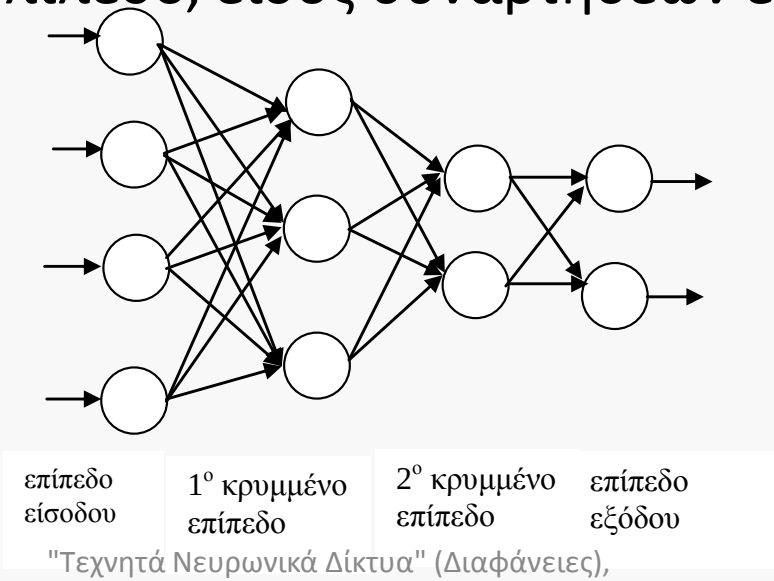
Μάθηση και Γενίκευση

Το Πολυεπίπεδο Perceptron (MultiLayer Perceptron (MLP))

- Έστω **σύνολο εκπαίδευσης D={(x^n,t^n)}**, n=1,…,N.
- $x^n=(x_{n1,\ldots,}x_{nd})^T$, $t^n=(t_{n1,\ldots,}t_{np})^T$
- Θα πρέπει το MLP να έχει d νευρώνες στο επίπεδο εισόδου και p νευρώνες στο επίπεδο εξόδου.
- Ο χρήστης καθορίζει: κρυμμένα επίπεδα, αριθμός κρυμμένων νευρώνων ανά επίπεδο, είδος συναρτήσεων ενεργοποίησης.

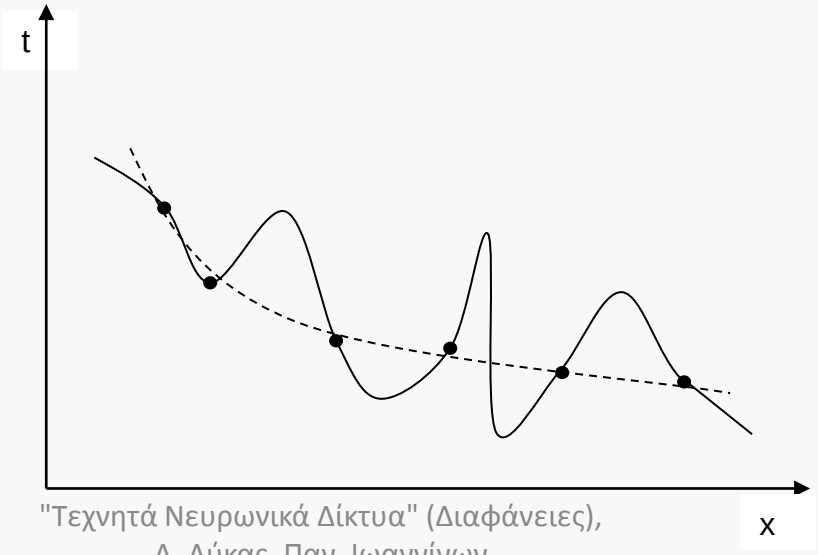


Ικανότητα Γενίκευσης

- Απώτερο στόχος της εκπαίδευσης είναι η η κατασκευή συστημάτων που να παρέχουν σωστές αποφάσεις για παραδείγματα που δεν έχουν χρησιμοποιηθεί κατά την εκπαίδευση: **ικανότητα γενίκευσης** (generalization).
- **Επιλογή αρχιτεκτονικής** στο MLP: με μεγάλο αριθμό κρυμμένων νευρώνων, ένα MLP μπορεί να εκπαιδευτεί ώστε να απεικονίζει με μεγάλη ακρίβεια όλα τα παραδείγματα του συνόλου εκπαίδευσης.
- ‘Μεγάλο’ MLP → συνήθως μικρή ικανότητα γενίκευσης: `απομνημονεύει' τα δεδομένα εκπαίδευσης και δεν παρουσιάζει καλές επιδόσεις σε νέα δεδομένα διότι, λόγω της μεγάλης ‘ευελιξίας’ του, δημιουργεί απεικονίσεις οι οποίες είναι συνήθως περισσότερο ‘πολύπλοκες’ απ’ ότι χρειάζεται.

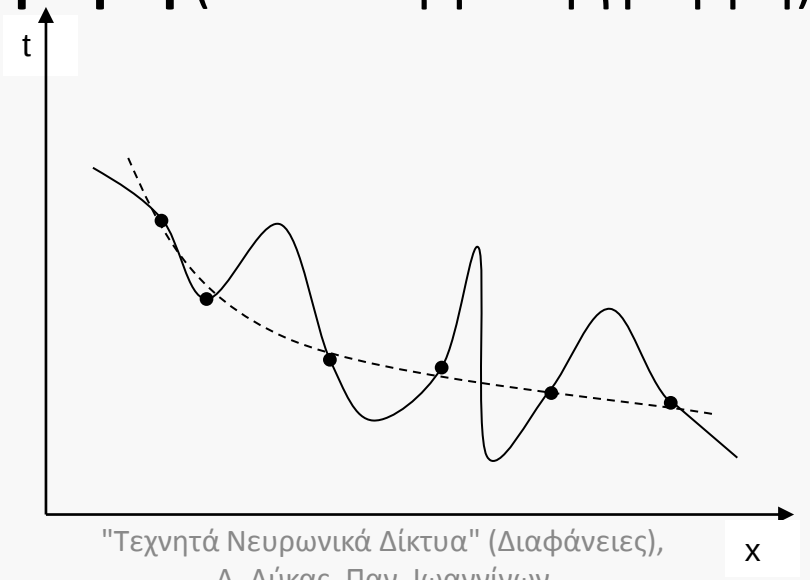
Ικανότητα Γενίκευσης (Παράδειγμα)

- Μονοδιάστατο πρόβλημα απεικόνισης: τα δεδομένα εκπαίδευσης αναπαρίστανται με τις μαύρες κουκίδες.
- Η συνάρτηση που αναπαρίσταται με συνεχή γραμμή, παρότι έχει μηδενικό σφάλμα εκπαίδευσης, είναι περισσότερο πολύπλοκη απότι χρειάζεται (**υπερεκπαίδευση**).
- Η συνάρτηση που αναπαρίσταται με διακεκομμένη γραμμή είναι πιο ομαλή και προτιμότερη ως λύση.



