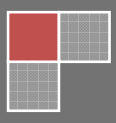
2012

Μηχανές διανυσματικής υποστήριξης (SVMs) και εφαρμογές σε πραγματικά σεισμολογικά δεδομένα

Διπλωματική εργασία της Μαριλένας Παπαδάκη
Επιβλέπων καθηγητής: Κος Κουκουβίνος Χρήστος, Καθηγητής Ε.Μ.Π.
Δ.Π.Μ.Σ.: Μαθηματική Προτυποποίηση στις Σύγχρονες Τεχνολογίες και την Οικονομία



Πίνακας Περιεχομένων

Περίληψη

Ο σκοπός της παρούσας εργασίας είναι να παρουσιάσει βασικές έννοιες των Μηχανών Διανυσματικής Υποστήριξης-SupportVector Machines (SVMs), καθώς πρόσφατες εφαρμογές και επεκτάσεις των SVMs. Η SVM τεχνική χρησιμοποιείται ευρέως και έχει απόδοση (δηλαδή, ο συντελεστής σφάλματος για τα σύνολα δοκιμών) είτε παρόμοια είτε σημαντικά καλύτερη από εκείνη των ανταγωνιστικών μεθόδων.

Το πρώτο κεφάλαιο αποτελεί μια εισαγωγή στις SVMs και σε σχετικές στατιστικές έννοιες όπως είναι η VC διάσταση, η ελαχιστοποίηση του διαρθρωτικού κινδύνου, τα σημεία προδιορισμού και ο αριθμός των παραμέτρων. Στο δεύτερο κεφάλαιο παρουσιάζονται οι λόγοι χρήσης των SVMs. Συγκεκριμένα, παρουσιάζονται το χαλαρό περιθώριο ταξινόμησης, ο έλεγχος πολυπλοκότητας και εφαρμογές των SVMs.

Στο τρίτο κεφάλαιο γίνεται αναφορά στις γραμμικές SVMs και στις προυποθέσεις Karush-Kuhn-Tucker (KKT). Στη συνέχεια γίνεται αναφορά στο βέλτιστο υπερσύνολο, στη δοκιμαστική φάση και στη μηχανολογική αναλογία. Στο τέταρτο κεφάλαιο παρουσιάζονται μη γραμμικές SVMs και η κατάσταση Mercer. Επίσης, στο τέταρτο κεφάλαιο παρουσιάζουμε μερικά παραδείγματα μη γραμμικών SVMs. Στο πέμπτο κεφάλαιο γίνεται λόγος για τη βελτιωμένη απόδοση των δεδομένων τόσο μέσω των wrappers όσο και μέσω των filters. Ακόμη, γίνεται ταξινόμηση μέσω SVMs καθώς και αξιολόγηση των αποτελεσμάτων και των παρατηρήσεων.

Στο έκτο κεφάλαιο αναφερόμαστε στην δυαδική κατηγοριοποίηση (Binary Classification) και στα soft margin. Επιπλέον αναφερόμαστε στην κανονικοποίηση, στο χειρισμό μη ισορροπημένου αριθμού δεδομένων, καθώς και παρουσιάζονται τα kernels, οι μέθοδοι cross validation και οι όροι sensitivity και specificity. Επιπροσθέτως, στο έβδομο κεφάλαιο γίνεται αναλυση και εξαγωγή αποτελεσμάτων των δεδομένων.

Τέλος στο όγδοο και τελικό κεφάλαιο μας παρουσιάζουμε τις τεχνικές εξόρυξης πληροφορίας (data mining) και το λογισμικό Clementine το οποίο και εφαρμόσαμε σε σεισμολογικά δεδομένα. Κατόπιν κάναμε μια σύντομη αναφορά στην Area Under the Curve (AUC) και τέλος προχωρήσαμε στην εφαρμογή καθώς και στην ερμηνεία των αποτελεσμάτων.

Abstract

The purpose of this paper is to present basic concepts of Support Vector Machines-(SVMs), their recent applications and some extensions of SVMs. The SVM technique is widely used and has performance (i.e., the error rate on test sets) either similar or significantly better than the competing methods.

The first chapter is an introduction to SVMs and related statistical concepts such as the VC dimension, the minimizing of the structural risk, the determination points and the number of parameters. The second chapter presents the reasons for using SVMs, specifically, the loose margin classification, the control complexity and applications of SVMs.

The third chapter refers to the linear SVMs and to the conditions Karush-Kuhn-Tucker (KKT). The following sections discuss the optimal hyperset, the test phase and the mechanical analogy. The fourth chapter presents non-linear SVMs, and the situation Mercer. In addition, in the fourth chapter we present some examples of nonlinear SVMs. The fifth chapter talks about the improved performance of data both through wrappers and filters methods. There has been also classification via SVMs and evaluation of the results and the observations.

In the sixth chapter we refer to the Binary Classification and to the soft margin. Furthermore, we refer to the normalization and handling of unbalanced number of data; present the kernels, cross validation methods and the terms of sensitivity and specificity. Moreover, in the seventh chapter, there has been analysis and export of data results.

Finally in the eighth and final chapter we present the data mining techniques and the Clementine software which we applied to seismic data. We then made a brief reference to the Area Under the Curve (AUC), and proceed with the implementation and interpretation of the results.

1. Μηχανική Μάθηση

Η Μηχανική Μάθηση θεωρείται ως ένα παρακλάδι της Τεχνητής Νοημοσύνης και ασχολείται με την ανάπτυξη τεχνικών και μεθόδων μέσω του υπολογιστή. Η ανάπτυξη των αλγορίθμων επιτρέπει στον υπολογιστή να ασκεί καθήκοντα και δραστηριότητες.

Η Support Vector Machine (SVM) αναπτύχθηκε για πρώτη φορά το 1992, και θεσπίστηκε από τους Boser, Guyon, και Vapnik σε COLT-92. Οι Μηχανές διανύσματος υποστήριξης (SVMs) είναι ένα σύνολο μεθόδων που χρησιμοποιούνται για την ταξινόμηση και παλινδρόμηση. Ανήκουν σε μια οικογένεια γενικευμένων γραμμικών ταξινομητών. Η Support Vector Machine (SVM) είναι ένα εργαλείο πρόβλεψης, ταξινόμησης και παλινδρόμησης που χρησιμοποιεί τη θεωρία της μηχανικής μάθησης για τη μεγιστοποίηση της προγνωστικής ακρίβειας ενώ αυτόματα αποφεύγονται τα υπερβολικά ταιριαστά δεδομένα. Η SVM ήταν αρχικά δημοφιλής στην κοινότητα του ερευνητικού κοινού και τώρα είναι ένα ενεργό μέρος της έρευνας της μηχανικής μάθησης σε όλο τον κόσμο. Η SVM έγινε ευρέως γνωστή όταν έγινε εισαγωγή χρήσης pixel, διότι δίνει ακρίβεια συγκρίσιμη με τα πολύπλοκα νευρωνικά δίκτυα με επεξεργασμένα στοιχεία σε μια εργασία αναγνώρισης χειρογράφου. Έχει, επίσης, χρησιμοποιηθεί από πολλές εφαρμογές, όπως η ανάλυση με το χέρι, η ανάλυση προσώπου και ούτω καθ 'εξής, ειδικά για την ταξινόμηση μοτίβου και παλινδρόμησης. Τα θεμέλια της SVM έχουν αναπτυχθεί από τον Vapnik και η SVM χρησιμοποιείται πλέον ευρέως λόγω των πολλών ελπιδοφόρων χαρακτηριστικών, όπως η καλύτερη εμπειρική απόδοση. Η ελαχιστοποίηση του κινδύνου στις SVMs (SRM) έχει αποδειχθεί ότι είναι ανώτερη από την παραδοσιακή εμπειρική ελαχιστοποίηση του κινδύνου (ERM), η οποία χρησιμοποιείται από τα συμβατικά νευρωνικά δίκτυα. Αυτή η διαφορά είναι πολύ σημαντική και εξοπλίζει την SVM με μια μεγαλύτερη δυνατότητα να γενικευθεί. Οι SVMs αναπτύχθηκαν για να λύσουν προβλήματα ταξινόμησης, αλλά πρόσφατα έχουν επεκταθεί στην επίλυση προβλημάτων παλινδρόμησης.

1.1 Περιορισμός σε σχέση με την γενικευμένη απόδοση

Υπάρχουν περιορισμοί που διέπουν τη σχέση μεταξύ της χωρητικότητας μιας SVM και της απόδοσής της. Η θεωρία αναπτύχθηκε από εκτιμήσεις, υπό ποιες περιστάσεις και πόσο γρήγορα ο μέσος όρος κάποιων εμπειρικών ποσοτήτων συγκλίνει ομοιόμορφα, καθώς ο αριθμός των σημείων των δεδομένων αυξάνει.

Ας υποθέσουμε ότι κάθε παρατήρηση αποτελείται από ένα ζεύγος: ένα διάνυσμα $x_i \in R_n$, i = 1,. . . , L , και y_i. Στο πρόβλημα του δέντρου αναγνώρισης, το x_i μπορεί να είναι ένα διάνυσμα τιμών pixel (π.χ. n = 256 για μια εικόνα), και το y_i θα είναι 1 αν η εικόνα περιέχει ένα δέντρο, και -1 διαφορετικά.

Τώρα υποτίθεται ότι υπάρχει κάποια άγνωστη κατανομή πιθανοτήτων P (x, y) από την οποία απεικονίζονται τα δεδομένα αυτά, δηλαδή, τα δεδομένα θεωρούνται "iid" (ανεξάρτητα σχεδιασμένα και με πανομοιότυπη διανομή). Θα συμβολίζουμε με P την αθροιστική κατανομή

πιθανότητας, και p την πυκνότητα τους. Σημειώνουμε ότι αυτή η υπόθεση είναι πιο γενική από το να συσχετίσεις ένα σταθερό y με κάθε x: επιτρέπει μια κατανομή του y για μια δεδομένη x.

Τώρα, ας υποθέσουμε ότι έχουμε ένα μηχάνημα, στόχος του οποίου είναι η χαρτογράφηση της απεικόνισης $x_i \to y_i$. Το μηχάνημα στην πραγματικότητα προσδιορίζεται από ένα σύνολο πιθανών αντιστοιχίσεων x → f (x, α), όπου οι λειτουργίες f (x, α) είναι αυτοχαρακτηρισμένες από τις παραμέτρους α. Το μηχάνημα υποτίθεται ότι είναι ντετερμινιστικό: για εισαγωγή δεδομένων (input) x, και επιλογή του α, θα δώσει πάντα το ίδιο αποτέλεσμα (output) f(x,α). Μια συγκεκριμένη επιλογή του α δημιουργεί αυτό που ονομάζουμε "εκπαιδευμένη μηχανή." (trained machine). Έτσι, για παράδειγμα, ένα νευρωνικό δίκτυο με συγκεκριμένη αρχιτεκτονική, με α να αντιστοιχεί στα βάρη, είναι μια μηχανική μάθησης. Η προσδοκία του σφάλματος για μια μηχανή μάθησης είναι ως εκ τούτου:

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

Σημειώστε ότι, όταν υπάρχει μια πυκνότητα p(x, y), η ποσότητα dP(x, y) μπορεί να γραφτεί p(x, y)dxdy. Αυτός είναι ένας τρόπος γραφής του πραγματικού μέσου λάθος, εκτός και αν έχουμε μια εκτίμηση του P (x, y). Η ποσότητα R (α) ονομάζεται ο αναμενόμενος κίνδυνος, ή απλά κίνδυνος. Εδώ θα το ονομάσουμε πραγματικό κίνδυνο, για να τονίσουμε ότι είναι η ποσότητα που τελικά μας ενδιαφέρει. Ο «Εμπειρικός κίνδυνος» $R_{emp}(a)$ ορίζεται να είναι ακριβώς το μέσο ποσοστό σφάλματος σε ένα σύνολο δεδομένων (για ένα σταθερό, πεπερασμένο αριθμό παρατηρήσεων)

$R_{emp}(a)$

$$R_{emp}(a) = \frac{1}{2l} \sum_{i=1}^{l} \left| y_i - f(x_i,a) \right|$$

$\{x_i, y_i\}$

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

$\{f(a)\}$

$\{f(a)\} \in \{-1,1\} \forall x, a$

$$R_{emp}(a) = \frac{1}{2l} \sum_{i=1}^{l} \left| y_i - f(x_i,a) \right|$$

Σημειώστε ότι δεν εμφανίζεται εδώ πιθανότητα κατανομής. Η $R_{emp}(a)$ είναι ένας σταθερός αριθμός για μια συγκεκριμένη επιλογή του α και για ένα συγκεκριμένο σύνολο δεδομένων

$\{x_i, y_i\}$.

Η ποσότητα $\frac{1}{2}|y_i - f(\mathbf{x}_i, \alpha)|$ ονομάζεται απώλεια. Η περίπτωση που περιγράφεται εδώ, μπορεί να λάβει μόνο τις τιμές 0 και 1. Τώρα επιλέγουμε κάποια η τέτοια ώστε $0 \leq \eta \leq 1$. Στη συνέχεια, για τις ζημίες που επιδέχονται αυτές οι τιμές, με πιθανότητα 1 - η, ισχύει το ακόλουθο:

$$R(a) \leq R_{emp}(a) + \sqrt{(\frac{h(\log(2l/h)+1) - \log(n/4)}{l})} \quad (1)$$

όπου h είναι ένας μη-αρνητικός ακέραιος που ονομάζεται Vapnik Chervonenkis (VC) διάσταση, και είναι ένα μέτρο της έννοιας της ικανότητας. Στη συνέχεια θα επικαλεστούμε το δικαίωμα της εξίσωσης (1), τον "Κίνδυνο περιορισμού". Ο δεύτερος όρος ονομάζεται «εμπιστοσύνη VC».

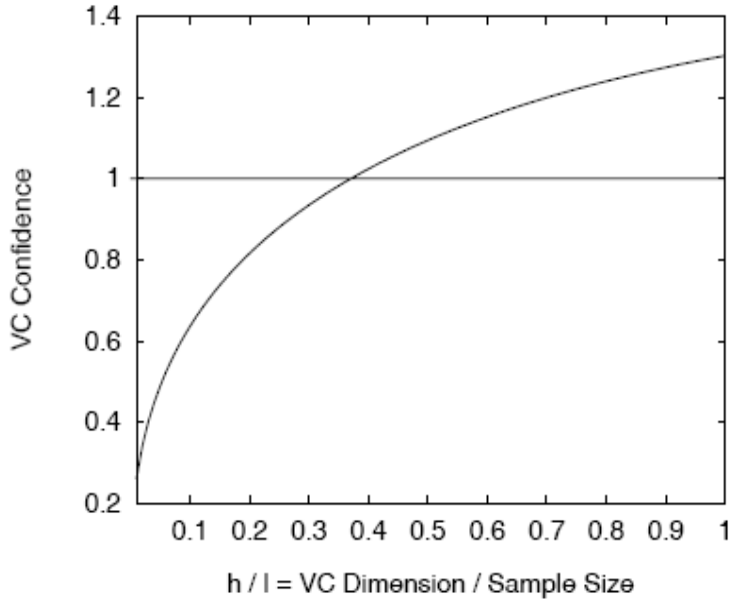
Σημειώνουμε τρία βασικά σημεία σχετικά με αυτό τον περιορισμό. Κατ 'αρχάς, είναι ανεξάρτητη του P (x, y). Υποθέτει μόνο ότι και τα δεδομένα εκπαίδευσης και τα δεδομένα των δοκιμών σχεδιάζονται ανεξάρτητα σύμφωνα με ορισμένες P (x, y).

Παρέχοντας περισσότερες διαφορετικές μηχανές μάθησης (υπενθυμίζουμε ότι «μηχανή μάθησης» είναι μια εναλλακτική ονομασία για μια οικογένεια των συναρτήσεων f (x, α)), και επιλέγοντας ένα σταθερό, αρκετά μικρό n, η μηχανή η οποία ελαχιστοποιεί την δεξιά πλευρά, επιλέγουμε αυτό το μηχάνημα που δίνει το χαμηλότερο άνω όριο για τον πραγματικό κίνδυνο. Αυτό παρέχει στον πειραματιστή μια μέθοδο για την επιλογή της μηχανής μάθησης για μια συγκεκριμένη εργασία, και είναι η βασική ιδέα ελαχιστοποίησης του κινδύνου.

1.1.1 Ελαχιστοποίηση του περιορισμού με την ελαχιστοποίηση h

Η Εικόνα 1 δείχνει πως ο δεύτερος όρος στη δεξιά πλευρά της παραπάνω εξίσωσης (1) διαφέρει κατά h, δεδομένου ότι το επιπέδο εμπιστοσύνης είναι 95% (n= 0.05) και υποθέτουμε ότι έχουμε ένα μεγέθος 10.000.

Έτσι, δίνεται κάποια επιλογή ανάμεσα στις μηχανές μάθησης των οποίων ο εμπειρικός κίνδυνος είναι μηδέν, όταν κάποιος θέλει να επιλέξει μια μηχανική μάθησης της οποίας τα σύνολα λειτουργιών έχουν ελάχιστη διάσταση VC. Αυτό θα οδηγήσει σε ένα καλύτερο άνω όριο για το πραγματικό λάθος. Σε γενικές γραμμές, για ένα μη μηδενικό εμπειρικό κίνδυνο, κάποιος θα πρέπει να επιλέξει τη μηχανή **η οποία ελαχιστοποιεί τη δεξιά πλευρά της εξίσωσης (1)**. Να σημειωθεί ότι όσον αφορά την υιοθέτηση αυτής της στρατηγικής, μπορούμε να χρησιμοποιήσουμε μόνο την Εξ. (1) ως οδηγό. Η Εξ. (1) δίνει με συγκεκριμένη πιθανότητα ένα άνω όριο για τον πραγματικό κίνδυνο. Σημειώστε επίσης ότι το διάγραμμα δείχνει ότι για h / l> 0.37 (και για n= 0.05 και l= 10,000), η VC εμπιστοσύνη υπερβαίνει την ενότητα.



Εικόνα 1

1.2 Στατιστική Θεωρία

Η στατιστική θεωρία παρέχει ένα πλαίσιο για τη μελέτη του προβλήματος για την απόκτηση γνώσης, κάνοντας

- προβλέψεις,
- λήψη αποφάσεων

από ένα σύνολο δεδομένων.

Στη στατιστική θεωρία το πρόβλημα της εποπτευόμενης μάθησης έχει ως εξής. Μας δίνεται ένα $R^n \times R$ σύνολο δεδομένων εκπαίδευσης $\{(\mathbf{x}_1,y_1)... (\mathbf{x}_l,y_l)\}$ και η δειγματοληψία γίνεται σύμφωνα με μια άγνωστη κατανομή πιθανοτήτων P (x, y), και μια απώλεια λειτουργίας V (y, f(x)) η οποία μετρά το λάθος, για ένα δεδομένο (x, f(x)) που είναι «προβλέψιμο» αντί της πραγματικής αξίας y. Το πρόβλημα συνίσταται στην εύρεση μίας συνάρτησης f που ελαχιστοποιεί την προσδοκία του σφάλματος σχετικά με τα νέα δεδομένα.

$$\int V(y, f(\mathbf{x}))\, P(\mathbf{x}, y)\, d\mathbf{x}\, dy$$

1.3 Η VC διάσταση

Η VC διάσταση είναι μια ιδιότητα ενός συνόλου λειτουργιών $\{f(a)\}$

(και πάλι, χρησιμοποιούμε α ως γενικό σύνολο παραμέτρων: η επιλογή του α καθορίζει μια συγκεκριμένη λειτουργία), μπορεί να οριστεί για διάφορες κατηγορίες της λειτουργιών f. Εδώ θα εξετάσουμε μόνο τις λειτουργίες που αντιστοιχούν σε δύο κατηγορίες, έτσι ώστε $\{f(a)\} \in \{-1,1\} \forall x, a$. Τώρα, ένα δεδομένο σύνολο με i σημεία μπορεί να επισημανθεί με 2^l πιθανούς τρόπους, και για κάθε επισήμανση, ένα μέλος του συνόλου $\{f(a)\}$ μπορεί να βρει ποιό αποδίδει σωστά τις ετικέτες, τότε λέμε ότι αυτό το σύνολο των σημείων είναι ένα σύνολο λειτουργιών. Η VC διάσταση για το σύνολο των λειτουργιών $\{f(a)\}$ ορίζεται ως ο μέγιστος αριθμός των σημείων κατάρτισης που μπορεί να έχουν προσδιοριστεί από τον $\{f(a)\}$.

Αρχικά, θα αποδειχθεί ότι η VC διάσταση των SVMs μπορεί να είναι πολύ μεγάλη (ακόμα και άπειρη). Στη συνέχεια, θα διερυνηθεί με διάφορα επιχειρήματα ως προς το γιατί, παρ 'όλα αυτά, οι SVMs συνήθως παρουσιάζουν καλή απόδοση. Επί του παρόντος δεν υπάρχει θεωρία η οποία εγγυάται ότι μια δεδομένη οικογένεια SVMs έχει υψηλή ακρίβεια σε ένα δεδομένο πρόβλημα. Θα καλούμε κάθε πυρήνα που πληρεί τις προϋποθέσεις Mercer ως θετικό πυρήνα, καθώς και τον αντίστοιχο Η χώρο, ως χώρο ενσωμάτωσης. Επίσης, καλούμε κάθε χώρο ενσωμάτωσης με ελάχιστη διάσταση για ένα συγκεκριμένο πυρήνα ως «ελάχιστο χώρο εμφύτευσης». Έχουμε τα εξής:

Θεώρημα 1.

'Εστω K ένας θετικός πυρήνας που αντιστοιχεί σε ελάχιστο χώρο ενσωμάτωσης H. Στη συνέχεια, η VC διάσταση του αντίστοιχου SVM είναι dim(H) + 1.

Ας δούμε δύο παραδείγματα.

(α) VC διάσταση

Σκεφτείτε μία SVM με ομοιογενή πολυωνυμικό πυρήνα, που ενεργεί για δεδομένα σε R^{d_L} :

$$K(x_1, x_2) = (x_1 \circ x_2)^p, \quad x_1, x_2 \in R^{d_L}$$

Όπως και στην περίπτωση $d_L = 2$ ο πυρήνας δεν είναι τετραγωνικός, και μπορεί κανείς να κατασκευάσει τον Φ χάρτη. Θέτοντας $z_i = x_{1i}x_{2i}$, έτσι ώστε $K(x_1, x_2) = (z_1 + ... + z_{d_L})^p$, βλέπουμε ότι κάθε διάσταση Η ανταποκρίνεται σε δοσμένες δυνάμεις z_i. Στην πραγματικότητα, εάν επιλέξουμε να αναφερόμαστε στον Φ(x) με αυτόν τον τρόπο, μπορούμε να γράψουμε τη χαρτογράφηση για κάθε p και d_L:

$$\Phi_{r_1 r_2}....(x) = \sqrt{(\frac{p!}{r_1!r_2!...r_{d_L}!})}x_1^{r_1}...x_{d_L}^{r_{d_L}}, \quad \sum_{i=1}^{d_L} r_i = p, r_i \geq 0$$

Αυτό οδηγεί στο παρακάτω θεώρημα:

Θεώρημα 2.

Αν ο χώρος των δεδομένων έχει διάσταση d_L , η διάσταση του ελάχιστου χώρου ενσωμάτωσης, για ομοιογενείς πολυωνυμικούς πυρήνες βαθμού $p(K(x_1, x_2) = (x_1 \bullet x_2)^p, \quad x_1, x_2 \in R^{d_L}$

είναι $\begin{pmatrix} d_L + p - 1 \\ p \end{pmatrix}$ Έτσι, η VC διάσταση των SVMs με αυτούς τους πυρήνες είναι $\begin{pmatrix} d_L + p - 1 \\ p \end{pmatrix} + 1$. Όπως προαναφέρθηκε, αυτό παίρνει πολύ μεγάλες διαστάσεις πολύ γρήγορα.

(β) VC διάσταση

Θεώρημα 3 .

θεωρήστε την κατηγορία Mercer πυρήνων για την οποία $K(x_1, x_2) \to 0$ καθώς και $\|x_1 - x_2\| \to \infty$ και για το οποίο K(x,x) είναι O (1), και υποθέτουν ότι τα δεδομένα μπορούν να επιλεγούν αυθαίρετα από R_d. Στη συνέχεια, η οικογένεια των ταξινομητών που αποτελείται από SVMs χρησιμοποιεί τους πυρήνες, για τους οποίους το λάθος επιτρέπεται να λάβει όλες τις τιμές, έχει άπειρη VC διάσταση.

Αυτή η υπόθεση μας δίνει μια δεύτερη ευκαιρία να παρουσιάσουμε μια κατάσταση όπου η λύση SVM μπορεί να υπολογιστεί αναλυτικά, γεγονός που συνιστά επίσης μια δεύτερη, εποικοδομητική απόδειξη του θεωρήματος.

Θεώρημα 4.

Συγκεκριμένα θα πάρουμε την υπόθεση για Gaussian πυρήνες RBF της μορφής $K(x_1, x_2) = e^{-\|x_1 - x_2\|^2/2\sigma^2}$ *. Ας επιλέξουμε τα σημεία κατάρτισης, έτσι ώστε η μικρότερη απόσταση μεταξύ οποιονδήποτε δύο σημείων να είναι πολύ μεγαλύτερη από το πλάτος σ. Η λειτουργική απόφαση αξιολογείται με βάση τον SV* s_j*:*

$$f(s_j) = \sum_i a_i y_i e^{-\|s_i - s_j\|^2/2\sigma^2} + b$$

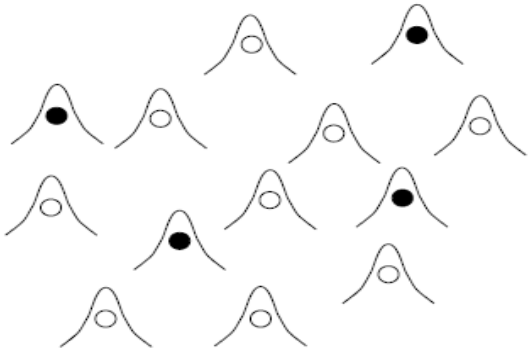
Το ποσό στην δεξιά πλευρά θα κυριαρχείται σε μεγάλο βαθμό από τον όρο *i=j*. Στην πραγματικότητα η αναλογία του όρου αυτού μπορεί να γίνει αυθαίρετα μεγάλη, επιλέγοντας τα σημεία να είναι σε αυθαίρετα μεγάλη απόσταση μεταξύ τους. Για να βρούμε τη λύση SVM, θα υποθέσουμε προς στιγμήν ότι κάθε σημείο της κατάρτισης γίνεται διάνυσμα υποστήριξης (SV). Έστω ότι υπάρχουν $N_+(N_-)$ θετικά (αρνητικά) πολικά σημεία. Θα υποθέσουμε ότι όλα τα θετικά(αρνητικά) πολικά σημεία έχουν την ίδια αξία $a_+(a_-)$ για τον πολλαπλασιαστή Lagrange. Στη συνέχεια :

$$a_+ + b = 1$$
$$-a_- + b = -1$$
$$N_+ a_+ - N_- a_- = 0$$

που ικανοποιούνται από

$$a_+ = \frac{2N_-}{N_- + N_+}$$
$$a_- = \frac{2N_+}{N_- + N_+}$$
$$b = \frac{N_+ - N_-}{N_- - N_+}$$

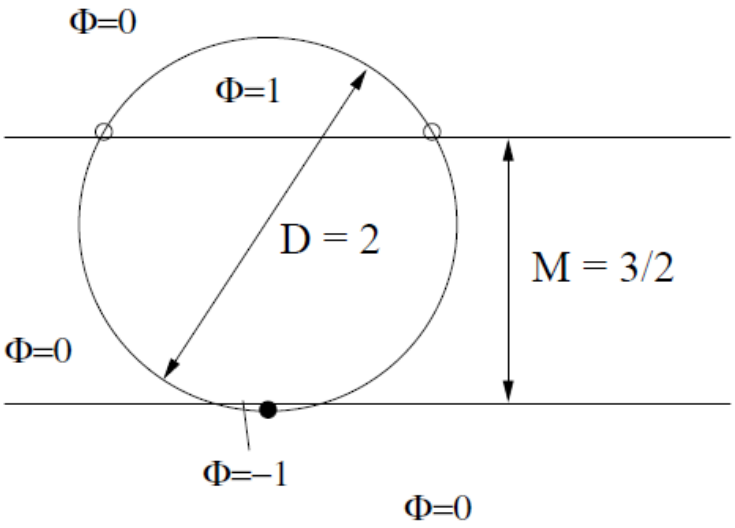
Έτσι, εφόσον τα αποτελέσματα a_i είναι επίσης θετικά, πληρούνται όλες οι προϋποθέσεις και οι περιορισμοί KKT. Δεδομένου ότι ο αριθμός των σημείων κατάρτισης, καθώς και η επισήμανσή τους, είναι αυθαίρετη, έχουν διαχωριστεί, χωρίς σφάλμα, η διάσταση VC είναι άπειρη. Η κατάσταση συνοψίζεται σχηματικά στην Εικόνα 2. Τώρα είμαστε αντιμέτωποι με ένα εντυπωσιακό αίνιγμα. Ακόμα κι αν η VC διάσταση τους είναι άπειρη, η SVM RBFs μπορεί να έχει εξαιρετική απόδοση . Μια παρόμοια ιστορία ισχύει και για πολυώνυμο SVMs.



Εικόνα 2. Gaussian SVMs RBF με αρκετά μικρό πλάτος, και κατά συνέπεια έχουν άπειρη διάσταση VC

- VC διάσταση του επιτρεπτού κενού των ταξινομητών

Σκεφτείτε μια οικογένεια ταξινομητών, την οποία θα ονομάσουμε «gap tolerant classifiers". Ένας ιδιαίτερος ταξινομητής $\varphi \in \Phi$ που καθορίζεται από τη θέση και τη διάμετρο



Εικόνα 3. Ένας gap tolerant classifier σε δεδομένα R²

μιας μπάλας σε R^d, και από δύο υπερσύνολα, με παράλληλες, επίσης σε R^d. Καλούμε το σύνολο των σημείων που βρίσκονται μεταξύ, αλλά όχι επάνω, στο υπερσύνολο ως " περιθώριο συνόλου." Οι λειτουργίες φ ορίζονται ως εξής: τα σημεία που βρίσκονται μέσα στη σφαίρα, αλλά όχι στο περιθώριο του συνόλου, έχουν εκχωρηθεί στην κατηγορία $\{\pm 1\}$, ανάλογα σε ποιά

πλευρά του περιθωρίου πέφτουν. Όλα τα άλλα σημεία απλά ορίζονται να είναι «σωστά», δηλαδή, δεν έχει δοθεί άλλη κατηγορία από τον ταξινομητή, και δεν συμβάλλουν σε οποιοδήποτε κίνδυνο. Η κατάσταση συνοψίζεται για d = 2. Αυτή η μάλλον περίεργη οικογένεια ταξινομητών, μαζί με μια κατάσταση που θα επιβληθεί, θα έχει ως αποτέλεσμα τα συστήματα να είναι παρόμοια με τις SVMs.

Βάζουμε ετικέτα στη διάμετρο της σφαίρας D και στη κάθετη απόσταση μεταξύ των δύο υπερσυνόλων M. Η διάσταση VC ορίζεται όπως και πριν να είναι ο μέγιστος αριθμός των σημείων που μπορούν να τραβηχτούν από την οικογένεια, αλλά με την έννοια «τραβηχτούν» εννοούμε ότι είναι τα σημεία που μπορεί να προκύψουν ως λάθη με όλους τους δυνατούς τρόπους. Σαφώς μπορούμε να ελέγξουμε τη VC διάσταση της οικογένειας αυτών των ταξινομητών από τον έλεγχο των ελάχιστων M περιθωρίων και μέγιστη διάμετρο D που επιτρέπεται.

Για παράδειγμα, θεωρούμε την οικογένεια του gap tolerant classifier σε R² με διάμετρο $D = 2$. Με M περιθώριο $M \geq 3/2$ μπορούν να χωριστούν κατά τρεις κατηγορίες. Αν $3/2 < M < 2$, τότε μπορεί να χωριστεί σε 2, και αν $M \geq 2$ μπορούν να χωριστούν μόνο σε ένα. Κάθε μία από αυτές τις τρεις οικογένειες ταξινομητών αντιστοιχεί σε ένα από τα σετ των ταξινομητών, με μόλις τρία υποσύνολα λειτουργιών, και με $h_1 = 1, h_2 = 2, h_3 = 3$. Αυτές οι ιδέες μπορούν να χρησιμοποιηθούν για να δείξουμε πώς ο gap tolerant classifier εφαρμόζεται για την ελαχιστοποίηση των διαρθρωτικών κινδύνων. Η επέκταση του παραπάνω παραδείγματος στους χώρους της αυθαίρετης διάστασης βρίσκεται σε ένα (τροποποιημένο) θεώρημα του:

Θεώρημα 5.

Για τα δεδομένα στο R^d, η VC διάσταση h των gap tolerant classifiers με ελάχιστο περιθώριο Mmin και μέγιστη διάμετρο Dmax οριοθετείται από $\min\left\{\left[D_{\max}^2 / M_{\min}^2\right], d\right\} + 1$

Λήμμα 1.

Έστω $n \leq d+1$ σημεία που βρίσκονται σε μια σφαίρα $B \in R^d$. Ας αφήσουμε τα σημεία να είναι χωριστά από το gap tolerant classifier με απόσταση M. Στη συνέχεια, προκειμένου να μεγιστοποιηθεί η M, τα σημεία πρέπει να βρίσκονται σε κορυφές ενός (n-1) διαστάσεων συμμετρικό simplex, και πρέπει να βρίσκονται επίσης στην επιφάνεια της σφαίρας.

Ως εκ τούτου, σε γενικές γραμμές η VC διάσταση του gap tolerant classifier πρέπει να πληρούν τις εξής προϋποθέσεις:

- $h \leq \left[\frac{D_{\max}^2}{M_{\min}^2}\right] + 1$

- $h \le d+1$.

1.3.1 Ελαχιστοποίηση του διαρθρωτικού κινδύνου

Μπορούμε να συνοψίσουμε τώρα την αρχή της ελαχιστοποίησης του διαρθρωτικού κινδύνου (Structural Risk Minimization-SRM).

Σημειώστε ότι ο VC όρος εμπιστοσύνης εξαρτάται από μια επιλεγμένη κατηγορία λειτουργιών, όπου ο εμπειρικός κίνδυνος και ο πραγματικός κίνδυνος εξαρτώνται από μια ιδιαίτερη λειτουργία η οποία επιλέγεται από τη διαδικασία της κατάρτισης. Θα θέλαμε να βρούμε αυτό το υποσύνολο του επιλεγμένου συνόλου λειτουργιών, έτσι ώστε το όριο του κίνδυνου σε αυτό το υποσύνολο είναι το ελάχιστο δυνατό. Σαφώς δεν μπορούμε να κανονίσουμε τα πράγματα έτσι ώστε να η VC διάσταση h να ποικίλλει ομαλά, δεδομένου ότι είναι ακέραιος. Αντ'αυτού, εισάγαμε μια "δομή" όπου διαιρέσαμε το σύνολο της κατηγορίας των λειτουργιών σε υποσύνολα. Για κάθε υποσύνολο, πρέπει να είμαστε σε θέση είτε να υπολογίσουμε το h, ή να πάρουμε ένα όριο για το ίδιο το h. Η SRM στη συνέχεια συνίσταται από τη διαπίστωση ότι ένα υποσύνολο των λειτουργιών ελαχιστοποιεί το όριο του πραγματικού κινδύνου.

1.3.2 Ο αριθμός των παραμέτρων

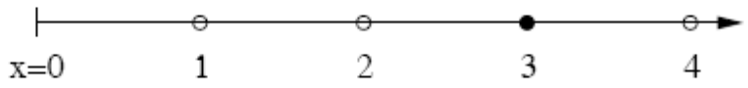
Η VC διάσταση συγκεκριμενοποιεί την ικανότητα μιας δεδομένης ομάδας των λειτουργιών. Διαισθητικά, θα μπορούσε να οδηγήσει κάποιον να αναμένει ότι οι SVMs με πολλές παραμέτρους θα έχουν υψηλή διάσταση VC, σε αντίθεση με αυτές που έχουν λίγες παραμέτρους. Υπάρχει ένα εντυπωσιακό αντιπαράδειγμα σε αυτό, της E. Levin και J.S. Denker:

Μια μηχανή μάθησης με μία μόνο παράμετρο, αλλά με άπειρη διάσταση VC . Προσδιορίζεται το βήμα της συνάρτησης

$$\theta(x), x \in R : \{\theta(x) = 1 \forall x > 0, \;\; \theta(x) = -1 \;\; \forall x \le 0\}.$$

Σκεφτείτε μία παράμετρο σαν οικογένεια των λειτουργιών, που ορίζονται με $f(x,a) \equiv \theta(\sin(ax)), \;\; x, a \in R$

Μπορείτε να επιλέξετε κάποιον “l” αριθμό, με στόχο να βρείτε “l” σημεία τα οποία μπορούν να περιοριστούν. Επιλέγω να είναι:



Εικόνα 4.

$x_i = 10^{-i}, i = 1,...,l$

Προσδιορίζουμε όποια σημεία θέλουμε

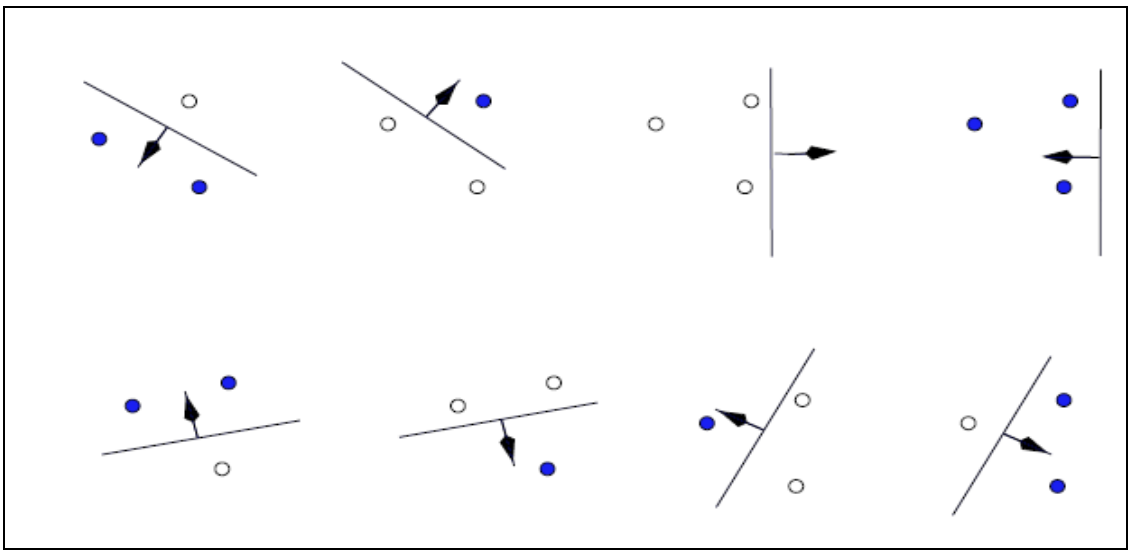
$y_1,...,y_{l,},y_i \in \{-1,1\}$

$$a = \pi(1 + \sum_{i=1}^{l} \frac{(1-y_i)10^i}{2})$$

Έτσι, η VC διάσταση αυτής της μηχανής είναι το άπειρο. Είναι ενδιαφέρον το γεγονός ότι ακόμα κι αν περιορίσουμε ένα μεγάλο αριθμό σημείων, μπορούμε να βρούμε τέσσερα σημεία που δεν μπορούν. Απλά πρέπει να είναι ίση απόσταση μεταξύ τους, και η απόδοσή τους, όπως φαίνεται στην Εικόνα 4.

1.3.3 Σημεία προσδιορισμού στον R^n

Ας υποθέσουμε ότι ο χώρος στον οποίο υπάρχουν τα δεδομένα είναι διάστασης R², και το σύνολο $\{f(a)\}$ αποτελείται από ευθείες γραμμές, έτσι ώστε για μια δεδομένη γραμμή, όλα τα σημεία από τη μία πλευρά τους να αποδίδονται στη κατηγορία 1, και όλα τα σημεία από την άλλη πλευρά, στην τάξη -1. Ο προσανατολισμός φαίνεται στην Εικόνα 5, από ένα βέλος, προσδιορίζοντας σε ποια πλευρά της γραμμής είναι τα σημεία που θα διαθέτουν το σημείο 1. Ενώ είναι δυνατόν να βρεθούν τρία σημεία που μπορούν να περιοριστούν από αυτό το σύνολο λειτουργιών, ωστόσο δεν είναι δυνατό να βρεθούν τέσσερα. Έτσι, η VC διάσταση του συνόλου των προσανατολισμένων γραμμών στην R² είναι τρεις.



Εικόνα 5.

1.3.4 Δύο παραδείγματα

Εξετάστε το k-οστό ως τον πλησιέστερο γειτονικό ταξινομητή, με k = 1.
Αυτό το σύνολο λειτουργιών έχει άπειρηVC διάσταση και μηδενικό εμπειρικό κίνδυνο, δεδομένου ότι οποιοσδήποτε αριθμός σημείων επισημαίνεται αυθαίρετα.
Έτσι, ο περιορισμός δεν παρέχει καμία πληροφορία.Στην πραγματικότητα, για κάθε ταξινόμηση με άπειρη διάσταση VC, το όριο δεν ισχύει.
Ωστόσο, ακόμη και αν ο περιορισμός δεν είναι έγκυρος, ο πλησιέστερος γειτονικός ταξινομητής μπορεί ακόμα να αποδώσει καλά. Έτσι, αυτό το πρώτο παράδειγμα είναι μια ιστορία όπου η άπειρη "χωρητικότητα" δεν εγγυάται την κακή απόδοση.

