρωσης δεν χρησιμοποιούνται.
- Υπολογίζουμε το σφάλμα ελέγχου του τελικού μοντέλου χρησιμοποιώντας τα δεδομένα ελέγχου.

Πίνακας 4.1.1: Τα σφάλματα ελέγχου (%) για τα συνθετικά δεδομένα με 400 δείγματα εκπαίδευσης.

Trial	1	2	3	4	5	6	7	8	9	10
Cox	26.6	27.1	26.3	29.6	27.4	27.1	28.3	28.7	27.4	26.9
pSVM linear	25.7	**22.6**	**25.0**	**27.5**	**24.2**	26.5	26.1	**26.0**	25.6	**26.1**
pSVM rbf	**24.6**	25.7	25.8	27.9	25.7	**25.4**	25.7	26.9	26.2	26.8
LUPI	25.2	25.5	25.6	29.6	25.7	25.5	**25.6**	27.2	**25.0**	26.5

Πίνακας 4.1.2: Τα σφάλματα ελέγχου (%) για τα συνθετικά δεδομένα με 250 δείγματα εκπαίδευσης.

Trial	1	2	3	4	5	6	7	8	9	10
Cox	30.1	29.6	28.0	**27.6**	30.1	30.3	28.9	30.1	29.3	28.3
pSVM linear	**28.6**	**25.8**	**27.6**	28.1	29.8	**26.8**	**28.0**	28.1	27.3	29.0
pSVM rbf	28.9	26.9	30.4	**27.6**	30.5	28.1	27.5	**26.8**	27.7	28.1
LUPI	30.0	28.0	29.3	29.8	29.9	27.6	30.6	30.0	**25.0**	26.3

Πίνακας 4.1.3: Τα σφάλματα ελέγχου (%) για τα συνθετικά δεδομένα με 100 δείγματα εκπαίδευσης.

Trial	1	2	3	4	5	6	7	8	9	10
Cox	35.6	**31.3**	34.0	32.3	27.7	**30.6**	**30.6**	33.5	31.4	**28.4**
pSVM linear	**32.5**	33.0	33.5	30.0	**25.1**	33.5	36.9	**30.4**	31.4	30.8
pSVM rbf	**32.5**	32.0	33.8	**29.3**	32.2	32.2	34.2	31.4	33.1	29.9

LUPI	33.6	37.1	**32.0**	32.0	26.0	41.0	33.6	37.0	**30.9**	29.3

Πίνακας 4.1.4: Τα σφάλματα ελέγχου (%) για τα συνθετικά δεδομένα με 50 δείγματα εκπαίδευσης.

Trial	1	2	3	4	5	6	7	8	9	10
Cox	35.0	**31.6**	37.5	39.3	33.7	46.5	40.2	41.2	**33.9**	42.1
pSVM linear	**34.3**	35.1	37.6	34.3	34.8	40.3	41.8	40.9	35.7	**38.1**
pSVM rbf	35.8	**31.6**	37.5	33.1	**34.1**	**38.0**	**38.1**	**35.5**	35.8	39.1
LUPI	37.8	35.4	**35.5**	32.0	39.4	41.3	41.5	39.3	38.4	42.0

Η SVM+/LUPI έχει τρεις ρυθμιζόμενες παραμέτρους, τις $C, \gamma \; και \; \sigma$. Αυτές οι παράμετροι υπολογίστηκαν με τη χρήση των δεδομένων επικύρωσης και θεωρούμε το C στο διάστημα $[10^{-1}, 10^{2}]$, το γ στο $[10^{-3}, 10^{1}]$, και το σ στο $[2^{-2}, 2^{2}]$ για την επιλογή μοντέλου. Για την pSVM με πυρήνα RBF θεωρούμε C και $\tilde{C}$ σε διάστημα $[10^{-1}, 10^{2}]$ και το σ στο $[2^{-2}, 2^{2}]$.

Επιπλέον, το πείραμα πραγματοποιείται δέκα φορές με διαφορετικές τυχαίες υλοποιήσεις των δεδομένων εκπαίδευσης, επικύρωσης και ελέγχου. Σε αυτό το πείραμα, το μέσο ποσοστό της αποκομμένης παρατήρησης είναι 16.1% (ή περίπου 64 παρατηρήσεις στο σύνολο εκπαίδευσης αποκόπτονται). Τα σφάλματα ελέγχου για δέκα δοκιμές φαίνονται στον Πίνακα 4.1.1. Ο μέσος όρος των σφαλμάτων ελέγχου σε ποσοστό (μαζί με τις τυπικές αποκλίσεις) για το μοντέλο του Cox, την pSVM με γραμμικό πυρήνα, την pSVM με πυρήνα RBF και την LUPI είναι $27.5 \pm 1.0, 25.6 \pm 1.4, 26.1 \pm 0.9$ και 26.2 ± 1.4 αντίστοιχα.

