ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ
ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

Διπλωματική Εργασία

ΜΗΧΑΝΕΣ ΔΙΑΝΥΣΜΑΤΩΝ ΥΠΟΣΤΗΡΙΞΗΣ ΚΑΙ ψ-ΜΑΘΗΣΗ ΓΙΑ ΤΗΝ ΤΑΞΙΝΟΜΗΣΗ ΔΕΔΟΜΕΝΩΝ

ΚΩΝΣΤΑΝΤΗ Θ. ΠΑΝΑΓΙΩΤΑ

Επιβλέπων: Χρήστος Κουκουβίνος
Καθηγητής ΕΜΠ

Αθήνα, 2016

NATIONAL TECHNICAL UNIVERSITY OF ATHENS

School of Applied Mathematics and Physical Science

SUPPORT VECTOR MACHINES AND ψ-LEARNING FOR DATA CLASSIFICATION

by
CONSTANTI T. PANAYIOTA

Athens, Greece 2016

Περίληψη

Στην επιστήμη και την τεχνολογία η ταξινόμηση γίνεται ολοένα και πιο σημαντική ως εργαλείο για την εξαγωγή πληροφοριών. Διάφορες μέθοδοι κατά καιρούς έχουν προταθεί για την βελτίωση της ταξινόμησης και την απόδοση καλύτερων αποτελεσμάτων. Η παρούσα διπλωματική επικεντρώνεται σε δυο τεχνικές ταξινόμησης περιθωρίου-βάσης, τη δημοφιλή SVM (Μηχανή Διανυσμάτων Υποστήριξης) καθώς και μια νέα τεχνική τη λεγόμενη ψ-Learning. Παρά το γεγονός ότι και οι δυο τεχνικές μοιάζουν εκ τον υστέρων θα δούμε ότι η ψ-μάθηση πλεονεκτεί έναντι της SVM τόσο σε αριθμητικό όσο και σε θεωρητικό επίπεδο.

Στο πρώτο κεφάλαιο γίνεται εισαγωγή στην έννοια της μηχανικής μάθησης, στον τρόπο λειτουργίας της καθώς επίσης και στις τεχνικές ταξινόμησης που έχουν αναπτυχθεί παράλληλα στην στατιστική και την μηχανική μάθηση. Επίσης γίνεται μια εκτενέστερη αναφορά στους ταξινομητές μέγιστου περιθωρίου.

Στο δεύτερο κεφάλαιο αναπτύσσουμε την ιδέα των μηχανών διανυσμάτων υποστήριξης (SVMs) για την επίλυση των προβλημάτων διαχωρισμού των δεδομένων με μεγάλο περιθώριο σε δυαδικά προβλήματα ταξινόμησης καθώς επίσης και σε πράξεις στο επίπεδο του πυρήνα. Στην συνέχεια επεκτείνουμε την δυαδική ταξινόμηση στην περίπτωση πολλαπλών κατηγοριών με την χρήση δυο πολύ σημαντικών συστημάτων ταξινόμησης.

Στο τρίτο κεφάλαιο παρουσιάζουμε την τεχνική μάθησης ψ-μάθηση και δείχνουμε ότι σε μη διαχωρίσιμες περιπτώσεις αποτελεί λύση στα προβλήματα που αντιμετωπίζει η αντίστοιχη SVM. Αναπτύσσουμε την ιδέα της, η οποία βασίζεται σε ένα μη κυρτό πρόβλημα ελαχιστοποίησης και προτείνουμε μεθόδους επίλυσης της ελαχιστοποίησης αυτής. Επίσης αναφερόμαστε στο θεωρητικό της υπόβαθρο το οποίο της παρέχει ισχυρά πλεονεκτήματα έναντι των άλλων μεθόδων ταξινόμησης. Τέλος επεκτείνουμε την δυαδική ταξινόμηση σε ταξινόμηση πολλαπλών κατηγοριών και βλέπουμε ότι η ψ-learning διατηρεί την ερμηνεία των περιθωρίων και τις ιδιότητες της αντίστοιχης δυαδικής ταξινόμησης.

Στο τέταρτο κεφάλαιο αρχικά παρουσιάζουμε τα μέτρα αξιολόγησης ενός μοντέλου. Στην συνέχεια γίνονται εφαρμογές σε πραγματικά δεδομένα τα οποία πάρθηκαν από το UCI Machine Learning Repository. Αρχικά γίνεται μια εφαρμογή

στην R σε δεδομένα στα οποία έχουμε ένα πρόβλημα δυαδικής ταξινόμησης (Parkinsons Data Set). Εφαρμόζουμε την ταξινόμηση με την χρήση των μηχανών διανυσμάτων υποστήριξης για τους τέσσερις διαφορετικούς πυρήνες και κάνουμε συγκρίσεις των αποτελεσμάτων αυτών σύμφωνα με τα κριτήρια απόδοσης. Τέλος γίνονται ακόμη τρεις εφαρμογές σε πραγματικά δεδομένα για την σύγκριση των αποτελεσμάτων της τεχνικής SVM και ψ-μάθησης

Στο πέμπτο και τελευταίο κεφάλαιο γίνεται ένας επίλογος που παραθέτουμε τα γενικά μας συμπεράσματα.

Abstract

Classification is increasingly becoming an important tool for extracting information in both science and technology. Various methods have been proposed from time to time to improve the classification and yield better results. This thesis focuses on two Margin-based classification techniques, the popular Support Vector Machine (SVM) and a new technique called ψ-Learning. Although, the two techniques seem to be similar at first, it is shown afterwards that ψ-learning outweighs SVM not only in numerical but also in theoretical level.

The first chapter is the introduction to the concept of machine learning, its operations as well as the classification techniques that have been developed alongside in statistics and engineering learning. There is also a more extensive reference to the maximum margin classifiers.

In the second chapter the idea of SVMs is developed for solving problems of data separation by a large margin in binary classification problems as well as in operations to the kernel level. Then the binary classification is expanded in the case of multiple categories using two very important classification systems.

In the third chapter the ψ-Learning technique is presented and it is shown that in non-separable cases this method is the solution to problems the corresponding SVM faces. Its concept, which is based on a non-convex minimization problem, is developed and we suggest methods to solve this minimization. There is also a reference to its theoretical background which gives this technique powerful advantages over other classification methods. Finally, the binary classification is expanded into a multicategory classification and it is clearly seen that ψ-learning maintains the interpretation of the margins and the qualities of the corresponding binary classification.

In the fourth chapter we initially present the evaluation measures of a model. Then applications are contacted to real data that have been obtained from the UCI Machine Learning Repository. Originally data with a binary classification problem (Parkinsons Data Set) is applied to R. The classification is applied with the use of support vector machines for four different kernels and we compare these results according to the performance criteria. Finally, there are three more applications to real data which can enhance the comparison of the results between the technical SVM and ψ-learning.

In the fifth and final chapter the overall conclusions are presented.

Ευχαριστίες

Πρωτίστως θα ήθελα να ευχαριστήσω θερμά τον Καθηγητή του Εθνικού Μετσόβιου Πολυτεχνείου κ. Χρήστο Κουκουβίνο για την ανάθεση της παρούσας διπλωματικής καθώς επίσης και για την πολύτιμη καθοδήγηση του και για όλη την προσφορά του καθ' όλη την διάρκεια των σπουδών μου.

Τις θερμές μου ευχαριστίες θα ήθελα να εκφράσω στην υποψήφια διδάκτορα Κρυσταλλένια Δρόσου, για την πολύτιμη βοήθεια της και τη συνεχή υποστήριξη της κατά τη διάρκεια εκπόνησης της διπλωματικής μου εργασίας.

Επίσης θέλω να εκφράσω τις βαθύτατες ευχαριστίες μου στην οικογένεια μου, στον πατέρα μου Θεόδωρο και στην μητέρα μου Αντωνία καθώς επίσης και στις δυο μικρές μου αδελφές Χαρίκλεια και Κωνσταντίνα για την υποστήριξη τους καθ' όλη τη διάρκεια των σπουδών μου, και την αμέριστη συμπαράσταση που μου πρόσφερα όλα αυτά τα χρόνια.

Τέλος θα ήθελα να ευχαριστήσω του φίλους μου αλλά κυρίως του συμφοιτητές μου για αυτά τα υπέροχα πέντε χρόνια που περάσαμε μαζί.

Παναγιώτα Κωνσταντή

Εθνικό Μετσόβιο Πολυτεχνείο,

Σχολή Εφαρμοσμένων Μαθηματικών

και Φυσικών Επιστημών

Αθήνα, 2016

Περιεχόμενα

Περιεχόμενα Σχημάτων

Περιεχόμενα Πινάκων

Κεφάλαιο 1

Μηχανική Μάθηση και Ταξινόμηση

(Machine Learning and Classification)

1.1 Μηχανική Μάθηση (Machine Learning)

Η μηχανική μάθηση αποτελεί περιοχή της τεχνητής νοημοσύνης, η οποία γνωρίζει σημαντικότατη ανάπτυξη τις τελευταίες δεκαετίες, με εφαρμογές που κυμαίνονται από τον χώρο της ανάκτησης πληροφορίας μέχρι τον χώρο των πολυμέσων. Ένας ορισμός που θα μπορούσαμε να δώσουμε στον όρο μηχανική μάθηση είναι ότι αποτελεί την δημιουργία μοντέλων ή προτύπων από ένα σύνολο δεδομένων, από ένα υπολογιστικό σύστημα. Θα μπορούσε κανείς να πει ότι η μηχανική μάθηση επικαλύπτεται σημαντικά από τη στατιστική, αφού και τα δύο πεδία μελετούν την ανάλυση δεδομένων. Μερικές τεχνικές προέρχονται από τις δεξιότητες που διδάσκονται στα βασικά μαθήματα στατιστικής και άλλες είναι στενά συνδεδεμένες με το είδος της μηχανικής μάθησης που έχει προκύψει από την επιστήμη των υπολογιστών. Κοινές μέθοδοι όπως τα δέντρα αποφάσεων

(decision trees) καθώς και η μέθοδος του πλησιέστερου – γείτονα (nearest-neighbor) για την ταξινόμηση, έχουν αναπτυχθεί παράλληλα στη στατιστική και τη μηχανική μάθηση με απώτερο σκοπό τη βελτίωση των επιδόσεων της ταξινόμησης καθώς και για να καταστεί η διαδικασία πιο αποτελεσματική υπολογιστικά.

Γενικά έχουν αναπτυχθεί πολλές τεχνικές μηχανικής μάθησης που χρησιμοποιούνται ανάλογα με τη φύση του προβλήματος και εμπίπτουν σε ένα από τα παρακάτω δυο είδη:

- Μάθηση με επίβλεψη ή εποπτευόμενη μάθηση (supervised learning) ή μάθηση με παραδείγματα (learning from examples). Στην μάθηση με επίβλεψη μπορούμε να γνωρίζουμε με βεβαιότητα ότι υπάρχουν τόσες πολλές κλάσεις, και ο στόχος μας είναι να καθιερώσουμε έναν κανόνα σύμφωνα με τον οποίο να μπορούμε να ταξινομήσουμε μια νέα παρατήρηση σε μια από τις είδη υπάρχουσες κλάσεις. Πιο συγκεκριμένα τα δεδομένα εκπαίδευσης (training set) συνοδεύονται από ετικέτες για την κλάση στην οποία ανήκει το καθένα, δηλαδή η διαδικασία μάθησης οδηγείται από την παρουσία των αποτελεσμάτων της κλάσης. Σε αυτή την κατηγορία ανήκουν τα δέντρα αποφάσεων (decision trees), η λογιστική παλινδρόμηση (logistic regression), τα νευρωνικά δίκτυα (neural networks) καθώς και οι μηχανές διανυσμάτων υποστήριξης (SVM) και ψ-μάθηση τις οποίες θα αναφέρουμε εκτενέστερα σε επόμενα κεφάλαια.

- Μάθηση χωρίς επίβλεψη ή μη εποπτευόμενη μάθηση (unsupervised learning) ή μάθηση από παρατήρηση (learning from observation). Στη μάθηση χωρίς επίβλεψη μπορούμε να δώσουμε ένα σύνολο παρατηρήσεων με σκοπό να αποδείξουμε την ύπαρξη των κλάσεων ή ομάδων στα δεδομένα. Πιο συγκεκριμένα δεν είναι γνωστό σε ποια κλάση ανήκουν τα δεδομένα εκπαίδευσης, δηλαδή γνωρίζουμε μόνο τις τιμές των χαρακτηριστικών και όχι την τιμή του αποτελέσματος.

Στην πράξη συναντάμε περισσότερο προβλήματα που ανήκουν στην κατηγορία της μάθηση με επίβλεψη όπου υπάρχει ένα σύνολο δεδομένων εκπαίδευσης (training set) στο οποίο γνωρίζουμε την τιμή του αποτελέσματος και τις τιμές των χαρακτηριστικών που μας ενδιαφέρουν, και προσπαθούμε με βάση αυτά τα δεδομένα να κατασκευάσουμε ένα μοντέλο πρόβλεψης. Αυτό το μοντέλο στη συνέχεια θα το χρησιμοποιήσουμε για να προβλέψουμε το αποτέλεσμα νέων συνόλων δεδομένων εξέτασης, στα οποία είναι γνωστές οι τιμές των χαρακτηριστικών αλλά δεν είναι γνωστή η τιμή της τάξης (κλάσης/κατηγορίας).

Το σύστημα πρέπει να "μάθει" επαγωγικά μια συνάρτηση που ονομάζεται συνάρτηση στόχος (target function) και αποτελεί έκφραση του μοντέλου που περιγράφει τα δεδομένα. Η συνάρτηση στόχος χρησιμοποιείται για την πρόβλεψη της τιμής μιας μεταβλητής, που ονομάζεται εξαρτημένη μεταβλητή (predictors), βάσει των τιμών ενός συνόλου μεταβλητών, που ονομάζονται ανεξάρτητες μεταβλητές.

Στην μάθηση με επίβλεψη διακρίνονται δυο είδη προβλημάτων (learning tasks), τα προβλήματα ταξινόμησης και τα προβλήματα παλινδρόμησης.

- Ταξινόμηση(classification): αφορά στη δημιουργία μοντέλων πρόβλεψης διακριτών τάξεων (κλάσεων/κατηγοριών) (π.χ. ομάδα αίματος).

- Παλινδρόμηση (regression): αφορά στη δημιουργία μοντέλων πρόβλεψης αριθμητικών τιμών (π.χ. πρόβλεψη ισοτιμίας νομισμάτων ή τιμής μετοχής).

Μέσω της μηχανικής μάθησης μπορούμε να χρησιμοποιήσουμε τους υπολογιστές για να ανακαλύψουμε και να περιγράψουμε τα μοτίβα που βασίζονται στις συμπεριφορές των διάφορων φαινομένων που παρατηρούμε γύρω μας. Ο ευκολότερος τρόπος για να περιγράψουμε τα φαινόμενα αυτά είναι μέσω της ταξινόμησης.

Συνήθως όμως οι ταξινομήσεις δεν πραγματοποιούνται τόσο απλά και σε γενικές γραμμές δεν έχουμε εύκολη πρόσβαση στις διαδικασίες όπου προστίθεται ετικέτα στα αντικείμενα. Αντιθέτως μπορούμε να παρατηρήσουμε μόνο τις συνέπειες αυτών των διαδικασιών δηλαδή μόνο τις παρατηρήσιμες ετικέτες για κάθε αντικείμενο. Ο στόχος της μηχανικής μάθησης είναι να υπολογίσει ένα κατάλληλο μοντέλο για μια διαδικασία σήμανσης (labelling process) που προσεγγίζει την αρχική διαδικασία όσο το δυνατόν περισσότερο.

Για να γίνει αυτή η διαδικασία πιο αντιληπτή δίνουμε ένα ορισμό της μηχανικής μάθησης:

Έστω X είναι το σύνολο των δεδομένων και S⊂X είναι το σύνολο του δείγματος, με

$$f: X \to \{-1,1\}$$

να αποτελεί την συνάρτηση στόχος (διαδικασία σήμανσης) και,

$$D = \{(x, y) |\ x \in S \ και \ y = f(x)\}$$

το σύνολο εκπαίδευσης D.

Τότε υπολογίζεται μια συνάρτηση

$$\hat{f}: X \to \{-1,1\}$$

χρησιμοποιώντας το σύνολο εκπαίδευσης D τέτοια ώστε

$$\hat{f}(x) \cong f(x)$$

για κάθε $x \in X$.

Εδώ το σύνολο των δεδομένων είναι το σύνολο που μας ενδιαφέρει. Το σύνολο του δείγματος S είναι αναγκαίο, δεδομένου ότι οι περισσότερες συλλογές αντικειμένων που μας ενδιαφέρουν τείνουν να είναι πολύ μεγαλύτερες ή ίσως άπειρες, και η οικοδόμηση των μοντέλων μπορεί να είναι πολύ αργή για μεγάλα σύνολα δεδομένων και αδύνατη για άπειρα σύνολα δεδομένων. Ως εκ τούτου, το σύνολο του δείγματος S ενεργεί ως εκπρόσωπος του συνόλου δεδομένων, προκειμένου να καταστεί εφικτή η διαδικασία οικοδόμησης των μοντέλων. Η συνάρτηση στόχος (target function) f είναι η διαδικασία που παρέχει τις παρατηρήσιμες ετικέτες (observable labels). Υποτίθεται ότι η f είναι σε θέση να παρέχει μια κατάλληλη τιμή {-1,1} για κάθε στοιχείο του X όταν το στοιχείο είναι παρατηρούμενο. Έτσι, ακόμη κι αν δεν έχουμε άμεση πρόσβαση στην ίδια τη διαδικασία, είμαστε πάντα σε θέση να παρατηρήσουμε τις ετικέτες που η διαδικασία εκχωρεί στα στοιχεία του συνόλου δεδομένων. Χρησιμοποιούμε αυτήν την ιδιότητα της συνάρτησης στόχος για να κατασκευάσουμε το σύνολο εκπαίδευσης D παρατηρώντας τις ετικέτες των αντικειμένων στο σύνολο του δείγματος S. Το είδος της μηχανική μάθηση που κάνει χρήση των επισημασμένων δεδομένων εκπαίδευσης (labelled training data) όπως έχουμε αναφέρει και πιο πάνω είναι η επιβλεπόμενη μάθηση. Τέλος, η μάθηση μπορεί να θεωρηθεί ως ο υπολογισμός της προσέγγιση της συνάρτησης $\hat{f}$ ή ένα μοντέλο της αρχικής διαδικασίας f με βάση τα παραδείγματα εκπαίδευσης στο D. Δηλαδή, το αποτέλεσμα της μηχανικής μάθησης είναι ένα μοντέλο της αρχικής συνάρτησης επισήμανσης. Ωστόσο, για ευκολία συνηθίζουμε να λέμε ότι η $\hat{f}$ είναι ένα μοντέλο των δεδομένων εκπαίδευσης D. Αυτό είναι συμβατό με την επίσημη άποψη, αφού τα στοιχεία στο σύνολο εκπαίδευσης είναι ζεύγη εισόδου-εξόδου της αρχικής συνάρτησης στόχου:

$$\{(x, y) \text{ με } x \in S \text{ και } y = f(x)\}$$

και αυτό σημαίνει ότι η μοντελοποίηση της συνάρτησης f και η μοντελοποίηση των δεδομένων εκπαίδευσης D είναι ένα και το αυτό.

Θα μπορούσαμε επίσης να εξετάσουμε τα προβλήματα ταξινόμησης με περισσότερες από δύο κλάσεις. Η μόνη διαφορά από την παραπάνω διαδικασία θα ήταν ότι το πεδίο τιμών της αρχικής συνάρτησης στόχου f και το μοντέλο της

$\hat{f}$ θα ήταν ένα σύνολο που περιλαμβάνει ένα κατάλληλο αριθμό διαφορετικών ετικετών. Περεταίρω αναφορά για προβλήματα ταξινόμησης πολλαπλών κατηγοριών θα γίνει σε επόμενη παράγραφο.

Τέλος, με την απόκτηση ενός μοντέλου της αρχικής διαδικασίας επισήμανσης, μπορούν να επιτευχθούν δύο ενδιαφέροντα πράγματα. Πρώτον, μπορούμε να χρησιμοποιήσουμε το μοντέλο για να υπολογίσουμε ή να προβλέψουμε την ετικέτα ενός στοιχείου στο σύνολο δεδομένων Χ χωρίς να χρειάζεται να παρατηρήσουμε αυτό το στοιχείο. Δεύτερον, το μοντέλο μπορεί να παρέχει κάποια στοιχεία σχετικά με την αρχική διαδικασία σήμανσης, δηλαδή, ένα μοντέλο κατέχει κάποιες επεξηγηματικές ικανότητες.

1.2 Ταξινόμηση (Classification)

Στην στατιστική και στην μηχανική μάθηση, η ταξινόμηση (classification) είναι μια απαραίτητη διαδικασία ώστε να αναγνωρίσουμε σε ποιο σετ κατηγοριών (κλάσεων) ανήκει μια νέα παρατήρηση, με βάση ένα σύνολο δεδομένων εκπαίδευσης που περιέχει τις παρατηρήσεις. Οι επιμέρους παρατηρήσεις αναλύονται σε ένα σύνολο μετρήσιμων ιδιοτήτων γνωστά ως χαρακτηριστικά (features). Αυτές οι παρατηρήσεις μπορεί να είναι κατηγορικές, αριθμητικές, ακέραιες ή πραγματικές. Ένας αλγόριθμος που υλοποιεί την ταξινόμηση, ειδικά σε μια συγκεκριμένη εφαρμογή, είναι γνωστός ως ένα ταξινομητής (classifier). Ο όρος “ταξινομητής” μερικές φορές αναφέρεται επίσης στην μαθηματική συνάρτηση, που υλοποιείται από έναν αλγόριθμο ταξινόμησης, η οποία αντιστοιχίζει τα δεδομένα εισόδου σε μια κατηγορία.

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

Στην επιστήμη και την τεχνολογία, η ταξινόμηση γίνεται ολοένα και πιο σημαντική ως εργαλείο για την εξαγωγή πληροφοριών. Σε επιβλεπόμενη μάθηση, υπάρχουν τέσσερις βασικές συνιστώσες της στατιστικής ταξινόμησης: ένας χώρος εισόδου S, ένας χώρος εξόδου X, μια συνάρτηση απόφασης f, και ένα σύνολο εκπαίδευσης. Το σύνολο εκπαίδευσης συνήθως συμβολίζεται με

$$D = \{(x_1, y_1), \dots, (x_n, y_n)\} \subseteq (S \times X)^n$$

όπου n είναι ορισμένα υποδείγματα. Το σύνολο n ζευγών εισόδου / εξόδου $(x_i, y_i)_{i=1}^n$ αποτελεί το δείγμα όπου τα x_i εισόδου τα βλέπουμε ως εξαρτημένες μεταβλητές (predictors) και την έκβαση y_i ως την ετικέτα του δείγματος που υποδεικνύει σε ποια κλάση ανήκει. Χρησιμοποιώντας αυτά τα δεδομένα χτίζουμε ένα μοντέλο πρόβλεψης, το οποίο μας επιτρέπει να προβλέψουμε την τάξη για τα νέα «κρυμμένα» αντικείμενα με δεδομένο x∈S εισόδου. Τυπικά, η στατιστική ταξινόμηση γίνεται με την κατασκευή του διανύσματος συνάρτησης απόφασης $\boldsymbol{f} = (f_1, f_2, \dots, f_k)$, με f_j να εκπροσωπεί την κατηγορία j, χαρτογραφώντας από το σύνολο $\mathrm{S} \subset \mathrm{R}^{\mathrm{d}}$ έως R^1, δηλαδή $f : \mathrm{S} \subset \mathrm{R}^{\mathrm{d}} \to \mathrm{R}^1$. Ένας ταξινομητής $argmax_{j=1,\dots,k} = f_j(x)$, που επάγεται από την f, χρησιμοποιείται για να οριστεί μια τιμή της ετικέτας σε κάθε διάνυσμα εισόδου x∈S. Με άλλα λόγια, το x αντιστοιχίζεται σε μια κατηγορία με την υψηλότερη τιμή του $f_j(x)$, j=1,..,k, με $f_j(x)$, να δείχνει ότι το x ανήκει στην κατηγορία j. Ένας ταξινομητής εκπαιδεύεται μέσω ενός δείγματος εκπαίδευσης $\{(x_i, y_i)$: i=1,...,n} ανεξάρτητων ταυτόσημων κατανεμημένων σύμφωνα με κάποια άγνωστη από κοινού κατανομή P(x,y). Σε ένα πρόβλημα k-τάξης, ένας ταξινομητής διαχωρίζει το χώρο εισόδου S σε k ξένα υποσύνολα, $S_1, \dots, S_k$ τέτοια ώστε για ένα δείγμα με είσοδο $x \in S_j$, η προβλεπόμενη τάξη να είναι j. Μια καλή ταξινόμηση είναι αυτή που προβλέπει με ακρίβεια την κατηγορία y για δεδομένο x. Η απόδοση του ταξινομητή μπορεί να αξιολογηθεί χρησιμοποιώντας ένα ανεξάρτητο δείγμα, το οποίο ονομάζεται δείγμα δοκιμής ή δείγμα επικύρωσης (validation sample).

Η ταξινόμηση μπορεί να εφαρμοστεί σε πολλά διαφορετικά πεδία όπως:

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

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

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

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

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

1.2.1 Στατιστικές Μέθοδοι (Statistical Methods)

Υπάρχουν πολλές τεχνικές ταξινόμησης, όπως η μέθοδος ελαχίστων τετραγώνων (Least Squares Method), η γραμμική διακριτική ανάλυση (Linear Discriminant Analysis) (LDA), η τετραγωνική διακριτική ανάλυση (Quadratic Discriminant Analysis) (QDA), διαχωρίζοντα υπερεπίπεδα (Separating hyperplanes), Bagging, και η ανάλυση συστάδων (Cluster Analysis). Θα εξετάσουμε εν συντομία κάποιες γνωστές τεχνικές ταξινόμησης και στη συνέχεια θα επικεντρωθούμε στη μηχανή διανυσμάτων υποστήριξης και ψ-μάθηση.

1.2.1.1 Ταξινομητές Πλησιέστερου Γείτονα (Nearest-Neighbor Classifier)

Η μέθοδος πλησιέστερου γείτονα χρησιμοποιεί μια μετρική για απόσταση, όπως είναι η Ευκλείδεια απόσταση. Ο κανόνας k-NN, του Fix και Hodges (1951), που εξαρτάται από μια μόνο ρυθμιστική παράμετρο (tuning parameter), δηλαδή το μέγεθος του γείτονα k, ταξινομεί το σύνολο δοκιμών με βάση την εκπαίδευση και ορίζεται ως εξής: Για κάθε σημείο στο σύνολο δοκιμών, βρίσκει τα k πλησιέστερα σημεία στο δείγμα εκπαίδευσης και στην συνέχεια με πλειοψηφία προβλέπει την τάξη. Ο αριθμός των γειτόνων, k, επιλέγεται συνήθως από διασταυρωμένη επικύρωση (cross-validation) στο σύνολο εκπαίδευσης. Ένας αριθμός τιμών του

k μπορεί να «δοκιμάσει» και τον αριθμό με το μικρότερο βαθμό σφάλματος διασταυρωμένης επικύρωσης στο σύνολο δοκιμής που έχει επιλεγεί.

1.2.1.2 Δέντρα Ταξινόμησης (Classification Trees)

Τα δέντρα ταξινόμησης είναι μια δημοφιλής μέθοδος ταξινόμησης δέντρου-βάσης. Η ταξινόμηση δέντρου-βάσης χωρίζει το χώρο των χαρακτηριστικών (feature space) σε ένα σύνολο ορθογωνίων, και στη συνέχεια ταιριάζει το κάθε ένα σε ένα απλό μοντέλο. Τα δυαδικά δέντρα ταξινόμησης είναι κατασκευασμένα από επανειλημμένες διασπάσεις των υποσυνόλων (κόμβων) του χώρου εισόδου S σε δύο υποσύνολα, ξεκινώντας με το ίδιο S. Σε κάθε τερματικό υποσύνολο (τελικοί κόμβοι) εκχωρείται μια ετικέτα κλάσης (class label) και η προκύπτουσα διαμέριση του χώρου εισόδου S αντιστοιχεί σε ένα ταξινομητή. Ωστόσο, τα δέντρα ταξινόμησης τείνουν να είναι ασταθή.

1.2.1.3 Σύνολο Ταξινομητών (Aggregating Classifiers)

Το σύνολο ταξινομητών είναι η κατασκευή πολλών ταξινομητών όπου συνδυάζονται για να κάνουν μια τελική απόφαση. Η ορθότητα της ασταθούς πρόβλεψης όπως στον ταξινομητή δέντρου-βάσης (tree-based classifiers) μπορεί να βελτιωθεί με το σύνολο των ταξινομητών. Παραδείγματα του συνόλου ταξινομητών περιλαμβάνουν Bagging (Breiman 1996) και Boosting (Freund και Schapire 1997). Τόσο το Bagging όσο και το Boosting είναι αναδειγματοληψία δεδομένων (resample data) για την κατασκευή πολλών ταξινομητών και στη συνέχεια συνδυάζοντας τους να αποκτήσουν μεγαλύτερη ακρίβεια.

1.2.1.4 Ταξινόμηση Boosting (Boosting Classification)

Ο στόχος της ταξινόμησης Boosting (Freund και Schapire, 1997) είναι να βελτιώσει την ακρίβεια του κάθε δεδομένου αλγόριθμου ταξινόμησης. Ισχύει διαδοχικά ο «αδύναμος» αλγόριθμος ταξινόμησης (weak classification algorithm) με επανάληψη τροποποιημένων εκδόσεων των δεδομένων, δημιουργώντας έτσι μια ακολουθία από «αδύναμους» ταξινομητές. Οι προβλέψεις απ' όλα αυτά, στη συνέχεια, σε συνδυασμό με μια πλειοψηφία που γίνεται παράγουν τον τελικό

ταξινομητή, ο οποίος μπορεί να επιτύχει μεγαλύτερη ακρίβεια απ' ότι οι «αδύναμοι» ταξινομητές στην ακολουθία.

1.2.1.5 Ταξινόμηση Περιθωρίου-βάσης (Margin-based Classification)

Οι ταξινομήσεις περιθωρίου-βάσης έχουν αποκτήσει πρόσφατα τεράστια δημοτικότητα. Αυτές οι τεχνικές έχουν αποδειχθεί αποτελεσματικές και πέτυχαν απόδοση με την τελευταία λέξη της τεχνολογίας (state-of-the-art). Παραδείγματα περιλαμβάνουν τις μηχανές διανυσμάτων υποστήριξης (Support Vector Machines) και ψ-μάθηση (Shen, Tseng, Zhang, και Wong, 2003), μεταξύ άλλων. Αυτές οι τεχνικές έχουν αποδειχθεί επιτυχείς σε μια σειρά από μελέτες που κυμαίνονται από γονιδιωματική ταξινόμηση του καρκίνου σε χειρόγραφη αναγνώρισης χαρακτήρων.

1.3 Ταξινομητές (Classifiers)

1.3.1 Η έννοια του περιθωρίου (Margin)

Στην γραμμική ταξινόμηση, για την διαχωρίσιμη περίπτωση τα θετικά υποδείγματα (με $y_i = +1$) στα δεδομένα εκπαίδευσης μπορούν να διαχωριστούν πλήρως από τα αρνητικά υποδείγματα (με $y_i = -1$).

Τα υπερεπίπεδα που χρησιμοποιούνται ως συνάρτηση απόφασης ορίζονται ως

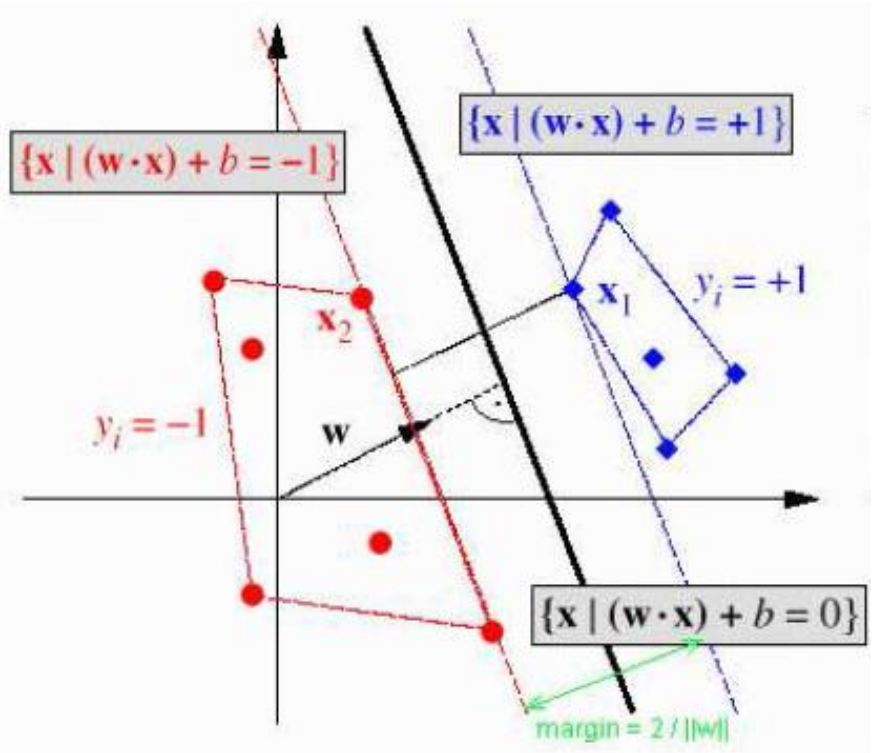
$$f(\boldsymbol{x}) = \langle \boldsymbol{w}, \boldsymbol{x} \rangle + b = \sum_{i=1}^{d} w_i x_i + b$$

όπου $(\boldsymbol{w}, b) \in R^d \times R$ και $\langle \cdot, \cdot \rangle$ είναι το σύνηθες εσωτερικό γινόμενο στον αντίστοιχο Ευκλείδειο χώρο $\mathrm{R^d}$.

Η συνάρτηση περιθωρίου ενός παραδείγματος (x_i, y_i), σε σχέση με ένα υπερεπίπεδο (**w**, b) ορίζεται ως η ποσότητα $\gamma_i = y_i f(x_i)$. Εδώ να σημειωθεί ότι το θετικό περιθώριο γ_i προϋποθέτει σωστή ταξινόμηση του (x_i, y_i). Τα

περιθώρια $\{\gamma_i\}_{i=1}^n$ είναι πολύ χρήσιμα για να δείχνουν την απόδοση ενός ταξινομητή, επειδή περιγράφουν το βαθμό διαχωρισμού των περιπτώσεων, καθώς και των δύο κλάσεων. Ένας ταξινομητής περιθωρίου -βάσης ορίζεται από μια αντικειμενική συνάρτηση η οποία είναι η συνάρτηση των περιθωρίων $\{\gamma_i\}_{i=1}^n$. Η γεωμετρική του περιθωρίου ορίζεται ως περιθώριο λειτουργιών ενός κανονικοποιημένου διανύσματος βάρους $\left(\frac{1}{\|\boldsymbol{w}\|}\boldsymbol{w}, \frac{1}{\|\boldsymbol{w}\|}b\right)$ το οποίο, ως εκ τούτου μετρά την Ευκλείδεια απόσταση των σημείων από το όριο απόφασης στο χώρο εισόδου. Εδώ $\|\boldsymbol{w}\|$ είναι η συνήθης L_2 νόρμα.

Τέλος, το περιθώριο ενός συνόλου εκπαίδευσης ορίζεται ως το άθροισμα της μικρότερης απόστασης από το υπερεπίπεδο στο πλησιέστερο θετικό υπόδειγμα και το πλησιέστερο αρνητικό υπόδειγμα. Όταν το πιο κοντινό θετικό υπόδειγμα βρίσκεται στο $f(x) = +1$ και ομοίως, το πιο κοντινό αρνητικό υπόδειγμα βρίσκεται στο $f(x) = -1$, το περιθώριο τότε δίνεται από $\frac{2}{\|\boldsymbol{w}\|}$.



Σχήμα 1.1: Το περιθώριο ενός συνόλου εκπαίδευσης

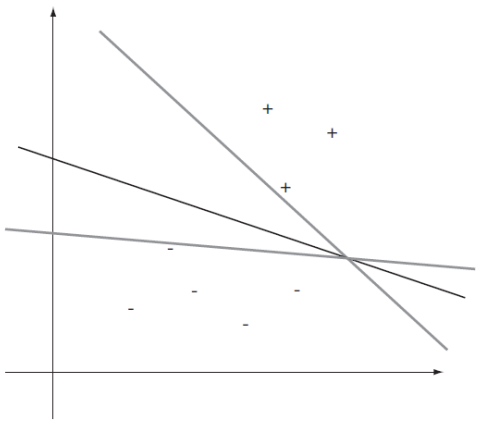
1.3.2 Ταξινομητές Μέγιστου Περιθωρίου (Maximum Margin Classifiers)

Οι διάφορες προσεγγίσεις μάθησης που έχουν αναπτυχθεί σε δυαδικά προβλήματα ταξινόμησης μπορεί να οδηγήσουν σε επιφάνειες απόφασης που υποδηλώνουν πιθανή κακή ταξινόμηση των σημείων που δεν αποτελούν μέρος του συνόλου εκπαίδευσης. Στρεβλώσεις της επιφάνειας απόφασης μπορεί να προκύψουν εξαιτίας των ακραίων τιμών. Αυτές οι στρεβλώσεις μπορούν να οδηγήσουν σε λανθασμένες ταξινομήσεις.