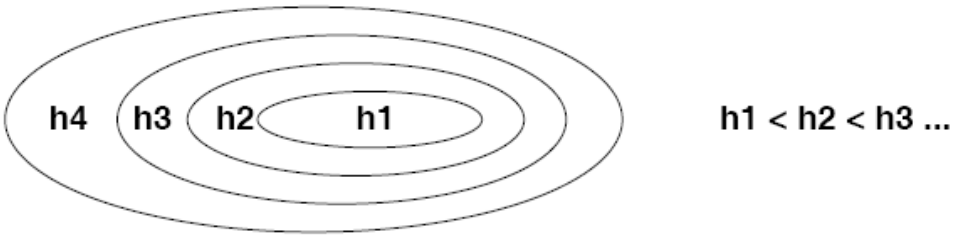
Ας ακολουθήσουμε τον παραδοσιακό χρόνο αντίληψης των πραγμάτων και να δούμε αν μπορούμε να καταλήξουμε σε μια ταξινόμηση που παραβιάζει τον περιορισμό.
Θέλουμε η αριστερή πλευρά της εξίσωσης να είναι όσο το δυνατόν μεγαλύτερη, και η δεξιά πλευρά να είναι όσο το δυνατόν μικρότερη. Θέλουμε συγγενείς ταξινομητές οι οποίοι να δίνουν το χειρότερο δυνατό πραγματικό κίνδυνο του 0.5, μηδέν εμπειρικό κίνδυνο σε κάποιες παρατηρήσεις, εκ των οποίων η VC διάσταση είναι εύκολο να υπολογιστεί και να είναι μικρότερη από l (έτσι ώστε το όριο να είναι μη τετριμμένο).
Ένα παράδειγμα είναι το εξής, το οποίο αποκαλούμε "ταξινομητή σημειωματάριου". Ο ταξινομητής αποτελείται από ένα σημειωματάριο με αρκετό χώρο για να γράψουμε τις τάξεις των m παρατηρήσεων, όπου m≤l. Για όλα τα επόμενα σχέδια, ο ταξινομητής λέει απλά ότι όλα τα σχέδια έχουν την ίδια τάξη.
Ας υποθέσουμε, επίσης, ότι τα δεδομένα έχουν πιο πολλά θετικά (y = +1) και αρνητικά (y = ⊂1) παραδείγματα, και ότι τα δείγματα επιλέγονται τυχαία. Ο ταξινομητής θα έχει μηδενικό εμπειρικό κίνδυνο μέχρι και για m παρατηρήσεις, 0.5 λάθος για όλες τις επόμενες παρατηρήσεις, 0.5 πραγματικό σφάλμα, και VC διάσταση h= m. Αντικαθιστώντας αυτές τις τιμές το όριο γίνεται

$$\frac{m}{4l} \le \ln(2l/m) + 1 - (1/m)\ln(\eta/4)$$

το οποίο ισχύει αν

$$f(z) = (\frac{z}{2})\exp^{(z/4-1)} \le 1, z \equiv (m/l), \quad 0 \le z \le 1.$$

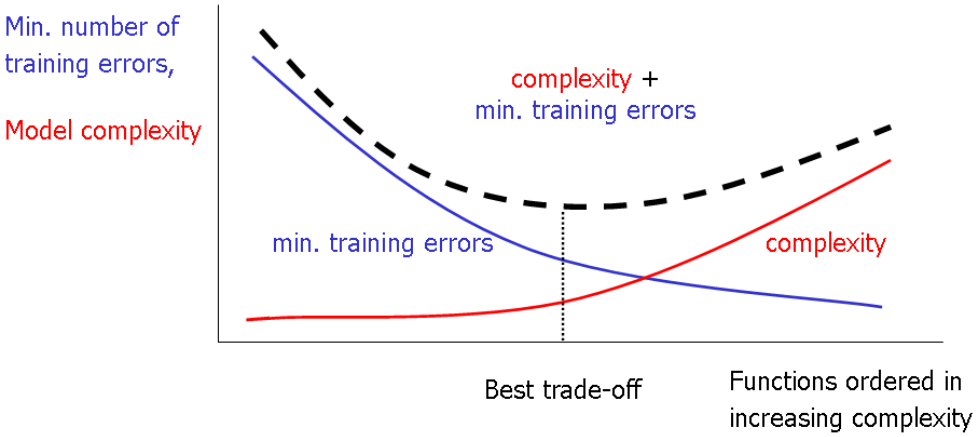
Η παραπάνω σχέση ισχύει εφόσον $f(z)$ είναι μονοτονικά αυξανόμενη και f(z=1)=0.236



Εικόνα 6.

1.3.5. Μάθηση και Γενίκευση

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

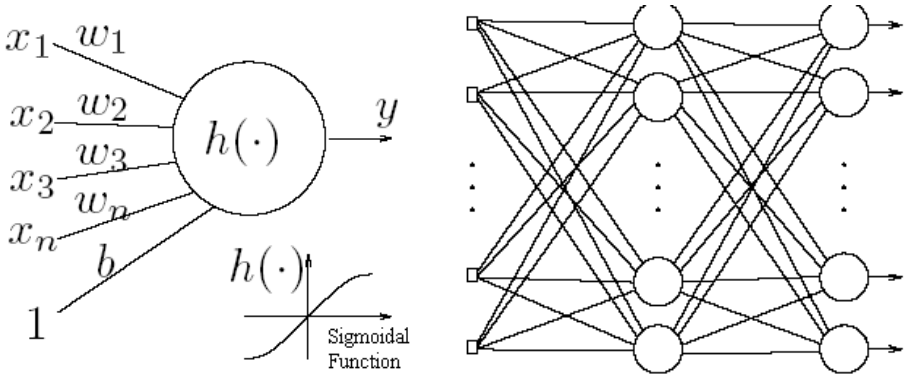


Εικόνα 6: Αριθμός Πολυπλοκότητας

2. Γιατί SVMs

Οι MLP(Multilayer perceptron)χρησιμοποιούν επαναλαμβανόμενα δίκτυα.

Οι MLP ιδιότητες περιλαμβάνουν καθολική προσέγγιση των συνεχών μη γραμμικών συναρτήσεων και επίσης περιλαμβάνουν προηγμένες αρχιτεκτονικές δικτύων με πολλαπλές εισόδους και εξόδους.



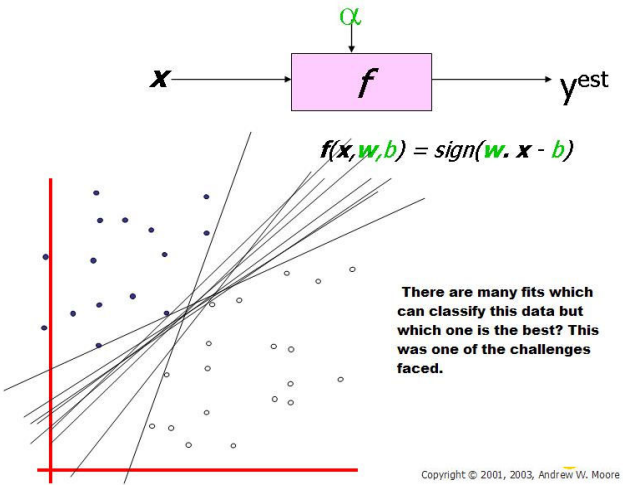
Εικόνα 7. a) Απλά Νευρωνικά Δικτύα b) Multilayer Perceptron δίκτυο

Μπορεί να υπάρχουν κάποια παρατηρούμενα ζητήματα:

- Μερικά από αυτά έχουν πολλά τοπικά ελάχιστα,
- Πόσα νευρωνικά ενδέχεται να χρειαστούν για μια εργασία

  είναι ένα άλλο ζήτημα που καθορίζει το κατά πόσον οι βελτιστοποίησεις της έχουν επιτευχθεί. Ένα άλλο πράγμα που αξίζει να σημειωθεί είναι ότι ακόμα και αν στο νευρωνικό δίκτυο που χρησιμοποιείται οι λύσεις τείνουν να συγκλίνουν, αυτό δεν μπορεί να οδηγήσει σε μια μοναδική λύση.

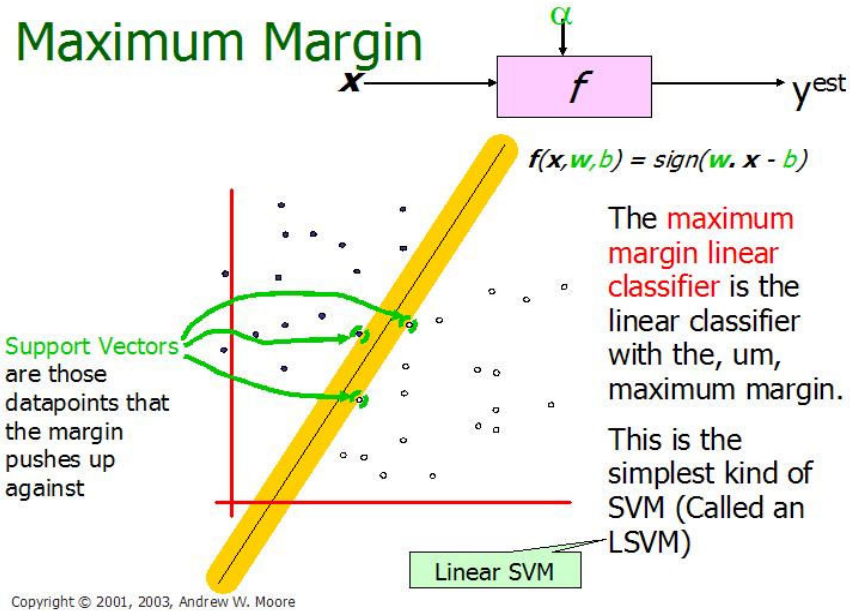
  Τώρα ας δούμε ένα άλλο παράδειγμα, όπου έχουμε ένα plot δεδομένων και να προσπαθήσουμε να τα ταξινομήσουμε.



Εικόνα 8. Υπάρχουν πολλά υπερσύνολα που μπορούν να χαρακτηρίσουν τα δεδομένα

Από το παραπάνω παράδειγμα, υπάρχουν πολλοί γραμμικοί ταξινομητές (υπερσύνολα) που διαχωρίζουν τα δεδομένα. Ωστόσο μόνο ένα από αυτά επιτυγχάνει το διαχωρισμό.

Η επόμενη εικόνα παρέχει το μέγιστο παράδειγμα ταξινόμησης περιθωρίου το οποίο παρέχει μια λύση στο παραπάνω πρόβλημα.



Εικόνα 9. Απεικόνιση της Γραμμικής SVM

Το μέγιστο περιθώριο δίνεται ως:

$$\text{margin} \equiv \arg\min_{\mathbf{x}\in D} d(\mathbf{x}) = \arg\min_{\mathbf{x}\in D} \frac{\left|\mathbf{x}\cdot\mathbf{w}+b\right|}{\sqrt{\sum_{i=1}^{d} w_i^2}}$$

Η παραπάνω εικόνα είναι ο μέγιστος γραμμικός ταξινομητής με το μέγιστο εύρος. Στο πλαίσιο αυτό, είναι ένα παράδειγμα ενός απλού γραμμικού ταξινομητή SVM.

Ένα άλλο ενδιαφέρον ερώτημα είναι γιατί μέγιστο περιθώριο; Υπάρχουν μερικές καλές εξηγήσεις που περιλαμβάνουν τη βελτίωση της εμπειρικής απόδοσης. Ένας λόγος είναι ότι ακόμα και αν έχουμε κάνει ένα μικρό λάθος στην τοποθεσία του ορίου, αυτό μας δίνει μικρότερη πιθανότητα να προκαλέσουμε εσφαλμένη κατάταξη.

Το άλλο πλεονέκτημα θα ήταν να αποφεύγονται τοπικά ελάχιστα και καλύτερη κατάταξη. Οι στόχοι της SVM έχουν το διαχωρισμό των δεδομένων από το υπερσύνολο και να επεκταθεί σε μη γραμμικά όρια χρησιμοποιώντας το τέχνασμα του πυρήνα. Βλέπουμε ότι ο στόχος είναι να ταξινομήσει σωστά όλα τα δεδομένα.

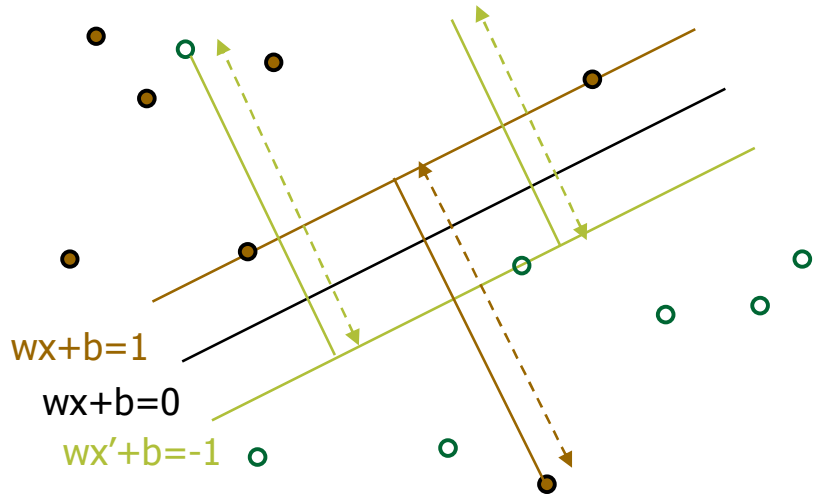
$[a]\,\varepsilon\acute{\alpha}\nu\ \mathrm{Y}_i = +1;\ wx_i + b \geq 1$

$[b]\,\varepsilon\acute{\alpha}\nu\ \mathrm{Y}_i = -1;\ wx_i + b \leq 1$

$[c]\,i;\ y_i(w_i + b) \geq 1$

Σε αυτήν την εξίσωση το *x* είναι ένα διανυσματικό σημείο και *w* είναι το βάρος και είναι επίσης ένα διάνυσμα. Έτσι, για το διαχωρισμό των δεδομένων [a], θα πρέπει πάντα να είναι μεγαλύτερα από το μηδέν. Ανάμεσα σε όλα τα πιθανά υπερσύνολα, η SVM επιλέγει εκείνο όπου η απόσταση του επιπέδου είναι όσο το δυνατόν μεγαλύτερη.

Το επιθυμητό επίπεδο που μεγιστοποιεί το περιθώριο διχοτομεί επίσης τις γραμμές μεταξύ πλησιέστερων σημείων στο κυρτό των δύο συνόλων δεδομένων. Έτσι έχουμε τα [a], [b] και [c].

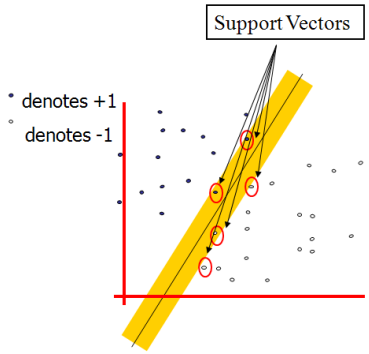


Εικόνα 10. Παρουσίαση υπερεπιπέδων

Έτσι, με την επίλυση και αφαίρεση των δύο αποστάσεων θα έχουμε τη σύνοψη της απόστασης του χωριζόμενου υπερεπιπέδου των πλησιέστερων σημείων με Μέγιστο Περιθώριο = M = 2 / ||w||.

Μεγιστοποιώντας το περιθώριο είναι ίδιο με το ελάχιστο. Έχουμε ένα τετραγωνικό πρόβλημα βελτιστοποίησης και πρέπει να το επιλύσουμε για *w* και b. Για να λυθεί αυτό θα πρέπει να βελτιστοποιηθεί η τετραγωνική συνάρτηση με γραμμικούς περιορισμούς. Η λύση περιλαμβάνει την κατασκευή ενός διπλού προβλήματος όπου είναι συνδεδεμένος ο πολλαπλασιαστής α_i Langlier . Πρέπει να βρούμε w και b τέτοια ώστε η Φ (w) =½ |w'||w| να ελαχιστοποιείται και για όλα τα $\{(x_i, y_i)\}$ να ισχύει η σχέση y_i (w * x_i + *b*) ≥ 1.

Τώρα για την επίλυση έχουμε ότι w =Σα_i * x_i; *b*= y_k- w *x_k για οποιαδήποτε x_k όπου $ak \neq 0$. Η λειτουργία ταξινόμησης θα έχει την ακόλουθη μορφή: *f*(x) = Σα_i y_i x_i * x + *b*



Εικόνα 11. Παρουσίαση Support Vectors

- SVM Παρουσίαση

Παρουσιάζουμε τις διατυπώσεις για την ταξινόμηση SVM .Αυτή είναι μια απλή απεικόνιση.

SV κατάταξη:

$\min_{f,\xi_i} \|f\|_K^2 + C\sum_{i=1}^{l} \xi_i$, για όλα τα $y_i f(\mathbf{x}_i) \geq 1 - \xi_i$, για όλα τα i, με $\xi_i \geq 0$

SVM ταξινόμηση, Διπλή διατύπωση:

$$\min_{\alpha_i} \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l}\sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \quad 0 \leq \alpha_i \leq C, \text{ για όλα τα i με}$$

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

Οι μεταβλητές ξ_i ονομάζονται χαλαρές μεταβλητές και μετρούν το σφάλμα στο σημείο (xi, yi).

Η SVM γίνεται αρκετά δύσκολη όταν ο αριθμός των σημείων εκπαίδευσης είναι μεγάλος.

2.1 Χαλαρό Περιθώριο Ταξινόμησης

Στο πραγματικό κόσμο το πρόβλημα δεν είναι πιθανό να πάρει μια ακριβή ξεχωριστή γραμμή που χωρίζει τα δεδομένα μέσα στο χώρο και μπορεί να έχουμε ένα κυρτό όριο απόφασης. Μπορεί να έχουμε ένα υπερεπίπεδο που θα μπορούσε να διαχωρίζει ακριβώς τα δεδομένα, αλλά αυτό μπορεί να μην είναι επιθυμητό, αν στα δεδομένα περιέχεται θόρυβος. Είναι καλύτερα για το χαλαρό όριο να αγνοήσει μερικά σημεία δεδομένων από το να είναι κυρτά ή να οδηγήσουν σε βρόχους, σε ακραίες τιμές.

Αυτό αντιμετωπίζεται με διαφορετικό τρόπο. Εισάγουμε λοιπόν τον όρο χαλαρές μεταβλητές. Έχουμε, $y_i(w'x + b) \geq 1 - S_k$. Αυτό επιτρέπει σε ένα σημείο να είναι σε μικρή απόσταση Sk στη λάθος πλευρά του υπερσυνόλου χωρίς να παραβιάζεται ο περιορισμός. Θα μπορούσαμε να καταλήξουμε έχοντας τεράστιες χαλαρές μεταβλητές που να επιτρέπουν σε κάθε γραμμή το διαχωρισμό των δεδομένων, έτσι ώστε σε αυτά τα σενάρια έχουμε την Lagrangian μεταβλητή η οποία ζημιώνει τη μεγάλη βραδύτητα.

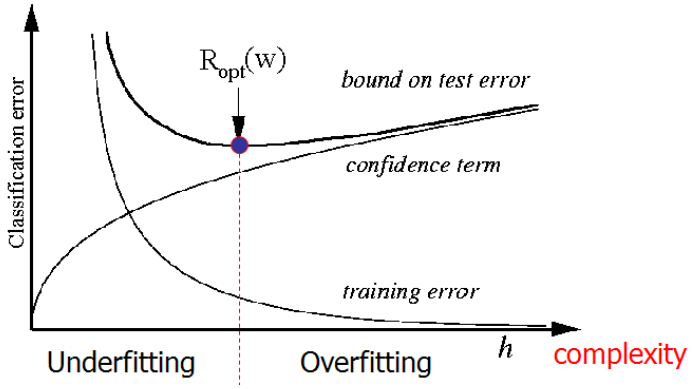
$\min L = ½\ w'w - \sum \lambda_k (y_k (w'x_k + b) + s_k - 1) + \alpha \sum s_k$

Σε περίπτωση μείωσης του α επιτρέπει σε περισσότερα δεδομένα να βρίσκονται στη λάθος πλευρά του υπερσυνόλου και θεωρούνται ως ακραίες τιμές που δίνουν ομαλότερο όριο απόφασης .

2.2 Έλεγχος Πολυπλοκότητας στην SVM: Ανταλλαγή

Η SVM είναι ισχυρή για την προσέγγιση των δεδομένων εκπαίδευσης, και γενικεύει καλύτερα τα δοσμένα σύνολα δεδομένων. Η πολυπλοκότητα από άποψη πυρήνα επηρεάζει την απόδοση των νέων δεδομένων. Η SVM υποστηρίζει τις παραμέτρους για τον έλεγχο της πολυπλοκότητας και πάνω απ όλα η SVM δεν μας λέει πώς να ρυθμίσουμε αυτές τις παραμέτρους και ότι πρέπει να είμαστε σε θέση να καθορίσουμε αυτές τις παραμέτρους μέσω διασταυρωμένης επικύρωσης (cross validation).

Το διάγραμμα που ακολουθεί δίνει μια καλύτερη εικόνα.



Εικόνα 12. Έλεγχος πολυπλοκότητας

2.2.1 Η SVM για την ταξινόμηση

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

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

2.2.2 Η SVM για την παλινδρόμηση

Οι SVMs μπορούν επίσης να εφαρμοστούν σε προβλήματα παλινδρόμησης με την εισαγωγή μιας εναλλακτικής λειτουργίας απώλειας . Η λειτουργία απώλειας πρέπει να τροποποιηθεί ώστε να συμπεριλάβει το μέτρο της απόστασης. Η παλινδρόμηση μπορεί να είναι γραμμική και μη γραμμική.Τα Γραμμικά μοντέλα αποτελούνται κυρίως από τις ακόλουθες λειτουργίες απώλειας, εντατικές λειτουργίες απώλειας, τετραγωνική και Huber λειτουργίες απώλ ιας. Ομοίως με τα προβλήματα ταξινόμησης, ένα μη γραμμικό μοντέλο συνήθως απαιτεί επαρκή δεδομένα. Με τον ίδιο τρόπο, μια μη γραμμική χαρτογράφηση μπορεί να χρησιμοποιηθεί για να χαρτογραφήσει τα δεδομένα σε ένα υψηλό διαστάσεων χώρο χαρακτηριστικών, όπου η γραμμική παλινδρόμηση εκτελείται. Η προσέγγιση του πυρήνα και πάλι χρησιμοποιείται για την αντιμετώπιση της διάστασης. Στη μέθοδο παλινδρόμησης υπάρχουν εκτιμήσεις που βασίζεται σε προγενέστερη γνώση του προβλήματος και τη διανομή του θορύβου.

2.3 Εφαρμογές της SVM

Η SVM έχει βρεθεί να είναι επιτυχής όταν χρησιμοποιείται για προβλήματα ταξινόμησης. Η εφαρμογή της προσέγγισης του Support Vector σε ένα συγκεκριμένο πρακτικό πρόβλημα αφορά την επίλυση μια σειράς από ερωτήματα με βάση τον ορισμό του προβλήματος και τον σχεδιασμό που εμπλέκονται με αυτό. Μία από τις μεγαλύτερες προκλήσεις είναι η επιλογή κατάλληλου πυρήνα για τη συγκεκριμένη εφαρμογή. Υπάρχουν επιλογές, όπως η **Gaussian** ή **το πολυώνυμο του πυρήνα** που είναι προεπιλεγμένες, αλλά αν αυτά αποδειχθούν αναποτελεσματικά ή αν οι είσοδοι είναι διακριτές δομές θα χρειαστούν πιο περίτεχνοι πυρήνες. Με σιωπηρό καθορισμό ενός χώρου χαρακτηριστικών, ο πυρήνας παρέχει τη χαρακτηριστική γλώσσα που χρησιμοποιείται από τη μηχανή για την προβολή των δεδομένων. Μετά την επιλογή του κριτηρίου του πυρήνα και τη βελτιστοποίηση έχουν τεθεί οι βασικές συνιστώσες του συστήματος.

Το έργο της κατηγοριοποίησης του κειμένου είναι η κατάταξη των φυσικών εγγράφων του κειμένου σε ένα σταθερό αριθμό από προκαθορισμένες κατηγορίες με βάση το περιεχόμενό τους. Δεδομένου ότι ένα έγγραφο μπορεί να αποδοθεί σε περισσότερες από μία κατηγορίες αυτό δεν είναι ένα πολυδιάστατο πρόβλημα ταξινόμησης, αλλά μπορεί να θεωρηθεί ως μια σειρά από διμερή προβλήματα ταξινόμησης, ένα για κάθε κατηγορία. Ένα από τα πρότυπα αναπαράστασης του κειμένου για τους σκοπούς της ανάκτησης πληροφοριών παρέχει μια ιδανική χαρτογράφηση χαρακτηριστικών για την κατασκευή ενός πυρήνα Mercer . Πράγματι, οι πυρήνες ενσωματώνουν κατά κάποιο τρόπο ένα μέτρο ομοιότητας μεταξύ των περιπτώσεων, και είναι λογικό να υποθέτουμε ότι οι εμπειρογνώμονες που εργάζονται στον συγκεκριμένο τομέα εφαρμογής έχουν ήδη εντοπίσει τη λήψη αποτελεσματικών μέτρων ομοιότητας, ιδίως σε τομείς όπως η ανάκτηση πληροφοριών και η γενεσιουργία μοντέλων.

Οι παραδοσιακές προσεγγίσεις ταξινόμησης έχουν κακές επιδόσεις όταν συνεργάζονται άμεσα λόγω των υψηλών διαστάσεων των δεδομένων, αλλά οι Support Vector Machines μπορούν να αποφύγουν τις παγίδες των πολύ υψηλών διαστάσεων αναπαραστάσεων. Μια παρόμοια προσέγγιση για τις τεχνικές που περιγράφονται για την κατηγοριοποίηση κειμένου μπορεί επίσης

να χρησιμοποιηθεί για την ταξινόμηση της εικόνας. Η πρώτη εργασία στην οποία οι SVM εξετάστηκαν ήταν το πρόβλημα της **χειρόγραφης αναγνώρισης χαρακτήρων**. Επιπλέον, οι πολυδιάστατες SVMs έχουν δοκιμαστεί σε αυτά τα δεδομένα. Είναι ενδιαφέρον, όχι μόνο να συγκριθούν οι SVMs με άλλους ταξινομητές, αλλά και να συγκρίνουν τις διαφορές των SVMs μεταξύ τους. Έχει αποδειχθεί ότι έχουν περίπου την ίδια απόδοση, και, επιπλέον, ότι μοιράζονται τους περισσότερους από τους φορείς υποστήριξης(SV), ανεξάρτητα από την επιλογή του πυρήνα. Το γεγονός ότι η SVM μπορεί να εκτελέσει, το ίδιο καλά με τα συστήματα αυτά χωρίς να περιλαμβάνει καμία λεπτομερή εκ των προτέρων γνώση είναι, οπωσδήποτε αξιόλογη.

2.3.1 Η δύναμη και η αδυναμία της SVM

Οι βασικές δυνατότητες της SVM είναι ότι η εκπαίδευση είναι σχετικά εύκολη. Κλιμακώνεται σε σχετικά καλές υψηλές διαστάσεις των δεδομένων και η εξισορρόπηση μεταξύ της ταξινόμησης της πολυπλοκότητας και του λάθους μπορεί να ελεγχθεί ρητά. Περιλαμβάνει την ανάγκη για μια καλή λειτουργία του πυρήνα.

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

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

Στα προβλήματα ταξινόμησης επιτυγχάνεται γενικευμένος έλεγχος με τη μεγιστοποίηση του περιθωρίου κέρδους, το οποίο αντιστοιχεί στην ελαχιστοποίηση του διανύσματος βάρους σε ένα κανονικό πλαίσιο. Η ελαχιστοποίηση του διανύσματος βάρους μπορεί να χρησιμοποιηθεί ως κριτήριο σε προβλήματα παλινδρόμησης, με μια τροποποιημένη λειτουργία απώλειας. Οι μελλοντικές κατευθύνσεις περιλαμβάνουν, **μια τεχνική για την επιλογή της λειτουργίας του πυρήνα και επιπλέον έλεγχος ικανότητας**. Τέλος, οι νέες κατευθύνσεις που αναφέρονται στη νέα SVM σχετίζονται με σκευάσματα μάθησης που προτάθηκαν πρόσφατα από τον Vapnik .

3 Γραμμικές SVMs

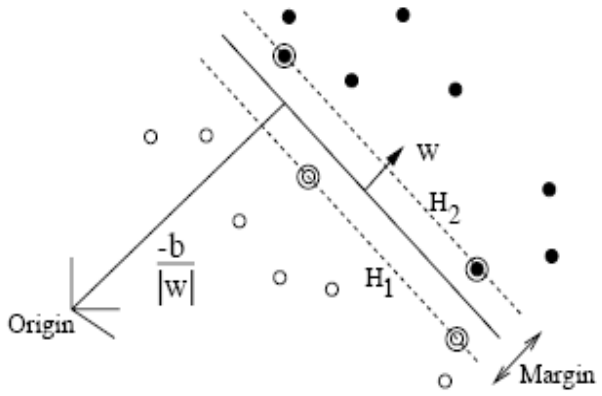
3.1 Η Διαχωριστική υπόθεση

Θα ξεκινήσουμε με την πιο απλή υπόθεση. Οι γραμμικές μηχανές ειδικεύονται στο να διαχωρίζουν τα δεδομένα (όπως θα δούμε, στην ανάλυση της γενικής περίπτωσης - μη γραμμικές μηχανές που ειδικεύονται σε μη διαχωρίσιμα δεδομένα- οδηγεί σε ένα παρόμοιο τετραγωνικό προγραμματιστικό πρόβλημα). Και πάλι προσδιορίζουμε τα δεδομένα $\{x_i, y_i\}, i = \ldots, l,\ y_i \in \{-1,1\},\ x_i \in R^d$. Ας υποθέσουμε ότι έχουμε κάποιο υπερσύνολο το οποίο διαχωρίζει τα θετικά από τα αρνητικά.Τα σημεία x που βρίσκονται στο υπερσύνολο που ικανοποιεί w * x+b=0, όπου W είναι κάθετη προς το υπερσύνολο, $|b|/\|w\|$ είναι η κάθετη απόσταση από το υπερσύνολο με το σημείο αναφοράς, και $\|w\|$ είναι η Ευκλείδεια νόρμα του w. Η $d_+ + d_-$ είναι η μικρότερη απόσταση μεταξύ του υπερσυνόλου και του πλησιέστερου θετικού (αρνητικού) παραδείγματος .

Ορίζουμε το "περιθώριο" ενός διαχωριστικού υπερσυνόλου ως $d_+ + d_-$.

Για την γραμμικά διαχωρίσιμη υπόθεση, ο SV αλγόριθμος ψάχνει το υπερσύνολο με το μεγαλύτερο περιθώριο. Αυτό μπορεί να διατυπωθεί ως εξής: ας υποθέσουμε ότι όλα τα δεδομένα ικανοποιούν τους παρακάτω περιορισμούς:

$$x_i \cdot w + b \geq +1,\ \gamma\iota\alpha\ \ y_i = +1$$
$$x_i \cdot w + b \leq +1,\ \gamma\iota\alpha\ \ y_i = -1$$



Εικόνα 13

Αυτά μπορούν να συνδυαστούν σε ένα ενιαίο σύνολο ανισοτήτων

$$y_i(x_i \cdot w + b) - 1 \geq 0, \ \forall i$$

Τα σημεία τα οποία η πρώτη ισότητα κατέχει βρίσκονται στο υπερσύνολο $H_1 : x_i \cdot w + b = 1$ με κανονική w και την κάθετη απόσταση $|1-b| / \|w\|$.

Ομοίως, τα σημεία για την δεύτερη ισότητα βρίσκονται επί του υπερσυνόλου $H_2 : x_i \cdot w + b = -1$, με κανονική και πάλι w, και κάθετη απόσταση $|-1-b| / \|w\|$.

Ως εκ τούτου $d_+ = d_- = 1/\|w\|$ και το περιθώριο είναι απλά $2/\|w\|$.

Σημειώστε ότι οι H_1 και H_2 είναι παράλληλες (έχουν την ίδια κανονική κατανομή) και ότι δεν βρίσκονται μεταξύ τους σημεία. Έτσι μπορούμε να βρούμε το ζεύγος των υπερσυνόλων που δίνει το μέγιστο περιθώριο με την ελαχιστοποίηση $\|w\|^2$, βάση των περιορισμών.

Έτσι, αναμένουμε ότι η λύση για μια τυπική περίπτωση δύο διαστάσεων έχει τη μορφή που υποδεικνύεται στην Εικόνα 13. Αυτά τα σημεία τα οποία η ισότητα κατέχει και των οποίων η αφαίρεση θα άλλαζε την λύση που βρέθηκε, ονομάζονται διανύσματα υποστήριξης (SV) και απεικονίζονται στην Εικόνα 13 με επιπλέον κύκλους. Τώρα θα μεταβούμε σε μια Lagrangian διατύπωση του προβλήματος.

Υπάρχουν δύο λόγοι για να γίνει αυτό. Ο πρώτος είναι ότι οι περιορισμοί θα αντικατασταθούν από περιορισμούς Πολλαπλασιαστών Lagrange , οι οποίοι θα είναι πολύ πιο εύκολοι στον χειρισμό. Ο δεύτερος είναι ότι με την παρούσα αναδιατύπωση του προβλήματος, τα δεδομένα θα εμφανίζονται μόνο με τη μορφή των κουκκίδων προϊόντων μεταξύ των διανυσμάτων. Αυτό είναι κρίσιμο διότι θα μας επιτρέψει να γενικευθεί η διαδικασία με μη γραμμική περίπτωση. Έτσι, εισάγουμε θετικούς πολλαπλασιαστές Lagrange $a_i, i = 1, ..., l$ ένα για κάθε ένα από τους περιορισμούς . Υπενθυμίζεται ότι ο κανόνας αυτός είναι για περιορισμούς της μορφής $c_i \geq 0$, οι εξισώσεις περιορισμού πολλαπλασιάζονται με θετικούς πολλαπλασιαστές Lagrange και αφαιρούνται από την αντικειμενική συνάρτηση, για να σχηματίσουν τη Lagrangian. Για τους περιορισμούς της ισότητας, οι Lagrange πολλαπλασιαστές είναι απεριόριστοι. Αυτό δίνει την Lagrangian:

$$L_p \equiv \frac{1}{2}\|w\|^2 - \sum_{i=1}^{l} a_i y_i (x_i \cdot w + b) + \sum_{i=1}^{l} a_i$$

Πρέπει να ελαχιστοποιήσουμε τώρα την L_p ως προς w, b, και ταυτόχρονα απαιτούμε τα παράγωγα του L_p σε σχέση με όλες τις a_i, όλα υπόκεινται στους περιορισμούς $a_i \geq 0$ (ας ονομάσουμε αυτό το συγκεκριμένο σύνολο περιορισμών C_1). Αυτό είναι ένα κυρτό τετραγωνικό προγραμματιστικό πρόβλημα, δεδομένου ότι η αντικειμενική

συνάρτηση είναι η ίδια κυρτή, και τα σημεία που ικανοποιούν τους περιορισμούς αποτελούν επίσης μια κυρτή σειρά. Αυτό σημαίνει ότι μπορούμε να λύσουμε ισοδύναμα το ακόλουθο: «διπλό» πρόβλημα: μεγιστοποίηση L_p, μια επιφύλαξη που εξαφανίζεται μέσω των περιορισμών της κλίσης του L_p ως προς w και b, και επίσης στους περιορισμούς $a_i \geq 0$ (ας ονομάσουμε αυτό το συγκεκριμένο σύνολο περιορισμοί C_2).
Αυτή η συγκεκριμένη διπλή διατύπωση του προβλήματος ονομάζεται διπλής Wolfe. Έχει την ιδιότητα ότι το μέγιστο του L_p, με την επιφύλαξη των περιορισμών C_2, εμφανίζεται στις ίδιες τιμές των w, b και α, ως το ελάχιστο του L_p, που υπόκεινται στους περιορισμούς C_1.

Η απαίτηση ότι η κλίση του L_p σε σχέση με w και b θα δώσει τις προϋποθέσεις:

$$w = \sum_i a_i y_i x_i$$

$$\sum_i a_i y_i = 0$$

Δεδομένου ότι πρόκειται για περιορισμούς ισότητας , μπορούμε να τα αντικαταστήσουμε:

$$L_D = \sum_i a_i - \frac{1}{2}\sum_{i,j} a_i a_j y_i y_j x_i \cdot x_j$$

Σημειώστε τους Lagrangian προσδιορισμούς (P για την πρωταρχική, D για τη διπλή) για να τονίσουμε ότι είναι διαφορετικά: οι L_p και L_D προκύπτουν από τον ίδια λειτουργία αλλά με διαφορετικούς περιορισμούς και η λύση βρίσκεται με την ελαχιστοποίηση L_p ή από μεγιστοποίηση L_D. Σημειώστε επίσης ότι αν διατυπώσουμε το πρόβλημα με b = 0, το οποίο απαιτεί όλα τα υπερσύνολα να περιέχουν την προέλευση, ο περιορισμός δεν εμφανίζεται. Αυτός είναι ένα ήπιος περιορισμός για υψηλών διαστάσεων χώρους, διότι ισοδυναμεί με μείωση του αριθμού των βαθμών ελευθερίας. Οι SV (γραμμική περίπτωση) ανέρχονται συνεπώς, για τη μεγιστοποίηση L_D σε σχέση με το a_i, υπόκειται στους περιορισμούς και στη θετικότητα του a_i. Σημειώστε ότι υπάρχει ένας πολλαπλασιαστής Lagrange a_i για κάθε σημείο της κατάρτισης. Σε τελική μορφή, αυτά τα σημεία για τα οποία $a_i > 0$ αποκαλούνται "φορείς υποστήριξης" (SV), και βρίσκονται στα υπερσύνολα H_1, H_2.
Όλα τα άλλα σημεία έχουν $a_i = 0$ και βρίσκονται είτε στο H_1 ή στο H_2.
Για αυτά τα μηχανήματα, οι SV είναι τα κρίσιμα στοιχεία του συνόλου των δεδομένων. Βρίσκονται πιο κοντά στο όριο της απόφασης. Αν όλα τα σημεία είχαν αφαιρεθεί (ή αν μετακινούνταν έτσι ώστε να μην διασχίζουν τα H_1 ή H_2), και αν επαναλαμβανόταν, θα είχε βρεθεί το ίδιο υπερσύνολο.

3.2 Οι Προϋποθέσεις Karush-Kuhn-Tucker (KKT)

Οι Karush-Kuhn-Tucker (KKT) συνθήκες παίζουν κεντρικό ρόλο τόσο στην θεωρία όσο και πρακτική της βελτιστοποίησης. Για το παραπάνω πρωταρχικό πρόβλημα, οι όροι KKT μπορεί να αποδοθούν ως εξής :

$$\frac{\partial}{\partial w_v} L_P = w_v - \sum_i a_i y_i x_{iv} = 0, \quad v = 1, ..., d$$

$$\frac{\partial}{\partial b} L_P = -\sum_i a_i y_i = 0$$

$$y_i (x_i \cdot w + b) - 1 \geq 0, \quad i = 1, ..., l$$

$$a_i \geq 0, \forall i$$

$$a_i (y_i (w \cdot x_i + b) - 1) = 0, \forall i$$

Οι προϋποθέσεις KKT πληρούνται στην επίλυση περιορισμένων προβλημάτων βελτιστοποίησης (κυρτά ή μη), με κάθε είδους περιορισμούς, με την προϋπόθεση ότι η τομή του συνόλου των εφικτών κατευθύνσεων σε σχέση με το σύνολο των κατευθύνσεων καθόδου, συμπίπτει στο σημείο τομής.
Αυτή η τεχνική υπόθεση κανονικότητας ισχύει για όλες τις SVMs, δεδομένου ότι οι περιορισμοί είναι πάντοτε γραμμικοί. Επιπλέον, το πρόβλημα για SVMs είναι κυρτό (κυρτή αντικειμενική συνάρτηση, με περιορισμούς που δίνουν μία κυρτή περιοχή), και για κυρτά προβλήματα (αν ισχύει η κανονικότητα), οι KKT προϋποθέσεις είναι ικανές και αναγκαίες για τις w, b, α έτσι ώστε να υπάρχει λύση.

Έτσι, η επίλυση του προβλήματος SVM είναι ισοδύναμη με την εξεύρεση λύσης σε KKT συνθήκες.
Ως άμεση εφαρμογή, σημειώστε ότι, ενώ το w καθορίζεται ρητά από την πειραματική διαδικασία, το b όριο δεν καθορίζεται, αν και καθορίζεται σιωπηρά. Ωστόσο το b βρίσκεται εύκολα με τη χρήση του KKT «συμπληρωματικής» κατάστασης, επιλέγοντας i για το οποίο $a_i \neq 0$ και υπολογίζουμε το b (σημειώστε ότι είναι αριθμητικά ασφαλέστερο να λάβουμε τη μέση τιμή τουb που προκύπτει από όλες αυτές τις εξισώσεις).

3.3 Βέλτιστο υπερσύνολο: Ένα Παράδειγμα

Ο κύριος στόχος είναι να διερευνήσουμε ένα μη τετριμμένο πρόβλημα αναγνώρισης προτύπων, όπου η λύση του διανύσματος υποστήριξης(SV) μπορεί να βρεθεί αναλυτικά. Για το συγκεκριμένο πρόβλημα θεωρείται ότι, κάθε σημείο θα είναι ένα διάνυσμα υποστήριξης, το οποίο είναι ένας λόγος που μπορούμε να βρούμε τη λύση αναλυτικά. Εξετάστε το ενδεχόμενο n + 1 συμμετρικά τοποθετημένα σημεία να βρίσκονται σε μια σφαίρα S^{n-1}ακτίνας R: πιο συγκεκριμένα, τα σημεία αποτελούν τις κορυφές ενός n διαστάσεων συμμετρικού συμπλέγματος. Είναι βολικό να ενσωματώσουμε τα σημεία στο R^{n+1} με τέτοιο

τρόπο ώστε να βρίσκονται όλα στο υπερσύνολο που περνά μέσα από το σημείο αναφοράς, το οποίο είναι κάθετο προς το (n+1)-φορέα (1,…,1). Οι συντεταγμένες δίνονται από

$$x_{i\mu} = -(1-\delta_{i,\mu})\sqrt{\frac{R}{n(n+1)}} + \delta_{1,\mu}\sqrt{\frac{R_n}{n+1}}$$

όπου ο δέλτα Kronecker $\delta_{i,\mu}$ ορίζεται από $\delta_{i,\mu} = 1$ αν μ=i, αλλιώς 0.

Έτσι, για παράδειγμα, οι φορείς για τρία ισαπέχοντα σημεία στο μοναδιαίο κύκλο δίνονται ως εξής :

$$x_1 = (\sqrt{\frac{2}{3}}, \frac{-1}{\sqrt{6}}, \frac{-1}{\sqrt{6}})$$
$$x_2 = (\frac{-1}{\sqrt{6}}, \sqrt{\frac{2}{3}}, \frac{-1}{\sqrt{6}})$$
$$x_3 = (\frac{-1}{\sqrt{6}}, \frac{-1}{\sqrt{6}}, \sqrt{\frac{2}{3}})$$

Μια συνέπεια της συμμετρίας είναι ότι η γωνία μεταξύ όποιουδήποτε ζεύγους φορέων είναι η ίδια (και είναι ίση με $\arccos(-1/n)$):

$$\|x_i\|^2 = R^2 \quad \textbf{ή} \quad \frac{x_i \cdot x_j}{R^2} = \delta_{i,j} - (1-\delta_{i,j})\frac{1}{n}$$

$$x_i \cdot x_j = -R^2/n$$

Εκχωρώντας μία κλάση $C \in \{+1,-1\}$ αυθαίρετα σε κάθε σημείο, θέλουμε να βρούμε το υπερσύνολο που χωρίζει τις δύο κατηγορίες με το μεγαλύτερο περιθώριο κέρδους. Έτσι, πρέπει να μεγιστοποιήσουμε την L_D, με την επιφύλαξη ότι $a_i \geq 0$. Η στρατηγική μας είναι να λύσουμε απλώς το πρόβλημα σαν να μην υπήρχε κανένας περιορισμός. Αν το αποτέλεσμα που προκύπτει ανταποκρίνεται στην πραγματικότητα $a_i \geq 0, \forall i$, τότε θα έχουμε βρει τη γενική λύση, δεδομένου ότι το πραγματικό μέγιστο της L_D θα βρίσκεται στην εφικτή περιοχή, υπό την προϋπόθεση των περιορισμών παραπάνω. Προκειμένου να επιβάλουμε τον περιορισμό της ισότητας εισαγάγουμε ένα πρόσθετο πολλαπλασιαστή Lagrange λ. Τον οποίο προσπαθούμε να μεγιστοποιήσουμε

$$L_D \equiv \sum_{i=1}^{n+1} a_i - \frac{1}{2}\sum_{i,j=1}^{n+1} a_i H_{ij} a_j - \lambda \sum_{i=1}^{n+1} a_i y_i$$

Εισάγουμε την ανίσωση Hessian $H_{ij} = y_i y_j x_i \cdot x_j$.

Θέτουμε $\frac{\partial L_D}{\partial a_i} = 0$ και μας δίνει $(Ha)_i + \lambda y_i = 1, \forall i$.

Η ανίσωση H έχει μια πολύ απλή δομή: τα μη διαγώνια στοιχεία είναι $-y_i y_j R^2 / n$, και τα διαγώνια στοιχεία είναι R². Το γεγονός ότι όλα τα εκτός της διαγωνίου στοιχεία διαφέρουν μόνο από τους παράγοντες του y_i μας παροτρύνει στο να ψάξουμε για μια λύση η οποία να έχει τη μορφή

$$a_i = (\frac{1+y_i}{2})a + (\frac{1+y_i}{2})b$$

όπου ο α και ο b είναι άγνωστοι. Ενσωματώνοντας έχουμε:

$$(\frac{n+1}{n})(\frac{a+b}{2}) - \frac{y_i p}{n}(\frac{a+b}{2}) = \frac{1-\lambda y_i}{R^2}$$

όπου $p \equiv \sum_{i=1}^{n+1} y_i$ και $a+b = \frac{2n}{R^2(n+1)}$.

