ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ

ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ ΕΚΠΑΙΔΕΥΣΗΣ ΑΠΟ ΑΠΟΣΤΑΣΗ

ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ	**:**	*ΠΛΗΡΟΦΟΡΙΚΗ*
ΘΕΜΑΤΙΚΗ ΕΝΟΤΗΤΑ P-INF-003	**:**	*ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ ΚΑΙ ΓΕΝΕΤΙΚΟΙ ΑΛΓΟΡΙΘΜΟΙ*

ΕΚΠΑΙΔΕΥΤΙΚΟ ΥΛΙΚΟ

ΤΕΤΑΡΤΟ ΚΕΦΑΛΑΙΟ

ΣΥΓΓΡΑΦΕΙΣ : ***Σ. ΛΥΚΟΘΑΝΑΣΗΣ***
ΕΠ. ΚΑΘΗΓΗΤΗΣ
ΤΜΗΜΑΤΟΣ ΜΗΧ/ΚΩΝ Η/Υ ΚΑΙ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΠΑΝΕΠΙΣΤΗΜΙΟΥ ΠΑΤΡΩΝ

Ε. ΓΕΩΡΓΟΠΟΥΛΟΣ
ΜΗΧΑΝΙΚΟΣ Η/Υ ΚΑΙ ΠΛΗΡΟΦΟΡΙΚΗΣ

- ΠΑΤΡΑ 1999 -

4. ΜΕΛΕΤΗ ΠΕΡΙΠΤΩΣΗΣ: ΑΡΧΕΣ ΚΑΙ ΠΕΡΙΟΡΙΣΜΟΙ ΣΧΕΔΙΑΣΜΟΥ ΤΕΧΝΗΤΩΝ ΝΕΥΡΩΝΙΚΩΝ ΔΙΚΤΥΩΝ

Σκοπός

Στο τρίτο κεφάλαιο είδαμε τους τρεις βασικούς αλγορίθμους εκπαίδευσης Τεχνητών Νευρωνικών Δικτύων. Συγκεκριμένα, πρώτα είδαμε δύο αλγορίθμους για εκπαίδευση απλών Ν.Δ. ενός επιπέδου, δηλαδή τον κανόνα του Perceptron και τον αλγόριθμο Ελάχιστων μέσων Τετραγώνων. Για την περίπτωση των Ν.Δ. πολλών επιπέδων, παρουσιάσαμε τον αλγόριθμο Πίσω Διάδοσης του λάθους και μια γενίκευσή του, τον Γενικευμένο Δέλτα κανόνα. Το κύριο βάρος στην παρουσίαση αυτών των αλγορίθμων, δόθηκε στην παραγωγή τους και άλλα θεωρητικά θέματα. Οι παραπάνω αλγόριθμοι εργάζονται για προκαθορισμένη διαμόρφωση (τοπολογία) του δικτύου.

Σε αυτό το κεφάλαιο θα ασχοληθούμε με θέματα που αφορούν την υλοποίηση Τ.Ν.Δ. πολλών επιπέδων. Θα συζητήσουμε πρώτα το πρόβλημα της αρχικοποίησης, μερικούς τρόπους αποτελεσματικότερης εκτέλεσης των αλγορίθμων και την εφαρμογή τους σαν συστήματα ταξινόμησης. Όμως, το βασικό πρόβλημα στην υλοποίηση ενός Ν.Δ., για την επίλυση ενός πραγματικού προβλήματος είναι ότι δεν είναι εκ των προτέρων γνωστή η διαμόρφωση του δικτύου. Δηλαδή δεν γνωρίζουμε τον ακριβή αριθμό κρυφών επιπέδων καθώς και τον αριθμό των κόμβων ανά κρυφό επίπεδο. Δεν υπάρχει κάποια γενική θεωρία που επιλύει αυτό το πρόβλημα. Σε αυτό το κεφάλαιο θα προσπαθήσουμε να δώσουμε μερικές ιδέες για το πως αντιμετωπίζεται ο καθορισμός του βέλτιστου δικτύου, με τη βοήθεια μιας μελέτης περίπτωσης. Αυτή η μελέτη θα γίνει με τη βοήθεια προσομοίωσης σε ηλεκτρονικό υπολογιστή.

Συνοψίζοντας, αναφέρουμε ότι σκοπός αυτού του κεφαλαίου είναι να βοηθήσει τον αναγνώστη στη σχεδίαση και υλοποίηση Τ.Ν.Δ., δίνοντάς του ιδέες για τον τρόπο αντιμετώπισης των πρακτικών προβλημάτων που προκύπτουν.

Προσδοκώμενα Αποτελέσματα:

Όταν θα έχετε τελειώσει τη μελέτη αυτού του κεφαλαίου, θα μπορείτε να:

- σχεδιάσετε ένα Τ.Ν.Δ. κατάλληλο για την επίλυση ενός πρακτικού προβλήματος,
- να υλοποιήσετε ένα τέτοιο δίκτυο,
- να αντιμετωπίσετε τα επιμέρους προβλήματα που παρουσιάζονται,
- να βελτιστοποιήσετε τη δομή ενός δικτύου.

Έννοιες Κλειδιά:

- αρχικοποίηση
- αναπαράσταση εξόδου
- κανόνας απόφασης
- βέλτιστη δομή
- γενίκευση

Εισαγωγικές Παρατηρήσεις:

Το βασικό πλεονέκτημα των Perceptrons πολλών επιπέδων είναι ότι μπορούν να διαχωρίσουν μη-γραμμικά διαχωριζόμενα πρότυπα. Θεωρητικά λοιπόν, μπορούν να αντιμετωπίσουν οποιοδήποτε πρόβλημα ταξινόμησης προτύπων. Δυστυχώς όμως, στην πράξη τα πράγματα δεν είναι τόσο απλά. Όπως ήδη αναφέραμε, το βασικό πρόβλημα ενός τέτοιου δικτύου είναι ο καθορισμός της βέλτιστης δομής που είναι κατάλληλη για την επίλυση ενός συγκεκριμένου προβλήματος. Έχοντας επιλέξει σωστά το εκπαιδευτικό σύνολο, αυτόματα έχουμε καθορίσει τον αριθμό των εισόδων και των εξόδων του δικτύου. Δεν έχουμε όμως καμία εκ των προτέρων γνώση για την εσωτερική αρχιτεκτονική του δικτύου (αριθμός κρυφών επιπέδων και αριθμός κόμβων ανά κρυφό επίπεδο). Ένας προφανής τρόπος, για να αντιμετωπίσουμε αυτό το πρόβλημα, είναι με τη μέθοδο δοκιμής και λάθους (trial and error). Με βάση κάποιους εμπειρικούς κανόνες, υλοποιούμε μια συγκεκριμένη αρχιτεκτονική δικτύου και το εκπαιδεύουμε, ελπίζοντας σε καλή γενίκευση. Αν μετά την εκπαίδευση το λάθος στην έξοδο του δικτύου είναι σχετικά μεγάλο, τροποποιούμε την δομή του (προσθέτοντας ή αφαιρώντας κρυφά επίπεδα ή/και κόμβους) και το εκπαιδεύουμε πάλι. Έτσι, με διαδοχικές δοκιμές, ελπίζουμε να επιτύχουμε το (σχεδόν) βέλτιστο δίκτυο. Είναι προφανές ότι αυτή η διαδικασία είναι χρονοβόρα και επίπονη. Γι' αυτό το λόγο λέγεται ότι μαθαίνει κανείς μαζί με το δίκτυο.

Τα τελευταία χρόνια, έχουν εμφανιστεί στη βιβλιογραφία, πιο έξυπνες μέθοδοι για την επίλυση αυτού του προβλήματος. Τέτοιες μέθοδοι είναι η μέθοδος του κλαδέματος

(pruning) [1], της ανάπτυξης (growing) [1] και της εξέλιξης Ν.Δ. [3], [4]. Η παρουσίαση αυτών των αλγορίθμων είναι πέρα από τους σκοπούς αυτού του μαθήματος και ο αναγνώστης παραπέμπεται στη σχετική βιβλιογραφία.

Στις ενότητες που ακολουθούν, θα παρουσιάσουμε τις πιο γνωστές ευριστικές μεθόδους, με τις οποίες μπορεί κανείς να αντιμετωπίσει τα παραπάνω προβλήματα.

4.1 Αρχικοποίηση

Το πρώτο βήμα της μάθησης πίσω-διάδοσης είναι να αρχικοποιήσουμε το δίκτυο. Μια καλή επιλογή για της αρχικές τιμές των ελεύθερων παραμέτρων του δικτύου θα βοηθήσει στην επιτυχία του σχεδιασμού και της λειτουργίας του δικτύου. Σε περιπτώσεις που έχουμε προηγούμενες πληροφορίες, θα ήταν καλύτερα να χρησιμοποιήσουμε τις προηγούμενες πληροφορίες για να μαντεύσουμε της αρχικές τιμές των ελεύθερων παραμέτρων. Πως όμως μπορούμε να αρχικοποιήσουμε το δίκτυο αν δεν έχουμε προηγούμενη πληροφορία; Η συνηθισμένη πρακτική είναι να θέσουμε στις ελεύθερους παραμέτρους του δικτύου τυχαίους αριθμούς που είναι ομοιόμορφα κατανεμημένοι μέσα σε μία μικρή περιοχή τιμών (π.χ. –0.1 μέχρι +0.1).

Η λανθασμένη επιλογή των αρχικών βαρών μπορεί να οδηγήσει σ'ένα φαινόμενο γνωστό ως 'πρόωρος κορεσμός' (Lee et al., 1991[1]). Αυτό το φαινόμενο αναφέρεται σε μία κατάσταση όπου το στιγμιαίο άθροισμα των τετραγωνικών λαθών $\boldsymbol{E(n)}$ παραμένει σχεδόν σταθερό για μία περίοδο χρόνου κατά τη διάρκεια μάθησης. Ένα τέτοιο φαινόμενο δεν μπορεί να θεωρηθεί σαν ένα τοπικό ελάχιστο γιατί το τετραγωνικό λάθος συνεχίζει να μειώνεται το τέλος αυτής της περιόδου.

Όταν εφαρμόζεται ένα δείγμα μάθησης στο επίπεδο εισόδου ενός πολυεπίπεδου perceptron, οι τιμές εξόδου του δικτύου υπολογίζονται μέσω μίας σειράς από εμπρός υπολογισμός που περιλαμβάνει εσωτερικά γινόμενα και σιγμοειδείς μετατροπές. Αυτό ακολουθείται με μία σειρά από πίσω υπολογισμούς των σημάτων λάθους και της σχετικής κλίσης της σιγμοειδούς συνάρτησης ενεργοποίησης και καταλήγει στις ρυθμίσεις των συναπτικών βαρών. Ας υποθέσουμε ότι για ένα συγκεκριμένο δείγμα μάθησης υπολογίστηκε το εσωτερικό επίπεδο ενεργοποίησης του δικτύου, για έναν νευρώνα εξόδου, να έχει μεγάλο μέγεθος. Τότε, θεωρώντας ότι η σιγμοειδής συνάρτηση ενεργοποίησης του νευρώνα έχει τις περιορισμένες τιμές -1 και +1, βρίσκουμε ότι η αντίστοιχη κλίση της συνάρτησης ενεργοποίησης για αυτόν τον νευρώνα θα είναι πολύ μικρή και η τιμή εξόδου του νευρώνα θα πλησιάζει το -1 ή +1.

Σε αυτή την περίπτωση λέμε ότι ο νευρώνας είναι σε κόρο. Αν η τιμή εξόδου πλησιάζει το +1 όταν η επιθυμητή απόκριση είναι -1, ή αντίστροφα, τότε λέμε ότι ο νευρώνας είναι κατά λάθος κορεσμένος. Όταν συμβαίνει αυτό, οι μεταβολές που γίνονται πάνω στα συναπτικά βάρη του νευρώνα θα είναι μικρές (ας είναι μεγάλο το μέγεθος του σχετικού σήματος λάθους) και το δίκτυο θα αργήσει να ξεφύγει από αυτή τη περίπτωση (Lee et al., 1991[1]).

Στην αρχική φάση της πίσω-διάδοσης μάθησης, ανάλογα με τις επικρατούσες συνθήκες, νευρώνες που είναι ή δεν είναι κορεσμένοι, μπορούν να υπάρχουν στο επίπεδο εξόδου του δικτύου. Καθώς η διαδικασία μάθησης συνεχίζεται, τα συναπτικά βάρη που σχετίζονται με τους μη κορεσμένους νευρώνες εξόδου αλλάζουν γρήγορα, γιατί τα αντίστοιχα σήματα λάθους και κλίσης έχουν σχετικά μεγάλο μέγεθος, και για αυτό το λόγο προκαλούν μία ελάττωση του στιγμιαίου αθροίσματος των τετραγωνικών λαθών $\boldsymbol{E(n)}$. Αν ωστόσο, σ' αυτό το σημείο οι κατά λάθος κορεσμένοι νευρώνες εξόδου μείνουν κορεσμένοι για μερικά δείγματα μάθησης, τότε το φαινόμενο του πρόωρου κορεσμού μπορεί να εμφανιστεί με το $\boldsymbol{E(n)}$ να παραμένει σταθερό.

Στην αναφορά Lee et al. (1991), μια φόρμουλα για την πιθανότητα του πρόωρου κορεσμού στην μάθηση πίσω-διάδοσης έχει παραχθεί για το σωρηδόν τρόπο ενημέρωσης, και έχει επαληθευτεί χρησιμοποιώντας εξομοίωση με υπολογιστή. Η ουσία αυτού του τύπου συνοψίζεται παρακάτω:

1. Αν επιλέξουμε της αρχικές τιμές για τα συναπτικά βάρη και για τα επίπεδα κατωφλίων του δικτύου να είναι ομοιόμορφα κατανεμημένες μέσα σε μία μικρή περιοχή τιμών τότε αποφεύγεται ο κατά λάθος κορεσμός.
2. Ο κατά λάθος κορεσμός έχει μικρότερη πιθανότητα να συμβεί όταν έχουμε μικρό αριθμό από κρυμμένους νευρώνες και έχουμε ικανοποιητική λειτουργία του δικτύου.
3. Ο κατά λάθος κορεσμός σπάνια συμβαίνει όταν οι νευρώνες του δικτύου λειτουργούν σε γραμμικές περιοχές.

Για τον τρόπο ενημέρωσης των βαρών πρότυπο προς πρότυπο, τα αποτελέσματα εξομοιώσεων δείχνουν παρόμοιες τάσεις με τη σωρηδόν λειτουργία που αναφέρθηκε μέχρι εδώ. Ο Russo (1991) [1] συνιστά ένα εμπειρικό τύπο για της αρχικές τιμές των βαρών για να αποφεύγεται ο κορεσμός των νευρώνων. Αυτό το τύπο τον περιγράφουμε

στο σημείο 3 της ενότητας 4.3.

Άσκηση αυτοαξιολόγησης 4.1 /1:

Ο πρόωρος κορεσμός είναι η κατάσταση όπου στο Ν.Δ.:

1. το στιγμιαίο άθροισμα των τετραγωνικών λαθών παραμένει σταθερό για μια χρονική περίοδο και η κλίση της συνάρτησης ενεργοποίησης του νευρώνα είναι πολύ μικρή,
2. η κλίση της συνάρτησης ενεργοποίησης του νευρώνα είναι πολύ μεγάλη,
3. τα σήματα λάθους έχουν μεγάλο μέγεθος,
4. το μέσο τετραγωνικό λάθος παραμένει σταθερό.

Απάντηση: Η σωστή απάντηση είναι η 1.

Άσκηση αυτοαξιολόγησης 4.1 / 2:

Ποιές είναι οι βασικές αρχές, για να αποφύγουμε τον πρόωρο κορεσμό;

Απάντηση: Είναι οι τρεις αρχές που αναφέρονται στο τέλος της ενότητας 4.1.

