ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ

Τμήμα Ηλεκτρολόγων Μηχ. Και Μηχ. Υπολογιστών

Στατιστική Μοντελοποίηση και Αναγνώριση Προτύπων (ΤΗΛ311)

Φυλλάδιο Ασκήσεων 1

Οδηγίες:

1. Σας παρακαλώ να σεβαστείτε τον παρακάτω κώδικα τιμής τον οποίον θα θεωρηθεί ότι προσυπογράφετε μαζί με τη συμμετοχή σας στο μάθημα και τις εργασίες του:
   a) *Οι απαντήσεις, ο κώδικας και γενικά οτιδήποτε αφορά τις εργασίες, τα φυλλάδια ασκήσεων και τις εξετάσεις του μαθήματος θα είναι προϊόν δικής μου δουλειάς.*
   b) *Δεν θα διαθέσω κώδικα, απαντήσεις και εργασίες μου σε κανέναν άλλο.*
   c) *Δεν θα εμπλακώ σε άλλες ενέργειες με τις οποίες ανέντιμα θα βελτιώνω τα αποτελέσματα μου ή ανέντιμα θα αλλάζω τα αποτελέσματα άλλων.*
2. Η εργασία αυτή είναι ατομική
3. Ημερομηνία παράδοσης: **Παρασκευή, 20/4/2018 στις 20:00**
4. **Ανεβάστε στο courses τα ακόλουθα παραδοτέα σε μορφή zip:**
   Παραδοτέα: α) Κώδικας και β) Αναφορά με τις λύσεις/απαντήσεις, παρατηρήσεις, πειράματα, αποτελέσματα και οδηγίες χρήσης του κώδικα.

Θέμα 1: Principal Component Analysis (PCA)

Σε αυτή την άσκηση, θα χρησιμοποιήσετε την μέθοδο κυρίων συνιστωσών – Principal Component Analysis (PCA) για να κάνετε μείωση διαστάσεων στα δεδομένα σας. Θα δοκιμάσετε πρώτα τεχνητά δεδομένα δύο διαστάσεων 2D για να δείτε τον τρόπο λειτουργίας του PCA και στη συνέχεια, θα εφαρμόσετε το PCA σε ένα μεγαλύτερο σύνολο πραγματικών δεδομένων από 5000 εικόνες προσώπων.

Το σύνολο των δισδιάστατων δεδομένων (2D), έχει μία κατεύθυνση μεγάλης διακύμανσης και μία μικρότερης διακύμανσης.

Τρέξτε το matlab/octave script ex1_1_pca.m, αφού προσθέσετε τον απαραίτητο κώδικα για να υλοποιήσετε τις συναρτήσεις που καλούνται από το συγκεκριμένο script. **ΠΡΟΣΟΧΗ: Δεν επιτρέπεται να χρησιμοποιήσετε έτοιμες υλοποιήσεις των αλγορίθμων στο Matlab.**

a) Απεικονίστε τα 2D δείγματα των δεδομένων που διαβάζετε από το αρχείο «ex1_1_data1.mat» χρησιμοποιώντας τον κώδικα που σας δίνεται.
b) Συμπληρώστε τον απαραίτητο κώδικα στη συνάρτηση *featureNormalize.m* η οποία θα πρέπει να κανονικοποιεί τα δείγματα που θα της δίνετε στην είσοδο ώστε κάθε χαρακτηριστικό τους να έχει μέση τιμή μηδέν $\mu = 0$ και διασπορά $\sigma = 1$.
c) Συμπληρώστε τον απαραίτητο κώδικα στην συνάρτηση *principalComponentAnalysis.m* η οποία θα πρέπει να υπολογίζει τις κύριες συνιστώσες (principal components) από τα δείγματα στην είσοδο της.
d) Συμπληρώσετε τον κώδικα στη συνάρτηση *projectData.m* η οποία θα μειώνει τη διάσταση των δειγμάτων στην είσοδο, χρησιμοποιώντας τις κύριες συνιστώσες που υπολογίσατε στο

προηγούμενο βήμα. Συγκεκριμένα, για ένα σύνολο δεδομένων X, και κύριες συνιστώσες U θα πρέπει να προβάλετε κάθε δείγμα X στις K=1 πρώτες κύριες συνιστώσες του U εφαρμόζοντας τον κατάλληλο μετασχηματισμό.

e) Αφού προβάλετε τα δεδομένα σε χώρο μικρότερων διαστάσεων, μπορείτε να ανακτήσετε προσεγγιστικά τα δεδομένα αναπαράγοντάς τα ξανά στον αρχικό χώρο υψηλών διαστάσεων. Υλοποιήστε το recoverData.m για να προβάλετε κάθε παράδειγμα στο Z πίσω στον αρχικό χώρο και να επιστρέψετε την ανακτημένη προσέγγιση στο X_rec.

f) Παρατηρήστε τι συμβαίνει όταν εφαρμόζετε τα παραπάνω χρησιμοποιώντας τα πραγματικά δεδομένα από εικόνες προσώπων που σας δίνονται.