Προκειμένου να βρούμε τον α και b προχωράμε ως εξής:

$$a = \frac{n}{R^2(n+1)}(1 - \frac{p}{n+1}), b = \frac{n}{R^2(n+1)}(1 + \frac{p}{n+1})$$

το οποίο μας δίνει την λύση

$$a_i = \frac{n}{R^2(n+1)}(1 - \frac{y_i p}{n+1}).$$

Επίσης,

$$(Ha)_i = 1 - \frac{y_i p}{n+1}$$

Επομένως,

$$\|w\|^2 = \sum_{i,j=1}^{n+1} a_i a_j y_i y_j x_i \cdot x_j = a^T H_a$$

$$= \sum_{i=1}^{n+1} a_i (1 - \frac{y_i p}{n+1}) = \sum_{i=1}^{n+1} a_i = (\frac{n}{R^2})(1 - (\frac{p}{n+1})^2)$$

Σημειώστε ότι αυτή είναι μια από εκείνες τις περιπτώσεις όπου ο πολλαπλασιαστής Lagrange λ μπορεί να παραμένει απροσδιόριστος (αν και ο προσδιορισμός του είναι τετριμμένος). Έχουμε λύσει τώρα το πρόβλημα, δεδομένου ότι όλα τα a_i είναι σαφώς θετικά ή μηδέν (στην πραγματικότητα ο a_i θα είναι μηδέν αν έχουν όλα τα σημεία την ίδια κατηγορία). Σημειώστε ότι ο $\|w\|$ εξαρτάται μόνο από τον αριθμό των θετικών (αρνητικών) πολικών σημείων, και όχι από το πώς κατηγοριοποιούνται τα σημεία .Αυτό σαφώς δεν ισχύει στον w, ο οποίος δίνεται από

$$w = \frac{n}{R^2 (n+1)} \sum_{i=1}^{n+1} (y_i - \frac{p}{n+1}) x_i .$$

Το περιθώριο $M = 2 / \|w\|$ δίνεται από

$$M = \frac{2R}{\sqrt{n(1 - (p/(n+1))^2)}} .$$

Έτσι, όταν ο αριθμός των σημείων είναι n + 1 , το ελάχιστο περιθώριο εμφανίζεται όταν p = 0 (ίσος αριθμός θετικών και αρνητικών παραδειγμάτων), στην οποία περίπτωση το περιθώριο είναι $M_{\min} = 2R / \sqrt{n}$. Εάν n + 1 είναι περιττός, το ελάχιστο περιθώριο εμφανίζεται όταν $p = \pm 1$, στην οποία περίπτωση $M_{\min} = 2R(n+1) / (n\sqrt{n+2})$.

Και στις δύο περιπτώσεις, το ανώτατο περιθώριο δίνεται από $M_{\max} = R(n+1)/n$.

Έτσι, για παράδειγμα, για το δισδιάστατο simplex που αποτελείται από τρία σημεία που βρίσκονται στον S^1 (και που εκτείνονται στον R^2), και με την επισήμανση, ότι δεν έχουν και τα τρία σημεία την ίδια πολικότητα, το μέγιστο και το ελάχιστο περιθώριο είναι 3R/2.

3.4. Δοκιμαστική Φάση

Εμείς απλά καθορίζουμε σε ποια πλευρά του ορίου (το υπερσύνολο βρίσκεται ακριβώς στη μέση μεταξύ H_1 και H_2 και παράλληλα με αυτά) ένα δεδομένο διάγραμμα δοκιμής x βρίσκεται και του ανατίθεται η αντίστοιχη τάξη, δηλαδή παίρνουμε την κλάση του x να είναι $\mathrm{sgn}(w \cdot x + b)$.

3.5 Η μη διαχωρίσιμη υπόθεση

Ο παραπάνω αλγόριθμος για διαχωρίσιμα δεδομένα, όταν εφαρμόζεται σε μη διαχωρίσιμα δεδομένα, δε θα βρει καμία εφικτή λύση: αυτό αποδεικνύεται από την

αντικειμενική συνάρτηση (δηλαδή το διπλό Lagrangian) που μεγαλώνει αυθαίρετα. Θα θέλαμε να χαλαρώσουμε τους περιορισμούς, αλλά μόνο όταν αυτό είναι απαραίτητο.
Θα θέλαμε να εισάγουμε ένα περαιτέρω κόστος (δηλαδή αύξηση της πρωταρχικής λειτουργικής συνάρτησης).
Αυτό μπορεί να γίνει με την εισαγωγή θετικών μεταβλητών $\xi_i, i = 1,...,l$ στους περιορισμούς, οι οποίοι στη συνέχεια γίνονται:

$$x_i \cdot w + b \geq +1 - \xi_i, \quad \gamma\iota\alpha\, y_i = +1$$
$$x_i \cdot w + b \leq -1 + \xi_i, \quad \gamma\iota\alpha\, y_i = -1$$

και

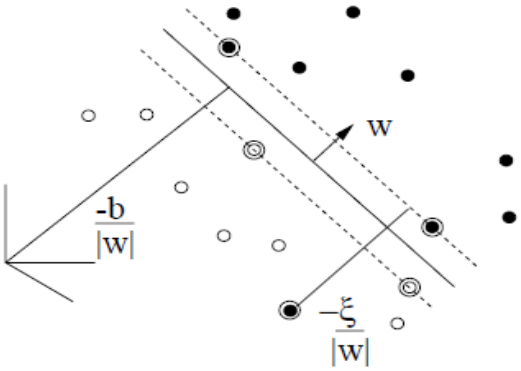
$$\xi_i \geq 0, \forall i$$

Έτσι για να συμβεί ένα λάθος , το αντίστοιχο ξ_i πρέπει να υπερβαίνει την ενότητα, έτσι το $\sum_i \xi_i$ είναι ένα ανώτερο όριο για τον αριθμό των λαθών . Ως εκ τούτου, ένας φυσικός τρόπος για να δοθεί έναν επιπλέον κόστος για τα λάθη είναι να αλλάξει η αντικειμενική συνάρτηση και το ελάχιστο $\|w\|^2 / 2\mu\acute{\epsilon}\chi\rho\iota \|w\|^2 / 2 + C(\sum_i \xi_i)^k$ όπου C είναι μια παράμετρος που θα επιλεγεί από το χρήστη, ένα μεγαλύτερο C το οποίο ανταποκρίνεται σε μεγαλύτερες «ποινές» των λαθών. Όπως έχουν τα πράγματα, αυτό είναι ένα κυρτό πρόβλημα προγραμματισμού για κάθε θετικό k ακέραιο, για k = 2 και k = 1,το οποίο είναι επίσης ένα τετραγωνικό πρόβλημα προγραμματισμού, και η επιλογή k = 1 έχει το επιπλέον πλεονέκτημα ότι ούτε το ξ_i , ούτε οι πολλαπλασιαστές Lagrange, εμφανίζονται στο Wolfe διπλό πρόβλημα, το οποίο γίνεται:

$$L_D \equiv \sum_i a_i - \frac{1}{2}\sum_{i,j} a_i a_j y_i y_j x_i \cdot x_j$$
$$0 \leq a_i \leq C$$
$$\sum_i a_i y_i = 0$$

Η λύση δίνεται από την εξίσωση $w = \sum_{i=1}^{N_S} a_i y_i x_i$ όπου N_S είναι ο αριθμός των φορέων υποστήριξης.
Έτσι, η μόνη διαφορά από τη βέλτιστο υπερσύνολο είναι η περίπτωση του a_i όπου έχει ένα άνω όριο του C.Η κατάσταση συνοψίζεται σχηματικά στην Εικόνα 14. Θα χρειαστούμε τις Karush-Kuhn-Tucker προϋποθέσεις για το πρόβλημα. Το κυρίαρχο Lagrangian είναι

$$L_p \equiv \frac{1}{2}\|w\|^2 + C\sum_i \xi_i - \sum_{i=1}^{l} a_i \{y_i(x_i \cdot w + b) - 1 + \xi_i\} - \sum_i \mu_i \xi_i$$



Εικόνα 14

όπου μ_i οι πολλαπλασιαστές Lagrange που εισήχθησαν για την επιβολή θετικότητας του ξ_i.
Οι ΚΚΤ προϋποθέσεις για το αρχικό πρόβλημα είναι ως εκ τούτου:

$$\frac{\partial L_P}{\partial w_v} = w_v - \sum_i a_i y_i x_{iv} = 0$$

$$\frac{\partial L_P}{\partial b} = -\sum_i a_i y_i = 0$$

$$\frac{\partial L_P}{\partial \xi_i} = C - a_i - \mu_i = 0$$

$$y_i (x_i \cdot w + b) - 1 + \xi_i \geq 0$$

$$\xi_i \geq 0$$

$$a_i \geq 0$$

$$\mu_i \geq 0$$

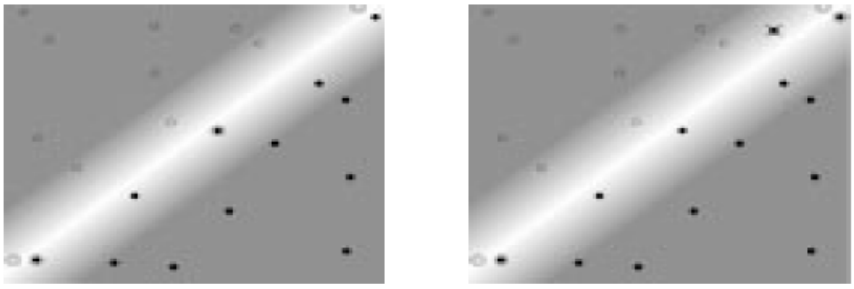
$$a_i \{ y_i (x_i \cdot w + b) - 1 + \xi_i \} = 0$$

$$\mu_i \xi_i = 0$$

Όπως και πριν, μπορούμε να χρησιμοποιήσουμε τις ΚΚΤ συνθήκες συμπληρωματικότητας, για να καθοριστεί το όριο b. Εφόσον $\xi_i = 0$ και $a_i < C$.
Έτσι μπορούμε να πάρουμε απλά οποιοδήποτε σημείο της κατάρτισης, για την οποία $0 < a_i < C$, για χρήση και για τον υπολογισμό του b.

3.6 Μηχανολογική αναλογία

Σκεφτείτε την περίπτωση κατά την οποία τα δεδομένα είναι σε R². Ας υποθέσουμε ότι το i -στο διάνυσμα υποστήριξης ασκεί μια δύναμη $F_i = a_i y_i \hat{w}$ που βρίσκεται κατά μήκος της επιφάνειας απόφασης (το "φύλλο απόφαση")



Εικόνα 15

Στη συνέχεια, η λύση πληρεί τους όρους της μηχανικής ισορροπίας:

$$\sum Forces = \sum_i a_i y_i \hat{w}$$

$$\sum Torques = \sum_i s_i \wedge (a_i y_i \hat{w}) = \hat{w} \wedge w = 0$$

(Εδώ η s_i είναι τα διανύσματα υποστήριξης, και ^ δηλώνει το προϊόν του διανύσματος.) Για τα δεδομένα στην R^n, σαφώς υπάρχει η πιθανότητα ότι το άθροισμα των δυνάμεων να μην υπάρχει.

4 Μη γραμμικά Μηχανήματα διανυσματικής Υποστήριξης (SVMs)

Πρώτον, ο μόνος τρόπος με τον οποίο τα δεδομένα εμφανίζεται στο πρόβλημα, στις εξισώσεις είναι με τη μορφή των dot προϊόντων, $x_i * x_j$.
Τώρα ας υποθέσουμε ότι έχουμε χαρτογραφήσει τα αρχεία για κάποιο άλλο (ενδεχομένως άπειρων διαστάσεων) Ευκλείδειο Η χώρο, χρησιμοποιώντας
μια χαρτογράφηση που έχουμε καλέσει Φ: $\Phi : R^d \mapsto H$.
Τότε βέβαια ο αλγόριθμος θα εξαρτάται μόνο από τα στοιχεία των dot προϊόντων στην Η, δηλαδή $\Phi(x_i) \cdot \Phi(x_j)$. Τώρα, αν υπήρχε μια «λειτουργία πυρήνα" Κ όπως , $\mathrm{K}(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j)$ θα έπρεπε μόνο να χρησιμοποιούσαμε την Κ στον αλγόριθμο, και δεν θα έπρεπε να γνωρίζουμε τι είναι η Φ. Ένα παράδειγμα είναι η $\mathrm{K}(x_i, x_j) = e^{-\|x_i - x_j\|^2 / 2\sigma^2}$.

Στο συγκεκριμένο παράδειγμα, η Η είναι άπειρων διαστάσεων, οπότε δεν θα ήταν πολύ εύκολο να συνδυαστεί με Φ . Ωστόσο, εάν κάποιος αντικαταστήσει $x_i \cdot x_j$ με $\mathrm{K}(x_i, x_j)$ παντού στον αλγόριθμο, ο αλγόριθμος θα παράγει μια SVM που υπάρχει σε ένα άπειρο τρισδιάστατο χώρο και επιπλέον το κάνει σε περίπου ίδιο χρόνο
με την εκπαίδευση των μη χαρτογραφημένων δεδομένων. Όλες οι προηγούμενες εκτιμήσεις ισχύουν, δεδομένου ότι κάνουμε ακόμα ένα γραμμικό διαχωρισμό, αλλά σε διαφορετικό χώρο. Αλλά πώς μπορούμε να χρησιμοποιήσουμε αυτό το μηχάνημα; Μετά από όλα, χρειαζόμαστε w, που να ανήκει στον Η. Αλλά σε αυτή τη φάση δοκιμάσαμε SVM με

$$f(x) = \sum_{i=1}^{N_s} a_i y_i \Phi(s_i) \cdot \Phi(x) + b = \sum_{i=1}^{N_S} a_i y_i K(s_i, x) + b$$

όπου η s_i είναι τα διανύσματα υποστήριξης(SV). Έτσι, και πάλι μπορούμε να αποφύγουμε τον υπολογισμό $\Phi(x)$ και να χρησιμοποιήσουμε το $K(s_i, x) = \Phi(s_i) \cdot \Phi(x)$.

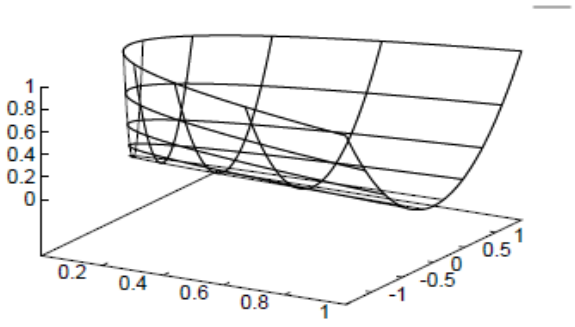
Ας καλέσουμε τον χώρο στον οποίο βρίσκονται τα δεδομένα, L.

Σημειώστε επίσης ότι είναι εύκολο να βρεθεί πυρήνας (για παράδειγμα, οι πυρήνες που είναι λειτουργίες των προϊόντων βαθμωτών dot του x_i σε L) τέτοια ώστε ο αλγόριθμος και η λύση να είναι ανεξάρτητες από τη διάσταση των δύο, L και Η.

Ένα πολύ απλό παράδειγμα ενός επιτρέποντα πυρήνα, για το οποίο μπορούμε να κατασκευάσουμε την χαρτογράφηση του Φ είναι το παρακάτω.

Ας υποθέσουμε ότι τα δεδομένα μας είναι διανύσματα στην R^2, και μπορούμε να επιλέξουμε $\mathrm{K}(x_i, x_j) = (x_i \cdot x_j)^2$.

Τότε είναι εύκολο να βρούμε ένα χώρο Η, και να χαρτογραφήσουμε το Φ, έτσι ώστε , επιλέγουμε Η=R³ και



Εικόνα 16. Χαρτογράφηση του Φ

$$\Phi(x) = \begin{pmatrix} x_1^2 \\ \sqrt{2}x_1x_2 \\ x_2^2 \end{pmatrix}$$

(σημειώστε ότι εδώ οι δείκτες αναφέρονται σε διάνυσμα). Για τα δεδομένα σε L προσδιορίζονται στο τετράγωνο $(-1,1)x(-1,1) \in R^2$.

Σημειώστε ότι ούτε η χαρτογράφηση Φ ούτε ο χώρος Η είναι μοναδικά για ένα δεδομένο πυρήνα.Θα μπορούσαμε να επιλέξουμε εξίσου ο Η να ανήκει R³ και

$$\Phi(x) = \frac{1}{\sqrt{2}} \begin{bmatrix} (x_1^2 - x_2^2) \\ 2x_1x_2 \\ (x_1^2 + x_2^2) \end{bmatrix}$$

ή ο Η να ανήκει R⁴

$$\Phi(x) = \begin{bmatrix} x_1^2 \\ x_1x_2 \\ x_1x_2 \\ x_2^2 \end{bmatrix}$$

Η βιβλιογραφία σχετικά με SVMs αναφέρεται συνήθως στην Η χώρο ως ένα χώρο Hilbert. Μπορείτε να σκεφτείτε ένα χώρο Hilbert ως γενίκευση του Ευκλείδειου χώρου. Συγκεκριμένα, είναι οποιοσδήποτε γραμμικός χώρος, με εσωτερικό προϊόν που ορίζεται, το οποίο είναι επίσης πλήρες σε σχέση με την αντίστοιχη νόρμα (δηλαδή κάθε ακολουθία Cauchy από τα σημεία συγκλίνει σε ένα σημείο στο χώρο).

Είναι μια γενίκευση κυρίως επειδή το εσωτερικό γινόμενο του μπορεί να είναι οποιοδήποτε εσωτερικό προϊόν, όχι μόνο το βαθμωτό("dot").

4.1 Κατάσταση Mercer

Υπάρχει μια χαρτογράφηση Φ και μια επέκταση για ποιούς πυρήνες υπάρχει ζευγάρι $\{\mathrm{H}, \Phi\}$, με τις ιδιότητες που περιγράφονται κάτω:

$$\mathrm{K}(x, y) = \sum_i \Phi(x)_i \cdot \Phi(y)_j$$

αν και μόνο αν υπάρχει g(x) τέτοιο ώστε $\int g(x)^2 dx$ και έπειτα $\int \mathrm{K}(x, y) g(x) g(y) dx dy \geq 0$.

Σημειώστε ότι για ειδικές περιπτώσεις, μπορεί να μην είναι εύκολο να ελεγχθεί εάν η κατάσταση της Mercer ικανοποιείται. Η παραπάνω εξίσωση πρέπει να ισχύει για κάθε g με πεπερασμένη L_2 νόρμα. Ωστόσο, μπορούμε εύκολα να αποδείξουμε ότι η προϋπόθεση αυτή πληρείται για θετικές αναπόσπαστες δυνάμεις σχετικά με το προϊόν βαθμωτό dot $\mathrm{K}(x, y) = (x \cdot y)^p$. Πρέπει να αποδείξουμε ότι $\int (\sum_{i=1}^{d} x_i y_i)^p g(x) g(y) dx dy \geq 0$

Η τυπική διάρκεια επέκτασης $(\sum_{i=1}^{d} x_i y_i)^p$ συμβάλλει ένας όρος της μορφής

$$\frac{p!}{r_1! r_2! \ldots (p - r_1 - r_2 \ldots)!} \int x_1^{r_1} x_2^{r_2} \ldots y_1^{r_1} y_2^{r_2} \ldots g(x) g(y) dx dy$$

το αριστερό μέρος της εξίσωσης η οποία παραγοντοποιείται

$$\frac{p!}{r_1! r_2! \ldots (p - r_1 - r_2 \ldots)!} (\int x_1^{r_1} x_2^{r_2} \ldots g(x) dx)^2 \geq 0.$$

Μία απλή συνέπεια είναι ότι οποιοσδήποτε πυρήνας μπορεί να εκφραστεί ως

$$\mathrm{K}(x, y) = \sum_{p=0}^{\infty} c_p (x \cdot y)^p,$$

όπου το c_p είναι θετικοί πραγματικοί συντελεστές και η σειρά είναι ομοιόμορφα συγκλίνουσα, η οποία ικανοποιεί τη Κατάσταση Mercer.

Σε γενικές γραμμές, ενδέχεται να υπάρχουν στοιχεία τέτοια που η Hessian να είναι απροσδιόριστη, καθώς και για τις οποίες το τετραγωνικό πρόβλημα προγραμματισμού δεν θα έχει καμία λύση (η διττή λειτουργία μπορεί να γίνει αυθαίρετα μεγάλη).

4.2 Μερικές Σημειώσεις για την Φ και Η

Η κατάσταση Mercer μας λέει αν μια προοπτική του πυρήνα είναι πραγματικά ή όχι ένα προϊόν βαθμωτό dot σε κάποιο χώρο, αλλά δεν μας λέει πώς να κατασκευάσουμε την Φ ή ακόμα και τι είναι η Η. Ωστόσο,όπως με τον ομοιογενή (δηλαδή, ομοιογενής στο βαθμωτό dot προϊόν σε L) σε τετραγωνικό πολυωνυμικό πυρήνα που αναφέρθηκε παραπάνω, μπορούμε να κατασκευάσουμε τη χαρτογράφηση για κάποιους πυρήνες.

Η εξίσωση παρακάτω δείχνει ότι μπορούμε να επεκταθούμε σε αυθαίρετους ομοιογενείς πολυωνυμικούς πυρήνες, και ότι το αντίστοιχο διάστημα Η είναι ένας Ευκλείδειος χώρος διάστασης

$$\begin{bmatrix} d+p-1 \\ p \end{bmatrix}.$$

Έτσι, για παράδειγμα, για p = 4 πολυωνύμο, και για τα δεδομένα που αποτελείται από 16 με 16 εικόνες (d= 256), dim (H) είναι 183.181.376. Συνήθως, η χαρτογράφηση των δεδομένων σε ένα "χώρο των χαρακτηριστικών" με ένα τεράστιο αριθμό διαστάσεων, Θα μπορούσε να χαρακτηριστεί επικίνδυνη για την γενικευμένη εκτέλεση της μηχανής αποτελέσματος. Έπειτα, το σύνολο όλων των υπερσυνόλων $\{w,b\}$ είναι παραμετροποιημένο από τους dim (H) +1 αριθμούς. Τα περισσότερα συστήματα με δισεκατομμύρια, ή ακόμα και άπειρους, αριθμούς παραμέτρων δεν θα καταστήσουν δυνατή την έναρξη.

Θα μπορούσε κανείς να υποστηρίξει ότι, δεδομένης της μορφή της λύσης, υπάρχουν το πολύ l+ 1 ρυθμιζόμενες παραμέτρους (όπου l είναι ο αριθμός των δειγμάτων).

Υπάρχει απαίτηση μας σχετικά με το μέγιστο περιθώριο υπερσυνόλου. Από την χαρτογράφηση της επιφάνειας είναι εγγενής διάσταση dim (L), εκτός αν dim (L) =dim (H). Η χαρτογράφηση δεν χρειάζεται να είναι ένα προς ένα π.χ. $x_1 \to -x_1, x_2 \to -x_2$. Η εικόνα του Φ δεν πρέπει να είναι διανυσματικός χώρος: και πάλι, λαμβάνοντας υπόψη το παραπάνω απλό τετραγωνικό παράδειγμα, το διάνυσμα –Φ(x) δεν είναι κατ 'εικόνα του Φ εκτός αν x=0. Επιπλέον, σε σχέση με τον ανομοιογενή πυρήνα

$$\mathrm{K}(x_i, x_j) = (x_i \cdot x_j + 1)^2$$

θα σας πείσει ότι η αντίστοιχη Φ μπορεί να χαρτογραφήσει δύο διανύσματα που είναι γραμμικά εξαρτημένα με το L σε δύο διανύσματα που είναι γραμμικά ανεξάρτητα στο Η. Κάποιος μπορεί να αρχίσει με την Φ και στη συνέχεια να προχωρήσει στην κατασκευή του αντίστοιχου πυρήνα. Για παράδειγμα, αν L = R^1, τότε η επέκταση Fourier στα δεδομένα x, σταματάει μετά από N όρους, και έχει τη μορφή

$$f(x) = \frac{a_0}{2} + \sum_{r=1}^{N} (a_{1r} \cos(rx) + a_{2r} \sin(rx))$$

Και αυτό μπορεί να γίνει αντιληπτό σαν dot βαθμωτό προϊόν μεταξύ δύο διανυσμάτων στην $R^{2N+1} : a = (\frac{a_0}{\sqrt{2}}, a_{11}, ..., a_{21}, ...)$ με χαρτογράφηση

$$\Phi(x) = (\frac{1}{\sqrt{2}}, \cos(x), \cos(2x), ..., \sin(x), \sin(2x), ...) .$$

Έπειτα ο ανταποκρινόμενος πυρήνας (Dirichlet) Μπορεί να γραφτεί ως

$$\Phi(x_i) \cdot \Phi(x_j) = \mathrm{K}(x_i, x_j) = \frac{\sin((N+1/2)(x_i - x_j))}{2\sin((x_i - x_j)/2)}$$

Αυτό φαίνεται εύκολα αν $\delta \equiv x_i - x_j$.

$$\Phi(x_i) \cdot \Phi(x_j) = \frac{1}{2} + \sum_{r=1}^{\mathrm{N}} \cos(rx_i)\cos(rx_j) + \sin(rx_i)\sin(rx_j)$$
$$= -\frac{1}{2} + \sum_{r=0}^{\mathrm{N}} \cos(r\delta) = -\frac{1}{2} + \mathrm{Re}\left\{ \sum_{r=0}^{N} e^{(ir\delta)} \right\}$$
$$= -\frac{1}{2} + \mathrm{Re}\left\{ (1 - e^{i(\mathrm{N}+1)\delta)}) / (1 - e^{(i\delta)}) \right\}$$
$$= (\sin((N+1/2)\delta)) / 2\sin(\delta/2)$$

Τέλος, είναι σαφές ότι το ανώτερο τέχνασμα της χαρτογράφησης θα λειτουργήσει για οποιοδήποτε αλγόριθμο σε οποιαδήποτε δεδομένα εμφανίζονται μόνο ως dot βαθμωτά προϊόντα. Το γεγονός αυτό έχει χρησιμοποιηθεί για τον υπολογισμό μη γραμμικής έκδοσης του PCA (principal component analysis).

4.2.1 Μερικά παραδείγματα των Μη Γραμμικών SVMs

Οι πρώτοι πυρήνες για το πρόβλημα αναγνώρισης προτύπων είναι οι εξής:

$$\mathrm{K}(x, y) = (x \cdot y + 1)^p$$

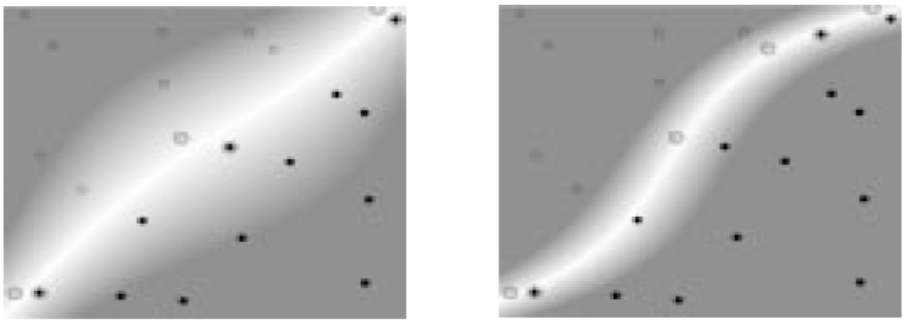
$$\mathrm{K}(x, y) = e^{-\|x-y\|^2/2\sigma^2}$$

$$\mathrm{K}(x, y) = \tanh(kx \cdot y - \delta)$$

Η εξίσωση οδηγεί σε μια ταξινόμηση η οποία είναι ένα πολυώνυμο βαθμού p. Δίνεται μια Gaussian συνάρτηση ταξινόμησης, και ένα ιδιαίτερο είδος διπλής στοιβάδας νευρωνικό

δίκτυο. Για την περίπτωση RBF, ο αριθμός των κέντρων N_s, τα κέντρα τους (η s_i), τα βάρη (a_i), και το όριο (b), όλα αυτά είναι παραγόμενα αυτόματα από την SVM και δίνουν άριστα αποτελέσματα σε σύγκριση με την κλασική RBFs, για την περίπτωση του Gaussian RBFs.

Για την περίπτωση των νευρωνικών δικτύων, η πρώτη στοιβάδα αποτελείται από N_s σύνολα από βάρη, κάθε σύνολο αποτελείται από d_L (η διάσταση των δεδομένων) βάρη, και η δεύτερη στοιβάδα αποτελείται από N_s βάρη (τα a_i), έτσι ώστε η αξιολόγηση απλώς απαιτεί σταθμισμένο άθροισμα, αξιολογείται από dot βαθμωτά προϊόντα των δεδομένων από τα διανύσματα υποστήριξης (SV). Έτσι, για το νευρωνικό δίκτυο, η αρχιτεκτονική (αριθμός των βαρών) καθορίζεται από την κατάρτιση της SVM.



Εικόνα 17

Σημειώστε, ωστόσο, ότι το υπερβολικά εφαπτόμενο κέντρο ικανοποιεί μόνο τον όρο Mercer για ορισμένες τιμές των παραμέτρων κ και δ (και των δεδομένων $\|x\|^2$).

Αυτή ήταν η πρώτη παρατήρηση πειραματικά, ωστόσο ορισμένες αναγκαίες προϋποθέσεις για αυτές τις παραμέτρους και για τη θετικότητα είναι επίσης γνωστές. Η Εικόνα 17 δείχνει τα αποτελέσματα για το ίδιο πρόβλημα αναγνώρισης προτύπων, όπου το κέντρο (πυρήνας) επιλέχθηκε να είναι ένα κυβικό πολυώνυμο. Παρατηρήστε ότι, παρόλο που ο αριθμός των βαθμών ελευθερίας είναι υψηλότερος, για τη γραμμικά διαχωρίσιμη περίπτωση (αριστερή πλευρά), η λύση είναι περίπου γραμμική, που δείχνει ότι η πληρότητα ελέγχεται, ηγραμμικά μη διαχωρίσιμη περίπτωση (δεξιά πλευρά) έχει διαχωριστ εί. Τέλος, σημειώστε ότι αν και οι ταξινομητές SVM που περιγράφονται παραπάνω είναι δυαδικοί ταξινομητές ,εύκολα συνδυάζονται για τη διαχείριση μιας πολυταξικής υπόθεσης. Ένας απλός, αποτελεσματικός συνδυασμός με N να είναι ένας-εναντίον-υπόλοιπων ταξινομητών (ας πούμε, "ένα" θετικό "υπόλοιπο" αρνητικό) για την περίπτωση N-τάξης παίρνει την τάξη για ένα σημείο δοκιμής που αντιστοιχεί στην μεγαλύτερη θετική απόσταση.

4.3. "Παγκόσμιες" λύσεις και Μοναδικότητα

Με το όρο "παγκόσμιο", εννοούμε ότι δεν υπάρχει άλλο σημείο της εφικτής περιοχής στο οποίο η αντικειμενική συνάρτηση να παίρνει μια χαμηλότερη τιμή. Θα αντιμετωπίσουμε δύο τρόπους για τους οποίους η μοναδικότητα δεν μπορεί να υπάρξει:

- λύσεις για τις οποίες τα $\{w, b\}$ είναι μοναδικές,
- αλλά για τα οποία η επέκταση του w δεν είναι,

  και οι λύσεις των οποίων διαφέρουν.

Και οι δύο έχουν ενδιαφέρον ακόμα και αν το ζευγάρι $\{w, b\}$ είναι μοναδικό, αν τα a_i δεν είναι, μπορεί να υπάρχουν ισοδύναμες επεκτάσεις του w που απαιτούν λιγότερα διανύσματα υποστήριξης (ένα ασήμαντο παράδειγμα αυτού δίνεται παρακάτω), και τα οποία ως εκ τούτου απαιτούν λιγότερες οδηγίες κατά τη διάρκεια της δοκιμαστικής φάσης. Αποδεικνύεται ότι κάθε τοπική λύση είναι επίσης παγκόσμια. Αυτό είναι μια ιδιότητα για κάθε κυρτό πρόβλημα προγραμματισμού. Επιπλέον, η μοναδικότητα της λύσης είναι εγγυημένη για την περίπτωση που η αντικειμενική λειτουργία είναι αυστηρά κυρτή, η οποία στην περίπτωσή μας σημαίνει ότι η Hessian και πρέπει να είναι θετικά ορισμένη. Ωστόσο, ακόμη και αν η Hessian είναι θετικά ημιορισμένη, η λύση μπορεί ακόμα να είναι μοναδική, σκεφτείτε δύο σημεία κατά μήκος της πραγματικής ευθείας με συντεταγμένες x_1= 1 και x_2 = 2, και με τη σωστή πολικότητα + και -.

Εδώ η Hessian είναι θετικά ημιορισμένη, αλλά η λύση ($w = -2, b = 3, \xi_i = 0$) είναι μοναδική. Είναι επίσης εύκολο να βρούμε λύσεις οι οποίες δεν είναι μοναδικές, με την έννοια ότι η a_i στην επέκταση του w δεν είναι μοναδική: Για παράδειγμα, το πρόβλημα τεσσάρων σημείων που διαχωρίζονται σε ένα τετράγωνο στον

R^2 : $x_1 = [1,1]$, $x_2 = [-1,1]$, $x_3 = [-1,-1]$, $x_4 = [1,-1]$,
με πολικότητες $[+,-,-,+]$ αντίστοιχα.

Μια λύση είναι, $w = [1,0], b = 0, a = [0.25, 0.25, 0.25, 0.25]$. Άλλο έχει την ίδια w και b, αλλά $a = [0.5, 0.5, 0, 0]$ (και οι δύο λύσεις ικανοποιούν τους περιορισμούς $a_i > 0$ και $\sum_i a_i y_i = 0$). Λαμβάνοντας υπόψη κάποια λύση α, επιλέξαμε ένα α΄ το οποίο ανήκει στο μηδενικό χώρο της Hessian $H_{ij} = y_i y_j x_i \cdot x_j$, και απαιτεί ο α΄ να είναι ορθογώνιος, το διάνυσμα του οποίου έχει όλες τις συνιστώσες να είναι 1. Στη συνέχεια, προσθέτοντας τον α΄ στον α θα αφήσει την L_D αμετάβλητη. Αν $0 \le a_i + a'_i \le C$ τότε α΄ + α είναι επίσης λύση.

Το παρακάτω πολύ απλό θεώρημα δείχνει ότι αν μη μοναδικές λύσεις προκύψουν, τότε η λύση σε ένα βέλτιστο σημείο είναι συνεχώς διαμορφωμένη, με τέτοιο τρόπο ώστε όλα τα ενδιάμεσα σημεία είναι επίσης, λύσεις.

Θεώρημα 6.

Υπάρχει μεταβλητή X για το ζεύγος των μεταβλητών $\{w,b\}$. Ας λάβουμε τη Hessian για το πρόβλημα να είναι θετικά ημιορισμένη, έτσι ώστε η αντικειμενική συνάρτηση να είναι κυρτή. Έστω X_0 και η X_1 είναι δύο σημεία στα οποία η συνάρτηση επιτυγχάνει την ελάχιστη τιμή της. Τότε υπάρχει υπάρχει $X = X(\tau) = (1-\tau)\mathrm{X}_0 + \tau\mathrm{X}_1, \tau \in [0,1]$, έτσι ώστε $X(\tau)$ είναι μια λύση για όλα τα τ.

Αν και απλό, αυτό το θεώρημα είναι αρκετά κατατοπιστικό. Για παράδειγμα, θα πίστευε κανείς ότι το πρόβλημα που απεικονίζεται στην Εικόνα 18 έχει πολλές διαφορετικές βέλτιστες λύσεις (για την υπόθεση SVM's). Ωστόσο, δεδομένου ότι δεν μπορεί κανείς να κινήσει ομαλά το υπερεπίπεδο από την προτεινόμενη λύση σε ένα άλλο χωρίς τη δημιουργία υπερσυνόλου που δεν αποτελεί λύση. Γνωρίζουμε ότι αυτές οι προτεινόμενες λύσεις δεν είναι στην πραγματικότητα λύσεις σε όλα. Στην πραγματικότητα, για κάθε μία από αυτές τις περιπτώσεις, η βέλτιστη λύση είναι μοναδική στο w = 0, με κατάλληλη επιλογή του b (η οποία έχει ως αποτέλεσμα την ανάθεση της ίδιας ετικέτας σε όλες τις μονάδες). Σημειώστε ότι αυτό είναι μια τέλεια

Εικόνα 18. Προβλήματα με μη μοναδικές λύσεις

αποδεκτή λύση στο πρόβλημα ταξινόμησης: κάθε προτεινόμενο υπερεπίπεδο (με $w \neq 0$) θα προκαλέσει την πρωταρχική λειτουργία να λάβει υψηλότερη τιμή.

Τέλος, σημειώστε το γεγονός ότι η SVM βρίσκει πάντα μια παγκόσμια λύση σε αντίθεση με την περίπτωση των νευρωνικών δικτύων, όπου υπάρχουν πολλά τοπικά ελάχιστα συνήθως.

5. Βελτιωμένη απόδοση Δεδομένων

Οι κατατάξεις συνεχίζουν να ενδιαφέρουν τους ερευνητές και κάθε χρόνο αρκετοί νέοι αλγόριθμοι προτείνονται προκειμένου να βελτιωθεί η ακρίβεια. Ενώ η πλειοψηφία των αλγορίθμων μάθησης έχουν καλές επιδόσεις, μειώνονται σε καταστάσεις όπως:

- δεδομένα υψηλά σε περιεκτικότητα σε θόρυβο,
- μικρά μεγέθη δείγματος σε σχέση με ορισμένα χαρακτηριστικά,
- άσχετες ή περιττές πληροφορίες
- και μη γραμμικές.

Ο Markovitch προσδιορίζει ως άσχετα, το θόρυβο και τις περιττές πληροφορίες ως επιζήμια στοιχεία που οδηγούν σε ανακρίβειες στην πρόβλεψη.

Η αποτελεσματικότητα της μάθησης αλγορίθμων μειώνεται σε τομείς με διαφορετικές και περιττές λειτουργίες. Επιπλέον, δεδομένου ότι ο αριθμός των στοιχείων που χρησιμοποιούνται για την ταξινόμηση εργασίας αυξάνεται, ο αριθμός των δειγμάτων εκπαίδευσης που απαιτείται για τη στατιστική προσαρμογή του μοντέλου ή και επίβλεψη των συστημάτων μάθησης αυξάνεται εκθετικά, μια κατάσταση άκρως ανεπιθύμητη σε συνθήκες χαμηλού μεγέθους του δείγματος.
Βελτιωμένη απόδοση μπορεί να επιτευχθεί μέσω της παράκαμψης, όπως των θορυβώδη, άνευ σημασίας και περιττών πληροφοριών . Προτείνεται η χρήση προεπεξεργασιών μεγάλου μήκους, όπως ***εξαγωγή χαρακτηριστικών,*** κατασκευή και λειτουργία, επιλογή χαρακτηριστικών.

Συστήματα εξαγωγής χαρακτηριστικών (όπως το Principal Component Analysis, Γραμμική διακριτική ανάλυση, Τοπικά γραμμική ενσωμάτωση, Isomap, Πολυδιάστατη κλιμάκωση, κ.λπ.) έχουν σαν αποτέλεσμα τη γραμμική / μη γραμμική μετατροπή των δεδομένων και μετατρέπουν σε χαμηλότερο τρισδιάστατο χώρο με τέτοιο τρόπο, ώστε οι περισσότερες από τις πληροφορίες διατηρούνται, ενώ απορρίπτονται τα θορυβώδη δεδομένα.

Η κατασκευή επιχειρεί να απλοποιήσει την υπόθεση αναζήτησης με την προσθήκη νεότερων στοιχείων με πρόσθετες πληροφορίες.

Αυτές οι δύο προσεγγίσεις προσπαθούν να λύσουν το πρόβλημα από διαφορετικές πληροφορίες στο χώρο των χαρακτηριστικών με την αλλαγή της παρουσίασης.

Η επιλογή χαρακτηριστικού είναι μια ειδική περίπτωση εξαγωγής που αφορά την επιλογή ενός υποσυνόλου χαρακτηριστικών που περιγράφουν την υπόθεση, τουλάχιστον όσο και το πρωτότυπο. Η εξαγωγή χαρακτηριστικών κάνει N μετρήσεις για να αποκτήσει M διαστάσεων δεδομένα (N>> M). Η επιλογή χαρακτηριστικών, από την άλλη πλευρά, οι

απορρίψεις (N-M) διαφορετικών χαρακτηριστικών, απαιτούν τη συλλογή μόνο σχετικών ιδιοτήτων μειώνοντας το κόστος των δεδομένων.

Τα οφέλη της επιλογής χαρακτηριστικών περιλαμβάνουν τη μείωση του όγκου των δεδομένων που απαιτούνται για την επίτευξη μάθησης, βελτιωμένη προγνωστική ακρίβεια, πιο συμπαγής και εύκολα κατανοητή και μείωση του χρόνου εκτέλεσης. Οι τελευταίοι δύο παράγοντες έχουν ιδιαίτερη σημασία στον τομέα της εμπορικής και βιομηχανικής εξόρυξης δεδομένων, καθιστώντας την πιο επιθυμητή. Είναι επιθυμητό το γενικό σύστημα να πρέπει επίσης να είναι σε θέση να χειρίζεται θορυβώδη χαρακτηριστικά.

5.1Wrappers-filters

Η βιβλιογραφία σχετικά με τις μεθόδους επιλογής χαρακτηριστικών και εφαρμογές είναι ευρέως διαδεδομένη σε πολλούς τομείς, συμπεριλαμβανομένων αυτών της ταξινόμησης εγγράφων, εξόρυξης δεδομένων, αναγνώρισης αντικειμένων, βιομετρικών στοιχείων, τηλεπισκόπησης και υπολογιστικής όρασης. Είναι σχετικό με οποιαδήποτε εργασία όπου ο αριθμός των χαρακτηριστικών είναι μεγαλύτερος από τον αριθμό των δειγμάτων εκπαίδευσης, ή πολύ μεγάλος για να είναι υπολογιστικά εφικτός. Οι υφιστάμενες μέθοδοι επιλογής χαρακτηριστικών για τη μηχανική μάθηση συνήθως εμπίπτουν σε δύο ευρείες κατηγορίες:

- *περιτυλίγματα* **(wrappers)**
- και *φίλτρα* ***(filters)***.

Wrappers

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

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

- **Filters**

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

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

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

5.2. Προτεινόμενο σύστημα-Συμβολισμός

Περιγράφεται ένα αποτελεσματικό και ισχυρό σχήμα που απορρίπτει τις θορυβώδεις, άσχετες και περιττές πληροφορίες στα παρούσα δεδομένα, ενώ διατηρεί την ξεχωριστεί δύναμη των στοιχείων. Ο συνδυασμός των filters και wrappers προσεγγίσεων προτείνεται για να πάρει τη βέλτιστη ακρίβεια, αποτελεσματικότητα και την καλύτερη γενίκευση.

Εδώ το filter παρέχει ένα πολύ καλό αρχικό υποσύνολο χαρακτηριστικών για ένα wrapper-μια διαδικασία που είναι πιθανόν να οδηγήσει σε μια μικρότερη, και ως εκ τούτου ταχύτερη αναζήτηση για το wrapper. Η προτεινόμενη μέθοδος που εφαρμόζεται είναι αυτή του συμβολισμού των δεδομένων για την επίλυση του διπλού προβλήματος του φιλτραρίσματος και μείωσης του θορύβου.