Ικανότητα Γενίκευσης (Παράδειγμα)

- Η πραγματική λύση από την οποία προέκυψαν τα δεδομένα εκπαίδευσης θα μπορούσε να είναι και η πολύπλοκη συνάρτηση. Αν ίσχυε κάτι τέτοιο τα παραδείγματα εκπαίδευσης που έχουμε στη διάθεσή μας δεν είναι αντιπροσωπευτικά.
- Για το συγκεκριμένο παράδειγμα, αφού και οι δύο συναρτήσεις ταιριάζουν επαρκώς στα δεδομένα, **η προτιμότερη λύση είναι η ομαλότερη συνάρτηση** (διακεκομμένη γραμμή).



Occam’s razor

- Ενα δίκτυο MLP με πολύ λίγους κρυμμένους νευρώνες πιθανόν να μην **έχει την απαιτούμενη ‘ευελιξία’** ώστε να μπορεί να ορίσει πολύπλοκες περιοχές απόφασης ή να προσεγγίσει συναρτήσεις με πολύπλοκη γραφική παράσταση (**υποεκπαίδευση**).
- Γενικά υπάρχει μια βέλτιστη αρχιτεκτονική ενός MLP, για την παρουσιάζει τις καλύτερες επιδόσεις γενίκευσης. Η βέλτιστη αυτή αρχιτεκτονική έχει το χαρακτηριστικό ότι είναι τόσο ‘πολύπλοκη’ (π.χ. σε αριθμό βαρών) όσο ‘πολύπλοκο’ είναι και το πρόβλημα μάθησης.
- **Βασική εμπειρική αρχή μηχανικής μάθησης (occam’s razor)**
  - Προτιμούμε το **απλούστερο μοντέλο** που μπορεί να **μάθει επαρκώς** τα παραδείγματα εκπαίδευσης.

Εκτίμηση της Γενικευτικής Ικανότητας

- Δεν έχει αντιμετωπιστεί επαρκώς με τη χρήση μαθηματικών μεθόδων. Καταφεύγουμε σε **εμπειρικές προσεγγίσεις:** χρήση **συνόλου παραδειγμάτων ελέγχου (test set).**
- Σύνολο ελέγχου: υποσύνολο των παραδειγμάτων που έχουμε στη διάθεσή μας, τα οποία **δεν τα χρησιμοποιούμε** κατά την εκπαίδευση του ΤΝΔ, η οποία γίνεται χρησιμοποιώντας τα υπόλοιπα παραδείγματα.
- Μετά την εκπαίδευση, εφαρμόζουμε τα παραδείγματα του συνόλου ελέγχου ως εισόδους στο ΤΝΔ και υπολογίζουμε τα αντίστοιχα σφάλματα στις εξόδους του.
- **Σφάλμα γενίκευσης: Η μέση τιμή (ή το ποσοστό) των σφαλμάτων ενός ΤΝΔ για τα παραδείγματα του συνόλου ελέγχου.**

Εκτίμηση της Γενικευτικής Ικανότητας

- Μικρό σφάλμα γενίκευσης συνεπάγεται υψηλή ικανότητα γενίκευσης και αντίστροφα.
- Για την αξιολόγηση της ικανότητας γενίκευσης απαιτείται ο **χωρισμός** του συνόλου των διαθέσιμων παραδειγμάτων σε δύο (ξένα μεταξύ τους) υποσύνολα:
  - το **σύνολο εκπαίδευσης (training set)** που το χρησιμοποιούμε για τον καθορισμό των βαρών του ΤΝΔ
  - το **σύνολο ελέγχου (test set)** που χρησιμοποιείται για τον υπολογισμό του σφάλματος γενίκευσης του δικτύου που προκύπτει από την εκπαίδευση.
- Πώς θα γίνει ο χωρισμός; Ποια παραδείγματα θα χρησιμοποιηθούν για εκπαίδευση και ποια για έλεγχο;

Hold-out

- Εάν τα παραδείγματα **είναι πολλά** δεν έχουμε ιδιαίτερο πρόβλημα (π.χ. τα χωρίζουμε τυχαία σε ποσοστό 70-30%) (μέθοδος **hold-out**).
- Εάν τα παραδείγματα **δεν είναι πολλά** χρειάζονται πιο πολύπλοκες προσεγγίσεις.
- Πολλαπλό hold-out:
  - Μπορούμε να επαναλάβουμε αρκετές φορές τη διαδικασία hold-out: τυχαία διάσπαση σε σύνολα εκπαίδευσης και ελέγχου, εκπαίδευση του ΤΝΔ και υπολογισμός του σφάλματος γενίκευσης.
  - Η τελική εκτίμηση για το σφάλμα γενίκευσης προκύπτει ως ο μέσος όρος των επιμέρους σφαλμάτων που υπολογίσαμε.

Cross-Validation

- **Διασταυρωμένη επικύρωση Κ-τμημάτων (Κ-fold cross-validation (**Κ-CV):
  - διαίρεση του συνόλου παραδειγμάτων D σε Κ *ξένα* μεταξύ τους υποσύνολα (folds) D_1 ,…, D_K (συνήθως Κ=10).
  - Για κάθε υποσύνολο D_i (i=1 ,…, Κ), εκπαιδεύουμε ένα ΤΝΔ θεωρώντας ως σύνολο εκπαίδευσης τα παραδείγματα των υπολοίπων Κ-1 υποσυνόλων (D-D_i) και υπολογίζουμε το σφάλμα γενίκευσης ge_i χρησιμοποιώντας ως σύνολο ελέγχου τα παραδείγματα του υποσυνόλου D_i.
  - Εκτιμούμε το σφάλμα γενίκευσης (ge) ως το μέσο όρο των επιμέρους σφαλμάτων ge_i
- Είναι πιο συστηματική, χρησιμοποιείται πολύ συχνά.

