Αναφορά 1ου σετ ασκήσεων στο μάθημα Στατιστική Μοντελοποίηση και Αναγνώριση Προτύπων

Λεωνίδας Μπακοπουλος
2018030036
2/5/22

θέμα πρώτο

Στην πρώτη άσκηση, ζητήθηκε η υλοποίηση της PCA ανάλυσης, για την μείωση αρχικά των διαστάσεων από 2-d δεδομένα σε 1-d, και έπειτα από 1024 διαστάσεις σε σημαντικά λιγότερες(ενδεικτικά 50-300)

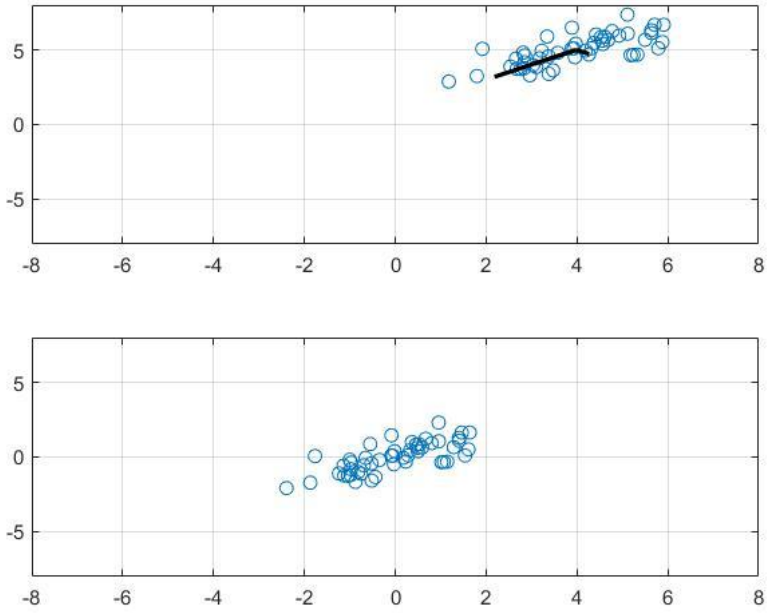
Αλγόριθμος PCA

Αρχικά, αξίζει να σημειωθεί ότι ο σκοπός του αλγορίθμου, είναι να βρεί ένα επίπεδο(εν προκειμένω μια ευθεία), το οποίο ελαχιστοποιεί το άθροισμα των αποστάσεων των δεδομένων από αυτό, και κατα συνέπεια (βάσει του πυθαγορείου θεωρήματος) μεγιστοποιεί την διασπορά των δεδομένων όταν αυτά προβάλλονται σε αυτό.

Βήματα αλγορίθμου

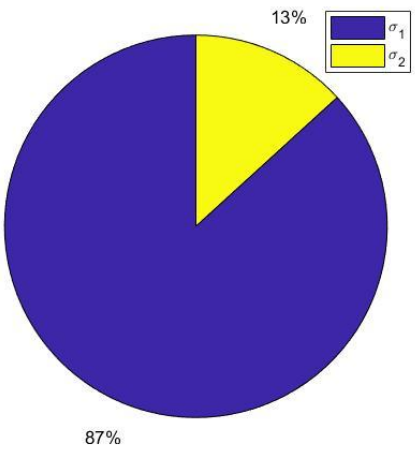
Step 0. Κανονικοποίηση δεδομένων, δηλ απεικόνιση των δεδομένων με “κέντρο” το σημείο [0,0] (featureNormalize.m)
Step 1. Δημιουργία του πίνακα συνδιασποράς των κανονικοποιημένων δεδομένων, και υπολογισμός των ιδιοδιανυσμάτων και των ιδιοτιμών (myPCA.m)
Step 2. Επιλογή των κ ιδιονυσμάτων με βάση της κ μεγαλύτερες ιδιοτιμές (εν προκειμένω κ=1 ή κ=50-300). (myPCA.m)
Step 3. Προβολή των δεδομένων στα προαναφερόμενα διανύσματα. (projectData.m)

