ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

[THL 311]- Στατιστική Μοντελοποίηση & Αναγνώριση Προτύπων

Αναφορά 1ης εργαστηριακής άσκησης

Διδάσκων Θεωρίας:
Ζερβάκης Μιχαήλ

Υπεύθυνος Εργαστηρίου:
Διακολουκάς Βασίλειος

Ιωάννης Περίδης
A.M. 2018030069

1 Μαίου 2022

Περιεχόμενα & Οδηγίες Χρήσης:

Στην αναφορά αυτή, θα γίνει παρουσίαση και επεξήγηση των ασκήσεων 1,3,4 και 5 και των αποτελεσμάτων τους. Οι ασκήσεις αυτές χρειάστηκαν να επιλυθούν στην MATLAB. Ο κώδικας που υλοποιήθηκε έχει δωθεί σε 4 διαφορετικά αρχεία με ονόματα exercise1_1/3/4/5. Κάθε τέτοιο αρχείο έχει μέσα πολλές συναρτήσεις και μόνο ένα εκτελέσιμο συνολικό αρχείο.m , με ονόματα ex1_1_pca για την άσκηση 1, ex1_3_Ida για την 3,ex1_4a και ex1_4e δύο εκτελέσιμα για την 4 και exercise1_5 για την 5.

Ο κώδικας είναι εμπλουτισμένος με σχόλια σε όλα τα σημεία για καλύτερη κατανόηση και μεγαλύτερη ευκολία ανάγνωσης. Για να τρέξετε τον κώδικα, πρέπει να παρακολουθείτε συνεχώς την κονσόλα για να βλέπετε τα αριθμητικά αποτελέσματα και τις γραφικές παρστάσεις στα figures. Μετά από τα αποτελέσματα του κάθε ερωτήματος η ροή του προγράμματος σταματάει με το pause και το πρόγραμμα περιμένει να πατήσετε ένα πλήκτρο για να συνεχίσει στα επόμενα.

Οι υπόλοιπες ασκήσεις 2,4,6 και 7 (η 4 χρειάστηκε και τα δύο), επιλύθηκαν με το χέρι στο χαρτί με μαθηματικό τρόπο και έχουν παραδωθεί σκαναρισμένες σε μορφή pdf σε ξεχωριστό φάκελο.

Θέμα 1: Principal Component Analysis (PCA):

Μέρος 1:

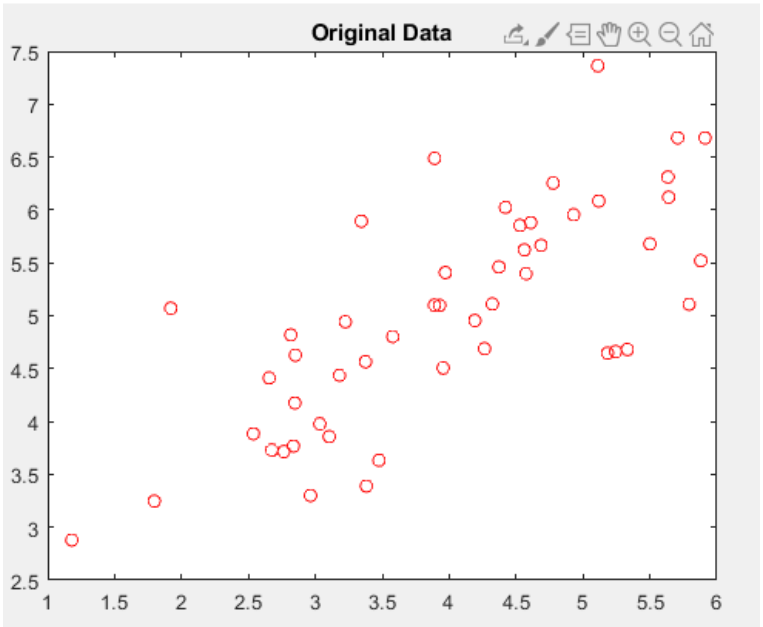
Στην 1η άσκηση, κληθήκαμε να εφαρμόσουμε την μέθοδο Principal Component Analysis σε τυχαία δεδομένα, με σκοπό να μειώσουμε τις διαστάσεις τους.

