Κεφάλαιο 5: Γραμμικές Διακρίνουσες Συναρτήσεις

5.1 Εισαγωγή

Στο τρίτο κεφάλαιο θεωρήθηκε ότι η μορφή των συναρτήσεων πυκνότητας πιθανότητας είναι γνωστή και ότι τα δείγματα εκπαίδευσης χρησιμοποιούνται για τον υπολογισμό των τιμών των άγνωστων παραμέτρων τους. Στο κεφάλαιο αυτό θα υποτεθεί ότι είναι γνωστές οι κατάλληλες μορφές των διακρίνουσων συναρτήσεων και θα χρησιμοποιηθούν τα δείγματα εκπαίδευσης για τον υπολογισμό των παραμέτρων του ταξινομητή. Θα διερευνηθούν διάφορες διαδικασίες προσδιορισμού των διακρίνουσων συναρτήσεων, κάποιες από τι οποίες είναι στατιστικές και κάποιες όχι. Καμία πάντως από αυτές δεν απαιτεί τη γνώση της μορφής των αντίστοιχων συναρτήσεων κατανομής πιθανότητας και υπό αυτή την έννοια μπορούν να θεωρηθούν ως μη παραμετρικές.

Σε αυτό το κεφάλαιο θα ασχοληθούμε με διακρίνουσες συναρτήσεις οι οποίες είναι είτε γραμμικές ως προς τα στοιχεία του x είτε γραμμικές ως προς ένα δεδομένο σύνολο συναρτήσεων του x. Οι γραμμικές διακρίνουσες συναρτήσεις έχουν πολλές ευπρόσδεκτες αναλυτικές ιδιότητες. Όπως παρουσιάστηκε στο δεύτερο κεφάλαιο, μπορεί να είναι βέλτιστες εάν οι αντίστοιχες κατανομές είναι συνεργατικές, όπως για παράδειγμα Gaussian με ίδιες διασπορές. Τέτοιου είδους κατανομές μπορούν να προκύψουν μέσα από εύστοχη επιλογή των ανιχνευτών χαρακτηριστικών. Ακόμα όμως και στην περίπτωση όπου δεν είναι βέλτιστες, συνήθως επιλέγονται διότι η απλότητά τους αποτελεί πολύ σημαντικό πλεονέκτημα. Οι γραμμικές διακρίνουσες συναρτήσεις είναι σχετικά απλές στους υπολογισμούς, ενώ οι γραμμικοί ταξινομητές, όταν δεν υπάρχει κάποια συγκεκριμένη εκ των προτέρων γνώση, είναι ελκυστικοί υποψήφιοι για χρήση ως αρχικοί, δοκιμαστικοί ταξινομητές.

Το πρόβλημα της εύρεσης μιας γραμμικής διακρίνουσας συνάρτησης θα παρουσιαστεί ως το πρόβλημα της ελαχιστοποίησης μιας συνάρτησης κριτήριο. Η προφανής συνάρτηση κριτήριο για προβλήματα ταξινόμησης είναι το λάθος εκπαίδευσης - το μέσο κόστος που οφείλεται στην ταξινόμηση του συνόλου των δειγμάτων εκπαίδευσης. Το κριτήριο αυτό όμως, εάν και φαίνεται ιδιαίτερα ελκυστικό, έχει πολλά προβλήματα. Ο αντικειμενικός μας σκοπός είναι η σωστή ταξινόμηση καινούριων δειγμάτων και ένα μικρό λάθος κατά την εκπαίδευση δεν εγγυάται ένα αντίστοιχα μικρό λάθος κατά τη λειτουργία και τον έλεγχο καινούριων δειγμάτων. Όπως θα δούμε στη συνέχεια, είναι δύσκολο να προκύψει η γραμμική διακρίνουσα συνάρτηση του ελάχιστου ρίσκου, και γι' αυτό το λόγο συνήθως αναλύονται διάφορες σχετικών κριτηρίων συναρτήσεις οι οποίες έχουν καλύτερες αναλυτικές ιδιότητες.

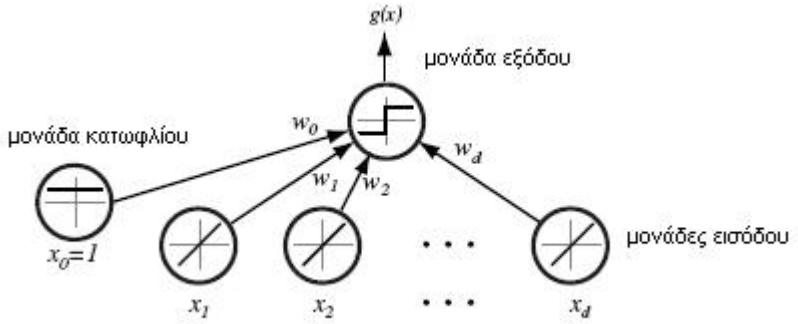
Ιδιαίτερη προσοχή θα δοθεί στη μελέτη των ιδιοτήτων σύγκλισης και της υπολογιστικής πολυπλοκότητας διαφόρων διαδικασιών κλίσης καθόδου (gradient descent) για την ελαχιστοποίηση συναρτήσεων κριτηρίου.

5.2 Γραμμικές Διακρίνουσες Συναρτήσεις και Επιφάνειες Απόφασης

Μια διακρίνουσα συνάρτηση που αποτελεί ένα γραμμικό συνδυασμό των στοιχείων του x μπορεί να γραφεί ως εξής:

$$g(x) = w^t x + w_0 \tag{2.1}$$

όπου το w είναι το διάνυσμα των βαρών και το w_0 είναι το βάρος κατωφλίου. Όπως παρουσιάστηκε στο δεύτερο κεφάλαιο, στη γενική περίπτωση θα υπάρχουν c διαφορετικές διακρίνουσες συναρτήσεις, όπου καθεμιά αντιστοιχεί σε μία από τις c διαφορετικές κατηγορίες. Αρχικά, θα διερευνηθεί η απλούστερη περίπτωση όπου υπάρχουν μόνο δύο κατηγορίες.



Εικόνα 5.1: Ένας απλός γραμμικός ταξινομητής που έχει d μονάδες εισόδου, καθεμιά από τις οποίες αντιστοιχεί στις τιμή ενός στοιχείου του διανύσματος εισόδου. Κάθε είσοδος (τιμή χαρακτηριστικού) x_i πολλαπλασιάζεται με το αντίστοιχο βάρος της w_i. Η συνολική είσοδος του νευρώνα εξόδου είναι το άθροισμα όλων των γινομένων $\sum w_i x_i$. Καθεμιά από τις d μονάδες εισόδου είναι γραμμική δίνοντας στην έξοδό της ακριβώς την αντίστοιχη τιμή της εισόδου της. Η μονάδα εξόδου δίνει a + 1 εάν $w^t x + w_0 > 0$ ή a – 1 διαφορετικά.

5.2.1 Η Περίπτωση Δύο Κατηγοριών

Για μια συνάρτηση της μορφής της εξίσωσης 5.1, ένας ταξινομητής δύο κατηγοριών υλοποιεί τον ακόλουθο κανόνα απόφασης: Αποφάσισε ω_1 εάν $g(x) > 0$ και ω_2 εάν $g(x) < 0$. Έτσι, το x ταξινομείται στην ω_1 εάν το εσωτερικό γινόμενο $w^t x$ υπερβαίνει το όριο w_0 και στην ω_2 διαφορετικά. Εάν $g(x) = 0$, το x μπορεί να ταξινομηθεί σε οποιαδήποτε από τις δύο κατηγορίες. Παρ' όλ' αυτά, στο κεφάλαιο αυτό η περίπτωση αυτή θα αντιμετωπίζεται ως απροσδιόριστη. Στην εικόνα 5.1 παρουσιάζεται μια τυπική υλοποίηση, ένα ξεκάθαρο παράδειγμα της γενικής δομής ενός συστήματος αναγνώρισης προτύπων.

Η εξίσωση $g(x) = 0$ προσδιορίζει την επιφάνεια απόφασης που διαχωρίζει τα σημεία που ταξινομούνται στην ω_1 από τα σημεία που ταξινομούνται στην ω_2. Όταν η g(x) είναι γραμμική, η περιοχή απόφασης είναι ένα υπερεπίπεδο. Εάν τα x_1 και x_2 βρίσκονται πάνω στην περιοχή απόφασης, τότε

$$w^t x_1 + w_0 = w^t x_2 + w_0$$

ή

$$w^t (x_1 - x_2) = 0$$

Αυτό δείχνει ότι το διάνυσμα w είναι κανονικό σε κάθε διάνυσμα που βρίσκεται στο υπερεπίπεδο. Γενικά, το υπερεπίπεδο Η διαιρεί το χώρο των χαρακτηριστικών σε δύο υποχώρους: την περιοχή απόφασης R_1 για την ω_1 και την περιοχή απόφασης R_2 για την ω_2. Επειδή ισχύει $g(x) > 0$ εάν το x ανήκει στην R_1, ως αποτέλεσμα το κανονικό διάνυσμα w δείχνει στην R_1. Πολλές φορές αναφέρεται ότι κάθε x που ανήκει στην R_1 βρίσκεται στη θετική πλευρά του Η ενώ κάθε x που ανήκει στην R_2 βρίσκεται στην αρνητική πλευρά του. Η διακρίνουσα συνάρτηση g(x) παρέχει ένα αλγεβρικό μέτρο

της απόστασης του x από το υπερεπίπεδο. Ίσως ο πιο απλός τρόπος για να δειχθεί αυτό είναι να εκφραστεί το x με τη μορφή

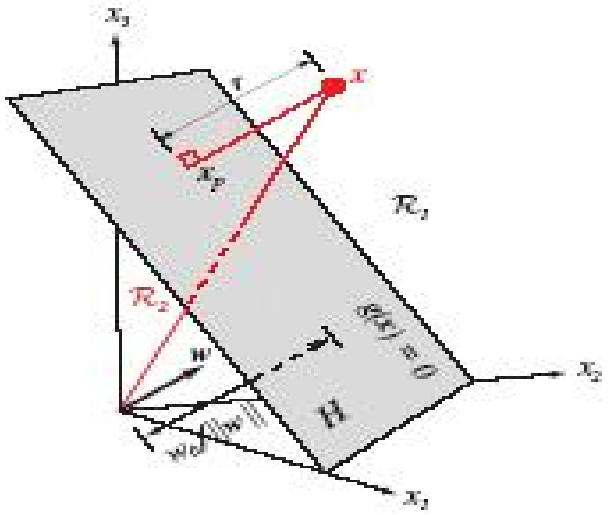
$$x = x_p + r\frac{w}{\|w\|}$$

όπου το x_p είναι η κανονική προβολή του x πάνω στο H και το r είναι η επιθυμητή αλγεβρική απόσταση - θετική εάν το x είναι στη θετική πλευρά και αρνητική εάν το x βρίσκεται στην αρνητική πλευρά του H. Άρα, αφού $g(x_p) = 0$,

$$g(x) = w^t x + w_0 = r\|w\|$$

ή

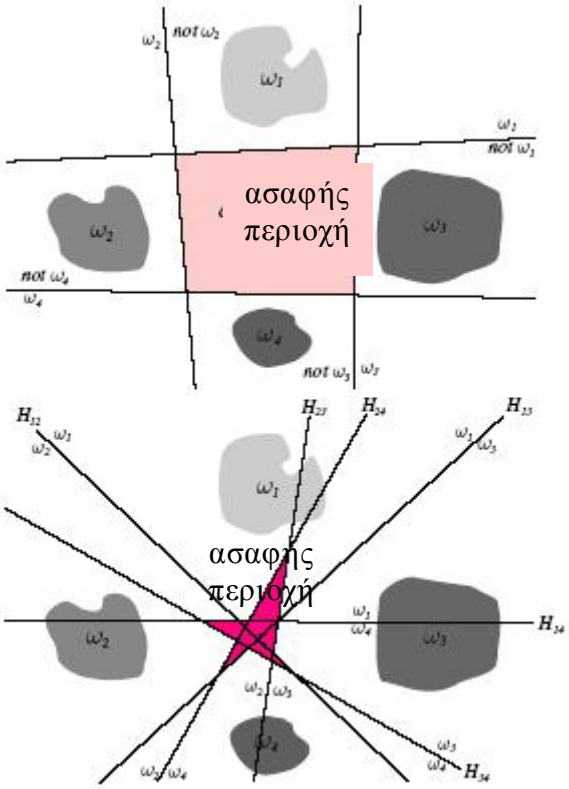
$$r = \frac{g(x)}{\|w\|}$$



Εικόνα 5.2: Το γραμμικό όριο απόφασης H, όπου $g(x) = w^t x + w_0 = 0$, διαχωρίζει το χώρο των χαρακτηριστικών σε δύο υποχώρους R_1 (όπου $g(x) > 0$) και R_2 (όπου $g(x) < 0$).

Πιο συγκεκριμένα, η απόσταση από την αρχή των αξόνων μέχρι το H δίνεται από το $w_0/\|w\|$. Εάν $w_0 > 0$, η αρχή των αξόνων βρίσκεται στη θετική πλευρά του H, ενώ εάν $w_0 < 0$ είναι στην αρνητική πλευρά. Εάν $w_0 = 0$, η g(x) παίρνει την ομογενή μορφή $w^t x$ και το υπερεπίπεδο διέρχεται από την κορυφή των αξόνων. Μια γεωμετρική αναπαράσταση των αλγεβρικών αυτών αποτελεσμάτων φαίνεται στην εικόνα 5.2.

Συνοψίζοντας, μία γραμμική διακρίνουσα συνάρτηση διαιρεί το χώρο των χαρακτηριστικών με μία επιφάνεια απόφασης που έχει τη μορφή υπερεπίπεδου. Ο προσανατολισμός της επιφάνειας καθορίζεται από το κανονικό διάνυσμα w, ενώ η θέση της επιφάνειας καθορίζεται από το βάρος του κατωφλίου w_0. Η διακρίνουσα συνάρτηση g(x) είναι ανάλογη της προσημασμένης απόστασης του x από το υπερεπίπεδο ενώ ισχύει $g(x) > 0$ εάν το x είναι στη θετική πλευρά και $g(x) < 0$ εάν το x βρίσκεται στην αρνητική πλευρά του H.



Εικόνα 5.3: Γραμμικά όρια απόφασης για ένα πρόβλημα τεσσάρων κατηγοριών. Στο πάνω σχήμα φαίνονται οι ω_i / όχι ω_i διχοτομήσεις, ενώ στο κάτω οι ω_i / ω_j διχοτομήσεις με τα αντίστοιχα H_{ij} όρια απόφασης.

5.2.3 Η Περίπτωση Πολλών Κατηγοριών

Υπάρχουν πολλοί τρόποι με τους οποίους μπορεί να αντιμετωπιστεί ένα πρόβλημα ταξινόμησης πολλών κατηγοριών όπου οι ταξινομητές χρησιμοποιούν γραμμικές διακρίνουσες συναρτήσεις. Για παράδειγμα, μπορεί να απλοποιηθεί σε c προβλήματα δύο κατηγοριών, όπου το i-οστό πρόβλημα επιλύεται από τη γραμμική διακρίνουσα συνάρτηση που διαχωρίζει τα σημεία που ταξινομούνται στην ω_i από αυτά που δεν ταξινομούνται στην ω_i. Μια άλλη προσέγγιση είναι να χρησιμοποιηθούν c(c - 1)/2 γραμμικές διακρίνουσες συναρτήσεις, μία για κάθε ζεύγος κατηγοριών. Όπως φαίνεται στην εικόνα 5.3 και οι δύο αυτές προσεγγίσεις μπορεί να οδηγήσουν σε περιοχές όπου η ταξινόμηση δεν μπορεί να οριστεί. Το πρόβλημα αυτό αντιμετωπίζεται υιοθετώντας την προσέγγιση που έγινε στο δεύτερο κεφάλαιο, δηλαδή ορίζοντας c γραμμικές διακρίνουσες συναρτήσεις

$$g_i(x) = w_i^t x + w_{i0} \qquad i = 1,...,c \tag{2.2}$$

και αναθέτοντας το x στην ω_1 εάν $g_i(x) > g_j(x)$ για κάθε $j \neq i$. Στις περιπτώσεις ισοπαλιών, η ταξινόμηση αφήνεται ως αδιευκρίνιστη. Ο ταξινομητής που προκύπτει καλείται γραμμική μηχανή. Μία γραμμική μηχανή διαιρεί το χώρο των χαρακτηριστικών σε c περιοχές απόφασης, με την $g_i(x)$ να είναι η μεγαλύτερη διακρίνουσα συνάρτηση εάν το x βρίσκεται στην περιοχή R_i. Εάν οι περιοχές R_i και R_j είναι όμορες, το όριο ανάμεσά τους είναι ένα μέρος του υπερεπιπέδου H_{ij} που ορίζεται από τη σχέση

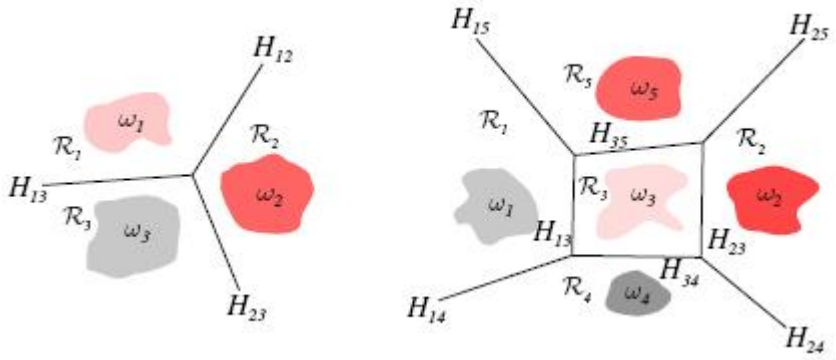
$$g_i(x) = g_j(x)$$

ή

$$\left(w_i - w_j\right)^t x + \left(w_{i0} - w_{j0}\right) = 0$$

Προκύπτει άμεσα ότι το w_i - w_j είναι κανονικό ως προς το H_{ij} και η προσημασμένη απόσταση από το x προς το H_{ij} δίνεται από το $\left(g_i(x) - g_j(x)\right) / \left\|w_i - w_j\right\|$. Έτσι, σε μια γραμμική μηχανή αυτό που έχει σημασία δεν είναι τα ίδια τα διανύσματα των βαρών αλλά οι διαφορές τους. Ενώ υπάρχουν c(c - 1)/2 ζεύγη περιοχών δεν είναι απαραίτητο να είναι σε όλα οι περιοχές όμορες μεταξύ τους και ως αποτέλεσμα ο συνολικός αριθμός των επιφανειών απόφασης είναι συχνά μικρότερος από c(c - 1)/2, όπως φαίνεται και στην εικόνα 5.4.

Είναι εύκολο να αποδειχθεί ότι οι περιοχές απόφασης σε μία γραμμική μηχανή είναι κυρτές. Αυτός ο περιορισμός περιορίζει φυσικά την προσαρμοστικότητα και την ακεραιότητα του ταξινομητή. Πιο συγκεκριμένα, κάθε περιοχή απόφασης είναι απλά συνδεδεμένη με αποτέλεσμα μια γραμμική μηχανή να είναι περισσότερο αποτελεσματική για προβλήματα στα οποία οι υπό συνθήκη συναρτήσεις πυκνότητας πιθανότητας $p(x/\omega_i)$ έχουν μόνο ένα ακρότατο (unimodal). Αυτό όμως δεν ισχύει πάντα. Υπάρχουν κατανομές με πολλά ακρότατα (multimodal) για τις οποίες η χρήση γραμμικών διακρίνουσων συναρτήσεων δίνει εξαιρετικά αποτελέσματα και κατανομές με ένα ακρότατο για τις οποίες δίνει πολύ άσχημα αποτελέσματα ταξινόμησης.



Εικόνα 5.4: Όρια απόφασης που δημιουργήθηκαν από μία γραμμική μηχανή για ένα πρόβλημα τριών κατηγοριών και πέντε κατηγοριών αντίστοιχα.

5.3 Γενικευμένες Γραμμικές Διακρίνουσες Συναρτήσεις

Η γραμμική διακρίνουσα συνάρτηση g(x) μπορεί να γραφεί ως

$$g(x) = w_0 + \sum_{i=1}^{d} w_i x_i \tag{5.3}$$

όπου οι συντελεστές w_i αποτελούν τα στοιχεία του διανύσματος των βαρών w. Προσθέτοντας επιπλέον όρους που εμπεριέχουν γινόμενα από ζεύγη στοιχείων του x προκύπτει η τετραγωνική διακρίνουσα συνάρτηση

$$g(x) = w_0 + \sum_{i=1}^{d} w_i x_i + \sum_{i=1}^{d}\sum_{j=1}^{d} w_{ij} x_i x_j \tag{5.4}$$

Επειδή ισχύει $x_i x_j = x_j x_i$, χωρίς απώλεια της γενικότητας μπορεί κάποιος να θεωρήσει ότι $w_{ij} = w_{ji}$. Έτσι, η τετραγωνική διακρίνουσα συνάρτηση έχει επιπλέον d(d + 1)/2 συντελεστές με τους οποίους μπορεί να δημιουργήσει πιο πολύπλοκες επιφάνειες διαχωρισμού. Η επιφάνεια διαχωρισμού που ορίζεται από την g(x) = 0 είναι μία δευτέρου βαθμού ή υπερτετραγωνική επιφάνεια. Εάν ο συμμετρικός πίνακας W =

$[w_{ij}]$ δεν είναι μη ομαλός (det(W)=0) οι γραμμικοί όροι της g(x) μπορούν να παραληφθούν με μεταφορά των αξόνων. Το βασικό χαρακτηριστικό της επιφάνειας διαχωρισμού μπορεί να περιγραφεί με όρους του κλιμακωτού πίνακα $\overline{W} = W/\left(w^t W^{-1} w - 4w_0\right)$. Εάν ο $\overline{W}$ είναι ένα θετικό πολλαπλάσιο του ταυτοτικού πίνακα, η επιφάνεια διαχωρισμού είναι μία υπερσφαίρα. Εάν ο $\overline{W}$ είναι θετικά ορισμένος, η επιφάνεια διαχωρισμού είναι ένα υπερελλειψοειδές. Εάν κάποιες από τις ιδιοτιμές του $\overline{W}$ είναι θετικές και άλλες είναι αρνητικές, η επιφάνεια θα είναι μία από την ποικιλία των διαφορετικών τύπων υπερυπερβολοειδών. Όπως αναφέρθηκε στο δεύτερο κεφάλαιο, αυτές οι επιφάνειες εμφανίζονται στη γενικής μορφής Gaussian περίπτωση πολλών μεταβλητών.

