ΠΟΛΥΤΕΧΝΕΙΟ
ΚΡΗΤΗΣ

1η σειρά ασκήσεων

Σωτήριος Μιχαήλ
2015030140

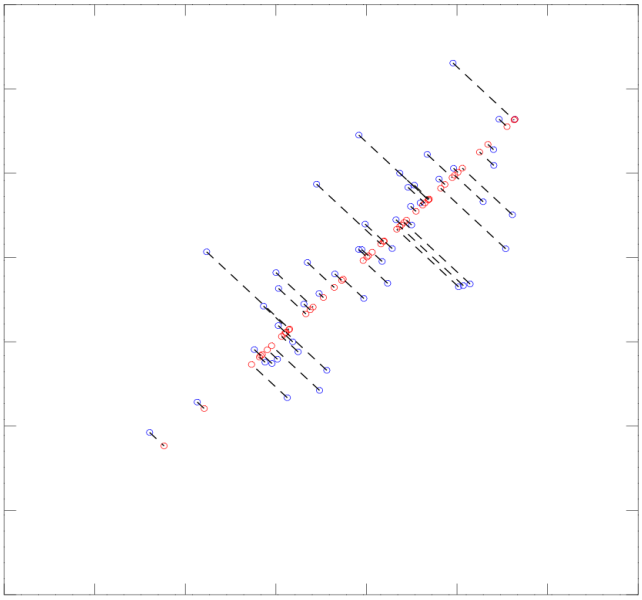
Στατιστική μοντελοποίηση και αναγνώριση προτύπων

Θέμα 1ο - Principal Component Analysis

Χρησιμοποιούμε τη μέθοδο Principal Component Analysis (PCA) για να μειώσουμε τις διαστάσεις δεδομένων, αρχικά δεδομένων δύο διαστάσεων κ στη συνέχεια, εφαρμόζουμε την ίδια μέθοδο σε ένα μεγαλύτερο σύνολο δεδομένων 5000 προσώπων.

Καθώς η μέθοδος PCA υπολειτουργεί όταν υπάρχει μεγάλη διαφορά μεταξύ των τιμών των χαρακτηριστικών, κρίθηκε απαραίτητη η κανονικοποίηση των δεδομένων τα οποία επεξεργάσθησαν.

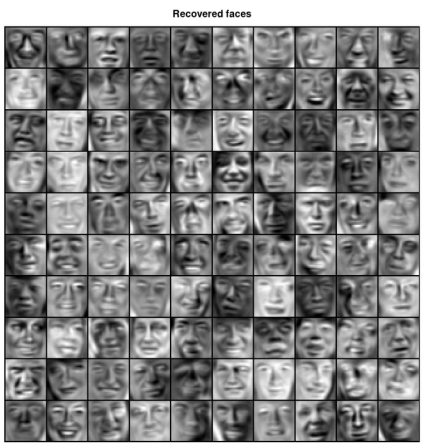
Στη συνέχεια, εφαρμόσθηκε η μέθοδος PCA, με την εύρεση των κατευθύνσεων της μέγιστης διασποράς καθώς κ το ποσοστό της ολικής διασποράς το οποίο έχει η κάθε κατεύθυνση. Παρατίθεται το αποτέλεσμα ης ανάκτησης των δεδομένων μέσω της μεθόδου:



***Εικόνα 1**: Κανονικοποιημένα δεδομένα στο πρώτο PC*

Είναι εμφανές πως το πρώτο principal component μας δείχνει την κατεύθυνση της μεγαλύτερης διασποράς των δεδομένων.

Στο δεύτερο μέρος του πρώτου θέματος, εφαρμόζουμε τη μέθοδο PCA στα δεδομένα 5000 προσώπων που δίνονται. Ακολουθώντας την ίδια διαδιακασία με το πρώτο μέρος της άσκησης, παρατίθεται μία σύγκριση ανάμεσα στα αρχικά κ τα ανεκτημένα δεδομένα:



Τα ιδιοδιανύσματα παρέχουν αρκετή πληροφορία, δηλαδή τα κύρια χαρακτηριστικά που διαχωρίζει μία εικόνα από μία άλλη, έτσι ώστε να σχηματιστεί κ πάλι η αρχική εικόνα, δηλαδή, να ανακτηθούν τα δεδομένα, όπως ζητείται.

