Αναγνώριση προτύπων
Εργασία 5
Σάββας Λιάπης 57403

Άσκηση

Το IRIS data set (δες http://en.wikipedia.org/wiki/Iris_flower_data_set) περιέχει μετρήσεις της μορφής: (μήκος σέπαλου, πλάτος σέπαλου, μήκος πετάλου, πλάτος πετάλου) σε cm για 150 φυτά iris (είδος κρίνου, αγριόκρινο). Από αυτά τα 150 φυτά, 50 είναι Iris Setosa (ω1), 50 είναι Iris Versicolour (ω2) και 50 είναι Iris Virginica (ω3). Γνωρίζουμε ότι μόνο η μία (Iris Setosa) από τις άλλες δυο κλάσεις είναι γραμμικά διαχωρίσιμη. Εάν δεν μπορείτε να προχωρήσετε με 4 χαρακτηριστικά, λύστε την άσκηση χρησιμοποιώντας μόνο τα τελευταία 2 χαρακτηριστικά (μήκος πετάλου, πλάτος πετάλου)

A. Να βρεθεί ένας γραμμικός ταξινομητής που να χωρίζει την Iris Setosa από τις άλλες 2 κατηγορίες με το batch perceptron (αλγόριθμος 3), και με το batch relaxation with margin (αλγόριθμος 6).
B. Να βρεθεί ένας γραμμικός ταξινομητής που να χωρίζει την Iris Setosa (ω1) από τις άλλες 2 κατηγορίες (ω2, ω3) χρησιμοποιώντας την μέθοδο των ελαχίστων τετραγώνων με χρήση του ψευδοαντιστρόφου, καθώς και με την επαναληπτική μέθοδο LMS (Windrow-Hopf) (αλγόριθμος 8).
C. Να βρεθεί ένας γραμμικός ταξινομητής που να χωρίζει την Iris Versicolour (ω2) από την Iris Virginica (ω3) χρησιμοποιώντας την μέθοδο των ελαχίστων τετραγώνων με χρήση του ψευδοαντιστρόφου (LS) καθώς και με την επαναληπτική μέθοδο του Ho-Kashyap (αλγόριθμος 9).
D. Να βρείτε τους γραμμικούς ταξινομητές και των τριών κατηγοριών χρησιμοποιώντας την μέθοδο των ελαχίστων τετραγώνων με χρήση του ψευδοαντιστρόφου (LS) και όλα τα χαρακτηριστικά (1,2,3,4)
E. Επαναλάβατε το D για τους χώρους (1,2,3) και (2,3,4) (1=μήκος σέπαλου, 2=πλάτος σέπαλου, 3=μήκος πετάλου,4= πλάτος πετάλου) και δείξτε τα υπερεπίπεδα διαχωρισμού στον χώρο που έχετε καλύτερα αποτελέσματα.
F. Προσπαθήστε να βρείτε τους γραμμικούς ταξινομητές και των τριών κατηγοριών, χρησιμοποιώντας την δομή Kesler. Σχολιάστε τα αποτελέσματά σας (να δώσετε και τα αρχεία λογισμικού που χρησιμοποιήσατε).

Για να αναλύσουμε το iris dataset πρέπει αρχικά να το εισάγουμε. Το dataset εισάγεται σε μορφή xlx (υπήρχε πρόχειρο από πέρσι) και αρχικοποιείται στην meas (από measurements). Μετά από αυτό μετασχηματίζουμε κατάλληλα τον πίνακα αυτόν για να μπορεί να χρησιμοποιηθεί αναλόγως σε κάθε ερώτημα.

Επίσης για τα ερωτήματα A, B, C θα χρησιμοποιούμε μόνο τα χαρακτηριστικά μήκος πετάλου, πλάτος πετάλου δηλαδή τα χαρακτηριστικά των 2 τελευταίων στηλών.

