Αναγνώριση Προτύπων

Ομαδοποίηση - Βασικές Έννοιες

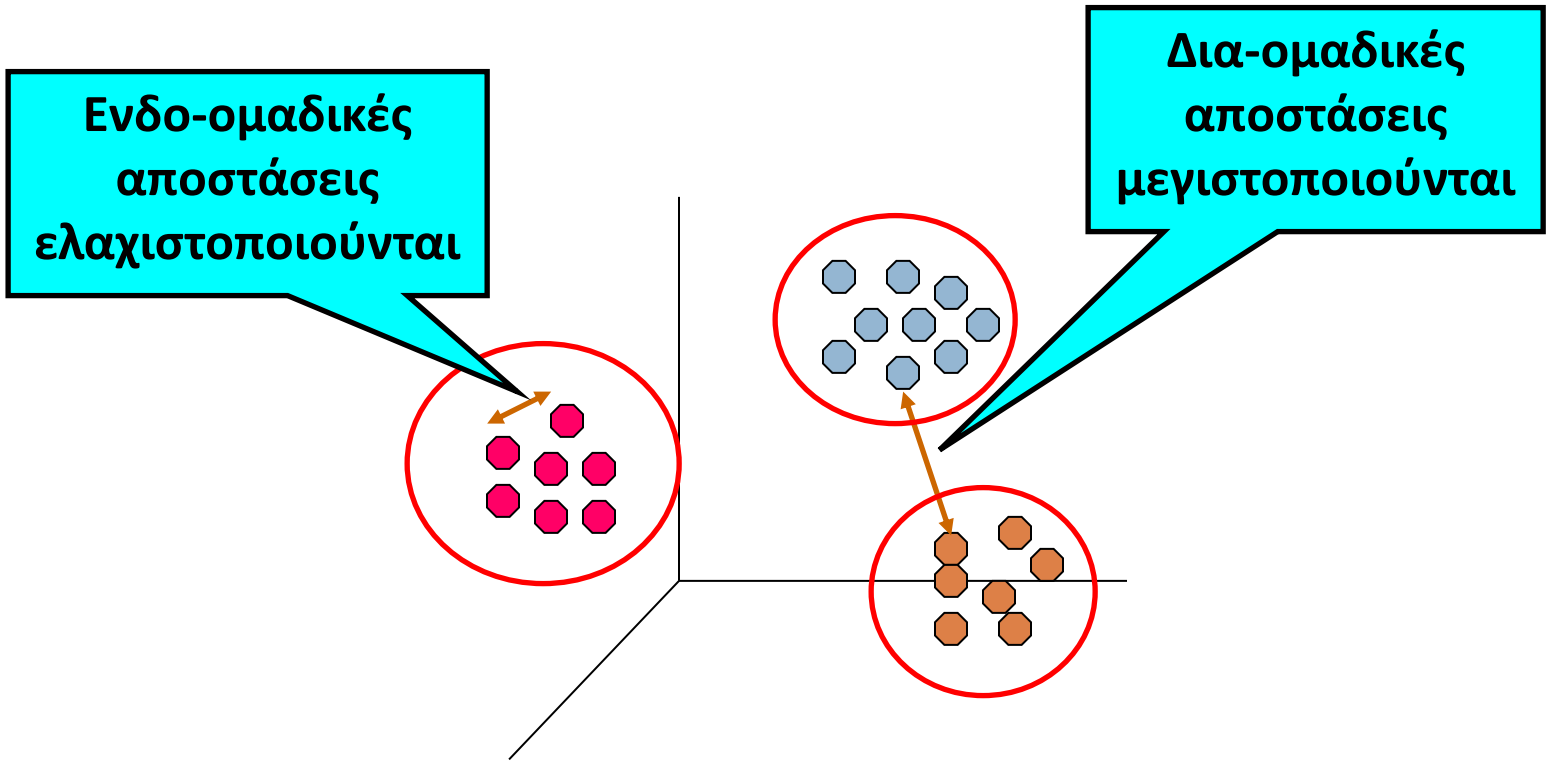
Ανδρέας Λ. Συμεωνίδης
Αν. Καθηγητής
Τμήμα Ηλεκτρολόγων Μηχ/κών
&
Μηχ/κών Υπολογιστών, Α.Π.Θ.
Email: asymeon@eng.auth.gr

Διάρθρωση διάλεξης

- Ομαδοποίηση: ορισμός και είδη
- Ομαδοποίηση διαχωρισμού
- Ο K-means και οι παραλλαγές του
- K-medoids
- Αξιολόγηση ομαδοποίησης
- Μετρικές αξιολόγησης

Τί είναι Ομαδοποίηση;

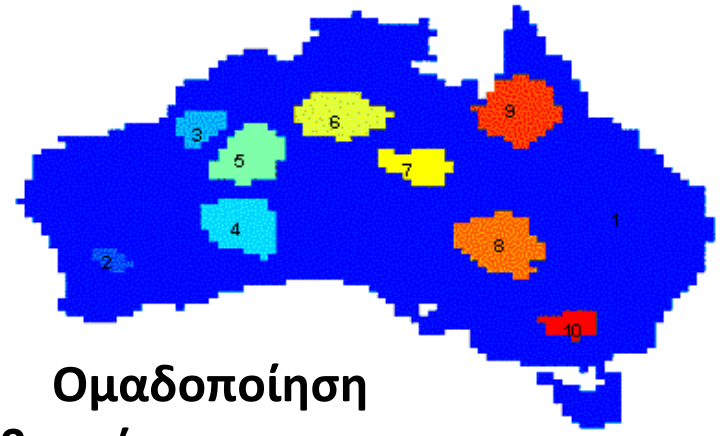
- Η εύρεση ομάδων αντικειμένων ώστε τα αντικείμενα σε μια ομάδα να είναι όμοια (ή να σχετίζονται) μεταξύ τους και ανόμοια (ή ασυσχέτιστα) απ'ο αντικείμενα σε άλλες ομάδες



Εφαρμογές της Ομαδοποίησης

- **Κατανόηση**
  - Ομαδοποίηση συναφών εγγράφων, ομαδοποίηση γονιδιωμάτων και πρωτεϊνών με παρόμοια λειτουργικότητα, ομαδοποίηση μετοχών που έχει παρόμοια διακύμανση στις τιμές τους
- **Σύνοψη**
  - Μείωση του μεγέθους μεγάλων σετ δεδομένων

	Discovered Clusters	*Industry Group*
1	Applied-Matl-DOWN,Bay-Network-Down,3-COM-DOWN, Cabletron-Sys-DOWN,CISCO-DOWN,HP-DOWN, DSC-Comm-DOWN,INTEL-DOWN,LSI-Logic-DOWN, Micron-Tech-DOWN,Texas-Inst-Down,Tellabs-Inc-Down, Natl-Semiconduct-DOWN,Oracl-DOWN,SGI-DOWN, Sun-DOWN	Technology1-DOWN
2	Apple-Comp-DOWN,Autodesk-DOWN,DEC-DOWN, ADV-Micro-Device-DOWN,Andrew-Corp-DOWN, Computer-Assoc-DOWN,Circuit-City-DOWN, Compaq-DOWN, EMC-Corp-DOWN, Gen-Inst-DOWN, Motorola-DOWN,Microsoft-DOWN,Scientific-Atl-DOWN	Technology2-DOWN
3	Fannie-Mae-DOWN,Fed-Home-Loan-DOWN, MBNA-Corp-DOWN,Morgan-Stanley-DOWN	Financial-DOWN
4	Baker-Hughes-UP,Dresser-Inds-UP,Halliburton-HLD-UP, Louisiana-Land-UP,Phillips-Petro-UP,Unocal-UP, Schlumberger-UP	Oil-UP



Ομαδοποίηση βροχόπτωσης στην Αυστραλία

Τί δεν είναι Ομαδοποίηση;

- Η επιβλεπόμενη ταξινόμηση
  - Η υπάρξη πληροφοριών για το χαρακτηριστικό-κλάση
- Η απλή κατάτμηση
  - Ο διαχωρισμός των φοιτητών σε ομάδες με βάση το επίθετο
- Τα αποτελέσματα ενός ερωτήματος
  - Η ομαδοποίηση είναι αποτέλεμσα μιας εξωτερικής προδιαγραφής
- Ο διαχωρισμός γράφων
  - Υπάρχουν κοινά χαρακτηριστικά, αλλά δεν είναι το ίδιο.

Αμφισημία της έννοιας Ομάδα...

Πόσες ομάδες;

Έξι ομάδες

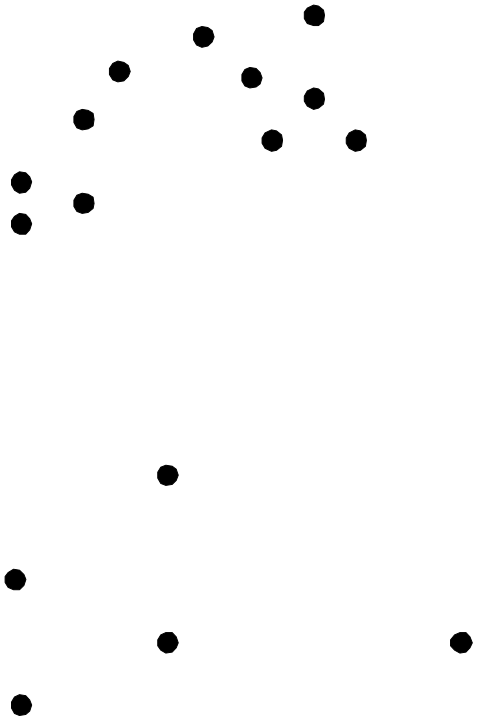
Δυο ομάδες

Τέσσερις ομάδες

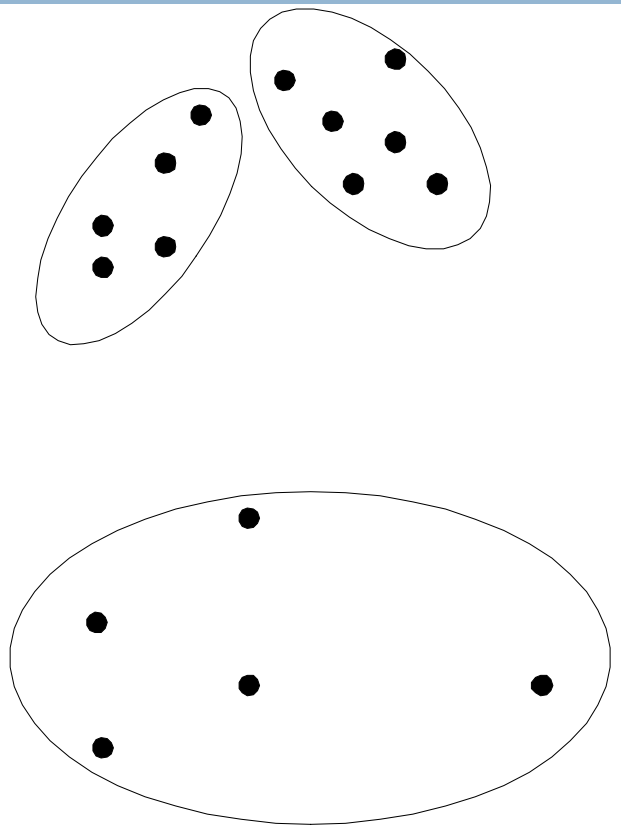
Τύποι Ομαδοποίησης

- Η ομαδοποίηση είναι ένα σετ από ομάδες
- Βασικός διαχωρισμός ανάμεσα σε ιεραρχικά σετ από ομάδες και σε σετ διαχωρισμού
- Ομαδοποίηση Διαχωρισμού
  - Ο διαχωρισμός των αντικειμένων (data objects) σε μη-επικαλυπτόμενα υποσέτ (ομάδες), ώστε κάθε αντικείμενο να ανήκει μόνο σε ένα υποσέτ.
- Ιεραρχική ομαδοποίηση
  - Ένα σετ από εμφωλευμένες ομάδες, οργανωμένες σε ένα ιεραρχικό δένδρο.