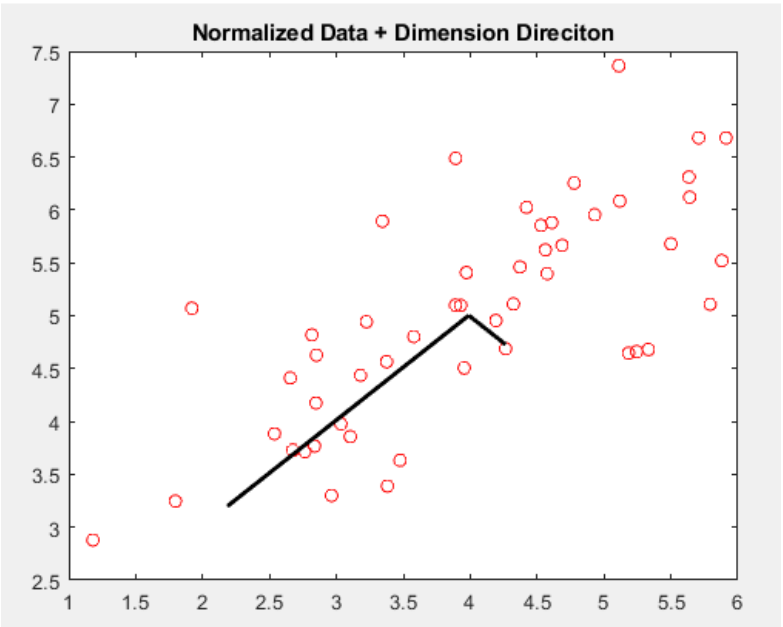
Αρχικά, στο 1ο μέρος μας δόθηκαν δεδομένα 2D για να γίνει πιο εύκολα κατανοητή η λειτουργία του PCA.

Ο αλγόριθμος που υλοποιήθηκε σε βήματα είναι ο εξής:

1. Φορτώνω το dataset που περιέχει τα δεδομένα μου (δίνεται έτοιμο).
2. Εφαρμόζω κανονικοποίηση (standardization) στο dataset, διότι ο PCA αντιμετωπίζει προβλήματα όταν υπάρχουν πολύ μεγάλες διαφορές μεταξύ των τιμών των χαρακτηριστικών.Προβλήματα λόγω κλίμακας μετρήσεων όπως να μην αλλάζει σε καμία περίπτωση η κατεύθυνση της μέγιστης διασποράς.Το παραπάνω επιτεύχθηκε με χρήση της συνάρτησης featureNormalize που επιστρέφει ένα κανονικοποιημένο διάνυσμα χ με μέση τιμή μ=0 και διασπορά σ=1.
3. Υπολογίζω τον πίνακα συνδιασποράς (covariance) και τα ιδιοδιανύσματα (eigenvectors) και τις ιδιοτιμές (eigenvalues) του παραπάνω πίνακα, μέσω της συνάρτησης myPCA.
4. Ταξινομώ με φθίνουσα σειρά όλες τις ιδιοτιμές που υπολόγισα στο προηγούμενο βήμα.
5. Κρατάω τις K πρώτες ιδιοτιμές που μου χρειάζονται, με K να είναι ο αιρθμός της δεδομένης διάστασής μου.
6. Προβάλω τα δεδομένα μου στις K διαστάσεις με χρήση της συνάρτησης projectData και ύστερα τα απεικονίζω. Για να γίνει η απεικόνιση πρέπει πρώτα να γίνει ανακατασκευή των δεδομένων , με τις ιδιοτιμές που κρατήθηκαν, αυτό γίνεται μέσω της συνάρτησης recoverData.

Παρακάτω φαίνονται οι γραφικές παραστάσεις των αρχικών δεδομένων και των δεδομένων μετά την κανονικοποίηση , μαζί με την κατεύθυνση της μέγιστης διασποράς:

