Αναφορά αποτελεσμάτων HomeWork 1

Μπανέλας Δημήτριος: 2018030140

Μάθημα: Στατιστική Μοντελοποίηση και Αναγνώριση Προτύπων

Διδάσκων θεωρίας: Ζερβάκης Μιχαήλ

Διδάσκων εργαστηρίου: Διακολουκάς Βασίλειος

Περιεχόμενα αναφοράς και οδηγίες εκτέλεσης κώδικα

Η παρούσα αναφορά θα παρούσιασει και θα εξηγήσει τα αποτελέσματα των ασκήσεων 1, 3 και 5, δηλαδή των ασκήσεων που είχαν σχέση με κώδικα. Οι υπόλοιπες ασκήσεις έχουν λυθεί στο χαρτί και θα βρεθούν στον αντίστοιχο φάκελο.

Ο κώδικας που αναπτύχθηκε για τις ανάγκες των ασκήσεων βρίσκεται στα αρχεία με κώδικα που έχουν ήδη δοθεί. Ο τρόπος εκτέλεσης έμεινε ίδιος με πριν. Πιο συγκεκριμένα, μετά απο κάθε ενέργεια, το πρόγραμμα γίνεται pause και περιμένει το πάτημα ενός πλήκτρου από τον χρήστη για να συνεχιστεί. Σημαντικό είναι, να υπάρχει παρατήρηση του console καθ΄ όλη τη διάρκεια εκτέλεσης του κάθε προγράμματος, καθώς εκεί εμφανίζονται σημαντικές πληροφορίες.

Σε πολλά σημεία υπάρχουν blocks κώδικα που βρίσκονται μέσα σε σχόλια. Ο κώδικας αυτός γράφτηκε για την καλύτερη κατανόηση των εννοιών και δεν είναι απαραίτητος για την ορθή λειτουργία του κάθε προγράματος. Τέλος, ο κώδικας περιέχει επεξηγηματικά σχόλια έτσι ώστε να μπορεί να διαβαστεί με σχετική ευκολία.

Θέμα 1: Principal Component Analysis (PCA)

Μέρος 1

Στο πρώτο μέρος ασχοληθήκαμε με τυχαία 2D δεδομένα στα οποία εφαρμόσαμε PCA έτσι ώστε να κατανοήσουμε τη λειτουργία του.

Αρχικά χρειάστηκε να κανονικοποιήσουμε τα δεδομένα, αφού ο PCA αντιμετωπίζει προβλήματα όταν υπάρχουν μεγάλες διαφορές μεταξύ των τιμών των χαρακτηριστικών. Π.χ έαν τα χαρακτηριστικά για τα οποία είχαμε δεδομένα ήταν το ύψος (mm) και το βάρος (Kg) ενός ανθρώπου, τότε η κατεύθυνση της μέγιστης διασποράς θα ήταν πάντα αυτή του ύψους, λόγω της κλίμακας στην οποία έγιναν οι μετρήσεις.

Παρακάτω παρατίθενται τα plots των δεδομένων πριν και μετά την κανονικοποίηση.