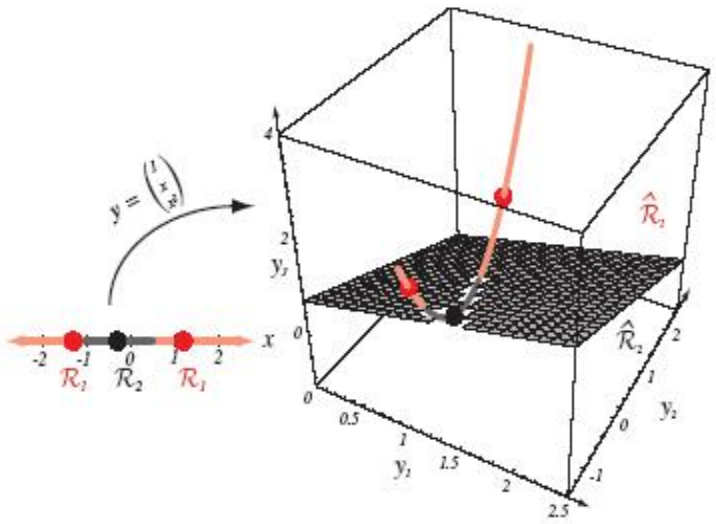
Συνεχίζοντας την προσθήκη όρων όπως οι $w_{ijk}x_i x_j x_k$, προκύπτει η κατηγορία των πολυωνυμικών διακρίνουσων συναρτήσεων. Αυτές οι συναρτήσεις μπορούν να θεωρηθούν ως στρογγυλοποιημένες επεκτάσεις σειρών κάποιας αυθαίρετης συνάρτησης g(x), τα οποία με τη σειρά τους υποδηλώνουν τη γενικευμένη γραμμική διακρίνουσα συνάρτηση

$$g(x) = \sum_{i=1}^{\hat{d}} \alpha_i y_i(x) \tag{5.5}$$

ή

$$g(x) = a^t y \tag{5.6}$$

όπου το a είναι ένα $\hat{d}$-διάστατο διάνυσμα βαρών και οι $\hat{d}$ συναρτήσεις $y_i(x)$ - οι οποίες καλούνται και συναρτήσεις φ - μπορούν να είναι οποιεσδήποτε συναρτήσεις του x. Τέτοιας μορφής συναρτήσεις μπορούν να υπολογιστούν από ένα υποσύστημα ανίχνευσης χαρακτηριστικών. Επιλέγοντας αυτές τις συναρτήσεις με συνετό τρόπο και επιτρέποντας στο $\hat{d}$ να είναι αρκετά μεγάλο, μπορεί να υπολογιστεί οποιαδήποτε επιθυμητή διακρίνουσα συνάρτηση με μια τέτοια expansion. Η διακρίνουσα συνάρτηση που προκύπτει δεν είναι γραμμική ως προς το x, αλλά είναι γραμμική ως προς το y. Οι $\hat{d}$ συναρτήσεις $y_i(x)$ κυρίως αντιστοιχούν σημεία από το d-διάστατο χώρο των x σε σημεία στο $\hat{d}$-διάστατο χώρο των y. Η ομογενής διακρίνουσα $a^t y$ διαχωρίζει τα σημεία σε αυτόν τον μετασχηματισμένο χώρο χρησιμοποιώντας ένα υπερεπίπεδο το οποίο διέρχεται από την αρχή των αξόνων. Έτσι, η αντιστοίχηση από το x στο y ανάγει το πρόβλημα στην εύρεση μιας ομογενούς γραμμικής διακρίνουσας συνάρτησης.



Εικόνα 5.5: Η απεικόνιση $y = (1, x, x^2)$ παίρνει μια γραμμή και τη μετασχηματίζει σε μία παραβολή στον τρισδιάστατο χώρο. Ένα επίπεδο χωρίζει τον προκύπτοντα y – χώρο σε περιοχές που αντιστοιχούν στις δύο κατηγορίες R_1 και R_2.

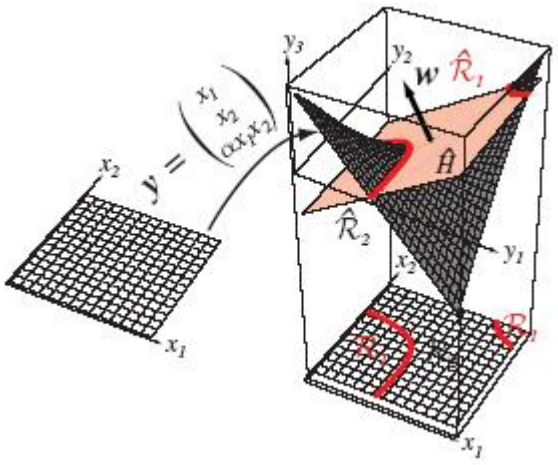
Στη συνέχεια θα παρουσιαστούν κάποια από τα σημαντικότερα πλεονεκτήματα και μειονεκτήματα αυτής της προσέγγισης χρησιμοποιώντας ένα απλό παράδειγμα. Έστω ότι η τετραγωνική διακρίνουσα συνάρτηση έχει τη μορφή

$$g(x) = \alpha_1 + \alpha_2 x + \alpha_3 x^2 \tag{5.7}$$

έτσι ώστε το τριών διαστάσεων διάνυσμα y να δίνεται από τη σχέση

$$y = \begin{pmatrix} 1 \\ x \\ x^2 \end{pmatrix} \tag{5.8}$$

Η αντιστοίχηση από το x στο y παρουσιάζεται στην εικόνα 5.5. Τα δεδομένα παραμένουν εγγενώς μονοδιάστατα, επειδή το μεταβλητό x αναγκάζει το y να ιχνηλατεί μία καμπύλη στις τρεις διαστάσεις. Έτσι, ένα πρώτο πράγμα που πρέπει να παρατηρήσει κανείς είναι ότι εάν το x καθορίζεται από ένα πιθανοτικό κανόνα p(x), η προκληθείς συνάρτηση πυκνότητας πιθανότητας $\hat{p}(y)$ θα εκφυλιστεί, θα είναι μηδέν παντού εκτός από την καμπύλη, όπου είναι άπειρη. Αυτό είναι ένα σύνηθες πρόβλημα όταν $\hat{d} > d$ και η αντιστοίχηση γίνεται από σημεία ενός χώρου με χαμηλότερη διάσταση σε σημεία ενός χώρου με υψηλότερη διάσταση. Το επίπεδο $\hat{H}$ που ορίζεται από την $a^t y = 0$ διαιρεί τον χώρο των y σε δύο περιοχές απόφασης $\hat{R}_1$ και $\hat{R}_2$. Στην εικόνα 5.6 φαίνεται το επίπεδο που αντιστοιχεί στο $a = (-1, 1, 2)^t$, οι περιοχές απόφασης $\hat{R}_1$ και $\hat{R}_2$, και οι αντίστοιχες περιοχές απόφασης R_1 και R_2 στον αρχικό χώρο των x. Η τετραγωνική διακρίνουσα συνάρτηση $g(x) = -1 + x + 2x^2$ είναι θετική εάν $x < -1$ ή εάν $x > 0.5$ και επομένως η R_1 είναι πολλαπλά συνδεδεμένη. Έτσι, εάν και οι περιοχές απόφασης στον χώρο των y είναι κυρτές, αυτό δε συμβαίνει αντίστοιχα στο χώρο των x. Γενικότερα, ακόμα και όταν οι συναρτήσεις $y_i(x)$ είναι σχετικά απλές, οι επιφάνειες απόφασης induced σε έναν χώρο των x μπορεί να είναι ιδιαίτερα πολύπλοκες.



Εικόνα 5.6: Ο δισδιάστατος χώρος εισόδων x απεικονίζεται μέσω της πολυωνυμικής συνάρτησης f στον χώρο y. Η απεικόνιση είναι $y_1 = x_1$, $y_2 = x_2$ και $y_3 = \alpha x_1 x_2$. Ένας γραμμικός διαχωριστής σε αυτόν τον μετασχηματισμένο χώρο είναι ένα υπερεπίπεδο, το οποίο τέμνει την επιφάνεια. Τα σημεία στη θετική πλευρά του $\hat{H}$ αντιστοιχούν στην κατηγορία ω_1, ενώ αυτά που βρίσκονται στην αρνητική αντιστοιχούν στην κατηγορία ω_2.

Μία πλήρης τετραγωνική διακρίνουσα συνάρτηση περιέχει $\hat{d} = (d+1)(d+2)/2$ όρους. Εάν το d είναι πολύ μεγάλο, π.χ. d=50, απαιτείται ο υπολογισμός πάρα πολλών όρων. Η εισαγωγή κυβικών και γενικότερα k-διάστατων τμημάτων στο πολυώνυμο οδηγεί σε όρους με $O(\hat{d}^k)$ υπολογιστικές απαιτήσεις. Επιπλέον, τα $\hat{d}$ στοιχεία του διανύσματος των βαρών a πρέπει να καθοριστούν από τα δείγματα εκπαίδευσης. Εάν το $\hat{d}$ αντιμετωπιστεί ως ο αριθμός των βαθμών ελευθερίας για τη διακρίνουσα συνάρτηση, είναι φυσικό να απαιτείται ο αριθμός των δειγμάτων να μην είναι μικρότερος από τον αριθμό των βαθμών ελευθερίας. Προφανώς, μία γενική επέκταση σειράς της g(x) μπορεί να οδηγήσει σε εντελώς μη ρεαλιστικές απαιτήσεις για τους υπολογισμούς και τα δεδομένα. Αυτό το μειονέκτημα μπορεί να αντιμετωπιστεί θέτοντας τον περιορισμό να υπάρχουν μεγάλα κενά ανάμεσα στα πρότυπα εκπαίδευσης. Σε αυτή την περίπτωση, δεν προσπαθούμε να προσεγγίσουμε όλες τις ελεύθερες παραμέτρους. Βασιζόμαστε στη θεώρηση ότι η αντιστοίχηση σε έναν χώρο υψηλότερης διάστασης δεν δημιουργεί spurius δομή ή σχέσεις μεταξύ των σημείων εκπαίδευσης. Μια εναλλακτική προσέγγιση στο πρόβλημα αυτό εφαρμόζεται με τη χρήση Νευρωνικών Δικτύων, τα οποία χρησιμοποιούν πολλά αντίγραφα μιας δεδομένης μη γραμμικής συνάρτησης των χαρακτηριστικών εισόδου.
Εάν και είναι σχετικά δύσκολο να συνειδητοποιήσει κανείς τα πιθανά πλεονεκτήματα της χρήσης μιας γενικευμένης γραμμικής διακρίνουσας συνάρτησης, μπορεί τουλάχιστον να εκμεταλλευτεί σε πρώτη φάση τη δυνατότητα να γράψει τη g(x) στην ομογενή μορφή $a^t y$. Πιο συγκεκριμένα, στην περίπτωση της γραμμικής διακρίνουσας συνάρτησης ισχύει

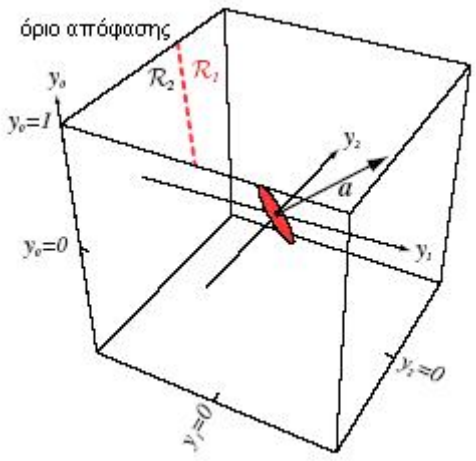
$$g(x) = w_0 + \sum_{i=1}^{d} w_i x_i = \sum_{i=0}^{d} w_i x_i \tag{5.9}$$

όπου το x_0 έχει τεθεί ίσο με 1. Έτσι, προκύπτει

$$y = \begin{bmatrix} 1 \\ x_1 \\ . \\ . \\ . \\ x_d \end{bmatrix} = \begin{bmatrix} 1 \\ \\ \\ x \\ \\ \\ \end{bmatrix} \tag{5.10}$$

όπου το y καλείται πολλές φορές επαυξημένο διάνυσμα χαρακτηριστικών. Ομοίως, ένα επαυξημένο διάνυσμα βαρών μπορεί να γραφεί ως:

$$a = \begin{bmatrix} w_0 \\ w_1 \\ . \\ . \\ . \\ w_d \end{bmatrix} = \begin{bmatrix} w_0 \\ \\ w \\ \\ \end{bmatrix} \tag{5.11}$$



Εικόνα 5.7: Ένας τρισδιάστατος επαυξημένος χώρος χαρακτηριστικών y και το επαυξημένο διάνυσμα των βαρών a. Το σύνολο των σημείων για τα οποία $a^t y = 0$, είναι ένα επίπεδο, (ή πιο γενικά ένα υπερεπίπεδο) κάθετο στο a το οποίο περνά από την αρχή των αξόνων του χώρου y. Ένα τέτοιο επίπεδο δεν περνάει υποχρεωτικά από την αρχή των αξόνων του δισδιάστατου χώρου χαρακτηριστικών του προβλήματος. Επομένως, υπάρχει ένα επαυξημένο διάνυσμα βαρών a το οποίο οδηγεί άμεσα σε αποφάσεις στο χώρο x.

Αυτή η αντιστοίχηση από το d-διάστατο χώρο των x στο (d +1)-διάστατο χώρο των y, εάν και από μαθηματικής πλευράς είναι τετριμμένη, είναι ιδιαίτερα βολική. Η προσθήκη ενός σταθερού όρου στο x αφήνει ανεπηρέαστες όλες τις αποστάσεις μεταξύ των δειγμάτων. Τα διανύσματα του y που προκύπτουν βρίσκονται όλα σε ένα d-διάστατο υποχώρο, ο οποίος δεν είναι άλλος από τον ίδιο το χώρο των x. Η επιφάνεια απόφασης μορφής υπερεπιπέδου $\hat{H}$ που ορίζεται από το $a^t y = 0$ διέρχεται από την αρχή των αξόνων του χώρου των y, εάν και το αντίστοιχο υπερεπίπεδο H μπορεί να βρίσκεται σε οποιαδήποτε θέση στο χώρο των x. Η απόσταση του y από το $\hat{H}$ δίνεται από τον τύπο $|a^t y|/\|a\|$ ή $|g(x)|/\|a\|$. Επειδή ισχύει $\|a\| \geq \|w\|$, η απόσταση αυτή είναι μικρότερη ή το πολύ ίση με την απόσταση του x από το H. Χρησιμοποιώντας αυτή την αντιστοίχηση το πρόβλημα της εύρεσης ενός διανύσματος βαρών w και ενός βάρους κατωφλίου w_0 ανάγεται στην εύρεση ενός μόνο διανύσματος βαρών a (εικόνα 5.7).

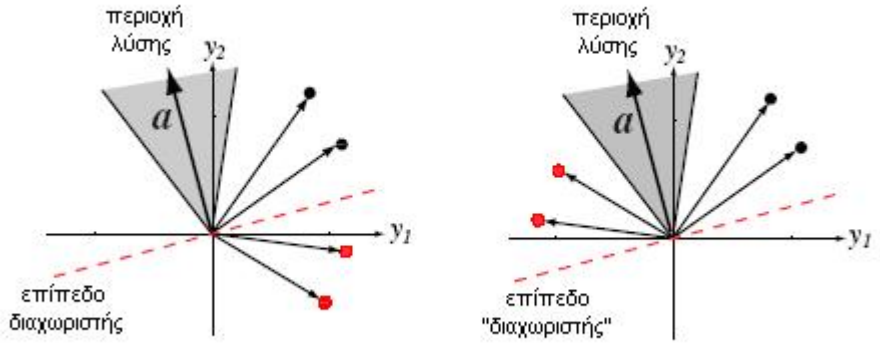
5.4 Η Περίπτωση Δύο Γραμμικά Διαχωριζόμενων Κατηγοριών

Έστω τώρα ότι υπάρχει ένα σύνολο από n δείγματα $y_1, ..., y_n$, κάποια από τα οποία έχουν ετικέτα ω_1 (υποτίθεται δηλαδή ότι ανήκουν στην κατηγορία ω_1) ενώ τα

υπόλοιπα έχουν ετικέτα ω_2 (υποτίθεται δηλαδή ότι ανήκουν στην κατηγορία ω_2). Πρέπει χρησιμοποιώντας τα δείγματα αυτά να καθοριστούν τα βάρη του διανύσματος a μιας γραμμικής διακρίνουσας συνάρτησης $g(x) = a^t y$. Έστω ότι υπάρχει κάποια ένδειξη ότι υπάρχει μία λύση για την οποία η πιθανότητα του λάθους είναι πολύ χαμηλή. Προφανώς μια λογική προσέγγιση είναι να ψάξει κανείς για ένα διάνυσμα βαρών το οποίο θα ταξινομεί σωστά όλα τα δείγματα. Εάν τελικά βρεθεί ένα τέτοιο διάνυσμα βαρών, τα δείγματα λέγεται ότι είναι γραμμικά διαχωρίσιμα.
Ένα δείγμα y_i ταξινομείται σωστά εάν ισχύει $a^t y_i > 0$ και το y_i έχει ετικέτα ω_1 ή εάν ισχύει $a^t y_i < 0$ και το y_i έχει ετικέτα ω_2. Αυτή η διαδικασία υποδηλώνει μια κανονικοποίηση, η οποία απλοποιεί την αντιμετώπιση της περίπτωσης των δύο κατηγοριών. Η κανονικοποίηση αυτή δεν είναι τίποτα παραπάνω από την αντικατάσταση όλων των δειγμάτων που έχουν ετικέτα ω_2 από τις αντίθετες τιμές τους. Χρησιμοποιώντας αυτή την κανονικοποίηση κάποιος μπορεί να παραλείψει τις ετικέτες και να περιορίσει την αναζήτησή του σε ένα διάνυσμα a τέτοιο ώστε $a^t y_i > 0$ για όλα τα δείγματα. Ένα τέτοιο διάνυσμα καλείται διάνυσμα διαχωρισμού ή πιο γενικά διάνυσμα λύσης.

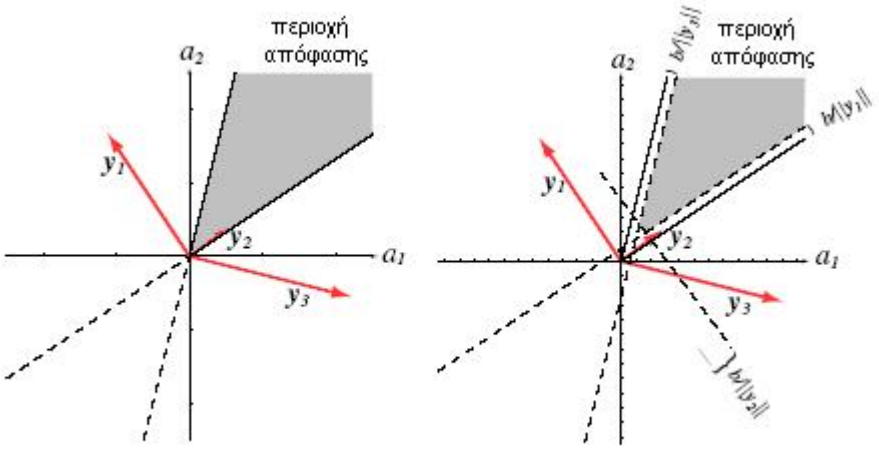
5.4.1 Γεωμετρία και Ορολογία

Το διάνυσμα των βαρών a μπορεί να θεωρηθεί ως ο καθορισμός ενός σημείου στο χώρο των βαρών. Κάθε δείγμα y_i θέτει ένα περιορισμό στην πιθανή θέση ενός διανύσματος λύσης. Η εξίσωση $a^t y_i = 0$ ορίζει ένα υπερεπίπεδο που διέρχεται από την αρχή των αξόνων του χώρου των βαρών έχοντας το y_i ως κανονικό διάνυσμα. Το διάνυσμα λύσης - εάν υπάρχει - πρέπει να βρίσκεται στη θετική πλευρά κάθε υπερεπιπέδου. Έτσι, ένα διάνυσμα λύσης πρέπει να βρίσκεται στην τομή n ημιχώρων. Πράγματι, κάθε διάνυσμα που βρίσκεται σε αυτή την περιοχή αποτελεί ένα διάνυσμα λύσης. Η αντίστοιχη περιοχή καλείται περιοχή λύσεων και δεν πρέπει να τη συγχέει κανείς με την περιοχή απόφασης του χώρου των χαρακτηριστικών που αντιστοιχεί σε κάποια συγκεκριμένη κατηγορία. Στην εικόνα 5.8 μπορεί να δει κανείς ένα παράδειγμα δύο διαστάσεων στο οποίο παρουσιάζεται η περιοχή λύσης τόσο για την κανονικοποιημένη όσο και για την μη κανονικοποιημένη περίπτωση.
Από τα παραπάνω, είναι προφανές ότι το διάνυσμα λύσης - εάν φυσικά υπάρχει - δεν είναι μοναδικό. Υπάρχουν διάφοροι τρόποι για να τεθούν επιπλέον απαιτήσεις οι οποίες να περιορίζουν το διάνυσμα λύση. Ένας τρόπος είναι η αναζήτηση για ένα μοναδιαίου μήκους διάνυσμα βαρών το οποίο μεγιστοποιεί την ελάχιστη απόσταση των δειγμάτων από το επίπεδο διαχωρισμού. Ένας άλλος τρόπος είναι η αναζήτηση για το ελάχιστου μήκους διάνυσμα βαρών που ικανοποιεί την $a^t y_i \geq b$ για όλα τα i, όπου το b είναι μία θετική σταθερά που καλείται όριο (margin). Όπως φαίνεται στην εικόνα 5.9 η περιοχή λύσης που προκύπτει από τις τομές των ημιχώρων για τους οποίους ισχύει $a^t y_i \geq b > 0$ εμπεριέχεται στην προηγούμενη περιοχή λύσης, διαφοροποιημένη σε σχέση με τα παλιά όρια κατά μια απόσταση ίση με $b/\|y_i\|$.



Εικόνα 5.8: Τέσσερα δείγματα εκπαίδευσης και η περιοχή λύσης στο χώρο των χαρακτηριστικών. Το σχήμα στα αριστερά, δείχνει τα αρχικά δεδομένα. Τα διανύσματα της λύσης οδηγούν σε ένα επίπεδο το οποίο χωρίζει τα πρότυπα των δύο κατηγοριών. Στο σχήμα στα δεξιά, τα σημεία της μιας κατηγορίας έχουν «κανονικοποιηθεί», δηλαδή έχει αλλαχτεί το πρόσημό τους. Το διάνυσμα λύσης οδηγεί σε ένα επίπεδο που τοποθετεί όλα τα «κανονικοποιημένα» σημεία στην ίδια πλευρά.

Το κίνητρο που βρίσκεται πίσω από τις προσπάθειες για την εύρεση ενός διανύσματος λύσης πλησιέστερου στο μέσον της περιοχής λύσης, είναι η πεποίθηση ότι η λύση που θα προκύψει έχει περισσότερες πιθανότητες να ταξινομεί σωστά νέα δείγματα ελέγχου. Παρ' όλ' αυτά, στις περισσότερες από τις περιπτώσεις που θα εξεταστούν, θα αρκεί να βρεθεί μία λύση που βρίσκεται αυστηρά μέσα στα όρια της περιοχής λύσης. Αυτό που ενδιαφέρει περισσότερο είναι να αποδειχθεί ότι κάθε επαναληπτική διαδικασία από αυτές που θα χρησιμοποιηθούν δε συγκλίνει σε ένα οριακό σημείο στην περιοχή του ορίου. Το πρόβλημα αυτό μπορεί πάντα να αντιμετωπιστεί με την εισαγωγή ενός ορίου, δηλαδή, απαιτώντας να ισχύει $a^t y_i \geq b > 0$ για κάθε i.