Ο συμβολισμός δεδομένων περιλαμβάνει διαφοροποίηση των χαρακτηριστικών στοιχείων σε ένα περιορισμένο σύνολο τιμών που ονομάζονται σύμβολα, τα οποία διατηρούν ντετερμινιστικά χαρακτηριστικά, ενώ μειώνουν το θόρυβο μέτρησης. Επιπλέον, η υπό όρους εντροπία της

ετικέτας της κλάσης σε σχέση με το λειτουργικό χαρακτηριστικό (μετατρέπεται σε συμβολική μορφή) υπολογίζεται για να καθορίσει αν η λειτουργία σχετίζεται με την κατηγορία ή όχι. Εδώ χαμηλά είναι υπό όρους εντροπία, υψηλότερα είναι η σύζευξη. Κατά τον ίδιο τρόπο μπορούμε να βρούμε το βαθμό σύζευξης ενός χαρακτηριστικού σε άλλα χαρακτηριστικά. Ποσότητες όπως ο συντελεστής συσχέτισης ή η συνάρτηση συσχετισμού συχνά δεν παρέχουν σαφείς ενδείξεις του συνδετικού χαρακτηριστικού των μεταβλητών και πληροφορίες ανά κατηγορία (δεδομένου ότι αυτά μπορεί να είναι αραιά και θορυβώδη). Τα συστήματα συμβολισμού, από την άλλη πλευρά, λειτουργούν ακόμη και παρουσία εξωτερικού θορύβου. Η μέθοδος του συμβολισμού δεδομένων μπορεί να εφαρμοστεί σε ντετερμινιστικό ή στοχαστικό, σε γραμμικά ή μη-γραμμικά συστήματα, χωρίς καμία εκ των προτέρων υπόθεση για τη φύση των δυναμικών διαδικασιών και έχει ένα πρακτικό πλεονέκτημα της απλοποίησης και της επιτάχυνσης στους μετέπειτα υπολογισμούς καθώς ο τύπος των δεδομένων άλλαξε από το συνεχή σε διακριτή μορφή. Παρακάτω περιγράφεται η σύνοψη του προτεινόμενου συστήματος.

1. Μετατροπή των δεδομένων σε συμβολική μορφή.

2. Υπολογισμός της υπό όρους εντροπίας πληροφοριών κατηγορίας σε σχέση με όλα τα χαρακτηριστικά ένα προς ένα.

Εδώ η υπό όρους εντροπία χρησιμοποιείται ως εκ τούτου σαν filter. Για αυτό το λόγο διατηρούμε τη συνάφεια στις τιμές για να διαιρείται το χαρακτηριστικό σε σχετικά και άσχετα χαρακτηριστικά. Αυτό μπορεί να γίνει είτε με τη τιμή υπό όρους εντροπίας απευθείας ή μέσω επιλέγοντας τις χαμηλότερες τιμές n και απορρίπτοντας τα υπόλοιπα χαρακτηριστικά.

3. Υπολογισμός της υπό όρους εντροπία του χαρακτηριστικού (με υψηλότερη σύζευξη πληροφοριών κατηγορίας) σε σχέση με όλα τα υπόλοιπα χαρακτηριστικά ένα προς ένα.

Χαρακτηριστικά που παρουσιάζουν υψηλή συσχέτιση (χαμηλές τιμές υπό όρους εντροπίας) απορρίπτονται. Και εδώ χρησιμοποιούμε είτε απευθείας κατώτατο όριο ή επιλέγουμε τις υψηλότερες τιμές n και απορρίπτουμε τις υπόλοιπες λειτουργίες.

4.Τακτοποίηση ώστε να διαθέτει σε αύξουσα σειρά τα δεδομένα εντροπίας με πληροφορίες κατηγορίας.

Πάρτε δεδομένα με τη δεύτερη χαμηλότερη εντροπία και υπολογίστε όρους της εντροπίας σε σχέση με τα χαρακτηριστικά, με όρους εντροπίας υψηλότερους από αυτό. Τα χαρακτηριστικά επιλέγονται με παρόμοιο τρόπο όπως στο Βήμα 3.

5. Επαναλάβετε το Βήμα 4 μέχρι και την τελευταία δυνατότητα.

6. Οι λειτουργίες που λαμβάνονται από το Βήμα 5 χρησιμοποιούνται ως πρώτη ύλη για το σύστημα wrapper χρησιμοποιώντας SVMs σαν αλγόριθμο μάθησης.

Οι SVMs βασίζονται στην αυστηρή στατιστική θεωρία μάθησης η οποία έχει πολλές επιθυμητές ιδιότητες, όπως τη μη γραμμική μάθηση, βελτίωση της γενικευμένης απόδοσης κλπ.. Εδώ η SVM υποστηρίζεται από το γενετικό αλγόριθμο και τον αλγόριθμο Newton για τη βέλτιστη ρύθμιση των παραμέτρων.

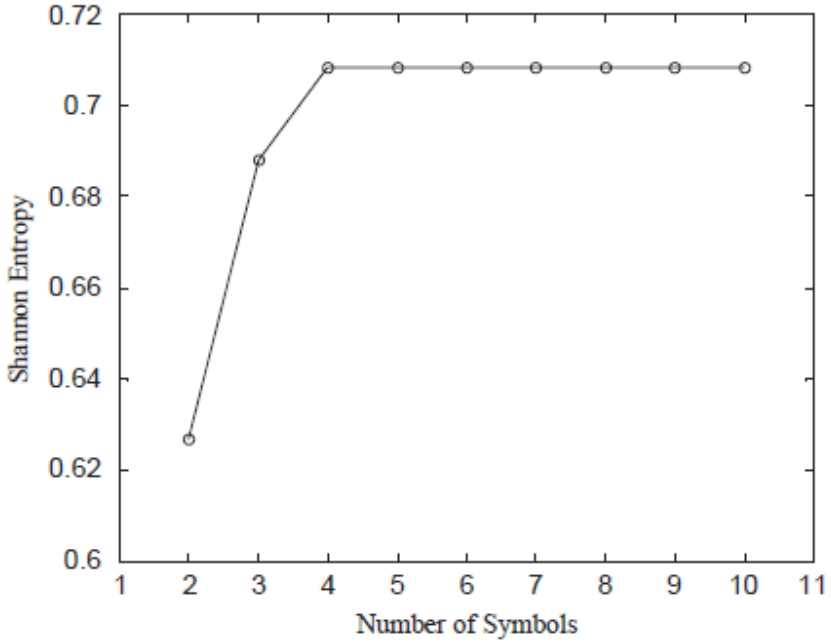
5.3. Συμβολισμός

Ο συμβολισμός προϋποθέτει διασκορπισμό (scatter). Συνήθως το εύρος για κάθε πρωτότυπο χαρακτηριστικό (ή το φάσμα κάποιων μετατρέψιμων αρχικών δεδομένων, όπως οι πρώτες διαφορές μεταξύ των διαδοχικών τιμών) είναι χωρισμένο σε έναν πεπερασμένο αριθμό διακριτών κελιών και γίνεται ανάθεση διαφορετικών σύμβολων σε κάθε κελί. Κάθε αρχική τιμή ενός χαρακτηριστικού είναι μοναδικά αντίστοιχη σε ένα συγκεκριμένο σύμβολο ανάλογα με τον τομέα στον οποίο η μέτρηση πέφτει. Έτσι

$$S_i = \begin{cases} 1 x_{\min} < x_i < x_{c_1} \\ 2 x_{c_1} < x_i < x_{c_2} \\ 3 x_{c_2} < x_i < x_{c_3} < x_{\max} \\ . \\ . \\ ...n\sigma\acute{\upsilon}\mu\beta o\lambda\alpha \end{cases}$$

Εδώ τα $x_{c_1}, x_{c_2}, x_{c_3}$ είναι κρίσιμα σημεία (σύμβολα), καθορίζοντας τα όρια των κελιών, 1, 2, 3,. . . , N. Ο αριθμός των συμβόλων που χρησιμοποιούνται, n, αναφέρεται ως το σύμβολο-set. Στην απλούστερη (δυαδική) περίπτωση n = 2.

Ο αριθμός των συμβόλων καθορίζει πόσο οι πρωτότυπες πληροφορίες διατηρούνται. Μία υψηλότερη τιμή του n λαμβάνει υπόψη περισσότερες λεπτομέρειες των αρχικών δεδομένων, μαζί με τις επιπτώσεις του θορύβου των μετρήσεων που θα μπορούσαν να είναι παρούσες. Για παράδειγμα, όταν το n ισούται με τον αριθμό των διακριτών τιμών στα δεδομένα, τα δεδομένα σύμβολα και τα αρχικά



Εικόνα 19. Shannon εντροπία vs αριθμό των συμβόλων για ίσα διαστήματα αποθήκευσης

στοιχεία είναι ισοδύναμα με την έννοια ότι περιέχουν την ίδια πληροφορία (δηλαδή, δεν υπάρχει απώλεια πληροφοριών λόγω συμβολισμού). Επίσης, για οποιοδήποτε μέγεθος συνόλου συμβόλων, η τοποθέτηση των κρίσιμων σημείων επηρεάζει τα χαρακτηριστικά της συμβολικής περιγραφής των δεδομένων. Ο διαχωρισμός των δεδομένων πρέπει να γίνει προσεκτικά, αφού η κακή επιλογή των σημείων διαχωρισμού μπορεί να οδηγήσει σε απώλεια σημαντικών πληροφοριών. Έτσι, ακόμη και αν ο συμβολισμός ελαχιστοποιεί τις επιπτώσεις του θορύβου στα δεδομένα, προκαλεί επίσης την απώλεια σημαντικών πληροφοριών κατά τη διάρκεια της διαδικασίας. Είναι απαραίτητο η απώλεια πληροφοριών κατά τη διάρκεια της διαδικασίας να είναι στο ελάχιστο. Εφαρμόζουμε αυτή την ανταλλαγή (μεταξύ της μείωσης του θορύβου και της απώλειας σημαντικών πληροφοριών) με τον διαχωρισμό των δεδομένων σε συνδυασμό με τις εντροπίες πληροφορίες ανάλυσης , όπως περιγράφεται στην συνέχεια. Για την ανάλυση, η σειρά συμβόλων μετατρέπεται σε συμβολική ακολουθία με τον καθορισμό ενός πεπερασμένου μήκους (L) που μπορούν να μετακινούνται κατά μήκος , συμβολοσειρές με ένα βήμα κάθε φορά, κάθε βήμα αποκαλύπτοντας νέα σειρά. Για διευκόλυνση της αναφοράς και αναγνώρισης κάθε σύντομη ακολουθία μοναδικά συμβολίζεται με ένα μόνο ακέραιο

$$l = \sum_{i=1}^{L} M^{L-iS_i}$$

όπου M είναι ο αριθμός των διαφορετικών συμβόλων και L είναι το μήκος της συμβολικής ακολουθίας. Αυτή η σειρά ακολουθία συμβόλων (ή κωδικοποιημένη σειρά) μπορεί να χαρακτηριστεί με πληροφορίες θεωρητικών μέτρων, όπως η Shannon εντροπία ορίζεται ως

$$E = -\frac{1}{L}\sum_{l} P_l \ln P_l$$

όπου P_l είναι η πιθανότητα να βρει μια συγκεκριμένη ακολουθία l. Ορίζεται ως ο αριθμός των φορών αυτής της ακολουθίας που μπορεί να βρεθεί σε συμβολοσειρά και διαιρείται με τον αριθμό των σύντομων ακολουθιών. Η Shannon εντροπία είναι ένα μέτρο για να μετρήσει το περιεχόμενο των πληροφοριών στη συμβολοσειρά. Ο βέλτιστος αριθμός των συμβόλων που θα πρέπει να χρησιμοποιούνται για τη μεγιστοποίηση του περιεχομένου των πληροφοριών και την ελαχιστοποίηση του αποτελέσματος του θορύβου μπορεί να επιτευχθεί με τη μεγιστοποίηση της εντροπίας Ε, με σεβασμό

(α) στον αριθμό των κρίσιμων σημείων και

(β) στις τοποθετήσεις αυτών των κρίσιμων σημείων.

Η όλη διαδικασία του συμβολισμού παρουσιάζεται παρακάτω για τα δεδομένα με ένα μόνο χαρακτηριστικό:

0,4966	0,8998	0,8216	0,6449
0,8180	0,6602	0,3420	0,2897
0,3412	0,5341	0,7271	0,3093
0,8385	0,5681	0,3704	0,7027
0,5466	0,4449	0,6946	0,6213

Η διαδικασία παρουσιάζεται για τον αριθμό των κελιών ίσο με n = 3 (ίσου μεγέθους), και για μήκος ακολουθίας L = 4. Τα παραπάνω δεδομένα μπορούν να μετατραπούν σε σύμβολα δεδομένων για τρία ξεχωριστά κελιά ως

2 3 3 2 3 2 1 1 1 2 3 1 3 2 1

3 2 1 2 2

Τώρα για L = 4, οι ακολουθίες:

2 3 3 2; 3 3 2 3;. . . ; 2 1 2 2

Ο αντίστοιχος κωδικός για την πρώτη ακολουθία που λαμβάνεται:

(3 ∧ 3)* 2 + (3 ∧ 2)* 3 + (3 ∧ 1)* 3 + (3 ∧ 0)* 2 = 92

Με την επανάληψη αυτού του βήματος κωδικοποίησης για όλες τις υπόλοιπες ακολουθίες, παίρνουμε τις παρακάτω κωδικοποιημένες σειρές:

92 117 110 88 103 67 41 45 55

87 101 61 105 74 61 104 71

Όλη αυτή η διαδικασία επαναλαμβάνεται για n = (2, 3, 4,..., 10). Η συνισταμένη εντροπία υπολογίζεται χρησιμοποιώντας την εξίσωση, δείχνει ότι μία μέγιστη εντροπία (0,7083) έχει ληφθεί για τον αριθμό των σύμβολων n = 4. Έτσι, τα συμβολισμένα δεδομένα για n=4 θα είναι βέλτιστα. Σε περίπτωση πολλαπλών δυνατοτήτων διαχείρισης δεδομένων, πρέπει να ακολουθήσουμε την ίδια διαδικασία για κάθε χαρακτηριστικό ανεξάρτητα από άλλα χαρακτηριστικά. Τα δεδομένα των δοκιμών (ή σύνδεση δεδομένων) συμβολίζονται με ίδια όρια κελιών που χρησιμοποιούνται για την εκπαίδευση. Σε περίπτωση που υπερβαίνει τα δεδομένα δοκιμών η περιοχή που καλύπτεται από δεδομένα εκπαίδευσης (και στις δύο πλευρές), τότε λαμβάνει το υψηλότερο και το χαμηλότερο σύμβολο σύμφωνα με τα όρια.

5.3.1. Υπό όρους εντροπίας

Για δύο σήματα {X} και {Z}, η υπό όρους εντροπία ορίζεται ως

$$E\left(\frac{Z}{X}\right) = -\frac{1}{Nx}\sum_{Sx}\sum_{Sz} P\left(\frac{Sz}{Sx}\right)\ln P\left(\frac{Sz}{Sx}\right) E\left(\frac{Z}{X}\right) = -\frac{1}{Nx}\sum_{Sx}\sum_{Sz} P\left(\frac{Sz}{Sx}\right)\ln P\left(\frac{Sz}{Sx}\right)$$
$$E\left(\frac{Z}{X}\right) = -\frac{1}{Nx}\sum_{Sx}\sum_{Sz} P\left(\frac{Sz}{Sx}\right)\ln P\left(\frac{Sz}{Sx}\right)$$

όπου Nx είναι ο συνολικός αριθμός των διαφορετικών Sx τιμών που παρατηρήθηκαν και $P\left(\frac{Sz}{Sx}\right)$ είναι η πιθανότητα για τη μεταβλητή Z να λάβει συμβολική αξία του Sz όταν η μεταβλητή X καταλαμβάνει τη συμβολική αξία του Sx.

Κάποιος μπορεί να ερμηνεύσει την υπό όρους εντροπία ως το ποσό που έχει παρατηρηθεί ότι παραμένει αβέβαιο για τον Z μετά τον X. Όσο χαμηλότερα από την υπό όρους εντροπία μεγαλύτερη είναι η συσχέτιση μεταξύ δύο μεταβλητών.

5.4. SVM ταξινόμηση

Οι SVMs με βάση τις αρχές της στατιστικής θεωρίας χρησιμοποιούνται συνήθως για διάφορα διμερή και πολύ καταμερισμένα καθήκοντα κατάταξης σε διάφορους τομείς. Οι υπολογιστικοί βιολόγοι απασχολούνται επίσης με SVM για την εκτέλεση σημαντικών εργασιών όπως οι διαρθρωτικές ταξινομήσεις των πρωτεϊνών.

Η μεθοδολογία, οι αλγόριθμοι και το λογισμικό είναι πλέον εύκολα διαθέσιμα. Έχουμε μια πολύ σύντομη παράθεση του δυαδικού αλγορίθμου ταξινόμησης SVM σε αυτή την ενότητα. Ξεκινώντας με ένα σύνολο των εισροών-εκροών στα ζεύγη εκπαίδευσης

$$(x_1, y_1), (x_2, y_2), ..., (x_N, y_N) x \in R^d, y \in R$$

Η λειτουργία απόφασης της SVM όσον αφορά μία κατάλληλα καθορισμένη λειτουργία του πυρήνα μπορεί να ληφθεί ως:

$$f(x) = \sum_{i=1}^{N} y_i a_i K_\theta(x_i, x) + b,$$

όπου N είναι το μέγεθος του δείγματος και το $K_\theta(x_i, x)$ είναι ο πυρήνας λειτουργίας της χαρτογράφησης των διανυσμάτων εισόδου σε ένα χώρο χαρακτηριστικών και θ ένα σύνολο παραμέτρων και b είναι η σταθερά προκατάληψης. Οι συντελεστές a_i είναι που λαμβάνεται με την επίλυση του τετραγωνικής βελτιστοποίησης προβλήματος

$$w(\alpha) = \sum_{i=1}^{N} a_i - \left(\frac{1}{2}\right) \sum_{i=1}^{N} \sum_{j=1}^{N} a_i a_j y_i y_j K_\theta(x_i, x_j)$$

με την επιφύλαξη των περιορισμών

$$0 \le a_i, i = 1, ..., N$$

$$\sum_{i=1}^{N} a_i y_i = 0.$$

Αν στο παραπάνω αυστηρό όριο της SVM του προβλήματος βελτιστοποίησης (δεν υπάρχει ρητή διάταξη για την επιβολή κυρώσεων για λάθη κατάρτισης) η ισότητα ικανοποιείται για τα σημεία x_i με το αντίστοιχο $a_i > 0$, τότε αυτά τα σημεία ονομάζονται μη μηδενικά SV. Αν το υπερεπίπεδο που διαχωρίζει επιτρέπεται να διέρχεται από την αρχή με τη λήψη b = 0, τότε

ο περιορισμός της ισότητας εξαφανίζεται και η διατύπωση του προβλήματος, ονομάζεται αυστηρό περιθώριο της SVM χωρίς όριο.

Σε περίπτωση μη διαχωρίσιμων προτύπων εκπαίδευσης, τα λάθη κατάρτισης επιτρέπονται και η διατύπωση του προβλήματος στην περίπτωση αυτή καλείται ως χαλαρό SVM περιθώριο.

Ο περιορισμός των ανισοτήτων είναι ελαφρώς τροποποιημένος ως $0 \le a_i \le C$, όπου C θεωρείται σταθερή για τα λάθη κατάρτισης (παράμετρος). Το χαλαρό SVM περιθώριο μπορεί επίσης να θεωρηθεί ως μια ειδική περίπτωση του αυστηρού SVM περιθωρίου με την τροποποιημένη λειτουργία του πυρήνα, όπως $K \leftarrow K + \frac{1}{C}$.

Η λειτουργία του πυρήνα που περιλαμβάνεται στο πρόβλημα μπορεί να επιλεγεί χρησιμοποιώντας το θεώρημα Mercers. Έχουμε χρησιμοποιήσει την Gaussian λειτουργία (RBF) του πυρήνα της μορφής

$$K(x_i, x_j) = \exp(-\sum_i \frac{(x_i - x_j)^2}{2\sigma_i^2})$$

όπου σ είναι η παράμετρος πλάτους του πυρήνα. Αν πάρουμε σ να είναι σταθερή και να επιτρέπει κάποιο λάθος εκπαίδευσης, τότε η παράμετρος πυρήνα C μπορεί να βελτιστοποιηθεί μαζί με τη σ, για να ελαχιστοποιηθεί το σφάλμα γενίκευσης.

Ο Vapnik, κάνει χρήση αλγόριθμου καθοδικής κλίσης, ενώ ο Keerthi χρησιμοποιεί τις Newton ενημερώσεις για αυτόματο συντονισμό. Και οι δύο αυτές μέθοδοι μπορούν να συγκλίνουν σε βέλτιστες τοπικές λύσεις , που απαιτούν τη χρήση των καλύτερων μεθόδων όπως η χρήση ενός υβριδικού πλαισίου των Newton (με BFGS αναβάθμιση) και γενετικών αλγόριθμων.

5.5. Πειράματα-Αξιολόγηση

Προκειμένου να αξιολογηθεί η προτεινόμενη μέθοδος παραθέτουμε τρία πειράματα πάνω σε τρία διαφορετικά σύνολα δεδομένων.

Δύο σύνολα στοιχείων τα "Ιονόσφαιρα Δεδομένα", τα "δεδομένα Αναγνώρισης κρασιού" επιλέχθηκαν από το αποθετήριο της UCI μηχανής μάθησης της βάσης δεδομένων (***http://www.ics.uci.edu/~mlearn/MLrepository***). Τα ιονόσφαιρα σύνολα δεδομένων έχουν 34 χαρακτηριστικά για ένα σύνολο 351 περιπτώσεων για ένα δυαδικό έργο ταξινόμησης που αντιστοιχεί σε "Καλή" επιστροφή και "Κακή" επιστροφή. Το σύνολο των δεδομένων είναι χωρισμένο τυχαία σε δύο ομάδες:150 παρατηρήσεις για εκπαίδευση και 201 για δοκιμή. Το σύνολο δεδομένων κρασιού αντιστοιχεί σε 13 χημικά συστατικά από 178 ιταλικά κρασιά που προέρχονται από τρεις διαφορετικές ποικιλίες. Έχουμε λύσει εδώ όλα τα προβλήματα δύο κλάσεων, με αποτέλεσμα την απόρριψη των 48 περιπτώσεων που αντιστοιχεί

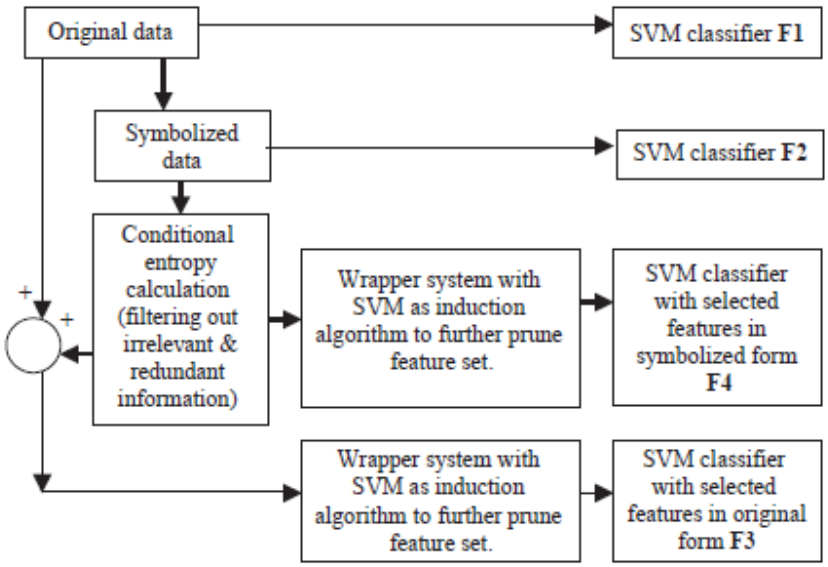
στη Γ'τάξη στα δεδομένα κρασιού. Οι υπόλοιπες 130 μονάδες δεδομένων χωρίζονται σε 80 για εκπαίδευση και 50 περιπτώσεις δοκιμών.

Το τρίτο σύνολο δεδομένων αφορά τον καρκίνο του παχέος εντέρου (***http://microarray.princeton.edu/oncology)***. Το σύνολο δεδομένων το οποίο αφορά τον καρκίνο του παχέος εντέρου αποτελείται από 62 δείγματα του παχέος εντέρου από επιθηλιακά κύτταρα από καρκίνο του παχέος εντέρου ασθενών. Τα δείγματα αποτελούνται από βιοψίες των όγκων που συλλέχθηκαν από όγκους, ενώ οι κανονικές βιοψίες συλλέγονται από υγιές τμήμα του παχέος εντέρου του ίδιου ασθενή. Ο αριθμός των γονιδίων στο σύνολο δεδομένων είναι 2000. Το σύνολο των δεδομένων είναι χωρισμένο σε δύο ομάδες: 30 σχέδια για την εκπαίδευση και 32 για τη δοκιμή.

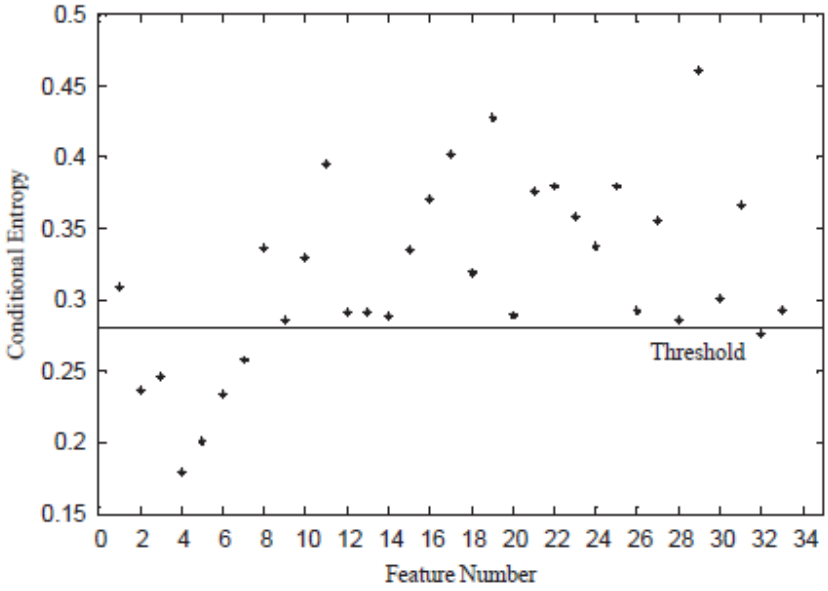
5.6. Αποτελέσματα και Παρατηρήσεις

Η μεθοδολογία που ίσχυσε για την ταξινόμηση των δεδομένων απεικονίζεται παρακάτω. Τα αρχικά δεδομένα μπορούν να να υποβληθούν σε απευθείας κατάταξη σύμφωνα με τον SVM αλγόριθμο. Αυτός ο ταξινομητής έχει οριστεί ως **F1** στη ροή του διαγράμματος. Τα αρχικά δεδομένα μπορεί να περιέχουν κάποιο θόρυβο και ακραίες τιμές, οι οποίες μπορεί να αφαιρεθούν μέσω της διαδικασίας του συμβολισμού.

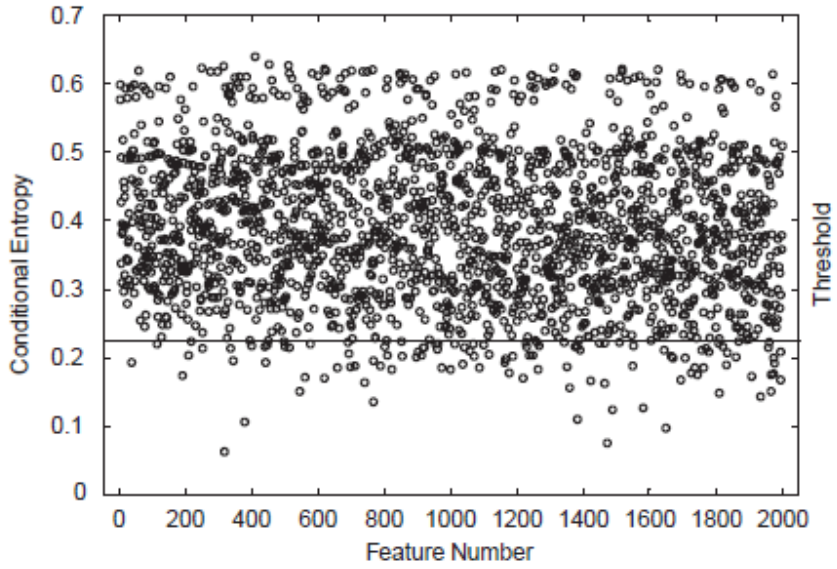
Τα δεδομένα που προκύπτουν μπορεί μετά να ταξινομηθούν και να χαρακτηριστούν ως **F2**. Τα συμβολισμένα δεδομένα μπορούν να υποβληθούν σε περαιτέρω επεξεργασία και να προσδιοριστούν οι άσχετες και περιττές λειτουργίες από τους υπολογισμούς υπό όρους εντροπίας. Οι Εικόνες 21, 22 και 23 παρουσιάζουν τα γραφήματα plot της υπό όρους εντροπίας των πληροφοριών της κάθε κατηγορίας σε σχέση με όλα τα δεδομένα που διαθέτει ένα προς ένα, για τα τρία σύνολα δεδομένων. Οπως αναφέρθηκε προηγουμένως, η υπό όρους εντροπία χρησιμοποιείται ως συναφές filter.



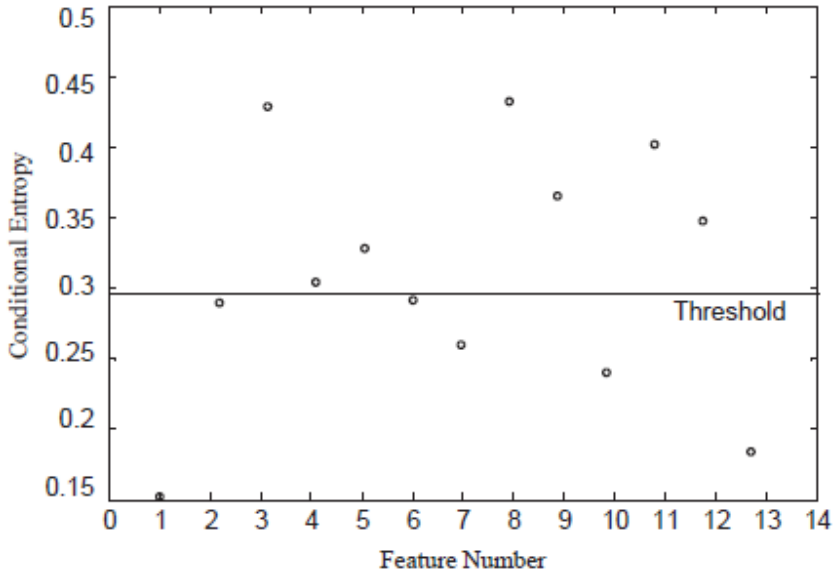
Εικόνα 20. Διάγραμμα ροής που απεικονίζει τη μεθοδολογία



Εικόνα 21. Υπό όρους εντροπίας της πληροφόρησης της ταξινόμησης σε σχέση με όλα τα δεδομένα που διαθέτει ένα-προς-ένα για ιονόσφαιρα δεδομένα



Εικόνα 22. Υπό όρους εντροπίας της πληροφόρησης ταξινόμησης σε σχέση με όλα τα δεδομένα που διαθέτει ένα-προς-ένα για δεδομένα του καρκίνου του παχέος εντέρου



Εικόνα 23. Υπό όρους εντροπίας της πληροφόρησης ταξινόμησης σε σχέση με όλα τα δεδομένα που διαθέτει ένα-προς-ένα για τα δεδομένα κρασιού

Οι χαμηλότερες τιμές της υπό όρους εντροπίας δίνουν τη σημασία που διαθέτουν αυτά τα χαρακτηριστικά και αυτά που υπερβαίνουν ένα ορισμένο όριο μπορούν να θεωρηθούν ως αλυσιτελής. Στις παραπάνω Εικόνες οι οριζόντιες συμπαγείς γραμμές αντιπροσωπεύουν το όριο που καθορίζεται από το χρήστη. Στην Εικόνα 21, η υπό όρους εντροπία είναι ίση με 0,27 αντιπροσωπεύοντας το καθορισμένο από το χρήστη όριο και τα χαρακτηριστικά {2, 3, 4,

5, 6, 7} με όρους εντροπίας χαμηλότερους από το όριο που αποτελεί τη σχετική σειρά. Έτσι το σχετικό filter και μόνο έχει μειώσει τον αριθμό των χαρακτηριστικών από 34 σε 6. Στην Εικόνα 23, η υπό όρους εντροπία είναι ίση με 0.295 και αντιπροσωπεύει το όριο που καθορίζεται από το χρήστη και έχει χαρακτηριστικά {1, 2, 6, 7, 10, 13} και αποτελεί το σχετικό σύνολο. Σε αυτή την περίπτωση επίσης η μείωση των σχετικών χαρακτηριστικών είναι σημαντική από 13 σε 6. Παρόμοια τάση παρατηρείται για τα στοιχεία του καρκίνου του παχέος εντέρου, όπου η υπό όρους εντροπία είναι ίση με το 0,23 και αντιπροσωπεύει το όριο που καθορίζεται από το χρήστη. Κατά τον ίδιο τρόπο απομακρύνεται ο πλεονασμός με υπολογισμό ανά χαρακτηριστικό υπό όρους εντροπίας. Μερικές φορές ακόμη και μετά την αφαίρεση των άσχετων και περιττών πληροφοριών έχουμε μείνει με ένα μεγάλο αριθμό χαρακτηριστικών.

Table 1
SVM Classifier results for non-noisy case

Classifier type	Ionosphere data	Wine data	Colon cancer data
	Test error (%)	Test error (%)	Test error (%)
F1	12.44	5	18.75
F2	12.44	0	18.75
F3	5.97	2.5	15.63
F4	4.98	0	9.38

Table 2
SVM Classifier results for noisy case (SNR = 3)

Classifier type	Ionosphere data	Wine data	Colon cancer data
	Test error (%)	Test error (%)	Test error (%)
F1	16.42	12.5	15.63
F2	16.42	2.5	12.50
F3	16.42	12.5	15.63
F4	14.93	2.5	15.63

Πίνακας 1 και 2

Αυτά μπορούν περαιτέρω να παραμεριστούν χρησιμοποιώντας **wrapper** με SVM ως αλγόριθμο επαγωγής για την απόκτηση αυτών επιλέγουμε λίγα που είναι πιο σημαντικά. Μόλις οι σχετικές ιδιότητες εντοπιστούν μπορούμε να χρησιμοποιήσουμε τις αριθμητικές τιμές που συνδέονται με αυτές για τους σκοπούς της ταξινόμησης.

Ο ταξινομητής έχει οριστεί ως F3 λαμβάνοντας υπόψη ότι η χρήση των συμβόλων για αυτά τα χαρακτηριστικά δίνει ταξινομητή F4. Στην περίπτωση των δεδομένων κρασιού ταυτίζουμε τις ιδιότητες 1, 10 και 13 ενώ για τα ιονόσφαιρα δεδομένα τα γνωρίσματα 2,4 και 5 βρέθηκαν να είναι σημαντικά. Ομοίως τα στοιχεία του καρκίνου του παχέος εντέρου έχουν δώσει χαρακτηριστικά (560, 1745, 765) ως βέλτιστα σύνολα. Για την εκτέλεση συμβολοποίησης χρησιμοποιήσαμε τα ισομήκη διαστήματα κατά τη διάρκεια μιας σειράς χαρακτηριστικών και ως

εκ τούτου δεν φτάσαμε τις βέλτιστες θέσεις των κρίσιμων σημείων. Η Shannon εντροπία έτσι μεγιστοποιείται σε σχέση με τον αριθμό των συμβόλων. Τα αποτελέσματα δείχνουν ότι ακόμη και αυτή η απλουστευμένη προσέγγιση έδωσε άριστα αποτελέσματα.

Προκειμένου να μελετηθεί η επίδραση του θορύβου στην απόδοση της ταξινόμησης, τυχαίος θόρυβος δημιουργήθηκε και προστέθηκε σε κάθε χαρακτηριστικό στο σύνολο των δεδομένων πριν από την εφαρμογή των αλγόριθμων μάθησης. Τα αποτελέσματα ταξινόμησης για διάφορα σύνολα δεδομένων για τις θορυβώδη και μη θορυβώδη περιπτώσεις που παρουσιάζονται στους Πίνακες 1 και 2. Από αυτά τα αποτελέσματα είναι σαφές ότι η διαδικασία επιλογής χαρακτηριστικών δίνει σημαντική βελτίωση στην απόδοση του ταξινομητή του, τόσο σε θορυβώδη και μη θορυβώδη περιπτώσεις. Επίσης, η ταξινόμηση είναι γενικά καλύτερη για τα δεδομένα συμβόλων από ό,τι για τα αρχικά δεδομένα. Ο λόγος για τις βελτιώσεις αυτές είναι ότι τα άσχετα, περιττά και θορυβώδη στοιχεία των δεδομένων απομακρύνονται από το συνδυασμένο αποτέλεσμα της επιλογής χαρακτηριστικών και ο συμβολισμός που ελαχιστοποιεί τις επιπτώσεις του μικρού εύρους των στοιχείων των μετρήσεων που δεν έχουν σχέση με την δυναμική που κυριαρχεί στης μεγάλης κλίμακας εκδηλώσεις.

Table 3
SVM Classifier CPU-TIME for non-noisy case: (seconds)

Classifier type	Ionosphere data	Wine data	Colon cancer data
F1	2.8290	0.6560	3.0000
F2	2.7810	0.5630	2.9530
F3	2.1250	0.4370	0.2340
F4	2.0320	0.4360	0.2190

Table 4
SVM Classifier CPU-TIME for noisy case: (seconds)

Classifier type	Ionosphere data	Wine data	Colon cancer data
F1	2.8600	0.5160	3.1250
F2	2.7970	0.4380	2.9220
F3	2.0470	0.5000	0.2350
F4	1.7350	0.4220	0.2340

Πίνακας 3 και 4

Table 5
KNN Classifier results for non-noisy case

Classifier type	Ionosphere data	Wine data	Colon cancer data
	Test error (%)	Test error (%)	Test error (%)
F1	11.99	0	9
F3	8.06	1	8

Πίνακας 5

Αυτό έχει πρακτικό πλεονέκτημα την απλούστευση και επιτάχυνση των μετέπειτα υπολογισμών (π.χ. βήμα wrapper). Ο χρόνος που απαιτείται για να πραγματοποιήσουμε μία φορά κατάρτιση και έλεγχο των ταξινομητών SVM για τα τρία σύνολα δεδομένων και στις δύο μη θορυβώδεις και θορυβώδεις περιπτώσεις παρουσιάζεται στους Πίνακες 3 και 4. Είναι σαφές από τα αποτελέσματα ότι η επιλογή χαρακτηριστικών και ο συμβολισμός έχει μειώσει σημαντικά τον χρόνο υπολογισμού.

Η υπολογιστικός χρόνος που έχει αναφερθεί είναι ο χρόνος CPU που απαιτείται για την εκτέλεση μιας επανάληψης της κατάρτισης και τη δοκιμή του ταξινομητή της SVM. Όλες οι προσομοιώσεις πραγματοποιήθηκαν σε Pentium IV, μνήμη RAM 512 MB μηχανή. Για να ελέγξουμε τη γενίκευση της μεθόδου ταξινομούμε τα δεδομένα χρησιμοποιώντας τους k πλησιέστερους γείτονες (KNN) αλγόριθμους με επιλεγμένα χαρακτηριστικά. Σημειώστε ότι τα δεδομένα που χρησιμοποιούνται με KNN είναι χωρίς συμβολισμό, καθώς ο συμβολισμός είναι άσχετος στην περίπτωση της απόστασης με βάση τη μέθοδο KNN. Τα αποτελέσματα για KNN παρουσιάζονται στους Πίνακες 5 και 6. Όπως φαίνεται, η απόδοση του ταξινομητή KNN είναι συγκρίσιμη και για τις δύο περιπτώσεις. Η κύρια συνεισφορά είναι η ταυτόχρονη εκτέλεση της επιλογής χαρακτηριστικών και μείωση του θορύβου με την καλύτερη γενικευμένη έννοια της επιλογής χαρακτηριστικών.

Table 6
KNN Classifier results for noisy case

Classifier type	Ionosphere data	Wine data	Colon cancer data
	Test error (%)	Test error (%)	Test error (%)
F1	15.42	3	9
F3	16.92	3	10

Πίνακας 6

Ο υπολογισμός του μήκους κατηγορίας, το χαρακτηριστικό γνώρισμα των συνδέσμων και οι υψηλότερες συσχετίσεις τάξης αγνοούνται . Αυτά μπορούν να συμπεριληφθούν αλλά χρειάζονται μεγάλα δείγματα δεδομένων, μεγάλος χρόνος υπολογισμού και κόστος.

5.7. Συμπεράσματα

Τα οφέλη από μια τέτοια προεπεξεργασία περιλαμβάνουν τη μείωση του όγκου των δεδομένων που απαιτούνται για να επιτευχθεί η μάθηση, βελτίωση της ακρίβειας και αποτελεσματικότητα. Επίσης, το υβριδικό σύστημα είναι πιο γενικό από το filter και wrapper, όπως καταδεικνύεται από τα αποτελέσματα KNN.

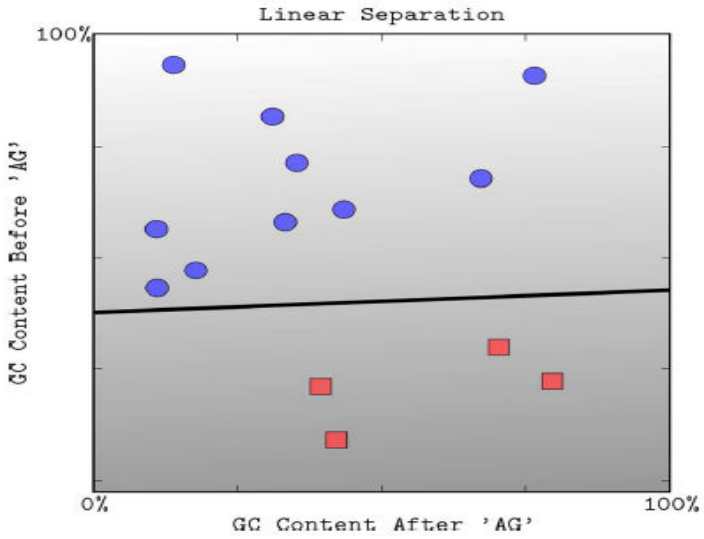
6. Δυαδική κατηγοριοποίηση (Binary Classification)

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

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

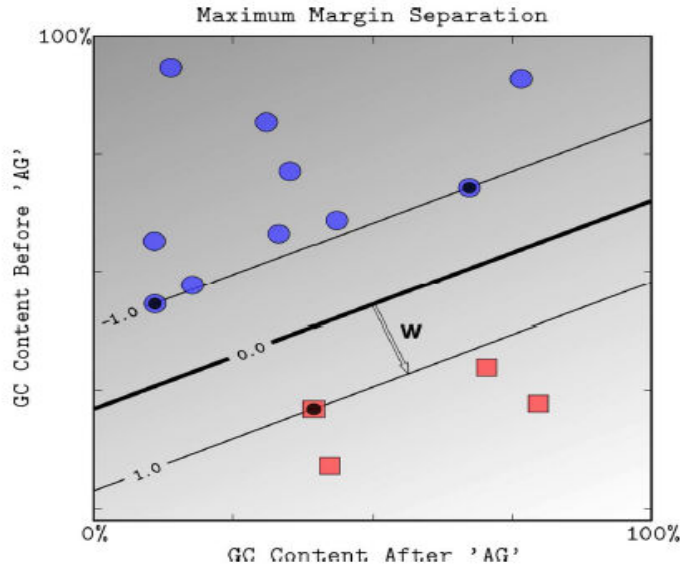
β) πράξεις στο επίπεδο των kernels (πυρήνων) **(kernel functions**).