4.2 Το πρόβλημα της XOR

Στο απλό perceptron δεν υπάρχουν κρυφοί νευρώνες. Κατά συνέπεια, δεν μπορεί να ταξινομήσει πρότυπα εισόδου τα οποία δεν είναι γραμμικά διαχωρίσιμα. Όμως μη-γραμμικά διαχωρίσιμα πρότυπα εμφανίζονται συνήθως στα προβλήματα. Για παράδειγμα αυτό εμφανίζεται στο πρόβλημα Exclusive OR (XOR), το οποίο μπορεί να θεωρηθεί σαν μια ειδική περίπτωση ενός πιο γενικού προβλήματος, αυτού της ταξινόμησης σημείων στον μοναδιαίο υπερκύβο. Κάθε σημείο στον υπερκύβο είναι είτε στην τάξη 0 είτε στην τάξη 1. Ωστόσο, στην ειδική περίπτωση του προβλήματος XOR, χρειάζεται να λάβουμε υπόψη μας μόνο τις τέσσερις γωνίες (άκρα) του μοναδιαίου τετραγώνου, οι οποίες αντιστοιχούν στα πρότυπα εισόδου (0,0), (0,1), (1,1) και (1,0). Το πρώτο και τρίτο πρότυπο εισόδου είναι στην τάξη 0 όπως φαίνεται από την :

$$0 \text{ XOR } 0 = 0$$

και:

1 XOR 1 = 0

Τα πρότυπα εισόδου (0,0) και (1,1) είναι σε αντίθετες γωνίες του μοναδιαίου τετραγώνου, και παρόλα αυτά παράγουν την ταυτόσημη έξοδο 0. Από την άλλη μεριά, τα πρότυπα εισόδου (0,1) και (1,0) είναι, επίσης σε αντίθετες γωνίες του τετραγώνου, αλλά είναι στην τάξη 1, όπως φαίνεται από την :

0 XOR 1 = 1

και :

1 XOR 0 = 1

Αρχικά αναγνωρίζουμε ότι η χρήση ενός απλού νευρώνα με δυο εισόδους έχει ως αποτέλεσμα μια ευθεία γραμμή για όριο απόφασης στον χώρο εισόδου. Για όλα τα σημεία στη μια πλευρά αυτής της γραμμής ο νευρώνας δίνει έξοδο 1, ενώ για όλα τα σημεία στην άλλη πλευρά της γραμμής δίνει έξοδο 0. Η θέση και ο προσανατολισμός της γραμμής στον χώρο εισόδου καθορίζονται από τα συναπτικά βάρη του νευρώνα με τα οποία είναι συνδεδεμένος με τους κόμβους εισόδου, και το κατώφλι που εφαρμόζεται στο νευρώνα. Με τα πρότυπα εισόδου (0,0) και (1,1) τοποθετημένα σε αντίθετες γωνίες του μοναδιαίου τετραγώνου και παρομοίως για τα άλλα δυο πρότυπα εισόδου (0,1) και (1,0), είναι προφανές ότι δεν μπορούμε να κατασκευάσουμε μια ευθεία γραμμή για ένα όριο απόφασης έτσι ώστε (0,0) και (1,1) να βρίσκονται στην μια περιοχή απόφασης και (0,1) και (1,0) να βρίσκονται στην άλλη περιοχή απόφασης. Με άλλα λόγια ένα στοιχειώδες perceptron δεν μπορεί να επιλύσει το πρόβλημα της XOR.

Μπορεί να επιλύσουμε το πρόβλημα της XOR χρησιμοποιώντας ένα απλό κρυφό επίπεδο με δυο νευρώνες, όπως στο σχήμα 1.a (Touretzky and Pomerleau, 1989) [1]. Το διάγραμμα ροής σήματος του δικτύου φαίνεται στο σχήμα 1.b. Οι υποθέσεις που έγιναν εδώ είναι οι ακόλουθες :

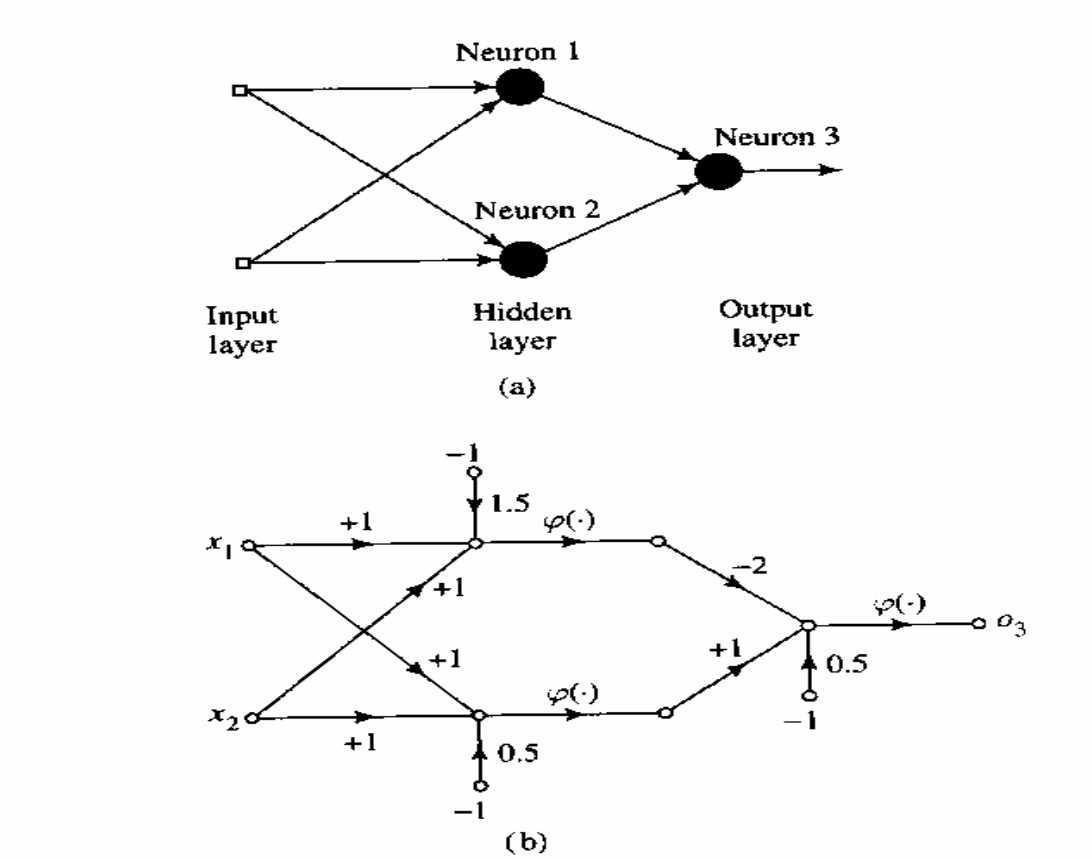
- Κάθε νευρώνας αναπαρίσταται από ένα μοντέλο McCulloch-Pitts.
- Τα ψηφία 0 και 1 αναπαρίστανται από τα επίπεδα 0 και +1 αντίστοιχα.

Ο νευρώνας κορυφής, που επιγράφεται ως 1 στο κρυφό επίπεδο, χαρακτηρίζεται από τις ακόλουθές σχέσεις :

$$w_{11} = w_{12} = +1$$

$$\theta_1 = +3/2$$

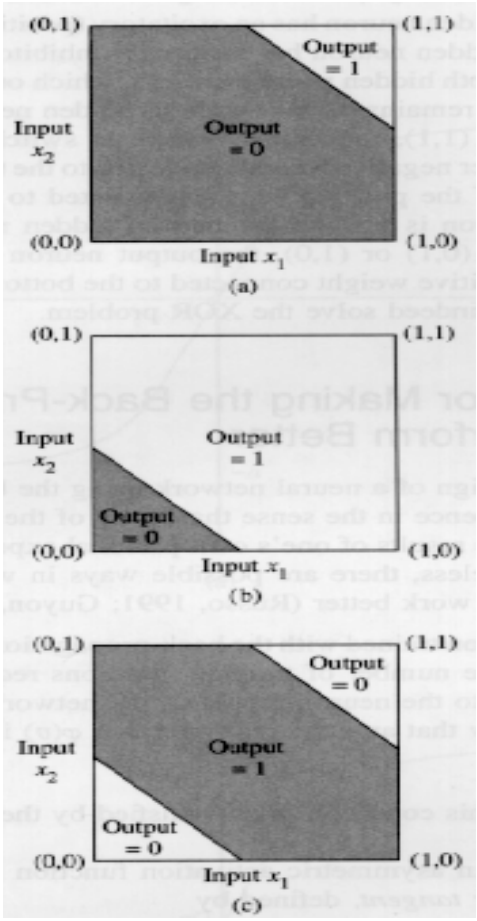
Η κλίση του ορίου απόφασης που κατασκευάζεται από αυτόν τον κρυφό νευρώνα ισούται με -1, και τοποθετείται όπως στο σχήμα 2.b. Ο νευρώνας πυθμένα, που επιγράφεται ως 2 στο κρυφό επίπεδο, χαρακτηρίζεται από τις ακόλουθες σχέσεις :



Σχήμα 1:(α) Ο αρχιτεκτονικός γράφος δικτύου για την επίλυση του XOR προβλήματος. (β) Το διάγραμμα ροής σημάτων του δικτύου.

$w_{21} = w_{22} = +1$

$\theta_2 = +1/2$



Σχήμα 2: (α) Όριο απόφασης που κατασκευάστηκε από τον κρυμμένο νευρώνα 1 του δικτύου στο σχήμα 1. (β) Όριο απόφασης που κατασκευάστηκε από τον κρυμμένο νευρώνα 2 του δικτύου. (γ) Όρια απόφασης που κατασκευάστηκαν από το συνολικό δίκτυο.

Ο προσανατολισμός και η θέση του ορίου απόφασης που κατασκευάζεται από αυτό το δεύτερο κρυφό νευρώνα είναι όπως φαίνεται στο σχήμα 2.b.

Ο νευρώνας εξόδου, που επιγράφεται ως 3 στο σχήμα 1.a, χαρακτηρίζεται από τις ακόλουθες σχέσεις :

$$w_{31} = -2$$

$$w_{32} = +1$$

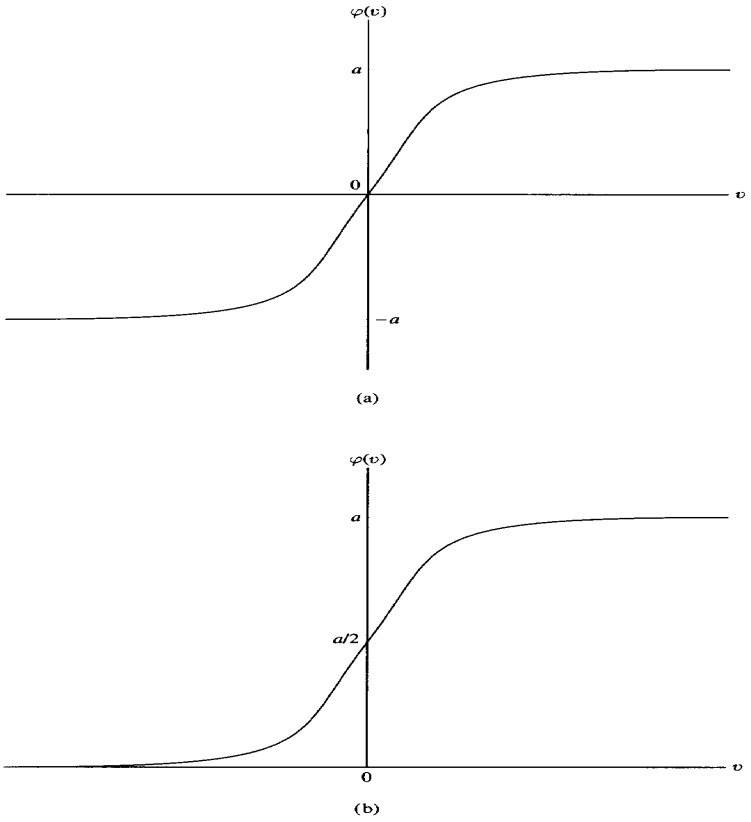
$$\theta_3 = +1/2$$

Η λειτουργία του νευρώνα εξόδου είναι να κατασκευάζει ένα γραμμικό συνδυασμό των ορίων απόφασης που διαμορφώνονται από τους δύο κρυφούς νευρώνες. Το αποτέλεσμα αυτού του υπολογισμού φαίνεται στο σχήμα 2.c. Ο κρυφός νευρώνας στο κάτω μέρος έχει μια διεγερτική (θετική) σύνδεση με τον νευρώνα εξόδου ενώ ο

νευρώνας στο πάνω μέρος έχει μια δυνατότερη κωλυτική (αρνητική) σύνδεση με το νευρώνα εξόδου. Όταν και οι δύο κρυφοί νευρώνες είναι off, το οποίο συμβαίνει όταν το πρότυπο εισόδου είναι (0,0), ο νευρώνας εξόδου παραμένει off. Όταν και οι δύο κρυφοί νευρώνες είναι on, το οποίο συμβαίνει όταν το πρότυπο εισόδου είναι (1,1), ο νευρώνας εξόδου γίνεται πάλι off, επειδή η κωλυτική επίδραση του μεγαλύτερου αρνητικού βάρους, που είναι συνδεδεμένο με τον κρυφό νευρώνα κορυφής υπερκαλύπτει την διεγερτική επίδραση του θετικού βάρους που είναι συνδεδεμένο με τον κρυφό νευρώνα στον πυθμένα. Όταν ο κρυφός νευρώνας κορυφής είναι off και ο κρυφός κάτω νευρώνας είναι on, το οποίο συμβαίνει όταν το πρότυπο εισόδου είναι (0,1) ή (1,0), ο νευρώνας εξόδου τίθεται on εξαιτίας της διεγερτικής επίδρασης του θετικού βάρους που είναι συνδεδεμένο με τον κάτω κρυφό νευρώνα. Επομένως το δίκτυο του σχήματος 1.a επιλύει πραγματικά το πρόβλημα της XOR.

4.3 Μερικοί τρόποι αποτελεσματικότερης εκτέλεσης του αλγορίθμου

Πολλές φορές λέγεται πως ο σχεδιασμός ενός νευρωνικού δικτύου χρησιμοποιώντας τον αλγόριθμο πίσω διάδοσης είναι περισσότερο τέχνη παρά επιστήμη, με την έννοια πως πολλοί από τους πολυάριθμους παράγοντες που περιλαμβάνονται στον σχεδιασμό είναι πράγματι το αποτέλεσμα της προσωπικής πείρας του καθενός μας. Πράγματι υπάρχει κάποια αλήθεια στην προηγούμενη πρόταση. Εν τούτοις όμως υπάρχουν τρόποι με τους οποίους είναι δυνατόν να κάνουμε τον αλγόριθμο πίσω διάδοσης να δουλεύει καλύτερα. (Russo, 1991; Guyon, 1991) [1].



Σχήμα 3: (α) ασυμμετρική συνάρτηση ενεργοποίησης (υπερβολική) (b) μη συμμετρική συνάρτηση ενεργοποίησης (λογιστική)

1. Ένα πολυεπίπεδο perceptron εκπαιδευμένο με τον αλγόριθμο πίσω διάδοσης θα μπορούσε, στην γενική περίπτωση να μαθαίνει γρηγορότερα (με την έννοια του αριθμού των επαναλήψεων για μάθηση που χρειάζονται) όταν η σιγμοειδής συνάρτηση ενεργοποίησης που έχει υλοποιηθεί στο μοντέλο του νευρώνα του δικτύου είναι ασυμμετρική (asymmetric) παρά όταν είναι μη συμμετρική. Εμείς λέμε πως μια συνάρτηση ενεργοποίησης φ(υ) είναι ασυμμετρική όταν:

$$\varphi(-\upsilon) = -\varphi(\upsilon)$$

όπως απεικονίζεται στο σχήμα 3.a. Αυτή η συνθήκη δεν ικανοποιείται από την λογιστική συνάρτηση, που απεικονίζεται στο σχήμα 3.b.