Για το λόγο αυτό είναι απαραίτητη μια νέα προσέγγιση που προσπαθεί να αποφύγει τις ελλείψεις αυτές. Η προσέγγιση αυτή βασίζεται στην αναζήτηση μιας επιφάνειας απόφασης που ισαπέχει στα όρια της κατηγορίας, όπου οι δύο κατηγορίες είναι πιο κοντά η μια στην άλλη και μεγιστοποιεί επίσης τις αποστάσεις σε αυτά τα όρια των κλάσεων. Με την τοποθέτηση της επιφάνειας απόφασης ακριβώς στη μέση μεταξύ των δύο ορίων της κατηγορίας και μεγιστοποιώντας τις αποστάσεις από τα όρια της κατηγορίας, αυτή η νέα προσέγγιση μειώνει την πιθανότητα εσφαλμένης ταξινόμησης. Καλούμε τέτοια μοντέλα ταξινομητές μέγιστου περιθωρίου.

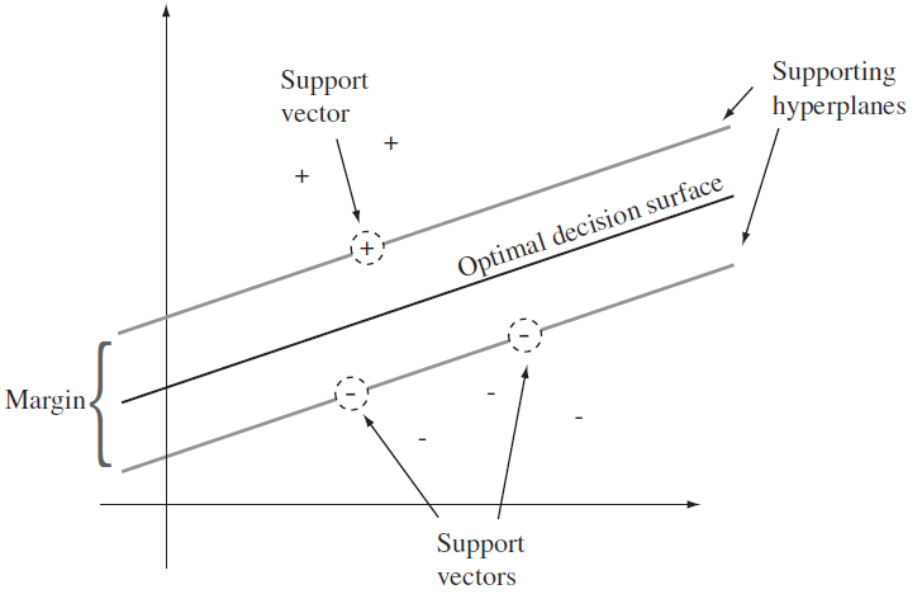
Το γεγονός ότι οι ταξινομητές μέγιστου περιθωρίου αναζητούν μια επιφάνεια απόφασης «μέγιστης απόστασης» οδηγεί σε ένα πρόβλημα βελτιστοποίησης, όπου πιο συγκεκριμένα από την κατασκευή του ταξινομητή μέγιστου περιθωρίου είναι ένα κυρτό πρόβλημα βελτιστοποίησης.

Όπως έχει αναφερθεί και πιο πάνω σε ένα γραμμικό διαχωρίσιμο σύνολο εκπαίδευσης για ένα δυαδικό πρόβλημα ταξινόμησης η τοποθέτηση της επιφάνειας απόφασης σε ίση απόσταση από τα αντίστοιχα όρια των κλάσεων αποτελεί την βέλτιστη επιφάνεια απόφασης.



Σχήμα 1.2: Η μαύρη γραμμή αντιπροσωπεύει βέλτιστη επιφάνεια απόφασης, και οι γκρι γραμμές αντιπροσωπεύουν μη βέλτιστες επιφάνειες απόφασης σε ένα δυαδικό πρόβλημα ταξινόμησης.

Πιο συγκεκριμένα μια επιφάνεια σ' ένα δυαδικό πρόβλημα ταξινόμησης είναι βέλτιστη αν είναι σε ίση απόσταση από τα δύο υπερεπίπεδα υποστήριξης (supporting hyperplanes) και μεγιστοποιεί το περιθώριο τους, όπου το υπερεπίπεδο υποστήριξης καλείται να είναι το υπερεπίπεδο που αποτελεί μια κλάση η οποία είναι παράλληλη προς την (γραμμική) επιφάνεια απόφασης και όλα τα σημεία της αντίστοιχης κλάσης του είναι πάνω ή κάτω από αυτήν.



Σχήμα 1.3: Βέλτιστος διαχωρισμός υπερεπιπέδου με δύο υπερεπίπεδα υποστήριξης. Τα δύο υπερεπίπεδα υποστήριξης είναι έτσι ώστε να αγγίζουν απλώς τα αντίστοιχα όρια της κατηγορίας τους. Η απόσταση μεταξύ των υπερεπιπέδων είναι το περιθώριο, και η βέλτιστη επιφάνεια απόφασης βρίσκεται στο κέντρο του περιθωρίου. Τα κυκλωμένα σημεία σε κάθε κατηγορία ονομάζονται διανύσματα υποστήριξης.

Ένας ταξινομητής μέγιστου περιθωρίου λειτουργεί ως εξής:

Έστω ένα γραμμικά διαχωρίσιμο σύνολο εκπαίδευσης

$$D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \dots, (\boldsymbol{x}_l, y_l)\} \subseteq R^n \times \{+1, -1\}$$

μπορούμε να υπολογίσει μια επιφάνεια απόφασης μέγιστου περιθωρίου

$$\boldsymbol{w}^* \bullet \boldsymbol{x} = b^*$$

με την ελαχιστοποίηση

$$min\phi(\boldsymbol{w}, b) = \min_{\boldsymbol{w},b} \frac{1}{2} \|\boldsymbol{w}\|^2$$

Για να αποφύγουμε τα σημεία μας να «πέσουν» στο περιθώριο προσθέτουμε τους ακόλουθους περιορισμούς:

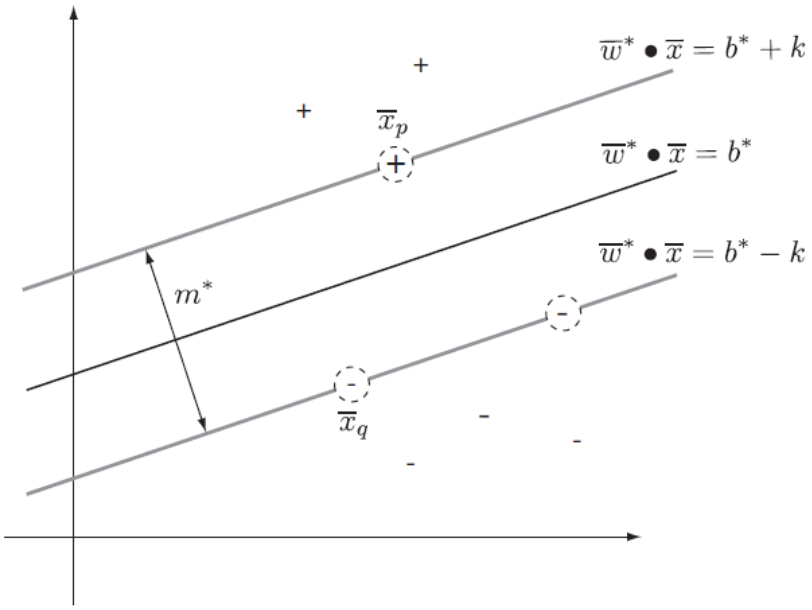
$$\boldsymbol{w} \bullet x - b \geq 1, \text{για } x_i \text{ στην πρώτη κλάση}$$
$$\text{ή}$$
$$\boldsymbol{w} \bullet x - b \leq -1, \text{για } x_i \text{ στη δεύτερη κλάση.}$$

Αυτοί οι περιορισμοί μπορούν να ξαναγραφτούν μαζί ως

$$y_i(\boldsymbol{w} \bullet x - b) \geq 1, \text{για } 1 \leq i \leq l.$$

Έτσι για την ελαχιστοποίηση μας παίρνουμε τους περιορισμούς:

$$\boldsymbol{w} \bullet (y_i \boldsymbol{x}_i) \geq 1 + y_i b \text{ για κάθε } (y_i \boldsymbol{x}_i) \in D$$

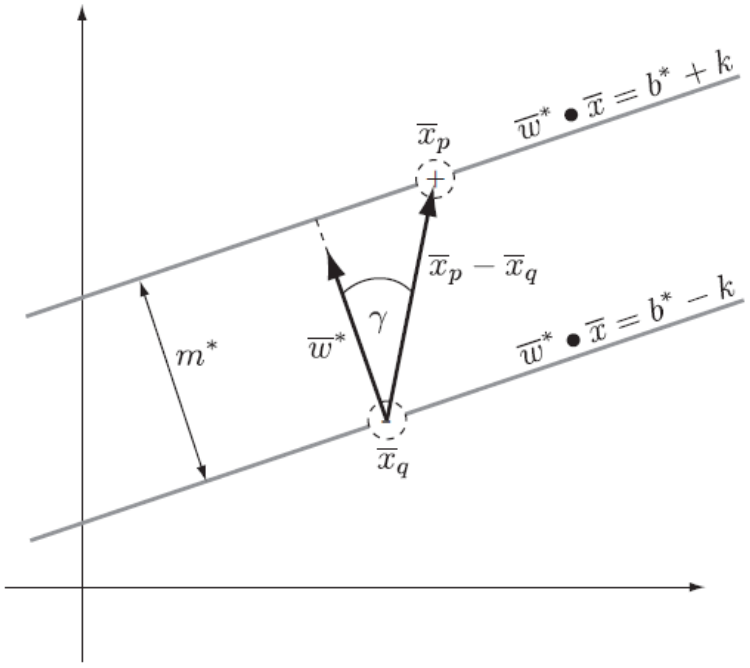


Σχήμα 1.4: Μέγιστο περιθώριο m* της βέλτιστης επιφάνειας $w^* \bullet x = b^*$ (όπου $\overline{w}^*$ είναι το διάνυσμα του w^*) με τα δύο υπερεπίπεδα υποστήριξης να είναι σε ίση απόσταση από την επιφάνεια απόφασης με $w^* \bullet x = b^* + k$ όπου (k=1) να αποτελεί το πρώτο υπερεπίπεδο υποστήριξης για την κατηγορία +1 το οποίο βρίσκεται πάνω από την επιφάνεια απόφασης και το δεύτερο υπερεπίπεδο υποστήριξης $w^* \bullet x = b^* - k$, για την κατηγορία -1 το οποίο βρίσκεται κάτω από την επιφάνεια απόφασης. Το $(x_p, +1)$ είναι διάνυσμα υποστήριξης για την +1 κατηγορία και το $(x_q, -1)$ είναι διάνυσμα υποστήριξης για την -1 κατηγορία τα οποία ικανοποιούν αντίστοιχα τις εξισώσεις των υπερεπιπέδων υποστήριξης όπου ανήκουν.

Τέλος, ο υπολογισμός του μέγιστου περιθωρίου γίνεται με την προβολή της διαφοράς μεταξύ των δύο διανυσμάτων υποστήριξης $\boldsymbol{x}_p - \boldsymbol{x}_q$ στην κατεύθυνση του κανονικού διανύσματος $\boldsymbol{w}^*$ της επιφάνειας απόφασης, δηλαδή

$$\begin{aligned} m^* &= |\boldsymbol{x}_p - \boldsymbol{x}_q| cos\gamma \\ &= \frac{\boldsymbol{w}^* \bullet (\boldsymbol{x}_p - \boldsymbol{x}_q)}{|\boldsymbol{w}^*|} \\ &= \frac{\boldsymbol{w}^* \bullet \boldsymbol{x}_p - \boldsymbol{w}^* \bullet \boldsymbol{x}_q}{|\boldsymbol{w}^*|} \\ &= \frac{(b^* + k) - (b^* - k)}{|\boldsymbol{w}^*|} \\ &= \frac{2k}{|\boldsymbol{w}^*|} \end{aligned}$$

όπου οι δύο πρώτες σχέσεις προκύπτουν από τον ορισμό της προβολής, η τρίτη λόγο γραμμικότητας και η τέταρτη από τις εξισώσεις των υπερεπιπέδων υποστηριξης.



Σχήμα 1.5: Ο υπολογισμός του περιθωρίου m* μεταξύ των δύο υπερεπιπέδων υποστήριξης

1.3.3 Ταξινομητές «Μαλακού» και «Σκληρού» Περιθωρίου (Soft-Margin Classifiers and Hard-Margin Classifiers)

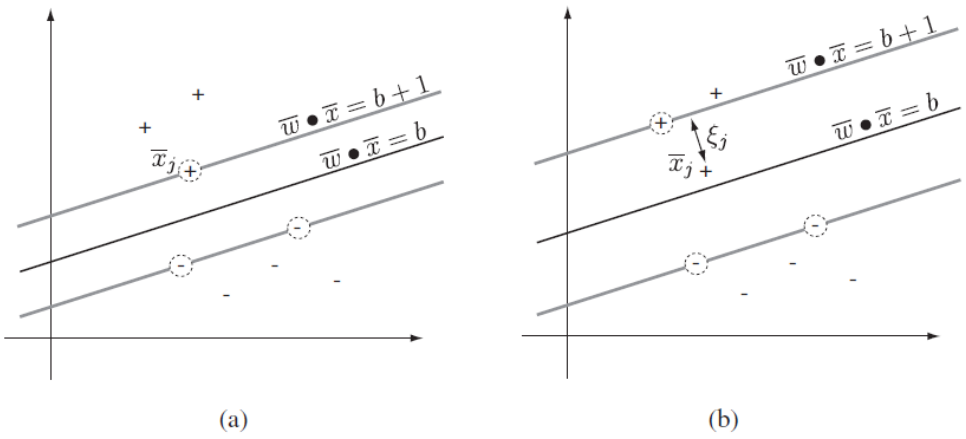
Από το γεγονός ότι στον πραγματικό κόσμο τα σύνολα εκπαίδευσης δεν είναι τέλεια και περιέχουν θορύβους οι οποίοι μπορεί να προκαλέσουν ένα εξαιρετικά περίπλοκο όριο μεταξύ των κλάσεων ενός προβλήματος ταξινόμησης, ένας ταξινομητής που δεν επιτρέπεται να κάνει λάθη, θα πρέπει να διαμορφώσει αυτό το περίπλοκο όριο άψογα, δίνοντας αφορμή για μια εξαιρετικά περίπλοκη επιφάνεια απόφασης. Για να αποφευχθεί η επιλογή μιας περίπλοκης επιφάνειας απόφασης σε ορισμένες περιπτώσεις επιτρέπεται ο ταξινομητής να αγνοήσει αυτά τα σημεία θορύβου, για να μπορεί να υπολογίσει μια πολύ πιο απλούστερη επιφάνεια απόφασης. Αυτό είναι πολύ ελκυστικό, δεδομένου ότι απλές επιφάνειες απόφασης έχουν πολύ μεγαλύτερη πιθανότητα σωστής ταξινόμησης των σημείων που δεν αποτελούν μέρος του συνόλου εκπαίδευσης. Με άλλα λόγια, απλές επιφάνειες απόφασης τείνουν να γενικεύσουν καλύτερα. Αυτή η επιτρεπτή αγνόηση των σημείων που βρίσκονται στην «λάθος» πλευρά των αντίστοιχων υπερεπιπέδων υποστήριξης γίνεται μέσω της εισαγωγής μίας μεταβλητής χαλάρωσης.

Αυτό οδήγησε στον διαχωρισμό των ταξινομητών μέγιστου περιθωρίου σε δύο κατηγορίες, στους ταξινομητές μέγιστου περιθωρίου οι οποίοι ενσωματώνουν μια μεταβλητή χαλάρωσης και καλούνται ταξινομητές «μαλακού» περιθωρίου (soft-margin) και στους ταξινομητές μέγιστου περιθωρίου οι οποίοι δεν ενσωματώνουν μεταβλητή χαλάρωσης και καλούνται ταξινομητές «σκληρού» περιθωρίου (hard-margin). Οι ταξινομητές «σκληρού» περιθωρίου λειτουργούν ακριβώς με το ίδιο τρόπο όπως έχουμε αναλύσει τους ταξινομητές μέγιστου περιθωρίου. Τώρα θα αναλύσουμε τους ταξινομητές «μαλακού» περιθωρίου.

Ταξινομητές «μαλακού» περιθωρίου

Η βελτιστοποίηση κατασκευάζει ένα ταξινομητή μέγιστου περιθωρίου από την τοποθέτηση των υπερεπίπεδων υποστήριξης (supporting hyperplanes), όσο το δυνατό πιο μακριά από την επιφάνεια απόφασης έτσι ώστε να αγγίζουν απλώς τα αντίστοιχα όρια της κλάση τους. Αυτό εν μέρει δεν είναι επιτυχές στην περίπτωση όπου υπάρχουν σφάλματα λόγω μετρήσεων ή ακόμη και λόγω εσφαλμένης καταχώρηση των δεδομένων. Οι ταξινομητές μαλακού περιθωρίου μειώνουν τον αντίκτυπο που αυτά τα σημεία έχουν στο μέγεθος του περιθωρίου, επιτρέποντάς τους να βρίσκονται στην «λάθος» πλευρά των αντίστοιχων υπερεπίπεδων υποστήριξης με την εισαγωγή μεταβλητών χαλάρωσης. Οι μεταβλητές χαλάρωσης είναι όροι σφάλματος που μετρούν πόσο μακριά βρίσκεται ένα συγκεκριμένο σημείο το οποίο βρίσκεται στη λάθος πλευρά των

αντίστοιχων υπερεπίπεδων υποστήριξης. Πιο συγκεκριμένα μια μεταβλητή χαλάρωσης μέτρα πόσο σφάλμα διαπράττεται από το γεγονός ότι επιτρέπεται στα σημεία να βρίσκονται σε λάθος πλευρά χωρίς να περιορίζονται από το συγκεκριμένο σημείο. Το πιο κάτω σχήμα (Σχήμα 1.6) απεικονίζει ένα ταξινομητή «μαλακού» περιθωρίου.



Σχήμα 1.6: Ταξινομητής «μαλακού» περιθωρίου: Στο τμήμα (α) βλέπουμε έναν ταξινομητή μέγιστου περιθωρίου με το περιθώριο να περιορίζεται από το σημείο $(x_j, +1)$. Στο τμήμα (β) τα υπερεπίπεδα υποστήριξης $\overline{w} \bullet x = b + 1$ δεν περιορίζονται από το σημείο εκπαίδευση $(x_j, +1)$. Εδώ το σημείο εκπαίδευση επιτρέπεται να βρίσκεται στη λάθος πλευρά του υπερεπιπέδου υποστήριξης και η ποσότητα του σφάλματος μετριέται από την αντίστοιχη μεταβλητή χαλάρωσης ξ_j .

Σε αυτή την περίπτωση όμως ο περιορισμός

$$\boldsymbol{w} \bullet \boldsymbol{x}_j - b - 1 \geq 0$$

από το πρόβλημα ελαχιστοποίησης παραβιάζεται. Ωστόσο, μπορούμε να ανακτήσουμε ένα λογικό περιορισμό λαμβάνοντας υπόψη τη μεταβλητή χαλάρωση, ως εξής,

$$\boldsymbol{w} \bullet \boldsymbol{x}_j - b + \xi_j - 1 \geq 0.$$

Δηλαδή, η μεταβλητή χαλάρωσης ξ_j του σημείου $(\boldsymbol{x}_j, +1)$ δημιουργεί την «ψευδαίσθηση» ότι το σημείο βρίσκεται ακριβώς πάνω στο υπερεπίπεδο υποστήριξης και ως εκ τούτου φαίνεται να ικανοποιεί τον αρχικό περιορισμό. Εδώ πρέπει να σημειωθεί ότι $\xi_j \geq 0$, δηλαδή, το σφάλμα είναι πάντα θετικό και μετράται ως θετική ποσότητα. Τώρα, αν εισάγουμε μια μεταβλητή χαλάρωσης για κάθε σημείο εκπαίδευσης $(\boldsymbol{x}_i, y_i)$ οι αντίστοιχοι τροποποιημένοι περιορισμοί που προκύπτουν είναι

$$y_i(\boldsymbol{w} \bullet \boldsymbol{x}_i - b) + \xi_i - 1 \geq 0 \text{ με } \xi_i \geq 0.$$

Είναι εύκολο να παρατηρήσουμε ότι, αν κάποια στιγμή το $\boldsymbol{x}_i$ δεν αποτελεί εμπόδιο για το αντίστοιχο υπερεπίπεδο υποστήριξης, η αντίστοιχη μεταβλητή χαλάρωσης $\xi_i = 0$ αφού δεν υπάρχει τίποτα για να διορθώσει. Σε αυτή την περίπτωση λαμβάνουμε τον αρχικό μας περιορισμός για αυτό το σημείο

$$y_i(\boldsymbol{w} \bullet \boldsymbol{x}_i - b) - 1 \geq 0.$$

Αν, από την άλλη πλευρά, το σημείο $\boldsymbol{x}_i$ βρίσκεται στη λάθος πλευρά των αντίστοιχων υπερεπιπέδων υποστήριξης, οι πιο πάνω περιορισμοί οδηγούν σε ένα νέο περιορισμό με $\xi_i > 0$.

Συνοψίζοντας τα πιο πάνω, ένας ταξινομητής «μαλακού» περιθωρίου λειτουργεί ως εξής:

Έστω ένα σύνολο εκπαίδευσης

$$D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \dots, (\boldsymbol{x}_l, y_l)\} \subseteq R^n \times \{+1, -1\}$$

μπορούμε να υπολογίσουμε μια επιφάνεια απόφασης «μαλακού» περιθωρίου

$$\boldsymbol{w}^* \bullet \boldsymbol{x} = b^*$$

με την ελαχιστοποίηση

$$min\phi(\boldsymbol{w}, \boldsymbol{\xi}, b) = \min_{\boldsymbol{w},\boldsymbol{\xi},b} \left(\frac{1}{2} \|\boldsymbol{w}\|^2 + C \sum_{i=1}^{l} \xi_i \right)$$

με τους εξής περιορισμούς

$$y_i(\boldsymbol{w} \bullet \boldsymbol{x}_i - b) + \xi_i - 1 \geq 0$$

$$\xi_i \geq 0$$

με $i = 1, \dots, l$, $\boldsymbol{\xi} = (\xi_1, \dots, \xi_l)$ και C>0 να είναι η παράμετρος που ελέγχει το trade-off μεταξύ της ποινής της μεταβλητής χαλάρωσης και του μεγέθους του περιθωρίου.

Αξιοσημείωτο είναι το γεγονός ότι οι ταξινομητές μέγιστου περιθωρίου δίνουν την δυνατότητα στους γραμμικούς ταξινομητές που βασίζονται σε μηχανές διανυσμάτων υποστήριξης (support vector machine) να μπορούν εύκολα να επεκταθούν και σε μη γραμμικούς ταξινομητές, διευρύνοντας έτσι την δυνατότητα εφαρμογής των μηχανών διανυσμάτων υποστήριξης.

Κεφάλαιο 2

Μηχανές Διανυσμάτων Υποστήριξης

(Support Vector Machine)

2.1 Εισαγωγή

Οι Μηχανές Διανυσμάτων Υποστήριξης σαν ιδέα δημιουργήθηκαν από τον Cortes και τον Vapnik (2000). Στη μηχανική μάθηση οι μηχανές διανυσμάτων υποστήριξης είναι μοντέλα μάθησης με επίβλεψη, με τους σχετικούς αλγόριθμους εκμάθησης που αναλύουν τα δεδομένα για να αναγνωρίσουν τα πρότυπα και τα οποία χρησιμοποιούνται για την ταξινόμηση.

Η SVM τεχνική βασίζεται στην στατιστική θεωρία της μάθησης και μπορεί να χρησιμοποιηθεί για την πρόβλεψη μελλοντικών δεδομένων. Αποτελεί ένα ισχυρό εργαλείο για δυαδική ταξινόμηση, έχει αποκτήσει τεράστια δημοτικότητα και προσέλκυσε μεγάλο ενδιαφέρον λόγω των θεωρητικών της πλεονεκτημάτων και την επιτυχία της σε πραγματικές εφαρμογές, όπως γονιδιωματική ταξινόμηση του καρκίνου (cancer genomics classification), τα δεδομένα γονιδιακής έκφρασης

για την ταξινόμηση κειμένου και την χειρόγραφη αναγνώριση χαρακτήρων. Η SVM έχει επίσης επιτύχει την απόδοση state-of-art σε τομείς όπως η κατηγοριοποίηση κειμένων.

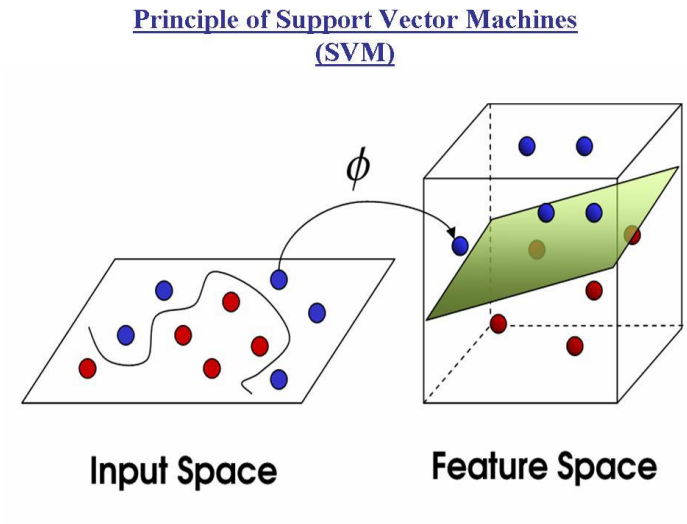
Με μια πρώτη μάτια οι μηχανές διανυσμάτων υποστήριξης δεν μοιάζουν με τις τυπικές μεθόδους στατιστικής ανάλυσης αφού η ανάπτυξη τους είναι εντελώς διαφορετική από τους αλγόριθμους που χρησιμοποιούνται για τη μάθηση και έτσι η SVM τεχνική παρέχει μια νέα άποψη μάθησης.

Τέλος τα τέσσερα πιο σημαντικά χαρακτηριστικά της SVM τεχνικής είναι η δυαδικότητα, οι πυρήνες, η κυρτότητα και η σποραδικότητα.

2.2 SVM για δυαδικά προβλήματα ταξινόμησης

Η SVM είναι μια χρήσιμη τεχνική για την ταξινόμηση των δεδομένων. Μια διαδικασία ταξινόμησης περιλαμβάνει συνήθως δεδομένα εκπαίδευσης, όπου κάθε παρατήρηση ανήκει σε μια κατηγορία της μεταβλητής απόκρισης. Ο αλγόριθμος εκπαίδευσης για τις SVM παράγει ένα μη πιθανοθεωρητικό μοντέλο το οποίο ταξινομεί τις παρατηρήσεις του συνόλου εξέτασης στις αντίστοιχες κατηγορίες απόκρισης.

Ένα μοντέλο SVM είναι μια αναπαράσταση του συνόλου εκπαίδευσης (training set) ως σημεία στο χώρο, τα οποία χαρτογραφούνται έτσι ώστε τα δεδομένα των επιμέρους κατηγοριών να χωρίζονται από ένα σαφές κενό που είναι όσο το δυνατόν ευρύτερο. Τα νέα δεδομένα στη συνέχεια αντιστοιχίζονται με το ίδιο διάστημα και προβλέπεται αν ανήκουν σε μια κατηγορία με βάση σε ποια πλευρά του χάσματος θα πέσουν. Με άλλα λόγια η εύρεση του κατάλληλου μοντέλου ταυτίζεται με την εύρεση του βέλτιστου διαχωριστικού υπερεπιπέδου ή αλλιώς του ορίου απόφασης (decision boundary) χρησιμοποιώντας την πληροφορία που παρέχεται από τις επεξηγηματικές μεταβλητές ούτως ώστε ο διαχωρισμός των παρατηρήσεων σε κατηγορίες να είναι όσο το δυνατό πιο ομογενής. Η διαδικασία ταξινόμησης ολοκληρώνεται όταν βρεθεί το διαχωριστικό επίπεδο που απέχει την μεγαλύτερη δυνατή απόσταση από τις πιθανές κατηγορίες.



Σχήμα 2.1: Παρατηρούμε τη βασική αρχή στην οποία βασίζονται οι μηχανές διανυσματικής υποστήριξης (SVM). Υποδεικνύει το διαχωριστικό υπερεπίπεδο όταν εφαρμόζεται στο χώρο των χαρακτηριστικών (feature space).

Η απλούστερη μορφή επίλυσης ενός προβλήματος πρόβλεψης είναι η δυαδική κατηγοριοποίηση (binary classification), όπου πρέπει να γίνει ένας διαχωρισμός σε αντικείμενα που ανήκουν σε μία από τις δύο κατηγορίες οι οποίες συμβολίζονται με θετικό (+1) ή αρνητικό (-1) πρόσημο. Οι SVMs χρησιμοποιούν για την επίλυση του προβλήματος:

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

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

Το βέλτιστο διαχωριστικό υπερεπίπεδο διαχωρίζει τις δύο κλάσεις και μεγιστοποιεί την απόσταση στο πλησιέστερο σημείο από κάθε κατηγορία. Δηλαδή παρέχει μία μοναδική λύση στο πρόβλημα του διαχωριστικού υπερεπιπέδου, με τη μεγιστοποίηση του περιθωρίου μεταξύ των δύο κατηγοριών για τα δεδομένα εκπαίδευσης, που οδηγεί σε καλύτερη απόδοση ταξινόμησης για τα δεδομένα δοκιμών.

Μέχρι στιγμής, έχουμε αναφερθεί σε γραμμικό διαχωρισμό των δεδομένων σε κατηγορίες των δύο διαστάσεων. Στην επόμενη παράγραφο θα παρουσιαστεί και η περίπτωση όπου ο διαχωρισμός δεν είναι γραμμικός, καθώς επίσης και σε επόμενη ενότητα θα δούμε ότι εκτός από δυαδικά προβλήματα ταξινόμησης οι δυαδικές SVM μπορούν να χρησιμοποιηθούν και για την ταξινόμηση των αντικειμένων σε ένα αυθαίρετο αριθμό κατηγοριών.

2.2.1 Γραμμικά Διαχωρίσιμα Δεδομένα

Έστω ένα σύνολο εκπαίδευσης

$$D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \dots, (\boldsymbol{x}_l, y_l)\} \subseteq R^n \times \{+1, -1\}$$

δηλαδή, έχουμε L σημεία εκπαίδευσης, όπου κάθε είσοδος x_i έχει n χαρακτηριστικά και είναι σε μια από τις δύο κατηγορίες $y_i = -1$ ή $y_i = +1$.

Υποθέτουμε ότι τα δεδομένα είναι γραμμικά διαχωρίσιμα και θέλουμε να βρούμε το υπερεπίπεδο μέγιστου περιθωρίου που χωρίζει τα σημεία που έχουν $y_i = -1$ από αυτά που έχουν $y_i = +1$.

Όπως αναφέραμε και στο πρώτο κεφάλαιο αυτό το υπερεπίπεδο μπορεί να γραφτεί ως

$$\boldsymbol{w} \bullet \boldsymbol{x} - b = 0$$

όπου $\boldsymbol{w}$ είναι το κάθετο διάνυσμα προς το υπερεπίπεδο και η παράμετρος

$$\frac{b}{\|\boldsymbol{w}\|}$$

είναι η κάθετη απόσταση από το υπερεπίπεδο προς την αρχή.

Η υλοποίηση SVM στηρίζεται στην επιλογή των μεταβλητών $\boldsymbol{w}$ και b, έτσι ώστε τα δεδομένα εκπαίδευσης να μπορούν να περιγραφούν με:

$$y_i(\boldsymbol{w} \bullet \boldsymbol{x}_i - b) - 1 \geq 0$$

δηλαδή, βρισκόμαστε στην περίπτωση «σκληρού» περιθωρίου χωρίς κάποια μεταβλητή χαλάρωσης.

Από αυτά που έχουμε αναφέρει και στο πρώτο κεφάλαιο έχουμε το πρόβλημα ελαχιστοποίησης

$$min\phi(\boldsymbol{w}, b) = \min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2$$

με τους εξής περιορισμούς

$$y_i(\boldsymbol{w} \bullet \boldsymbol{x}_i - b) - 1 \geq 0$$

για κάθε i=1,...,l.

Εμείς χρειάζεται να υπολογίσουμε αυτό το πρόβλημα ελαχιστοποίησης.

Προκειμένου να ληφθεί μέριμνα για τους περιορισμούς σε αυτή την ελαχιστοποίηση, θα πρέπει να κατανέμουμε σε αυτούς, πολλαπλασιαστές

Lagrange $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_l)$. Έτσι με την εισαγωγή αυτών των πολλαπλασιαστών προκύπτει το αντίστοιχο Lagrangian ως

$$L(\boldsymbol{\alpha}, \boldsymbol{w}, b) = \frac{1}{2}\|\boldsymbol{w}\|^2 - \alpha[y_i(\boldsymbol{w} \bullet \boldsymbol{x}_i - b) - 1] \quad \text{για κάθε i=1,..,l}$$

$$= \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{l} \alpha_i (y_i(\boldsymbol{w} \bullet \boldsymbol{x}_i - b) - 1)$$

$$= \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{l} \alpha_i y_i \boldsymbol{w} \bullet \boldsymbol{x}_i + b\sum_{i=1}^{l} \alpha_i y_i + \sum_{i=1}^{l} \alpha_i \qquad (2.2.1)$$

Θέλουμε να βρούμε το $\boldsymbol{w}^*$ και το b^* τα οποία ελαχιστοποιούν, και το $\boldsymbol{\alpha}$ το οποίο μεγιστοποιεί το πρόβλημα μας, δηλαδή

$$\max_{\boldsymbol{\alpha}} \min_{\boldsymbol{w},b} L(\boldsymbol{\alpha}, \boldsymbol{w}, b) = L(\boldsymbol{\alpha}^*, \boldsymbol{w}^*, b^*)$$

με τους περιορισμούς

$$\alpha_i \geq 0.$$

Αυτό επιτυγχάνεται διαφορίζοντας την $L(\boldsymbol{\alpha}, \boldsymbol{w}, b)$ ως προς $\boldsymbol{w}$ και b, και θέτοντας τις παραγώγους ίσες με το μηδέν:

$$\frac{\partial L}{\partial \boldsymbol{w}}(\boldsymbol{\alpha}, \boldsymbol{w}^*, b) = \boldsymbol{w}^* - \sum_{i=1}^{l} \alpha_i y_i \boldsymbol{x}_i = 0 \Longrightarrow \boldsymbol{w}^* = \sum_{i=1}^{l} \alpha_i y_i \boldsymbol{x}_i \qquad (2.2.2)$$

$$\frac{\partial L}{\partial b}(\boldsymbol{\alpha}, \boldsymbol{w}, b^*) = \sum_{i=1}^{l} \alpha_i y_i = 0 \qquad (2.2.3)$$

Παρατηρούμε ότι, η μερική παράγωγος του L σε σχέση με το b δεν αποφέρει μια έκφραση για b*, αλλά αντ' αυτού, μας παρέχει τον περιορισμό

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

Αντικαθιστώντας την (2.2.2) και (2.2.3) στην (2.2.1) παίρνουμε μια άλλη μορφή η οποία εξαρτάται από το $\boldsymbol{\alpha}$, και τότε πρέπει να μεγιστοποιήσουμε την:

$$L(\boldsymbol{\alpha}, \boldsymbol{w}^*, b^*) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i \bullet \boldsymbol{x}_j$$

με τους περιορισμούς

$$\alpha_i \geq 0 \text{ για κάθε i=1,..,l}$$

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

Το οποίο ισοδυναμεί με

$$L(\boldsymbol{\alpha}, \boldsymbol{w}^*, b^*) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} \alpha_i H_{ij}\, \alpha_j \quad \text{όπου} \quad H_{ij} = y_i y_j \boldsymbol{x}_i \bullet \boldsymbol{x}_j$$

$$\Longrightarrow \qquad L(\boldsymbol{\alpha}, \boldsymbol{w}^*, b^*) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2}\alpha^T H \alpha$$

με περιορισμούς

$$\alpha_i \geq 0 \text{ για κάθε i=1,..,l}$$

$$\sum_{i=1}^{l} \alpha_i y_i = 0 \qquad (2.2.4)$$

Αυτή η νέα σύνθεση $L(\boldsymbol{\alpha}, \boldsymbol{w}^*, b^*)$ αναφέρεται ως η διπλή μορφή της πρωτοβάθμιας $L(\boldsymbol{\alpha}, \boldsymbol{w}, b)$. Αξίζει να σημειωθεί ότι η διπλή μορφή απαιτεί μόνο να υπολογιστεί το γινόμενο των διανυσμάτων εισόδου x_i . Αυτό είναι πολύ σημαντικό για το τέχνασμα του πυρήνα που περιγράφεται στην επόμενη παράγραφο.

Το πρόβλημα μας έχει μετατοπιστεί πλέον από την ελαχιστοποίηση της $L(\boldsymbol{\alpha}, \boldsymbol{w}, b)$ στην μεγιστοποίηση της $L(\boldsymbol{\alpha}, \boldsymbol{w}^*, b^*)$, δηλαδή πρέπει να βρεθεί:

$$\max_{\boldsymbol{\alpha}} \left[\sum_{i=1}^{l} \alpha_i - \frac{1}{2}\alpha^T H \alpha \right]$$

με περιορισμούς

$$\alpha_i \geq 0 \text{ για κάθε i=1,..,l}$$

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

Αυτό είναι ένα κυρτό τετραγωνικό πρόβλημα βελτιστοποίησης και διατρέχουμε μια QP επίλυση η οποία θα επιστρέψει το $\boldsymbol{\alpha}$ και η (2.2.2) θα μα δώσει το $\boldsymbol{w}^*$. Για τον υπολογισμό του b^* αντικαθιστούμε στην (2.2.2) την (2.2.3) και

χρησιμοποιώντας την $y_i(\boldsymbol{w} \bullet \boldsymbol{x}_i - b) - 1 \geq 0$ για κάθε i=1,..,l και παίρνοντας το μέσο όρο όλων των x_s βρίσκουμε το b^*.