Η ιδέα του large margin separation μπορεί να αναπαρασταθεί εύκολα όταν διαχωρίζονται σημεία σε δύο διαστάσεις. Ένας απλός τρόπος να διαχωριστούν αυτά τα σημεία είναι να σχεδιάσουμε μια ίσια γραμμή που τα χωρίζει, και να αποκαλέσουμε τα σημεία που βρίσκονται στη μία πλευρά θετικά, και αυτά που βρίσκονται στην άλλη πλευρά αρνητικά. Αν τα δύο σύνολα χωριστούν σωστά, γίνεται προσπάθεια να σχεδιαστεί ξανά η γραμμή, αλλά αυτή τη φορά όσο πιο μακριά γίνεται από τα σημεία και των δύο συνόλων. Αυτή η επιλογή σχεδίασης αποτελεί την ιδέα του διαχωρισμού με το μεγαλύτερο δυνατό περιθώριο (large margin separation).



Εικόνα 24.Ένας γραμμικός κατηγοριοποιητής ο οποίος διαχωρίζει δύο κατηγορίες σημείων (τετράγωνα και κύκλους), σχεδιασμένα σε δύο διαστάσεις

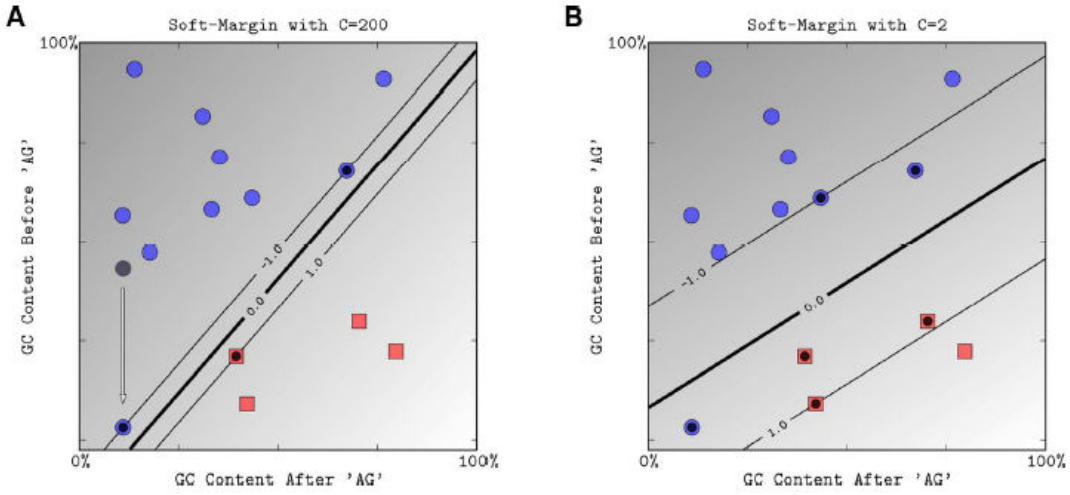
Η διαχωριστική γραμμή χωρίζει τον χώρο σε δύο σύνολα ανάλογα με το πρόσημο. Η απόχρωση του γκρι αναπαριστά την τιμή της διακρίνουσας εξίσωσης: σκούρο για χαμηλές τιμές και ανοιχτό για υψηλές.



Εικόνα 25. Το μέγιστο περιθώριο όπως υπολογίζεται από μία γραμμική SVM

Η περιοχή ανάμεσα στις δύο λεπτές γραμμές ορίζει την περιοχή του περιθωρίου (margin area) όπου $-1 \le \langle w, x \rangle + b \le 1$.

Τα σημεία των δεδομένων που έχουν μαύρα κέντρα είναι τα support vectors (SV), δηλαδή τα σημεία που είναι κοντά στο σύνορο απόφασης (**decision boundary**). Αυτά ορίζουν το περιθώριο το οποίο χωρίζει τις δύο κατηγορίες. Στο παραπάνω σχήμα φαίνονται τρία support vectors πάνω στα σύνορα (όπου $f(\mathbf{x}) = -1$ or $f(\mathbf{x}) = +1$).

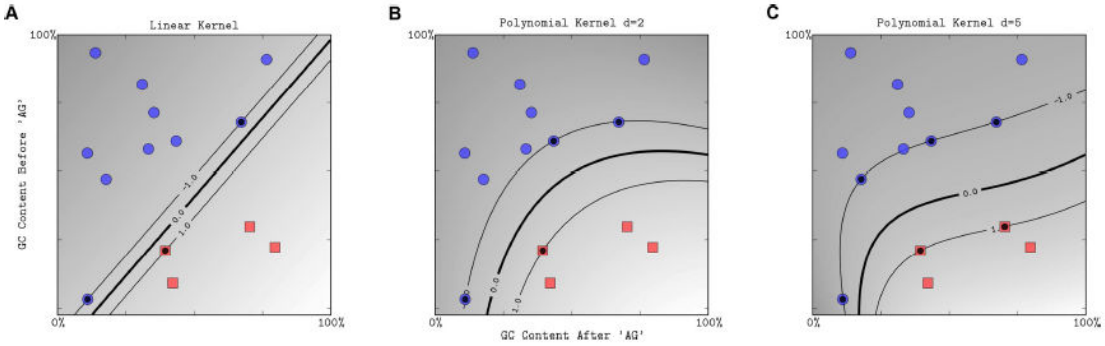


Εικόνα 26. Η επιρροή της σταθεράς του soft-margin, C, στο decision boundary

Αν τροποποιήσουμε τα δεδομένα μετακινώντας το σημείο στη νέα θέση που δείχνει το τόξο, αυτό θα μειώσει σημαντικά το περιθώριο με το οποίο τα hard margin SVM μπορούν να

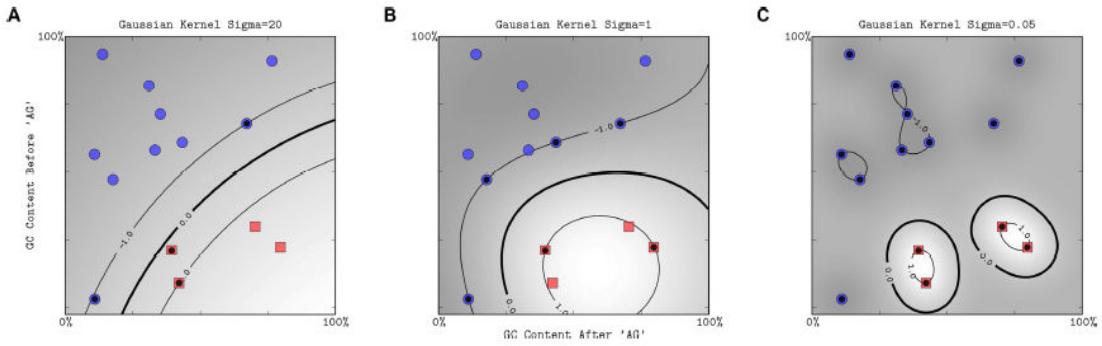
κατηγοριοποιήσουν τα δεδομένα. Η δημιουργία του περιθωρίου χρησιμοποιώντας μια πολύ μεγάλη τιμή του C, μιμείται τη συμπεριφορά του hard margin SVM, το οποίο μας δείχνει ότι κάποια σημεία μπορεί να αποτελούν λάθη και έχουν μεγάλο κόστος στο σχεδιασμό του περιθωρίου. Μία μικρότερη τιμή του C μας επιτρέπει να αγνοήσουμε κάποια σημεία πολύ κοντά στο περιθώριο και έτσι έχει την ευχέρεια να μεγαλώσει το σύνορο. Το περιθώριο επιλογής ανάμεσα στα θετικά και τα αρνητικά σημεία είναι η παχιά γραμμή που φαίνεται στο σχήμα, ενώ οι λεπτές γραμμές αντιπροσωπεύουν το σύνορο (-1 έως +1).

Αντί της ιδέας των σημείων στο χώρο, μπορούμε να σκεφτούμε ότι τα σημεία μας αναπαριστούν αντικείμενα, τα οποία αντιπροσωπεύονται από ένα σύνολο ιδιοτήτων το οποίο προήλθε από μετρήσεις σε κάθε ένα από αυτά. Όταν χρησιμοποιήθηκε μεγάλο σύνορο, αποδείχτηκε ότι η σχετική θέση των σημείων ή η ομοιότητα των σημείων είναι πιο σημαντική από την ακριβή θέση τους. Στην απλούστερη περίπτωση της γραμμικής κατηγοριοποίησης, η ομοιότητα δύο αντικειμένων υπολογίζεται από το εσωτερικό γινόμενο των αντίστοιχων διανυσμάτων (vectors). Για να οριστούν άλλα μέτρα ομοιότητας που οδηγούν σε **μη γραμμική κατηγοριοποίηση**, μπορεί να επεκταθεί η ιδέα του εσωτερικού γινομένου ανάμεσα στα σημεία με τη βοήθεια πράξεων σε επίπεδο **kernel**. Τα kernel (πυρήνες) υπολογίζουν την ομοιότητα ανάμεσα σε δύο σημεία και είναι η δεύτερη σημαντική ιδέα των SVMS και των πράξεων σε επίπεδο kernel.



Εικόνα 27. Η επίδραση του βαθμού ενός πολυωνυμικού kernel

Ένα πολυωνυμικό kernel με βαθμό 1 οδηγεί σε γραμμικό διαχωρισμό. Πολυωνυμικοί kernel υψηλότερου βαθμού επιτρέπουν πιο ευέλικτο σύνορο διαχωρισμού.



Εικόνα 28. Η επιρροή της παραμέτρου του πλάτους του Gaussian kernel (σ) για μια σταθερή τιμή του C.

Για μεγαλύτερες τιμές του σ, το σύνορο διαχωρισμού είναι σχεδόν γραμμικό. Όσο μικραίνει το «σ», η ευελιξία του συνόρου αυξάνεται. Μικρές τιμές του «σ» οδηγούν σε over fitting.
Όπως θα δούμε αργότερα, ορίζοντας ένα κατάλληλο kernel μεταξύ των αντικειμένων, αυτό έχει δύο πλεονεκτήματα:

α) την ικανότητα να παράγει μη-γραμμικά σύνορα διαχωρισμού χρησιμοποιώντας μεθόδους για σχεδιασμένες για μη γραμμικούς κατηγοριοποιητές, και

β) την πιθανότητα να εφαρμοστεί ένας κατηγοριοποιητής σε δεδομένα χωρίς προφανές διανυσματικό διάστημα.

6.1. Large Margin Separation

6.1.1. Γραμμικός διαχωρισμός με hyperplanes (υπερεπίπεδο).

Οι SVMs είναι ένα παράδειγμα ενός γραμμικού κατηγοριοποιητή δύο κλάσεων. Τα δεδομένα για ένα πρόβλημα εκμάθησης περιλαμβάνουν αντικείμενα τα οποία έχουν ένα από τα δύο: +1 (θετικά αντικείμενα) και -1 (αρνητικά αντικείμενα). Έστω το **x** αντιπροσωπεύει ένα διάνυσμα με M στοιχεία, όπου το x_j, j = 1,…,M αναπαριστά ένα σημείο σε ένα διανυσματικό πεδίο M διαστάσεων. Το x_i αναπαριστά το *i-οστό διάνυσμα των δεδομένων* $\{(x_i y_i)\}_{i=1}^{n}$, ενώ το y_i είναι η ετικέτα (label) το οποίο σχετίζεται με το αντικείμενο x_i, και n ο αριθμός των αντικειμένων. Τα αντικείμενα x_i ονομάζονται πρότυπα, είσοδοι ή παραδείγματα. Απαραίτητο για τον ορισμό ενός γραμμικού κατηγοριοποιητή (linear classifier) είναι το εσωτερικό γινόμενο ανάμεσα στα δύο διανύσματα $\langle w, x \rangle = \sum_{j=1}^{M} w_j x_j$. Ο γραμμικός κατηγοριοποιητής βασίζεται σε μία γραμμική διακρίνουσα εξίσωση της μορφής $f(x) = \langle w, x \rangle + b$.

Η διακρίνουσα εξίσωση $f(x)$ χρησιμοποιείται για να ληφθεί η απόφαση ως προς τον τρόπο κατηγοριοποίησης. Το διάνυσμα w είναι γνωστό ως weight factor, και ο βαθμωτός *b* ονομάζεται bias. Σε πρόβλημα δύο διαστάσεων, τα σημεία που ικανοποιούν την εξίσωση $\langle w, x \rangle = 0$,αντιστοιχούν σε μια γραμμή η οποία περνά από την αρχή των αξόνων, σε ένα επίπεδο τριών διαστάσεων, και πιο γενικά ένα υπέρ-επίπεδο (hyperplane). Το bias μετατοπίζει το hyperplane σε σχέση με την αρχή των αξόνων. Το hyperplane διαιρεί το χώρο στα δύο, και σύμφωνα με το πρόσημο της $f(\mathbf{x})$, γνωρίζουμε σε ποια πλευρά του hyperplane θα βρίσκεται κάθε σημείο. Αν η $f(\mathbf{x})>0$, τότε το σημείο βρίσκεται στη θετική πλευρά. Το σύνορο ανάμεσα στις δύο περιοχές οι οποίες έχουν διαχωριστεί ως θετικές ή αρνητικές ονομάζεται σύνορο απόφασης (decision boundary) του κατηγοριοποιητή. Το σύνορο απόφασης ορισμένο από ένα υπέρ-επίπεδο θεωρείται γραμμικό, επειδή είναι γραμμικό στα δεδομένα.

6.1.2. Κατηγοριοποίηση με μεγάλο περιθώριο (Classification with large margin)

Όταν έχουμε να δουλέψουμε με ένα σύνολο δεδομένων το οποίο μπορεί να διαχωριστεί γραμμικά, δεν υπάρχει μόνο ένα hyperplane το οποίο κατηγοριοποιεί σωστά όλα τα σημεία. Εδώ προκύπτει ο προβληματισμός επιλογής ενός μόνο hyperplane, το οποίο όχι μόνο χωρίζει τα σημεία σωστά, αλλά το κάνει με μεγάλο περιθώριο. Το περιθώριο ενός γραμμικού κατηγοριοποιητή, ορίζεται ως η απόσταση του κοντινότερου σημείου στο σύνορο απόφασης. Το hard-margin SVM, μπορεί να εφαρμοστεί σε όλα τα δεδομένα που χωρίζονται γραμμικά, και η ιδιότητα του είναι να κατηγοριοποιήσει σωστά όλα τα σημεία. Αυτή η ιδιότητα δίνει μεγάλη ακρίβεια στον classifier, αλλά έχει χαμηλή απόδοση σε σχέση με τα soft margin SVM.

6.2. Soft margin

Πρακτικά, τα δεδομένα δεν γίνεται πάντα να διαχωριστούν γραμμικά, και ακόμα και αν γίνεται μπορεί να επιτευχθεί μεγαλύτερο περιθώριο (margin) αν εσκεμμένα κατηγοριοποιηθούν λάθος κάποια σημεία . Η θεωρία και τα αποτελέσματα έρευνας έχουν δείξει ότι το μεγάλο περιθώριο θα αποδώσει καλύτερα από το hard margin SVM. Για να επιτραπούν σφάλματα στην κατηγοριοποίηση, μετατρέπουμε την εξίσωση ως εξής:

$$y_i(\langle w, x_i \rangle + b) \geq 1 - \xi_i, \ \ i = 1, ..., n$$

όπου $\xi i \geq 0$ είναι οι μεταβλητές που επιτρέπουν σε ένα σημείο να τοποθετηθεί μέσα στο σύνορο ή σε λάθος κατηγορία. Για να αποτραπεί η υπερβολική χρήση των λάθος τοποθετημένων σημείων, ορίζεται η σταθερά C η οποία θέτει τους όρους για τη μεγιστοποίηση του περιθωρίου και την ελαχιστοποίηση των λάθος κατηγοριοποιήσεων. Η μέθοδος αυτή ονομάζεται soft-margin SVM.
Για μεγαλύτερες τιμές του C, δίνεται βάρος στην αποφυγή των λανθασμένων κατηγοριοποιήσεων, όπου τα δύο σημεία που βρίσκονται κοντά στο hyperplane επηρεάζουν άμεσα τον προσανατολισμό του και το φέρνουν πολύ κοντά στα υπόλοιπα σημεία.

6.3. Κανονικοποίηση δεδομένων (Normalization)

Οι κατηγοριοποιητές οι οποίοι λειτουργούν με μεγάλο περιθώριο, είναι ευαίσθητοι ως προς τον τρόπο εισαγωγής των δεδομένων. Για αυτό το λόγο είναι πολλές φορές απαραίτητο να γίνει κανονικοποίηση των δεδομένων. Η κανονικοποίηση μπορεί να πραγματοποιηθεί είτε σε επίπεδο δεδομένων ή σε επίπεδο kernel, ή και στα δύο. Όταν μετρώνται δεδομένα σε διαφορετικές κλίμακες, γίνεται προσπάθεια να καθοριστούν οι τιμές έτσι ώστε να είναι της ίδιας κλίμακας, π.χ. για κάθε τιμή αφαιρούμε τη μέση τιμή και τη διαιρούμε με την τυπική απόκλιση. Μία εναλλακτική του να κανονικοποιήσουμε κάθε αντικείμενο ξεχωριστά, είναι να μετατρέψουμε τα αντικείμενα σε μοναδιαία διανύσματα. Σε γενικές γραμμές η κανονικοποίηση συχνά προσφέρει βελτιωμένη απόδοση σε γραμμικούς και μη γραμμικούς Kernels, και μπορεί επίσης να οδηγήσει σε ταχύτερη σύγκλιση.

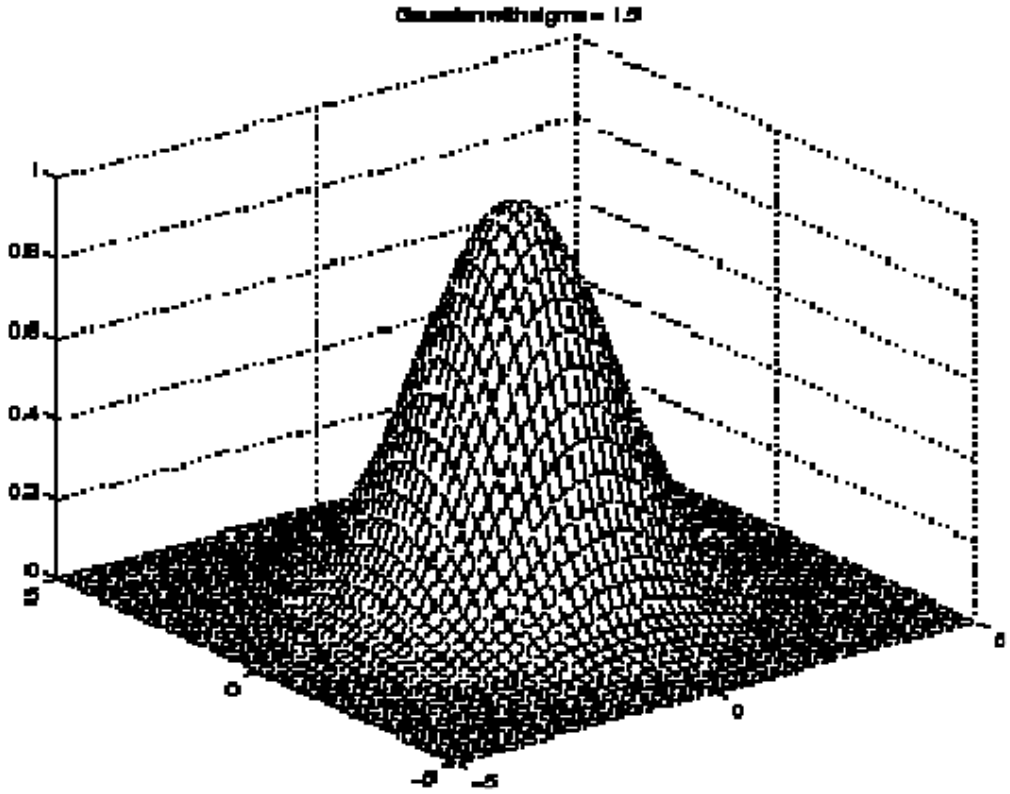
6.4. Χειρισμός μη ισορροπημένου αριθμού δεδομένων

Πολλά σύνολα δεδομένων που συναντάμε δεν είναι ισορροπημένα, δηλαδή η μία κατηγορία αποτελείται από πολύ λιγότερα αντικείμενα σε σχέση με την άλλη. Τέτοια σύνολα δεδομένων δυσκολεύουν πολύ τη δουλειά του κατηγοριοποιητή. Όταν ένα σύνολο δεδομένων δεν είναι ισορροπημένο, το κόστος της λάθος κατηγοριοποίησης δεν είναι και αυτό ισορροπημένο, μιας και ένα λάθος στην κατηγορία με τις λιγότερες τιμές θα κοστίσει πολύ περισσότερο από ένα λάθος στην άλλη κατηγορία. Για την αντιμετώπιση αυτού του προβλήματος δίνονται διαφορετικοί παράμετροι για κάθε κατηγορία, ώστε να μην επιτρέπει να γίνονται εύκολα λάθη στην κατηγορία με τις λιγότερες τιμές .

6.5. Επιλογή Kernel

Ο αλγόριθμος των SVM επιτρέπει τη χρήση kernels, δίνοντας έτσι τη δυνατότητα να υπολογιστούν εσωτερικά γινόμενα σε μη γραμμικό χώρο. Δεν υπάρχει γενικός κανόνας για την επιλογή kernel, μιας και όλα εξαρτώνται από τα δεδομένα. Δοκιμάζεται πρώτα ένας γραμμικός kernel και μετά εξετάζεται αν ένας πολυωνυμικός kernel αν μπορεί να βελτιώσει την απόδοση. Μετά δοκιμάζονται διαφορετικές τιμές για τις παραμέτρους του κάθε kernel.

Τα kernels που χρησιμοποιούνται πιο συχνά είναι τα εξής:
1. Linear
2. Polynomial
3. Gaussian Radial Basis Function



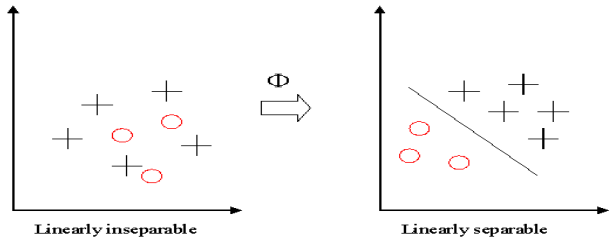
Εικόνα 29.Γραφική αναπαράσταση ενός radial basis function kernel

Πολλές φορές δεν δοκιμάζεται μόνο ένας kernel, αλλά συνδυασμός διαφορετικών kernels, με σκοπό να αυξηθεί η απόδοση του κατηγοριοποιητή.

Kernel

Εάν τα δεδομένα είναι γραμμικά, ίσως χρησιμοποιηθεί ένα διαχωριστικό υπερσύνολο για να χωρίσει τα δεδομένα. Ωστόσο, συχνά τα δεδομένα δεν είναι γραμμικά και τα σύνολα δεδομένων είναι αδιαχώριστα. Για να καταστεί δυνατό αυτό στους πυρήνες χρησιμοποιούνται σε μη γραμμικούς χάρτες τα δεδομένα εισόδου σε ένα υψηλό-διάστατο χώρο. Η νέα χαρτογράφηση στη συνέχεια είναι γραμμικά διαχωρίσιμη.

Ένα πολύ απλό παράδειγμα φαίνεται στην Εικόνα 30.



Εικόνα 30. Χρησιμότητα των kernels

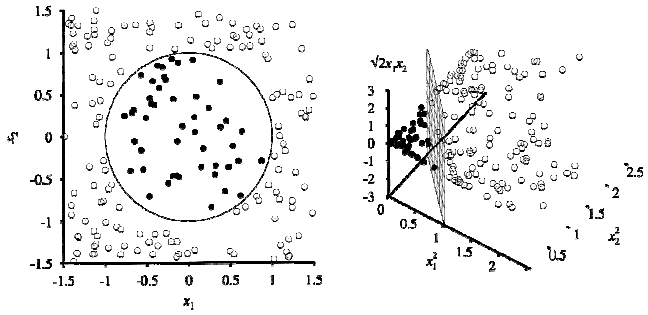
Αυτή η χαρτογράφηση ορίζεται από τον πυρήνα

$$K(x,y) = \Phi(x) \cdot \Phi(y)$$

<u>Ο Χώρος των χαρακτηριστικών</u>

Ο μετασχηματισμός των δεδομένων στο χαρακτηριστικό διάστημα επιτρέπει να ορίσουμε ένα μέτρο ομοιότητας με βάση το προϊόν βαθμωτό dot. Εάν ο χώρος χαρακτηριστικών έχει επιλεγεί κατάλληλα, η αναγνώριση προτύπων μπορεί να είναι εύκολη

$$\langle x_1 \cdot x_2 \rangle \leftarrow K(x_1, x_2) = \langle \Phi(x_1) \cdot \Phi(x_2) \rangle$$



Εικόνα 31. Χαρακτηριστικό χώρος αναπαράστασης

Στην τεχνική kernel, θα δούμε ότι, όταν w, b λαμβάνονται, το πρόβλημα λύνεται με ένα απλό γραμμικό σενάριο στο οποίο τα δεδομένα χωρίζονται από ένα υπερεπίπεδο. Το τέχνασμα επιτρέπει στη SVM να έχει μη γραμμικά όρια. Τα βήματα που εμπλέκονται στο τέχνασμα του kernel δίνονται παρακάτω.

- Ο αλγόριθμος εκφράζεται χρησιμοποιώντας μόνο τα εσωτερικά γινόμενα των συνόλων των δεδομένων. Αυτό καλείται επίσης και ως διπλό πρόβλημα.

- Τα αυθεντικά στοιχεία περνάνε από μη γραμμικούς χάρτες για να σχηματίσουν νέα δεδομένα όσον αφορά τις νέες διαστάσεις με την προσθήκη ενός ζεύγους προϊόντων.

- Αντί για ένα εσωτερικό προϊόν αποθηκεύονται σε πίνακες και στη συνέχεια σε πίνακες αναζήτησης, όπου μπορούν να αντιπροσωπευθούν από dot βαθμωτά προϊόντα των δεδομένων και μετά να γίνει η μη γραμμική χαρτογράφηση τους. Αυτή είναι η λειτουργία του πυρήνα (kernel function).

Kernal Τέχνασμα: Λιπλό πρόβλημα

Πρώτα μετατρέπουμε το πρόβλημα με βελτιστοποίηση στη διπλή μορφή με την οποία προσπαθούμε να εξαλείψουμε το w, και η Lagrangian είναι πλέον η συνάρτηση της λ_i. Για την επίλυση του προβλήματος θα πρέπει να μεγιστοποιηθεί η L_D σε σχέση με την λ_i. Η διπλή μορφή απλοποιεί τη βελτιστοποίηση και βλέπουμε ότι το μεγαλύτερο επίτευγμα είναι το βαθμωτό dot προϊόν που προέρχεται από αυτό.

Kernal Τέχνασμα: περίληψη Εσωτερικού γινομένου

Εδώ βλέπουμε ότι πρέπει να εκπροσωπήσουμε το βαθμωτό dot προϊόν των χρησιμοποιούμενων φορέων των δεδομένων. Το βαθμωτό dot προϊόν των μη γραμμικά χαρτογραφημένων στοιχείων μπορεί να είναι «ακριβό». Το τέχνασμα kernel παίρνει μόνο μια κατάλληλη λειτουργία, η οποία αντιστοιχεί σε βαθμωτά dot προϊόντα κάποιων μη γραμμικών χαρτογραφήσεων. Μερικές από τις πιο συχνές επιλεγόμενες λειτουργίες του πυρήνα, είναι οι ακόλουθες . Ένας ιδιαίτερος πυρήνας επιλέγεται από μια δοκιμή και από το σφάλμα στο σύνολο δοκιμής. Η επιλογή του σωστού πυρήνα με βάση το πρόβλημα ή την εφαρμογή θα βελτιώσει την απόδοση της SVM.

Συναρτήσεις kernel

Η ιδέα της λειτουργίας του πυρήνα είναι να ενεργοποιήσει τις λειτουργίες που πρέπει να εκτελεστούν στο χώρο εισόδου, αντί των δυνητικά υψηλών διαστάσεων του χώρου των χαρακτηριστικών. Εξ' ου και το εσωτερικό προϊόν δεν πρέπει να αξιολογηθεί στο χώρο των χαρακτηριστικών.

Θέλουμε η λειτουργία να εκτελέσει τη χαρτογράφηση των χαρακτηριστικών του χώρου εισόδου στο χώρο των χαρακτηριστικών.

Η λειτουργία του πυρήνα παίζει έναν κρίσιμο ρόλο στην SVM και στην απόδοσή της. Βασίζεται στην αναπαραγωγή των Kernel Χώρων Hilbert .

$$K(x, x') = \langle \varphi(x), \varphi(x') \rangle$$

όπου K είναι μια συμμετρική θετικά ορισμένη λειτουργία, με βάση τις προϋποθέσεις Mercer

$$K(x,x') = \sum_{m}^{\infty} a_m \varphi_m(x)\varphi_m(x'), \quad a_m \geq 0$$

$$\iint K(x,x')g(x)g(x')dxdx' > 0, \quad g \in L_2$$

Στη συνέχεια, ο πυρήνας αποτελεί θεμιτό εσωτερικό γινόμενο στο χώρο των χαρακτηριστικών. Το σύνολο εκπαίδευσης δεν είναι γραμμικά διαχωρίσιμο στο χώρο εισόδου.

Το σύνολο εκπαίδευσης είναι γραμμικά διαχωρίσιμο στο χώρο των χαρακτηριστικών. Αυτό ονομάζεται «τέχνασμα πυρήνα» (“kernel trick”).

Οι διάφορες λειτουργίες του πυρήνα αναφέρονται παρακάτω:

- **Πολυωνυμική** :

  Η χαρτογράφηση πολυωνύμου είναι μια δημοφιλής μέθοδος για τη μη γραμμική μοντελοποίηση. Ο δεύτερος πυρήνας είναι συνήθως προτιμότερος καθώς αποφεύγει τα προβλήματα με τη hessian ώστε να μη μηδενιστεί ($\neq 0$).

  $$K(x,x') = \langle x,x' \rangle^d$$
  $$K(x,x') = (\langle x,x' \rangle + 1)^d$$

- **Gaussian Radial Λειτουργία Βάσης**:

  $$K(x,x') = \exp(-\frac{\|x-x'\|^2}{2\sigma^2})$$

- **Η Εκθετική Radial Λειτουργία Βάσης**:
  Μια ακτινική συνάρτηση βάσης παράγει μια τμηματικά γραμμική λύση, η οποία μπορεί να είναι ελκυστική όταν οι ασυνέχειες είναι αποδεκτές.

  $$K(x,x') = \exp(-\frac{\|x-x'\|}{2\sigma^2})$$

- **Multi-Layer Perceptron**:
  Η μακρά καθιερωμένη MLP, με ένα μόνο κρυφό επίπεδο, έχει επίσης μια έγκυρη αναπαράσταση του πυρήνα.

  $$K(x,x') = \tanh(\rho\langle x,x' \rangle + \rho)$$

Υπάρχουν πολλοί περισσότερες συμπεριλαμβανομένων Fourier, splines, B splines,πυρήνες πρόσθετης ύλης και ο τανυστής των προϊόντων.

6.6. Μέθοδοι για Cross-validation

Το cross-validation είναι μια τεχνική η οποία ελέγχει τον τρόπο που θα χρησιμοποιηθούν τα δεδομένα σε μια στατιστική ανάλυση. Χρησιμοποιείται σε περιπτώσεις όπου στόχος είναι η πρόβλεψη, και επιχειρείται να βρεθεί πόσο ακριβή είναι τα αποτελέσματα. Έτσι ένα σύνολο δεδομένων χωρίζεται σε δύο ή περισσότερα υποσύνολα, χρησιμοποιώντας το ένα υποσύνολο για την εκπαίδευση του αλγορίθμου (training dataset) και το άλλο για την αξιολόγηση του (testing dataset). Για πιο ασφαλή αποτελέσματα, γίνονται πολλές επαναλήψεις του cross-validation χρησιμοποιώντας διαφορετικά υποσύνολα κάθε φορά, και βγαίνει ο μέσος όρος από όλες τις επαναλήψεις. Οι πιο κοινές μέθοδοι για cross-validation είναι οι εξής:

• ***K-fold cross-validation:*** Σε αυτή τη μέθοδο το αρχικό σύνολο δεδομένων χωρίζεται σε k υποσύνολα. Από τα k υποσύνολα ένα χρησιμοποιείται για επαλήθευση (test dataset) και τα υπόλοιπα (k-1) υποσύνολα χρησιμοποιούνται για την εκπαίδευση (training data). Η διαδικασία αυτή επαναλαμβάνεται k φορές (όσες και τα folds), όπου κάθε ένα από τα k υποσύνολα χρησιμοποιείται μία φορά σαν δεδομένο επαλήθευσης. Το μεγαλύτερο πλεονέκτημα αυτής της μεθόδου είναι ότι όλα τα αντικείμενα του συνόλου δεδομένων χρησιμοποιούνται και για εκπαίδευση αλλά και για επαλήθευση.

• ***2-fold cross-validation:*** Αυτή είναι η απλούστερη μορφή του k-fold cross validation και όπως λέει ο τίτλος το σύνολο δεδομένων χωρίζεται σε δύο υποσύνολα. Κάθε fold λαμβάνει αντικείμενα με τυχαίο τρόπο ώστε και τα δύο folds να έχουν ίδιο αριθμό αντικειμένων. Έπειτα το ένα από τα δύο χρησιμοποιείται για εκπαίδευση και το άλλο για επαλήθευση.

• ***Leave-one-out cross-validation:*** Όπως δηλώνει και το όνομα η μέθοδος leave one-out cross-validation (LOOCV), χρησιμοποιεί ένα μόνο αντικείμενο από το αρχικό σύνολο δεδομένων για επαλήθευση και όλα τα υπόλοιπα αντικείμενα χρησιμοποιούνται για εκπαίδευση. Αυτή η διαδικασία επαναλαμβάνεται μέχρι να χρησιμοποιηθούν όλα τα αντικείμενα από μία τουλάχιστον φορά για επαλήθευση. Η διαδικασία είναι ίδια με αυτή του **K-fold cross-validation**, απλά στην προκειμένη περίπτωση ο αριθμός K (folds) είναι ίσος με τον αριθμό των αντικειμένων. Η μέθοδος leave-one-out συνήθως δίνει τα καλύτερα αποτελέσματα, αλλά έχει μεγάλο κόστος σε υπολογιστική ισχύ λόγω των πολλών επαναλήψεων που απαιτούνται για την ολοκλήρωση της εκπαίδευσης.

6.7. Sensitivity και specificity

Τα μέτρα sensitivity (ευαισθησία) και specificity (ειδικότητα) είναι στατιστικά μέτρα απόδοσης-αξιολόγησης μιας μεθόδου δυαδικής κατηγοριοποίησης. Τα αποτελέσματα που λαμβάνουμε χωρίζονται σε τέσσερις κατηγορίες:

1. ***True positives***: Θετικοί (+1) οι οποίοι επαληθεύτηκαν σωστά.

2. ***False positives***: Αρνητικοί (-1) οι οποίοι επαληθεύτηκαν λανθασμένα ως θετικοί.
3. True negatives: Αρνητικοί (-1) οι οποίοι επαληθεύτηκαν σωστά.
4. False negatives: Θετικοί (+1) οι οποίοι επαληθεύτηκαν λανθασμένα ως αρνητικοί.

Sensitivity

Ο όρος sensitivity ορίζεται ως εξής:

$$\text{sensitivity} = \frac{\text{number of true positives}}{\text{number of true positives} + \text{number of false negatives}}$$

Specificity

Ο όρος specificity ορίζεται ως εξής:

$$\text{specificity} = \frac{\text{number of true negatives}}{\text{number of true negatives} + \text{number of false positives}}$$

7. Ανάλυση δεδομένων και εξαγωγή Αποτελεσμάτων

Θα αναλύσουμε τα περιεχόμενα και τη δομή του συνόλου δεδομένων, τον τρόπο χειρισμού τους, την εξαγωγή, την ανάλυση και την επεξήγηση των αποτελεσμάτων του κώδικα. Για την εκπόνηση της εργασίας και την εξαγωγή αποτελεσμάτων χρησιμοποιήθηκε περιβάλλον MATLAB σε συνδυασμό με το LIBSVM toolbox.

7.1. Περιεχόμενα των συνόλων δεδομένων

Τα σύνολα δεδομένων που χρησιμοποιήσαμε αποτελούνται από μετρήσεις 13 σημάτων, σε 4 διαφορετικές κυτταροσειρές. Τα σήματα που χρησιμοποιήσαμε για την εκπαίδευση των SVM, προέρχονται από δοκιμή 7 διαφορετικών φαρμάκων, το αποτέλεσμα των οποίων είναι γνωστό.

Οι τέσσερις κυτταροσειρές από τις οποίες έχουμε λάβει μετρήσεις είναι οι εξής:

- Human hepatoma Hep3B cell line
- Human hepatoma Hepg2 cell line
- Human hepatoma Huh7 cell line
- Mhc cell line

Σε αυτές τις κυτταροσειρές χορηγήθηκαν τα εξής φάρμακα τα αποτελέσματα των οποίων είναι γνωστά, και φαίνονται παρακάτω:

- Sunitinib: pass (label: +1)
- Sorafenib: pass (label: +1)
- Lapatinib: fail (label: -1)
- Gefitinib: fail (label: -1)
- Erlotenib: pass (label:+1)
- Bortezomib: fail (label:-1)
- DMSO: fail (label: -1)

Μετά τη χορήγηση των φαρμάκων, τα σήματα που μετρήθηκαν στα κύτταρα είναι τα εξής:

- AKT
- CREB
- ERK12
- HSP27
- ikb
- IRB
- IRS1
- JNK
- MEK1
- P38
- P70S6
- cmet
- igfr

7.2. Χειρισμός των συνόλων δεδομένων από τον κώδικα

Τα dataset χρησιμοποιήθηκαν με δύο διαφορετικούς τρόπους:

- ***Integrated dataset:*** περιέχει δεδομένα από τη δοκιμή φαρμάκων σε όλες τις κυτταροσειρές, και τα folds διαιρέθηκαν ανά φάρμακο. Τα folds χωρίστηκαν ανά φάρμακο, άρα κάθε fold περιέχει 4 γραμμές με δεδομένα, με το ίδιο φάρμακο αλλά διαφορετική κυτταροσειρά σε κάθε γραμμή των δεδομένων. Συνεπώς έχουμε 7 διαφορετικά folds. Σε κάθε τρέξιμο χρησιμοποιήσαμε 6 folds για εκπαίδευση (train dataset) και ένα για επαλήθευση (test dataset), όπου μετά από 7 loops του κώδικα όλα τα fold χρησιμοποιήθηκαν ακριβώς μία φορά για επαλήθευση.

- ***Ξεχωριστά datasets για κάθε κυτταροσειρά:*** Με αυτά τα dataset κάθε κυτταροσειρά εξετάστηκε μόνη της. Κάθε μία από τις 7 γραμμές του dataset αντιπροσωπεύει μετρήσεις από κάθε φάρμακο. Όπως και στο integrated dataset έχουμε 7 folds, με τη διαφορά ότι κάθε fold περιέχει μία μόνο γραμμή και αντιπροσωπεύει τις μετρήσεις από τη χορήγηση ενός φαρμάκου στην εξεταζόμενη κυτταροσειρά. Κατά την εκτέλεση ο κώδικας έχει την ιδιότητα να δοκιμάζει χρησιμοποιώντας όλα τα σήματα του συνόλου δεδομένων αλλά και μερικό σύνολο αυτών. Επαναληπτικά αφαιρείται κάθε φορά μία στήλη, η οποία περιέχει τις μετρήσεις ενός σήματος, και μετά προκύπτουν δύο διαφορετικά σύνολα δεδομένων. Το ένα περιέχει μόνο τη στήλη που αφαιρέθηκε και το άλλο όλες τις υπόλοιπες στήλες . Αυτό γίνεται για τους εξής λόγους:

  - Για να δούμε πόσο σωστά γίνεται και αν επηρεάζεται η επαλήθευση όταν ένα σήμα αφαιρεθεί από τα δεδομένα
  - Για να δούμε πόσο σωστά μπορεί να επαληθεύσει το μοντέλο με ένα μόνο σήμα.

7.3. Επιλογή dataset και εξαγωγή αποτελεσμάτων

Για να επιδιώξουμε τα καλύτερα δυνατά αποτελέσματα χρησιμοποιώντας τα παραπάνω δεδομένα, χρησιμοποιήθηκαν 5 datasets, τα οποία προέκυψαν από διαφορετικό τρόπο κανονικοποίησης του αρχικού dataset, και στη συνέχεια δοκιμάστηκαν και συγκρίθηκαν ώστε να διαλέξουμε ένα με το οποίο θα συνεχίσουμε τις δοκιμές. Τα datasets είναι τα εξής:

- Bool_0_45
- Bool_0_60
- Relmax
- Relmax_diff
- Relmax_collapse_timepoints

Στους πίνακες των αποτελεσμάτων εμφανίζεται στις γραμμές η κυτταροσειρά μαζί με το φάρμακο που έχει χορηγηθεί, ενώ στις στήλες εμφανίζονται το σήμα ή τα σήματα για τα οποία γίνεται η κατηγοριοποίηση.

7.4. Παρουσίαση πινάκων αποτελεσμάτων και περιεχόμενα

Ο χρωματικός διαχωρισμός στους πίνακες δείχνει τα αποτελέσματα της επαλήθευσης:

- **Πράσινο**: Σωστή επαλήθευση (**true negative ή true positive**)
- **Κόκκινο**: λανθασμένη επαλήθευση (**false negative ή false positive**)
-

Κάτω από κάθε πίνακα αναρτώνται χρήσιμα δεδομένα σχετικά με το kernel που χρησιμοποιήθηκε, την ακρίβεια, το sensitivity και το specificity της κατηγοριοποίησης.

