ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ

Τμήμα Ηλεκτρολόγων Μηχ. Και Μηχ. Υπολογιστών

Στατιστική Μοντελοποίηση και Αναγνώριση Προτύπων (ΤΗΛ311)

$1^{η}$ Σειρά Ασκήσεων

Οδηγίες:

1. Παρακαλείστε να σεβαστείτε τον παρακάτω <u>κώδικα τιμής</u> τον οποίον θα θεωρηθεί ότι προσυπογράφετε μαζί με τη συμμετοχή σας στο μάθημα και τις εργασίες του:
   1) *Οι απαντήσεις, ο κώδικας και γενικά οτιδήποτε αφορά τις εργασίες, τα φυλλάδια ασκήσεων και τις εξετάσεις του μαθήματος θα είναι προϊόν δικής μου δουλειάς.*
   2) *Δεν θα διαθέσω κώδικα, απαντήσεις και εργασίες μου σε κανέναν άλλο.*
   3) *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 θα γίνει σε **ομάδες έως 2 ατόμων**
3. Ημερομηνία παράδοσης: **<u>Κυριακή, 19/5/2024 στις 23:00</u>**
4. **<u>Ανεβάστε στο eclass τα ακόλουθα παραδοτέα σε μορφή zip:</u>**
   <u>Παραδοτέα:</u> α) Κώδικας και β) Αναφορά με τις απαντήσεις, παρατηρήσεις, πειράματα, αποτελέσματα και οδηγίες χρήσης του κώδικα.
5. **<u>ΠΡΟΣΟΧΗ: Δεν επιτρέπεται να χρησιμοποιήσετε έτοιμες υλοποιήσεις των αλγορίθμων που σας ζητούνται εκτός αν αναφέρεται ρητά.</u>**
6. **Την <u>εβδομάδα μετά την παράδοση</u> θα γίνει προφορική εξέταση κάθε ομάδας πάνω στα παραδοτέα. Λεπτομέρειες θα ανακοινωθούν στο eclass.**

Θέμα 1: Principal Component Analysis (PCA)

Σε αυτή την άσκηση, θα χρησιμοποιήσετε την μέθοδο Principal Component Analysis (PCA) αρχικά πάνω σε ένα σύνολο δεδομένων που περιγράφουν καλοήθεις και κακοήθεις όγκους στο στήθος και στη συνέχεια θα χρησιμοποιήσετε το PCA σε ένα μεγαλύτερο σύνολο δεδομένων 5000 εικόνων προσώπων.

Μέρος 1

Συμπληρώστε κατάλληλα τα matlab/octave scripts που σας δίνονται έτσι ώστε εκτελώντας το αρχικό script `ex1_1_pca.m` αυτό να κάνει τα ακόλουθα:

a) Αρχικά θα διαβάζει το αρχείο `'data/breast_cancer_data.csv'` το οποίο περιέχει 569 δείγματα το καθένα εκ των οποίων περιγράφεται με ένα διάνυσμα 30 χαρακτηριστικών. Η τελευταία τιμή στο διάνυσμα είναι η κατηγοριοποίηση του δείγματος σε καλοήθες (0) και κακοήθες (1).
b) Επιλέξτε ζεύγη χαρακτηριστικών και απεικονίστε τα δείγματα στον 2D χώρο. Χρησιμοποιήστε διαφορετικό χρώμα για κάθε κατηγορία και παρατηρήστε πόσο διαφοροποιούνται οι κλάσεις ανάλογα με τα ζεύγη των χαρακτηριστικών.
a) Εφαρμόζει standardization (κανονικοποίηση με μέση τιμή μηδέν και διασπορά 1) στα αρχικά δείγματα (προσθέστε κώδικα στο αρχείο featureNormalize.m). Σχεδιάστε τα κανονικοποιημένα δείγματα σε διαφορετική εικόνα και συγκρίνετε με τα αρχικά.

