Εργασια 3

Ον/νυμο : Αντώνης Μαρκάκης

Α.Μ.: 2016010017

Ημερ/νια: 19/12/18

**Φόρτωση δεδομένων**

Τα δεδομένα φορτώνονται στο πρόγραμμα «matlab» με τις εξής εντολές:

filename = 'breast' ;

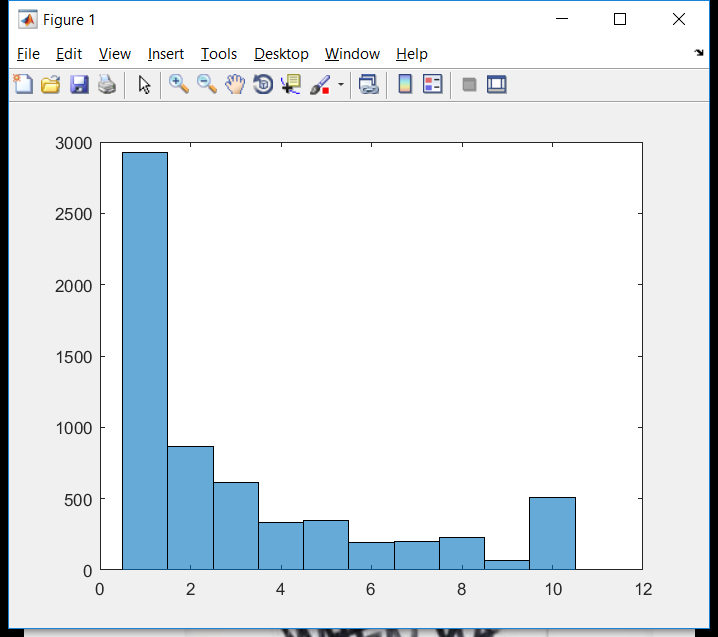
[x] = xlsread(filename,'B1:J699') ;

[t] = xlsread(filename, 'K1:K699') ;

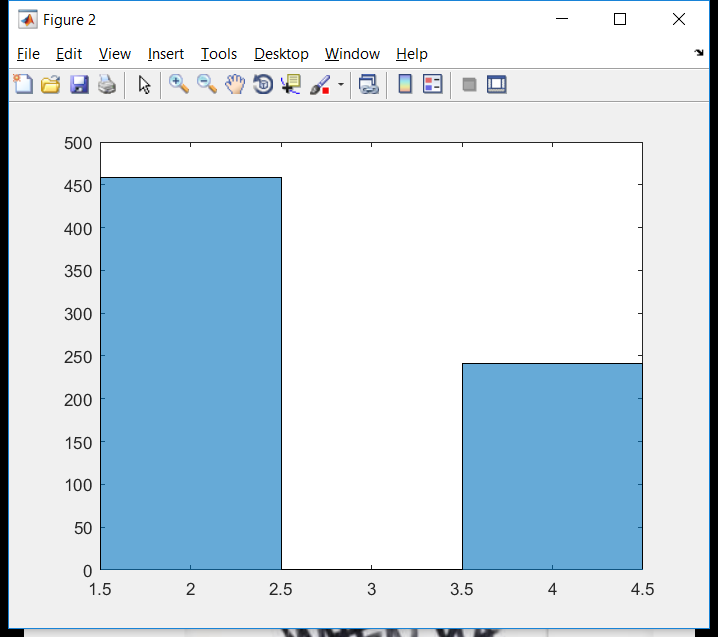
Τα δοθέν δεδομένα, τα οποία ήταν σε αρχείο μορφής «.csv», φορτώθηκαν στην εφαρμογή excel για καλύτερο χειρισμό τους.

Τα δεδομένα των εξετάσεων φορτώνονται σε έναν πίνακα 9x699 , και κατόπιν αναστρέφεται για την χρήση του στις εντολές machine learning. Το ίδιο ισχύει για τα αποτελέσματα μόνο που μετασχηματίζεται από διάνυσμα 1x699, με 2 και 4 ανάλογα το είδος του όγκου, σε πίνακα 2x699, όπου η πρώτη γραμμή περιέχει άσσο σε κάθε θέση στην οποία τα δεδομένα υποδεικνύουν ότι υπάρχει καλοήθης όγκος, και αντίστοιχα η δεύτερη για δεδομένα που αντιστοιχούν σε κακοήθη όγκο.