Απάντηση 5.Α:

Αρχικά δημιουργούμε έναν νέο πίνακα ο οποίος έχει στην πρώτη στήλη 1 και -1 και οι υπόλοιπες είναι ο πίνακα meas. Στην ουσία τα στοιχεία της πρώτης στήλης σηματοδοτούν της 2 κλάσεις που θέλουμε να χωρίσουμε τα λουλούδια.
Η συνάρτηση αυτή προσπαθεί να ελαχιστοποιήσει την συνάρτηση κριτηρίου:

$$\mathbf{J_p(a)} = \sum_{\mathbf{y \in Y}} (-\mathbf{a^t y})$$

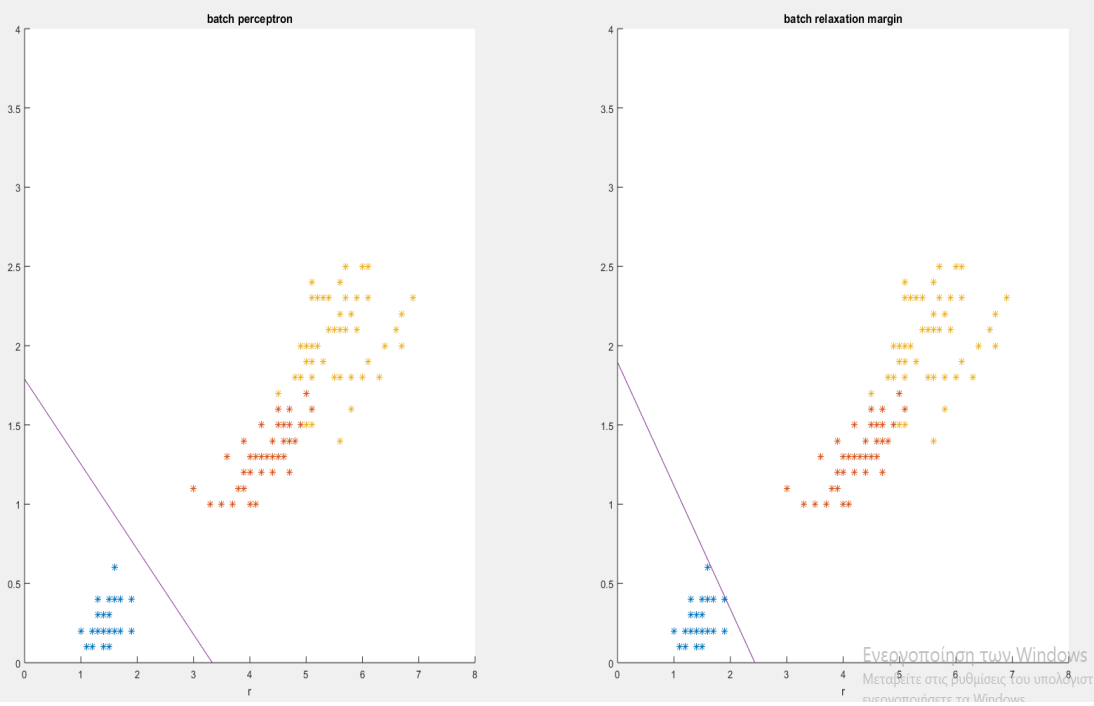
με διάνυσμα κλίσεων $\sum_{\mathbf{y \in Y}} -\mathbf{y}$ χρησιμοποιώντας την αναδρομική σχέση

$$\boldsymbol{\alpha(\kappa + 1) = \alpha(\kappa) + \eta(\kappa) \sum_{y \in Y_k} (y)}$$

όπου Υκ το σύνολο των λάθος ταξινομημένων δειγμάτων. Δηλαδή παίρνω τα σημεία που είναι λάθος ταξινομημένα και αθροίζω το εσωτερικό τους γινόμενο με το y. Όσο το εσωτερικό γινόμενο είναι >0 έχουμε λάθος ταξινομημένα δείγματα και η διαδικασία συνεχίζει. Για να αποτυπώσουμε το διάνυσμα σύγκλισης απλά προσθέτουμε στο διάνυσμα βαρών τις τιμές σύγκλισης

Ο αλγόριθμος batch relaxation with margin στην ουσία πάλι ελέγχει αν όλα τα εσωτερικά γινόμενα περνούν από ένα κατώφλι απλά αυτό το κατώφλι δεν είναι το τόσο αυστηρό μηδεν άλλα ένας μικρός αριθμός ο οποίος δεν θα προσθέσει μεγάλο σφάλμα στην ταξινόμηση.

Τρέχοντας τον αλγόριθμο exA παίρνουμε :



Βλέπουμε ότι για γραμμικώς διαχωρίσιμες κλάσεις παίρνουμε αρκετά καλά αποτελέσματα. Από τον αλγόριθμο batch perceptron παίρνουμε μηδενικό σφάλμα ενώ από το αλγόριθμο batch relaxation with margin μπορεί να υπάρχει ένα μικρό οριακό σφάλμα αλλα έχουμε γλυτώσει υπολογιστική ισχύ.