b) Υπολογίζει τις κύριες συνιστώσες του αλγορίθμου PCA και τις σχεδιάζει στην ίδια εικόνα μαζί με τα αρχικά δείγματα (προσθέστε κώδικα στο αρχείο myPCA.m). Παρατηρήστε τι μορφή έχουν οι κύριες συνιστώσες και σχολιάστε σχετικά. Για να υπολογίσετε τον πίνακα συνδιασποράς κανονικοποιημένων δεδομένων χρησιμοποιήστε τον τύπο $\Sigma = \frac{1}{m} X^T X$, όπου $X \in \mathbb{R}^{m\times n}$ είναι ένας πίνακας όπου κάθε γραμμή αντιστοιχεί σε ένα παράδειγμα με n χαρακτηριστικά. Για να υπολογίσετε ιδιοτιμές και ιδιοδιανύσματα χρησιμοποιήστε την συνάρτηση του Matlab `eig()`.

c) Υπολογίζει την συνεισφορά που έχει κάθε κύρια συνιστώσα (explained variance) στην συνολική διακύμανση και ακολούθως εφαρμόζει τον αλγόριθμο PCA στα δείγματα για να μειώσει τη διάστασή τους από 2D σε 1D. Θα πρέπει να συμπληρώσετε τον κώδικα στο projectData.m έτσι ώστε αυτό να επιστρέφει τα δείγματα Z μειωμένης διάστασης. Η είσοδος του projectData θα είναι το σύνολο δεδομένων X, οι κύριες συνιστώσες U και ο επιθυμητός αριθμός διαστάσεων K. Θα πρέπει να προβάλλετε κάθε δείγμα $\boldsymbol{x}_i$ στις K κύριες συνιστώσες με την εφαρμογή του γραμμικού μετασχηματισμού (προβολή χαμηλότερης διάστασης) $\boldsymbol{z_i} = U^T \boldsymbol{x}_i$. Σημείωση: οι K κύριες συνιστώσες δίνονται από τις πρώτες K στήλες του U.

d) Αφού προβάλλετε τα δεδομένα σε χώρο μικρότερων διαστάσεων, μπορείτε να ανακτήσετε προσεγγιστικά τα δεδομένα αναπαράγοντάς τα ξανά στον αρχικό χώρο υψηλών διαστάσεων (2D). Αυτό γίνεται με την προβολή τους πάνω σε όλες τις κύριες συνιστώσες (principal components) με τον μετασχηματισμό $\boldsymbol{x}_{rec} = U\boldsymbol{z}$. Υλοποιήστε το recoverData.m για να προβάλετε κάθε δείγμα στο $Z \in \mathbb{R}^{n\times K}$ πίσω στον αρχικό χώρο και να επιστρέψετε την ανακτημένη προσέγγιση στον $X_{rec} \in \mathbb{R}^{n\times m}$ ο οποίος περιέχει τα ανακτημένα δείγματα. Αν χρησιμοποιήστε την πρώτη κύρια συνιστώσα πως εμφανίζονται τα δείγματα στον 2D χώρο;

e) Επαναλάβετε την ανάλυση PCA χρησιμοποιώντας αυτή τη φορά όλο το διάνυσμα χαρακτηριστικών (30 χαρακτηριστικά). Ακολούθως επιλέξτε τις δύο πρώτες κύριες συνιστώσες και σχεδιάστε την προβολή των δειγμάτων σε αυτές. Τι παρατηρείτε; Σχεδιάστε τις κύριες συνιστώσες στο ίδιο γράφημα με τα χαρακτηριστικά που εμφανίζουν την μεγαλύτερη διακύμανση και σχολιάστε σε τι διαφέρουν σε σχέση με την μορφή που είχαν στο βήμα (b).

Μέρος 2

Επαναλάβετε τα παραπάνω χρησιμοποιώντας τα δεδομένα 5000 εικόνων προσώπων που σας δίνονται (`load('data\faces.mat')`). Συγκεκριμένα χρησιμοποιώντας τον κώδικα που φτιάξατε πιο πάνω υπό την προϋπόθεση ότι μπορεί να εφαρμοστεί σε γενικευμένα dataset:

f) Σχεδιάστε τα πρώτα 100 πρόσωπα από το dataset χρησιμοποιώντας τη συνάρτηση displayData() που σας δίνεται.

g) Εφαρμόστε standardization εφαρμόζοντας την featureNormalize() πάνω στα δείγματα που έχετε και ακολούθως υπολογίστε τις κύριες συνιστώσες εφαρμόζοντας τη συνάρτηση myPCA() που φτιάξατε. Σχεδιάστε σε μια νέα εικόνα τις πρώτες 36 κύριες συνιστώσες που βρήκατε με την συνάρτηση displayData(). Τι παρατηρείτε;

h) Μειώστε την διάσταση των δειγμάτων σας χρησιμοποιώντας τις 100 πρώτες κύριες συνιστώσες με την projectData().