Leave-one-out

- **Ένα παράδειγμα ελέγχου κάθε φορά (Leave-one-out) (LOT)** Ειδική περίπτωση της διασταυρωμένης επικύρωσης K τμημάτων όταν θέσουμε K=N, όπου N ο αριθμός όλων των παραδειγμάτων του συνόλου D που έχουμε στη διάθεσή μας.
- Για κάθε (x^i,t^i) του συνόλου D κατασκευάζουμε ένα ΤΝΔ θεωρώντας ως σύνολο εκπαίδευσης ολόκληρο το D εκτός από το συγκεκριμένο παράδειγμα. Στη συνέχεια εκτιμούμε το σφάλμα γενίκευσης ge_i υπολογίζοντας το σφάλμα του ΤΝΔ για το συγκεκριμένο παράδειγμα που αγνοήσαμε κατά την εκπαίδευση.
- Επαναλαμβάνοντας τη διαδικασία για όλα τα (x^i,t^i), (i=1,…,N) εκτιμούμε το σφάλμα γενίκευσης ως το μέσο όρο των ge_i.
- Πιο αξιόπιστη (δεν έχει τυχαιότητα), αλλά αυξημένη πολυπλοκότητα.

Εκτίμηση της Γενικευτικής Ικανότητας

- Δύο ερωτήματα: αν εκπαιδεύσουμε πολλά ΤΝΔ (π.χ. 10-fold CV) για να εκτιμήσουμε την ικανότητα γενίκευσης και να επιλέξουμε τη βέλτιστη αρχιτεκτονική δικτύου:
  - α) πώς θα κατασκευάσουμε **το τελικό ΤΝΔ** που θα αποτελεί τη λύση στο πρόβλημά μας;
  - β) ποια θα είναι η ικανότητα γενίκευσης της τελικής αυτής λύσης;
- Απαντήσεις: α) κατασκευάζουμε το τελικό ΤΝΔ χρησιμοποιώντας την βέλτιστη αρχιτεκτονική που έχουμε βρεί και **όλα τα διαθέσιμα** παραδείγματα εκπαίδευσης.

  β) Η ικανότητα γενίκευσης του τελικού ΤΝΔ έχει ήδη υπολογιστεί από την μέθοδο εκτίμησης της ικανότητας γενίκευσης για τη βέλτιστη αρχιτεκτονική.

Εκτίμηση της Γενικευτικής Ικανότητας

<u>Άσκηση:</u>

Έστω ότι χρησιμοποιώντας τη μέθοδο 10-CV εξετάσαμε δύο αρχιτεκτονικές MLP για ένα πρόβλημα ταξινόμησης. Η πρώτη έχει σφάλμα γενίκευσης 5% και η δεύτερη 8%.

- α) Ποια από τις δύο αρχιτεκτονικές θα χρησιμοποιήσουμε για να κατασκευάσουμε την τελική λύση;
- β) Ποια θα είναι η ικανότητα γενίκευσης της τελικής λύσης;

Υπερεκπαίδευση (overtraining)

- Στην περίπτωση που η αρχιτεκτονική του MLP είναι **μεγαλύτερη (πιο ευέλικτο δίκτυο)** από την απαιτούμενη, τότε η διαδικασία ελαχιστοποίησης του τετραγωνικού σφάλματος είναι πολύ πιθανό να οδηγήσει στην εμφάνιση του φαινομένου της **υπερεκπαίδευσης**:
- **Υπερεκπαίδευση (overtraining) ονομάζουμε το φαινόμενο κατά το οποίο ένα ΤΝΔ έχει 'απομνημονεύσει' τα παραδείγματα εκπαίδευσης, (δηλαδή παρέχει ακριβείς τιμές εξόδου για τα παραδείγματα εκπαίδευσης), αλλά παρουσιάζει μειωμένη ικανότητα γενίκευσης.**
- Εχουν προταθεί κάποιες απλές τεχνικές για την αποφυγή (σε κάποιο βαθμό) του φαινομένου της υπερεκπαίδευσης κατά την εκπαίδευση ενός MLP.

Η μέθοδος της φθοράς των βαρών

- Ενας τρόπος για να περιορίσουμε την ‘ευελιξία’ ενός MLP είναι περιορίζοντας την αρχιτεκτονική του, δηλαδή ουσιαστικά των αριθμό των βαρών του δικτύου.
- Ένας εναλλακτικός τρόπος περιορισμού της ευελιξίας ενός MLP είναι **περιορίζοντας τις τιμές** που μπορούν να πάρουν τα βάρη κατά τη διάρκεια της εκπαίδευσης. Η ιδέα αυτή ονομάζεται **κανονικοποίηση (regularization)**.
- Ο πιο απλός τρόπος για να επιτύχουμε κανονικοποίηση βασίζεται στην προσθήκη ενός **όρου τιμωρίας (penalty term)** στη συνάρτηση τετραγωνικού σφάλματος που ελαχιστοποιούμε κατά την εκπαίδευση του δικτύου.

Η μέθοδος της φθοράς των βαρών

- Πιο συγκεκριμένα, ένας **όρος κανονικοποίησης** που χρησιμοποιείται συχνότερα είναι **το άθροισμα των τετραγώνων των τιμών των βαρών** (όπου L ο αριθμός των βαρών)

$$R(w)=\sum_{i=1}^{L} w_i^2$$

- Η συνάρτηση που ελαχιστοποιείται κατά την εκπαίδευση γίνεται:

$$E_R(w)=E(w)+rR(w)=E(w)+r\sum_{i=1}^{L} w_i^2$$

E(w) είναι η συνάρτηση τετραγωνικού σφάλματος εκπαίδευσης.

- Η παράμετρος r καθορίζει το σχετικό βάρος των δύο στόχων της εκπαίδευσης: αφενός ελαχιστοποίηση του E(w), αφετέρου διατήρηση μικρών απόλυτων τιμών των βαρών του δικτύου.