Η διευκρίνιση των παραμέτρων είναι της μορφής “t/d/g” όπου:

- **t: ο τύπος kernel που χρησιμοποιήθηκε** (0: linear, 1: polynomial, 2: Gaussian rbf)
- **d: Ο βαθμός του πολυωνύμου**
- **g: το gamma variation**

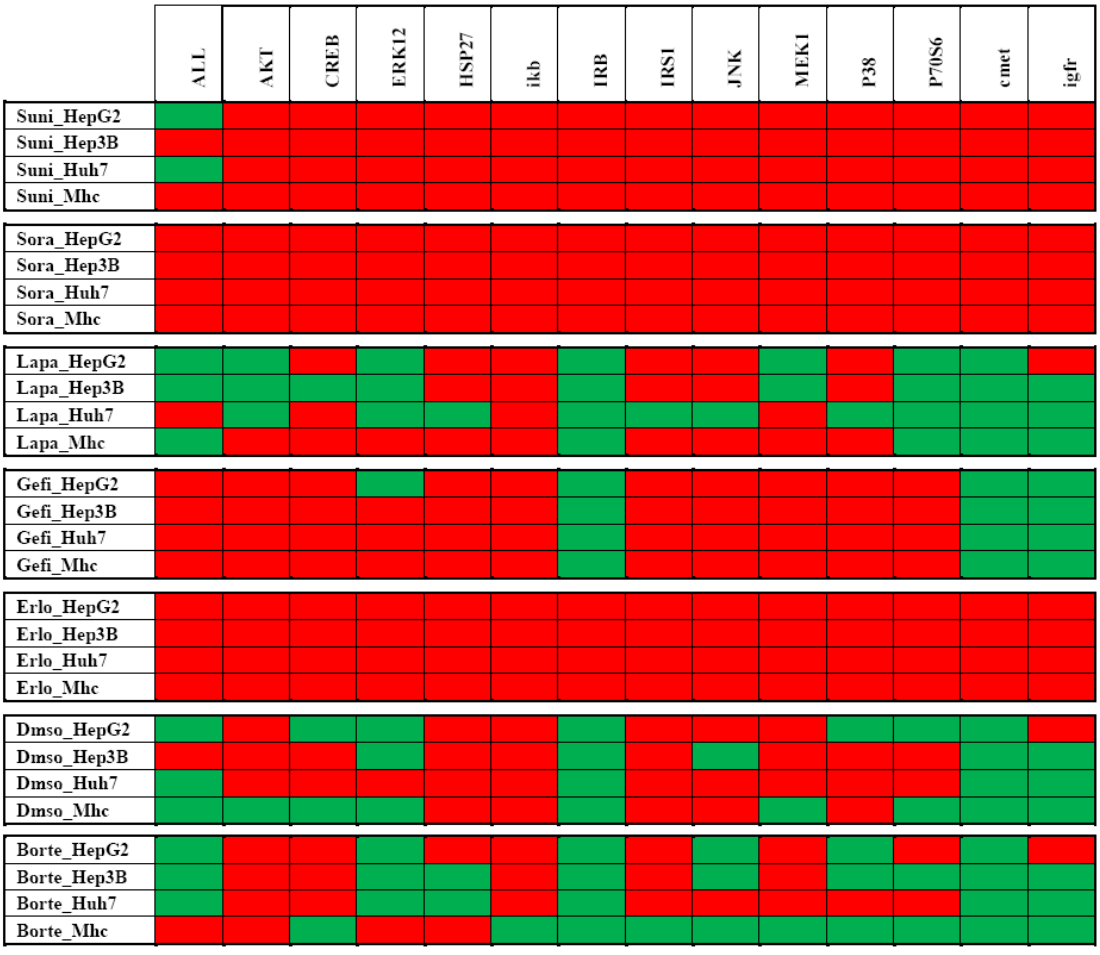
Οι πίνακες που θα συναντήσουμε παρακάτω είναι οι εξής:

- Πίνακας 1: Σύγκριση των 5 διαφορετικών dataset
- Πίνακες 2 - 6: Επαλήθευση δεδομένων σε integrated dataset αλλά και κάθε
- κυτταροσειρά μόνη της με τη μέθοδο one-only.
- Πίνακες 7 - 11: Επαλήθευση δεδομένων σε integrated dataset αλλά και κάθε
- κυτταροσειρά μόνη της με τη μέθοδο one-only.

7.5. Πίνακες αποτελεσμάτων εκπαίδευσης

	bool_0_45	bool_0_60	relmax	relmax_diff_1	relmax_col_t
Suni_HepG2					
Suni_Hep3B					
Suni_Huh7					
Suni_Mhc					
Sora_HepG2					
Sora_Hep3B					
Sora_Huh7					
Sora_Mhc					
Lapa_HepG2					
Lapa_Hep3B					
Lapa_Huh7					
Lapa_Mhc					
Gefi_HepG2					
Gefi_Hep3B					
Gefi_Huh7					
Gefi_Mhc					
Erlo_HepG2					
Erlo_Hep3B					
Erlo_Huh7					
Erlo_Mhc					
Dmso_HepG2					
Dmso_Hep3B					
Dmso_Huh7					
Dmso_Mhc					
Borte_HepG2					
Borte_Hep3B					
Borte_Huh7					
Borte_Mhc					
kernel	0	0	0	1/2/0,1	1/2/0,1
accuracy	0.321428571	0.392857143	0.357142857	0.571428571	0.571428571
sensitivity	0.4375	0.5625	0.5	1	1
specificity	0.166666667	0.166666667	0.166666667	0	0

Πίνακας 1. Σύγκριση των 5 διαφορετικών integrated datasets και επιλογή ενός από αυτά για χρήση στον στον κώδικα.



Kernel	0	0	0	0	0	0	2/2/0.1	0	0	0	0	0	1/2/.1	0
Accuracy	0.393	0.143	0.143	0.357	0.107	0.036	0.571	0.071	0.179	0.143	0.179	0.286	0.571	0.464
Sensitivity	0.563	0.25	0.25	0.625	0.188	0.063	1	0.125	0.313	0.25	0.313	0.5	1	0.813
Specificity	0.167	0	0	0	0	0	0	0	0	0	0	0	0	0

Πίνακας 2. Αποτελέσματα κατηγοριοποίησης με integrated dataset, χρησιμοποιώντας όλες τις κυτταροσειρές και τη μέθοδο one-only, όπου τα folds χωρίστηκαν κατά φάρμακο.

	ALL	AKT	CREB	ERK12	HSP27	ikb	IRB	IRS1	JNK	MEK1	P38	P70S6	cmet	igfr
Suni_Hep3B														
Sora_Hep3B														
Lapa_Hep3B														
Gefi_Hep3B														
Erlo_Hep3B														
Dmso_Hep3B														
Borte_Hep3B														

Kernel	ALL	1/2/0.1	0	1/2/0.1	0	0	0	1/2/0.1	0	1/2/0.1	ALL	0	0	0
Accuracy	0	0.143	0.286	0.286	0.143	0.143	0.571	0.143	0.143	0.286	0	0.429	0.571	0.571
Sensitivity	0	0.25	0.5	0.5	0.25	0.25	1	0.25	0.25	0.5	0	0.75	1	1
Specificity	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Πίνακας 3. Αποτελέσματα κατηγοριοποίησης χρησιμοποιώντας δεδομένα από την κυτταροσειρά Hep3B και τη μέθοδο one-only, μετά από δοκιμή όλων των παραμέτρων και εύρεση των βέλτιστων.

	ALL	AKT	CREB	ERK12	HSP27	ikb	IRB	IRS1	JNK	MEK1	P38	P70S6	cmet	igfr
Suni_HepG2														
Sora_HepG2														
Lapa_HepG2														
Gefi_HepG2														
Erlo_HepG2														
Dmso_HepG2														
Borte_HepG2														

Kernel	0	0	1/2/0.1	0	0	0	0	0	1/2/0.1	1/2/0.1	1/2/0.1	1/2/0.1	0	0
Accuracy	0.714	0.143	0.286	0.286	0.571	0.429	0.571	0.429	0.571	0.143	0.429	0.286	0.571	0.143
Sensitivity	0.75	0.25	0.5	0.5	1	0.75	1	0.75	1	0.25	0.75	0.5	1	0.25
Specificity	0.667	0	0	0	0	0	0	0	0	0	0	0	0	0

Πίνακας 4. Αποτελέσματα κατηγοριοποίησης χρησιμοποιώντας δεδομένα από την κυτταροσειρά HepG2 και τη μέθοδο one-only, μετά από δοκιμή όλων των παραμέτρων και εύρεση των βέλτιστων.

	ALL	AKT	CREB	ERK12	HSP27	ikb	IRB	IRS1	JNK	MEK1	P38	P70S6	cmet	igfr
Suni_huh7														
Sora_huh7														
Lapa_huh7														
Gefi_huh7														
Erlo_huh7														
Dmso_huh7														
Borte_huh7														

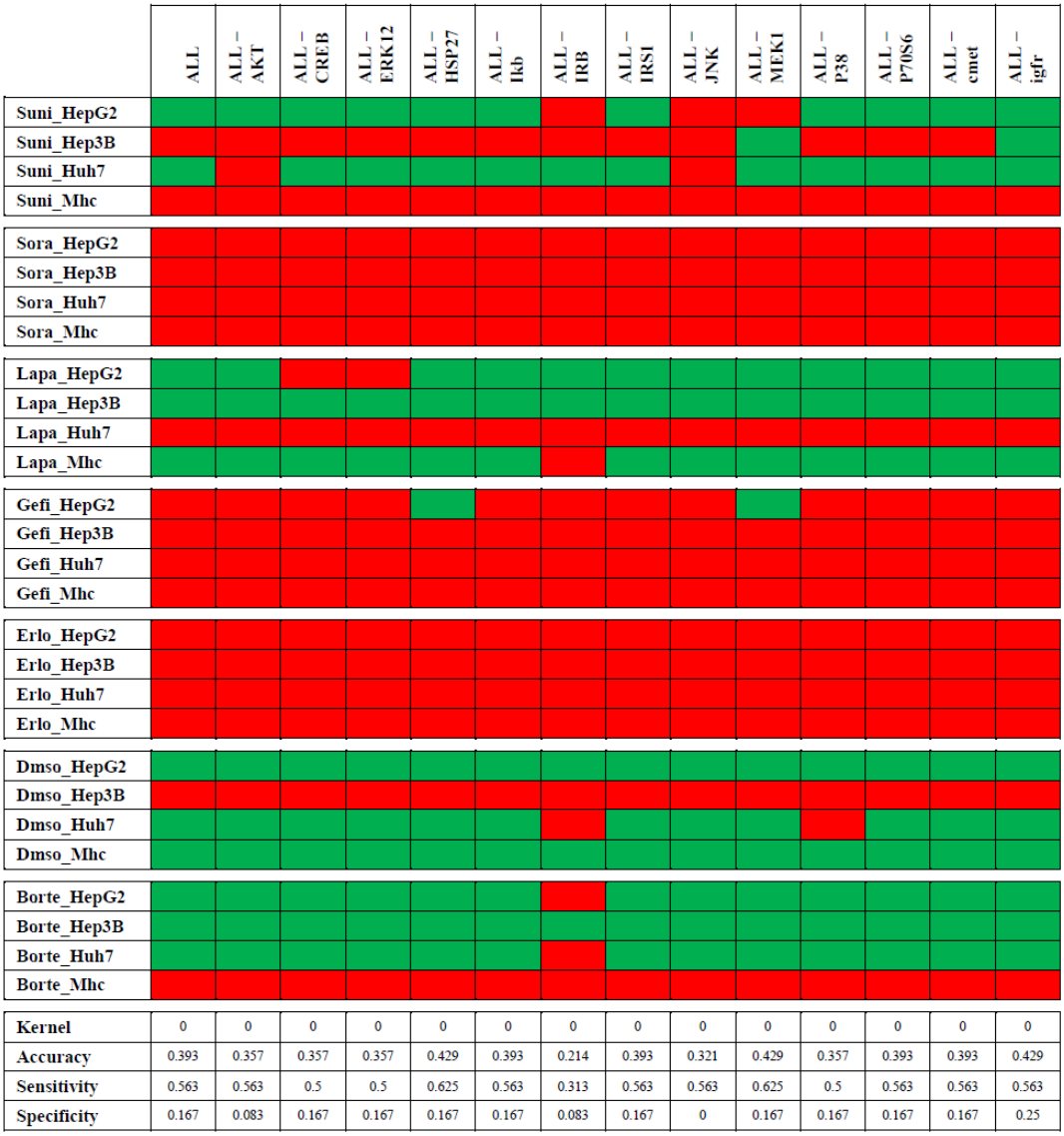
Kernel	0	1/2/0.1	0	0	0	0	0	0	0	0	0	1/2/0.1	0	0
Accuracy	0.143	0.286	0.143	0.286	0.286	0.143	0.286	0.286	0.143	0.143	0.143	0.286	0.571	0.571
Sensitivity	0.25	0.5	0.25	0.5	0.5	0.25	0.5	0.5	0.25	0.25	0.25	0.5	1	1
Specificity	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Πίνακας 5. Αποτελέσματα κατηγοριοποίησης χρησιμοποιώντας δεδομένα από την κυτταροσειρά huh7 και τη μέθοδο one-only, μετά από δοκιμή όλων των παραμέτρων και εύρεση των βέλτιστων.

	ALL	AKT	CREB	ERK12	HSP27	ikb	IRB	IRS1	JNK	MEK1	P38	P70S6	cmet	igfr
Suni_mhc														
Sora_mhc														
Lapa_mhc														
Gefi_mhc														
Erlo_mhc														
Dmso_mhc														
Borte_mhc														

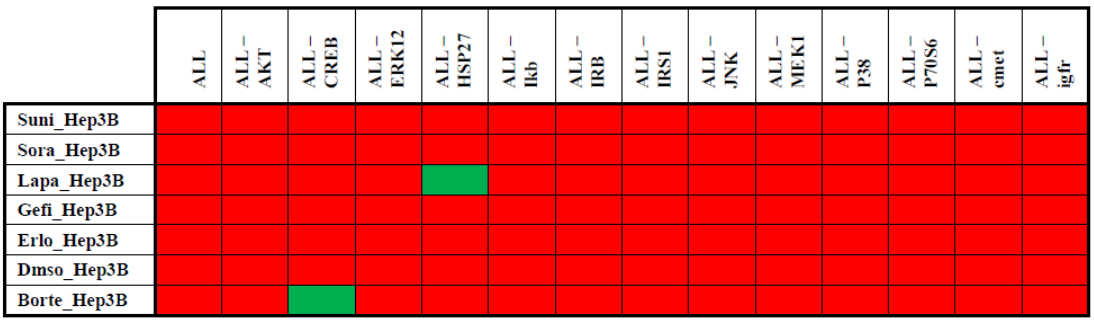
Kernel	0	1/2/0.1	0	0	0	0	0	0	0	0	0	0	0	0
Accuracy	0.714	0.286	0.571	0.143	0.143	0.571	0.571	0.143	0.429	0.429	0.571	0.286	0.571	0.571
Sensitivity	0.75	0.5	1	0.25	0.25	1	1	0.25	0.75	0.75	1	0.5	1	1
Specificity	0.667	0	0	0	0	0	0	0	0	0	0	0	0	0

Πίνακας 6. Αποτελέσματα κατηγοριοποίησης χρησιμοποιώντας δεδομένα από την κυτταροσειρά mhc και τη μέθοδο one-only, μετά από δοκιμή όλων των παραμέτρων και εύρεση των βέλτιστων.



Kernel	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Accuracy	0.393	0.357	0.357	0.357	0.429	0.393	0.214	0.393	0.321	0.429	0.357	0.393	0.393	0.429
Sensitivity	0.563	0.563	0.5	0.5	0.625	0.563	0.313	0.563	0.563	0.625	0.5	0.563	0.563	0.563
Specificity	0.167	0.083	0.167	0.167	0.167	0.167	0.083	0.167	0	0.167	0.167	0.167	0.167	0.25

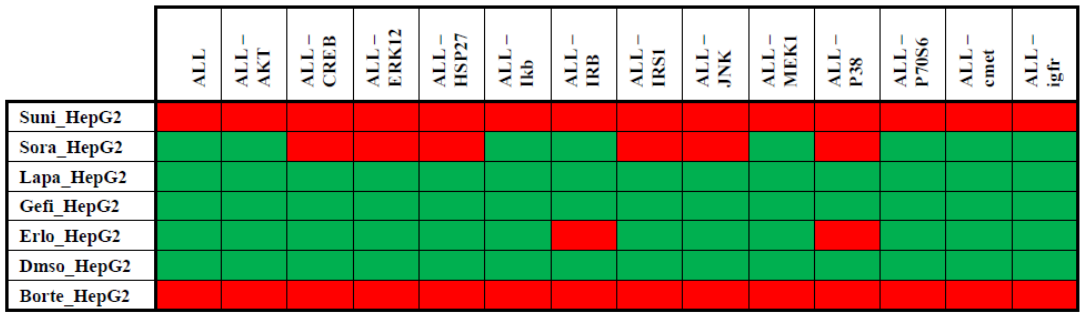
Πίνακας 7. Αποτελέσματα κατηγοριοποίησης με integrated dataset χρησιμοποιώντας όλες της κυτταροσειρές και τη μέθοδο one-out, όπου τα folds χωρίστηκαν κατά φάρμακο.



	ALL	ALL – AKT	ALL – CREB	ALL – ERK12	ALL – HSP27	ALL – Ikb	ALL – IRB	ALL – IRS1	ALL – JNK	ALL – MEK1	ALL – P38	ALL – P70S6	ALL – cmet	ALL – igfr
Suni_Hep3B														
Sora_Hep3B														
Lapa_Hep3B														
Gefi_Hep3B														
Erlo_Hep3B														
Dmso_Hep3B														
Borte_Hep3B														

Kernel	ALL	ALL	1/2/0.1	ALL	0	ALL	ALL	ALL	ALL	ALL	ALL	ALL	ALL	ALL
Accuracy	0	0	0.143	0	0.143	0	0	0	0	0	0	0	0	0
Sensitivity	0	0	0.25	0	0.25	0	0	0	0	0	0	0	0	0
Specificity	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Πίνακας 8. Αποτελέσματα κατηγοριοποίησης χρησιμοποιώντας δεδομένα από την κυτταροσειρά Hep3B και τη μέθοδο one-out, μετά από δοκιμή όλων των παραμέτρων και εύρεση των βέλτιστων.



	ALL	ALL – AKT	ALL – CREB	ALL – ERK12	ALL – HSP27	ALL – Ikb	ALL – IRB	ALL – IRS1	ALL – JNK	ALL – MEK1	ALL – P38	ALL – P70S6	ALL – cmet	ALL – igfr
Suni_HepG2														
Sora_HepG2														
Lapa_HepG2														
Gefi_HepG2														
Erlo_HepG2														
Dmso_HepG2														
Borte_HepG2														

Kernel	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Accuracy	0.714	0.714	0.571	0.571	0.571	0.714	0.571	0.571	0.571	0.714	0.429	0.714	0.714	0.714
Sensitivity	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75
Specificity	0.667	0.667	0.333	0.333	0.333	0.667	0.333	0.333	0.333	0.667	0	0.667	0.667	0.667

Πίνακας 9. Αποτελέσματα κατηγοριοποίησης χρησιμοποιώντας δεδομένα από την κυτταροσειρά HepG2 και τη μέθοδο one-out, μετά από δοκιμή όλων των παραμέτρων και εύρεση των βέλτιστων.

	ALL	ALL – AKT	ALL – CREB	ALL – ERK12	ALL – HSP27	ALL – Ikb	ALL – IRB	ALL – IRS1	ALL – JNK	ALL – MEK1	ALL – P38	ALL – P70S6	ALL – cmet	ALL – igfr
Suni_huh7														
Sora_huh7														
Lapa_huh7														
Gefi_huh7														
Erlo_huh7														
Dmso_huh7														
Borte_huh7														

Kernel	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Accuracy	0.143	0.143	0.143	0.143	0.143	0.143	0.143	0.143	0.143	0.143	0.143	0.143	0.143	0.143
Sensitivity	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25
Specificity	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Πίνακας 10. Αποτελέσματα κατηγοριοποίησης χρησιμοποιώντας δεδομένα από την κυτταροσειρά huh7 και τη μέθοδο one-out, μετά από δοκιμή όλων των παραμέτρων και εύρεση των βέλτιστων.

	ALL	ALL – AKT	ALL – CREB	ALL – ERK12	ALL – HSP27	ALL – Ikb	ALL – IRB	ALL – IRS1	ALL – JNK	ALL – MEK1	ALL – P38	ALL – P70S6	ALL – cmet	ALL – igfr
Suni_mhc														
Sora_mhc														
Lapa_mhc														
Gefi_mhc														
Erlo_mhc														
Dmso_mhc														
Borte_mhc														

Kernel	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Accuracy	0.714	0.714	0.714	0.714	1	0.714	0.714	0.714	0.571	0.714	0.714	0.714	0.714	0.714
Sensitivity	0.75	0.75	0.75	0.75	1	0.75	0.75	0.75	0.5	0.75	0.75	0.75	0.75	0.75
Specificity	0.667	0.667	0.667	0.667	1	0.667	0.667	0.667	0.667	0.667	0.667	0.667	0.667	0.667

Πίνακας 11. Αποτελέσματα κατηγοριοποίησης χρησιμοποιώντας δεδομένα από την κυτταροσειρά mhc και τη μέθοδο one-out, μετά από δοκιμή όλων των παραμέτρων και εύρεση των βέλτιστων.

8. Τεχνικές εξόρυξης πληροφορίας

8.1 ROC Καμπύλες-AUC

8.1.1 ROC γραφήματα

Ένα γράφημα ROC είναι μια τεχνική για την οπτικοποίηση, οργάνωση και ταξινόμηση με βάση την απόδοση των δεδομένων. Τα ROC γραφήματα εδώ και καιρό έχουν χρησιμοποιηθεί στη θεωρία ανίχνευσης σημάτων για να περιγράψουν την σχέση μεταξύ των συντελεστών και ποσοστών εσφαλμένων συναγερμών ή ταξινομητές. Η ROC ανάλυση έχει εκτεταμένη χρήση στην απεικόνιση και στην ανάλυση της συμπεριφοράς των διαγνωστικών συστημάτων. Η ιατρική κοινότητα στη λήψη αποφάσεων έχει εκτεταμένη βιβλιογραφία σχετικά με τη χρήση των γραφημάτων ROC για τις διαγνωστικές δοκιμές.

Ένας από τους πρώτους που υιοθέτησαν τα ROC γραφήματα στη μηχανική μάθηση ήταν Spackman (1989), ο οποίος απέδειξε την αξία των καμπύλων ROC για την αξιολόγηση και σύγκριση αλγορίθμων. Τα τελευταία χρόνια παρατηρείται μια αύξηση στη χρήση των γραφημάτων ROC στην κοινότητα της μηχανικής μάθησης. Εκτός του ότι είναι μία χρήσιμη μέθοδος γενικά στην απόδοση γραφημάτων, έχει και ιδιότητες που την καθιστούν ιδιαίτερα χρήσιμη για τους τομείς με ασύμμετρη κατανομή τάξης και στην άνιση κατανομή σφάλματος. Αυτά τα χαρακτηριστικά τωνROC διαγραμμάτων έχουν γίνει όλο και πιο σημαντικά, καθώς η έρευνα συνεχίζεται σε περιοχές μάθησης που έχουν ευαίσθητο κόστος και στην μάθηση με την παρουσία μιας μη ισορροπημένης τάξης.

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

Η βιβλιογραφία που σχετίζεται με την εξόρυξη δεδομένων και τη μηχανική μάθηση περιλαμβάνει μια συνοπτική μόνο περιγραφή της μεθόδου των ROC καμπύλων με αποτέλεσμα να ανακύπτουν δυσκολίες και πολυπλοκότητες κατά την ερευνητική διαδικασία καθώς και παρανοήσεις όταν χρησιμοποιούνται στην πράξη. Ωστόσο στατιστικά πακέτα όπως το SAS, το STATA και το SPSS έχουν έτοιμες ρουτίνες ανάλυσης καμπύλων ενώ στο δίκτυο αλλά και στη βιβλιογραφία υπάρχουν δημοσιευμένες συναρτήσεις για τα πακέτα S-plus, Matlab και Mathematica.

Κυκλοφορούν επίσης και εξειδικευμένα λογισμικά ανάλυσης καμπύλων ROC με δημοφιλέστερο αυτό του τμήματος Ραδιολογίας του Πανεπιστημίου του Chicago, που είναι διαθέσιμο μέσω δικτύου (www-radiology.uchicago.edu/krl/toppagell.htm). Οι κυριότερες σειρές που αποτελούν το επιστημονικό forum ανάπτυξης της θεωρίας των καμπύλων ROC είναι οι: ‘Statistics in Medicine’, ‘Biometrics’, ‘Biometrica’, ‘Medical Decision Making’ και ‘Academic Radiology’,

ενώ κλασσικά εγχειρίδια είναι των Green, Swets (1974), Egan (1975) και Swets, Pickett (1982). Σημειώνουμε εδώ ότι οι καμπύλες ROC μελετώνται διεξοδικά από τη σκοπιά της διαγνωστικής Ιατρικής σε βιβλίο που εκδόθηκε μόλις τον Αύγουστο του 2002 (Zhou et al 2002).

		Αληθής Ταξινόμηση	
Υποθετική ταξινόμηση	Y	Αληθώς Θετικά	Ψευδώς Θετικά
	N	Ψευδώς Αρνητικά	Αληθώς Αρνητικά

Σύνολο στηλών : P N

$$\text{TP rate } = \frac{TP}{P} = \text{Ανάκληση} \approx \frac{\text{αληθώς θετικά}}{\text{σύνολο θετικών}}$$

Το ψευδώς θετικό ποσοστό (False Positive rate) ενός ταξινομητή είναι:

$$\text{FP rate } = \frac{FP}{N} \approx \frac{\text{ψευδώς αρνητικά}}{\text{σύνολο αρνητικών}}$$

$$\text{Ακρίβεια} = \frac{TP}{TP + FP}$$

$$F - score = \text{Ακρίβεια } x \text{ Ανάκληση}$$

$$\text{Επιπολασμός} = \frac{TP + TN}{P + N}$$

8.1.2 Ταξινόμηση απόδοσης

Αρχίζουμε με την εξέταση των προβλημάτων χρησιμοποιώντας μόνο δύο κατηγορίες. Τυπικά, κάθε περίπτωση I θα αντιστοιχίζεται σε ένα στοιχείο του συνόλου $\{p, n\}$ των θετικών και αρνητικών κατηγοριών. Ένα μοντέλο κατάταξης (ταξινόμησης), είναι μια χαρτογράφηση από τις περιπτώσεις στις προβλεπόμενες κατηγορίες. Για να γίνει διάκριση μεταξύ της πραγματικής τάξη

και της προβλεπόμενης τάξης χρησιμοποιούμε την ετικέτα $\{Y, N\}$ για τις προβλέψεις που παράγονται από ένα μοντέλο.

Δεδομένου ενός ταξινομητή και ενός παραδείγματος, υπάρχουν τέσσερα πιθανά αποτελέσματα. Αν η περίπτωση είναι θετική και είναι ταξινομημένη ως θετική, υπολογίζεται ως μια αληθώς θετική, εφόσον έχει ταξινομηθεί ως αρνητική, αυτό υπολογίζεται ως ψευδώς αρνητική. Αν η περίπτωση είναι αρνητική και έχει ταξινομηθεί ως αρνητική, αυτή υπολογίζεται ως ένα αληθώς αρνητική, εφόσον έχει ταξινομηθεί ως θετική, προσμετρείται ως ψευδώς θετική. Δεδομένου ενός ταξινομητή και μια σειρά από περιπτώσεις (το σύνολο δοκιμής), ένας 2x2 πίνακας (ονομάζεται επίσης πίνακας συνάφειας) μπορεί να κατασκευαστεί αντιπροσωπεύοντας τις διατάξεις του συνόλου των περιπτώσεων. Αυτή η μήτρα αποτελεί τη βάση για πολλές μετρήσεις. Οι αριθμοί κατά μήκος των κυρίων διαγωνίων αντιπροσωπεύουν τις σωστές αποφάσεις, και οι αριθμοί εκτός της διαγωνίου αντιπροσωπεύουν τα λάθη -τη σύγχυση- μεταξύ των διαφόρων κατηγοριών. Το Αληθώς Θετικό ποσοστό (True Positive rate) (ονομάζεται επίσης ποσοστό επιτυχίας και ανάκληση) ενός ταξινομητή υπολογίζεται ως εξής:

$$\text{TP rate} \approx \frac{\text{αληθώς θετικά}}{\text{σύνολο θετικών}}$$

Το ψευδώς θετικό ποσοστό (False Positive rate) ενός ταξινομητή είναι:

$$\text{FP rate} \approx \frac{\text{ψευδώς αρνητικά}}{\text{σύνολο αρνητικών}}$$

Επιπλέον όροι σχετιζόμενοι με καμπύλες ROC:

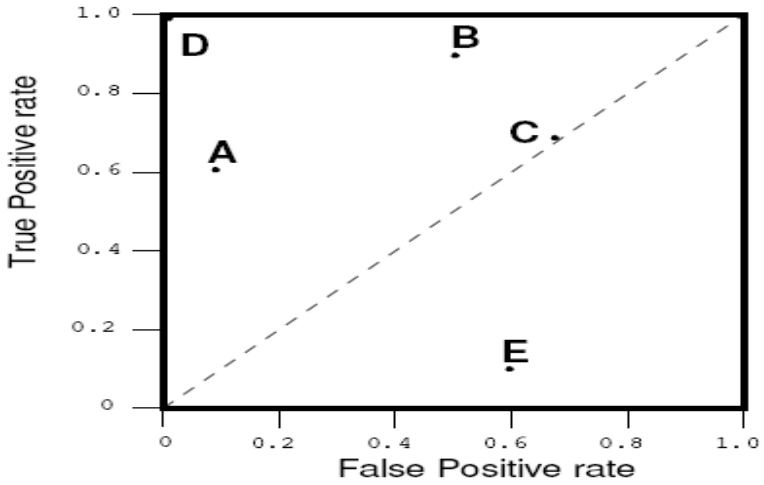
$$\text{Ευαισθησία}(sensitivity) = \text{Ανάκληση}$$

$$Ειδικότητα(specificity) = \frac{\text{Αληθώς αρνητικά}}{\text{Ψευδώς θετικά} + \text{Αληθώς Αρνητικά}} = 1 - F\,\text{Pr}\,ate$$

$$PPV\,(Positive\ predictive\ value) = \text{Ακρίβεια}$$

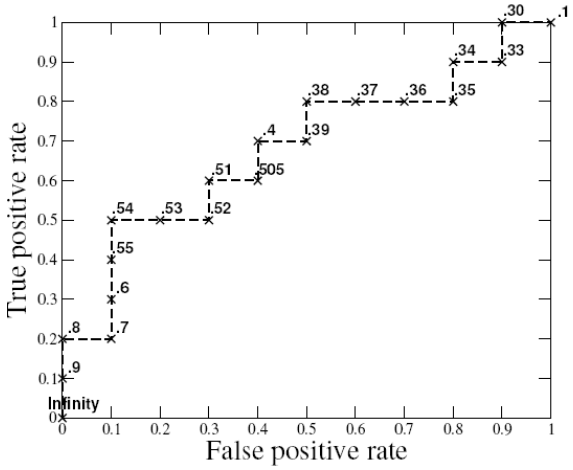
8.1.3 Τυχαία απόδοση

Η διαγώνια γραμμή y = x αντιπροσωπεύει τη στρατηγική της τυχαίας υπόθεσης μιας τάξης. Έτσι, μια τυχαία κατάταξη θα παράγει ένα σημείο ROC το οποίο μετακινείται εμπρός και πίσω στη διαγώνιο με βάση τη συχνότητα με την οποία εικάζει τη θετική τάξη. Για να ξεφύγουμε από τη διαγώνιο στο άνω τριγωνική περιοχή, η ταξινόμηση πρέπει να εκμεταλλευτεί κάποιες πληροφορίες όσον αφορά τα δεδομένα. Στην εικόνα παρακάτω η απόδοση του C είναι σχεδόν τυχαία. Στο (0.7, 0.7), το C μπορεί να ειπωθεί ότι εικάζει το 70% του χρόνου, την θετική ταξινόμηση.



Εικόνα 32.Classifiers TP-FP

Κάθε ταξινομητής που εμφανίζεται στο κάτω δεξιό τρίγωνο εκτελεί δυσμενέστερα από ό, τι η τυχαία εικασία. Αυτό το τρίγωνο είναι ως εκ τούτου συνήθως άδειο στα ROC γραφήματα. Ωστόσο, σημειώστε ότι ο χώρος απόφασης είναι συμμετρικός σχετικά με τη διαγώνιο



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

Εικόνα 33.Η ROC καμπύλη δημιουργήθηκε από ένα σύνολο ορίων ελέγχου.

Ο πίνακας στα δεξιά δείχνει είκοσι δεδομένα και το σκορ ανατεθεί σε κάθε ένα τη βαθμολόγηση. Το γράφημα στα αριστερά δείχνει την αντίστοιχη καμπύλη ROC με κάθε σημείο χαρακτηρισμένο από το όριο που το παράγει.

που χωρίζει τα δύο τρίγωνα. Αν αντιστρέψετε μια ταξινόμηση δηλαδή, η αντίστροφη της ταξινόμησης των αποφάσεων σε κάθε περίπτωση, οι αληθώς θετικά ταξινομήσεις γίνονται ψευδώς θετικά λάθη και ψευδώς θετικά γίνονται αληθώς θετικά. Ως εκ τούτου, οποιαδήποτε ταξινόμηση που παράγει ένα σημείο στο κάτω δεξιά τρίγωνο μπορεί να εξαλειφθεί για να

παραχθεί ένα σημείο στο επάνω αριστερό τρίγωνο. Στο παραπάνω σχήμα, το Ε εκτελεί πολύ χειρότερα από ό, τι το τυχαίο, και είναι στην πραγματικότητα η άρνηση του Α.

8.1.4 Η περιοχή κάτω από την ROC καμπύλη (AUC)

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

Η AUC έχει μια σημαντική στατιστική ιδιότητα: η AUC ενός ταξινομητή είναι ισοδύναμη με την πιθανότητα που ο ταξινομητής θα ταξινομήσει ένα τυχαίο επιλεγμένο θετικό παράδειγμα υψηλότερο από ένα τυχαία επιλεγμένο αρνητικό παράδειγμα. Αυτό είναι ισοδύναμο με τη δοκιμή Wilcoxon των βαθμίδων (Hanley & McNeil, 1982). Η AUC είναι επίσης στενά συνδεδεμένη με το δείκτη Gini (Breiman, Friedman, Olshen, & Stone, 1984), ο οποίος είναι διπλάσιος από το χώρο ανάμεσα στην διαγώνιο και στην καμπύλη ROC. Ο Hand και Till (2001) επισημαίνουν ότι Gini + 1 = 2 x AUC.

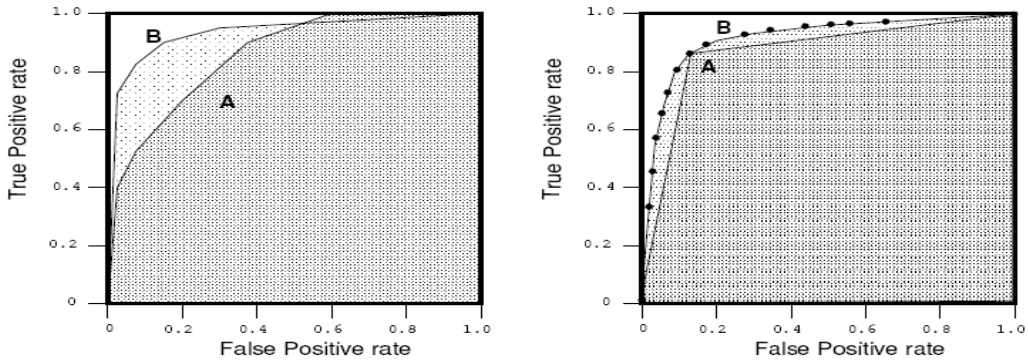
Η παρακάτω εικόνα δείχνει τις περιοχές κάτω από τις δύο ROC καμπύλες, Α και Β. Ο ταξινομητής Β έχει μεγαλύτερη έκταση και συνεπώς καλύτερη μέση απόδοση. Επίσης δείχνει την περιοχή κάτω από την καμπύλη ενός δυαδικού ταξινομητή Α και ένα αθροιστικό ταξινομητή Β. Ο ταξινομητής Α αντιπροσωπεύει την απόδοση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

Είναι δυνατόν για ένα υψηλής-AUC ταξινομητή να έχει χειρότερες επιδόσεις σε μια περιοχή της ROC από ένα χαμηλό-AUC ταξινομητής. AUC μπορεί να υπολογιστεί εύκολα χρησιμοποιώντας τον αλγόριθμο 4.

Algorithm 4 Calculating the area under an ROC curve

Inputs: L, the set of test instances; $f(i)$, the probabilistic classifier's estimate that instance i is positive.
Outputs: A, the area under the ROC curve.

```
1:  L_sorted ⇐ L sorted decreasing by f scores
2:  FP ⇐ TP ⇐ 0
3:  FP_prev ⇐ TP_prev ⇐ 0
4:  A ⇐ 0
5:  f_prev ⇐ −∞
6:  for i ∈ L_sorted do
7:    if f(i) ≠ f_prev then
8:      A ⇐ A + TRAP_AREA(FP, FP_prev, TP, TP_prev)        /* See A.3 for TRAP_AREA */
9:      f_prev ⇐ f(i)
10:     FP_prev ⇐ FP
11:     TP_prev ⇐ TP
12:   if i is a positive example then
13:     TP ⇐ TP + 1
14:   else
15:     FP ⇐ FP + 1
16: A ⇐ A + TRAP_AREA(1, FP_prev, 1, TP_prev)
17: A ⇐ A/(P · N)                                          /* scale from P × N onto the unit square */
18: end
```



Εικόνα 34.Δύο ROC γραφήματα.

Το γράφημα στα αριστερά δείχνει την περιοχή κάτω από δύο καμπύλες ROC. Το γράφημα στα δεξιά δείχνει την περιοχή κάτω από τις καμπύλες του διακριτού ταξινομητή A και του πιθανού ταξινομητή B.

Αντί της συλλογής ROC σημείων, ο αλγόριθμος προσθέτει διαδοχικά περιοχές τραπεζοειδών στην Περιοχή.

8.2 Εισαγωγή στο πρόβλημα

Πρόκειται για μία βάση δεδομένων με 10.333 (n) εγγραφές και 11 επεξηγηματικές μεταβλητές.

Οι 11 αυτές επεξηγηματικές μεταβλητές που μελετήθηκαν είναι οι εξής :

X_1	**Χρόνια (years)**
X_2	**Νομός (1-54)**
X_3	**Γεωγραφικό μήκος (longitude)**
X_4	**Γεωγραφικό πλάτος (latitude)**
X_5	**Ένταση (1-12)**
X_6	**Απόσταση από το σεισμό (km)**
X_7	**Hyper distance (degrees)**
X_8	**Αζιμούθιο (degrees)**
X_9	**Επίκεντρο, στον άξονα x (τεταγμένη)**
X_{10}	**Επίκεντρο, στον άξονα y (τετμημένη)**
X_{11}	**Βάθος (0-700 km)**
Y	**Magnitude (0(<6.5), 1(>6.5))**

Όπου τις χωρίζουμε αναλόγως σε συνεχείς (range) και κατηγορικές (ordinal).

- Συνεχείς (Range)

X_1	**Χρόνια (years)**
X_3	**Γεωγραφικό μήκος (longitude)**
X_4	**Γεωγραφικό πλάτος (latitude)**
X_6	**Απόσταση από το σεισμό (km)**
X_7	**Hyper distance (degrees)**
X_8	**Αζιμούθιο (degrees)**
X_9	**Επίκεντρο, στον άξονα x (τεταγμένη)**
X_{10}	**Επίκεντρο, στον άξονα y (τετμημένη)**
X_{11}	**Βάθος (0-700 km)**

- Κατηγορικές (ordinal)

X_2	**Νομός (1-54)**
X_5	**Ένταση (1-12)**

8.2.1 Εφαρμογή Clementine

ΒΗΜΑΤΑ :

1. Άνοιγμα του προγράμματος Clementine από το Windows Start menu
2. Φόρτωση του αρχείου με τις 10.333 εγγραφές και τις 11 επεξηγηματικές μεταβλητές .
3. Τοποθέτηση στο stream canvas ενός type node με σκοπό να διαβαστούν οι τύποι των τιμών των πεδίων. Άρα καθορίζεται ο τύπος των δεδομένων (type) για κάθε πεδίο και επιπρόσθετα καθορίζεται η κατεύθυνση (direction) που επιδεικνύει τον ρόλο που παίζει κάθε πεδίο στην μοντελοποίση.

Πιο αναλυτικά :

Ο τύπος πληροφορίας για κάθε πεδίο πρέπει να τεθεί πριν τα πεδία χρησιμοποιηθούν στα διάφορα modeling nodes. Το Clementine διακρίνει τους εξής τύπους δεδομένων:

- Εύρος (range)
  Το range χρησιμοποιείται για να περιγράψει συνεχείς αριθμητικές τιμές, ένα σύνολο ή μία κλίμακα 0-100 ή 0.75-1.25. ή Μία τιμή range μπορεί να είναι ακέραιος, πραγματικός αριθμός ή ημερομηνία/ώρα.
- Διακριτοποίηση (discrete)
  Το discrete χρησιμοποιείται για να περιγράψει αλφαριθμητικές τιμές όταν ένας ακριβής αριθμός διαφορετικών τιμών είναι άγνωστος (π.χ 1,5,8)
- Δίτιμη παράμετρος-λογική παράμετρος τύπου Boolean (flag)
  Το flag χρησιμοποιείται από δεδομένα με δύο μόνο τιμές yes/no ή 0/1 ή 1/2.
- Σύνολο (set)
  Το set χρησιμοποιείται για να περιγράψει δεδομένα με πολλαπλές διακεκριμένες τιμές όπου η καθεμιά αντιμετωπίζεται σαν μονάδα ενός συνόλου, ή διακεκριμένες κατηγορίες όπως small/medium/large.
- Ανένταχτος τύπος (typeless)
  Το typeless χρησιμοποιείται για δεδομένα που δεν εντάσσονται σε καμία πό τις παραπάνω κατηγορίες ή για δεδομένα τύπου set με πάρα πολλές διακεκριμένες τιμές. Η επιλογή του τύπου typeless ορίζεται αυτόματα το πεδίο του direction se none, δηλαδή το πεδίο που δεν μπορεί να χρησιμοποιηθεί σε μοντέλα.

Τα δεδομένα εντάσσονται αρχικά σε μία από τις παραπάνω κατηγορίες με το πού εισέρχονται στο σύστημα. Για παράδειγμα, ο discrete τύπος δίνεται προσωρινά σε κατηγορικές μεταβλητές μέχρι να μπορεί να προσδιορισθεί αν πρόκειται για set ή flag τύπο και ο τύπος range δίνεται σε όλες τις αριθμητικές μεταβλητές.

Η τιμή του direction ενός πεδίου σχετίζεται μόνο με τη μοντελοποίηση. Υπάρχουν τέσσερις δυνατές κατευθύνσεις :

- IN : το πεδίο χρησιμοποιείται σαν input, δηλαδή είναι μία τιμή που θα βοηθήσει στην πρόβλεψη

- OUT: το πεδίο χρησιμοποιείται σαν output-στόχος της τεχνικής μοντελοποίησης. Δηλαδή είναι το πεδίο που θα προβλέψουμε.
- BOTH: το πεδίο επιτρέπεται να είναι και input και output σε κανόνα συσχέτισης (association rule). Όλες οι άλλες τεχνικές μοντελοποίησης αγνοούν αυτό το πεδίο.
  NONE: το πεδίο δεν χρησιμοποιείται στην μοντελοποίηση

Στο δικό μας dialog box του type node οι τύποι των 11 επεξηγηματικών μεταβλητών ήταν τύπου Range. Η απόκριση y =magnitude αποτελεί τον στόχο πρόβλεψης ,δηλαδή ο στόχος μας στην παρούσα μελέτη ορίζεται να είναι με άλλα λόγια το μέγεθος (σφοδρότητα) του σεισμού. Η y λοιπόν είναι τύπου Flag εφόσον είναι δίτιμη με τιμές y=0 (<6.5)και y=1 (>6.5).

Στην δική μας εφαρμογή όσον αφορά την τιμή του direction έχουμε direction input για τις 11 επεξηγηματικές μεταβλητές που το In δείχνει ότι θα χρησιμοποιηθούν σαν μεταβλητές πρόβλεψης και και direction output για το y=magnitude που το Out δείχνει ότι πρόκειται για το πεδίο που θέλεις να προβλέψεις.

