ΑΝΑΓΝΩΡΙΣΗ ΠΡΟΤΥΠΩΝ

ΕΡΓΑΣΙΑ 8

ΑΛΒΑΝΑΚΗ ΠΑΡΑΣΚΕΥΗ

Α.Μ 57286

8.1 K-MEANS

Χρησιμοποιώντας τον αλγόριθμο K-means προσπαθήστε να ταξινομήσετε τα φυτά σε 3 ομάδες. Ποιό είναι το λάθος που βρίσκετε?

Απάντηση

Ένας αλγόριθμος διαχωρισμού όπου κάθε ομάδα (cluster) συνδέεται με ένα centroid (κέντρο). Κάθε σημείο αποδίδεται στην ομάδα στης οποίας το κέντρο βρίσκεται πιο κοντά. Ο αριθμός των ομάδων K στις οποίες χωρίζονται τα σημεία πρέπει να καθοριστεί από πριν. Ο αλγόριθμος έχει ως εξής:

- Επιλογή τυχαία των αρχικών K points ως αρχικά κέντρα.

- Επανάληψη: Δημιουργία K clusters αναθέτοντας κάθε σημείο στο κοντινότερο σε αυτό centroid. Επαναϋπολογισμός των centroids κάθε cluster.

- Έως ότου τα centroids σταματήσουν να αλλάζουν για 2 συνεχόμενες επαναλήψεις.

Για κάθε σημείο, το σφάλμα του είναι η απόστασή του από το κοντινότερο σε αυτό κέντρο. Το άθροισμα των τετραγώνων όλων των σφαλμάτων ονομάζεται SSE.

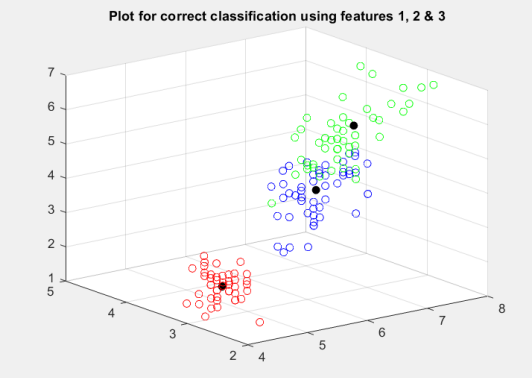
Για την υλοποίηση του k-means χρησιμοποιούμε την έτοιμη συνάρτηση k_means από το βιβλίο στην οποία δηλώνουμε τα αρχικά κέντρα που θα ξεκινήσει από αυτά να τρέχει ο αλγόριθμος και τα samples στα οποία θα κάνουμε την ομαδοποίηση. Τα αρχικά κέντρα τα τοποθετούμε τυχαία με τη συνάρτηση randi.

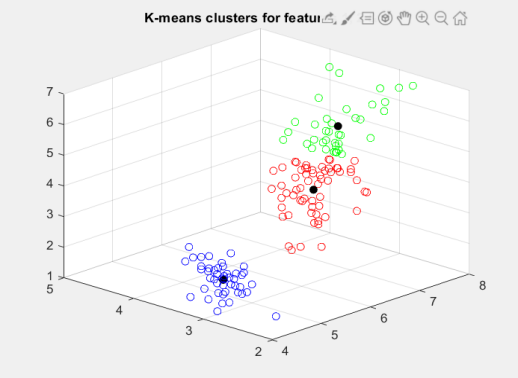
Έχοντας αρχικοποιήσει τα κέντρα τυχαία μπορούμε να οδηγηθούμε σε τοπικό και όχι σε global ελάχιστο της απόστασης μεταξύ ενός σημείου και του κέντρου του cluster στο οποίο έχει ανατεθεί(πχ αν υπάρξει κέντρο σε outlier κατά την αρχικοποίηση).Επιπλέον όταν υπάρχουν μεγάλες διαφορές στο scale των τιμών 2 features το feature με τις μεγαλύτερες τιμές θα υπερέχει αφού δεν γίνεται standardization.Το σφάλμα στο clustering υπολογίζεται από το άθροισμα του τετραγώνου των σφαλμάτων για όλα τα δείγματα(Sum of Squared Error).Ουσιαστικά η μετρική αυτή υπολογίζεται με βάση την απόσταση κάθε δείγματος από το κοντινότερο του κέντρο.