Εικόνα 5.9: Το αποτέλεσμα της χρήσης διαστήματος στο διάνυσμα λύσης. Στο αριστερό σχήμα φαίνεται η περίπτωση όπου δεν υπάρχει διάστημα (b = 0) και αντιστοιχεί στο αριστερό σχήμα της εικόνας 5.8. Στο σχήμα στα δεξιά, φαίνεται η περίπτωση όπου υπάρχει κάποιο διάστημα (b > 0) και η περιοχή λύσης περιορίζεται κατά τα διαστήματα $b/\|y_i\|$.

5.4.2 Διαδικασίες Κλίσης Καθόδου (Gradient Descent)

Η προσέγγιση που θα ακολουθηθεί για την εύρεση μιας λύσης για το σύνολο των γραμμικών ανισοτήτων $a^t y_i > 0$ είναι να οριστεί μία συνάρτηση κριτηρίου J(a) η οποία ελαχιστοποιείται εάν το a αποτελεί ένα διάνυσμα λύσης. Αυτό ανάγει το αρχικό πρόβλημα σε ένα πρόβλημα ελαχιστοποίησης μιας συνάρτησης πρώτου βαθμού - πρόβλημα το οποίο μπορεί συχνά να επιλυθεί από μια διαδικασία κλίσης καθόδου. Η βασική διαδικασία κλίσης καθόδου είναι πολύ απλή. Στην αρχή επιλέγεται αυθαίρετα ένα διάνυσμα βαρών a(1) και υπολογίζεται το διάνυσμα κλίσης $\nabla J(a(1))$. Η επόμενη τιμή a(2) υπολογίζεται με μετακίνηση από το a(1) κατά μία συγκεκριμένη απόσταση στη διεύθυνση της κλίσης καθόδου, δηλαδή κατά μήκος της αρνητικής κλίσης. Γενικότερα, το a(k + 1) υπολογίζεται από το a(k) με βάση την εξίσωση

$$a(k+1) = a(k) - \eta(k)\nabla J(a(k)) \tag{5.12}$$

όπου το η είναι ένας θετικός παράγοντας, ο οποίος καλείται παράμετρος μάθησης και καθορίζει το μέγεθος του βήματος. Τελικός σκοπός είναι αυτή η ακολουθία των διανυσμάτων των βαρών να συγκλίνει σε μία λύση η οποία να ελαχιστοποιεί το J(a). Η μορφή του αλγορίθμου είναι η εξής:

Αλγόριθμος 1: Βασικός Gradient Descent
1 αρχή αρχικοποίησε a, κατώφλι θ, η(·), k ← 0
2 κάνε k ← k + 1
3 a ← a – η(k) ∇J(a)
4 μέχρι | η(k) ∇J(a)| < θ
5 επέστρεψε a
6 τέλος

Τα πολλά προβλήματα που σχετίζονται με τις διαδικασίες κλίσης καθόδου είναι λίγο πολύ γνωστά. Ευτυχώς, οι συναρτήσεις που πρέπει να μεγιστοποιηθούν συνήθως επιλέγονται έτσι ώστε να επιτρέπουν την αποφυγή των σημαντικότερων από τα προβλήματα αυτά. Κάτι όμως που θα αποτελεί πάντα σημαντικό πρόβλημα, είναι η επιλογή της τιμής της παραμέτρου μάθησης η(k). Εάν η τιμή της η(k) είναι πολύ μικρή, η σύγκλιση είναι αναίτια πολύ αργή, ενώ εάν η τιμή της είναι πολύ μεγάλη, η διαδικασία διόρθωσης είναι τόσο έντονη ώστε πολλές φορές οδηγεί σε απόκλιση.
Στη συνέχεια θα περιγραφεί μία πρωταρχική μέθοδος για τον καθορισμό του ρυθμού μάθησης. Έστω ότι η συνάρτηση κριτηρίου μπορεί να υπολογιστεί ικανοποιητικά από την επέκταση δεύτερου βαθμού γύρω από μία τιμή a(k) ως εξής:

$$J(a) \cong J(a(k)) + \nabla J^t (a - a(k)) + \frac{1}{2}(a - a(k))^t H (a - a(k)) \tag{5.13}$$

όπου το H είναι ο Hessian πίνακας των δεύτερων μερικών παραγώγων $\partial^2 J / \partial a_i \partial a_j$ που αποτιμούνται στο a(k). Έπειτα, αντικαθιστώντας το a(k + 1) από την εξίσωση 5.12 στην εξίσωση 5.13 προκύπτει

$$J\left(a(k+1)\right) \cong J\left(a(k)\right) - \eta(k)\left\|\nabla J^{2}\right\| + \frac{1}{2}\eta^{2}(k)\nabla J^{t} H \nabla J$$

Από αυτό συνεπάγεται ότι το J(a(k + 1)) μπορεί να ελαχιστοποιηθεί επιλέγοντας

$$\eta(k) = \frac{\left\|\nabla J\right\|^{2}}{\nabla J^{t} H \nabla J} \tag{5.14}$$

όπου το H εξαρτάται από το a και επομένως έμμεσα και από την τιμή του k. Αυτή η τιμή αποτελεί τη βέλτιστη επιλογή για το η(k) δεδομένων των υποθέσεων που έγιναν. Να σημειωθεί ότι εάν η συνάρτηση κριτηρίου J(a) είναι τετραγωνική μέσα στην περιοχή που ενδιαφέρει, τότε το H είναι σταθερό και το η είναι μία σταθερά ανεξάρτητη από την τιμή του k.

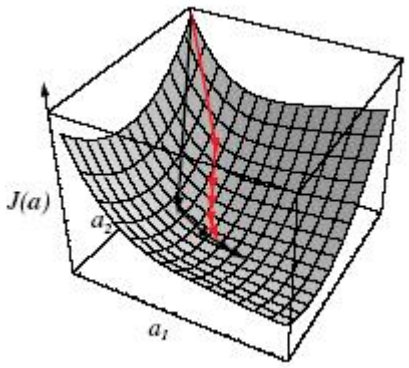
Μια εναλλακτική προσέγγιση, που προκύπτει αγνοώντας την εξίσωση 5.12 και επιλέγοντας το a(k + 1) για ελαχιστοποίηση της επέκτασης δεύτερου βαθμού, είναι ο αλγόριθμος του Newton. Ο αλγόριθμος αυτός προκύπτει από τον 1^{o} αλγόριθμο αντικαθιστώντας την τρίτη γραμμή με το εξής:

$$a(k+1) = a(k) - H^{-1}\nabla J \tag{5.15}$$

γεγονός που οδηγεί στον παρακάτω αλγόριθμο:

Αλγόριθμος 2: Newton Descent
1 αρχή αρχικοποίησε a, κατώφλι θ
2 κάνε
3 $a \leftarrow a - H^{-1} \nabla J(a)$
4 μέχρι $|H^{-1} \nabla J(a)| < \theta$
5 επέστρεψε a
6 τέλος

Στην εικόνα 5.10 συγκρίνεται αλγόριθμος του απλού gradient descent και ο αλγόριθμος του Newton.



Εικόνα 5.10: Η ακολουθία των διανυσμάτων των βαρών που παράγεται από μία απλή μέθοδο κλίσης καθόδου και από το δευτέρου βαθμού αλγόριθμο του Newton. Η μέθοδος του Newton οδηγεί συνήθως σε μεγαλύτερη βελτίωση ανά βήμα ακόμα και στην περίπτωση όπου χρησιμοποιούνται οι βέλτιστοι ρυθμοί μάθησης για τις δύο μεθόδους. Παρ' όλ' αυτά, το σημαντικό υπολογιστικό βάρος που προσθέτει ο υπολογισμός του Hessian πίνακα στη μέθοδο του Newton, έχει ως αποτέλεσμα πολλές φορές να επιλέγονται οι μέθοδοι κλίσης καθόδου.

Γενικότερα, ο αλγόριθμος του Newton προσφέρει συνήθως μία μεγαλύτερη βελτίωση ανά βήμα σε σχέση με τον απλό αλγόριθμο κλίσης καθόδου, ακόμα και στην περίπτωση όπου το η(k) έχει τη βέλτιστη τιμή του. Παρ' όλ' αυτά, ο αλγόριθμος του Newton δεν μπορεί να εφαρμοστεί εάν ο Hessian πίνακας H είναι μη ομαλός (det(H)=0). Επιπλέον, ακόμα και όταν ο H δεν είναι singular, ο $O(d^3)$ χρόνος που απαιτείται για την αντιστροφή του πίνακα σε κάθε επανάληψη συνηγορεί υπέρ του απλού descent αλγόριθμου. Στην πραγματικότητα, συνήθως απαιτείται λιγότερος χρόνος για να τεθεί το η(k) σε μία σταθερή τιμή η η οποία είναι μικρότερη από την απαιτούμενη και στη συνέχεια να γίνουν οι απαραίτητες διορθώσεις, σε σχέση με τον υπολογισμό του βέλτιστου η(k) σε κάθε βήμα.

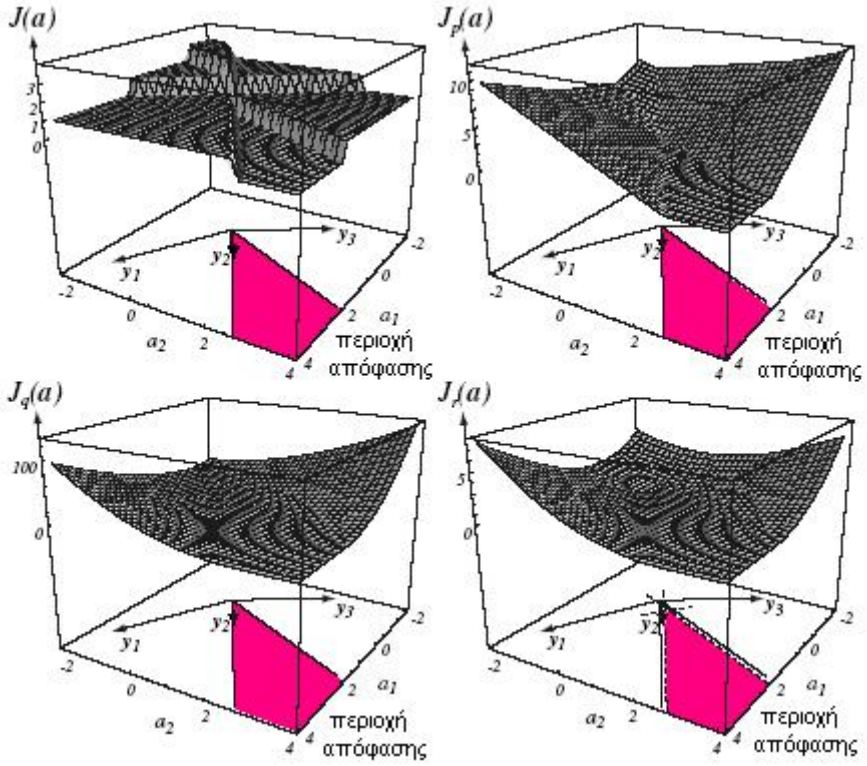
5.5 Ελαχιστοποιώντας τη Συνάρτηση Κριτηρίου του Αλγόριθμου του Perceptron

5.1 Η Συνάρτηση Κριτηρίου του Αλγόριθμου του Perceptron

Έστω το πρόβλημα της κατασκευής μιας συνάρτησης κριτηρίου για την επίλυση των γραμμικών ανισοτήτων $a^t y_i > 0$. Η πιο προφανής επιλογή είναι να οριστεί ως $J(a; y_1, ..., y_n)$ το πλήθος των δειγμάτων που ταξινομούνται λανθασμένα από το a. Παρ' όλ' αυτά, επειδή η παραπάνω συνάρτηση είναι κατά διαστήματα σταθερή, δεν είναι προφανώς κατάλληλη για αναζήτηση κλίσης. Μία καλύτερη επιλογή είναι η συνάρτηση κριτηρίου του αλγόριθμου του Perceptron:

$$J_p(a) = \sum_{y \in Y} \left(-a^t y\right) \tag{5.16}$$

όπου το Y(a) είναι το σύνολο των δειγμάτων που ταξινομούνται λανθασμένα από το a. Εάν όλα τα δείγματα ταξινομούνται ορθά, το Y είναι άδειο και το J_p τίθεται ίσο με το μηδέν. Επειδή ισχύει $a^t y \leq 0$ όταν το y ταξινομείται λανθασμένα, το Jp(a) δεν γίνεται ποτέ αρνητικό ενώ είναι ίσο με το μηδέν μόνο στην περίπτωση όπου το a είναι ένα διάνυσμα λύσης ή βρίσκεται πάνω στο όριο απόφασης. Από γεωμετρική σκοπιά, το $J_p(a)$ είναι ανάλογο με το άθροισμα των αποστάσεων των λανθασμένα ταξινομημένων δειγμάτων από το όριο απόφασης. Στην εικόνα 5.11 παρουσιάζεται η μορφή του J_p για ένα απλό παράδειγμα δύο διαστάσεων.



Εικόνα 5.11: Τέσσερα κριτήρια ως συνάρτηση των βαρών σε ένα γραμμικό ταξινομητή. Στο πάνω αριστερά σχήμα το κριτήριο είναι ο συνολικός αριθμός των μη σωστά ταξινομημένων προτύπων, ο οποίος είναι κατά διαστήματα σταθερός και επομένως μη αποδεκτός για διαδικασίες κλίσης καθόδου. Στο πάνω δεξιό σχήμα, είναι το κριτήριο του Perceptron (εξίσωση 5.16) το οποίο είναι κατά διαστήματα γραμμικό και επομένως αποδεκτό για διαδικασίες κλίσης καθόδου. Το κάτω αριστερά είναι το τετραγωνικό λάθος (εξίσωση 5.32) το οποίο έχει αναλυτικές ιδιότητες και είναι χρήσιμο ακόμα και σε περιπτώσεις όπου τα πρότυπα δεν είναι γραμμικά διαχωρίσιμα. Το κάτω δεξιά σχήμα αντιστοιχεί στο τετραγωνικό λάθος με διάστημα (εξίσωση 5.33).

Επειδή το j-οστό στοιχείο της κλίσης του J_p ισούται με $\partial J_p / \partial a_j$, από την εξίσωση 5.16 προκύπτει ότι

$$\nabla J_p = \sum_{y \in Y} (-y) \tag{5.17}$$

και επομένως ο κανόνας ενημέρωσης παίρνει τη μορφή

$$a(k+1) = a(k) + \eta(k) \sum_{y \in Y_k} y \tag{5.18}$$

όπου το Y_k είναι το σύνολο των δειγμάτων που ταξινομείται λανθασμένα από το a(k). Ο αλγόριθμος του Perceptron είναι ο εξής:

Αλγόριθμος 3: Batch Perceptron

1 αρχή αρχικοποίησε a, κριτήριο θ, η(·), k ← 0

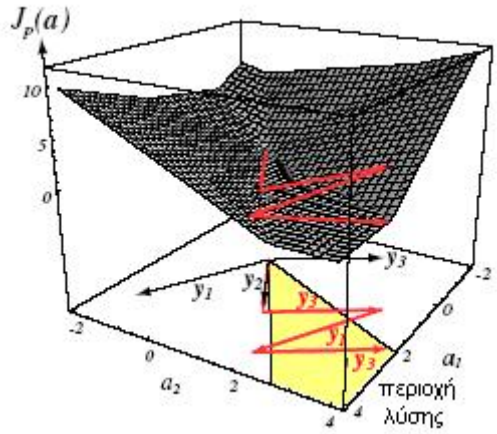
2 κάνε k ← k + 1

3 a ← a + η(k) $\sum_{y \in Y_k} y$

4 μέχρι | η(k) $\sum_{y \in Y_k} y$ | < θ

5 επέστρεψε a
6 τέλος

Επομένως, ο σωρηδόν (batch) αλγόριθμος του Perceptron για την εύρεση ενός διανύσματος λύσης μπορεί να περιγραφεί με τον εξής απλό τρόπο: Το επόμενο διάνυσμα των βαρών υπολογίζεται με την προσθήκη κάποιου πολλαπλασίου του αθροίσματος των λανθασμένα ταξινομημένων δειγμάτων στο τρέχον διάνυσμα των βαρών. Ο όρος σωρηδόν χρησιμοποιείται για να δείξει ότι, στη γενική περίπτωση, χρησιμοποιείται ένα μεγάλο σύνολο δειγμάτων για κάθε βήμα υπολογισμού της ενημέρωσης των βαρών. Στην εικόνα 5.12 φαίνεται ο τρόπος με τον οποίο ο αλγόριθμος καταλήγει σε ένα διάνυσμα λύσης για ένα απλό παράδειγμα δύο διαστάσεων με a(1) = 0 και η(k) = 1. Στη συνέχεια θα αποδειχθεί ότι ο αλγόριθμος του Perceptron καταλήγει σε λύση για κάθε γραμμικά διαχωριζόμενο πρόβλημα.



Εικόνα 5.12: Το κριτήριο του Perceptron, $J_p(a)$, απεικονίζεται γραφικά ως συνάρτηση των βαρών a_1 και a_2 για ένα πρόβλημα τριών προτύπων. Το διάνυσμα των βαρών ξεκινάει από το 0 και ο αλγόριθμος σειριακά προσθέτει σε αυτό διανύσματα ίσα με τα «κανονικοποιημένα» μη σωστά ταξινομημένα. Στο παράδειγμα, η ακολουθία αυτή είναι η y_2, y_3, y_1, y_3 μετά την οποία το διάνυσμα βρίσκεται μέσα στην περιοχή λύσης και η επαναλήψεις τερματίζονται.

5.5.2 Απόδειξη Σύγκλισης για την Μέθοδο Διόρθωσης Ενός Δείγματος

Η διερεύνηση των ιδιοτήτων σύγκλισης του αλγόριθμου του Perceptron θα γίνει θεωρώντας αρχικά μια παραλλαγή η οποία είναι ευκολότερο να αναλυθεί. Αντί να ελέγχεται το a(k) σε όλα τα δείγματα και να βασίζεται η διόρθωση στο σύνολο Y_k των λανθασμένα ταξινομημένων δειγμάτων, τα δείγματα τοποθετούνται σε μία ακολουθία και το διάνυσμα των βαρών διορθώνεται κάθε φορά που ταξινομεί λανθασμένα ένα δείγμα. Για την απόδειξη της σύγκλισης, η φύση της διαδικασίας δεν έχει σημασία, αρκεί όλα δείγματα να εμφανίζονται στην ακολουθία οσοδήποτε συχνά. Ο πιο απλός τρόπος για να σιγουρευτεί αυτό είναι η επανάληψη των δειγμάτων κυκλικά, εάν και η τυχαία επιλογή προτιμάται περισσότερο από πρακτικής πλευράς. Προφανώς, ούτε η σωρηδόν ούτε η ενός δείγματος εκδοχή του αλγόριθμου Perceptron είναι on-line επειδή και στις δύο απαιτείται η αποθήκευση και η πιθανή επαναχρησιμοποίηση όλων των δειγμάτων εκπαίδευσης.

Δύο επιπλέον απλοποιήσεις βοηθούν σημαντικά στην clarify περιγραφή. Αρχικά, η ανάλυση θα περιοριστεί στην περίπτωση όπου το η(k) είναι σταθερό - περίπτωση

σταθερής αύξησης. Είναι προφανές από την εξίσωση 5.18 ότι εάν το η(t) είναι σταθερό, περισσότερο χρησιμεύει για κλιμάκωση των δειγμάτων. Έτσι, στην περίπτωση της σταθερής αύξησης μπορεί χωρίς χάσιμο της γενικότητας να τεθεί η(t) = 1. Η δεύτερη απλοποίηση είναι κυρίως θέμα συμβολισμού. Όταν τα δείγματα θεωρούνται σειριακά, κάποια θα ταξινομηθούν λανθασμένα. Επειδή οι αλλαγές γίνονται μόνο στο διάνυσμα των βαρών όταν υπάρχει κάποιο λάθος, στην πραγματικότητα δε χρειάζεται να ασχοληθεί κανείς με τα λανθασμένα ταξινομημένα δείγματα. Έτσι, η ακολουθία των δειγμάτων συμβολίζεται με τη χρήση εκθέτη - δηλαδή ως $y^1, y^2, ..., y^k, ...$, όπου κάθε y^k είναι κάποιο από τα n δείγματα $y_1, y_2, ..., y_n$ και είναι λανθασμένα ταξινομημένο. Για παράδειγμα, εάν τα δείγματα y_1, y_2 και y_3 θεωρούνται κυκλικά και εάν τα παρακάτω σημειωμένα δείγματα

$$\ddot{y}_1, y_2, \ddot{y}_3, \ddot{y}_1, \ddot{y}_2, y_3, y_1, \ddot{y}_2, ... \tag{5.19}$$

είναι λανθασμένα ταξινομημένα, η ακολουθία $y^1, y^2, y^3, y^4, y^5, ...$ συμβολίζει την ακολουθία $y_1, y_3, y_1, y_2, y_2, ...$ Χρησιμοποιώντας τον παραπάνω συμβολισμό, ο κανόνας της σταθερής αύξησης για τη δημιουργία μιας ακολουθίας διανυσμάτων βαρών μπορεί να γραφεί ως

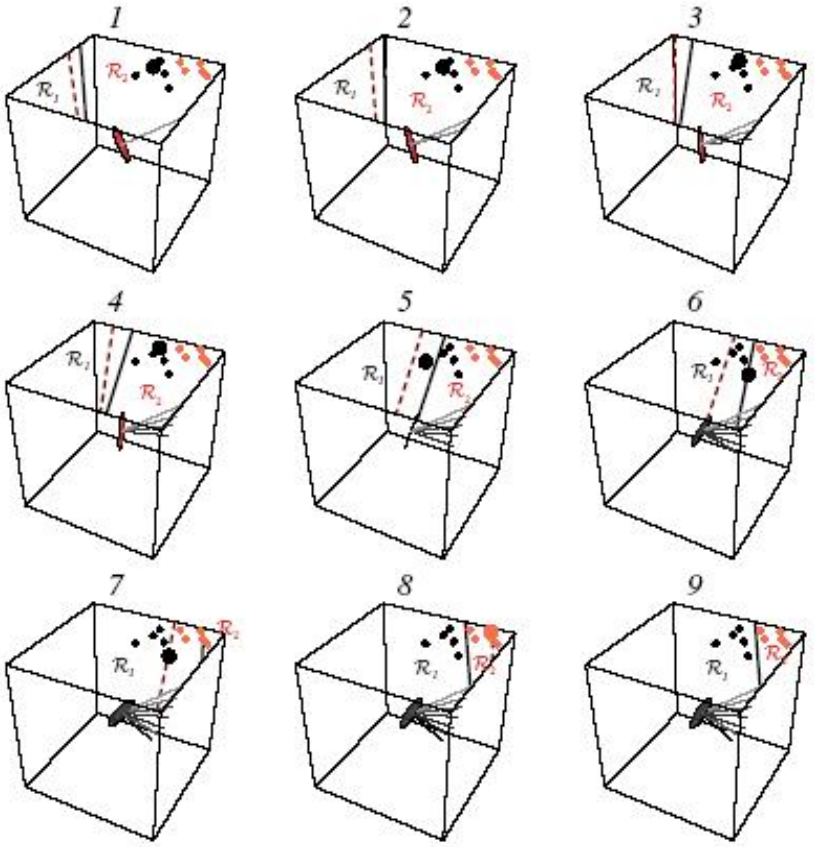
$$\begin{aligned} a(1) &\qquad \tau\upsilon\chi\alpha\acute{\iota}\alpha \\ a(k+1) &= a(k) + y^k \qquad k \geq 1 \end{aligned} \tag{5.20}$$

όπου $a^t(k)y^k \leq 0$ για όλα τα k. Εάν με n συμβολιστεί ο συνολικός αριθμός των προτύπων, ο αλγόριθμος έχει την εξής μορφή:

Αλγόριθμος 4: Σταθερής Αύξησης Ενός Δείγματος Perceptron

```
1 αρχή αρχικοποίησε a, k ← 0
2          κάνε   k ← (k + 1) mod n
3                 εάν το y^k ταξινομηθεί λάθος από το a τότε a ← a + y^k
4          μέχρι  όλα τα πρότυπα να ταξινομηθούν σωστά
5     επέστρεψε a
6 τέλος
```

Ο κανόνας Perceptron σταθερής αύξησης είναι από τους πιο απλούς κανόνες που έχουν προταθεί για την επίλυση συστημάτων γραμμικών ανισώσεων. Από γεωμετρική άποψη, η ερμηνεία του στο χώρο των βαρών είναι ιδιαίτερα προφανής. Επειδή το a(k) ταξινομεί λανθασμένα το y^k, το a(k) δε βρίσκεται στη θετική πλευρά του y^k υπερεπιπέδου $a^t y^k = 0$. Η πρόσθεση του y^k στο a(k) μετακινεί το διάνυσμα των βαρών προς και ίσως κατά μήκος αυτού του υπερεπιπέδου. Ανεξάρτητα από το εάν το υπερεπίπεδο μετακινείται κατά μήκος ή όχι, το νέο εσωτερικό γινόμενο $a^t(k+1)y^k$ είναι μεγαλύτερο από το παλιό εσωτερικό γινόμενο $a^t(k)y^k$ κατά την ποσότητα $\|y^k\|^2$, και επομένως η διόρθωση μετακινεί το διάνυσμα των βαρών προς την "καλή" κατεύθυνση (εικόνα 5.13).



