Αναφορά Εργασίας Υπολογιστικής Νοημοσύνης ΜΥΕ-035

Παναγιώτης Ζούμπας 4873
‘Αγγελος Μπιριντζης 4741
Αδαμάντιος Δημήτριος Κάλλης 4685

Άσκηση 2 – Σύνολα Δεδομένων Ομαδοποίησης (ΣΔΟ) και k-means

Σύνοψη Κλάσεων – Άσκηση 2 (ΣΔΟ + k-means)

SDOLoader: Φορτώνει τα δεδομένα του συνόλου ομαδοποίησης από αρχείο CSV.
Τα δεδομένα δεν περιέχουν ετικέτες, καθώς το πρόβλημα είναι μη επιβλεπόμενης μάθησης.
Χρησιμοποιείται αποκλειστικά από τον αλγόριθμο k-means.

KMeans: Υλοποιεί τον αλγόριθμο k-means. Περιλαμβάνει τα βήματα ανάθεσης σημείων στα κοντινότερα κέντρα, ενημέρωσης των κέντρων και υπολογισμού του σφάλματος SSE.
Η Ευκλείδεια απόσταση χρησιμοποιείται ως μέτρο ομοιότητας, όπως ζητείται στην εκφώνηση.

KMeansExperimentMain (ask2): Η κλάση **KMeansExperimentMain** της Άσκησης 2 εκτελεί τον k-means για διαφορετικές τιμές του k. Για κάθε τιμή, υπολογίζει το σφάλμα SSE και επιτρέπει τη σύγκριση των αποτελεσμάτων, ώστε να μελετηθεί η επίδραση του αριθμού ομάδων στην ποιότητα της ομαδοποίησης.

GenerateSDO: Δημιουργεί το Σύνολο Δεδομένων Ομαδοποίησης (ΣΔΟ).
Παράγει σημεία στο επίπεδο (x_1, x_2)χωρίς ετικέτες κατηγορίας, καθώς το πρόβλημα αφορά μη επιβλεπόμενη μάθηση. Τα δεδομένα αυτά χρησιμοποιούνται αποκλειστικά από τον αλγόριθμο k-means για τη μελέτη της ομαδοποίησης και του σφάλματος SSE.

Point: Αναπαριστά ένα σημείο στο δισδιάστατο επίπεδο.
Αποθηκεύει τις συντεταγμένες x_1και x_2και παρέχει βοηθητικές μεθόδους, όπως τον υπολογισμό της τετραγωνικής Ευκλείδειας απόστασης από ένα άλλο σημείο.
Χρησιμοποιείται τόσο στην υλοποίηση του k-means όσο και στον υπολογισμό του σφάλματος ομαδοποίησης.

KMeansResult: Χρησιμοποιείται για την αποθήκευση των αποτελεσμάτων του αλγορίθμου k-means. Περιέχει τα τελικά κέντρα των ομάδων, την ανάθεση κάθε σημείου σε ομάδα και την τιμή του σφάλματος SSE. Η ύπαρξή της διευκολύνει την αποθήκευση, σύγκριση και αξιολόγηση διαφορετικών εκτελέσεων του k-means για διαφορετικές τιμές του k.

Φόρτωση δεδομένων ΣΔΟ

Η φόρτωση των δεδομένων πραγματοποιείται στην κλάση **SDOLoader**. Η κλάση αυτή διαβάζει τα δεδομένα από αρχείο CSV και δημιουργεί ένα σύνολο από σημεία χωρίς καμία πληροφορία κατηγορίας. Κάθε παράδειγμα αποθηκεύεται ως αντικείμενο της κλάσης **Point**, το οποίο περιέχει τις συντεταγμένες του σημείου στο επίπεδο.

Αναπαράσταση σημείων

Η κλάση **Point** χρησιμοποιείται για την αναπαράσταση κάθε παραδείγματος του ΣΔΟ. Κάθε αντικείμενο Point περιέχει τις συντεταγμένες του σημείου και παρέχει βοηθητικές μεθόδους για τον υπολογισμό αποστάσεων, οι οποίες είναι απαραίτητες για τον αλγόριθμο k-means.

Υλοποίηση αλγορίθμου k-means

Ο ίδιος ο αλγόριθμος k-means υλοποιείται στην κλάση **KMeans**. Η κλάση αυτή είναι υπεύθυνη για:

- την τυχαία αρχικοποίηση των κέντρων των ομάδων,
- την ανάθεση κάθε σημείου στην κοντινότερη ομάδα βάσει Ευκλείδειας απόστασης,
- την επαναϋπολογισμό των κέντρων ως μέσο όρο των σημείων κάθε ομάδας,
- την επαναληπτική εκτέλεση της διαδικασίας μέχρι τη σύγκλιση.

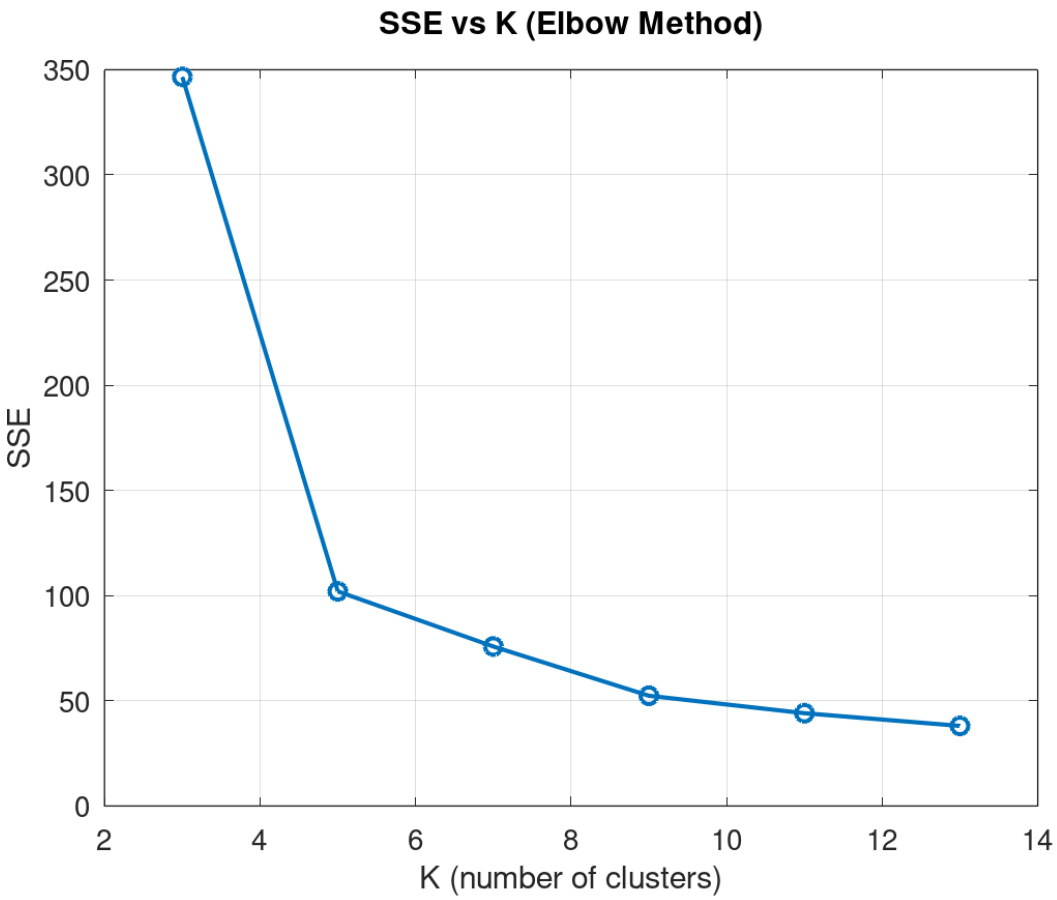
Υπολογισμός σφάλματος ομαδοποίησης (SSE)

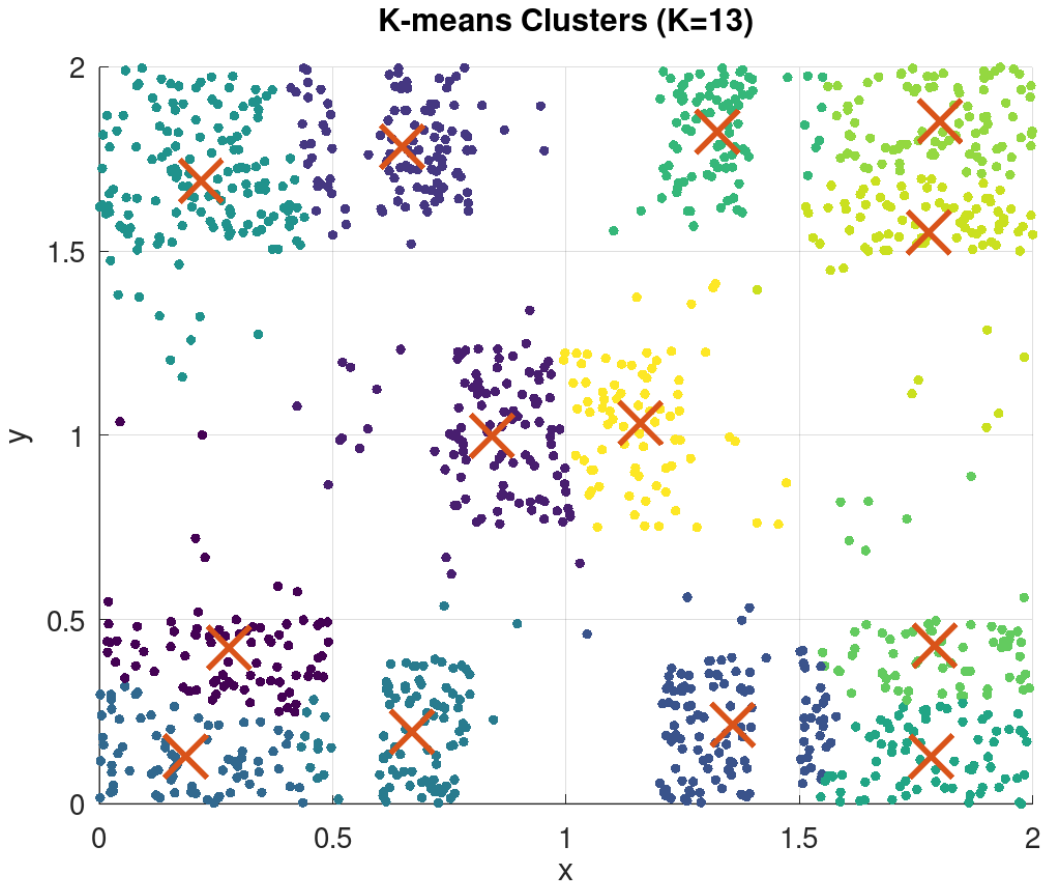
Ο υπολογισμός του σφάλματος **SSE (Sum of Squared Errors)** πραγματοποιείται επίσης στην κλάση **KMeans**. Για κάθε σημείο υπολογίζεται η τετραγωνική Ευκλείδεια απόσταση από το κέντρο της ομάδας στην οποία έχει ανατεθεί και όλες οι αποστάσεις αθροίζονται. Ο υπολογισμός αυτός ακολουθεί πιστά τον ορισμό της εκφώνησης και χρησιμοποιείται ως μέτρο αξιολόγησης της ποιότητας της ομαδοποίησης.

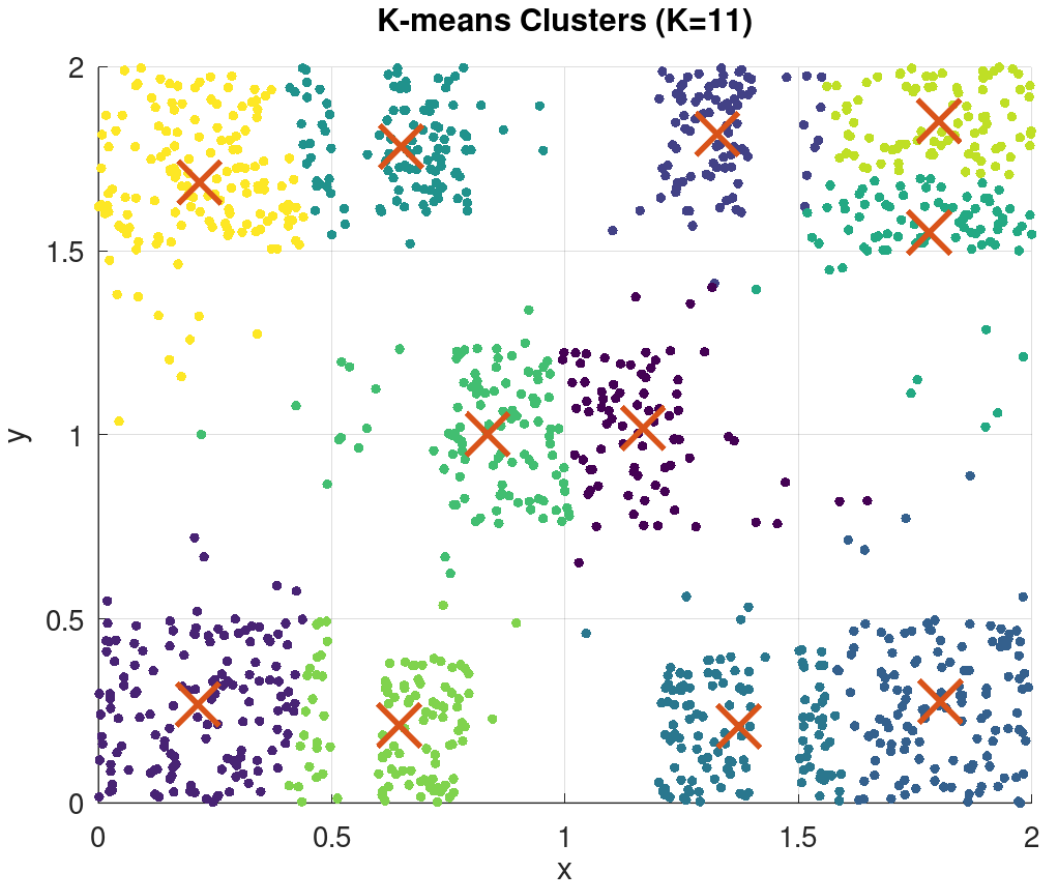
Εκτέλεση πειραμάτων για διαφορετικές τιμές k

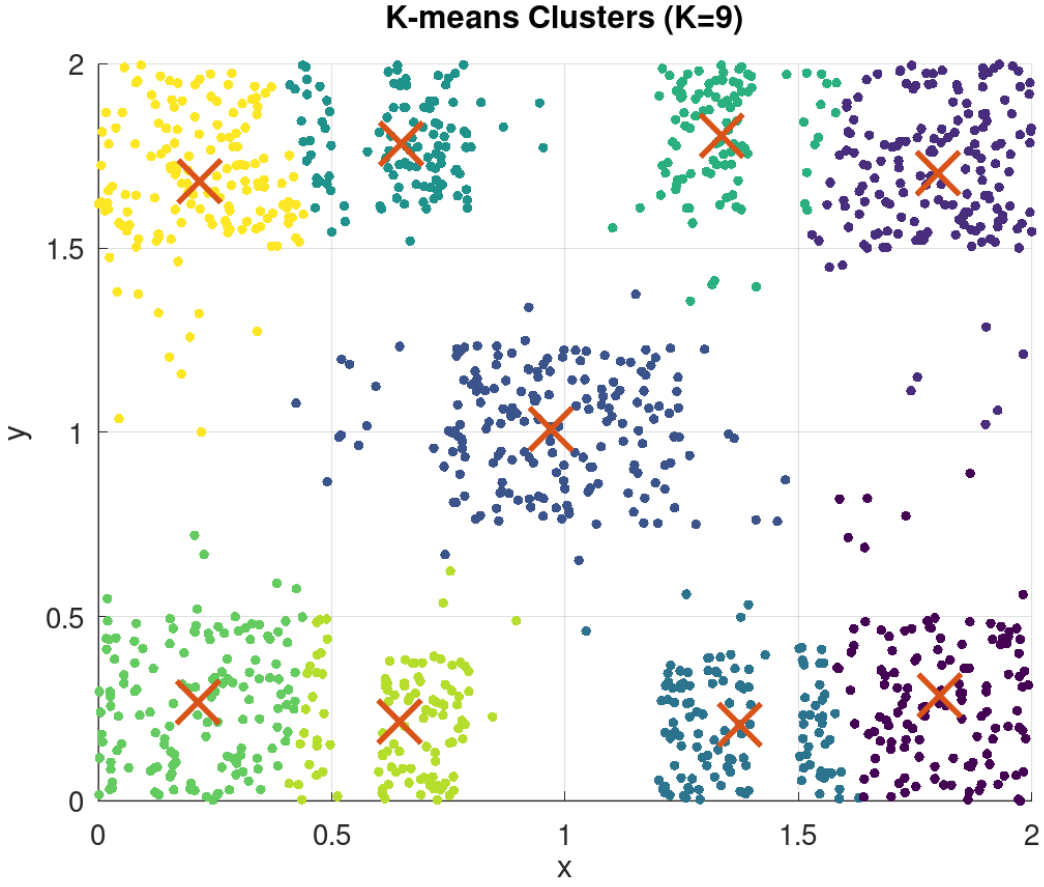
Η εκτέλεση του k-means για διαφορετικές τιμές του αριθμού ομάδων kπραγματοποιείται στην κλάση **ExperimentMain** της Άσκησης 2. Η κλάση αυτή καλεί τον αλγόριθμο k-means, συλλέγει τα αποτελέσματα και καταγράφει το αντίστοιχο SSE για κάθε τιμή του k, ώστε να είναι δυνατή η σύγκριση των λύσεων.

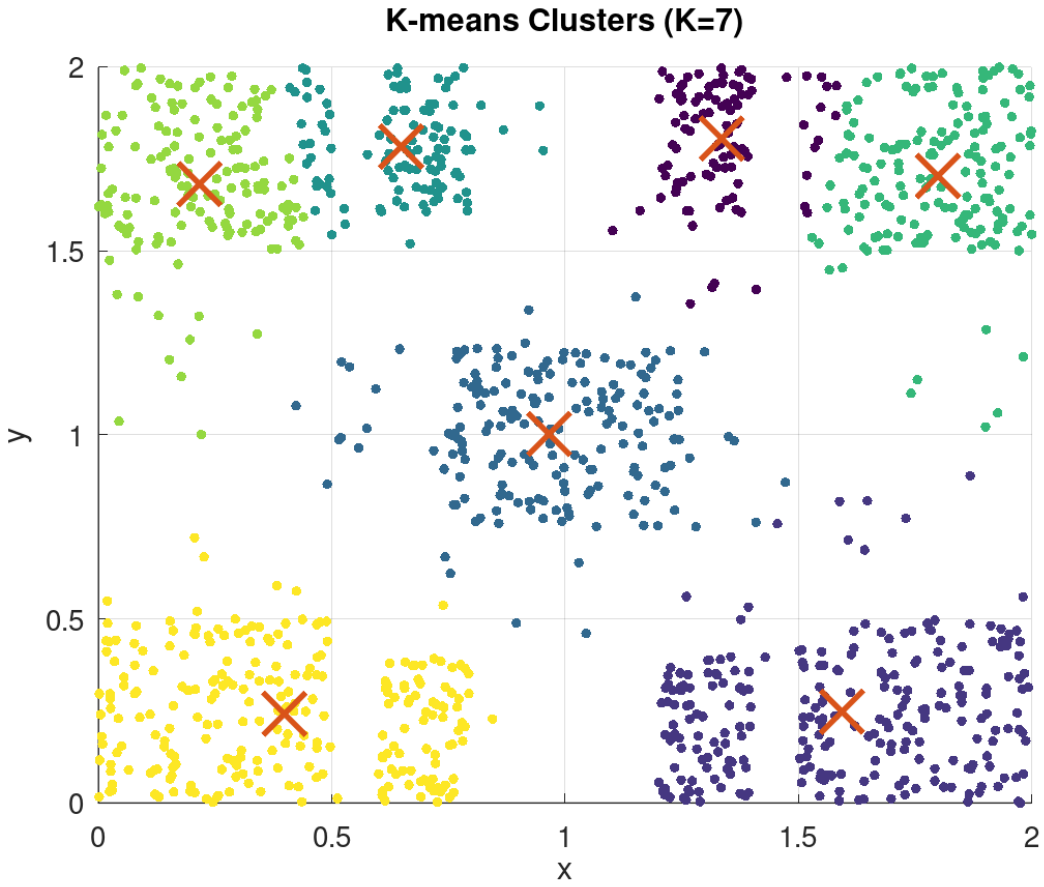
Παρακάτω ακολούθουν τα plot όπου τα παραδείγματα συμβολίζονται με data points και τα κέντρα με (X).

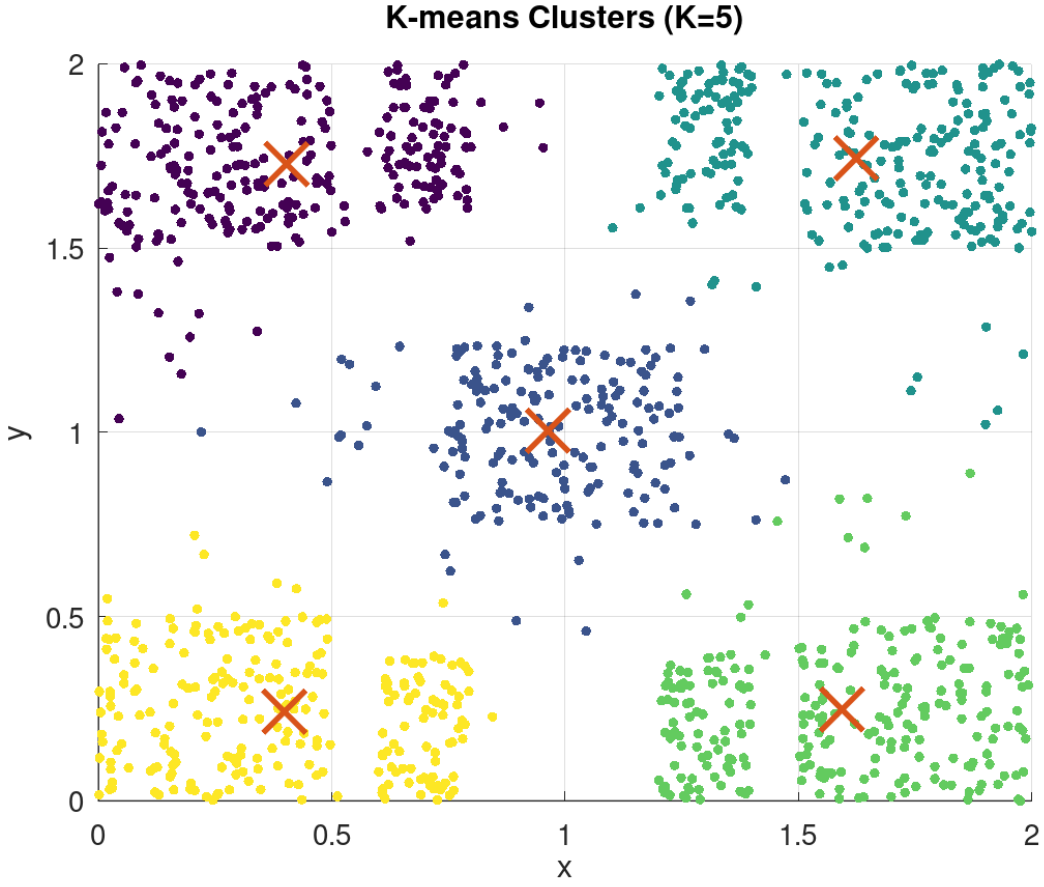

SSE vs K (Elbow Method)
SSE
K (number of clusters)
350
300
250
200
150
100
50
0
2
4
6
8
10
12
14

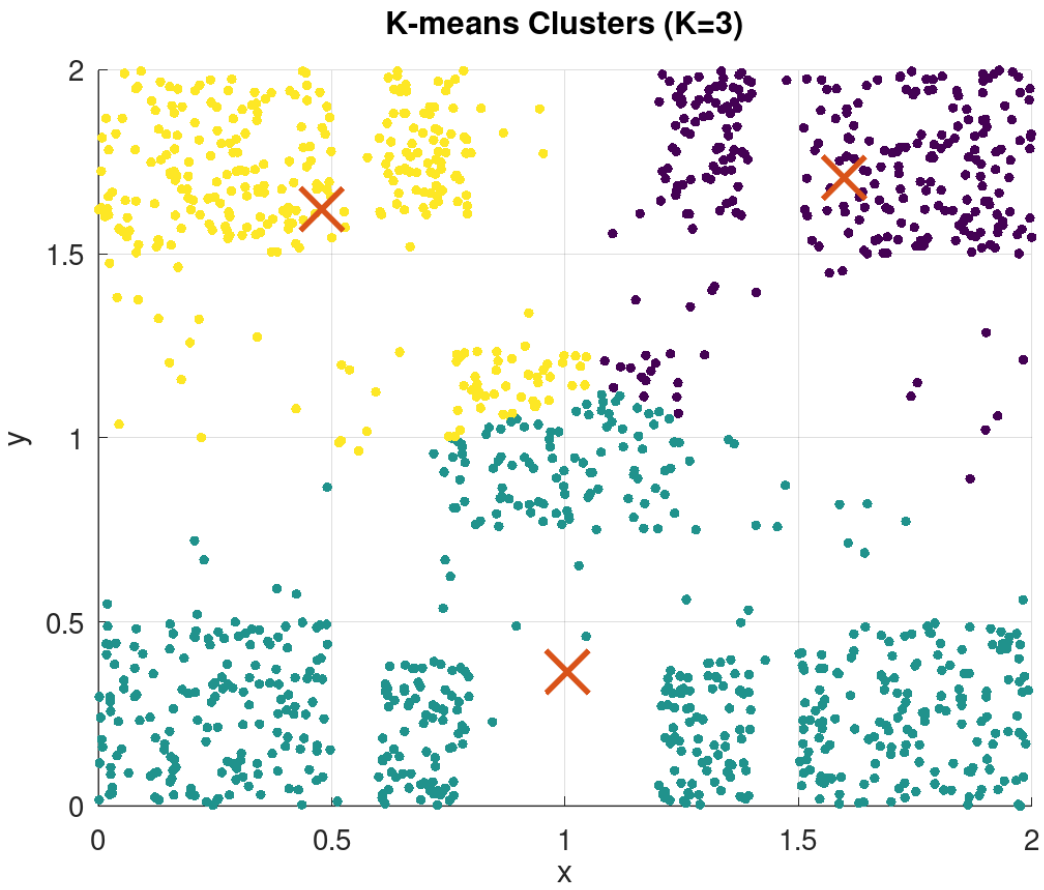

K-means Clusters (K=13)
y
x
2
1.5
1
0.5
0
0
0.5
1
1.5
2


K-means Clusters (K=11)
2
1.5
1
0.5
0
y
0
0.5
1
1.5
2
x


K-means Clusters (K=9)
2
1.5
1
0.5
0
y
0
0.5
1
1.5
2
x


K-means Clusters (K=7)
2
1.5
1
0.5
0
y
0
0.5
1
1.5
2
x


K-means Clusters (K=5)
2
1.5
1
0.5
0
y
0
0.5
1
1.5
2
x



Παρατηρείται ότι όσο αυξάνεται ο αριθμός των ομάδων, το σφάλμα SSE μειώνεται, καθώς τα σημεία κατανέμονται σε περισσότερες και μικρότερες ομάδες. Ωστόσο, η μείωση αυτή δεν είναι γραμμική. Συνήθως παρατηρείται ένα σημείο στο οποίο η μείωση του σφάλματος αρχίζει να γίνεται πολύ μικρότερη.

Στην παρούσα άσκηση, όπου ο πραγματικός αριθμός ομάδων του συνόλου ΣΔΟ είναι 9, αναμένεται ότι το διάγραμμα SSE–k θα παρουσιάζει μια αισθητή αλλαγή κλίσης κοντά στην τιμή k = 9. Συνεπώς, το σφάλμα ομαδοποίησης μπορεί να χρησιμοποιηθεί ως εργαλείο για την εκτίμηση του πραγματικού αριθμού ομάδων, αν και η εκτίμηση αυτή δεν είναι πάντα απολύτως ακριβής και βασίζεται κυρίως σε εμπειρική παρατήρηση.

Για την εκτελεση του προγραμμάτος εκτελείτε μέσα στον φάκελο src :

javac -encoding UTF-8 ask2/*.java

java ask2.KMeansExperiment