2.2.2 Μη Γραμμικά Διαχωρίσιμα Δεδομένα

Όπως αναφέραμε και στο πρώτο κεφάλαιο οι ταξινομητές μέγιστου περιθωρίου δίνουν την δυνατότητα στους γραμμικούς ταξινομητές που βασίζονται σε SVM να μπορούν να επεκταθούν σε μη γραμμικούς ταξινομητές.

Η υλοποίηση SVM σε αυτή την περίπτωση στηρίζεται στην επιλογή των μεταβλητών $\boldsymbol{w}$, $\boldsymbol{\xi}$ και b, έτσι ώστε τα δεδομένα εκπαίδευσης να μπορούν να περιγραφούν με:

$$y_i(\boldsymbol{w} \bullet \boldsymbol{x}_i - b) - 1 + \boldsymbol{\xi}_i \geq 0$$

δηλαδή, βρισκόμαστε στην περίπτωση «μαλακού» περιθωρίου με την ύπαρξη μεταβλητής χαλάρωσης ξ_i.

Έτσι στα μη γραμμικά διαχωρίσιμα δεδομένα ισχύει ότι έχουμε πει στο πρώτο κεφάλαιο για τους ταξινομητές «μαλακού» περιθωρίου, δηλαδή έχουμε την ελαχιστοποίηση

$$min\phi(\boldsymbol{w}, \boldsymbol{\xi}, b) = \min_{\boldsymbol{w},\boldsymbol{\xi},b} \left(\frac{1}{2} \|\boldsymbol{w}\|^2 + C \sum_{i=1}^{l} \xi_i \right)$$

με τους εξής περιορισμούς

$$y_i(\boldsymbol{w} \bullet \boldsymbol{x}_i - b) + \xi_i - 1 \geq 0$$

$$\xi_i \geq 0$$

με $i = 1, \dots, l$, $\boldsymbol{\xi} = (\xi_1, \dots, \xi_l)$ και C>0 να είναι η παράμετρος που ελέγχει το trade-off μεταξύ της ποινής της μεταβλητής χαλάρωσης και του μεγέθους του περιθωρίου.

Η αναδιατύπωση ως Lagrangian, η οποία όπως και πριν θα πρέπει να ελαχιστοποιηθεί σε σχέση με τα $\boldsymbol{w}$, b και ξ_i για να μεγιστοποιηθεί ως προς $\boldsymbol{\alpha}$, έχει ως εξής:

$$L(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{w}, \boldsymbol{\xi}, b) = \frac{1}{2} \|\boldsymbol{w}\|^2 + C \sum_{i=1}^{l} \xi_i - \sum_{i=1}^{l} \alpha_i [(\boldsymbol{w} \bullet \boldsymbol{x}_i - b) + \xi_i - 1] - \sum_{i=1}^{l} \beta_i \xi_i$$

όπου η μεταβλητή $\boldsymbol{\beta} = (\beta_1, \ldots., \beta_l)$ αποτελεί του πολλαπλασιαστές Lagrange για τους περιορισμού $\xi_i \geq 0$.

Τώρα θέλουμε να βρούμε το $\boldsymbol{w}^*$ το $\boldsymbol{\xi}^*$ και το b^* τα οποία ελαχιστοποιούν, και τα $\boldsymbol{\alpha}$ και $\boldsymbol{\beta}$ τα οποία μεγιστοποιούν το πρόβλημα μας, δηλαδή

$$\max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \min_{\boldsymbol{w},\boldsymbol{\xi},b} L(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{w}, \boldsymbol{\xi}, b)$$

υπό τους περιορισμούς

$$\alpha_i \geq 0$$

$$\beta_i \geq 0$$

για κάθε i=1,...,l.

Αυτό επιτυγχάνεται διαφορίζοντας την $L(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{w}, \boldsymbol{\xi}, b)$ ως προς $\boldsymbol{w}$, b και ξ_i και θέτοντας τις παραγώγους ίσες με το μηδέν:

$$\frac{\partial L}{\partial \boldsymbol{w}}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{w}^*, \boldsymbol{\xi}, b) = \boldsymbol{w}^* - \sum_{i=1}^{l} \alpha_i y_i \boldsymbol{x}_i = 0 \Longrightarrow \boldsymbol{w}^* = \sum_{i=1}^{l} \alpha_i y_i \boldsymbol{x}_i$$

$$\frac{\partial L}{\partial b}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{w}, \boldsymbol{\xi}, b^*) = \sum_{i=1}^{l} \alpha_i y_i = 0 \Longrightarrow \text{προκύπτει ο περιορισμός } \sum_{i=1}^{l} \alpha_i y_i = 0$$

όπως και στην περίπτωση του «σκληρού» περιθωρίου

$$\frac{\partial L}{\partial \xi_i}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{w}, \xi_i^*, b) = 0 \Longrightarrow C - \alpha_i - \beta_i = 0$$

$\Longrightarrow$ μας δίνει τους νέους περιορισμούς $\alpha_i = C - \beta_i$ (2.2.5)

Αντικαθιστώντας τους όρους που λαμβάνονται από τις πιο πάνω μερικές διαφοροποιήσεις και με την εφαρμογή των περιορισμών παίρνουμε την εξής αντικειμενική συνάρτηση:

$$L(\boldsymbol{\alpha}, \boldsymbol{w}, b^*) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i \bullet \boldsymbol{x}_j$$

η οποία έχει την ίδια μορφή με την αντικειμενική συνάρτηση του ταξινομητή «σκληρού» περιθωρίου που δίνεται στην (2.2.4). Έτσι οδηγούμαστε στο συμπέρασμα ότι η βασική φύση του προβλήματος δεν αλλάζει άλλα αυτό που αλλάζει είναι μόνο οι περιορισμοί, δηλαδή τώρα έχουμε το εξής πρόβλημα μεγιστοποίησης:

$$\max_{\alpha}\left[\sum_{i=1}^{l}\alpha_i - \frac{1}{2}\alpha^T H\alpha\right]$$

με τους περιορισμούς

$$0 \le \alpha_i \le C$$

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

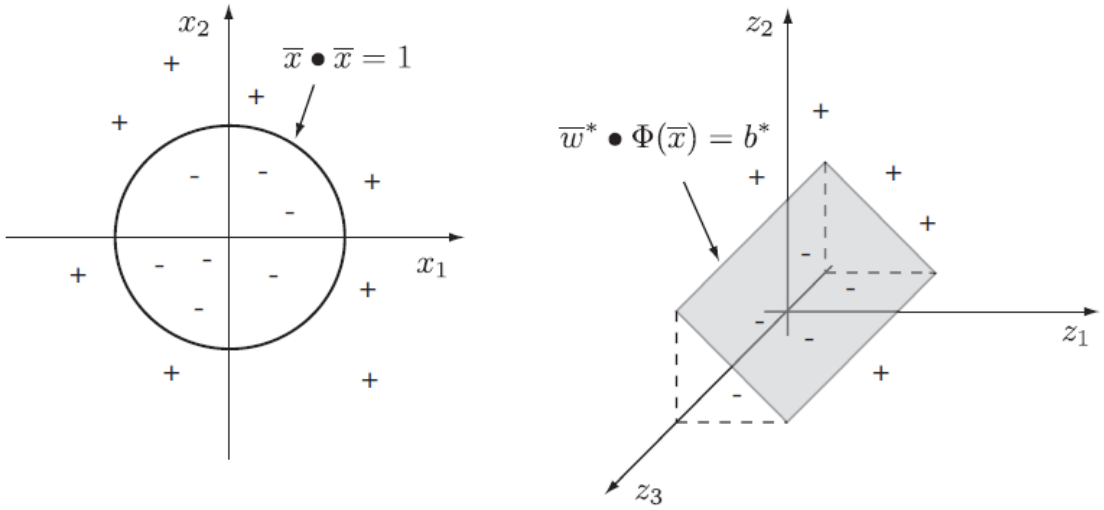
Τέλος είναι σημαντικό να σημειωθεί ότι, σε αυτές τις μη γραμμικά διαχωρίσιμες περιπτώσεις γίνεται ουσιαστικά μετατροπή του χώρου εισόδου, όπου το σύνολο των δεδομένων δεν είναι γραμμικά διαχωρίσιμο σε ένα χώρο υψηλότερων διαστάσεων που ονομάζεται χώρος των χαρακτηριστικών, όπου τα δεδομένα είναι γραμμικά διαχωρίσιμα. Οι συναρτήσεις που σχετίζονται με αυτούς τους μετασχηματισμούς ονομάζονται συναρτήσεις του πυρήνα (kernel functions), και η διαδικασία χρήσης αυτών των συναρτήσεων, για να περάσουμε από μια μη γραμμική σε μια γραμμική SVM ονομάζεται τέχνασμα του πυρήνα (kernel tricks), τα οποία θα αναπτύξουμε στην αμέσως επόμενη παράγραφο.

2.3 Συναρτήσεις Πυρήνα (Kernel Functions)

Η μέθοδος των συναρτήσεων πυρήνα είναι από τις πλέον δημοφιλείς και αποτελεσματικές στον τομέα της μάθησης. Το σκεπτικό τους βασίζεται στο τέχνασμα του πυρήνα (kernel tricks) το οποίο μπορεί να εφαρμοστεί σε κάθε γραμμικό αλγόριθμο που βασίζεται σε δεδομένα από την άποψη των εσωτερικών γινομένων μεταξύ των παρατηρήσεων.

Η ιδέα ακολουθεί ως εξής:

Έστω τα δεδομένα που καθορίζονται στο Σχήμα 2.2(α).



Σχήμα 2.2: Χαρτογράφηση ενός μη γραμμικού συνόλου δεδομένων (α) με $\overline{x} = (x_1, x_2) \in R^2$ σε ένα χώρο χαρακτηριστικών (β) με $\overline{z} = (x_1, x_2) \in R^3$, χρησιμοποιώντας τον μετασχηματισμό Φ: $R^2 \rightarrow R^3$.

Σε αυτή την περίπτωση προφανώς δεν υπάρχει επιφάνεια απόφασης της μορφής $\boldsymbol{w} \bullet \boldsymbol{x} = b$ που να διαχωρίζει τις δύο κλάσεις, χωρίς τυχόν λάθη. Σε αντίθεση, η μη γραμμική επιφάνεια απόφασης

$$\boldsymbol{x} \bullet \boldsymbol{x} = 1 \tag{2.3.1}$$

με $\boldsymbol{x} \in R^2$ διαχωρίζει το σύνολο των δεδομένων, όπως φαίνεται στο Σχήμα 2.2(α).

Ο κύριος στόχος είναι η κατασκευή μιας συνάρτηση απόφασης που στηρίζεται σε μια επιφάνεια απόφασης στο χώρο εισόδου, δηλαδή που χαρτογραφεί τα σημεία $\boldsymbol{x} \in R^2$ σε κάποιο υψηλότερης διαστάσεων χώρο με εσωτερικό γινόμενο, τον λεγόμενο R^3.

Με την χρήση του μετασχηματισμού Φ: $R^2 \rightarrow R^3$, δηλαδή με την χαρτογράφηση από ένα δισδιάστατο χώρο σε ένα τρισδιάστατο χώρο, οποιοδήποτε σημείο της μη γραμμικής επιφανείας απόφασης στο χώρο εισόδου αντιστοιχεί σε ένα επίπεδο στο χώρο των χαρακτηριστικών της μορφής

$$\boldsymbol{w}^* \bullet \Phi(\boldsymbol{x}) = b^* \tag{2.3.2}$$

Να σημειωθεί εδώ ότι όλα τα σημεία του χώρου εισόδου που επισημαίνονται με +1 θα χαρτογραφηθούν σε σημεία πάνω από αυτό το επίπεδο στο χώρο των

χαρακτηριστικών, και τυχόν σημεία που επισημαίνονται με -1 στο χώρο εισόδου θα πρέπει να αντιστοιχίζονται με τα σημεία κάτω από το επίπεδο. Το γεγονός ότι το επίπεδο (2.3.2) χωρίζει τις κλάσεις στον χώρο των χαρακτηριστικών σημαίνει ότι το επίπεδο είναι μια γραμμική επιφάνεια απόφασης. Αυτό απεικονίζεται στο Σχήμα 2.2(β). Αυτό σημαίνει ότι η χαρτογράφηση Φ μετασχηματίζει το μη γραμμικό πρόβλημα στο χώρο εισόδου σε ένα γραμμικό πρόβλημα στο χώρο των χαρακτηριστικών.

Με την χρήση της συνάρτησης Φ, αντί να λάβουμε μια συνάρτηση της οποίας η πολυπλοκότητα είναι ανάλογη με τις διαστάσεις του χώρου των χαρακτηριστικών, παίρνουμε μια έκφραση της οποίας η πολυπλοκότητα είναι ανάλογη με τον αριθμό των διανυσμάτων υποστήριξης.

Συναρτήσεις της μορφής

$$k(\boldsymbol{x}, \boldsymbol{y}) = \Phi(\boldsymbol{x}) \bullet \Phi(\boldsymbol{y}) \qquad (2.3.3)$$

όπου $\boldsymbol{x}, \boldsymbol{y} \in R^n$, ονομάζονται συναρτήσεις πυρήνα. Οι συναρτήσεις πυρήνα αξιολογούν ένα εσωτερικό γινόμενο στο χώρο των χαρακτηριστικών και το καθοριστικό χαρακτηριστικό του πυρήνα είναι ότι η τιμή του εσωτερικού γινομένου αυτού υπολογίζεται στην πραγματικότητα στο χώρο εισόδου.

Η συνάρτηση απόφασης που προκύπτει με την χρήση του μετασχηματισμού Φ και με αντικατάσταση των συναρτήσεων της μορφής (2.3.3) είναι γνωστή ως το τέχνασμα του πυρήνα. Δηλαδή, η χρήση οποιασδήποτε κατάλληλης συνάρτησης πυρήνα που μπορεί να επωφεληθεί από τις χαρτογραφήσεις στους χώρους των χαρακτηριστικών χωρίς να χρειάζεται να «πληρώσει» το τίμημα που πραγματικά απαιτείται για να υπολογιστή η ρητή χαρτογράφηση. Επιλέγοντας τη συνάρτηση πυρήνα, μπορούμε να ελέγξουμε την πολυπλοκότητα αυτού του μοντέλου. Το τέχνασμα έγκειται στην εύρεση του κατάλληλου πυρήνα, προκειμένου να κατασκευάσει ένα μοντέλο για ένα συγκεκριμένο σύνολο δεδομένων.

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

$$H_{ij} = y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) = y_i y_j \boldsymbol{x}_i \bullet \boldsymbol{x}_j$$

Ο $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ είναι ένα παράδειγμα μιας οικογένειας συναρτήσεων που είναι γνωστός ως γραμμικός πυρήνας.

Άλλοι δημοφιλείς πυρήνες για ταξινόμηση είναι:

- Πυρήνας ακτινικής βάσης (Gaussian Radial Basis Kernel)

$$K(\boldsymbol{x}, \boldsymbol{y}) = e^{-\left(\frac{\|\boldsymbol{x}-\boldsymbol{y}\|^2}{2\sigma^2}\right)}$$

- Σιγμοειδής πυρήνας (Neural Kernel)

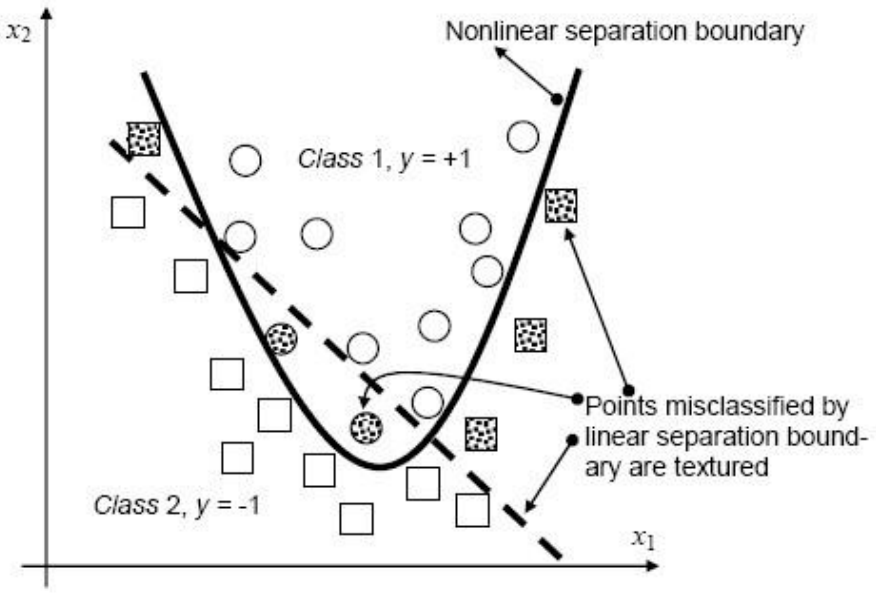
$$K(\boldsymbol{x}, \boldsymbol{y}) = tanh(\alpha \boldsymbol{x} \bullet \boldsymbol{y} - b)$$

όπου α και b είναι οι παράμετροι που καθορίζουν την συμπεριφορά του πυρήνα

- Πολυωνυμικός πυρήνας (Polynomial Kernle)

$$K(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x} \bullet \boldsymbol{y} + \alpha)^b$$

όπου α και b είναι οι παράμετροι που καθορίζουν την συμπεριφορά του πυρήνα.



Σχήμα 2.3: Μη γραμμικό διαχωριστικό σύνορο με τη βοήθεια της μεθόδου των πυρήνων. Χρησιμοποιήθηκε ο πολυωνυμικός πυρήνας.

2.4 Μέθοδοι επιλογής μοντέλου/παραμέτρων για τις Μηχανές Διανυσμάτων Υποστήριξης

Η απόδοση των μηχανών διανυσμάτων υποστήριξης (SVM) επηρεάζεται σημαντικά από τις παραμέτρους του μοντέλου. Μια κοινώς χρησιμοποιούμενη μέθοδος επιλογής παραμέτρων SVM, είναι το πλέγμα αναζήτησης (GS), η οποία είναι πολύ χρονοβόρα.

Η αναζήτηση πλέγματος (grid search) αναφέρεται σε μία εξαντλητική αναζήτηση μέσω ενός υποσυνόλου του παραμετρικού (hyperparameter) χώρου του αλγορίθμου μάθησης για να λύσει το πρόβλημα της επιλογής μοντέλου ή της βελτιστοποίησης των παραμέτρων.

Εφόσον σε αυτή τη μέθοδο ταξινόμησης έχουμε παραγωγή ενός ακριβούς και ταυτόχρονα σύντομου σε εκτέλεση μοντέλου, κατά καιρούς αρκετές επιστημονικές προτάσεις έχουν προταθεί για την μείωση του υπολογιστικού χρόνου των SVM.

Οι Ou et al. (2003) πρότειναν ένα μηχανισμό για τη μείωση των δεδομένων με σκοπό την επισκευή της διαδικασίας επιλογής μοντέλου στην SVM μέθοδο. Τα πειραματικά αποτελέσματα δείχνουν ότι ο προτεινόμενος μηχανισμός είναι σε θέση να μειώσει σημαντικά το χρόνο για να πραγματοποιηθεί η επιλογή μοντέλου με το ελάχιστο κόστος.

Στην εργασία των G.Lebrun et. al (2006) προτείνεται μια νέα μέθοδος μάθησης για την κατασκευή μιας δίτιμης συνάρτησης αποφάσεων (Binary Decision function (BDF)) στις μηχανές διανυσμάτων υποστήριξης (SVMs) μειώνοντας την πολυπλοκότητα και καθιστώντας αποτελεσματική τη γενίκευση. Στόχος είναι η κατασκευή ενός γρήγορου και αποτελεσματικού SVM ταξινομητή. Ορίζεται ένα κριτήριο για την αξιολόγηση της ποιότητας της συνάρτησης αποφάσεων (DFQ, Decision function Quality), η οποία λαμβάνει υπ' όψιν το ποσοστό αναγνώρισης και την πολυπλοκότητα της BDF. Η επιλογή μοντέλου γίνεται με βάση την επιλογή του απλούστερου επιπέδου, ενός υποσυνόλου χαρακτηριστικών και των παραμέτρων του SVM (hyperparameters) και εκτελείται για την βελτιστοποίηση της DFQ.

Οι Hwang, et al. (2007) πρότειναν ένα ομοιόμορφό σχεδιασμό μεθοδολογιών (UD, uniform design) για την αποτελεσματική, ισχυρή και αυτόματη επιλογή μοντέλου

για τις μηχανές διανυσμάτων υποστήριξης. Η προτεινόμενη μέθοδος εφαρμόζεται για να επιλεγεί το σύνολο των υποψήφιων παραμέτρων και εκτελείται μια k-fold διασταυρωμένη επικύρωση (cross-validation) για να αξιολογηθεί η γενικευμένη απόδοση του κάθε συνδυασμού παραμέτρων.

Γενικά, η κακή επιλογή των ρυθμιστικών παραμέτρων μπορεί να μειώσει δραματικά την απόδοση των SVMs. Θα ήταν επιθυμητό να έχουμε ένα αποτελεσματικό και αυτόματο σύστημα επιλογής μοντέλου κάνοντας έτσι τα SVMs πρακτικά σε εφαρμογές της πραγματικής ζωής, ιδιαίτερα, για τους ανθρώπους που δεν είναι εξοικειωμένοι με τις παραμέτρους (parameters tuning) στα SVMs.

2.5 Ταξινόμηση στις SVMs για προβλήματα πολλαπλών κατηγοριών (Multiclass Classification)

Το πρότυπο της θεωρίας των μηχανών διανυσμάτων υποστήριξης υποστηρίζει μόνο δυαδικά προβλήματα ταξινόμησης. Ωστόσο, πολλά προβλήματα του πραγματικού κόσμου ασχολούνται με την ταξινόμηση των αντικειμένων σε περισσότερες από δυο κλάσεις. Για παράδειγμα, η χειρόγραφη ψηφιακή αναγνώριση θεωρεί σύνολα δέκα κλάσεων, π.χ. ψηφία 0 έως 9. Υπάρχουν πολλοί τρόποι να επεκτείνουμε τα SVMs για τέτοιες περιπτώσεις. Στη συνέχεια, θα εξεταστούν 2 τέτοιες μέθοδοι με τις οποίες μπορούμε να χρησιμοποιήσουμε δυαδικές μηχανές διανυσμάτων υποστήριξης (binary support vector machines) για την ταξινόμηση των αντικειμένων σε ένα αυθαίρετο αριθμό κατηγοριών.

2.5.1 Σύστημα ταξινόμησης «ένας-εναντίον-των-υπολοίπων» (One-against-the-rest Classification)

Μέχρι στιγμής η ταξινόμηση one-against-the-rest είναι η πιο δημοφιλής τεχνική για ταξινόμηση πολλαπλών κατηγοριών, χρησιμοποιώντας την δυαδική μηχανή διανυσμάτων υποστήριξης.

Το σύστημα ταξινόμησης one-against-the-rest απαιτεί την κατασκευή ενός αριθμού μηχανών διανυσμάτων υποστήριξης όσες είναι και οι κλάσεις του προβλήματος ταξινόμησης.

Έστω το σύνολο εκπαίδευσης

$$D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \dots, (\boldsymbol{x}_l, y_l)\} \subset R^n \times \{1, \dots, M\}$$

όπου y_i είναι η ετικέτα για κάθε παρατήρηση και μπορεί να πάρει οποιαδήποτε τιμή στο {1, 2,…,M} με M>2.

Στην τεχνική του one-against-the-rest, δεδομένου M κατηγοριών/κλάσεων, κατασκευάζουμε M δυαδικά, διανύσματα-βάσης υποστήριξη (support vector-based) των επιφανειών απόφασης, $g^1, \dots, g^M$. Κάθε επιφάνεια απόφασης έχει εκπαιδευτεί να διαχωρίζει μία κατηγορία από τις υπόλοιπες. Δηλαδή, η επιφάνεια απόφασης g^1 έχει εκπαιδευτεί να διαχωρίζει την κατηγορία με την ένδειξη 1 από όλες τις άλλες κατηγορίες, η επιφάνεια απόφαση g^2 έχει εκπαιδευτεί να διαχωρίζει την κατηγορία με την ένδειξη 2 από όλες τις άλλες κατηγορίες, και ούτω καθεξής. Για να ταξινομηθεί ένα άγνωστο σημείο χρησιμοποιούμε ένα σύστημα ψηφοφορίας (voting scheme) βάσει ποιων από τις επιφάνειες απόφασης M επιστρέφουν τη μεγαλύτερη τιμή για αυτό το άγνωστο σημείο. Στη συνέχεια, χρησιμοποιούν την επιφάνεια απόφασης που επιστρέφει τη μεγαλύτερη τιμή για το άγνωστο σημείο για να εκχωρήσει αυτό το σημείο σε μια κατηγορία.

Για να εκπαιδεύσουμε τις επιφάνειες απόφασης κατασκευάζουμε M δυαδικά σύνολα εκπαίδευσης

$$D^p = D_+^p \cup D_-^p$$

όπου

$$D_+^p = \{(\boldsymbol{x}, +1) | (\boldsymbol{x}, y) \in D \; με \; y = p\}$$

και

$$D_-^p = \{(\boldsymbol{x}, -1) | (\boldsymbol{x}, y) \in D \; με \; y \neq p\}$$

με p=1,...,M. Το σύνολο D_+^p περιέχει όλες τις παρατηρήσεις του D που είναι μέλη της κατηγορίας p, και το σύνολο D_-^p περιέχει όλες τις υπόλοιπες παρατηρήσεις. Στο σύνολο εκπαίδευσης D^p έχουμε ετικέτες {+1,-1}. Η ετικέτα +1 χρησιμοποιείται για τις παρατηρήσεις στην κατηγορία p, και η ετικέτα -1 χρησιμοποιείται για τις παρατηρήσεις που δεν είναι στην κατηγορία p.

Στην συνέχεια εκπαιδεύουμε κάθε επιφάνεια απόφασης g^p για τα αντίστοιχα δεδομένα συνόλου D^p . Η επιφάνεια απόφασης $g^p: R^n \to R$ επιστρέφει ένα πρόσημο πραγματικής τιμής που μπορεί να ερμηνευθεί ως η απόσταση από κάποιο σημείο $\boldsymbol{x} \in R^n$ στην επιφάνεια απόφασης. Εάν η τιμή που επιστρέφεται είναι θετική, το σημείο **x** είναι πάνω από την επιφάνεια απόφασης και θεωρείται ότι είναι ένα μέλος της κλάσης +1 σε σχέση με την επιφάνεια απόφασης και, εάν η τιμή που επιστρέφεται είναι αρνητική, το σημείο είναι κάτω από την επιφάνεια απόφασης και θεωρείται ότι είναι ένα μέλος της κλάσης -1 σε σχέση με την επιφάνεια απόφαση. Μπορούμε επίσης να ερμηνεύσουμε την τιμή που επιστρέφεται ως τιμή εμπιστοσύνης, δηλαδή όσο μεγαλύτερη είναι η τιμή που επιστρέφεται από μια επιφάνεια απόφασης για κάποιο σημείο, τόσο πιο σίγουροι είμαστε ότι αυτό το σημείο ανήκει στην κατηγορία +1 σε σχέση με αυτή την επιφάνεια απόφασης. Αυτό σημαίνει ότι αν η επιφάνεια απόφασής μας επιστρέφει μια μεγάλη αρνητική τιμή για κάποιο σημείο, είμαστε σίγουροι ότι το σημείο αυτό ανήκει στην κατηγορία -1. Εάν, από την άλλη πλευρά, η επιφάνεια απόφασης επιστρέφει μια μεγάλη θετική τιμή, είμαστε πολύ σίγουροι ότι το σημείο ανήκει στην κατηγορία +1.

Υποθέτουμε τώρα ότι στο σύνολο εκπαίδευσης D^p για την επιφάνεια απόφασης g^p όλες οι παρατηρήσεις στην κατηγορία p είναι θετικά παραδείγματα, δηλαδή, $(\boldsymbol{x}_i, p) \in D$ και άρα $(\boldsymbol{x}, +1) \in D_+^p$, τότε προκύπτει ότι μια επιφάνεια απόφασης g^p που επιστρέφει τη μεγαλύτερη τιμή για κάποιο σημείο $\boldsymbol{x}$ μεταξύ όλων των άλλων επιφανειών απόφαση $g^1, \dots, g^M$ εκχωρεί αυτό το σημείο στην τάξη m με $m \in \{1, \dots, M\}$.

Με βάση αυτά κατασκευάζουμε μια συνάρτηση απόφασης $\hat{f}: N \to \{1, \dots, M\}$ για πρόβλημα ταξινόμησης πολλαπλών κατηγοριών ως εξής:

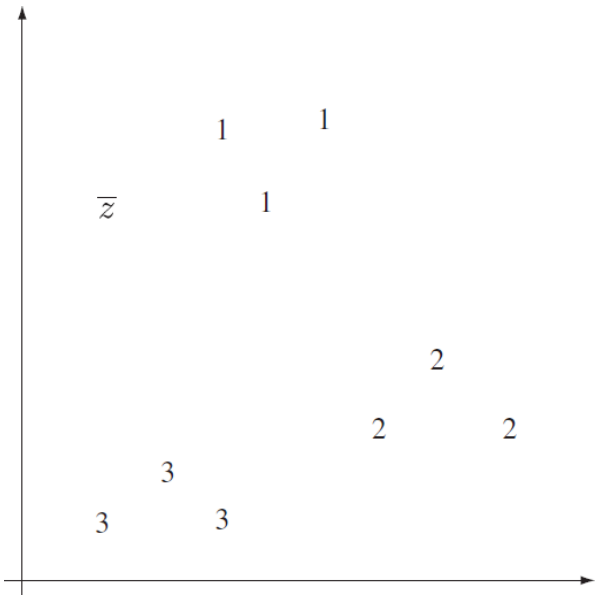
$$\hat{f}(\boldsymbol{x}) = arg \max_p g^p(\boldsymbol{p})$$

όπου $p \in \{1, \dots, M\}$. Η συνάρτηση απόφασης επιστρέφει την ετικέτα της επιφάνειας απόφασης που αναθέτει κάποιο σημείο $\boldsymbol{x} \in R^n$ στην +1 κατηγορία με την υψηλότερη εμπιστοσύνη. Για να γίνει πιο αντιληπτή η λειτουργία αυτής της τεχνικής παραθέτουμε το κάτωθι παράδειγμα:

Παράδειγμα: Έστω ένα πρόβλημα ταξινόμησης με τρεις τάξεις, όπου το σύνολο εκπαίδευσης D ορίζεται ως

$$D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \dots, (\boldsymbol{x}_l, y_l)\} \subset R^2 \times \{1,2,3\}$$

με $l = 9$.



Σχήμα 2.4: Μια γραφική αναπαράσταση του συνόλου δεδομένων, όπου κάθε παρατήρηση αντιπροσωπεύεται από την αντίστοιχη ετικέτα. Αποτελεί ένα πρόβλημα πολλαπλής κατηγορία ταξινόμησης με τρεις διακριτές κατηγορίες των παρατηρήσεων, το καθένα επισημασμένο με την κατάλληλη ετικέτα. Στο Σχήμα βλέπουμε επίσης ένα σημείο $\bar{z}$ που θα θέλουμε να ταξινομηθεί.

Τότε τα τρία σύνολα εκπαίδευσης έχουν ως εξής:

$$D^1 = D^1_+ \cup D^1_-$$

$$D^2 = D^2_+ \cup D^2_-$$

$$D^3 = D^3_+ \cup D^3_-$$

όπου

$$D^1_+ = \{(\boldsymbol{x}, +1) | (\boldsymbol{x}, y) \in D \; με \; y = 1\}$$

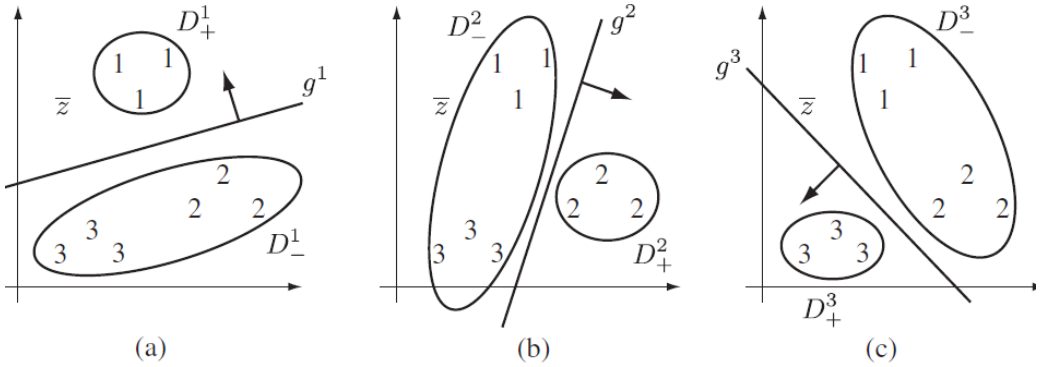
$$D^1_- = \{(\boldsymbol{x}, -1) | (\boldsymbol{x}, y) \in D \; με \; y \neq 1\}$$

$$D^2_+ = \{(\boldsymbol{x}, +1) | (\boldsymbol{x}, y) \in D \; με \; y = 2\}$$

$$D^2_- = \{(\boldsymbol{x}, -1) | (\boldsymbol{x}, y) \in D \; με \; y \neq 2\}$$

$$D^3_+ = \{(\boldsymbol{x}, +1) | (\boldsymbol{x}, y) \in D \; με \; y = 3\}$$

$$D^3_- = \{(\boldsymbol{x}, +1) | (\boldsymbol{x}, y) \in D \; με \; y \neq 3\}$$



Σχήμα 2.5: Η εκπαίδευση των επιφανειών απόφασης, g^1, g^2 και g^3, για τα αντίστοιχα σύνολα δεδομένων. Το τμήμα (α) δείχνει την επιφάνεια απόφασης g^p με την υπόθεση ότι p=1, το τμήμα (b) για p=2, και το τμήμα (c) για p=3. Παρατηρούμε ότι το κανονικό διάνυσμα για κάθε επιφάνεια απόφασης δείχνει πάντα προς την αντίστοιχη κατηγορία g^p που διαχωρίζεται από τα υπόλοιπα. Δηλαδή, τα σημεία που ανήκουν σε αυτή την κατηγορία θεωρούνται πάντα να είναι πάνω από την επιφάνεια απόφασης. Εδώ το σημείο z (που ισοδυναμεί με το διάνυσμα του z) ανήκει στην κατηγορία 1, επειδή η επιφάνεια απόφασης g^1 επιστρέφει τη μεγαλύτερη τιμή για αυτό το σημείο.

Θα κατασκευάσουμε τώρα τη συνάρτηση απόφαση $\hat{f}: R^2 \rightarrow \{1,2,3\}$

$$\hat{f}(\boldsymbol{x}) = arg\max_p g^p(\boldsymbol{p})$$

με p =1,2,3 και $\boldsymbol{x} \in R^2$. Αν εφαρμόσουμε αυτή τη συνάρτηση απόφασης στο σημείο **z** από το Σχήμα 2.5, βλέπουμε ότι $\hat{f}(z) \rightarrow 1$, (όπου $\boldsymbol{z}$=$\bar{z}$) επειδή η επιφάνεια απόφασης g^1 επιστρέφει τη μεγαλύτερη τιμή για το σημείο αυτό. Αυτό επαληθεύεται εύκολα από το Σχήμα 2.5 μιας και το σημείο $\boldsymbol{z}$ βρίσκεται πλησιέστερα προς τα σημεία της κατηγορίας 1 και, συνεπώς, θα πρέπει να εκχωρηθεί στην εν λόγω κατηγορία.

2.5.2 Σύστημα ταξινόμησης «ένας-εναντίον-ενός» (One-against-one Classification)

Αν και η τεχνική ταξινόμησης one-against-the-rest έχει αποδειχθεί ότι είναι ισχυρή σε πραγματικές εφαρμογές, εμφανίζει δύο σημαντικά προβλήματα. Το πρώτο πρόβλημα είναι η ομαδοποίηση κατηγοριών που κάνουμε σε κάθε ένα από τα M δυαδικά προβλήματα μπορεί να οδηγήσει σε ασυνεπή αποτελέσματα διότι επιτρέπει στις παρατηρήσεις να κατατάσσονται ταυτόχρονα σε διάφορες κατηγορίες. Το δεύτερο πρόβλημα εντοπίζεται στο ότι τα σύνολα εκπαίδευσης που χρησιμοποιούνται σε κάθε δυαδικό πρόβλημα δεν είναι «ισορροπημένα».

Για να διορθωθεί το θέμα που προκύπτει με την ισορροπία των συνόλων εκπαίδευσης, οι Lee et al. (2001) πρότειναν μια παραλλαγή της τεχνικής one-against-the-rest την λεγόμενη one-against-one.

Η τεχνική ταξινόμησης one-against-one αποφεύγει αυτή την κατάσταση με την κατασκευή επιφανειών απόφασης για κάθε ζεύγος των κατηγοριών. Η ταξινόμηση ενός άγνωστου σημείου επιτυγχάνεται και σε αυτή την περίπτωση με ένα σύστημα ψηφοφορίας (voting scheme).

Έστω ένα πρόβλημα ταξινόμησης με το σύνολο εκπαίδευσης

$$D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \dots, (\boldsymbol{x}_l, y_l)\} \subset R^n \times \{1, \dots, M\}.$$

Στην τεχνική ταξινόμησης one-against-one θα πρέπει να κατασκευάσουμε $M(M-1)/2$ επιφάνειες απόφασης, μία επιφάνεια απόφασης για κάθε δυνατό ζεύγος των κατηγοριών. Η επιφάνεια απόφασης που χωρίζει το ζεύγος των κατηγοριών p και q με $p \neq q$ και $\{p, q\} \subset \{1, .., M\}$ ορίζεται ως

$$g^{p,q}: R^n \to \{p, q\}$$

Εκπαιδεύουμε κάθε επιφάνεια απόφασης $g^{p,q}$ στο σύνολο των δεδομένων

$$D^{p,q} = D^p \cup D^q$$

όπου

$$D^p = \{(\boldsymbol{x}, y) | (\boldsymbol{x}, y) \in D \ με \ y = p\}$$

και

$$D^q = \{(\boldsymbol{x}, y) | (\boldsymbol{x}, y) \in D \ με \ y = q\}.$$

Το σύνολο D^p αποτελείται από το σύνολο των παρατηρήσεων του D με την ετικέτα p και το συνόλου D^q αποτελείται από το σύνολο των παρατηρήσεων του D με την ετικέτα q. Η εκπαίδευση του συνόλου $D^{p,q}$ για το ζεύγος των κατηγοριών p και q είναι απλά η ένωση αυτών των δύο συνόλων.

