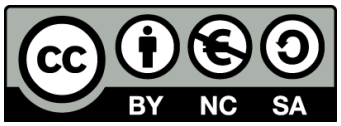
ΑΡΙΣΤΟΤΕΛΕΙΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΘΕΣΣΑΛΟΝΙΚΗΣ

ΑΝΟΙΚΤΑ
ΑΚΑΔΗΜΑΪΚΑ
ΜΑΘΗΜΑΤΑ

Αποθήκες Δεδομένων και Εξόρυξη Δεδομένων

Ενότητα 7: Ομαδοποίηση – Μέρος Α΄

Αναστάσιος Γούναρης, Επίκουρος Καθηγητής
Τμήμα Πληροφορικής ΑΠΘ



Άδειες Χρήσης

- Το παρόν εκπαιδευτικό υλικό υπόκειται σε άδειες χρήσης Creative Commons.
- Για εκπαιδευτικό υλικό, όπως εικόνες, που υπόκειται σε άλλου τύπου άδειας χρήσης, η άδεια χρήσης αναφέρεται ρητώς.

Χρηματοδότηση

- Το παρόν εκπαιδευτικό υλικό έχει αναπτυχθεί στα πλαίσια του εκπαιδευτικού έργου του διδάσκοντα.
- Το έργο «Ανοικτά Ακαδημαϊκά Μαθήματα στο Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης» έχει χρηματοδοτήσει μόνο την αναδιαμόρφωση του εκπαιδευτικού υλικού.
- Το έργο υλοποιείται στο πλαίσιο του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» και συγχρηματοδοτείται από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο) και από εθνικούς πόρους.

ΑΡΙΣΤΟΤΕΛΕΙΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΘΕΣΣΑΛΟΝΙΚΗΣ

ΑΝΟΙΚΤΑ
ΑΚΑΔΗΜΑΪΚΑ
ΜΑΘΗΜΑΤΑ

Ομαδοποίηση – Μέρος Α΄

Λειτουργία της ομαδοποίησης, τύποι συστάδων, αλγόριθμος k-means

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ ΚΑΙ ΘΡΗΣΚΕΥΜΑΤΩΝ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ
ΕΣΠΑ 2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Περιεχόμενα ενότητας

1. Αλγόριθμοι ομαδοποίησης:
   i. Αλγόριθμοι τμηματοποίησης.

Σκοποί ενότητας

- Παρουσίαση της λειτουργίας της ομαδοποίησης.
- Παρουσίαση των τύπων συστάδων.
- Περιγραφή του k-means αλγόριθμου, καθώς και παραλλαγές του.

Τί είναι ομαδοποίηση – ανάλυση συστάδων

- Συστάδα: συλλογή από αντικείμενα
  - που είναι όμοια ή σχετίζονται με κάποιο τρόπο με τα υπόλοιπα αντικείμενα της ομάδας,
  - και που δεν είναι (τόσο) όμοια ή δεν σχετίζονται με τα αντικείμενα άλλων ομάδων.
- Ανάλυση συστάδων:
  - Ομαδοποίηση αντικειμένων σε συστάδες.
- Η ομαδοποίηση μπορεί να θεωρηθεί ως μη επιβλεπόμενη κατηγοριοποίηση: δεν υπάρχουν προκαθορισμένες κατηγορίες.
- Τυπικές εφαρμογές:
  - Ως αυτόνομο εργαλείο ανάλυσης και κατανόησης δεδομένων.
  - Ως προεπεξεργασία δεδομένων.

Παραδείγματα εφαρμογών (1/2)

- **Χρήση γης**: Εντοπισμός περιοχών με παρόμοια χρήση γης σε μία βάση δεδομένων με φωτογραφίες από δορυφόρους.
- **Χωροταξικά σχέδια**: Εντοπισμός ομάδων κτιρίων σύμφωνα με τον τύπο κατοικίας, αξία, γεωγραφική περιοχή, κλπ.
- **Σεισμογραφικές μελέτες**: Ομαδοποίηση βάσει επικέντρων και άλλων χαρακτηριστικών των σεισμών.

Παραδείγματα εφαρμογών (2/2)

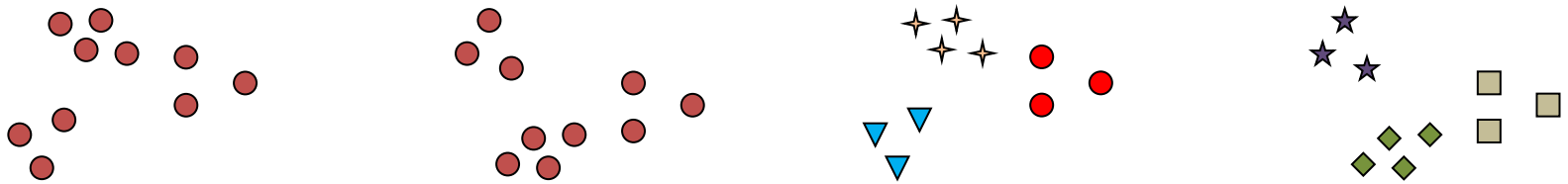
- **Ανάλυση προτύπων – επεξεργασία εικόνας**.
- **Δεδομένα παγκοσμίου ιστού**: ομαδοποίηση ιστοσελίδων, ομαδοποίηση συμπεριφορών χρήσης.
- **Marketing**: Βοήθεια στην ανακάλυψη ομάδων πελατών για στοχευμένη διαφήμιση/marketing.

Πότε η ομαδοποίηση είναι καλή;

- Μία καλή μέθοδος ομαδοποίησης παράγει συστάδες:
  - Με μεγάλη ομοιότητα μεταξύ των αντικειμένων στην ίδια συστάδα.
  - Και μικρή ομοιότητα μεταξύ των αντικειμένων σε διαφορετικές συστάδες.
- Η ποιότητα εξαρτάται από:
  - Τη μέθοδο υλοποίησης της τεχνικής ομαδοποίησης.
  - Το μέτρο ομοιότητας που χρησιμοποιείται.
- Επίσης εξαρτάται από το πόσο μπορεί να ανακαλύψει κρυφά πρότυπα στα δεδομένα.

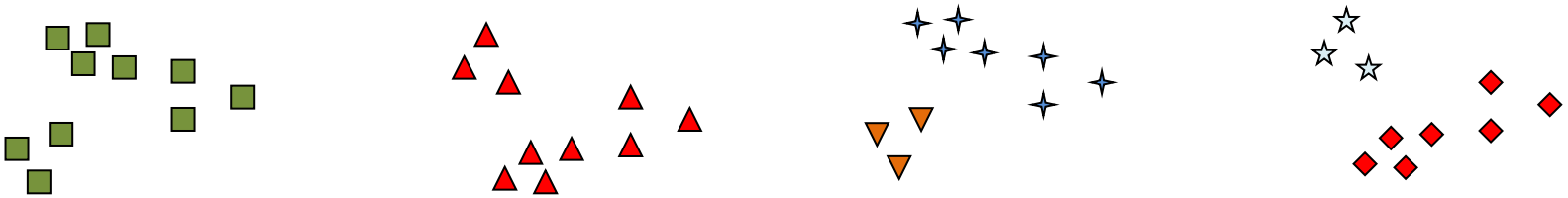
Έννοια συστάδας

- Αλλά η έννοια της συστάδας δεν είναι πάντοτε αντικειμενική.