Ένα γνωστό παράδειγμα μιας ασυμμετρικής συνάρτησης ενεργοποίησης είναι η σιγμοειδής μη γραμμική με την μορφή της υπερβολικής εφαπτομένης, που ορίζεται ως

$$\varphi(\upsilon) = a \tanh(b\upsilon)$$

όπου τα a και b είναι σταθερές. Σημειώνουμε πως η υπερβολική εφαπτομένη είναι απλά η λογιστική συνάρτηση πολωμένη όπως δείχνεται παρακάτω:

$$a \tanh(b\upsilon) = \alpha\left[\frac{1-\exp(-b\upsilon)}{1+\exp(-b\upsilon)}\right] = \frac{2a}{1+\exp(-b\upsilon)} - \alpha \qquad (1)$$

Αντίστοιχα, οι αλλαγές που γίνονται στην διατύπωση του αλγορίθμου πίσω διάδοσης χρησιμοποιώντας αυτήν την μορφή της σιγμοειδούς μη γραμμικότητας είναι μικρής σημασίας.

Κατάλληλες τιμές για τις σταθερές α και b είναι (Guyon 1991)[1]:

$\alpha = 1.716$ και $b = 2/3$

2. Είναι σημαντικό οι τιμές στόχος (επιθυμητή απόκριση) να επιλεγούν μέσα στο διάστημα της σιγμοειδούς συνάρτησης ενεργοποίησης. Πιο ειδικά, η επιθυμητή απόκριση dj για τον νευρώνα j στο επίπεδο εξόδου του πολυεπίπεδου perceptron θα πρέπει να μετατοπιστεί κατά κάποια ποσότητα **ε** μακριά από την περιοριστική τιμή της σιγμοειδούς συνάρτησης ενεργοποίησης , εξαρτώμενη από το εάν η περιοριστική τιμή είναι θετική ή αρνητική. Στην αντίθετη περίπτωση , ο αλγόριθμος πίσω διάδοσης έχει την τάση να οδηγήσει τις ελεύθερες παραμέτρους του δικτύου στο άπειρο και ως εκ τούτου επιβραδύνει την διαδικασία μάθησης κατά τάξεις μεγέθους. Πιο ειδικά θεωρήστε την ασυμμετρική συνάρτηση ενεργοποίησης του σχήματος 3.a. Για την πεπερασμένη τιμή **+α** εμείς θέτουμε dj= α - ε όπου ε είναι μια κατάλληλη θετική σταθερά. Για την επιλογή α = 1.716 που αναφέρθηκε νωρίτερα, μπορούμε να θέσουμε ε = 0.716 , στην οποία περίπτωση η επιθυμητή τιμή dj μπορεί να επιλεγεί καταλλήλως να είναι $\pm$ 1.

3. Η αρχικοποίηση των συναπτικών βαρών και των επιπέδων κατωφλιού του δικτύου θα πρέπει να είναι ομοιόμορφα κατανεμημένα σ' ένα μικρό διάστημα. Ο λόγος που μικραίνουμε το διάστημα είναι να μειώσουμε την πιθανότητα των νευρώνων στο δίκτυο να κορεστούν και να παράγουν μικρές κλίσεις λάθους. Εν τούτοις όμως, το διάστημα δεν θα πρέπει να γίνει πολύ μικρό διότι αυτό μπορεί να έχει αποτέλεσμα οι κλίσεις λάθους να είναι πολύ μικρές και δια τούτο η μάθηση αρχικά να είναι πολύ αργή. Για μια ασυμμετρική συνάρτηση ενεργοποίησης με την μορφή της υπερβολικής

εφαπτομένης για την οποία οι σταθερές a και b προσδιορίστηκαν παραπάνω μια πιθανή δυνατότητα αρχικοποίησης είναι να πάρουμε τυχαία τιμές για τα συναπτικά βάρη και επίπεδα κατωφλιού οι οποίες είναι ομοιόμορφα κατανεμημένες στο διάστημα:

$$\left(-\frac{2.4}{Fi},+\frac{2.4}{Fi}\right)$$

όπου το Fi είναι το fan-in (δηλ. ο συνολικός αριθμός των εισόδων) του νευρώνα i μέσα στο δίκτυο, με άλλα λόγια, η αρχικοποίηση βαρών γίνεται από νευρώνα σε νευρώνα.

4. Όλοι οι νευρώνες του πολυεπίπεδου perceptron είναι επιθυμητό να μαθαίνουν με τη ίδια ταχύτητα. Χαρακτηριστικά τα τελευταία επίπεδα τείνουν να έχουν μεγαλύτερες τοπικές κλίσεις σε σχέση με τα επίπεδα στην αρχή του δικτύου. Δια τούτου στην παράμετρο $\boldsymbol{\eta}$ του ρυθμού μάθησης θα πρέπει να εκχωρηθεί μια μικρότερη τιμή στα τελευταία επίπεδα σε σχέση με τα εμπρός επίπεδα. Νευρώνες με πολλές εισόδους θα πρέπει να έχουν μικρότερη παράμετρο ρυθμού μάθησης σε σχέση με νευρώνες που έχουν λίγες εισόδους.

5. Για on-line λειτουργία, η ενημέρωση πρότυπο προς πρότυπο είναι προτιμότερο σε σχέση με σωρηδόν ενημέρωση για την ρύθμιση των βαρών. Για προβλήματα ταξινόμησης δειγμάτων που περιλαμβάνουν μια μεγάλη και πλεονάζουσα βάση δεδομένων, η πρότυπο προς πρότυπο ενημέρωση τείνει να είναι τάξεις μεγέθους γρηγορότερη σε σχέση με σωρηδόν ενημέρωση. Ωστόσο όμως η πρότυπο προς πρότυπο ενημέρωση είναι πιο δύσκολο να εκτελεστεί παράλληλα.

5. Η σειρά με την οποία εμφανίζονται τα παραδείγματα εκπαίδευσης στο δίκτυο θα πρέπει να καθορίζεται τυχαία από κύκλο σε κύκλο. Η μορφή της τυχαιοποίησης είναι κρίσιμη για την παραπέρα βελτίωση στην ταχύτητα της σύγκλισης. Επίσης η χρήση ουρών μπορεί να βελτιώσει την εκπαίδευση αποδοτικά (Baum, 1991) [1].

7. Η εκμάθηση από ένα σύνολο δειγμάτων εκπαίδευσης σχετίζεται με μια άγνωστη συνάρτηση f(.) η οποία αντιστοιχίζει την είσοδο στην έξοδο. Ως εκ τούτου, η διαδικασία μάθησης εκμεταλλεύεται την πληροφορία που περιλαμβάνεται στα

παραδείγματα σχετικά με την συνάρτηση f(.) και εξάγει μια προσεγγιστική υλοποίηση της. Η διαδικασία μάθησης από παραδείγματα μπορεί να γενικευτεί, ώστε να συμπεριλάβει μάθηση από υπαινιγμούς, το οποίο πετυχαίνεται επιτρέποντας πληροφορία εκ των προτέρων που έχουμε για την συνάρτηση f(.) να συμπεριληφθεί στην διαδικασία μάθησης (Abu-Mostafa, 1990; Al-Mashouq and Reed, 1991) [1].

Άσκηση αυτοαξιολόγησης 4.3 / 3:

Να αναφέρετε τους βασικούς τρόπους για την αποτελεσματικότερη εκτέλεση του αλγορίθμου. Ποιοί από αυτούς είναι οι πιο σημαντικοί και γιατί.

Απάντηση: Οι βασικοί τρόποι αναφέρονται σε αυτή την ενότητα. Οι πιο σημαντικοί είναι η επιλογή της συνάρτησης ενεργοποίησης, η αρχικοποίηση των βαρών και οι κατάλληλες τιμές των παραμέτρων μάθησης και ορμής.

4.4 Αναπαράσταση εξόδου και κανόνα απόφασης

Στην θεωρία για ένα m - κλάσεων πρόβλημα ταξινόμησης στο οποίο η ένωση των m διακριτών κλάσεων σχηματίζει το συνολικό χώρο εισόδου εμείς χρειαζόμαστε m εξόδους για να αναπαραστήσουμε όλες τις πιθανές αποφάσεις ταξινόμησης, όπως απεικονίζεται στο σχήμα 4. Στο σχήμα αυτό, το διάνυσμα xj αναφέρεται στο j-στο προτότυπο (δηλαδή μοναδικό δείγμα) ενός p-διαστάσεων τυχαίου διανύσματος x το οποίο πρέπει να ταξινομηθεί από ένα πολυεπίπεδο perceptron. Το k-στο από τις m δυνατές κλάσεις στις οποίες μπορεί να ανήκει το x το συμβολίζουμε με c_k. Έστω $y_{k,j}$ είναι η k-στη έξοδος του δικτύου που παράγεται σε απόκριση του πρωτοτύπου x_j, όπως δείχνεται από την σχέση:

$$y_{k,j} = F_k(x_j), \quad k = 1,2, \ldots ,m \qquad (2)$$

όπου η συνάρτηση $F_k(.)$ ορίζει την απεικόνιση την οποία έχει μάθει το δίκτυο της εισόδου στην k-στη έξοδο. Για λόγους ευκολίας στην αναπαράσταση , έστω:

$$y_j = [\, y_{1,j} , y_{2,j} , \ldots , y_{m,j} \,]^T = [\, F_1(x_j) , F_2(x_j) , \ldots , F_m(x_j) \,]^T = \mathbf{F(\, x_j)} \qquad (3)$$

όπου η **F(.)** είναι συνάρτηση που παίρνει ως τιμές διανύσματα. Μια βασική ερώτηση την οποία θέλουμε να μελετήσουμε στην παράγραφο αυτή είναι η ακόλουθη :

Μετά την εκπαίδευση του πολυεπίπεδου perceptron ποιος θα πρέπει να είναι ο βέλτιστος κανόνας απόφασης για την ταξινόμηση των m εξόδων του δικτύου;

Σίγουρα κάθε λογικός κανόνας απόφασης εξόδου οφείλει να βασιστεί στην γνώση της διανυσματικής συνάρτησης:

$$\mathbf{F}: R^p \ni x \rightarrow y \in R^m \qquad (4)$$

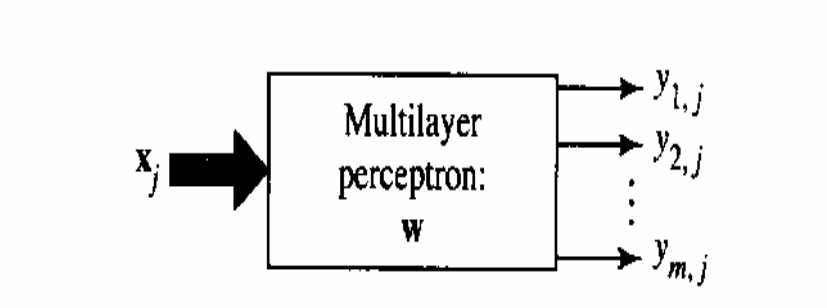
Γενικά το βέβαιο για την διανυσματική συνάρτηση **F( )** είναι ότι είναι μια συνεχή συνάρτηση η οποία ελαχιστοποιεί το μέσο τετραγωνικό σφάλμα το οποίο ορίζεται ως η τιμή της συνάρτησης κόστους:

$$L(\mathbf{F}) = \frac{1}{2N}\sum_{j=1}^{N} || d_j - \mathbf{F}(x_j) ||^2 \qquad (5)$$

όπου το d_j είναι το επιθυμητό (στοχευόμενο) δείγμα εξόδου για το πρότυπο x_j , $||\bullet||$ είναι η Ευκλείδια νόρμα του διανύσματος και N είναι ο συνολικός αριθμός των δειγμάτων εισόδου - εξόδου που παρουσιάζονται στο δίκτυο για εκπαίδευση. Η ουσία του κριτηρίου του μέσου τετραγωνικού σφάλματος της εξίσωσης (5) είναι ίδια με αυτή της συνάρτησης κόστους της εξίσωσης (11) της ενότητας 3.2.1. Η συνάρτηση **F(.)** είναι στενά εξαρτημένη από την επιλογή των ζευγαριών εισόδου - εξόδου (x_j , y_j) τα οποία χρησιμοποιούνται στην εκπαίδευση του δικτύου, έτσι ώστε διαφορετικές τιμές των (x_j , y_j) να οδηγήσουν πράγματι σε διαφορετικές διανυσματικές συναρτήσεις **F(.)**. Σημειώνουμε ότι η ορολογία (x_j , y_j) που χρησιμοποιούμε εδώ είναι η ίδια όπως η $[x(j),y(j)]$ που χρησιμοποιήσαμε προηγουμένως. Υποθέστε τώρα ότι το δίκτυο εκπαιδεύεται με δυαδικές τιμές (οι οποίες συμπτωματικά ανταποκρίνονται στο πάνω και κάτω όριο της εξόδου του δικτύου όταν χρησιμοποιούμε τη λογιστική συνάρτηση), γραμμένες ως εξής:

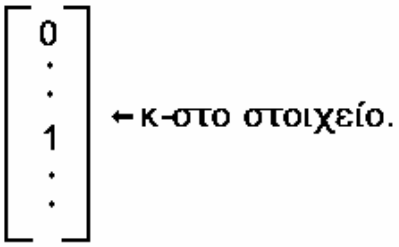
$$d_{k,j} = \begin{cases} 1 & \text{όταν το πρότυπο } x_j \text{ ανήκει στην κλάση } \mathbb{C}_k \\ \end{cases} \qquad (6)$$

0 όταν το πρότυπο x_j δεν ανήκει στην κλάση Ck



Σχήμα 4: Μπλοκ διάγραμμα ενός ταξινομητή προτύπων

Βασισμένη σε αυτή την σημειογραφία, η κλάση Ck αντιπροσωπεύεται από ένα m-διάστατο διάνυσμα:



Είναι δελεαστικό να υποθέσουμε ότι ένας πολυεπίπεδοw perceptron ταξινομητής εκπαιδευμένος με τον αλγόριθμο πίσω διάδοσης πάνω σε ένα πεπερασμένο σύνολο ανεξάρτητων όμοια κατανεμημένων παραδειγμάτων μπορεί να μας οδηγήσει σε ασυμπτωτική προσέγγιση των βασικών εκ των υστέρων πιθανοτήτων κλάσεων. Αυτή η ιδιότητα μπορεί να δικαιολογηθεί βασιζόμενη στους ακόλουθους λόγους (White, 1989; Richard and Lippmann, 1991) [1]:

- Επικαλούμαστε τον νόμο των μεγάλων αριθμών για να δείξουμε ότι καθώς το μέγεθος του συνόλου εκπαίδευσης N, προσεγγίζει το άπειρο το διάνυσμα βαρών w το οποίο ελαχιστοποιεί την συνάρτηση κόστους L(**F**) της εξίσωσης (5) προσεγγίζει το βέλτιστο διάνυσμα βαρών w* το οποίο ελαχιστοποιεί την προσδοκία της τυχαίας ποσότητας ½ || d - **F(**w,x**)** $||^2$, όπου το d είναι το επιθυμητό διάνυσμα απόκρισης και το **F(**w,x**)** είναι η προσέγγιση πραγματοποιημένη από ένα πολυεπίπεδο perceptron με διάνυσμα βαρών w και διάνυσμα x ως εισόδου (White, 1989) [1]. Η συνάρτηση **F(**w,x**)** ,που δείχνει σαφή εξάρτηση στο διάνυσμα βαρών w , είναι ίδια με την **F(**x**)** που χρησιμοποιήσαμε προηγουμένως.

- Το βέλτιστο διάνυσμα βαρών w* έχει την ιδιότητα ότι το αντίστοιχο διάνυσμα των παρόντων εξόδων του δικτύου, **F(**w*,x**)** είναι μια μέση τετραγωνική προσέγγιση ελαχιστοποίησης λάθους σε σχέση με την υπό συνθήκη προσδοκία E[d|x] (White, 1989) [1].