Ομαδοποίηση διαχωρισμού

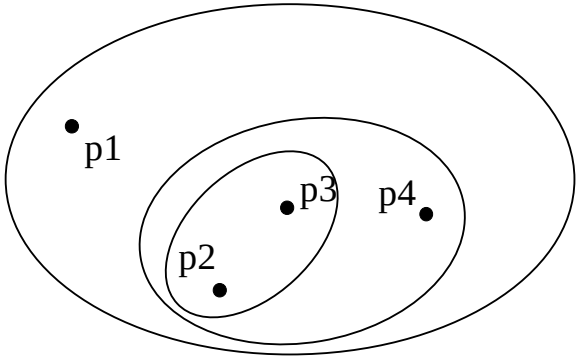


Αρχικά αντικείμενα

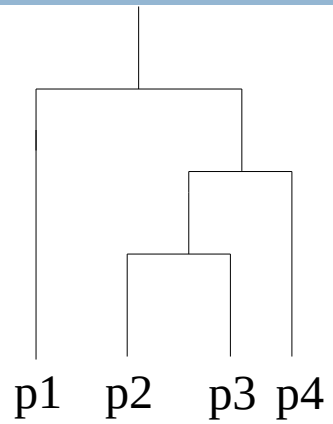


Ομαδοποίηση διαχωρισμού

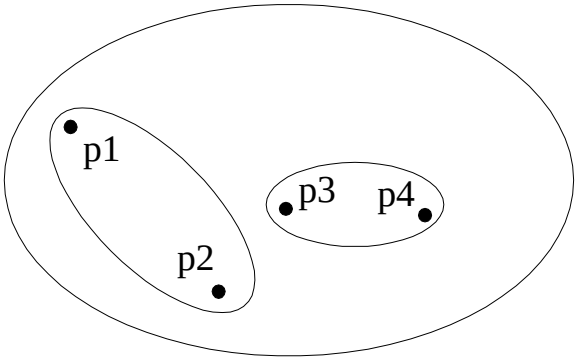
Ιεραρχική ομαδοποίηση



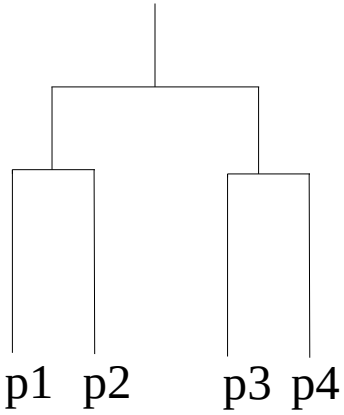
Κλασική ιεραρχική ομαδοποίηση



Κλασικό δενδρόγραμμα



Υβριδική ιεραρχική ομαδοποίηση



Υβριδικό δενδρόγραμμα

Άλλες κατηγοριοποιήσεις ομαδοποίησης

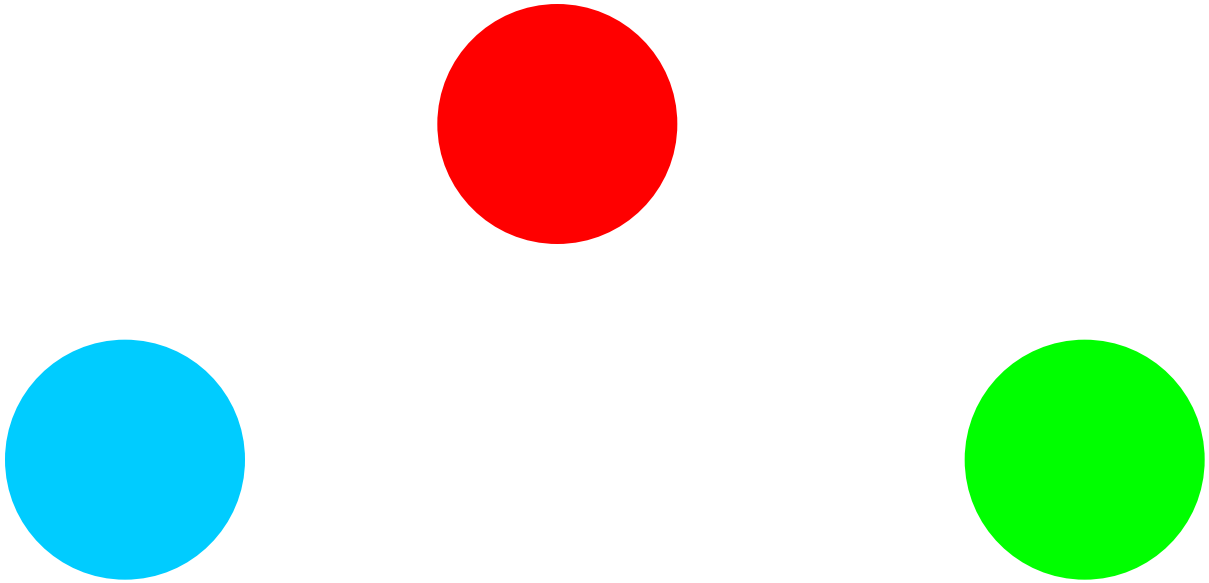
- Αποκλειστική versus μη αποκλειστική
  - Σε μη-αποκλειστικές ομαδοποιήσεις, τα σημεία μπορεί να ανήκουν σε πολλά σημεία.
  - Μπορεί να αναπαριστούν πολλαπλές κλάσεις ή σημεία ‘σύνορα’ (‘border’ points)
- Ασαφής versus μη-ασαφής
  - Στην ασαφή ομαδοποίηση, ένα σημείο ανήκει σε κάθε ομάδα με ένα βάρος ανάμεσα στο 0 και το 1
  - Τα βάρη συναθροίζονται στο 1
  - Η πιθανοτική ομαδοποίηση έχει παρόμοια χαρακτηριστικά
- Μερική versus γενική
  - Σε ορισμένες περιπτώσεις, θέλουμε/μπορούμε να ομαδοποιήσουμε μέρος του σετ δεδομένων
- Ετερογενής versus ομογενής
  - Ομάδες διαφορετικών μεγεθών, σχημάτων και πυκνοτήτων

Τύποι Ομάδων

- Καλά διαχωρισμένες ομάδες
- Κεντρικές ομάδες
- Συνεχόμενες ομάδες
- Πυκνωτικές ομάδες
- Ομάδες κοινής ιδιότητας ή Εννοιολογικές ομάδες
- Ομάδες που περιγράφονται από κοινή συνάνρτηση

Καλά διαχωρισμένες ομάδες

- Η ομάδα είναι ένα σύνολο από σημεία για τα οποία κάθε σημείο της ομάδας είναι πιο κοντά (ή πιο όμοιο) από κάθε άλλο σημείο που δε βρίσκεται στην ομάδα.



3 καλά διαχωρισμένες ομάδες

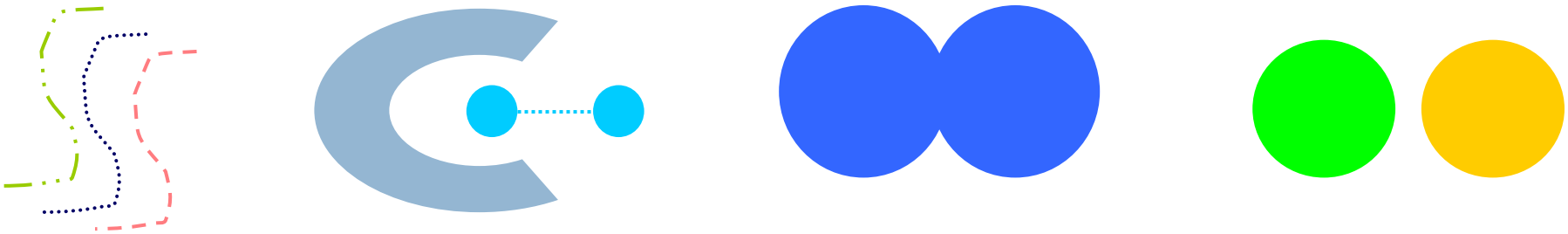
Κεντρικές ομάδες

- Η ομάδα είναι ένα σύνολο από σημεία για τα οποία κάθε σημείο της ομάδας είναι πιο κοντά (ή πιο όμοιο) στο κέντρο της ομάδας, παρά στο κέντρο οποιασδήποτε άλλης ομάδας.
- Το κέντρο της ομάδας είναι συχνά το centroid, ο μέσος όρος των σημείων της ομάδας, ή το medoid, το πιο “αντιπροσωπευτικό” σημείο της ομάδας.

4 κεντρικές ομάδες

Συνεχόμενες ομάδες

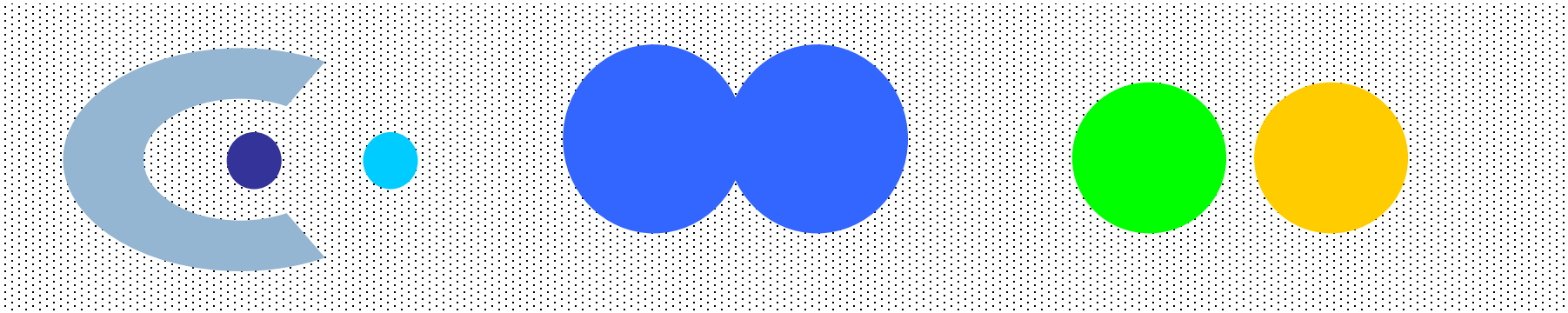
- Η ομάδα είναι ένα σύνολο από σημεία για τα οποία κάθε σημείο της ομάδας είναι πιο κοντά (ή πιο όμοιο) σε ένα ή περισσότερα της ομάδας στο κέντρο της ομάδας, παρά οποιοδήποτε σημείο άλλης ομάδας.
- Ονομάζονται και μεταβατικές (transitive), ή κοντινότερου γείτονα (nearest neighbor).



8 συνεχόμενες ομάδες

Πυκνωτικές ομάδες

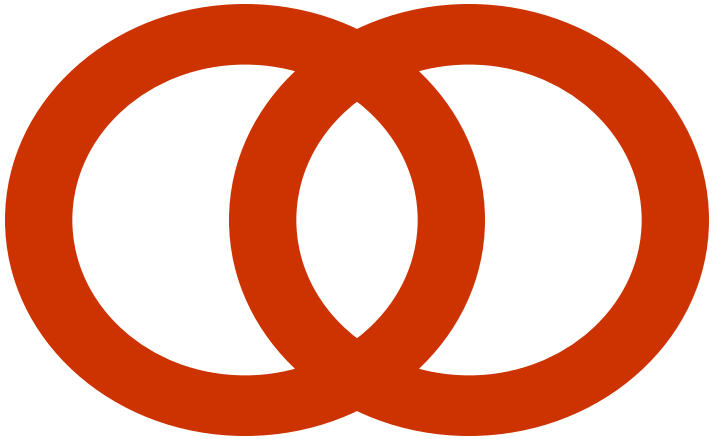
- Μια ομάδα είναι μια πυκνή περιοχή από σημεία, η οποία διαχωρίζεται από τις άλλες ομάδες από αραιές περιοχές από σημεία.
- Χρησιμοποιούνται όταν οι ομάδες είναι μη κανονικές είναι πεπλεγμένες, και όταν υπάρχει θόρυβος και εξωκείμενες τιμές.



6 πυκνωτικές ομάδες

Εννοιολογικές ομάδες

- Ομάδες οι οποίες μοιράζονται κάποιο κοινό χαρακτηριστικό ή αναπαριστούν την ίδια έννοια.