Μέρος πρώτο



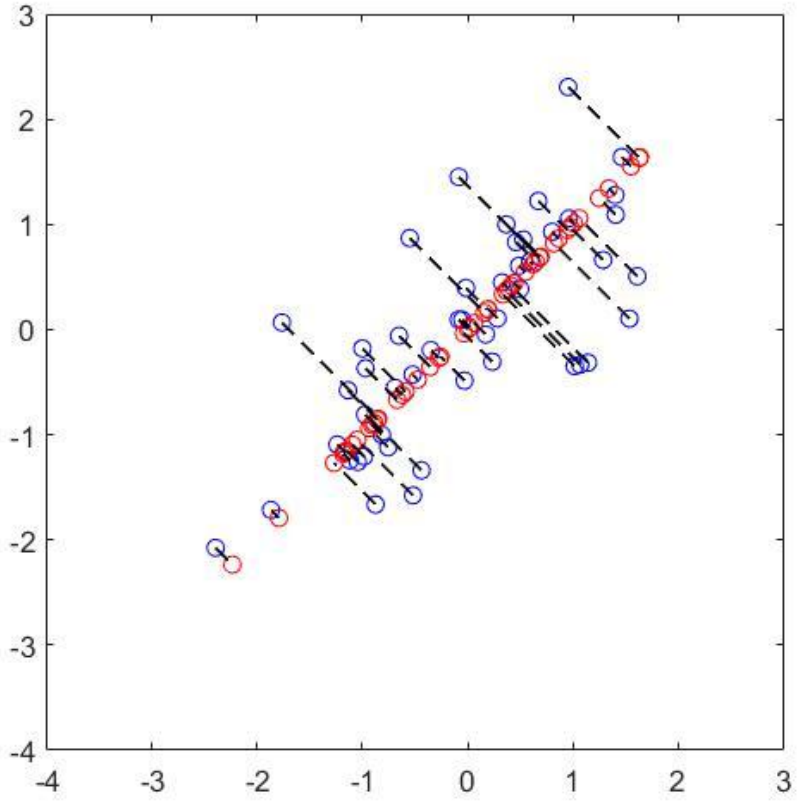
Διάγραμμα 1

Στο διάγραμμα 1, διακρίνονται τα δεδομένα όπως αυτά φορτώθηκαν μέσω του αρχείου ex1_1_data1.mat, αρχικά χωρίς την κανονικοποίηση. Επίσης στο πρώτο διάγραμμα φαίνονται και τα ιδιοδιανύσματα που προέκυψαν από τον πίνακα συνδιασποράς. Αυτά ορίζουν τους άξονες που θα προβληθούν τα δεδομένα. Διαπιστώνεται, ότι αν χρησιμοποιηθεί ο “οριζόντιος” μαύρος άξονας, η διασπορά των δεδομένων θα είναι μεγαλύτερη. Στο δεύτερο διάγραμμα φαίνονται τα κανονικοποιημένα πλέον δεδομένα.



Διάγραμμα 2

Στο διάγραμμα 2, φαίνεται το ποσοστό της διαφοροποιησης των τιμών από τα αρχικά δεδομένα, όταν αυτά προβληθούν στον πρώτο και στο δεύτερο ιδιοδυανισμα αντίστοιχα.Από τα παραπάνω, είναι εμφανές ότι το “οριζόντιο” ιδιοδιάνυσμα παρέχει μεγαλύτερη διαφοροποίηση στα δεδομένα απ ότι το άλλο.



Διάγραμμα 3

Όπως αναφέρθηκε και προηγουμένως, επιλέχθηκε το οριζόντιο διάνυσμα για την προβολή των δεδομένων. Πλέον τα δεδομένα, βρίσκονται σε μία διάσταση, χωρίς όμως να έχουν χαθεί σε μεγάλο βαθμό, οι σχετικές τους αποστάσεις.
Σε περίπτωση που τα δεδομένα είχαν μεγαλύτερη διακύμανση στον y’y αξονα απ’οτι στον x’x, ο αλγόριθμος θα είχε επιλέξει το κατακόρυφο ιδιοδιάνυσμα.

Μέρος δεύτερο

Στό δεύτερο μέρος της άσκησης, υλοποιήθηκε ο ίδιος αλγόριθμος, με δεδομένα 1024 διαστάσεων και μειωση στις 10,50,100,200,300 διαστάσεις.

Διάγραμμα 4

Στο παραπάνω διάγραμμα, διακρίνονται τα πρώτα 36 πρόσωπα, όταν οι διαστάσεις των δεδομένων, έχουν μειωθεί στις 150. Παρατηρείται ότι τα πρόσωπα έχουν διατηρήσει κάποια χαρακτηριστικά τους και το σημαντικότερο είναι, ότι τα πρόσωπα μεταξύ τους φαίνονται διαφορετικά.

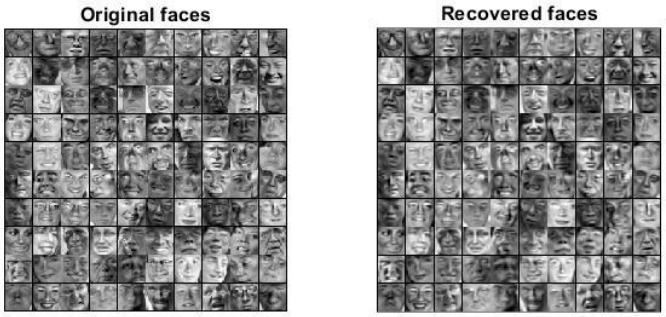
K	% of variance
10	67.27
50	86.79
100	93.19
200	97.31
300	98.73

Πίνακας 1

Στον πίνακα 1, απεικονίζονται τα ποσοστά της διακύμανσης των δεδομένων, όταν η μείωση των διαστάσεων φτάνει τα 10, 50, 100 κτλ. Αξίζει να σημειωθεί ότι με σχεδόν το 1% των διαστάσεων, η διακριτικότητα των δεδομένων είναι αρκετά υψηλή (≃ 67%), κάτι το οποίο αποδεικνύει την αποτελεσματικότητα τους αλγορίθμου.