- Για το πρόβλημα ταξινόμησης *l* εκ των m δειγμάτων, το k-στο στοιχείο του επιθυμητού διανύσματος απόκρισης d είναι ίσο με ένα εάν το διάνυσμα εισόδου ανήκει στην κλάση ck και μηδέν αλλιώς. Κάτω από αυτήν την συνθήκη , η υπό συνθήκη προσδοκία E[d|x] είναι ίση με την εκ των υστέρων πιθανότητα κλάσης P(ck | x), k=1,2, ... ,m (Richard and Lippmann, 1991) [1].

Άρα συμπεραίνουμε πως ένας πολυεπίπεδος perceptron ταξινομητής (που χρησιμοποιεί την λογιστική συνάρτηση για μη γραμμικότητα) πράγματι προσεγγίζει τις εκ των υστέρων πιθανότητες κλάσεων , υπό τον όρο ότι το μέγεθος του συνόλου εκπαίδευσης είναι αρκετά μεγάλο και η διαδικασία της μάθησης της πίσω διάδοσης δεν κολλάει σε ένα τοπικό ελάχιστο. Τώρα μπορούμε να προχωρήσουμε στο να απαντήσουμε στην ερώτηση που είχαμε θέση πριν. Ειδικότερα μπορούμε να πούμε πως ο πιο κατάλληλος κανόνας απόφασης είναι:

Ταξινόμησε το τυχαίο διάνυσμα x ως ανήκων στην κλάση ck εάν

$$F_k(x) > F_j(x) \text{ για όλα τα } j \neq k \qquad (7)$$

όπου Fk(x) και Fj(x) είναι στοιχεία της διανυσματικής συνάρτησης:

$$F(x) = \begin{bmatrix} F_1(x) \\ F_2(x) \\ \cdot \\ \cdot \\ F_3(x) \end{bmatrix}$$

Όταν οι βασικές εκ των υστέρων κατανομές κλάσεων είναι διακεκριμένες τότε υπάρχει μια μοναδική μέγιστη τιμή εξόδου με πιθανότητα ένα. Για αυτό το λόγο, αυτός ο κανόνας απόφασης έχει το πλεονέκτημα της προσφοράς αναμφίβολων αποφάσεων

πάνω στο κοινό ad hoc κανόνα της επιλογής κλάσης βασισμένος στην γενική ιδέα της εκπυρσοκρότησης της εξόδου. Δηλαδή το διάνυσμα x είναι εκχωρημένο μέλος σε μια συγκεκριμένη κλάση εάν η αντίστοιχη τιμή εξόδου είναι μεγαλύτερη από κάποιο σταθερό κατώφλι (συνήθως 0.5 για την λογιστική συνάρτηση ενεργοποίησης) το οποίο μπορεί να οδηγήσει σε πολλαπλές εκχωρήσεις κλάσεων.

Επίσης είναι ενδιαφέρον να σημειώσουμε πως όταν ένα όριο απόφασης καθορίζεται συγκρίνοντας τις εξόδους του πολυεπίπεδου perceptron απέναντι σε κάποιες φιξαρισμένες τιμές, η ολική μορφή και ο προσανατολισμός του ορίου απόφασης θα μπορούσε να εξηγηθεί ευριστικά (για την περίπτωση ενός κρυμμένου επιπέδου) σε όρους του αριθμού των κρυμμένων νευρώνων και των αναλογιών των συναπτικών βαρών συνδεμένων με αυτά (Lui, 1989) [1]. Παρόλ' αυτά μια τέτοια ανάλυση δεν είναι εφαρμόσιμη σε ένα όριο απόφασης σχηματισμένο σύμφωνα με τον κανόνα απόφασης εξόδου της εξίσωσης (7). Μια πιο κατάλληλη προσέγγιση είναι να θεωρήσουμε τους κρυμμένους νευρώνες ως μη γραμμικούς ανιχνευτές χαρακτηριστικών, οι οποίοι προσπαθούν να απεικονίσουν κλάσεις από το πρωτότυπο χώρο εισόδου R^P, όπου οι κλάσεις μπορεί να μην είναι γραμμικά διαχωρίσιμες, στο χώρο του κρυμμένου επιπέδου ενεργοποιήσεων, όπου είναι πιθανότερο για αυτές να είναι γραμμικά διαχωρίσιμες.(Yee, 1992) [1].

Άσκηση αυτοαξιολόγησης 4.4 / 4:

Με ποιά έννοια ένας κανόνας απόφασης θεωρείται βέλτιστος;

Απάντηση: Όταν ελαχιστοποιεί μια προκαθορισμένη συνάρτηση κόστους. Έτσι ελαχιστοποιείται η πιθανότητα λάθους ταξινόμησης.

Άσκηση αυτοαξιολόγησης 4.4 / 5:

Να αιτιολογήσετε γιατί ο κανόνας απόφασης που ορίζεται από τη σχέση (7) της ενότητας 4.4, είναι ο πιο κατάλληλος κανόνας απόφασης.

Απάντηση:

Για το πρόβλημα ταξινόμησης *l* εκ των m δειγμάτων, το k-στο στοιχείο του επιθυμητού διανύσματος απόκρισης d είναι ίσο με ένα εάν το διάνυσμα εισόδου ανήκει στην κλάση ck και μηδέν αλλιώς. Κάτω από αυτήν την συνθήκη , η υπό συνθήκη προσδοκία E[d|x] είναι ίση με την εκ των υστέρων πιθανότητα κλάσης P(ck | x), k=1,2, ... ,m (Richard

and Lippmann, 1991) [1].

4.5 Προσομοίωση σε ηλεκτρονικό υπολογιστή

Σ' αυτό το κεφάλαιο χρησιμοποιούμε ένα πείραμα με υπολογιστή για να δείξουμε τη συμπεριφορά εκμάθησης ενός πολυεπίπεδου perceptron νευρωνικού δικτύου χρησιμοποιούμενου ως ταξινομητή δειγμάτων. Το αντικείμενο του πειράματος είναι ο διαχωρισμός μεταξύ δύο κλάσεων από αλληλοεπικαλυπτόμενα, δύο διαστάσεων, Gaussian κατανεμημένων δειγμάτων, που ονομάσαμε 1 και 2. Έστω $\mathbb{C}1$ και $\mathbb{C}2$ τα σύνολα των γεγονότων για τα οποία ένα τυχαίο διάνυσμα x ανήκει στα δείγματα 1 και 2, αντίστοιχα. Μπορούμε τότε να εκφράσουμε τις υπό συνθήκη συναρτήσεις πυκνότητας πιθανότητας για τις δύο κλάσεις ως εξής :

Κλάση $\mathbb{C}1$: $$f(x|\mathbb{C}1)= \frac{1}{2\pi\sigma_1^2}\exp\left(-\frac{1}{2\sigma_1^2}\|x-\mu_1\|^2\right) \quad (8)$$

όπου:

μ_1=μέσο διάνυσμα=$[0,0]^T$

σ_1^2 =διασπορά=1

Κλάση $\mathbb{C}2$: $$f(x|\mathbb{C}2)= \frac{1}{2\pi\sigma_2^2}\exp\left(-\frac{1}{2\sigma_2^2}\|x-\mu_2\|^2\right) \quad (9)$$

όπου:

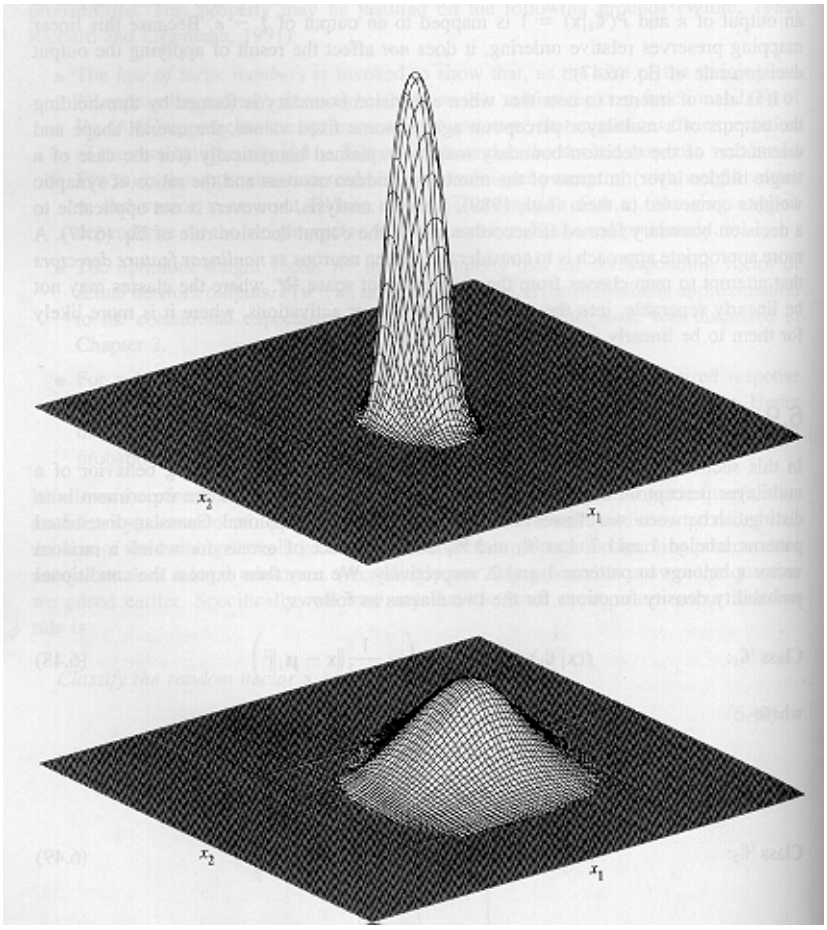
$\mu_2=[2,0]^T$

σ_2 =4

Οι δύο κλάσεις θεωρούνται ισοπίθανες , δηλαδή:

$P(\mathbb{C}1)=P(\mathbb{C}2)=0.5$

Το σχήμα 5 δείχνει τρισδιάστατα γραφήματα των δύο Gaussian κατανομών που ορίστηκαν από τις εξισώσεις (8) και (9). Μια υπέρθεεση των διαγραμμάτων δείχνει καθαρά ότι οι δύο κατανομές επικαλύπτουν η μία την άλλη σημαντικά .



Σχήμα 5: Πάνω: Συνάρτηση πυκνότητας πιθανότητας f(**x**|C1)

Κάτω: Συνάρτηση πυκνότητας πιθανότητας f(**x**|C_2)

4.5.1 Όριο απόφασης του Bayes

Χρησιμοποιώντας το κριτήριο του Bayes, ένα βέλτιστο όριο απόφασης για το όριο των δύο κλάσεων μπορεί να βρεθεί εφαρμόζοντας το τεστ του λόγου πιθανοτήτων :

$$\Lambda(x) \underset{C_2}{\overset{C_1}{\lessgtr}} \lambda \qquad (10)$$

όπου Λ(x) είναι ο λόγος πιθανοτήτων, ορισμένος ως :

$$\Lambda(x) = \frac{f(x|C_2)}{f(x|C_1)} \qquad (11)$$

και λ είναι το κατώφλι του τεστ, που ορίζεται ως :

$$\lambda = \frac{P(c1)}{P(c2)} \qquad (12)$$

Για το συγκεκριμένο παράδειγμά μας ,έχουμε :

$$\Lambda(x)=\frac{\sigma_1^2}{\sigma_2^2}\exp\left(-\frac{1}{2\sigma_2^2}\|x-\mu_2\|^2+\frac{1}{2\sigma_1^2}\|x-\mu_1\|^2\right)$$

και:

$$\lambda=1$$

Το βέλτιστο (Bayesian) όριο απόφασης ορίζεται επομένως από την:

$$\frac{\sigma_1^2}{\sigma_2^2}\exp\left(-\frac{1}{2\sigma_2^2}\|x-\mu_2\|^2+\frac{1}{2\sigma_1^2}\|x-\mu_1\|^2\right)=1$$

ή ισοδύναμα:

$$\frac{1}{\sigma_1^2}\|x-\mu_1\|^2+\frac{1}{\sigma_2^2}\|x-\mu_2\|^2=4\ln\left(\frac{\sigma_2}{\sigma_2}\right) \qquad (13)$$

Χρησιμοποιώντας απλούς χειρισμούς , μπορούμε να επαναορίσουμε το βέλτιστο όριο απόφασης της εξίσωση (13) ως:

$$||x-x_c||^2=r^2 \qquad (14)$$

όπου:

$$x_c=\frac{\sigma_2^2\mu_1-\sigma_1^2\mu_2}{\sigma_2-\sigma_1} \qquad (15)$$

και:

$$r^2=\frac{\sigma_1^2\sigma_2^2}{\sigma_2^2-\sigma_1^2}\left[\frac{\|\mu_1+\mu_2\|^2}{\sigma_2^2-\sigma_1^2}+4\ln\left(\frac{\sigma_2}{\sigma_1}\right)\right] \qquad (16)$$

Η εξίσωση (14) παριστάνει κύκλο με κέντρο x_c και ακτίνα r. Έστω ότι με Ω_1 συμβολίζουμε την περιοχή εντός του κύκλου. Τότε, ο κανόνας της Bayesian ταξινόμησης για το δεδομένο πρόβλημα μπορεί να διατυπωθεί ως ακολούθως :

Κατέταξε το παρακολουθούμενο διάνυσμα x σαν να ανήκει στην κλάση c1 *αν* $x \in \Omega_1$ *και στην κλάση* c2 *διαφορετικά* .

Για τις ιδιαίτερες παραμέτρους αυτού του πειράματος, έχουμε ένα κυκλικό όριο απόφασης του οποίου το κέντρο βρίσκεται στο:

$$x_c=[-2/3,0]$$

και η ακτίνα του είναι:

$$\rho\cong 2.34$$

Έστω c το σύνολο των σωστών ταξινομήσεων, και e το σύνολο των λαθεμένων ταξινομήσεων . Η μέση πιθανότητα λάθους, P_e ,ενός ταξινομητή που λειτουργεί σύμφωνα με τον κανόνα απόφασης του Bayes,είναι:

$$P_e=P(e|C1)P(C1)+P(e|C2)P(C2) \qquad (17)$$

όπου P(e|C1) είναι η υπό συνθήκη πιθανότητα λάθους δεδομένου ότι τα δεδομένα εισόδου του ταξινομητή πάρθηκαν από την κατανομή της κλάσης C1, και παρόμοια για το P(e|C2). Για το συγκεκριμένο πρόβλημα , βρίσκουμε ότι:

$$P(e|C1) \cong 0.1056$$

και:

$$P(e|C2) \cong 0.2642$$

Η μέση πιθανότητα λάθους είναι επομένως:

$$P_e \cong 0.1849$$

Ισοδύναμα , η μέση πιθανότητα σωστής ταξινόμησης είναι:

$$P_c=1-P_e \cong 0.8151$$

4.6 Πειραματικός καθορισμός του βέλτιστου πολυεπίπεδου Perceptron

Ο πίνακας 1 περιέχει τη λίστα των μεταβλητών παραμέτρων ενός MLP το οποίο περιέχει ένα μόνο επίπεδο κρυμμένων νευρώνων και εκπαιδεύεται με τον αλγόριθμο της πίσω διάδοσης . Καθώς ο βασικός στόχος ενός ταξινομητή δειγμάτων είναι να πετύχει έναν αποδεκτό βαθμό ταξινόμησης, αυτό το κριτήριο χρησιμοποιείται για να κρίνει πότε οι μεταβλητές παράμετροι του MLP (το οποίο χρησιμοποιείται ως ταξινομητής προτύπων) είναι βέλτιστες.

Άσκηση αυτοαξιολόγησης 4.5 / 6:

Πως καθορίζεται το βέλτιστο (Bayesian) όριο απόφασης για δύο κλάσεις, όταν οι υπό συνθήκη συναρτήσεις πυκνότητας πιθανότητας ακολουθούν κανονική κατανομή.

