ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ
TECHNICAL UNIVERSITY OF CRETE

Αναφορά 1ης Σειράς Ασκήσεων

Authors

Θωμάς Λάγκαλης, 2021030079

Ιωάννης Λεμονάκης, 2021030138

ΤΗΛ311 Στατιστική Μοντελοποίηση και Αναγνώριση Προτύπων

Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Πολυτεχνείο Κρήτης

Απρίλιος 2023

1 Εισαγωγή

Στην παρούσα αναφορά παρουσιάζονται τα απαραίτητα σχόλια για τα θέματα του πρώτου σετ ασκήσεων όπου αυτό ζητείται (δηλαδή, στα θέματα 1, 5, 7). Για τα σχόλια των υπολοίπων ερωτημάτων, η τεκμηρίωση παρεμβάλλεται μέσα στον κώδικα ή στις σκαναρισμένες εικόνες (αρχείο pdf). Συγκεκριμένα, στο θέμα 1 σχολιάζεται η μορφή των συνόρων απόφασης για τις διάφορες τιμές των πιθανοτήτων που δίνονται στην εκφώνηση. Στο θέμα 5 σχολιάζεται η επαναφορά των δεδομένων στην αρχική τους διάσταση και τέλος στο θέμα 7 αναλύεται η σύγκριση των μεθόδων μείωσης διαστάσεων LDA και PCA.

2 Ανάλυση Ερωτημάτων

2.1 Θέμα 1

Σημείωση: Ο κώδικας για το θέμα 1 είναι καλύτερο να τρέξει στο GNU Octave.

Όπως γίνεται αντιληπτό από το σχήμα 1 και 2 για διαφορετικές τιμές των πιθανοτήτων $P(\omega_1)$ (και αντίστοιχα $P(\omega_2)$), τα σύνορα απόφασης παρουσιάζονται ελαφρώς μετατοπισμένα. Σύμφωνα με τους υπολογισμούς έχει βρεθεί η γενική εξίσωση του συνόρου απόφασης για διαφορετικούς πίνακες συνδιασποράς:

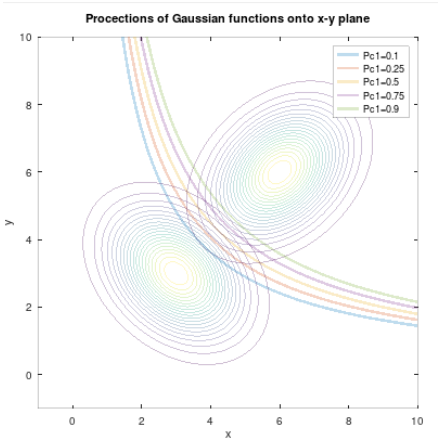
$$x_0 = \frac{2.56 \ln(\frac{P(\omega_1)}{1-P(\omega_1)}) + 28.2}{1.6y} \qquad (1)$$

Για ίδιους πίνακες δυνδιασποράς το σύνορο απόφασης έχει εξίσωση ευθείας:

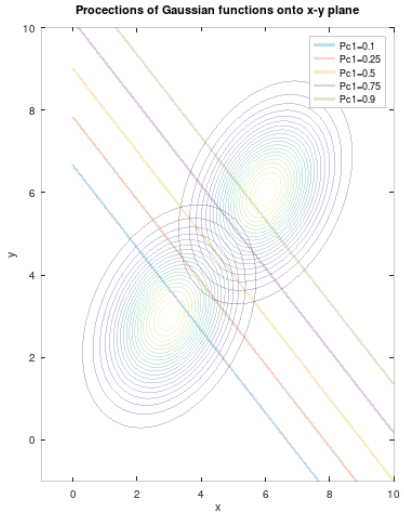
$$x_0 = -y + 9 - 1.0666 \ln(\frac{P(\omega_1)}{1 - P(\omega_1)}) \qquad (2)$$

Γραφικά ο κανόνας απόφασης είναι ο εξής: αν το δείγμα (μέτρηση) έχει συντεταγμένες (features) που βρίσκεται κάτω από το σύνορο απόφασης τότε κατατάσσεται στην κλάση 1 αλλοώς στην κλάση 2.

Όσο η τιμή του $P(\omega_1)$ αυξάνεται τόσο αυξάνεται και το x_0 και η γραφική παράσταση του συνόρου απόφασης μετατοπίζεται δεξιά (και πάνω). Αυτό το γεγονός ισχύει και για τις δύο περιπτώσεις των πινάκων συνδιασποράς (και για τα δύο σχήματα). Επομένως, με την αύξηση της a priori πιθανότητας της κλάσης 1 το σύνορο μετατοπίζεται πιο κοντά στη κλάση 2 και για ίδιους πίνακες συνδιασποράς το σύνορο είναι ευθεία ενώ για διαφορετικούς είναι καμπύλη.



Σχήμα 1: Σύνορα απόφασης και ισοϋψείς καμπύλες των γκαουσσιανών κατανομών για διαφορετικούς πίνακες συνδιασποράς



Σχήμα 2: Σύνορα απόφασης και ισοϋψείς καμπύλες των γκαουσσιανών κατανομών για ίδιους πίνακες συνδιασποράς.