-0.3054	-0.1614	0.0490	0.1353	0.1120	0.0853	0.3597	0.1016	-0.1776	-0.2948	-0.2557	-0.0151	0.1429	0.2563	0.0497
-0.1418	0.0275	0.1071	0.1291	0.0653	-0.0176	-0.0770	-0.0405	0.3725	0.0787	-0.0874	-0.0336	0.0145	-0.0260	0.0198
-0.0423	0.0317	0.2193	0.0233	-0.2583	-0.3020	0.0654	0.6087	-0.0095	-0.2732	-0.2307	-0.1007	0.0138	-0.0225	-0.0280
0.4688	0.0708	-0.1640	-0.0995	0.0326	0.0544	-0.0692	-0.1970	-0.2931	-0.2937	0.1158	0.3262	-0.0222	0.0571	0.1240
-0.0953	-0.0842	0.0027	-0.0620	-0.0921	-0.0519	-0.0152	-0.0098	-0.0858	-0.1045	-0.1310	-0.1887	-0.2057	-0.1709	-0.1245
0.0521	0.0842	-0.0287	-0.1648	-0.2578	-0.1680	-0.0947	0.6202	0.5684	0.2387	-0.3314	-0.5209	-0.3327	-0.1107	0.0642
0.3507	0.0127	-0.1276	-0.0149	-0.0255	-0.0272	-0.0933	-0.1186	0.0036	-0.0311	-0.0856	-0.0661	0.0292	0.1638	0.1963
-0.5426	-0.3645	-0.1376	0.1037	0.2512	0.2551	0.0395	0.0623	0.2890	0.1795	0.0875	0.0016	0.0084	-0.0120	0.0063
-0.3196	-0.1554	0.2531	0.6354	0.4954	-0.0117	-0.2741	-0.5330	-0.5026	-0.3890	-0.1357	-0.0056	0.1859	0.2023	0.1358
0.0655	0.0492	0.0217	-0.0106	-0.0821	-0.1693	-0.1762	-0.1066	-0.0813	-0.0603	-0.0669	-0.0653	-0.0269	-0.0306	-0.0715
-0.0181	0.0598	-0.0291	-0.1503	-0.1459	-0.0169	0.0728	-0.0830	-0.1823	-0.1250	-0.0993	-0.2190	-0.2381	-0.1563	-0.0928
0.0217	0.0213	0.0085	0.0844	-0.0695	-0.0900	-0.0275	0.0758	0.1871	0.1563	0.0498	0.0244	0.0673	0.0360	0.0076
-0.2925	-0.3602	0.2873	0.8338	0.1706	-0.3017	-0.0115	0.0381	0.2627	0.2213	0.3471	0.3571	0.2471	0.1295	0.2624
-0.0431	0.0896	0.2892	0.1424	0.0412	-0.1606	-0.2400	-0.0634	0.1784	0.1276	0.2098	0.0927	-0.1137	-0.1047	-0.1123
-0.0286	-0.0987	-0.2445	-0.3615	-0.3235	-0.1451	0.3185	0.3259	0.2779	-0.0363	-0.0799	0.0780	0.2537	0.0256	-0.0165
0.0226	-0.0501	-0.1453	-0.0440	0.0259	0.0753	-0.0137	-0.0020	0.1348	0.1455	-0.0233	-0.1488	-0.2543	-0.1070	0.0059
-0.2783	0.0884	0.0504	0.1073	-0.2760	-0.5318	-0.6262	0.5875	0.4841	-0.6581	-0.5847	-0.0601	0.5403	0.6387	0.3317
0.0816	0.0855	0.1210	0.2471	0.1639	-0.1975	-0.2478	-0.0666	0.0477	0.1181	0.1444	0.2062	0.1248	-0.0417	-0.1181
-0.1461	-0.0996	0.0731	0.0706	0.0042	-0.0584	0.0536	0.3811	0.3306	-0.1507	-0.2706	-0.1847	-0.1412	-0.0366	0.1080
0.0074	-0.1585	-0.1573	-0.1316	-0.0949	-0.1530	-0.2106	-0.2062	-0.0326	0.2000	-0.0155	-0.2508	0.1487	0.4750	0.5967
0.4292	0.2181	0.0921	-0.1585	-0.2999	-0.2302	-0.1021	0.0430	0.2192	0.1802	0.0429	-0.1003	-0.0601	-0.0127	0.0594
-0.0916	-0.0184	-0.0216	-0.0163	-0.0900	0.0138	0.0925	0.0526	0.0513	-0.0369	-0.0922	-0.1638	-0.0929	-0.0517	-0.0980
0.2642	0.2593	0.1949	0.0201	-0.1890	-0.3809	-0.3399	-0.2259	-0.0630	0.1171	0.1621	0.1610	0.1166	-0.0967	-0.2362