i) Σχεδιάστε τα δείγματα μειωμένης διάστασης αφού προηγουμένως τα προβάλετε στον αρχικό χώρο χρησιμοποιώντας τη συνάρτηση recoverData(). Τι παρατηρείτε; Δοκιμάστε να επαναλάβετε την διαδικασία χρησιμοποιώντας διαφορετικό αριθμό από κύριες συνιστώσες (10, 50, 200)

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

Ένα σύνολο δεδομένων έχει προκύψει από δύο κατηγορίες ω_A, ω_B , οι κατανομές των οποίων θεωρούνται Γκαουσιανές και η a-Priori πιθανότητα της πρώτης κατηγορίας είναι: $\mathrm{P}(\omega_A) = 0.4$. Οι πίνακες συνδιασποράς και οι μέσες τιμές έχουν εκτιμηθεί από τα δεδομένα ως:

$$\mu_A = \begin{bmatrix} 0 \\ 2 \end{bmatrix}, \ \ \mu_B = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$$

$$\Sigma_A = \begin{bmatrix} 3 & 1 \\ 1 & 2 \end{bmatrix}, \ \ \Sigma_B = \begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix}$$

1) Εφαρμόστε τη μέθοδο LDA αναλυτικά για να υπολογίσετε το διάνυσμα προβολής w, κάνοντας τις πράξεις με το χέρι. Για να αντιστρέψετε τον πίνακα σκέδασης S_w (Within-Class Scatter Matrix) χρησιμοποιήστε τον τύπο:
$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$
2) Υπολογίστε την προβολή των δύο παρακάτω δειγμάτων, ένα από κάθε κατηγορία, στον άξονα που ορίζεται από το διάνυσμα w:
$$x_A = \begin{bmatrix} 1 \\ 3 \end{bmatrix}, \ \ x_B = \begin{bmatrix} 3 \\ -1 \end{bmatrix}$$

Θέμα 3: Linear Discriminant Analysis (LDA) vs PCA

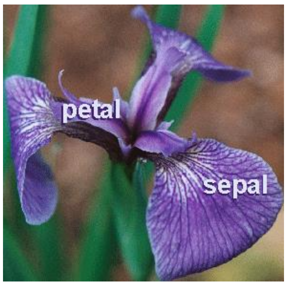
Ο σκοπός της άσκησης είναι να εφαρμόσετε Linear Discriminant Analysis για να μειώσετε τη διάσταση ενός feature vector και να συγκρίνετε τα αποτελέσματα σας με τη μέθοδο PCA.

Μέρος 1

Αρχικά θα εφαρμόσετε τον αλγόριθμο σε 2D τεχνητά δεδομένα δύο κλάσεων. Συγκεκριμένα για την άσκηση αυτή θα χρειαστεί να συμπληρώσετε κατάλληλα τον κώδικα στο αρχείο Matlab/Octave script ex1_2_lda.m και να το τρέξετε έτσι ώστε αυτό να κάνει τα ακόλουθα:

a) Φορτώνει τα δείγματα ως εξής: `load('data\data2.mat')`. Θα δημιουργηθούν οι μεταβλητές c που περιέχει την κλάση και η X που περιέχει τα χαρακτηριστικά διανύσματα.
b) Θα πρέπει να κάνετε standardization (κανονικοποίηση με $\mu = 0$ και $\sigma = 1$) χρησιμοποιώντας τη συνάρτηση featureNormalize που είχατε φτιάξει για το θέμα 1. Σχεδιάστε με διαφορετικό χρώμα τα δείγματα της κάθε κλάσης στην ίδια εικόνα.
c) Προσθέστε τον απαραίτητο κώδικα στο αρχείο «fisherLinearDiscriminant.m» έτσι ώστε αυτό να υλοποιεί τον αλγόριθμο LDA για προβλήματα 2 κλάσεων.
d) Προσθέστε τον κατάλληλο κώδικα στο αρχείο «projectDataLDA.m» ώστε τρέχοντας τη συνάρτηση να δημιουργείτε δείγματα μειωμένης διάσταση (1D)
e) Προσθέστε τον κατάλληλο κώδικα στο αρχείο «recoverDataLDA.m» ώστε να κάνετε ανακατασκευή των δειγμάτων μειωμένης διάστασης στον 2D χώρο προβάλλοντας τα πάνω στην κατεύθυνση του διανύσματος προβολής LDA. Σχεδιάστε πάνω στην ίδια εικόνα τα δείγματα μειωμένης διάστασης όπως προβάλλονται στον 2D χώρο μαζί με τα αρχικά σας δείγματα.
f) Εφαρμόστε την αντίστοιχη διαδικασία μείωσης διάστασης με τη μέθοδο PCA όπως την εφαρμόσατε στο θέμα 1. Σχεδιάστε τα δείγματα που προκύπτουν με τη μέθοδο PCA και συγκρίνετε την προβολή τους με την αντίστοιχη LDA.