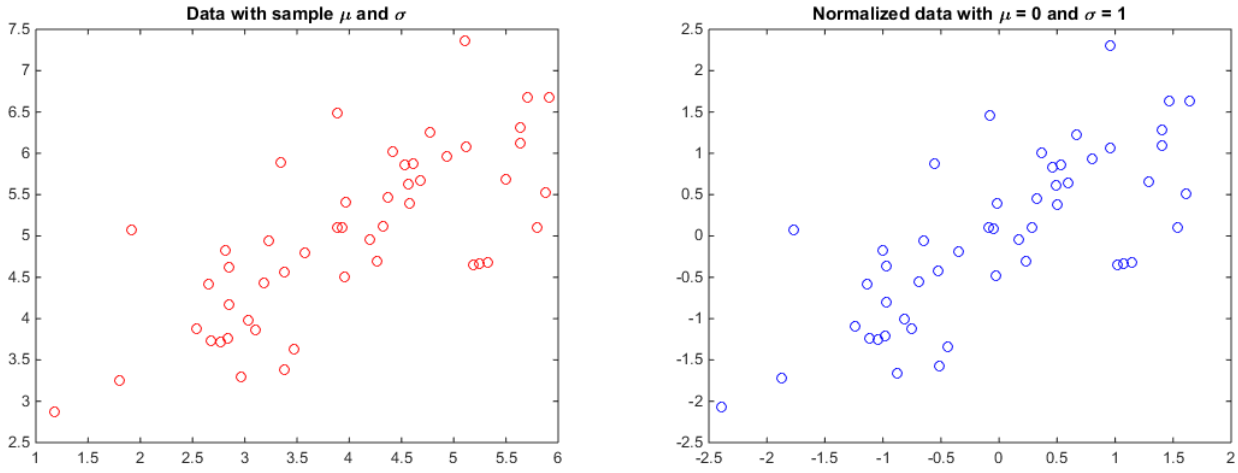


Figure 1: Data before and after normalization

Έπειτα, εφαρμόσαμε τον PCA βρίσκοντας τις κατευθύνσεις της μέγιστης διασποράς, καθώς και τι ποσοστό της ολικής διασποράς ΄κουβαλάει΄ η κάθε κατεύθυνση. Παρακάτω παρατίθενται τα plots των Principal Components πάνω στα δεδομένα, καθώς και η ανάκτηση των δεδομένων από τον 1D χώρο στον 2D, πάνω στο πρώτο PC.

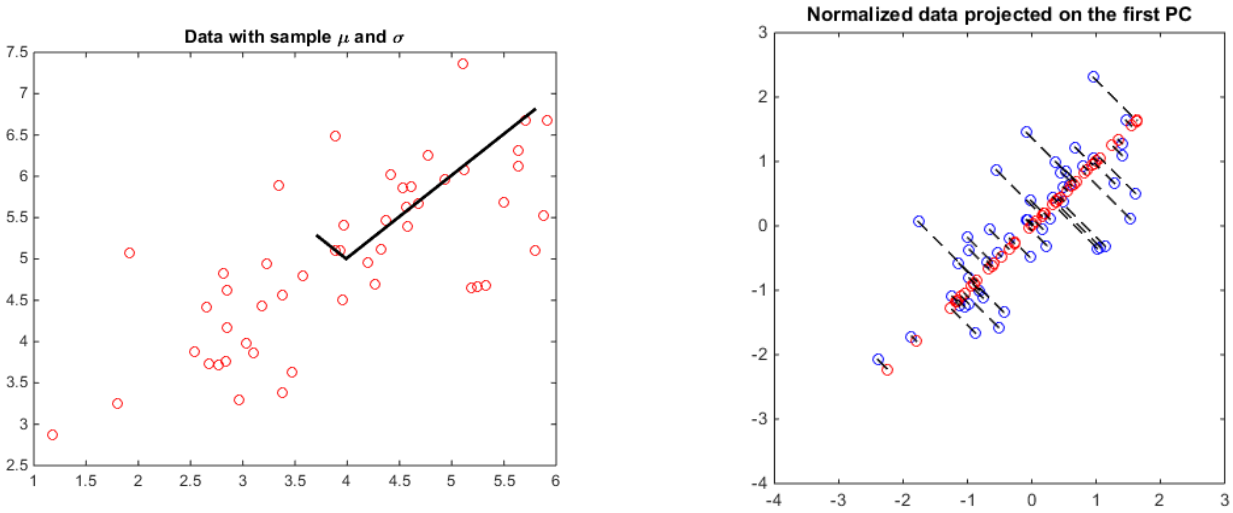


Figure 2: PC1_variance = 86.77% , PC2_variance = 13.22%

Παρατηρούμε πως πράγματι το πρώτο PC μας δίνει την κατεύθυνση στην οποία τα δεδομένα έχουν την μεγαλύτερη διασπορά. Επίσης, μπορούμε να δούμε ότι μετά την προβολή των δεδομένων πάνω στο πρώτο PC έχουμε καταφέρει πάμε στον 1D χώρο διατηρώντας ταυτόχρονα, αρκετά μεγάλο ποσοστό της αρχικής διασποράς.