Πόσα clusters?

6 Clusters



2 Clusters

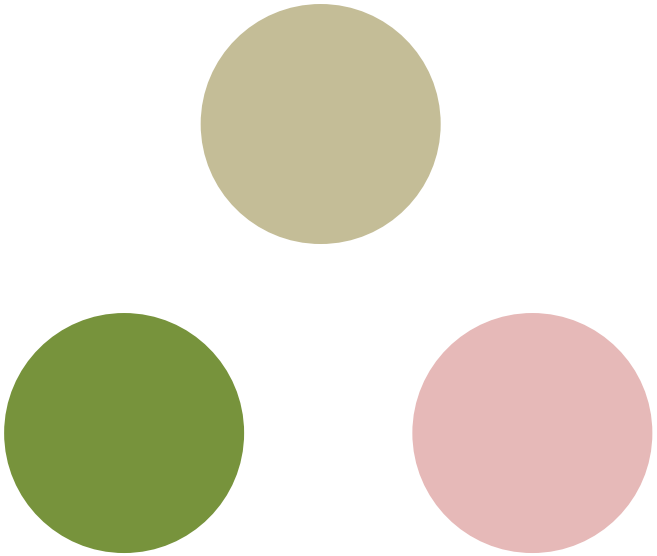
4 Clusters

Διαχωρισμοί μεταξύ ομαδοποιήσεων

- Ιεραρχική vs Διαχωριστική (Partitional) ομαδοποίηση:
  - Η ιεραρχική ομαδοποίηση παράγει εμφωλευμένες συστάδες που ανήκουν σε μια ιεραρχία.
  - Η διαχωριστική παράγει μη επικαλυπτόμενες συστάδες.
- Αποκλειστική vs μη-αποκλειστική vs ασαφής.
- Πλήρης vs μερική.

Τύποι Συστάδων: Καλά διαχωρισμένες

- Καλά διαχωρισμένες (Well-Separated) συστάδες:
  - Οι συστάδες είναι σύνολα σημείων όπου κάθε σημείο είναι πιο κοντά σε κάθε άλλο σημείο της ίδιας συστάδας παρά σε κάποιο σημείο άλλης συστάδας.



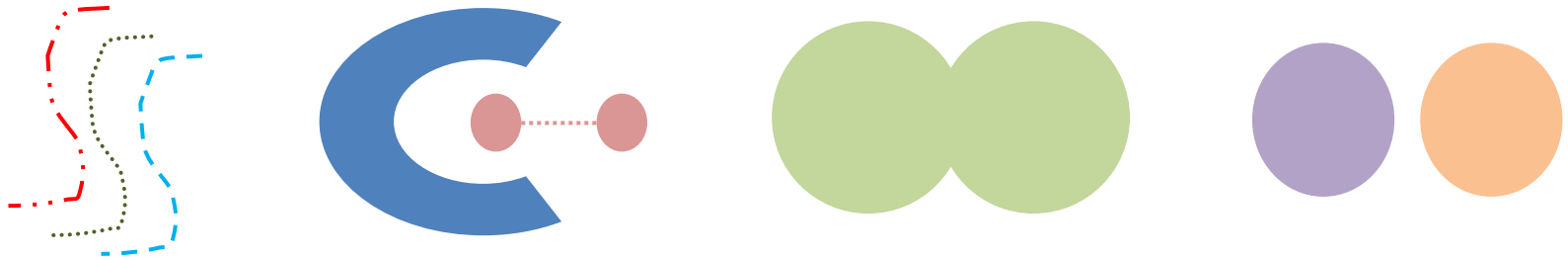
3 καλά διαχωρισμένες συστάδες

Τύποι Συστάδων: Βασισμένες στο κέντρο

- Η συστάδα περιέχει τα αντικείμενα που είναι κοντύτερα στο δικό της «κέντρο».
  - Centroid vs medoid

Άλλοι Τύποι Συστάδων

- Βάσει γειτνίασης
  - Κάθε σημείο της συστάδας είναι πιο κοντά σε τουλάχιστον ένα άλλο σημείο της συστάδας παρά σε κάθε σημείο εκτός συστάδας.
  - Οι αντίστοιχες τεχνικές βασίζονται κυρίως σε θεωρίες γράφων.
  - Παράδειγμα με 8 συστάδες:



  - Βάσει πυκνότητας.
  - Πιο γενικά, βάσει κοινών ιδιοτήτων:

Γενικές Απαιτήσεις στην ΕΔ

- Κλιμάκωση.
- Δυνατότητα χειρισμού διαφορετικών τύπων δεδομένων.
- Ανακάλυψη συστάδων με οποιοδήποτε σχήμα.
- Ελάχιστη ή μηδαμινή απαίτηση γνώσης πεδίου (domain knowledge).
- Ικανότητα χειρισμού θορύβου και ανωμαλιών.
- Όχι εξάρτηση από την σειρά των δεδομένων.
- Χειρισμός πολλών διαστάσεων.
- Εισαγωγή περιορισμών.
- Ευχρηστία και εύκολη κατανόηση/επεξήγηση.

Μέτρα ομοιότητας/απόστασης: επανάληψη

- Ιδιότητες μέτρων ομοιότητας:
  - Ανακλαστική.
  - Συμμετρική.
- ΑΠΟΣΤΑΣΗ=*ΜΟΝΟΤΟΝΑ_ΦΘΗΝΟΥΣΑ*(ΟΜΟΙΟΤΗΤΑ).
- Ιδιότητες μετρικού απόστασης:
  - Θετικότητα: $d(x, y) \geq 0$
  - Ανακλαστική: $d(x, y) = 0 \Leftrightarrow x = y$
  - Συμμετρική: $d(x, y) = d(y, x)$
  - Τριγωνική ανισότητα: $d(x, y) \leq d(x, z) + d(z, y)$

Αλγόριθμοι Τμηματοποίησης (Partitioning Algorithms)

- **Μεθοδολογία**: Δημιουργία τμηματοποίησης μίας ΒΔ ***D*** με ***n*** αντικείμενα σε ένα σύνολο ***k*** συστάδων.
- Δεδομένου του *k*, πρέπει να βρεθεί η τμηματοποίηση που βελτιστοποιεί το κριτήριο τμηματοποίησης.
  - Βέλτιστη λύση: πρέπει να εξεταστούν όλες οι περιπτώσεις.
  - Ευρετικές μέθοδοι: *k-means, k-medoids , k-nn* .
  - ***k-means*** (MacQueen'67): Κάθε συστάδα αντιπροσωπεύεται από το κέντρο της.
  - ***k-medoids*** or PAM (Partition around medoids) (Kaufman & Rousseeuw'87): Κάθε συστάδα αντιπροσωπεύεται από ένα αντικείμενό της.