Η μέθοδος της φθοράς των βαρών

- Η προσθήκη του όρου κανονικοποίησης στην ουσία παρεμποδίζει τα βάρη να λάβουν υψηλές (κατ' απόλυτη τιμή) τιμές κατά την εκπαίδευση.
- Μερικές φορές οδηγεί κάποιες τιμές των βαρών να γίνουν σχεδόν μηδέν, δηλαδή στην ουσία είναι σαν οι αντίστοιχες συνδέσεις να αφαιρούνται από το δίκτυο.
- Μπορούμε δηλαδή να θεωρήσουμε ότι οι τιμές των βαρών 'φθείρονται' κατά τη διάρκεια της εκπαίδευσης, για το λόγο αυτό η μέθοδος ονομάζεται **εκπαίδευση με φθορά βαρών (weight decay)** .
- Ενημέρωση των βαρών: $$w_i(t+1)=w_i(t)-\eta\frac{\partial E_R}{\partial w_i}$$

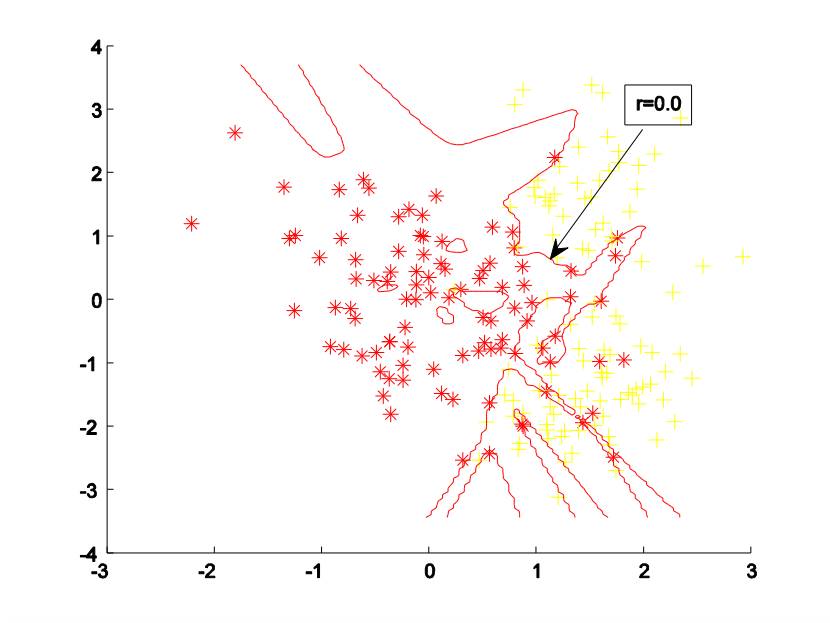
$$w_i(t+1)=w_i(t)-\eta\left(\frac{\partial E}{\partial w_i}+2rw_i(t)\right)$$

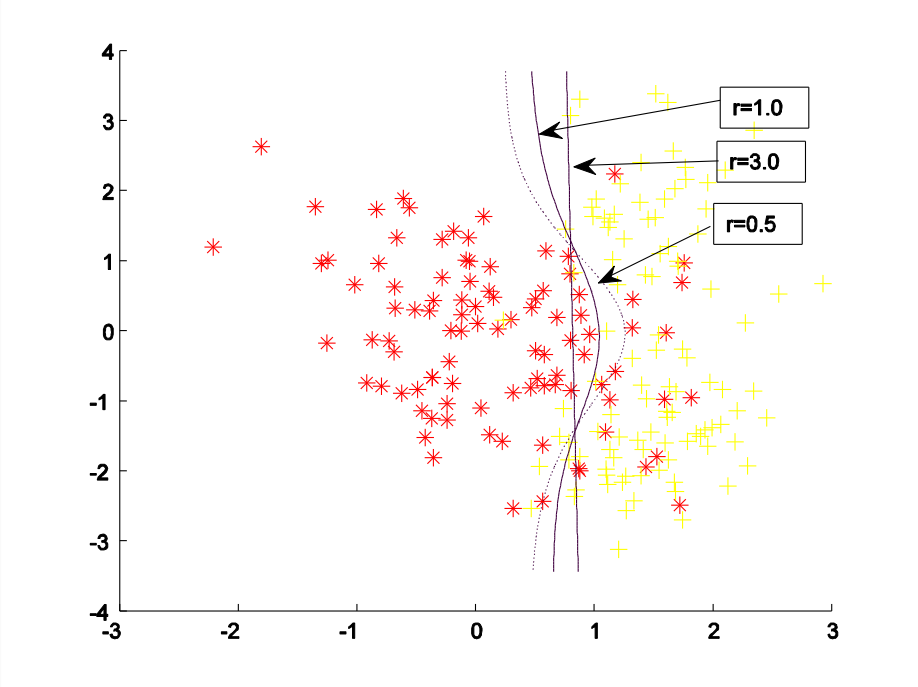
Η μέθοδος της φθοράς των βαρών

- Εάν η παράμετρος r έχει καθοριστεί σωστά και το μέγεθος του δικτύου είναι μεγαλύτερο απ’ ότι απαιτείται, στο τέλος της εκπαίδευσης προκύπτουν συνήθως δίκτυα με καλύτερες δυνατότητες γενίκευσης.
- Εάν η παράμετρος r είναι μεγάλη τότε παρεμποδίζεται η προσαρμογή του δικτύου στα παραδείγματα εκπαίδευσης.
- Εάν η παράμετρος r τείνει στο μηδέν τότε είναι σαν να εκπαιδεύουμε το δίκτυο χωρίς κανονικοποίηση. Η σωστή ρύθμιση της παραμέτρου r αποτελεί το βασικό πρόβλημα αυτής της μεθόδου.

Η μέθοδος της φθοράς των βαρών

- MLP με 1 κρυμμένο επίπεδο με 20 νευρώνες





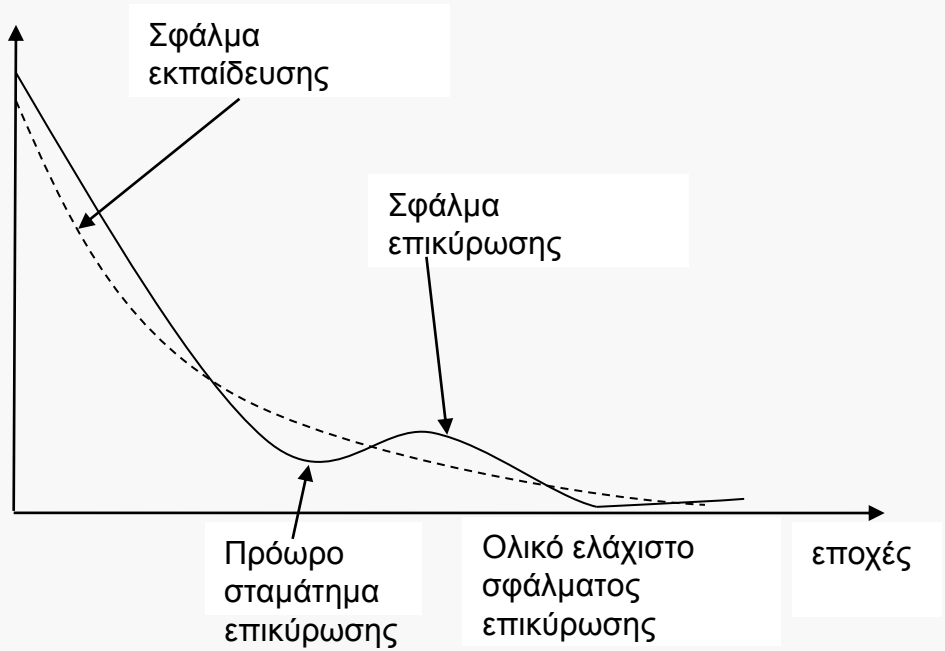
Πρόωρο Σταμάτημα (early stopping)

