Επεξεργασία Δεδομένων και Αλγόριθμοι Μάθησης

Αναφορά Σετ Ασκήσεων 3

Όνομα: Μακάριος Χρηστάκης

Περιεχόμενα

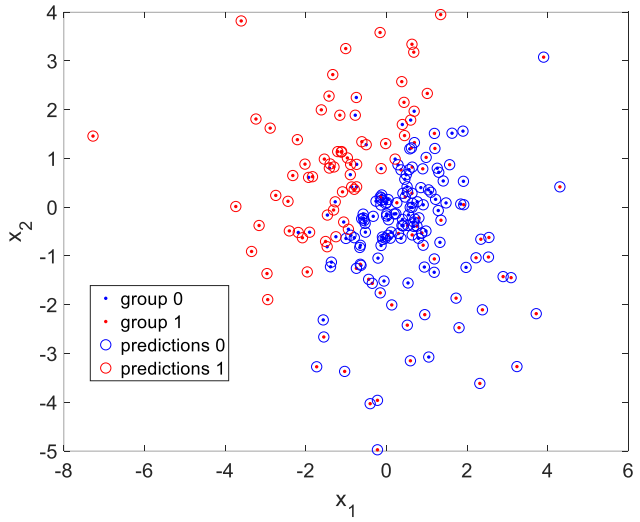
Πρόβλημα 1

Ερώτημα (α)

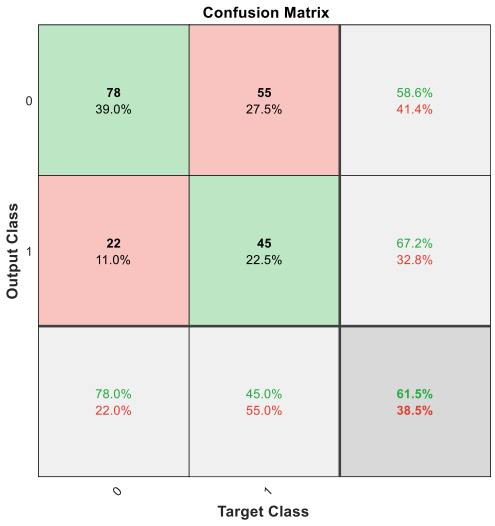
Στόχος μας είναι να ομαδοποιήσουμε τα δεδομένα μας σε 2 σύνολα, χωρίς να έχουμε κάποια πληροφορία για τα στατιστικά χαρακτηριστικά των δυο συνόλων αυτών. Χρησιμοποιούμε λοιπόν τον αλγόριθμο K-means παίρνοντας σαν αρχικά κέντρα δυο τυχαία δείγματα από το dataset μας. Τρέχουμε τον αλγόριθμο 10 φορές μέχρι να συγκλίνει με threshold 10^{-6} και κοιτάμε τη συνολική διασπορά όλων των ομάδων, δηλαδή το άθροισμα των τετραγώνων των ευκλείδειων αποστάσεων των δειγμάτων κάθε κατηγορίας από τα αντίστοιχα κέντρα της κατηγορίας τους. Από τις 10 επαναλήψεις του πειράματος (με διαφορετικά αρχικά κέντρα) κρατάμε αυτή με την ελάχιστη συνολική διασπορά.

Για να αξιολογήσουμε την απόδοση του αλγορίθμου χρησιμοποιούμε την κρυφή πληροφορία για να παρατηρήσουμε πόσα διανύσματα ταξινομούνται λανθασμένα και πόσα σωστά. Φυσικά φροντίζουμε οι κατηγορίες που προκύπτουν να είναι αριθμημένες έτσι ώστε να έχουμε λιγότερα σφάλματα απ' ότι σωστές κατηγοριοποιήσεις, μιας και ο K-means απλά ομαδοποιεί τα δεδομένα, χωρίς να ξέρει ποιο αυθαίρετο label δώσαμε στην κάθε ομάδα δεδομένων.

Τα αποτελέσματα φαίνονται στα Σχήμα 1 και Σχήμα 2.



