# ΑΝΑΓΝΩΡΙΣΗ ΠΡΟΤΥΠΩΝ: ΠΡΑΚΤΙΚΗ ΑΣΚΗΣΗ 2

**Θέμα 1**

Σκοπός αυτού του τμήματος της άσκησης είναι η εξοικείωση με τη χρήση ταξινομητών μέσω του λογισμικού που παρέχει το WEKA Machine Learning Project. Από την ιστοσελίδα http://www.cs.waikato.ac.nz/~ml/weka/index.html, να κατεβάσετε και να εγκαταστήσετε το σχετικό λογισμικό. Αν χρησιμοποιείτε Windows, κατεβάστε την έκδοση που περιλαμβάνει και την JAVA VM. Διαβάστε τη σχετική τεκμηρίωση που παρέχει ο ιστοτόπος του project για τη χρήση της WEKA γενικά (http://weka.wikispaces.com/Frequently+Asked+Questions) και για τον λεγόμενο "Explorer" της WEKA, ώστε να εξοικειωθείτε με τις βασικές έννοιες (http://www.cs.waikato.ac.nz/~ml/weka/gui_explorer.html) .

Θα εργαστείτε με δεδομένα που θα κατεβάσετε από το UCI Machine Learning Repository (http://mlearn.ics.uci.edu/MLRepository.html). Πρόκειται για ταμιευτήρα δεδομένων που περιλαμβάνει τα κυριότερα τυποποιημένα προβλήματα (benchmarks) που χρησιμοποιούνται σε προβλήματα μηχανικής μάθησης. Ειδικότερα, τα προβλήματα ταξινόμησης που θα αντιμετωπίσετε τα βρίσκετε ενεργοποιώντας τον σύνδεσμο "Summary page" και είναι τα εξής:

1) IRIS PLANT DATABASE (ταξινόμηση φυτών Iris σε τρία είδη).
2) MOLECULAR BIOLOGY DATABASES, απ' όπου επιλέγετε την PROMOTER GENE SEQUENCE DATABASE (ταξινόμηση ακολουθιών γονιδίων). Να μετατρέψετε εξ αρχής τα ονόματα των βάσεων σε αριθμητικές μεταβλητές (π.χ. a=0, c=1, g=2, t=3).

Για διευκόλυνσή σας, τα δεδομένα που χρειάζονται επισυνάπτονται και εδώ, στο eclass (αρχείο UCIdata.rar)

Εξοικειωθείτε με τη φύση των προβλημάτων ταξινόμησης, μέσω των αρχείων με επέκταση ".names". Τα δεδομένα βρίσκονται στα αρχεία με επέκταση ".data". Για να χρησιμοποιήσετε τα αρχεία αυτά στη WEKA, πρέπει να κάνετε μικρές αλλαγές, ώστε να μετατραπούν σε format ARFF (http://weka.wikispaces.com/ARFF).

Στη συνέχεια, αφού εκκινήσετε το πρόγραμμα WEKA, επιλέξτε από τα "Applications" τον "Explorer" και φορτώστε διαδοχικά τα αρχεία δεδομένων ARFF που δημιουργήσατε. Για τα δύο προβλήματα ταξινόμησης, να χρησιμοποιήσετε τους ταξινομητές Naïve Bayes, SVM (αναφέρεται ως SMO στις επιλογές που παρέχει το πρόγραμμα), και δίκτυο ακτινικής βάσης (RBFnetwork). Για όλους τους ταξινομητές, να κάνετε χρήση της μεθόδου tenfold crossvalidation για την αποτίμηση των αποτελεσμάτων.

Να πειραματιστείτε με τις παραμέτρους των δύο τελευταίων ταξινομητών, ώστε να επιτύχετε την καλύτερη δυνατή γενικευτική ικανότητα, όπως μετριέται από την ποσότητα «Correctly classified instances».

- Στην περίπτωση του SMO οι βασικές παράμετροι είναι ο τύπος του πυρήνα (γραμμικός, πολυωνυμικός, RBF), καθώς και η σταθερά C που ρυθμίζει την έκταση του περιθωρίου. Όσον αφορά τους πυρήνες: Ο γραμμικός πυρήνας προκύπτει από τον πολυωνυμικό (Polykernel) με επιλογή της παραμέτρου E ίση με 1. Πειραματιστείτε και με καθαρά πολυωνυμικούς πυρήνες με E ίσο με 2 και 3. Στην περίπτωση του RBF πυρήνα, η παράμετρος Gamma ρυθμίζει την τυπική απόκλιση των γκαουσιανών συναρτήσεων.
- Στην περίπτωση του RBFnetwork, οι βασικές παράμετροι είναι το πλήθος των ενδιάμεσων νευρώνων (numClusters) και η ελάχιστη τυπική απόκλιση των γκαουσιανών συναρτήσεων (minStdDev).

Να περιγράψετε τα πειράματά σας και να δώσετε συγκριτικούς πίνακες αξιολόγησης των ταξινομητών για τα διάφορα προβλήματα ταξινόμησης που εξετάσατε.