- Εκπαιδεύουμε το MLP (ενημερώνουμε τα βάρη του) μέσω της ελαχιστοποίησης του σφάλματος εκπαίδευσης.
- Σε τακτά χρονικά διαστήματα (π.χ. κάθε 10 εποχές) **‘παγώνουμε’ τη διαδικασία εκπαίδευσης και** με τις τρέχουσες τιμές των βαρών **υπολογίζουμε μια εκτίμηση του σφάλματος γενίκευσης** σε ένα ανεξάρτητο σύνολο παραδειγμάτων (διαφορετικό από το σύνολο εκπαίδευσης και το σύνολο ελέγχου).
- Το τρίτο αυτό σύνολο παραδειγμάτων που χρησιμοποιούμε ονομάζεται **σύνολο επικύρωσης (validation set)** και το αντίστοιχο σφάλμα ονομάζεται **σφάλμα επικύρωσης**.
- Κατόπιν συνεχίζουμε τη διαδικασία εκπαίδευσης και της ενημέρωσης των βαρών μέχρι το επόμενο χρονικό σημείο υπολογισμού του σφάλματος επικύρωσης.

Πρόωρο Σταμάτημα (early stopping)

- Στις αρχικές επαναλήψεις της εκπαίδευσης και όσο προχωρεί η εκπαίδευση, μειώνεται το σφάλμα εκπαίδευσης και συγχρόνως μειώνεται και το σφάλμα επικύρωσης.
- Υπάρχει συνήθως ένα **χρονικό σημείο (ειδικά στις περιπτώσεις μεγάλων δικτύων) πέρα από το οποίο περαιτέρω μείωση του σφάλματος εκπαίδευσης οδηγεί σε αύξηση του σφάλματος επικύρωσης**, διότι αρχίζει να εμφανίζεται το φαινόμενο της υπερεκπαίδευσης.
- Στο σημείο αυτό μπορούμε να σταματήσουμε την εκπαίδευση του δικτύου (**πρόωρο σταμάτημα**).

Πρόωρο Σταμάτημα (early stopping)



Πρόωρο Σταμάτημα (early stopping)

- Εναλλακτικά, μπορούμε, αντί να σταματήσουμε πρόωρα, να εκτελέσουμε τον αλγόριθμο εκπαίδευσης μέχρι να τερματίσουμε σε τοπικό ελάχιστο, φροντίζοντας όμως να **αποθηκεύουμε κάθε φορά το διάνυσμα βαρών w_{val} που παρέχει το μικρότερο σφάλμα επικύρωσης που έχουμε υπολογίσει μέχρι στιγμής** κατά τη διάρκεια της εκπαίδευσης.
  - Η τιμή των βαρών w_{val} στο τέλος της εκπαίδευσης αποτελεί και το τελικό διάνυσμα βαρών για το MLP, διότι παρέχει την ελάχιστη τιμή του σφάλματος επικύρωσης.

Πρόωρο Σταμάτημα (early stopping)

- Συνοψίζοντας, στη μέθοδο του πρόωρου σταματήματος:
  - α) Το MLP πρέπει να είναι σχετικά μεγάλο.
  - β) ενημερώνουμε τα βάρη χρησιμοποιώντας τα παραδείγματα του συνόλου εκπαίδευσης
  - γ) επιλέγουμε ως τελική λύση για τα βάρη αυτή με την μικρότερη τιμή του σφάλματος που υπολογίζουμε χρησιμοποιώντας τα παραδείγματα του συνόλου επικύρωσης.
- **Τίμημα:** θα πρέπει να αφαιρέσουμε ένα ποσοστό των παραδειγμάτων από το σύνολο εκπαίδευσης και να τα βάλουμε στο σύνολο επικύρωσης. Πρόβλημα εάν τα παραδείγματα είναι λίγα. Εξάρτηση από τον διαμερισμό.
- Δεν επιτρέπεται τα σύνολα εκπαίδευσης, επικύρωσης και ελέγχου να έχουν κοινά παραδείγματα.

Επιλογή MLP με ελαχιστοποίηση σφάλματος επικύρωσης

- Εστω ότι θεωρούμε MLPs με ένα κρυμμένο επίπεδο.
- Εστω ότι **το σφάλμα γενίκευσης εκτιμάται από το σφάλμα επικύρωσης.**

1. Καθορισμός αρχικού αριθμού κρυμμένων νευρώνων M (π.χ. M=2), μέγιστου αριθμού κρυμμένων νευρώνων M_{max} και παραμέτρων εκπαίδευσης (π.χ. ρυθμός μάθησης).
2. Διαίρεση του συνόλου παραδειγμάτων στα τρία σύνολα εκπαίδευσης, επικύρωσης και ελέγχου.
3. Εκπαίδευση του δικτύου MLP με M κρυμμένους νευρώνες με ελαχιστοποίηση του **σφάλματος επικύρωσης** και εύρεση της λύσης MLP(M) της οποία το σφάλμα επικύρωσης είναι ge(M).
4. Αύξηση του αριθμού των κρυμμένων νευρώνων, π.χ. M:=M+1 και επιστροφή στο βήμα 3 εάν $M \leq M_{max}$.

Επιλογή MLP με ελαχιστοποίηση σφάλματος επικύρωσης

5. Επιλογή ως τελικής λύσης εκείνου του δικτύου MLP(M^*) με το μικρότερο σφάλμα επικύρωσης: $ge(M^*) \leq ge(M)$
6. Υπολογισμός του **σφάλματος γενίκευσης** του τελικού δικτύου χρησιμοποιώντας τα παραδείγματα του **συνόλου ελέγχου**.

