ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΤΜΗΜΑ ΜΗΧ/ΚΩΝ Η/Υ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ ΚΑΙ ΔΙΚΤΥΩΝ

'ΑΝΑΓΝΩΡΙΣΗ ΠΡΟΤΥΠΩΝ' (ΗΥ345)

Ακαδ. Έτος **2012 - 2013**

ΕΡΓΑΣΙΑ MATLAB

Παρατηρήσεις:

1. Στόχος της εργασίας είναι η εμπέδωση σημαντικών εννοιών του μαθήματος αλλά και γενικότερα η εξοικείωση με τις εφαρμογές του περιβάλλοντος MATLAB στην 'Αναγνώριση Προτύπων'.
2. Η εργασία μετράει κατά 25/100 της συνολικής βαθμολογίας τους μαθήματος και μπορεί να παραδοθεί από ομάδες μέχρι 3 ατόμων.
3. Ομάδες 2 ατόμων πρέπει να απαντήσουν στα Μέρη Α-Δ της εργασίας για να πάρουν πλήρη βαθμολογία.
4. Ομάδες 3 ατόμων πρέπει να απαντήσουν και στο Μέρος Ε της εργασίας για να πάρουν πλήρη βαθμολογία.
5. Καταληκτική ημερομηνία παράδοσης: Έως και την ημέρα/ώρα του τελικού διαγωνίσματος.
6. Καμία παράταση δεν θα δοθεί.
7. Οι εργασίες ίσως να εξεταστούν και προφορικά. Εάν αυτό κριθεί απαραίτητο, λεπτομέρειες θα ανακοινωθούν έγκαιρα.
8. Η βαθμολογία μιας εργασίας θα εξαρτηθεί από την ορθότητα των αποτελεσμάτων και τον τρόπο παρουσίασής τους. Θα αξιολογηθούν ακόμα και ημιτελείς προσπάθειες αρκεί να έχουν παραδοθεί εμπρόθεσμα οι σχετικές αναφορές.
9. Ο κώδικας που έχει χρησιμοποιηθεί θα πρέπει να εκτυπωθεί και να αποτελέσει μέρος της αναφοράς (με σχόλια).
10. Η αναφορά πρέπει να παραδοθεί σε εκτυπωμένη μορφή.
11. Ηλεκτρονικός τρόπος παράδοσης (με email) δεν είναι δεκτός.

ΜΕΡΟΣ Α: Έστω 4 <u>*ισοπίθανες*</u> κλάσεις, ω_1, ω_2, ω_3, και ω_4 ($P(\omega_i) = 1/4$), με διανύσματα χαρακτηριστικών σε δύο διαστάσεις ($\mathbf{x} \in \mathcal{R}^2$) που ακολουθούν για κάθε κλάση 2-D Gaussian κατανομές $p(\mathbf{x} \,|\, \omega_i)$ με μέσες τιμές

$$\underline{\mu_1} = [\,0\,,0\,]\,, \;\; \underline{\mu_2} = [\,2\,,2\,]\,, \;\; \underline{\mu_3} = [\,2\,,0\,]\,, \;\; \underline{\mu_4} = [\,4\,,2\,]\,, \tag{1}$$

αντίστοιχα, και κοινό <u>*διαγώνιο*</u> μητρώο συνδιασποράς, ίσο με

$$\mathbf{S} = \begin{bmatrix} 0.6 & 0.0 \\ 0.0 & 0.6 \end{bmatrix} \;. \tag{2}$$

Στη συνέχεια:

A.1 Δημιουργήστε ένα σύνολο εκπαίδευσης <u>*με επίβλεψη*</u> που αποτελείται από $N = 1000$ σημεία $\mathbf{x}_n$ στον χώρο $\mathcal{R}^2$ που «πηγάζουν» από τις παραπάνω 4 κλάσεις και κατανομές. Χρησιμοποιείστε για τον σκοπό αυτό την εντολή `mvnrnd` του Matlab με κατάλληλη παραλλαγή και παραμέτρους. Χρήσιμη είναι επίσης η αρχικοποίηση της γεννήτριας τυχαίων αριθμών με συγκεκριμένο τρόπο, π.χ. με την εντολή `randn('seed',0)`, ώστε να είναι εφικτή η επανάληψη του πειράματος και αναπαραγωγή των αποτελεσμάτων.

A.2 Σχεδιάστε τα δεδομένα στον 2-D χώρο. Χρησιμοποιείστε διαφορετικές « ετικέτες » (labels) στο διάγραμμα για κάθε κλάση.

A.3 Από τα δεδομένα $\mathbf{x}_n$ του συνόλου εκπαίδευσης που ανήκουν στην κάθε κλάση, εκτιμήστε τις μέσες τιμές και μητρώα συνδιασποράς των 4 υπό συνθήκη συναρτήσεων πυκνότητας πιθανότητας, χρησιμοποιώντας εκτίμηση μέγιστης πιθανοφάνειας (maximum likelihood estimation).