Συνολικός αριθμός ομαδοποιήσεων (1/3)

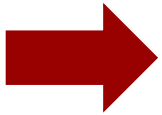
- S(n,k): αριθμός ομαδοποιήσεων n σημείων σε k ομάδες:
  - S(n-1,k-1) περιπτώσεις για k-1 ομάδες με n-1 σημεία. Το n-οστό σημείο δημιουργεί νέα ομάδα που συνδυάζεται με κάθε μία από τις S(n-1,k-1) ώστε να προκύψουν k ομάδες με n σημεία.
    - Προκύπτουν συνολικά *S(n-1,k-1)* περιπτώσεις.
  - S(n-1,k) περιπτώσεις για k ομάδες με n-1 σημεία. Το n-οστό σημείο εντάσσεται σε κάθε μία από τις ομάδες ώστε να προκύψουν k ομάδες με n σημεία.
    - Προκύπτουν συνολικά *k*S(n-1,k)* περιπτώσεις.

Συνολικός αριθμός ομαδοποιήσεων (2/3)

- Παράδειγμα: A,B,C,D σημεία. S(4,3) = ?

(A,B) (C)
(A,C) (B)
(A)(B,C)

S(3,2) = 3

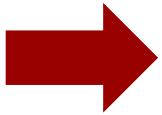


(A,B) (C) (D)
(A,C) (B) (D)
(A) (B,C) (D)

S(3,2) περιπτώσεις

(A) (B) (C)

S(3,2) = 3



(A,D) (B) (C)
(A) (B, D) (C)
(A) (B) (C, D)

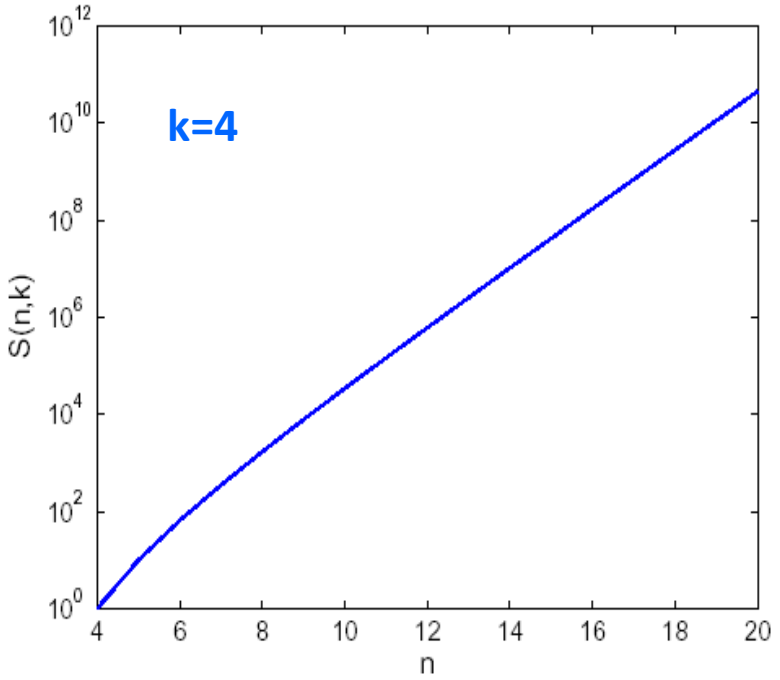
3*S(3,3) περιπτώσεις

Συνολικός αριθμός ομαδοποιήσεων (3/3)