Η μέθοδος PCA δεν είναι βέλτιστη σε περίπτωση που είναι αναγκαία η διατήρηση χαρακτηριστικών απαραίτητων για τη διαφοροποίηση συγκεκριμένων κλάσεων, αλλά, σε αυτή τη περίπτωση, η μέθοδος είναι χρήσιμη, καθώς με ένα ποσοστό των αρχικών διαστάσεων, μπορούμε να ανακτήσουμε την εικόνα σε ικανοποιητικό βαθμό.

Θέμα 2ο – Ταξινομητής LDA

$$\Sigma_w = p_{\omega_1}\cdot\Sigma_1 + p_{\omega_2}\cdot\Sigma_2 = \frac{1}{2}\cdot\begin{pmatrix}11 & 9\\ 9 & 11\end{pmatrix}+\frac{1}{2}\cdot\begin{pmatrix}2 & 0\\ 0 & 2\end{pmatrix}=\begin{pmatrix}\frac{13}{2} & \frac{9}{2}\\ \frac{9}{2} & \frac{13}{2}\end{pmatrix}$$

$$det(\Sigma_w)=\left(\frac{13}{2}\right)^2-\left(\frac{9}{2}\right)^2=\frac{169}{4}-\frac{81}{4}=\frac{88}{4}=22$$

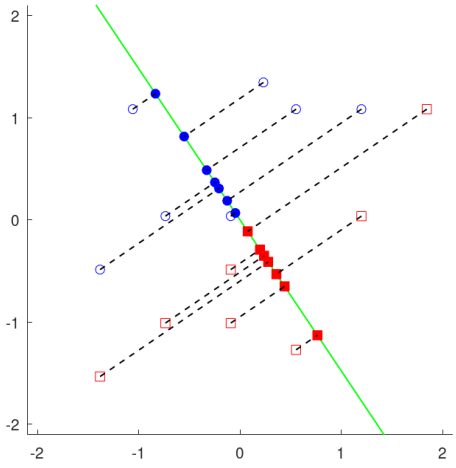
$$\Sigma_w^{-1}=\frac{1}{det(\Sigma_w)}\cdot\begin{pmatrix}\frac{13}{2} & -\frac{9}{2}\\ -\frac{9}{2} & \frac{13}{2}\end{pmatrix}=\begin{pmatrix}\frac{13}{44} & -\frac{9}{44}\\ -\frac{9}{44} & \frac{13}{44}\end{pmatrix}$$

$$w=\Sigma_w^{-1}\cdot(\mu_1-\mu_2)=\begin{pmatrix}\frac{13}{44} & -\frac{9}{44}\\ -\frac{9}{44} & \frac{13}{44}\end{pmatrix}\cdot\begin{pmatrix}-15\\ -10\end{pmatrix}=\begin{pmatrix}-\frac{105}{44}\\ \frac{5}{44}\end{pmatrix}$$

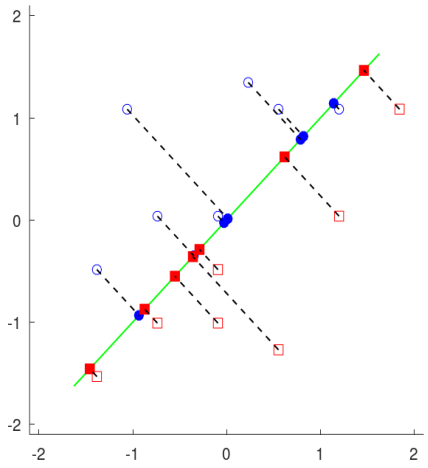
Θέμα 3ο – Linear Discriminant Analysis (LDA) vs PCA

Εφαρμόζουμε Linear Discriminant Analysis για τη μείωση της διάστασης ενός διανύσματος χαρακτηριστικών, κ συγκρίνουμε τα αποτελέσματα με αυτά της μεθόδου PCA. Αρχικά, εφαρμόζουμε τις μεθόδους σε τεχνητά δισδιάστατα δεδομένα δύο κλάσεων.

Μετά τη κανονικοποίηση των δεδομένων, τα προβάλλουμε στις ευθείες των δύο μεθόδων:



***Εικόνα 2**: Προβολή δεδομένων σε LDA*

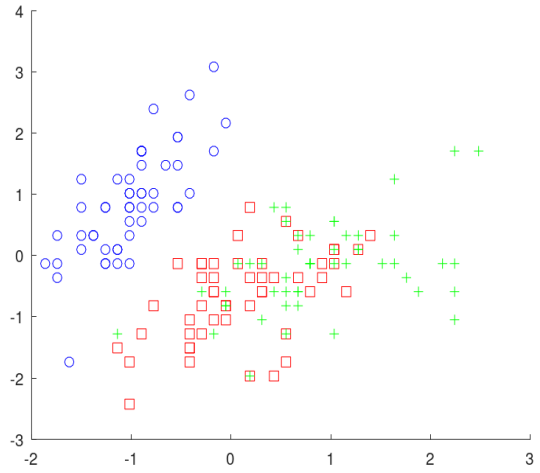


***Εικόνα 3**: Προβολή δεδομένων σε PCA*

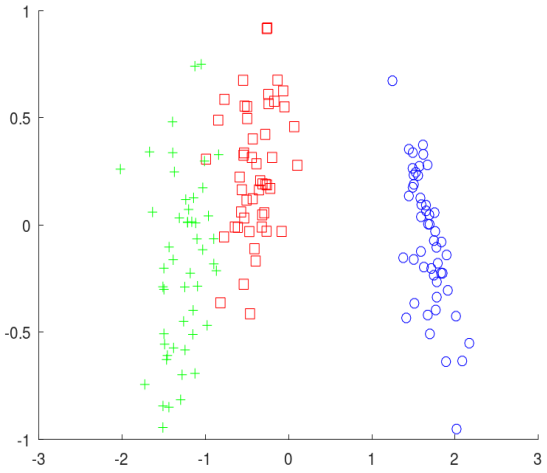
Καθώς η μέθοδος PCA βρίσκει μόνο τη κατεύθυνση της μέγιστης διαφοράς κ δε χρησιμοποιεί πληροφορίες για τις ετικέτες κάθε κλάσσης, είναι φανερό πως η διαχωρισιμότητα των κλάσεων υστερεί κ έτσι η μέθοδος δεν ενδείκνυται για την υλοποίηση κάποιου αλγορίθμου ταξινόμησης.

Αντιθέτως, ο LDA, χρησιμοποιώντας τις ετικέτες των κλάσεων, βρίσκει τη βέλτιστη κατεύθυνση στην οποία οι μέσες τιμές των κλάσεων έχουν τη μέγιστη απόσταση μεταξύ του, με ελάχιστο within-class variance. Το αποτέλεσμα είναι καλύτερο αυτού της μεθόδου PCA, και επομένως, η μέθοδος LDA θα μπορούσε να χρησιμοποιηθεί ως ταξινομητής.

Στο δεύτερο μέρος της άσκησης αυτής, εφαρμόζουμε τη μέθοδο LDA, στη βάση δεδομένων Fisher's Iris.



***Εικόνα 4**: Πριν την εφαρμογή LDA*



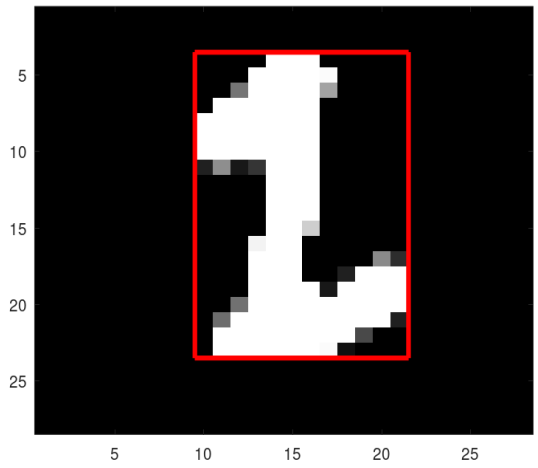
***Εικόνα 5**: Μετά την εφαρμογή LDA*

Παρατηρείται πως τα δεδομένα είναι αρκετά διαχωρίσιμα μετά την εφαρμογή της μεθόδου LDA, η οποία μειώνει τις διαστάσεις κ προσφέρει τις κατευθύνσεις της μέγιστης διαχωρισιμότητος.

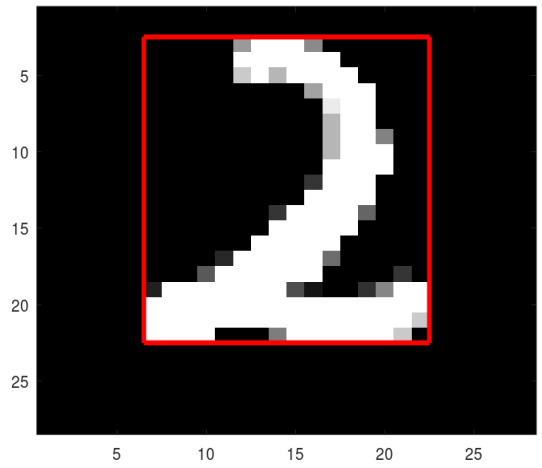
Θέμα 5ο – Εξαγωγή χαρακτηριστικών κ Bayes Classification

Μέσω ενός ταξινομητή Bayes, εξάγουμε δεδομένα από από χειρόγραφες εικόνες αριθμητικών ψηφίων, συγκεκριμένα, των ψηφίων 1 κ 2 όπως ζητήθηκε από την εκφώνηση της ασκήσεως.

Επιλέγουμε τον λόγο διαστάσεων των μη-μηδενικών εικονοστοιχείων ως χαρακτηριστικό για την ταξινόμηση. Χρησιμοποιώντας τον κανόνα Bayes, υποθέτουμε πως τα δείγματα των λόγων διαστάσεων θα έχουν κανονική κατανομή κ υπολογίζουμε τις εκ των υστέρων πιθανότητες.



***Εικόνα 6**: Ψηφίο 1*



***Εικόνα 7**: Ψηφίο 2*

Θέμα 6ο – Ελάχιστο ρίσκο

Έχουμε:
$$l_1 = \lambda_{11}\cdot P(\omega_1)\cdot P(x|\omega_1)+\lambda_{21}\cdot P(\omega_2)\cdot P(x|\omega_2) \Rightarrow l_1 = P(\omega_2)\cdot P(x|\omega_2)$$
$$l_2 = \lambda_{12}\cdot P(\omega_1)\cdot P(x|\omega_1)+\lambda_{22}\cdot P(\omega_2)\cdot P(x|\omega_2) \Rightarrow l_2 = \frac{1}{2}P(\omega_1)\cdot P(x|\omega_1)$$

Όπου: $L=\begin{pmatrix}\lambda_{11} & \lambda_{12}\\ \lambda_{21} & \lambda_{22}\end{pmatrix}=\begin{pmatrix}0 & 0.5\\ 1 & 0\end{pmatrix}$ και $\begin{matrix}Αν\ l_1<l_2, τότε\ \omega_1\\ Αν\ l_1>l_2, τότε\ \omega_2\end{matrix}$

$$\Rightarrow P(\omega_2)\cdot P(x|\omega_2) = \frac{1}{2}P(\omega_1)\cdot P(x|\omega_1)$$
$$\Rightarrow \frac{X_0}{2^2}\cdot e^{-\frac{X_0^2}{8}}=\frac{1}{2}\cdot X_0\cdot e^{-\frac{X_0^2}{2}}$$
$$\Rightarrow \ln\left(\frac{1}{2}\cdot e^{-\frac{X_0^2}{8}}\right)=\ln\left(e^{-\frac{X_0^2}{2}}\right)$$
$$\Rightarrow -\frac{X_0^2}{8}-\ln 2=-\frac{X_0^2}{2}$$
$$\Rightarrow -X_0^2-8\ln 2=-4X_0^2$$
$$\Rightarrow 3X_0^2=8\ln 2$$
$$\Rightarrow X_0=\sqrt{\frac{8\ln 2}{3}}$$