Σχήμα 1: Δεδομένα κάθε ομάδας(τελείες) και Προβλέψεις(κύκλοι)



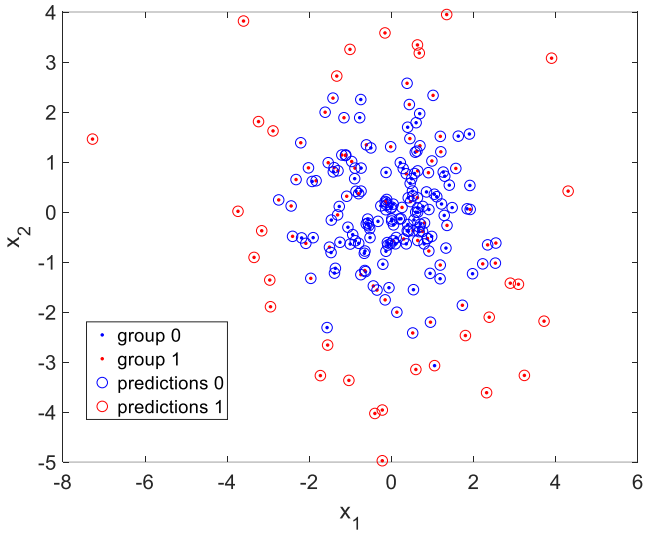
Σχήμα 2: Πίνακας σύγχυσης για την κατηγοριοποίηση των 2-D δεδομένων

Όπως φαίνεται ο αλγόριθμος K-means στα 2D δεδομένα μας είχε <u>ποσοστό σφάλματος</u> ***38,5%***

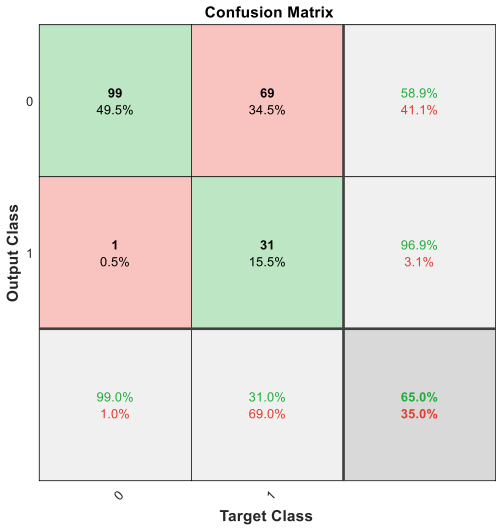
Ερώτημα (β)

Χρησιμοποιώντας έμμεσα την μυστική πληροφορία παρατηρείται ότι τα σημεία της μιας κατηγορίας είναι πιο κοντά στην αρχή των αξόνων απ' ότι τα σημεία της άλλης. Έτσι, θέλοντας να αυξήσουμε τεχνητά τη διάσταση του dataset μας για να βελτιώσουμε την απόδοση του K-means, ορίζουμε σαν $3^{η}$ διάσταση για κάθε διάνυσμα να είναι η νόρμα στο τετράγωνο του εκάστοτε αρχικού 2-D διανύσματος.

Ξανατρέχοντας με τον ίδιο τρόπο τον αλγόριθμο K-means παίρνουμε τα αποτελέσματα των Σχήμα 3 και Σχήμα 4.



Σχήμα 3: Δεδομένα κάθε ομάδας(τελείες) και Προβλέψεις(κύκλοι)



Σχήμα 4: Πίνακας σύγχυσης για την κατηγοριοποίηση των 3-D δεδομένων

Όπως φαίνεται ο αλγόριθμος K-means στα 3D δεδομένα είχε <u>ποσοστό σφάλματος</u> ***35%*** μόλις 3.5% χαμηλότερο από την περίπτωση των 2D δεδομένων.

Παρατηρούμε ωστόσο πως για τα δεδομένα της ομάδας 0 που είναι πιο κοντά στην αρχή των αξόνων έγινε μόνο 1 σφάλμα στα 100 παραδείγματα, ενώ αυξήθηκαν οι λανθασμένες κατηγοριοποιήσεις στα δεδομένα της ομάδας 1. Κατά μια έννοια θα μπορούσε να ειπωθεί ότι ο κατηγοριοποιητής αυτός τείνει να κατηγοριοποιεί περισσότερα παραδείγματα στην ομάδα 0 παρά στην 1, λόγω της έξτρα διάστασης που προστέθηκε. Τέλος, παρατηρήθηκε ότι ο K-means με 3-D δεδομένα συνέκλινε με λιγότερες επαναλήψεις από την περίπτωση των 2D δεδομένων (5 επαναλήψεις για τα 3D δεδομένα σε σχέση με 12 επαναλήψεις για τα 2D δεδομένα).