Θέμα 2: Σχεδιάστε ένα ταξινομητή LDA (Linear Discriminant Analysis)

Ένα σύνολο δεδομένων έχει προκύψει από δύο ισοπίθανες κατηγορίες ω_1, ω_2, οι κατανομές των οποίων θεωρούνται Γκαουσιανές. Οι πίνακες συνδιασποράς και οι μέσες τιμές έχουν εκτιμηθεί από τα δεδομένα ως:

$$\mu_1 = \begin{bmatrix} -5 \\ 5 \end{bmatrix}, \quad \mu_2 = \begin{bmatrix} 10 \\ 15 \end{bmatrix}$$

$$\Sigma_1 = \begin{bmatrix} 11 & 9 \\ 9 & 11 \end{bmatrix}, \quad \Sigma_2 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

Να υπολογίσετε το διάνυσμα προβολής w αναλυτικά, κάνοντας τις πράξεις με το χέρι. Για να αντιστρέψετε τον πίνακα Σ_w, χρησιμοποιήστε τον τύπο:

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

Θέμα 3: Linear Discriminant Analysis (LDA) vs PCA

Για να κατανοήσετε την διαφορά ανάμεσα στον αλγόριθμο PCA και τον LDA τρέξτε το Matlab/Octave script ex1_3_lda.m Θα χρειαστεί να συμπληρώσετε τις επόμενες συναρτήσεις.

a) fisherLinearDiscriminant.m
b) projectDataLDA.m
c) recoverDataLDA.m

Σε ότι αφορά την εφαρμογή του PCA χρησιμοποιήστε τον κώδικα που δημιουργήσατε στο θέμα 1.
Σχολιάστε τα αποτελέσματα και συγκρίνετε τους δύο αλγορίθμους.

Θέμα 4: Bayes

Έστω ότι σε ένα πρόβλημα κατηγοριοποίησης σε δύο κλάσεις ω_1 και ω_2 οι εκ των προτέρων πιθανότητες είναι $P(\omega_1)$ και $P(\omega_2)$ αντίστοιχα. Τα δείγματα x που πρέπει να κατηγοριοποιηθούν είναι δισδιάστατα (2D) και οι κλάσεις περιγράφονται από τις ακόλουθες κανονικές κατανομές:

$$P(x|\omega_1) = \mathcal{N}(\mu_1, \mathbf{\Sigma}_1), \quad P(x|\omega_2) = \mathcal{N}(\mu_2, \mathbf{\Sigma}_2)$$

Όπου

$$\mu_1 = \begin{pmatrix} 3 \\ 3 \end{pmatrix} \quad \boldsymbol{\Sigma}_1 = \begin{pmatrix} 1.2 & -0.4 \\ -0.4 & 1.2 \end{pmatrix}$$
$$\mu_2 = \begin{pmatrix} 6 \\ 6 \end{pmatrix} \quad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 1.2 & 0.4 \\ 0.4 & 1.2 \end{pmatrix}$$

a) Βρείτε μια έκφραση για το σύνορο απόφασης (διαχωρισμού)
b) Σχεδιάστε μερικές ισοϋψείς καμπύλες των δεσμευμένων πιθανοτήτων $p(x|\omega_i) \in \mathbb{R}^2$ για κάθε κλάση i χρησιμοποιώντας Matlab/Octave ή άλλο σχετικό λογισμικό.
c) Σχεδιάστε στην ίδια εικόνα τα σύνορα απόφασης θεωρώντας τις ακόλουθες τιμές της εκ των προτέρων πιθανότητας $P(\omega_1) = 0.1, 0.25, 0.5, 0.75, 0.9$.
d) Σχολιάστε τα αποτελέσματα. Ποια είναι η μορφή των συνόρων απόφασης και γιατί; Πως επηρεάζεται το σύνορο απόφασης από τις διαφορετικές τιμές της εκ των προτέρων πιθανότητας.
e) Επαναλάβετε τις παραπάνω ερωτήσεις υποθέτοντας ότι οι πίνακες συνδιασποράς είναι ίδιοι:

$$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \begin{pmatrix} 1.2 & 0.4 \\ 0.4 & 1.2 \end{pmatrix}$$

Θέμα 5: Minimum risk

Έστω ότι σε ένα πρόβλημα κατηγοριοποίησης σε δύο κλάσεις ω_1 και ω_2 οι εκ των προτέρων πιθανότητες είναι ίσες $P(\omega_1) = P(\omega_2)$. Τα δείγματα x που πρέπει να κατηγοριοποιηθούν είναι μονοδιάστατα (1D) και ακολουθούν κατανομή Rayleigh με συνάρτηση πυκνότητας πιθανότητας που δίνεται από την ακόλουθη σχέση:

$$p(x|\omega_i) = \begin{cases} \dfrac{x}{\sigma_i^2} e^{-\frac{x^2}{2\sigma_i^2}} & x \geq 0 \\ 0, & x < 0 \end{cases}$$

Με $\sigma_1 = 1$ και $\sigma_2 = 2$. Υπολογίστε το όριο απόφασης x_0 που έχει το μικρότερο ρίσκο, δεδομένου ότι ο πίνακας ρίσκου είναι:

$$\boldsymbol{L} = \begin{pmatrix} 0 & 0.5 \\ 1.0 & 0 \end{pmatrix}$$