Μέρος 2



Σ’ αυτό το μέρος θα εφαρμόσετε τον αλγόριθμο LDA σε ένα πολύ δημοφιλές dataset στο χώρο της Μηχανικής Μάθησης, τη βάση δεδομένων Iris. Η βάση δεδομένων Iris ή Fisher’s Iris, περιέχει 50 δείγματα από 3 διαφορετικά είδη της οικογένειας του λουλουδιού του είδους Iris (Iris Setosa, Iris Versicolor, Iris Virginica). Κάθε δείγμα αποτελείται από τέσσερα χαρακτηριστικά/μετρήσεις των λουλουδιών και συγκεκριμένα το μήκος και το πλάτος (σε cm) των σέπαλων και των ανθόφυλλων. Για το συγκεκριμένο dataset προσθέστε κατάλληλα κώδικα ο οποίος θα κάνει τα ακόλουθα:

a) Φορτώνει τα δείγματα iris ως εξής: `load('fisheriris.mat').` Δεν χρειάζεται να σας δοθεί το αρχείο fisheriris.mat, υπάρχει έτοιμο στο Matlab. Με την παραπάνω εντολή θα δημιουργηθεί η μεταβλητή $meas$ που περιέχει τα 150 δείγματα (χαρακτηριστικά διανύσματα) και η μεταβλητή $species$ που περιέχει την κλάση για κάθε δείγμα (είναι σε μορφή cell).
b) Εφαρμόστε standardization στα δείγματα και ακολούθως σχεδιάστε τα 2 πρώτα χαρακτηριστικά (features) σε ένα 2D γράφημα χρησιμοποιώντας ξεχωριστά χρώματα για κάθε κλάση.
c) Προσθέστε τον απαραίτητο κώδικα στη συνάρτηση myLDA η οποία θα δέχεται στην είσοδο τα κανονικοποιημένα δείγματα, τα labels της κλάσης στην οποία ανήκουν και τον αριθμό των διαστάσεων ($NewDim$) στις οποίες θέλουμε να μειώσουμε το χώρο των χαρακτηριστικών. Επιλέξτε η νέα διάσταση του διανύσματος των χαρακτηριστικών να είναι $NewDim$=2. Αυτό που θα πρέπει να επιστρέφει είναι ένας πίνακας που να περιέχει στις στήλες του τα διανύσματα προβολής LDA. Για να υλοποιήσετε τη συνάρτηση θα πρέπει να μπορεί να υπολογίζει από τα δεδομένα εισόδου:
    a. Τις prior πιθανότητες των Κλάσεων
    b. Τις μέσες τιμές των Κλάσεων
    c. Τον ολικό μέσο
    d. Τον πίνακα σκέδασης S_w (Within-Class Scatter Matrix)
    e. Τον πίνακα σκέδασης S_b (Between-Class Scatter Matrix)
    f. Τον πίνακα $S_w^{-1}S_b$ του γενικευμένου συστήματος ιδιοτιμών στον οποίο θα πρέπει να εφαρμόσετε eigendecomposition. Δηλαδή υπολογίστε ιδιοτιμές και ιδιοδιανύσματα, ταξινομήστε τα με φθίνουσα σειρά ιδιοτιμής και κρατήστε τα ιδιοδιανύσματα που αντιστοιχούν στις μεγαλύτερες $NewDim = 2$ ιδιοτιμές
d) Στο τέλος εφαρμόστε τα διανύσματα προβολής που υπολογίσατε με την myLDA πάνω στα αρχικά σας δείγματα με τη συνάρτηση projectDataLDA για να μειώσετε τη διάσταση του σε 2. Σχεδιάστε τα σε ένα γράφημα στο επίπεδο (2D) χρησιμοποιώντας ξεχωριστά χρώματα για τις διαφορετικές κλάσεις.