**Θέμα 2**

Ο σκοπός μας εδώ είναι να δημιουργήσουμε ένα νευρωνικό δίκτυο που θα προσεγγίζει τη συνάρτηση $y(x) = \sin(6\pi x)$ στο διάστημα [0,1].

Η εργασία θα γίνει με χρήση του Neural Network Toolbox του Matlab. Για να εξοικειωθείτε, διαβάστε πρώτα το help του Neural Network toobox, και ιδιαίτερα το κεφάλαιο Backpropagation, όπου υπάρχουν αρκετά παραδείγματα που θα σας βοηθήσουν.

α) Να δημιουργήσετε ένα εκπαιδευτικό σύνολο 200 προτύπων για την εκπαίδευση του δικτύου. Οι είσοδοι του δικτύου καθορίζονται παίρνοντας τυχαία ομοιόμορφα κατανεμημένα σημεία $x_i$ στο διάστημα [0,1], ενώ οι αντίστοιχες έξοδοι-στόχοι καθορίζονται από τη σχέση

$$y_i = \sin(6\pi x_i) + \varepsilon r_i$$

όπου ο τελευταίος όρος αντιπροσωπεύει θόρυβο. $\varepsilon$ είναι ένας μικρός πραγματικός αριθμός (χρησιμοποιήστε ε=0.2) και ο $r_i$ παίρνει τυχαίες τιμές ομοιόμορφα κατανεμημένες μεταξύ -1 και 1.

β) Χρησιμοποιώντας τα παραπάνω πρότυπα, να εκπαιδεύσετε ένα δίκτυο χωρίς ανατροφοδότηση με μία είσοδο, ένα ενδιάμεσο στρώμα νευρώνων και ένα νευρώνα εξόδου. Όσον αφορά τις συναρτήσεις ενεργοποίησης, χρησιμοποιήστε σιγμοειδή συνάρτηση υπερβολικής εφαπτομένης στο ενδιάμεσο στρώμα νευρώνων και γραμμική συνάρτηση για το νευρώνα εξόδου. Να χρησιμοποιήσετε τρείς διαφορετικούς αλγορίθμους εκπαίδευσης:

ι) Backpropagation με όρο ορμής
ιι) Μέθοδο συζυγών κλίσεων (conjugate gradient)
ιιι) Μέθοδο Levenberg-Marquardt

Να εκπαιδεύσετε δίκτυα με 5, 10, 15 και 20 ενδιάμεσους νευρώνες. Επιλέξτε εμπειρικά μόνοι σας κατάλληλο κριτήριο τερματισμού της εκπαίδευσης και αριθμό επαναλήψεων. Για κάθε αριθμό ενδιάμεσων νευρώνων, να επαναληφθεί η εκπαίδευση 20 φορές ξεκινώντας απο διαφορετικά αρχικά βάρη και να καταγραφεί το μέσο τετραγωνικό σφάλμα που επιτυγχάνεται για κάθε μέθοδο εκπαίδευσης και αριθμό ενδιάμεσων νευρώνων.

γ) Να γίνει αποτίμηση της ικανότητας γενίκευσης των δικτύων που εκπαιδεύσατε με χρήση ενός συνόλου ελέγχου (test set) που θα αποτελείται από 500 πρότυπα ελέγχου. Και πάλι οι είσοδοι του δικτύου καθορίζονται παίρνοντας τυχαία ομοιόμορφα κατανεμημένα σημεία $x_i$ στο διάστημα [0,1], ενώ οι αντίστοιχες έξοδοι-στόχοι καθορίζονται από τη σχέση

$$y_i = \sin(6\pi x_i)$$

Προσέξτε ότι εδώ δεν υπάρχει θόρυβος, εφόσον στόχος μας είναι να δούμε πόσο καλά έχει προσεγγίσει το δίκτυο τη συνάρτησή μας. Για κάθε αριθμό ενδιάμεσων νευρώνων και για κάθε μέθοδο εκπαίδευσης, να χρησιμοποιηθεί ως κριτήριο για την αξιολόγηση της ικανότητας γενίκευσης των δικτύων το μέσο τετραγωνικό σφάλμα για το σύνολο των προτύπων του συνόλου ελέγχου (μέσος όρος για τα 20 δίκτυα που έχουν εκπαιδευτεί με διαφορετικά αρχικά βάρη).

δ) Να επαναλάβετε τα παραπάνω χωρίζοντας το αρχικό σύνολο των 200 προτύπων σε σύνολο εκπαίδευσης 150 προτύπων και σύνολο επικύρωσης (validation set) 50 προτύπων. Να αποτιμήσετε και πάλι την ικανότητα γενίκευσης των δικτύων σας χρησιμοποιώντας ως κριτήριο αξιολόγησης το μέσο τετραγωνικό σφάλμα πάνω στο ίδιο σύνολο ελέγχου των 500 προτύπων όπως και προηγουμένως.