A.4 Ταξινομήστε τα δεδομένα $\mathbf{x}_n$ του συνόλου εκπαίδευσης στις 4 κλάσεις με βάση τον ταξινομητή Ευκλείδειας απόστασης, χρησιμοποιώντας τις εκτιμήσεις των παραμέτρων των $p(\mathbf{x} \,|\, \omega_i)$ από το Βήμα A.3. Υπολογίστε το λάθος ταξινόμησης (%), συγκρίνοντας τις αποφάσεις του ταξινομητή με τις ετικέτες των δεδομένων (από το Βήμα A.1).

A.5 Επαναλάβατε το πείραμα χρησιμοποιώντας τον ταξινομητή Mahalanobis απόστασης, χρησιμοποιώντας τις εκτιμήσεις των παραμέτρων των $p(\mathbf{x} \,|\, \omega_i)$ από το Βήμα A.3. Ειδικά για τον πίνακα συνδιασποράς, χρησιμοποιείστε έναν <u>*κοινό*</u> πίνακα $\hat{\mathbf{S}}$ ως τον σταθμισμένο μέσο των εκτιμήσεων των μητρώων συνδιασποράς του Βήματος A.3 για τις 4 κλάσεις. Υπολογίστε το λάθος ταξινόμησης και σχολιάστε το αποτέλεσμα σε σχέση με αυτό του Βήματος A.4.

A.6 Επαναλάβετε χρησιμοποιώντας τον Bayesian ταξινομητή, με βάση και πάλι τις εκτιμήσεις των παραμέτρων των $p(\mathbf{x} \,|\, \omega_i)$ του Βήματος A.3. Εν αντιθέσει με το Βήμα A.5, χρησιμοποιείστε τις εκτιμήσεις του πίνακα συνδιασποράς για κάθε κλάση ξεχωριστά. Υπολογίστε το λάθος ταξινόμησης και σχολιάστε το αποτέλεσμα σε σχέση με αυτά των Βημάτων A.4 και A.5.

ΜΕΡΟΣ Β: Επαναλάβετε τα παραπάνω για κλάσεις με κοινό αλλά *μη διαγώνιο* μητρώο συνδιασποράς, ίσο με

$$\mathbf{S} = \begin{bmatrix} 0.6 & 0.4 \\ 0.4 & 0.6 \end{bmatrix} \tag{3}$$

(Βήματα Β.1-Β.6). Οι πρότερες πιθανότητες των κλάσεων, $P(\omega_i)$, και οι μέσες τιμές των $p(\mathbf{x}\,|\,\omega_i)$ παραμένουν ως έχουν στο Μέρος Α (εξ. (1)). Σχολιάστε τις διαφοροποιήσεις μεταξύ των ταξινομητών Ευκλείδειας απόστασης, απόστασης Mahalanobis, και του Bayesian ταξινομητή σε σχέση με αυτές του Μέρους Α της εργασίας.

ΜΕΡΟΣ Γ: Επαναλάβετε τα παραπάνω στη συνέχεια για *μη ισοπίθανες κλάσεις*, και συγκεκριμένα για πρότερες (a-priori) πιθανότητες κλάσεων

$$P(\omega_1) = \frac{1}{2}\ , \quad P(\omega_2) = \frac{1}{16}\ , \quad P(\omega_3) = \frac{1}{16}\ , \quad P(\omega_4) = \frac{3}{8} \tag{4}$$

(βήματα Γ.1-Γ.6). Ισχύουν οι εξ. (1) (από το Μέρος Α) και εξ. (3) (από το Μέρος Β) για τις μέσες τιμές και μητρώα συνδιασποράς αντίστοιχα. Όπως και πριν, σχολιάστε τις διαφοροποιήσεις μεταξύ των ταξινομητών Ευκλείδειας απόστασης, απόστασης Mahalanobis, και του Bayesian ταξινομητή σε σχέση με αυτές των Μερών Α και Β της εργασίας.