Απάντηση 5.Β:

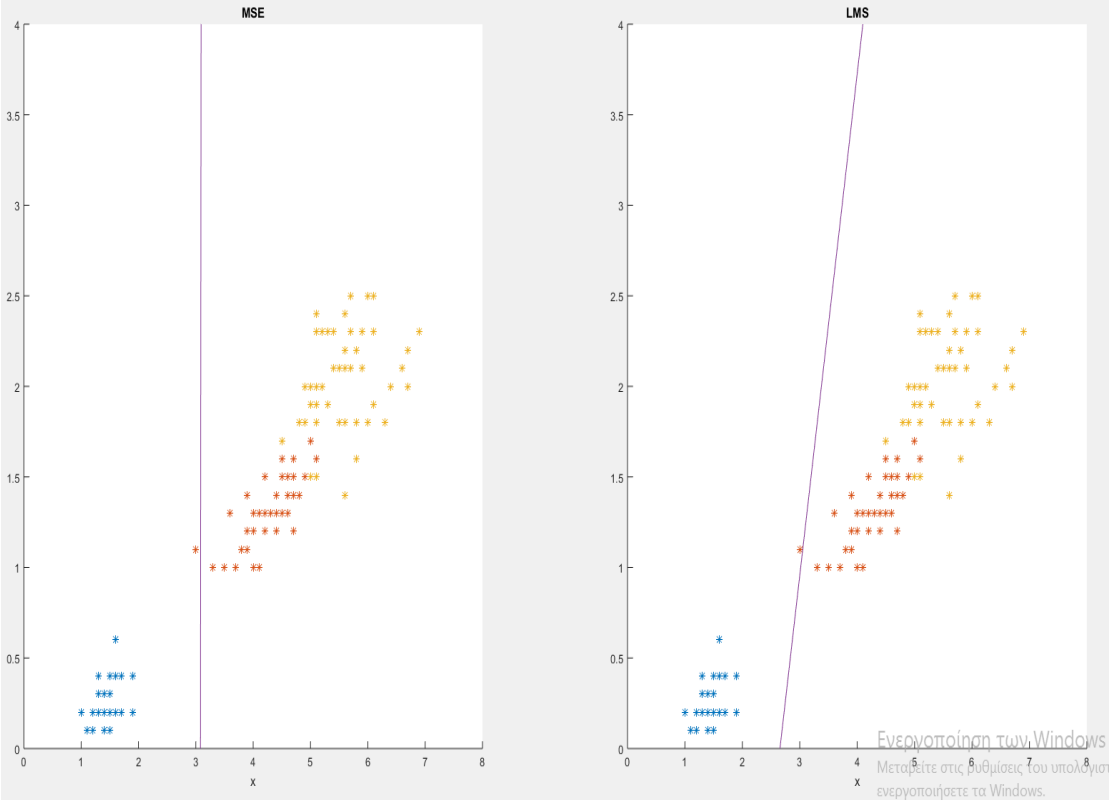
Η συνάρτηση MSE φτιάχνει σαν πρόβλημα γραμμικού συστήματος και υπολογίζει το διάνυσμα των βαρών σύμφωνα με τη σχέση :

$$\boldsymbol{a} = (\boldsymbol{Y}'\boldsymbol{Y})^{-1}\boldsymbol{Y}^{t}\boldsymbol{b}$$

Η MSE δέχεται τον πίνακα με τα χαρακτηριστικά που θέλουμε να ταξινομήσουμε και το διάνυσμα άσσων b. Και εκτελεί την παραπάνω πράξη.

Όσον αφορά την LMS θέλουμε τα βάρη να συγκλίνου στην $\boldsymbol{2Y'(Ya - b) * step.}$ Η διαδικασία αυτή πραγματοποιείται εως ότου η τιμή σύγκλισης γίνει μικρότερη από το κατώφλι που έχουμε ορίσει.