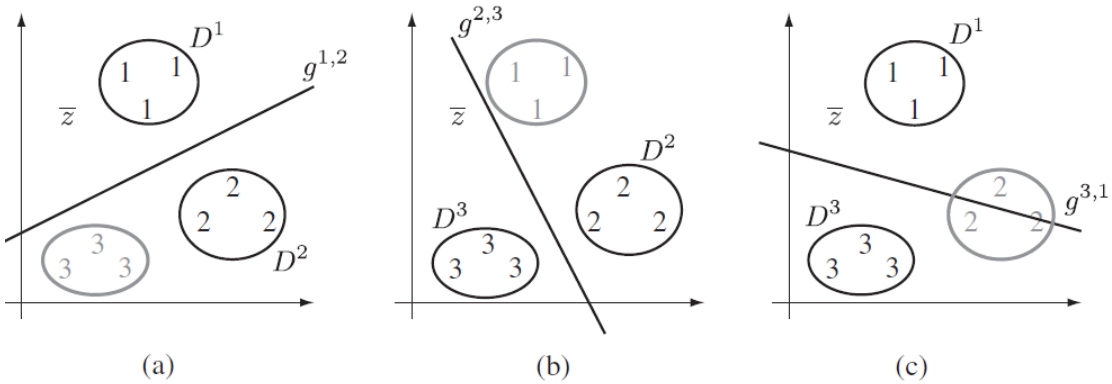
Μόλις κατασκευάσουμε όλες τις επιφάνειες απόφασης ανά ζεύγη $g^{p,q}$, χρησιμοποιώντας τα αντίστοιχα σύνολα εκπαίδευσης $D^{p,q}$, μπορούμε να ταξινομήσουμε ένα άγνωστο σημείο με την εφαρμογή κάθε M(M-1)/2 επιφάνειας απόφασης σε αυτό το σημείο, εντοπίζοντας πόσες φορές το άγνωστο σημείο εκχωρήθηκε και σε ποια ετικέτα κατηγορίας. Η ετικέτα κατηγορίας με την

υψηλότερη μέτρηση θεωρείται τότε η ετικέτα για το άγνωστο σημείο. Για να γίνουν όλα αυτά πιο αντιληπτά δίνεται το εξής παράδειγμα.

Παράδειγμα: Έστω ένα πολλαπλής κατηγορίας σύνολο δεδομένων D και ένα σημείο $\boldsymbol{z}$ του οποίου η ετικέτα είναι άγνωστη.

$$D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \dots, (\boldsymbol{x}_l, y_l)\} \subset R^2 \times \{1,2,3\}$$

με $\mathrm{l} = 9$.



Σχήμα 2.6: Εδώ παρουσιάζονται τα τρία σύνολα εκπαίδευσης $D^p \cup D^q$, η κατασκευή και η εφαρμογή των τριών ανά ζεύγη επιφανειών απόφασης $g^{p,q}$ με $\{g, q\} \subset \{1, 2, 3\}$. Το τμήμα (α) δείχνει την επιφάνεια απόφασης $g^{p,q}$ με την υπόθεση ότι p=1 και q=2, το τμήμα (b) για p=2 και q=3 και το τμήμα (c) για p=3 και q=1. Το σημείο z, (όπου z=$\overline{z}$) είναι το σημείο που θέλουμε να ταξινομηθεί, και εδώ όπως φαίνεται ανήκει στην κατηγορία 1, επειδή έχει την υψηλότερη μέτρηση.

Όπως παρατηρούμε και στο Σχήμα 2.6 στο τμήμα (α) βλέπουμε την επιφάνεια απόφασης $g^{1,2}$, όταν εφαρμόζεται για το σημείο $\boldsymbol{z}$, είναι σαφές ότι θα εκχωρηθεί στην κατηγορία 1. Στο τμήμα (b) κατασκευάζουμε επιφάνεια απόφασης $g^{2,3}$ και στη συνέχεια την εφαρμόζουμε για το σημείο $\boldsymbol{z}$. Σε αυτή την περίπτωση $\boldsymbol{z}$ έχει εκχωρηθεί στην κατηγορία 3. Τέλος, στο τμήμα (c) κατασκευάζουμε απόφασης επιφάνεια $g^{3,1}$. Όταν εφαρμόσουμε αυτήν την επιφάνεια απόφασης στο σημείο $\boldsymbol{z}$ αποδίδεται στην ετικέτα 1. Συνοψίζοντας τα πιο πάνω έχουμε

Κατηγορία 1	**Κατηγορία 2**	**Κατηγορία 3**
2	0	1

Με βάση τον πιο πάνω πίνακα εκχωρούμε στο σημείο ***z*** την ετικέτα 1 αφού η κατηγορία 1 έχει την μεγαλύτερη τιμή. Αυτό φαίνεται διαισθητικά, δεδομένου ότι το σημείο ***z*** βρίσκεται πλησιέστερα προς τα σημεία της κατηγορίας 1 όπως φαίνεται στο Σχήμα 2.6.

Στο σύστημα ψηφοφορίας της ταξινόμησης one-against-one υπάρχει η πιθανότητα της ισοπαλίας. Μπορούμε να αντιμετωπίσουμε την ισοπαλία ερμηνεύοντας τις πραγματικές τιμές που επιστρέφονται από τις επιφάνειες απόφασης ως τιμές εμπιστοσύνης (confidence values). Όταν προσθέτουμε τις απόλυτες τιμές των τιμών εμπιστοσύνης σε καθεμία από τις ετικέτες, θεωρούμε ότι η ετικέτα με το μεγαλύτερο άθροισμα των τιμών εμπιστοσύνη είναι η επικρατέστερη.

Φαίνεται ότι η ταξινόμηση one-against-one λύνει το πρόβλημα ασύμμετρων συνόλων δεδομένων. Ωστόσο, λύνει αυτό το πρόβλημα σε βάρος της εισαγωγής μιας νέας επιπλοκής, το γεγονός ότι για τις κατηγορίες Μ έχουμε να κατασκευάσουμε Μ(Μ-1)/2 επιφάνειες απόφασης. Για προβλήματα ταξινόμησης με ένα μικρό αριθμό κατηγοριών η διαφορά μεταξύ του αριθμού των επιφανειών απόφασης που πρέπει να οικοδομήσει το σύστημα ταξινόμησης one-against-the-rest και one-against-one δεν είναι και τόσο μεγάλη. Σκεφτείτε ένα πρόβλημα ταξινόμησης με Μ = 4. Εδώ θα έχουμε να κατασκευάσουμε τέσσερις επιφάνειες απόφασης για την ταξινόμηση one-against-the-rest και έξι επιφάνειες απόφασης στην ταξινόμησης one-against-one. Ωστόσο, κατά την εξέταση των προβλημάτων ταξινόμησης με έναν μεγάλο αριθμό διαφορετικών κατηγοριών, η διαφορά μπορεί να είναι αρκετά μεγάλη. Στην περίπτωση του Μ=10, θα πρέπει να κατασκευάσουμε 10 επιφάνειες απόφασης για την ταξινόμηση one-against-the-rest ενώ στην ταξινόμηση one-against-one θα πρέπει να κατασκευάσουμε 45 επιφάνειες απόφασης.

2.5.3 Εναλλακτικοί Μέθοδοι

Πέραν από τις δύο μεθόδους ταξινόμησης πολλαπλών κατηγοριών που έχουμε αναφέρει υπάρχουν δύο άλλες μέθοδοι που αναφέρονται συχνά στη βιβλιογραφία. Η πρώτη καλείται ταξινόμηση error-correcting output code και η δεύτερη καλείται multi-objective μηχανή διανυσμάτων υποστήριξης. Η ταξινόμηση error-correcting output code επεκτείνει την μέθοδο ταξινόμησης one-against-the-rest επιτρέποντας τον λεπτομερή έλεγχο του αριθμού των επιφανειών απόφασης που κατασκευάζονται και στη συνέχεια χρησιμοποιείται στην ταξινόμηση των αγνώστων σημείων. Η multi-objective μηχανή διανυσμάτων υποστήριξης επεκτείνει την θεωρία των μηχανών διανυσμάτων

υποστήριξης απευθείας από δυαδικά μοντέλα σε μοντέλα πολλαπλών κατηγοριών, με αποτέλεσμα ένα multi-objective πρόβλημα βελτιστοποίησης ως αλγόριθμο εκπαίδευσης. Και οι δύο αυτές προσεγγίσεις έχουν θεωρητικές ιδιότητες, αλλά δεν χρησιμοποιούνται συχνά στην πράξη, λόγω της αυξημένης υπολογιστικής πολυπλοκότητας τους.

Κεφάλαιο 3

ψ-Μάθηση

(ψ-Learning)

3.1 Εισαγωγή

Η έννοια του μέγιστου περιθωρίου έχει αναγνωριστεί ως μια σημαντική αρχή στην ανάλυση των μεθόδων μάθησης. Ωστόσο, η έννοια αυτή από μόνη της δεν είναι επαρκής για τη μάθηση σε μη διαχωρίσιμες περιπτώσεις. Από το προηγούμενο κεφάλαιο είδαμε ότι η θεωρία της SVM είναι καλά ανεπτυγμένη για τον διαχωρισμό των περιπτώσεων βάση της ιδέας του «σκληρού» περιθωρίου. Έτσι από τα θεμέλια της γίνεται λιγότερο σταθερή όταν επεκταθεί σε μη διαχωρίσιμες περιπτώσεις. Σε αυτές τις περιπτώσεις, όπου τα σφάλματα γενίκευσης γίνονται πολύ πιο σημαντικά, δεν είχαν ληφθεί πλήρως υπόψη κατά τη διαμόρφωση της SVM. Για την αντιμετώπιση αυτού του προβλήματος που παρουσιάζεται στις μη διαχωρίσιμες περιπτώσεις εισάγουμε μια νέα τεχνική μάθησης την λεγόμενη ψ-μάθηση. Διατηρώντας παράλληλα την ερμηνεία του

μέγιστου περιθωρίου, η ψ-μάθηση παρέχει βελτιωμένη απόδοση για τις μη διαχωρίσιμες περιπτώσεις ελέγχοντας κατάλληλα τα λάθη εκπαίδευσης.

Η τεχνική ψ-μάθηση προτάθηκε από τους Shen et al. (2003). Αποτελεί μια τεχνική άμεσης εξέτασης των σφαλμάτων γενίκευσης και αποδείχθηκε ότι στο πλαίσιο της ταξινόμησης αποδίδει μεγαλύτερη ακρίβεια γενίκευσης από την SVM. Στην μηχανική μάθηση η ψ-μάθηση είναι ένα μοντέλο μάθησης με επίβλεψη και ανήκει στην κατηγορία της ταξινόμησης μέγιστου περιθωρίου, όπως και η SVM. Κύριος στόχος της είναι η εύρεση ενός ταξινομητή, $(Sign(f))$ ο οποίος ελαχιστοποιεί τα σφάλματα γενίκευσης ή ισοδύναμα μεγιστοποιεί την ακρίβεια γενίκευσης, με αποτέλεσμα την καλύτερη ικανότητα γενίκευσης.

Εκτός από τα αριθμητικά της πλεονεκτήματα, η ψ-μάθηση έχει αποδεικτή πως πλεονεκτεί και σε θεωρητικό επίπεδο. Μια θεωρία έχει αναπτυχθεί για την ποσοτικοποίηση της ακρίβειας της μάθησης ως συνάρτηση του μεγέθους του δείγματος εκπαίδευσης και της κατηγορίας των υποψηφίων συναρτήσεων απόφασης. Η θεωρία αυτή, εξηγεί γιατί μια ψ-μάθηση αναμένεται να έχει υψηλότερη ακρίβεια έναντι της SVM, καθώς επίσης και ένα πρόσθετο πλεονέκτημα της, ότι αποκαλύπτει το trade-off μεταξύ της επιλογή μιας ρυθμιστικής παραμέτρου και του μεγέθους της τάξης της υποψηφίας συνάρτησης.

Παρά το γεγονός ότι η ψ-μάθηση έχει τη δυνατότητα να παράγει υψηλή απόδοση, η υπολογιστική πτυχή της απαιτεί ιδιαίτερη προσοχή, διότι η ελαχιστοποίηση της που προκύπτει είναι ένα μη κυρτό πρόβλημα ελαχιστοποίησης.

3.2 ψ-Μάθηση για δυαδικά προβλήματα ταξινόμηση και μη κυρτή ελαχιστοποίηση

Στα δυαδικά προβλήματα ταξινόμησης της SVM είχαμε:

Α) Για την διαχωρίσιμη περίπτωση

$$\min_{\boldsymbol{w}} \frac{1}{2} \|\boldsymbol{w}\|^2$$

με τους περιορισμούς

$$y_i(\boldsymbol{w} \bullet \boldsymbol{x}_i - b) - 1 \geq 0$$

Β) Για την μη διαχωρίσιμη περίπτωση

$$\min_{\boldsymbol{w},\boldsymbol{\xi},b} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n}\xi_i$$

με τους περιορισμούς

$$y_i(\boldsymbol{w} \bullet \boldsymbol{x}_i - b) + \xi_i - 1 \geq 0$$

$$\xi_i \geq 0.$$

Από τις συνθήκες Karush-Kuhn-Tucket (KKT) της θεωρία της βελτιστοποίησης η λύση της SVM πληροί τους εξής περιορισμούς

$$\xi_i = (1 - y_i(f(x_i)) \geq 0$$

ή

$$\xi_i = 0$$

εάν $1 - y_i f(x_i) < 0$

για i=1,...,n όπου $f(x_i) = \boldsymbol{w} \bullet \boldsymbol{x}_i - b.$

Έτσι αποδίδεται μια ισοδύναμη μορφή της SVM χωρίς περιορισμούς που έχει ως εξής:

$$\min_{b,\boldsymbol{w}} \left(\frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n}\psi_{SVM}(y_i(f(x_i))\right) \qquad (3.2.1)$$

όπου

$$\psi_{SVM}\big(y_i(f(x_i)\big) = \begin{cases} \psi_{SVM}\big(y_i(f(x_i)\big) = 0 & \text{εάν } y_i(f(x_i) \geq 1 \\ \psi_{SVM}(y_i(f(x_i)) = 1 - y_i(f(x_i) & \text{διαφορετικά} \end{cases}$$

Επίσης, C>0 είναι μια ρυθμιστική παράμετρος που ελέγχει την ισορροπία μεταξύ των δεδομένων προσαρμογής, $\frac{1}{2}\|\boldsymbol{w}\|^2$ είναι η ποινή και ψ_{SVM} είναι η συνάρτηση απώλειας των μηχανών διανυσμάτων υποστήριξης (hinge loss function).

Εδώ να σημειωθεί ότι η ρυθμιστική παράμετρος εξασφαλίζει ότι το μοντέλο έχει μια καλή προσαρμογή στα δεδομένα εκπαίδευσης, ο όρος ποινής αποφεύγει την

υπερπροσαρμογή του μοντέλου που προκύπτει και η «hinge loss» συνάρτηση των SVMs χρησιμοποιείται για την μεγιστοποίηση του περιθωρίου ταξινόμησης.

Το πιο φυσικό μέτρο της προσαρμογής των δεδομένων είναι το σφάλμα της ταξινόμησης που βασίζεται σε 0-1 απώλεια (0-1 loss) ή ισοδύναμα στην συνάρτηση $1 - \text{Sign}$ για τα δεδομένα εκπαίδευσης. Ωστόσο η βελτιστοποίηση της $1 - \text{Sign}$ είναι πολύ δύσκολη και για αυτό το λόγο αντικαθιστάται από κυρτές απώλειες όπου στην συγκεκριμένη περίπτωση είναι η συνάρτηση απώλειας (hinge loss) ψ_{SVM}.

Ένας επιθυμητός ταξινομητής είναι αυτός με την καλύτερη ικανότητα γενίκευσης, η οποία μετράται από το σφάλμα γενίκευσης (GE). Το σφάλμα γενίκευσης (generalization error) ορίζεται ως η πιθανότητα εσφαλμένης ταξινόμησης και γράφεται ως

$$Err(f) = P(Yf(X) < 0) = \frac{1}{2}E\left(1 - Sign\big(Yf(X)\big)\right)$$

ενώ

$$(2n)^{-1}\sum_{i=1}^{n}(1 - Sign(Y_i f(X_i)))$$

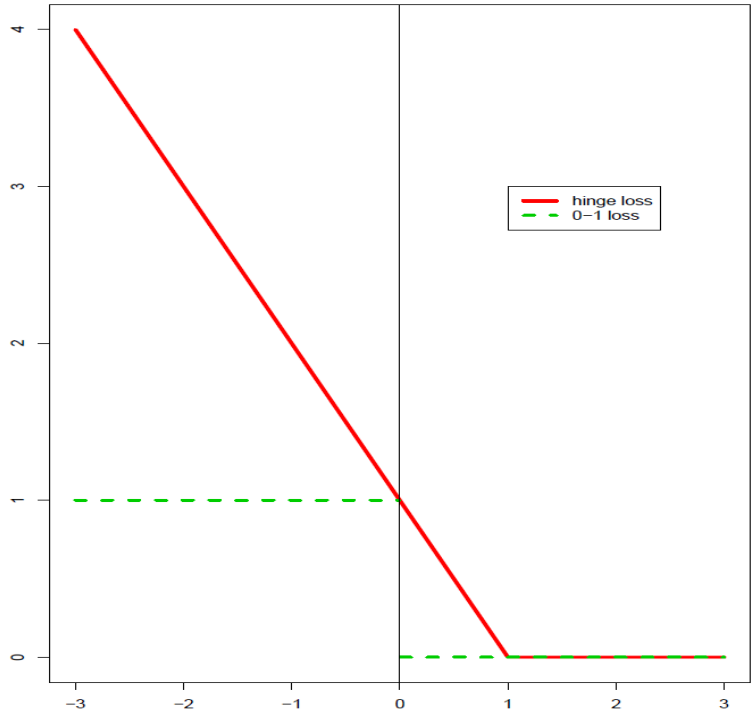
αποτελεί το εμπειρικό σφάλμα γενίκευσης (EGE) (empirical generalization error) όπου $Sign(f)$ καλείται «ταξινομητής».

Ένας ταξινομητής με μια συνάρτηση απώλειας ψ_{SVM} επιδιώκει να ελαχιστοποιήσει την GE μέσω της ψ_{SVM} . Στην περίπτωση της μηχανής διανυσμάτων υποστήριξης, όπως έχουμε ορίσει και πιο πάνω η $\psi_{SVM}(u)$ είναι ίση με $\psi_{SVM}(u) = [1 - u]_+$, που αποτελεί μια τμηματικά γραμμική κυρτή συνάρτηση, η οποία είναι το πιο πλησιέστερο δυνατό υποκατάστατο της 1-Sign όπως φαίνεται στο Σχήμα 3.1.

Ωστόσο επειδή η συνάρτηση Sign είναι αναλλοίωτης κλίμακας δηλαδή οποιοσδήποτε θετικός μετασχηματισμός αφήνει το πρόσημο της f αμετάβλητο, για την εξίσωση (3.2.1) με $\psi_{SVM}(u) = 1 - Sign(u)$, δηλαδή

$$\min_{b,\boldsymbol{w}}\left(\frac{1}{2}\|\boldsymbol{w}\|^2 + C\left(\sum_{i=1}^{n}1 - Sign(y_i(f(x_i))\right)\right)$$

προκύπτει το δεύτερο μέρος να είναι ίσο με μηδέν.



Σχήμα 3.1: Γραφική παράσταση που παρουσιάζει την συνάρτηση απώλεια των SVMs (hinge loss) και την συνάρτηση απώλειας 0-1 ή ισοδύναμα $1 - Sign$.

Για να εξαλείψουμε αυτό το πρόβλημα που παρουσιάζεται, εισάγουμε μια ομάδα από συναρτήσεις απώλειας ψ(x) οι οποίες ικανοποιούν τις εξής ιδιότητες:

$$\begin{cases} U \geq \psi(x) > 0 & \text{εάν } x \in (0, \tau] \\ \psi(x) = (1 - Sign(x)) & \text{διαφορετικά} \end{cases} \tag{3.2.2}$$

όπου $0 < \tau \leq 1$ και $U > 0$ είναι κάποιες σταθερές.

Έτσι οι θετικές τιμές της $\psi(x)$ εξαλείφουν το πρόβλημα της κλιμάκωσης της συνάρτησης Sign και αποφεύγουν την συσσώρευση πάρα πολλών σημείων γύρω από το όριο απόφασης. Για κάθε x_i υποδείγματα (instances) τέτοια ώστε $y_i(f(x_i)) \geq 0$ η ψ συνάρτηση ωθείται προς την κατεύθυνση $y_i(f(x_i)) \geq \tau$ λόγο του ότι η ψ εκχωρεί μια θετική ποινή σε οποιαδήποτε τιμή στο εύρος του $(0, \tau]$.

Έτσι για την περίπτωση γραμμικών προβλημάτων ταξινόμησης της τεχνική ψ-μάθηση προκύπτει η αντικειμενική συνάρτηση:

$$s(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n}\psi(y_i f(\boldsymbol{x}_i)) \tag{3.2.3}$$

όπου C>0 είναι μια παράμετρος συντονισμού η οποία θα πρέπει να εξαρτάται από το n και να επιλέγεται από τα δεδομένα στην πράξη.

Η γραμμική ταξινόμηση της ψ-μάθησης επιδιώκει την εύρεση των $(\boldsymbol{w}, b)$ που ελαχιστοποιούν την (3.2.3), δηλαδή

$$\min_{b,\boldsymbol{w}} s(\boldsymbol{w}) = \min_{b,\boldsymbol{w}} \left(\frac{1}{2} \|\boldsymbol{w}\|^2 + C \sum_{i=1}^{n} \psi(y_i f(\boldsymbol{x}_i)) \right) \qquad (3.2.4)$$

Έτσι από την ελαχιστοποίηση της SVM οδηγηθήκαμε στην ελαχιστοποίηση της τεχνικής ψ-μάθησης. Αξιοσημείωτο είναι το γεγονός ότι η βελτιστοποίηση που προέκυψε σε αυτή την περίπτωση είναι μια μη κυρτή ελαχιστοποίηση σε αντίθεση με αυτή της SVM που ήταν κυρτή.

Για μη γραμμική ταξινόμηση η αντικειμενική συνάρτηση που προκύπτει για την ψ-μάθηση είναι:

$$s(\boldsymbol{w}) = \frac{1}{2} \|g\|_K^2 + C \sum_{i=1}^{n} \psi(y_i f(\boldsymbol{x}_i)) \qquad (3.2.5)$$

όπου η συνάρτηση απόφασης $f(x) = g(x) + b$ με

$$g(x) = \sum_{i=1}^{n} w_i K(x, x_i)$$

και

$$\|g\|_K^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j K(x_i, x_j)$$

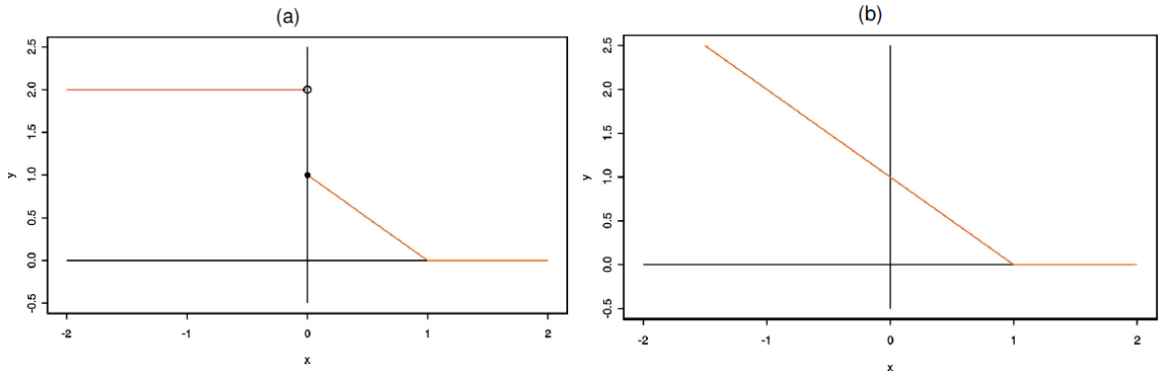
όπου $K(\cdot,\cdot)$ είναι ο γνωστός γραμμικός πυρήνας. Έτσι η αντίστοιχη ελαχιστοποίηση που προκύπτει είναι:

$$\min_{\boldsymbol{w},b} s(\boldsymbol{w}) = \min_{\boldsymbol{w},b} \left(\frac{1}{2} \|g\|_K^2 + C \sum_{i=1}^{n} \psi(y_i f(\boldsymbol{x}_i)) \right) \qquad (3.2.6)$$

Η βασική ιδέα για την επιλογή της συνάρτησης ψ είναι ότι θα πρέπει να είναι όσο το δυνατόν πιο κοντά στην 1-Sign. Μια επιλογή της συνάρτησης ψ είναι η λεγόμενη απλή γραμμική συνάρτηση $\psi_0(x)$ που ορίζεται να είναι:

$$\psi_0 = \begin{cases} 0 & \text{εάν } x \geq 1 \\ 1 - x & \text{εάν } 0 \leq x \leq 1 \\ 2 & \text{διαφορετικά} \end{cases}$$

Υπάρχουν και άλλες δυνατές επιλογές των ψ συναρτήσεων, όπως ψ(x)=2(1-x) για $0 \leq x \leq 1$.



Σχήμα 3.2: Στην γραφική παράσταση (a) βλέπουμε την ψ_0 συνάρτηση ενώ στην γραφική παράσταση (b) βλέπουμε την συνάρτηση ψ_{SVM}.

3.2.1 Ιδιότητες της ψ

Οι συναρτήσεις απώλειας ψ(x) που έχουμε εισαγάγει και που ικανοποιούν την (3.2.2) παρουσιάζουν δυο πολύ σημαντικές ιδιότητες οι οποίες μας δίνουν σημαντικές πληροφορίες σχετικά με την απόδοση της ψ-μάθησης.

Προτού αναπτύξουμε τις ιδιότητες αυτές είναι σημαντικό να κατανοήσουμε τον στόχο της ψ-μάθησης. Η ψ-μάθηση όπως έχουμε προαναφέρει επιδιώκει την εύρεση του ιδανικού ταξινομητή ο οποίος ελαχιστοποιεί τα σφάλματα γενίκευσης (GE). Ο ιδανικός βέλτιστος ταξινομητής που λαμβάνεται από την ελαχιστοποίηση $E(1 - Sign(Yf(X)))$ για κάθε f, είναι ο ταξινομητής του Bayes και ορίζεται ως εξής:

$$\bar{f} = Sign(f^*)$$

όπου $f^* = P(Y = 1|x) - 1/2$, είναι η συνάρτηση απόφασης του Bayes.

Έτσι η πρώτη ιδιότητα των ψ συναρτήσεων προκύπτει από το γεγονός ότι για κάθε ψ συνάρτηση που ικανοποιεί την (3.2.2) ο ταξινομητής Bayes $\bar{f}$ ελαχιστοποιεί τις $E\psi(Yf(X)$ και $E(1 - Sign(Yf(X))$ δηλαδή:

$$E\psi\big(Yf(X)\big) \geq E\psi\left(Y\bar{f}(X)\right) = E\left(1 - Sign\left(Y\bar{f}(X)\right)\right)$$
$$\leq E\left(1 - Sign\big(Yf(X)\big)\right).$$

Το πόρισμα αυτής της έκφρασης είναι ότι η μέθοδος ψ-μάθηση εκτιμά το ταξινομητή Bayes $\bar{f}$, αντί για την συνάρτηση απόφασης f^* του Bayes. Αυτό το χαρακτηριστικό της ψ είναι απαραίτητο, διότι ουσιαστικά η βέλτιστη απόδοση του ταξινομητή $\bar{f}$ πραγματοποιείται με τη χρήση της συνάρτησης ψ.

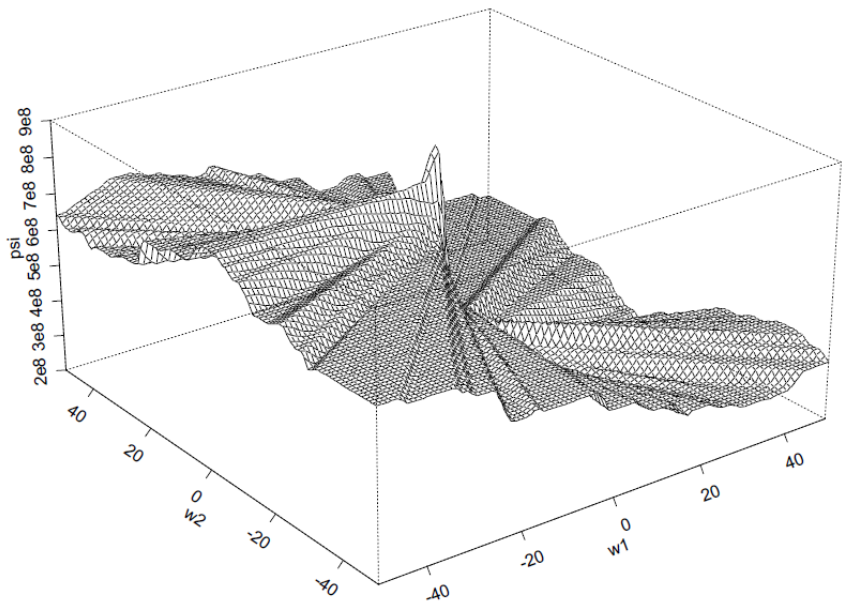
Η δεύτερη ιδιότητα προκύπτει από το γεγονός ότι στις διαχωρίσιμες περιπτώσεις (3.2.1) και (3.2.4) αποδίδεται η ίδια λύση όταν $C \to \infty$. Δηλαδή μια ψ-μάθηση για μεγάλες τιμές της παραμέτρου C είναι ισοδύναμη με το «σκληρό» περιθώριο της SVM. Όταν η ρυθμιστική παράμετρος C είναι μεγάλη, η ψ επιβάλλει "0" σφάλμα εκπαίδευσης σε διαχωρίσιμες περιπτώσεις, το οποίο δρα με ένα παράλληλο τρόπο με αυτόν της αντικειμενικής συνάρτησης του «σκληρού» περιθωρίου της SVM.

Συνοπτικά, η ψ-μάθηση και SVM βασίζονται σε διαφορετικές αρχές. Είναι ενδιαφέρον να σημειωθεί ότι $\psi_{SVM}(x) = \psi_0(x)$ όταν δεν υπάρχει σφάλμα εκπαίδευσης, όπως στις διαχωρίσιμες περιπτώσεις. Στις μη διαχωρίσιμες περιπτώσεις, το πλεονέκτημα της ψ επί της ψ_{SVM} είναι ότι βελτιώνει την ακρίβεια μάθησης, και το μειονέκτημα της είναι ότι η ελαχιστοποίηση που προκύπτει είναι μη κυρτή. Επίσης είναι πολύ σημαντικό το γεγονός ότι το κέρδος σε ποσοστά σφάλματος που προσφέρεται από τη χρήση των μη κυρτών αντικειμενικών συναρτήσεων είναι πολύ σημαντικό και δείχνει ότι η έρευνα στη ελαχιστοποίηση των συναρτήσεων της μορφής (3.2.3) θα πρέπει να αποτελεί υψηλή προτεραιότητα.

3.3 Υπολογιστικοί μέθοδοι για υψηλότερη ακρίβεια γενίκευσης

Η θεωρητική και αριθμητική ανάλυση έχει δείξει ότι η ψ-μάθηση έχει καλές ιδιότητες γενίκευση και μπορεί να εφαρμοστεί σε οποιοδήποτε πρόβλημα μάθησης για την ταξινόμηση των δεδομένων. Η λογική πίσω από την μεθοδολογία της είναι ότι η ταξινόμηση στη φύση της είναι ταξινόμηση μη κυρτών προβλημάτων και, τελικά θα πρέπει να αντιμετωπίζεται μέσω των μη κυρτών αντικειμενικών συναρτήσεων. Παρ' όλα αυτά, ένα σημαντικό πρακτικό πρόβλημα που προκύπτει είναι πώς να ανταποκριθούμε στην υπολογιστική πρόκληση της μη κυρτής ελαχιστοποίησης.

Στις ελαχιστοποιήσεις (3.2.4) και (3.2.6) στην γραμμική και μη γραμμική περίπτωση αντίστοιχα, που έχουν προκύψει είναι γενικά πολύ δύσκολο να εντοπίσουμε το ολικό ελάχιστο σε μια κατάσταση υψηλών διαστάσεων.



Σχήμα 3.3: Απεικονίζει το επίπεδο των δυσκολιών των ελαχιστοποιήσεων (3.2.4) και (3.2.6). Αποτελεί ένα διάγραμμα της εξίσωσης (3.2.3) ως συνάρτηση των $\boldsymbol{w} = (\boldsymbol{w_1}, \boldsymbol{w_2})$ και b=0,25 με n=50 και $\boldsymbol{C} = \mathbf{10^7}$.

Για την αντιμετώπιση αυτής της δυσκολίας εισάγουμε κάποιους αποδοτικούς αλγόριθμους όπου με την χρήση τους εξαλείφουμε την πολυπλοκότητα της λύσης της μη κυρτή ελαχιστοποίησης. Με την αξιοποίηση της ιδιότητα της τεχνικής της διαφοράς των κυρτών συναρτήσεων (Difference of Convex functions) (DC), θα αναπτύξουμε δυο υπολογιστικές στρατηγικές οι οποίες αποδίδουν μεγαλύτερη ακρίβεια γενίκευσης για την ψ-μάθηση στην πράξη και οι οποίες επεκτείνουν τις εφαρμογές της.

Η πρώτη στρατηγική χρησιμοποιεί τον αλγόριθμο της διαφοράς των κυρτών συναρτήσεων, (Differenced Convex Algorithm) που είναι γνωστή ως DCA (An και Tao, 1997), η οποία λύνει την μη κυρτή ελαχιστοποίηση μέσω μιας ακολουθίας τετραγωνικού προγραμματισμού (Sequential Quadratic Programming) (SQP). Η δεύτερη στρατηγική καθίσταται δυνατή μόνο με την χρήση των πολυωνυμικών αλγορίθμων (polynomial time algorithms) για πολυεδρικούς υπολογισμούς, οι οποίοι χρησιμοποιούν μια κατασκευή εξωτερικής προσέγγισης αποδίδοντας μια ολική ελαχιστοποίηση της ψ-μάθησης μέσω μιας ακολουθίας κυρτών

ελαχιστοποιήσεων που λαμβάνεται από την κορυφή απαρίθμησης (vertex enumeration).

Εδώ να σημειωθεί ότι στην τεχνική DC χρησιμοποιούμε μια συγκεκριμένη επιλογή της ψ συνάρτησης η οποία ορίζεται ως εξής:

$$\psi(x) = \begin{cases} 0 & \text{εάν } x \geq 1 \\ 2(1-x) & \text{εάν } 0 \leq x \leq 1 \\ 2 & \text{διαφορετικά} \end{cases}$$

3.3.1 D.C Αλγόριθμος ελαχιστοποίησης

Προκειμένου να εφαρμοστεί ο DCA, εφαρμόζουμε τον προγραμματισμό D.C ο οποίος χρησιμοποιεί μια ανάλυση της αντικειμενικής συνάρτησης σε μια διαφορά δύο κυρτών συναρτήσεων, βάση των οποίων οι προσεγγίσεις της αντικειμενικής συνάρτησης είναι κατασκευασμένες. Μια τέτοια ανάλυση διαδραματίζει εξαιρετικά σημαντικό ρόλο στον προσδιορισμό της ολικής αναζητούμενης λύσης, της ταχύτητας της σύγκλισης καθώς και της σταθερότητας.

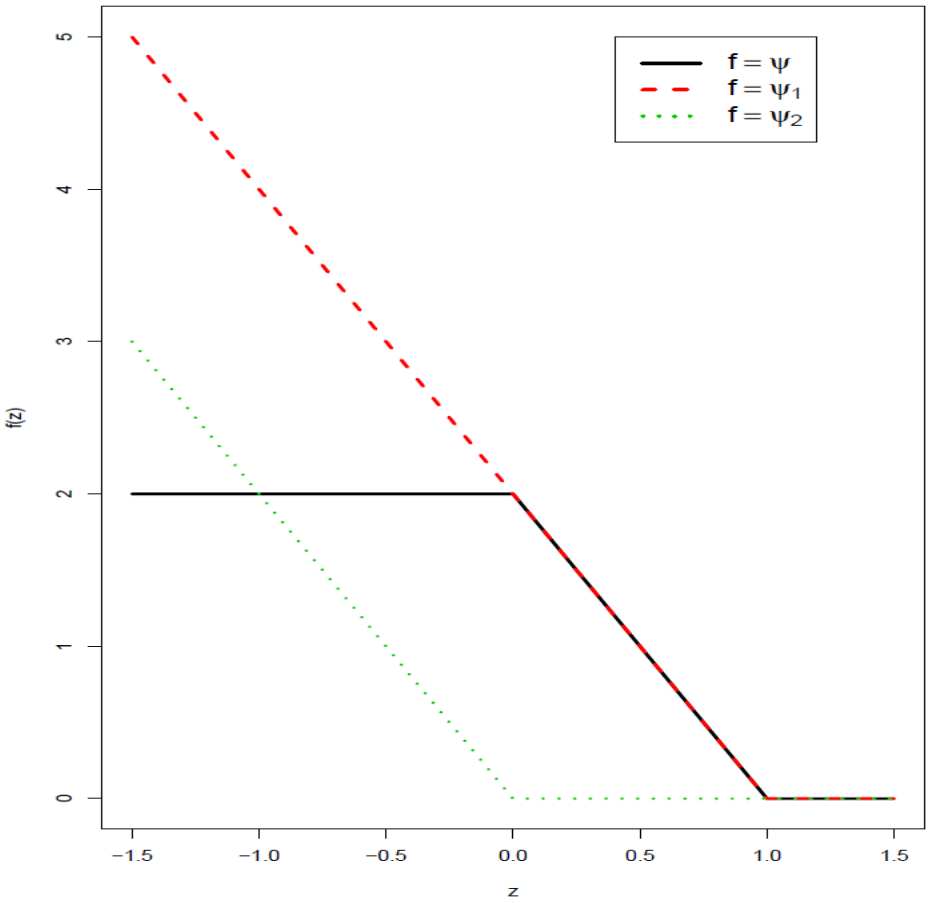
Για την επίλυση της (3.2.4) (κατά παρόμοιο τρόπο και της (3.2.6)), πρέπει πρώτα να αναλύσουμε την ψ συνάρτηση σε ψ_1 και ψ_2 δηλαδή, $\psi = \psi_1 - \psi_2$

όπου

$$\psi_1(x) = \begin{cases} 0 & \text{εάν } z \geq 1 \\ 2(1-z) & \text{διαφορετικά} \end{cases}$$

$$\psi_2(x) = \begin{cases} 0 & \text{εάν } z \geq 0 \\ -2z & \text{διαφορετικά.} \end{cases}$$

Τόσο η ψ_1 όσο και η ψ_2 είναι κυρτές.