2.2 Θέμα 2

Στο μέρος 2 του θέματος 5 εφαρμόστηκε η μέθοδος PCA για μείωση διαστάσεων σε ένα σύνολο 5000 δειγμάτων με 1024 features το καθ' ένα. Συγκεκριμένα, αφού υπολογίστηκαν οι ιδιοτιμές και τα ιδιοδιανύσματα του πίνακα συνδιασποράς στη συνάρτηση myPCA, επιλέχθηκαν οι πρώτες K κύριες συνιστώσες (principal components). Έπειτα, έγινε προβολή των δεδομένων πάνω στο K-διάστατο επίπεδο με τη συνάρτηση projectData (μείωση διάστασης). Τέλος, γίνεται ανάκτηση των δεδομένων στην αρχική διάσταση με τη συνάρτηση recoverData και αποτυπώνονται μερικά δείγματα των αρχικών δεδομένων και των δεδομένων μετά την ανάκτηση για να γίνει η σύγκριση. Παρακάτω φαίνονται τα ανακατεμένα δεδομένα για διαφορετικές τιμές K.

Σχήμα 3: Αναρτημένες εικόνες για K=10

Όπως γίνεται αντιληπτό και στα σχήματα 3 , 4 και 5, με την αύξηση των principal components αυξάνεται και η ποιότητα των ανακτημένων δεδομένων. Στο σχήμα 3 φαίνονται οι εικόνες των προσώπων θολωμένες ενώ για μεγάλο αριθμό κύριων συνιστωσών τα πρόσωπα είναι αρκετά ευδιάκριτα, όχι όμως όσο είναι τα πρόσωπα των αρχικών δεδομένων. Επομένως, χρησιμοποιώντας τους γραμμικούς μετασχηματισμούς για την προβολή και ανάκτηση δεδομένων σε μικρότερη και την αρχική τους διάσταση αντίστοιχα προστίθεται ένα σφάλμα στις τιμές τους το οποίο μειώνεται με την αύξηση του αριθμού των κύριων συνιστωσών (K).

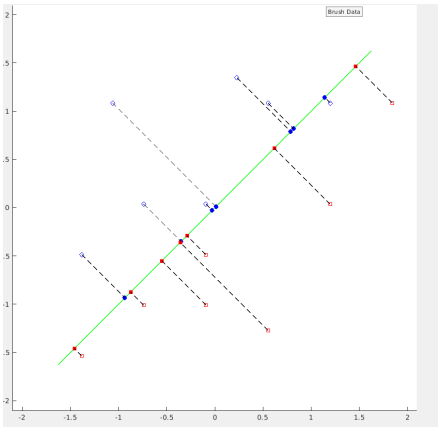
Σχήμα 4: Αναρτημένες εικόνες για K=50

Σχήμα 5: Αναρτημένες εικόνες για K=200

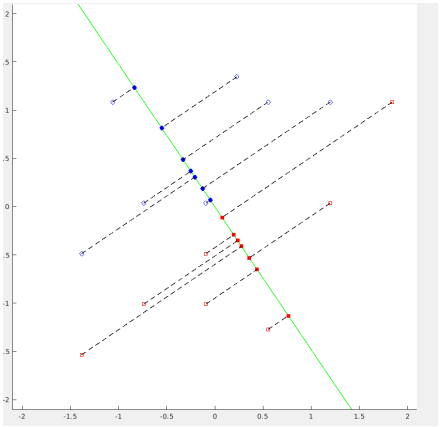
2.3 Θέμα 7

Στο μέρος αυτό σχολιάζεται το 1ο κομμάτι του θέματος 7 και ιδιαίτερα οι διαφορές μεταξύ των μεθόδων μείωσης διαστάσεων PCA και LDA σε ένα δείγμα με 2D δεδομένα. Δηλαδή, εφαρμόστηκαν οι δύο μέθοδοι για την προβολή των σημείων από το επίπεδο πάνω σε ένα διάνυσμα την κατεύθυνση του οποίου καθορίζει η μέθοδος που εφαρμόστηκε (PCA ή LDA). Επίσης, επισημαίνεται ότι τα δεδομένα είναι labeled δηλαδή μαζί με τα δεδομένα δηλώνεται και η κλάση στην οποία ανήκουν.

Από τα σχήματα 6 και 7 γίνεται αντιληπτό



Σχήμα 6: Προβολή των δεδομένων πάνω σε διάνυσμα με την PCA



Σχήμα 7: Προβολή των δεδομένων πάνω σε διάνυσμα με την LDA

πως η μέθοδος LDA είναι πιο αποτελεσματική στο να διαχωρίζει τις δύο κλάσεις στη μειωμένη διάσταση (σημειώνεται, πως τα σημεία διαφορετικού χρώματος ανήκουν σε διαφορετικές κλάσεις). Αυτό οφείλεται στο γεγονός ότι η μέθοδος LDA επιλέγει το διάνυσμα προβολής με βάση τη μέγιστη διασπορά μεταξύ των κλάσεων. Από την άλλη η PCA επιλέγει το διάνυσμα προβολής ώστε να μεγιστοποιήσει τη διασπορά μεταξύ των σημείων. Δηλαδή, η LDA χρησιμοποιείται για labeled δεδομένα ενώ η PCA για unlabeled. Συμπερασματικά, στο δοθέν παράδειγμα της άσκησης, που υπάρχουν labeled δεδομένα, η LDA είναι πιο αποτελεσματική.

3 Πηγές

Για την ανάπτυξη των ασκήσεων χρησιμοποιήθηκε η θεωρία από τις διαφάνειες του Δρ.Γ.Καρυστινού και του Δρ. Β. Διακολουκά. Επίσης, για τον κώδικα matlab/GNU Octave χρησιμοποιήθηκε το επίσημο documentation της matlab.