Πρόβλημα 2

Ερώτημα (α)

Στο πρόβλημα αυτό θα εφαρμόσουμε τη μέθοδο expectation/maximization για να εκτιμήσουμε τις παραμέτρους μιας μίξης δυο 2D Gaussian κατανομών με βάση τα δεδομένα μας. Θεωρούμε το σύνολο παραμέτρων θ της μίξης αυτής:

$$\theta = \{w_1, w_2, \boldsymbol{\mu_1}, \boldsymbol{\mu_2}, \boldsymbol{\Sigma_1}, \boldsymbol{\Sigma_2}\}, \quad \boldsymbol{\mu_i} \in \mathbb{R}^2 \, , \boldsymbol{\Sigma_i} \in \mathbb{R}^{2\times 2}, w_i \in \mathbb{R}$$

Και επιπλέον θεωρούμε πως η πυκνότητα πιθανότητας που ακολουθούν τα διανύσματα $\boldsymbol{x}$ των παραδειγμάτων μας είναι άθροιση 2 διαφορετικών πυκνοτήτων μέσω μιας «κρυφής» τυχαίας μεταβλητής j. Έτσι θα έχουμε:

$$f(\boldsymbol{x}, j|\theta) = w_j \frac{e^{-1/2(\boldsymbol{x}-\boldsymbol{\mu_j})^T \boldsymbol{\Sigma_j^{-1}} (\boldsymbol{x}-\boldsymbol{\mu_j})}}{\sqrt{(2\pi)^2 |\boldsymbol{\Sigma_j}|}}$$

Και η εκ των υστέρων πιθανότητα ενός διανύσματος $\boldsymbol{x_i}$ να ανήκει στην j Gaussian κατανομή είναι:

$$Q_i(j|\theta) = p_{i,j} = \frac{f(\boldsymbol{x_i}, j|\theta)}{f(\boldsymbol{x_i}|\theta)} = \frac{\frac{w_j e^{-1/2(\boldsymbol{x_i}-\boldsymbol{\mu_j})^T \boldsymbol{\Sigma_j^{-1}} (\boldsymbol{x_i}-\boldsymbol{\mu_j})}}{\sqrt{(2\pi)^2 |\boldsymbol{\Sigma_j}|}}}{\sum_{j=1}^{2} \frac{w_j e^{-1/2(\boldsymbol{x_i}-\boldsymbol{\mu_j})^T \boldsymbol{\Sigma_j^{-1}} (\boldsymbol{x_i}-\boldsymbol{\mu_j})}}{\sqrt{(2\pi)^2 |\boldsymbol{\Sigma_j}|}}} \tag{1}$$

Εξίσωση 1: Εκτίμηση της a posteriori πιθανότητας ενός διανύσματος να προέρχεται από την εκάστοτε συνιστώσα του μοντέλου.

Στόχος μας είναι να μεγιστοποιήσουμε την συνολική πιθανοφάνεια των δειγμάτων του dataset:

$$\max_{\theta} \sum_{i=1}^{N} \log\left(f(\boldsymbol{x_i}|\theta)\right)$$

Πρόβλημα το οποίο μπορεί να μετασχηματιστεί όπως απεδείχθη στη θεωρία του μαθήματος σε μια επαναληπτική μέθοδο, όπου στην επανάληψη t με σύνολο παραμέτρων θ_t υπολογίζουμε το σύνολο των καινούριων παραμέτρων μέσω της παράστασης:

$$\max_{\theta} \sum_{i=1}^{N} \sum_{j=1}^{2} Q_i(j|\theta_t) \log \frac{f(\boldsymbol{x_i}, j|\theta)}{Q_i(j|\theta_t)}$$