$$S(n,k) = S(n-1,k-1) + kS(n-1,k)$$

$$S(n,1) = 1, S(n,n) = 1, S(n,k) = 0, k > n$$

$$S(n,k) = \frac{1}{k!}\sum_{i=1}^{k}(-1)^{k-1}\binom{k}{i}i^n$$

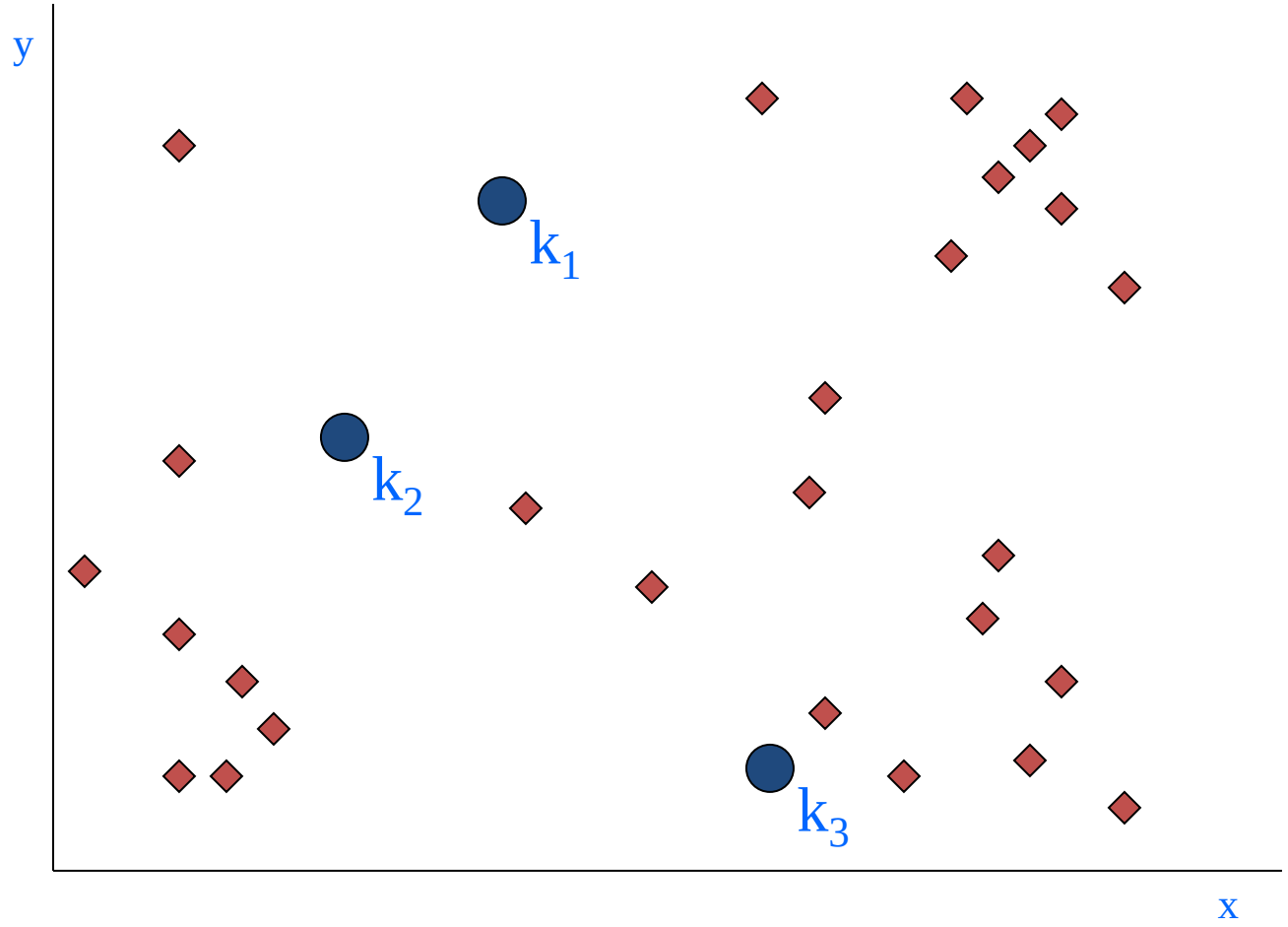


K-means

1. Διάλεξε k τυχαία κέντρα. (Τα κέντρα μπορεί να μην αντιστοιχούν σε ένα από τα δεδομένα αντικείμενα).
2. Ανάθεσε κάθε αντικείμενο στο πλησιέστερο προς αυτό κέντρο.
3. Για κάθε μία από τις k ομάδες, υπολόγισε το νέο κέντρο.
4. Αν όλα τα νέα κέντρα συμπίπτουν με τα προηγούμενα (δηλαδή, δεν υπήρξε μεταβολή), τότε τερμάτισε γιατί ο αλγόριθμος έχει συγκλίνει. Αλλιώς, επανάλαβε το βήμα 2.

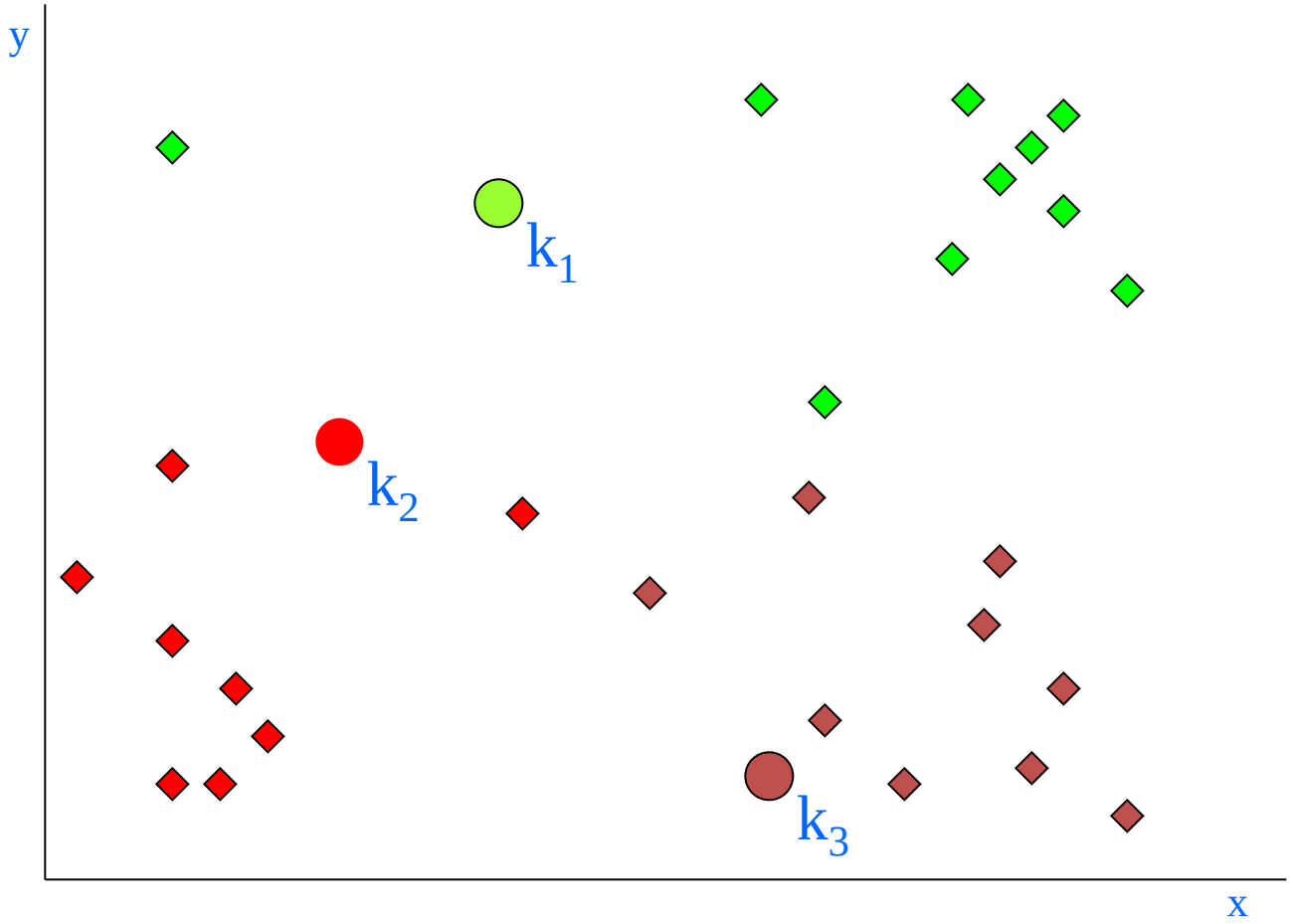
Παράδειγμα K-means: Βήμα 1

- Διάλεξε τα αρχικά 3 κέντρα των συστάδων (τυχαία).