2 εννοιολογικές ομάδες

Ομάδες με κοινή συνάρτηση

- Ομάδες οι οποίες ελαχιστούν ή μεγοστοποιούν μια συνάρτηση.
- Απαριθμούν όλες τις δυνατές περιπτώσεις διαχωρισμού των σημείων σε ομάδες και υπολογίζουν την **‘αξία’** κάθε δυνητικού σετ από ομάδες, με βάση τη συνάρτηση.
- Μπορεί να υπάρχει ολικά ή και τοπικά μέγιστα/ελάχιστα
  - Οι ιεραρχικοί αλγόριθμοι έχουν συνήθως τοπικά μέγιστα/ελάχιστα.
  - Οι αλγόριθμοι διαχωρισμού έχουν συνήθως ολικά μέγιστα/ελάχιστα.

Ομάδες με κοινή συνάρτηση (συν.)

- Αντιστοίχιση του προβλήματος ομαδοποίησης σε ένα άλλο πεδίο εφαρμογής και επίλυση του προβλήματος στο πεδίο αυτό
  - Πίνακας γειτνίασης ο οποίος ορίζει ένα γράφο με βάρη, όπου οι κόμβοι είναι τα σημεία που ομαδοποιούνται και οι ακμές με βάρη την απόσταση ανάμεσα στα σημεία
  - Η ομαδοποίηση είναι ισοδύναμη με το διαχωρισμό του γράφου σε συνδεδεμένα τμήματα, ένα για κάθε ομάδα.
  - Στόχος είναι η ελαχιστοποίηση του βάρων των ακμών μέσα στις ομάδες και η μεγιστοποίηση του βάρων των ακμών ανάμεσα στις ομάδες

Σημασία των χαρακτηριστικών των δεδομένων εισόδου

- Ο τύπος της μετρικής γειτνίασης ή πυκνότητας
  - Ορίζεται κατά περίπτωση, αλλά είναι ιδιαίτερα σημαντικό στην ομαδοποίηση
- Αραιά δεδομένα
  - Υπαγορεύει τον τύπο ομοιότητας
  - Προσθέτει στην αποδοτικότητα του αλγορίθμου
- Τύπος μεταβλητών του σετ εισόδου
  - Υπαγορεύει τον τύπο ομοιότητας
- Διαστάσεις του σετ εισόδου
- Θόρυβος και εξωκείμενες τιμές
- Τύπος κατανομής

Ομάδες αλγορίθμων ομαδοποίησης

- Ο K-means και οι παραλλαγές του
- Ιεραρχικοί αλγόριθμοι
- Πυκνωτικοί αλγόριθμοι

K-means Ομαδοποίηση

- Αλγόριθμος διαχωρισμού
- Κάθε ομάδα συνδέεται με ένα centroid (κέντρο)
- Κάθε σημείο αποδίδεται στην ομάδα με το πιο κοντινό κέντρο
- Ο αριθμός των ομάδων, K, **πρέπει να καθοριστεί**
- Πολύ απλός αλγόριθμος

1: Select K points as the initial centroids.
2: **repeat**
3: Form K clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: **until** The centroids don't change

K-means Ομαδοποίηση – Πιο αναλυτικά

- Τα αρχικά κέντρα συνήθως επιλέγονται τυχαία
  - Οι ομάδες που προκύπτουν αλλάζουν σε κάθε εκτέλεση του αλγορίθμου
- Το κέντρο είναι (τυπικά) ο μέσος όρος των σημείων της ομάδας
- Η **'Εγγύτητα'** μετράται ως η Ευκλείδεια απόσταση, η ομοιότητα συνημίτονου, η συσχέτιση, κτλ
- Ο K-means συνήθως συγκλίνει για τις παραπάνω μετρικές
- Η σύγκλιση συμβαίνει συνήθως στις πρώτες επαναλήψεις
- Η πολυπλοκότητα είναι: **O(n*K*I*d)**
  - *n* = αριθμός σημείων, *K* = αριθμός ομάδων, *I* = αριθμός επαναλήψεων, *d* = αριθμός μεταβλητών

Αξιολόγηση των ομάδων του K-means

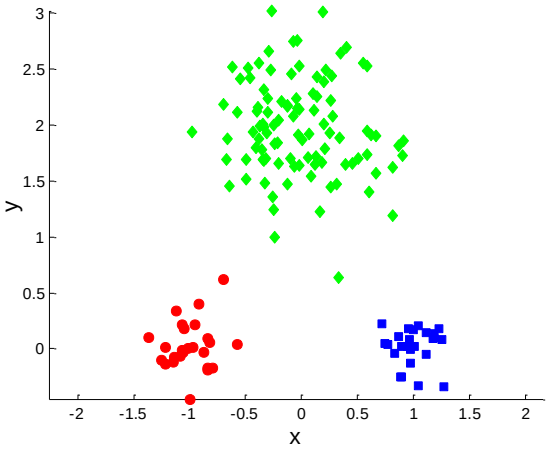
- Η πιο κλασική μετρική είναι άθροισμα του τετραγώνου σφαλμάτων (Sum of Squared Error - SSE)
  - Για κάθε σημείο, το σφάλμα είναι η απόσταση από το κέντρο της κοντινότερης ομάδας
  - Το SSE είναι το άθροισμα του τετραγώνου σφαλμάτων αυτών:

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$

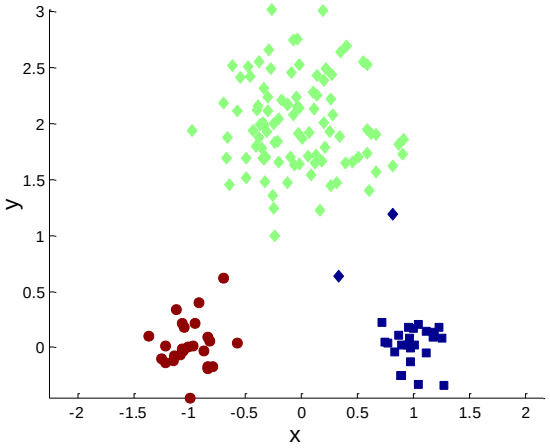
όπου x είναι ένα σημείο στην ομάδα C_i και m_i είναι το αντιπροσωπευτικό σημείο (το κέντρο) για την ομάδα C_i

  - Δεδομένων δυο ομάδων, μπορούμε να επιλέξουμε αυτήν με το μικρότερο σφάλμα
  - Ένας εύκολος τρόπος για τη μείωση του SSE είναι η αύξηση του K, του αριθμού των ομάδων
    - Μια καλή ομαδοποίηση με μικρότερο K μπορεί να έχει μικρότερο SSE από μια κακή ομαδοποίηση με μεγαλύτερο K