Οπού αντικαθιστώντας τις παραπάνω ποσότητες γίνεται:

$$\max_{\theta}\left\{\sum_{j=1}^{2}\log w_j\sum_{i=1}^{N}p_{i,j}(t)-\frac{1}{2}\sum_{j=1}^{2}\sum_{i=1}^{N}p_{i,j}(t)\left(\boldsymbol{x_i}-\boldsymbol{\mu_j}\right)^T\boldsymbol{\Sigma_j^{-1}}\left(\boldsymbol{x_i}-\boldsymbol{\mu_j}\right)\right.$$
$$\left.-\frac{1}{2}\sum_{j=1}^{2}\log\left|\boldsymbol{\Sigma_j}\right|\sum_{i=1}^{N}p_{i,j}(t)\right\} \qquad (2)$$

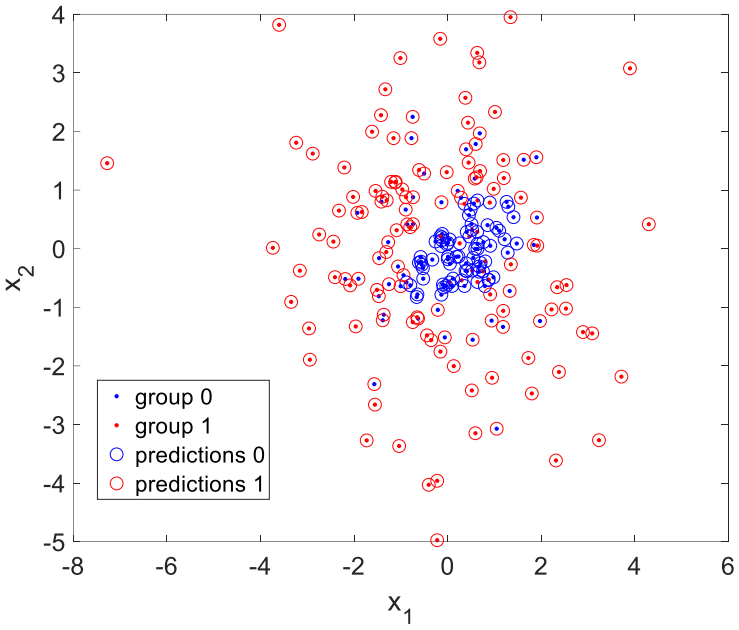
Εξίσωση 2: Το γενικό πρόβλημα βελτιστοποίησης για εκτίμηση παραμέτρων Gaussian Mixture Model

Παραγωγίζοντας ως προς τις παραμέτρους του στατιστικού μας μοντέλου παίρνουμε τις σχέσεις ανανέωσης των παραμέτρων αυτών, όπως αναπτύχθηκε στη θεωρία του μαθήματος. Χρησιμοποιώντας τη νέα εκτίμηση για το σύνολο παραμέτρων θ του μοντέλου μας μπορούμε μέσω της Εξίσωση 1 να υπολογίσουμε τις νέες εκ των υστέρων πιθανότητες ως εξής:

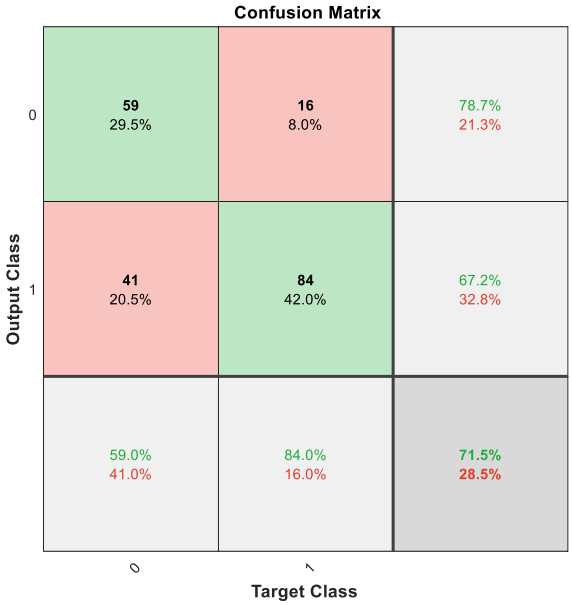
$$p_{i,j}(t+1)=\frac{\frac{w_j(t+1)e^{-1/2\left(\boldsymbol{x_i}-\boldsymbol{\mu_j}(t+1)\right)^T\boldsymbol{\Sigma_j^{-1}}(t+1)\left(\boldsymbol{x_i}-\boldsymbol{\mu_j}(t+1)\right)}}{\sqrt{(2\pi)^2\left|\boldsymbol{\Sigma_j}(t+1)\right|}}}{\sum_{j=1}^{2}\frac{w_j(t+1)e^{-1/2\left(\boldsymbol{x_i}-\boldsymbol{\mu_j}(t+1)\right)^T\boldsymbol{\Sigma_j^{-1}}(t+1)\left(\boldsymbol{x_i}-\boldsymbol{\mu_j}(t+1)\right)}}{\sqrt{(2\pi)^2\left|\boldsymbol{\Sigma_j}(t+1)\right|}}}$$