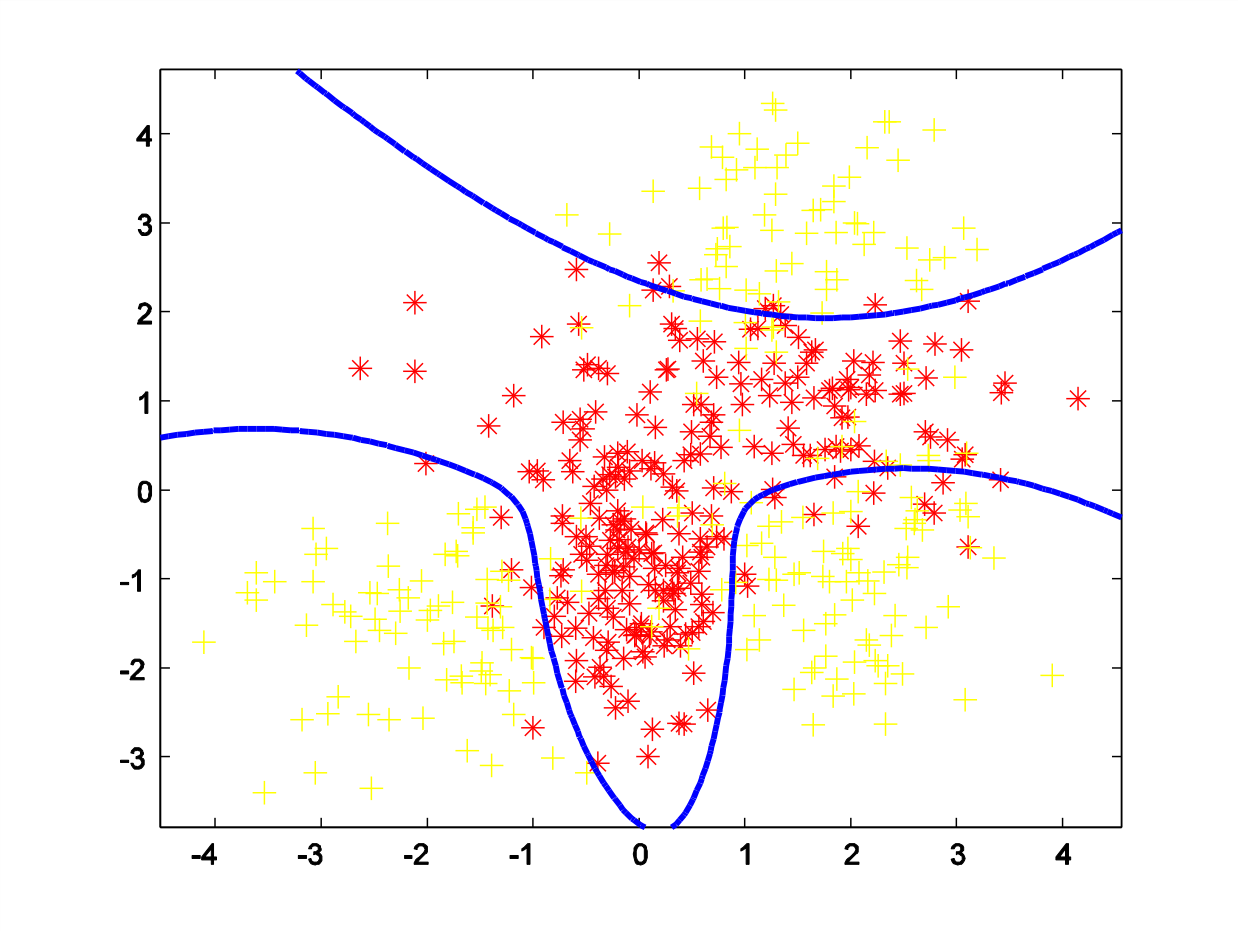
- Εξάρτηση από το συγκεκριμένο διαχωρισμό των παραδειγμάτων στα τρία επιμέρους σύνολα.
- Συνήθως δεν χρειάζεται να φτάσουμε μέχρι την τιμή M_{max}. Ενώ αρχικά το εκτιμώμενο σφάλμα γενίκευσης μειώνεται αυξάνοντας το M, υπάρχει κάποια τιμή του M πέρα από την οποία το σφάλμα γενίκευσης αρχίζει να αυξάνεται.
- Μπορούμε τότε να σταματήσουμε να αυξάνουμε το M υποθέτοντας ότι εντοπίσαμε την κατάλληλη αρχιτεκτονική.

Επιλογή MLP με cross-validation (K-CV)

1. Καθορισμός αρχικού M (π.χ. M=2), M_{max}, αριθμού folds K (π.χ. K=10) και παραμέτρων εκπαίδευσης (π.χ. ρυθμός μάθησης).
2. Διαμερισμός του συνόλου παραδειγμάτων D σε υποσύνολα D_1, …, D_K για την εφαρμογή της τεχνικής K-CV.
3. Υπολογισμός με (K-CV) του σφάλματος γενίκευσης ge(M) για M κρυμμένους νευρώνες.
4. Αύξηση του αριθμού των κρυμμένων νευρώνων, π.χ. M:=M+1 και επιστροφή στο βήμα 3 εάν $M \leq M_{max}$.
5. Επιλογή ως βέλτιστης αρχιτεκτονικής εκείνης με το μικρότερο σφάλμα γενίκευσης: $ge(M^*) \leq ge(M)$
6. Εκπαίδευση του MLP με M^* κρυμμένους νευρώνες σε όλο το σύνολο παραδειγμάτων και εύρεση της τελικής λύσης.