Τέλος, αξίζει να σημειωθεί οτι ο επανυπολογισμός των αρχικών δεδομένων, είναι θεωρητικά εφικτός αν τα δεδομένα προβληθούν ξανά σε όλα τα eigenvectors. Στην πράξη, όπως φαίνεται και στο παραπάνω πίνακα (ο οποίος απεικονίζει τη διαφορά x-x_recover), υπάρχει μικρή διαφορά ανάμεσα στα αρχικά και στα επαναπολογισμένα δεδομένα, η οποία δεν αποτυπώνεται εν προκειμένω σε σοβαρές διαφορές στα πρόσωπα οπως φαίνεται στο παρακάτω διάγραμμα, αλλά μπορεί εν γενει να αποτελέσει πρόβλημα.



Θέμα τρίτο

Για την ολοκλήρωση του τρίτου θέματος, ζητήθηκε η υλοποίηση και εφαρμογή του LDA και η εφαρμογή PCA στα ίδια σετ δεδομένων (ένα σετ εξάσκησης 2 διαστάσεων και το iris data set 4 διαστάσεων), καθώς και η σύγκριση των δύο αυτών αλγορίθμων.

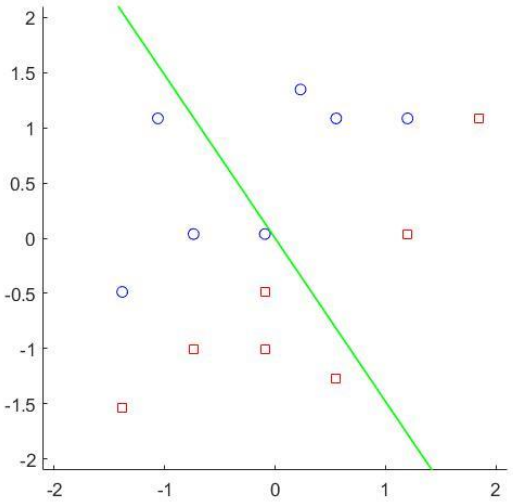
- Υλοποίηση του LDA

Step 0. Κανονικοποίηση των δεδομένων γύρω από το μηδέν.
Step 1. Υπολογισμός των πινάκων συνδιασποράς (και για τα δεδομένα της κάθε κλάσης αλλά και μεταξύ των N κλάσεων, όταν N≥3)
Step 2. Υπολογισμός του διανύσματος προβολής .
Step 3. Προβολή πάνω στο προηγούμενο διάνυσμα ,ή (υπερ)επίπεδο αν $N_{\text{μειωμενο}} > 1$

- Σκοπός του LDA

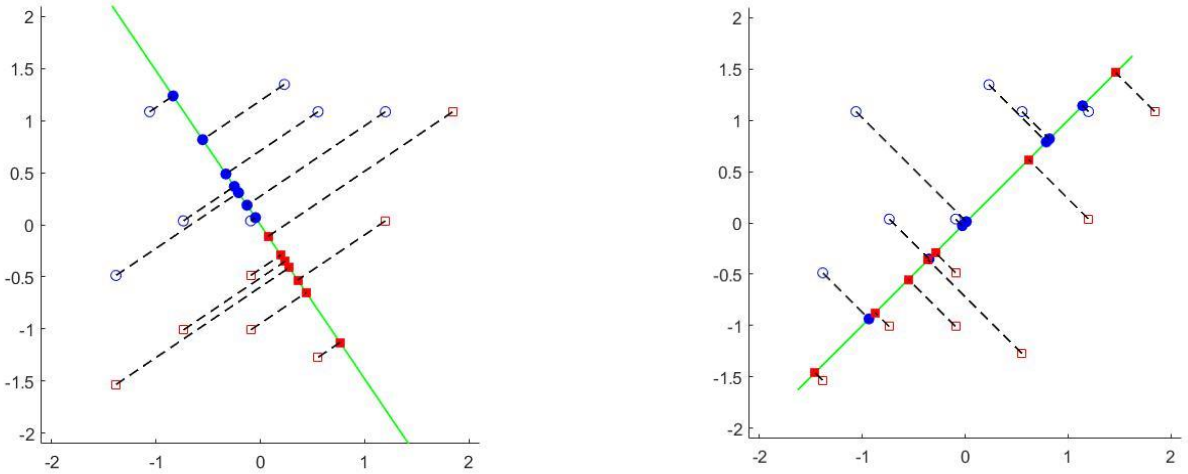
Η μείωση των διαστάσεων των δεδομένων κάποιων μεταβλητων (που ανήκουν σε συγκεκριμένες κλάσεις), προσπαθόντας τη μεγιστή διακρισιμότητα μεταξύ των κλάσεων (άρα επιχειρώντας προβολή με τρόπο ώστε να επιτευχθεί μεγάλη απόσταση μεταξύ των μέσων τιμών και μικρή διακύμανση των δεδομένων της κάθε κλάσης)