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

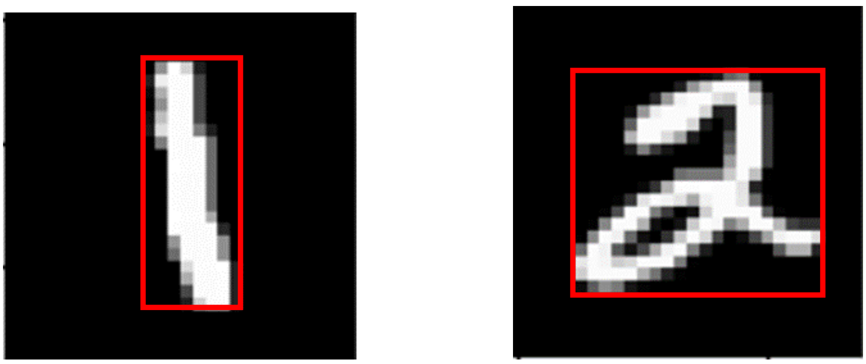
$$\mu_1 = \begin{pmatrix} 2 \\ 3 \end{pmatrix} \qquad \mathbf{\Sigma}_1 = \begin{pmatrix} 2 & 0.5 \\ 0.5 & 1 \end{pmatrix}$$
$$\mu_2 = \begin{pmatrix} 4 \\ 4 \end{pmatrix} \qquad \mathbf{\Sigma}_2 = \begin{pmatrix} 1.5 & -0.3 \\ -0.3 & 0.8 \end{pmatrix}$$

1) Υπολογίστε την εξίσωση του συνόρου απόφασης (Decision Boundary)
2) Χρησιμοποιήστε Matlab ή Python για να σχεδιάσετε τις ισοϋψείς καμπύλες των δύο κατανομών $p(x|\omega_i) \in \mathbb{R}^2$ για κάθε κλάση i.
3) Δεδομένου ότι $P(\omega_2) = 1 - P(\omega_1)$, σχεδιάστε στην ίδια εικόνα τα σύνορα απόφασης για τις ακόλουθες τιμές της εκ των προτέρων πιθανότητας $P(\omega_1) = 0.1, 0.25, 0.5, 0.75, 0.9$. 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.
4) Επαναλάβετε τις παραπάνω ερωτήσεις υποθέτοντας ότι οι πίνακες συνδιασποράς είναι ίδιοι:
$$\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \begin{pmatrix} 1.2 & 0.4 \\ 0.4 & 1.2 \end{pmatrix}$$

Θέμα 5: Εξαγωγή χαρακτηριστικών και Bayes Classification.

Στον κατάλογο exercise1_5/data υπάρχουν τα αρχεία *mnist_train.csv* και το *mnist_test.csv τα* οποία περιέχουν χειρόγραφες εικόνες των αριθμητικών ψηφίων [0,1,2,3,4,5,6,7,8,9] χωρισμένες σε δείγματα εκπαίδευσης το πρώτο και δείγματα ελέγχου το δεύτερο (*mnist_test.csv*). Σε αυτή την άσκηση θα προσπαθήσουμε να δημιουργήσουμε έναν απλό ταξινομητή Bayes ο οποίος θα μπορεί να ταξινομεί δείγματα σε δύο κλάσεις και συγκεκριμένα στις κλάσεις C_1 και C_2των ψηφίων **1** και **2** αντίστοιχα. Συνεπώς θα αγνοήσουμε αρχικά τα υπόλοιπα δείγματα.