Εικόνα 5.13: Τα δείγματα από δύο κατηγορίες ω_1 και ω_2 απεικονίζονται στον επαυξημένο χώρο των χαρακτηριστικών μαζί με το επαυξημένο διάνυσμα των βαρών a. Σε κάθε βήμα σε έναν κανόνα σταθερής αύξησης, ένα από τα μη σωστά ταξινομημένα πρότυπα, y^k, απεικονίζεται με τη μεγάλη κουκίδα. Μία διόρθωση Δa (ανάλογη με το διάνυσμα y^k) προστίθεται στο διάνυσμα των βαρών – ωθώντας προς ένα ω_1 σημείο ή απομακρύνοντας από ένα ω_2 σημείο. Αυτό αλλάζει το όριο απόφασης από την διακεκομμένη γραμμή στην συνεχή.

Προφανώς ο αλγόριθμος μπορεί να τερματιστεί μόνο εάν τα δείγματα είναι γραμμικά διαχωρίσιμα. Στη συνέχεια θα παρουσιαστεί η απόδειξη της σύγκλισης του αλγόριθμου του Perceptron υπό αυτή τη συνθήκη.

Θεώρημα 5.1: Σύγκλιση του Αλγόριθμου του Perceptron
Εάν τα δείγματα είναι γραμμικά διαχωρίσιμα, τότε η ακολουθία των διανυσμάτων των βαρών που παράγεται από τον 4^o Αλγόριθμο θα τερματίζεται σε ένα διάνυσμα λύσης.

Απόδειξη:
Προφάνως θα δειχθεί ότι κάθε διόρθωση μετακινεί το διάνυσμα των βαρών πιο κοντά στο διάνυσμα λύσης. Δηλαδή, ότι εάν το $\hat{a}$ είναι ένα διάνυσμα λύσης, τότε το $\|a(k+1)-\hat{a}\|$ είναι μικρότερο από το $\|a(k)-\hat{a}\|$. Εάν και αυτό δεν ισχύει γενικότερα (βήματα 6 και 7 της εικόνας 5.13), θα δειχθεί ότι είναι πραγματικότητα για διανύσματα λύσης τα οποία έχουν αρκετά μεγάλο μήκος.

Έστω ότι με $\hat{a}$ συμβολίζεται κάποιο διάνυσμα λύση, τέτοιο ώστε το $\hat{a}^t y_i$ να είναι αυστηρά θετικό για κάθε i, και έστω ότι με α συμβολίζεται ένας θετικός βαθμωτός παράγοντας. Από την εξίσωση 5.20 προκύπτει

$$a(k+1) - \alpha\hat{a} = (a(k) - \alpha\hat{a}) + y^k$$

και επομένως

$$\|a(k+1) - \alpha\hat{a}\|^2 = \|a(k) - \alpha\hat{a}\|^2 + 2(a(k) - \alpha\hat{a})^t y^k + \|y^k\|^2$$

Επειδή το y^k ταξινομήθηκε λανθασμένα, $a^t(k)y^k \leq 0$ και έτσι

$$\|a(k+1) - \alpha\hat{a}\|^2 \leq \|a(k) - \alpha\hat{a}\|^2 - 2(a(k) - \alpha\hat{a})^t y^k + \|y^k\|^2$$

Επειδή το $\hat{a}^t y^k$ είναι αυστηρά θετικό, ο δεύτερος όρος θα υπερισχύσει του τρίτου εάν το α είναι αρκετά μεγάλο. Πιο συγκεκριμένα, εάν το β είναι το μέγιστο μήκος ενός διανύσματος προτύπων,

$$\beta^2 = \max_i \|y_i\|^2 \tag{5.21}$$

και το γ είναι το μικρότερο εσωτερικό γινόμενο του διανύσματος λύσης με οποιοδήποτε διάνυσμα προτύπων, δηλαδή,

$$\gamma = \min_i [\hat{a}^t y_i] > 0 \tag{5.22}$$

τότε προκύπτει η ανίσωση

$$\|a(k+1) - \alpha\hat{a}\|^2 \leq \|a(k) - \alpha\hat{a}\|^2 - 2\alpha\gamma + \beta^2$$

Εάν επιλεγεί

$$\alpha = \frac{\beta^2}{\gamma} \tag{5.23}$$

προκύπτει

$$\|a(k+1) - \alpha\hat{a}\|^2 \leq \|a(k) - \alpha\hat{a}\|^2 - \beta^2$$

Έτσι, η τετραγωνική απόσταση από το a(k) στο $\alpha\hat{a}$ μειώνεται το λιγότερο κατά β^2 σε κάθε επανάληψη. Μετά από k επαναλήψεις προκύπτει

$$\|a(k+1) - \alpha\hat{a}\|^2 \leq \|a(1) - \alpha\hat{a}\|^2 - k\beta^2 \tag{5.24}$$

Επειδή αυτή η τετραγωνική απόσταση δεν μπορεί να γίνει αρνητική, οι διορθώσεις πρέπει να ολοκληρώνονται μετά από k_0 διορθώσεις, όπου

$$k_0 = \frac{\|a(1) - \alpha\hat{a}\|^2}{\beta^2} \tag{5.25}$$

Επειδή μία διόρθωση συμβαίνει κάθε φορά που ένα διάνυσμα ταξινομείται λανθασμένα και επειδή κάθε δείγμα μπορεί να εμφανίζεται οσοδήποτε συχνά στην ακολουθία, συνεπάγεται ότι όταν σταματήσουν οι διορθώσεις το διάνυσμα των βαρών που προκύπτει πρέπει να είναι σε θέση να ταξινομεί όλα τα δείγματα σωστά.

Το k_0 παρέχει ένα άνω όριο για τον αριθμό των επαναλήψεων. Εάν a(1) = 0, προκύπτει η ακόλουθη ιδιαίτερα απλή έκφραση για το k_0:

$$k_0 = \frac{\alpha^2\|\hat{a}\|^2}{\beta^2} = \frac{\beta^2\|\hat{a}\|^2}{\gamma^2} \frac{\max_i \|y_i\|^2 \|\hat{a}\|^2}{\min_i [y_i^t \hat{a}]^2} \tag{5.26}$$

Ο dominator στην εξίσωση 5.26 δείχνει ότι η δυσκολία του προβλήματος καθορίζεται κυρίως από τα δείγματα που είναι σχεδόν ορθοκανονικά προς το διάνυσμα λύσης. Δυστυχώς, δεν παρέχει καμία βοήθεια όταν αντιμετωπίζεται ένα άλυτο πρόβλημα, διότι στην περίπτωση αυτή το όριο εκφράζεται με όρους ενός άγνωστου διανύσματος

λύσης. Είναι προφανές ότι τα γραμμικά διαχωρίσιμα προβλήματα μπορούν να γίνουν πολύ δύσκολα εάν τα δείγματα γίνουν σχεδόν συνεπίπεδα. Παρ' όλ' αυτά, εάν τα δείγματα εκπαίδευσης είναι γραμμικά διαχωρίσιμα, ο κανόνας της σταθερής αύξησης θα οδηγήσει σε λύση μετά από ένα πεπερασμένο αριθμό βελτιώσεων.

5.5.3 Κάποιες Άμεσες Γενικεύσεις

Ο κανόνας της σταθερής αύξησης μπορεί να γενικευτεί για να παραχθεί μια ποικιλία από σχετικούς αλγόριθμους. Θα γίνει αναφορά σε δύο παραλλαγές που εμφανίζουν ιδιαίτερο ενδιαφέρον. Η πρώτη παραλλαγή εισάγει μια μεταβλητή αύξηση η(k) και ένα όριο b και εκτελεί μια διόρθωση κάθε φορά που το $a^t(k)y^k$ αποτυγχάνει να υπερβεί το όριο. Η εξίσωση ενημέρωσης των βαρών παίρνει τη μορφή

$$\begin{array}{ll} a(1) & \text{τυχαία} \\ a(k+1) = a(k) + \eta y^k & k \geq 1 \end{array} \tag{5.27}$$

όπου το $a^t(k)y^k \leq b$ για κάθε k. Έτσι, για n πρότυπα, ο αλγόριθμος παίρνει τη μορφή:

Αλγόριθμος 5: Μεταβλητής Αύξησης Perceptron με Διάστημα
1 αρχή αρχικοποίησε a, κριτήριο θ, διάστημα b, η(·), k ← 0
2 κάνε k ← (k + 1) mod n
3 εάν $a^t y^k \leq b$ τότε a ← a + η(k)y^k
4 μέχρι $a^t y^k > b$ για όλα τα k
5 επέστρεψε a
6 τέλος

Μπορεί να δειχθεί ότι εάν τα δείγματα είναι γραμμικά διαχωρίσιμα και εάν

$$\eta(k) \geq 0 \tag{5.28}$$

$$\lim_{m\to\infty} \sum_{k=1}^{m} \eta(k) = \infty \tag{5.29}$$

και

$$\lim_{m\to\infty} \frac{\sum_{k=1}^{m} \eta^2(k)}{\left(\sum_{k=1}^{m} \eta^2(k)\right)^2} = 0 \tag{5.30}$$

τότε το a(k) συγκλίνει σε ένα διάνυσμα λύσης a το οποίο ικανοποιεί τη σχέση $a^t y_i > b$ για κάθε i. Πιο συγκεκριμένα, αυτές οι συνθήκες για το η(k) ικανοποιούνται εάν το η(k) είναι μία θετική σταθερά ή εάν μειώνεται αντιστρόφως ανάλογα με το k.
Μία άλλη παραλλαγή που παρουσιάζει ενδιαφέρον είναι ο αρχικός αλγόριθμος κλίσης καθόδου για το J_p,

$$\begin{array}{ll} a(1) & \text{τυχαία} \\ a(k+1) = a(k) + \eta(k) \sum_{y\in Y_k} y & k \geq 1 \end{array} \tag{5.31}$$

όπου το Y_k είναι το σύνολο των δειγμάτων εκπαίδευσης που ταξινομούνται λανθασμένα από το a(k). Είναι εύκολο να δει κανείς ότι ο αλγόριθμος αυτός θα οδηγήσει επίσης σε μία λύση, αρκεί να παρατηρήσει ότι εάν το â είναι ένα διάνυσμα λύσης για τα $y_1, ..., y_n$ τότε θα ταξινομεί σωστά και το διάνυσμα διόρθωσης

$$y^k = \sum_{y\in Y_k} y$$

Ο αλγόριθμος παρατίθεται με περισσότερες λεπτομέρειες στη συνέχεια:

Αλγόριθμος 6: Μεταβλητής Αύξησης Batch Perceptron

1 αρχή αρχικοποίησε a, η(·), k ← 0
2 κάνε k ← (k + 1) mod n
3 $Y_k = \{\}$
4 j ← 0
5 κάνε j ← j + 1
6 εάν το y_j ταξινομείται λάθος τότε βάλε το y_j στο Y_k
7 μέχρι j = n
8 $a \leftarrow a + \eta(k) \sum_{y \in Y_k} y$
9 μέχρι $Y_k = \{\}$
10 επέστρεψε a
11 τέλος

Το πλεονέκτημα του σωρηδόν αλγόριθμου κλίσης καθόδου είναι ότι η τροχιά του διανύσματος λύσης είναι εξομαλυμμένη σε σχέση με αυτή που αντιστοιχεί στους διάφορους ενός δείγματος αλγόριθμους, επειδή σε κάθε ενημέρωση χρησιμοποιείται το πλήρες σύνολο των λανθασμένα ταξινομημένων προτύπων. Έτσι, εάν τα δείγματα είναι γραμμικά διαχωρίσιμα, όλα τα πιθανά διανύσματα διόρθωσης σχηματίζουν ένα γραμμικά διαχωρίσιμο σύνολο, και εάν το η(k) ικανοποιεί τις εξισώσεις 5.28 - 5.30, η ακολουθία των διανυσμάτων των βαρών που παράγεται από τον αλγόριθμο κλίσης καθόδου για το $J_p(\cdot)$ θα συγκλίνει πάντα σε ένα διάνυσμα λύσης. Είναι ενδιαφέρον να σημειωθεί ότι οι συνθήκες για το η(k) ικανοποιούνται εάν το η(k) είναι μία θετική σταθερά, ή εάν μειώνεται αντιστρόφως ανάλογα με το k, ή ακόμα και εάν αυξάνεται ανάλογα με το k. Γενικότερα, είναι προτιμότερο να μειώνεται το η(k) με το πέρασμα του χρόνου. Αυτό ισχύει κυρίως για περιπτώσεις όπου υπάρχει υπόνοια ότι το σύνολο των δειγμάτων δεν είναι γραμμικά διαχωρίσιμο, επειδή ακριβώς μειώνει τα καταστροφικά αποτελέσματα κάποιων "κακών" δειγμάτων. Πάντως, στη γραμμικά διαχωρίσιμη περίπτωση, το η(k) μπορεί να αυξάνεται και παρ' όλ' αυτά να οδηγηθεί ο αλγόριθμος σε λύση.

Η παρατήρηση αυτή φέρνει στην επιφάνεια μία από τις βασικές διαφορές ανάμεσα στη θεωρία και την πράξη. Από θεωρητικής πλευράς, είναι ενδιαφέρον ότι μπορεί να προκύψει λύση σε ένα πεπερασμένο αριθμό βημάτων για οποιοδήποτε πεπερασμένο σύνολο διαχωρίσιμων δειγμάτων, για οποιοδήποτε αρχικό διάνυσμα βαρών a(1), για οποιοδήποτε μη αρνητικό διάστημα b και για οποιοδήποτε κλιμακωτό παράγοντα η(k) που ικανοποιεί τις εξισώσεις 5.28 - 5.30. Από πρακτικής άποψης, πρέπει να γίνουν όσο το δυνατόν καλύτερες επιλογές για αυτές τις ποσότητες. Θεωρείστε για παράδειγμα το διάστημα b. Εάν το b είναι πολύ μικρότερο από το $\eta(k)\left\|y^k\right\|^2$, δηλαδή την ποσότητα κατά την οποία μεταβάλλει μία διορθωτική κίνηση το $a^t(k)y^k$, είναι προφανές ότι θα επηρεάζει πολύ λίγο τη διαδικασία μάθησης, εάν την επηρεάζει. Εάν είναι πολύ μεγαλύτερο από το $\eta(k)\left\|y^k\right\|^2$, θα απαιτηθούν πολλές διορθώσεις για να ικανοποιηθεί η συνθήκη $a^t(k)y^k > b$. Μία τιμή κοντά στο $\eta(k)\left\|y^k\right\|^2$ αποτελεί συχνά μία καλή επιλογή. Εκτός από την επιλογή των τιμών για το η(k) και το b, τα αποτελέσματα μπορεί επίσης να επηρεαστούν σημαντικά από την κλιμάκωση των στοιχείων του y^k. Προφανώς, η ύπαρξη ενός θεωρήματος σύγκλισης δεν απαλείφει τους περιορισμούς που πρέπει να ληφθούν υπόψη όταν εφαρμόζονται αυτές οι τεχνικές.

Ένας κοντινός απόγονος του αλγόριθμου του Perceptron είναι ο αλγόριθμος του Winnow, ο οποίος βρίσκει εφαρμογή σε διαχωρίσιμα δεδομένα εκπαίδευσης. Η βασική διαφορά τους είναι ότι ενώ στον αλγόριθμο του Perceptron το διάνυσμα των βαρών έχει ως στοιχεία τα α_i (i = 0, ...,d), στον αλγόριθμο του Winnow τα στοιχεία αυτά είναι κλιμακούμενα με βάση το $2\sinh[\alpha_i]$. Σε μία εκδοχή, τον εξισορροπημένο αλγόριθμο του Winnow, υπάρχουν ξεχωριστά "θετικά" και "αρνητικά" διανύσματα βαρών, a^+ και a^-, και το καθένα σχετίζεται με μία από τις δύο υπάρχουσες κατηγορίες ταξινόμησης. Διορθώσεις στο διάνυσμα των θετικών βαρών γίνονται εάν και μόνο εάν ένα πρότυπο εκπαίδευσης της ω_1 έχει ταξινομηθεί λανθασμένα. Αντίστοιχα, διορθώσεις στο διάνυσμα των αρνητικών βαρών γίνονται εάν και μόνο εάν ένα πρότυπο εκπαίδευσης της ω_2 έχει ταξινομηθεί λανθασμένα.

Αλγόριθμος 7: Balanced Winnow
1 αρχή αρχικοποίησε a^+, a^-, $\eta(\cdot)$, $k \leftarrow 0$, $\alpha > 1$
2 εάν $\text{sgn}[a^{+t}y_k - a^{-t}y_k] \neq z_k$ (λάθος ταξινόμηση)
3 τότε εάν $z_k = +1$ τότε $a_i^+ \leftarrow a^{+y_i}a_i^+$; $a_i^- \leftarrow a^{-y_i}a_i^-$ για όλα τα i
4 εάν $z_k = -1$ τότε $a_i^+ \leftarrow a^{-y_i}a_i^+$; $a_i^- \leftarrow a^{+y_i}a_i^-$ για όλα τα i
5 επέστρεψε a^+, a^-
6 τέλος

Αυτή η εκδοχή του αλγόριθμου του Winnow παρουσιάζει δύο σημαντικά πλεονεκτήματα. Το πρώτο είναι ότι κατά τη διάρκεια της εκπαίδευσης και τα δύο διανύσματα των βαρών κινούνται ομοιόμορφα και επομένως, για διαχωρίσιμα δεδομένα, η απόσταση, που ορίζεται από τα δύο αυτά διανύσματα, παραμένει σταθερή. Αυτή η διαπίστωση οδηγεί σε μια απόδειξη σύγκλισης η οποία είναι πιο γενική από το θεώρημα σύγκλισης του Perceptron. Το δεύτερο πλεονέκτημα είναι ότι η σύγκλιση επιτυγχάνεται γρηγορότερα σε σχέση με τον αλγόριθμο του Perceptron, επειδή εάν επιλεγεί ο κατάλληλος ρυθμός μάθησης, κάθε constituent βάρος δεν overshoot την τελική του τιμή. Αυτό το πλεονέκτημα είναι ιδιαίτερα εμφανές όταν υπάρχει μεγάλος αριθμός από μη σχετικά ή πλεονάζοντα χαρακτηριστικά.

5.6 Διαδικασίες Χαλάρωσης (Relaxation)

Στις προηγούμενες ενότητες παρουσιάστηκε ο τρόπος με τον οποίο ένας γραμμικός ταξινομητής εκπαιδεύεται μέσω της ελαχιστοποίησης του κριτηρίου του Perceptron της εξίσωσης 5.16. Αυτή η προσέγγιση μπορεί να γενικευτεί για να περιλαμβάνει μία ευρύτερη κλάση συναρτήσεων κριτηρίου και μεθόδων για την ελαχιστοποίησή τους.