Ενδεικτικά παρατίθενται παρακάτω τα αποτελέσματα 2 runs μαζί με τις γραφικές τους

1ο Run

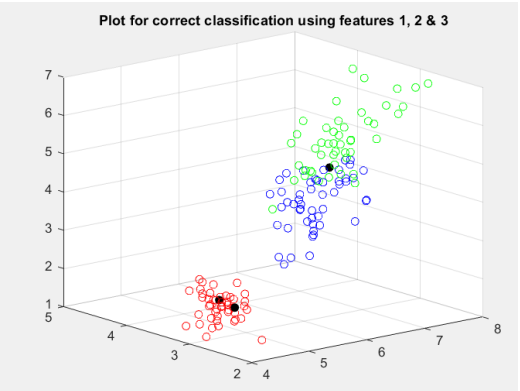
```
    37   121    67

The number of samples in cluster 1 is:
    50

The number of samples in cluster 2 is:
    39

The number of samples in cluster 3 is:
    61

The SSE value is 78.9451
```

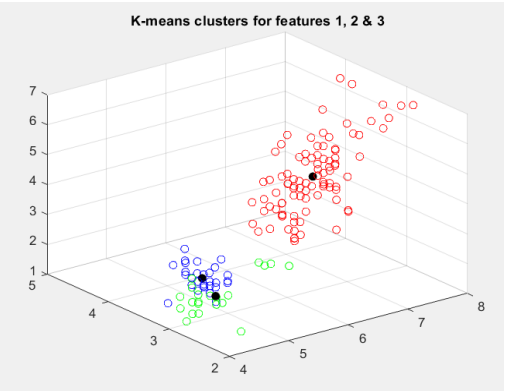




2ο Run

```
   113    22     3

The number of samples in cluster 1 is:
    96

The number of samples in cluster 2 is:
    30

The number of samples in cluster 3 is:
    24

The SSE value is 142.8593
```





<u>Συμπεράσματα</u>

Όπως αναφέρθηκε και προηγουμένως και επιβεβαιώνεται και από τα runs ο αλγόριθμος επηρεάζεται αρκετά από την τυχαία αρχικοποίηση των κέντρων των κλάσεων. Το ένα run να συγκλίνει στην ορθή ομαδοποίηση ενώ το άλλο σε λανθασμένη. Για την επίλυση αυτού

του προβλήματος εφαρμόζουμε τον k-means πολλές φορές.Σε αρκετά από τα runs ο αλγόριθμος κατάφερνε να ομαδοποιήσει σωστα την κλάση που είναι γραμμικά διαχωρίσιμη από τις άλλες.Ωστόσο στις άλλες 2 κλάσεις, που είναι μη γραμμικά διαχωρίσιμες, τα αποτελέσματα δεν ήταν τόσο καλά καθώς με τον k-means δημιουργούνται ομάδες σφαιρικού σχήματος εξ αιτίας της ευκλείδιας απόστασης που χρησιμοποιείται .
Ο αριθμός των δειγμάτων που ομαδοποιούνται λανθασμένα ποικίλει για τους λόγους που αναφέρθηκαν προηγουμένως.

8.2 FUZZY C-MEANS

Χρησιμοποιώντας τον αλγόριθμο fuzzy C-means προσπαθήστε να ταξινομήσετε τα φυτά σε 3 ομάδες. Ποιό είναι το λάθος που βρίσκετε?

Απάντηση 8.2

Ο fuzzy C-means έιναι μία παραλλαγή του K-means στην οποία κάθε στοιχείο μπορεί να ανήκει σε περισσότερα από ένα clusters. Ο αλγόριθμος διαφέρει σε σχέση με τον k-means καθώς εφαρμόζει soft clustering σε αντιθεση με το hard clustering του k-means.Κάθε σημείο του dataset ανήκει σε όλες τις ομάδες με ένα βαθμό συμμετοχής.Εισάγεται μία νέα συνάρτηση membership uj.

Η διαδικασία που ακολουθείται έχει ως εξής:

- Ορίζεται ένας αριθμός clusters.

- Τυχαία ανατίθενται coefficients σε κάθε data point για το membership τους σε κάθε cluster.

- Επανάληψη ως ότου για 2 συνεχόμενα iterations η αλλαγή των coefficients δεν είναι μεγαλύτερη από μία ανεκτικότητα ε:

Υπολογισμός του centroid κάθε cluster γίνεται σύμφωνα με τον τύπο που δίνεται παρακάτω, όπου m η hyper-parameter που ορίζει πόσο fuzzy θα είναι το κάθε cluster.