Μέρος πρώτο



Διάγραμμα 5

Στο παραπάνω διάγραμμα, απεικονίζονται τα δεδομένα, καθώς και το διάνυσμα που αποφάσισε ο αλγόριθμος LDA, ο οποίος ακολούθησε τα προαναφερόμενα βήματα.

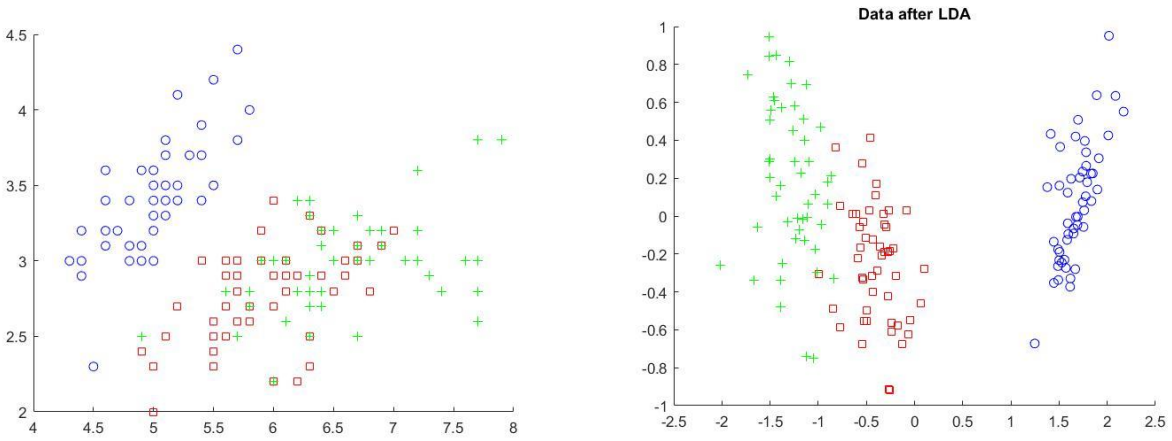


Διάγραμμα 6

Στα παραπάνω διαγράμματα, απεικονίζονται οι προβολές των πειραματικών δεδομένων, όταν σε αυτά εφαρμόστηκε ο LDA (αριστερά) και ο PCA (δεξιά).

Παρατηρείται απόλυτη διακρισιμότητα μεταξύ των κλάσεων σε περίπτωση του LDA, κάτι το οποίο αποτελεί χαρακτηριστικό του LDA, ενώ στα δεξιά παρατηρείται, μεγάλη διασπορά των δεδομένων καθ'όλο το μήκος του πράσινου διανύσματος.

Μέρος δεύτερο



Διάγραμμα 7

Στό δεύτερο μέρος της άσκησης, ζητήθηκε η εφαρμογή του lda, στο τεσσάρων διαστάσεων iris dataset. Αριστερά, απεικονίζονται οι πρώτες δύο διαστάσεις των 150 δειγμάτων, και δεξιά απεικονίζονται τα δείγματα όταν οι διαστάσεις τους έχουν μειωθεί στις δύο, μέσω του αλγορίθμου LDA. Παρατηρείται, ότι η μείωση των διαστάσεων, επέφερε ικανή διαχωρισιμότητα μεταξύ των τριών κλάσεων, κάτι το οποίο ήταν αδύνατο, επιλέγοντας "χειροκινητα" τα πρώτα δύο χαρακτηριστικά ή χρησιμοποιώντας τον αλγόριθμο PCA, κάτι το οποίο αποδείχθηκε στο πρώτο μέρος της άσκησης.

