

ΑΝΑΦΟΡΑ 2ΗΣ ΑΣΚΗΣΗΣ ΣΤΗΝ ΥΠΟΛΟΓΙΣΤΙΚΗ
ΝΟΗΜΟΣΥΝΗ

ΟΝΟΜΑΤΕΠΩΝΥΜΟ: ΣΤΥΛΙΑΝΟΣ ΣΥΡΡΟΣ

ΑΜ:4805

ΕΤΟΣ:4ο

ΟΝΟΜΑΤΕΠΩΝΥΜΟ:ΘΕΟΦΑΝΗΣ ΜΠΟΥΡΑΪΜΗΣ

ΑΜ:4745

ΕΤΟΣ:4ο

ΟΝΟΜΑΤΕΠΩΝΥΜΟ: ΕΛΕΥΘΕΡΙΟΣ-ΧΡΗΣΤΟΣ
ΔΡΙΤΣΩΝΑΣ

ΑΜ:4668

ΕΤΟΣ:4ο

ΗΜΕΡΟΜΗΝΙΑ:3/01/2023

Εισαγωγή

Στην συγκεκριμένη αναφορά θα περιγράψουμε και θα αναλύσουμε τα αποτελέσματα(οπτική αναπαράσταση των παραδειγμάτων ,θέση των κέντρων καθώς και το σφάλμα ομαδοποίησης) που μας δίνει ο αλγόριθμος k-means με βάση την υλοποίηση που κάναμε στο πρόγραμμα. Επίσης, θα αναφέρουμε και κάποιες παρατηρήσεις πάνω σε αυτά τα δεδομένα.

Αρχικά, πριν περάσουμε στην εξήγηση των αποτελεσμάτων πρέπει να αναφέρουμε ότι το πρόγραμμα στο οποίο υλοποιείται ο αλγόριθμος k-means έχει χρησιμοποιηθεί η γλώσσα προγραμματισμού C , ενώ τα διαγράμματα στα οποία φαίνεται η ομαδοποίηση ανάλογα με τον αριθμό των ομάδων που επιθυμεί ο χρήστης να γίνει ομαδοποίηση των δεδομένων έγινε στην γλώσσα προγραμματισμού Python.

ΕΚΤΕΛΕΣΗ ΠΡΟΓΡΑΜΜΑΤΩΝ

Προφανώς για να μπορέσουμε να δουλέψουμε στην άσκηση πρέπει να δημιουργήσουμε το σύνολο δεδομένων με τα παραδείγματα. Έτσι , για να το πετύχουμε αυτό , μεταγλωττίζουμε το πρόγραμμα `examples.c`.

Γράφουμε την εντολή **gcc examples.c** . Αφού μεταγλωττιστεί επιτυχώς γράφουμε την εντολή **./a.out** στο τερματικό για να μπορέσουμε να το τρέξουμε. Πλέον, μετά την εκτέλεση της τελευταίας εντολής έχει δημιουργηθεί το `examplesSDO.txt` που είναι το αρχείο κειμένου με τα παραδείγματα για την άσκηση 2.

Μετά από αυτό προχωράμε στα υπόλοιπα αρχεία.

Για να τρέξουμε το πρόγραμμα με τον k-means γράφουμε αρχικά την εντολή στο τερματικό :

gcc kmeans.c -lm για την μεταγλώττιση του προγράμματος.

Έπειτα γράφουμε την εντολή **./a.out** για να τρέξουμε το πρόγραμμα έπειτα από την επιτυχής μεταγλώττιση του. Το συγκεκριμένο πρόγραμμα δημιουργεί 2 αρχεία κειμένου (.txt) όπου το ένα περιέχει τις συντεταγμένες κάθε παραδείγματος και την ομάδα στην οποία ανήκει το συγκεκριμένο παράδειγμα(**teams_data.txt**) , ενώ το άλλο αρχείο περιέχει τις συντεταγμένες των κέντρων μετά το τέλος του k-means (**centers_data.txt**).

Μετά από αυτό εκτελούμε το python αρχείο(**script.py**) που μας εμφανίζει οπτικά τα παραδείγματα ομαδοποιημένα σε ομάδες ανάλογα με τον αριθμό ομάδων που θέλει ο χρήστης και τέλος το διάγραμμα που μας δείχνει το πως μεταβάλλεται το σφάλμα ομαδοποίησης ανάλογα με τον αριθμό των ομάδων.

Συγκεκριμένα γράφουμε την εντολή: **python script.py**

Τέλος κάτι που αξίζει να αναφέρουμε είναι ότι για να μεταβάλλουμε τον αριθμό των ομάδων πρέπει να πάμε και να αλλάξουμε την 5^η γραμμή από το αρχείο **kmeans.c**

καθώς ο αριθμός των ομάδων προσδιορίζεται με την χρήση της εντολής **#define** όπως αναφέρεται στην εκφώνηση.

Τα συγκεκριμένα βήματα θα πρέπει να εκτελούνται με την παραπάνω σειρά έτσι ώστε να βλέπουμε τις αλλαγές που γίνονται για κάθε διαφορετικό αριθμό ομάδων.

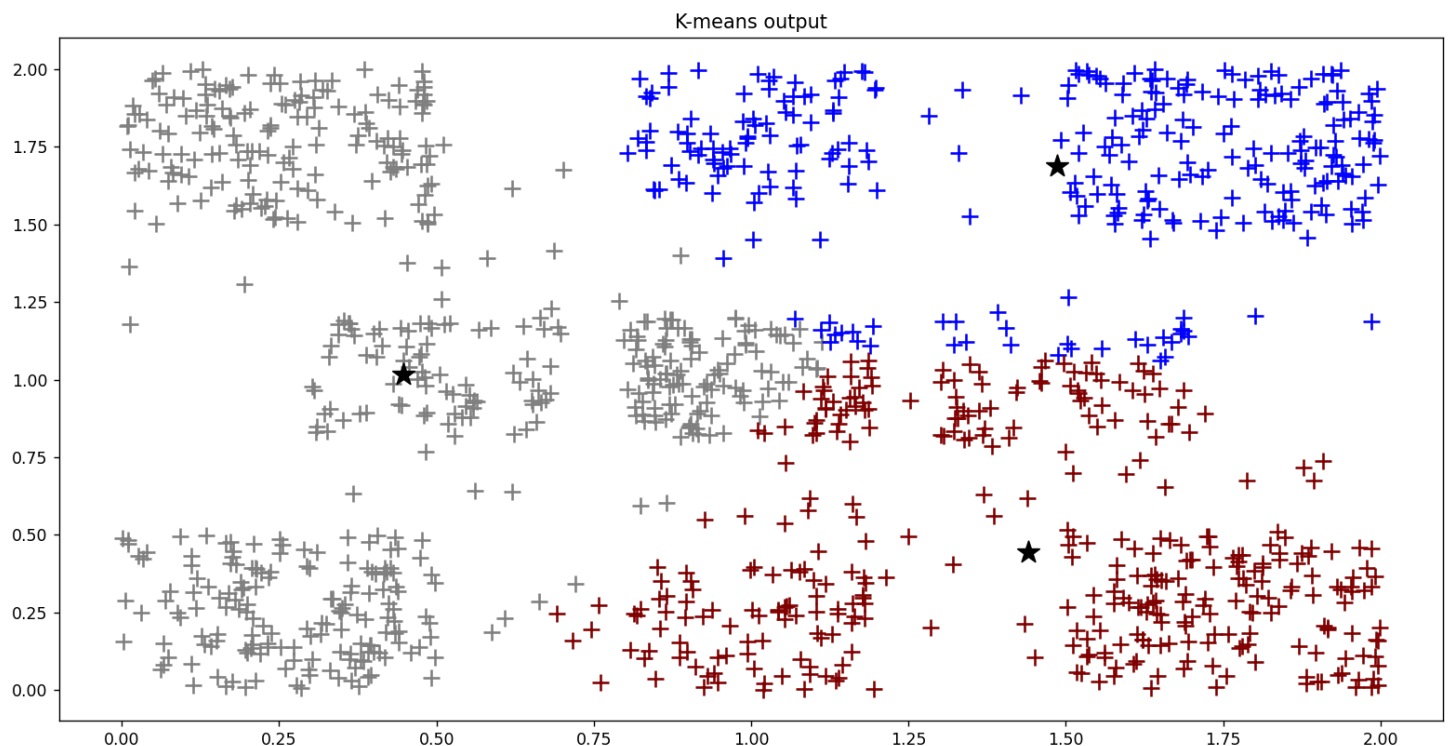
ΕΞΗΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΩΝ

Τώρα για κάθε τιμή του $M(3,6,9,12)$ πραγματοποιήσαμε 15 τρεξίματα και κρατήσαμε το σχήμα με το μικρότερο σφάλμα ομαδοποίησης, όπως δηλαδή αναφέρεται ακριβώς στην εκφώνηση.

Έτσι ακολουθούν 4 screenshots το καθένα για διαφορετική τιμή του M .

Σε όλα τα διαγράμματα που θα παρουσιαστούν τα παραδείγματα απεικονίζονται με το σήμα του '+' ενώ το κέντρο κάθε ομάδας απεικονίζεται με το σύμβολο του άστρου (*) με μαύρο χρώμα.

M=3



Για $M=3$ το μικρότερο σφάλμα ομαδοποίησης που βρήκαμε ήταν το 621.960449 και η λύση που δημιουργείται από τον k-means για το αντίστοιχο σφάλμα φαίνεται παραπάνω.

Έτσι στο παραπάνω σχήμα βλέπουμε ότι όλα τα παραδείγματα έχουν ομαδοποιηθεί σε 3 ομάδες (αφού $M=3$) και τα οποία μπορούμε και τα ξεχωρίζουμε από το

διαφορετικό χρώμα που έχει κάθε παράδειγμα. Συνεπώς, διακρίνουμε με εύκολο τρόπο και το κέντρο κάθε ομάδας που έχει το σήμα του αστεριού όπως αναφέραμε. Το κέντρο κάθε ομάδας έχει μαύρο χρώμα έτσι ώστε να μπορούμε να το διακρίνουμε με μεγαλύτερη ευκολία. Επίσης, το χρώμα κάθε ομάδας επιλέγεται με τυχαίο τρόπο. Τα κέντρα είναι 3 επειδή ο αριθμός των ομάδων είναι 3.

```
The new centers are:0.447286,1.015170  
The new centers are:1.486594,1.688645  
The new centers are:1.440802,0.443343  
Iterations:9  
The total team error is:621.960449
```

Στην παραπάνω φωτογραφία βλέπουμε τις τελικές θέσεις που έχουν τα κέντρα μετά το τέλος του k-means τις οποίες και επιβεβαιώνουμε στο σχήμα που προηγήθηκε.

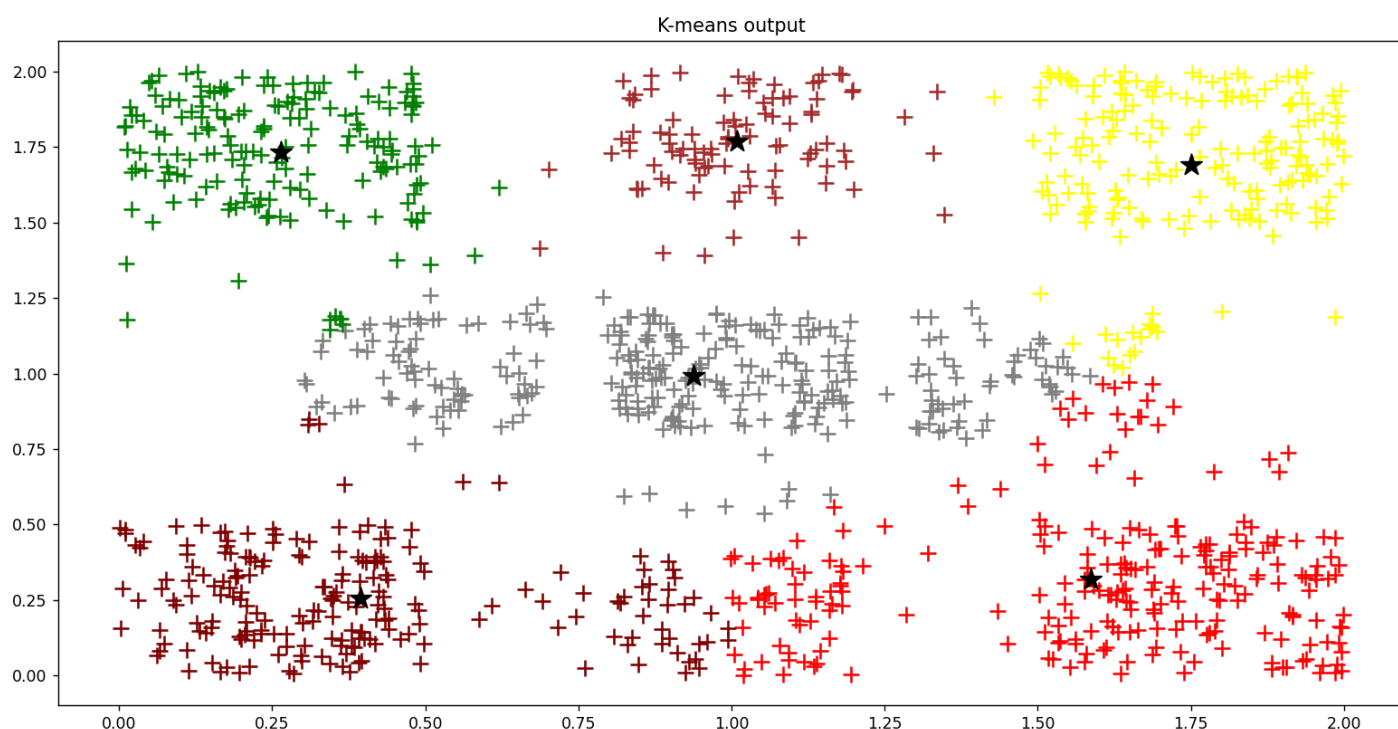
Ο k-means για $M=3$ για το μικρότερο δυνατό σφάλμα ομαδοποίησης που βρήκαμε συγκλίνει έπειτα από 9 επαναλήψεις. Επίσης, μια παρατήρηση που μπορούμε να κάνουμε είναι ότι το σφάλμα στην συγκεκριμένη περίπτωση είναι το μικρότερο από τα 15 τρεξίματα που κάναμε επειδή η τυχαία επιλογή των παραδειγμάτων(αρχικοποίηση των κέντρων) που έγινε πριν την έναρξη του αλγορίθμου ήταν σε σημεία στα οποία βρίσκονται μεγάλο μέρος παραδειγμάτων κάθε ομάδας με αποτέλεσμα ο αλγόριθμος να συγκλίνει πιο γρήγορα. Αυτό αποτυπώνεται από το πλήθος επαναλήψεων που χρειάστηκαν (Iterations:9).

Η ομαδοποίηση των παραδειγμάτων σε 3 ομάδες είναι ένας μικρός αριθμός για το συγκεκριμένο σύνολο δεδομένων καθώς κάθε ομάδα έχει πάρα πολλά παραδείγματα και συνεπώς πολύ μεγάλες αποστάσεις από το κέντρο της ομάδας με αποτέλεσμα μια επόμενη λύση του k-means με μεγαλύτερο αριθμό ομάδων να αναμένουμε να δώσει καλύτερο οπτικό αποτέλεσμα.

Πλέον συνεχίζουμε την ανάλυση για $M=6$ ομάδες.

M=6

Στην συγκεκριμένη περίπτωση το μικρότερο σφάλμα που βρήκαμε μετά από 15 εκτελέσεις του αλγορίθμου είναι 325.041260.



Στο παραπάνω σχήμα φαίνεται ο διαχωρισμός των παραδειγμάτων σε 6 ομάδες. Σε σύγκριση με τον διαχωρισμό των παραδειγμάτων σε 3 ομάδες μπορούμε να πούμε ότι η ομαδοποίηση των παραδειγμάτων για $M=6$, σε οπτικό επίπεδο, είναι αρκετά καλύτερη. Αυτό προκύπτει βέβαια και από το σφάλμα ομαδοποίησης το οποίο είναι αρκετά μειωμένο, συγκεκριμένα περίπου κατά ήμισυ, σε σχέση με το σφάλμα που είχε προκύψει για $M=3$. Συνεπώς, μπορούμε να αναφέρουμε ότι η ομαδοποίηση για $M=6$ στα συγκεκριμένα παραδείγματα είναι προτιμότερη από $M=3$ επειδή το σφάλμα είναι μικρότερο και τα παραδείγματα είναι καλύτερα ομαδοποιημένα. Το σφάλμα μειώνεται επειδή ο αριθμός των ομάδων αυξάνεται και συνεπώς τα παραδείγματα κάθε ομάδας έχουν μικρότερη απόσταση από το κέντρο της ομάδας στην οποία ανήκουν και άρα η γενικότερη λύση είναι καλύτερη.

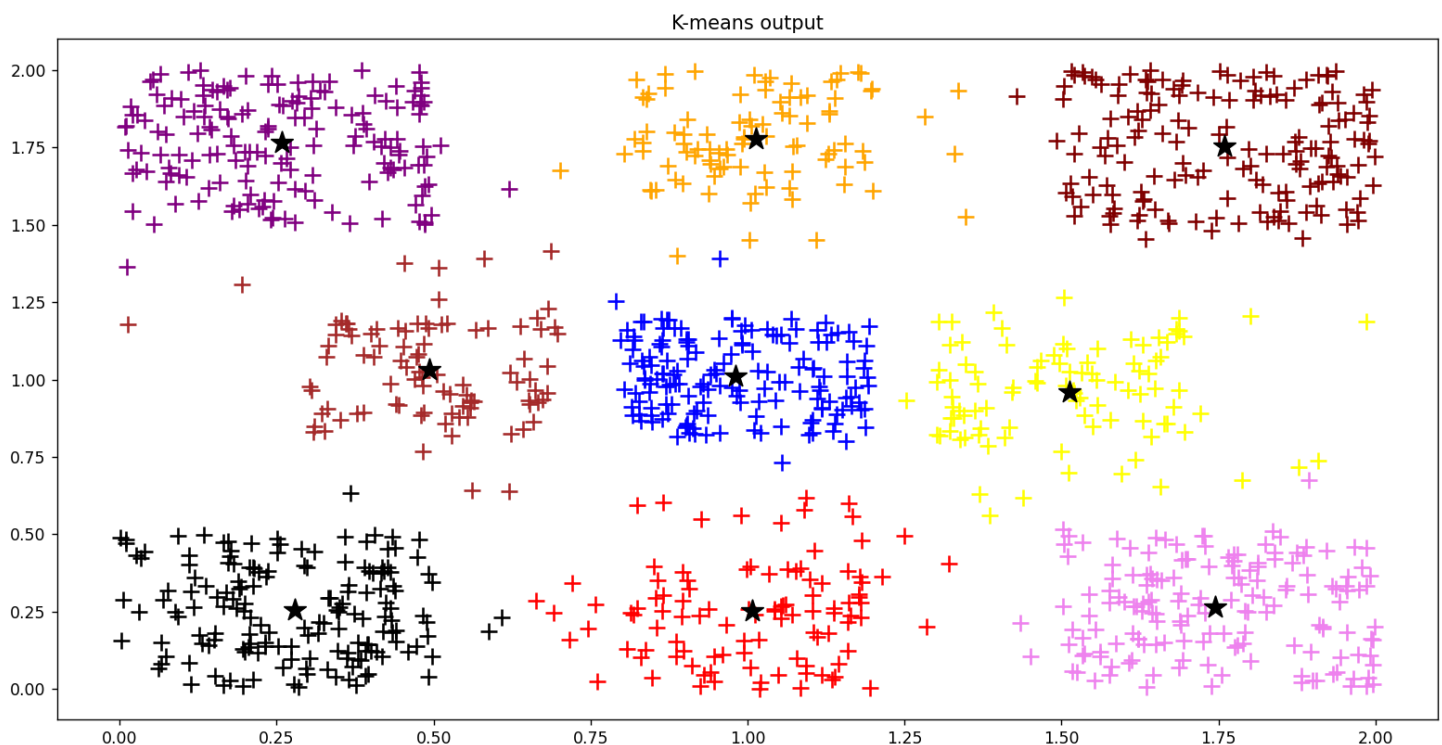
Προφανώς ο αριθμός των κέντρων είναι 6 επειδή ο αριθμός των ομάδων είναι 6 και σε κάθε ομάδα αντιστοιχεί ένα κέντρο. Το κάθε κέντρο έχει το σχήμα του αστεριού με μαύρο χρώμα. Παρατηρούμε ότι η θέση κάθε κέντρου είναι σε αρκετά καλό σημείο επειδή γύρω από αυτό το σημείο βρίσκονται τα περισσότερα παραδείγματα που ανήκουν στην ομάδα αυτή.

```
The new centers are:1.008691,1.769222
The new centers are:0.938330,0.992284
The new centers are:1.751197,1.691754
The new centers are:1.586398,0.317268
The new centers are:0.394175,0.257125
The new centers are:0.263829,1.732576
Iterations:12
The total team error is:325.041260
```

Στην παραπάνω φωτογραφία βλέπουμε ότι χρειάστηκαν 12 επαναλήψεις για να τελειώσει ο αλγόριθμος. Επιπλέον, φαίνονται και οι θέσεις των 6 κέντρων καθώς μπορούν να επιβεβαιωθούν από το σχήμα που παρουσιάστηκε.

Προχωράμε για $M=9$ ομάδες.

M=9



Στην συγκεκριμένη περίπτωση το μικρότερο σφάλμα που βρήκαμε μετά από 15 εκτελέσεις του αλγορίθμου είναι 225.237823. Στο παραπάνω σχήμα, παρατηρούμε ότι η ομαδοποίηση είναι η καλύτερη από όλες έχουμε εξετάσει μέχρι τώρα από οπτική πλευρά διότι τα παραδείγματα κάθε ομάδας έχουν πολύ μικρή απόσταση μεταξύ τους με αποτέλεσμα ο σχηματισμός των ομάδων να είναι πολύ καλός. Αυτή η ιδιότητα δεν ίσχυε σε τόσο μεγάλο βαθμό για $M=3$ καθώς βλέπαμε ότι οι αποστάσεις μεταξύ των παραδειγμάτων κάθε ομάδας ήταν αρκετά μεγαλύτερες. Για $M=6$ είχαμε ένα ικανοποιητικό αποτέλεσμα αλλά εδώ και μικρότερο σφάλμα έχουμε και μικρότερες αποστάσεις από το κέντρο κάθε ομάδας για κάθε

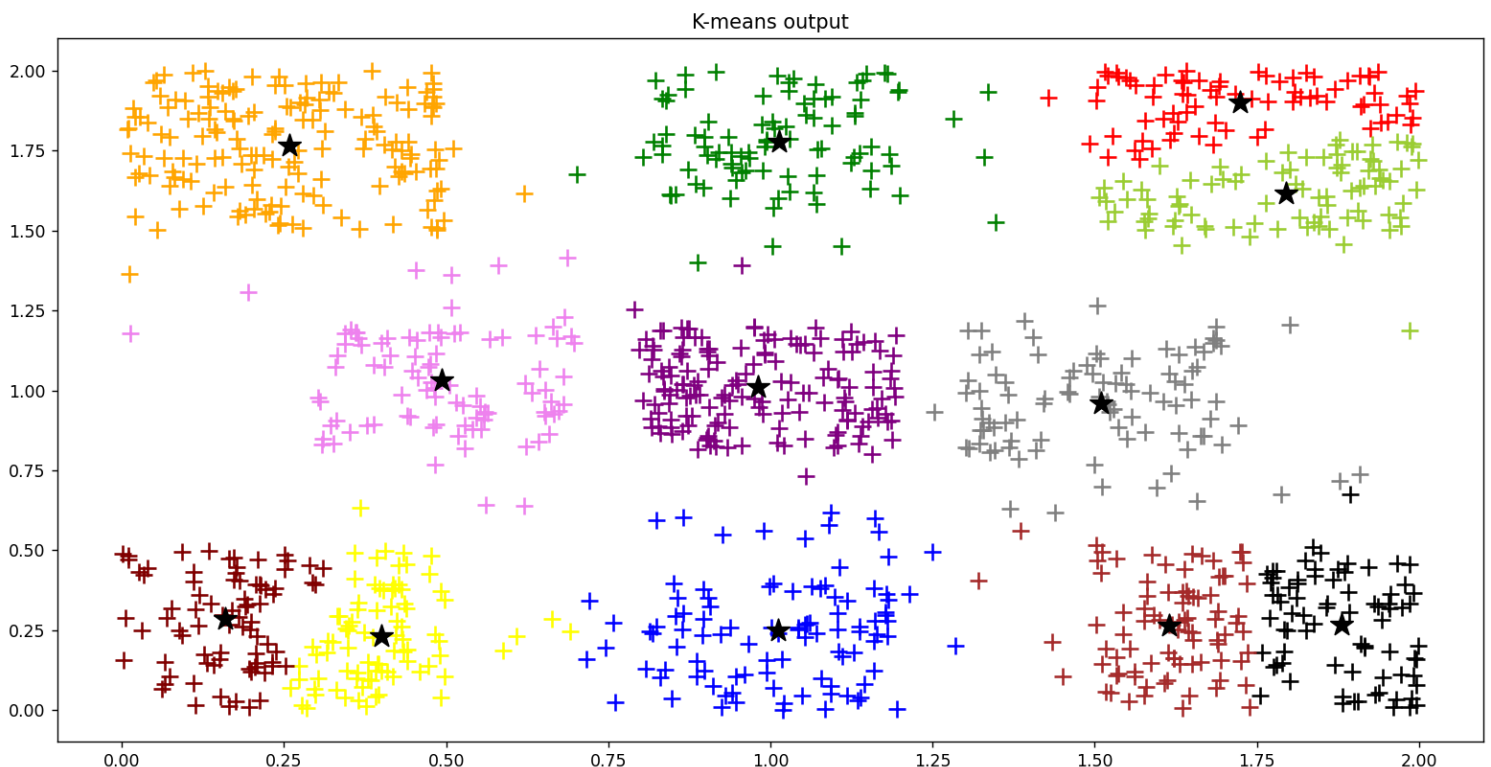
παράδειγμα. Φυσικά και η θέση των κέντρων είναι η καλύτερη μέχρι τώρα αφού τα 9 κέντρα(9 αστέρια=9 ομάδες) βρίσκονται σε σημεία όπου γύρω από αυτά έχουμε το μεγαλύτερο μέρος , αν όχι όλο, των παραδειγμάτων κάθε ομάδας.

```
The new centers are:1.012693,1.777018
The new centers are:1.760244,1.753262
The new centers are:1.744733,0.264486
The new centers are:0.981278,1.009853
The new centers are:1.513102,0.958775
The new centers are:0.278839,0.256650
The new centers are:0.257894,1.765539
The new centers are:1.007975,0.251484
The new centers are:0.493576,1.032394
Iterations:15
The total team error is:225.237823
```

Οι επαναλήψεις που χρειάστηκαν για $M=9$ για να έχουμε την τελική λύση ήταν 15. Επιπλέον, φαίνονται και οι τελικές θέσεις των 9 κέντρων όπως και το ελάχιστο σφάλμα ομαδοποίησης. Προχωράμε για $M=12$ ομάδες.

M=12

Τώρα επειδή έχουμε μεγαλύτερο αριθμό ομάδων αναμένουμε το σφάλμα ομαδοποίησης να μειωθεί και άλλο.



Μετά από 15 εκτελέσεις βρίσκουμε μικρότερο σφάλμα ομαδοποίησης ίσο με 201.638275. Το συγκεκριμένο σφάλμα είναι μειωμένο σε σχέση με το σφάλμα που είχαμε για $M=9$ αλλά δεν έχει πάρα πολύ μεγάλη απόκλιση διότι η λύση που είχαμε για $M=9$ ήταν μια πάρα πολύ καλή λύση , αν όχι η καλύτερη, στο συγκεκριμένο σύνολο δεδομένων. Συνεπώς , στο παραπάνω σχήμα με τις 12 ομάδες

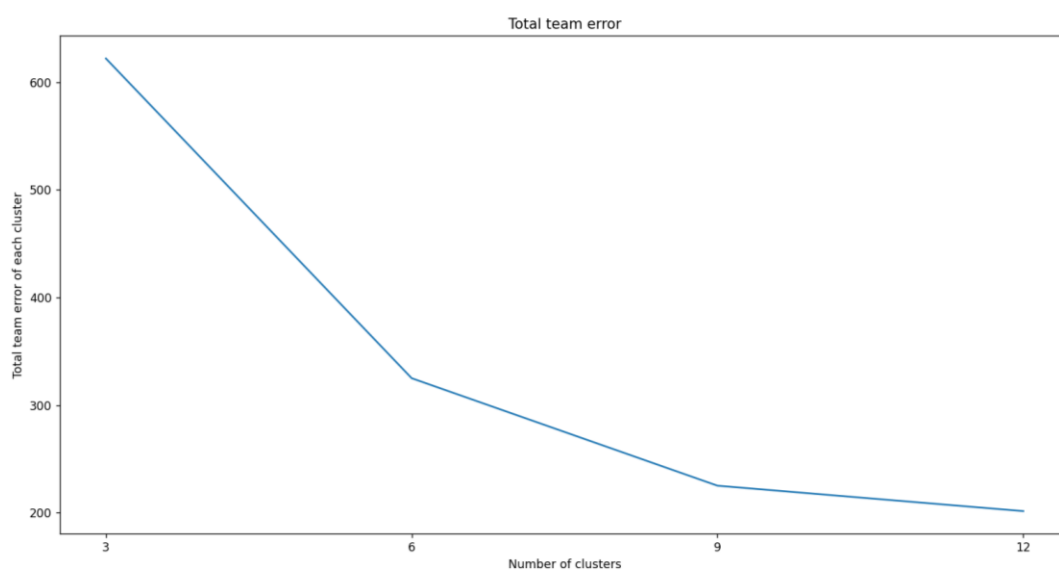
(12 κέντρα=12 αστέρια=12 διαφορετικά χρώματα) βλέπουμε ότι και πάλι στις περισσότερες ομάδες η απόσταση των παραδειγμάτων από το κέντρο της ομάδας τους είναι αρκετά μικρή και άρα η λύση είναι αρκετά καλή. Όμως ο αριθμός των ομάδων που γίνεται ομαδοποίηση του συνόλου παραδειγμάτων είναι αρκετά μεγάλος και η λύση για $M=9$ σε οπτική αναπαράσταση έδινε καλύτερο αποτέλεσμα. Τα παραδείγματα έτσι όπως είναι τοποθετημένα στο σχήμα έχουν την μορφή «νέφων» και άρα είναι προτιμότερο να έχουμε σαν ξεχωριστή ομάδα κάθε νέφος, όπου αυτό επιτυγχάνεται για $M=9$. Άρα για $M=9$ έχουμε την καλύτερη λύση από αυτές που εξετάσαμε.

```
The new centers are:0.399878,0.230936
The new centers are:1.880679,0.268516
The new centers are:1.615410,0.265941
The new centers are:1.509468,0.960565
The new centers are:0.493576,1.032394
The new centers are:0.981278,1.009853
The new centers are:0.257894,1.765539
The new centers are:1.723904,1.901000
The new centers are:0.158673,0.284553
The new centers are:1.795014,1.615880
The new centers are:1.012693,1.777018
The new centers are:1.011440,0.249650
Iterations:9
The total team error is:201.638275
```

Για την σύγκλιση του αλγορίθμου για $M=12$ χρειάστηκαν συνολικά 9 επαναλήψεις. Επίσης, βλέπουμε και τις τελικές θέσεις των 12 κέντρων με το τέλος του k-means που επιβεβαιώνονται στο παραπάνω σχήμα.

Τώρα με βάση τα παραπάνω αποτελέσματα ακολουθεί το διάγραμμα μεταξύ του σφάλματος ομαδοποίησης και τον αριθμό ομάδων.

ΔΙΑΓΡΑΜΜΑ ΣΦΑΛΜΑΤΟΣ ΟΜΑΔΟΠΟΙΗΣΗΣ- ΑΡΙΘΜΟ ΟΜΑΔΩΝ



Τα σημεία που βάλαμε στο παραπάνω διάγραμμα είναι τα:

(3, 621.960449) , (6, 325.041260) , (9, 225.237823) , (12, 201.638275) . Στον x άξονα έχουμε τον αριθμό των ομάδων και στον y άξονα το ελάχιστο σφάλμα ομαδοποίησης ανά cluster. Η μορφή που έχει η παραπάνω γραφική παράσταση φαίνεται στο παραπάνω διάγραμμα. Παρατηρούμε , ότι το σφάλμα ομαδοποίησης αρχικά μειώνεται με πάρα πολύ γρήγορο ρυθμό (από $M=3$ έως και $M=9$) ενώ από $M=9$ έως και $M=12$ και έπειτα αρχίζει να παρουσιάζει μια πιο ομαλή συμπεριφορά.

Αυτό που μπορούμε να επιβεβαιώσουμε με βάση το παραπάνω διάγραμμα είναι ότι όντως ο πραγματικός αριθμός των ομάδων στο συγκεκριμένο σύνολο δεδομένων είναι 9(αναφέρεται και στην εκφώνηση) καθώς από εκείνο το σημείο και έπειτα η καμπύλη τείνει να πάρει την μορφή μιας ευθείας γραμμής και άρα δεν υπάρχει κάποια αξιοσημείωτη παρατήρηση για $M=9$ και μετά, επειδή ουσιαστικά έχουμε βρει τον κατάλληλο αριθμό ομάδων για το συγκεκριμένο ΣΔΟ.

Γενικεύοντας, μπορούμε να αναφέρουμε ότι χρησιμοποιώντας το σφάλμα ομαδοποίησης **μπορούμε να βρούμε τον πραγματικό αριθμό ομάδων** εστιάζοντας την προσοχή μας στο σημείο της καμπύλης όπου από εκεί και πέρα το σφάλμα μειώνεται με πάρα πολύ αργό ρυθμό σε σχέση με πριν(όπου μειωνόταν πάρα πολύ γρήγορα) . Στην περίπτωση μας από 225 περίπου μειώνεται στο 201 ενώ πριν πάρει την τιμή 225 ήταν στο 325 και η άρα η απόκλιση ήταν αρκετά μεγαλύτερη.

Βέβαια το σφάλμα ομαδοποίησης δεν μπορεί να μας προσφέρει κάποια χρήσιμη πληροφορία όταν έχουμε τον ελάχιστο αριθμό ομάδων ($K=1$) δηλαδή όλα τα δεδομένα να ανήκουν σε μια ομάδα όπου εκεί το σφάλμα ομαδοποίησης έχει πάρα πολύ μεγάλη τιμή , τείνει ουσιαστικά στο άπειρο ή αντιθέτως όταν έχουμε τόσες ομάδες όσα είναι και τα παραδείγματα όπου πρακτικά κάθε παράδειγμα είναι και μια ομάδα με συνέπεια το σφάλμα ομαδοποίησης να είναι 0. Άρα σε αυτές τις 2 ακραίες περιπτώσεις το σφάλμα ομαδοποίησης δεν θα μπορεί να μας βοηθήσει ουσιαστικά.

ΚΑΤΑΛΟΓΟΣ ΓΙΑ TURNIN

Στον κατάλογο που θα υποβάλλουμε για την 2^η άσκηση υπάρχει το αρχείο **examples.c** το οποίο δημιουργεί το αρχείο examplesSDO.txt με τα παραδείγματα.

Περιέχεται επίσης το αρχείο **kmeans.c** που περιέχει την υλοποίηση του k_means.

Περιέχεται και το αρχείο **script.py** που είναι υπεύθυνο για την δημιουργία και την παρουσίαση των σχημάτων και των διαγραμμάτων.

Φυσικά περιέχεται και το report σε μορφή pdf.