Απάντηση:Εφαρμόζουμε το τεστ του λόγου πιθανοτήτων και καταλήγουμε στη σχέση (13).

Άσκηση αυτοαξιολόγησης 4.5 / 7:

Αν P(c1)=1/3 και P(c2)=2/3, να υπολογίσετε την μέση πιθανότητα λάθους, για τον παραπάνω ταξινομητή Bayes.

Απάντηση:

Υπολογίζεται εύκολα από τη σχέση (17).

4.6.1 Βέλτιστος αριθμός κρυφών νευρώνων

Αντανακλώντας πρακτικές προσεγγίσεις του προβλήματος καθορισμού του βέλτιστου αριθμού κρυμμένων νευρώνων M , το κριτήριο είναι ο μικρότερος αριθμός κρυμμένων νευρώνων που αποφέρουν απόδοση κοντά σ' αυτή του βέλτιστου Bayesian ταξινομητή - ας πούμε περίπου 1 %. Έτσι , η πειραματική μελέτη ξεκινά με δύο κρυμμένους νευρώνες ως σημείο εκκίνησης, για τα αποτελέσματα της εξομοίωσης που συγκεντρώνονται στον πίνακα 2. Καθώς ο σκοπός της πρώτης ομάδας των εξομοιώσεων είναι απλώς να εξακριβώσει την επάρκεια ή όχι των δύο νευρώνων, ο ρυθμός εκμάθησης η και η σταθερά ορμής α τίθενται αυθαίρετα σε κάποιες ονομαστικές τιμές. Για κάθε εκτέλεση εξομοίωσης, ένα σύνολο εκπαίδευσης από ζεύγη εισόδου-εξόδου, τυχαία παραγόμενων από τις Gaussian κατανομές για τις κλάσεις $C1$ και $C2$ με ίση πιθανότητα, επαναλαμβανόμενα ανακυκλώνονται διαμέσου του δικτύου, με κάθε κύκλο εκπαίδευσης να αναπαριστά ένα κύκλο (epoch) . Ο αριθμός των κύκλων επιλέχθηκε έτσι ώστε ο συνολικός αριθμός των δειγμάτων εκπαίδευσης που χρησιμοποιήθηκαν για κάθε εκτέλεση να είναι σταθερός. Μ' αυτό τον τρόπο όλα τα δυνατά αποτελέσματα που προκύπτουν από τις μεταβολές του μεγέθους των συνόλων εκπαίδευσης εξισώνονται .

ΠΙΝΑΚΑΣ 1 Μεταβλητές Παράμετροι του Πολυεπίπεδου Perceptron

Παράμετρος	Σύμβολο	Τυπικό Εύρος
Αριθμός κρυμμένων νευρώνων	**M**	**(2,∞)**
Παράμετρος ρυθμού εκμάθησης	**η**	**(0,1)**
Σταθερά ορμής	**α**	**(0,1)**

Στον πίνακα 2 και στους επόμενους πίνακες, το μέσο τετραγωνικό λάθος υπολογίζεται ακριβώς όπως στην εξίσωση (5) . Πρέπει να τονιστεί ότι το μέσο τετραγωνικό λάθος περιλήφθηκε σ' αυτούς τους πίνακες μόνο για λόγους καταγραφής, καθώς *ένα μικρό μέσο τετραγωνικό λάθος δε συνεπάγεται απαραίτητα και καλή γενίκευση* (δηλαδή, καλή απόδοση με δεδομένα που δεν εισήχθησαν παλαιότερα).

Μετά τη σύγκλιση ενός δικτύου που εκπαιδεύτηκε με έναν συνολικό αριθμό από N δείγματα , η πιθανότητα σωστής ταξινόμησης μπορεί , θεωρητικά , να υπολογιστεί ως ακολούθως :

$$P(c; N)=P(c; N|C1)P(C1)+P(c; N|C2)P(C2) \quad (18)$$

όπου:

$$P(c; N|C1)= \int_{\Omega 1(N)} f(x|C1)dx \quad (19)$$

$$P(c; N|C2) = 1- \int_{\Omega 1(N)} f(x|C2)dx \quad (20)$$

και $\Omega_1(N)$ είναι η περιοχή του πεδίου απόφασης στην οποία το MLP (που εκπαιδεύτηκε με N δείγματα) ταξινομεί το διάνυσμα x τοποθετώντας το στην κλάση C1. Συνήθως, η περιοχή αυτή βρίσκεται πειραματικά με υπολογισμό της συνάρτησης αντιστοίχησης που μαθεύτηκε από το δίκτυο και εφαρμόζοντας τον κανόνα απόφασης εξόδου της εξίσωσης (7). Δυστυχώς, ο αριθμητικός υπολογισμός των P(c; N|C1) και P(c; N|C2) είναι προβληματικός, επειδή κλειστές μορφές εκφράσεων που να περιγράφουν το όριο απόφασης $\Omega_1(N)$ είναι δύσκολο να βρεθούν .

Συνεπώς, καταφεύγουμε στη λύση μιας πειραματικής προσέγγισης η οποία περιλαμβάνει έλεγχο του εκπαιδευμένου MLP χρησιμοποιώντας άλλα ανεξάρτητα σύνολα από ζεύγη εισόδου-εξόδου, τα οποία επίσης παίρνονται τυχαία από τις κατανομές των κλάσεων C1 και C2 με ίση πιθανότητα . Έστω A μια τυχαία μεταβλητή η οποία μετράει τον αριθμό των δειγμάτων από τα N δείγματα δοκιμής τα οποία ταξινομούνται σωστά . Τότε ο λόγος:

$$P_N = \frac{A}{N} \qquad (21)$$

είναι μια τυχαία μεταβλητή η οποία δίνει την μεγίστης-πιθανότητας (maximum-likelihood) αμερόληπτη εκτίμηση της πραγματικής απόδοσης ταξινόμησης p του δικτύου. Θεωρώντας ότι το p είναι σταθερό πάνω στα N ζεύγη εισόδου-εξόδου, μπορούμε να εφαρμόσουμε το όριο του Chernoff (Devroye, 1991) [1], στον εκτιμητή p_N του p ,παίρνοντας:

$$\text{Prob}(|p_N - p| > \varepsilon) < 2\exp(-2\varepsilon^2 N) = \delta \qquad (22)$$

Η εφαρμογή του ορίου Chernoff δίνει N≅26500 για ε= 0,01 και δ=0,01 (δηλ. , 99 τις εκατό σιγουριά ότι η εκτίμηση p έχει την δεδομένη ανοχή). Έτσι , πήραμε ένα σύνολο δοκιμής μεγέθους N=32000. Η τελευταία στήλη του πίνακα 2 παρουσιάζει τη μέση πιθανότητα σωστής ταξινόμησης που εκτιμήθηκε γι' αυτό το μέγεθος του συνόλου δοκιμής.

ΠΙΝΑΚΑΣ 2 Αποτελέσματα Εξομοίωσης για 2 Κρυφούς Νεύρωνες (η=0,1 & α=0)

Αριθμός Εκτελέσεων	Μέγεθος Συνόλου Εκπαίδευσης	Αριθμός Κύκλων	Μέσο Τετραγ. Λάθος	Πιθανότητα Σωστής ταξινόμησης, P_c
1	**500**	**320**	**0,2331**	**79,84%**
2	**2000**	**80**	**0,2328**	**80,34%**
3	**8000**	**20**	**0,2272**	**80,23%**

Η μέση απόδοση ταξινόμησης που δείχνεται στον πίνακα 5.2 για ένα MLP με δύο κρυμμένους νευρώνες είναι ήδη κοντά σε λογικό βαθμό στην Bayesian απόδοση P_c=81,51 τις εκατό. Πάνω σ' αυτή τη βάση, μπορούμε να καταλήξουμε στο ότι το πρόβλημα ταξινόμησης δειγμάτων που περιγράφηκε εδώ, η χρήση δύο κρυμμένων νευρώνων είναι επαρκής. Για να τονίσουμε αυτό το αποτέλεσμα, στον πίνακα 3 παρουσιάζουμε τα αποτελέσματα εξομοιώσεων που επαναλήφθηκαν για την περίπτωση τεσσάρων κρυμμένων νευρώνων, με όλες τις άλλες παραμέτρους να

παραμένουν οι ίδιες. Παρ' όλο που κατά μέσο όρο το μέσο τετραγωνικό λάθος στον πίνακα 3 για τους τέσσερις κρυμμένους νευρώνες είναι ελαφρώς μικρότερο από αυτό του πίνακα 2 για τους δύο κρυμμένους νευρώνες, ο μέσος βαθμός σωστού διαχωρισμού δε δείχνει σημαντική βελτίωση. Συνεπώς, για το υπόλοιπο του πειράματος που περιγράφεται εδώ, ο αριθμός των κρυμμένων νευρώνων κρατιέται στους δύο.

ΠΙΝΑΚΑΣ 3 Αποτελέσματα Εξομοίωσης για 4 Κρυμμένους Νεύρωνες(η=0,1& α=0)

Αριθμός Εκτελέσεων	Μέγεθος Συνόλου Εκπαίδευσης	Αριθμός Κύκλων	Μέσο Τετραγ. Λάθος	Πιθανότητα Σωστής ταξινόμησης, P_c
1	500	320	0,2175	80,43%
2	2000	80	0,2175	80,45%
3	8000	20	0,2195	80,99%

4.6.2 Βέλτιστες σταθερές μάθησης και ορμής

Για τις βέλτιστες τιμές του ρυθμού μάθησης ***η*** και της σταθεράς ορμής ***α*** μπορούμε να χρησιμοποιήσουμε οποιονδήποτε από τους παρακάτω ορισμούς :

1. Τα ***η*** και ***α*** τα οποία, στη μέση περίπτωση, επιφέρουν σύγκλιση σ' ένα τοπικό ελάχιστο της επιφάνειας λάθους του δικτύου με τον μικρότερο αριθμό κύκλων.

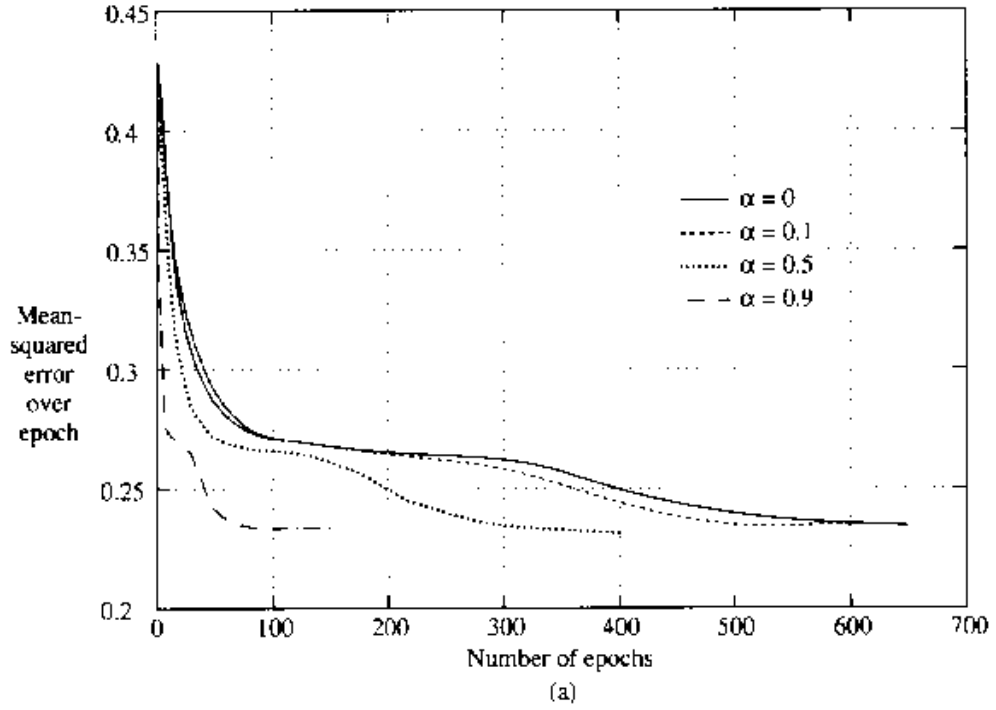
2. Τα ***η*** και ***α*** τα οποία, στη μέση ή την χειρότερη περίπτωση, επιφέρουν σύγκλιση στο ολικό ελάχιστο της επιφάνειας λάθους με το μικρότερο αριθμό κύκλων.

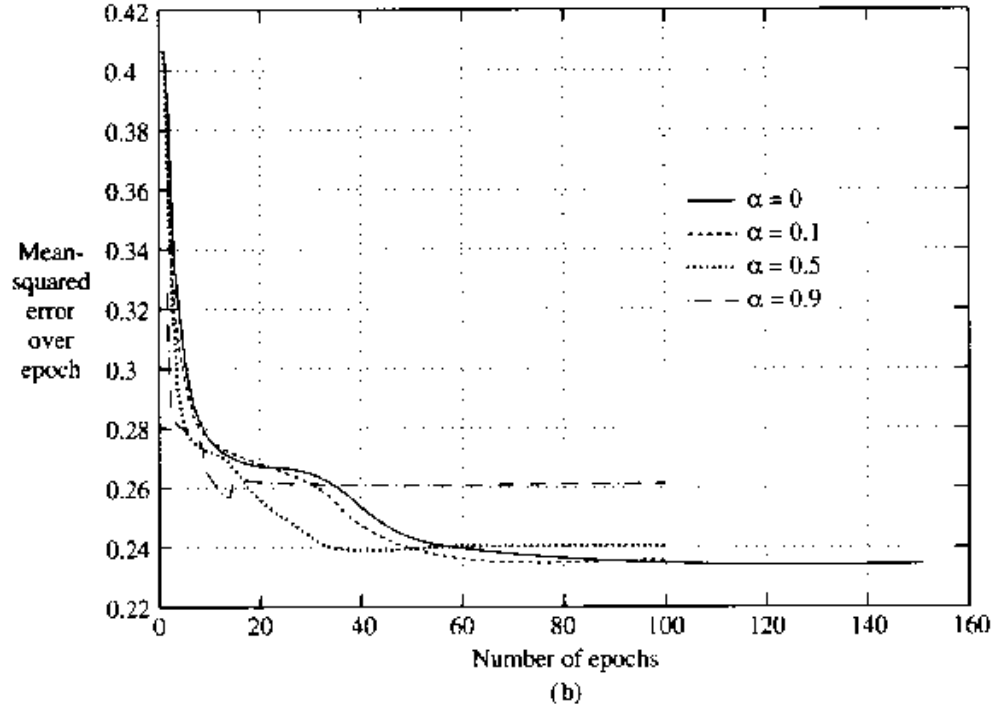
3. Τα ***η*** και ***α*** τα οποία, στη μέση περίπτωση, επιφέρουν σύγκλιση σ' εκείνη τη διαμόρφωση του δικτύου η οποία έχει την καλύτερη γενίκευση, πάνω στο συνολικό πεδίο εισόδων, με το μικρότερο αριθμό κύκλων.

Η μέση και χειρότερη περίπτωση αναφέρονται στην κατανομή των ζευγών εισόδου-

εξόδου της εκπαίδευσης. Ο ορισμός 3 είναι ο ιδανικός. Στην πράξη όμως είναι δύσκολο να εφαρμοστεί καθώς η ελαχιστοποίηση του μέσου τετραγωνικού λάθους είναι συνήθως το μαθηματικό κριτήριο για βελτιστοποίηση κατά τη διάρκεια της εκπαίδευσης του δικτύου και, όπως σημειώθηκε πριν από λίγο, ένα μικρότερο τετραγωνικό λάθος πάνω στο σύνολο εκπαίδευσης δε συνεπάγεται απαραίτητα και καλή γενίκευση. Από ερευνητικής απόψεως, ο ορισμός 2 έχει προκαλέσει μεγαλύτερο ενδιαφέρον από τον ορισμό 1. Για παράδειγμα, στο Luo (1991) [1], παρουσιάζονται αυστηρά αποτελέσματα για τη βέλτιστη προσαρμογή του ρυθμού μάθησης $\boldsymbol{\eta}$ έτσι ώστε να απαιτείται ο μικρότερος αριθμός κύκλων για το MLP να προσεγγίσει τον συνολικά βέλτιστο πίνακα συναπτικών βαρών σε μια επιθυμητή ακρίβεια, εκτός από την ειδική περίπτωση των γραμμικών νευρώνων. Γενικά, ωστόσο, ευριστικές και πειραματικές διαδικασίες χρησιμοποιούν τελευταία τον ορισμό 1 για τη βέλτιστη επιλογή των $\boldsymbol{\eta}$ και $\boldsymbol{\alpha}$. Εμείς για το παρόν πείραμα θα χρησιμοποιήσουμε τον ορισμό 1.