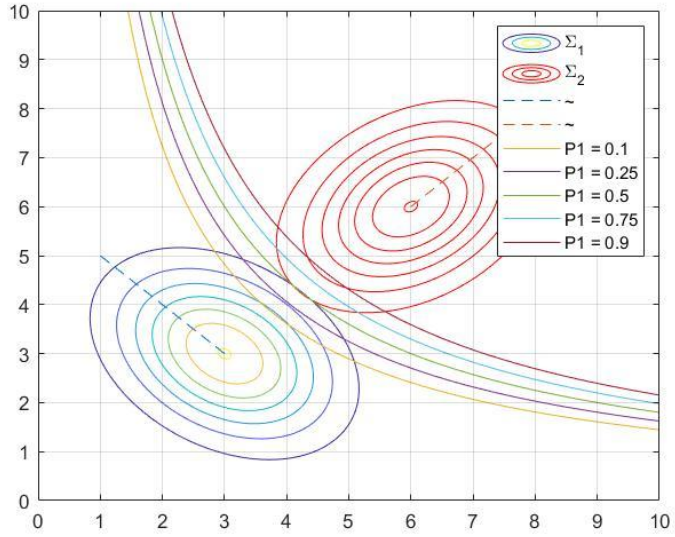
Συμπεράσματα

Αν και οι δύο αλγόριθμοι -PCA και LDA- χρησιμοποιούνται για μείωση των διαστάσεων ενός σετ δεδομένων, ο κάθε ένας έχει διαφορετικά κριτήρια. Πιο συγκεκριμένα, ο PCA προσπαθεί να αυξήσει τη διασπορά μεταξύ των δεδομένων-"απλώνοντάς" τα έτσι καλύτερα στον χώρο.Αντίθετα, ο LDA, προσπαθεί να μειώσει την διακύμανση των δεδομένων της κάθε κλάσης και να αυξήσει τη απόσταση των μέσων τιμών (της κάθε κλάσης), δηλαδή να αυξήσει την απόσταση μεταξύ των κλάσεων, επιτυγχάνοντας έτσι μέγιστη διαχωρισιμότητα.

Θέμα τέταρτο

Ερώτημα c,d)

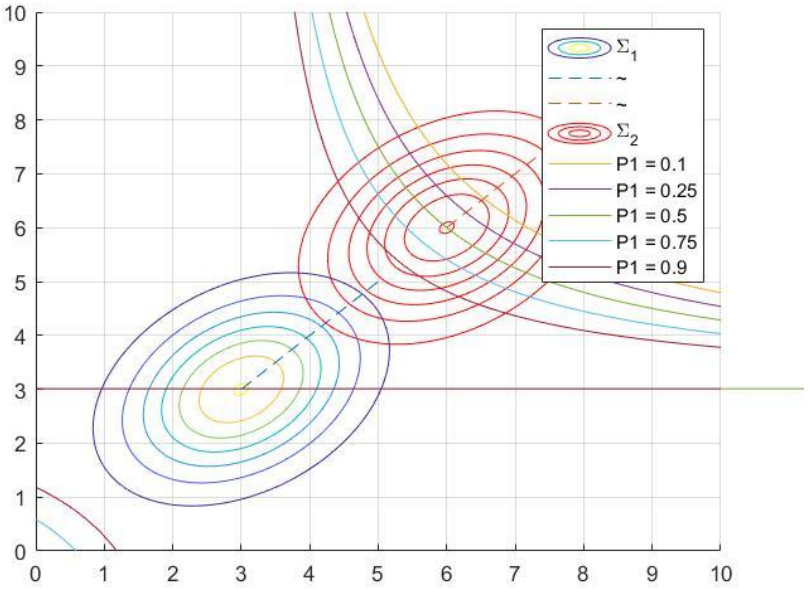
Στην τέταρτη άσκηση, ζητήθηκε η εύρεση των ορίων απόφασης, σε ένα πρόβλημα κατηγοριοποίησης δύο κλάσεων, αποτελούμενων από δύο μεταβλητές. Αφού υπολογίστηκε το όριο απόφασης συναρτήσει του λόγου πιθανοτήτων (οι πράξεις αναλυτικά στις χειρόγραφες απαντήσεις), σχεδιάστηκαν οι ισοϋψείς καμπύλες και τα όρια απόφασης. Παρατηρούμε αρχικά για τις ισουψείς καμπύλες ότι έχουν ορθά ως σημείο μέγιστης πιθανότητα την μέση τιμή (δηλ το σημειο [3,3] για την πολύχρωμη και [6,6] για την κόκκινη). Επίσης παρατηρείται, ότι η κατεύθυνσή των δύο κατανομών (διακεκομμένες γραμμές), βρίσκεται στην ευθεια $y = x$ και $y = -x$ όπως αναμενόταν από τον πίνακα συνδιασποράς. Εν συνεχεία, όπως αναλύεται και στο θέμα πέντε,η απόσταση μεταξύ κατανομής 1 από το σύνορο απόφασης και την πιθανότητα p_1 είναι ανάλογες, γιατί όσο πιθανότερη είναι apriori η περίπτωση να είσαχθεί στον classifier ένα στοιχείο της κλάσης 1, τόσο ευνοϊκότερα πρέπει ο classifier να αποφασίζει υπέρ της. Τέλος λόγω της μορφής του ορίου απόφασης ($x_2 = \frac{1}{x_1} + c$), το σχήμα των ορίων είναι ελλειπτικο.