Οι a posteriori πιθανότητες αυτές θα χρησιμοποιηθούν στο επόμενο iteration του αλγορίθμου για να κάνουν την εκτίμηση των παραμέτρων του μοντέλου.

Εφαρμόζοντας την επαναληπτική μέθοδο μέχρι να συγκλίνει το συνολικό log-likelihood με κατώφλι 10^{-6} και χρησιμοποιώντας ίσες αρχικές σταθερές αναλογίας $w_j(t_0)=0.5$, μέσες τιμές για τις δυο gaussian δυο τυχαία παραδείγματα από το dataset και σαν αρχικό πίνακα διασποράς και για τις δυο κατηγορίες τον διαγώνιο πίνακα με στοιχεία τις διασπορές του πλήρους dataset μας για κάθε μια από τις 2 διαστάσεις του. Τα αποτελέσματα φαίνονται στα Σχήμα 5 και Σχήμα 6 παρακάτω:



Σχήμα 5: Δεδομένα κάθε ομάδας(τελείες) και Προβλέψεις(κύκλοι)



Σχήμα 6: Πίνακας σύγχυσης για την κατηγοριοποίηση με χρήση Gaussian Mixture Model

Χρησιμοποιώντας το gaussian mixture model για να κατηγοριοποιήσουμε τα δεδομένα μέσω των εκ των υστέρων πιθανοτήτων με τον κανόνα:

$$class\{x_i\} = \arg\max_{j} p_{i,j}\,(t_{final})$$

καταλήγουμε να παίρνουμε <u>ποσοστό σφάλματος</u> **28.5%**.

Ερώτημα (β)

Έχοντας σαν δεδομένο πως:

$$\boldsymbol{\mu_j} = [0 \quad 0]^T$$
$$w_j = \frac{1}{2}$$
$$\boldsymbol{\Sigma_j} = \begin{bmatrix} \sigma_j^2 & 0 \\ 0 & \sigma_j^2 \end{bmatrix}$$

Για κάθε j = {1,2}, η Εξίσωση 2 γίνεται:

$$\max_{\sigma_j^2} \left\{ \sum_{j=1}^{2} \log 0.5 \sum_{i=1}^{N} p_{i,j}(t) - \frac{1}{2} \sum_{j=1}^{2} \sum_{i=1}^{N} p_{i,j}(t) \mathbf{x_i^T} \frac{1}{\sigma_j^2} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \mathbf{x_i} - \frac{1}{2} \sum_{j=1}^{2} \log \sigma_j^4 \sum_{i=1}^{N} p_{i,j}(t) \right\}$$
$$\Leftrightarrow \max_{\sigma_j^2} \left\{ -\frac{1}{2} \sum_{j=1}^{2} \frac{1}{\sigma_j^2} \sum_{i=1}^{N} p_{i,j}(t) \mathbf{x_i^T x_i} - \frac{1}{2} \sum_{j=1}^{2} \log \sigma_j^4 \sum_{i=1}^{N} p_{i,j}(t) \right\} = \max_{\sigma_j^2} C\left(\sigma_j^2\right)$$