Μέρος 2

Στο δεύτερο μέρος εφαρμόσαμε PCA σε έναν μεγάλο όγκο δεδομένων απο εικόνες προσώπων μειώνοντας τις διαστάσεις, και παρατηρήσαμε τα αποτελέσματα. Σε αυτό το πείραμα, αξίζει να σταθούμε στα εξής δύο πράγματα:

- Eigenfaces:

  Εφαρμόζοντας PCA στα δεδομένα πήραμε έναν πίνακα με 1024 PCs. Παρατίθεται το plot των πρώτων 36:

Figure 3: Faces from 36 first PCs

Παρατηρούμε πως το κάθε ιδιοδιάνυσμα, περιέχει πληροφορία η οποία είναι αρκετή για να σχηματιστεί μια εικόνα. Η πληροφορία αυτή είναι τα κύρια χαρακτηριστικά, τα οποία διαχωρίζουν τις εικόνες μεταξύ τους. Πχ το πρώτο eigenface, οπως φαίνεται στο παραπάνω figure, περιέχει πληροφορία για την φωτεινότητα των εικόνων. Οι είκόνες των eigenfaces αποτελούν τις "base" εικόνες των πραγματικων εικόνων. Οι πραγματικές εικόνες μπορούν να κατασκευαστούν από έναν γραμμικό των eigenfaces.

- Ποσοστό συνολικής διασποράς σε K Principal Components:

  Παρατίθεται ο πίνακας με τις μετρήσεις το ποσοστό της συνολικής διασποράς πάνω σε K PCs.

Table 1: Percentage of total variance in K PCs

K	% of variance
50	86.79
100	93.19
150	95.86
200	97.30
300	98.72

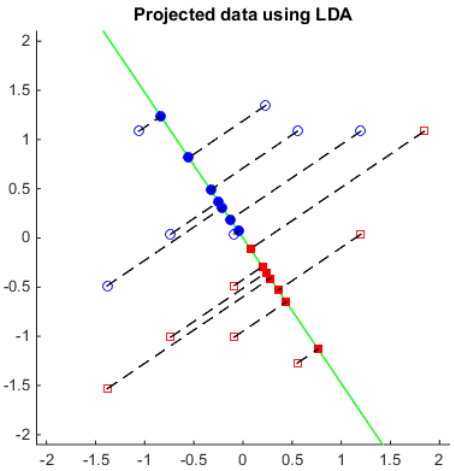
Μπορούμε να παρατηρήσουμε ότι ρίχνοντας τις διαστάσεις από τις αρχικές 1024 στις 50 διατηρείται το 86.79% της αρχικής πληροφορίας. Σίγουρα ένα τέτοιο ποσοστό δεν είναι αρκετό για να είναι αξιόπιστα τα τελικά δεδομένα, ωστόσο είναι αποδεικτικό του πόσο αποδοτικός είναι ο PCA στον συγκεκριμένο τύπο προβλημάτων. Επίσης, γίνεται αντιληπτό πως με μόλις 200/1024 διαστάσεις μπορεί να διατηρηθεί το 97+% της αρχικής πληροφορίας, κάτι το οποίο είναι πολύ σημαντικό αφού η μείωση αυτή των δεδομένων θα επιτρέψει στους machine learning αλγορίθμους να εκπαιδευτούν πολύ πιο γρήγορα, χωρίς σημαντική μείωση της αξιοπιστίας των δεδομένων. Τέλος, αξίζει να αναφέρουμε πως σε περίπτωση που αναγκαία θα ήταν η διατήρηση των χαρακτηριστικών τα οποία διαφοροποιούν κάποιες κλάσεις, ο PCA θα αποτύγχανε, αφού κοιτάζει να μεγίστοποιήσει μόνο τη διασπορά των δεδομένων.