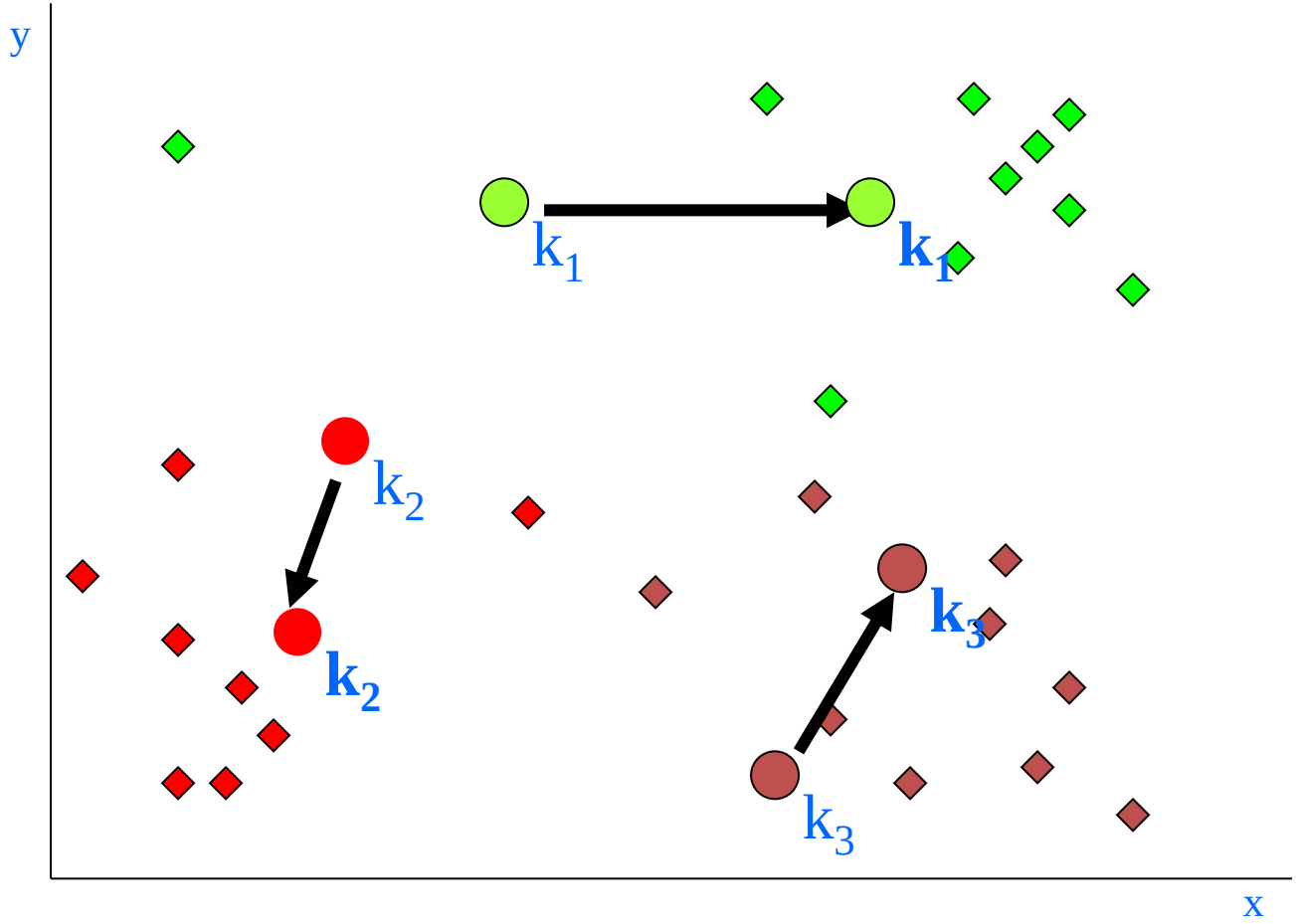
Παράδειγμα K-means: Βήμα 2

- Ανάθεσε κάθε σημείο στο κοντινότερο κέντρο.



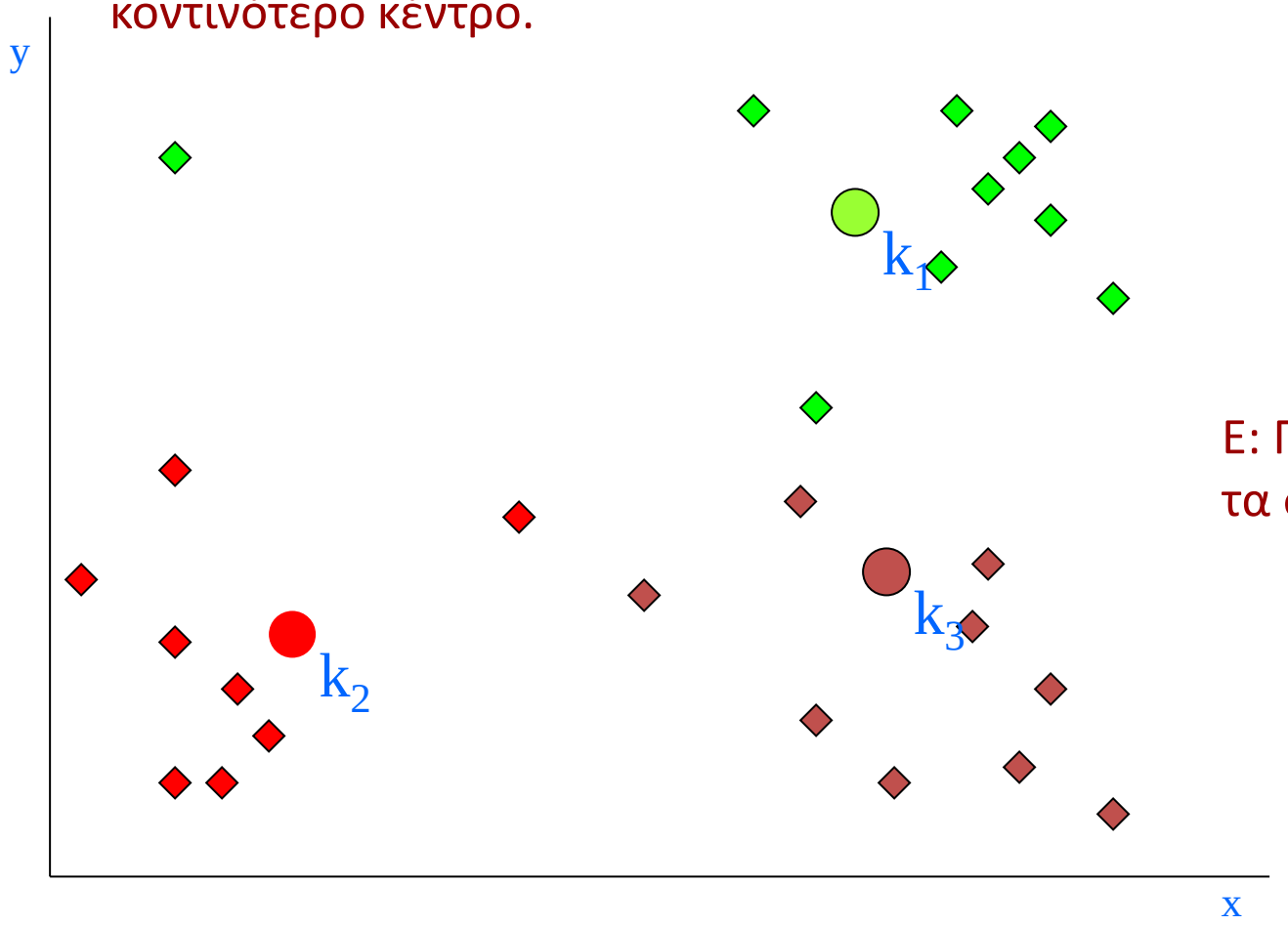
Παράδειγμα K-means: Βήμα 3

- Μετακίνησε κάθε κέντρο στο μέσο της συστάδας.