4. Τοποθέτηση στο stream canvas ενός Partition node με σκοπό να χωριστούν τα δεδομένα σε δεδομένα εκπαίδευσης(training set) και δεδομένα ελέγχου-εξέτασης (test dataset). Πιο αναλυτικά :

Σε ένα τυπικό πρόβλημα του data mining, έχουμε ένα σύνολο δεδομένων εκπαίδευσης (training set) στο οποίο γνωρίζουμε την τιμή του αποτελέσματος και τις τιμές των χαρακτηριστικών που μας ενδιαφέρουν, και προσπαθούμε με βάση αυτά τα δεδομένα να κατασκευάσουμε ένα μοντέλο πρόβλεψης. Το μοντέλο αυτό θα το χρησιμοποιήσουμε στη συνέχεια για να προβλέψουμε το αποτέλεσμα νέων συνόλων δεδομένων εξέτασης (test set), στα οποία σύνολα είναι γνωστές οι τιμές των χαρακτηριστικών αλλά δεν είναι γνωστή η τιμή του αποτελέσματος, δηλαδή η τιμή της τάξης. Στην περίπτωση τώρα όπου ο αλγόριθμος που εφαρμόζουμε στηρίζεται σε κατασκευή και εκτίμηση μοντέλου, τα δεδομένα διαχωρίζονται σε δύο υποσύνολα: 1) τα δεδομένα εκπαίδευσης(training data) τα οποία χρησιμοποιούνται για την προσαρμογή του μοντέλου και 2) τα δεδομένα ελέγχου (test data) που χρησιμοποιούνται για τον υπολογισμό της γενικευμένης τιμής σφάλματος του τελικά επιλεγμένου μοντέλου. Καθένα από αυτά τα σύνολα θα πρέπει να επιλεγεί ανεξάρτητα. Αυτός είναι ένας τρόπος να μεγιστοποιήσουμε τα δεδομένα που παράγουν το μοντέλο με το οποίο θα ασχοληθούμε στην πραγματικότητα. Αυτό που είναι σημαντικό, είναι ότι το ποσοστό σφάλματος δεν καθορίζεται με βάση κανένα από αυτά τα δεδομένα.
Γενικά, μπορούμε να πούμε ότι η ποιότητα του μοντέλου είναι ανάλογη του όγκου των διαθέσιμων δεδομένων, αν και συχνά βαίνει φθίνουσα όταν ο όγκος του συνόλου εκπαίδευσης υπερβαίνει κάποιο όριο. Επίσης, και η αξιοπιστία της εκτίμησης του σφάλματος είναι ανάλογη του όγκου των δεδομένων ελέγχου. Τα προβλήματα αρχίζουν όταν δεν υπάρχει επαρκής όγκος δεδομένων και επομένως περιορίζεται το ποσό των δεδομένων που μπορεί να χρησιμοποιηθεί ως σύνολο εκπαίδευσης και σύνολο ελέγχου. Σε τέτοια σύνολα δεδομένων ένα μέρος των δεδομένων χρησιμοποιείται για τον έλεγχο και το υπόλοιπο για την εκπαίδευση. Αυτή η διαδικασία ονομάζεται διαδικασία παρακράτησης (holdout procedure) και το δίλημμα που προκύπτει τώρα είναι πώς διαχωρίσουμε το αρχικό σύνολο έτσι ώστε και τα δύο σύνολα να είναι μεγάλα.

Είναι δύσκολο να δώσουμε ένα γενικό κανόνα σχετικά με το πώς επιλέγεται ο αριθμός των παρατηρήσεων που καταχωρείται σε καθένα από αυτά τα τρία σύνολα, καθώς εξαρτάται από το ποσοστό θορύβου στα δεδομένα και το μέγεθος του δείγματος εκπαίδευσης. Μία τυπική διάκριση που χρησιμοποιείται είναι το 75% στο σύνολο εκπαίδευσης και από 25% στα σύνολα ελέγχου αντίστοιχα.

Εφαρμόζουμε λοιπόν τον παραπάνω διάκριση στις δικές μας 10.333 εγγραφές και προκύπτουν:

- Training set : $\approx 75\% \times 10.333 = 7.749,75$ υποδείγματα
- Test set : $\approx 25\% \times 10.333 = 2.583,25$ υποδείγματα

5. Τοποθέτηση στο stream canvas ενός Feature selection node  τον οποίο συνδέουμε στον Partition node με σκοπό να επιλεγούν για επίπεδο σημαντικότητας α=0.05 οι σημαντικές μεταβλητές
   Πιο αναλυτικά :

Στο δικό μας πρόβλημα εξόρυξης δεδομένων, όπως συμβαίνει και στην πλειοψηφία των προβλημάτων εξόρυξης δεδομένων, εμπεριέχονται εκατοντάδες πεδία-μεταβλητές τα οποία είναι πιθανόν να χρησιμοποιηθούν με σκοπό την πρόβλεψη. Σαν αποτέλεσμα, χρειάζεται να ξοδευτεί αρκετός χρόνος και προσπάθεια για να εξεταστεί ποια από αυτά τα πεδία πρέπει να συμπεριληφθούν στο μοντέλο. Για να μειώσουμε στο ελάχιστο τις πιθανές επιλογές, ο αλγόριθμος της επιλογής των χαρακτηριστικών (Feature Selection Algorithm) μπορεί να χρησιμοποιηθεί για να προσδιορίσει τα πεδία εκείνα τα οποία είναι πιο σημαντικά για την δεδομένη ανάλυση.
Η επιλογή χαρακτηριστικών αποτελείται από τρία βήματα :

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

- Ranking (Στοίχιση)
  Σε αυτό το βήμα ξεχωρίζονται οι εναπομείναντες μεταβλητές πρόβλεψης και καθορίζονται ranks βασισμένα στην σημαντικότητα.

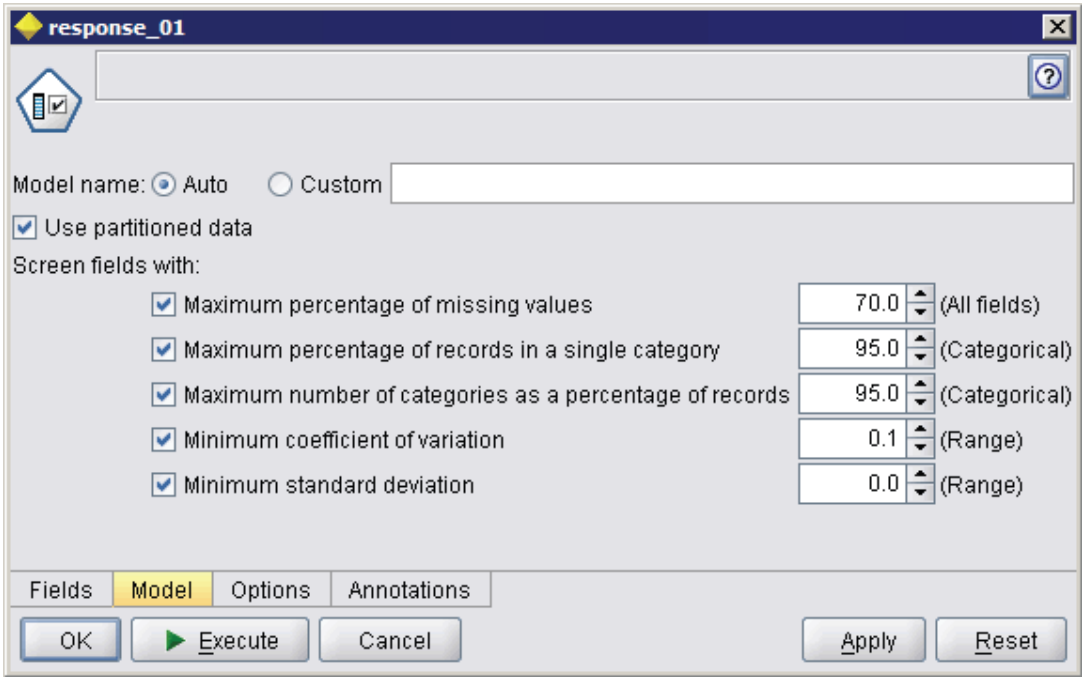
- Επιλογή
  Σε αυτό το βήμα αναγνωρίζεται το υποσύνολο των χαρακτηριστικών που θα χρησιμοποιηθεί στα μοντέλα που ακολουθούν κρατώντας μόνο τις πιο σημαντικές μεταβλητές πρόβλεψης και φιλτράροντας ή αποκλείοντας όλες τις υπόλοιπες.

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

Model tab-Options tab

Στον κόμβο της επιλογής χαρακτηριστικών υπάρχει ένα model tab το οποίο περιλαμβάνει βασικές επιλογές για το μοντέλο καθώς και ρυθμίσεις που επιτρέπουν την εύρεση κριτηρίων για το κρισάρισμα των μεταβλητών πρόβλεψης.

Model tab



Model name : (auto) το όνομα του μοντέλου παράγεται αυτόματα.
Use partitioned data : (ν) το τικάρουμε διότι αυτή η επιλογή μας εξασφαλίζει οτί τα δεδομένα μόνο από το το training set χρησιμοποιούντα για την κατασκευή του μοντέλου.
Τα πεδία κρισάρονται με την βοήθεια των παρακάτω κτριτηρίων :

- Μέγιστο ποσοστό ελλειπουσών τιμών
  Κρισάρει τα πεδία με μεγάλο αριθμό ελλειπουσών τιμών που προσφέρουν ελάχιστη πληροφορία πρόβλεψης
- Μέγιστο ποσοστό εγγγραφών σε μια απλή κατηγορία
  Κρισάρει τα πεδία τα οποία έχουν πάρα πολλές εγγραφές να ανήκουν στην ίδια κατηγορία
- Μέγιστος αριθμός των κατηγοριών ως ποσοστό των εγγραφών

Κρισάρει τα πεδία με πολλές κατηγορίες συγκριτικά με τον συνολικό αριθμό των εγγραφών δηλαδή εάν ένα μεγάλο ποσοστό των κατηγοριών περιέχει μόνο μία περίπτωση, το πεδίο δεν μπορεί παρά να χρησιμοποιηθεί ελάχιστα.

- Ελάχιστος συντελεστής διακύμανσης
  Κρισάρει τα πεδία με συντελεστή βάρους μικρότερο ή ίσο από το καθορισμένο ελάχιστο όριο. Εάν η τιμή είναι κοντά στο 0 , δεν υπάρχει μεγάλη μεταβλητότητα στις τιμές της μεταβλητής.
- Ελάχιστη τυπική απόκλιση
  Κρισάρει τα πεδία με τυπική απόκλιση μικρότερη ή ίση από το καθορισμένο ελάχιστο όριο.

Οι εγγραφές οι οποίες έχουν ελλειπούσες τιμές για το πεδίο στόχου ή ελλειπούσες τιμές για όλες τις μεταβλητές πρόβλεψης , αποκλείονται αυτόματα από όλους τους υπολογισμούς μέσα στο rankings.

Στον κόμβο της επιλογής χαρακτηριστικών υπάρχει ένα options tab το οποίο σου επιτρέπει να καθορίσεις τις default (εξ'ορισμού) ρυθμίσεις για την επιλογή ή τον αποκλεισμό των πεδίων πρόβλεψης του μοντέλου.
Σε αυτό το βήμα θεωρείται μία μεταβλητή πρόβλεψης την φορά για να εξεταστεί πόσο καλά κάθε μεταβλητή πρόβλεψης ξεχωριστά προβλέπει την μεταβλητή στόχο. Οι μεταβλητές πρόβλεψης ιεραρχούνται σύμφωνα με το κριτήριο που καθορίζεται από τον πειραματιστή.
Η τιμή σημαντικότητας κάθε μεταβλητής ή διαφορετικά ένα μέτρο το οποίο χρησιμοποιείται για να βάλει σε σειρά τα πεδία ή τα αποτελέσματα σε ποσοστιαία κλίμακα ορίζεται ως $(1-p)$ όπου p είναι η τιμή p value του κατάλληλου στατιστικού τεστ της σχέσης μεταξύ της υποψήφιας μεταβλητής πρόβλεψης και της μεταβλητής στόχο .
Στην δική μας εφαρμογή χρησιμοποιήσαμε τιμή p value βασισμένη στο στατιστικό του Pearson , το Pearson chi-square το οποίο εξετάζει την ανεξαρτησία του στόχου και της μεταβλητής πρόβλεψης χωρίς να δείχνει την δύναμη ή την κατεύθυνση οποιασδήποτε υπάρχουσας σχέσης .
Αναλυτικά :
Το Pearson chi-square είναι ένα τεστ ανεξαρτησίας μεταξύ X και Y το οποίο περιλαμβάνει την διαφορά μεταξύ των παρατηρούμενων και των αναμενόμενων συχνοτήτων. Τα αναμενόμενα κελιά συχνότητας κάτω από την μηδενική υπόθεση υπολογίζονται-εκτιμώνται από τον τύπο

$$\hat{N}_{ij} = N_{i.}N_{.j} / N$$

Κάτω από την μηδενική υπόθεση, το Pearson chi-square συγκλίνει ασυμπτωτικά σε μία κατανομή chi-square ${\chi_d}^2$ με $d = (I-1)(J-1)$ βαθμούς ελευθερίας.

Το p value το οποίο βασίζεται στο Pearson chi-square X^2 υπολογίζεται ως

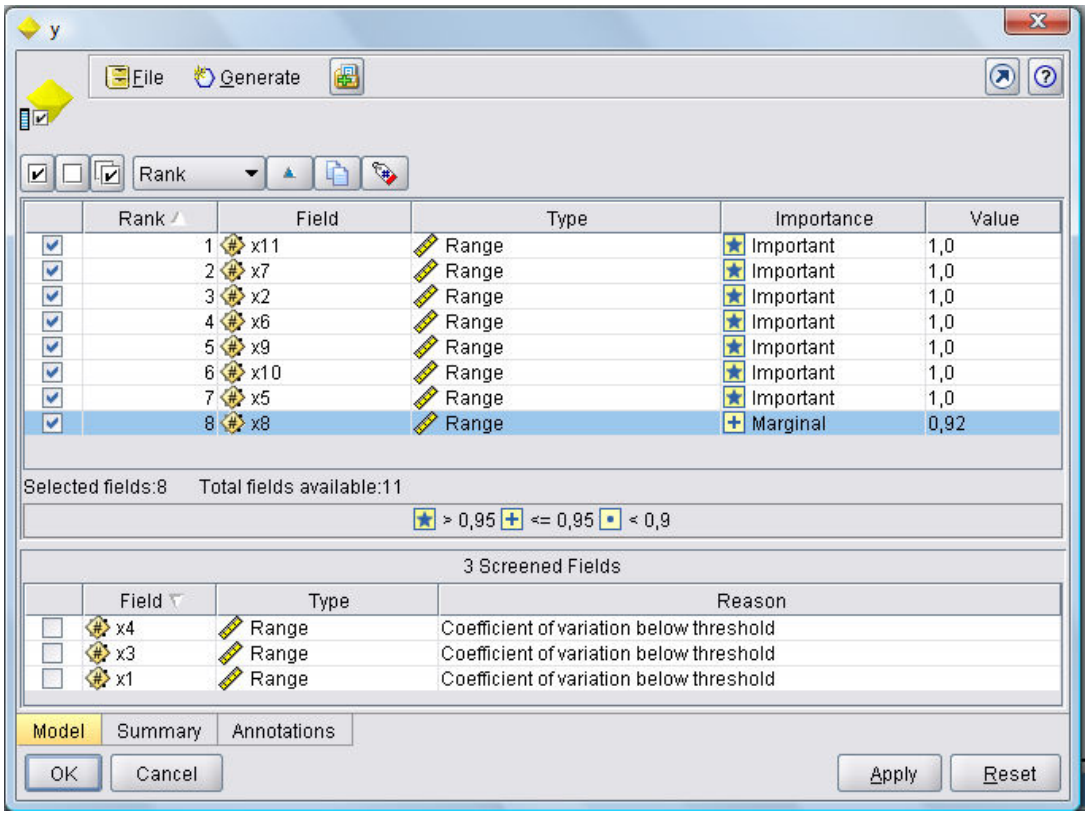
p value $=\text{Prob}({x_d}^2 \succ X^2)$ όπου το $X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\left(N_{ij} - \hat{N}_{ij}\right)^2 / \hat{N}_{ij}$.

Όπου το X η μεταβλητή πρόβλεψης με την θεώρηση ότι έχουμε I κατηγορίες
Y η μεταβλητή στόχος με J κατηγορίες
N ο συνολικός αριθμός των περιπτώσεων
N_{ij} ο αριθμός των περιπτώσεων όπου $X = i$ και $Y = j$

$N_{i.}$ ο αριθμός των περιπτώσεων όπου $X = i$ και $N_{i.} = \sum_{j=1}^{J} N_{ij}$

$N_{.j}$ ο αριθμός των περιπτώσεων όπου $Y = j$ και $N_{.j} = \sum_{i=1}^{I} N_{ij}$

Feature Selection Options tab



Μετά από όλη αυτή την διαδικασία επιλογής μεταβλητών παρατηρούμε ότι οι επεξηγηματικές μου μεταβλητές από 11 που ήταν αρχικά μειώνονται στις 8.
Για επίπεδο σημαντικότητας α=0.05 οι σημαντικές 8 αυτές μεταβλητές με p value =1,0 παρουσιάζονται παραπάνω.

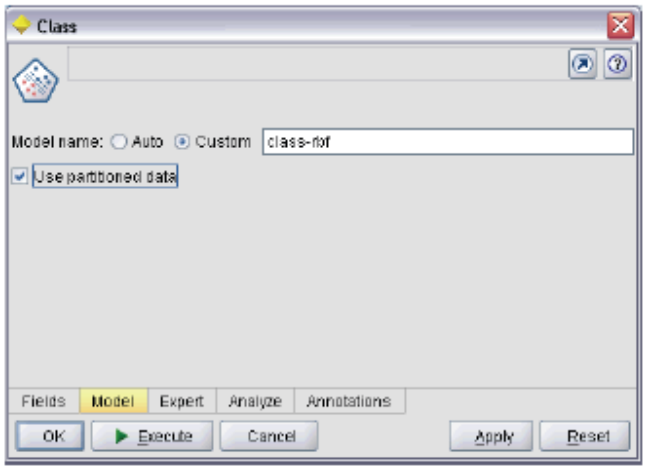
6. Συνδέουμε το Partition node στο Feature selection model  .

Ο κόμβος SVM προσφέρει τη δυνατότητα επιλογής των λειτουργιών του πυρήνα για την εκτέλεση της επεξεργασίας. Δεδομένου ότι δεν υπάρχει εύκολος τρόπος για να γνωρίζει ποια λειτουργία αποδίδει καλύτερα, θα επιλέξουμε διαφορετικές λειτουργίες και θα συγκρίνουμε τα αποτελέσματα. Ας αρχίσουμε με την προεπιλογή, RBF (ακτινική συνάρτηση).

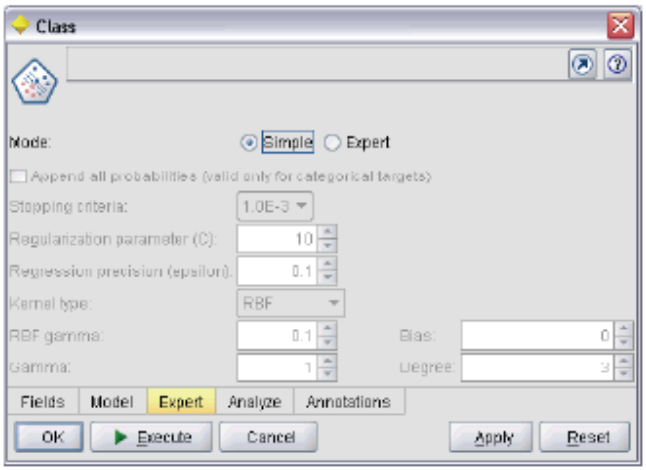
Τοποθέτηση στο stream canvas ενός SVM node, συνδέουμε τον κόμβο SVM στον κόμβο τύπου. Feature selection model. Ανοίγουμε τον κόμβο SVM. Στην καρτέλα model name, κάντε κλικ στην επιλογή Custom για το όνομα του μοντέλου και την επιλογή τύπου RBF.

Ρύθμιση των επιλογών του μοντέλου. Ρυθμίζοντας τις simple επιλογές .

Model tab settings



Expert tab settings for Simple mode

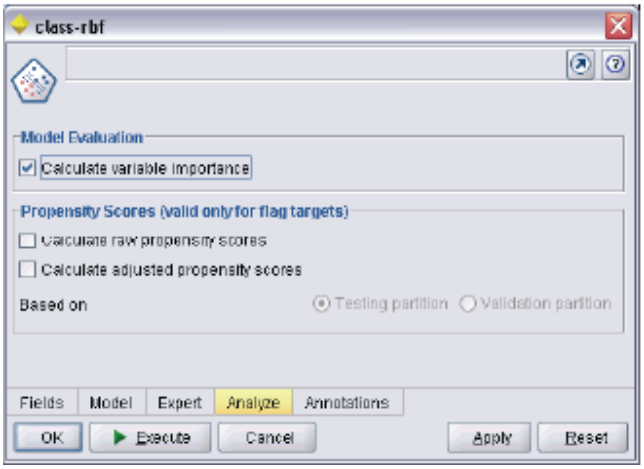


Στην καρτέλα expert, αφήνουμε τη ρύθμιση Λειτουργίας ως simple και σημειώνουμε ότι ο τύπος του πυρήνα έχει οριστεί σε RBF από προεπιλογή.

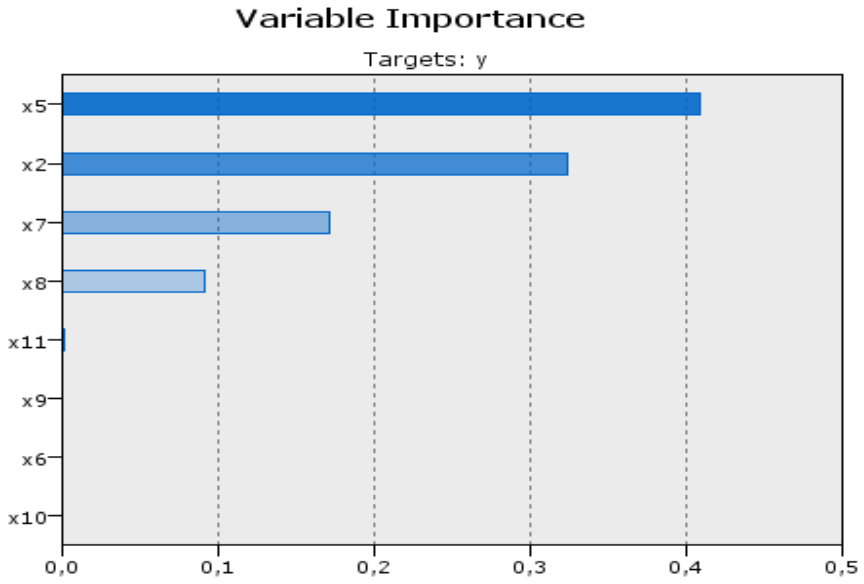
Στην καρτέλα Ανάλυση, επιλέγουμε το calculate variable importance και τα prospensity scores για τα flag.

Στη συνέχεια επιλέγουμε εκτέλεση (execute).

Analyze tab settings



Το παρακάτω γράφημα σημαντικότητας μεταβλητών δείχνει την σχετική επίδραση των διαφόρων πεδίων και μεταβλητών(επεξηγηματικών) στην πρόβλεψη(απόκρισης) y. το γράφημα αυτό μας δείχνει ότι η μεταβλητή x_5 έχει τη μεγαλύτερη επίδραση ενώ οι μεταβλητές x_2, x_7 και x_8 είναι επίσης αρκετά σημαντικές. Η μεταβλητή x_{11} είναι που επηρεάζει λιγότερο την πρόβλεψη.



X_5	0,41
X_2	0,324
X_7	0,172

X_8	0,092
X_{11}	0,002
X_9	0
X_6	0
X_{10}	0

Στη συνέχεια συνδέουμε τον κόμβο table και εκτελούμε.

Table (14 fields, 10.333 records) #1

File Edit Generate

	y	x2	x5	x6	x7	x8	x9	x10	x11	Partition	$S-y	$SP-y	$SRP-y	$SAP-y
1	0....	1....	3....	153....	160....	95.0...	22....	38.1...	47....	1_Training	1.0...	0.875	0.875	0.865
2	0....	2....	3....	75.0...	88.0...	314....	22....	38.1...	47....	1_Training	0.0...	0.661	0.339	0.252
3	0....	2....	3....	74.0...	88.0...	317....	22....	38.1...	47....	1_Training	0.0...	0.660	0.340	0.253
4	0....	2....	3....	67.0...	81.0...	295....	22....	38.1...	47....	2_Testing	0.0...	0.697	0.303	0.228
5	0....	2....	3....	127....	136....	308....	22....	38.1...	47....	1_Training	1.0...	0.627	0.627	0.571
6	0....	2....	3....	66.0...	81.0...	311....	22....	38.1...	47....	1_Training	0.0...	0.698	0.302	0.227
7	0....	3....	3....	125....	133....	180....	22....	38.1...	47....	1_Training	1.0...	0.692	0.692	0.662
8	0....	3....	3....	126....	134....	192....	22....	38.1...	47....	1_Training	1.0...	0.686	0.686	0.653
9	0....	3....	3....	135....	143....	192....	22....	38.1...	47....	1_Training	1.0...	0.731	0.731	0.714
10	0....	3....	3....	120....	129....	197....	22....	38.1...	47....	1_Training	1.0...	0.655	0.655	0.611
11	0....	4....	4....	52.0...	70.0...	39.0...	22....	38.1...	47....	1_Training	1.0...	0.699	0.699	0.672
12	0....	4....	4....	54.0...	72.0...	64.0...	22....	38.1...	47....	1_Training	1.0...	0.671	0.671	0.633
13	0....	5....	3....	180....	186....	154....	22....	38.1...	47....	2_Testing	1.0...	0.900	0.900	0.883
14	0....	5....	3....	129....	137....	162....	22....	38.1...	47....	1_Training	1.0...	0.736	0.736	0.720
15	0....	3....	3....	118....	127....	183....	22....	38.1...	47....	2_Testing	1.0...	0.657	0.657	0.614
16	0....	3....	3....	128....	136....	196....	22....	38.1...	47....	2_Testing	1.0...	0.693	0.693	0.663
17	0....	3....	3....	106....	116....	196....	22....	38.1...	47....	1_Training	1.0...	0.581	0.581	0.507
18	0....	3....	3....	105....	115....	181....	22....	38.1...	47....	1_Training	1.0...	0.590	0.590	0.520
19	0....	6....	3....	69.0...	84.0...	106....	22....	38.1...	47....	1_Training	1.0...	0.517	0.517	0.425
20	0....	6....	3....	44.0...	64.0...	95.0...	22....	38.1...	47....	1_Training	0.0...	0.588	0.412	0.312
21	0....	6....	3....	87.0...	99.0...	108....	22....	38.1...	47....	1_Training	1.0...	0.604	0.604	0.540
22	0....	6....	3....	86.0...	98.0...	106....	22....	38.1...	47....	2_Testing	1.0...	0.601	0.601	0.536
23	0....	6....	3....	90.0...	101....	113....	22....	38.1...	47....	1_Training	1.0...	0.608	0.608	0.544
24	0....	6....	3....	101....	111....	113....	22....	38.1...	47....	1_Training	1.0...	0.664	0.664	0.624
25	0....	2....	3....	65.0...	80.0...	238....	22....	38.1...	47....	1_Training	0.0...	0.514	0.486	0.388
26	0....	2....	3....	75.0...	88.0...	243....	22....	38.1...	47....	1_Training	1.0...	0.529	0.529	0.439
27	0....	2....	3....	74.0...	87.0...	250....	22....	38.1...	47....	2_Testing	1.0...	0.523	0.523	0.432
28	0....	2....	3....	51.0...	70.0...	230....	22....	38.1...	47....	2_Testing	0.0...	0.565	0.435	0.334
29	0....	2....	3....	71.0...	85.0...	209....	22....	38.1...	47....	1_Training	1.0...	0.520	0.520	0.429
30	0....	2....	3....	86.0...	98.0...	190....	22....	38.1...	47....	1_Training	1.0...	0.599	0.599	0.532
31	0....	1....	4....	22.0...	52.0...	252....	22....	38.1...	47....	1_Training	0.0...	0.536	0.464	0.364
32	0....	1....	3....	31.0...	56.0...	247....	22....	38.1...	47....	1_Training	0.0...	0.712	0.288	0.218
33	0....	1....	3....	45.0...	65.0...	239....	22....	38.1...	47....	1_Training	0.0...	0.573	0.427	0.326

Έχουν δημιουργηθεί τέσσερα πεδία:

1) S-y είναι η τιμή για το y, προβλεπόμενο από το μοντέλο (εκτιμώμενο), όταν το S-y συμπίπτει με το πραγματικό y έχει γίνει σωστά η πρόβλεψη μας.

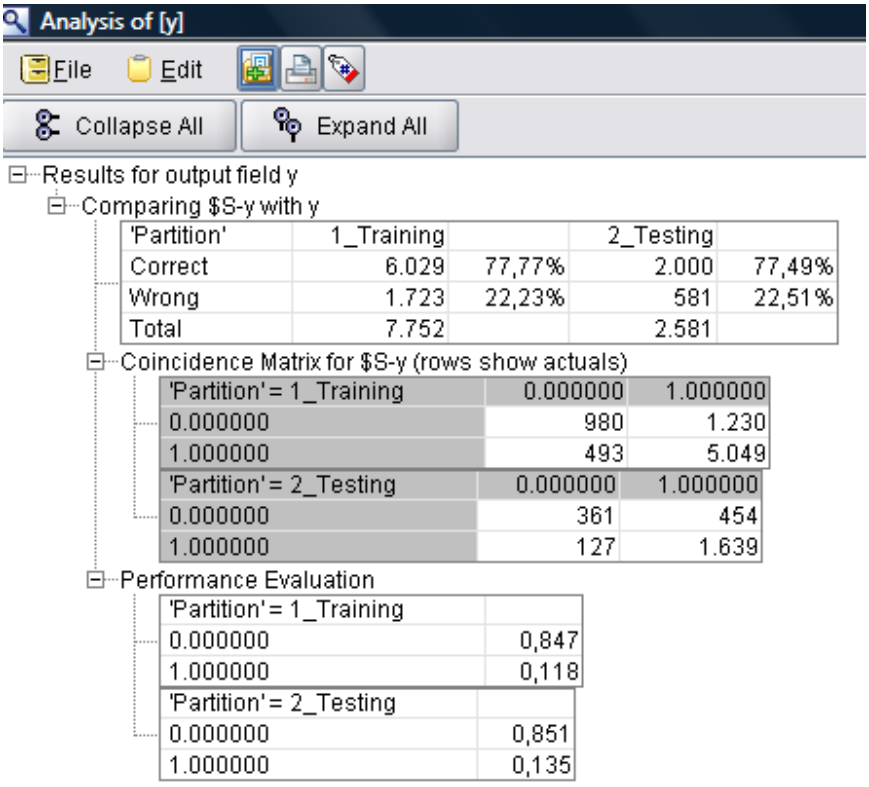
2) SP-y είναι η πιθανότητα η οποία δεδομένων των μεταβλητών που μας δίνονται μας δίνει 0 ή 1. Δηλαδή η πιθανότητα της πρόβλεψης να είναι σημαντική (0-1), κοντά στο 0,5 είναι χαμηλή και μη αποδεκτή.

3) SRP-y αφορά το raw propensity score

4) SAP-y αφορά το adjusted propensity score

Συγκρίνοντας το y με το S-y βλέπουμε αν το μοντέλο έχει κάνει σωστές ή όχι προβλέψεις ανεξάρτητα αν το propensity score είναι σχετικά υψηλό(μπορεί το propensity score να είναι υψηλό αλλά να έχουμε λάθος πρόβλεψη)

Επειδή έχουμε χαμηλά propensity scores, θα βάλουμε Analysis και άρα πρέπει να δοκιμάσουμε κι άλλους kernels.



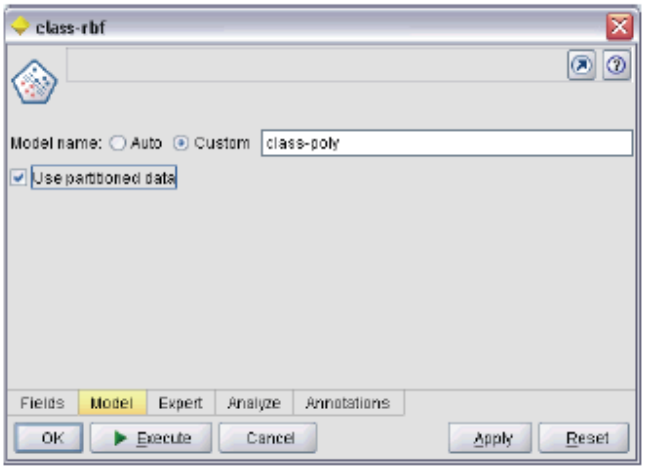
'Partition'	1_Training		2_Testing	
Correct	6.029	77,77%	2.000	77,49%
Wrong	1.723	22,23%	581	22,51%
Total	7.752		2.581	

'Partition' = 1_Training	0.000000	1.000000
0.000000	980	1.230
1.000000	493	5.049
'Partition' = 2_Testing	0.000000	1.000000
0.000000	361	454
1.000000	127	1.639

'Partition' = 1_Training	
0.000000	0,847
1.000000	0,118
'Partition' = 2_Testing	
0.000000	0,851
1.000000	0,135

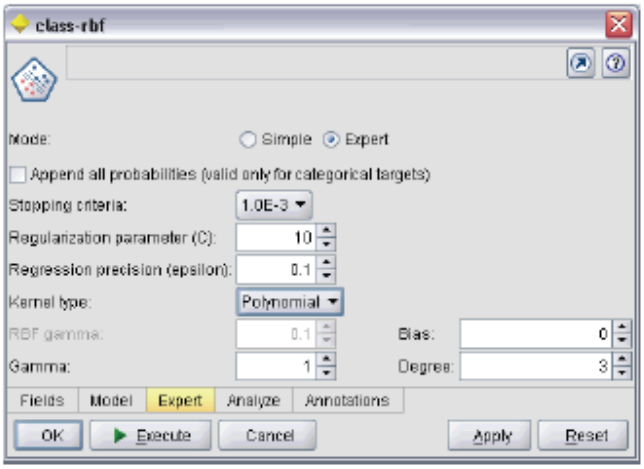
Ρύθμιση των επιλογών του μοντέλου. Ρυθμίζοντας τις expert επιλογές .

Setting a new name for the model



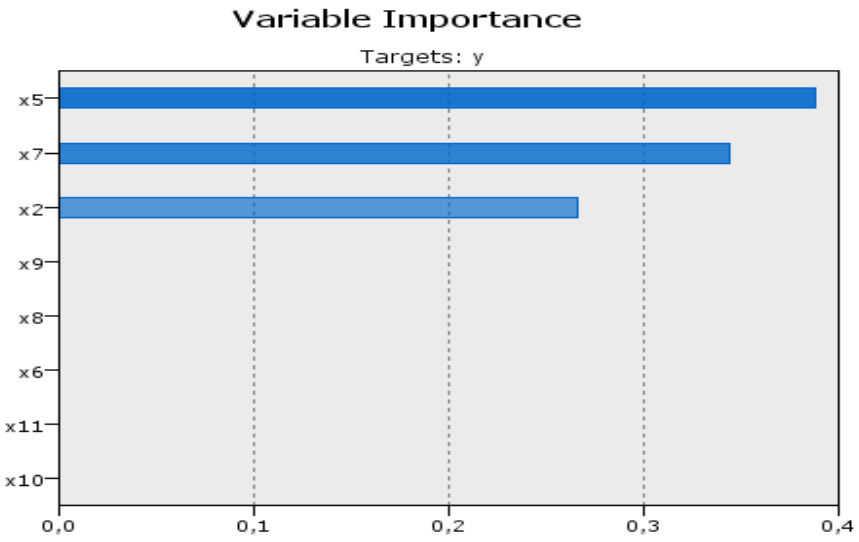
- Polynomial

Expert tab settings for Polynomial



Παρατηρούμε ότι υπάρχουν stopping criteria και ότι RBF gamma είναι ίσο με 0,1, ίδιο με το simple. Στην polynomial το gamma είναι ίσο με 1.

Το παρακάτω γράφημα σημαντικότητας μεταβλητών δείχνει την σχετική επίδραση των διαφόρων πεδίων και μεταβλητών(επεξηγηματικών) στην πρόβλεψη(απόκρισης) y. το γράφημα αυτό μας δείχνει ότι η μεταβλητή x_5 έχει τη μεγαλύτερη επίδραση ενώ οι μεταβλητές x_7, x_2 είναι επίσης αρκετά σημαντικές.



X_5	0,389
X_7	0,345
X_2	0,267

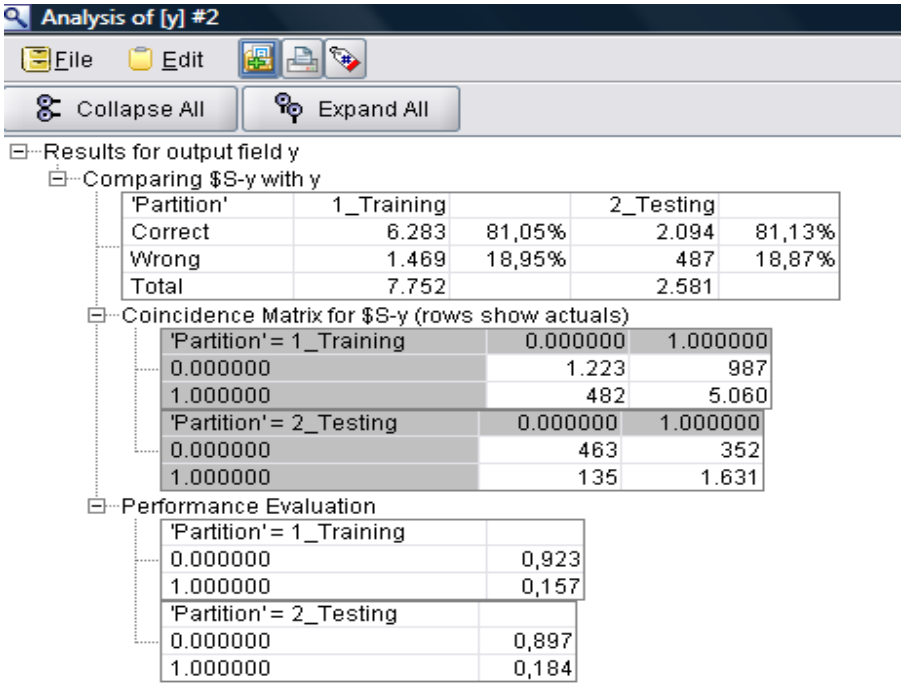
Εργαζόμαστε όπως και στην simple περίπτωση με κόμβο table.

Table (14 fields, 10.333 records) #3

File Edit Generate

	y	x2	x5	x6	x7	x8	x9	x10	x11	Partition	$S-y	$SP-y	$SRP-y	$SAP-y
1	0....	1....	3....	153....	160....	95.0...	22....	38.1...	47....	1_Training	1.0...	0.735	0.735	0.725
2	0....	2....	3....	75.0...	88.0...	314....	22....	38.1...	47....	1_Training	0.0...	0.763	0.237	0.188
3	0....	2....	3....	74.0...	88.0...	317....	22....	38.1...	47....	1_Training	0.0...	0.755	0.245	0.192
4	0....	2....	3....	67.0...	81.0...	295....	22....	38.1...	47....	2_Testing	0.0...	0.855	0.145	0.145
5	0....	2....	3....	127....	136....	308....	22....	38.1...	47....	1_Training	1.0...	0.635	0.635	0.591
6	0....	2....	3....	66.0...	81.0...	311....	22....	38.1...	47....	1_Training	0.0...	0.824	0.176	0.158
7	0....	3....	3....	125....	133....	180....	22....	38.1...	47....	1_Training	0.0...	0.563	0.437	0.342
8	0....	3....	3....	126....	134....	192....	22....	38.1...	47....	1_Training	0.0...	0.565	0.435	0.340
9	0....	3....	3....	135....	143....	192....	22....	38.1...	47....	1_Training	0.0...	0.505	0.495	0.406
10	0....	3....	3....	120....	129....	197....	22....	38.1...	47....	1_Training	0.0...	0.604	0.396	0.302
11	0....	4....	4....	52.0...	70.0...	39.0...	22....	38.1...	47....	1_Training	1.0...	0.568	0.568	0.499
12	0....	4....	4....	54.0...	72.0...	64.0...	22....	38.1...	47....	1_Training	0.0...	0.503	0.497	0.408
13	0....	5....	3....	180....	186....	154....	22....	38.1...	47....	2_Testing	1.0...	0.730	0.730	0.719
14	0....	5....	3....	129....	137....	162....	22....	38.1...	47....	1_Training	0.0...	0.509	0.491	0.402
15	0....	3....	3....	118....	127....	183....	22....	38.1...	47....	2_Testing	0.0...	0.609	0.391	0.297
16	0....	3....	3....	128....	136....	196....	22....	38.1...	47....	2_Testing	0.0...	0.553	0.447	0.352
17	0....	3....	3....	106....	116....	196....	22....	38.1...	47....	1_Training	0.0...	0.696	0.304	0.229
18	0....	3....	3....	105....	115....	181....	22....	38.1...	47....	1_Training	0.0...	0.693	0.307	0.231
19	0....	6....	3....	69.0...	84.0...	106....	22....	38.1...	47....	1_Training	0.0...	0.762	0.238	0.188
20	0....	6....	3....	44.0...	64.0...	95.0...	22....	38.1...	47....	1_Training	0.0...	0.867	0.133	0.140
21	0....	6....	3....	87.0...	99.0...	108....	22....	38.1...	47....	1_Training	0.0...	0.653	0.347	0.260
22	0....	6....	3....	86.0...	98.0...	106....	22....	38.1...	47....	2_Testing	0.0...	0.655	0.345	0.259
23	0....	6....	3....	90.0...	101....	113....	22....	38.1...	47....	1_Training	0.0...	0.650	0.350	0.263
24	0....	6....	3....	101....	111....	113....	22....	38.1...	47....	1_Training	0.0...	0.575	0.425	0.330
25	0....	2....	3....	65.0...	80.0...	238....	22....	38.1...	47....	1_Training	0.0...	0.681	0.319	0.239
26	0....	2....	3....	75.0...	88.0...	243....	22....	38.1...	47....	1_Training	0.0...	0.626	0.374	0.282
27	0....	2....	3....	74.0...	87.0...	250....	22....	38.1...	47....	2_Testing	0.0...	0.619	0.381	0.288
28	0....	2....	3....	51.0...	70.0...	230....	22....	38.1...	47....	2_Testing	0.0...	0.745	0.255	0.197
29	0....	2....	3....	71.0...	85.0...	209....	22....	38.1...	47....	1_Training	0.0...	0.677	0.323	0.242
30	0....	2....	3....	86.0...	98.0...	190....	22....	38.1...	47....	1_Training	0.0...	0.598	0.402	0.307
31	0....	1....	4....	22.0...	52.0...	252....	22....	38.1...	47....	1_Training	0.0...	0.781	0.219	0.178
32	0....	1....	3....	31.0...	56.0...	247....	22....	38.1...	47....	1_Training	0.0...	0.912	0.088	0.125
33	0....	1....	3....	45.0...	65.0...	239....	22....	38.1...	47....	1_Training	0.0...	0.814	0.186	0.162

Παλί έχουν προκύψει τέσσερα νέα πεδία και στη συνέχεια προχωράμε σε analysis



'Partition'	1_Training		2_Testing	
Correct	6.283	81,05%	2.094	81,13%
Wrong	1.469	18,95%	487	18,87%
Total	7.752		2.581	

'Partition' = 1_Training	0.000000	1.000000
0.000000	1.223	987
1.000000	482	5.060

'Partition' = 2_Testing	0.000000	1.000000
0.000000	463	352
1.000000	135	1.631

'Partition' = 1_Training	
0.000000	0,923
1.000000	0,157

'Partition' = 2_Testing	
0.000000	0,897
1.000000	0,184

Με βάση τον παρακάτω συγκεντρωτικό πίνακα παρατηρούμε ότι ο kernel Polynomial είναι πολύ καύτερος απ'ότι ο RBF.