$$c_k = \frac{\sum_x u_k(x)^m x}{\sum u_k(x)^m}$$

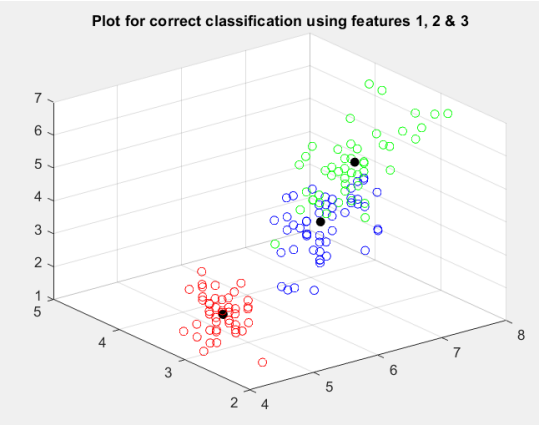
Όσο μεγαλύτερη η τιμή του m τόσο πιο fuzzy το cluster. Για κάθε σημείο γίνεται υπολογισμός των coefficients των memberships για κάθε cluster.

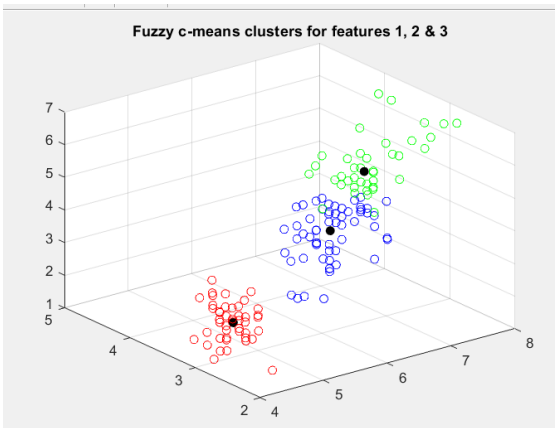
Για την υλοποίηση του fuzzy c-means χρησιμοποιούμε την έτοιμη συνάρτηση fzzy_c_means από το βιβλίο στην οποία δηλώνουμε τον αριθμό των clusters,τον fuzzifier και τα samples στα οποία θα κάνουμε την ομαδοποίηση.Η συνάρτηση αυτή επιστρέφει εκτός από τα τελικά κέντρα και τον βαθμό συμμετοχής του κάθε δείγματος σε κάθε ομάδα.Έτσι μπορούμε να βρούμε για κάθε samle το μεγαλύτερο βαθμό συμμετοχής του και συνεπώς σε ποια ομάδα ταξινομειται.

Ενδεικτικά παρατίθενται παρακάτω τα αποτελέσματα 2 runs μαζί με τις γραφικές τους

1ο Run (q=3)

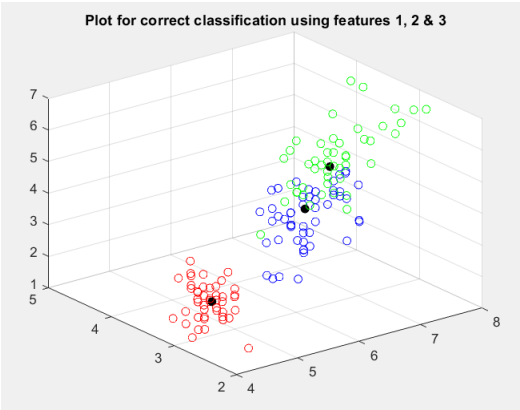
```
Iteration count = 22, obj. fcn = 29.110245
The number of samples in cluster 1 is:
    59

The number of samples in cluster 2 is:
    41

The number of samples in cluster 3 is:
    50
```

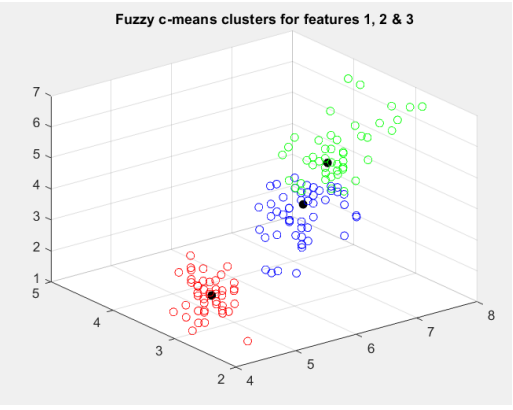




2ο Run (αύξηση fuzzifier q=9)

```
Iteration count = 11, obj. fcn = 0.061248
The number of samples in cluster 1 is:
    50

The number of samples in cluster 2 is:
    51

The number of samples in cluster 3 is:
    49
```





<u>Συμπεράσματα</u>

Παρατηρούμε και από τον αριθμό των samples σε κάθε κλάση και από το σφάλμα fcn πως έχουμε καλύτερα από τελέσματα σε σχέση με τον k-means.Αυξάνοντας τον fuzzifier έχουμε φανερά καλύτερη ομαδοποίηση καθώς μειώνεται σημαντικά το σφάλμα.Τέλος ο fuzzy c-means κάνει διαχωρισμό της γραμμικά διαχωρίσιμης κλάσεις καιι τα πάει αρκετά καλά και στον διαχωρισμό των μη γραμμικών.