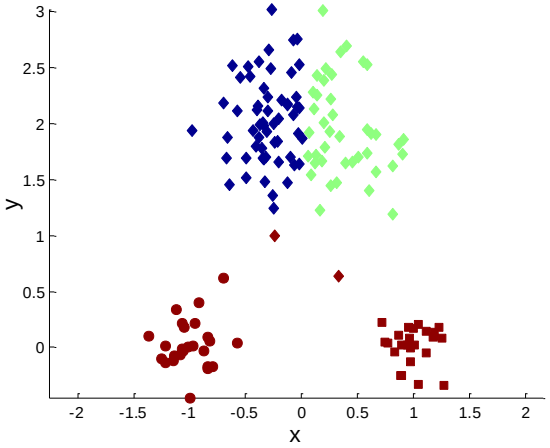
Δυο K-means ομαδοποιήσεις



Αρχικά σημεία

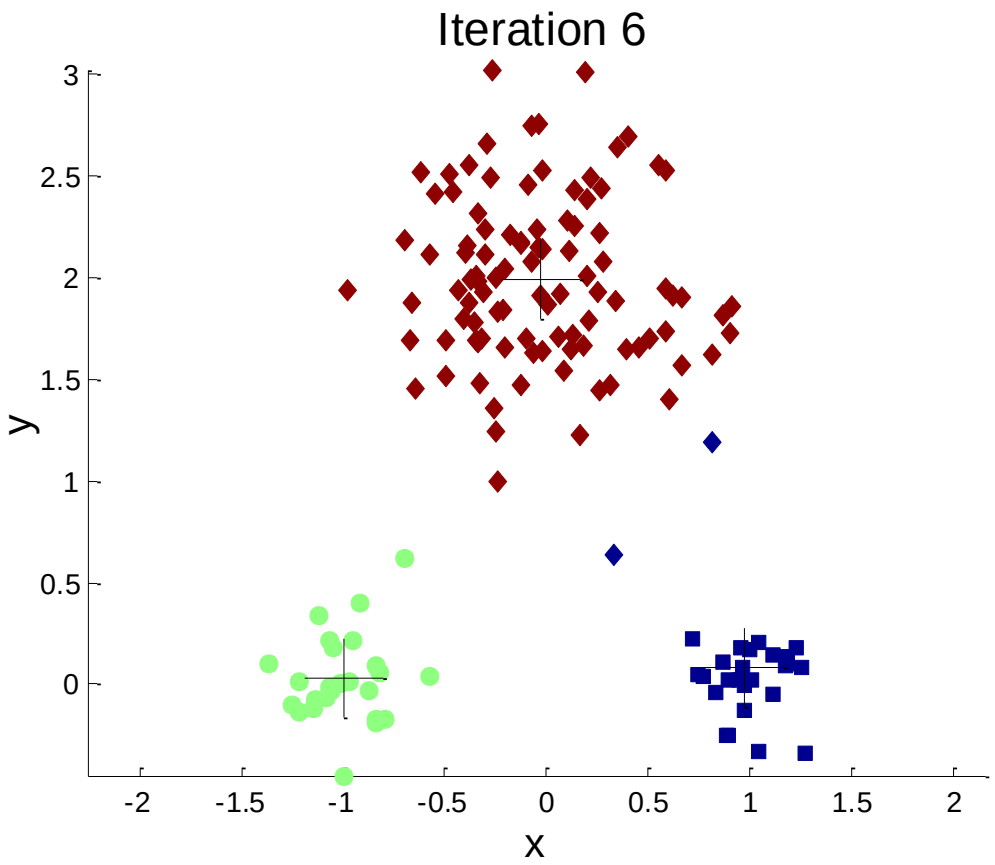


Βέλτιστη ομαδοποίηση

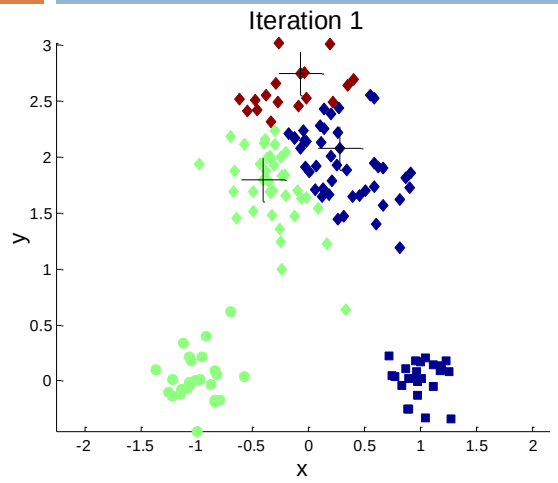


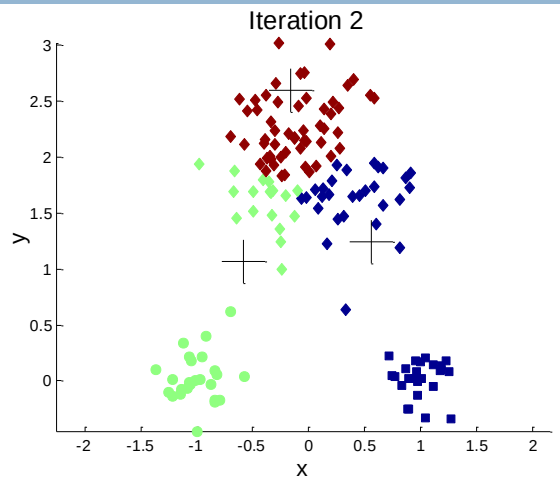
Υποβέλτιστη ομαδοποίηση

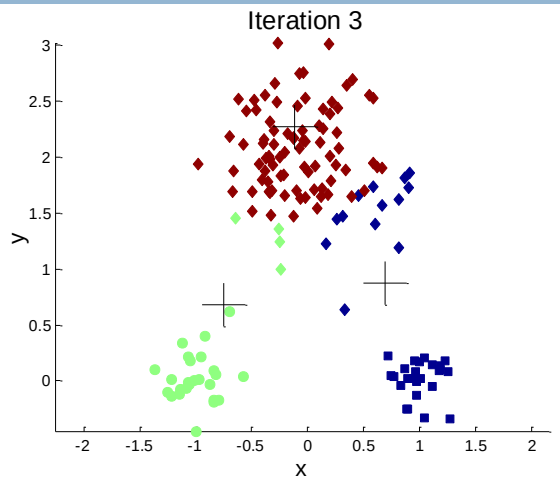
Η σημασία επιλογής των αρχικών κέντρων

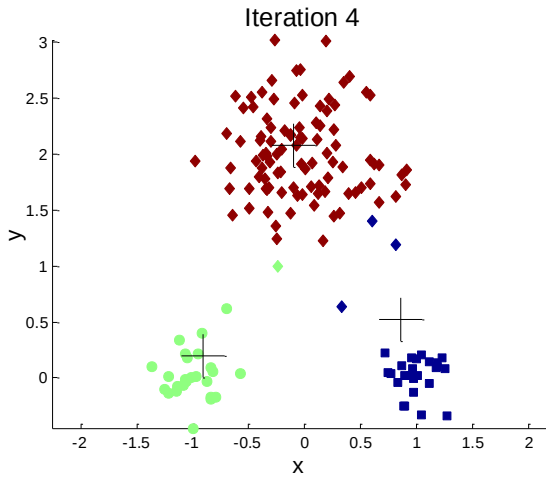


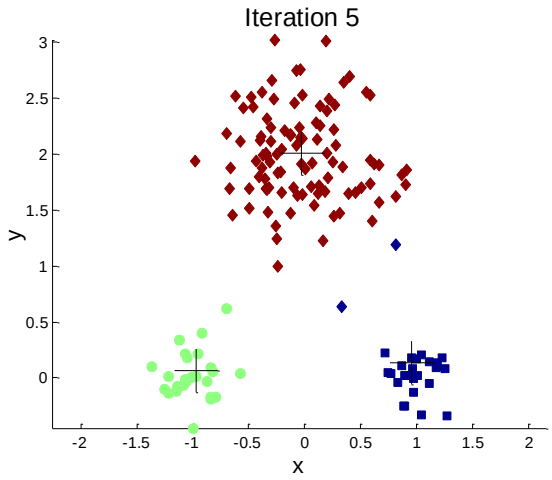
Η σημασία επιλογής των αρχικών κέντρων

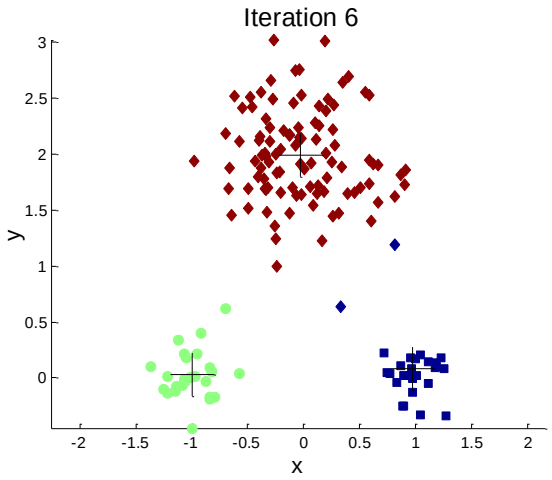




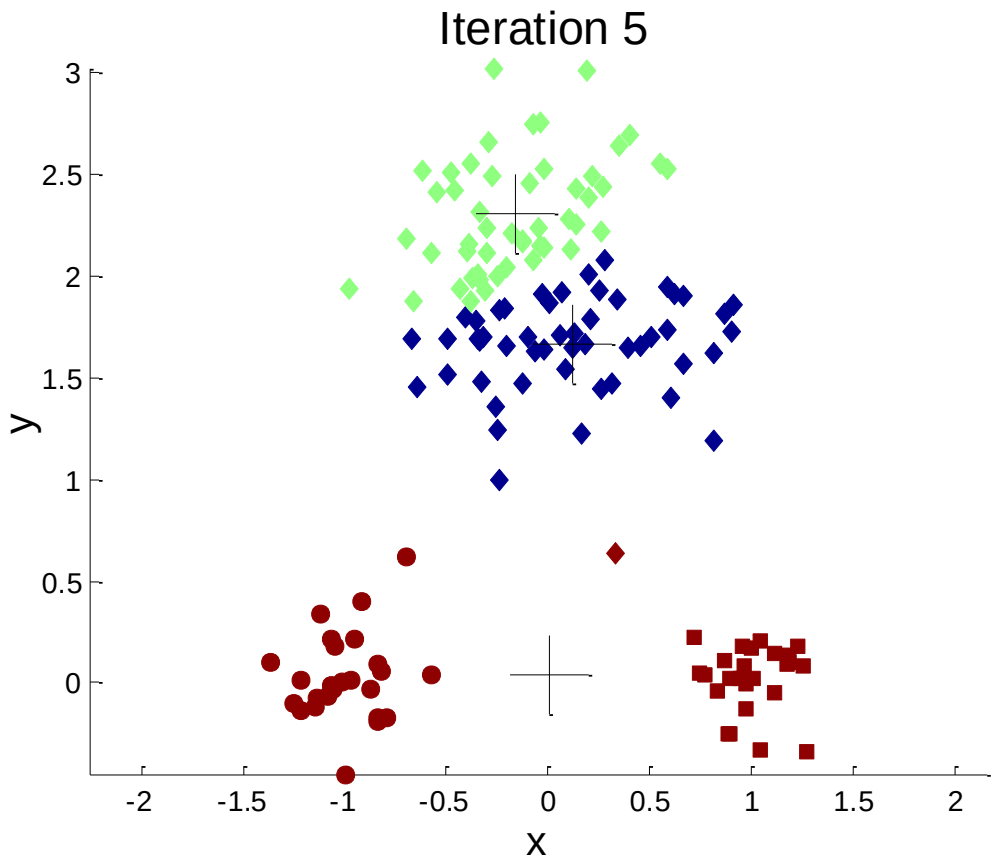




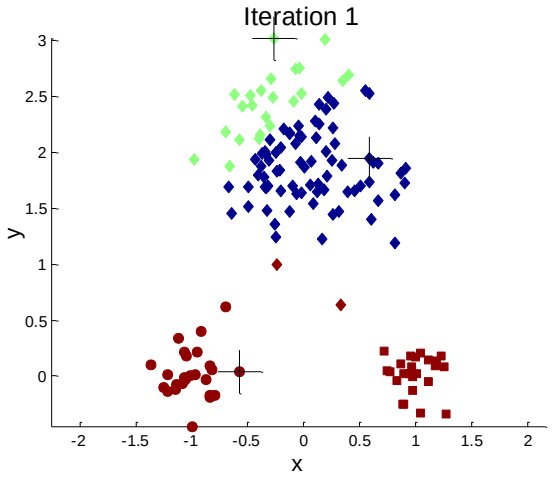


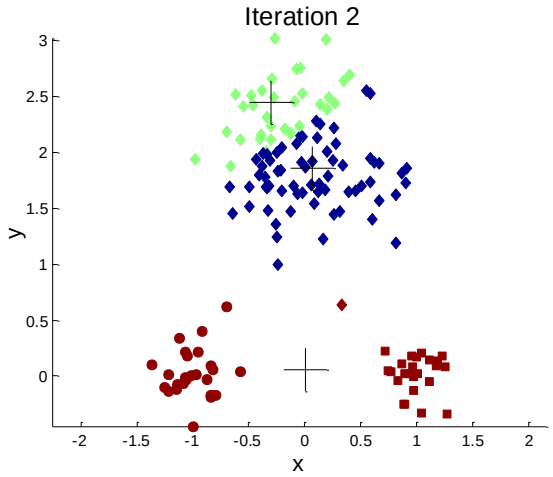


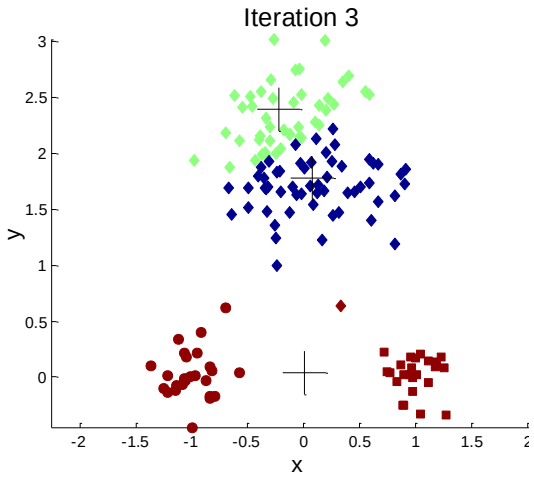
Η σημασία επιλογής των αρχικών κέντρων

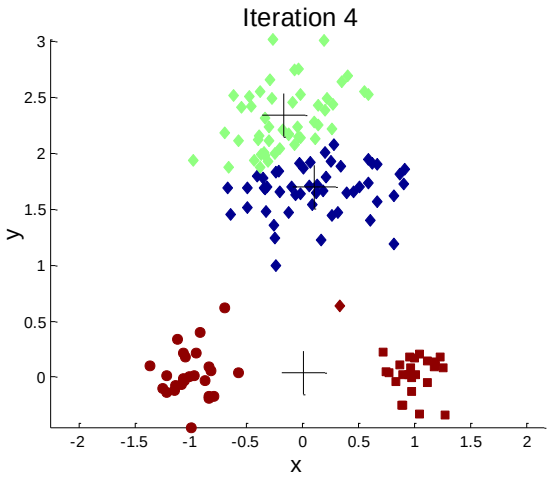


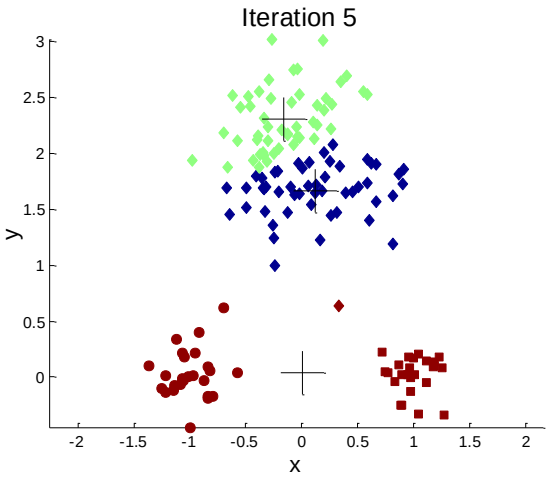