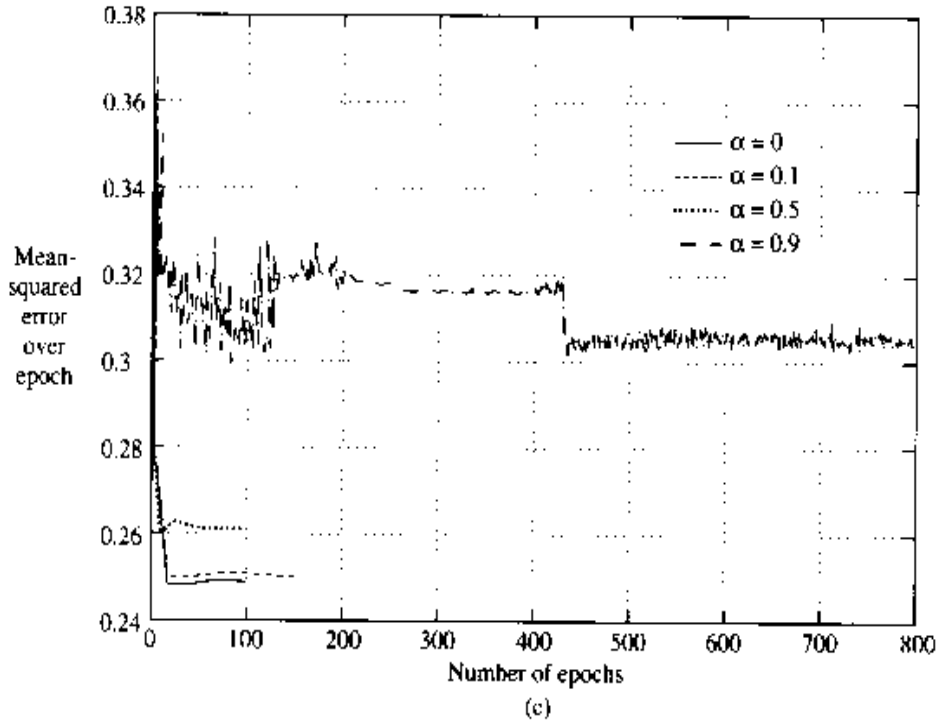
Χρησιμοποιώντας ένα MLP με δύο κρυμμένους νευρώνες , διάφοροι συνδυασμοί των $\boldsymbol{\eta} \in \{0,01\ ,\ 0,1\ ,\ 0,5\ ,\ 0,9\}$ και $\boldsymbol{\alpha} \in \{0,0\ ,\ 0,1\ ,\ 0,5\ ,\ 0,9\}$ εξομοιώνονται για να παρατηρήσουμε τα αποτελέσματά τους στη σύγκλιση του δικτύου. Κάθε συνδυασμός εκπαιδεύεται με το ίδιο σύνολο αρχικών τυχαίων βαρών και το ίδιο σύνολο από 500 δείγματα εισόδου-εξόδου, έτσι ώστε τα αποτελέσματα του πειράματος να μπορούν να συγκριθούν άμεσα. Όπως ειπώθηκε προηγουμένως, ένα δίκτυο θεωρείται ότι έχει συγκλίνει όταν ο απόλυτος ρυθμός αλλαγής του μέσου τετραγωνικού λάθους ανά κύκλο είναι επαρκώς μικρός . Σ' αυτό το πείραμα επιλέξαμε ο ρυθμός αυτός να είναι

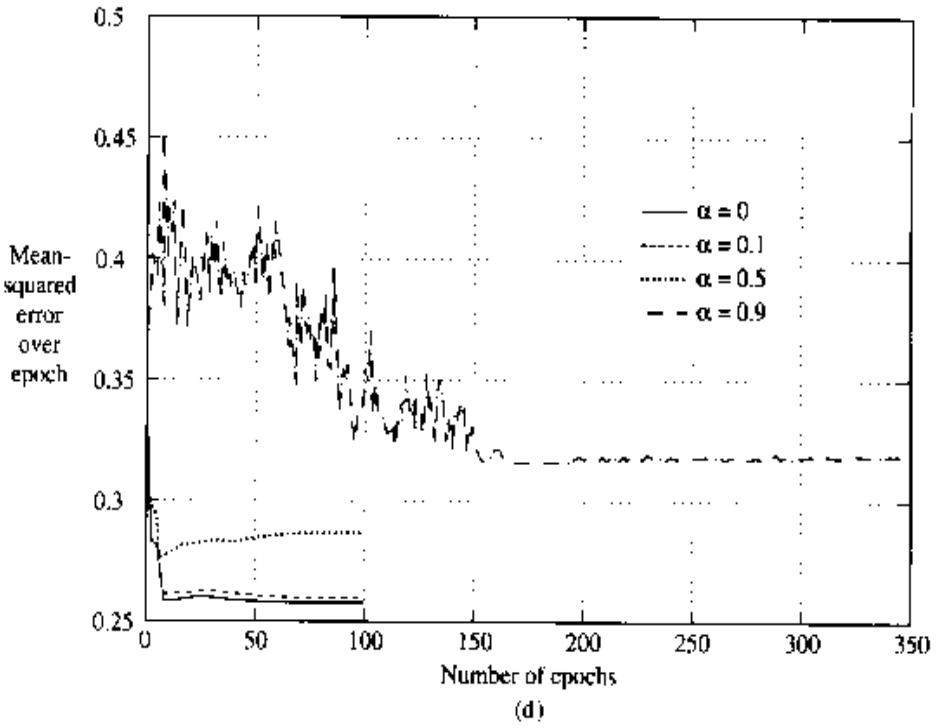




Σχήμα 6: Καμπύλες μάθησης για διάφορες τιμές της σταθεράς ορμής α και τις ακόλουθες τιμές της παραμέτρου ρυθμού μάθησης (α) η=0.01 ; (b) η=0.1.

είναι μικρότερος από 0,01 τις εκατό. Οι καμπύλες μάθησης που υπολογίστηκαν μ' αυτό τον τρόπο σχεδιάστηκαν στα σχήματα 6.α ,b και 7.c, δ [1], ένα σχήμα για κάθε διαφορετικό $\boldsymbol{\eta}$.





Σχήμα 7: (συνέχεια) : (c) η=0.5 ; (d) η=0.9

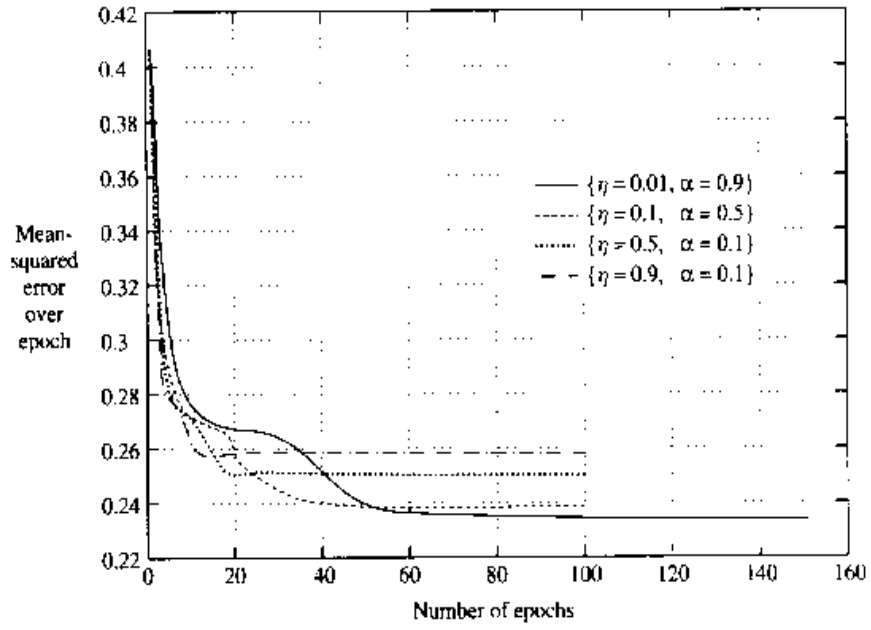
Οι παραπάνω πειραματικές καμπύλες μάθησης οδηγούν στα παρακάτω συμπεράσματα:

- Καθώς, γενικά, μικρότερος ρυθμός μάθησης ***η*** έχει ως αποτέλεσμα πιο αργή σύγκλιση, μπορεί να εντοπίσει ″βαθύτερο″ τοπικό ελάχιστο στην επιφάνεια λάθους

απ' ότι ένα μεγαλύτερο η. Αυτό είναι και διαισθητικά σωστό, καθώς μικρότερο η συνεπάγεται ότι η αναζήτηση για ελάχιστο θα καλύψει μεγαλύτερο μέρος της επιφάνειας λάθους απ' ότι στην περίπτωση ενός μεγαλύτερου η.

- Για $\eta \rightarrow 0$, η χρήση $\alpha \rightarrow 1$ προκαλεί αύξηση στην ταχύτητα σύγκλισης. Από την άλλη, για $\eta \rightarrow 1$, η χρήση $\alpha \rightarrow 0$ απαιτείται για να σιγουρέψει τη σταθερότητα μάθησης.

- Η χρήση των σταθερών η = { 0,5 , 0,9 } και α = 0,9 προκαλεί ταλαντώσεις στο μέσο τετραγωνικό λάθος κατά την διάρκεια μάθησης και μια υψηλότερη τιμή για το τελικό μέσο τετραγωνικό λάθος της σύγκλισης, δύο αποτελέσματα τα οποία είναι ανεπιθύμητα .



Σχήμα 8: Οι καλύτερες καμπύλες μάθησης που επιλέχθηκαν από το σχήμα 7.

Στο σχήμα 8 βλέπουμε τις γραφικές παραστάσεις των καλύτερων καμπυλών μάθησης από κάθε ομάδα των καμπυλών μάθησης που αναπαρίστανται γραφικά στο σχήμα 7, έτσι ώστε να καθορίσουμε μία ''συνολικά'' καλύτερη καμπύλη μάθησης. Από το σχήμα 8 φαίνεται ότι η βέλτιστη παράμετρος ρυθμού μάθησης η_{opt} είναι περίπου 0.1 και η βέλτιστη σταθερά ορμής α_{opt} είναι περίπου 0.5. Έτσι ο πίνακας 4 συνοψίζει τις

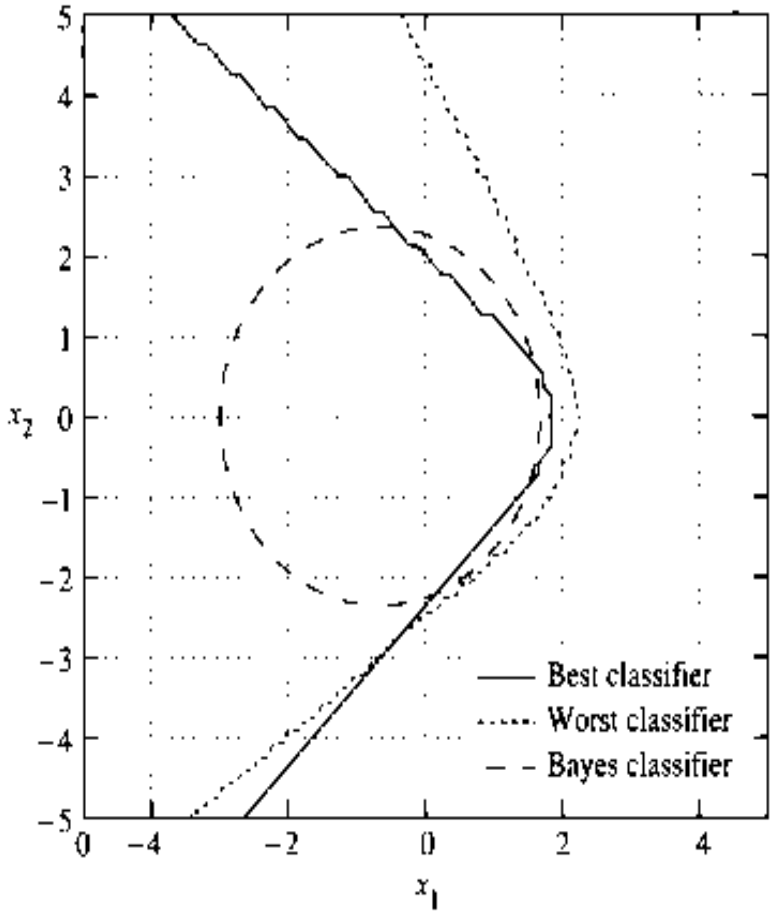
‘’βέλτιστες’’ τιμές των παραμέτρων του δικτύου που χρησιμοποιούνται στο υπόλοιπο μέρος του πειράματος. Το γεγονός ότι τα τελικά μέσα τετραγωνικά σφάλματα των καμπυλών στο σχήματος 8 δεν διαφέρουν σημαντικά για τα διάφορα $\boldsymbol{\eta}$ και $\boldsymbol{\alpha}$ υποδηλώνει μία ‘’καλά-συμπεριφερόμενη’’ (σχετικά λεία) επιφάνεια λάθους για το ίδιο το πρόβλημα.

4.6.3 Εκτίμηση του βέλτιστου σχεδιασμού του δικτύου

Δοσμένου του ‘’βέλτιστου’’ πολυεπίπεδου perceptron που έχει τις παραμέτρους που συνοψίζονται στον πίνακα 4 το δίκτυο τελικά αξιολογείται για να καθορίσει το όριο απόφασης του, την τελική μέση καμπύλη μάθησης του και την μέση πιθανότητα σωστής ταξινόμησης. Με πεπερασμένου μεγέθους εκπαιδευτικά σύνολα η συνάρτηση που το δίκτυο έμαθε με τις βέλτιστες παραμέτρους είναι ‘’στοχαστική’’ από την φύση της. Κατά αντιστοιχία αυτά τα μέτρα απόδοσης είναι ένας μέσος συνόλου για πάνω από 20 ανεξάρτητα εκπαιδευμένα δίκτυα. Κάθε εκπαιδευτικό σύνολο αποτελείται από 1000 ζεύγη εισόδου-εξόδου, παρμένα από τις κατανομές για τις κλάσεις $\mathcal{C}1$ και $\mathcal{C}2$ με ίση πιθανότητα, και τα οποία παρουσιάζονται στο δίκτυο με τυχαία σειρά. Όπως προηγουμένως, το κριτήριο που χρησιμοποιείται για την σύγκλιση του δικτύου είναι ο απόλυτος ρυθμός αλλαγής του μέσου τετραγωνικού σφάλματος να είναι μικρότερος από 0.01% ανά κύκλο. Για τον πειραματικό καθορισμό των μέσων πιθανοτήτων σωστής ταξινόμησης, το ίδιο τεστ από 32,000 ζεύγη εισόδου-εξόδου που χρησιμοποιήθηκε προηγουμένως, χρησιμοποιείται μία φορά ακόμη.