8.3 ISODATA

Χρησιμοποιώντας τον αλγόριθμο ISODATA προσπαθήστε να ταξινομήσετε τα φυτά σε ομάδες. Πόσες ομάδες βρίσκετε ? Ποιό είναι το λάθος που βρίσκετε ?

Απάντηση 8.3

Ο ISODATA είναι ουσιαστικά ο k-means αλγόριθμος με το επιπλέον χαρακτηριστικό ότι μπορεί αυτόματα να επιλέξει πλέον το πλήθος των κλάσεων.

Οι παράμετροι που επιλέγονται είναι οι εξής:

NMIN_EX -> ελάχιστο πλήθος δειγμάτων ανά cluster.

ND ->επιθυμητό πλήθος clusters.

σ^2 -> μέγιστη διασπορά για διαχωρισμό clusters.

DMERGE -> μέγιστη απόσταση για ένωση των clusters.

NMERGE -> μέγιστο πλήθος clusters που μπορούν να ενωθούν.

Η διαδικασία έχει ως εξής:

- Επιλέγονται αυθαίρετα τα κέντρα και τα σημεία ανατίθενται στο κοντινότερο σε αυτά cluster.
- Το standard deviation κάθε cluster και η απόσταση των clusters μεταξύ τους υπολογίζονται.
- Τα clusters χωρίζονται αν ένα ή περισσότερα deviations είναι μεγαλύτερα από το προκαθορισμένο threshold.
- Τα clusters ενώνονται αν η απόσταση μεταξύ τους είναι μικρότερη από το προκαθορισμένο threshold.
- Εφαρμόζεται περεταίρω iterations των παραπάνω με τα καινούρια πλέον cluster centers έως ότου: ο μέσος όρος των inner-center distances είναι μικρότερος από το προκαθορισμένο threshold ,ο μέσος όρος των αλλαγών στα inner-center distances μεταξύ των iterations είναι μικρότερος από το προκαθορισμένο threshold ή έχει πραγματοποιηθεί ο μέγιστος επιτρεπόμενος αριθμός iterations.

Χρησιμοποιήθηκε η wfIsodata_ND η οποία επιστρέφει τα τελικά κέντρα μαζί με τα στοιχεία που αποτελούν κάθε ομάδα, τον αριθμό ομάδων και την ομάδα στην οποία αποδίδεται κάθε sample.

Υπολογίζουμε το σφάλμα SSEμε βάση το τετράγωνο της απόστασης των σημείων από το κέντρο της ομάδας και βλέπουμε με βάση αυτό και τον αριθμό των ταξινομημένων δειγμάτων αν έγινε σωστά η ομαδοποίηση.
Ενδεικτικά παρατίθενται παρακάτω τα αποτελέσματα 2 runs μαζί με τις γραφικές τους

1ο Run (μέγιστες ομάδες 6)

```
Total Number of Clusters:
     4

The number of samples in cluster 1 is:
    50

The number of samples in cluster 2 is:
    40

The number of samples in cluster 3 is:
    32

The number of samples in cluster 4 is:
    28

The SSE value is 57.3179
```

2ο Run (μέγιστες ομάδες 4)

```
Total Number of Clusters:
     3

The number of samples in cluster 1 is:
    38

The number of samples in cluster 2 is:
    50

The number of samples in cluster 3 is:
    62

The SSE value is 78.9408
```

Συμπεράσματα

Στην 1η περίπτωση και για μεγαλύτερες τιμές από 6 παρατηρήθηκε πως ο αλγόριθμος συγκλίνει στις 4 κλάσεις ενώ για μικρότερες μέγιστες ομάδες συγκλίνει στις 3.Στις 4 κλάσεις παρατηρούμε πως έχουμε μικρότερο σφάλμα από τον k-means.Επειδή ο αλγόριθμος βασίζεται στον k-means στις ομάδες που είναι μη γραμμικά διαχωρίσιμες ουσιαστικά τις σπάει σε παραπάνω προκυμμένου να δημιουργηθούν σφαιρικές κλάσεις. Όταν μειώσαμε τον αριθμό τον κλάσεων στο 2ο run παρατηρούμε πως έχουμε τα ίδια αποτελέσματα με τον k-means.