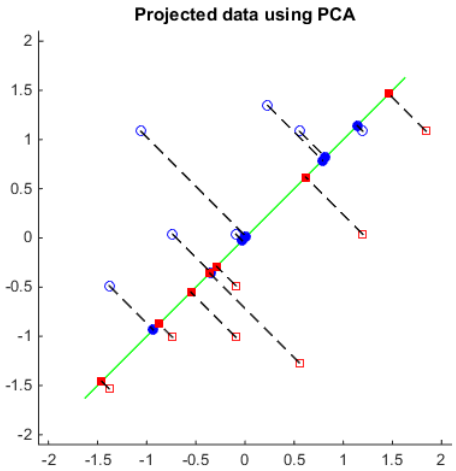
Θέμα 3: Linear Discriminant Analysis(LDA) vs PCA

Μέρος 1

Στο πρώτο μέρος συγκρίναμε τους PCA και LDA και παρατηρήσαμε τις διαφορές τους, εφαρμόζοντας τους σε τεχνητά 2D δεδομένα 2 κλάσεων.

Αφού κανονικοποιήσαμε τα δεδομένα, βρήκαμε την ευθεία την οποία παράγει ο LDA, καθώς και αυτή που παράγει ο PCA.Επειτα, προβάλλαμε τα δεδομένα μας πάνω στις ευθείες αυτές, και αυτά ήταν τα αποτελέσματα:





Στις παραπάνω εικόνες μπορούμε να δούμε ξεκάθαρα τις διαφορές μεταξύ των αλγορίθμων, καθώς έχουν επιλέξει διαφορετικές ευθείες για την προβολή των δεδομένων.

- PCA:

  Ειναι φανερό πως ο PCA, μην έχοντας πληροφορίες για τα class labels, προσπαθεί να βρεί μόνο την κατεύθυνση στην οποία η διασπορά των δεδομένων είναι η μέγιστη. Επιτυγχάνοντας το αυτό, χάνεται η διαχωρισιμότητα μεταξύ των δύο κλάσεων, επομένως δεν θα μπορούσαμε να τρέξουμε εύκολα κάποιον classification αλγόριθμο αργότερα.

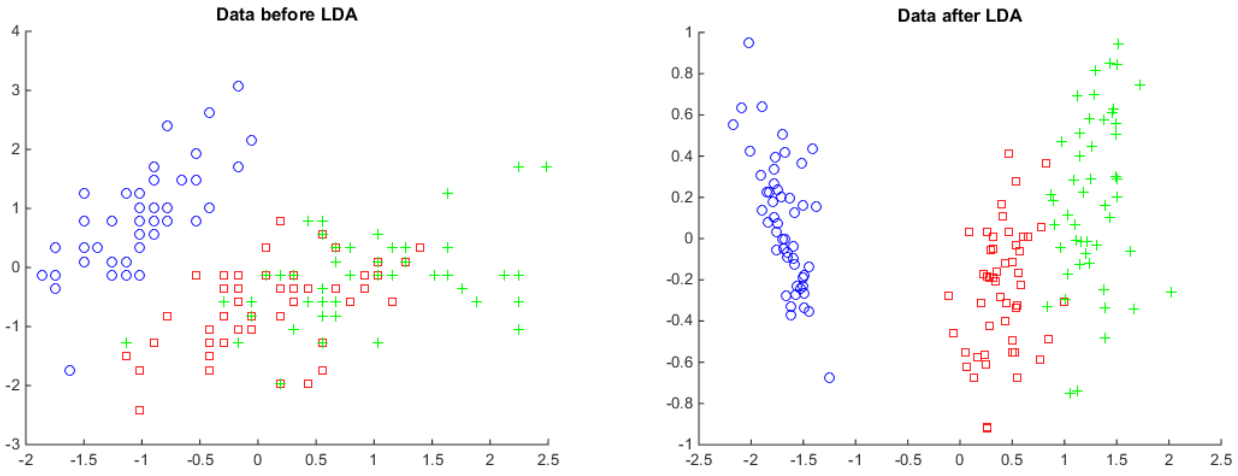
- LDA:

  Ο LDA, γνωρίζοντας τα class labels, προσπαθεί μία βέλτιστη κατεύθυνση, στην οποία τα means των κλάσεων έχουν τη μέγιστη απόσταση μεταξύ τους, και οι within class variances είναι ελάχιστες. Παρατήρουμε ότι το αποτέλεσμα είναι αρκετά καλό, και θα κάνει το classification αρκετά εύκολο. Αυτός είναι ο λόγος για το γεγονός ότι ο LDA μπορεί να χρησιμοποιηθεί και σαν classifier.