Η σημασία επιλογής των αρχικών κέντρων











Το πρόβλημα της επιλογής των αρχικών κέντρων

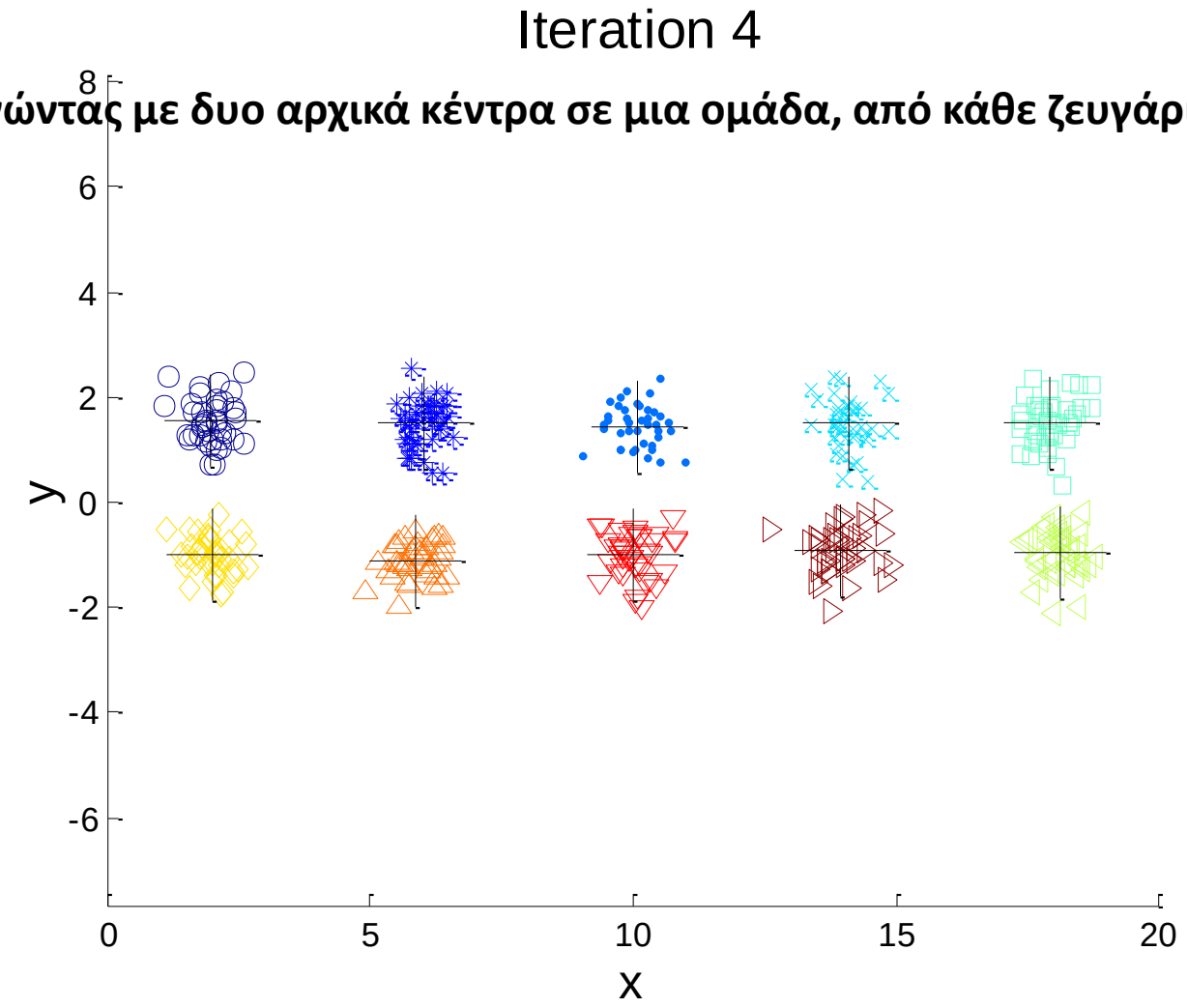
- Αν υπάρχουν K 'πραγματικές' ομάδες, τότε η πιθανότητα επιλογής ενός κέντρου από κάθε ομάδα είναι πολύ μικρή
  - Αν οι ομάδες είναι ισομεγέθεις (n σημεία), τότε

$$P = \frac{\text{number of ways to select one centroid from each cluster}}{\text{number of ways to select } K \text{ centroids}} = \frac{K!n^K}{(Kn)^K} = \frac{K!}{K^K}$$

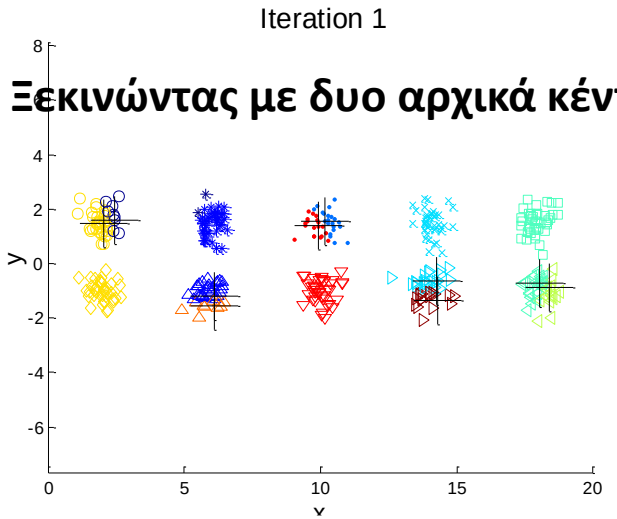
Για παράδειγμα, αν K = 10, τότε η πιθανότητα επιλογής ενός αρχικού κέντρου από κάθε πραγματική ομάδα =
= $10!/10^{10} = 0.00036$

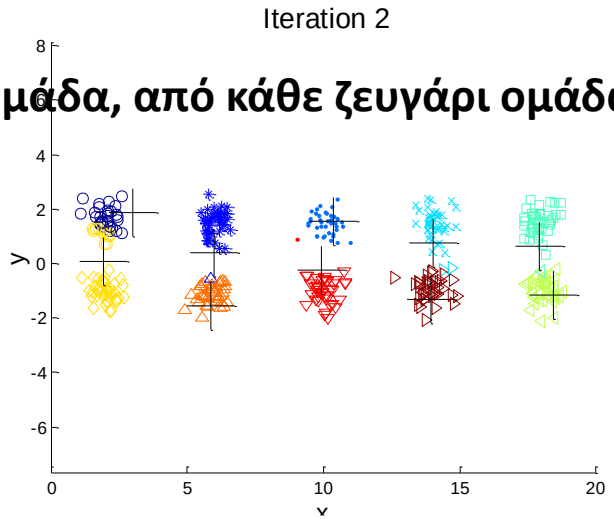
- Αρκετές φορές, τα αρχικά κέντρα θα προσαρμοστούν με τον «σωστό» τρόπο. Άλλες πάλι όχι…