Διάγραμμα 8

Ερώτημα e)

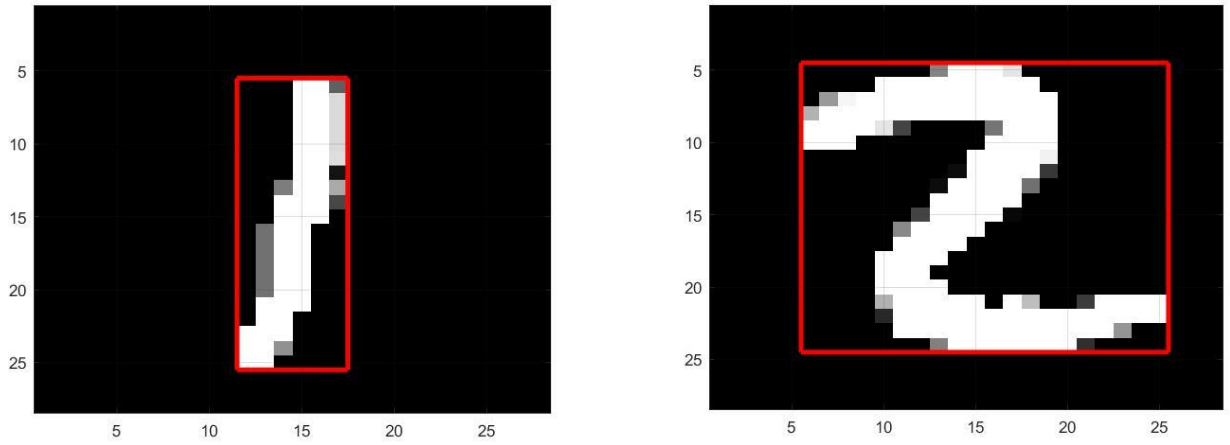
Στο παρακάτω διάγραμμα απεικονίζονται οι ισοϋψείς καμπύλες και ο περιοχές απόφασης για τις δύο κατανομές. Από τη στιγμή που οι δύο κατανομές είχαν ίδιο πίνακα συνδιασποράς αναμενόταν οι περιοχές απόφασης να ήταν ευθείες γραμμές περίπου στο μέσο των δύο κατανομών. Λόγω πιθανού σφάλματος στις πράξεις κάτι τέτοιο δεν ήταν εφικτό.



Διάγραμμα 9

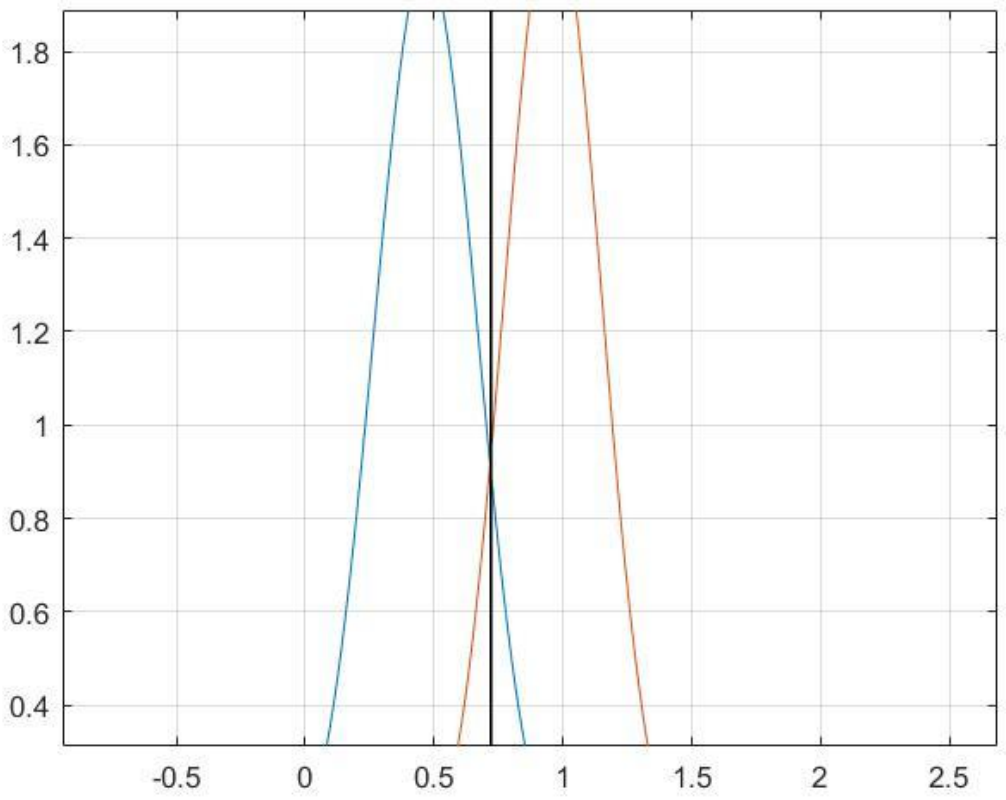
Θέμα πέμπτο

Για την επίλυση του πέμπτου θέματος, ζητήθηκε η δημιουργία ενός bayes classifier, ο οποίος θα διαχωρίζει χειρόγραφια ψηφίας (μόνο τους αριθμούς ένα και δύο), χρησιμοποιώντας ως τυχαία μεταβλητή το λόγο $aspect\ ratio = \frac{width}{height}$.