Μέρος 2

Στο δεύτερο μέρος, εφαρμόσαμε LDA πάνω σε ένα dataset της βάσης IRIS και παρατηρήσαμε τα αποτελέσματα.

Το dataset περιείχε 4 χαρακτηριστικά-διαστάσεις, και ο LDA μας επέστρεψε δεδομένα $C - 1 = 2$ διαστάσεων, όπου C ο αριθμός των κλάσεων. Παρακάτω παρατίθενται τα plots των χαρακτηριστικών, πριν και μετά την εφαρμογή του LDA.

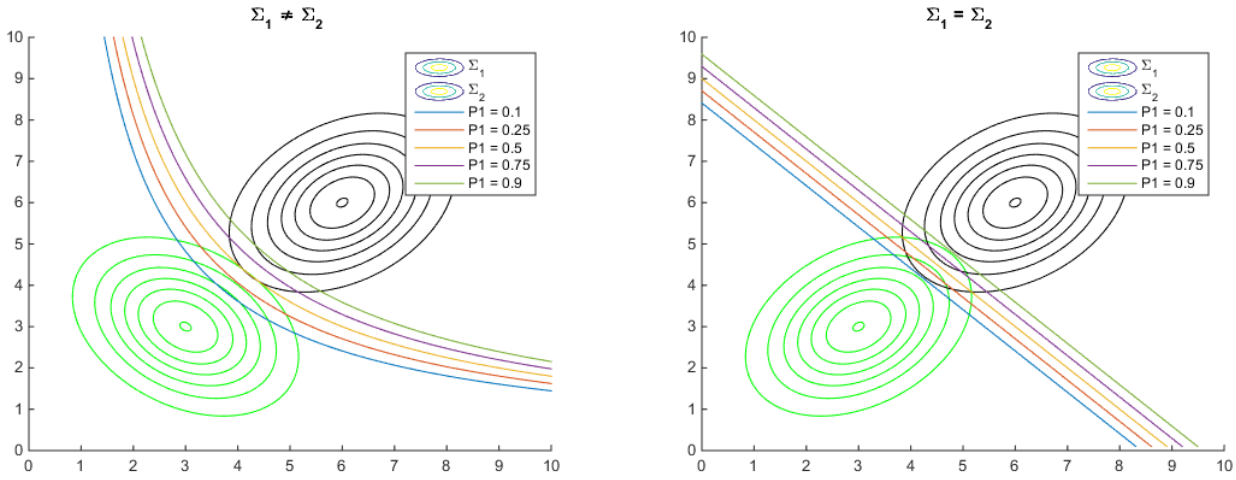


Παρατηροούμε ότι πλέον υπάρχει καλύτερη διαχωρισιμότητα μεταξύ των δεδομένων, χάρη στον LDA, ο οποίος εκτός απο dimensionality reduction, μας έδωσε και τις κατευθύνσεις στις οποίες υπάρχει η μέγιστη διαχωρισιμότητα.

Συμπερασματικά, και οι 2 παραπάνω αλγόριθμοι πραγματοποιούν μείωση των διαστάσεων του dataset, με διαφορετικό τρόπο ο καθένας. Το αν και πως θα χρησιμοποιήσουμε τον κάθε αλγόριθμο εξαρτάται από τις ιδιότητες του dataset. Υπάρχουν περιπτώσεις όπου μπορούν να χρησιμοποιηθούν και οι 2 αλγόριθμοι. Πρώτα ο PCA για dimensionality reduction, έπειτα ο LDA για να μετασχηματίσει τα δεδομένα έτσι ώστε να έχουν καλύτερο seperability, και τέλος ένα μοντέλο για classification.