5.6.1 Ο Αλγόριθμος Descent

Η συνάρτηση κριτηρίου $J_p(\cdot)$ είναι αδιαμφισβήτητα η μοναδική συνάρτηση που μπορεί να κατασκευαστεί έτσι ώστε να ελαχιστοποιείται όταν το a αποτελεί ένα διάνυσμα λύσης. Μία αντίστοιχη διακριτή συνάρτηση είναι η

$$J_q(a) = \sum_{y \in Y} \left(a^t y\right)^2 \tag{5.32}$$

όπου το Y(a) δηλώνει το σύνολο των δειγμάτων εκπαίδευσης που ταξινομούνται λανθασμένα από το a. Τόσο το J_p όσο και το J_q δίνουν έμφαση στα μη σωστά ταξινομημένα δείγματα. Η βασική διαφορά τους είναι ότι η κλίση του J_q είναι συνεχής ενώ η κλίση του J_p δεν είναι. Έτσι, το J_q αντιπροσωπεύει μια πιο ομαλή επιφάνεια αναζήτησης (εικόνα 5.11). Δυστυχώς, το J_q είναι τόσο ομαλό κοντά στο

όριο της περιοχής λύσης ώστε η ακολουθία των διανυσμάτων των βαρών να υπάρχει περίπτωση να συγκλίνει σε ένα σημείο πάνω στο όριο. Είναι εντελώς άσκοπο να χαθεί υπολογιστικός χρόνος για να βρεθεί το οριακό σημείο a = 0. Ένα άλλο πρόβλημα με το J_q είναι ότι η τιμή του μπορεί να καθοριστεί αποκλειστικά από τα διανύσματα δειγμάτων με το μεγαλύτερο μήκος. Τα δύο αυτά προβλήματα αντιμετωπίζονται εάν χρησιμοποιηθεί η ακόλουθη συνάρτηση κριτηρίου:

$$J_r(a) = \frac{1}{2}\sum_{y\in Y}\frac{\left(a^t y - b\right)^2}{\|y\|^2} \tag{5.33}$$

όπου το Y(a) είναι το σύνολο των δειγμάτων για τα οποία $a^t y \le b$. Εάν το Y(a) είναι άδειο, το J_r ορίζεται ίσο με το μηδέν. Έτσι, το $J_r(a)$ δεν είναι ποτέ αρνητικό, ενώ είναι ίσο με το μηδέν εάν και μόνο εάν $a^t y \le b$ για όλα τα δείγματα εκπαίδευσης. Η κλίση του J_r δίνεται από

$$\nabla J_r = \sum_{y\in Y}\frac{a^t y - b}{\|y\|^2} y$$

ενώ ο κανόνας ενημέρωσης είναι ο

$$a(1) \qquad \textit{τυχαία}$$
$$a(k+1) = a(k) + \eta(k)\sum_{y\in Y_k}\frac{b - a^t y}{\|y\|^2} y \qquad k \ge 1 \tag{5.34}$$

Έτσι, ο αλγόριθμος χαλάρωσης παίρνει τη μορφή:

Αλγόριθμος 8: Σωρηδόν Χαλάρωση με Διάστημα
1 αρχή αρχικοποίησε a, η(·), b, k ← 0
2 κάνε k ← (k + 1) mod n
3 Y_k = {}
4 j ← 0
5 κάνε j ← j + 1
6 εάν $a^t y^j \le b$ τότε βάλε το y^j στο Y_k
7 μέχρι j = n
8 $a \leftarrow a + \eta(k)\sum_{y\in Y_k}\frac{b - a^t y}{\|y\|^2} y$
9 μέχρι Y_k = {}
10 επέστρεψε a
11 τέλος

Όπως και προηγουμένως, είναι ευκολότερο να αποδειχθεί η σύγκλιση όταν τα δείγματα αντιμετωπίζονται ένα κάθε χρονική στιγμή παρά όλα μαζί. Επίσης, το ενδιαφέρον θα περιοριστεί στην περίπτωση της σταθερής αύξησης, η(k) = η. Επομένως, σχηματίζεται πάλι μία ακολουθία $y_1, y_2, \ldots$ η οποία αποτελείται από τα δείγματα που χρησιμοποιούνται για τη διόρθωση του διανύσματος των βαρών. Ο κανόνας διόρθωσης του μονού δείγματος, σε αντιστοιχία με την εξίσωση 5.33 είναι

$$a(1) \qquad \textit{τυχαία}$$
$$a(k+1) = a(k) + \eta\frac{b - a^t(k)y^k}{\|y^k\|^2} y^k \qquad k \ge 1 \tag{5.35}$$

όπου $a^t(k)y^k \le b$ για όλα τα k. Ο αλγόριθμος έχει την παρακάτω μορφή:

Αλγόριθμος 9: Ενός Δείγματος Χαλάρωση με Διάστημα

1 αρχή αρχικοποίησε a, η(·), k ← 0

2 κάνε k ← (k + 1) mod n

3 εάν $a^t y^k \le b$ τότε $a \leftarrow a + \eta(k)\dfrac{b - a^t y}{\left[y^k\right]^2} y^k$

4 μέχρι $a^t y^k > b$ για όλα τα y^k

5 επέστρεψε a

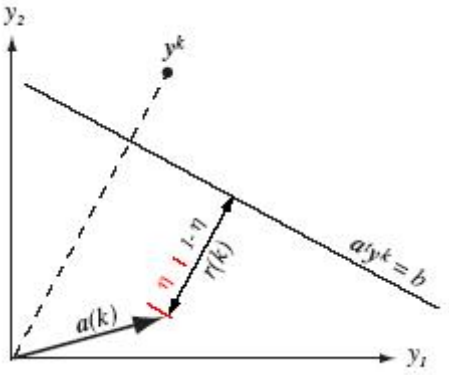
6 τέλος

Ο παραπάνω αλγόριθμος είναι γνωστός ως ο ενός δείγματος κανόνας χαλάρωσης με διάστημα και έχει μια απλή γεωμετρική ερμηνεία. Η ποσότητα

$$r(k) = \frac{b - a^t(k) y^k}{\left\| y^k \right\|} \tag{5.36}$$

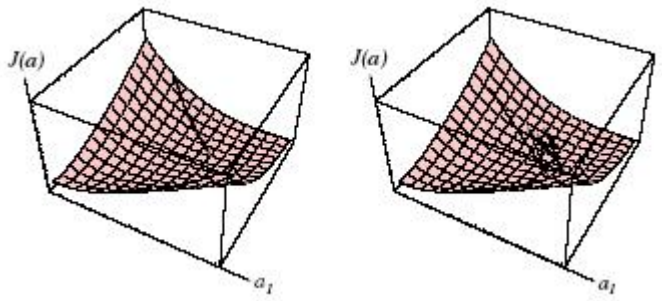
είναι η απόσταση από το a(k) στο υπερεπίπεδο $a^t y^k = b$. Επειδή το $y^k / \left\| y^k \right\|$ αποτελεί το μοναδιαίο κανονικό διάνυσμα για το υπερεπίπεδο, η εξίσωση 5.35 αναγκάζει το a(k) να μετατοπιστεί κατά ένα συγκεκριμένο κλάσμα η της απόστασης του a(k) από το υπερεπίπεδο. Εάν η = 1, το a(k) μετακινείται ακριβώς προς το υπερεπίπεδο, έτσι ώστε η ένταση που δημιουργήθηκε από την ανίσωση $a^t(k) y^k \le b$ να χαλαρώνει (εικόνα 5.14). Από την εξίσωση 5.35, μετά από μια μικρή διόρθωση, προκύπτει

$$a^t(k+1) y^k - b = (1 - \eta)\left(a^t(k) y^k - b\right) \tag{5.37}$$



Εικόνα 5.14: Σε κάθε βήμα του βασικού αλγόριθμου χαλάρωσης, το διάνυσμα των βαρών μετακινείται κατά η φορές την απόσταση προς το υπερεπίπεδο που ορίζεται από την $a^t y^k = b$.

Εάν η < 1, τότε το $a^t(k+1) y^k$ είναι μικρότερο από το b, ενώ εάν η > 1, το $a^t(k+1) y^k$ είναι μεγαλύτερο. Αυτές οι συνθήκες είναι γνωστές ως υποχαλάρωση (underrelaxation) ή υπερχαλάρωση (overrelaxation) αντίστοιχα. Γενικότερα, το η θα περιοριστεί στο διάστημα 0 < η < 2 (εικόνες 5.14 και 5.15).



Εικόνα 5.15: Στο αριστερό σχήμα, υποχαλάρωση (η < 1) οδηγεί σε μη απαραίτητη αργή σύγκλιση, ακόμα και σε αποτυχία. Στο δεξιό σχήμα, η υπερχαλάρωση (1 < η < 2) οδηγεί σε βεβιασμένη προσπάθεια για σύγκλιση. Παρ' όλ' αυτά, η σύγκλιση θα επιτευχθεί τελικά.

5.6.2 Απόδειξη Σύγκλισης

Όταν εφαρμόζεται ο κανόνας χαλάρωσης σε ένα σύνολο γραμμικά διαχωρίσιμων δειγμάτων, ο αριθμός των διορθώσεων μπορεί να είναι αλλά μπορεί και να μην είναι πεπερασμένος. Στην περίπτωση που είναι πεπερασμένος, τότε προφανώς έχουμε καταλήξει σε ένα διάνυσμα λύσης. Εάν δεν είναι πεπερασμένος, το a(k) συγκλίνει σε ένα οριακό διάνυσμα που βρίσκεται στο όριο της περιοχής λύσης. Επειδή η περιοχή στην οποία ισχύει $a^t y \geq b$ περιέχεται σε μία μεγαλύτερη περιοχή για την οποία ισχύει $a^t y>0$ εάν b>0, αυτό υποδηλώνει ότι το a(k) θα βρεθεί σε αυτή τη μεγαλύτερη περιοχή τουλάχιστον μία φορά, και θα παραμείνει εκεί για όλες τις τιμές του k που είναι μεγαλύτερες από κάποια καθορισμένη τιμή k_0.

Η απόδειξη βασίζεται στο ότι εάν το $\hat{a}$ είναι οποιοδήποτε διάνυσμα στην περιοχή λύσεων, δηλαδή οποιοδήποτε διάνυσμα που ικανοποιεί τη σχέση $\hat{a}^t y_i > b$ για όλα τα i, τότε σε κάθε βήμα, το a(k) πλησιάζει στο $\hat{a}$. Αυτό προκύπτει άμεσα από την εξίσωση 5.35 επειδή

$$\left\|a(k+1)-\hat{a}\right\|^2 = \left\|a(k)-\hat{a}\right\|^2 - 2\eta\frac{\left(b-a^t(k)y^k\right)}{\left\|y^k\right\|^2}(\hat{a}-a(k))^t y^k + \eta^2\frac{\left(b-a^t(k)y^k\right)^2}{\left\|y^k\right\|^2} \quad (5.38)$$

και

$$(\hat{a}-a(k))^t y^k > b-a^t(k)y^k \geq 0 \quad (5.39)$$

έτσι ώστε

$$\left\|a(k+1)-\hat{a}\right\|^2 \leq \left\|a(k)-\hat{a}\right\|^2 - \eta(2-\eta)\frac{\left(b-a^t(k)y^k\right)^2}{\left\|y^k\right\|^2} \quad (5.40)$$

Επειδή το η έχει περιοριστεί στο διάστημα $0<\eta<2$, προκύπτει ότι $\left\|a(k+1)-\hat{a}\right\| \leq \left\|a(k)-\hat{a}\right\|$. Έτσι, τα διανύσματα της ακολουθίας a(1), a(2), ... πλησιάζουν όλο και περισσότερο στο $\hat{a}$. Οριακά, για k τείνοντος στο άπειρο η απόσταση $\left\|a(k+1)-\hat{a}\right\|$ πλησιάζει μια οριακή απόσταση $r(\hat{a})$. Αυτό σημαίνει ότι για k τείνοντος στο άπειρο, το a(k) περιορρίζεται στην επιφάνεια μιας υπερσφαίρας με κέντρο το $\hat{a}$ και ακτίνα το $r(\hat{a})$. Επειδή αυτό ισχύει για οποιοδήποτε $\hat{a}$ που βρίσκεται μέσα στην περιοχή λύσης, το οριακό a(k) περιορίζεται στην τομή των υπερσφαιρών που έχουν ως κέντρα όλα τα πιθανά διανύσματα λύσης.

Στην συνέχεια θα αποδειχθεί ότι η τομή αυτών των υπερσφαιρών είναι ένα σημείο πάνω στο όριο της περιοχής λύσεων. Έστω αρχικά ότι υπάρχουν τουλάχιστον δύο διαφορετικά σημεία a' και a'' πάνω στην τομή των υπερσφαιρών. Τότε $\|a' - \hat{a}\| = \|a'' - \hat{a}\|$ για κάθε $\hat{a}$ που βρίσκεται στην περιοχή λύσεων. Αυτό όμως υποδηλώνει ότι η περιοχή λύσεων περιέχεται στο $(\hat{d} - 1)$-διάστατο υπερεπίπεδο των σημείων που απέχουν ίση απόσταση από τα a' και a''. Όμως είναι γνωστό ότι η περιοχή λύσεων είναι $\hat{d}$-διάστατη. Εάν $\hat{a}^t y_i > 0$ για i = 1, ..., n, τότε για οποιοδήποτε $\hat{d}$-διάστατο διάνυσμα v, ισχύει $(\hat{a} + \varepsilon v)^t y > 0$ για i = 1, ..., n εάν το ε είναι αρκετά μικρό. Επομένως, το a(k) συγκλίνει σε ένα μοναδικό σημείο a. Αυτό το σημείο βεβαίως δεν είναι μέσα στην περιοχή λύσεων, διότι εάν ήταν η ακολουθία θα ήταν πεπερασμένη. Δεν είναι όμως ούτε έξω από αυτή, διότι κάθε διόρθωση αναγκάζει το διάνυσμα των βαρών να μετακινείται η φορές την απόστασή του από το επίπεδο του ορίου, γεγονός που δεν επιτρέπει στο διάνυσμα να είναι φραγμένο μακριά από το όριο για πάντα. Επομένως, το οριακό σημείο πρέπει να βρίσκεται πάνω στο όριο.

5.7 Μη Διαχωρίσιμη Συμπεριφορά

Ο αλγόριθμος του Perceptron και οι υπόλοιπες διαδικασίες χαλάρωσης παρέχουν έναν αριθμό από απλές μεθόδους για την εύρεση ενός διανύσματος που έχει την ικανότητα να διαχωρίζει δείγματα που είναι γραμμικά διαχωρίσιμα. Αυτές οι μέθοδοι καλούνται διαδικασίες διόρθωσης λάθους επειδή προκαλούν μια τροποποίηση του διανύσματος των βαρών όταν και μόνο όταν εμφανίζεται κάποιο λάθος. Η επιτυχία τους στην επίλυση γραμμικά διαχωρίσιμων προβλημάτων οφείλεται κυρίως σε αυτή την αέναη αναζήτηση για μία απαλλαγμένη από λάθη λύση. Πρακτικά, αυτές οι μέθοδοι μπορούν να χρησιμοποιηθούν μόνο στην περίπτωση όπου υπάρχει πεποίθηση ότι ο ρυθμός λάθους για τη βέλτιστη γραμμική διακρίνουσα συνάρτηση είναι πολύ μικρός.

Φυσικά, ακόμα και στην περίπτωση όπου προκύπτει ένα διάνυσμα διαχωριστής για τα δείγματα εκπαίδευσης, αυτό δε συνεπάγεται ότι ο αντίστοιχος ταξινομητής θα έχει καλή απόδοση σε κάποια άλλα ανεξάρτητα δεδομένα ελέγχου. Στην πραγματικότητα, οποιοδήποτε σύνολο με λιγότερα από $2\hat{d}$ δείγματα έχει πάρα πολλές πιθανότητες να είναι γραμμικά διαχωρίσιμο. Έτσι, πρέπει να γίνεται προσπάθεια ώστε ο ταξινομητής που προκύπτει να έχει εξίσου καλή απόδοση και στα δεδομένα εκπαίδευσης και στα δεδομένα ελέγχου. Δυστυχώς, εάν ένα σύνολο σχεδιασμού είναι αρκετά μεγάλο είναι σχεδόν σίγουρο ότι δεν είναι γραμμικά διαχωρίσιμο. Αυτό σημαίνει ότι είναι χρήσιμο να είναι γνωστή η συμπεριφορά των διαδικασιών διόρθωσης όταν αυτές εφαρμόζονται σε μη γραμμικά διαχωρίσιμα δείγματα.

Επειδή, εξ ορισμού, κανένα διάνυσμα βαρών δεν μπορεί να ταξινομήσει σωστά όλα τα δείγματα ενός μη γραμμικά διαχωρίσιμου συνόλου, είναι προφανές ότι οι διορθώσεις που κάνει μια διαδικασία διόρθωσης λαθών δεν μπορούν να σταματήσουν ποτέ. Κάθε αλγόριθμος παράγει μία μη πεπερασμένη ακολουθία από διανύσματα βαρών, καθένα από τα οποία μπορεί να παρέχει αλλά μπορεί και να μην παρέχει μία χρήσιμη λύση. Η ακριβής συμπεριφορά αυτών των κανόνων, σε περιπτώσεις μη γραμμικά διαχωρίσιμες, έχει μελετηθεί επαρκώς για κάποιες ειδικές περιπτώσεις. Είναι γνωστό για παράδειγμα, ότι το μήκος των διανυσμάτων των βαρών που παράγονται από τον κανόνα σταθερής αύξησης είναι φραγμένο. Πολλοί εμπειρικοί κανόνες για τον τερματισμό της διαδικασίας διόρθωσης βασίζονται πάνω σ' αυτήν ακριβώς την τάση που έχει το μήκος του διανύσματος των βαρών να κυμαίνεται

κοντά σε μία οριακή τιμή. Από θεωρητικής άποψης, εάν τα στοιχεία των δειγμάτων είναι ακέραιοι αριθμοί, η καθορισμένης αύξησης ρουτίνα καταλήγει σε μία διαδικασία πεπερασμένης κατάστασης. Εάν η διαδικασία διόρθωσης τερματίζεται σε κάποιο αυθαίρετο σημείο, το διάνυσμα των βαρών μπορεί να μη βρίσκεται σε κάποια καλή κατάσταση. Παίρνοντας το μέσο όρο των διανυσμάτων των βαρών που παράγονται από τον κανόνα διόρθωσης, ελαχιστοποιείται η πιθανότητα να καταλήξει η διαδικασία σε μία κακή λύση επειδή επιλέχτηκε μια λανθασμένη στιγμή τερματισμού.
Ένας σημαντικός αριθμός από ευρετικές παραλλαγές στους υπάρχοντες κανόνες διόρθωσης λαθών έχουν προταθεί και μελετηθεί εμπειρικά. Ο σκοπός των παραλλαγών αυτών είναι να προκύψει αποδεκτή απόδοση στα γραμμικά μη διαχωρίσιμα προβλήματα ενώ παράλληλα θα διατηρείται η ικανότητα εύρεσης ενός διανύσματος διαχωριστή για τα γραμμικά διαχωρίσιμα προβλήματα. Μια συνήθης τακτική είναι η χρήση μιας μεταβλητής τιμής για το η(k), με το η(k) να τείνει στο μηδέν όταν το k τείνει στο άπειρο. Ο ρυθμός με τον οποίο το η(k) πλησιάζει το μηδέν είναι ιδιαίτερα σημαντικός. Εάν είναι πολύ μικρός, τα αποτελέσματα θα παραμένουν ευαίσθητα σε αυτά τα δείγματα εκπαίδευσης που κάνουν το σύνολο γραμμικά μη διαχωρίσιμο. Εάν είναι πολύ μεγάλος, το διάνυσμα των βαρών μπορεί να συγκλίνει πολύ νωρίς καταλήγοντας σε μη βέλτιστη λύση. Μια άλλη επιλογή για την τιμή του η(k) είναι να είναι μία συνάρτηση της πρόσφατης απόδοσης και συγκεκριμένα να ελαττώνεται καθώς η απόδοση βελτιώνεται. Μια άλλη επιλογή είναι το η(k) να ισούται με η(1) / k.

5.8 Διαδικασίες Ελάχιστου Τετραγωνικού Λάθους

Η συναρτήσεις κριτηρίου που αναφέρθηκαν μέχρι τώρα βασίζονται στη διόρθωση των λανθασμένα ταξινομημένων δειγμάτων. Στην ενότητα αυτή θα παρουσιαστεί μία συνάρτηση κριτηρίου που περιλαμβάνει όλα τα δείγματα. Εκεί που προηγουμένως υπήρχε ένα διάνυσμα βαρών a που έκανε όλα τα εσωτερικά διανύσματα $a^t y_i$, τώρα θα ισχύει $a^t y_i = b_i$, όπου τα b_i είναι κάποιες αυθαίρετα ορισμένες θετικές σταθερές. Έτσι, το πρόβλημα της εύρεσης λύσης σε ένα σύνολο από γραμμικές ανισώσεις αντικαθίσταται από το πρόβλημα εύρεσης λύσης σε ένα σύνολο από γραμμικές εξισώσεις.

