Επεξεργασία Δεδομένων και Αλγόριθμοι Μάθησης
3ο Σετ Ασκήσεων

Πρόβλημα 1: Το αρχείο hw3-1data.mat περιέχει $N = 200$ διανύσματα μήκους 2, τα οποία επιθυμούμε να ομαδοποιήσουμε σε δύο ομάδες. Σας δίνεται η **ΜΥΣΤΙΚΗ πληροφορία την οποία ΑΠΑΓΟΡΕΥΕΤΑΙ να χρησιμοποιήσετε στη μέθοδο ομαδοποίησης που θα δημιουργήσετε:** ότι τα πρώτα 100 διανύσματα ανήκουν σε μια ομάδα και τα δεύτερα 100 σε άλλη.

α) Εφαρμόστε την K-means ώστε να κάνετε την ομαδοποίηση. Αφού συγκλίνει η μέθοδος μετρήστε το ποσοστό των σφαλμάτων που κάνατε κάνοντας χρήση της μυστικής πληροφορίας. *Προσοχή!!!: Το τι είναι πρώτη ομάδα και τι δεύτερη δεν είναι καθορισμένο. Οπότε από τις δύο δυνατές επιλογές θα διαλέξετε εκείνη που κάνει τα λιγότερα λάθη! Επίσης έχετε υπόψη σας ότι εάν επιλέξετε εντελώς τυχαία την ομάδα σε κάθε διάνυσμα η πιθανότητα να κάνετε λάθος είναι 0.5. Επομένως εάν η μέθοδός σας σαν δίνει ποσοστό σφάλματος κοντά στο 0.5 τότε φυσικά δεν είναι καλή.*

β) Ένας τρόπος να βελτιώσετε την απόδοση της K-means είναι να παρατηρήσετε ότι έχει την τάση να δημιουργεί γραμμικά όρια. Επομένως εάν μεγαλώσετε με τεχνητό τρόπο τη διάσταση των δεδομένων σας αυτό μπορεί να βελτιώσει τα αποτελέσματα. Δημιουργείστε μια τρίτη συντεταγμένη όπου κάθε δισδιάστατο X_i θα αντικατασταθεί από το τρισδιάστατο $\{X_i, \|X_i\|^2\}$. Με άλλα λόγια, σημεία X_i που είναι κοντά στην αρχή των αξόνων στις τρεις διαστάσεις θα είναι πιο κοντά στο οριζόντιο επίπεδο από ό,τι σημεία που βρίσκονται μακρύτερα. Φυσικά εδώ κάνουμε μια ελαφριά χρήση της μυστικής πληροφορίας μιας και παρατηρούμε ότι τα σημεία της μιας ομάδας βρίσκονται συγκεντρωμένα γύρω από την αρχή των αξόνων, αλλά δεν πειράζει. Επαναλάβατε την K-means με τα σημεία στον τρισδιάστατο χώρο και υπολογίστε πάλι το σφάλμα ομαδοποίησης.

Για να δείτε τα κατορθώματά σας, τοποθετείστε τα δεδομένα στο επίπεδο σαν τελείες με διαφορετικό χρώμα κάθε ομάδα, χρησιμοποιώντας τη μυστική πληροφορία που έχετε. Κατόπιν σε κάθε σημείο βάλτε ένα κύκλο με το χρώμα της ομάδας που επιλέγετε για κάθε σημείο. Το κάνετε αυτό για κάθε μία από τις δύο μεθόδους σε χωριστές εικόνες. Σε όσο πιο πολλά σημεία οι τελείες και οι κύκλοι έχουν το ίδιο χρώμα, τόσο καλύτερη είναι η αντίστοιχη μέθοδος.

Πρόβλημα 2: Χρησιμοποιούμε πάλι τα δεδομένα του αρχείου hw3-1data.mat. Αλλά τώρα σας δίνεται η πληροφορία ότι και οι δύο ομάδες διανυσμάτων είναι υλοποιήσεις δύο διαφορετικών Gaussian διανυσμάτων. Το δεδομένα του αρχείου αποτελούν δηλαδή μια Gaussian μίξη. Οι αρχικές πιθανότητες των δύο Gaussian είναι w_1, w_2, οι μέσες τιμές μ_1, μ_2 και οι μήτρες συνδιασποράς Σ_1, Σ_2.

α) Εφαρμόστε τη μέθοδο expectation/maximization για την εκτίμηση όλων των παραμέτρων της μίξης δηλαδή των $\{w_1, w_2, \mu_1, \mu_2, \Sigma_1, \Sigma_2\}$ καθώς και *τις (εκ των υστέρων) πιθανότητες p_{ij} με τις οποίες κάθε διάνυσμα X_i ανήκει στην ομάδα $j = 1, 2$.* "Τρέξτε" τον αλγόριθμο έως ότου συγκλίνει και χρησιμοποιώντας τις (εκ των υστέρων) πιθανότητες κάθε διανύσματος αποφασίστε σε ποια ομάδα ανήκει το κάθε διάνυσμα επιλέγοντας την ομάδα με τη μεγαλύτερη (εκ των υστέρων) πιθανότητα. Μετρήστε τα σφάλματα χρησιμοποιώντας τη μυστική πληροφορία.

β) Θεωρείστε τώρα ότι σας δίνεται επίσης πως $w_1 = w_2 = 0.5$, $\mu_1 = \mu_2 = 0$ και $\Sigma_1 = \sigma_1^2 I$, $\Sigma_2 = \sigma_2^2 I$ με μόνα τα σ_1^2, σ_2^2 άγνωστα. Εφαρμόστε την ίδια ιδέα του expectation/maximization και προτείνετε εκτιμητές για τα σ_1^2, σ_2^2 καθώς και για τις (εκ των υστέρων) πιθανότητες με τις οποίες κάθε διάνυσμα ανήκει σε κάθε ομάδα. Τρέξτε τον αλγόριθμο μέχρι να συγκλίνει και κάντε ομαδοποίηση χρησιμοποιώντας τις (εκ των υστέρων) πιθανότητες όπως και προηγουμένως. Μετρήστε πάλι τα σφάλματα.

Σχεδιάστε όπως και προηγουμένως στο επίπεδο τα σημεία με τελείες και την ομαδοποίηση που κάνετε με κύκλους δημιουργώντας μια εικόνα για κάθε μέθοδο. *Δίνεται ότι το Bayes τεστ που ξέρει τα πάντα έχει πιθανότητα σφάλματος περίπου ίση με 0.263*. Πως συγκρίνονται τα σφάλματα των 4 μεθόδων (των Προβλημάτων 1 και 2) με την ιδανική πιθανότητα σφάλματος;