Σχήμα 3.4: Γραφική παράσταση των συναρτήσεων ψ_1 και ψ_2 όπου $\psi = \psi_1 - \psi_2$ είναι μια D.C ανάλυση της ψ.

Αυτό αποδίδει μια D.C ανάλυση της αντικειμενικής συνάρτησης (3.2.3) σε:

$$s(\boldsymbol{w}) = s_1(\boldsymbol{w}) - s_2(\boldsymbol{w}) \tag{3.3.1}$$

όπου

$$s_1(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n}\psi_1(y_i f(\boldsymbol{x}_i)) \tag{3.3.2}$$

$$s_2(\boldsymbol{w}) = C\sum_{i=1}^{n}\psi_2(y_i f(\boldsymbol{x}_i)) \tag{3.3.3}$$

είναι και οι δύο κυρτές στο $\boldsymbol{w}$.

Έτσι με βάση αυτή την D.C ανάλυση η DCA κατασκευάζει μια μη αύξουσα άνω περιβάλλουσα της αντικειμενικής συνάρτησης η οποία αποδίδει μια ακολουθία από κυρτά υποπροβλήματα. Αυτό επιτρέπει την ανάπτυξη αποδοτικών αλγορίθμων για την ψ-μάθηση, ιδιαίτερα για προβλήματα μεγάλης κλίμακας. Υπάρχου δύο είδη της DCA, η κανονική και η απλοποιημένη. Εμείς εδώ θα αναφερθούμε μόνο στην απλοποιημένη περίπτωση.

Πριν προχωρήσουμε στην ανάλυση της DCA ορίζουμε την εκτίμηση της συνάρτησης απόφασης να είναι $\hat{f}(x) = \langle \widehat{\boldsymbol{w}}, x \rangle$ όπου $\widehat{\boldsymbol{w}}$ ελαχιστοποιεί τις εξισώσεις (3.2.3) και (3.2.5), ισοδύναμα είναι η λύση των (3.2.4) και (3.2.6). Εμείς επιθυμούμε την εξής ελαχιστοποίηση:

$$\min_{\boldsymbol{w}} s(\boldsymbol{w}) = \min_{\boldsymbol{w}} (s_1(\boldsymbol{w}) - s_2(\boldsymbol{w}))$$

Η βασική ιδέα είναι να κατασκευάσουμε μια ακολουθία προβλημάτων, τα οποία λαμβάνονται από την αντικατάσταση της $s_2(\boldsymbol{w})$ με βάση την συνάρτηση

$$s_2\left(\boldsymbol{w}^{(k)}\right) + \langle \boldsymbol{w} - \boldsymbol{w}^{(k)}, \nabla s_2\left(\boldsymbol{w}^{(k)}\right) \rangle$$

και την επίλυση της επαναληπτικά, όπου $\nabla s_2\left(\boldsymbol{w}^{(k)}\right)$ είναι η κλίση της $s_2(\boldsymbol{w})$ στο $\boldsymbol{w}^{(k)}$.

Έτσι αν ορίσουμε σαν λύση του k-οστού προβλήματος την $\left(\mathbf{w}^{(\mathrm{k})}, \nabla \mathrm{s}_2\left(\mathbf{w}^{(\mathrm{k})}\right)\right)$, τότε το (k+1) πρόβλημα ορίζεται ως εξής:

$$s_1(\boldsymbol{w}) - \left(s_2\left(\boldsymbol{w}^{(k)}\right) + \langle \boldsymbol{w} - \boldsymbol{w}^{(k)}, \nabla s_2\left(\boldsymbol{w}^{(k)}\right) \rangle\right).$$

Αυτά τα προβλήματα που προκύπτουν παρέχουν μια ακολουθία από άνω προσεγγίσεις του αρχικού προβλήματος ελαχιστοποίησης και έτσι οδηγούν στην σύγκλιση του $\boldsymbol{w}^{(k)}$ και άρα στην επιθυμητή λύση.

Πιο συγκεκριμένα η κλίση $\nabla s_2\left(\boldsymbol{w}^{(k)}\right)$ ορίζεται να είναι

$$V^{(k)} = \left(V_1^{(k)}, V_2^{(k)}\right)$$

όπου

$$V_1^{(k)} = C \sum_{i=1}^{n} \nabla \psi_2\left(y_i f^{(k)}(x_i)\right) y_i x_i \quad \text{ως προς } \boldsymbol{w},$$

$$V_2^{(k)} = C \sum_{i=1}^{n} \nabla \psi_2\left(y_i f^{(k)}(x_i)\right) y_i \quad \text{ως προς } b,$$

$$f^{(k)}(x_i) = \langle \boldsymbol{w}^{(k)}, x_i \rangle$$

και

$$\nabla \psi_2(x) = \begin{cases} 0 & \text{εάν } z \geq 0 \\ -2 & \text{διαφορετικά} \end{cases}$$

Έτσι ο αλγόριθμος μας λύνει μια ακολουθία από προβλήματα.

Στην (k+1) επανάληψη, αφού αγνοήσουμε τις σταθερές τιμές, το πρόβλημα που προκύπτει είναι ισοδύναμο με:

$$\min_{\boldsymbol{w}}\left(s_1(\boldsymbol{w}) - \langle \boldsymbol{w}, \nabla s_2\left(\boldsymbol{w}^{(k)}\right)\rangle\right)$$

Έτσι με αντικατάσταση της $s_1(\boldsymbol{w})$ και της $\nabla s_2\left(\boldsymbol{w}^{(k)}\right)$ τελικά το πρόβλημα παίρνει την εξής μορφή:

$$\min_{\boldsymbol{w}}\left(\frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n}\psi_1(y_i f(x_i)) - \langle \boldsymbol{w}, V_1^{(k)}\rangle - \langle b, V_2^{(k)}\rangle\right) \qquad (3.3.4)$$

Αυτό το πρόβλημα ελαχιστοποίησης λύνεται μέσω του τετραγωνικού προγραμματισμού (QP). Η διαδικασία που ακολουθούμε έχει ως εξής:

Στην (k+1) επανάληψη της SQP το πρόβλημα της (3.3.4) γράφεται ισοδύναμα χωρίς περιορισμούς ως εξής:

$$\min_{\boldsymbol{w},b}\left[\sum_{i=1}^{n}[2(1 - y_i f(x_i))]_+ + \frac{\lambda}{2}\langle \boldsymbol{w}, \boldsymbol{w}\rangle - \langle \boldsymbol{w}, V_1^{(k)}\rangle - \langle b, V_2^{(k)}\rangle\right]$$

όπου λ=1/C και $2(1 - y_i f(x_i))$ προκύπτει από τον ορισμό της συνάρτησης $\psi_1(y_i f(x_i))$.

Για αυτό το νέο πρόβλημα εισάγουμε μεταβλητές χαλάρωσης ξ_i για την $\psi_1(y_i f(x_i))$, με $\xi_i \geq 0$. Τότε προκύπτει η ελαχιστοποίηση

$$\min_{\boldsymbol{w},b}\sum_{i=1}^{n}\xi_i + \frac{\lambda}{2}\langle \boldsymbol{w}, \boldsymbol{w}\rangle - \langle \boldsymbol{w}, V_1^{(k)}\rangle - \langle b, V_2^{(k)}\rangle \qquad (3.3.5)$$

με τους περιορισμούς

$$2(1 - y_i f(x_i) \leq \xi_i$$

$$\xi_i \geq 0$$

για i=1,...,n.

Στην συνέχεια κατανέμουμε στο νέο πρόβλημα (3.3.5) που προέκυψε, τους πολλαπλασιαστές Lagrange, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$ και $\boldsymbol{r} = (r_1, \ldots, r_n)$. Τότε το πρόβλημα μας παίρνει την εξής μορφή:

$$L_p = \sum_{i=1}^{n}\xi_i + \frac{\lambda}{2}\langle \boldsymbol{w}, \boldsymbol{w}\rangle - \langle \boldsymbol{w}, V_1^{(k)}\rangle - \langle b, V_2^{(k)}\rangle + \sum_{i=1}^{n}a_i\left(1 - y_i f(x_i) - \frac{1}{2}\xi_i\right) - \sum_{i=1}^{n}r_i\xi_i \qquad (3.3.6)$$

όπου λ=1/C, $2(1 - y_i f(x_i) \leq \xi_i, a_i \geq 0$ και $r_i \geq 0$.

Εμείς αναζητούμε τα **w**,b και **ξ** που ελαχιστοποιούν το πρόβλημα (3.3.6) και τα **α** τα οποία το μεγιστοποιούν.

Αυτό επιτυγχάνεται διαφορίζοντας το αντίστοιχο πρόβλημα Lagrange (3.3.6) ως προς **w**,b και **ξ**, και θέτοντας τις παραγώγους ίσες με το μηδέν:

$$\frac{\partial L_p}{\partial \boldsymbol{w}} = 0 \implies \lambda \boldsymbol{w} - V_1^{(k)} - \sum_{i=1}^{n} a_i y_i x_i = 0 \implies \boldsymbol{w} = \frac{1}{\lambda}\left(\sum_{i=1}^{n} a_i y_i x_i + V_1^{(k)}\right) \quad (3.3.7)$$

$$\frac{\partial L_p}{\partial b} = 0 \implies \sum_{i=1}^{n} a_i y_i + V_2^{(k)} = 0 \implies \sum_{i=1}^{n} a_i y_i - = V_2^{(k)} \quad (3.3.8)$$

$$\frac{\partial L_p}{\partial \xi} = 0 \implies 1 - \frac{1}{2}\alpha_i - r_i = 0 \implies 1 - \frac{1}{2}\alpha_i = r_i \quad (3.3.9)$$

Τότε με αντικατάσταση των (3.3.7), (3.3.8) και (3.3.9) στην εξίσωση (3.3.6) προκύπτει:

$$L = \left[\sum_{i=1}^{n} \alpha_i \left[1 - y_i \langle V_1^{(k)}, x_i \rangle\right] - \frac{\lambda}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle\right]$$

με τους περιορισμούς

$$\sum_{i=1}^{n} a_i y_i = -V_2^{(k)}$$

$$2 \geq \alpha_i \geq 0$$

για i=1,...,n.

Έτσι το πρόβλημα μας πλέον γίνεται:

$$\max_{\boldsymbol{\alpha}} \left[\sum_{i=1}^{n} \alpha_i \left[1 - y_i \langle V_1^{(k)}, x_i \rangle\right] - \frac{\lambda}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle\right] \quad (3.3.10)$$

με τους περιορισμούς

$$\sum_{i=1}^{n} a_i y_i = -V_2^{(k)}$$

$$2 \geq \alpha_i \geq 0$$

για i=1,...,n.

Θεώρημα 3.3.1: (Γραμμική περίπτωση)

Με την εισαγωγή των πολλαπλασιαστών Lagrange $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, .., \alpha_n)$,το (k+1) πρόβλημα της (3.3.4) προκύπτει να είναι:

$$\max_{\boldsymbol{\alpha}} W(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i \left[1 - y_i \langle V_1^{(k)}, x_i \rangle\right] - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle \quad (3.3.11)$$

με τους εξής περιορισμούς

$$\sum_{i=1}^{n} \alpha_i y_i = -V_2^{(k)}$$

$$2C \geq \alpha_i$$

$$\alpha_i \geq 0$$

για i=1,…,n.

Τότε η λύση $\left(w_1^{(k+1)}, .., w_d^{(k+1)}\right)$ της (3.3.4) είναι $V_1^{(k)} + \sum_{i=1}^{n} \alpha_i^{(k+1)} y_i x_i$ και ισχύουν τα εξής:

$w_{d+1}^{(k+1)}$ ικανοποιεί την συνθήκη ΚΚΤ, $y_i \langle w^{(k+1)}, x_i \rangle = 1$ για i=1,..,n,

$2C > \alpha_i^{(k+1)} > 0$ και $\left\{\alpha_i^{(k+1)}\right\}_{i=1}^{n}$ είναι η λύση του προβλήματος (3.3.11).

Θεώρημα 3.3.2: (Μη γραμμική περίπτωση)

Το (k+1) πρόβλημα της (3.3.4) είναι η (3.3.11), αντικαθιστώντας το $\langle V_1^{(k)}, x_i \rangle$ με

$$C \sum_{j=1}^{n} \nabla \psi_2 \left(y_j f^{(k)}(x_j)\right) y_j K(x_i, x_j)$$

και το $\langle x_i, x_j \rangle$ με $K(x_i, x_j)$, δηλαδή

$$\max_{\boldsymbol{\alpha}} W(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i - \alpha_i y_i C \nabla \psi_2 \left(y_j f^{(k)}(x_j)\right) y_j K(x_i, x_j) - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K \langle x_i, x_j \rangle$$

Η λύση $\left\{\alpha_i^{(k+1)}\right\}_{i=1}^{n}$ αποδίδει την λύση

$$w_j^{(k+1)} = y_j \left(\alpha_j^{(k+1)} + C \nabla \psi_2 \left(y_j f^{(k)}(x_j)\right)\right)$$

του προβλήματος (3.3.4) για j=1,...,n, με $w_{n+1}^{(k+1)}$ να ικανοποιεί την συνθήκη ΚΚΤ $y_i\langle w^{(k+1)}, x_i\rangle = 1$ για i=1,..,n και $2C > \alpha_i^{(k+1)} > 0$.

Με βάση τα πιο πάνω θεωρήματα διατρέχοντας μια QP επίλυση παίρνουμε την επιθυμητή λύση για το κυρτό πλέον πρόβλημα (3.3.11) αποδίδοντας την επιθυμητή λύση του αντίστοιχου μη κυρτού προβλήματος (3.2.4) (αντίστοιχα του (3.2.6) για την μη γραμμική περίπτωση).

3.3.2 Εξωτερική Μέθοδος Προσέγγισης (Outer Approximation Method)

Οι εξωτερικές μέθοδοι προσέγγισης είναι ισχυρά εργαλεία που έχουν αναπτυχθεί για την μη κυρτή ελαχιστοποίηση και με την σωστή χρήση τους, μπορούν να προσφέρουν υψηλές αποδόσεις. Παρ όλα αυτά, αυτές οι μέθοδοι σε προβλήματα μεγάλης κλίμακας μπορούν να γίνουν ακατόρθωτοι ενώ για μικρά προβλήτα είναι πιο αποτελεσματικοί.

Στόχος μας είναι να λύσουμε το μη κυρτό πρόβλημα ελαχιστοποίησης (3.2.4) (αντίστοιχα (3.2.6)) με την ελαχιστοποίηση μιας ακολουθίας από κυρτές άνω προσεγγίσεις της αντικειμενικής συνάρτησης (3.2.3). Η τεχνική αυτή υλοποιείται μέσω κατασκευής μιας εξωτερικής προσέγγισης. Αυτή η τεχνική λύνει ένα D.C. πρόβλημα ελαχιστοποίησης μέσω της ελαχιστοποίησης μιας κατώτερης περιβάλλουσας (lower envelopes) ακολουθίας καθώς επίσης και μίας ακολουθίας από κυρτές ελαχιστοποιήσεις μέσω της κορυφής απαρίθμησης (vertex enumeration). Ως αποτέλεσμα, παράγεται μια ολική ελαχιστοποίηση.

Πιο συγκεκριμένα η μέθοδος, κατασκευάζει μια κατώτερη περιβάλλουσα για την $s_1(\boldsymbol{w})$ (3.3.2), αποδίδοντας μια ακολουθία από μη συνεχής κατώτερες περιβάλλουσες $\{L^{(k)}\}$ της $s(w)$ (3.3.1). Η περιβάλλουσα $L^{(k)}$ αποδίδει το ελάχιστο $\{\boldsymbol{w}^{(k+1)}\}$, το οποίο «οδηγεί» με τη σειρά του την $L^{(k)}$ στο να δώσει μια καλύτερη κατώτερη περιβάλλουσα $L^{(k+1)}$. Αυτή η διαδικασία επαναλαμβάνεται μέχρι το κριτήριο της διακοπή να πληρούται. Έτσι $\boldsymbol{w}^{(k)}$ συγκλίνει στο ολικό ελάχιστο. Με βάση αυτά οδηγούμαστε σε δύο πολύ κρίσιμα βήματα:

- Στην κατασκευή της $\{L^{(k)}\}$
- Στον τρόπο επίλυσης της ελαχιστοποίησης της $L^{(k)}$.

Πρόσφατα, οι Blanquero και Carrizosa (2000) έδειξαν ότι η υπολογιστική επιβάρυνση των εξωτερικών μεθόδων προσέγγισης μειώνεται σημαντικά όταν υπάρχει μια καλή D.C. ανάλυση.

Το νέο πρόβλημα της ελαχιστοποίησης της $L^{(k)}$ που προκύπτει περιλαμβάνει την κυρτή ελαχιστοποίηση της $-s_2(\boldsymbol{w})$ (3.3.3) με ορισμένους γραμμικούς περιορισμούς που ορίζονται από την $L^{(k)}$. Για τον υπολογισμό, χρησιμοποιούμε μια διακριτή ελαχιστοποίηση για την εύρεση της ελάχιστης κορυφής μεταξύ των κορυφών ενός πολυέδρου, που ορίζεται από αυτούς τους γραμμικούς περιορισμούς. Εδώ να σημειωθεί ότι ένα πολύεδρο είναι το σύνολο των λύσεων σε ένα σύστημα γραμμικών ανισοτήτων.

Η ελαχιστοποίηση της (3.2.3) (αντίστοιχα για την μη γραμμική περίπτωση(3.2.5)) με την μέθοδο της εξωτερικής προσέγγισης γίνεται ως εξής:

Αρχικά κατασκευάζουμε την κατώτερη περιβάλλουσα $\{L^{(k)}\}$ της

$$s(\boldsymbol{w}) = s_1(\boldsymbol{w}) - s_2(\boldsymbol{w})$$

ως εξής:

Κατά την (k+1) επανάληψη με δεδομένο $w^{(j)}$ για j=1,...,k, θέτουμε

$$L^{(k)} = \max_{1 \le j \le k} L_j$$

όπου L_j είναι γραμμική συνάρτηση και ορίζεται ως εξής:

$$L_j(w) = -s_2(\boldsymbol{w}) + s_1\left(\boldsymbol{w}^{(j)}\right) + \langle \nabla s_1\left(\boldsymbol{w}^{(j)}\right), \boldsymbol{w} - \boldsymbol{w}^{(j)} \rangle, \text{ j=1,...,k}$$

με

$$\nabla s_1\left(\boldsymbol{w}^{(j)}\right) = \boldsymbol{w}^{(j)} + C \sum_{i=1}^{n} \left(\nabla \psi_1\left(y_i f(x_i)\right) y_i\right) x_i$$

να είναι η κλίση του $s_1(\boldsymbol{w})$ στο $\boldsymbol{w}^{(j)}$ για j=1,...,k.

Δηλαδή γίνεται αντικατάσταση της $s_1(\boldsymbol{w})$ με βάση την συνάρτηση

$$s_1\left(\boldsymbol{w}^{(j)}\right) + \langle \nabla s_1\left(\boldsymbol{w}^{(j)}\right), \boldsymbol{w} - \boldsymbol{w}^{(j)} \rangle.$$

Τότε στην (k+1) επανάληψη προκύπτει ότι

$$\boldsymbol{w}^{(k+1)} = arg \min_{\boldsymbol{w}} L^{(k)}(\boldsymbol{w})$$

και έτσι το L_{k+1} ορίζεται να είναι

$$L_{k+1}(\boldsymbol{w}) = -s_2(\boldsymbol{w}) + s_1\left(\boldsymbol{w}^{(k+1)}\right) + \langle \nabla s_1\left(\boldsymbol{w}^{(k+1)}\right), \boldsymbol{w} - \boldsymbol{w}^{(k+1)} \rangle$$

οδηγώντας έτσι στην $L^{(k+1)}$ η οποία προκύπτει,

$$L^{(k+1)}(\boldsymbol{w}) = max\big(L_{k+1}(\boldsymbol{w}), L^{(k)}(\boldsymbol{w})\big).$$

Από την κατασκευή της $L^{(k)}$ προκύπτουν οι εξής ανισότητες:

$$s(\boldsymbol{w}) \geq L^{(k+1)} \geq L^{(k)} \geq \cdots \geq L^{(1)}$$

Για την εύρεση του $\boldsymbol{w}^{(k+1)}$ στην (k+1) επανάληψη λύνουμε την ακόλουθη κυρτή ελαχιστοποίηση

$$\min_{(\boldsymbol{w},t)\in B^{(k)}} -s_2(\boldsymbol{w}) + t \qquad (3.3.12)$$

όπου

$$B^{(k)} = \big\{(\boldsymbol{w}, t): t \geq s_1\big(\boldsymbol{w}^{(j)}\big) + \langle \nabla s_1\big(\boldsymbol{w}^{(j)}\big), \boldsymbol{w} - \boldsymbol{w}^{(j)}\rangle, j = 1, \dots, k\big\}$$

είναι ένα πολύεδρο και ισχύει ότι $B^{(k+1)} \subset B^{(k)} \subset \cdots \subset B^{(0)}$. Για να αποφύγουμε τα εκφυλισμένα πολύεδρα αρχίζουμε με ένα φραγμένο πολύεδρο $B^{(0)}$, όπου $B^{(0)}$ κατασκευάζεται από $2(k(d+1)+1)$ γραμμικούς περιορισμούς που ορίζονται από τα αρχικά όρια του $(\boldsymbol{w}, t)$. Αφού $s_1(\boldsymbol{w}) \geq 0$ τότε ισχύει ότι, $t \geq 0$, το οποίο χρησιμοποιείται σαν ένα κάτω όριο του t.

Η εύρεση της κυρτής ελαχιστοποίησης $\boldsymbol{w}^{(k+1)}$ επιτυγχάνεται σε μια κορυφή του φραγμένου πολύεδρου $B^{(k)}$. Με την προσθήκη του νέου γραμμικού περιορισμού

$$t \geq s_1\big(\boldsymbol{w}^{(j)}\big) + \langle \nabla s_1\big(\boldsymbol{w}^{(j)}\big), \boldsymbol{w} - \boldsymbol{w}^{(j)}\rangle$$

η ελαχιστοποίηση (3.3.6) μειώνεται σε ένα πρόβλημα εύρεσης νέων κορυφών του $B^{(k)}$, δηλαδή γίνεται ένα πρόβλημα διακριτής ελαχιστοποίησης (discrete minimization) για όλες τις νέες κορυφές του $B^{(k)}$. Ο νέος αυτός γραμμικός περιορισμός που δημιουργείται από την (k+1) επανάληψη αποδίδει νέες κορυφές για την (k+2) επανάληψη κ.ο.κ. Έτσι επαναληπτικά οδηγούμαστε στην σύγκλιση της ακολουθίας $\boldsymbol{w}^{(k)}$ στο ολικό ελάχιστο, δηλαδή

$$\lim_{k\to\infty} s\big(\boldsymbol{w}^{(k+1)}\big) = \min_{\boldsymbol{w}} s(\boldsymbol{w})$$

και έτσι έχουμε την επιθυμητή λύση.

Εν κατακλείδι μια καλή αρχική τιμή $\boldsymbol{w}^{(0)}$ επισπεύδει τη σύγκλιση αλλά δεν επηρεάζει την τελική λύση. Έτσι οποιαδήποτε τιμή $\boldsymbol{w}^{(0)}$ μπορεί να επιλεγεί μέσα στο πολύεδρο $B^{(0)}$ ως αρχική τιμή. Στην πραγματικότητα, αυτή η τεχνική δεν είναι ευαίσθητη στην επιλογή των αρχικών τιμών. Η επιλογή όμως από την άλλη του $B^{(0)}$ είναι σημαντική για την ταχύτητα σύγκλισης. Εάν η περιοχή του $B^{(0)}$ είναι πολύ μεγάλη τότε έχουμε αργή σύγκλιση. Παρ' όλα αυτά για να περιέχει την

ολική ελαχιστοποίηση που επιθυμούμε το πολύεδρο $B^{(0)}$ θα πρέπει να είναι αρκετά μεγάλο.

3.4 Άλλοι τρόποι βελτίωσης της ψ-μάθησης

Εκτός από τις προαναφερθέντες μεθόδους βελτίωσης της ψ-μάθησης για περισσότερη ακρίβεια γενίκευσης έχει αναπτυχθεί και μια άλλη μέθοδος η οποία μετατρέπει την μη κυρτή ελαχιστοποίηση της ψ-μάθησης σε ένα πρόβλημα μικτού ακέραιου προγραμματισμού (mixed integer programming) (MIP). Ο νέος αυτός αλγόριθμος μπορεί να λύση την κυρτή ελαχιστοποίηση της ψ-μάθησης με μια τμηματικά γραμμική απώλεια χωρίς να απαιτείται η συνάρτησης απώλειας να είναι συνεχής.

Ένα πρόβλημα MIP είναι η ελαχιστοποίηση μιας γραμμικής ή τετραγωνικής συνάρτησης που υπόκεινται σε περιορισμούς γραμμικούς με ορισμένες από τις μεταβλητές να είναι ακέραιοι. Όταν η αντικειμενική συνάρτηση είναι τετραγωνική, γίνεται ένα μικτό ακέραιο πρόβλημα τετραγωνικού προγραμματισμού (MIQP). MIP είναι ένα σημαντικό πρόβλημα βελτιστοποίησης στο πεδίο της έρευνας και έχει μελετηθεί για πολλά χρόνια. Μια κοινή προσέγγιση για να λύσουμε ένα MIP είναι ο αλγόριθμος Branch και Bound. Ο αλγόριθμος Branch and Bound αποτελείται από δύο λειτουργίες την οριοθέτηση (bounding) και την υποδιαίρεση (subdivision). Η λειτουργία της οριοθέτησης κατασκευάζει τα άνω και κάτω όρια της s(**w**), ενώ η λειτουργία της υποδιαίρεσης χωρίζει τις περιοχές. Η ουσία αυτού του αλγόριθμου είναι η επίλυση του αρχικού γραμμικού προβλήματος χαλάρωσης χωρίς την απαίτηση ακέραιου. Αν η λύση ικανοποιεί τους ακέραιους περιορισμούς, τότε η βέλτιστη λύση βρέθηκε. Διαφορετικά μέσω διακλάδωσης δημιουργούμε δύο νέα υποπροβλήματα μιας κλασματικής μεταβλητή η οποία απαιτείται να είναι ένας ακέραιος. Ο αλγόριθμος σταματά όταν επιτυγχάνεται η βέλτιστη λύση.

Η πολυπλοκότητα της MIP ποικίλλει, ανάλογα με το μέγεθος του προβλήματος, τα αριθμητικά χαρακτηριστικά των δεδομένων καθώς και από τον αλγόριθμο που χρησιμοποιείται. Κάποια προβλήματα MIP με εκατοντάδες χιλιάδες μεταβλητές και αντίστοιχους περιορισμούς μπορούν να επιλυθούν μέσα σε λίγα λεπτά. Από την άλλη πλευρά, υπάρχουν μικρά προβλήματα MIP με μερικές εκατοντάδες μεταβλητές που δεν έχουν ακόμα επιλυθεί. Όταν δεν είναι δυνατό

να υπολογίσουμε τη βέλτιστη λύση, μπορούμε να διευθετήσουμε για μια καλή λύση που δεν είναι η βέλτιστη.

Μετατρέποντας το πρόβλημα ελαχιστοποίησης της ψ-μάθησης σε MIP αυτό μας επιτρέπει να κάνουμε χρήση των υφιστάμενων εργαλείων του MIP για την αντιμετώπιση της ψ-μάθησης. Η μετατροπή του προβλήματος ελαχιστοποίησης σε MIP καλείται απεριόριστη ψ-μάθηση. Παρά το γεγονός ότι ορισμένα προβλήματα MIP δεν μπορούν να λυθούν σε πολυωνυμικό χρόνο (polynomial time,), πολλοί MIP μπορούν να οδηγήσουν σε μια καλή δυνατή λύση που λαμβάνεται μέσα σε ένα συγκεκριμένο χρονικό διάστημα.

3.5 Θεωρία Στατιστικής Μάθησης

Οι Shen et al. (2003) ανέπτυξαν μια θεωρία μάθησης για την τεχνική ψ-μάθηση. Αυτή η θεωρία μάθησης διερευνά την δυνατότητα γενίκευσης της ψ-μάθησης σε θεωρητικό επίπεδο παρέχοντας της θεωρητικά πλεονεκτήματα. Πιο συγκεκριμένα η θεωρία αυτή παρέχει την πιθανότητα ακρίβειας της γενίκευσης της τεχνικής ψ-μάθησης δίνοντας ένα καλύτερο άνω φράγμα στο σφάλματος γενίκευσης (ή ισοδύναμα σφάλμα δοκιμών) καθώς επίσης και μια καλύτερη απόδοση με την καθοδήγηση της επιλογής της ρυθμιστικής παραμέτρου C και την εύρεση του βέλτιστου δείκτη σύγκλισης.

Σε αυτή την παράγραφο θα αναπτύξουμε μια θεωρία για την ποσοτικοποίηση της ακρίβειας της γενίκευσης της μάθησης η οποία μετριέται μέσω της διαφοράς της πραγματικής και ιδανικής απόδοσης της συνάρτησης f ως συνάρτηση του μεγέθους του δείγματος εκπαίδευσης, n και της κατηγορίας των υποψηφίων συναρτήσεων απόφασης $\mathcal{F}$. Αυτή η θεωρία δείχνει ότι η ψ-μάθηση επιτυγχάνει ουσιαστικά βέλτιστα ποσοστά σύγκλισης.

3.5.1 Στατιστικές Ιδιότητες

Η απόδοση ενός ταξινομητή που ορίζεται από την συνάρτηση απόφασης f μετριέται από το μέγεθος του σφάλματος γενίκευσης GE, πιο συγκεκριμένα, η ακρίβεια της μάθηση της f μετριέται από την διαφορά,

$$e(f, f^*) = Err(f) - Err(f^*)$$

$$= E|f^*(X)|\big|Sign\big(f(X)\big) - Sign\big(f^*(X)\big)\big| \geq 0 \quad (3.5.1)$$

που αντιπροσωπεύει τη διαφορά μεταξύ της πραγματικής απόδοσης και της ιδανικής απόδοσης.

Αρχικά επιθυμούμε την ιδανική βέλτιστη συνάρτηση απόφασης του Bayes f^* που ελαχιστοποιεί την $Err(f) = E[1 - Sign(Yf(X))]$, δηλαδή

$$\inf_{f \neq 0} Err(f) = Err(f^*) = E\left(\frac{1}{2} - |f^*(X)|\right)$$

όπου $f^*(x) = P(Y = 1|x) - 1/2$.

Εν συνέχεια, θα αναπτύξουμε την θεωρία μάθησης για την ποσοτικοποίηση της διαφοράς $e(f, f^*)$, δηλαδή της ακρίβειας της μάθησης ως συνάρτηση του n, σε σχέση με την τιμή της ρυθμιστικής παραμέτρου C και του μεγέθους του συνόλου

$$G(\mathcal{F}) = \{G_f = \{x: f(x) \geq 0\}: f \in \mathcal{F}\}$$

όπου $G(\mathcal{F})$ είναι η κατηγορία των υποψήφιων συνόλων ταξινόμησης, που δημιουργείται από την κλάση $\mathcal{F}$ η οποία περιλαμβάνει όλες τις υποψήφιες συναρτήσεις απόφασης.

3.5.2 Θεωρία Μάθησης

Σκοπός της θεωρίας μας είναι η ποσοτικοποίηση της ακρίβειας της μάθησης καθώς και να αποκαλύψουμε το καλύτερο trade-off που σχετίζεται με την επιλογή της ρυθμιστικής παραμέτρου C και του μεγέθους της τάξης της υποψήφιας συνάρτησης. Για να επιτευχθούν αυτά αρχικά δίνουμε ένα άνω φράγμα του σφάλματος γενίκευσης (GE) σε σχέση με την πολυπλοκότητα της κλάσης των υποψήφιων συναρτήσεων απόφασης. Η θεωρία μάθησης διαμορφώνεται με βάση το μέγεθος του συνόλου $G(\mathcal{F})$, που ορίζεται να μετριέται με την μετρική εντρόπια. Για την συγκεκριμένη θεωρία, η ιδανική βέλτιστη ταξινόμηση του συνόλου

$$G_{f^*} = \{x \in S: f^*(x) \geq 0\}$$

δεν απαιτείται να ανήκει στο σύνολο $G(\mathcal{F})$. Σε αντίθεση, υποθέτουμε ότι ο βέλτιστος ταξινομητής $\bar{f} = Sign(f^*)$, μπορεί να προσεγγιστεί καλά από την $\mathcal{F}$. Για την ποσοτικοποίηση αυτής της προσέγγισης, θεωρούμε

$$e_\psi(f, \bar{f}) = \frac{1}{2}\left(E\psi(Yf(X)) - E\psi(Y\bar{f}(X))\right)$$

που μετρά το σφάλμα προσέγγισης.

Έστω $J_0 = max(J(f_0), 1)$

Κάνουμε τις ακόλουθες τέσσερις υποθέσεις.

Υπόθεση Α: (Σφάλμα Προσέγγισης) Για κάποια θετική ακολουθία $s_n \to 0$ με $n \to \infty$ υπάρχει $f_0 \in \mathcal{F}$ τέτοια ώστε

$$e_\psi(f_0, \bar{f}) \leq s_n.$$

Ισοδύναμα,

$$\inf_{\{f \in \mathcal{F}\}} e_\psi(f, \bar{f}) \leq s_n.$$

Εάν $\mathcal{F}$ και $\bar{f}$ είναι ανεξάρτητα του n

$\Longrightarrow$ $$\inf_{\{f \in \mathcal{F}\}} e_\psi(f, \bar{f}) = 0$$

Εδώ να σημειωθεί ότι ο ταξινομητής του Bayes $\bar{f}$ ελαχιστοποιεί τις $E\psi(Yf(X))$ και $E\left(1 - Sign(Yf(X))\right)$ αυτό οδηγεί στην εξής ανισότητα:

$$E\psi(Yf(X)) \geq E\psi\left(Y\bar{f}(X)\right) = E\left(1 - Sign\left(Y\bar{f}(X)\right)\right) \leq E\left(1 - Sign(Yf(X))\right)$$

Από την ανισότητα και από τον ορισμό του σφάλματος προσέγγισης e_ψ προκύπτει ότι

$$e(f_0, f^*) \leq e_\psi(f_0, \bar{f}) \leq s_n.$$

Υπόθεση Β: (Συνοριακή Συμπεριφορά) Έστω $0 < \alpha \leq +\infty$ και $c_1 > 0$ κάποιες σταθερές τέτοιες ώστε

$$P(x \epsilon S: |f^*(x)| \leq \delta) \leq c_1 \delta^\alpha$$

για κάθε αρκετά μικρό $\delta \geq 0$.

Η υπόθεση Β περιγράφει τη συμπεριφορά της f^* κοντά στο όριο απόφασης $\{x: f^*(x) = 0\}$.

Προτού ορίσουμε την Υπόθεση Γ, θα ορίσουμε τη μετρική εντρόπια.

Ορισμός 1: Για μια δεδομένη κλάση $\mathcal{B}$ υποσυνόλων του S και για κάθε $\varepsilon > 0$, ένα ε-bracketing σύνολο καλείται το σύνολο

$$S(\varepsilon, m) = \left\{\left(G_1^l, G_1^u\right), \dots, \left(G_m^l, G_m^u\right)\right\}$$

του $\mathcal{B}$ εάν για κάθε G ∈ $\mathcal{B}$ υπάρχει ένα j τέτοιο ώστε

$$G_j^l \subset G \subset G_j^u \text{ και } \max_{1 \le j \le m} d\left(G_j^u, G_j^l\right) \le \varepsilon$$

όπου $d(\cdot,\cdot)$ είναι η απόσταση για οποιαδήποτε δύο σύνολα $G_i \in S$ και ορίζεται ως εξής:

$$d(G_1, G_2) = \int_{\{G_1 \Delta G_2\}} dP = P(G_1 \Delta G_2)$$

με

$$G_1 \Delta G_2 = (G_1 \backslash G_2) \cup (G_2 \backslash G_1)$$

να αποτελεί το σύνολο της διαφορά μεταξύ των G_i.