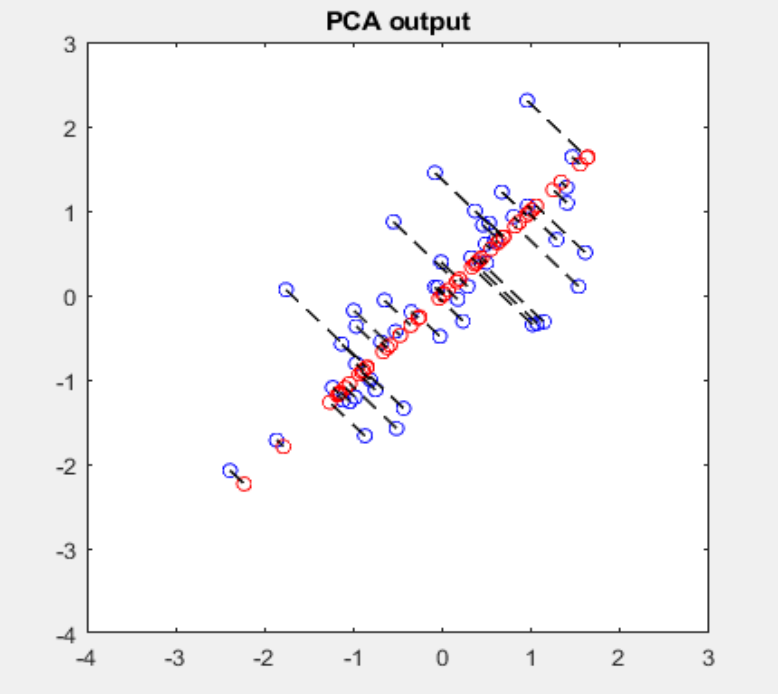




Το ποσοστό της ολικής διασποράς που έχιε η κάθε κατεύθυνση είναι:

$$PC1_{VARIANCE} = 86,776\% \quad και \quad PC2_{VARIANCE} = 13,223\%$$

Παρακάτω φαίνεται το αποτέλεσμα εξόδου του πρώτου PCA:



Παρατηρείται πως όντως το πρώτο PC μας δίνει την κατεύθυνση την οποία τα δεδομένα έχουν την μεγαλύτερη διασπορά. Ακόμη, καταφέραμε να πάμε από τον 2D χώρο στον 1D, χωρίς να χάσουμε κάποιο μεγάλο ποσοστό της διασποράς που είχαμε αρχικά.

Μέρος 2:

Στο 2ο μέρος, μας δίνεται ένας μεγάλος όγκος δεδομένων απο εικόνες προσώπων σαν είσοδος.Ακολουθείται ξανά η ίδια ανάλυση PCA με προιγουμένος.

Παρακάτω φαίνεται το γράφιμα των πρώτων 36 PCs (eigenfaces), από τον πίνακα που πήραμε ο οποίος ήταν μεγέθους 1024 PCs:

Παρατηρείται πως ακόμα και τα eigenfaces μόνα τους αρκούν για να σχηματιστούν εικόνες που διαφέρουν ανάμεταξύ τους σε κάποιον βαθμό. Αυτό γίνεται καθώς και τα ιδιοδιανύσματα παρέχουν πληροφορία όπως είναι στην περίπτωσή μας ο φωτισμός (σκίαση), η οποία αποτελεί την βάση της κάθε πραγματικής εικόνας. Για να δημιουργηθούν φυσικά οι πραγματικές εικόνες, πρέπει να χρησιμοποιηθεί ένας γραμμικός συνδιασμός των eigenfaces.

Παρακάτω φαίνοται τα αποτελέσματα των αρχικών δεδομένων (εικόνων) σε σχέση με αυτές που ανακτήθηκαν:

Σημαντικό στο σημείο αυτό είναι να αναλυθεί το ποσοστό της συνολικής διασποράς σε σχέση με τον αριθμο των PCs για διάφορες τιμές.Δίνεται ο πίνακας των αποτελεσμάτων παρακάτω.