Τρέχοντας τον αλγόριθμο exB παίρνουμε τα εξής αποτελέσματα :



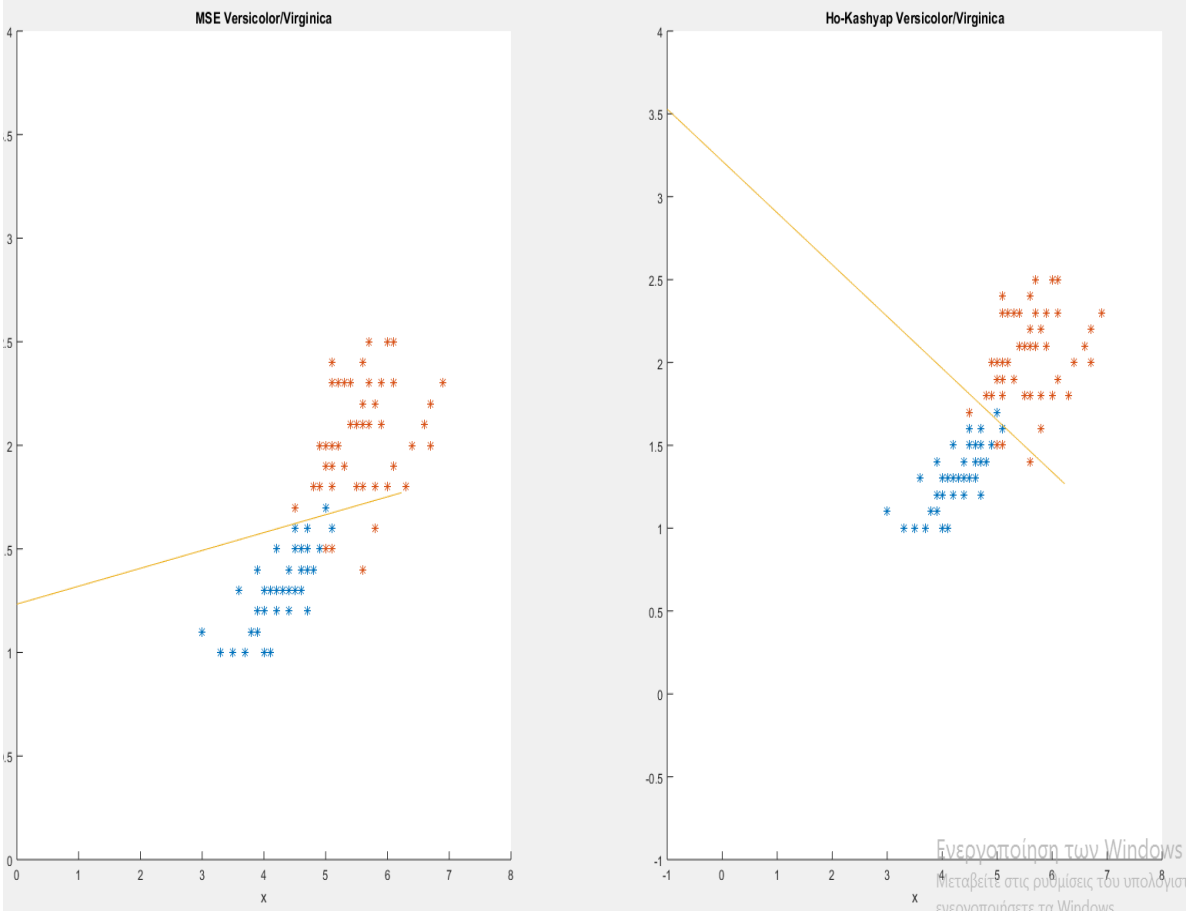
Βλέπουμε ότι ο γραμμικός διαχωρισμός των 2 κλάσεων δουλεύει αρκετά αποτελεσματικά με αυτές τις 2 μεθόδους με ένα πολύ μικρό σφάλμα.

Απάντηση 5.C:

Σε αυτό το ερώτημα εφαρμόζουμε ουσιαστικά την MSE για μη γραμμικά διαχωρίσιμες κλάσεις. Στην ουσία αλλάζουμε κατάλληλα τα στοιχεία που εισάγουμε στην MSE για να ταξινομήσει τα στοιχεία των κλάσεων virginica και versicolor.

Όσον αφορά την μέθοδο Ho-Kashyap στοχεύει στην ελαχιστοποίηση του $||\boldsymbol{Ya}-\boldsymbol{b}||^2$ . . Σε κάθε επανάληψη υπολογίζουμε το **Ya-b** και από αυτό υπολογίζουμε το $\boldsymbol{e^t}=\frac{\boldsymbol{e}+|\boldsymbol{e}|}{\boldsymbol{2}}$ στη συνέχεια τα βάρη μας υπολογιστούν από τη σχέση $\boldsymbol{w}=\boldsymbol{w}+\boldsymbol{2\eta(k)e^t}$

Τρέχοντας τον αλγόριθμο exC παίρνουμε τα εξής αποτελέσματα :



Παρατηρούμε ότι παρόλο που οι κλάσεις είναι μη γραμμικά διαχωρίσιμες οι ταξινομητές πετυχαίνουν ένα αρκετά καλό αποτέλεσμα με με 6% σφάλμα και 5%σφάλμα.