Παράδειγμα: 10 ομάδες…

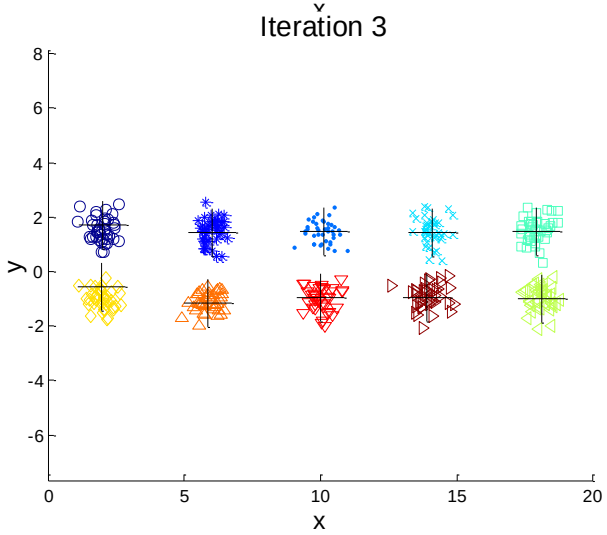


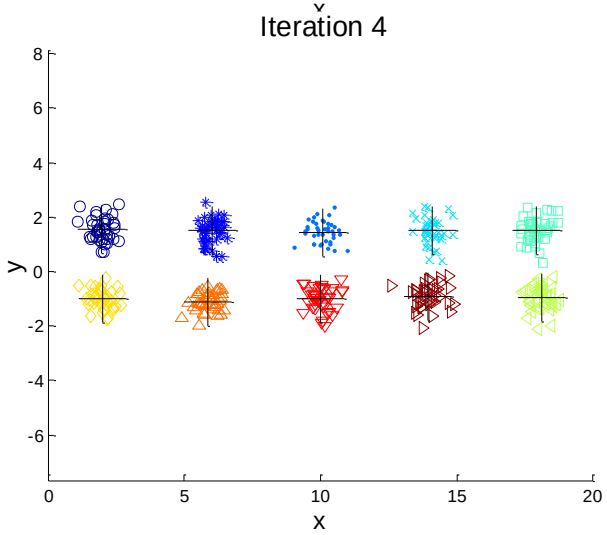
Παράδειγμα: 10 ομάδες…



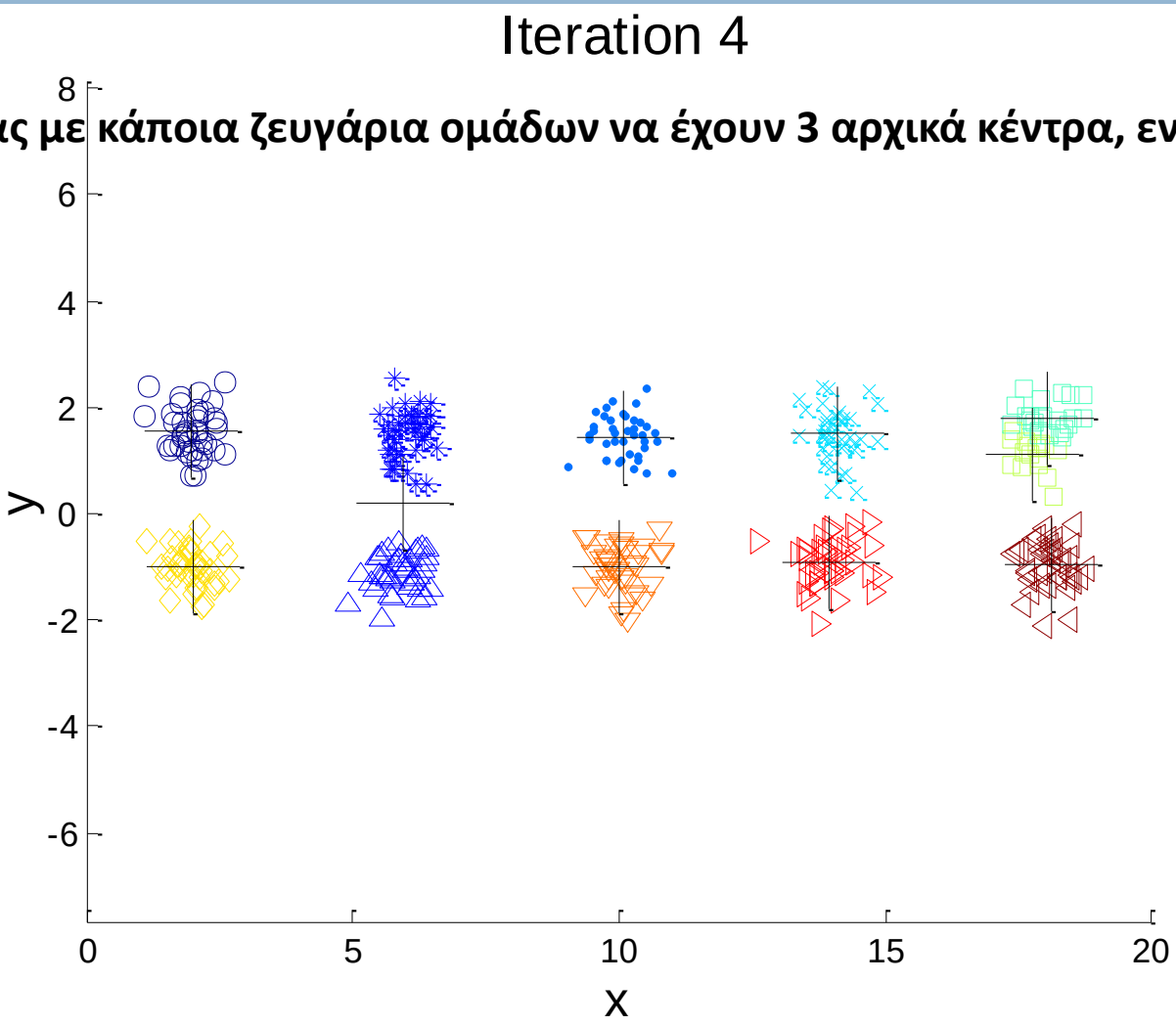


Ξεκινώντας με δυο αρχικά κέντρα σε μια ομάδα, από κάθε ζευγάρι ομάδων

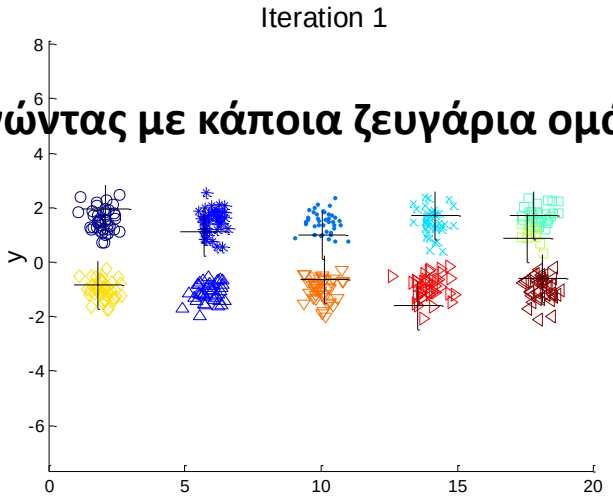


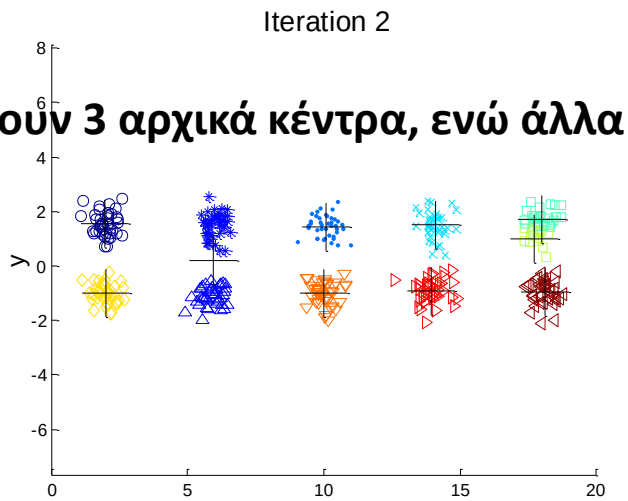


Παράδειγμα: 10 ομάδες…

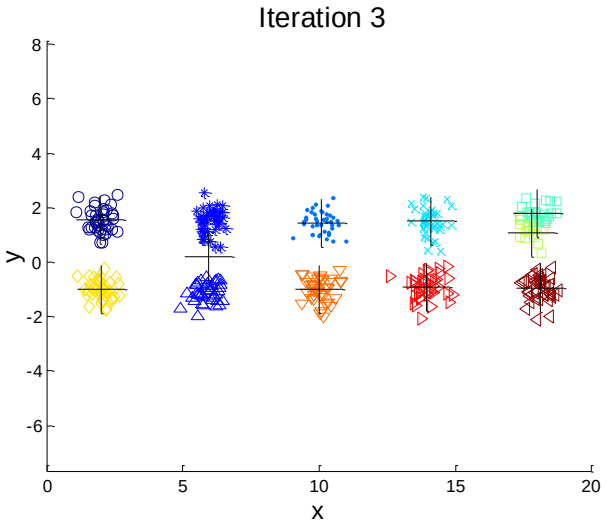


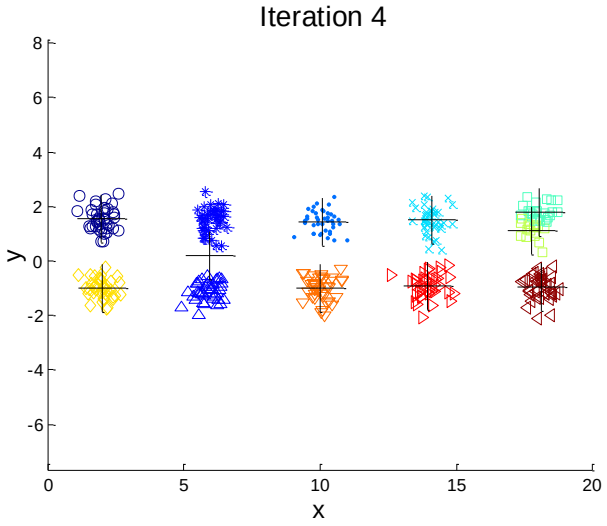
Παράδειγμα: 10 ομάδες…





Ξεκινώντας με κάποια ζευγάρια ομάδων να έχουν 3 αρχικά κέντρα, ενώ άλλα μόνο 1





Λύσεις στο πρόβλημα των αρχικών κέντρων

- Πολλές εφαρμογές του αλγορίθμου
  - Βοηθούν, αλλά οι πιθανότητες δεν είναι με το μέρος σας ☹
- Κάντε δειγματοληψία και εφαρμόστε ιεραρχική ομαδοποίηση για να καθορίσετε τα αρχικά κέντρα
- Επιλέξτε περισσότερα από K αρχικά κέντρα και στη συνέχεια διαλέξτε από αυτά
  - Επιλέξετε αυτά που είναι όσο πιο καλά διαχωρισμένα
- Μετ-επεξεργασία
- Διχοτομικός K-means
  - Δεν επηρεάζεται τόσο από θέματα αρχικοποίησης

Διαχείριση άδειων ομάδων

- Ο απλός K-means μπορεί να γεννήσει άδειες ομάδες
- Στρατηγικές επίλυσης:
  - Επιλογή ως αρχικό κέντρο του σημείου που παράγει το μεγαλύτερο SSE
  - Επιλογή ως αρχικό κέντρο ενός σημείου από την ομάδα με το μεγαλύτερο SSE
  - Αν υπάρχουν πολλές άδειες ομάδες, η διαδικασία επαναλαμβάνεται πολλές φορές

Προοδευτική ανανέωση κέντρων

- Στον βασικό K-means, τα κέντρα ανανεώνονται αφού όλα τα σημεία έχουν αντιστοιχιστεί σε ένα κέντρο
- Εναλλακτικά, μπορούν να ανανεώνονται μετά από κάθε αντιστοίχιση σημείου (προοδευτική προσέγγιση)
  - Κάθε αντιστοίχιση ανανεώνει 0 ή 2 κέντρα
  - Υπολογιστικά πιο ακριβή
  - Εισάγει μια εξάρτηση τάξης
  - Δε δημιουργεί ποτέ άδειες ομάδες
  - Μπορούν να χρησιμοποιηθούν “βάρη” για να αλλάξουν οι συσχετίσεις

Προ-επεξεργασία και μετ-επεξεργασία

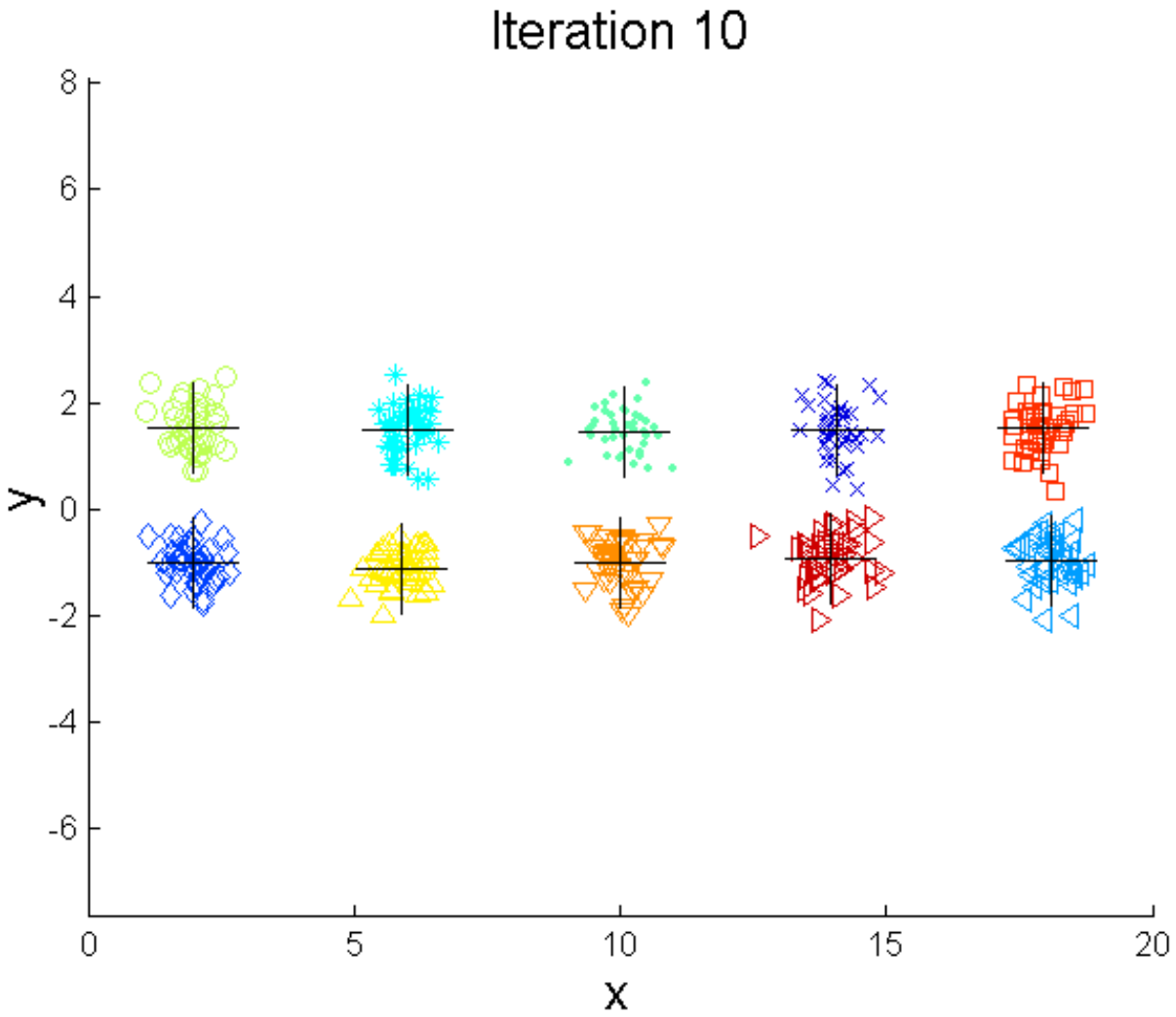
- Προ-επεξεργασία
  - Κανονικοποίηση δεδομένων
  - Απαλοιφή εξωκείμενων τιμών
- Μετ-επεξεργασία
  - Διαγραφή των μικρών ομάδων που μπορεί να αναπαριστούν εξωκείμενες τιμές
  - Διαχωρισμός των ‘χαλαρών’ ομάδων (ομάδες με σχετικά ψηλό SSE)
  - Ενοποίηση ομάδων που είναι ‘κοντά’ και έχουν σχετικά χαμηλό SSE
  - Τα βήματα αυτά μπορούν να γίνουν κατά τη διαδικασία ομαδοποίησης
    - ISODATA αλγόριθμος

Διχοτομικός K-means

- Παραλλαγή του K-means που μπορεί να παράξει ομαδοποίηση διαχωρισμού ή ιεραρχική

1: Initialize the list of clusters to contain the cluster containing all points.
2: **repeat**
3: Select a cluster from the list of clusters
4: **for** $i = 1$ to *number_of_iterations* **do**
5: Bisect the selected cluster using basic K-means
6: **end for**
7: Add the two clusters from the bisection with the lowest SSE to the list of clusters.
8: **until** Until the list of clusters contains K clusters

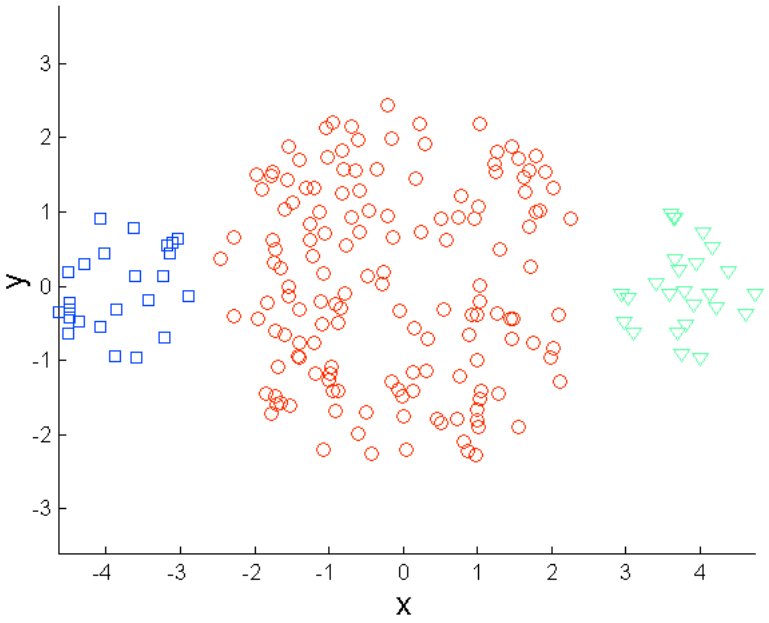
Διχοτομικός K-means: παράδειγμα



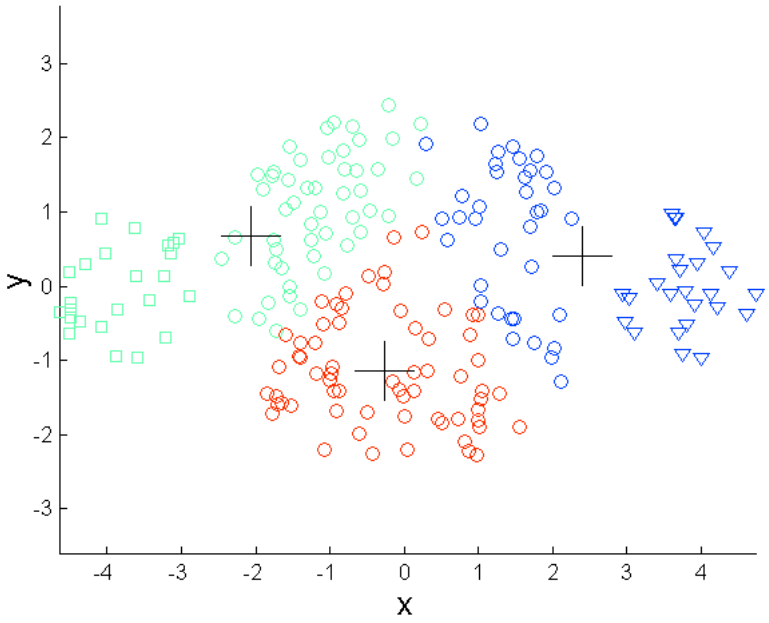
Περιορισμοί του K-means

- Ο K-means αντιμετωπίζει προβλήματα όταν οι ομάδες:
  - διαφέρουν σε μέγεθος
  - διαφέρουν σε πυκνότητα
  - έχουν μη-σφαιρικά σχήματα
- Ο K-means αντιμετωπίζει προβλήματα όταν τα δεδομένα έχουν εξωκείμενες τιμές