**Ιστογράμματα χαρακτηριστικών και κλάσης**



Το παραπάνω ιστόγραμμα αντιστοιχεί στα χαρακτηριστικά της κάθε μέτρησης. Το πλάτος κάθε μπάρας αντιστοιχεί σε ένα εύρος τιμών και ύψος στην ποσότητα των τιμώ που περιέχονται σε αυτό. Οι περισσότερες τιμές των χαρακτηριστικών βρίσκονται στο εύρος ~0,25-~1,5 ενώ οι λιγότερες στο εύρος 7,5-8,5 .



Το γράφημα των αποτελεσμάτων, ουσιαστικά παρουσιάζει το ύψος του ποσού κάθε αποτελέσματος. Οι ασθενείς με καλοήθη όγκο είναι λίγο περισσότεροι από 450 ενώ με κακοήθη είναι περίπου 240.

**Αρχιτεκτονική δικτύου**

Η αρχιτεκτονική του νευρωνικού δικτύου της εργασία υόποιείται με τις εξής εντολές :

net = feedforwardnet([5 4 2]);

net.layers{1}.transferFcn = 'hardlim' ;

net.layers{2}.transferFcn = 'tansig' ;

net.layers{3}.transferFcn = 'tansig' ;

Με την εντολή «net = feedforwardnet([5 4 2]);»

Κατασκευάζεται το δίκτυο, το χωρίζει σε 3 layers, όπου η πρώτη έχει 5 γραμμές δεδομένων η δεύτερη 4 και η τρίτη 2. Με τις εντολές

net.layers{1}.transferFcn = 'hardlim' ;

net.layers{2}.transferFcn = 'tansig' ;

net.layers{3}.transferFcn = 'tansig' ;

ορίζουμε τον τρόπο σύνδεσης μεταξύ των layers.

**Εκπαίδευση δικτύου**

To νευρωνικό δίκτυο εκπαιδεύτηκε με την εντολή:

[net,tr] = train(net,input,target) ;

Όπου ως «input» οι μετρήσεις των εξετάσεων, «target» τα αποτελέσματα των εξετάσεων και ταυτόχρονα οι τιμές-στόχος του δικτύου, ενώ ως «net», το δίκτυο και «tr» τα χαρακτηριστικά εκμάθησης του νευρωνικού, . Οι τιμές των παραμέτρων ορίστηκαν ως εξής:

net.divideParam.trainRatio = 66/100;

net.divideParam.valRatio = 17/100;

net.divideParam.testRatio = 17/100;

Χωριζόμενες σε 66% τα δεδομένα προς εκπαίδευση, 17% για επαλήθευση και 17% για να τεστάρει τα αποτελέσματα, το νευρωνικό δίκτυο.

**Αποτελέσματα test set**

Με την χρήση των εντολών :

testX = input(:,tr.testInd) ;

testT = target(:,tr.testInd) ;

testY = net(testX) ;

δημιουργούνται 3 πίνακες, σχετικοί με το τεστ του νευρωνικού δικτύου. Ο πίνακας «testX» είναι 9x119 και περιέχει τις τιμές των εξετάσεων του 17% των δεδομένων που επιλέχθηκαν για εξέταση, «testT» τα πραγματικά αποτελέσματα αυτών και «testY» τα αποτελέσματα του δικτύου. ΜΕ την χρήση των εντολών

dif = (testT - testY) ;

er = zeros(1,119);

for i=1:119

if (dif(i) > 0.8 || dif(i)< -0.8)

j = j + 1 ;

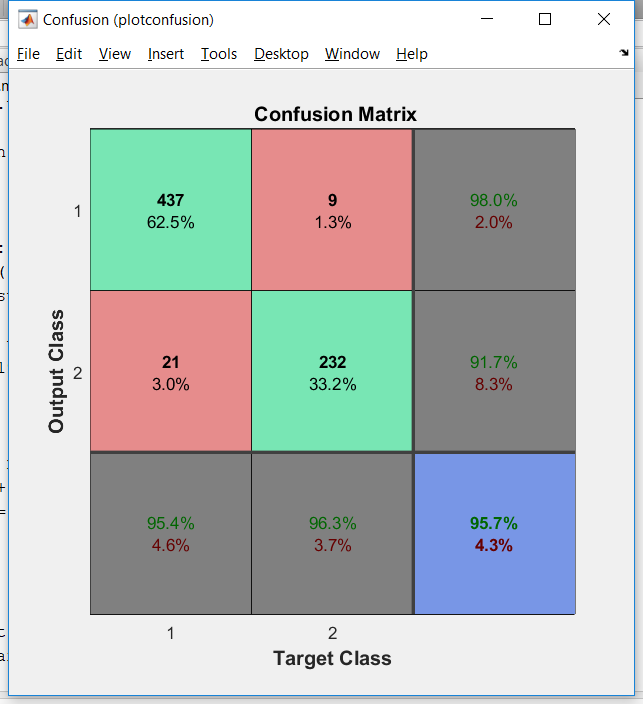
er(i) = 1 ;