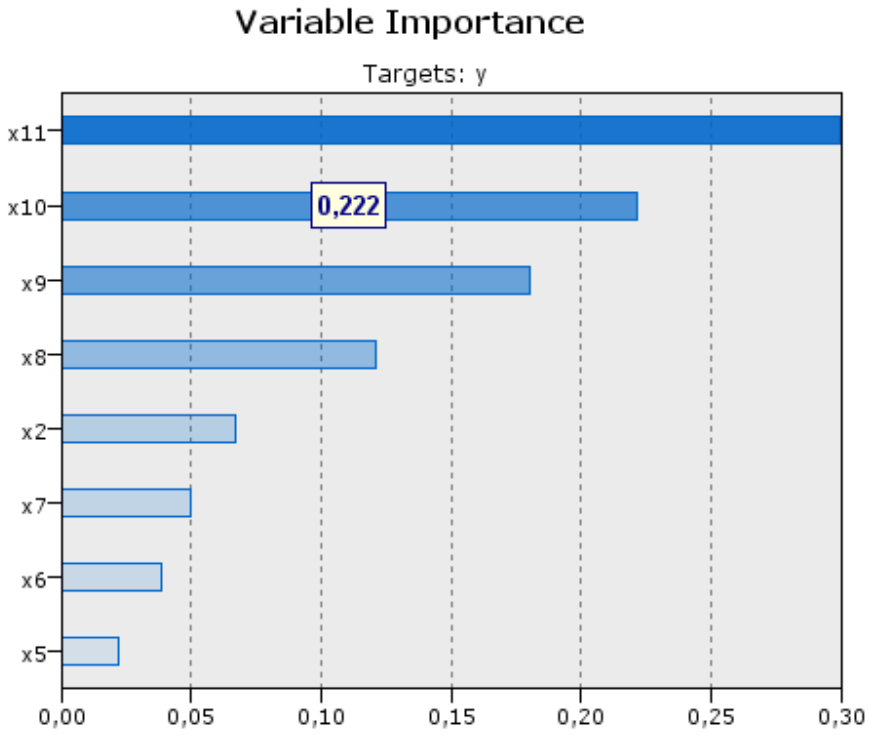
Correct	Training	Testing
RBF	77,77%	77,49%
Polynomial	81,05%	81,03%

Συνεχίζουμε τεστάροντας και τις υπόλοιπες περιπτώσεις στο expert Model

- Sigmoid

Παρατηρούμε ότι το gamma είναι ίσο με 1.

Το παρακάτω γράφημα σημαντικότητας μεταβλητών δείχνει την σχετική επίδραση των διαφόρων πεδίων και μεταβλητών(επεξηγηματικών) στην πρόβλεψη(απόκρισης) y. το γράφημα αυτό μας δείχνει ότι η μεταβλητή x_{11} έχει τη μεγαλύτερη επίδραση ενώ οι μεταβλητές x_{10}, x_9, x_8, x_2, x_7, x_6, x_5 είναι επίσης αρκετά σημαντικές.



X_{11}	0,299
X_{10}	0,222
X_9	0,181
X_8	0,121
X_2	0,067
X_7	0,01
X_6	0,038
X_5	0,022

Συνεχίζουμε με table και analysis

Table (14 fields, 10.333 records) #4

File Edit Generate

	y	x2	x5	x6	x7	x8	x9	x10	x11	Partition	$S-y	$SP-y	$SRP-y	$SAP-y
1	0....	1....	3....	153....	160....	95.0...	22....	38.1...	47....	1_Training	1.0...	0.759	0.759	0.698
2	0....	2....	3....	75.0...	88.0...	314....	22....	38.1...	47....	1_Training	1.0...	0.739	0.739	0.691
3	0....	2....	3....	74.0...	88.0...	317....	22....	38.1...	47....	1_Training	1.0...	0.739	0.739	0.691
4	0....	2....	3....	67.0...	81.0...	295....	22....	38.1...	47....	2_Testing	1.0...	0.736	0.736	0.689
5	0....	2....	3....	127....	136....	308....	22....	38.1...	47....	1_Training	1.0...	0.751	0.751	0.695
6	0....	2....	3....	66.0...	81.0...	311....	22....	38.1...	47....	1_Training	1.0...	0.736	0.736	0.690
7	0....	3....	3....	125....	133....	180....	22....	38.1...	47....	1_Training	1.0...	0.750	0.750	0.694
8	0....	3....	3....	126....	134....	192....	22....	38.1...	47....	1_Training	1.0...	0.750	0.750	0.694
9	0....	3....	3....	135....	143....	192....	22....	38.1...	47....	1_Training	1.0...	0.752	0.752	0.695
10	0....	3....	3....	120....	129....	192.0	22....	38.1...	47....	1_Training	1.0...	0.748	0.748	0.694
11	0....	4....	4....	52.0...	70.0...	39.0...	22....	38.1...	47....	1_Training	1.0...	0.755	0.755	0.696
12	0....	4....	4....	54.0...	72.0...	64.0...	22....	38.1...	47....	1_Training	1.0...	0.752	0.752	0.695
13	0....	5....	3....	180....	186....	154....	22....	38.1...	47....	2_Testing	1.0...	0.766	0.766	0.700
14	0....	5....	3....	129....	137....	162....	22....	38.1...	47....	1_Training	1.0...	0.754	0.754	0.696
15	0....	3....	3....	118....	127....	183....	22....	38.1...	47....	2_Testing	1.0...	0.748	0.748	0.694
16	0....	3....	3....	128....	136....	196....	22....	38.1...	47....	2_Testing	1.0...	0.750	0.750	0.695
17	0....	3....	3....	106....	116....	196....	22....	38.1...	47....	1_Training	1.0...	0.744	0.744	0.693
18	0....	3....	3....	105....	115....	181....	22....	38.1...	47....	1_Training	1.0...	0.744	0.744	0.693
19	0....	6....	3....	69.0...	84.0...	106....	22....	38.1...	47....	1_Training	1.0...	0.741	0.741	0.691
20	0....	6....	3....	44.0...	64.0...	95.0...	22....	38.1...	47....	1_Training	1.0...	0.734	0.734	0.689
21	0....	6....	3....	87.0...	99.0...	108....	22....	38.1...	47....	1_Training	1.0...	0.746	0.746	0.693
22	0....	6....	3....	86.0...	98.0...	106....	22....	38.1...	47....	2_Testing	1.0...	0.746	0.746	0.693
23	0....	6....	3....	90.0...	101....	113....	22....	38.1...	47....	1_Training	1.0...	0.746	0.746	0.693
24	0....	6....	3....	101....	111....	113....	22....	38.1...	47....	1_Training	1.0...	0.750	0.750	0.694
25	0....	2....	3....	65.0...	80.0...	238....	22....	38.1...	47....	1_Training	1.0...	0.757	0.757	0.697
26	0....	2....	3....	75.0...	88.0...	243....	22....	38.1...	47....	1_Training	1.0...	0.760	0.760	0.698
27	0....	2....	3....	74.0...	87.0...	250....	22....	38.1...	47....	2_Testing	1.0...	0.760	0.760	0.698
28	0....	2....	3....	51.0...	70.0...	230....	22....	38.1...	47....	2_Testing	1.0...	0.754	0.754	0.696
29	0....	2....	3....	71.0...	85.0...	209....	22....	38.1...	47....	1_Training	1.0...	0.758	0.758	0.697
30	0....	2....	3....	86.0...	98.0...	190....	22....	38.1...	47....	1_Training	1.0...	0.761	0.761	0.698
31	0....	1....	4....	22.0...	52.0...	252....	22....	38.1...	47....	1_Training	1.0...	0.752	0.752	0.695
32	0....	1....	3....	31.0...	56.0...	247....	22....	38.1...	47....	1_Training	1.0...	0.741	0.741	0.691
33	0....	1....	3....	45.0...	65.0...	239....	22....	38.1...	47....	1_Training	1.0...	0.750	0.750	0.695

Analysis of [y] #3

File Edit

Collapse All Expand All

Results for output field y

Comparing $S-y with y

'Partition'	1_Training		2_Testing	
Correct	5.508	71,05%	1.761	68,23%
Wrong	2.244	28,95%	820	31,77%
Total	7.752		2.581	

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

'Partition' = 1_Training	0.000000	1.000000
0.000000	11	2.199
1.000000	45	5.497

'Partition' = 2_Testing	0.000000	1.000000
0.000000	6	809
1.000000	11	1.755

Performance Evaluation

'Partition' = 1_Training	
0.000000	-0,117
1.000000	-0,002

'Partition' = 2_Testing	
0.000000	0,111
1.000000	0,0

Και παρατηρούμε ότι οι τιμές είναι χαμηλότερες από τους προηγούμενους δύο kernels.

Συγκεντρωτικά έχουμε:

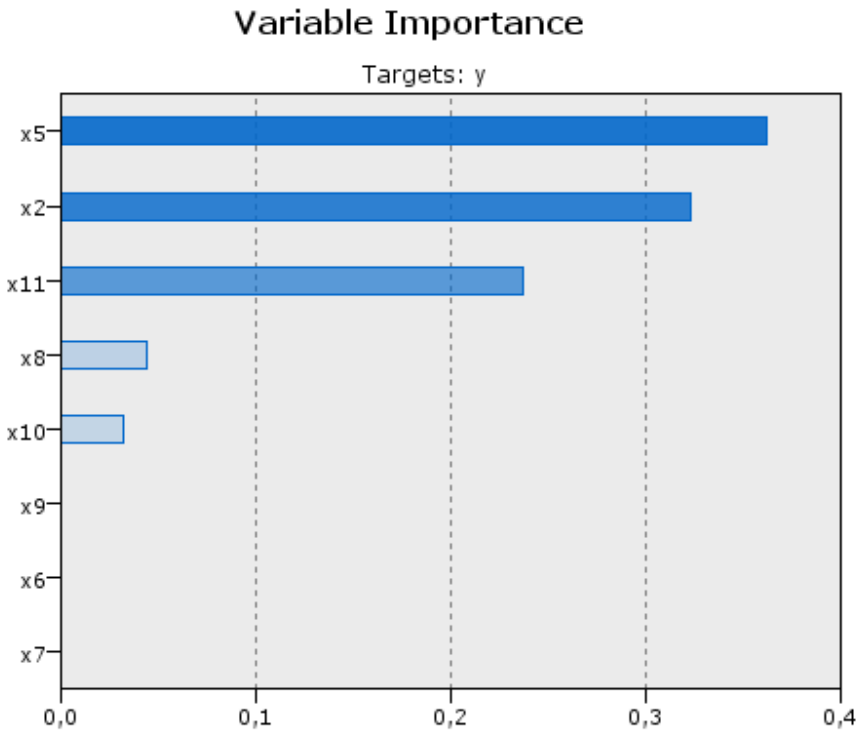
Correct	Training	Testing
RBF	77,77%	77,49%
Polynomial	81,05%	81,03%
Sigmoid	71,05%	68,23%

Συνεχίζουμε με τον τελευταίο kernel.

- Linear

Παρατηρούμε ότι το gamma είναι ίσο με 1.

Το παρακάτω γράφημα σημαντικότητας μεταβλητών δείχνει την σχετική επίδραση των διαφόρων πεδίων και μεταβλητών(επεξηγηματικών) στην πρόβλεψη(απόκρισης) y. το γράφημα αυτό μας δείχνει ότι η μεταβλητή x_5 έχει τη μεγαλύτερη επίδραση ενώ οι μεταβλητές x_2, x_{11}, x_8, x_{10} είναι επίσης αρκετά σημαντικές.



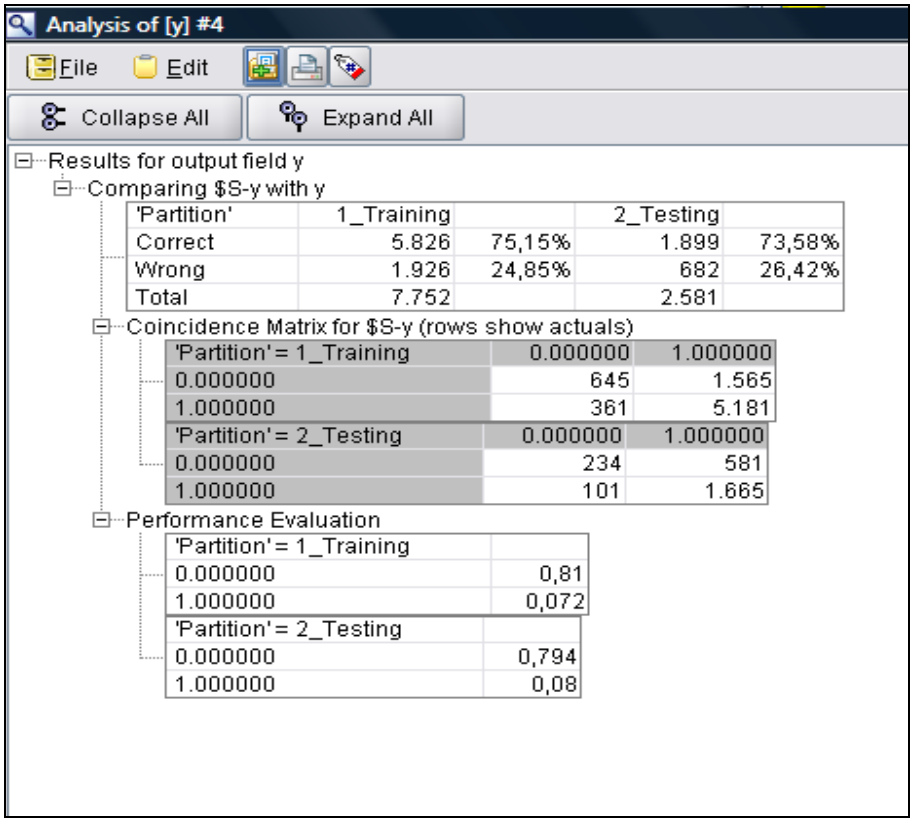
X_5	0,362

X_2	0,323
X_{11}	0,238
X_8	0,044
X_{10}	0,033

Table (14 fields, 10.333 records) #5

File Edit Generate

	y	x2	x5	x6	x7	x8	x9	x10	x11	Partition	$S-y	$SP-y	$SRP-y	$SAP-y
1	0....	1....	3....	153....	160....	95.0...	22....	38.1...	47....	1_Training	1.0...	0.812	0.812	0.803
2	0....	2....	3....	75.0...	88.0...	314....	22....	38.1...	47....	1_Training	1.0...	0.599	0.599	0.556
3	0....	2....	3....	74.0...	88.0...	317....	22....	38.1...	47....	1_Training	1.0...	0.615	0.615	0.578
4	0....	2....	3....	67.0...	81.0...	295....	22....	38.1...	47....	2_Testing	1.0...	0.592	0.592	0.546
5	0....	2....	3....	127....	136....	308....	22....	38.1...	47....	1_Training	1.0...	0.726	0.726	0.716
6	0....	2....	3....	66.0...	81.0...	311....	22....	38.1...	47....	1_Training	1.0...	0.604	0.604	0.563
7	0....	3....	3....	125....	133....	180....	22....	38.1...	47....	1_Training	1.0...	0.735	0.735	0.726
8	0....	3....	3....	126....	134....	192....	22....	38.1...	47....	1_Training	1.0...	0.736	0.736	0.727
9	0....	3....	3....	135....	143....	192....	22....	38.1...	47....	1_Training	1.0...	0.765	0.765	0.759
10	0....	3....	3....	120....	129....	197....	22....	38.1...	47....	1_Training	1.0...	0.731	0.731	0.722
11	0....	4....	4....	52.0...	70.0...	39.0...	22....	38.1...	47....	1_Training	1.0...	0.792	0.792	0.785
12	0....	4....	4....	54.0...	72.0...	64.0...	22....	38.1...	47....	1_Training	1.0...	0.794	0.794	0.786
13	0....	5....	3....	180....	186....	154....	22....	38.1...	47....	2_Testing	1.0...	0.869	0.869	0.848
14	0....	5....	3....	129....	137....	162....	22....	38.1...	47....	1_Training	1.0...	0.762	0.762	0.755
15	0....	3....	3....	118....	127....	183....	154.0	38.1...	47....	2_Testing	1.0...	0.727	0.727	0.718
16	0....	3....	3....	128....	136....	196....	22....	38.1...	47....	2_Testing	1.0...	0.742	0.742	0.734
17	0....	3....	3....	106....	116....	196....	22....	38.1...	47....	1_Training	1.0...	0.699	0.699	0.685
18	0....	3....	3....	105....	115....	181....	22....	38.1...	47....	1_Training	1.0...	0.699	0.699	0.685
19	0....	6....	3....	69.0...	84.0...	106....	22....	38.1...	47....	1_Training	1.0...	0.692	0.692	0.676
20	0....	6....	3....	44.0...	64.0...	95.0...	22....	38.1...	47....	1_Training	1.0...	0.694	0.694	0.679
21	0....	6....	3....	87.0...	99.0...	108....	22....	38.1...	47....	1_Training	1.0...	0.703	0.703	0.689
22	0....	6....	3....	86.0...	98.0...	106....	22....	38.1...	47....	2_Testing	1.0...	0.700	0.700	0.686
23	0....	6....	3....	90.0...	101....	113....	22....	38.1...	47....	1_Training	1.0...	0.695	0.695	0.679
24	0....	6....	3....	101....	111....	113....	22....	38.1...	47....	1_Training	1.0...	0.717	0.717	0.705
25	0....	2....	3....	65.0...	80.0...	238....	22....	38.1...	47....	1_Training	1.0...	0.723	0.723	0.713
26	0....	2....	3....	75.0...	88.0...	243....	22....	38.1...	47....	1_Training	1.0...	0.722	0.722	0.712
27	0....	2....	3....	74.0...	87.0...	250....	22....	38.1...	47....	2_Testing	1.0...	0.717	0.717	0.706
28	0....	2....	3....	51.0...	70.0...	230....	22....	38.1...	47....	2_Testing	1.0...	0.745	0.745	0.737
29	0....	2....	3....	71.0...	85.0...	209....	22....	38.1...	47....	1_Training	1.0...	0.733	0.733	0.724
30	0....	2....	3....	86.0...	98.0...	190....	22....	38.1...	47....	1_Training	1.0...	0.754	0.754	0.747
31	0....	1....	4....	22.0...	52.0...	252....	22....	38.1...	47....	1_Training	1.0...	0.865	0.865	0.845
32	0....	1....	3....	31.0...	56.0...	247....	22....	38.1...	47....	1_Training	1.0...	0.739	0.739	0.731
33	0....	1....	3....	45.0...	65.0...	239....	22....	38.1...	47....	1_Training	1.0...	0.756	0.756	0.749

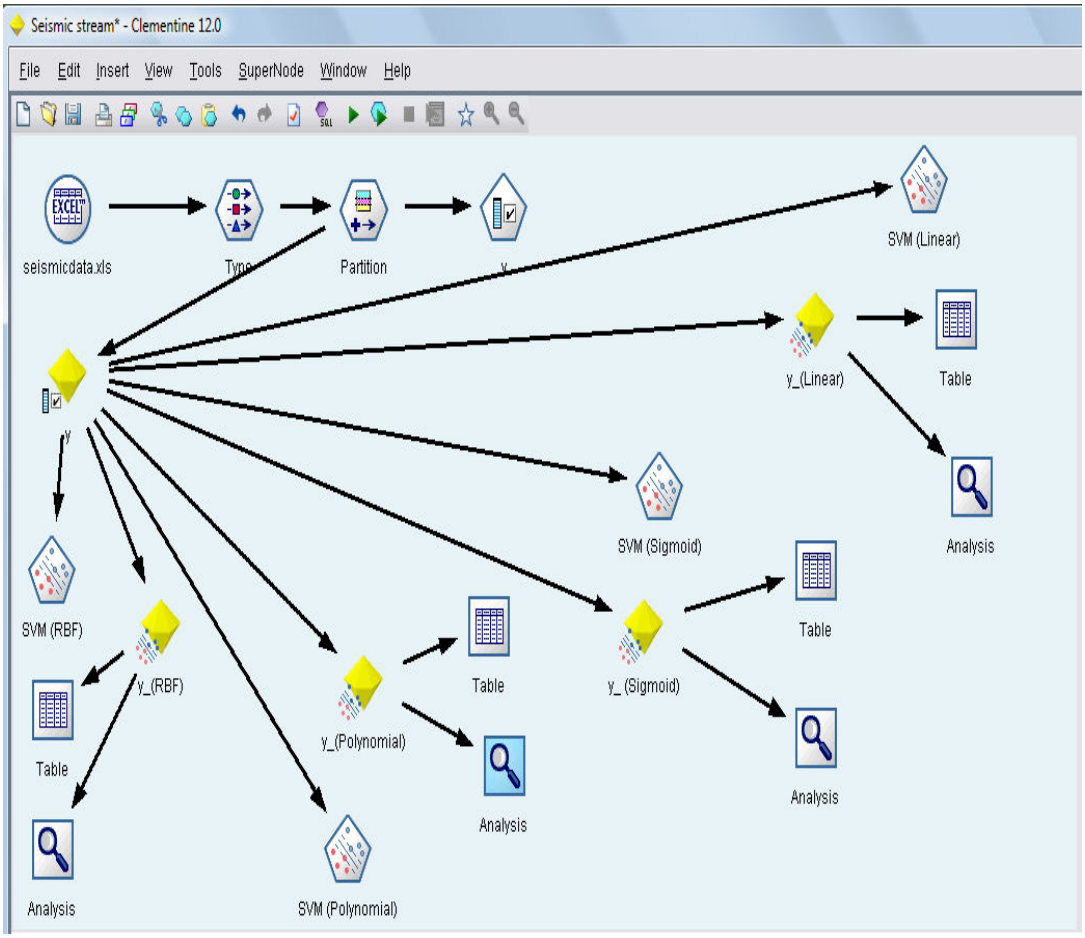


Και παρατηρούμε ότι οι τιμές είναι παρόμοιες με τον πυρήνα RBF.

Συγκεντρωτικά έχουμε:

Correct	Training	Testing
RBF	77,77%	77,49%
Polynomial	81,05%	81,03%
Sigmoid	71,05%	68,23%
Linear	75,15%	73,58%

Η τελική εικόνα του stream canvas μετά την εφαρμογή των αλγορίθμων με την βοήθεια του Clementine είναι :



Στη συνέχεια θα κάνουμε ταξινόμηση με Binary Classifier προκειμένου να συγκρίνουμε την μέθοδο SVM με τα νευρωνικά δίκτυα.

8.2.2 Μέτρα αξιολόγησης διαγνωστικών τεστ

Διαγνωστικό τεστ	**Αποτέλεσμα**	**Αποτέλεσμα**	
	Magnitude High (+)	Magnitude Min(-)	Σύνολο
Θετικό (+)	a (πραγματικά θετικά)	b (λανθασμένα θετικά)	a+b
Αρνητικό (-)	c (λανθασμένα αρνητικά)	d (πραγματικά αρνητικά)	c+d
Σύνολο	a+c	b+d	a+b+c+d

Όπου :

- (sensitivity) ευαισθησία = $\frac{a}{a+c}$
- (specificity) ειδικότητα = $\frac{d}{b+d}$
- θετική προγνωστική αξία (Θ.Π.Α) = $\frac{a}{a+b}$
- αρνητική προγνωστική αξία (Α.Π.Α) = $\frac{d}{d+c}$
- ακρίβεια = $\frac{a+d}{a+c+b+d}$

Θα υπολογίσουμε τα μέτρα αυτά αξιολόγησης στα σύνολα εκπαίδευσης (training data) και ελέγχου (test data) ξεχωριστά και θα τα απεικονίσουμε αναλυτικά σε μορφή ποσοστών .

1. **RBF kernel**

- Για το σύνολο εκπαίδευσης (training) στην simple mode με πυρήνα RBF έχω :

		0 (-)	1(+)
0	Magnitude Min(-)	980	1.230
1	Magnitude High (+)	493	5.049

Άρα a=(+ +)= πραγματικά θετικά = 5.049
b = (+ -) =λανθασμένα θετικά = 493
c =(- +) =λανθασμένα αρνητικά= 1.230
d = (- -) = πραγματικά αρνητικά= 980
και τα ποσοστά των μέτρων % είναι αναλυτικά :

- ευαισθησία = 80,41 (%)
- ειδικότητα = 66,53 (%)
- θετική προγνωστική αξία (Θ.Π.Α) = 91,10 (%)
- αρνητική προγνωστική αξία (Α.Π.Α)= 44,34 (%)
- ακρίβεια = 77,77 (%)

- Για το σύνολο ελέγχου (test) στην αντίστοιχη περίπτωση RBF simple mode έχω :

		0 (-)	1(+)
0	Magnitude Min (-)	361	454
1	Magnitude High (+)	127	1.639

Άρα a=(+ +)= πραγματικά θετικά = 1639
b = (+ -) =λανθασμένα θετικά = 127
c =(- +) =λανθασμένα αρνητικά= 454
d = (- -) = πραγματικά αρνητικά= 361

και τα ποσοστά των μέτρων % είναι αναλυτικά :

- ευαισθησία = 78,31 (%)
- ειδικότητα = 73,98 (%)
- θετική προγνωστική αξία (Θ.Π.Α) = 92,81 (%)
- αρνητική προγνωστική αξία (Α.Π.Α) = 44,29 (%)
- ακρίβεια = 77,49 (%)

1. Polynomial - expert mode

- Για το σύνολο εκπαίδευσης (training) στην expert mode με πυρήνα polynomial έχω :

		0 (-)	1(+)
0	Magnitude Min(-)	1.223	987
1	Magnitude High (+)	482	5.060

Άρα a=(+ +)= πραγματικά θετικά = 5.060
b = (+ -) =λανθασμένα θετικά = 482
c =(- +) =λανθασμένα αρνητικά= 987
d = (- -) = πραγματικά αρνητικά= 1.223
και τα ποσοστά των μέτρων % είναι αναλυτικά :

- ευαισθησία = 83,68 (%)
- ειδικότητα = 71,73 (%)
- θετική προγνωστική αξία (Θ.Π.Α) = 91,30 (%)
- αρνητική προγνωστική αξία (Α.Π.Α)= 55,34 (%)
- ακρίβεια = 81,05 (%)

- Για το σύνολο ελέγχου(test) στην αντίστοιχη περίπτωση polynomial expert mode έχω :

		0 (-)	1(+)
0	Magnitude Min (-)	463	352
1	Magnitude High (+)	135	1.631

Άρα a=(+ +)= πραγματικά θετικά = 1631
b = (+ -) =λανθασμένα θετικά = 135
c =(- +) =λανθασμένα αρνητικά= 352
d = (- -) = πραγματικά αρνητικά= 463

και τα ποσοστά των μέτρων % είναι αναλυτικά :

- ευαισθησία = 82,25 (%)
- ειδικότητα = 77,42 (%)
- θετική προγνωστική αξία (Θ.Π.Α) = 92,36 (%)
- αρνητική προγνωστική αξία (Α.Π.Α) = 56,81 (%)
- ακρίβεια = 81,13 (%)

2. Sigmoid - expert mode

- Για το σύνολο εκπαίδευσης (training) στην expert mode με πυρήνα sigmoid έχω :

		0 (-)	1(+)
0	Magnitude Min(-)	11	2.199
1	Magnitude High (+)	45	5.497

Άρα a=(+ +)= πραγματικά θετικά = 5.497
b = (+ -) =λανθασμένα θετικά = 45
c =(- +) =λανθασμένα αρνητικά= 2.199
d = (- -) = πραγματικά αρνητικά= 11

και τα ποσοστά των μέτρων % είναι αναλυτικά :

- ευαισθησία = 71,43 (%)
- ειδικότητα = 19,64 (%)
- θετική προγνωστική αξία (Θ.Π.Α) = 99,19 (%)
- αρνητική προγνωστική αξία (Α.Π.Α)= 0,50 (%)
- ακρίβεια = 71,05 (%)

- Για το σύνολο ελέγχου(test) στην αντίστοιχη περίπτωση sigmoid expert mode έχω :

		0 (-)	1(+)
0	Magnitude Min (-)	6	809
1	Magnitude High (+)	11	1.755

Άρα a=(+ +)= πραγματικά θετικά = 1.755
b = (+ -) =λανθασμένα θετικά = 11
c =(- +) =λανθασμένα αρνητικά= 809
d = (- -) = πραγματικά αρνητικά= 6

και τα ποσοστά των μέτρων % είναι αναλυτικά :

- ευαισθησία = 68,45 (%)
- ειδικότητα = 35,29 (%)
- θετική προγνωστική αξία (Θ.Π.Α) = 99,38 (%)
- αρνητική προγνωστική αξία (Α.Π.Α) = 0,74 (%)

- ακρίβεια = 68,23 (%)

3. Linear - expert mode

- Για το σύνολο εκπαίδευσης (training) στην expert mode με πυρήνα linear έχω :

		0 (-)	1(+)
0	Magnitude Min(-)	645	1.565
1	Magnitude High (+)	361	5.181

Άρα a=(+ +)= πραγματικά θετικά = 5.181
b = (+ -) =λανθασμένα θετικά = 361
c =(- +) =λανθασμένα αρνητικά= 1.565
d = (- -) = πραγματικά αρνητικά= 645
και τα ποσοστά των μέτρων % είναι αναλυτικά :

- ευαισθησία = 76,80 (%)
- ειδικότητα = 64,12 (%)
- θετική προγνωστική αξία (Θ.Π.Α) = 93,49 (%)
- αρνητική προγνωστική αξία (Α.Π.Α)= 29,19 (%)
- ακρίβεια = 75,15 (%)

- Για το σύνολο ελέγχου (test) στην αντίστοιχη περίπτωση linear expert mode έχω :

		0 (-)	1(+)
0	Magnitude Min (-)	234	581
1	Magnitude High (+)	101	1.665

Άρα a=(+ +)= πραγματικά θετικά = 1.755
b = (+ -) =λανθασμένα θετικά = 11
c =(- +) =λανθασμένα αρνητικά= 809
d = (- -) = πραγματικά αρνητικά= 6

και τα ποσοστά των μέτρων % είναι αναλυτικά :

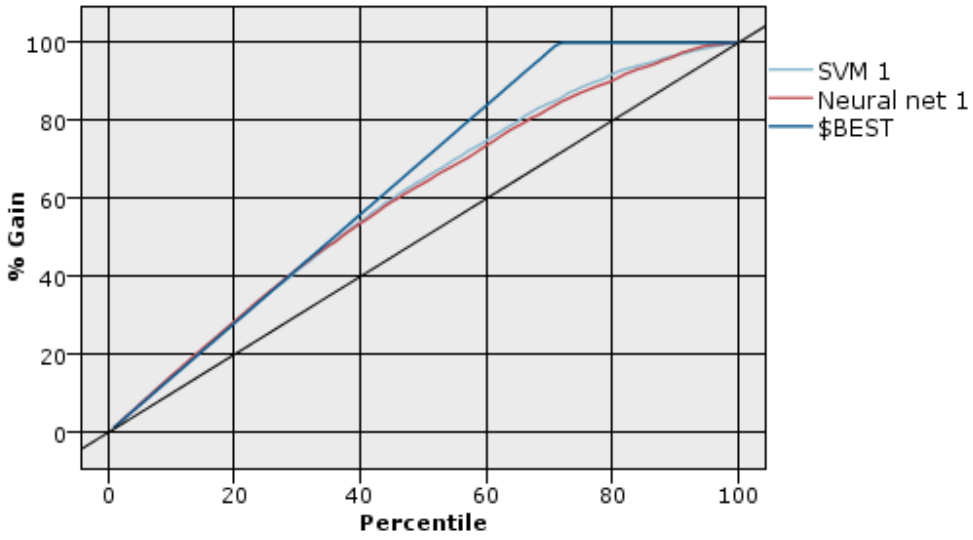
- ευαισθησία = 74,13 (%)
- ειδικότητα = 69,85 (%)
- θετική προγνωστική αξία (Θ.Π.Α) = 94,28 (%)
- αρνητική προγνωστική αξία (Α.Π.Α) = 28,71 (%)
- ακρίβεια = 73,58 (%)

8.2.3 Binary Classifier

Περαιτέρω συγκρίναμε την απόδοση των SVMs με την αυτή των Νευρωνικών δικτύων. Εξετάστηκαν σε δύο τομείς και συγκεκριμένα η σύγκριση των μοντέλων έγινε βάση της 1) Overall accuracy και του 2) Area Under the curve.

1) Overall accuracy

Graph	Model	Build Time (mins)	Max Profit	Max Profit Occurs in (%)	Lift (Top 30%)	Overall Accuracy (%)	No. Fields Used	Area Under Curve
	SVM 1	< 1	5.970	73	1,384	77,489	8	0,826
	Neural net 1	< 1	5.725	74	1,401	74,7	8	0,81

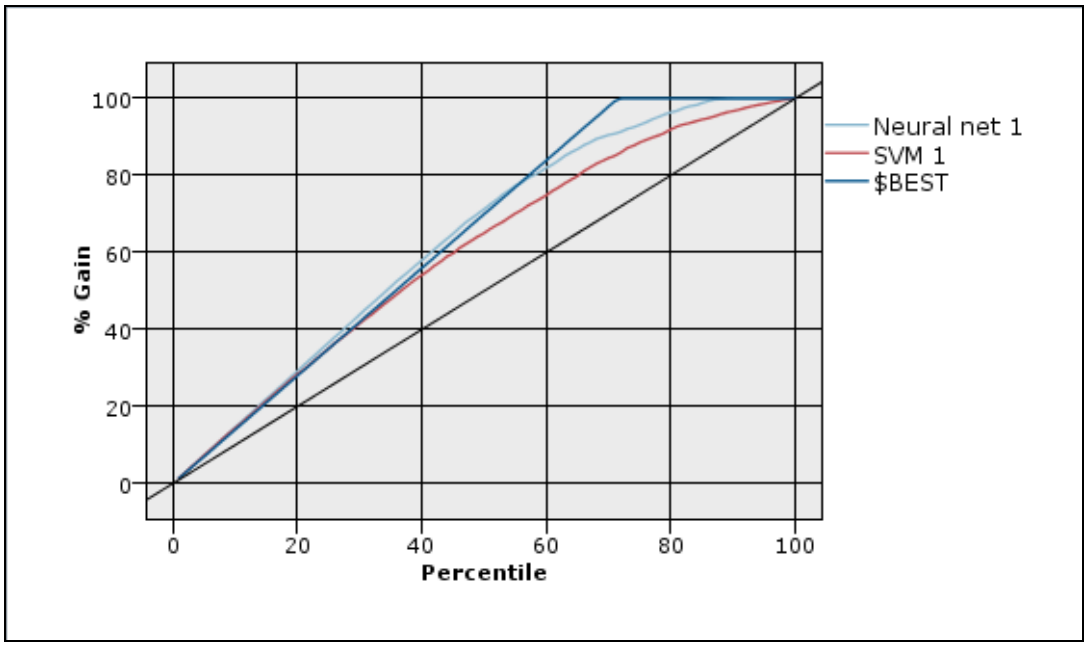


Όπως παρατηρούμε η απόδοση είναι κοντα με τα προηγούμενα τεστ. Οι τιμές της SVM και των NN είναι κοντά, ωστόσο η SVM παρουσιάζει καλύτερες τιμές.

Να προσθέσουμε ότι δεν κρατάμε μοντέλα με accuracy χαμηλότερη του 60%. Σε αυτή τη περίπτωση και οι δύο έχουν υψηλότερη ακρίβεια (accuracy).

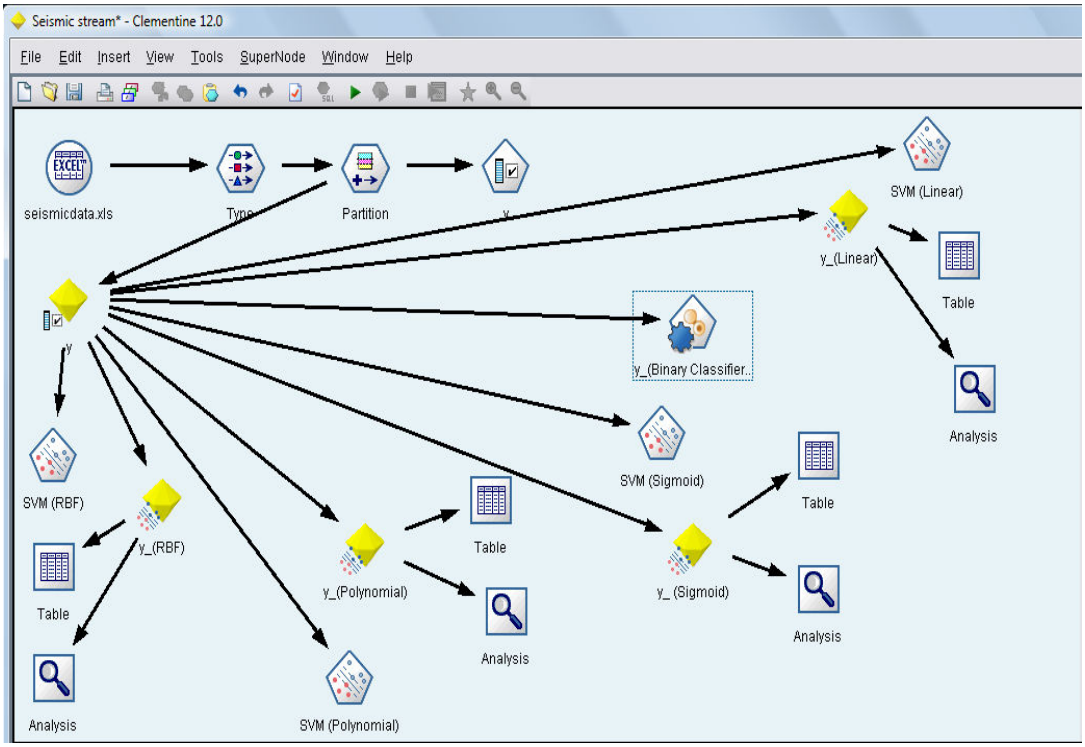
2) Area Under the Curve (AUC)

Graph	Model	Build Time (mins)	Max Profit	Max Profit Occurs in (%)	Lift (Top 30%)	Overall Accuracy (%)	No. Fields Used	Area Under Curve
	Neural net 1	< 1	7.025	68	1,461	84,657	8	0,935
	SVM 1	< 1	5.970	73	1,384	77,489	8	0,826



Οι τιμές της SVM και των NN είναι κοντά, ωστόσο τα NNs παρουσιάζουν καλύτερες τιμές.

Η τελική εικόνα του stream canvas μετά την εφαρμογή των αλγορίθμων με την βοήθεια του Clementine είναι :



Ευχαριστίες

Η εκπόνηση της συγκεκριμένης εργασίας θα ήταν πραγματικά αδύνατη χωρίς την σημαντική βοήθεια και ουσιαστική καθοδήγηση του Καθηγητή του Εθνικού Μετσόβιου Πολυτεχνείου κ. Χ. Κουκουβίνου, τον οποίο και ευχαριστώ θερμά. Επίσης θα ήθελα να ευχαριστήσω την υποψήφια διδάκτορα Χριστίνα Πάρπουλα για τη συμπαράστασή της και το συνεχές ενδιαφέρον της καθ όλη τη διάρκεια της εκπόνησης της διπλωματικής εργασίας, καθώς και τη διπλωματούχο της σχολής Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών Σοφία Παναγιωτοπούλου για την πολύτιμη βοήθειά της. Χωρίς τη συμβολή τον παραπάνω ανθρώπων θα ήταν αδυνατη η δημιουργία και παρουσίαση αυτής της εργασίας.

Βιβλιογραφία

1. I. Aydin, M. Karakose and E. Akin, (2009). The Prediction Algorithm Based on Fuzzy Logic Using Time Series Data Mining Method. World Academy of Science, *Engineering and Technology*, **51**, 91-98.

2. A. Berson, S. Smith and K.Thearling, (1999). *Building data mining applications for CRM*, 1st edt., Mc Graw-Hill Professional.

3. B. E. Boser, I. M. Gyon and V.N. Vapnik, (1992). A training algorithm for optimal margin classifiers, *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, ACM Press, pp. 144-152.

4. A. P. Bradley, (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms, *Pattern Recognition*, **30**, 1145-1159.

5. L. Breiman, J. H. Friedman, R. A. Olshen and C. J. Stone, (1984). *Classification and Regression Trees,* Monterey, Calif., U.S.A.: Wadsworth, Inc.

6. C. J. C. Burges, (1998). *A tutorial on Support Vector Machines for pattern recognition*, Kluwer Academic Publishers, Boston.

7. P. Cortez, (2010). Data mining with neural networks and support vector machines using the R/rminer tool, *Proceedings of the 10th industrial conference on Advances in data mining: applications and theoretical aspects*, pp. 572-583.

8. M. Deighton and M. Petrou, (2008). Data mining for large scale 3D seismic data analysis. *Machine Vision and Applications*, **20** (1), 11-22.

9. A. G. Eapen, (2004). *Application of Data mining in Medical Applications.* A thesis presented to the University of Waterloo in fulfillment of the thesis requirement for the degree of Master of Applied Science in Systems Design Engineering.

10. T. Fawcett, (2003). *ROC Graphs: Notes and Practical Considerations for Data Mining Researchers*, Intelligent Enterprise Technologies Laboratory, HP Laboratories Palo Alto.

11. G. Fung, M. Dundar, B. Krishnapuram, and R. B. Rao, (2006). Multiple instance algorithms for computer aided diagnosis. In *Advances in Neural Information Processing Systems*.

12. D. M. Green and J. A. Swets, (1974). *Signal detection theory and psychophysics*, New York, Wiley & Sons.

13. I. Guyon, V.N. Vapnik, B. E. Boser, L. Bottou and S. A. Solla, (1992). Structural Risk Minimization for Character Recognition**,** *Advances in Neural Information Processing Systems*, **4**, Morgan Kaufman, Denver.

14. D. J. Hand and R. J. Till, (2001). A Simple Generalization of the Area Under the ROC, *Machine Learning*, **45** (2), 71-186.

15. J. A. Hanley and B. J. McNeil, (1983). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology,* **143***,* 29-36.

16. V. Jakkula. Tutorial on Support Vector Machine (SVM). School of EECS, Washington State University, Pullman 99164.

17. R. Kumar, V. K. Jayaraman and B. D. Kulkarni, (2005). An SVM classifier incorporating simultaneous noise reduction and feature selection: illustrative case examples, *Pattern Recognition*, **38**, 41-49.

18. J. P. Lewis, (2004). *Tutorial on SVM*, CGIT Lab, USC.

19. A. Y. Ng, (2008). Support Vector Machines, Lecture notes*,* Part V, pp. 1-25.

20. P. Poncelet, F. Masseglia and M. Teisseire, (2007). *Data Mining Patterns: New Methods and Applications*, pp. 1-324, doi:10.4018/978-1-59904-162-9.

21. K. A. Spackman, (1989). Signal detection theory: Valuable tools for evaluating inductive learning. In *Proceedings of the Sixth International Workshop on Machine Learning*, pp. 160-163, San Mateo, CA. Morgan Kaufman.

22. J.A. Swets and R. M. Pickett, (1982). *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory*. Academic Press, New York.

23. Z. Wang, A. R. Childress, J. Wang, and J. A. Detre, (2007). Support vector machine learning-based fMRI data group analysis, *Neuroimage*, **36** (4), 1139-1151.

24. R. Zhou, D.-Y. Hu, L.-M, Liu, X.-W. Zhou, (2002). Protective effects of apocynin on "two-hit" injury induced by hemorrhagic shock and lipopolysaccharide, *Acta Pharmacol Sin,* **23** (11), 1023-1028.

25. Ε. Γεωργίου , (2009). *Χρήση των Μηχανών Διανυσμάτων υποστήριξης στην εκτίμηση τιμών ακινήτων*. Διπλωματικη εργασία Πανεπιστήμιο Κύπρου.

26. Σ.Ουγιάρογλου, (2006). *Classification based on dynamic number of Nearest Neighbours*. Διπλωματικη εργασία στο ΑΠΘ.

27. Μ.Παπαδάκη, (2010). *Τεχνικές εξόρυξης πληροφορίας, γενικευμένα γραμμικά μοντέλα και εφαρμογές με χρήση στατιστικών πακέτων*. Διπλωματικη εργασία στο ΕΜΠ.

28. Clementine tutorial

- Clementine® 12.0 Applications Guide

- Clementine® 12.0 Algorithms Guide
- Clementine® 12.0 Modeling Nodes

29. Ιστιακοί Τόποι

- ics.uci.edu/ ~ mlearn / MLrepository
- microarray.princeton.edu/oncology
- radiology.uchicago.edu/krl/toppagell.htm
- spss.com
- statsoft.com
- wikipedia.com