ΜΕΡΟΣ Δ: Επανέρχεστε στο Μέρος Β, δηλαδή στις 4 ισοπίθανες κλάσεις με υπό συνθήκη συναρτήσεις πυκνότητας πιθανότητας των χαρακτηριστικών που είναι 2-D Gaussian με μέσες τιμές της εξ. (1) και κοινά, μη διαγώνια μητρώα συνδιασποράς της εξ. (3). Στο μέρος αυτό της εργασίας θα θεωρήσετε ότι έχετε δύο κλάσεις που προκύπτουν από την ένωση των προηγούμενων κλάσεων, και θα ταξινομήσετε τα δεδομένα χρησιμοποιώντας την προβολή τους σε κατάλληλη ευθεία, με βάση τους μετασχηματισμούς PCA και LDA. Για τον σκοπό αυτό:

- Δ.1 Όπως και στα προηγούμενα μέρη, δημιουργήστε το σύνολο εκπαίδευσης 1000 σημείων, που «πηγάζουν» από τις 4 κλάσεις και κατανομές.
- Δ.2 Συνδυάστε τις κλάσεις σε δύο, δηλαδή την $\Omega_1 = \omega_1 \cup \omega_2$ και $\Omega_2 = \omega_3 \cup \omega_4$.
- Δ.3 Σχεδιάστε τα δεδομένα των δύο κλάσεων στον 2-D χώρο.
- Δ.4 Θεωρήστε ότι οι δύο κλάσεις μοντελοποιούνται με μία 2-D Gaussian η καθεμία (δηλ. η υπό συνθήκη συνάρτηση πυκνότητας πιθανότητας των χαρακτηριστικών δοθείσης της κλάσης). Υπολογίστε στη συνέχεια τις μέσες τιμές και μητρώα συνδιασποράς των δύο Gaussians χρησιμοποιώντας εκτίμηση μέγιστης πιθανοφάνειας.
- Δ.5 Ταξινομήστε τα δεδομένα στις δύο κλάσεις με βάση τον ταξινομητή Ευκλείδειας απόστασης και τον Bayesian ταξινομητή. Σε κάθε περίπτωση, υπολογίστε το σφάλμα ταξινόμησης (%).

Δ.6 Στη συνέχεια, βρείτε την μονοδιάστατη προβολή των δεδομένων του συνόλου εκπαίδευσης κατά μήκος του ιδιο-διανύσματος που αντιστοιχεί στην μεγαλύτερη ιδιοτιμή με βάση τον μετασχηματισμό PCA. Σχεδιάστε τα σημεία που προκύπτουν. Τι παρατηρείτε;

Δ.7 Ταξινομήστε τα σημεία που προκύπτουν χρησιμοποιώντας τον μονοδιάστατο ταξινομητή Ευκλείδειας απόστασης και τον Bayesian ταξινομητή (για τον τελευταίο χρησιμοποιείστε 1-D Gaussians για κάθε κλάση). Υπολογίστε το σφάλμα ταξινόμησης.

Δ.8 Στη συνέχεια, βρείτε πάλι την μονοδιάστατη προβολή των δεδομένων του συνόλου εκπαίδευσης, αλλά αυτήν τη φορά με βάση τον μετασχηματισμό LDA. Σχεδιάστε τα σημεία που προκύπτουν. Τι παρατηρείτε σε σχέση και με το Βήμα Δ.6;

Δ.9 Ταξινομήστε τα σημεία που προκύπτουν χρησιμοποιώντας τον μονοδιάστατο ταξινομητή Ευκλείδειας απόστασης και τον Bayesian ταξινομητή (για τον τελευταίο χρησιμοποιείστε 1-D Gaussians για κάθε κλάση). Υπολογίστε το σφάλμα ταξινόμησης. Τι παρατηρείτε σε σχέση και με το Βήμα Δ.7;

ΜΕΡΟΣ Ε: (Υποχρεωτικό μόνο για ομάδες 3 ατόμων) Το μέρος αυτό ακολουθεί το Μέρος Δ της εργασίας, αλλά στοχεύει σε μοντελοποίηση με βάση μείγμα από Gaussians (GMM).

Ε.1 Επαναλάβατε τα βήματα Δ.1-Δ.3.

Ε.2 Θεωρήστε ότι οι δύο κλάσεις μοντελοποιούνται με ένα μείγμα από δύο 2-D Gaussians η καθεμία. Τρέξτε τον αλγόριθμο Expectation-Maximization για ένα αριθμό επαναλήψεων, αρχικοποιώντας τις μέσες τιμές με μία μικρή διατάραξη (perturbation) από τις τιμές του βήματος Δ.4 (Gaussian mixture splitting).

Ε.3 Σε κάθε επανάληψη του αλγόριθμου υπολογίστε την συνολική πιθανοφάνεια των δεδομένων, όπως και το σφάλμα ταξινόμησης του προκύπτοντα Bayesian ταξινομητή. Τι παρατηρείτε;

Ε.4 Επαναλάβατε τα βήματα Ε.2 και Ε.3 αρχικοποιώντας τις συνολικά 4 συνιστώσες (δύο συνιστώσες μειγμάτων για κάθε μία από τις δύο κλάσεις) με τις αντίστοιχες μέσες τιμές της εξ. (1) και μητρώο συνδιασποράς της εξ. (3). Τι παρατηρείτε σε σχέση με τα διαγράμματα του Βήματος Ε.3;