5.8.1 Ελάχιστο Τετραγωνικό Λάθος και Ψευδοαντίστροφος Πίνακας

Η αντιμετώπιση γραμμικών εξισώσεων απλοποιείται με τη χρήση πινάκων. Έστω ότι με Y συμβολίζεται ο $n \times \hat{d}$ πίνακας ($\hat{d} = d + 1$) του οποίου η i-οστή σειρά είναι το διάνυσμα y_i^t και έστω ότι με b συμβολίζεται το διάνυσμα στήλη $b = (b_1, ..., b_n)^t$. Τότε, το πρόβλημα απλοποιείται στην εύρεση ενός διανύσματος βαρών a που να ικανοποιεί τη σχέση

$$\begin{pmatrix} y_{10} & y_{11} & \cdots & y_{1d} \\ y_{20} & y_{21} & \cdots & y_{2d} \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ y_{n0} & y_{n1} & \cdots & y_{nd} \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \cdots \\ \alpha_d \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \cdots \\ \cdots \\ \cdots \\ b_n \end{pmatrix} \quad \text{ή} \quad Ya = b \tag{5.41}$$

Εάν ο Y ήταν ομαλός, η λύση θα προέκυπτε άμεσα και θα ήταν ίση με $a = Y^{-1}b$. Όμως, ο Y είναι ορθογώνιος, και έχει συνήθως περισσότερες γραμμές από στήλες. Όταν υπάρχουν περισσότερες εξισώσεις από αγνώστους, δεν υπάρχει μία μόνο ακριβής λύση για το a. Παρ' όλ' αυτά, μπορεί να βρεθεί ένα διάνυσμα βαρών a το οποίο να ελαχιστοποιεί κάποια συνάρτηση του λάθους ανάμεσα στο Ya και το b. Εάν οριστεί το διάνυσμα λάθους e ως

$$e = Ya - b \tag{5.42}$$

τότε μία προσέγγιση είναι να ελαχιστοποιηθεί το τετραγωνικό μήκος του διανύσματος λάθους. Αυτή η προσέγγιση είναι ισοδύναμη με την ελαχιστοποίηση της συνάρτησης κριτηρίου του αθροίσματος των τετραγώνων των λαθών

$$J_s(a) = \|Ya - b\|^2 = \sum_{i=1}^{n} \left(a^t y_i - b_i\right)^2 \tag{5.43}$$

Το πρόβλημα της ελαχιστοποίησης του αθροίσματος του τετραγωνικού λάθους αποτελεί ένα κλασσικό πρόβλημα. Μπορεί να επιλυθεί με χρήση μιας διαδικασίας αναζήτησης κλίσης, όπως θα παρουσιαστεί παρακάτω. Μια απλή κλειστής μορφής λύση μπορεί επίσης να δοθεί εάν σχηματιστεί η κλίση

$$\nabla J_s = \sum_{i=1}^{n} 2\left(a^t y_i - b_i\right) y_i = 2Y^t\left(Ya - b\right) \tag{5.44}$$

και τεθεί ίση με το μηδέν. Από αυτό προκύπτει η απαραίτητη συνθήκη

$$Y^t Ya = Y^t b \tag{5.45}$$

και έτσι το πρόβλημα μετατρέπεται από την επίλυση του $Ya = b$ στην επίλυση του $Y^t Ya = Y^t b$. Η δεύτερη εξίσωση έχει το σημαντικό πλεονέκτημα ότι ο $\hat{d} \times \hat{d}$ πίνακας Y^tY είναι τετραγωνικός και συχνά ομαλός. Εάν είναι ομαλός, η εξίσωση μπορεί να λυθεί μοναδικά ως προς το a ως εξής:

$$a = (Y^tY)^{-1}Y^tb = Y^{\dagger}b \tag{5.46}$$

όπου ο $\hat{d} \times n$ πίνακας

$$Y^{\dagger} \equiv (Y^tY)^{-1}Y^t \tag{5.47}$$

καλείται ψευδοαντίστροφος πίνακας του Y. Να σημειωθεί ότι εάν ο Y είναι τετραγωνικός και ομαλός, ο ψευδοαντίστροφος ταυτίζεται με τον κανονικό αντίστροφο. Να σημειωθεί επίσης ότι στη γενική περίπτωση $Y^{\dagger} Y = I$ ενώ $YY^{\dagger} \neq I$. Παρ' όλ' αυτά, η λύση ελάχιστου τετραγωνικού λάθους (Minimum Squared Error – MSE) υφίσταται σε όλες τις περιπτώσεις. Πιο συγκεκριμένα, εάν ο $Y^{\dagger}$ έχει οριστεί πιο γενικά ως

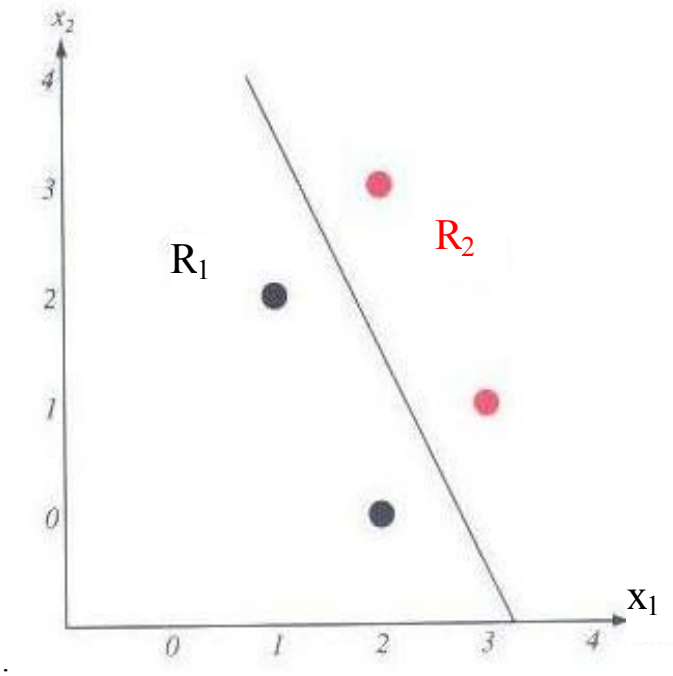
$$Y^{\dagger} \equiv \lim_{\varepsilon \to 0} \left(Y^tY + \varepsilon I\right)^{-1} Y^t \tag{5.48}$$

μπορεί να αποδειχθεί ότι το παραπάνω όριο υπάρχει πάντοτε και ότι το $a = Y^{\dagger}b$ είναι μία MSE λύση για την $Ya = b$.

Η MSE λύση εξαρτάται από το διάνυσμα b και μάλιστα διαφορετικές τιμές για το b δίνουν στη λύση διαφορετικές ιδιότητες. Εάν το b έχει επιλεγεί αυθαίρετα, δεν υπάρχει βεβαιότητα ότι η MSE λύση καταλήγει σε ένα διάνυσμα διαχωριστή για τη γραμμικά διαχωρίσιμα περίπτωση. Παρ' όλ' αυτά, είναι φυσικό να αναμένει κανείς ότι με την ελαχιστοποίηση του τετραγωνικού λάθους είναι πιθανόν να προκύψει μία χρήσιμη διακρίνουσα συνάρτηση τόσο για τη γραμμικά διαχωρίσιμη όσο και για τη μη γραμμικά διαχωρίσιμη περίπτωση.

<u>Παράδειγμα 1. Κατασκευή ενός γραμμικού ταξινομητή με χρήση του ψευδοαντίστροφου πίνακα</u>

Έστω ότι έχουμε τα ακόλουθα σημεία δύο διαστάσεων για δύο κατηγορίες ω_1 : $(1,2)^t$ και $(2,0)^t$, και ω_2 : $(3,1)^t$ και $(2,3)^t$, τα οποία φαίνονται στο ακόλουθο σχήμα.



Τα τέσσερα σημεία εκπαίδευσης και το όριο απόφασης $a^t \begin{pmatrix} 1 \\ x_1 \\ x_2 \end{pmatrix} = 0$ φαίνονται στο παραπάνω σχήμα. Το α έχει υπολογιστεί με την τεχνική του ψευδοαντίστροφου.

Με τη μέθοδο του ψευδοαντίστροφου πίνακα βρίσκονται τα εξής:

Ο πίνακας Y ισούται με: $Y = \begin{pmatrix} 1 & 1 & 2 \\ 1 & 2 & 0 \\ -1 & -3 & -1 \\ -1 & -2 & -3 \end{pmatrix}$

Με κάποιους απλούς υπολογισμούς προκύπτει ο ψευδοαντίστροφος πίνακας:

$$Y^{+} = (Y^t Y)^{-1} Y^t = \begin{pmatrix} 5/4 & 13/12 & 3/4 & 7/12 \\ -1/2 & -1/6 & -1/2 & -1/6 \\ 0 & -1/3 & 0 & -1/3 \end{pmatrix}$$

Έστω ότι $b=(1, 1, 1, 1)^t$, δηλαδή όλα τα διαστήματα είναι ίδια.
Η λύση που προκύπτει είναι η εξής:

$\alpha = Y^{+} b = (11/3, -4/3, -2/3)^t$ η οποία οδηγεί στο όριο απόφασης του παραπάνω σχήματος. Φυσικά, διαφορετικές επιλογές για το b θα είχαν ως αποτέλεσμα τον υπολογισμό διαφορετικών ορίων απόφασης.

5.8.2 Η Σχέση με τη Γραμμική Διακρίνουσα Συνάρτηση του Fisher

Στην ενότητα αυτή θα αποδειχθεί ότι με σωστή επιλογή για το διάνυσμα b, η MSE διακρίνουσα συνάρτηση $a^t y$ σχετίζεται άμεσα με τη γραμμική διακρίνουσα συνάρτηση του Fisher. Έστω ότι υπάρχει ένα σύνολο από n-διάστατα δείγματα x^1, …, x^n, n_1 από τα οποία ανήκουν στο υποσύνολο D_1 με ετικέτα ω_1, και n_2 από τα οποία ανήκουν στο υποσύνολο D_2 με ετικέτα ω_2. Επιπλέον, έστω ότι ένα δείγμα y_i σχηματίζεται από το x_i προσθέτοντας ένα κατώφλι $x_0 = 1$ για να δημιουργηθεί ένα επαυξημένο διάνυσμα προτύπων. Επίσης, εάν το δείγμα έχει ετικέτα ω_2, τότε ολόκληρο το διάνυσμα των προτύπων πολλαπλασιάζεται με -1 (η διαδικασία κανονικοποίησης που παρουσιάστηκε στην ενότητα 5.4.1). Χωρίς βλάβη της γενικότητας, θεωρείται ότι τα πρώτα n_1 δείγματα έχουν ετικέτα ω_1 και τα επόμενα n_2 έχουν ετικέτα ω_2. Τότε ο πίνακας Y μπορεί να χωριστεί ως εξής:

$$Y = \begin{bmatrix} 1_1 & X_1 \\ -1_2 & -X_2 \end{bmatrix}$$

όπου το 1_i είναι ένα διάνυσμα στήλη που αποτελείται από n_i μονάδες, και το X_i είναι ένας $n_i \times d$ πίνακας του οποίου οι γραμμές είναι τα δείγματα με ετικέτα ω_i. Αντίστοιχα χωρίζονται τα a και b ως

$$a = \begin{bmatrix} w_0 \\ w \end{bmatrix}$$

και

$$b = \begin{bmatrix} \frac{n}{n_1} 1_1 \\ \frac{n}{n_2} 1_2 \end{bmatrix}$$

Στη συνέχεια θα δειχθεί ότι αυτή η ειδική επιλογή για την τιμή του b συσχετίζει την MSE λύση με τη γραμμική διακρίνουσα συνάρτηση του Fisher.

Αρχικά, γράφεται η εξίσωση 5.45 για το a χρησιμοποιώντας τους διαχωρισμένους πίνακες:

$$\begin{bmatrix} 1_1^t & -1_2^t \\ X_1^t & -X_2^t \end{bmatrix} \begin{bmatrix} 1_1 & X_1 \\ -1_2 & -X_2 \end{bmatrix} \begin{bmatrix} w_0 \\ w \end{bmatrix} = \begin{bmatrix} 1_1^t & -1_2^t \\ X_1^t & -X_2^t \end{bmatrix} \begin{bmatrix} \frac{n}{n_1} 1_1 \\ \frac{n}{n_2} 1_2 \end{bmatrix} \tag{5.49}$$

Ορίζοντας τις μέσες τιμές των δειγμάτων m_i και τον πίνακα διασποράς των δειγμάτων S_w ως

$$m_i = \frac{1}{n_i} \sum_{x \in D_i} x \quad i = 1,2 \tag{5.50}$$

και

$$S_w = \sum_{i=1}^{2} \sum_{x \in D_i} (x - m_i)(x - m_i)^t \tag{5.51}$$

και πολλαπλασιάζοντας τους πίνακες της εξίσωσης 5.49 προκύπτει

$$\begin{bmatrix} n & (n_1 m_1 + n_2 m_2)^t \\ (n_1 m_1 + n_2 m_2) & S_w + n_1 m_1 m_1^t + n_2 m_2 m_2^t \end{bmatrix} \begin{bmatrix} w_0 \\ w \end{bmatrix} = \begin{bmatrix} 0 \\ n(m_1 - m_2) \end{bmatrix}$$

Η παραπάνω εξίσωση μπορεί να θεωρηθεί ως ένα ζευγάρι εξισώσεων, η πρώτη από τις οποίες μπορεί να επιλυθεί για το w_0 ως προς το w:

$$w_0 = -m^t w \tag{5.52}$$

όπου το m είναι η μέση τιμή όλων των δειγμάτων. Αντικαθιστώντας το παραπάνω στη δεύτερη εξίσωση και μετά από μερικούς αλγεβρικούς μετασχηματισμούς προκύπτει το εξής:

$$\left[\frac{1}{n}S_w + \frac{n_1 n_2}{n^2}(m_1 - m_2)(m_1 - m_2)^t\right] w = m_1 - m_2 \tag{5.53}$$

Επειδή το διάνυσμα $(m_1 - m_2)(m_1 - m_2)^t w$ είναι στη διεύθυνση του $m_1 - m_2$ για οποιαδήποτε τιμή του w, προκύπτει

$$\frac{n_1 n_2}{n^2}(m_1 - m_2)(m_1 - m_2)^t w = (1-\alpha)(m_1 - m_2)$$

όπου το α είναι ένας αριθμός. Η εξίσωση 5.53 παίρνει τη μορφή

$$w = \alpha n S_w^{-1}(m_1 - m_2) \tag{5.54}$$

η οποία, εκτός από τον ασήμαντο αριθμητικό παράγοντα, είναι όμοια με τη λύση που δίνει η γραμμική διακρίνουσα του Fisher. Έτσι, προκύπτει το βάρος κατωφλίου w_0 και ο ακόλουθος κανόνας απόφασης: Αποφάσισε ω_1 εάν $w^t(x - m) > 0$, διαφορετικά αποφάσισε ω_2.

5.8.3 Ασυμπτωτική Προσέγγιση σε μία Βέλτιστη Διακρίνουσα

Μία άλλη ιδιότητα της MSE λύσης η οποία ενισχύει τη χρήση της είναι ότι εάν $b = 1_n$, αυτή τείνει σε μια ελάχιστου μέσου τετραγωνικού λάθους προσέγγιση της διακρίνουσας συνάρτησης του Bayes όταν ο αριθμός των δειγμάτων τείνει στο άπειρο.

$$g_0(x) = P(\omega_1 / x) - P(\omega_2 / x) \tag{5.55}$$

Για να αποδειχθεί αυτή η ιδιότητα, θεωρείται ότι τα δείγματα είναι ανεξάρτητα ομοιόμορφα κατανεμημένα σύμφωνα με τον παρακάτω πιθανοτικό κανόνα

$$p(x) = p(x/\omega_1)P(\omega_1) + p(x/\omega_2)P(\omega_2) \tag{5.56}$$

Με όρους του επαυξημένου διανύσματος y, η MSE λύση καταλήγει στην ανάπτυξη της σειράς $g(x) = a^t y$, όπου y = y(x). Εάν ορίσουμε την προσέγγιση του μέσου τετραγωνικού λάθους ως

$$\varepsilon^2 = \int [a^t y - g_0(x)]^2 p(x) dx \tag{5.57}$$

τότε σκοπός είναι να αποδειχθεί ότι το ε_2 ελαχιστοποιείται από τη λύση της $a = Y^{\dagger} 1_n$. Η απόδειξη απλοποιείται εάν διατηρήσει κανείς τη διαφορά ανάμεσα στα δείγματα της κατηγορίας ω_1 και της κατηγορίας ω_2. Με όρους των μη κανονικοποιημένων δεδομένων, η συνάρτηση κριτηρίου J_s παίρνει τη μορφή

$$J_s(a) = \sum_{y \in Y_1}(a^t y - 1)^2 + \sum_{y \in Y_2}(a^t y + 1)^2 = n\left[\frac{n_1}{n}\frac{1}{n_1}\sum_{y \in Y_1}(a^t y - 1)^2 + \frac{n_2}{n}\frac{1}{n_2}\sum_{y \in Y_2}(a^t y + 1)^2\right] \tag{5.58}$$

Έτσι, σύμφωνα με τον κανόνα τον μεγάλων αριθμών, καθώς το n τείνει στο άπειρο το $(1/n)\cdot J_s(a)$ τείνει στο

$$\bar{J}(a) = P(\omega_1)E_1[(a^t y - 1)^2] + P(\omega_2)E_2[(a^t y + 1)^2] \tag{5.59}$$

με πιθανότητα ένα, όπου

$$E_1[(a^t y - 1)^2] = \int (a^t y - 1)^2 p(x/\omega_1) dx$$

και

$$E_2\left[\left(a^t y+1\right)^2\right]=\int\left(a^t y+1\right)^2 p(x/\omega_2)dx$$

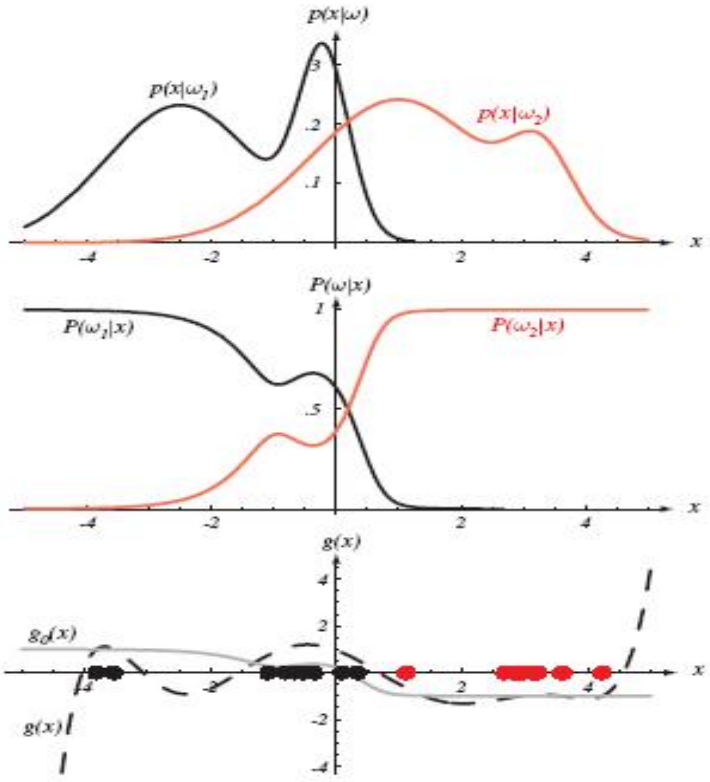
Τώρα, υπολογίζοντας από την εξίσωση 5.55 ότι

$$g_0(x)=\frac{p(x,\omega_1)-p(x,\omega_2)}{p(x)}$$

καταλήγει κανείς στο

$$\begin{aligned}\bar{J}(a)&=\int\left(a^t y-1\right)^2 p(x,\omega_1)dx+\int\left(a^t y+1\right)^2 p(x,\omega_2)dx\\&=\int\left(a^t y\right)^2 p(x)dx-2\int a^t y g_0(x)p(x)dx+1\\&=\underbrace{\int\left[a^t y-g_0(x)\right]^2 p(x)dx}_{\varepsilon^2}+\underbrace{1-\int g_0^2(x)p(x)dx}_{\text{ανεξάρτητο του a}}\end{aligned}\qquad(5.60)$$

Ο δεύτερος όρος στο παραπάνω άθροισμα, είναι ανεξάρτητος από το διάνυσμα των βαρών a. Επομένως, το a που ελαχιστοποιεί το J_s ελαχιστοποιεί επίσης και το ε_2, δηλαδή το μέσο τετραγωνικό λάθος ανάμεσα στο $a^t y$ και το g(x) (εικόνα 5.16).



Εικόνα 5.16: Το πάνω σχήμα δείχνει δύο υπό συνθήκη πυκνότητες πιθανότητας, ενώ το μεσαίο σχήμα τις εκ των υστέρων πιθανότητες, θεωρώντας ίσες τιμές για τις εκ των προτέρων πιθανότητες. Η ελαχιστοποίηση του MSE λάθους ελαχιστοποιεί επίσης και το μέσο τετραγωνικό λάθος ανάμεσα στο $a^t y$ και τη διακρίνουσα συνάρτηση g(x) (στο παράδειγμα ένα πολυώνυμο έβδομου βαθμού) μετρημένη πάνω στα δεδομένα της κατανομής, όπως φαίνεται στο κάτω σχήμα. Σημειώνεται ότι η g(x) υπολογίζει καλύτερα την $g_0(x)$ στις περιοχές όπου βρίσκονται τα σημεία των δεδομένων.

Το αποτέλεσμα αυτό παρέχει σημαντικές πληροφορίες για την MSE διαδικασία. Υπολογίζοντας την $g_0(x)$, η διακρίνουσα συνάρτηση $a^t y$ παρέχει άμεση πληροφόρηση για τις εκ των υστέρων πιθανότητες $P(\omega_1 / x) = (1 + g_0) / 2$ και $P(\omega_2 / x) = (1 - g_0) / 2$. Η ποιότητα των υπολογισμών βασίζεται στις συναρτήσεις $y_i(x)$ και τον αριθμό των

όρων της expansion $a^t y$. Δυστυχώς, το κριτήριο του μέσου τετραγωνικού λάθους δίνει έμφαση στα σημεία όπου η p(x) είναι μεγαλύτερη, παρά στα σημεία που βρίσκονται πλησίον της επιφάνειας απόφασης $g_0(x) = 0$. Άρα, η διακρίνουσα συνάρτηση που προσεγγίζει βέλτιστα τη διακρίνουσα συνάρτηση του Bayes δεν ελαχιστοποιεί απαραίτητα την πιθανότητα του λάθους. Ανεξάρτητα από το γεγονός αυτό, η MSE λύση εμφανίζει ενδιαφέρουσες ιδιότητες και έχει αποτελέσει πολλές φορές θέμα ανάλυσης στη διεθνή βιβλιογραφία.

5.8.4 Η Διαδικασία Widrow – Hoff ή Ελαχίστων Μέσων Τετραγώνων (Least Mean Squared – LMS)

Σε προηγούμενη ενότητα σημειώθηκε ότι το $J_s(a) = \|Ya - b\|^2$ μπορεί να ελαχιστοποιηθεί με χρήση μιας διαδικασίας κλίσης καθόδου. Μια τέτοια προσέγγιση έχει δύο πλεονεκτήματα όσον αφορά τον υπολογισμό του ψευδοαντίστροφου: 1) αποφεύγει τα προβλήματα που εμφανίζονται όταν ο Y^tY είναι μη ομαλός, 2) αποφεύγει την ανάγκη για υπολογισμούς με μεγάλους πίνακες. Επιπρόσθετα, οι υπολογισμοί που περιλαμβάνονται αποτελούν ένα σημαντικό πρόβλημα το οποίο όμως μπορεί να αντιμετωπιστεί άμεσα, όπως τα περισσότερα προβλήματα υπολογισμών που οφείλονται σε στρογγυλοποιήσεις και αποκοπές. Επειδή $\nabla J_s = 2Y^t(Ya - b)$ ο προφανής κανόνας ενημέρωσης είναι

$$a(1) \quad \text{τυχαία}$$
$$a(k+1) = a(k) + \eta(k)Y^t(Ya(k) - b) \quad k \geq 1 \qquad (5.61)$$

Εάν το η(k) ισούται με η(1) / k, όπου το η(1) μπορεί να είναι οποιαδήποτε θετική σταθερά, ο κανόνας αυτός δημιουργεί μια ακολουθία από διανύσματα βαρών η οποία συγκλίνει σε ένα οριακό διάνυσμα a που ικανοποιεί τη σχέση

$$Y^t(Ya(k) - b) = 0$$

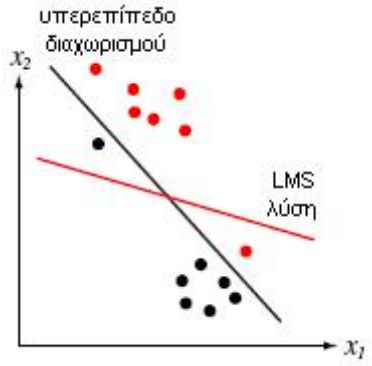
Έτσι, ο αλγόριθμος κλίσης καθόδου καταλήγει πάντοτε σε μια λύση ανεξάρτητα από το εάν ο πίνακας Y^tY είναι μη ομαλός ή όχι.