Για να κάνουμε ταξινόμηση θα χρειαστεί να δημιουργήσουμε χαρακτηριστικά για κάθε δείγμα εικόνας. Ένα από τα χαρακτηριστικά που μπορούν να χρησιμοποιηθούν σε ένα τέτοιο πρόβλημα ταξινόμησης των δειγμάτων στις δύο κλάσεις είναι ο λόγος όψεως (aspect ratio), ο οποίος ορίζεται ως ο λόγος $width/height$, όπου $width$ και $height$ είναι το μήκος και το ύψος του ελάχιστου ορθογωνίου που περικλείει ένα ψηφίο, όπως φαίνεται στα παρακάτω παραδείγματα.



Μέρος Α

Συμπληρώστε τον κώδικα που λείπει στο αρχείο python *HW1-Bayes-MNIST.py* έτσι ώστε:

a) Να υπολογίζει τον λόγο όψεως (aspect ratio) όπως ορίζεται πιο πάνω για κάθε εικόνα που ανήκει στις κλάσεις C_1 και C_2 στο train και test set. Τυπώστε την ελάχιστη και μέγιστη τιμή του aspect ratio. Συμπληρώστε τον κώδικα που λείπει στη συνάρτηση `aspect_ratio`
b) Σχεδιάστε δύο οποιαδήποτε δείγματα από το dataset σας (ένα από κάθε κλάση) μαζί με ένα παραλληλόγραμμο (bounding box) στα όρια του aspect ratio που υπολογίσατε (όπως στις παραπάνω εικόνες) για να βεβαιωθείτε ότι υπολογίζεται σωστά.

c) Πριν χρησιμοποιήσετε το χαρακτηριστικό, εφαρμόστε min-max κανονικοποίηση στο aspect ratio στο διάστημα [-1, 1] συμπληρώνοντας κατάλληλα τη συνάρτηση `min_max_scaling`.

d) Υπολογίστε από τα δείγματα τις εκ των προτέρων πιθανότητες $P(C_1)$ και $P(C_2)$ συμπληρώνοντας κατάλληλα τον κώδικα στη μέθοδο `train` της κλάσης "`MyBayesClassifier`"

e) Έστω ότι η κατανομή του aspect ratio χαρακτηριστικού σε κάθε κλάση ακολουθεί κανονική κατανομή με μέση τιμή $\bar{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i$ και τυπική απόκλιση $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{\mu})^2}$, όπου x_i τα δείγματα εκπαίδευσης της κλάσης των εικόνων. Εκπαιδεύστε έναν Bayes ταξινομητή στη μέθοδο `train`

f) Βρείτε ποια είναι η πρόβλεψη της κλάσης στα δείγματα ελέγχου χρησιμοποιώντας τον ταξινομητή που εκπαιδεύσατε (`predict`)

g) Εκτιμήστε την ακρίβεια ταξινόμησης στο σύνολο των δειγμάτων ελέγχου (test) και σχολιάστε το αποτέλεσμα.

h) Προσθέστε ως νέο χαρακτηριστικό τον αριθμό των μη μαύρων pixels που σχηματίζουν τον αριθμό σε κάθε εικόνα (`foreground_pixels`). Επαναλάβετε την διαδικασία εκπαίδευσης του Bayes ταξινομητή κάνοντας την υπόθεση ότι τα χαρακτηριστικά είναι ανεξάρτητα μεταξύ τους.

i) Προσθέστε ως τρίτο χαρακτηριστικό την τιμή του κεντροειδούς (κέντρο μάζας των μη μηδενικών pixels). Το κεντροειδές ορίζεται ως συντεταγμένες στο επίπεδο της εικόνας. Για να το μετασχηματίσετε σε μονοδιάστατο χαρακτηριστικό κάντε τον μετασχηματισμό ($centroid = centX + (centY * W)$), όπου W είναι το πλάτος της εικόνας, και centX, centY οι συντεταγμένες του κεντροειδούς που υπολογίσατε. Επαναλάβετε την διαδικασία εκπαίδευσης του Bayes ταξινομητή χρησιμοποιώντας όλα τα χαρακτηριστικά μαζί και κάνοντας την υπόθεση ότι τα χαρακτηριστικά είναι ανεξάρτητα μεταξύ τους.

j) Επαναλάβετε τα παραπάνω θεωρώντας ότι έχετε τρεις κλάσεις και συγκεκριμένα τα ψηφία 0, 1, 2. Πως μεταβάλλεται η ακρίβεια του ταξινομητή σ' αυτή την περίπτωση;

Θέμα 6: 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:

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

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:

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