Θέμα 4: Bayes

Σε αυτό το κομμάτι της αναφοράς θα παρουσιαστουν τα αποτελέσματα της 3ης άσκησης. **Τα ερωτήματα a και e που αφορούσαν τον υπολογισμό των ορίων απόφασης, έχουν λυθεί χειρόγραφα, και βρίσκονται στο αντίστοιχο αρχείο**.Παρατίθενται τα ζητούμενα plots, καθώς και τα αποτελέσματα των υπόλοιπων ερωτημάτων:



- για $\Sigma_1 \neq \Sigma_2$

  Παρατηρούμε πως τα όρια αποφάσεων είναι καμπύλες. Αυτό συμβαίνει γιατί ο όρος $(\mathbf{x} - \mu_\mathbf{i})^\intercal \Sigma_i^{-1} (\mathbf{x} - \mu_\mathbf{i})$ της pdf, περιέχει όρους της μορφής ax_1x_2, όπου x_1 και x_1 τα στοιχεία του $\mathbf{x}$. Γενικά, για διαφορετικά scatter matrices μπορούμε να περιμένουμε όρια που έχουν μορφή υπερβολής η παραβολής, εξαιτίας του τετραγωνικού όρου που υπάρχει στο εκθετικό της pdf.

- για $\Sigma_1 = \Sigma_2$

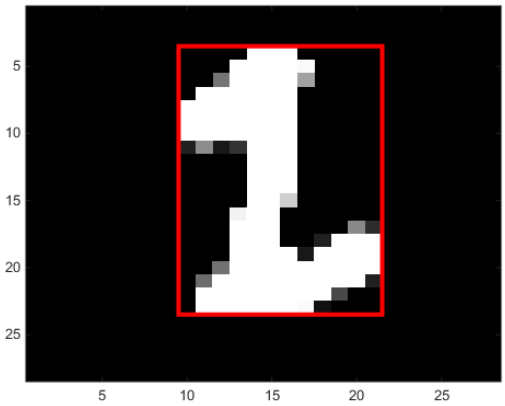
  Τα όρια απόφασης είναι ευθείες, αφού ο όρος της pdf που αναφέρθηκε παραπάνω δεν περιέχει ούτε τετραγωνικούς όρους, ούτε όρους της μορφής ax_1x_2.

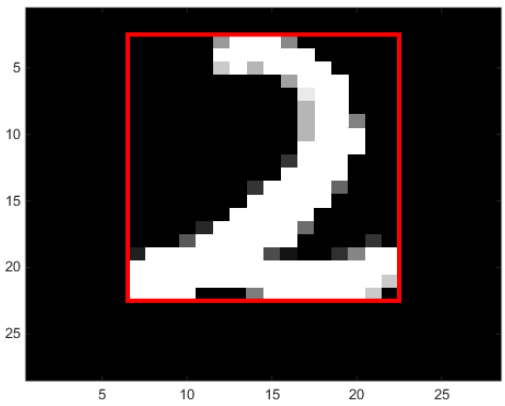
Και για τις 2 περιπτώσεις ισχύει ότι, όσο αν η a-priori πιθανότητα μιας κλάσης αυξάνεται, τόσο το όριο μετακινείται πιο κοντά στην άλλη κλάση. Διαισθητικά, μπορούμε να σκεφτούμε ότι ¨θέλουμε να αποφασίζουμε πιο συχνά την κλάση με την μεγαλύτερη a-priori πιθανότητα, και έτσι της δίνουμε μεγαλύτερη περιοχή απόφασης¨.

Θέμα 5: Εξαγωγή χαρακτηρηστικών και Bayes Classification

Στην 5η άσκηση ασχληθήκαμε με την εξαγωγή χαρακτηρηστικών και το classification των ψηφίων 1 και 2 του mnist database.

Το χαρακτηριστικό το οποίο συλλεξαμε από τις εικόνες είναι το aspect ratio της κάθε εικόνας, δηλαδή τον λόγο $\frac{width}{height}$ των μη μηδενικών pixels. Έπειτα, υποθέσαμε ότι τα δείγματα των aspect ratios προέρχονται από κανονικές κατανομές, με τα sample μ και σ των δειγμάτων εκπαίδευσης, και φτίαξαμε τις a-posteriori πιθανότητες για να κάνουμε classification, χρησιμοποιώντας τον κανόνα του Bayes. Παρατίθενται 2 ενδεικτικά ψηφία μαζί με τα ορθογώνια που ορίζουν τα aspect ratios.





Το ποσοστό λανθασμένης απόφασης του παραπάνω ταξινομητή ανέρχεται στο 10.93%.