Παράδειγμα K-means: Βήμα 4

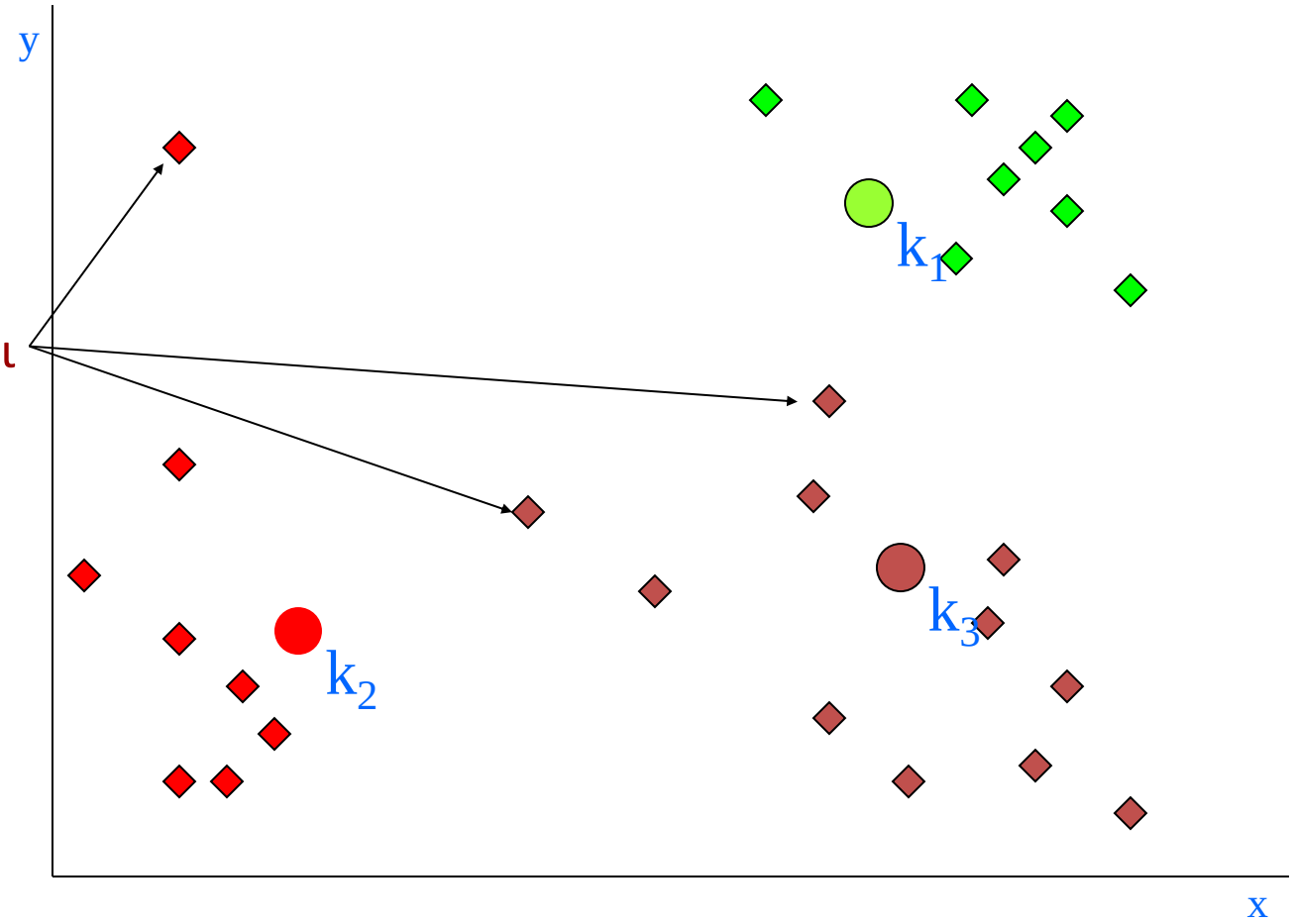
- Αναθεώρησε την ανάθεση των σημείων για τα οποία υπάρχει κοντινότερο κέντρο.



Ε: Ποια είναι αυτά τα σημεία;