Απάντηση 5.D:

Σε αυτό το ερώτημα εφαρμόζουμε 3 φορές τον διαχωρισμό της μία κλάσης από τις άλλες 2 ώς προς και τις 4 διαστάσεις. Για να το κάνουμε αυτό θα πρέπει να αλλάξουμε κατάλληλα τις εισόδους των συναρτήσεων, δηλαδή να βάλουμε και τα 4 είδη δεδομένων . Επίσης σε αυτό το ερώτημα θα εμφανίσουμε απλά τα διανύσματα διάκρισης καθώς δεν μπορούμε να σχεδιάσουμε διάγραμμα με 4 διαστάσεις.

Τρέχοντας τον αλγόριθμο παίρνουμε τα διανύσματα διάκρισης που εμφανίζονται δεξιά:

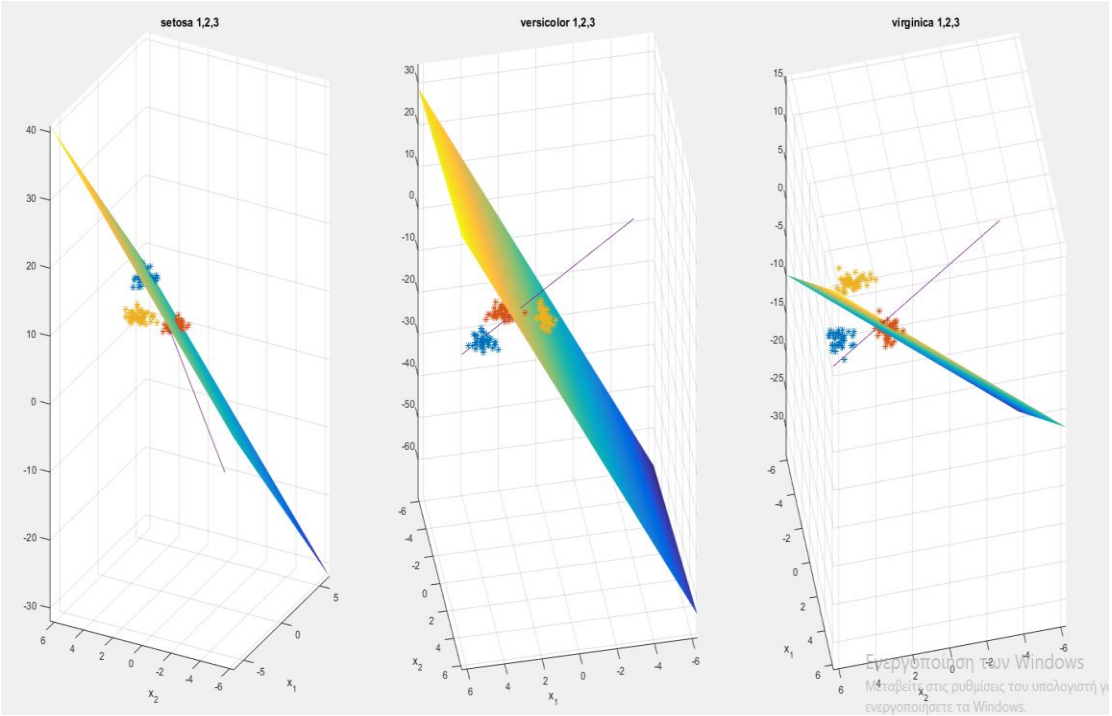
```
a_setosa =

   -0.7816
    0.1428
    0.4784
   -0.4544
   -0.1168
```

```
a_versicolor =

    2.1930
   -0.0659
   -0.8760
    0.4575
   -0.9901
```

```
a_virginica =

   -2.4114
   -0.0769
    0.3976
   -0.0031
    1.1069
```