Παράδειγμα Εκπαίδευσης

Σύνολο εκπαίδευσης (μέγιστη γενίκευση: 89%)

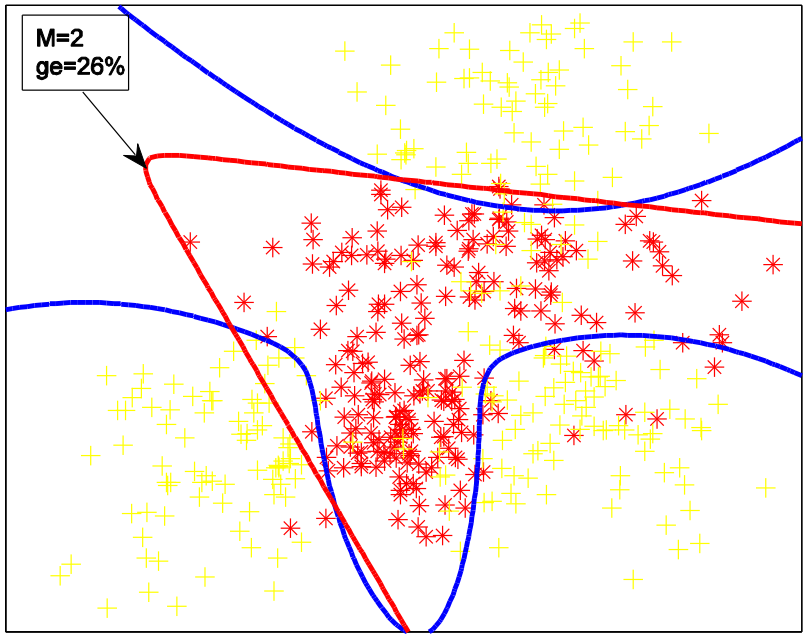


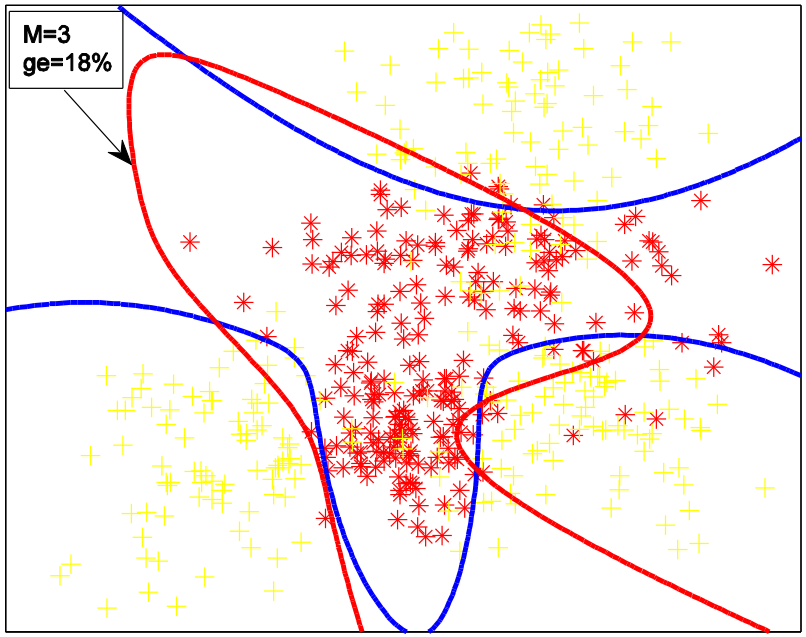
Παράδειγμα Εκπαίδευσης

Βέλτιστο M=5, gen=100-12=88%

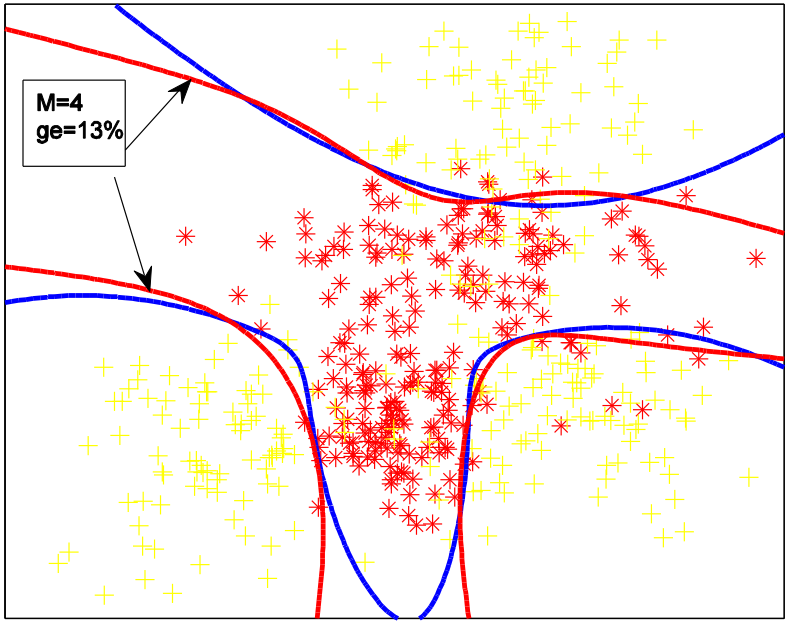
M	Σφάλμα Γενίκευσης (10-CV)
2	28%
3	18%
4	13%
5	12%
6	15%
7	15%