K	**%Variance**
50	86,79
100	93,19
150	95,86
200	97,30

Από τον πίνακα γίνεται αντιληπτό πως με έναν πολύ μικρό αριθμό PCs ίσο με 50 , καταφέρνουμε να διατηρείσουμε ένα μεγάλο ποσοστό ίσο με 86,79% της αρχικής πληροφορίας μας. Παρόλα αυτά, δεν είναι αρκτετό για να ανακτήσουμε τα δεδομένα μας με αξιοπιστία , όμως είναι ένα δείγμα της μεγάλης αποδοτικότητας της ανάλυσης PCA στο συγκεκριμένο πρόβλημα. Συνεχίζοντας, αυξάνοντας μονάχα στο 200 από τα συνολικά 1024 στοιχεία , διατηρείται ένα αποδεκτό (αρκετά αξιόπιστο) ποσοστό της τάξεως του 97% και άνω. Αυτό έχει ως αποτέλεσμα ένας αλγόριθμος τεχνητής μάθησης που θα το χρησιμοποιούσε να λειτουργεί εξαιρετικά πιο γρήγορα ,χωρίς φυσικά μεγάλο σφάλμα στην αποδοτικότητά του.
Τέλος, είναι φανερό πως ο PCA είναι τόσο πολύ αποδοτικός διότι στην περίπτωσή μας, δεν είναι αναγκαία η διατήρηση των χαρακτηρηστικών τα οποία διαφοροποιούν τις κλάσεις μεταξύ τους, παρά μόνο μας νοιάζει η μεγιστοποίηση της διασποράς των δεδομένων εισόδου.

Θέμα 3: Linear Discriminant Analysis (LDA) vs PCA:

Μέρος 1:

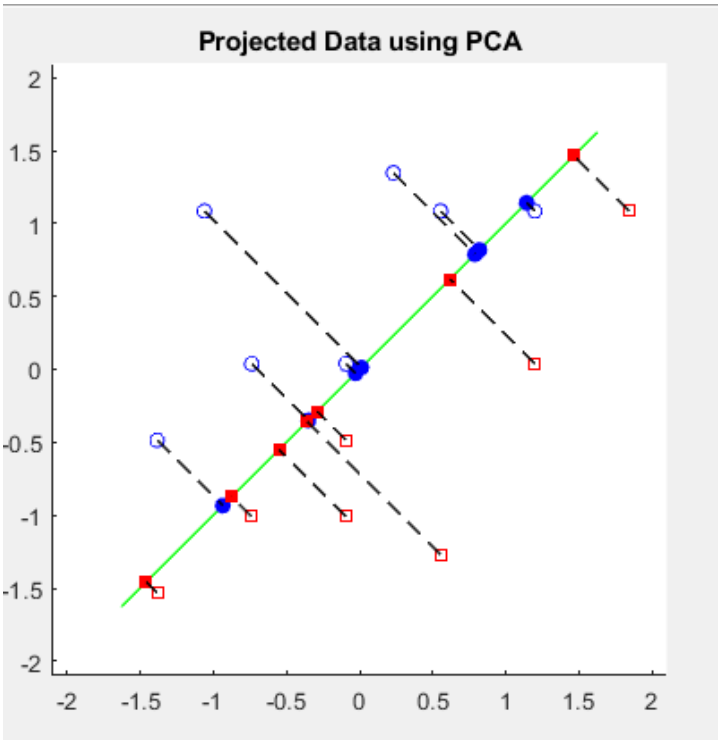
Στην άσκηση αυτή ξαναεφαρμόζουμε την τεχνική της μείωσης των διαστάσεων ενός feature vector, αλλά με διαφορετική μέθοδο, την supervised τεχνική LDA.
Στο 1ο μέρος δίνονται τεχνητά 2D δεδομένα 2 διαφορετικών κλάσεων και συγκρίνονται οι τεχνικές LDA και PCA.

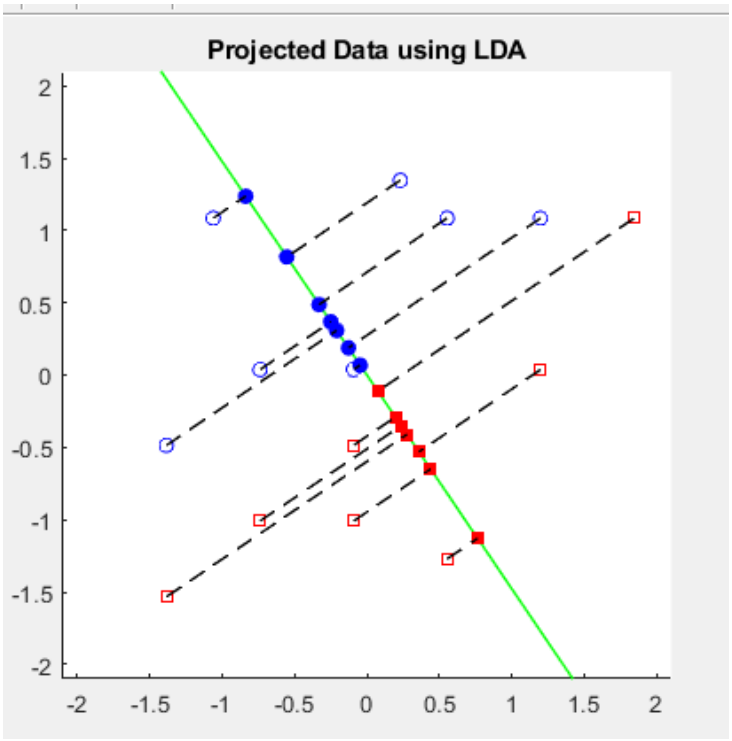
Ο αλγόριθμος που υλοποιήθηκε σε βήματα είναι ο εξής:

1. Φορτώνω το dataset που περιέχει τα δεδομένα (δίνεται έτοιμο).
2. Εφαρμόζω κανονικοποίηση αντίστοιχα όπως και πριν.
3. Υπολογίζω το μέσο (mean) και τον πίνακα συνδυασποράς.
4. Υπολογίζω τα within-scatter matrix με τον τύπο: $Sw = \frac{1}{2}(\Sigma 1 + \Sigma 2)$.
5. Υπολογίζω διάνυσμα προβολής με οτν τύπο: $w = Sw^{-1}(\mu 1 - \mu 2)$.
6. Προβάλω τα δεδομένα με την συνάρτηση projectDataLDA και τα επανακατασκευάζω με την recoverDataLDA.

Για τα βήματα 3 έως 5 χρησιμοποιώ την συνάρτηση fisherLinearDiscriminant.

Παρακάτω φαίνοται τα γραφήματα από τα προβαλόμενα δεδομένα πάνω στις ευθείες του PCA και οτυ LDA:





Είναι φανερό πως οι δύο αναλύσεις διαφέρουν σημαντικά καθώς και έχουν σχηματίσει διαφορετικά τις ευθείες για την προβολή των δεδομένων.

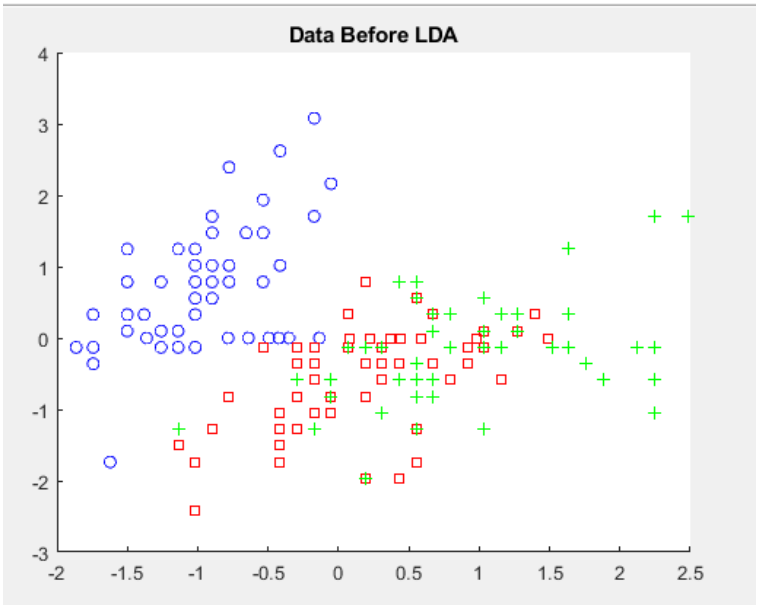
Ξεκινώντας με τον PCA , είναι εύκολο να αντιληφθούμε πως δεν διαχωρίζονται τα δεδομένα των δύο κλάσεων, επομένως και δεν θα μπορούσε να χρησιμοποιηθεί σαν classifier επιτυχώς. Αυτό συμβαίνει διότι γνωρίζουμε πως ο PCA δεν έχει πληροφορία για τα class labels και η λειτουργία του στηρίζεται μόνο στην κατεύθυνση που μεγιστοποιείται η διασπορά των δεδομένων εισόδου.