Με βάση τον ορισμό 1 είμαστε σε θέση να ορίσουμε την μετρική εντρόπια $H(\varepsilon, \mathcal{B})$ του $\mathcal{B}$:

Ορισμός 2: Η μετρική εντρόπια $H(\varepsilon, \mathcal{B})$ του $\mathcal{B}$ ορίζεται ως ο λογάριθμος της πληθικότητας του μικρότερου ε-bracketing σύνολο του $\mathcal{B}$, δηλαδή

$$H(\varepsilon, \mathcal{B}) = log(min\{m: S(\varepsilon, m) \text{ είναι ένα } \varepsilon\text{–bracketing σύνολο του } \mathcal{B}\})$$

Εν συνεχεία υποθέτουμε ότι

$$\mathcal{G}(k) = \left\{G_f = \{x: f(x) \ge 0\}: f \in \mathcal{F}, J(f) \le k\right\}$$

$$\subset G(\mathcal{F}) = \left\{G_f = \{x: f(x) \ge 0\}: f \in \mathcal{F}, J(f) < +\infty\right\}$$

όπου

$J(f) = \frac{1}{2}\|\boldsymbol{w}\|^2$ στην (3.2.3) και

$J(f) = \frac{1}{2}\|g\|_K^2$ στην (3.2.5).

Υπόθεση Γ: (Μετρική εντρόπια) Για κάποιες θετικές σταθερές $c_i, i = 2,3,4$ υπάρχει κάποια $\varepsilon_n > 0$ τέτοια ώστε

$$\sup_{\{k\geq 1\}} \phi(\varepsilon_n, k) \leq c_2 n^{1/2} \qquad (3.5.2)$$

όπου

$$\phi(\varepsilon_n, k) = \int_{c_4 L}^{\sqrt{c_3} L^{\frac{a}{2(a+1)}}} H^{1/2}(u^2/2, \mathcal{G}(k)) du/L$$

και

$$L = L(\varepsilon_n, C, k) = min\left(\varepsilon_n^2 + (Cn)^{-1} J_0 \left(\frac{k}{2} - 1\right), 1\right).$$

Υπόθεση Δ: (ψ συνάρτηση) Με σταθερές $0 < \tau \leq 1$ και U,

$$\begin{cases} U \geq \psi(x) \geq \left(1 - Sign(x)\right) & \text{εάν } x \in (0, \tau] \\ \psi(x) = (1 - Sign(x)) & \text{διαφορετικά} \end{cases}$$

Θεώρημα 3.5.1: Υποθέτουμε ότι οι υποθέσεις Α-Δ πληρούνται. Για κάθε ταξινομητή $Sign(\hat{f})$, της ψ-μάθηση υπάρχει μια σταθερή $c_5 > 0$ τέτοια ώστε

$$P\left(e(\hat{f}, f^*) \geq \delta_n^2\right) \leq 3.5 exp\left(-c_5 n (nC)^{-\frac{\alpha+2}{\alpha+1}} J_0^{\frac{\alpha+2}{\alpha+1}}\right)$$

υπό την προϋπόθεση ότι

$$Cn \geq 2\delta_n^{-2} J_0$$

όπου

$$\delta_n^2 = min(max(\varepsilon_n^2, 2s_n), 1)$$

Πόρισμα: Από τις προϋποθέσεις του πιο πάνω θεωρήματος προκύπτουν οι εξής δύο σχέσεις:

$$\left|e(\hat{f}, f^*)\right| = O_p(\delta_n^2)$$

$$E\left|e(\hat{f}, f^*)\right| = O(\delta_n^2)$$

υπό την προϋπόθεση ότι $n^{-\frac{1}{\alpha+1}}(C^{-1}J_0)^{\frac{\alpha+2}{\alpha+1}} \to 0$.

Εδώ να σημειωθεί ότι f^* είναι η βέλτιστη συνάρτηση απόφασης του Bayes και $\hat{f}$ είναι η εκτίμηση της συνάρτηση απόφασης.

Για την εφαρμογή αυτής της θεωρίας, θα πρέπει αρχικά να είμαστε βέβαιοι ότι οι υποθέσεις ικανοποιούνται για $s_n \to 0$ και $\varepsilon_n \to 0$. Έπειτα επιλέγουμε τη βέλτιστη trade-off τιμή για δ_n. Η βέλτιστη τιμή της ρυθμιστικής παραμέτρου C που αποδίδει την καλύτερη τιμή για το δ_n^2 καθορίζεται από τις εξής δύο σχέσεις:

1) $C/J_0 \geq 2n^{-1}\delta_n^{-1}$
2) $C/J_0 = O\left(n^{-\frac{1}{\alpha+2}}\right)$

Συνήθως, μια καλή επιλογή της παραμέτρου C είναι της τάξης του $n^{-1}\delta_n^{-1}J_0$. Σε αυτή την περίπτωση,

$$\begin{cases} \text{αν } \alpha \to \infty \implies 3.5exp(-c_5 n\delta_n^2) \\ \text{αν } \alpha \to 0 \implies 3.5exp(-c_5 n\delta_n^4) \end{cases}$$

δηλαδή παρατηρούμε ότι αν α → 0 τότε το όριο του θεωρήματος μειώνεται.

Με βάση τις υποθέσεις και το πιο πάνω θεώρημα κάνουμε τις εξής παρατηρήσεις:

- Παρά το γεγονός ότι η θεωρία μας λέει ότι το αποτέλεσμα ισχύει για κάθε ψ συνάρτηση που ικανοποιεί την Υπόθεση Δ, είναι σημαντικό να σημειωθεί ότι το σφάλμα προσέγγισης $e_\psi(f_0, \bar{f})$ μπορεί να διαφέρει σημαντικά ανάλογα με την επιλογή της ψ. Για παράδειγμα, αν

$$\psi(x) = \begin{cases} \dfrac{1}{1-x} & \text{για κάθε } 0 < x \leq 1 \\ 1 - Sign(x) & \text{διαφορετικά} \end{cases}$$

  τότε το σφάλμα προσέγγισης $e_\psi(f_0, \bar{f})$ θα μπορούσε να είναι πολύ μεγαλύτερο από ό,τι το $e_{\psi_0}(f_0, \bar{f})$. Επιπρόσθετα, η θεωρία επιτρέπει f_0, f^* και $\mathcal{F}$ να εξαρτώνται από το μέγεθος του δείγματος εκπαίδευσης n.

- Η Υπόθεση Β είναι στενά συνδεδεμένη με την (3.2.3), αλλά διαφορετική, δεδομένου ότι η υπόθεση αυτή επιβάλλεται μόνο στην συνάρτηση

απόφασης f^* του Bayes αντί για οποιαδήποτε συνάρτηση απόφασης $f \in \mathcal{F}$.

- Το ολοκλήρωμα της εξίσωσης (3.5.2) έχει χρησιμοποιηθεί για να ποσοτικοποιηθούν τα ποσοστά της σύγκλισης της μέγιστης πιθανοφάνειας των εκτιμητών.

Το θεώρημα μας παρέχει τέσσερεις πολύ σημαντικές συνέπειες. Η πρώτη είναι ότι παίρνουμε το όριο της πιθανότητας η ακρίβεια μάθησης $e(\hat{f}, f^*)$ να είναι μεγαλύτερη από την βέλτιστη τιμή δ_n^2, με το μικρότερο ε_n που ικανοποιεί την (3.5.2) και έτσι παίρνουμε ένα καλύτερο άνω όριο του σφάλματος γενίκευσης (GE). Η δεύτερη είναι ότι διαπιστώνουμε την ύπαρξη ενός trade-off μεταξύ της τιμής της παραμέτρου C και των αποδόσεων. Η τρίτη συνέπεια είναι ότι η καλύτερη απόδοση επιτυγχάνεται όταν η παράμετρος C δίνει την καλύτερη ισορροπία μεταξύ του μεγέθους της $G(\mathcal{F})$ των συναρτήσεων απόφασης και του μεγέθους του δείγματος n. Και τέλος παρέχει μια καθοδήγηση όσον αφορά την επιλογή της τιμής του C.

Εν κατακλείδι υπάρχουν δύο βασικές προσεγγίσεις για τη θεωρία της στατιστικής μάθησης για την ταξινόμηση. Η πρώτη προσέγγιση είναι η προσέγγιση του ορίου του σφάλματος γενίκευσης (GE) ενός ταξινομητή σε σχέση με το εμπειρικό σφάλμα εκπαίδευσης (EGE) και την πολυπλοκότητα της κλάση των υποψήφιων συναρτήσεων απόφασης. Ένα άνω όριο του $Err(\hat{f})$ μπορεί στη συνέχεια να ληφθεί για ένα συγκεκριμένο δείγμα εκπαίδευσης από την κατάλληλη επιλογή της σταθεράς δ που εξαρτάται από το n. Η δεύτερη προσέγγιση εκφράζει ένα άνω όριο της GE ενός ταξινομητή σε σχέση με την πολυπλοκότητα της κλάση των υποψήφιων συναρτήσεων απόφασης και της trade-off μεταξύ της πολυπλοκότητας και του σφάλματος εκπαίδευσης. Οι δύο προσεγγίσεις είναι συμπληρωματικές μεταξύ τους. Η πρώτη προσέγγιση είναι χρήσιμη όταν θέλουμε ένα όριο για το ποσοστό σφάλματος ταξινόμησης που βασίζεται σε ένα συγκεκριμένο παρατηρούμενο σύνολο δεδομένων. Παρ' όλα αυτά, επειδή αυτό το όριο είναι τυχαίο, δεν μπορεί να χρησιμοποιηθεί για τη σύγκριση διαφορετικών ταξινομητών εκ των προτέρων. Για τον σκοπό αυτό κυρίως βασιζόμαστε στην δεύτερη προσέγγιση.

3.5.3 Επεξηγηματικό Παράδειγμα

Τώρα θα μελετήσουμε ένα συγκεκριμένο παράδειγμα μάθησης για γραμμική ταξινόμησης και θα εφαρμόσουμε τη θεωρία της μάθησης όπως την αναπτύξαμε,

για να υπολογίσουμε την βέλτιστη σύγκλιση καθώς και το βέλτιστο άνω όριο της ακρίβεια μάθησης.

Παράδειγμα:

Έστω μια γραμμική ταξινόμηση που χρησιμοποιεί μια κατηγορία από υπερεπίπεδα,

$$\mathcal{F} = \{x \in S: f(x) = w \cdot x + b : w \in R^2\}$$

όπου

$$S = \{(x_1, x_2): x_1^2 + x_2^2 \leq 1\} \subset R^2$$

και η συνάρτηση απόφασης είναι $f_t(x) = x_1$ η οποία αποδίδει την κάθετη ευθεία μέχρι το όριο απόφασης.

Επιλέγουμε $f_0 = nf_t \in \mathcal{F}$, τότε:

$$e_\psi(f_0, \bar{f}) = \frac{1}{2}\left[E\psi(Yf_0(x)) - E\psi\left(Y\bar{f}(x)\right)\right]$$

$$\leq E\psi(Yf_0(X)) - E\left(1 - YSign(f^*(X))\right) \leq s_n = c_1 n^{-1}$$

για κάποια σταθερά $c_1 > 0$. Τότε η Υπόθεση Α πληρείται. Επίσης μέσω του κανόνα του Bayes

$$P(Y = 1|X = x) = \frac{\pi_1 f_1(x)}{\pi_1 f_1(x) + \pi_2 f_2(x)}$$

όπου $\pi_i, i = 1,2$ είναι οι πιθανότητες των θετικών και αρνητικών κλάσεων παίρνουμε ότι

$$P(x \in S: |f^*(x)| \leq \delta) = 0$$

για κάθε αρκετά μικρό $\delta > 0$. Τότε η υπόθεση Β ικανοποιείται για α =+∞.

Για να ελέγξουμε την Υπόθεση Γ, υπολογίζουμε την εντροπία της $\mathcal{G}_1(k)$, όπου $\mathcal{G}_1(k) = \mathcal{G}(k) \cap \{e(f, f_0) \leq 2u^2\}$, όπου $u \geq 0$ κάποια σταθερά. Εδώ να σημειωθεί ότι

$$e(f, f_0) \leq e_\psi(f, f_0) \leq 2u^2$$

τότε συνεπάγεται ότι $\|w - w_0\| \leq c'u^2$ για κάποια σταθερά $c' > 0$ όπου $f_0(x) = w_0 \cdot x$.

Επιπλέον,

$$\min_{1\le i\le n}|(w-w_0)\cdot x_i| \le |b-b_0| \le \max_{1\le i\le n}|(w-w_0)\cdot x_i|.$$

Αυτό ισχύει επειδή το b ελαχιστοποιεί την $\sum_{i=1}^{n}\psi\big(y_i f(x_i)\big)$ για οποιοδήποτε w. Ως εκ τούτου, για κάθε $f \in \mathcal{G}_1(k)$ από τον ορισμό του συνόλου $\mathcal{G}_1(k)$ προκύπτει ότι

$$|b-b_0| \le \|(w-w_0)\| \le c'u^2 \text{ και } \|w\| \le (2k)^{\frac{1}{2}}$$

Τότε οι υπολογισμοί αποδίδουν ότι

$$H(u^2, \mathcal{G}_1(k)) \le O\left(log\big(min\big(\sqrt{k_1}, c'u^2\big)/u^2\big)\right)$$

με $k_1 = \sqrt{(2k+\|w_0\|^2)}$.

Εδώ να επισημάνουμε ότι διαιρούμε με u^2 επειδή θέλουμε τον λογάριθμο της πληθικότητας.

Στην συνέχεια υποθέτουμε ότι,

$$\phi_1(\varepsilon_n, k) = \frac{\sqrt{H(\varepsilon_n^2, \mathcal{G}_1(k))}}{\sqrt{L}} = \frac{\sqrt{log\left(\frac{min\left(\sqrt{k_1}, c'\varepsilon_n^2\right)}{\varepsilon_n^2}\right)}}{\sqrt{min\left(\varepsilon_n^2 + (Cn)^{-1}J_0\left(\frac{k}{2}-1\right), 1\right)}}$$

όπου

$$L = min\left(\varepsilon_n^2 + (Cn)^{-1}J_0\left(\frac{k}{2}-1\right), 1\right)$$

Εύκολα προκύπτει ότι

$$\sup_{k\ge 1}\phi(\varepsilon_n, k) \le \phi_1(\varepsilon_n, 1) = \frac{\left(log\left(\frac{min\left((2+\|w_0\|^2)^{\frac{1}{4}}, c'\varepsilon_n^2\right)}{\varepsilon_n^2}\right)\right)^{\frac{1}{2}}}{\left(min\left(\varepsilon_n^2 + (Cn)^{-1}J_0\left(-\frac{1}{2}\right), 1\right)\right)^{\frac{1}{2}}} = c/\varepsilon_n$$

όπου $c > 0$ είναι μια σταθερά.

Τότε από την (3.5.2) προκύπτει ότι

$$\frac{c}{\varepsilon_n} = \frac{c_2}{n^{1/2}} \implies \varepsilon_n = n^{-1/2}$$

Έτσι όταν $C/max(J(f_0), 1)$ είναι μια αρκετά μεγάλη σταθερά προκύπτει ένας δείκτης $\varepsilon_n = n^{-1/2}$. Αυτός ο δείκτης που προέκυψε είναι ο βέλτιστος για τις διαχωρίσιμες περιπτώσεις. Ο δείκτης σφάλματος της γραμμικής SVM είναι $n^{-1/2}$ σε μη διαχωρίσιμες περιπτώσεις και n^{-1} σε διαχωρίσιμες περιπτώσεις. Σύμφωνα με Shen et al. (2003), η ψ-μάθηση επιτυγχάνει ένα ταχύτερο δείκτη σύγκλισης $n^{-1}\, log(n)$ στις μη διαχωρίσιμες περιπτώσεις. Αυτό είναι αξιοσημείωτο γιατί ένας τέτοιος δείκτης σύγκλισης επιτυγχάνεται από πολλούς ταξινομητές άλλα μόνο σε διαχωρίσιμες περιπτώσεις.

Μένει τώρα να προσδιορίσουμε το άνω όριο της ακρίβειας της μάθησης. Υπενθυμίζουμε ότι βρισκόμαστε στην περίπτωση όπου $\alpha = +\infty$ (από Υπόθεση Β) έτσι από το θεώρημα και την προϋπόθεση

$$\delta_n^2 = min(max(\varepsilon_n^2, 2s_n), 1) = min\left(max\left(\frac{1}{n}, 2s_n\right), 1\right) = 1/n$$

προκύπτει ότι $e(\hat{f}, f^*) \leq O(n^{-1}\, log(1/\delta))$ και $E(\hat{f}, f^*) = O(n^{-1})$, εκτός για ένα σύνολο μικρότερης πιθανότητας από δ, όπου $\delta > 0$ μια πολύ μικρή σταθερά.

Το αποτέλεσμα αυτό ισχύει γενικά για κάθε ψ που ικανοποιεί την υπόθεση Δ, συμπεριλαμβανομένου και της ψ_0.

Από τα αποτελέσματα του πιο πάνω παραδείγματος βλέπουμε ότι πράγματι η ψ-μάθηση επιτυγχάνει ταχύτερη σύγκλιση σε μη διαχωρίσιμες περιπτώσεις από την SVM. Επίσης μέσω της θεωρίας της ψ-μάθησης οδηγηθήκαμε σε ένα βέλτιστο άνω φράγμα για την ακρίβεια της μάθησης καθώς και σε μια βέλτιστη απόδοση του δ_n^2. Εδώ να σημειωθεί ότι βρισκόμαστε στην διαχωρίσιμη περίπτωση όπου η ακρίβεια της μάθησης προκύπτει να είναι ίση με το σφάλμα γενίκευσης της εκτίμησης της συνάρτησης απόφασης έτσι,

$$e(\hat{f}, f^*) = Err(\hat{f}) \leq O(n^{-1}\, log(1/\delta))$$

και άρα έχουμε το επιθυμητό άνω φράγμα της GE σε σχέση με την πολυπλοκότητα.

3.6 Γενίκευση της δυαδικής ψ-μάθησης στην περίπτωση πολλαπλών κατηγοριών

Οι περιπτώσεις ταξινόμησης πολλαπλών κατηγορών (multicategory) συχνά αντιμετωπίζονται χωριστά από την δυαδική ταξινόμηση γιατί διαφορετικά δεν είναι δυνατή η σωστή γενίκευση.

Οι τεχνικές ταξινόμησης περιθωρίου βάσης, αντί της απευθείας εκτίμηση των δεσμευμένων πιθανοτήτων, επικεντρώνονται στα όρια απόφασης αποδίδοντας έτσι την ταξινόμηση. Κατά συνέπεια, μια γενίκευση από την δυαδική στη περίπτωση πολλαπλών κατηγοριών είναι εξαιρετικά μη τετριμμένη.

Πρόσφατα, Liu και Shen (2004) γενίκευσαν την τεχνική της ψ-μάθησης από τη δυαδική ταξινόμηση στην περίπτωση πολλαπλών κατηγοριών, και έδειξαν ότι η γενίκευση αυτή διατηρεί την ερμηνεία των περιθωρίων καθώς και τις ιδιότητες της δυαδικής ταξινόμησης. Επιπλέον, έδειξαν ότι η ψ-μάθηση σε αντίθεση με την SVM δεν έχει κακές αποδόσεις στην περίπτωση όπου δεν υπάρχει κυρίαρχη τάξη (οι Lee, et. al (2004) επισήμαναν ότι στην περίπτωση συστήματος ταξινόμησης one-versus-the-rest μπορεί να έχουμε κακές επιδόσεις όταν δεν υπάρχει μια κυρίαρχη τάξη). Πιο συγκεκριμένα αποδείχθηκε ότι η ψ-μάθηση υπολογίζει άμεσα το πραγματικό όριο απόφασης ανεξάρτητα από την παρουσία ή απουσία της κυρίαρχης τάξης.

Η ταξινόμηση πολλαπλών κατηγοριών με την χρήση της ψ-μάθησης φαίνεται να είναι πιο ισχυρή σε ακραίες περιπτώσεις οι οποίες είναι εσφαλμένα ταξινομημένες από την αντίστοιχη ταξινόμηση της SVM. Για την γενίκευση της δυαδικής ψ-μάθησης στην περίπτωση πολλαπλών κατηγοριών και για την αντιμετώπιση όλων των τάξεων συγχρόνως, γενικεύουμε την έννοια των περιθωρίων και των διανυσμάτων υποστήριξης μέσω πολλαπλών συγκρίσεων μεταξύ των διαφόρων κατηγοριών. Επιπλέον θα δούμε ότι, στην γραμμική περίπτωση παρουσιάζεται κάποια συμπεριφορά που είναι ίδια με αυτήν της μη γραμμικής περίπτωσης σε σχέση με την ρυθμιστική παράμετρο, αλλά που διαφέρει από εκείνη της δυαδικής περίπτωσης.

3.6.1 Μεθοδολογία

Υπενθυμίζουμε ότι ο πρωταρχικός στόχος της ταξινόμησης είναι να προβλέψει την κλάση y για ένα δεδομένο διάνυσμα εισόδου $\boldsymbol{x} \in S$ μέσω ενός ταξινομητή,

όπου S είναι ένας χώρος εισόδου. Για k-κατηγορίες ταξινόμησης, ένας ταξινομητής χωρίζει το χώρο εισόδου S σε k ξένες περιοχές $S_1, \ldots, S_k$ με S_j να αντιστοιχεί στην κατηγορία j. Ένας καλός ταξινομητής είναι αυτός που προβλέπει με ακρίβεια την ετικέτα της κλάσης y για δεδομένο x, όπως μετράται από την ακρίβεια γενίκευσης.

Στην περίπτωση των πολλαπλών κατηγοριών κωδικοποιούμε την κλάση y ως $\{1, \ldots, k\}$ και ορίζουμε την $\boldsymbol{f} = (f_1, \ldots, f_k)$ ως το διάνυσμα της συνάρτησης απόφασης. Εδώ $f_j: S \to R$, και αντιπροσωπεύει την κατηγορία j, για j =1,. . . ,k. Το σφάλμα γενίκευσης (GE) σε αυτή τη περίπτωση ορίζεται να είναι

$$Err(f) = \frac{1}{2} E[1 - Sign(\boldsymbol{g}(\boldsymbol{f}(X), Y))]$$

όπου

$$\boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x}), y) = \left(f_y(\boldsymbol{x}) - f_1(\boldsymbol{x}), \ldots., f_y(\boldsymbol{x}) - f_{y-1}(\boldsymbol{x}), f_y(\boldsymbol{x}) - f_{y+1}(\boldsymbol{x}), \ldots, f_y(\boldsymbol{x}) - f_k(\boldsymbol{x})\right)$$

εκτελεί πολλαπλές συγκρίσεις της κλάσης y έναντι των υπολοίπων κλάσεων.

Έτσι το εμπειρικό σφάλμα γενίκευσης (EGE) προκύπτει να είναι:

$$(2n)^{-1} \sum_{i=1}^{n} \left(1 - Sign(\boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x}_i), y_i))\right)$$

Η συνάρτηση $\boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x}), y)$ διαδραματίζει σημαντικό ρόλο, γιατί περιγράφει τον τρόπο που εκτελούνται οι πολλαπλές συγκρίσεις για προβλήματα πολλαπλών κατηγοριών και συνδέεται άμεσα με τα γενικευμένα περιθώρια.

Για την γενίκευση από την δυαδική περίπτωση στην περίπτωση πολλαπλών κατηγοριών για ένα διάνυσμα $\boldsymbol{u} = (u_1, \ldots, u_{k-1})$, ορίζουμε την συνάρτηση πολλών μεταβλητών $Sign(\boldsymbol{u})$ ως εξής:

$$Sign(\boldsymbol{u}) = \begin{cases} 1 & \text{εάν } \boldsymbol{u}_{min} = min(u_1, \ldots, u_{k-1}) \geq 0 \\ -1 & \text{εάν } \boldsymbol{u}_{min} = min(u_1, \ldots, u_{k-1}) < 0 \end{cases}$$

Με τις συναρτήσεις $Sign(\cdot)$ και $\boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x}), y)$ στη θέση της f αποδίδεται ορθή ταξινόμηση για κάθε δεδομένο $(\boldsymbol{x}, y)$, εάν $\boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x}), y) \geq \boldsymbol{0}_{k-1}$ όπου $\boldsymbol{0}_{k-1}$ είναι ένα (k-1)-διάστατο διάνυσμα με μηδενικά.

3.6.2 Ταξινόμηση πολλαπλών κατηγοριών

Στην δυαδική περίπτωση ορίσαμε μια ψ συνάρτηση η οποία έπαιζε το ρόλο της καταργήσεως του προβλήματος της κλιμάκωσης της συνάρτησης Sign, η οποία ήταν αναλλοίωτη. Τώρα στη περίπτωση πολλαπλών κατηγοριών χρησιμοποιώντας την κωδικοποίηση {1,…,k} ορίζουμε μια συνάρτηση ψ πολλών μεταβλητών με k-1 ορίσματα ως εξής:

$$\begin{cases} U \geq \psi(\boldsymbol{u}) > 0 & \text{εάν } \boldsymbol{u}_{min} \in (0,\tau_1] \times \cdots \times (0,\tau_{k-1}] \\ \psi(\boldsymbol{u}) = 1 - Sign(\boldsymbol{u}) & \text{διαφορετικά} \end{cases} \tag{3.6.1}$$

όπου $0 < \tau_1 < \cdots < \tau_{k-1} \leq 1$, και $0 < U \leq 2$ είναι κάποιες σταθερές. Τότε η συνάρτηση $\psi(\boldsymbol{u})$, για κάθε u_j, $j = 1, \dots, k-1$ δεν αυξάνεται και έτσι η συνάρτηση πολλών μεταβλητών (3.6.1) διατηρεί τις επιθυμητές ιδιότητες της αντίστοιχης συνάρτησης (3.2.2) στην οποία είχαμε μόνο μια μεταβλητής. Πιο συγκεκριμένα, για κάθε υπόδειγμα x_i τέτοιο ώστε $Sign\big(\boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x}_i), y_i)\big) = 1$, δηλαδή η $\boldsymbol{f}$ ταξινομεί σωστά το x_i στην κατηγορία y_i, η $\psi(\boldsymbol{u})$ συνάρτηση εκχωρεί μια θετική ποινή σε κάθε περίπτωση μέσα στην περιοχή $(0,\tau_1] \times \cdots \times (0,\tau_{k-1}]$. Σε αυτή την περίπτωση χρησιμοποιούμε μια συγκεκριμένη συνάρτηση η οποία ικανοποιεί την (3.6.1) και ορίζεται ως εξής:

$$\psi(\boldsymbol{u}) = \begin{cases} 0 & \text{εάν } \boldsymbol{u}_{min} \geq 1 \\ 2 & \text{εάν } \boldsymbol{u}_{min} < 0 \\ 2(1 - \boldsymbol{u}_{min}) & \text{εάν } 0 \leq \boldsymbol{u}_{min} < 1 \end{cases} \tag{3.6.2}$$

Με βάση αυτά είμαστε σε θέση να ορίσουμε τις αντικειμενικές συναρτήσεις που προκύπτουν στην περίπτωση ταξινόμησης πολλαπλών κατηγοριών τόσο στην γραμμική όσο και στην μη γραμμική περίπτωση.

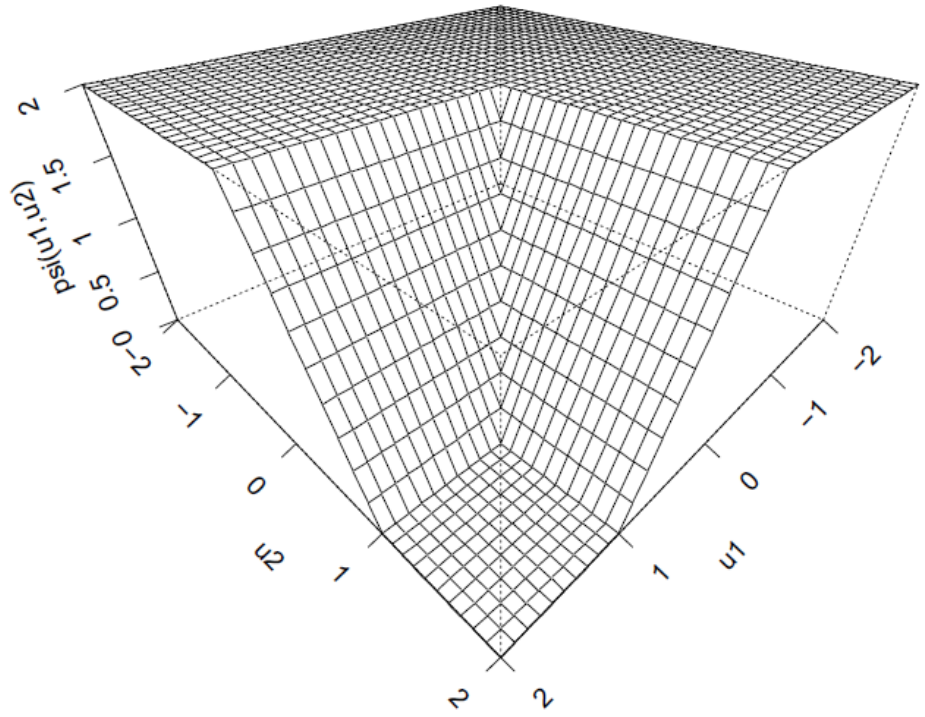
<u>Γραμμικά προβλήματα πολλαπλών κατηγοριών</u>

Στην γραμμική περίπτωση πολλαπλών κατηγοριών παίρνουμε την εξής ελαχιστοποίηση για την ψ-μάθηση:

$$\min_{b,\boldsymbol{w}} \left(\frac{1}{2} \sum_{j=1}^{k} \|\boldsymbol{w}_j\|^2 + C \sum_{i=1}^{n} \psi\big(\boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x}_i), y_i)\big) \right)$$

με τους περιορισμούς

$$\textstyle\sum_{j=1}^{k} f_j(\boldsymbol{x}) = 0 \ \forall\, x \in S \tag{3.6.3}$$



Σχήμα 3.5: Γράφημα 3-κλάσεων (k=3) της ψ συνάρτησης όπως την έχουμε ορίσει στην (3.6.2)

όπου $f_j(\boldsymbol{x}) = \langle \boldsymbol{w}_j, \boldsymbol{x} \rangle + b_j$, με $\boldsymbol{w}_j \in R^d$, $b_j \in R^1, j = 1, \dots, k$, $\boldsymbol{w} = vec(w_1, .., w_k)$ είναι ένα $(\mathrm{k} \cdot \mathrm{d})$ διαστάσεων διάνυσμα και $\boldsymbol{b} = (b_1, .., b_k)^T$ ένα k διαστάσεων διάνυσμα.

Όμως ο περιορισμός $\sum_{j=1}^{k} f_j(\boldsymbol{x}) = 0$ στην (3.6.3) περιλαμβάνει όλες τις τιμές του **x** στο S. Πρέπει συνεπώς να μειώσουμε τους άπειρους περιορισμούς για κάθε **x** στο S σε πεπερασμένους περιορισμούς για $\boldsymbol{x}_i$, $i = 1, \dots, n$. Αυτό επιτυγχάνεται με την χρήση του εξής θεωρήματος:

Θεώρημα 3.6.1: Ο περιορισμός

$$\sum_{j=1}^{k} f_j(\boldsymbol{x}) = 0 \ \forall \ \boldsymbol{x} \in S$$

είναι ισοδύναμος με τον περιορισμό

$$\tilde{X}\sum_{j=1}^{k}\tilde{\boldsymbol{w}}_j = 0$$

όπου

$\tilde{X} = [\mathbf{1}_n, X] = (\tilde{\boldsymbol{x}}_1, \dots, \tilde{\boldsymbol{x}}_n)^T$ είναι ένας $n \times (d+1)$ πίνακας , $X = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)^T$είναι $n \times d$ πίνακας σχεδιασμού, $\mathbf{1}_n = (1, \dots, 1)^T$ είναι ένα n-διάστατο διάνυσμα, $\tilde{\boldsymbol{x}}_i = \begin{bmatrix}1\\ \boldsymbol{x}_i\end{bmatrix}, i = 1, \dots, n$ είναι ένα διάνυσμα $(d+1)$ διαστάσεων, $\tilde{\boldsymbol{w}}_j = \begin{bmatrix}b_j\\ \boldsymbol{w}_j\end{bmatrix} \in R^{d+1}$, για $j = 1, \dots, k$ και $\tilde{\boldsymbol{w}} = (\tilde{\boldsymbol{w}}_1, \dots, \tilde{\boldsymbol{w}}_k) \in R^{k(d+1)}$.

Έτσι τελικά σε μια πιο απλή μορφή παίρνουμε τους εξής περιορισμούς:

$$\sum_{j=1}^{k} b_j \mathbf{1}_n + X\sum_{j=1}^{k}\boldsymbol{w}_j = 0$$

Συνοψίζοντας τα πιο πάνω και με βάση το θεώρημα καταλήγουμε στην εξής ελαχιστοποίηση για την γραμμική περίπτωση πολλαπλών κατηγοριών:

$$\min_{b,\boldsymbol{w}}\left(\frac{1}{2}\sum_{j=1}^{k}\left\|\boldsymbol{w}_j\right\|^2 + C\sum_{i=1}^{n}\psi\big(\boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x}_i), y_i)\big)\right)$$

με περιορισμούς

$$\textstyle\sum_{j=1}^{k} b_j \mathbf{1}_n + X\sum_{j=1}^{k}\boldsymbol{w}_j = 0 \qquad (3.6.4)$$

όπου η τιμή της παραμέτρου C (C>0) αντικατοπτρίζει τη σχετική σημασία μεταξύ του γεωμετρικού περιθωρίου και του εμπειρικού σφάλματος γενίκευσης (EGE).

Ορίζουμε τώρα το γενικευμένο περιθώριο λειτουργιών για υποδείγματα $(\boldsymbol{x}_i, y_i)$ ως

$$min(\boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x}_i), y_i)),$$

ενώ το γενικευμένο γεωμετρικό περιθώριο είναι

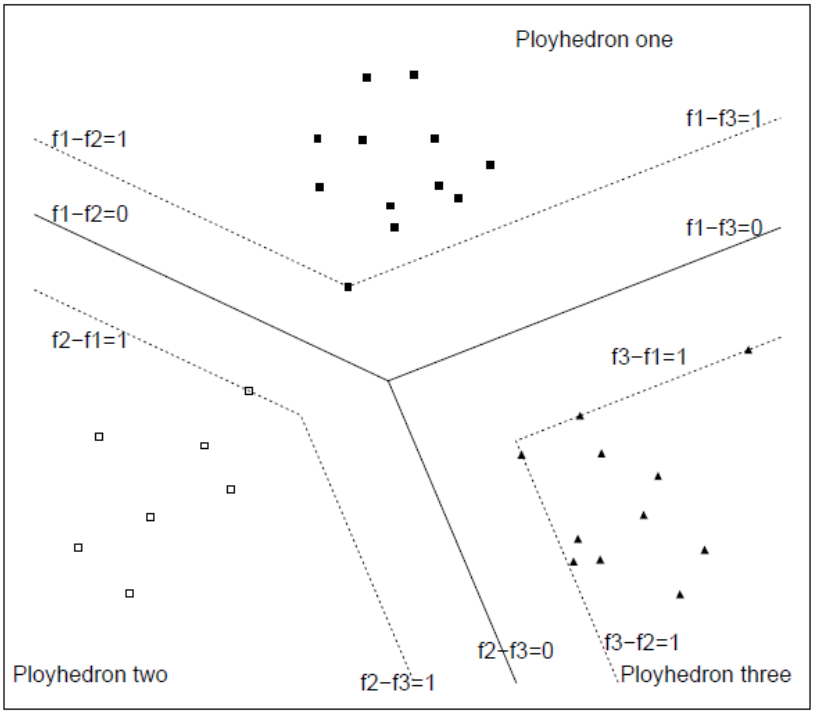
$$\gamma = \min_{1\le j_1<j_2\le k}\gamma_{j_1 j_2}$$

με $\gamma_{j_1 j_2} = \frac{2}{\|\boldsymbol{w}_{j1} - \boldsymbol{w}_{j2}\|}$ να είναι η κατακόρυφη Ευκλείδεια απόσταση μεταξύ των υπερεπίπεδων $f_{j1} - f_{j2} = \pm 1$. Αυτή η κατακόρυφη απόσταση $\gamma_{j_1 j_2}$ μέτρα τον διαχωρισμό μεταξύ των κλάσεων i και j.

Το Σχήμα 3.6 παρουσιάζει το ρόλο του γεωμετρικού περιθωρίου γ για την περίπτωση του k = 3 (όταν k = 2 βρισκόμαστε στη δυαδική περίπτωση (3.2.4)). Τα διανύσματα υποστήριξης είναι εκείνα τα υποδείγματα (instances) που καθορίζουν τα όρια απόφασης. Εξ ορισμού, τα διανύσματα υποστήριξης καθορίζουν μοναδικά τα όρια απόφασης των πολλαπλών κατηγοριών στην (3.6.4). Στην διαχωρίσιμη περίπτωση, τα υποδείγματα στα όρια του πολύεδρου D_j, είναι τα διανύσματα υποστήριξης. Εδώ να σημειωθεί ότι το πολύεδρο D_j είναι η συλλογή των λύσεων σε ένα πεπερασμένο σύστημα των γραμμικών ανισοτήτων που ορίζεται από

$$min\big(\boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x}_i), y_i)\big) \geq 1,\ j = 1, .., k.$$

Στην μη διαχωρίσιμη περίπτωση, τα διανύσματα υποστήριξης είναι τα υποδείγματα που ανήκουν στην κλάση j και που βρίσκονται στα όρια του D_j καθώς και στην περιοχή έξω από το D_j, για j=1,...,k.



