2η ΕΡΓΑΣΙΑ AIVC/AI and Machine Learning

Στην εργασία αυτή σας ζητείται να υλοποιήσετε στο Matlab τις παραλλαγές του αλγορίθμου επιβλεπόμενης διανυσματικής κβάντισης LVQ2 και LFM (Learning From Mistakes). Στη συνέχεια θα πρέπει να χρησιμοποιήσετε τους αλγορίθμους αυτούς προκειμένου να ταξινομήσετε τα εξής δεδομένα:

α) **Wine dataset**[1]: 178 διανύσματα εισόδου 13 χαρακτηριστικών από χημικές αναλύσεις κρασιών τριών τύπων. Η έξοδος για κάθε διάνυσμα εισόδου είναι η κωδικοποίηση του τύπου του κρασιού (μια εκ των τριών στηλών του μοναδιαίου πίνακα I_{3x3}).

β) **Iris dataset**: 150 διανύσματα εισόδου 4 χαρακτηριστικών (μήκος και πλάτος πετάλων και σεπάλων) ανθέων του φυτού iris. Ταξινόμηση σε τρία υποείδη με εντοπισμένη κωδικοποίηση των εξόδων όπως και στο wine dataset.

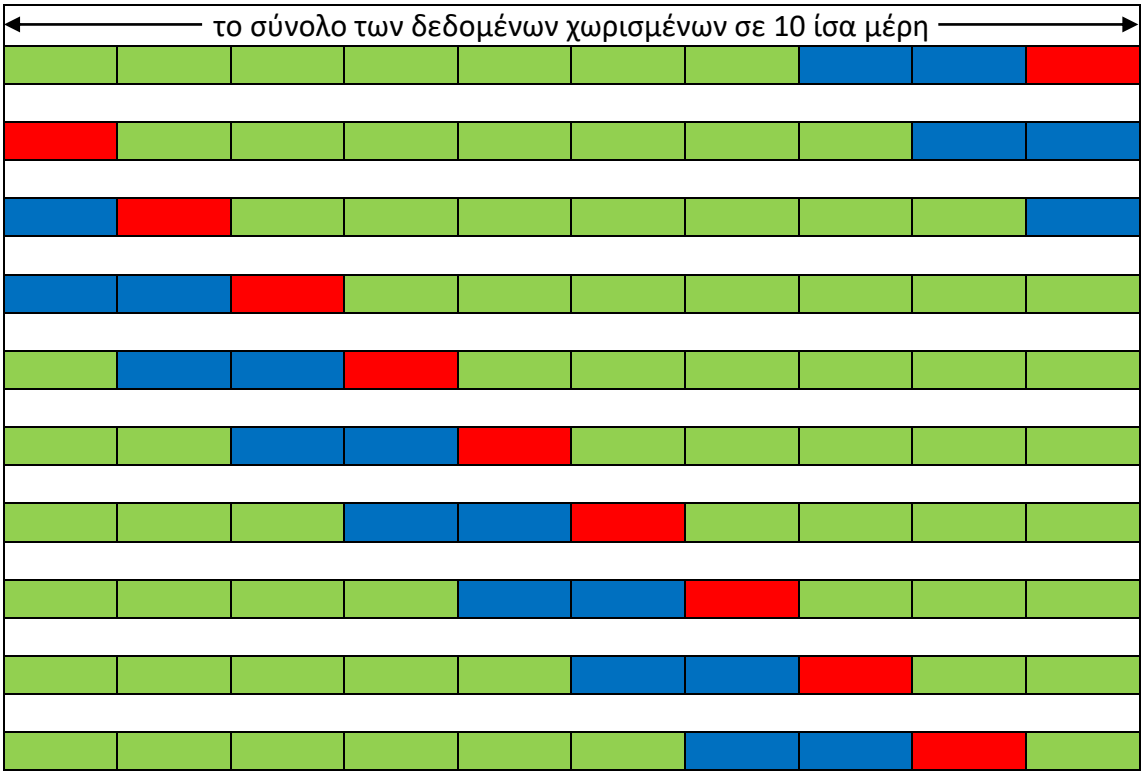
Πριν την εκπαίδευση των νευρωνικών δικτύων πρέπει να κάνετε μια τυχαία μετάθεση των δεδομένων και στη συνέχεια να τα παρουσιάζετε ακολουθιακά σύμφωνα με τη νέα τυχαία σειρά (η διαδικασία αυτή προσομοιώνει την τυχαία επιλογή των δεδομένων κατά την εκπαίδευση).

Για την αύξηση της γενικευτικής ικανότητας των νευρωνικών ταξινομητών και την εξαγωγή αξιόπιστων στατιστικών ως προς την επίδοσή τους στα δεδομένα της εφαρμογής, θα χρειαστεί να εξαχθούν τρία υποσύνολα δεδομένων: i) το σύνολο εκπαίδευσης (training set) που θα αποτελείται από το 70% των δεδομένων, ii) το σύνολο τεκμηρίωσης (validation set) που θα αποτελείται από το 20% των δεδομένων και iii) το σύνολο ελέγχου (test set) που θα αποτελείται από το υπόλοιπο 10% των δεδομένων. Η τυχαία σειρά των δεδομένων μετά την τυχαία μετάθεση του προηγούμενου βήματος εξασφαλίζει ότι τα τρία σύνολα δεδομένων θα έχουν παρόμοια στατιστικά.