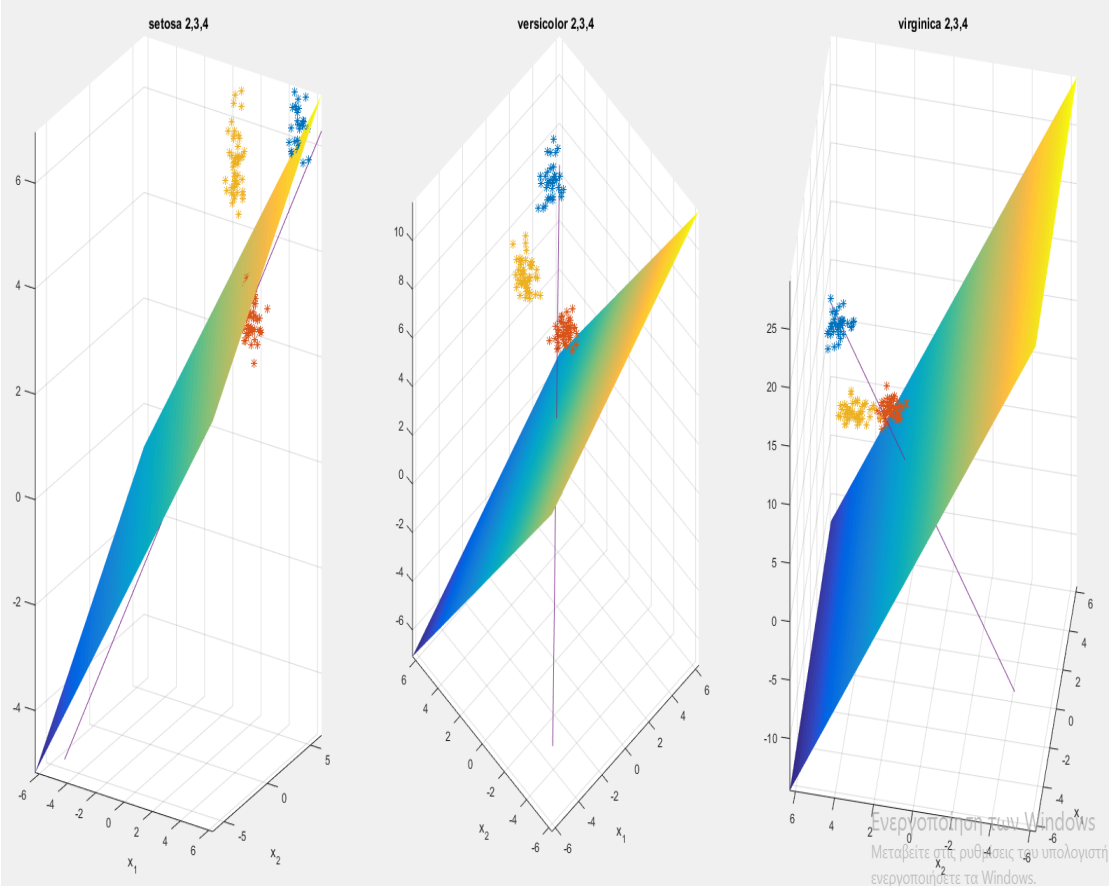
Απάντηση 5.E:

Η διαδικασία που θα ακολουθήσουμε θα είναι παρόμοια με αυτή που ακολουθήσαμε στο Δ. πλέον όμως έχουμε 3 χαρακτηριστικά ως προς τα οποία πρέπει να γίνει ο διαχωρισμός. Ως προς τις πρώτες 3 κλάσεις παίρνουμε τα εξής αποτελέσματα :



Παρατηρούμε ότι η επιφάνεια διαχωρισμού χωρίζει την μία κλάση από τις υπόλοιπες, ωστόσο δέχεται και στοιχεία άλλων κλάσεων στον ημιχώρο που θα έπρεπε να υπάρχει μόνο μία κλάση

Ως προς τις κλάσεις 2,3,4 παίρνουμε τα εξής αποτελέσματα :



Η ταξινόμηση λειτουργεί αρκετά αποτελεσματικά για την setosa ο διαχωρισμός ως προς τις άλλες 2 κλάσεις δεν δίνει καλά αποτελέσματα

Απάντηση 5.F:

Σε αυτόν το ερώτημα τρέχουμε τον αλγόριθμο kesler και παίρνουμε σαν αποτέλεσμα ότι το σφάλμα είναι 33% .

```
errKESLER =

    0.3333
```

Ωστόσο δεν μπορώ να επιστρέψω το διάγραμμα γιατί μου πετάει σφάλμα στο linspace ότι οι εσωτερικές διαστάσεις των πινάκων δεν είναι ίδιες ενώ είναι ίδιες και ενώ δοκίμασα πολλά πράγματα δεν βρήκα επίλυση, αλλά δεν υπήρχε πρόβλημα εξαρχής αφού οι διαστάσεις για τον πολλαπλασιασμό τον πινάκων είναι σωστές.