end

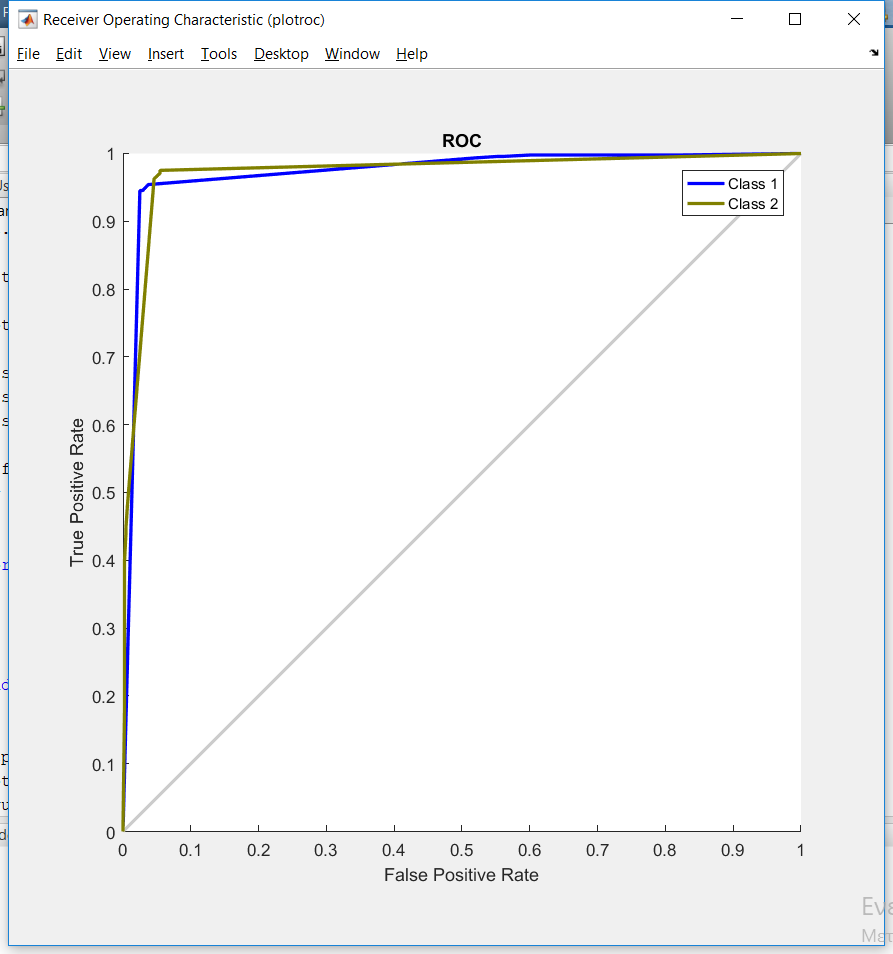
end

δημιουργείται ένας πίνακας «dif» ο οποίος μετράει την διαφορά μεταξύ των αποτελεσμάτων και των τιμών του δικτύου, ένας πίνακας «er» ο οποίος έχει άσσο σε κάθε τιμή για την οποία το νευρωνικό δίκτυο βρήκε λάθος αποτέλεσμα, και ένας μετρητής «j» του οποίου η τιμή δείχνει το ποσό των λαθών.

**Ανάλυση γραφημάτων αποτελεσμάτων**



Το παραπάνω γράφημα αντιστοιχεί στο confusion matrix του νευρωνικού δικτύου, το οποίο συγκρίνει ανά κλάση τα αποτελέσματα με τους στόχους. Η πρώτη γραμμή αντιστοιχεί στην πρώτη κλάση (καλοήθη όγκους) και δείχνει ότι το παρών νευρωνικό δίκτυο αποτυγχάνει κατά 2%, και η δεύτερη δείχνει την δεύτερη κλάση (κακοήθη όγκο), όπου αποτυγχάνει κατά 8.3%.Αντίστοιχα το ίδιο γίνεται και στις στήλες. Συνολικά το νευρωνικό προβλέπει σωστά το 95.7%.



Στο παραπάνω γράφημα φαίνεται η καμπύλη ROC του νευρωνικού δικτύου. Η μπλέ γραμμή αντιστοιχεί στην πρώτη κλάση, ενώ η δεύτερη κλάση αντιστοιχεί στην μπέζ γραμμή. Η γραμμή με το καλύτερο classification των δεδομένων πρέπει ναι «αγκαλιάζει» καλύτερα την γωνία, στην προκειμένη, η μπλέ.