Ενώ ο $\hat{d} \times \hat{d}$ πίνακας Y^tY είναι συνήθως μικρότερος από τον $\hat{d} \times n$ πίνακα Y^t, η αποθηκευτικές απαιτήσεις μπορούν να ελαττωθούν ακόμα περισσότερο εάν τα δείγματα αντιμετωπιστούν ακολουθιακά και χρησιμοποιηθεί ο Widrow – Hoff ή LMS αλγόριθμος:

Αλγόριθμος 10: LMS
1 αρχή αρχικοποίησε a, b, κατώφλι θ, η(·), k ← 0
2 κάνε k ← (k + 1) mod n
3 a ← a + η(k)($b_k - a^t y^k$)y^k
4 μέχρι |+ η(k)($b_k - a^t y^k$)y^k | < θ
5 επέστρεψε a
6 τέλος

Με μια πρώτη ματιά, αυτός ο αλγόριθμος κλίσης καθόδου φαίνεται να μοιάζει σημαντικά με τον κανόνα χαλάρωσης. Η βασική διαφορά τους είναι ότι ο κανόνας χαλάρωσης είναι ένας κανόνας διόρθωσης λάθους, έτσι ώστε το $a^t(k)y^k$ να μην ισούται με το b_k, και επομένως η διόρθωση να μην τερματίζεται ποτέ. Έτσι, το η(k) πρέπει να μειώνεται σε σχέση με το k για να επιτευχθεί σύγκλιση. Μια συνήθης επιλογή είναι το η(k) = η(1) / k. Η ακριβής ανάλυσης του Widrow – Hoff κανόνα για την ντετερμινιστική περίπτωση είναι αρκετά πολύπλοκη και περισσότερο υποδηλώνει

ότι η ακολουθία των διανυσμάτων των βαρών τείνει να συγκλίνει στην επιθυμητή λύση. Στην επόμενη ενότητα, αντί να αναλυθεί ο κανόνας αυτός, θα περιγραφεί ένας παρόμοιος κανόνας που παρουσιάζεται από μία στοχαστική διαδικασία κλίσης καθόδου. Σημειώνεται παρ' όλ' αυτά, ότι η λύση δεν απαιτείται να καταλήγει σε ένα διάνυσμα διαχωριστή, ακόμα και όταν υπάρχει τέτοιο, όπως φαίνεται και από την εικόνα 5.17.



Εικόνα 5.17: Ο αλγόριθμος LMS δε χρειάζεται να συγκλίνει σε υπερεπίπεδο διαχωρισμού, ακόμα και εάν υπάρχει τέτοιο. Επειδή η λύση του LMS ελαχιστοποιεί το άθροισμα των τετραγώνων των αποστάσεων των σημείων εκπαίδευσης από το υπερεπίπεδο, για το παραπάνω παράδειγμα το επίπεδο είναι περιστρεμμένο κατά τη φορά των δεικτών του ρολογιού σε σχέση με ένα υπερεπίπεδο διαχωρισμού.

5.8.5 Στοχαστικές Μέθοδοι Προσέγγισης

Όλες οι επαναληπτικές διαδικασίες κλίσης καθόδου που αναλύθηκαν μέχρι τώρα περιγράφηκαν με ντετερμινιστικούς όρους. Δεδομένου ενός συγκεκριμένου συνόλου από δείγματα, δημιουργείται μία συγκεκριμένη ακολουθία από διανύσματα βαρών. Στην ενότητα αυτή θα αναλυθεί μία MSE διαδικασία στην οποία τα δείγματα επιλέγονται τυχαία και παράγεται μια τυχαία ακολουθία από διανύσματα βαρών.

Έστω ότι επιλέγονται δείγματα με ανεξάρτητο τρόπο και με πιθανότητα $P(\omega_i)$ να ανήκουν στην κατάσταση της φύσης ωi και στη συνέχεια επιλέγεται ένα ένα x σύμφωνα με τον πιθανοτικό κανόνα $p(x/\omega_i)$. Για κάθε x έστω ότι το θ είναι η ετικέτα του, με θ = +1 εάν το x έχει ετικέτα ω_1 και θ = -1 εάν το x έχει ετικέτα ω_2. Τότε τα δεδομένα σχηματίζουν μια μη πεπερασμένη ακολουθία από ανεξάρτητα ζεύγη (x, θ_1), (x, θ_2), ..., (x_k, θ_k), …. Εάν και η μεταβλητή ετικέτας θ παίρνει μόνο δύο διακριτές τιμές, μπορεί να θεωρηθεί ως μία «θορυβώδης» εκδοχή της διακρίνουσας συνάρτησης του Bayes $g_0(x)$. Αυτό προκύπτει από την παρατήρηση ότι

$$P(\theta = 1/x) = P(\omega_1/x)$$

και

$$P(\theta = -1/x) = P(\omega_2/x)$$

έτσι ώστε η υπό συνθήκη μέση τιμή του θ δίνεται από τον τύπο:

$$E_{\theta/x}[\theta] = \sum_{\theta} \theta\, P(\theta/x) = P(\omega_1/x) - P(\omega_2/x) = g_0(x) \qquad (5.62)$$

Έστω ότι πρέπει να υπολογιστεί η $g_0(x)$ από την επέκταση της πεπερασμένης σειράς

$$g(x) = a^t y = \sum_{i=1}^{\hat{d}} a_i\, y_i(x)$$

όπου οι δύο βασικές συναρτήσεις $y_i(x)$ και ο αριθμός των όρων $\hat{d}$ είναι γνωστά. Τότε μπορεί να βρεθεί ένα διάνυσμα βαρών $\hat{a}$ το οποίο ελαχιστοποιεί το μέσο τετραγωνικό λάθος της προσέγγισης

$$\varepsilon^2 = E\left[\left(a^t y - g_0(x)\right)^2\right] \tag{5.63}$$

Η ελαχιστοποίηση του ε^2 φαίνεται να απαιτεί τη γνώση της διακρίνουσας του Bayes $g_0(x)$. Παρ' όλ' αυτά, μπορεί να αποδειχθεί ότι το διάνυσμα των βαρών $\hat{a}$ που ελαχιστοποιεί το ε^2 ελαχιστοποιεί επίσης και τη συνάρτηση κριτηρίου

$$J_m(a) = E\left[\left(a^t y - \theta\right)^2\right] \tag{5.64}$$

Αυτό προκύπτει επίσης και από τη θεώρηση ότι η το θ αποτελεί μία «θορυβώδη» εκδοχή της $g_0(x)$. Επειδή η κλίση ισούται με

$$\nabla J_m = 2E\left[\left(a^t y - \theta\right)y\right] \tag{5.65}$$

Προκύπτει η παρακάτω κλειστής μορφής λύση

$$\hat{a} = E\left[yy^t\right]^{-1} E\left[\theta\, y\right] \tag{5.66}$$

Επομένως, ένας τρόπος για να χρησιμοποιηθούν τα δείγματα είναι να υπολογιστούν τα $E\left[yy^t\right]$ και $E\left[\theta\, y\right]$ και χρησιμοποιώντας την εξίσωση 5.66 να υπολογιστεί η βέλτιστη MSE γραμμική διακρίνουσα συνάρτηση. Μια άλλη μέθοδος είναι να ελαχιστοποιηθεί το $J_m(a)$ χρησιμοποιώντας μια διαδικασία κλίσης καθόδου. Έστω ότι στη θέση της πραγματικής κλίσης τοποθετηθεί η θορυβώδης εκδοχή $2(a^t y^k - \theta_k)y^k$. Αυτό οδηγεί στον ακόλουθο κανόνα ενημέρωσης

$$a(k+1) = a(k) + \eta\left(\theta_k - a^t(k)y_k\right)y_k \tag{5.67}$$

ο οποίος στην πραγματικότητα είναι ο κανόνας Widrow – Hoff. Μπορεί να αποδειχθεί ότι εάν το $E\left[yy^t\right]$ είναι ομαλός και εάν οι συντελεστές η(k) ικανοποιούν τις συνθήκες

$$\lim_{m\to\infty} \sum_{k=1}^{m} \eta(k) = +\infty \tag{5.68}$$

και

$$\lim_{m\to\infty} \sum_{k=1}^{m} \eta^2(k) < \infty \tag{5.69}$$

τότε το a(k) συγκλίνει στο $\hat{a}$ όσον αφορά τη μέση τετραγωνική τιμή:

$$\lim_{k\to\infty} E\left[\left\|a(k) - \hat{a}\right\|^2\right] = 0 \tag{5.70}$$

Οι λόγοι για τους οποίους απαιτούνται αυτές οι συνθήκες για το η(k) είναι απλοί. Η πρώτη συνθήκη αποτρέπει το διάνυσμα των βαρών από το να συγκλίνει τόσο γρήγορα ώστε ένα συστηματικό λάθος να παραμείνει για πάντα χωρίς διόρθωση. Η δεύτερη συνθήκη εγγυάται ότι οι τυχαίες διακυμάνσεις εξομαλύνονται ομοιόμορφα. Και οι δύο συνθήκες ικανοποιούνται επιλέγοντας η(k) = 1 /k. Δυστυχώς, αυτή η σταθερή μείωση του η(k), που είναι ανεξάρτητη από το κάθε προς επίλυση πρόβλημα, οδηγεί συνήθως σε πολύ αργή σύγκλιση. Φυσικά, αυτός δεν είναι ούτε ο μοναδικός ούτε ο καλύτερος αλγόριθμος κλίσης καθόδου για την ελαχιστοποίηση του J_m. Για παράδειγμα, εάν ο πίνακας των δεύτερων μερικών παραγώγων για το J_m δίνεται από το

$$D = 2E\left[yy^t\right]$$

Ο κανόνας του Newton για την ελαχιστοποίηση του J_m (εξίσωση 5.15) γίνεται

$$a(k+1) = a(k) + E\left[yy^t\right]^{-1} E\left[\left(\theta - a^t y\right)y\right]$$

Μια στοχαστική ανάλογη έκφραση αυτού του κανόνα δίνεται από τον τύπο

$$a(k+1) = a(k) + R_{k+1}\left(\theta_k - a^t(k)y_k\right)y_k \tag{5.71}$$

Με

$$R_{k+1}^{-1} = R_k^{-1} + y_k y_k^t \tag{5.72}$$

ή αντίστοιχα

$$R_{k+1} = R_k - \frac{R_k y_k (R_k y_k)^t}{1 + y_k^t R_k y_k} \tag{5.73}$$

Ο κανόνας αυτός παράγει επίσης μία ακολουθία από διανύσματα βαρών η οποία συγκλίνει σε μία βέλτιστη λύση όσον αφορά τη μέση τετραγωνική τιμή. Η σύγκλισή του είναι γρηγορότερη αλλά απαιτεί περισσότερους υπολογισμούς ανά βήμα.

Αυτές οι διαδικασίες κλίσεις μπορούν να θεωρηθούν ως μέθοδοι για την ελαχιστοποίηση μιας συνάρτησης κριτηρίου ή την εύρεση της τιμής η οποία μηδενίζει την κλίση, παρουσία θορύβου. Στη βιβλιογραφία, συναρτήσεις όπως οι J_m και η ∇J_m οι οποίες έχουν τη μορφή $E\left[f(a,x)\right]$ καλούνται συναρτήσεις αναδρομής (regression procedures), ενώ οι επαναληπτικοί αλγόριθμοι καλούνται στοχαστικές προσεγγιστικές διαδικασίες (stochastic approximation procedures). Δύο αρκετά γνωστές από αυτές είναι η διαδικασία των Kiefer – Wolfowitz για την ελαχιστοποίηση μιας αναδρομικής συνάρτησης και η διαδικασία των Robbins – Monro για την εύρεση μιας ρίζας μιας αναδρομικής συνάρτησης. Συχνά, ο πιο εύκολος τρόπος για να καταλήξει κανείς στην απόδειξη της σύγκλισης μιας συγκεκριμένης descent ή προσεγγιστικής διαδικασίας είναι να δείξει ότι ικανοποιεί τις συνθήκες σύγκλισης για αυτές τις πιο γενικές διαδικασίες.

5.9 Οι Διαδικασίες Ho – Kashyap

Οι διαδικασίες που μελετήθηκαν έως τώρα διαφέρουν σε πολλά σημεία. Ο αλγόριθμος του Perceptron και οι διαδικασίες χαλάρωσης καταλήγουν σε διανύσματα διαχωριστές εάν τα δείγματα είναι γραμμικά διαχωρίσιμα, αλλά δεν συγκλίνουν σε περιπτώσεις όπου τα δείγματα είναι μη γραμμικά διαχωρίσιμα. Οι MSE διαδικασίες παράγουν ένα διάνυσμα βαρών ανεξάρτητα από το εάν τα δείγματα είναι γραμμικά διαχωρίσιμα ή όχι, όμως το διάνυσμα αυτό δεν είναι σίγουρο ότι αποτελεί ένα διάνυσμα διαχωριστή για την περίπτωση όπου τα δείγματα είναι γραμμικά διαχωρίσιμα (εικόνα 5.17). Εάν το διάνυσμα των ορίων b έχει επιλεγεί αυθαίρετα, το μόνο που είναι σίγουρο είναι ότι οι MSE διαδικασίες ελαχιστοποιούν το $\|Ya - b\|^2$. Εάν τα δείγματα εκπαίδευσης συμβαίνει να είναι γραμμικά διαχωρίσιμα, τότε υπάρχει ένα $\hat{a}$ και ένα $\hat{b}$ τέτοιο ώστε

$$Y\hat{a} = \hat{b} > 0$$

όπου με το $\hat{b} > 0$ εννοείται ότι κάθε στοιχείο του $\hat{b}$ είναι θετικό. Προφανώς, εάν κάποιος θέσει $b = \hat{b}$ και εφαρμόσει την MSE διαδικασία, θα καταλήξει σε ένα διάνυσμα διαχωριστή. Φυσικά, η τιμή του $\hat{b}$ δεν είναι συνήθως γνωστή εκ των προτέρων. Παρ' όλ' αυτά, στη συνέχεια θα παρουσιαστεί ο τρόπος με τον οποίο η MSE διαδικασία μπορεί να τροποποιηθεί ώστε να καταλήγει τόσο σε ένα διάνυσμα διαχωριστή a όσο και σε ένα διάνυσμα ορίου b. Η ιδέα στην οποία βασίζεται προκύπτει από την παρατήρηση ότι εάν τα δείγματα είναι διαχωρίσιμα, και εάν τόσο το a όσο και το b επιτρέπεται να παίρνουν διάφορες τιμές στη συνάρτηση κριτηρίου

$$J_s(a,b) = \|Ya - b\|^2 \tag{5.74}$$

υπό την προϋπόθεση πάντα ότι b > 0, τότε η ελάχιστη τιμή του J_s είναι το μηδέν, ενώ το a που επιτυγχάνει αυτό το ελάχιστο είναι ένα διάνυσμα διαχωριστής.

5.9.1 Η Διαδικασία Καθόδου

Για να ελαχιστοποιηθεί το J_s στην εξίσωση 5.74 πρέπει να χρησιμοποιηθεί μία τροποποιημένη διαδικασία κλίσης καθόδου. Η κλίση του J_s σε σχέση με το a δίνεται από τη

$$\nabla_a J_s = 2Y^t(Ya - b) \tag{5.75}$$

και το gradient του J_s σε σχέση με το b δίνεται από τη

$$\nabla_b J_s = -2(Ya - b) \tag{5.76}$$

Για οποιαδήποτε τιμή του b, μπορεί πάντοτε κανείς να θεωρήσει

$$a = Y^{\dagger}b \tag{5.77}$$

και καταλήγοντας στο $\nabla_a J_s = 0$ να ελαχιστοποιήσει το J_s σε σχέση με το a σε ένα βήμα. Για την τροποποίηση του b δεν υπάρχει τόσο μεγάλη ελευθερία, επειδή πρέπει να ικανοποιείται ο περιορισμός b > 0 και να αποφευχθεί η χρήση μίας διαδικασίας καθόδου η οποία να συγκλίνει στο b = 0. Ένας τρόπος να αποφευχθεί η σύγκλιση του b στο μηδέν είναι να ξεκινήσει κανείς με b > 0 και στη συνέχεια να μην ελαττώνει κανένα από τα στοιχεία του. Αυτό μπορεί να συμβαίνει και παράλληλα να γίνεται προσπάθεια να ακολουθηθεί η αρνητική κλίση εάν αρχικά τεθούν όλα τα θετικά στοιχεία του $\nabla_b J_s$ ίσα με το μηδέν. Έτσι, εάν με |v| συμβολιστεί το διάνυσμα του οποίου τα στοιχεία είναι τα μέτρα των αντίστοιχων στοιχείων του v, καταλήγει κανείς στον παρακάτω κανόνα ενημέρωσης για το διάνυσμα του ορίου

$$b(k+1) = b(k) - \eta\frac{1}{2}\left[\nabla_b J_s - |\nabla_b J_s|\right] \tag{5.78}$$

Χρησιμοποιώντας τις εξισώσεις 5.76 και 5.77 προκύπτει ο κανόνας των Ho – Kashyap για την ελαχιστοποίηση του $J_s(a, b)$:

$$b(1) > 0 \quad \text{αυθαίρετα}$$

$$b(k+1) = b(k) + 2\eta(k)e^+(k) \tag{5.79}$$

όπου το e(k) είναι το διάνυσμα του λάθους

$$e(k) = Ya(k) - b(k) \tag{5.80}$$

$e^+(k)$ είναι το θετικό κομμάτι του διανύσματος του λάθους

$$e^+(k) = \frac{1}{2}(e(k) + |e(k)|) \tag{5.81}$$

και

$$a(k) = Y^{\dagger}b(k) \quad k = 1,2,\ldots \tag{5.82}$$

Επομένως, εάν το b_{min} είναι ένα μικρό κριτήριο σύγκλισης και το Abs[e] συμβολίζει το θετικό κομμάτι του e, ο αλγόριθμος έχει την παρακάτω μορφή:

Αλγόριθμος 11: LMS

1 αρχή αρχικοποίησε a, b, η(·) < 1, κατώφλι b_{min}, k_{max}
2 κάνε $k \leftarrow (k + 1) \bmod n$
3 $e \leftarrow Ya - b$
4 $e^+ \leftarrow 1/2\ (e + Abs[e])$
5 $b \leftarrow b + 2\eta(k)e^+$
6 $a \leftarrow Y^{\dagger}b$

7 εάν Abs[e] ≤ b_{min} τότε επέστρεψε a, b και έξοδος
8 μέχρι k = k_{max}
9 εμφάνισε «ΔΕΝ ΒΡΕΘΗΚΕ ΛΥΣΗ»
10 τέλος

Επειδή το διάνυσμα των βαρών a(k) καθορίζεται πλήρως από το διάνυσμα ορίου b(k), ο παραπάνω αλγόριθμος είναι περισσότερο ένας αλγόριθμος για την παραγωγή μιας ακολουθίας από διανύσματα ορίων. Το αρχικό διάνυσμα b(1) είναι θετικό, και αν $\eta > 0$, όλα τα επόμενα διανύσματα b(k) θα είναι θετικά. Κάποιος θα μπορούσε να ισχυριστεί ότι υπάρχει ο κίνδυνος εάν κανένα από τα στοιχεία του e(k) δεν είναι θετικά, με αποτέλεσμα το b(k) να σταματήσει να μεταβάλλεται, ο αλγόριθμος να αποτύχει να καταλήξει σε λύση. Παρ' όλ' αυτά, στην περίπτωση αυτή είτε ισχύει e(k) = 0 και επομένως υπάρχει λύση, είτε ισχύει e(k) ≤ 0 και υπάρχει απόδειξη ότι τα δείγματα δεν είναι γραμμικά διαχωρίσιμα.

5.9.2 Απόδειξη Σύγκλισης

Στη συνέχεια θα αποδειχθεί ότι εάν τα δείγματα είναι γραμμικά διαχωρίσιμα, και εάν $0 < \eta < 1$, τότε ο αλγόριθμος των Ho – Kashyap καταλήγει σε ένα διάνυσμα λύσης σε πεπερασμένο αριθμό βημάτων. Για να τερματιστεί ο αλγόριθμος πρέπει να προστεθεί ένα κριτήριο τερματισμού το οποίο να δηλώνει ότι οι διορθώσεις σταματάνε όταν βρεθεί ένα διάνυσμα λύσης ή όταν έχει ολοκληρωθεί ένας αρκετά μεγάλος αριθμός από επαναλήψεις. Παρ' όλ' αυτά, από μαθηματικής άποψης είναι προτιμότερο να αφήσει κανείς τις διορθώσεις να συνεχίζονται και να δείξει ότι το διάνυσμα του λάθους e(k) είτε μηδενίζεται για κάποιο πεπερασμένο k, είτε συγκλίνει στο μηδέν όταν το k τείνει στο άπειρο.

Είναι προφανές ότι είτε ισχύει e(k) = 0 για κάποιο k, είτε δεν υπάρχει κανένα μηδενικό διάνυσμα στην ακολουθία e(1), e(2), …. Στην πρώτη περίπτωση, όταν εμφανιστεί ένα μηδενικό διάνυσμα λάθους, από εκείνο το σημείο και μετά τα a(k), b(k) και e(k) δεν αλλάζουν, και ισχύει Ya(k) = b(k) > 0 για όλα τα k ≥ k_0. Έτσι, εάν κατά τη διάρκεια της εκτέλεσης του αλγορίθμου προκύψει ένα μηδενικό διάνυσμα λάθους, ο αλγόριθμος τερματίζεται αυτόματα έχοντας βρει ένα διάνυσμα λύση.