Αντιθέτως, ο LDA , γνωρίζει και χρησιμοποιεί τα class labels. Σαν αποτέλεσμα αυτού, βρίσκει μια κατεύθυνση της ευθείας τέτοια ώστε τα μέσα (means) των κλάσεων να έχουν την μέγιστη απόσταση μεταξύ τους, αλλά ταυτόχρονα οι within –class διασπορές να είναι ελάχιστες. Από το σχήμα φαίνεται λοιπόν πως πετυχαίνει το επιθυμητό αποτέλεσμα του διαχωρισμού των δεδομένων των κλάσεων και θα μπορούσε άρα να χρησιμοποιηθεί σαν classifier.

Μέρος 2:

Στο μέρος αυτό, παίρνουμε τα δεδομένα μας από το Iris dataset, το οποίο έχει 3 κλάσεις δεδομένων, επομένως θα υλοποιηθεί η γενική μορφή αυτή τη φορά της τεχνικής LDA. Η υλοποίηση ήταν παρόμοια με πρίν με ορισμένες αλλαγές και προσθήκη της εύρεσης του global mean και του between-class scatter matrix. Για την διαδικασία χρησιμοποιήθηκε η συνάρτηση myLDA.

Παρακάτω φαίνοται τα γραφήματα από τα δεδομένα πριν και μετά την εφαρμογή του LDA:



Παρατηρείται από τα γραφήματα πως υπάρχει μια πολύ καλή διαχωρισμότητα μετά τον LDA. Αυτό, συμβαίνει διότι μετά από την μείωση της διάστασης, η ανάλυση αυτή δίνει σαν αποτέλεσμα και τις κατευθύνσεις οι οποίες έχουν την μέγιστη διαχωρισιμότητα.

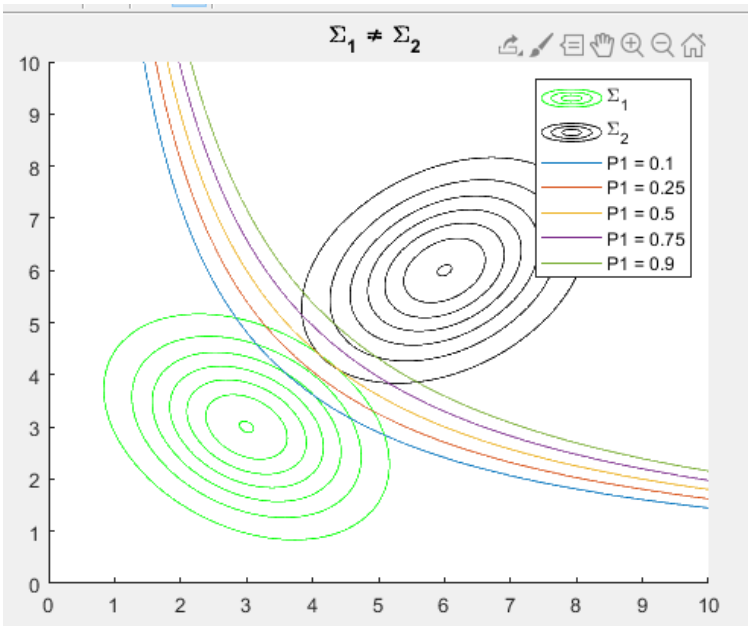
Τελικά, καταλαβαίνουμε πως και οι δύο αλγόριθμοι μπορούν να χρησιμοποιηθούν για μείωση διαστάσεων ενός dataset, με διαφορετικούς τρόπους ο καθένας.Παρόλα αυτά, ο LDA μπορεί να μετασχηματίζει τα δεδομένα , να τα διαχωρίζει , επομένως μπορεί να χρησιμοποιηθεί και σαν classifier. Δεν μπορούμε να ορίσουμε κάποιον από τους δύο καλύτερο από τον άλλο, απλά πρέπει να επιλέγουμε πότε θα τους προτιμάμε και πότε όχι , ανάλογα με τις ιδιότητες των δεδομένων μας.