Η pSVM με γραμμικό πυρήνα επιτυγχάνει το χαμηλότερο σφάλμα ελέγχου μεταξύ των μεθόδων στις περισσότερες δοκιμές. Συγκρίνοντας τη μέθοδο pSVM με διαφορετικούς πυρήνες, δεν αποτελεί έκπληξη να βρούμε ότι η pSVM με γραμμικό πυρήνα αποδίδει καλύτερα από ότι με RBF πυρήνα. Επειδή τα συνθετικά μας δεδομένα δημιουργούνται από ένα σχεδόν γραμμικό μοντέλο και υπάρχει εγγενής γραμμικότητα στα δεδομένα. Οι μέθοδοι με γραμμικό πυρήνα αναμένεται να αποδώσουν καλύτερα από ότι εκείνες με RBF πυρήνα.

Το μοντέλο του Cox έχει το μεγαλύτερο σφάλμα ελέγχου στις περισσότερες διαδρομές. Τα αποτελέσματα απεικονίζουν το δυνητικό πλεονέκτημα της χρήσης μιας μεθόδου που βασίζεται στην SVM. Σημειώνεται ότι οι μέθοδοι που βασίζονται στην SVM δίνουν παρόμοια ή ανώτερη απόδοση σε σχέση με τα κλασικά μοντέλα του Cox, παρόλο που τα δεδομένα εκπαίδευσης και ελέγχου παράγονται χρησιμοποιώντας την εκθετική κατανομή (για την οποία η μέθοδος του Cox είναι γνωστό ότι είναι στατιστικά άριστη).

A. Αριθμός Δειγμάτων Εκπαίδευσης

Για να διερευνηθεί η επίδραση του μεγέθους του δείγματος εκπαίδευσης στα σφάλματα ελέγχου, το μέγεθος του δείγματος εκπαίδευσης μειώνεται σε 250, 100 και 50. Τα μεγέθη των δειγμάτων επικύρωσης αλλάζουν αναλόγως. Τα αποτελέσματα αναφέρονται στους Πίνακες 4.1.2, 4.1.3 και 4.1.4.

Για τα 250 δείγματα εκπαίδευσης, ο μέσος όρος των σφαλμάτων ελέγχου για το μοντέλο του Cox, την pSVM με γραμμικό πυρήνα, την pSVM με RBF πυρήνα και την LUPI είναι $29.2 \pm 1.0, 27.9 \pm 1.1, 28.3 \pm 1.3$ *και* 28.7 ± 1.9 αντίστοιχα. Η pSVM με γραμμικό πυρήνα έχει την καλύτερη απόδοση σε πέντε δοκιμές. Οι σχετικές αποδόσεις μεταξύ των pSVM με RBF πυρήνα και LUPI είναι περίπου οι ίδιες. Ωστόσο η διαφορά απόδοσης μεταξύ του μοντέλου του Cox και της pSVM με γραμμικό πυρήνα κλείνει όταν το μέγεθος των δεδομένων εκπαίδευσης μειώνεται. Η παρατήρηση είναι πιο εμφανής όταν το μέγεθος του δείγματος μειώνεται στο 100. Για 100 δείγματα εκπαίδευσης, το μοντέλο του Cox έχει το χαμηλότερο σφάλμα ελέγχου σε τέσσερις δοκιμές, ενώ η pSVM με γραμμικό πυρήνα έχει την καλύτερη απόδοση σε τρεις μόνο δοκιμές.

Όταν το μέγεθος του δείγματος εκπαίδευσης μειώνεται επιπλέον σε 50, τόσο το μοντέλο του Cox, όσο και η pSVM με γραμμικό πυρήνα ξεπέρασαν σε απόδοση την pSVM με RBF πυρήνα. Αυτό μπορεί να αποδίδεται στην υψηλή διάσταση της εισόδου (λειτουργία) διανυσμάτων. Με τα υψηλών διαστάσεων διανύσματα εισόδου, οι μέθοδοι με γραμμικό πυρήνα αδυνατούν να συλλάβουν τη γραμμικότητα των δεδομένων όταν μόνο 50 δείγματα είναι διαθέσιμα για την εκπαίδευση. Αναμένεται επίσης ότι το εκτιμώμενο μοντέλο του Cox δεν είναι ακριβές, λόγω του μικρού μεγέθους του δείγματος.

Πίνακας 4.1.5: Σφάλματα ελέγχου ως συνάρτηση του μεγέθους του δείγματος εκπαίδευσης.

Training Size	50	100	250	400
Censoring	16.6%	15.9%	16.4%	16.1%
Cox	38.1± 4.6	**31.5± 2.4**	29.2± 1.0	27.5± 1.0
pSVM linear	37.3± 2.9	31.7± 3.1	**27.9± 1.1**	**25.6± 1.4**
pSVM rbf	**35.8± 2.4**	32.0± 1.5	28.3± 1.3	26.1± 0.9
LUPI	38.3± 3.2	33.2± 4.3	28.7± 1.9	26.2± 1.4

Πίνακας 4.1.6: Σφάλματα ελέγχου ως συνάρτηση του επιπέδου του θορύβου.

Noise Level	0	0.1	0.2	0.5
Censoring	15.9%	16.0%	17.2%	17.7%
Cox	**11.1± 0.4**	22.5± 1.7	28.7± 1.8	36.3± 1.3

pSVM linear	14.2± 1.0	**21.1± 1.8**	**27.1± 2.0**	34.8± 1.1
pSVM rbf	15.1± 1.5	22.5± 0.9	27.2± 2.1	36.0± 1.4
LUPI	14.3± 0.7	22.8± 1.7	27.5± 2.1	**34.7± 2.0**

Ο Πίνακας 4.1.5 δείχνει τη σχετική απόδοση των πέντε μεθόδων, ως συνάρτηση του μεγέθους του δείγματος. Η pSVM με γραμμικό πυρήνα ξεπερνά όλες τις άλλες μεθόδους, όταν το μέγεθος του δείγματος εκπαίδευσης είναι μεγαλύτερο από 250. Αυτό δεν αποτελεί έκπληξη, επειδή ο γραμμικός χώρος ταιριάζει με το συνθετικό μοντέλο δεδομένων. Όπως ήταν αναμενόμενο, με αύξηση του αριθμού του δείγματος εκπαίδευσης, το σχετικό πλεονέκτημα από τις μεθόδους που βασίζονται στην SVM είναι πιο αισθητό. Παρ' όλα αυτά, το μοντέλο του Cox είναι πιο ανταγωνιστικό για μέτρια μεγέθη δείγματος εκπαίδευσης (100).

Β. Επίπεδο θορύβου στο χρόνο επιβίωσης

Για να εξεταστεί η επίδραση του επιπέδου του θορύβου στο χρόνο επιβίωσης σε σφάλματα ελέγχου, οι θόρυβοι με διαφορετικές διακυμάνσεις προστίθενται στο χρόνο επιβίωσης. Η διακύμανση του θορύβου κυμαίνεται από 0 έως 0.5 και τα μεγέθη του δείγματος εκπαίδευσης και επικύρωσης διατηρούνται στο 250. Τα σφάλματα ελέγχου συνοψίζονται στον Πίνακα 4.1.6.

Είναι προφανές ότι τα σφάλματα ελέγχου μειώνονται σε όλες τις μεθόδους όταν η διακύμανση του θορύβου μειώνεται. Όταν δεν υπάρχει θόρυβος στο χρόνο επιβίωσης, τα δεδομένα παράγονται από μια κατανομή που ακολουθεί την υπόθεση του μοντέλου του Cox. Αναμένεται ότι το μοντέλο του Cox επιτυγχάνει το μικρότερο σφάλμα ελέγχου σύμφωνα με την υπόθεση χαμηλού θορύβου. Ωστόσο, η αύξηση του επιπέδου του θορύβου έχει πολύ μεγαλύτερη αρνητική επίδραση στην προσέγγιση μοντελοποίησης του Cox. Το σφάλμα ελέγχου αυξάνεται από 11% στο 36% όταν το επίπεδο θορύβου αυξηθεί από 0 σε 0.5. Εν τω μεταξύ, για τις ίδιες μεταβολές στα επίπεδα θορύβου, τα σφάλματα ελέγχου των προσεγγίσεων που βασίζονται στην SVM μέθοδο αυξάνονται από 14% σε 35%.

Εκτός από την υπόθεση ο θόρυβος να είναι μηδέν, η pSVM με γραμμικό πυρήνα επιτυγχάνει το χαμηλότερο μέσο όρο σφάλματος ελέγχου όταν η διακύμανση του θορύβου είναι μικρότερη από 0.2. Η LUPI, όμως, έχει την καλύτερη απόδοση όταν το επίπεδο θορύβου είναι μεγαλύτερο από 0.2. Μπορούμε να καταλήξουμε στο συμπέρασμα ότι οι μέθοδοι που βασίζονται στην SVM δείχνουν μεγαλύτερη αξιοπιστία σε θορυβώδη δεδομένα.

C. Ποσοστό αποκοπής

Έχουμε προσαρμόσει επίσης την αναλογία αποκοπής των δεδομένων εκπαίδευσης για να διερευνηθεί η επίδραση αποκοπής στα σφάλματα ελέγχου. Το ποσοστό των

αποκομμένων παρατηρήσεων στα δεδομένα εκπαίδευσης κυμαίνεται από 6% έως 46% στο πείραμά μας. Η διακύμανση του θορύβου έχει οριστεί έως 0.2 και τα μεγέθη των δειγμάτων εκπαίδευσης και επικύρωσης διατηρούνται στο 250. Τα αποτελέσματα του πειράματος συνοψίζονται στον Πίνακα 4.1.7.

Πίνακας 4.1.7: Σφάλματα ελέγχου ως συνάρτηση του ρυθμού διαγραφής.

Censoring	6.1%	30.6%	38.6%	46.0%
Cox	27.4± 2.0	33.8± 1.6	38.6± 2.2	42.0± 1.0
pSVM linear	**26.1± 1.6**	**31.5± 1.8**	36.8± 1.9	41.8± 2.4
pSVM rbf	26.9± 1.7	32.4± 2.5	**36.7± 1.3**	**39.9± 1.4**
LUPI	28.0± 2.7	32.5± 2.2	37.1± 2.1	41.3± 1.5

Όταν λιγότερο από το 30% των δεδομένων εκπαίδευσης αποκόπτονται, η γραμμική pSVM δίνει το χαμηλότερο σφάλμα ελέγχου. Αντιθέτως, εάν ένα μεγάλο μέρος των παρατηρήσεων αποκόπτονται (περίπου το 40% ή περισσότερο), η pSVM με πυρήνα RBF έχει καλύτερες επιδόσεις από όλες τις άλλες μεθόδους. Με περισσότερες αποκομμένες παρατηρήσεις στο σύνολο εκπαίδευσης, οι περισσότεροι παρατηρούμενοι χρόνοι επιβίωσης λαμβάνονται από το μη-γραμμικό φορέα 2.5.1. Άλλωστε, η γραμμικότητα μέσα στα δεδομένα δεν διατηρείται πλέον και οι μέθοδοι με μη-γραμμικές παραμετροποιήσεις (πυρήνας) αναμένεται να επιτύχουν καλύτερες επιδόσεις.

4.2 Ανάλυση Πραγματικών Δεδομένων

Αυτή η ενότητα περιγράφει εμπειρικές συγκρίσεις χρησιμοποιώντας τέσσερα σύνολα δεδομένων της πραγματικής ζωής από το πακέτο επιβίωσης της R (Therneau, 2013). Για όλες τις συγκρίσεις, ο κοινός χώρος απόφασης για την SVM+ χρησιμοποιεί τον γραμμικό πυρήνα, ενώ ο μοναδικός χώρος διόρθωσης χρησιμοποιεί τον RBF πυρήνα. Για την pSVM μέθοδο, διερευνήθηκαν τόσο ο γραμμικός, όσο και ο RBF πυρήνας. Σε όλα τα πειράματα, ο χρόνος ενδιαφέροντος r ορίστηκε με τη διάμεσο των παρατηρούμενων χρόνων επιβίωσης. Τα πειράματά μας για τα τέσσερα σύνολα ιατρικών δεδομένων ακολουθούν την παρακάτω διαδικασία (Liang et al., 2009, Cherkassky and Mulier, 2007):

- Χρησιμοποιούμε πέντε περιπτώσεις διασταυρωμένης επικύρωσης για να εκτιμήσουμε τα σφάλματα ελέγχου.

- Μέσα σε κάθε περίπτωση της εκπαίδευσης, η ρύθμιση των παραμέτρων (επιλογή μοντέλου) γίνεται μέσω μιας επαναδειγματοληψίας πέντε περιπτώσεων.

Η διευθέτηση του πειράματός μας χρησιμοποιεί διπλή διαδικασία επαναδειγματοληψίας (Cherkassky and Mulier, 2007). Το πρώτο επίπεδο επαναδειγματοληψίας χρησιμοποιείται για την εκτίμηση του σφάλματος ελέγχου μιας μεθόδου μάθησης και το δεύτερο επίπεδο είναι για τη ρύθμιση των παραμέτρων του μοντέλου (ή την επιλογή του μοντέλου). Κατά τη διάρκεια του σταδίου επιλογής του μοντέλου, οι πιθανές επιλογές των παραμέτρων συντονισμού είναι οι C και $\tilde{C}$ σε διάστημα $[10^{-1}, 10^{2}]$, γ σε διάστημα $[10^{-3}, 10^{1}]$ και σ σε διάστημα $[2^{-2}, 2^{2}]$. Δεδομένου ότι δεν υπάρχει οριστική ετικέτα για τις αποκομμένες παρατηρήσεις με $U_i < r$, τα σφάλματα ελέγχου παρουσιάζονται βάσει των δειγμάτων με σαφείς ετικέτες, δηλαδή ακριβείς παρατηρήσεις και αποκομμένες παρατηρήσεις με $U_i \geq r$. Επιπλέον, οι παράμετροι του μοντέλου επιλέγονται με βάση την απόδοση εκείνων των δειγμάτων με καλά ορισμένες ετικέτες.

1. Σύνολο Δεδομένων Veteran: Το σύνολο δεδομένων Veteran είναι από τη μελέτη διαχείρισης Veteran του καρκίνου του πνεύμονα η οποία είναι μια τυχαία δοκιμή δύο θεραπευτικών σχημάτων για τον καρκίνο του πνεύμονα. Στα veteran σύνολα δεδομένων υπάρχουν 137 περιπτώσεις (παρατηρήσεις) και κάθε περίπτωση έχει 10 χαρακτηριστικά. Λιγότερο από το 7% των περιπτώσεων αποκόπτονται. Μεταξύ των 9 αποκομμένων περιπτώσεων, η μια έχει τον παρατηρηθέντα χρόνο επιβίωσης μικρότερο από το χρόνο που μας ενδιαφέρει. Με άλλα λόγια, μόνο η μια περίπτωση συνδέεται με την αβέβαιη ετικέτα της κλάσης στο veteran σύνολο δεδομένων.
2. Σύνολο Δεδομένων για τον Πνεύμονα: Το σύνολο δεδομένων του πνεύμονα μελέτησε την επιβίωση και τις συνήθεις καθημερινές δραστηριότητες σε ασθενείς με προχωρημένο καρκίνο του πνεύμονα από την βόρεια κεντρική ομάδα θεραπείας του καρκίνου (NCCTG). Υπάρχουν 167 περιπτώσεις σε αυτό το σύνολο δεδομένων και κάθε περίπτωση έχει 8 χαρακτηριστικά. Περίπου το 28% των περιπτώσεων αποκόπτονται και 21 αποκομμένες περιπτώσεις συνδέονται με αβέβαιες ετικέτες κλάσεων.

Πίνακας 4.2.1: Περίληψη των συνόλων δεδομένων επιβίωσης και των αποτελεσμάτων του πειράματος.

Data set	Veteran	Lung	PBC	Stanford2
Size	137	167	258	157
Attributes	10	8	22	2
δ=0	9	47	147	55
Censored %	6.57	28.14	56.98	35.03
Uncertain cls	1	21	54	8
Cox	**23.4± 4.6**	43.3± 5.6	34.3± 7.1	51.9± 4.7
pSVM linear	27.2± 7.8	**38.3± 6.2**	26.2± 2.5	53.9± 7.4
pSVM rbf	32.0± 5.9	42.5± 8.0	**23.5± 5.2**	**34.3± 6.2**
LUPI	30.4± 4.5	38.3± 9.9	25.3± 10.6	42.4± 17.7

3. Σύνολο Δεδομένων PBC: Το σύνολο δεδομένων pbc είναι από την δοκιμή της Κλινικής Mayo στην πρωτοπαθή χολική κίρρωση (PBC) του ήπατος και πραγματοποιήθηκε μεταξύ του 1974 και του 1984. Το σύνολο δεδομένων pbc περιέχει 258 περιπτώσεις και κάθε περίπτωση έχει 22 χαρακτηριστικά. Περισσότερες από τις μισές περιπτώσεις αποκόπτονται και 54 αποκομμένες περιπτώσεις δεν έχουν τις οριστικές ετικέτες κλάσης.
4. Σύνολο Δεδομένων Stanford2: Το τέταρτο σύνολο δεδομένων είναι το stanford2 από τα δεδομένα μεταμόσχευσης καρδιάς του Stanford, το οποίο περιέχει 157 περιπτώσεις, η καθεμία με 2 χαρακτηριστικά. Περισσότερες από το 35% των περιπτώσεων αποκόπτονται και 8 από αυτές συνδέονται με αβέβαιες ετικέτες.

Οι περιγραφές των συνόλων δεδομένων συνοψίζονται στον Πίνακα 4.2.1. Η τέταρτη γραμμή δείχνει τις αναλογίες των αποκομμένων παρατηρήσεων στα σύνολα δεδομένων. Η πέμπτη σειρά δείχνει τον αριθμό των αποκομμένων παρατηρήσεων με $U_i < r$, όταν το r έχει οριστεί ως η διάμεσος των παρατηρούμενων χρόνων επιβίωσης. Ο Πίνακας 4.2.1 δείχνει επίσης τα σφάλματα ελέγχου από διαφορετικές μεθόδους που εφαρμόζονται στα τέσσερα σύνολα δεδομένων. Σημειώνεται ότι οι προσεγγίσεις που βασίζονται στην SVM μέθοδο επιτυγχάνουν το χαμηλότερο σφάλμα ελέγχου στα τρία από τα τέσσερα σύνολα δεδομένων. Από την άλλη πλευρά, το μοντέλο του Cox δίνει την καλύτερη απόδοση στα veteran σύνολα δεδομένων. Σε αυτά τα πειράματα, ο αριθμός των δειγμάτων εκπαίδευσης είναι σταθερός, οπότε δεν μπορούμε να κάνουμε οποιαδήποτε συμπεράσματα σχετικά με την επίδραση του μεγέθους του δείγματος στην επίδοση της μεθόδου. Ωστόσο, μπορούμε να εξάγουμε συμπεράσματα σχετικά με την εγγενή μη-γραμμικότητα σε μερικά από τα σύνολα δεδομένων. Για παράδειγμα, για το σύνολο δεδομένων stanford2, η μη γραμμική pSVM αποδίδει πολύ καλύτερα από τις άλλες μεθόδους χρησιμοποιώντας γραμμική παραμετροποίηση. Έτσι, μπορούμε να συμπεράνουμε ότι αυτό το σύνολο δεδομένων απαιτεί μη γραμμική μοντελοποίηση.

Αυτά τα αποτελέσματα δείχνουν την επίδραση της αποκοπής στην γενίκευση της επίδρασης. Για μικρό ποσοστό αποκοπής (όπως 6%) στα δεδομένα, το μοντέλο του

Cox δίνει το χαμηλότερο σφάλμα ελέγχου. Όμως οι μέθοδοι που βασίζονται στην SVM δείχνουν τα πλεονεκτήματά τους όταν η αναλογία αποκοπής αυξάνεται. Επιπλέον, σχετικά πλεονεκτήματα των προσεγγίσεων που βασίζονται στη μέθοδο SVM γίνονται αρκετά εμφανή για δεδομένα επιβίωσης υψηλότερων διαστάσεων.

Αυτά τα αποτελέσματα δείχνουν επίσης μεγάλη μεταβλητότητα στα εκτιμώμενα σφάλματα ελέγχου, λόγω της κατάτμησης των διαθέσιμων δεδομένων σε πέντε τμήματα (εκπαίδευσης, ελέγχου). Αυτή η μεταβλητότητα αντικατοπτρίζεται σε μεγάλες τυπικές αποκλίσεις των ποσοστών σφαλμάτων ελέγχου. Άμεσες συγκρίσεις δείχνουν ότι οι μέθοδοι που βασίζονται στην SVM δίνουν μικρότερο ή παρόμοιο σφάλμα ελέγχου σε καθένα τμήμα (εκπαίδευση, έλεγχος). Ένας άλλος λόγος για τη μεταβλητότητα των εκτιμήσεων του μοντέλου που βασίζεται στην SVM οφείλεται στην επιλογή του μοντέλου μέσω της αναδειγματοληψίας. Αξίζει να σημειωθεί ότι, οι τυπικές αποκλίσεις των ποσοστών σφάλματος για όλες τις μεθόδους που βασίζονται στην SVM και φαίνονται στον Πίνακα 4.2.1 είναι σταθερά υψηλότερες από τις τυπικές αποκλίσεις για το μοντέλο του Cox (το οποίο δεν έχει ρυθμιζόμενες παραμέτρους). Αυτό υπογραμμίζει τη σημαντικότητα των τεκμηριωμένων στρατηγικών επιλογής μοντέλου για τις μεθόδους που βασίζονται στην SVM, η οποία θα είναι το επίκεντρο των μελλοντικών εργασιών.

ΚΕΦΑΛΑΙΟ 5: ΣΥΜΠΕΡΑΣΜΑΤΑ

Αυτή η εργασία προτείνει προγνωστική μοντελοποίηση των υψηλών διαστάσεων δεδομένων επιβίωσης ως ένα δυαδικό πρόβλημα ταξινόμησης. Εφαρμόζουμε τη σύνθεση LUPI και την SVM με αβέβαιες ετικέτες κλάσης για να λύσουν το πρόβλημα. Και οι δύο μέθοδοι ενσωματώνουν την πληροφορία σχετικά με το χρόνο επιβίωσης για να εκτιμήσουν τον SVM ταξινομητή. Έχουμε απεικονίσει τα πλεονεκτήματα και τους περιορισμούς αυτών των προσεγγίσεων μοντελοποίησης με τη χρήση συνθετικών, αλλά και από την πραγματική ζωή, συνόλων δεδομένων.

Προηγμένες μέθοδοι που βασίζονται στην SVM φαίνονται πολύ αποτελεσματικές όταν η αναλογία των αποκομμένων δεδομένων εκπαίδευσης είναι μεγάλη, ή ο παρατηρούμενος χρόνος επιβίωσης δεν ακολουθεί τις κλασικές πιθανολογικές παραδοχές (Aalen et al., 2008, Cherkassky and Ma, 2002). Από την άλλη πλευρά, με λιγότερες αποκομμένες παρατηρήσεις η προσέγγιση του μοντέλου του Cox μπορεί να αποδώσει καλύτερα. Επιπλέον, η σχετική απόδοση των LUPI και pSVM εξαρτάται από την ουσιαστική γραμμικότητα/ μη γραμμικότητα των ίδιων των δεδομένων. Συγκεκριμένα, η ανώτερη απόδοση της pSVM με τον RBF πυρήνα για τα stanford2 δεδομένα δείχνουν μια εσωτερική μη γραμμικότητα του συνόλου δεδομένων.

Το ίδιο κόστος εσφαλμένης ταξινόμησης θεωρείται σε όλη αυτή την εργασία. Ωστόσο, ρεαλιστικές ιατρικές εφαρμογές χρησιμοποιούν άνισα έξοδα. Εμείς θα ενσωματώσουμε διαφορετικά κόστη εσφαλμένης ταξινόμησης στις προτεινόμενες συνθέσεις που βασίζονται στην SVM. Επιπλέον, η μεθοδολογία μας για την προγνωστική μοντελοποίηση των δεδομένων επιβίωσης μπορεί εύκολα να επεκταθεί και σε άλλες (μη ιατρικές) εφαρμογές, όπως στην πρόβλεψη επιχειρηματικής αποτυχίας (γνωστή και ως πτώχευση) ή στην πρόβλεψη της αποτυχίας ενός γάμου (γνωστό και ως διαζύγιο).

ΒΙΒΛΙΟΓΡΑΦΙΑ

Aalen, O. O. (1975). *Statistical Inference for a Family of Counting Processes*. PhD thesis, University of California, Berkeley.

Aalen, O., Borgan O., Gjessing H. (2008). *Survival and Event History Analysis: A Process Point of View*, series for Statistics for Biology and Health. Springer-Verlag, New York, 2008.

Andersen, P. K., Bentzon, M. W. and Klein, J. P. (1996), Estimating the survival function in the proportional hazards regression model: A study of the small sample size properties. Scandinavian Journal of Statistics, **23**, 1–12.

Andersen, P. K. and Gill, R. D. (1982), Cox's regression model for counting processes: A large sample study, *Annals of Statistics* **10**, 1100–1120.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford.

Borgan, O. and Liestol, K. (1990), A note on confidence-intervals and bands for the survival function based on transformations, *Scandinavian Journal of Statistics* **17**, 35–41.

Breslow, N. E. (1972), Discussion of Professor Cox's paper, *Journal of the Royal Statistical Society - Series B* **34**, 216–217.

Breslow, N. E. (1974). Covariance analysis of censored survival data, *Biometrics* **30**, 89–99.

Cherkassky, V. and Ma, Y. (2002). SVM-based Learning for Multiple Model Estimation. Submitted to IEEE Transaction on Neural Networks.

Cherkassky, V. and Mulier, F. (2007). *Learning from data: concepts, theory, and methods*. September 2007, Wiley-IEEE Press.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society - Series B*, Vol. 34, pp. 187–220.

Devijver, P.A., and Kittler J. (1982). *Pattern Recognition: A Statistical Approach*. Prentice-Hall, Inc., Englewood Cliffs, N. J.

Dietterich, T.G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, Vol. 7(10), pp. 1895-1924.

Duda, R. O. and Hart P. E. (1973). *Pattern Classification and Scene Analysis*. John Wiley & Sons, N. Y.

Drosou, K. (2013). *Statistical Methods for the Analysis of High Dimensional Data*. Graduate thesis. National Technical University of Athens.

Drosou, K. (2015). Class Imbalanced Problem With Support Vector Machines. Post-Graduate thesis. National Technical University of Athens.

Drosou, K., Georgiou, S., Koukouvinos, C., Stylianou, S. (2014). Support Vector Machines classification on class imbalanced data: a case study with real medical data, *Journal of Data Science,* Vol. 12, 727-754.

Drosou K. and Koukouvinos, C. (2017). Proximal support vector machine techniques on medical prediction outcome. *Journal of Applied Statistics,* Vol. 44, Issue 3, p.p. 533-553.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy R. (1996). *Advances in Knowledge Discovery and Data Mining*. AAAI/MIT Press.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, Vol. 7, pp. 179-188.

Friedman J. H. (1989). Regularized discriminant analysis. *Journal of the American Statistical Association*, Vol. 84(405), pp. 165-175.

Fu., K.-S. (1982). Syntactic Pattern Recognition and Applications. Englewood Cliffs, pp 76-90.

Fukunaga, K. (1972). *Introduction to Statistical Pattern Recognition*. Academic Press, Inc., Orlando, FL.

Glick, N. (1978). Additive estimators for probabilities of correct classification. *Pattern Recognition*, Vol. 10, pp 211-222.

Greenwood, M. (1926). The natural duration of cancer, in *Reports on Public Health and Medical Subjects* **33**, London: His Majesty's Stationery Office, pp. 1–26.

Halley, E. (1693). An estimate of the degrees of the mortality of mankind, drawn from curious tables of the births and funerals at the city of Breslaw with an attempt to ascertain the price of annuities upon lives. *Philosophical Transactions*, Vol. 17, pp. 596–610.

Holmström, L., Koistinen P., Laaksonen J., and Oja E. (1997). Neural and statistical classifiers – taxonomy and two case studies. *IEEE Transactions on Neural Networks*, Vol. 8(1), pp. 5-17.

James, G., Witten, D., Hastie, T., Tibshirani R. (2013). *An Introduction to Statistical Learning with Applications in R*. Springer Texts in Statistics 103, Series Editors: G. Casella, S. Fienberg, I. Olkin,Springer-Verlag New York, Heidelberg Dordrecht London.

Kaplan, E. and Meier, P. (1958). Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association*, Vol. 43, pp. 457–481.

Klein J. P., Houwelingen, H. C., Ibrahim, J. G., Scheike T. H. (2014). *Handbook of Survival Analysis*. 1st edition, New York, NY: Chapman & Hall/CRC Press, Taylor & Francis Group.

Klein, J. P. and Moeschberger, M. L. (2003). *Survival Analysis: Techniques for Censored and Truncated Data, Statistics for Biology and Health*, 2nd ed., Springer, New York.

Kittler, J., Hatef, M., Duin, R. P. W., Matas, J. (1998). On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 20(3), pp. 226-239.

Kuncheva, I. L., (2014). *Combining Pattern Classifiers: Methods and Algorithms.* 2nd Edition, John Wiley & Sons, Inc., New Jersey.

Liang L., Cai F., and Cherkassky V. (2009). Predictive learning with structured (grouped) data. *Neural Networks*, Vol. 22, no. 5-6, pp. 766–773.

Lippmann, R. P. (1991). A critical overview of neural network pattern classifiers. *In Proc. IEEE Workshop on Neural Networks for Signal Processing*, pp. 266-275.

Looney, C. G. *Pattern Recognition Using Neural Networks: theory and algorithms for engineers and scientists*. Oxford University Press, Oxford, Inc. New York, NY, USA 1997.

Nelson, W. (1969). Hazard plotting for incomplete failure data. *Journal of Quality Technology* **1**, 27–52.

Niaf, E., Flamary, R., Lartizien, C., and Canu, S. (2011). Handling uncertainties in SVM classification. *In Statistical Signal Processing Workshop (SSP)*, 2011 IEEE, pp. 757–760.

Parpoula, C., Drosou, K., Koukouvinos, C. (2013). Large-Scale Statistical Modelling via Machine Learning Classifiers, *Journal of Statistics Applications & Probability* 2, No. 3, 1-20.

Patrick, E. A. (1972). *Fundamentals of Pattern Recognition*. Prentice-Hall, Inc., Engelwood Cliffs, N. J.

Piatetsky-Shapiro, G., Frawley, W. J. (1991). *Knowledge Discovery in Databases*. AAAI/MIT Press.

Platt, J. C. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *In Advances in Large Margin Classifiers*. MIT Press, pp. 61–74.

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge.

Shiao, H.-T. and Cherkassky, V.(2013). SVM-based approaches for predictive modeling of survival data. *The 2013 International Conference on Data Mining.*

Silberschatz, A., Stonebraker, M., Ullman, J. D. (1995). Database research: Achievements and opportunities into the 21st century. In *Report of an NSF Workshop on the Future of Database Systems Research*, May 1995.

Therneau, T. M. (2013). A Package for Survival Analysis in R, r package version 2.37-4. Available: http://CRAN.R-project.org/package=survival.

Tou, J. T. and Gonzalez R. C. (1974). *Pattern Recognition Principles*. Addison-Wesley, Reading, MA.

Toussaint, G. T. (1974). Bibliography on estimation of misclassification. IEEE *Transactions on Information Theory*, Vol. 20, pp. 472-479.

Traven, H. G. C. (1991). A neural network approach to statistical pattern classification by "semiparametric" estimation of probability density functions. *IEEE Transaction on Neural Networks*, Vol. 2(3), pp. 366-377.

Tsiatis, A. A. (1981). A large sample study of Cox's regression model. *Annals of Statistics* **9**, 93–108.

van de Vijver, M. J., He, Y. D., van 't Veer, L. J., Dai, H., Hart, A. A. M., Voskuil, D. W., Schreiber, G. J., Peterse, J. L., Roberts, C., Marton, M. J., Parrish, M., Atsma, D., Witteveen, A., Glas, A., Delahaye, L., van der Velde, T., Bartelink, H., Rodenhuis, S., Rutgers, E. T., Friend, S. H. and Bernards, R. (2002). A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine* **347**, 1999–2009.

van Houwelingen, H. C., Bruinsma, T., Hart, A. A. M., van't Veer, L. J. and Wessels, L. F. A. (2006). Cross-validated Cox regression on microarray gene expression data. *Statistics in Medicine* **25**, 3201–3216.

van Houwelingen, H. C. and Putter, H. (2012). *Dynamic Prediction in Clinical Survival Analysis, Monographs on statistics and applied probability*, CRC Press, Boca Raton.

van't Veer, L. J., Dai, H. Y., van de Vijver, M. J., He, Y. D. D., Hart, A. A. M., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R. and Friend, S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530–536.

Vapnik, V. N. (2006). *Estimation of dependences based on empirical data*. Empirical inference science: afterword of 2006. Springer.

Vapnik V. and Vashist, A. (2009). Special issue: A new learning paradigm: Learning using privileged information. *Neural Networks*, Vol. 22, no. 5-6, pp. 544–557.

Zhou M., "Use software R to do survival analysis and simulation. a tutorial," http://www.ms.uky.edu/ mai/Rsurv.pdf.

Καρώνη - Ρίτσαρντσον, Χ. (2009). *Μοντέλα αξιοποίησης και επιβίωσης*. Έτος έκδοσης: 2009. Συμεών.