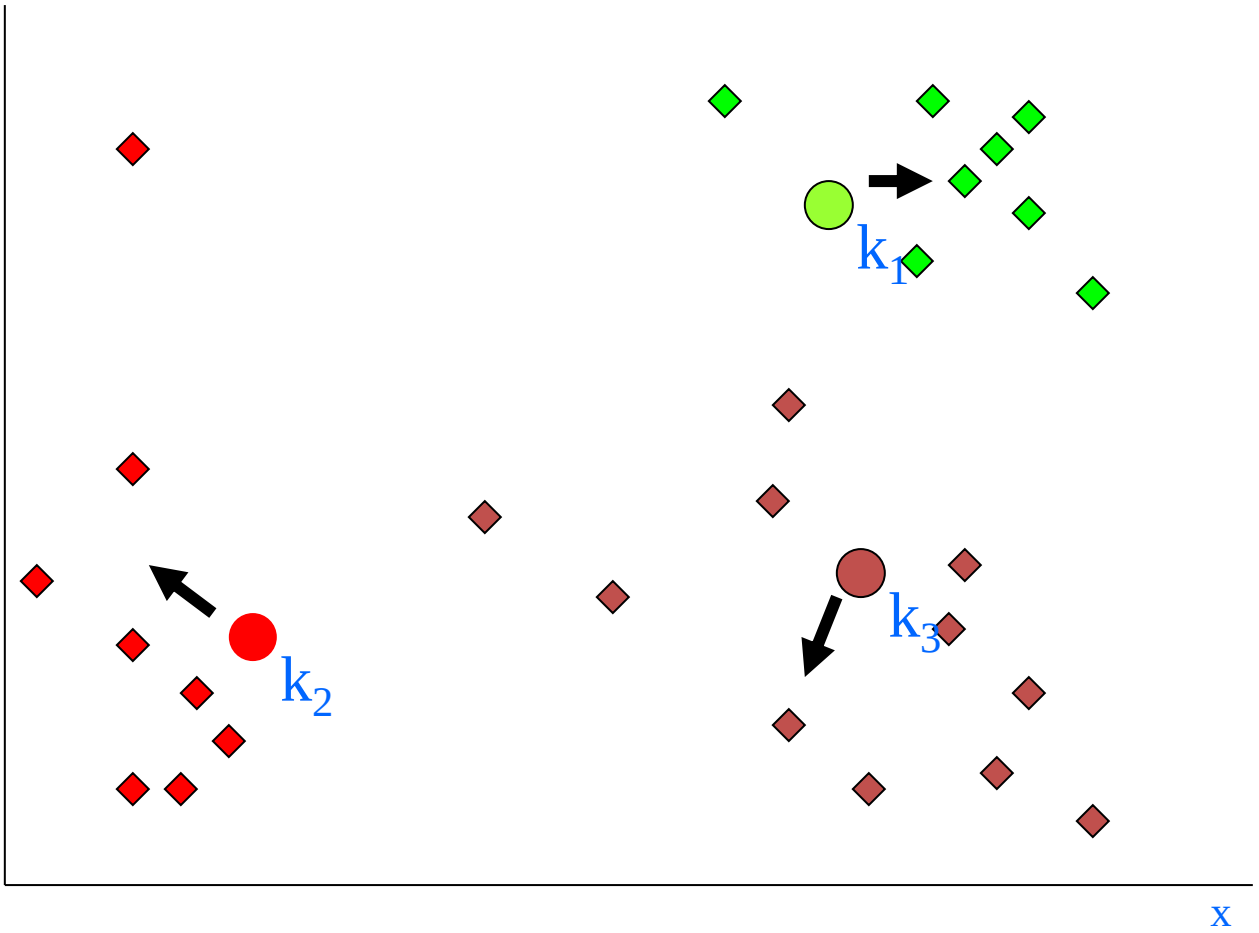
Παράδειγμα K-means: Βήμα 4α …

3 σημεία ανατίθενται διαφορετικά



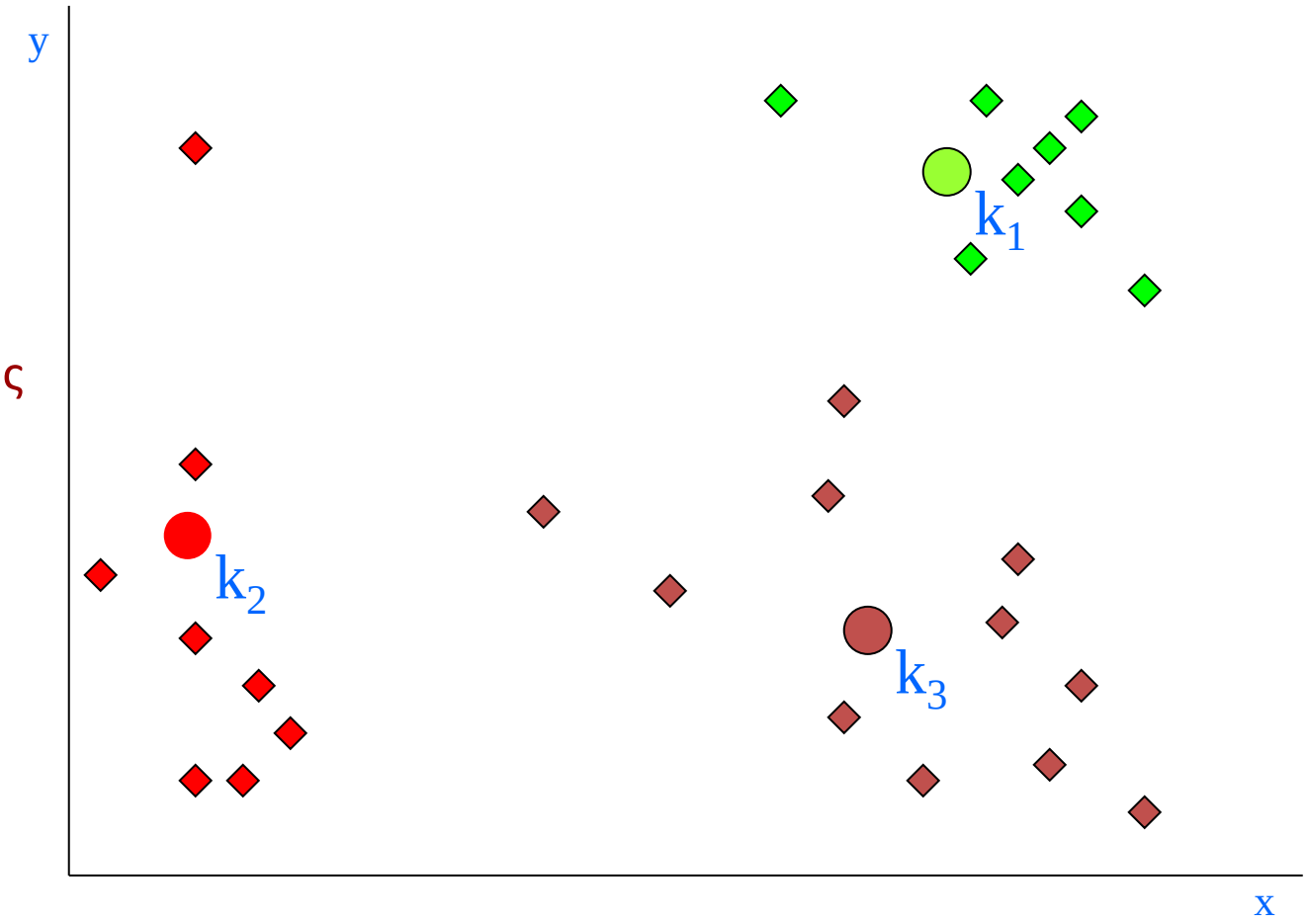
Παράδειγμα K-means: Βήμα 4β

Επανα-υπολογισμός μέσων

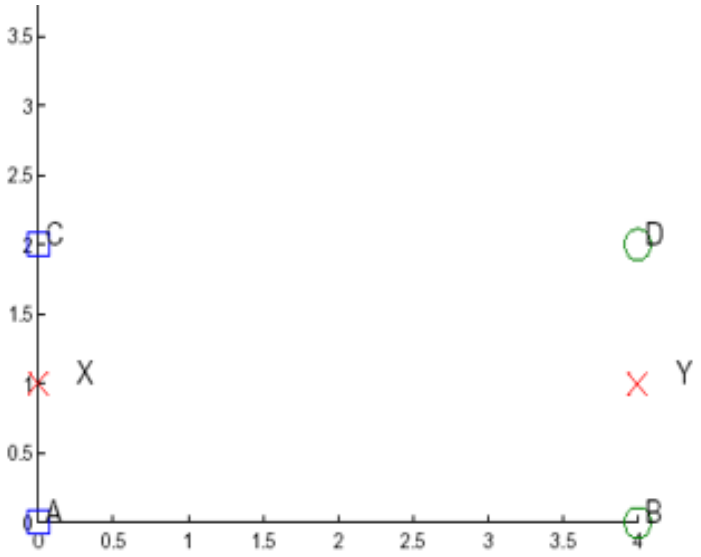


Παράδειγμα K-means: Βήμα 5

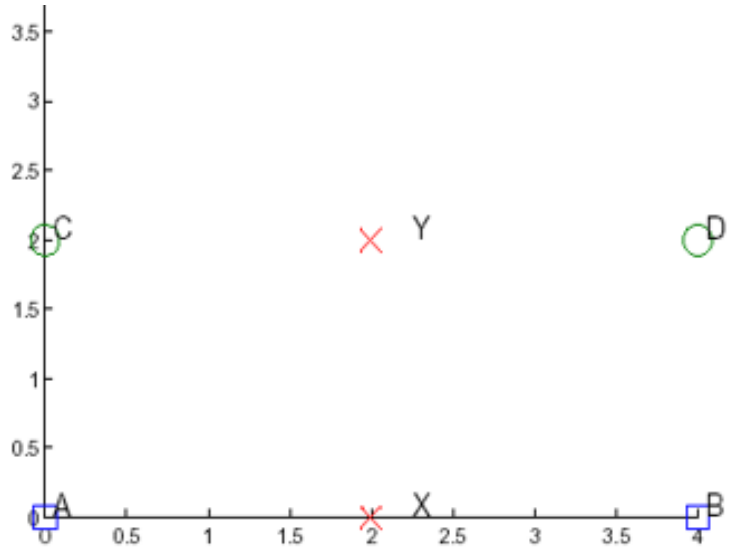
Επανα-υπολογισμός μέσων



Σύγκλιση σε τοπικά ελάχιστα



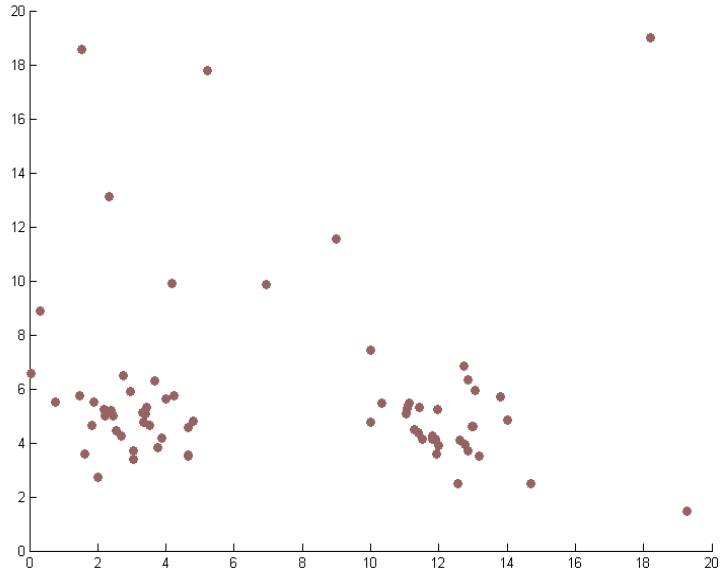
SSE = 4

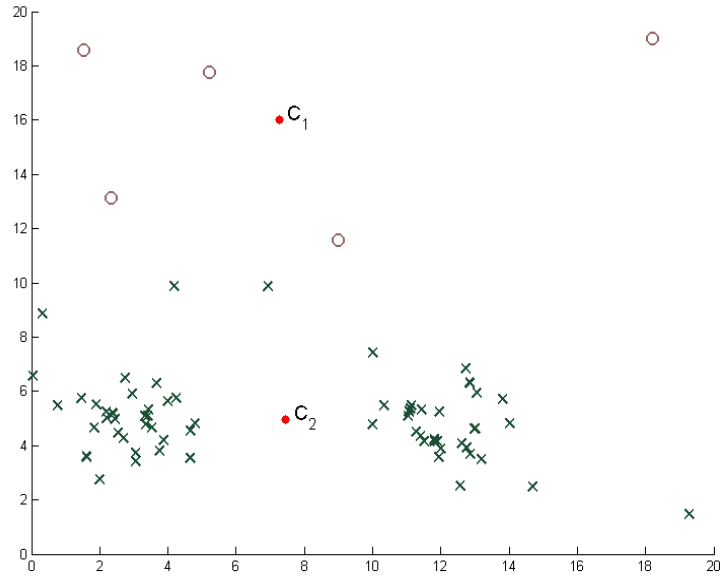


SSE = 16

Επανάληψη με άλλα αρχικά κέντρα

Ευαισθησία σε θόρυβο/outliers





Εναλλακτικές συνθήκες τερματισμού

- Μετά από συγκεκριμένο αριθμό επαναλήψεων.
- Το άθροισμα των μεταβολών των κέντρων είναι κάτω από ένα κατώφλι.

Πρόβλημα αρχικής επιλογής κέντρων

- Αν υπάρχουν K συστάδες πραγματικά με n σημεία η κάθε μία, η πιθανότητα επιλογής ενός κέντρου σε κάθε μία συστάδα είναι μικρή:

  - Για K = 10, τότε P = $10!/10^{10}$ = 0.00036.
  - Κάποιες φορές τα κέντρα θα αυτορυθμιστούν σωστά, και κάποιες φορές όχι.

Λύσεις

- Πολλαπλές επαναλήψεις.
- Δειγματοληψία και εφαρμογή ιεραρχικής ομαδοποίησης.
- Επιλογή περισσότερων αρχικών σημείων και κατόπιν επιλογής των πιο απομακρυσμένων.
- Εκ των υστέρων επεξεργασία.
- Διχοτομικός (Bisecting) K-means.

Χαρακτηριστικά K-means

Πλεονεκτήματα

- Απλός.
- Γρήγορος (O(nd)).

Μειονεκτήματα

- Επιλογή k.