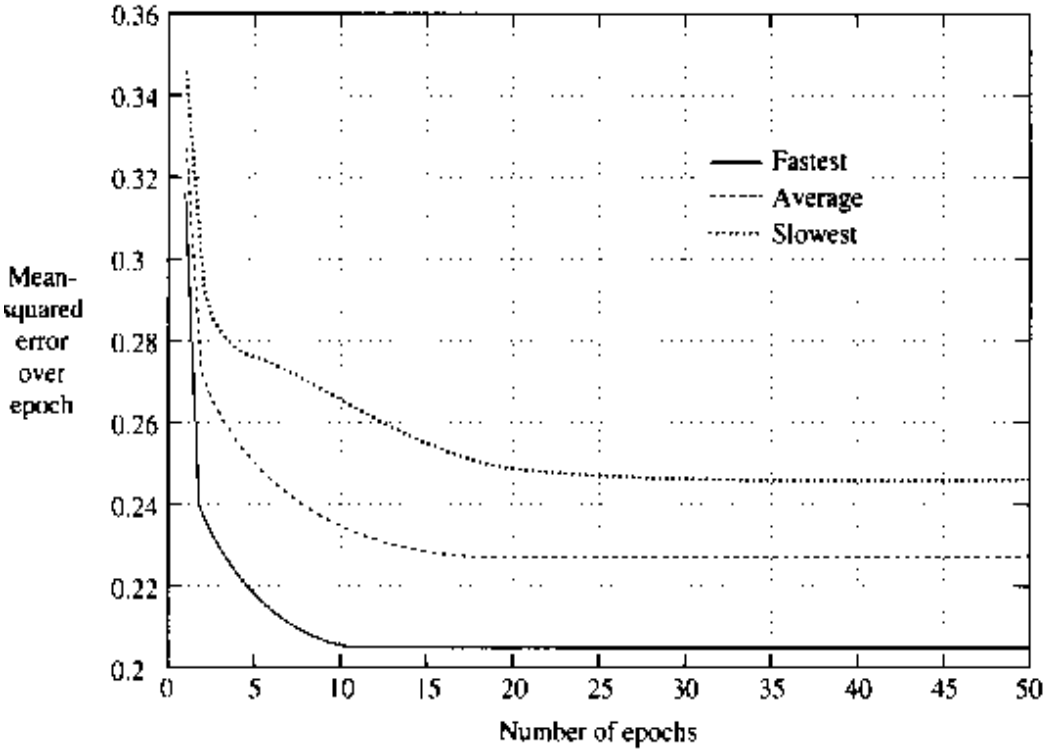
Πίνακας 4 Διαμόρφωση του βέλτιστου πολυεπίπεδου Perceptron

Παράμετρος	Σύμβολο	Τιμή
Βέλτιστος αριθμός κρυμμένων νευρώνων	M_{opt}	2
Βέλτιστη παράμετρος ρυθμού μάθησης	η_{opt}	0.1
Βέλτιστη σταθερά ορμής	α_{opt}	0.5

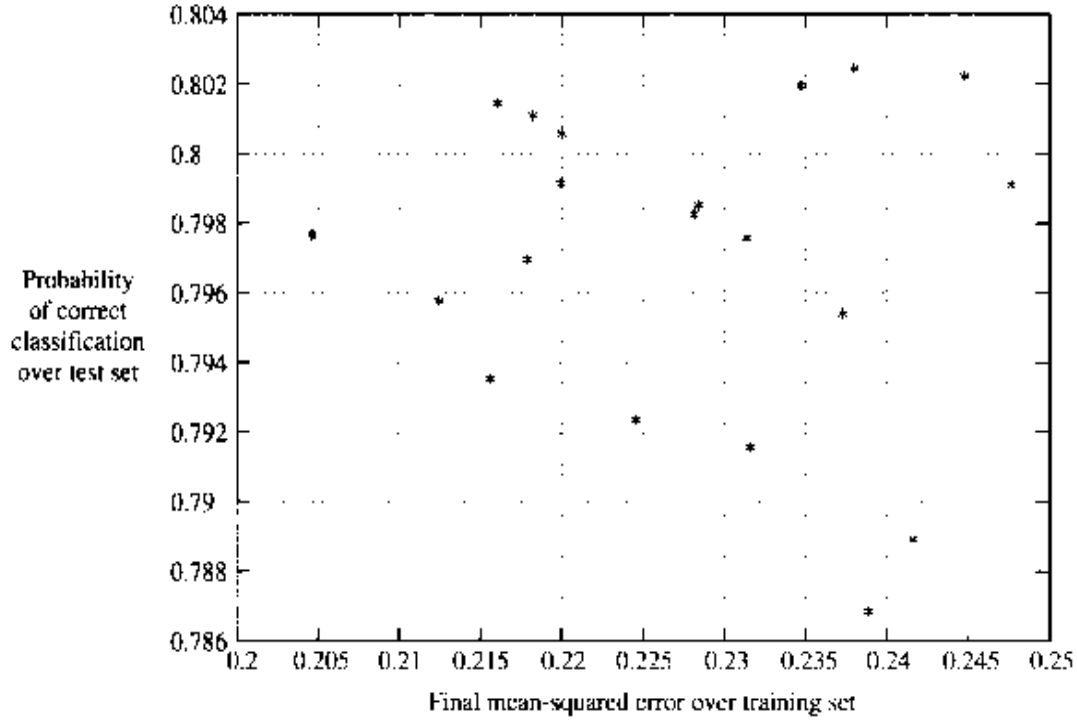


Σχήμα 9: Όρια απόφασης που κατασκευάστηκαν από το πολυεπίπεδο perceptron

Το σχήμα 9 δείχνει τα όρια απόφασης για τα δύο δίκτυα από τα είκοσι συνολικά, με την χειρότερη και καλύτερη απόδοση ταξινόμησης. Αυτή η εικόνα επίσης περιέχει, για αναφορά, το (κυκλικό) Bayesian όριο απόφασης. Και τα δύο όρια απόφασης (για την χειρότερη και καλύτερη απόδοση ταξινόμησης) είναι κυρτά με αναφορά στην περιοχή όπου ταξινομούν το διάνυσμα x όπως αυτό ανήκει στην κλάση $C1$ ή την κλάση $C2$. Κάθε όριο εμφανίζεται να περιέχει δύο γραμμικά τμήματα μ' ένα μη- γραμμικό τμήμα να τα συνδέει κοντά στην δική τους προβαλλόμενη τομή.



Σχήμα 10: Καμπύλες μάθησης - αργότερες και ταχύτερες αποδόσεις για το βελτιστοποιημένο δίκτυο



Σχήμα 11: Μέση σωστή πιθανότητα ταξινόμησης έναντι μέσου τετραγωνικού λάθους

Το σχήμα 10 δείχνει τις τελικές μέσες, τις αργότερες και ταχύτερες καμπύλες μάθησης που παρατηρήθηκαν κατά την διάρκεια της εκπαίδευσης του ‘’βελτιστοποιημένου’’ δικτύου. Παρόλο το ανώτερο τελικό μέσο τετραγωνικό σφάλμα του δικτύου με την ταχύτερη καμπύλη μάθησης (0.2047 έναντι 0.2477 για το δίκτυο με την αργότερη καμπύλη μάθησης), αυτό το προφανές πλεονέκτημα δεν παρέχει κανένα κέρδος στην απόδοση της ταξινόμησης. Στην πραγματικότητα, το δίκτυο που μαθαίνει ταχύτερα έχει μία οριακά χειρότερη απόδοση ταξινόμησης απ’ ότι το δίκτυο που μαθαίνει αργότερα (79.78 έναντι 79.90%). Μία παρόμοια κατάσταση συμβαίνει όταν συνολικά το μέσο τετραγωνικό σφάλμα του δικτύου με την καλύτερη απόδοση ταξινόμησης συγκρίνετε μ’ αυτό του δικτύου με την χειρότερη απόδοση ταξινόμησης. Το πρώτο έχει P_C της τάξης του 80.23% μ’ ένα τελικό μέσο τετραγωνικό σφάλμα της τάξης του 0.2380, ενώ το άλλο έχει ένα P_C της τάξης του 78.68% μ’ ένα τελικό μέσο τετραγωνικό σφάλμα της τάξης του 0.2391, όχι σημαντικά διαφορετικά από το πρώτο. Αυτά τα αποτελέσματα ενισχύουν την αντίληψη ότι ένα χαμηλότερο μέσο τετραγωνικό σφάλμα κατά μήκος του εκπαιδευτικού συνόλου δεν είναι από μόνο του μία ικανή συνθήκη για μία καλύτερη απόδοση ταξινόμησης.

Πράγματι , η γραφική παράσταση της πειραματικά καθοριζόμενης πιθανότητας της σωστής ταξινόμησης έναντι του τελικού μέσου τετραγωνικού σφάλματος που φαίνεται στο σχήμα 11 παρουσιάζει μόνο έναν αδύνατο αρνητικό συσχετισμό -0.1220 ανάμεσα στα δύο μέτρα απόδοσης του δικτύου. Οι συνολικές στατιστικές των μέτρων απόδοσης, μέση πιθανότητα σωστής ταξινόμησης και τελικό μέσο τετραγωνικό σφάλμα, υπολογιζόμενα για το εκπαιδευτικό σύνολο φαίνονται στον πίνακα 5.

Πίνακας 5: Στατιστικές συνόλου του μέτρου επίδοσης (μήκος συνόλου = 20)

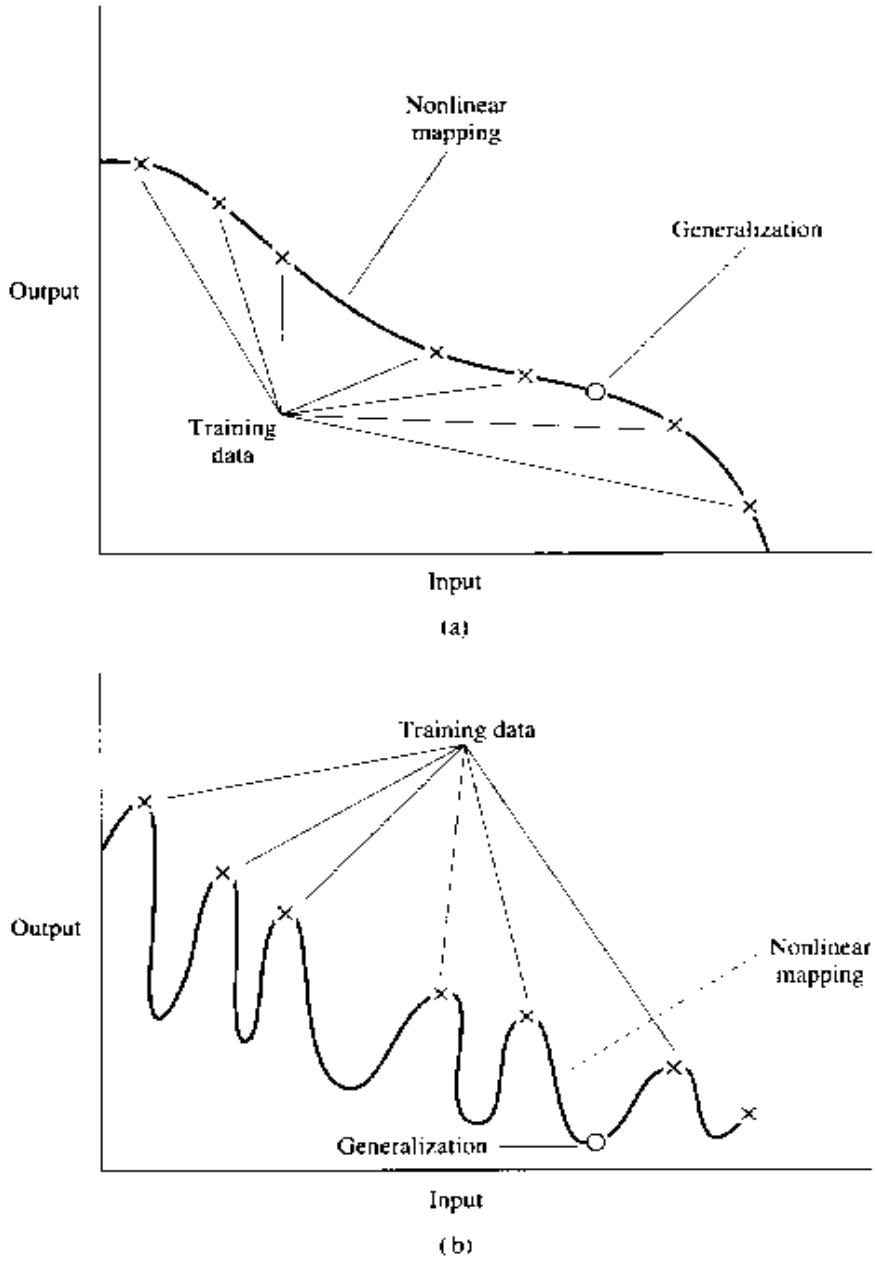
Μέτρο επίδοσης	Μέση τιμή	Τυπική απόκλιση
Μέση πιθανότητα σωστής ταξινόμησης	79.70%	0.44%
Τελικό μέσο τετραγωνικό λάθος	0.2277	0.0118

4.6.4 Γενίκευση

Στην μάθηση πίσω-διάδοσης ξεκινάμε μ’ ένα εκπαιδευτικό σύνολο και

χρησιμοποιούμε τον αλγόριθμο πίσω-διάδοσης για να υπολογίσουμε τα συναπτικά βάρη ενός πολυεπίπεδου perceptron φορτώνοντας (κωδικοποιώντας) όσα περισσότερα από τα εκπαιδευτικά παραδείγματα είναι δυνατόν μέσα στο δίκτυο. Η επιθυμία είναι ότι το νευρωνικό δίκτυο έτσι σχεδιασμένο θα γενικεύει. Ένα δίκτυο λέγεται ότι γενικεύει καλά όταν η σχέση εισόδου-εξόδου υπολογιζόμενη από το δίκτυο είναι σωστή (ή σχεδόν σωστή) για δείγματα εισόδου-εξόδου (δεδομένα ελέγχου) που ποτέ δεν χρησιμοποιήθηκαν στην δημιουργία ή στην εκπαίδευση του δικτύου. Ο όρος ‘’γενίκευση’’ είναι δανεισμένος από την ψυχολογία. Εδώ βέβαια υποτίθεται ότι τα δεδομένα ελέγχου προέρχονται από τον ίδιο πληθυσμό που χρησιμοποιήθηκε για να παράγει τα εκπαιδευτικά δεδομένα.

Η διαδικασία μάθησης (εκπαίδευσης ενός νευρωνικού δικτύου) μπορεί να θεωρηθεί σαν ένα πρόβλημα προσαρμογής καμπύλης. Το δίκτυο το ίδιο μπορεί να θεωρηθεί απλά σαν μία μη-γραμμική αντιστοίχηση εισόδων σε εξόδους. Τέτοια άποψη μας επιτρέπει να δούμε την γενίκευση όχι σαν μία μυστηριώδη ιδιότητα των νευρωνικών δικτύων αλλά απλούστερα σαν την επίδραση μίας καλής μη-γραμμικής παρεμβολής των δεδομένων εισόδου. Το δίκτυο εκτελεί χρήσιμη παρεμβολή κυρίως επειδή πολυεπίπεδα perceptrons με συνεχείς συναρτήσεις ενεργοποίησης οδηγούν σε εξωτερικές συναρτήσεις που είναι επίσης συνεχείς.



Σχήμα 12: (α) Κατάλληλα ταιριασμένα δεδομένα (καλή γενίκευση) (b) Υπερβολικά ταιριασμένα δεδομένα (φτωχή γενίκευση)

Το σχήμα 12.a επεξηγεί πως η γενίκευση μπορεί να συμβεί σ' ένα υποθετικό δίκτυο. Η μη-γραμμική αντιστοίχηση εισόδων σε εξόδους που παριστάνεται από την καμπύλη την απεικονιζόμενη σ' αυτό το σχήμα υπολογίζεται από το δίκτυο σαν αποτέλεσμα της εκμάθησης των σημείων με ετικέτα '' εκπαιδευτικά δεδομένα ''. Το σημείο που είναι μαρκαρισμένο στην καμπύλη σαν ''γενίκευση'', είναι αυτό που φαίνεται απλά ως το αποτέλεσμα της παρεμβολής που εκτελείται από το δίκτυο.