Για την εκπαίδευση των ταξινομητών θα χρησιμοποιείται το εκάστοτε σύνολο εκπαίδευσης. Επειδή, τα αποτελέσματα εξαρτώνται από την αρχικοποίηση των συναπτικών βαρών των ταξινομητών, τη σειρά εμφάνισης των δεδομένων εκπαίδευσης αλλά και τα ίδια τα δεδομένα του συνόλου εκπαίδευσης, προκειμένου να αυξήσουμε την σημαντικότητα των αποτελεσμάτων (ακρίβεια ταξινόμησης) επαναλαμβάνουμε τα πειράματά μας αρκετές φορές με διαφορετικές αρχικοποιήσεις συναπτικών βαρών, τυχαίες μεταθέσεις και διαχωρισμό των δεδομένων στα τρία υποσύνολα. Τα τελικά αποτελέσματα της επίδοσης των ταξινομητών θα δίνονται ως οι μέσοι όροι των επαναλήψεων.

Επιπλέον, λόγω του μικρού πλήθους δεδομένων, τα πειράματα πρέπει να γίνουν με την μέθοδο της διασταυρούμενης επικύρωσης (cross-validation). Σύμφωνα με την μέθοδο αυτή, θα πρέπει να χωρίσετε τα δεδομένα σε 10 ίσα (κατά προσέγγιση) μέρη (10-fold cross-validation) και στη συνέχεια να δημιουργείτε τα τρία σύνολα δεδομένων από συνεννώσεις διαδοχικών μερών.

Όπως φαίνεται και στην Εικ. 1, η κάθε τριάδα συνόλων θα δημιουργείται με κυκλική ολίσθηση των ίσων μερών κατά ένα μέρος τη φορά. Για τα 178 δεδομένα του συνόλου ταξινόμησης κρασιών το πλήθος των δεδομένων για τα τρία υποσύνολα θα είναι 124-36-18 ενώ για το iris dataset θα είναι 105-30-15.



	δεδομένα συνόλου εκπαίδευσης (70%)
	δεδομένα συνόλου τεκμηρίωσης (20%)
	δεδομένα συνόλου ελέγχου (10%)

Εικ 1: η τεχνική της 10πλής διασταυρούμενης επικύρωσης (10-fold cross-validation).

<u>Ζητούμενα</u>

α) Να δοθούν τα αποτελέσματα ταξινόμησης για τα δύο προβλήματα ως μέσοι όροι της ακρίβειας ταξινόμησης για κάθε ένα σύνολο χωριστά (δηλαδή, τα ποσοστά σωστών ταξινομήσεων για τα σύνολα εκπαίδευσης, τεκμηρίωσης και ελέγχου).

β) Η αρχικοποίηση των αντιπροσωπευτικών διανυσμάτων του δικτύου LVQ (δηλαδή, των διανυσμάτων βαρών των νευρωνίων) να γίνει με τυχαία επιλογή δεδομένων, της ίδιας κατηγορίας με το εκάστοτε αντιπροσωπευτικό διάνυσμα που αρχικοποιείται, από το σύνολο εκπαίδευσης.

γ) Να χρησιμοποιήσετε κέρδος προσαρμογής που ελαττώνεται σύμφωνα με τη συνάρτηση $\alpha(t) = \alpha 0/(1+K_\alpha t)$ για $\alpha 0 = 0.5$ και $K_\alpha = 0.1$ ή 0.01 (το K_α ρυθμίζει πόσο γρήγορα θα τείνει το $\alpha(t)$ στο μηδέν) και για τα δύο προβλήματα. Να επιλέξετε

διάρκεια εκπαίδευσης 5, 10 ή 15 εποχών. Για τον LVQ2 επιλέξτε μέγεθος παραθύρου που αντιστοιχεί σε s = 0.6.

δ) Να επαναλάβετε τα πειράματα για 1, 2 ή 5 νευρώνια ανά κατηγορία.

ε) Να εμφανίσετε τα αποτελέσματα του κάθε αλγορίθμου σε μορφή πίνακα για κάθε διαφορετική επιλογή των παραμέτρων.

Παραδοτέα εργασίας

Ένα αρχείο zip με το όνομα σας το οποίο θα ανεβάσετε στο eclass και το οποίο θα περιέχει τεχνική αναφορά με ένα listing του προγράμματος που υλοποιήσατε στο matlab/octave με πλήρη τεκμηρίωση και τα αποτελέσματα των πειραμάτων (ακρίβεια ταξινόμησης στο τέλος της εκπαίδευσης για κάθε ένα από τα τρία σύνολα δεδομένων) μαζί με τους χρόνους εκπαίδευσης σε μορφή πίνακα για κάθε LVQ μοντέλο ταξινόμησης. Επίσης, να δοθούν και τα διαγράμματα με την ακρίβεια ταξινόμησης ανά εποχή για το πείραμα με τη βέλτιστη επίδοση στα δεδομένα του συνόλου τεκμηρίωσης.

ΠΡΟΣΟΧΗ: ΟΙ ΑΝΤΙΓΡΑΦΕΣ ΘΑ ΜΗΔΕΝΙΖΟΝΤΑΙ

[1]Παράδειγμα διαβάσματος από το αρχείο 'wine.txt':

```
fid = fopen('wine.net');
junk = fgetl(fid);
junk = fscanf(fid,'%s',1);
nin = fscanf(fid,'%d',1);   %nin = number of inputs
junk = fscanf(fid,'%s',1);
nout = fscanf(fid,'%d',1);  %nout = number of outputs
junk = fscanf(fid,'%s',3);
nrpat = fscanf(fid,'%d',1); %nrpat = number of patterns
A = fscanf(fid,'%f',[nin+nout,Inf]); %A = [I/O pairs]
fclose(fid);

x = A(1:nin,:);   %x = input patterns as column vectors
d = A(nin+1:nin+nout,:);    %d = desired output vectors
```