- Τοπικά ελάχιστα.
- Ευαισθησία σε θόρυβο, outliers.
- Χαμηλή αποτελεσματικότητα όταν οι πραγματικές συστάδες είναι διαφορετικού μεγέθους, πυκνότητας, ή δεν έχουν σφαιρικό σχήμα.
  - Χρειάζονται αυξημένο k.

Παραλλαγές K-means

- **K-medians** – αντί για μέσες τιμές, χρησιμοποιούνται οι ενδιάμεσες.
  - Mean των 1, 3, 5, 7, 9: **5**
  - Mean των 1, 3, 5, 7, 1009: **205**
  - Median των 1, 3, 5, 7, 1009: **5**
  - Προτέρημα των Median τιμών: δεν επηρεάζονται από ακραίες τιμές.
- Δειγματοληψία (για μεγάλες ΒΔ).

Σημείωμα Αναφοράς

Διαθέσιμο από τη δικτυακή διεύθυνση:http://eclass.auth.gr/courses/OCRS182/

Σημείωμα Αδειοδότησης

ΑΡΙΣΤΟΤΕΛΕΙΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΘΕΣΣΑΛΟΝΙΚΗΣ

ΑΝΟΙΚΤΑ
ΑΚΑΔΗΜΑΪΚΑ
ΜΑΘΗΜΑΤΑ

Τέλος ενότητας

Επεξεργασία: Ανδρέας Κοσματόπουλος

Θεσσαλονίκη, Χειμερινό Εξάμηνο 2013-2014

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ ΚΑΙ ΘΡΗΣΚΕΥΜΑΤΩΝ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ
ΕΣΠΑ 2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Σημειώματα

Διατήρηση Σημειωμάτων

Οποιαδήποτε αναπαραγωγή ή διασκευή του υλικού θα πρέπει να συμπεριλαμβάνει:

- το Σημείωμα Αναφοράς
- το Σημείωμα Αδειοδότησης
- τη δήλωση Διατήρησης Σημειωμάτων
- το Σημείωμα Χρήσης Έργων Τρίτων (εφόσον υπάρχει)

μαζί με τους συνοδευόμενους υπερσυνδέσμους.