Ένα νευρωνικό δίκτυο που σχεδιάστηκε να γενικεύει καλά θα παράγει μία σωστή

αντιστοίχηση εισόδων σε εξόδους ακόμα και όταν η είσοδος είναι λίγο διαφορετική από τα παραδείγματα που χρησιμοποιήθηκαν για να εκπαιδευτεί το δίκτυο, όπως επεξηγείται στο σχήμα 12.a. Όταν όμως ένα νευρωνικό δίκτυο μαθαίνει αρκετές ειδικές σχέσεις εισόδου-εξόδου (υπερεκπαιδεύεται) το δίκτυο μπορεί να κρατάει στην μνήμη του τα δεδομένα εκπαίδευσης και συνεπώς να είναι λιγότερο ικανό να γενικεύει μεταξύ παρόμοιων δειγμάτων εισόδου-εξόδου. Όπως συνήθως, δεδομένα που φορτώνονται σ' ένα πολυεπίπεδο perceptron με τέτοιο τρόπο απαιτούν την χρησιμοποίηση περισσοτέρων κρυφών νευρώνων απ' ότι πραγματικά είναι απαραίτητο, με αποτέλεσμα ανεπιθύμητες καμπύλες στον χώρο του προβλήματος να αποθηκεύονται στα συναπτικά βάρη του δικτύου. Ένα παράδειγμα για το πως φτωχή γενίκευση οφειλόμενη σε 'μνημοποίηση' σ' ένα νευρωνικό δίκτυο μπορεί να συμβεί επεξηγείται στο σχήμα 12.b για τα ίδια δεδομένα που απεικονίζονται στο σχήμα 12.a. Η μνημοποίηση είναι βασικά ένας '' πίνακας ψαξίματος '' ('' look-up table '') που υπονοεί ότι η αντιστοίχηση εισόδων σε εξόδους που υπολογίζεται από το νευρωνικό δίκτυο δεν είναι '' λεία ''.

4.6.5 Κατάλληλο μέγεθος εκπαιδευτικού συνόλου για έγκυρη γενίκευση

Η γενίκευση επηρεάζεται από τρεις παράγοντες: το μέγεθος και την αποτελεσματικότητα του εκπαιδευτικού συνόλου , την αρχιτεκτονική του δικτύου και την φυσική πολυπλοκότητα του ίδιου του προβλήματος. Είναι φανερό ότι δεν έχουμε κανένα έλεγχο γύρω από τον τελευταίο παράγοντα. Σχετικά με τους άλλους δύο παράγοντες μπορούμε να δούμε το ζήτημα της γενίκευσης από δύο διαφορετικές σκοπιές (Hush και Horne , 1993) [1]:

- Η αρχιτεκτονική του δικτύου είναι φιξαρισμένη (ευτυχώς και σε αντιστοίχηση με την φυσική πολυπλοκότητα του δεδομένου προβλήματος) και το ζήτημα που πρέπει να επιλυθεί είναι αυτό του καθορισμού του μεγέθους του εκπαιδευτικού συνόλου που απαιτείται για να συμβεί μία καλή γενίκευση.

- Το μέγεθος του εκπαιδευτικού συνόλου είναι φιξαρισμένο και το ζήτημα που μας ενδιαφέρει είναι ο καθορισμός της καλύτερης αρχιτεκτονικής του δικτύου για την επίτευξη μίας καλής γενίκευσης.

Παρόλο που και οι δύο αυτές απόψεις είναι έγκυρες με τον δικό τους ξεχωριστό τρόπο η πρώτη άποψη είναι αυτή που πιο συχνά αντιμετωπίζουμε στην πράξη. Γι' αυτό θα συγκεντρωθούμε σ' αυτήν από δω και πέρα.

Πράγματι, η καταλληλότητα του μεγέθους του εκπαιδευτικού συνόλου είναι ένα θεωρητικό ζήτημα που έχει προσελκύσει ένα μεγάλο ποσό προσοχής στην βιβλιογραφία και συνεχίζει να το κάνει. Σ' αυτήν την υποενότητα περιληπτικά θα περιγράψουμε ένα χρήσιμο αποτέλεσμα προερχόμενο από τους Baum και Haussler (1989) [1], για την περίπτωση ενός νευρωνικού δικτύου που περιέχει ένα μόνο κρυφό επίπεδο και χρησιμοποιείται σαν δυαδικός ταξινομητής. Για να προετοιμάσουμε το έδαφος για μία πρόταση του μαθηματικού τύπου τους πρώτα παρουσιάζουμε κάποιους ορισμούς.

Ένα παράδειγμα ορίζεται σαν ζεύγος { x , d } όπου το διάνυσμα εισόδου $x \in R^P$ και η επιθυμητή έξοδος $d \in [-1,1]$. Με άλλα λόγια το δίκτυο δρα σαν δυαδικός ταξινομητής. Ένας κύκλος ορίζεται σαν μία σειρά από παραδείγματα που παίρνονται ανεξάρτητα και κατά τυχαίο τρόπο από κάποια κατανομή D. Έστω f μία συνάρτηση από το διάστημα R^P στο [-1,1] , με d=f(x). Ένα λάθος της συνάρτησης f , με αναφορά στην κατανομή D, ορίζεται ως η πιθανότητα η έξοδος y να είναι διαφορετική από το d για ένα ζεύγος (x, d) που επιλέχτηκε κατά τυχαίο τρόπο. Έστω M ο συνολικός αριθμός των κρυφών υπολογιστικών κόμβων. Έστω W ο συνολικός αριθμός των συναπτικών βαρών στο δίκτυο. Έστω N ο αριθμός των τυχαίων παραδειγμάτων που χρησιμοποιούνται για να εκπαιδεύσουν το δίκτυο. Έστω $\boldsymbol{\varepsilon}$ το φράγμα των λαθών που επιτρέπονται στο τεστ. Τότε βασιζόμενοι στους Baum και Haussler το δίκτυο σχεδόν βέβαια παρέχει γενίκευση αν και μόνο αν ικανοποιούνται οι δύο παρακάτω συνθήκες:

1. Το φράγμα των λαθών που δημιουργήθηκαν από το εκπαιδευτικό σύνολο είναι μικρότερο από $\boldsymbol{\varepsilon}/2$.

2. Ο αριθμός των παραδειγμάτων N, που χρησιμοποιήθηκαν στην εκπαίδευση είναι:

$$N \geq \frac{32W}{\varepsilon} \ln\left(\frac{32M}{\varepsilon}\right) \qquad (23)$$

όπου ln υποδηλώνει τον φυσικό λογάριθμο.

Η σχέση (23) παρέχει έναν ελεύθερο από κατανομές και χειρότερης περίπτωσης τύπο για την εκτίμηση του μεγέθους του εκπαιδευτικού συνόλου για ένα νευρωνικό δίκτυο ενός επιπέδου που είναι κατάλληλο για μία καλή γενίκευση. Λέμε ‘’χειρότερης περίπτωσης‘’ επειδή στην πράξη μπορεί να υπάρχει ένα τεράστιο αριθμητικό κενό ανάμεσα στο πραγματικό μέγεθος του εκπαιδευτικού συνόλου που απαιτείται και απ’ αυτό που προβλέπεται από το κριτήριο της σχέσης (23). Πρέπει όμως να υπογραμμιστεί ότι αυτό το κενό είναι απλά μία αντανάκλαση της φύσης της χειρότερης περίπτωσης του κριτηρίου.

Αγνοώντας τον λογαριθμικό παράγοντα στην σχέση (23) βλέπουμε ότι ο κατάλληλος αριθμός από εκπαιδευτικά παραδείγματα είναι, σε μία πρώτη προσέγγιση, ανάλογος του αριθμού των βαρών στο δίκτυο και αντιστρόφως ανάλογος της παραμέτρου ακριβείας ε. Πράγματι φαίνεται ότι στην πράξη όλο και όλο αυτό που χρειαζόμαστε για μία καλή γενίκευση είναι να ικανοποιείται η συνθήκη:

$$N \geq \frac{W}{\varepsilon} \qquad (24)$$

Έτσι μ’ ένα λάθος της τάξης του 10% λέμε ότι ο αριθμός των εκπαιδευτικών παραδειγμάτων πρέπει να είναι προσεγγιστικά 10 φορές ο αριθμός των συναπτικών βαρών στο δίκτυο.

Άσκηση αυτοαξιολόγησης 4.6 / 8:

Ποιά είναι η επίδραση του μεγέθους του εκπαιδευτικού συνόλου στην εκπαίδευση του νευρωνικού δικτύου.

Απάντηση:

Όσο μεγαλύτερο είναι το μέγεθος, τόσο μικρότερος είναι ο αριθμός των κύκλων και το μέσο τετραγωνικό λάθος.

Άσκηση αυτοαξιολόγησης 4.6 / 9:
Ποιο είναι το αποτέλεσμα της αύξησης του αριθμού των κρυφών νευρώνων;

Απάντηση:

Βελτιώνεται ελαφρώς το μέσο τετραγωνικό λάθος, αλλά δεν βελτιώνεται ο σωστός διαχωρισμός των προτύπων.

Άσκηση αυτοαξιολόγησης 4.6 / 10:

Ποιά είναι τα βασικά συμπεράσματα σχετικά με τη βέλτιστη τιμή των παραμέτρων η και α.

Απάντηση:
Είναι τρία και προκύπτουν από τα αποτελέσματα των προσομοιώσεων που παρουσιάζονται στα σχήματα της ενότητας 4.6.2.

Άσκηση αυτοαξιολόγησης 4.6 / 11:

Να αναφέρετε ένα κριτήριο σύγκλισης ενός Ν.Δ., δηλαδή πότε τερματίζει η εκτέλεση του αλγορίθμου.

Απάντηση:

Σε κάθε επανάληψη πρέπει να ισχύει η συνθήκη: $E(n+1) - E(n) \leq \varepsilon$, όπου ε οσοδήποτε μικρό.

Άσκηση αυτοαξιολόγησης 4.6 / 12:

Ποιες συνθήκες, και με ποιούς περιορισμούς, καθορίζουν το κατάλληλο μέγεθος εκπαιδευτικού συνόλου για έγκυρη γενίκευση;

Απάντηση:

Η σωστή απάντηση βρίσκεται στην ενότητα 4.6.5.

Δραστηριότητα 4/1:

Να σχεδιάστε και να υλοποιήσετε ένα πολυεπίπεδο Perceptron, το οποίο εκπαιδεύεται με το Γενικευμένο Δέλτα κανόνα. Να επιβεβαιώσετε τα αποτελέσματα που παρουσιάζονται στους πίνακες που περιέχονται στην ενότητα 4.6.

Υπόδειξη:

Μπορείτε να χρησιμοποιήσετε έτοιμα προγράμματα που υλοποιούν τον παραπάνω αλγόριθμο.

4.7 Σύνοψη κεφαλαίου

Σε αυτό το κεφάλαιο ασχοληθήκαμε με θέματα που αφορούν την υλοποίηση Τ.Ν.Δ. πολλών επιπέδων. Πρώτα συζητήσαμε το πρόβλημα της αρχικοποίησης, μερικούς τρόπους αποτελεσματικότερης εκτέλεσης των αλγορίθμων και την εφαρμογή τους σαν συστήματα ταξινόμησης. Όμως, το βασικό πρόβλημα στην υλοποίηση ενός Ν.Δ., για την επίλυση ενός πραγματικού προβλήματος είναι ότι δεν είναι εκ των προτέρων

γνωστή η διαμόρφωση του δικτύου. Δηλαδή δεν γνωρίζουμε τον ακριβή αριθμό κρυφών επιπέδων καθώς και τον αριθμό των κόμβων ανά κρυφό επίπεδο. Επειδή δεν υπάρχει κάποια γενική θεωρία που επιλύει αυτό το πρόβλημα προσπαθήσαμε να δώσουμε μερικές ιδέες για το πως αντιμετωπίζεται ο καθορισμός του βέλτιστου δικτύου, με τη βοήθεια μιας μελέτης περίπτωσης. Αυτή η μελέτη έγινε με τη βοήθεια προσομοίωσης σε ηλεκτρονικό υπολογιστή.

Το βασικό πλεονέκτημα των Perceptrons πολλών επιπέδων είναι ότι μπορούν να διαχωρίσουν μη-γραμμικά διαχωριζόμενα πρότυπα. Θεωρητικά λοιπόν, μπορούν να αντιμετωπίσουν οποιοδήποτε πρόβλημα ταξινόμησης προτύπων. Δυστυχώς όμως, στην πράξη τα πράγματα δεν είναι τόσο απλά. Όπως ήδη αναφέραμε, το βασικό πρόβλημα ενός τέτοιου δικτύου είναι ο καθορισμός της βέλτιστης δομής που είναι κατάλληλη για την επίλυση ενός συγκεκριμένου προβλήματος. Έχοντας επιλέξει σωστά το εκπαιδευτικό σύνολο, αυτόματα έχουμε καθορίσει τον αριθμό των εισόδων και των εξόδων του δικτύου. Δεν έχουμε όμως καμία εκ των προτέρων γνώση για την εσωτερική αρχιτεκτονική του δικτύου (αριθμός κρυφών επιπέδων και αριθμός κόμβων ανά κρυφό επίπεδο). Ένας προφανής τρόπος, για να αντιμετωπίσουμε αυτό το πρόβλημα, είναι με τη μέθοδο δοκιμής και λάθους (trial and error). Με βάση κάποιους εμπειρικούς κανόνες, υλοποιούμε μια συγκεκριμένη αρχιτεκτονική δικτύου και το εκπαιδεύουμε, ελπίζοντας σε καλή γενίκευση. Αν μετά την εκπαίδευση το λάθος στην έξοδο του δικτύου είναι σχετικά μεγάλο, τροποποιούμε την δομή του (προσθέτοντας ή αφαιρώντας κρυφά επίπεδα ή/και κόμβους) και το εκπαιδεύουμε πάλι. Έτσι, με διαδοχικές δοκιμές, ελπίζουμε να επιτύχουμε το (σχεδόν) βέλτιστο δίκτυο. Είναι προφανές ότι αυτή η διαδικασία είναι χρονοβόρα και επίπονη. Γι' αυτό το λόγο λέγεται ότι μαθαίνει κανείς μαζί με το δίκτυο.

Το πρόβλημα του πειραματικού προσδιορισμού του βέλτιστου πολυεπίπεδου Perceptron, αντιμετωπίστηκε σαν ένα σύνολο από υποπροβήματα. Αυτά ήταν ο καθορισμός του βέλτιστου αριθμού κρυφών νευρώνων, των βέλτιστων σταθερών μάθησης και ορμής και του βέλτιστου σχεδιασμού του δικτύου. Τέλος ασχοληθήκαμε με το πρόβλημα της γενίκευσης και του υπολογισμού του κατάλληλου μεγέθους του εκπαιδευτικού συνόλου για έγκυρη γενίκευση.

Εδώ ολοκληρώθηκε η παρουσίαση της ύλης των Τεχνητών Νευρωνικών Δικτύων. Στα επόμενα δύο κεφάλαια, θα γίνει η παρουσίαση του δεύτερου αντικειμένου αυτού του μαθήματος, δηλαδή των Γενετικών Αλγορίθμων.

4.8 Βιβλιογραφία

1. "NEURAL NETWORKS: A Comprehensive Foundation'', S. Haykin, Macmillan Publishing Company, N.Y., 1994 (ISBN 0-02-352761-7)

2. ''NEURAL NETWORK DESIGN'', M. T. Hagan, H. B. Demuth and m. Beal, PWS Publishing Company, Boston, 1996 (ISBN 0-534-94332-2)

3. 'Genetic Algorithms + Data Structures = Evolution Programs', Z. Michalewich, Springer - Verlang, 2nd ed., 1992.

4. ''GENETIC ALGORITHMS in Search, Optimization and Machine Learning'', D.E. Goldberg, Addison Wesley Publishing Company, Inc., 1989.