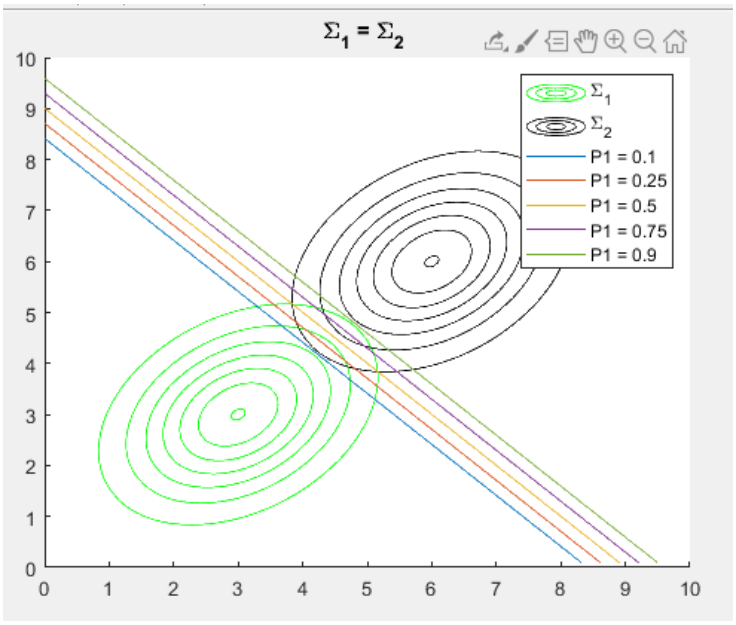
Θέμα 4: Bayes:

Στο θέμα αυτό, φαίνονται τα ερωτήματα b,c,d τα οποία και χρειάστηκαν την υλοποίηση τους στην MATLAB. Τα υπόλοιπα ερωτήματα a και e που σχετίζονται με τον υπολογισμό του

οριού απόφασης,έχουν λυθεί στο χαρτί και βρίσκονται στις σκαναρισμένες χειρόγραφες λύσεις.

Παρακάτω φαίνοται τα αποτελέσματα για Σ1 ίσο και διάφορο από το Σ2:





Αρχίζοντας απο την περίπτωση που $\Sigma 1 \neq \Sigma 2$, παρατηρούμε πως τα όρια των αποφάσεων είναι καμπύλες. Οι καμπύλες , αλλάζουν ανάλογα με τα scatter matrices που έχουμε, συνήθως είναι της μορφής υπερβολής ή παραβολής. Αυτό συμβαίνει λόγω του τετραγωνικού όρου που έχει το εκθετικό της pdf. Στην περίπτωσή μας, τα όρια είναι καμπυλωτά διότι ο όρος της pdf $(x-\mu i)^T \Sigma^{-1}(x-\mu i)$,περιέχει όρους της μορφής $ax1x2$.

Συνεχίζοντας, με την περίπτωση όπου $\Sigma 1 = \Sigma 2$, παρατηρούμε πως τα όρια των αποφάσεων είναι ευθείες. Αυτό συμβαίνει διότι δεν υπάρχουν στον όρο της pdf, ούτε όροι της μορφής $ax1x2$, αλλά ούτε και τετραγωνικοί όροι.

Το μέγεθος των περιοχών απόφασης και στις δύο περιπτώσεις εξαρτάται από το πόσο συχνά αποφασίζουμε να επιλέξουμε κάθε κλάση. Αυτό σημαίνει πως όσο αυξάνεται η πιθανότητα a-priori της μιας κλάσης,τόσο μετακινείται το όριο πιο κοντά στην άλλη κλάση.

Θέμα 5: Εξαγωγή χαρακτηριστικών και Bayes Classification:

Στην άσκηση αυτή, μας δίνεται το mnist database και καλούμαστε να ασχοληθούμε με την εξαγωγή χαρακτηρηστικών, κάνοντας classification στα ψηφία 1 και 2.

Αρχικά, συλλέχθηκε ο λόγος μήκος/ύψος του κάθε μη μηδενικού pixel της κάθε εικόνας. Ο λόγος αυτός είναι το λεγόμενο aspect ratio.Υποθέτοντας πως τα δείγματα των aspect ratios είναι δείγματα προερχόμενα από κανονική κατανομή , με sample των δειγμάτων εκπαίδευσης μ και σ, έγινε το classification. Για το classification, χρειάστηκε να δημιουργηθούν οι πιθανότητες a-priori και χρησιμοποιήθηκε ο κανόνας του Bayes.
Να σημειωθεί πως το ποσοστό λανθασμένης απόφασηςτου classifier έχει τιμή 10,9368%.

Παρακάτω φαίνοται τα αποτελέσματα τα ψηφία 1 και 2, με τα ορθογώνια τα οποία ορίζουν τα aspect ratios τους: