Problem Set 1

Τρίμας Χρήστος
AM:2016030054
Στατιστική Μοντελοποιήση και Αναγνώριση Προτύπων
Πολυτεχνειο Κρητης

April 7, 2020

Ερώτηση 1

Principal Component Analysis(PCA).

Στην 1η εργασία του μαθήματος, χρησιμοποιήθηκε η μέθοδος PCA, για την μείωση των διαστάσεων των δεδομένων.
Στο 1ο μέρος, δοκιμάστηκε η μέθοδος σε ενα dataset, έτσι ώστε να γίνει εξοικείωση τόσο με την μέθοδο, όσο και με την χρήση του matlab για τέτοιου είδους προβλήματα. Αρχικά, έγινε load και απεικόνιση του dataset στον 2-D χώρο.

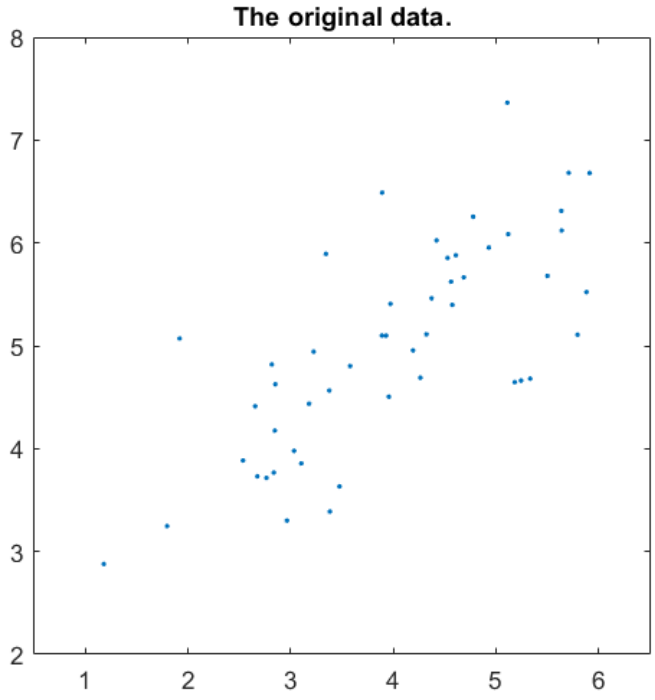

Figure 1: Visualized Dataset

Έπειτα, πραγματοποιήθηκε standarization, δηλαδή κανονικοποίηση με μέση τιμή μηδέν και διασπορά 1, στα αρχικά δείγματα.Στην εικόνα φαίνεται, η νέα "κατανομή" των δειγμάτων στον χώρο.

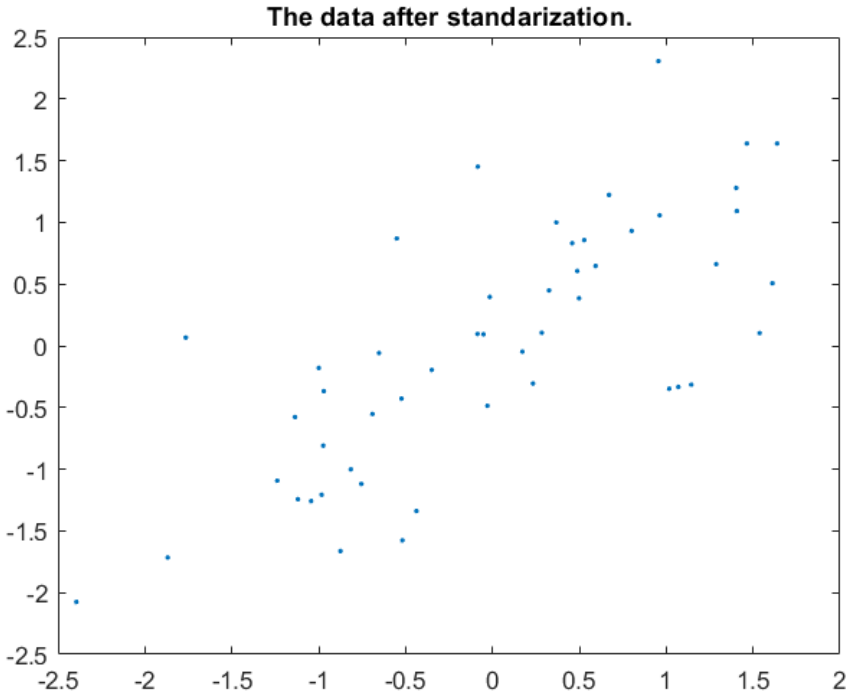

Figure 2: Dataset After Standarization

Στη συνέχεια, εφαρμόζεται ο PCA αλγόριθμος. Υπολογίζεται ο πίνακας συνδιασποράς των κανονικοποιημένων δεδομένων με την χρήση του τύπου $\Sigma = (1/m)X^T X$. Τέλος,επιστρέφουμε τους πίνακες των ιδιοδιανυσμάτων και των ιδιοτιμών, οι οποιοί υπολογίστηκαν με την χρήση της εντολής svd() του MatLab.

Κατά αυτόν τον τρόπο, έχει υπολογιστεί η συνεισφορά κάθε συνιστώσας στην συνολική διακύμανση και στη συνέχεια με προβολή και επαναφορά των δειγμάτων θα παρθεί το νέο αποτέλεσμα σε μειωμένη διάσταση (από 2-D σε 1-D).

Επιλογικά, τα βήματα που ακολουθήθηκαν για την υλοποιήση του αλγορίθμου είναι τα εξής:

1) Αφαίρεση της μέσης τιμής για την δημιουργία ενός dataset με μέση τιμή μηδέν.

2) Υπολογισμός του πίνακα συνδιασποράς.

3) Υπολογισμός των ιδιοδιανυσμάτων και των ιδοτιμών.

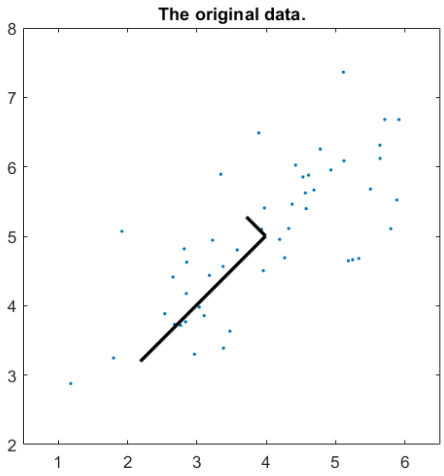

Figure 3: PCA in the original Samples

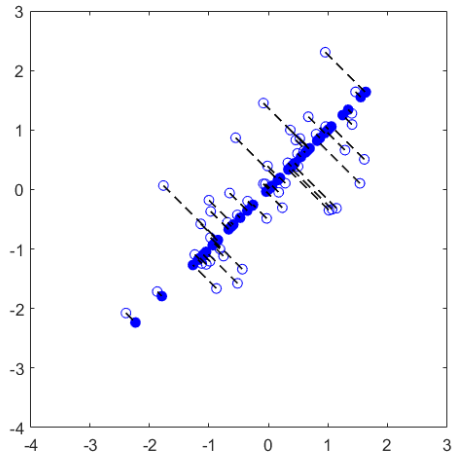

Figure 4: Reconstructed Dataset

4) Μείωση της διάστασης και δημιουργία του διανύσματος με την μεγαλύτερη ιδιοτιμή(Principal Component).

5) Δημιουργία "νέων" δεδομένων.

6) Επαναδημιουργία των αρχικών δεδομένων σε μειωμένη διάσταση.

Για το 2ο Μέρος της 1ης Άσκησης, δοκιμάστηκε ο αλγόριθμος PCA που υλοποιήθηκε, σε ένα πραγματικό dataset με εικόνες προσώπων.

Πιο συγκεκριμένα, αφού έγινε load και display των εικόνων, εφαρμόστηκε feature normalization και με την σειρά του ο PCA. Στην εικόνα No6, παρατηρούνται τα πρόσωπα των 36 πρώτων ιδιοδιανυσμάτων που βρέθηκαν. Από την απεικόνιση αυτή των δειγμάτων, παρατηρείται μια σημαντική απώλεια πληροφορίας, ή αλλιώς ένα ξεθόριασμα πάνω στις εικόνες.

Όπως και προηγουμένως, έτσι και τώρα, για να γίνει ορατή η μείωση των διαστάσεων, γίνεται προβολή των εικόνων στον "ιδιο-χώρο" με την χρήση των 10/50/100/500 ιδιοδιανυσμάτων.

Παρατηρείται, ότι όσο μεγαλώνει ο αριθμός των κύριων συνιστωσών, τόσο πιο καθαρά οπτικό είναι το αποτέλεσμα. Για K=10, το αποτέλεσμα δεν είναι καθόλου ξεκάθαρο, και επομένως η μείωση των διαστάσεων προκάλεσε απώλεια πληροφορίας. Όσο αυξάνεται λοιπόν το K, δηλαδή οι συνιστώσες, τόσο καλύτερο recover των εικόνων έχουμε, όπως φαίνεται και στις εικόνες παρακάτω.

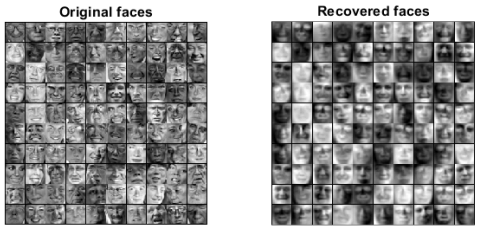

Figure 5: K=10

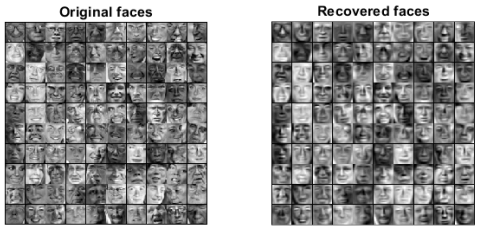

Figure 6: K=50

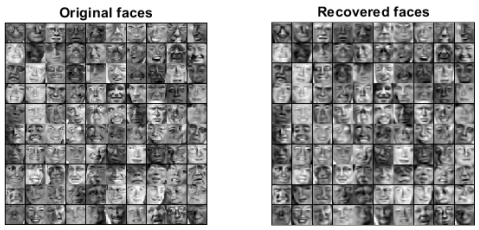

Figure 7: K=100

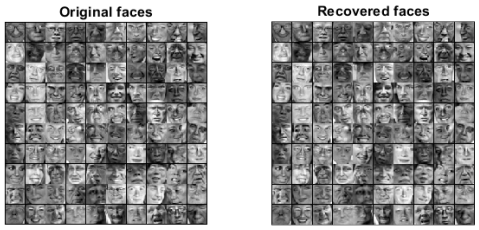

Figure 8: K=500

Ερώτηση 3

Linear Discriminant Analysis (LDA) vs (PCA).

Σκοπός της άσκησης, ήταν η εξοικείωση με τον LDA αλγόριθμο, η σύγκριση αυτού με τον PCA καθώς επίσης, και η εφαρμογή του multiclass LDA σε ενα πραγματικό dataset.

Στο 1ο μέρος, ο LDA ή Fisher's Linear Discriminant Analysis, εφαρμόστηκε σε ένα σύνολο από τεχνητά δεδομένα, τα οποία απεικονίζονται στην ακόλουθη εικόνα.

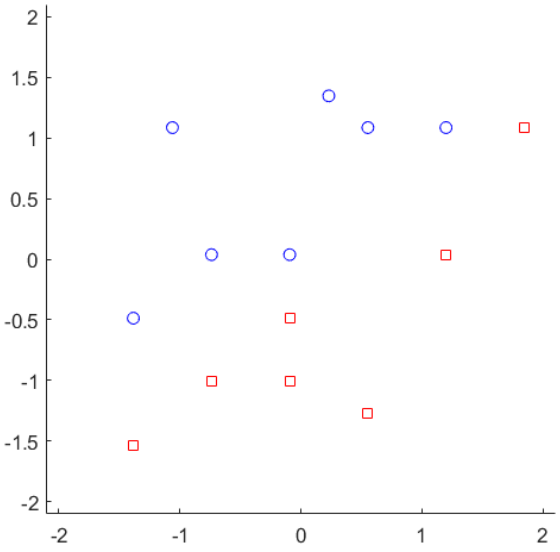

Figure 9: Visualization of the Samples.

Στη συνέχεια, εφαρμόζεται ο αλγόριθμος του Fisher, για τις δυο κλάσεις δειγμάτων, με σκοπό την μείωση των διαστάσεων, και την εύκολη διάκριση τους σε 2 κατηγορίες.

Αναλυτικά τα βήματα του Αλγορίθμου:

1) Υπολογισμός των μέσων κάθε κλάσης.

2) Υπολογισμός του within Scatter Matrix.

3) Εύρεση της βέλτιστης κατεύθυνσης για τον ξεχωρισμό των κλάσεων.

4) Normalization.

Όπως και στην άσκηση 1, έτσι και τώρα, γίνετε προβολή πάνω στο διάνυσμα που βρέθηκε από τον αλγόριθμο και επανακατασκευή των data πάνω στο διάνυσμα.

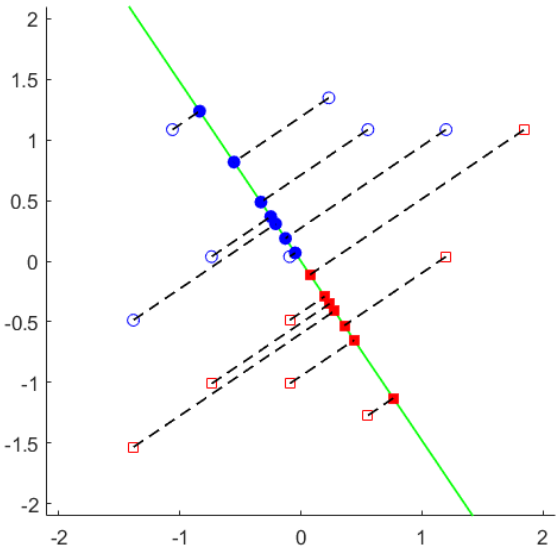

Figure 10: Reduced dimensions using LDA

Η διάκριση των δεδομένων σε κάθε κλάση, στην μειωμένη διάσταση, είναι αρκετά πιο "εύκολη" με την χρήση του LDA. Από την άλλη, με εφαρμογή PCA αλγορίθμου, η διάσταση μειώνεται, όμως η διάκριση των δεδομένων δεν είναι δυνατή.

Για το 2ο Μέρος της άσκησης, έγινε εφαρμογή του Multiclass LDA, πάνω στο IRIS dataset. Το dataset, περιλαμβάνει 50 δείγματα από 3 διαφορετικά είδη λουλουδιών. Σκοπός είναι η

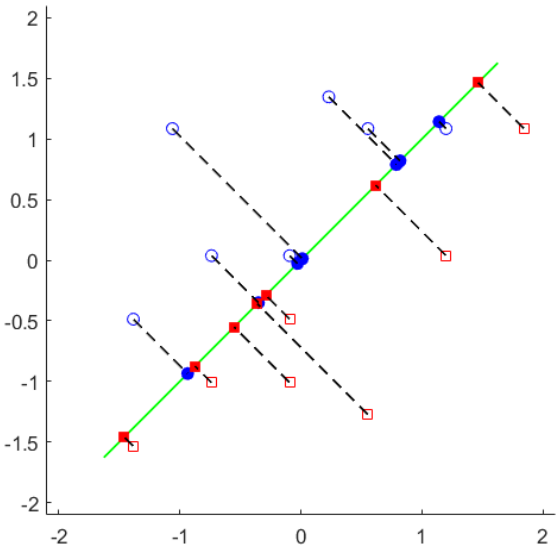

Figure 11: Reduced dimensions using PCA

μείωση των διαστάσεων όπως ακριβώ στο προηγούμενο μέρος.

Αφού έγινε load και display των δειγμάτων των 3ων κλάσεων, εφαρμόστηκε ο LDA αλγόριθμος.

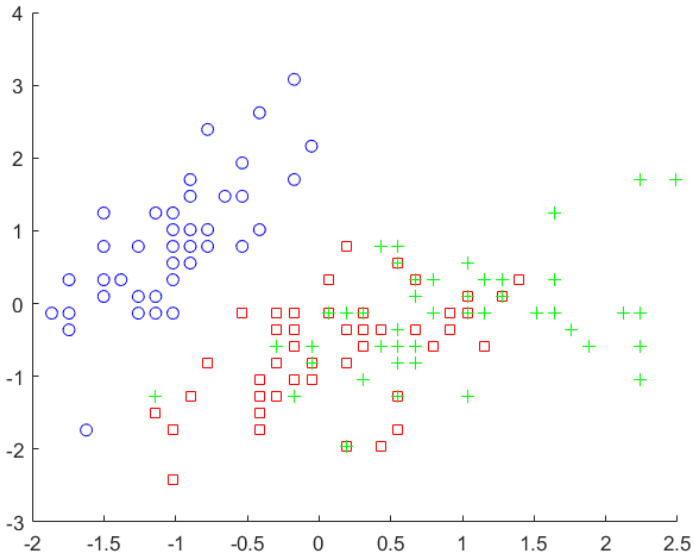

Figure 12: IRIS dataset

Multiclass LDA:

1) Εύρεση της μέσης τιμής για κάθε κλάση.