Έστω τώρα ότι το e(k) δε γίνεται ποτέ ίσο με το μηδέν για πεπερασμένο k. Για να δείξει κανείς ότι το e(k) πρέπει έτσι κι αλλιώς να συγκλίνει στο μηδέν, πρέπει αρχικά να θέσει το ερώτημα για το εάν μπορεί ή όχι να καταλήξει ο αλγόριθμος σε ένα e(k) το οποίο να μην έχει καθόλου θετικά στοιχεία. Εάν συνέβαινε αυτό θα ίσχυε Ya(k) ≤ b(k) και από τη στιγμή που το $e^+(k)$ θα ήταν ίσο με το μηδέν δε θα συνέβαιναν άλλες μετατροπές στα a(k), b(k) και e(k). Ευτυχώς, αυτό δε μπορεί να συμβεί ποτέ εάν τα δείγματα είναι γραμμικά διαχωρίσιμα. Η απόδειξη είναι απλή και βασίζεται στο ότι εάν ισχύει $Y^tYa(k)=Y^tb$, τότε $Y^te(k) = 0$. Εάν όμως τα δείγματα είναι γραμμικά διαχωρίσιμα, υπάρχει ένα $\hat{a}$ και ένα $\hat{b} > 0$ τέτοια ώστε

$$Y\hat{a} = \hat{b}$$

Έτσι,

$$e^t(k)Y\hat{a} = 0 = e^t(k)\hat{b}$$

και επειδή όλα τα στοιχεία του $\hat{b}$ είναι θετικά, είτε το e(k) = 0 είτε τουλάχιστον ένα από τα στοιχεία του e(k) πρέπει να είναι θετικό. Επειδή η περίπτωση e(k) = 0 έχει αποκλειστεί, προκύπτει ότι το $e^+(k)$ δεν μπορεί να ισούται με το μηδέν για πεπερασμένο k. Η απόδειξη του ότι το διάνυσμα του λάθους συγκλίνει πάντοτε στο μηδέν, εκμεταλλεύεται το ότι ο πίνακας YY^t είναι συμμετρικός, θετικά ημι-ορισμένος και ικανοποιεί τη σχέση

$$(YY^{\dagger})^t (YY^{\dagger}) = YY^{\dagger} \tag{5.83}$$

Αν και τα αποτελέσματα αυτά ισχύουν γενικότερα, για λόγους απλότητας θα παρουσιαστεί μόνο η περίπτωση όπου ο Y^tY είναι ομαλός. Στην περίπτωση αυτή, ισχύει $YY^{\dagger} = Y(Y^tY)^{-1}Y^t$ και η συμμετρία είναι αυταπόδεικτη. Επειδή ο Y^tY είναι θετικά ορισμένος προφανώς είναι και ο $(Y^tY)^{-1}$. Έτσι, ισχύει $bY(Y^tY)^{-1}Y^tb \geq 0$ για κάθε b και ο $YY^{\dagger}$ είναι τουλάχιστον θετικά ημι-ορισμένος. Τέλος, η εξίσωση 5.83 προκύπτει από την

$$(YY^{\dagger})^t (YY^{\dagger}) = [Y(Y^tY)^{-1}Y^t][\ Y(Y^tY)^{-1}Y^t]$$

Για να δειχθεί ότι το e(k) πρέπει υποχρεωτικά να συγκλίνει στο μηδέν, απαλείφεται το a(k) από τις εξισώσεις 5.80 και 5.82 και προκύπτει

$$e(k) = (YY^{\dagger} - I)b(k)$$

Στη συνέχεια, χρησιμοποιώντας σταθερή παράμετρο μάθησης και την εξίσωση 5.79 προκύπτει η αναδρομική σχέση

$$\begin{aligned} e(k+1) &= (YY^{\dagger} - I)(b(k) + 2\eta e^{+}(k)) \\ &= e(k) + 2\eta(YY^{\dagger} - I)e^{+}(k) \end{aligned} \tag{5.84}$$

έτσι ώστε

$$\frac{1}{4}\|e(k+1)\|^2 = \frac{1}{4}\|e(k)\|^2 + \eta e^t(k)(\ YY^{\dagger} - I)e^{+}(k) + \|\eta(YY^{\dagger} - I)e^{+}(k)\|^2$$

Ο δεύτερος και ο τρίτος όρος μπορούν να απλοποιηθούν σημαντικά. Επειδή $e^t(k)Y = 0$, ο δεύτερος όρος παίρνει τη μορφή

$$\eta e(k)(\ YY^{\dagger} - I)e^{+}(k) = -\eta e^t(k)e^{+t}(k) = -\eta\|e^{+}(k)\|^2$$

όπου τα μη μηδενικά στοιχεία του $e^{+}(k)$ είναι τα θετικά στοιχεία του e(k). Επειδή ο $YY^{\dagger}$ είναι συμμετρικός και ίσος με τον $(YY^{\dagger})^t(YY^{\dagger})$, ο τρίτος όρος παίρνει τη μορφή

$$\begin{aligned} \|\eta(YY^{\dagger} - I)e^{+}(k)\|^2 &= \eta^2 e^{+t}(k)(YY^{\dagger} - I)^t(YY^{\dagger} - I)e^{+}(k) \\ &= -\eta\|e^{+}(k)\|^2 - \eta^2 e^{+}(k)YY^{\dagger}e^{+}(k) \end{aligned}$$

και έτσι προκύπτει

$$\frac{1}{4}\left(\|e(k)\|^2 - \|e(k+1)\|^2\right) = \eta(1-\eta)\|e^{+}(k)\|^2 + \eta^2 e^{+t}(k)\ YY^{\dagger}e^{+}(k) \tag{5.85}$$

Επειδή το $e^{+}(k)$ έχει θεωρηθεί μη μηδενικό και επειδή ο $YY^{\dagger}$ είναι θετικά ημι-ορισμένος, $\|e(k)\|^2 > \|e(k+1)\|^2$ εάν $0 < \eta < 1$. Έτσι, η ακολουθία $\|e(1)\|^2, \|e(2)\|^2, \ldots$ αυξάνεται μονότονα και πρέπει να συγκλίνει σε κάποια οριακή τιμή $\|e\|^2$. Όμως, για να επιτευχθεί σύγκλιση, το $e^{+}(k)$ πρέπει να συγκλίνει στο μηδέν, έτσι ώστε όλα τα θετικά στοιχεία του e(k) να συγκλίνουν στο μηδέν. Επομένως, εάν $0 < \eta < 1$ και τα δείγματα είναι γραμμικά διαχωρίσιμα, το a(k) θα συγκλίνει σε ένα διάνυσμα λύσης όταν το k τείνει στο άπειρο.

Εάν ελέγχονται τα πρόσημα των στοιχείων του Ya(k) σε κάθε βήμα εκτέλεσης και ο αλγόριθμος τερματιστεί όταν βρεθούν όλα θετικά, θα προκύψει ένα διάνυσμα διαχωριστής σε ένα πεπερασμένο αριθμό βημάτων. Αυτό προκύπτει από το ότι $Ya(k) = b(k) + e(k)$ και τα στοιχεία του b(k) δεν ελαττώνονται ποτέ. Έτσι, εάν με b_{min} συμβολιστεί το μικρότερο στοιχείο του b(1) και εάν το e(k) συγκλίνει στο μηδέν, τότε το e(k) πρέπει να εισέλθει μέσα στην υπερσφαίρα $\|e(k)\| = b_{min}$ μετά από ένα πεπερασμένο αριθμό βημάτων, και μάλιστα στο σημείο όπου Ya(k) > 0. Εάν και για την απόδειξη παραλήφθηκαν αυτές οι συνθήκες τερματισμού, στην πράξη χρησιμοποιούνται πάντοτε.

5.9.3 Μη Διαχωρίσιμη Συμπεριφορά

Στην παραπάνω απόδειξη χρειάστηκε τα δείγματα να θεωρηθούν ότι είναι διαχωρίσιμα σε δύο σημεία. Αρχικά, το γεγονός ότι $e^t(k)\hat{b} = 0$ χρησιμοποιήθηκε για να δειχθεί ότι είτε το e(k) = 0 για κάποιο πεπερασμένο k, είτε το $e^+(k)$ δεν γίνεται ποτέ μηδέν και οι διορθώσεις συνεχίζονται έπ' άπειρο. Στη συνέχεια, ο ίδιος περιορισμός χρησιμοποιείται για να δειχθεί ότι εάν το $e^+(k)$ συγκλίνει στο μηδέν θα πρέπει και το e(k) να συγκλίνει υποχρεωτικά στο μηδέν.

Στην περίπτωση όπου τα δείγματα δεν είναι γραμμικά διαχωρίσιμα, δεν συνεπάγεται ότι εάν το $e^+(k)$ είναι μηδέν τότε και το e(k) θα πρέπει να είναι μηδέν. Πράγματι, σε ένα μη διαχωρίσιμο πρόβλημα, μπορεί να καταλήξει κανείς σε ένα μη μηδενικό διάνυσμα λάθους το οποίο να μην έχει κανένα θετικό στοιχείο. Εάν συμβεί αυτό, ο αλγόριθμος τερματίζεται και υπάρχει απόδειξη ότι τα δείγματα δεν είναι διαχωρίσιμα. Τι συμβαίνει όμως εάν τα δείγματα δεν είναι διαχωρίσιμα, αλλά το $e^+(k)$ δεν γίνεται ποτέ ίσο με το μηδέν; Στην περίπτωση αυτή ισχύει πάλι

$$e(k+1) = e(k) + 2\eta(YY^{\dagger} - I)e^+(k) \tag{5.86}$$

και

$$\frac{1}{4}\left(\|e(k)\|^2 - \|e(k+1)\|^2\right) = \eta(1-\eta)\|e^+(k)\|^2 + \eta^2 e^{+t}(k)\, YY^{\dagger} e^+(k) \tag{5.87}$$

Έτσι, η ακολουθία $\|e(1)\|^2, \|e(2)\|^2, \ldots$ πρέπει ξανά να συγκλίνει, αν και η οριακή τιμή $\|e\|^2$, αυτή τη φορά, δεν μπορεί να ισούται με το μηδέν. Η σύγκλιση απαιτεί να ισχύει $e^+(k) = 0$ για κάποιο πεπερασμένο k ή το $e^+(k)$ να συγκλίνει στο μηδέν ενώ το $\|e(k)\|$ είναι φραγμένο μακριά από το μηδέν. Έτσι, ο αλγόριθμος των Ho-Kashyap παρέχει ένα διάνυσμα διαχωριστή, στην περίπτωση διαχωρίσιμων δειγμάτων, και απόδειξη για μη διαχωρισιμότητα στην περίπτωση μη διαχωρίσιμων δειγμάτων. Όμως, δεν υπάρχει κάποιο άνω όριο στον αριθμό των βημάτων που απαιτούνται για να δειχθεί η μη διαχωρισιμότητα.

5.9.4 Κάποιες Σχετικές Διαδικασίες

Εάν κάποιος γράψει $Y^{\dagger} = (Y^tY)^{-1}Y^t$ και χρησιμοποιήσει το γεγονός ότι $Y^te(k) = 0$ ο κανόνας των Ho-Kashyap παίρνει την παρακάτω μορφή:

$$\begin{aligned} &b(1) > 0 \\ &a(1) = Y^{\dagger}b(1) \\ &b(k+1) = b(k) + \eta(e(k) + |e(k)|) \\ &a(k+1) = a(k) + \eta Y^{\dagger}|e(k)| \end{aligned} \tag{5.88}$$

όπου, ως συνήθως,

$$e(k) = Ya(k) - b(k) \tag{5.89}$$

Έτσι προκύπτει ο αλγόριθμος για καθορισμένο ρυθμό μάθησης:

Αλγόριθμος 12:

1 αρχή αρχικοποίησε a, b, η < 1, κατώφλι b_{min}, k_{max}
2 κάνε $k \leftarrow (k+1) \bmod n$
3 $e \leftarrow Ya - b$
4 $e^+ \leftarrow 1/2\,(e + Abs[e])$
5 $b \leftarrow b + 2\eta(k)(e + Abs[e])$
6 $a \leftarrow Y^{\dagger}b$
7 εάν $Abs[e] \le b_{min}$ τότε επέστρεψε a, b και έξοδος
8 μέχρι $k = k_{max}$

9 εμφάνισε «ΔΕΝ ΒΡΕΘΗΚΕ ΛΥΣΗ»
10 τέλος

Ο αλγόριθμος αυτός διαφέρει από τον αλγόριθμο του Perceptron και τους αλγόριθμους χαλάρωσης για την επίλυση γραμμικών ανισοτήτων σε τουλάχιστον τρία σημεία:

1. μεταβάλλει τόσο το διάνυσμα των βαρών a όσο και το διάνυσμα του ορίου b.
2. παρέχει απόδειξη στην περίπτωση μη διαχωρισιμότητας.
3. απαιτεί τον υπολογισμό του ψευδοαντίστροφου πίνακα του Y.

Εάν και ο τελευταίος αυτός υπολογισμός απαιτείται να γίνει μόνο μία φορά, μπορεί να πάρει πολύ χρόνο και απαιτεί ειδικό χειρισμό εάν ο Y^tY είναι μη ομαλός. Ένας ενδιαφέρων εναλλακτικός αλγόριθμος που μοιάζει με την εξίσωση 5.88 αλλά δεν απαιτεί τον υπολογισμό του $Y^\dagger$ είναι ο ακόλουθος:

$$\begin{aligned} &b(1) > 0 \\ &a(1) \text{ αυθαίρετο} \\ &b(k+1) = b(k) + (e(k) + |e(k)|) \\ &a(k+1) = a(k) + \eta R Y^\dagger |e(k)| \end{aligned} \tag{5.90}$$

όπου το R είναι ένας αυθαίρετος, σταθερός, θετικά ημι-ορισμένος $\hat{d} \times \hat{d}$ πίνακας. Στη συνέχεια θα δειχθεί ότι εάν η τιμή του η επιλεγεί σωστά, και αυτός ο αλγόριθμος καταλήγει σε ένα διάνυσμα λύσης σε πεπερασμένο αριθμό βημάτων, δεδομένου φυσικά ότι υπάρχει τουλάχιστον μία λύση. Επιπλέον, εάν δεν υπάρχει καμία λύση, το διάνυσμα $Y^\dagger|e(k)|$ είτε εξαφανίζεται, φανερώνοντας την μη διαχωρισιμότητα, είτε συγκλίνει στο μηδέν.

Η απόδειξη προκύπτει άμεσα. Ανεξάρτητα από το εάν τα δείγματα είναι διαχωρίσιμα ή όχι οι εξισώσεις 5.89 και 5.90 δείχνουν ότι

$$\begin{aligned} e(k+1) &= Ya(k+1) - b(k+1) \\ &= (\eta Y R Y^t - I)|e(k)| \end{aligned}$$

Στη συνέχεια προκύπτει ότι το τετράγωνο του μέτρου ισούται με

$$\|e(k+1)\|^2 = |e(k)|^t \left(\eta^2 Y R Y^t Y R Y - 2\eta Y R Y^t + I\right)|e(k)|$$

και επιπλέον

$$\|e(k)\|^2 - \|e(k+1)\|^2 = \left(Y^t|e(k)|\right)^t A\left(Y^t|e(k)|\right) \tag{5.91}$$

όπου

$$A = 2\eta R - \eta^2 R Y^t R \tag{5.92}$$

Προφανώς, εάν το η είναι θετικό, αλλά αρκετά μικρό, ο A θα ισούται περίπου με 2ηR και επομένως θα είναι θετικά ορισμένος. Έτσι, εάν $Y^t|e(k)| \neq 0$ προκύπτει $\|e(k)\|^2 > \|e(k+1)\|^2$.

Στο σημείο αυτό πρέπει να γίνει ένας διαχωρισμός μεταξύ της διαχωρίσιμης και της μη διαχωρίσιμης περίπτωσης. Στη διαχωρίσιμη περίπτωση υπάρχει ένα $\hat{a}$ και ένα $\hat{b} > 0$ που ικανοποιούν τη σχέση $Y\hat{a} = \hat{b}$. Έτσι, εάν $|e(k)| \neq 0$,

$$|e(k)|^t Y\hat{a} = |e(k)|^t \hat{b} > 0$$

και επομένως το $Y^te(k)$ δεν μπορεί να είναι ίσο με το μηδέν εκτός εάν το e(k) είναι ίσο με το μηδέν. Έτσι, η ακολουθία $\|e(1)\|^2, \|e(2)\|^2, \ldots$ μειώνεται μονότονα και πρέπει να συγκλίνει σε κάποια οριακή τιμή $\|e\|^2$. Όμως για να πραγματοποιηθεί η σύγκλιση,

το $Y^t e(k)$ πρέπει να συγκλίνει στο μηδέν, γεγονός που υπονοεί ότι το $|e(k)|$ και επομένως και το e(k) πρέπει να συγκλίνει στο μηδέν. Επειδή το e(k) έχει αρχικά θετική τιμή και η τιμή του ποτέ δεν ελαττώνεται, προκύπτει ότι το a(k) πρέπει να συγκλίνει σε ένα διάνυσμα διαχωριστή. Επιπλέον, χρησιμοποιώντας το ίδιο επιχείρημα με προηγουμένως, προκύπτει ότι θα καταλήξει σε ένα διάνυσμα λύση μετά από ένα πεπερασμένο αριθμό βημάτων. Στην μη διαχωρίσιμη περίπτωση, το e(k) δεν μπορεί ούτε να είναι ίσο με το μηδέν ούτε να συγκλίνει στο μηδέν. Σε κάποιο βήμα μπορεί να συμβεί το $Y^t|e(k)| = 0$, πράγμα που σημαίνει ότι υπάρχει μη διαχωρισιμότητα. Παρ' όλ' αυτά, υπάρχει περίπτωση η ακολουθία των διορθώσεων να συνεχίζεται έπ' άπειρο. Στην περίπτωση αυτή, προκύπτει ξανά ότι η ακολουθία $\|e(1)\|^2, \|e(2)\|^2, \ldots$ πρέπει να συγκλίνει σε μία οριακή τιμή $\|e\|^2 \neq 0$ και ότι το $Y^t|e(k)|$ πρέπει να συγκλίνει στο μηδέν. Έτσι, καταλήγει και πάλι σε απόδειξη της μη διαχωρισιμότητας στην περίπτωση μη διαχωρίσιμων δειγμάτων.

Κλείνοντας θα γίνει μια μικρή αναφορά στην επιλογή των τιμών για το η και το R. Η πιο απλή επιλογή για το R είναι ο ταυτοτικός πίνακας στην οποία περίπτωση θα ισχύει $A = 2\eta I - \eta^2 Y^t Y$. Ο πίνακας αυτός θα είναι θετικά ορισμένος, επιβεβαιώνοντας με αυτόν τον τρόπο τη σύγκλιση, εάν $0 < \eta < 2/\lambda_{max}$ όπου το λ_{max} είναι η μεγαλύτερη ιδιοτιμή του Y^tY. Επειδή το ίχνος του πίνακα Y^tY ισούται με το άθροισμα των ιδιοτιμών του αλλά και με το άθροισμα των τετραγώνων των στοιχείων του Y, μπορεί κάποιος να επιλέξει το απαισιόδοξο όριο $\hat{d}\lambda_{max} \leq \sum_i \|y_i\|^2$ για την επιλογή της τιμής του η.

Μια πιο ενδιαφέρουσα προσέγγιση είναι να μεταβάλλεται η τιμή του η σε κάθε βήμα, επιλέγοντας κάθε φορά την τιμή που μεγιστοποιεί το $\|e(k)\|^2 - \|e(k+1)\|^2$. Οι εξισώσεις 5.91 και 5.92 δίνουν

$$\|e(k)\|^2 - \|e(k+1)\|^2 = |e(k)|^t Y\left(2\eta R - \eta^2 R Y^t Y R\right) Y^t |e(k)| \tag{5.93}$$

Παίρνοντας την παράγωγο ως προς το η προκύπτει η βέλτιστη τιμή του:

$$\eta(k) = \frac{|e(k)|^t Y R Y^t |e(k)|}{|e(k)|^t Y R Y^t Y R Y^t |e(k)|} \tag{5.94}$$

η οποία, για R = I, ισούται με

$$\eta(k) = \frac{\left\|Y^t |e(k)|\right\|^2}{\left\|Y Y^t |e(k)|\right\|^2} \tag{5.95}$$

Η ίδια προσέγγιση μπορεί να ακολουθηθεί και για την επιλογή της τιμής του πίνακα R. Αντικαθιστώντας στην εξίσωση 5.93 το R με το συμμετρικό πίνακα R + δR και παραβλέποντας τους όρους δεύτερης τάξης, προκύπτει το εξής:

$$\delta\left(\|e(k)\|^2 - \|e(k+1)\|^2\right) = |e(k)| Y\left[\delta R^t \left(I - \eta Y^t Y R\right) + \left(I - \eta R Y^t Y\right)\delta R\right] Y^t |e(k)|$$

Έτσι, η μείωση στο διάνυσμα του τετραγώνου του λάθους μεγιστοποιείται από την επιλογή

$$R = \frac{1}{\eta}\left(Y^t Y\right)^{-1} \tag{5.96}$$

και επειδή $\eta R Y^t = Y^\dagger$, ο descent αλγόριθμος γίνεται στην πραγματικότητα ίδιος με τον αρχικό αλγόριθμο των Ho-Kashyap.

5.10 Βιβλιογραφία

[1] Henry D. Block and Simon A. Levin. On the bounded-ness of an iterative procedure for solving a system of linear inequalities. Proceedings of the American Mathematical Society, 26:229-235, 1970.

[2] Bernard E. Boser, Isabelle Guyon, and Vladimir Vapnik. A training algorithm for optimal margin classifiers, In David Haussler, editor. Proceedings of the 4th Workshop on Computational Learning Theory, pages 144-152, ACM Press, San Mateo, CA, 1992.

[3] Herve Bourlard and Yves Kamp. Auto-association by multilayer perceptions and singular value decomposition. Biological Cybernetics, 59:291-294, 1988.

[4] Vladimir Cherkassky and Filip Mulier. Learning from Data: Concepts, Theory, and Methods. Wiley, New York, 1998.

[5] Ronald A. Fisher. The use of multiple measurements in taxonomic problems. Annals of Eugenics, 7 Part II:179-188, 1936.

[6] David Gale, Harold W. Kuhn, and Albert W. Tucker. Linear programming and the theory of games. In Tjalling C. Koopmans, editor. Activity Analysis of Production and Allocation, pages 317-329. Wiley, New York, 1951.

[7] Adam I. Grave, Nicholas Littlestone, and Dale Schuurmans. General convergence results for linear discriminant updates. In Proceedings of the COLT 97, pages 171-183. ACM Press, 1997.

[8] Isabelle Guyon and David G. Stork. Linear discriminant and support vector classifiers. In Alex Smola, Peter Bartlett, Bernhard Scholkopf, and Dale Schuurmans, editors, Advances in large margin classifiers. MIT Press, Cambridge, MA, 1999.

[9] Wilbur H. Highleyman. Linear decision functions, with application to pattern recognition. Proceedings of the [RE, 50:1501-1514,1962.

[10] Nicholas Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. Machine Learning, 2(4):285-3l8, 1988.

[11] Nicholas Littlestone. Redundant noisy attributes, attribute errors, and linear-threshold learning using Winnow. In Manfred K. Warmuth and Leslie G. Valiant, editors, Proceedings of COLT 91, pages 147-156, Morgan Kaufmann, San Mateo, CA, 1991.

[12] Warren S. McCulloch and Walter Pitts. A logical calculus of ideas imminent in nervous activity. Bulletin of Mathematical Biophysics, 5:115-133, 1943.

[13] Marvin L. Minsky and Seymour A, Papert. Perceptrons: An Introduction to Computational Geometry, MIT Press, Cambridge. MA, 1969.

[14] Bernhard Scholkopf, Christopher J. C. Burges, and Alexander J. Smola, editors. Advances in Kernel Methods: Support Vector Learning. MIT Press, Cambridge, MA, 1999.

[15] Bernhard Scholkopf, Christopher J. C. Burges, and Vladimir Vapnik. Extracting support data for a given task. In Usama M. Fayyad and Ramasamy Uthurasamy, editors, Proceedings of the First International Conference on Knowledge Discovery ami Data Mining, pages 252-257. AAAI Press, Menlo Park, CA, 1995.

[16] Fred W. Smith. Design of multicategory pattern classifiers with two-category classifier design procedures. IEEE Transactions on Computers, C-18(6):548-551, 1969.

[17] Gilbert Strang. Introduction to Applied Mathematics. Wellesley-Cambridge Press, Wellesley, MA, 1986.

[18] Vladimir Vapnik. The Nature of Statistical Learning Theory. Springer-Verlag, New York, 1995.

[19] Sarunas Raudys. Evolution and generalization of a single neuron: I. Single-layer perceptron as seven statistical classifiers. Neural Networks, 11(2):283-296,1998.
[20] Sarunas Raudys. Evolution and generalization of a single neurone: II. Complexity of statistical classifiers and sample size considerations. Neural Networks, 11(2);297-313, 1998.
[21] Stephen S. Yau and John M. Schumpert. Design of pattern classifiers with the updating property using stochastic approximation techniques. IEEE Transactions on Computers, C-17(9): 861-872, 1968.