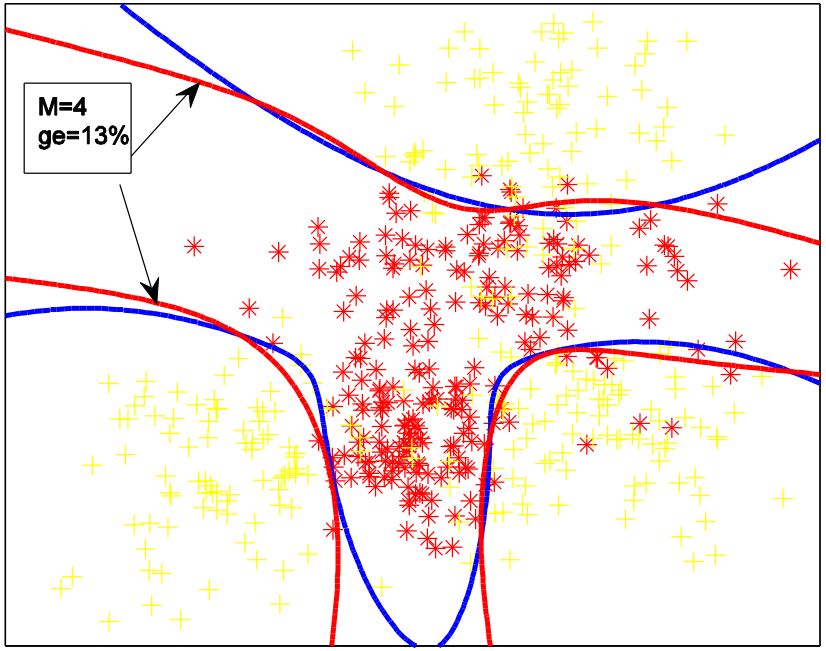
Παράδειγμα Εκπαίδευσης



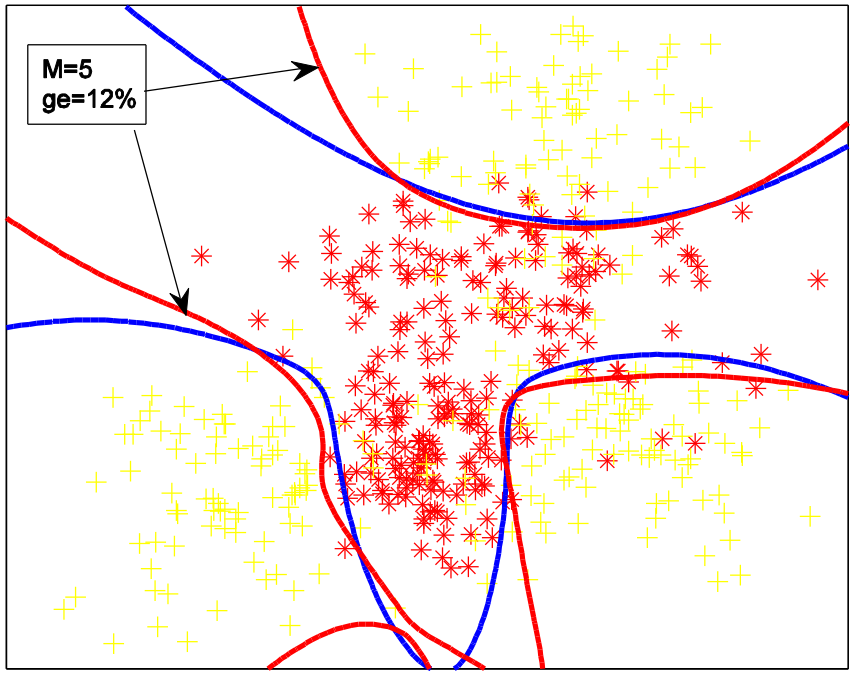


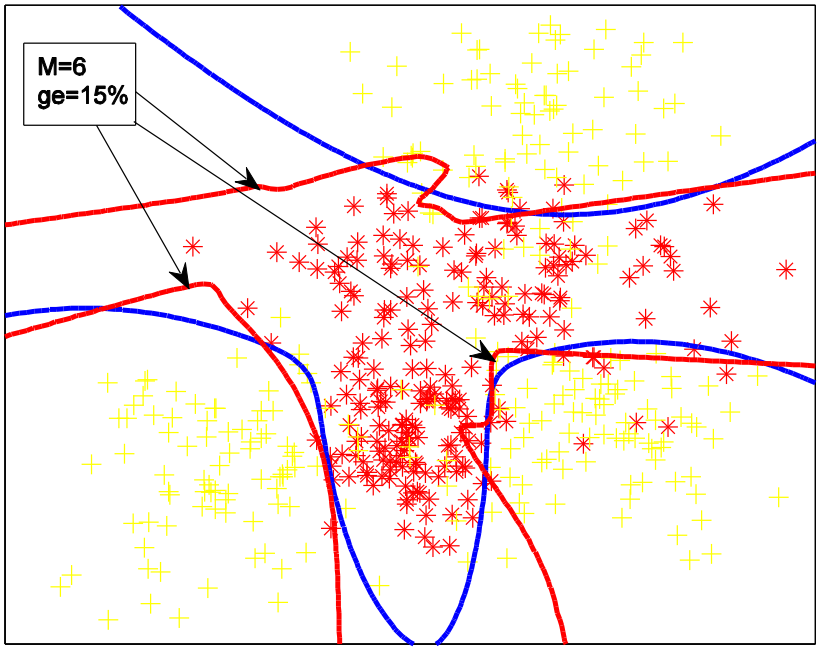
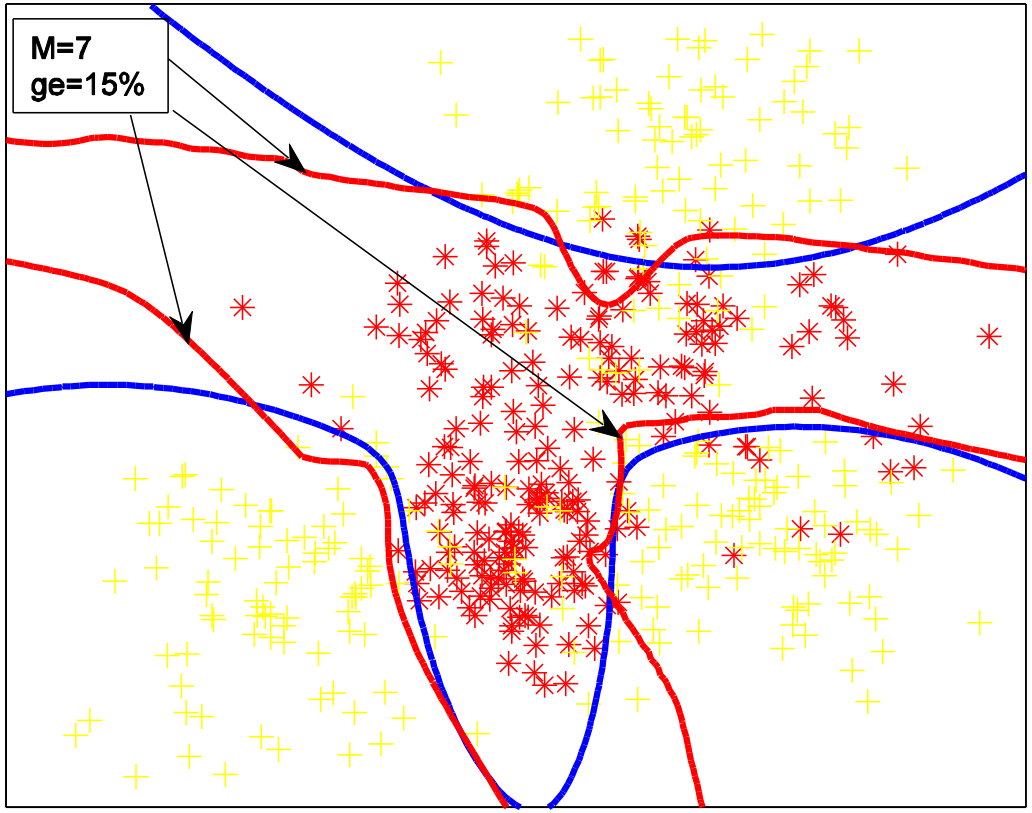
Παράδειγμα Εκπαίδευσης





Παράδειγμα Εκπαίδευσης





Παράδειγμα Εκπαίδευσης