2) Υπολογισμός των a-priori πιθανοτήτων.

3) Υπολογισμός του within Scatter matrix κάθε κλάσης.

4) Υπολογισμός του ολικού μέσου.

5) Υπολογισμός του between class Scatter matrix.

6) Eigenvalue Decomposition.

7) Ταξινόμηση των ιδιοτιμών και επιστροφή του πίνακα μειωμένων διαστάσεων.

Τέλος, γίνεται προβολή στην κατεύθυνση του πίνακα μειωμένων διαστάσεων.

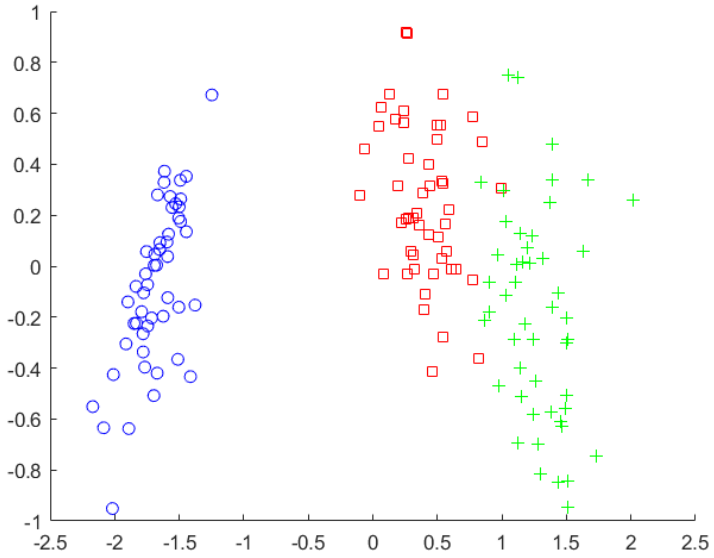

Figure 13: Every class after LDA applied.

Αρχικά η διάκριση των κλάσεων δεν ήταν καθόλου εύκολη. Με εφαρμογή του LDA, οι τρείς κλάσεις έγιναν διακριτές και η διάκριση τους πλέον είναι ευκολότερη.

Ερώτηση 4

Εξαγωγή χαρακτηριστικών και υπολογισμός των εκ των υστέρων πιθανοτήτων με τον κανόνα του Bayes.

Σε αυτή την εργασία, έγινε χρήση του mnist dataset. Σκοπός ήταν η δημιουργία ενός χώρου δειγμάτων και ο υπολογισμός κάποιον πιθανοτήτων, με την χρήση του Bayes theorem.

Αρχικά, υπολογίστηκε το aspect ratio κάθε εικόνας. Ως aspect ratio, ορίστηκε, από το ελάχιστο ορθογώνιο που περικλιεί έναν χειρόγραφο αριθμό στο mnist dataset, ο λόγος width/height. Στο συγκεκριμένο πρόβλημα, έγινε επεξεργασία μόνο των ψηφίων 1(κλάση 1) και 2(κλάση 2).

Αφού έγινε υπολογισμός του aspect ratio των δύο αυτών κλάσεων, δημιουργήθηκε ένα διάστημα τιμών, με εύρος $[minAspectRatio, maxAspectRatio]$, και από τις δύο κλάσεις. Έπειτα, το διάστημα αυτό, χωρίστηκε σε 3 ίσα υποδιαστήματα, και έγινε ταξινόμηση κάθε εικόνας σε ένα από αυτά τα L(ow), M(edium), H(igh), διστήματα, ανάλογα το aspect ratio της.

Αφού λοιπόν δημιουγήθηκε ο δειγματικός χώρος, υπολογίστηκαν οι εξείς πιθανότητες:

1) $P(C_1) = 0.1124$

2) $P(C_2) = 0.0993$

3) $P(L|C_1) = 0.0114$

4) $P(M|C_1) = 0.3908$

5) $P(H|C_1) = 0.5977$

6) $P(L|C_2) = 0$

7) $P(M|C_2) = 0$

8) $P(H|C_2) = 1$

9) $P(L) = 0.0013$

10) $P(M) = 0.0439$

11) $P(H) = 0.1665$

12) $P(C_1|L) = 1$

13) $P(C2|L) = 0$