Σχήμα 3.6: Παρουσιάζεται γραφικά η έννοια των περιθωρίων και των διανυσμάτων υποστήριξης σε 3 διαχωρίσιμες κατηγορίες. Οι περιπτώσεις για τις κατηγορίες 1-3 συμπίπτουν αντίστοιχα στα πολύεδρα $\boldsymbol{D_j}$ $\boldsymbol{j = 1, 2, 3}$ όπου $\boldsymbol{D_1 = \{x: f_1(x) - f_2(x) \geq 1, f_1(x) - f_3(x) \geq 1\}, D_2 = \{x: f_2(x) - f_1(x) \geq 1, f_2(x) - f_3(x) \geq 1\}}$ και $\boldsymbol{D_3 = \{x: f_3(x) - f_1(x) \geq 1, f_3(x) - f_2(x) \geq 1\}}$. Το γενικευμένο γεωμετρικά περιθώριο γ ορίζεται ως $\boldsymbol{min\{\gamma_{12}, \gamma_{12}, \gamma_{23}\}}$ που μεγιστοποιείται για να αποκτήσει το όριο απόφασης. Στο σχήμα υπάρχουν πέντε διανύσματα υποστήριξης στην κατηγορία 1, ένα στην κατηγορία 2, και τα άλλα τρία είναι από την κατηγορία 3.

Μη γραμμικά προβλήματα πολλαπλών κατηγοριών

Για μη γραμμικά προβλήματα, η f_j αναπαρίσταται ως

$$f_j = h_j(\boldsymbol{x}) + b_j$$

με $h_j = \sum_{i=1}^{n} v_{ji} K(\boldsymbol{x}_i, \boldsymbol{x})$, j=1,...,k, ορίζεται από ένα κατάλληλο πυρήνα $K(\cdot,\cdot)$, τον λεγόμενο γραμμικό από το $S \times S$ στο R. Έτσι σύμφωνα με το θεώρημα 3.6.1 στην μη γραμμική περίπτωση προκύπτουν οι περιορισμοί

$$\widetilde{\boldsymbol{K}} \sum_{j=1}^{k} \widetilde{\boldsymbol{v}}_j = 0$$

όπου $\widetilde{\boldsymbol{K}} = [1_n, K]$, $\boldsymbol{v}_j = \left(v_{j1}, \dots., v_{jn}\right)^T \in R^n$, $\boldsymbol{v} = (\boldsymbol{v}_1, \dots, \boldsymbol{v}_k) \in R^{nk}$,

$\widetilde{\boldsymbol{v}}_j = \begin{bmatrix} b_j \\ v_j \end{bmatrix} \in R^{n+1}$, και $\widetilde{\boldsymbol{v}} = (\widetilde{\boldsymbol{v}}_1, \dots, \widetilde{\boldsymbol{v}}_k) \in R^{k(n+1)}$.

Ισοδύναμα σε πιο απλή μορφή

$$\sum_{j=1}^{k} b_j \mathbf{1}_n + \boldsymbol{K} \sum_{j=1}^{k} \boldsymbol{v}_j = 0$$

Έτσι ο πυρήνας βάσης των πολλαπλών κατηγοριών της ψ-μάθησης λύνει το εξής πρόβλημα ελαχιστοποίησης:

$$\min_{\boldsymbol{b},\boldsymbol{v}} \frac{1}{2} \sum_{j=1}^{k} \left\| h_j \right\|_{H_K}^2 + C \sum_{i=1}^{n} \psi\left(\boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x}_i), y_i)\right)$$

με τους περιορισμούς

$$\textstyle\sum_{j=1}^{k} b_j \mathbf{1}_n + \boldsymbol{K} \sum_{j=1}^{k} \boldsymbol{v}_j = 0 \qquad (3.6.5)$$

Ειδικότερα χρησιμοποιώντας την ιδιότητα αναπαραγωγής του πυρήνα, $\left\| h_j \right\|_{H_K}^2$ μπορεί να γραφτεί ως $\boldsymbol{v}_j^T \boldsymbol{K} \boldsymbol{v}_j$.

Παρατηρούμε ότι και στις δυο περιπτώσεις προκύπτει μη κυρτή ελαχιστοποίηση όπως προέκυψε και στην δυαδική ταξινόμηση. Για την αντιμετώπιση των μη κυρτών ελαχιστοποιήσεων (3.6.4) και (3.6.5) εφαρμόζουμε τις ίδιες στρατηγικές με αυτές της δυαδικής περίπτωσης.

Αρχικά αναλύουμε την συνάρτηση ψ σε ψ_1 και ψ_2 , δηλαδή $\psi = \psi_1 + \psi_2$ αλλά τώρα θα έχουμε

$$\psi_1(\boldsymbol{u}) = \begin{cases} 0 & \text{εάν } \boldsymbol{u}_{min} \geq 1 \\ 2(1 - \boldsymbol{u}_{min}) & \text{διαφορετικά} \end{cases}$$

$$\psi_2(\boldsymbol{u}) = \begin{cases} 0 & \text{εάν } \boldsymbol{u}_{min} \geq 1 \\ 2\boldsymbol{u}_{min} & \text{διαφορετικά} \end{cases}$$

Εδώ οι ψ_1 και ψ_2 μπορεί να θεωρηθούν ως συναρτήσεις πολλών μεταβλητών οι οποίες γενικεύουν τις αντίστοιχες συναρτήσεις ψ_1 και ψ_2 της δυαδικής περίπτωσης. Έτσι έχουμε μια D.C ανάλυση και με την εφαρμογή του DCA ή των εξωτερικών προσεγγίσεων παίρνουμε την επιθυμητή λύση.

Κεφάλαιο 4

Αξιολόγηση του Μοντέλου και Εφαρμογές σε Πραγματικά Δεδομένα

4.1 Εισαγωγή

Η αξιολόγηση ενός μοντέλου είναι πολύ σημαντικό κεφάλαιο της στατιστικής ανάλυσης, γιατί κατευθύνει την επιλογή της μεθόδου εκμάθησης αλλά και την επιλογή του μοντέλου και μας δίνει ένα μέτρο ποιότητας του τελικώς επιλεγμένου μοντέλου. Σε αυτό το κεφάλαιο αρχικά θα αναλύσουμε θεωρητικά τα κριτήρια απόδοσης του μοντέλου Confusion Matrix τα οποία αποτελούν μέθοδοι αξιολόγησης της απόδοσης του επιλεγμένου μοντέλου. Έπειτα θα αξιολογήσουμε πειραματικά την εφαρμογή των Μηχανών Διανυσμάτων Υποστήριξης σε ένα πρόβλημα δυαδικής ταξινόμησης με βάση αυτά τα κριτήρια απόδοσης μέσω της R. Στην συνέχεια θα εφαρμόσουμε την θεωρία που έχουμε αναπτύξει στο τρίτο κεφάλαιο σε τρία πραγματικά δεδομένα. Μέσα από αυτές τις εφαρμογές, λαμβάνοντας τα σφάλματα δοκιμής για κάθε περίπτωση, θα έχουμε την

δυνατότητα να συγκρίνουμε τα αποτελέσματα των δυο μεθόδων SVM και ψ-μάθηση και κατ' επέκταση τις αποδόσεις τους.

4.2 Μέτρα αξιολόγησης

Σε ένα πρόβλημα δυαδικής ταξινόμησης υπάρχουν τέσσερις πιθανές εκβάσεις, όταν ένα μοντέλο εφαρμόζεται σε μια παρατήρηση:

TP	Αν η ετικέτα της παρατήρησης είναι θετική και είναι ταξινομημένη ως θετική, υπολογίζεται ως μια αληθώς θετική.
FN	Αν η ετικέτα της παρατήρησης είναι θετική και έχει ταξινομηθεί ως αρνητική, αυτό υπολογίζεται ως ψευδώς αρνητική.
TN	Αν η ετικέτα της παρατήρησης είναι αρνητική και έχει ταξινομηθεί ως αρνητική, αυτή υπολογίζεται ως μια αληθώς αρνητική.
FP	Αν η ετικέτα της παρατήρησης είναι αρνητική και έχει ταξινομηθεί ως θετική, τότε υπολογίζεται ως ψευδώς θετική.

Πίνακας 4.1: Τα τέσσερα πιθανά αποτελέσματα σε ένα πρόβλημα δυαδική ταξινόμησης

Ένας πίνακας σύγχυσης για ένα δυαδικό μοντέλο ταξινόμησης είναι ένας 2 × 2 πίνακας, στον οποίο εμφανίζονται οι παρατηρούμενες ετικέτες έναντι των προβλεπόμενων ετικετών για ένα σύνολο δεδομένων. Η πρώτη ετικέτα, y, οφείλεται στην παρατήρηση, και η δεύτερη ετικέτα, $\hat{y}$, οφείλεται στην πρόβλεψη του μοντέλου. Δηλαδή, για κάθε παρατήρηση έχουμε ένα ζεύγος ετικετών $(y, \hat{y})$. Αυτό το ζεύγος των ετικετών προσδιορίζει τις συντεταγμένες κάθε παρατήρησης μέσα στον πίνακα σύγχυσης, δηλαδή η πρώτη ετικέτα διευκρινίζει τη γραμμή του πίνακα και η δεύτερη ετικέτα διευκρινίζει τη στήλη του πίνακα, όπως φαίνεται στον πιο κάτω πίνακα:

Παρατηρούμενη ετικέτα (y)	Προβλεπόμενη ετικέτα ($\hat{y}$)	
	+1	-1
+1	Αληθώς θετική (TP)	Ψευδώς αρνητική (FN)
-1	Ψευδώς θετική (FP)	Αληθώς αρνητική (TN)

Πίνακας 4.2: Πίνακας σύγχυσης

Εδώ να σημειωθεί ότι στις περιπτώσεις όπου έχουμε ψευδώς θετική και ψευδώς αρνητική έκβαση, η τιμή εξόδου του μοντέλου δηλαδή η προβλεπόμενη ετικέτα, δεν ταιριάζει με την παρατηρούμενη ετικέτα, αυτό σημαίνει ότι έχουμε λάθος αποτελέσματα. Οι αριθμοί κατά μήκος των κυρίων διαγωνίων, του πίνακα σύγχυσης (Πίνακας 4.2) αντιπροσωπεύουν τις σωστές αποφάσεις, και οι αριθμοί εκτός της διαγωνίου αντιπροσωπεύουν τα λάθη της σύγχυσης μεταξύ των διαφόρων κατηγοριών. Έτσι για ένα μοντέλο το οποίο δεν διαπράττει τυχόν λάθη, όλες οι προβλέψεις θα πρέπει να χαρτογραφηθούν στα πάνω αριστερά και κάτω δεξιά πεδία.

Μέσα από τον πίνακα σύγχυσης (Πίνακας 4.2), μπορούμε να υπολογίσουμε την ακρίβεια (accuracy και precision) του μοντέλου. Η ακρίβεια (accuracy) του συστήματος μέτρησης είναι ο βαθμός της εγγύτητας των μετρήσεων της ποσότητας με την πραγματική (αληθής) τιμή της ποσότητας του. Από την άλλη η ακρίβεια (precision) του συστήματος μέτρησης, που σχετίζεται με την επαναληψιμότητα και την αναπαραγωγικότητα, είναι ο βαθμός στον οποίο οι επανειλημμένες μετρήσεις υπό αμετάβλητες συνθήκες δείχνουν τα ίδια αποτελέσματα. Η ακρίβεια (accuracy) χρησιμοποιείται επίσης ως ένα στατιστικό μέτρο του πόσο καλά ένα τεστ σε μία δυαδική ταξινόμηση προσδιορίζει σωστά ή αποκλείει μια κατάσταση. Έτσι ορίζουμε την ακρίβεια (accuracy και precision) του μοντέλου ως εξής:

Ακρίβεια (Accuracy) είναι το πηλίκο των αληθώς θετικών (TP) και αληθώς αρνητικών (TN) δια τα τέσσερα πιθανά αποτελέσματα,

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

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

Ακρίβεια (Precision) είναι το πηλίκο των αληθώς θετικών προβλέψεων δια του αθροίσματος όλων των θετικών αποτελεσμάτων (αληθώς θετικά και ψευδώς θετικά).

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

Εκτός από την ακρίβεια έχουμε δύο άλλες μετρικές που συνήθως χρησιμοποιούνται για να χαρακτηρίσουμε την επίδοση του μοντέλου:

- την ευαισθησία (sensitivity)
- την ειδικότητα (specificity).

Ευαισθησία (Sensitivity): είναι το πηλίκο των αληθώς θετικών προβλέψεων δια του αθροίσματος όλων των θετικών παρατηρήσεων,

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

Ειδικότητα (Specificity): είναι οι αληθώς αρνητικές προβλέψεις διαιρούμενες με το άθροισμα όλων των αρνητικών παρατηρήσεων,

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

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

Μια ειδικότητα του 1,0 για ένα μοντέλο σημαίνει ότι το μοντέλο προβλέπει όλες τις αρνητικές παρατηρήσεις σωστά, με άλλα λόγια, το μοντέλο δεν διαπράττει καμία ψευδώς θετική πρόβλεψη.

Τα ποσοστά αυτά, δηλαδή η ευαισθησία και η ειδικότητα καθώς και τα συμπληρωματικά τους (ποσοστό ψευδώς αρνητικών (FNR) και ψευδώς θετικών αποτελεσμάτων (FPR), αντίστοιχα) ονομάζονται πιθανοφάνειες (likelihood) της διαγνωστικής δοκιμασίας.

Προφανώς ισχύει:

$$TPR = 1 - FNR$$

όπου

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

Αυτές οι 3 μετρικές αποτελούν τα κριτήρια απόδοσης του Confusion Matrix. Για ένα δυαδικό πρόβλημα ταξινόμησης με Μηχανές Διανυσμάτων Υποστήριξης (γραμμικό ή μη), ένας ταξινομητής θα πρέπει να παρέχει υψηλές τιμές των

κριτηρίων απόδοσης δηλαδή, της ευαισθησίας (sensitivity), της ειδικότητας (specificity) και της ακρίβειας (accuracy).

Άλλες δύο χρήσιμες έννοιες που αφορούν στους διαγνωστικούς ελέγχους είναι:

- η θετική προγνωστική αξία (positive predictive value) που συμβολίζεται με PPV:

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

- η αρνητική προγνωστική αξία (negative predictive value) που συμβολίζεται με NPV:

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

		Πραγματικές τιμές		
		Ψευδές (F)	Αληθές (T)	
Προβλεπόμενη τιμή (Αποτέλεσμα του τεστ)	Αρνητικό (N)	TN	FN	⇨ Αρνητική προγνωστική αξία (NPV)
	Θετικό (P)	FP	TP	⇨ Θετική προγνωστική αξία (PPV)
		⇩ Ειδικότητα (Specificity)	⇩ Ευαισθησία (Sensitivity)	Ακρίβεια (Accuracy)

Πίνακας 4.3: Συγκεντρωτικός πίνακας σύγχυσης

4.3 Εφαρμογή

Σε αυτή την ενότητα θα αξιολογήσουμε πειραματικά την εφαρμογή των Μηχανών Διανυσμάτων Υποστήριξης (SVMs) σε πρόβλημα ταξινόμησης δύο κλάσεων.

Για να παρέχουμε μία αμερόληπτη εκτίμηση για την ποιότητα ταξινόμησης του κάθε μοντέλου χρησιμοποιώντας τη μέθοδο της διάκρισης (discrimination), οι τιμές των κριτηρίων απόδοσης υπολογίζονται από ένα σύνολο δεδομένων που δεν χρησιμοποιήθηκε στη διαδικασία μοντελοποίησης. Για το σκοπό αυτό χρησιμοποιούμε από το πραγματικό σύνολο δεδομένων, ένα μέρος (σύνολο δοκιμής) το οποίο αφήσαμε στην άκρη για αυτό το σκοπό. Ο διαχωρισμός γίνεται 75% με 25% αντίστοιχα για το σύνολο εκπαίδευσης και το σύνολο δοκιμής.

4.3.1 Χειρισμός δεδομένων στην R

Αρχικά εισάγουμε τα δεδομένα στην R σε μορφή πίνακα, αφού προηγουμένως τα αποθηκεύσουμε σε ένα αρχείο σε μορφή txt. Εν συνεχεία εγκαθιστούμε στην R τα πακέτα 'e1071' καθώς και 'caret' ,τα οποία μας χρειάζονται για την εφαρμογή των μεθόδων των SVMs, μέσω των εντολών:

install.packages('e1071')

install.packages('caret')

και καλούμε τις αντίστοιχες βιβλιοθήκες με τις εντολές:

library('e1071')

library('caret').

Έπειτα καθορίζουμε τον τύπο των δεδομένων (data.frame) και χωρίζουμε τα δεδομένα μας σε

i. **δεδομένα εκπαίδευσης** (training set) (75%): στα οποία γνωρίζουμε την τιμή του αποτελέσματος και προσπαθούμε να κατασκευάσουμε ένα μοντέλο πρόβλεψης.
ii. **δεδομένα δοκιμής** (test set) (25%): το μοντέλο που δημιουργήσαμε θα το χρησιμοποιήσουμε στη συνέχεια για να προβλέψουμε το αποτέλεσμα νέων

συνόλων δεδομένων δοκιμής (test set), στα οποία σύνολα είναι γνωστές οι τιμές των χαρακτηριστικών αλλά δεν είναι γνωστή η τιμή του αποτελέσματος, δηλαδή η τιμή της κλάσης.

Στην συνέχεια για κάθε εφαρμογή θα εξετάσουμε τη συμπεριφορά του ταξινομητή μας με καθένα από τους τέσσερις πυρήνες, γραμμικό, ακτινικό, πολυωνυμικό και σιγμοειδή τόσο στο σύνολο εκπαίδευσης όσο και στο σύνολο δοκιμής και θα πάρουμε τις αντίστοιχες τιμές της ακρίβειας (Accuracy), ευαισθησίας (Sensitivity) και ειδικότητας (Specificity).

4.3.2 Περιγραφή των δεδομένων

Από το UCI Machine Learning Repository, χρησιμοποιήσαμε το σύνολο δεδομένων Parkinsons Data Set το οποίο προέρχονται από

'[Max A. Little, Patrick E. McSharry, Eric J. Hunter, Lorraine O. Ramig (2008), 'Suitability of dysphonia measurements for telemonitoring of Parkinson's disease', IEEE Transactions on Biomedical Engineering (to appear).]'

Αυτό το σύνολο δεδομένων αποτελείται από μια σειρά των βιοϊατρικών μετρήσεων της φωνής 31 ατόμων, εκ τον οποίον τα 23 έπασχαν από την νόσο του Πάρκινσον (PD). Κάθε χαρακτηριστικό είναι ένα συγκεκριμένο μέτρο της φωνής. Το σύνολο των δεδομένων μας είναι 195, με 23 χαρακτηριστικά τα οποία πάρθηκαν για κάθε άτομο από μια ηχογράφηση της φωνής.

Κύριος στόχος είναι να διακρίνουμε τους ανθρώπους που είναι υγιείς από εκείνους που πάσχουν με την νόσο PD. Αυτό γίνεται σύμφωνα με το χαρακτηριστικό Status, δηλαδή με την κατάσταση της υγείας του ανθρώπου η οποία είναι: ένα (1) αν έχει την νόσο του Πάρκινσον και μηδέν (0) αν είναι υγιής.

Πίνακας Δεδομένων

X1	MDVP: Fo (Hz) : Μέση φωνητική θεμελιώδης συχνότητα		**Επεξηγηματικές Μεταβλητές**
X2	MDVP: FHI (Hz) : Μέγιστη φωνητική θεμελιώδης συχνότητα		
X3	MDVP: Flo (Hz) : Ελάχιστη φωνητική θεμελιώδης συχνότητα		
X4	MDVP: Jitter (%)	Μέτρα μεταβολής της θεμελιώδη συχνότητα	
X5	MDVP: Jitter (ABS)		
X6	MDVP: RAP		
X7	MDVP: PPO		
X8	Jitter: DDP		
X9	MDVP: Shimmer	Μέτρα μεταβολής του εύρους (πλάτους)	
X10	MDVP: Shimmer (dB)		
X11	Shimmer: APQ3		
X12	Shimmer: APQ5		
X13	MDVP: APQ		
X14	Shimmer: DDA		
X15	NHP	Μέτρα του δείκτη θορύβου στα τονικά στοιχεία της φωνή	
X16	HNR		
X17	RPDE	Μη γραμμικά δυναμικά μέτρα πολυπλοκότητα	
X18	D2		
X19	DFA: Σήμα fractal εκθετικής κλιμάκωσης		
X20	Spread1	Μη γραμμικά μέτρα των θεμελιώδη διακυμάνσεων της συχνότητας	
X21	Spread2		
X22	PPE		
Y	Status: Η κατάσταση της υγείας του ανθρώπου: ένα (1) αν έχει την νόσο του Πάρκινσον, μηδέν (0) αν είναι υγιής		**Μεταβλητή απόκρισης**

Πίνακας 4.4: Περιγραφή του συνόλου δεδομένων

Έτσι αφού δημιουργήσουμε το μοντέλο με τη βοήθεια του SVM συγκρίνουμε τις διαφορετικές τιμές των accuracy, sensitivity και specificity για τα δύο σύνολα αλλά και για τους διάφορους πυρήνες των SVM (Linear, Radial, Polynomial, Sigmoid). Τα αποτελέσματα παρουσιάζονται στον πιο κάτω Πίνακα 4.5. Εδώ να σημειωθεί ότι σε κάθε περίπτωση διατηρούμε την τιμή του κόστους σταθερή και ίση με 1.

	Accuracy		**Sensitivity**		**Specificity**	
	Train	**Test**	**Train**	**Test**	**Train**	**Test**
Linear	0.9252	0.8125	0.9735	0.9706	0.7647	0.4286
Radial	0.9184	0.8333	1.0000	1.0000	0.6471	0.4286
Polynomial	0.8844	0.7917	1.0000	0.9706	0.5000	0.3571
Sigmoid	0.8707	0.7917	0.9381	0.9412	0.6471	0.4286

Πίνακας 4.5: Πίνακας αξιολόγησης

Από τον πίνακα αξιολόγησης (Πίνακας 4.5) προκύπτουν τα εξής συμπεράσματα:

- Σε κάθε περίπτωση η ακρίβεια (accuracy) είναι ικανοποιητική. Παρατηρούμε ότι υψηλότερη ακρίβεια για το σύνολο εκπαίδευσης παρουσιάζει ο γραμμικός πυρήνας (linear) ενώ για το σύνολο δοκιμών ο ακτινικός πυρήνας (radial).

- Η ευαισθησία (sensitivity) του μοντέλου επίσης είναι ικανοποιητική. Ο ακτινικός πυρήνας (radial) τόσο στο σύνολο εκπαίδευσης όσο και στο σύνολο δοκιμών παίρνει την υψηλότερη τιμή (1.0). Σε κάθε περίπτωση ο δείκτης της ευαισθησίας είναι πάνω από 0.9 και αυτό οδηγεί στο συμπέρασμα ότι το μοντέλο μας διαπράττει ελάχιστες ψευδώς αρνητικές προβλέψεις.

- Ο δείκτης ειδικότητας (specificity) θα μπορούσαμε να πούμε ότι είναι χαμηλός σε κάθε περίπτωση. Αυτό οδηγεί στο συμπέρασμα ότι το μοντέλο μας διαπράττει αρκετές ψευδώς θετικές προβλέψεις.

Με μια πιο προσεκτική ματιά θα μπορούσαμε να πούμε ότι ο γραμμικός πυρήνας (linear) και ο ακτινικός (radial) παρουσιάζουν αυξημένη ακρίβεια και ευαισθησία

μεταξύ των υπολοίπων και άρα είναι οι καλύτεροι για την ταξινόμηση των δεδομένων μας.

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

Confusion Matrix and Statistics

	Reference	
Prediction	0	1
0	26	3
1	8	110

Από το πιο πάνω συμπεραίνουμε ότι το μοντέλο μας προέβλεψε 26 τιμές μηδέν σωστά (TN=26), 3 τιμές προέβλεψε μηδέν αλλά λάθος (FN=3), 8 τιμές προέβλεψε θετικές (δηλαδή με ετικέτα 1) αλλά λάθος (FP=8) και 110 τιμές προέβλεψε θετικές σωστά (TP=110). Έτσι, μπορούμε να βρούμε τις τιμές accuracy, sensitivity και specificity από τους τύπους που προαναφέραμε:

$$accuracy = \frac{TN + TP}{TN + FP + FN + TP} = \frac{26 + 110}{26 + 3 + 8 + 110} = 0{,}9252$$

$$sensitivity = \frac{TP}{TP + FN} = \frac{110}{110 + 3} = 0{,}9735$$

$$specificity = \frac{TN}{TN + FP} = \frac{26}{26 + 8} = 0{,}7647$$

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

```
          Accuracy : 0.9252
            95% CI : (0.8701, 0.9621)
   No Information Rate : 0.7687
   P-Value [Acc > NIR] : 4.697e-07

             Kappa : 0.7782
 Mcnemar's Test P-Value : 0.2278

        Sensitivity : 0.9735
        Specificity : 0.7647
     Pos Pred Value : 0.9322
     Neg Pred Value : 0.8966
         Prevalence : 0.7687
     Detection Rate : 0.7483
 Detection Prevalence : 0.8027
    Balanced Accuracy : 0.8691

     'Positive' Class : 1
```

Με τον ίδιο τρόπο εκτελούμε για το σύνολο εκπαίδευσης και το σύνολο δοκιμής για όλους του πυρήνες, παίρνουμε τα αποτελέσματα στην R όπως προηγουμένως και συμπληρώνουμε τον συγκεντρωτικό πίνακα αξιολόγησης (Πίνακας 4.5).

4.4 Συγκρίσεις των μεθόδων ψ-μάθηση και SVM

Τώρα θα συγκρίνουμε τις δυο τεχνικές μας SVM και ψ-μάθηση μέσα από 3 σύνολα δεδομένων τα οποία προέρχονται από το UCI Machine Learning Repository και αποτελούν προβλήματα για δυαδική ταξινόμηση. Αυτά τα δεδομένα θα τα χρησιμοποιήσουμε για να συγκρίνουμε τα αποτελέσματα των δυο μεθόδων όσον αφορά το σφάλμα δοκιμής, το πλήθος των διανυσμάτων υποστήριξης καθώς και το μέγεθος της βελτίωσης της ψ-μάθησης έναντι της SVM.

Εδώ να επισημάνουμε ότι το σφάλμα δοκιμής (ή ισοδύναμα σφάλμα γενίκευσης), είναι μια λειτουργία που μετρά πόσο καλά μια μηχανική μάθηση γενικεύεται σε πρωτοεμφανιζόμενα δεδομένα. Μετράται ως η απόσταση μεταξύ του σφάλματος σχετικά με το σύνολο εκπαίδευσης και το σύνολο των δοκιμών και είναι κατά μέσο όρο για το σύνολο των πιθανών στοιχείων κατάρτισης που μπορούν να δημιουργηθούν μετά από κάθε επανάληψη της διαδικασίας μάθησης. Έχει αυτό το όνομα επειδή η λειτουργία αυτή δείχνει την ικανότητα μιας μηχανής που μαθαίνει με τον προκαθορισμένο αλγόριθμο να συναγάγει έναν κανόνα (ή να γενικεύσει) που χρησιμοποιείται από το μηχάνημα να παράγει δεδομένα που βασίζονται μόνο σε μερικά παραδείγματα. Επίσης το μέγεθος της βελτίωσης της ψ-μάθησης έναντι της SVM ορίζεται ως εξής:

$$\frac{\big(T(SVM)-T(Bayes)\big)-\big(T(\psi)-T(Bayes)\big)}{T(SVM)-T(Bayes)} = \frac{T(SVM)-T(\psi)}{T(SVM)-T(Bayes)} \qquad (4.3.1)$$

όπου $T(SVM)$ είναι το σφάλμα δοκιμής της SVM, $T(\psi)$ το σφάλμα δοκιμής της ψ-μάθησης και $T(Bayes)$ είναι το σφάλμα δοκιμής του Bayes.

Όμως στην περίπτωση μας επειδή το σφάλμα δοκιμής του Bayes είναι απροσδιόριστο, το μέγεθος της βελτίωσης υπολογίζεται από τον τύπο:

$$\frac{T(SVM)-T(\psi)}{T(SVM)} \qquad (4.3.2)$$

Για την εφαρμογή των συγκρίσεων των δυο μεθόδων σε πραγματικά δεδομένα, όπως έχουμε προαναφέρει θα χρησιμοποιήσουμε 3 σύνολα δεδομένων από το UCI Machine Learning Repository:

i Wisconsin Breast Cancer
ii Liver-Disorders
iii Page-Block

4.4.1 Περιγραφή Δεδομένων

<u>Wisconsin Breast Cancer</u>

Αυτά τα δεδομένα συλλέχθηκαν από τον Dr. William H. Wolberg, του Πανεπιστημίου του Wisconsin Hospital, Madison και αφορούν ασθενής με καρκίνο του μαστού. Τα δείγματα συλλέχθηκαν περιοδικά από κλινικές περιπτώσεις και τα δείγματα αποτελούνται από ψηφιακή εικόνα με την οποία αξιολογήθηκαν τα πυρηνικά χαρακτηριστικά μέσω της αναρρόφησης λεπτής

βελόνας (FNAs) που λαμβάνονται από τα στήθη των ασθενών. Κάθε δείγμα έχει εκχωρηθεί σε ένα 9 διαστάσεων διάνυσμα. Κάθε συστατικό είναι στο διάστημα 1 έως 10, με την τιμή 1 να αντιστοιχεί σε κανονική κατάσταση ενώ η τιμή 10 σε μια πιο φυσιολογική κατάσταση. Το σύνολο των δεδομένων μας αποτελείται από 699 παρατηρήσεις και από 10 χαρακτηριστικά.

Κύριος στόχος είναι να διαπιστώσουμε αν ο καρκίνος είναι καλοήθης ή κακοήθης.

Πίνακας Δεδομένων

<table>
<tr><td>X1</td><td colspan="2">Πάχος της συστάδας:1 – 10</td><td rowspan="10">Επεξηγηματικές μεταβλητές</td></tr>
<tr><td>X2</td><td>Μέγεθος: 1 – 10</td><td rowspan="2">Ομοιομορφία των κυττάρων</td></tr>
<tr><td>X3</td><td>Σχήμα: 1 – 10</td></tr>
<tr><td>X4</td><td colspan="2">Οριακή προσκόλληση: 1 – 10</td></tr>
<tr><td>X5</td><td colspan="2">Μέγεθος ενιαίων επιθηλιακών κυττάρων: 1–10</td></tr>
<tr><td>X6</td><td colspan="2">Πυρήνες Bare: 1 – 10</td></tr>
<tr><td>X7</td><td colspan="2">Μειλίχιο χρωματίνης: 1 – 10</td></tr>
<tr><td>X8</td><td colspan="2">Κανονική πυρηνίσκοι: 1 – 10</td></tr>
<tr><td>X9</td><td colspan="2">Μιτώσεις: 1 – 10</td></tr>
<tr><td>Y</td><td colspan="2">Κατηγορία: 2 για καλοήθη και 4 για κακοήθη</td><td>Μεταβλητή απόκρισης</td></tr>
</table>

Πίνακας 4.6: Περιγραφή του συνόλου των δεδομένων Wisconsin Breast Cancer

Liver-Disorders

Αυτό το σύνολο δεδομένο αφορά της διαταραχές του ύπατος. Αποτελείται από 345 παρατηρήσεις για τις οποίες υπάρχουν 7 χαρακτηριστικά. Τα 5 πρώτα χαρακτηριστικά αφορούν εξετάσεις αίματος και πιστεύεται ότι αυτά τα χαρακτηριστικά είναι ευαίσθητα σε διαταραχές του ύπατος που μπορεί να προκληθούν από την υπερβολική κατανάλωση αλκοόλ. Κάθε εγγραφή αποτελείται μόνο από ένα αρσενικό άτομο.

Στόχος είναι να διαπιστώσουμε την παρουσία ή απουσία της διαταραχής ύπατος.

Πίνακας Δεδομένων

X1	MCV μέσος όγκος ερυθρών	**Επεξηγηματικές μεταβλητές**
X2	Alkphos: αλκαλική φωσφοτάσης	
X3	SGPT: αλαμίνη αμινοτρανσφεράσης	
X4	SGOT: ασπαρτική αμινοτρανσφεράσης	
X5	Gammagt: γ-γλουταμυλοτρανσφεράσης	
X6	ο αριθμός των μισών ποτηριών, των αλκοολούχων ποτών που ήπιε ανά ημέρα	
Y	Πρόβλεψη: ένα (1) αν έχουμε απουσία της διαταραχή και δυο (2) αν έχουμε παρουσία της διαταραχής	**Μεταβλητή απόκρισης**

Πίνακας 4.7: Περιγραφή του συνόλου των δεδομένων Liver-Disorders

Page-Block

Αυτό το δείγμα αποτελεί ένα πρόβλημα ταξινόμησης όλων των μπλοκ της σελίδας για την διάταξη ενός εγγράφου που έχει ανιχνευθεί από τον κατακερματισμό της διαδικασία. Οι πέντε κατηγορίες είναι:

(1) κείμενο
(2) οριζόντια γραμμή
(3) εικόνα
(4) κάθετη γραμμή
(5) γραφικές

Τα δεδομένα αποτελούνται από 5473 παρατηρήσεις και προέρχονται από 54 διακριτά έγγραφα, με 10 χαρακτηριστικά. Κάθε παρατήρηση αφορά ένα τετράγωνο και όλα τα χαρακτηριστικά είναι αριθμητικά.

Εμείς για την εν λόγω εφαρμογή επιλέγουμε τις οριζόντιες γραμμές (2) με 329 περιπτώσεις και τις εικόνες (3) με 115 περιπτώσεις περιορίζοντας το πρόβλημα μας σε ένα δυαδικό πρόβλημα ταξινόμησης.

Στόχος είναι να διακρίνουμε αν το μπλοκ αποτελείται από εικόνες (3) ή οριζόντιες γραμμές (2).

Πίνακας Δεδομένων

X1	height: Ύψος του μπλοκ.	**Επεξηγηματικές μεταβλητές**
X2	lenght: Μήκος του μπλοκ.	
X3	area: Περιοχή του μπλοκ (ύψος × μήκος)	
X4	eccen: Εκκεντρικότητα του μπλοκ (μήκος/ύψος)	
X5	blackpix: Συνολικός αριθμός μαύρων εικονοστοιχείων στην αρχική bitmap του μπλοκ.	
X6	p_black: Ποσοστό των μαύρων pixels εντός του μπλοκ (blackpix/area)	
X7	blackand: Συνολικός αριθμός μαύρων εικονοστοιχείων στο bitmap του μπλοκ μετά την RLSA.	
X8	wb_trans: Αριθμός λευκό-μαύρο εναλλαγών στο αρχικό bitmap του μπλοκ.	
X9	p_and: Ποσοστό των μαύρων pixels, μετά την εφαρμογή του RLSA (blackand/area)	
X10	mean_tr: Μέσος αριθμός των λευκών-μαύρων εναλλαγών (blackpix/wb_trans)	
Y	Τρία (3) αν είναι εικόνα και δυο (2) αν είναι οριζόντιες γραμμές	**Μεταβλητή απόκρισης**

Πίνακας 4.8: Περιεχόμενα του συνόλου των δεδομένων Page-Block

4.4.2 Αποτελέσματα

Τα 3 σύνολα δεδομένων που επιλέξαμε τα χωρίζουμε τυχαία σε δύο ίσα μέρη (50% με 50% διαχωρισμός) σε σύνολο εκπαίδευσης και σύνολο δοκιμών.

Σε κάθε περίπτωση ελαχιστοποιούμε την τιμή της ρυθμιστικής παραμέτρου C στο διάστημα $(0,10^4]$.

Εφαρμόζουμε τα 3 σύνολα δεδομένων σε 3 διαφορετικές περιπτώσεις:

i Σε μια δυαδική ταξινόμηση με την χρήση του γραμμικού πυρήνα

$$K(x, y) = \langle x, y \rangle$$

Αυτή την περίπτωση την εφαρμόζουμε στα δύο από τα τρία σύνολα δεδομένων μας, συγκεκριμένα στα σύνολα δεδομένων Wisconsin Breast Cancer και Liver-disorders. Τα αποτελέσματα που προκύπτουν σύμφωνα με Liu 2006 παρουσιάζονται στον πιο κάτω Πίνακα 4.9.

Δεδομένα		Testing Error	Πλήθος SV
WBC 699×9	SVM ψ-learning Βελτίωση	3.48(0.09)% 3.12(0.08)% 10.3%	32.49 15.76
Liver 345×6	SVM ψ-learning Βελτίωση	32.00(0.29)% 30.38(0.28)% 5.1%	123.46 51.69

Πίνακας 4.9: Μέσος όρος των σφαλμάτων δοκιμής της SVM και ψ-μάθησης καθώς και το πλήθος των διανυσμάτων υποστήριξης στην περίπτωση γραμμικού πυρήνα. Στις παρενθέσεις είναι τα τυπικά σφάλματα.

ii Σε μια δυαδική ταξινόμηση με την χρήση του πολυωνυμικού πυρήνα

$$K(x, y) = (1 + \langle x, y \rangle)^2$$

και την εφαρμόζουμε στο σύνολο δεδομένων Liver-disorders. Τα αποτελέσματα που προκύπτουν σύμφωνα και πάλι με Liu (2006) παρουσιάζονται στον Πίνακα 4.10.

Δεδομένα		Testing Error	Πλήθος SV
Liver 345×6	SVM ψ-learning Βελτίωση	27.46(0.225)% 26.63(0.210)% 3.0%	101.29 53.13

Πίνακας 4.10: Παρουσιάζονται τα σφάλματα δοκιμής και το πλήθος των διανυσμάτων υποστήριξης και σε παρένθεση τα τυπικά σφάλματα, με την χρήση του πολυωνυμικού πυρήνα.

iii Σε μια δυαδική ταξινόμηση με την χρήση του πυρήνα ακτινικής βάσης (Gaussian Radial Basis Kernel)

$$K(x,y) = \exp\left(-\frac{1}{\sigma^2}\|x-y\|^2\right)$$

όπου η παράμετρος σ είναι η διάμεσος της απόστασης μεταξύ των δυο κλάσεων, και την εφαρμόζουμε στο σύνολο δεδομένων Pages-Block. Τα αποτελέσματα που προκύπτουν σύμφωνα και πάλι με Liu (2006) παρουσιάζονται στον Πίνακα 4.11.