Διάγραμμα 8

Όπως μπορεί να παρατηρηθεί και από το διάγραμμα 8, το παραλληλόγραμμο που περικλείει τον αριθμό “ένα” είναι διαφορετικό από το αντίστοιχο του “δύο”. Πιο συγκεκριμένα, το aRatio για τα “ένα”, αναμένεται να είναι μικρότερο από τα αντίστοιχα του “δύο”.Βέβαια, υπήρξαν και χειρόγραφοι αριθμοί που ο συγκεκριμένος κανόνας δεν θα μπορούσε να αποτελέσει κριτήριο διαχωρισμού (γνωστά και ως corner cases). Χρησιμοποιώντας ως δεδομένα τα training data, και θεωρώντας ότι τα aRatio των δεδομένων ακολουθούν δύο κανονικές κατανομές, υπολογίστηκαν οι μέσες τιμές και η διακυμάνσεις και σχηματίστηκαν οι παρακάτω καμπύλες.

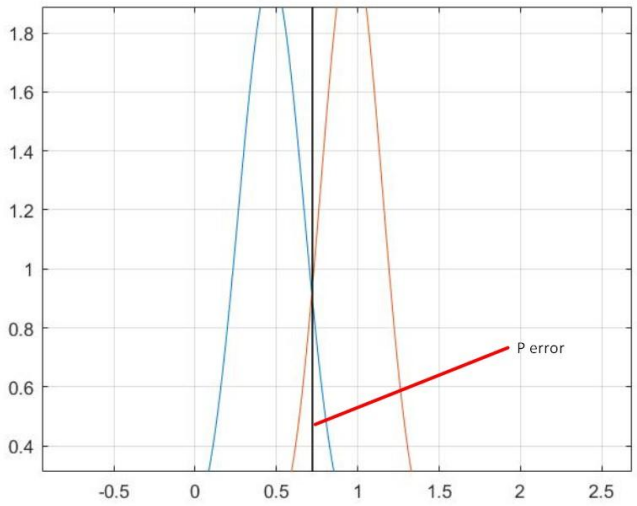


Διάγραμμα 9

Έχοντας υπολογίσει ότι η apriori πιθανότητα για να δοθεί στον classifier “η κλάση “ένα” είναι 0,53, υπολογίστηκε το decision bound ως το x για το οποίο ισχύει $p_1 p(x|\omega_1) = p_2 p(x|\omega_2)$ και σχεδιάστηκε στο διάγραμμα 9. Παρατηρείται ότι το οριο απόφασης είναι ελαφρώς πιο κοντά στην πορτοκαλί καμπύλη (pdf για τα χειρόγραφα “δύο”) . Το παραπάνω θα μπορούσε να δικαιολογηθεί και διαισθητικά.Πιο συγκεκριμένα, από το γεγονός ότι $p_1 > p_2$, η πιθανότητα να εισαχθεί ένα χειρόγραφο της κλάσης “ένα” στον classifier, είναι μεγαλύτερη και συνεπώς στα corner cases(δηλαδή όταν aRatio θα μπορούσε με παρόμοια πιθανότητα να άνηκε και στις δύο κλάσεις), θα έπρεπε να αποφασίζει ευνοικότερα για την κλάση “ένα”.

Τέλος, και αφού είχε υπολογιστεί το σημείο απόφασης, για κάθε καινούργιο δείγμα, υπολογίζονταν το aRatio και αναλόγως που βρισκόταν σε σχέση με το όριο, αποφασίζονταν αν ο χειρόγραφος αριθμός ανήκε στην κλάση “ένα” ή “δύο”. Με την παραπάνω μέθοδο, ο classifier λειτούργησε με ποσοστά λάθους 10.9368%.

Αξίζει να σημειωθεί, ότι θα μπορούσε να υπολογιστεί και θεωρητικά το ποσοστό σφάλματος του συγκεκριμένου classifier, βρίσκοντας το εμβαδόν στα σημεία που οι δύο καμπύλες επικαλύπτονται, δηλαδή την περιοχή όπου υποδεικνύει το διάγραμμα 10.



Διάγραμμα 10