Οπότε παραγωγίζοντας έχουμε:

$$\frac{\partial C\left(\sigma_j^2\right)}{\partial \sigma_j^2} = \frac{1}{2} \frac{1}{\sigma_j^4} \sum_{i=1}^{N} p_{i,j}(t) \mathbf{x_i^T x_i} - \frac{1}{2} \frac{2\sigma_j^2}{\sigma_j^4} \sum_{i=1}^{N} p_{i,j}(t)$$
$$= \frac{1}{2\sigma_j^4} \sum_{i=1}^{N} p_{i,j}(t) \mathbf{x_i^T x_i} - \frac{1}{\sigma_j^2} \sum_{i=1}^{N} p_{i,j}(t)$$

Οπότε θέτοντας την παράγωγο ίση με μηδέν έχουμε:

$$\frac{\partial C\left(\sigma_j^2\right)}{\partial \sigma_j^2} = 0 \Leftrightarrow \frac{1}{2} \sum_{i=1}^{N} p_{i,j}(t) \mathbf{x_i^T x_i} = \sigma_j^2 \sum_{i=1}^{N} p_{i,j}(t)$$
$$\Leftrightarrow \sigma_j^2 = \frac{\sum_{i=1}^{N} p_{i,j}(t) \mathbf{x_i^T x_i}}{2 \cdot \sum_{i=1}^{N} p_{i,j}(t)}$$

Οπότε καταλήγουμε στο:

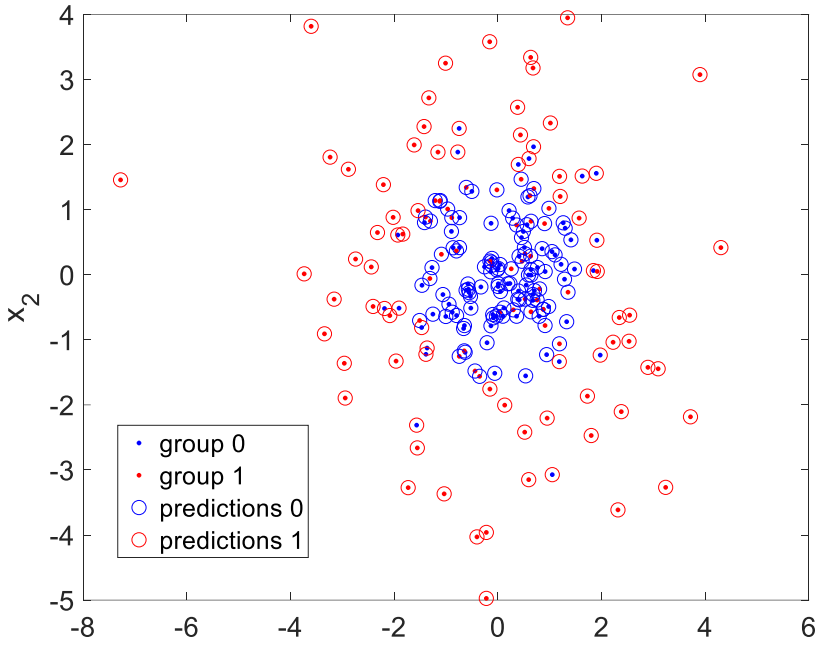
$$\boxed{\sigma_j^2 = \frac{\sum_{i=1}^{N} p_{i,j}(t) \|\mathbf{x_i}\|^2}{2 \cdot \sum_{i=1}^{N} p_{i,j}(t)}}$$

Συγκρίνοντας την παραπάνω σχέση με τον βέλτιστο εκτιμητή της διασποράς, δεδομένου ότι ξέρουμε ποια παραδείγματα ανήκουν στην εκάστοτε κατηγορία j = {1,2}, με μηδενική μέση τιμή για κάθε κατηγορία:

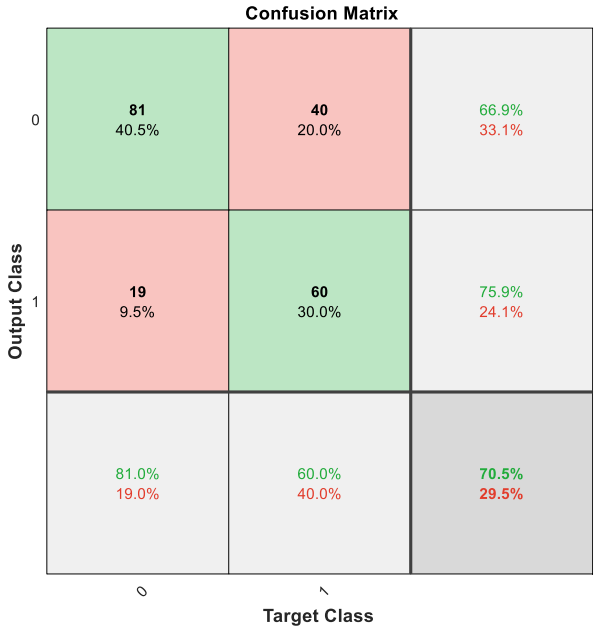
$$\boldsymbol{\Sigma_j} = \frac{1}{N_j} \sum_{i=1}^{N_j} \boldsymbol{x_i x_i^T}$$

Παρατηρούμε ότι η «σιγουριά» που έχουμε στην περίπτωση που ξέρουμε ποια παραδείγματα ανήκουν σε ποια κατηγορία μεταφράζεται στις a posteriori πιθανότητες $p_{i,j}(t)$ που μας λένε πόσο πιθανό είναι το εκάστοτε δείγμα $\boldsymbol{x_i}$ να ανήκει στην κατηγορία j.

Τρέχοντας τον επαναληπτικό αλγόριθμο με αρχικά σ_j^2 τυχαίους αριθμούς στο διάστημα (0,1) και κατηγοριοποιώντας με τον ίδιο τρόπο όπως στο προηγούμενο ερώτημα, παίρνουμε τα παρακάτω αποτελέσματα:



Σχήμα 7: Δεδομένα κάθε ομάδας(τελείες) και Προβλέψεις(κύκλοι)



Σχήμα 8: Πίνακας σύγχυσης για την κατηγοριοποίηση με χρήση Gaussian Mixture Model με άγνωστη παράμετρο τις διαγώνιες διασπορές.

Όπως φαίνεται παραπάνω το ποσοστό σφάλματος της κατηγοριοποίησης είναι **29.5%**.

Παρατηρούμε πως το ποσοστό είναι μεγαλύτερο κατά 1% σε σχέση με το προηγούμενο ερώτημα, πράγμα που ενδεχομένως να οφείλεται στο μικρό μέγεθος του dataset, που δεν μας επιτρέπει να εκτιμήσουμε με ακρίβεια τα στατιστικά χαρακτηριστικά της κάθε κατηγορίας.

Σύνοψη

Συνοπτικά τα αποτελέσματα των δυο προβλημάτων φαίνονται στον παρακάτω πίνακα:

	Bayes Test	2D K-Means	3D K-Means	GMM	GMM με άγνωστες διασπορές μόνο
Ποσοστό σφάλματος	26.3%	38.5%	35%	28.5%	29.5%
Επαναλήψεις μέχρι τη σύγκλιση	-	12	5	88	22

Πίνακας 1: Σύγκριση μεθόδων ομαδοποίησης

Παρατηρείται πως οι μέθοδοι expectation/maximization σε Gaussian Mixture Models έχουν ποσοστά σφάλματος κοντινότερα στο βέλτιστο ποσοστό σφάλματος κατά Bayes (26.3%). Αυτό ενδεχομένως να οφείλεται στο γεγονός ότι η μέθοδος αυτή κάνει την υπόθεση πως η κατανομή των δεδομένων μας είναι Gaussian ενώ ο K-means δεν κάνει κάποια υπόθεση για την κατανομή των δεδομένων μας.

Επιπλέον μπορούμε να δούμε πως για το ίδιο κατώφλι (10^{-6}) οι GMM μέθοδοι χρειάστηκαν περισσότερες επαναλήψεις απ' ότι οι K-means μέθοδοι. Ωστόσο η υπολογιστική πολυπλοκότητα του κάθε αλγορίθμου είναι διαφορετική οπότε δεν μπορούμε να αποφανθούμε χωρίς περεταίρω διερεύνηση για την αποδοτικότητα του εκάστοτε αλγορίθμου.