Δεδομένα		Testing Error	Πλήθος SV
Page Block 345×6	SVM ψ-learning Βελτίωση	5.06(0.202)% 4.96(0.222)% 1.8%	50.46 43.50

Πίνακας 4.11: Παρουσιάζονται τα σφάλματα δοκιμής και το πλήθος των διανυσμάτων υποστήριξης και σε παρένθεση τα τυπικά σφάλματα, με την χρήση του πυρήνα ακτινικής βάσης.

Μέσα από αυτά τα αποτελέσματα οδηγούμαστε στα εξής συμπεράσματα:

- Στον Πίνακα 4.9 βλέπουμε ότι η ψ-μάθηση έχει μικρότερο σφάλμα δοκιμής, σε κάθε περίπτωση, έτσι έχει την καλύτερη ικανότητα γενίκευσης σε σύγκριση με την SVM. Επιπρόσθετα, η ψ-μάθηση αποδίδει ένα μικρότερο αριθμό διανυσμάτων υποστήριξης. Αυτό έχει ως αποτέλεσμα η ψ-μάθηση να παρουσιάζει μια ισχυρότερη ικανότητα μείωσης των δεδομένων από εκείνη της SVM, δηλαδή για μεγαλύτερο αριθμό δεδομένων παρουσιάζει καλύτερα αποτελέσματα από ότι η SVM. Επιπλέον, κάθε ταξινομητής με απεριόριστη συνάρτηση απώλειας, όπως η SVM, «πάσχει» από τις ακραίες τιμές. Αυτό ενισχύει την άποψη ότι η ψ-μάθηση είναι πιο ισχυρή στην περίπτωση των ακραίων τιμών. Έτσι οδηγούμαστε στο συμπέρασμα ότι στην περίπτωση του γραμμικού πυρήνα η ψ-μάθηση έχει καλύτερα αποτελέσματα και μεγαλύτερη ικανότητα γενίκευσης από ότι η SVM. Αυτό συμφωνεί απόλυτα με τα θεωρητικά μας αποτελέσματα του τρίτου κεφαλαίου.

- Στον Πίνακα 4.10 βλέπουμε ότι και στην περίπτωση του πολυωνυμικού πυρήνα η ψ–μάθηση έχει μικρότερο σφάλμα δοκιμών και πλήθος διανυσμάτων, αλλά σε σχέση με τον γραμμικό πυρήνα η βελτίωση της ψ-μάθησης έναντι της SVM μειώνεται. (στο γραμμικό πυρήνα για το σύνολο δεδομένων Liver είχαμε 5.1% βελτίωση ενώ στον πολυωνυμικό 3.0%)

- Στον Πίνακα 4.11 βλέπουμε ότι και στην περίπτωση του πυρήνα ακτινικής βάσης και πάλι η ψ-μάθηση παρουσιάζει μικρότερο σφάλμα δοκιμών έναντι της SVM καθώς και μικρότερο πλήθος διανυσμάτων υποστήριξης.

Τα τελικά μας συμπεράσματα είναι ότι σε κάθε περίπτωση όπως ήταν αναμενόμενο από την θεωρία που έχουμε αναπτύξει, η ψ-μάθηση ξεπερνά την SVM από την άποψη ότι τα σφάλματα δοκιμών της προκύπτουν να είναι μικρότερα από αυτά της SVM το ίδιο και το πλήθος των διανυσμάτων υποστήριξης. Επίσης αξιοσημείωτο είναι το γεγονός ότι για μεγαλύτερα n φαίνεται να παρουσιάζει μια μεγαλύτερη βελτίωση έναντι της SVM. Στους Πίνακες 4.10 και 4.11 το ποσοστό βελτίωσης της ψ-μάθησης έναντι της SVM παρατηρούμε ότι είναι πολύ μέτριο. Αυτό μπορεί να οφείλεται στο γεγονός ότι το περίσσιο ποσοστό σφάλματος (4.3.1) της SVM πάνω από το ποσοστό σφάλματος του Bayes είναι πιθανόν μικρότερο από 10%. Όπως έχουμε προαναφέρει εμείς χρησιμοποιούμε τον τύπο (4.3.2) για το μέγεθος της βελτίωσης της ψ-μάθησης έναντι της SVM, το οποίο αποτελεί μόνο ένα κάτω φράγμα της (4.3.1). Αυτό έχει σαν αποτέλεσμα να εμφανίζεται ως βελτίωση της ψ-μάθησης μόνο ένα μικρό ποσοστό.

Κεφάλαιο 5

Επίλογος Γενικά Συμπεράσματα

Μέσα από την παρούσα διπλωματική έχουμε προτείνει μια νέα τεχνική μάθησης η οποία όπως έχουμε δει τόσο σε θεωρητικό όσο και σε αριθμητικό επίπεδο είναι καλύτερη από την γνωστή SVM. Αν και η ψ-μάθηση παρουσιάζει δυσκολία από την άποψη ότι η βελτιστοποίηση της που προκύπτει είναι ένα μη κυρτό πρόβλημα ελαχιστοποίησης παρ' όλα αυτά η βελτίωση που παρουσιάζει έναντι της SVM είναι αξιοσημείωτη. Μέσα από τις εφαρμογές μας διαπιστώνουμε ότι σε πραγματικά δεδομένα αυτή η νέα τεχνική αποδίδει καλύτερη γενίκευση, δηλαδή μικρότερα σφάλματα δοκιμών, γεγονός που δηλώνει ότι η ψ-μάθηση γενικεύεται καλύτερα σε πρωτοεμφανιζόμενα δεδομένα το οποίο είναι και το ζητούμενο σε μια μηχανική μάθηση. Αξιοσημείωτο είναι το γεγονός ότι για μεγάλα n, δηλαδή για μεγάλο αριθμό δεδομένων αυτή η τεχνική αποδίδει ακόμη καλύτερα αποτελέσματα και μια μεγαλύτερη βελτίωση.

Παράρτημα Α

(Αποδείξεις)

Απόδειξη θεωρήματος 3.3.1: Με την εισαγωγή n μεταβλητών χαλάρωσης $\xi_i, i = 1, \dots, n$, το k-οστό πρόβλημα (3.3.4) είναι ισοδύναμη με:

$$\min_{\boldsymbol{w}} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n}\xi_i - \langle \boldsymbol{w}, V_1^{(k)}\rangle - \langle b, V_2^{(k)}\rangle$$

με τους εξής περιορισμούς (A.1)

$$\xi_i \geq 2\big(1 - y_i f(x_i)\big)$$

$$\xi_i \geq 0$$

ή ισοδύναμα, $y_i(\langle \boldsymbol{w}, x_i\rangle) \geq 1 - \frac{1}{2}\xi_i\,,\ i = 1, \dots, n.$

Για την επίλυση της (A.1), εισάγουμε τους πολλαπλασιαστές Lagrange $\boldsymbol{\alpha} = (\alpha_1, .., \alpha_n)$ καθώς και $\boldsymbol{\beta} = (\beta_1, \dots, \beta_n)$ που προκύπτουν από τους περιορισμούς $\xi_i \geq 0$. Έτσι οδηγούμαστε στο εξής πρόβλημα Lagrangian

$L(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{w}, \xi, b) = \frac{1}{2}\langle \boldsymbol{w}, \boldsymbol{w}\rangle + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i\left[y_i(\langle \boldsymbol{w}, x_i\rangle) - 1 + \frac{1}{2}\xi_i\right] - \sum_{i=1}^{n}\beta_i\xi_i - \langle \boldsymbol{w}, V_1^{(k)}\rangle - \langle b, V_2^{(k)}\rangle$ (A.2)

όπου $\alpha_i \geq 0$ και $\beta_i \geq 0,\ i = 1, \dots, n.$

Στην συνέχεια διαφορίζουμε την $L(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{w}, \xi, b)$ ως προς $\boldsymbol{w}, b, \xi$ και θέτουμε τις παραγώγους ίσες με μηδέν:

$$\frac{\partial L(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{w}, \xi, b)}{\partial \boldsymbol{w}} = 0 \implies \boldsymbol{w} = V_1^{(k)} + \sum_{i=1}^{n}\alpha_i y_i x_i \quad \text{(A.3)}$$

$\frac{\partial L(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{w},\xi,b)}{\partial b} = 0 \implies C - \frac{1}{2}\alpha_i - \beta_i = 0$ (A.4)

$\frac{\partial L(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{w},\xi,b)}{\partial \xi} = 0 \implies \sum_{i=1}^{n} \alpha_i y_i + V_2^{(k)} = 0 \Longrightarrow \sum_{i=1}^{n} \alpha_i y_i = -V_2^{(k)}$ (A.5)

$\Longrightarrow$ από την (A.5) προκύπτει ο περιορισμός

$$\sum_{i=1}^{n} \alpha_i y_i = -V_2^{(k)}$$

Στην συνέχεια με αντικατάσταση των (A.3) και (A.4) στην (A.2) παίρνουμε:

$$L(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{w}, \xi, b) = \sum_{i=1}^{n} \alpha_i \left[1 - y_i \langle V_1^{(k)}, x_i \rangle\right] - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle\right) - \frac{1}{2}\left(V_1^{(k)}, V_1^{(k)}\right)$$

Αγνοώντας τον σταθερό όρο $\frac{1}{2}\left(V_1^{(k)}, V_1^{(k)}\right)$ τελικά παίρνουμε

$$L(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{w}, \xi, b) = \sum_{i=1}^{n} \alpha_i \left[1 - y_i \langle V_1^{(k)}, x_i \rangle\right] - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\, \alpha_j y_i y_j \langle x_i, x_j \rangle$$

που αποτελεί την εξίσωση (3.3.11).

Για τους απομένοντα περιορισμούς, από την (A.4)

$$C - \frac{1}{2}\alpha_i - \beta_i = 0$$

με $\beta_i \geq 0$ και $\alpha_i \geq 0$

προκύπτει ότι $0 \leq \alpha_i \leq 2C$.

Επιπλέον, για $\xi_i \neq 0$ συνεπάγεται ότι $\beta_i = 0$ και άρα $\alpha_i = 2C$.

Ως εκ τούτου, οι συνθήκες Karush-Kuhn-Tucker γίνονται

$$\alpha_i \left[y_i \langle \boldsymbol{w}, x_i \rangle - 1 + \frac{1}{2}\xi_i\right] = 0$$

$$\xi_i[\alpha_i - 2C] = 0 \quad \text{για i=1,..,n.}$$

Επιπλέον, από τις συνθήκες Karush-Kuhn-Tucker συνεπάγεται ότι $\xi_i \neq 0$ μπορεί να συμβεί μόνο όταν $\alpha_i = 2C$. Αυτό ολοκληρώνει την απόδειξη.

Απόδειξη του θεωρήματος 3.3.2: Η απόδειξη είναι ουσιαστικά η ίδια όπως εκείνη στο θεώρημα 3.3.1 με ελαφρές τροποποιήσεις, και έτσι παραλείπεται.

Απόδειξη του θεωρήματος 3.5.1: Αρχικά εισάγουμε κάποιους συμβολισμούς που είναι απαραίτητοι για την παρούσα απόδειξη.

Έστω

$$\tilde{l}_\psi(f, Z_i) = l_\psi(f, Z_i) + \lambda J(f)$$

είναι η αντικειμενική συνάρτηση που πρέπει να ελαχιστοποιηθεί, όπως στην (3.3), όπου

$$\tilde{l}_\psi(f, Z_i) = \psi\big(Y_i f(X_i)\big)$$

και $\lambda = 1/(C_n)$.

Έστω

$$\tilde{l}(f, Z_i) = l(f, Z_i) + \lambda J(f)$$

εναι η αντίστοιχη αντικειμενική συνάρτηση που καθορίζεται από Sign(·), όπου

$$l(f, Z_i) = \Big(1 - Sign\big(Y_i f(X_i)\big)\Big)$$

Ορίζουμε την εμπειρική διαδικασία κλιμάκωσης, $E_n\left(\tilde{l}(f, Z) - \tilde{l}_\psi(f_0, Z)\right)$, ως

$$n^{-1}\sum_{i=1}^{n}\left(\tilde{l}(f, Z_i) - \tilde{l}_\psi(f_0, Z_i) - E\left(\tilde{l}(f, Z_i) - \tilde{l}_\psi(f_0, Z_i)\right)\right) = E_n\left(l(f, Z) - l_\psi(f_0, Z)\right)$$

όπου Z=(X,Y).

Έστω,

$$A_{i,j} = \big\{f \in \mathcal{F}: 2^{i-1}\delta_n^2 \le e\big(f, \bar{f}\big) < 2^i\delta_n^2, 2^{j-1}max(J(f_0), 1) \le J(f) < 2^j max(J(f_0), 1)\big\}$$

και

$$A_{i,0} = \{f \in \mathcal{F}: 2^{i-1}\delta_n^2 \le e(f,\bar{f}) < 2^i\delta_n^2, J(f) < max(J(f_0),1)\}$$

για j=1,2,.... και i=1,2,...

Χωρίς βλάβη της γενικότητας, υποθέτουμε ότι $J(f_0) \ge 1$.

Επίσης, επειδή $e(f,\bar{f}) \ge 1/2$ για κάθε f, υποθέτουμε ότι $max(\varepsilon_n^2, 2s_n) < 1$.

Η προσέγγισή μας για την εύρεση ενός άνω φράγματος της πιθανότητα $P(e(\hat{f},\bar{f}) \ge \delta_n^2)$ είναι να μειωθεί το πρόβλημα μας σε μια ακολουθία εμπειρικών διαδικασιών που επάγονται από την αντικειμενική συνάρτηση $\tilde{l}$.

Σύμφωνα με το Θεώρημα 3 των Shen, X., and Wong, W. H. (1994) (Convergence Rate of Sieve Estimates), μία μεγάλη απόκλιση της ανισότητας για εμπειρικές διαδικασίες, για το όριο της πιθανότητα $P(A_{ij})$, $i,j = 1,\dots,n$ ελέγχεται από την μέση τιμή και την διασπορά, που ορίζονται από $l(f,Z_i)$ και την ποινή λ. Αυτό αποδίδει μία ανισότητα για την ακολουθία των εμπειρικών διαδικασιών και ως εκ τούτου για το $e(\hat{f},\bar{f})$.

Μέχρι στιγμής έχουμε δημιουργήσει μια σύνδεση μεταξύ της ακρίβειας τη μάθησης, $e(\hat{f},\bar{f})$ και των εμπειρικών διαδικασιών. Με την Υπόθεση Δ παίρνουμε,

$$\tilde{l}_\psi(f_0,Z_i) - \tilde{l}(f,Z_i) \ge \tilde{l}_\psi(f_0,Z_i) - \tilde{l}_\psi(f,Z_i) \qquad \text{(A.6)}$$

για i=1,...,n.

Από το γεγονός ότι η $\hat{f}$ είναι η μεγιστοποιεί της $-n^{-1}\sum_{i=1}^n \tilde{l}_\psi(f,Z_i)$,

$e(f_0,\bar{f}) \le e_\psi(f_0,\bar{f}) \le \delta_n^2$ και από την (A.6) προκύπτει ότι:

$$\{e(\hat{f},\bar{f}) \ge \delta_n^2\} \subset \left\{ \sup_{\{f\in\mathcal{F}: e(f,\bar{f})\ge\delta_n^2\}} n^{-1}\sum_{i=1}^n \left(\tilde{l}_\psi(f_0,Z_i) - \tilde{l}_\psi(f,Z_i)\right) \ge 0 \right\}$$
$$\subset \left\{ \sup_{\{f\in\mathcal{F}: e(f,\bar{f})\ge\delta_n^2\}} n^{-1}\sum_{i=1}^n \left(\tilde{l}_\psi(f_0,Z_i) - \tilde{l}(f,Z_i)\right) \ge 0 \right\}$$

Ως εκ τούτου,

$$P(e(\hat{f},\bar{f}) \ge \delta_n^2) \le P^* \left(\sup_{\{f\in\mathcal{F}: e(f,\bar{f})\ge\delta_n^2\}} n^{-1} \times \sum_{i=1}^n \left(\tilde{l}_\psi(f_0,Z_i) - \tilde{l}(f,Z_i)\right) \ge 0 \right) = I$$

όπου P* δηλώνει το εξωτερικό μέτρο πιθανότητας. Για τον όριο του I, αρκεί να φράξουμε την αντίστοιχη πιθανότητα του $A_{i,j}$ για κάθε i,j=1,.... Για το σκοπό

αυτό, χρειαζόμαστε κάποιες ανισότητες όσον αφορά τις πρώτες και δεύτερες ροπές της διαφορά $\tilde{l}(f,Z_i)-\tilde{l}_\psi(f_0,Z_i)$ για $f \in A_{i,j}$ (οι πρώτες και δεύτερες ροπές της κατανομής).

Εδώ να σημειωθεί ότι για την πρώτη ροπή έχουμε:

$$E\left(l(f,Z)-l_\psi(f_0,Z)\right)=E\left(l(f,Z)-l_\psi(\bar{f},Z)\right)-E\left(l_\psi(f_0,Z)-l_\psi(\bar{f},Z)\right)$$

το οποίο από το γεγονός ότι $El_\psi(\bar{f},Z)=El(\bar{f},Z)$ είναι ίσο με $2\left(e(f,\bar{f})-e_\psi(f_0,\bar{f})\right)$, δηλαδή $E\left(l(f,Z)-l_\psi(f_0,Z)\right)=2\left(e(f,\bar{f})-e_\psi(f_0,\bar{f})\right)$.

Από την Υπόθεση Α έχουμε ότι $2e_\psi(f_0,\bar{f})\le 2s_n\le\delta_n^2$. Στη συνέχεια, χρησιμοποιώντας την υπόθεση ότι $max(J(f_0),1)\lambda\le\delta_n^2/2$, για οποιουσδήποτε ακέραιους $i,j\ge 1$, και χρησιμοποιώντας το γεγονός ότι $2^i-1\ge 2^{i-1}$ παίρνουμε τα εξής:

$$\inf_{A_{i,j}} E\left(\tilde{l}(f,Z)-\tilde{l}_\psi(f_0,Z)\right)\ge M(i,j)=\left(2^{i-1}\delta_n^2\right)+\lambda\left(2^{j-1}-1\right)J(f_0) \quad \text{(A.7)}$$

και

$$\inf_{A_{i,0}} E\left(\tilde{l}(f,Z)-\tilde{l}_\psi(f_0,Z)\right)\ge\left(2^{i-1}-1/2\right)\delta_n^2\ge M(i,0)=2^{i-1}\delta_n^2 \quad \text{(A.8)}$$

Για την δεύτερη ροπή, από την σχέση (3.5.1) και την Υπόθεση Β παίρνουμε ότι για οποιαδήποτε $f\in\mathcal{F}$,

$$\begin{aligned} e(f,\bar{f}) &= E|f^*(X)|\left|Sign\left(Y\bar{f}(X)\right)-Sign(Yf(X))\right| \\ &\ge \delta E\left|Sign\left(Y\bar{f}(X)\right)-Sign(Yf(X))\right| I(|f^*(X)|\ge\delta) \\ &\ge \delta\left(E\left|Sign\left(Y\bar{f}(X)\right)-Sign(Yf(X))\right|-2c_1\delta^\alpha\right) \\ &\ge 2^{-1}(4c_1)^{-\frac{1}{\alpha}}\left(E\left|Sign\left(Y\bar{f}(X)\right)-Sign(Yf(X))\right|\right)^{\frac{\alpha+1}{\alpha}} \end{aligned}$$

με μια επιλογή του $\delta=\left(E\left|Sign\left(Y\bar{f}(X)\right)-Sign(Yf(X))\right|/4c_1\right)^{\frac{1}{\alpha}}$.

Έτσι έχουμε δημιουργήσει μια σύνδεση μεταξύ των πρώτων και δεύτερων ροπών. Επίσης από τις ιδιότητες της ψ προκύπτει ότι

$$E\left(\psi\left(Y\bar{f}(X)\right)-\left(1-Sign\left(Y\bar{f}(X)\right)\right)\right)=0.$$

Εδώ να σημειωθεί ότι $\psi(x) \geq \left(1 - Sign(x)\right) \forall\, x$ και

$$E\left|\psi\left(Yf_0(X)\right) - \left(1 - Sign\left(Yf_0(X)\right)\right)\right| = E\left(\psi f_0(X) - \left(1 - Sign\left(Yf_0(X)\right)\right)\right)$$
$$\leq e_\psi(f_0, \bar{f}).$$

Ως εκ τούτου, από την τριγωνική ανισότητα προκύπτει ότι,

$$E\left(l(f,Z) - l_\psi(f_0,Z)\right)^2 \leq UE\left|\left(1 - Sign\left(Yf(X)\right)\right) - \psi\left(Yf_0(X)\right)\right|$$
$$\leq U\left(E\left|Sign\left(Y\bar{f}(X)\right) - Sign\left(Yf(X)\right)\right|\right.$$
$$\left. + E\left|Sign\left(Y\bar{f}(X)\right) - Sign\left(Yf_0(X)\right)\right| + e_\psi\left(f_0, \bar{f}\right)\right)$$

Για κάθε $f \in A_{i,j}$, $e\left(f,\bar{f}\right)^{\frac{a}{a+1}} \geq (2^{-1}\delta_n^2)^{\frac{a}{a+1}} \geq 2^{-1}\delta_n^2 \geq e_\psi\left(f_0,\bar{f}\right)$ και $e\left(f,\bar{f}\right) \geq e\left(f_0,\bar{f}\right)$, πράγμα που σημαίνει ότι

$$E\left(l(f,Z) - l_\psi(f_0,Z)\right)^2 \leq U\left(2(4c_1)^{\frac{1}{a}}\left(e\left(f,\bar{f}\right)^{\frac{a}{a+1}} + e\left(f_0,\bar{f}\right)^{\frac{a}{a+1}}\right) + e_\psi\left(f_0,\bar{f}\right)\right)$$
$$\leq c_3\left(\frac{e(f,\bar{f})}{2}\right)^{\frac{a}{a+1}}$$

όπου $c_3 = 2^{\frac{1}{a}}Umax\left(4(2^{a+2}c_1)^{\frac{1}{a+1}} + 2\right),\ 8max(U,2)$

Κατά συνέπεια,

$$\sup_{A_{i,j}} E\left(l_\psi(f_0,Z) - l(f,Z)\right)^2 \leq \nu^2(i,j) = c_3 M(i,j)^{\frac{a}{a+1}}$$

για i=1,... j=0,.....

Με την υπόθεση ότι $max(J(f_0),1)\lambda \leq \delta_n^2/2$ και από την (Α.8) έχουμε

$$I \leq \sum_{i,j} P^*\left(\sup_{A_{i,j}} E_n\left(l_\psi(f_0,Z) - l(f,Z)\right) \geq M(i,j)\right)$$
$$+ \sum_i P^*\left(\sup_{A_{i,0}} E_n\left(l_\psi(f_0,Z) - l(f,Z)\right) \geq M(i,0)\right) = I_1 + I_2$$

Στη συνέχεια θα προχωρήσουμε στην εύρεση άνω φράγματος των I_i ξεχωριστά. Για τον υπολογισμό της μετρική εντροπίας ορίζουμε μια συνάρτηση bracketing για την διαφορά $l_\psi(f_0,Z) - l(f,Z)$.

Ορίζουμε ένα ε-bracketing σύνολο για το σύνολο

$$\{G_f :: G_f = \{x \in S: f(x) \geq 0\}, f \in A_{i,j}\}$$

να είναι $\{(G_1^l, G_1^u), \dots, (G_m^l, G_m^u)\}$.

Υποθέτουμε τώρα ότι $s_j^l(x)$ είναι -1 εάν $x \in G_j^u$ και 1 αλλιώς καθώς επίσης και $s_j^u(x)$ να είναι -1 εάν $x \in G_j^l$ και 1 αλλιώς, για j=1,...,m. Τότε $\{(s_1^l, s_1^u), \dots, (s_m^l, s_m^u)\}$ αποτελεί μια συνάρτηση ε-bracketing της $-Sign(f)$ για $f \in A_{i,j}$.

Αυτό σημαίνει ότι για κάθε $\varepsilon \geq M(i,j)$ και $f \in A_{i,j}$ υπάρχει ένα j με $(1 \leq j \leq m)$ τέτοιο ώστε

$$l_j^l(z) \leq l(f,z) - l_\psi(f_0,z) \leq l_j^u(z) \ \forall \ z = (x,y)$$

όπου

$$l_j^u(z) = 1 + \left(s_j^u(x)(1+y)/2 - s_j^l(x)(1-y)/2\right) - l_\psi(f_0,z)$$

$$l_j^l(z) = 1 + \left(s_j^l(x)(1+y)/2 - s_j^u(x)(1-y)/2\right) - l_\psi(f_0,z)$$

και $$\left(E\left(l_j^u - l_j^l\right)^2\right)^{1/2} = \left(E\left(s_j^u(x) - s_j^l(x)\right)^2\right)^{1/2} \leq 2^{1/2}\varepsilon^{1/2}$$

Ως εκ τούτου, $\left(E\left(l_j^u - l_j^l\right)^2\right)^{1/2} \leq min\left((2\varepsilon)^{1/2}, 2^{1/2}\right)$.

$\Longrightarrow H_B\left(\varepsilon, \mathcal{F}\left(2^j\right)\right) \leq H\left(\varepsilon^2/2, \mathcal{G}\left(2^j\right)\right) \ \forall \ \varepsilon > 0$ και j=0,....,

όπου $\mathcal{F}\left(2^j\right) = \left\{l(f,z) - l_\psi(f,z): f \in \mathcal{F}, J(f) \leq 2^j\right\}$

Χρησιμοποιώντας το γεγονός ότι

$$\int_{\alpha M(i,j)}^{\nu(i,j)} H^{1/2}\left(u^2/2, \mathcal{G}\left(2^j\right)\right) du/M(i,j)$$

δεν αυξάνεται στο i και $M(i,j)$ με i=1,... έχουμε

$$\int_{\alpha M(i,j)}^{\nu(i,j)} H^{1/2}\left(u^2/2, \mathcal{G}\left(2^j\right)\right) du/M(i,j)$$

$$\leq \int_{\alpha M(i,j)}^{c_3^{1/2} M(1,j)^{a/2(a+1)}} H^{1/2}\left(u^2/2, \mathcal{G}\left(2^j\right)\right) du/M(1,j) \leq \phi\left(\varepsilon_n, 2^j\right)$$

όπου $\alpha = \varepsilon/32$. Η Υπόθεση Γ με την επιλογή του $\varepsilon = 1/2$ και $c_i,\ i = 3{,}4$.

Επιλέγουμε $M(i,j)/v^2(i,j) \le 1/8max(U,2)$.

Για $0 < \delta_n \le 1$ και $\lambda max(J(f_0),1) \le \delta_n^2/2$. Από το θεωρήματος 3 του Shen και Wang (1994) με $M = n^{1/2}M(i,j), v = v^2(i,j), \varepsilon = 1/2$ και $T = max(U,2)$ προκύπτει ότι:

$$I_1 \le \sum_{j=1}^{\infty}\sum_{i=1}^{\infty} 3exp\left(-\frac{(1-\varepsilon)nM(i,j)^2}{2(4v^2(i,j)+M(i,j)T/3)}\right)$$

$$\le \sum_{j=1}^{\infty}\sum_{i=1}^{\infty} 3exp\left(-c_5 nM(i,j)^{\frac{a+2}{a+1}}\right)$$

$$\le \sum_{j=1}^{\infty}\sum_{i=1}^{\infty} 3exp\left(-c_5 n\left[2^{i-1}\delta_n^2 + \left(2^{j-1}-1\right)\lambda J(f_0)\right]^{\frac{a+2}{a+1}}\right)$$

$$\le \sum_{j=1}^{\infty}\sum_{i=1}^{\infty} 3exp\left(-c_5 n\left[\left(2^{i-1}\delta_n^2\right)^{\frac{a+2}{a+1}} + \left(\left(2^{j-1}-1\right)\lambda J(f_0)\right)^{\frac{a+2}{a+1}}\right]\right)$$

$$\le 3exp\left(-c_5 n\left(\lambda J(f_0)\right)^{\frac{a+2}{a+1}}\right) / \left[1 - exp\left(-c_5 n\left(\lambda J(f_0)^{\frac{a+2}{a+1}}\right)\right)\right]^2$$

όπου, c_5 είναι μια θετική σταθερά. Ομοίως, για το I_2 βρίσκουμε ένα άνω φράγμα.

Τελικά,

$$I \le 6exp\left(-c_5 n\left(\lambda J(f_0)\right)^{\frac{a+2}{a+1}}\right) / \left[1 - exp\left(-c_5 n\left(\lambda J(f_0)^{\frac{a+2}{a+1}}\right)\right)\right]^2$$

Αυτό σημαίνει ότι

$$I^{1/2} \le \left(\frac{5}{2} + I^{1/2}\right) exp\left(-c_5 n\left(\lambda J(f_0)\right)^{\frac{a+2}{a+1}}\right)$$

Το αποτέλεσμα προκύπτει από το γεγονός $I \le I^{1/2} \le 1$.

Απόδειξη του θεωρήματος 3.6.1: Αρκεί να δείξουμε ότι

$$\sum_{j=0}^{k} f_j^0(x) = \sum_{j=1}^{k} b_j^0 + \left(\boldsymbol{w}_j^0\right)^T x = 0$$

$\forall\, x \in S$ εάν $(\boldsymbol{b}^0, \boldsymbol{w}^0)$ είναι η ελαχιστοποίηση της

$$\min_{b,w}\left(\frac{1}{2}\sum_{j=1}^{k}\left\|w_j\right\|^2 + C\sum_{i=1}^{n}\psi\left(\boldsymbol{g}(\boldsymbol{f}(x_i), y_i)\right)\right)$$

με τους περιορισμούς

$$\tilde{X}\sum_{j=1}^{k}\tilde{w}_j = 0$$

Η απόδειξη γίνεται με εις άτοπο επαγωγή.

Έστω ότι

$$\sum_{j=0}^{k} f_j^0(\boldsymbol{x}^*) \neq 0$$

για κάποιο $\boldsymbol{x}^* \in S$.

Εν συνεχεία ορίζουμε τις εξής νέες συναρτήσεις απόφασης

$$f_j^1(\boldsymbol{x}) = b_j^1 + \left(\boldsymbol{w}_j^1\right)^T \boldsymbol{x} = \left(b_j^0 - \bar{b}^0\right) + \left(\boldsymbol{w}_j^0 - \bar{\boldsymbol{w}}_j^0\right)^T \boldsymbol{x}$$

για j=1,...,k όπου

$$\bar{b}^0 = \sum_{j=1}^{k}\frac{b_j^0}{k}$$

και

$\bar{\boldsymbol{w}}_j^0 = \sum_{j=1}^{k}\frac{\boldsymbol{w}_j^0}{k}$,

Σαφώς τόσο η f_j^0 όσο και η f_j^1 για j=1,..,k ικανοποιούν τον περιορισμό της (3.6.4).

Με αντικατάσταση τώρα στην αντικειμενική μας συνάρτηση παίρνουμε ότι:

$$\sum_{j=1}^{k}\langle \boldsymbol{w}_j^0 - \bar{\boldsymbol{w}}^0, \boldsymbol{w}_j^0 - \bar{\boldsymbol{w}}^0\rangle$$

$$= \sum_{j=1}^{k}\langle \boldsymbol{w}_j^0, \boldsymbol{w}_j^0\rangle - 2\sum_{j=1}^{k}\langle \boldsymbol{w}_j^0, \bar{\boldsymbol{w}}^0\rangle + k\langle \bar{\boldsymbol{w}}^0, \bar{\boldsymbol{w}}^0\rangle - \sum_{j=1}^{k}\langle \boldsymbol{w}_j^0, \boldsymbol{w}_j^0\rangle$$

$$- k\langle \bar{\boldsymbol{w}}^0, \bar{\boldsymbol{w}}^0\rangle \leq \sum_{j=1}^{k}\langle \boldsymbol{w}_j^0, \boldsymbol{w}_j^0\rangle$$

Η ανισότητα ισχύει επειδή $\sum_{j=0}^{k} f_j^0(\boldsymbol{x}^*) \neq 0$ για κάποιο $\boldsymbol{x}^* \in S$

Έτσι για $(b^1, \boldsymbol{w}^1)$ αποδίδεται μια μικρότερη τιμή της αντικειμενικής συνάρτησης από ότι για $(b^0, \boldsymbol{w}^0)$. Αυτό όμως έρχεται σε αντίθεση με την υπόθεση ότι $(b^0, \boldsymbol{w}^0)$ είναι η ελαχιστοποίηση της (3.6.4) και άρα οδηγούμαστε σε άτοπο έτσι έχουμε το επιθυμητό αποτέλεσμα.

Βιβλιογραφία

(Reference)

[1] Altman, D.G., Bland, J.M. (1994) 'Diagnostic tests 1: sensitivity and specificity,' *British Medical Journal*, Vol. 308, 1552.

[2] Altman, D.G., Bland, J.M. (1994) "Diagnostic tests 2: predictive values," *British Medical Journal*, vol 309, 102.

[3] An, L. T. H., and Tao, P.D. (1998). A D.C. optimization algorithm for solving the trust region subproblem. *Journal Optimization,* Vol. 8, No. 2, 467-505.

[4] Avis, D., and Fukuda, K. (1990). A pivoting algorithm for convex hulls and vertex enumeration of arrangements and polyhedra.

[5] Avis, D. (1999). A revised implementation of the reverse search vertex enumeration algorithm.

[6] Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. (2006). Convexity, Classification, and Risk Bounds. *Journal of the American Statistical Association*, Vol. 101, No. 473.

[7] Bradley, P. and Mangasarian, O. (1998). Feature selection via concave minimization and support vector machines.

[8] Crammer, K., and Singer, Y. (2001). On the Algorithmic Implementation of Multiclass Kernel-Based Vector Machines. *Journal of Machine Learning* Research 2, 265–292.

[9] David Meyer. (2012). Support Vector Machines, The Interface to libsvm in package e1071.

[10] Hastie, T., Tibshirani, R., and Zhu, J. (2004). The entire regularization path for the support vector machine. Journal of Machine Learning Research, 5, 1391-1415.

[11] Hastie T., Tibshirani R., J. Friedman. (2008). The Elements of Statistical Learning, Data Mining, Inference, and Prediction. Second Edition.

[12] Haykin S. απόδοση Ε. ΓΚΑΓΚΑΤΣΙΟΥ. (2010). Νευρωνικά Δίκτυα και Μηχανική Μάθηση. Τρίτη έκδοση.

[13] Koltchinskii V., and Panchenko D. (2002). Empirical Margin Distributions and Bounding the Generalization Error of Combined Classifier. *The Annals of Statistics*, 30, 1–50.

[14] Kuhn M. Contributions form Wing J., Weston S. and Williams A. (2008). Classification and Regression Training (The caret Package).

[15] Kuhn M., Ph.D. (2013). Predictive Modeling with R and the caret Package use R.

[16] Kuhn, M. (2015). A Short Introduction to the caret Package.

[17] Kuhn, M. (2008), "Building predictive models in R using the caret package, " *Journal of Statistical Software*,

[18] Lee Yoonkyung, Lin Yi and Wahba Grace. (2001). Multicategory Support Vector Machines, Department of statistics, *Technical Report* No. 1043.

[19] Liu. Y and Wu. Y. (2006). Optimizing ψ-learning via mixed integer programming. Statistica Sinica 16, 441-457

[20] Liu Y., M.S. (2004). Multicategory psi-learning and Support Vector Machine.

[21] Liu. Y and Shen. X. (2006). Multicategory ψ-learning. *Journal of the American Statistical Association,* Vol. 101, No. 474.

[22] Liu Y., Shen X., and Doss H. (2005). Multicategory ψ-Learning and Support Vector Machine: Computational Tools. *Journal of Computational and Graphical Statistics*, Volume 14, Number 1, Pages 219–236

[23] Lutz Hamel. (2009). Knowledge Discovery With Support Vector Machines.

[24] Mason, L., Bartlett, P., and Baxter, J. (1999). Improved Generalization Through Explicit Optimization of Margins. *Machine Learning*, 0, 1-11.

[25] Meyer D. [aut, cre], Dimitriadou E. [aut], Hornik K. [aut], Weingessel A. [aut], Leisch F. [aut]. (2013). Package 'e1071'.

[26] Sandrine Dudoit, Jane Fridlyand, and Terence P. Speed. (2002). Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data. Journal *of the American Statistical Association*, Vol. 97, No. 457.

[27] Shen, X., and Wong, W. H. (1994). Convergence Rate of Sieve Estimates. *The Annals of Statistics*, Vol. 22, No.2, 580–615.

[28] Shen X., Tseng G. C., Zhang X., and. Wong W. H. (2003). On ψ-learning. Journal *of the American Statistical Association*, Vol. 98, No. 463.

[29] Sijin Liu, M.S. (2006). Computational Development for psi-learning.

[30] Tao, P. D., An, L.T.H. (1997). Convex Analysis Approach to DC (diference of convex functions) Programming. Theory, Algorithms, Applications. ACTA Mathematica Vietnamica Volume 22, No 1, 289–355.

[31] Tsybakov, A. B. (2004). Optimal Aggregation of Classifiers in Statistical Learning. *The Annals of Statistics*, Volume 32, No 1, 135–166.

[32] Vapnik, V. (2000). The nature of Statistical Learning Theory. 2nd edition.

[33] Zhang, T. (2004). Statistical behaviour and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, Vol. 32, No. 1, 56–134

[34] Γιολάντα Π. Εγγλέζου. (2014). Επιλογή Μεταβλητών για την Ταξινόμηση με Μηχανές Διανυσματικής Υποστήριξης.

[35] Δρόσου Π. Κρυσταλλένια. (2013). Στατιστικές Μέθοδοι για την Ανάλυση Δεδομένων Υψηλής Διάστασης.

[36] Ευγενία Ι. Στουφή. (2015). Κριτήρια Πληροφορίας για την επιλογή μεταβλητών στις Μηχανές Διανυσματικής Υποστήριξης και Εφαρμογές.