Περιορισμοί του K-means: Διαφορετικά μεγέθη

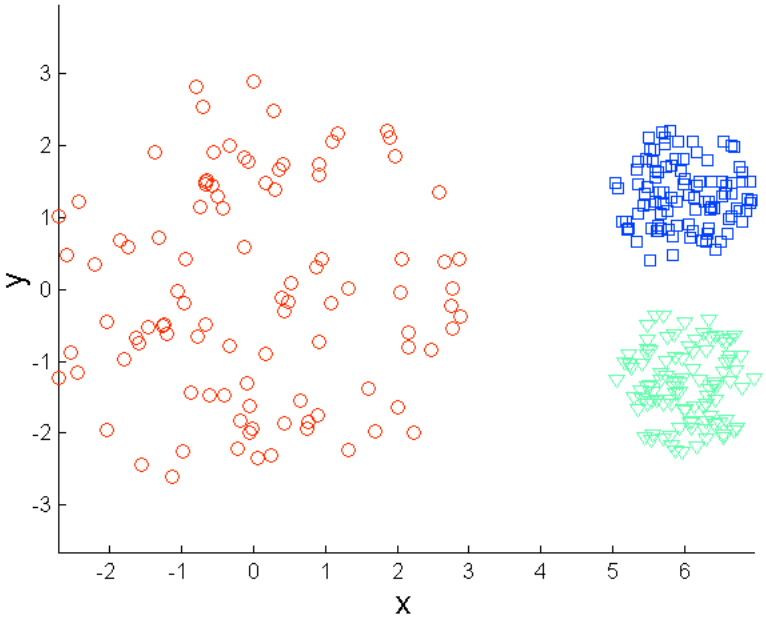


Αρχικά σημεία

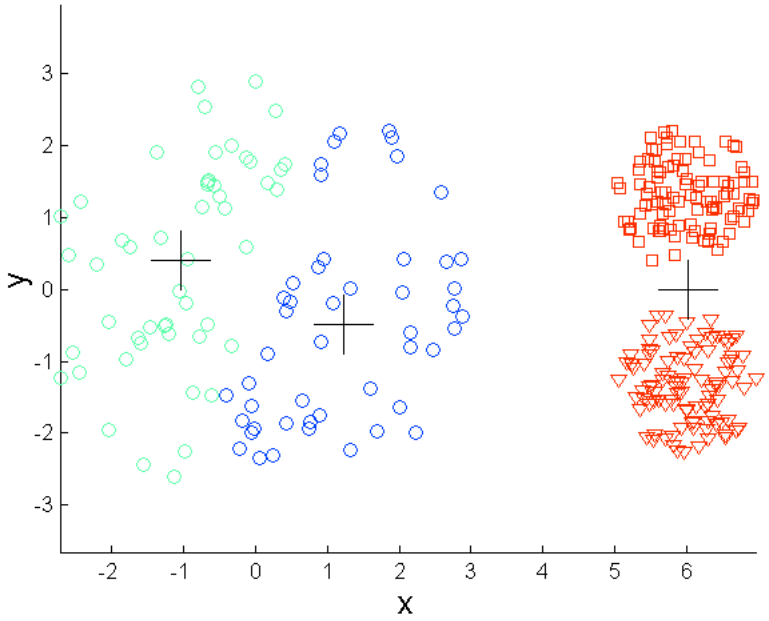


K-means (3 Ομάδες)

Περιορισμοί του K-means: Διαφορετική πυκνότητα

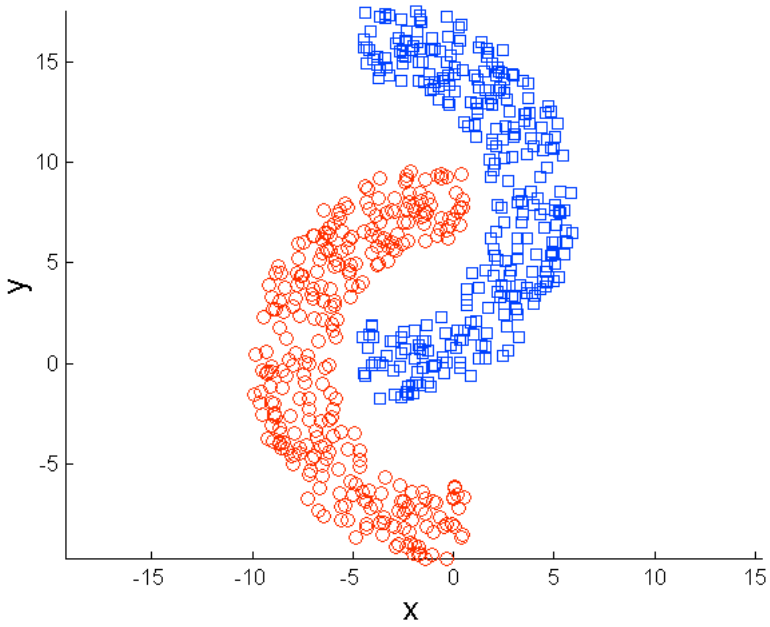


Αρχικά σημεία

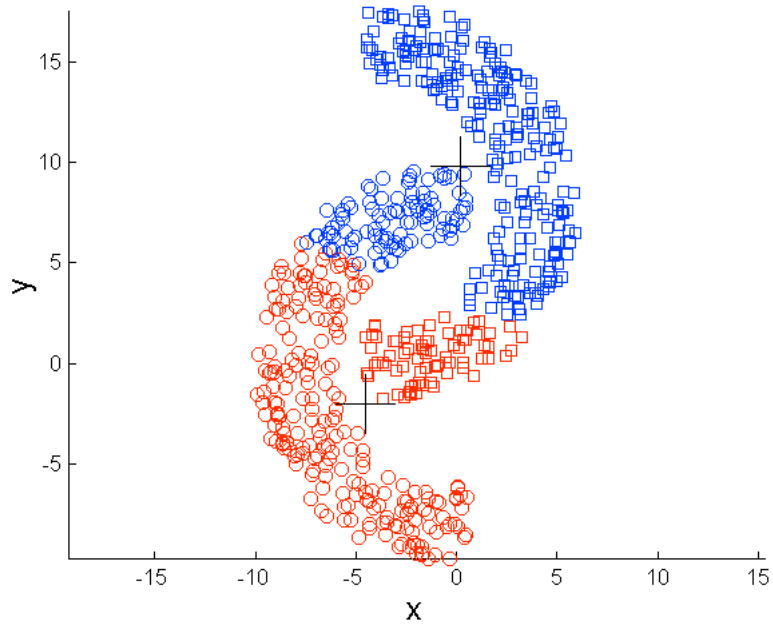


K-means (3 Ομάδες)

Περιορισμοί του K-means: Μη-σφαιρικά σχήματα



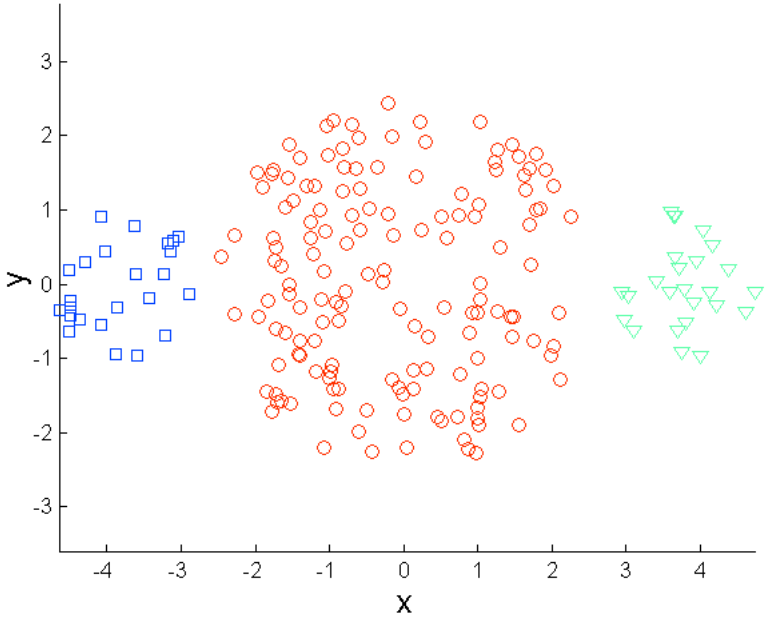
Αρχικά σημεία



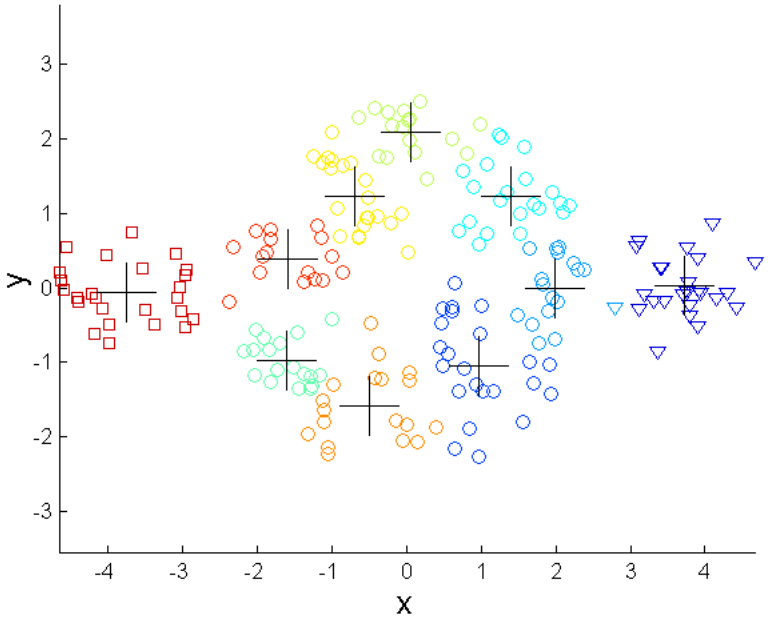
K-means (2 Ομάδες)

Ξεπερνώντας τους περιορισμούς του K-means

- Δημιουργία πολλών μικρών ομάδων και σύνθεσή τους σε επίπεδο μετ-επεξεργασίας

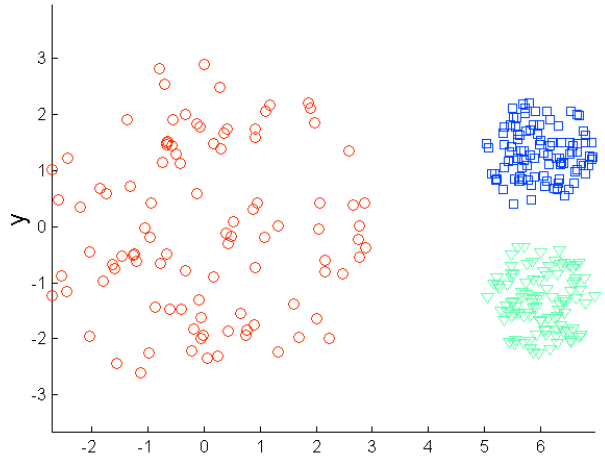


Αρχικά σημεία

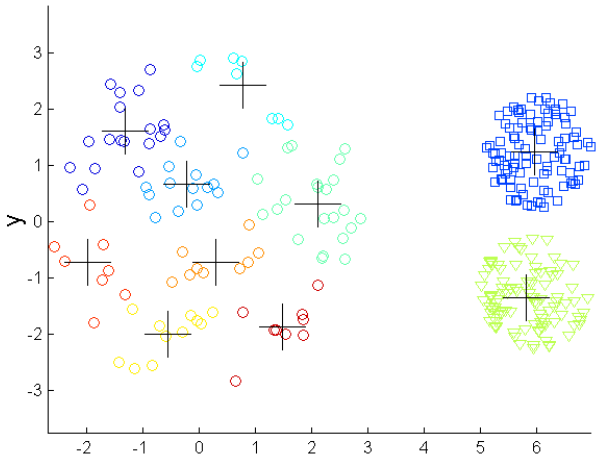


K-means

Ξεπερνώντας τους περιορισμούς του K-means



Αρχικά σημεία



K-means

K-medoids: K-means για κατηγορικά δεδομένα

- Ο K-means εφαρμόζεται μόνο σε αριθμητικά δεδομένα
- Για κατηγορικά δεδομένα (και διακριτές τιμές) --> K-medoids

K-medoids

- Κάθε ομάδα αναπαρίσταται από ένα σημείο που επιλέγεται ανάμεσα στα σημεία του X (medoid).
- Μια ομάδα περιέχει:
  - Το medoid της
  - Όλα τα σημεία στο X τα οποία:
    - Δε χρησιμοποιούνται ως medoids σε άλλες ομάδες
    - Βρίσκονται πιο κοντά στο medoid της ομάδας σε σχέση με τα medoids των άλλων ομάδων.

K-medoids

- Έστω Θ το σετ των medoids για όλες τις ομάδες, I_Θ το σετ των δεικτών των σημείων του X που αποτελούν το Θ και $I_{X-\Theta}$ το σετ των δεικτών των σημείων που δεν είναι medoids.
- Η εύρεση του σετ των medoids Θ το οποίο αναπαριστά καλύτερα το σετ δεδομένων X, είναι ισοδύναμη με την ελαχιστοποίηση της συνάρτησης κόστους:

$$J(\Theta, U) = \sum_{i \in I_{X-\Theta}} \sum_{j \in I_\Theta} u_{ij} d(\underline{x}_i, \underline{x}_j)$$

με

$$u_{ij} = \begin{cases} 1, & if \;\; d(\underline{x}_i, \underline{x}_j) = \min_{q \in I_\Theta} d(\underline{x}_i, \underline{x}_q) \\ 0, & otherwise \end{cases}, \quad i = 1, \ldots, N$$

Μέσες τιμές vs. medoids

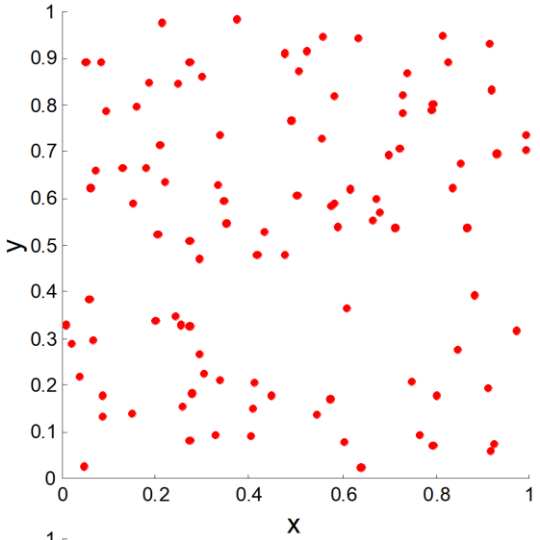
Αλγόριθμοι με Μέσες τιμές	Αλγόριθμοι με Medoids
1. Μόνο για αριθμητικά δεδομένα	1. Για αριθμητικά δεδομένα και διακριτά/κατηγορικά δεδομένα
2. Ευαίσθητοι σε εξωκείμενες τιμές	2. Λιγότερο ευαίσθητοι σε εξωκείμενες τιμές
3. Η μέση τιμή έχει μια σαφή γεωμετρική και στατιστική σημασία	3. Το medoid δεν έχει σαφή γεωμετρική σημασία
4. Απαιτείται μικρή υπολογιστική ισχύς	4. Απαιτείται μεγαλύτερη υπολογιστική ισχύς

Αξιολόγηση ομαδοποίησης

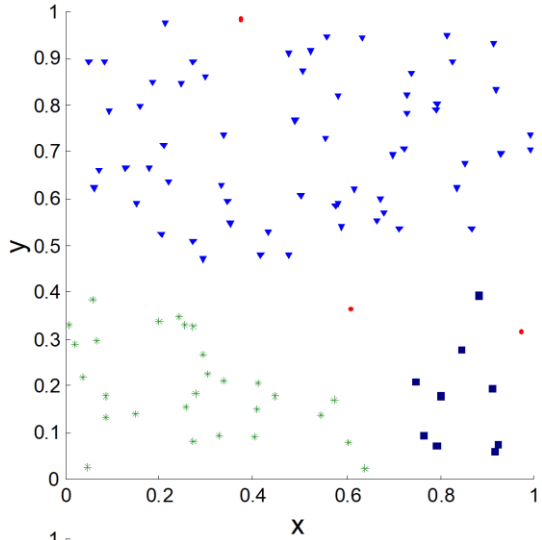
- Πώς αποτιμάται η “ποιότητα” μιας ομαδοποίησης;
- Το αποτέλεσμα της ομαδοποίησης αφήνεται για τον αναλυτή!!!
- Παρόλα αυτά πρέπει να γίνεται αξιολόγηση για να:
  - Αποφευχθεί η εύρεση προτύπων στο θόρυβο
  - Είναι δυνατή των αποτελεσμάτων της ομαδοποίησης με πρότερα γνωστά αποτελέσματα (ύπαρξη εξωτερικά ορισμένων τιμών κλάσεων)
  - Είναι δυνατή η σύγκριση αλγορίθμων ομαδοποίησης
  - Είναι δυνατή η σύγκριση διαφορετικών ομαδοποιήσεων
  - Είναι δυνατή η σύγκριση δυο ομάδων
  - Βρεθεί ο βέλτιστος αριθμός ομάδων στο σετ

Ομαδοποίηση σε τυχαία δεδομένα

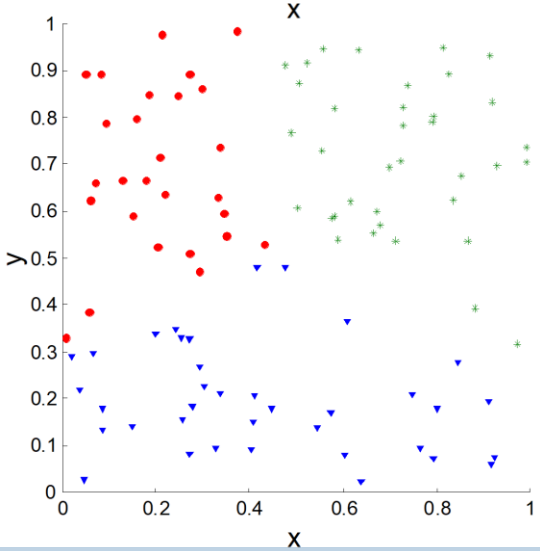
Τυχαία σημεία



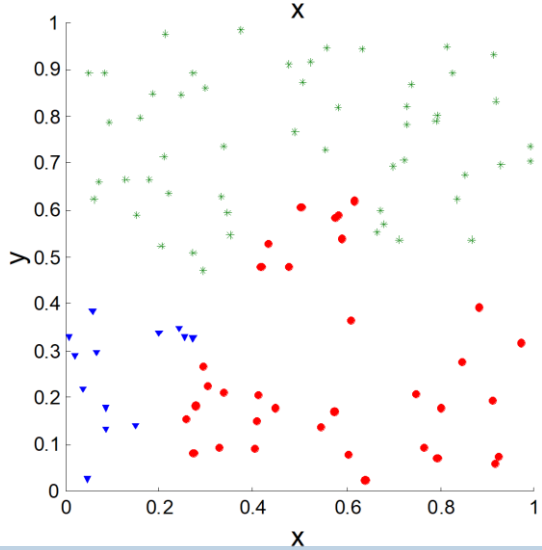
DBSCAN



K-means



Complete Link



Μετρικές αξιολόγησης ομάδων

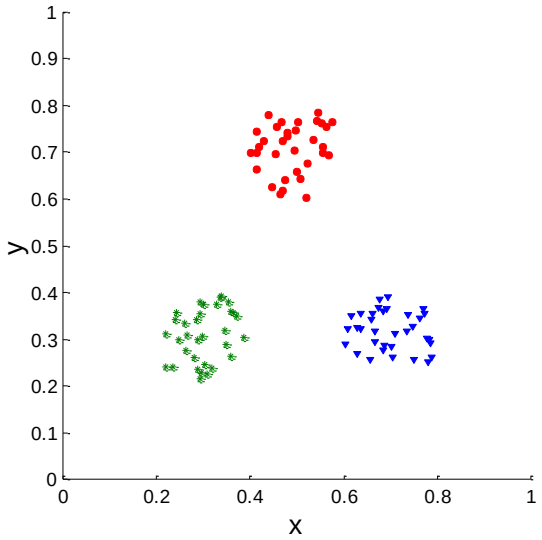
- Οι διάφορες αριθμητικές μετρικές που εφαρμόζονται για να αποτιμήσουν την ποιότητα της ομαδοποίησης, χωρίζονται σε 3 κατηγορίες:
  - Εξωτερικοί δείκτες: Χρησιμοποιούνται για να μετρήσουν τον βαθμό στον οποίο ετικέτες στις ομάδες μπορούν να αντιστοιχιστούν σε εξωτερικά ορισμένες κλάσεις (π.χ. Εντροπία)
  - Εσωτερικοί δείκτες: Χρησιμοποιούνται για να μετρήσουν την ποιότητα μιας ομαδοποίησης χωρίς αναφορά σε εξωτερική πληροφορία (π.χ. SSE)
  - Σχετικοί δείκτες: Χρησιμοποιούνται για να συγκρίνουν δυο διαφορετικές ομαδοποιήσεις ή δυο ομάδες

Αξιολόγηση ομάδων μέσω συσχέτισης

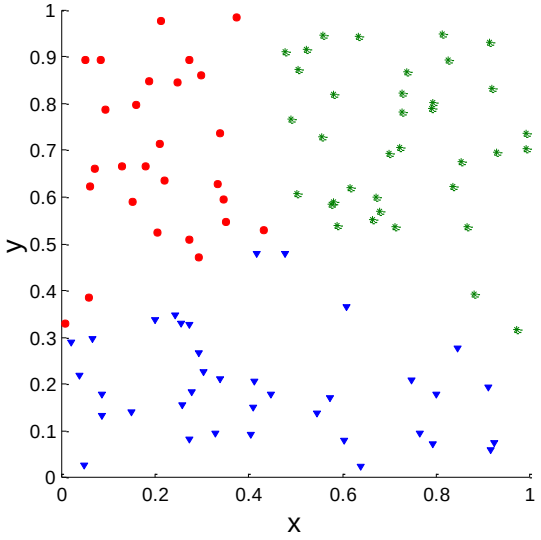
- Δυο πίνακες
  - Πίνακας γειτνίασης
  - Πίνακας “Γεγονότων”
    - Μια γραμμή και μια στήλη για κάθε σημείο
    - Η τιμή είναι 1 αν το συγκεκριμένο ζεύγος σημείων ανήκει στην ίδια ομάδα, και 0 διαφορετικά
- Υπολογισμός της συσχέτισης ανάμεσα στους δυο πίνακες
  - Οι πίνακες είναι συμμετρικοί. Κατά συνέπεια, υπολογίζεται η συσχέτιση μόνο των $n(n-1)/2$ εγγραφών
- Υψηλή συσχέτιση υποδεικνύει ότι τα σημεία που ανήκουν στην ίδια ομάδα είναι κοντά μεταξύ τους
- Δεν είναι καλή μετρική για πυκνωτικές ή συνεχείς ομαδοποιήσεις

Αξιολόγηση ομάδων μέσω συσχέτισης

- Συσχέτιση των πινάκων γειτνίασης και γεγονότων για την ομαδοποίηση με K-means για τα παρακάτω σετ δεδομένων



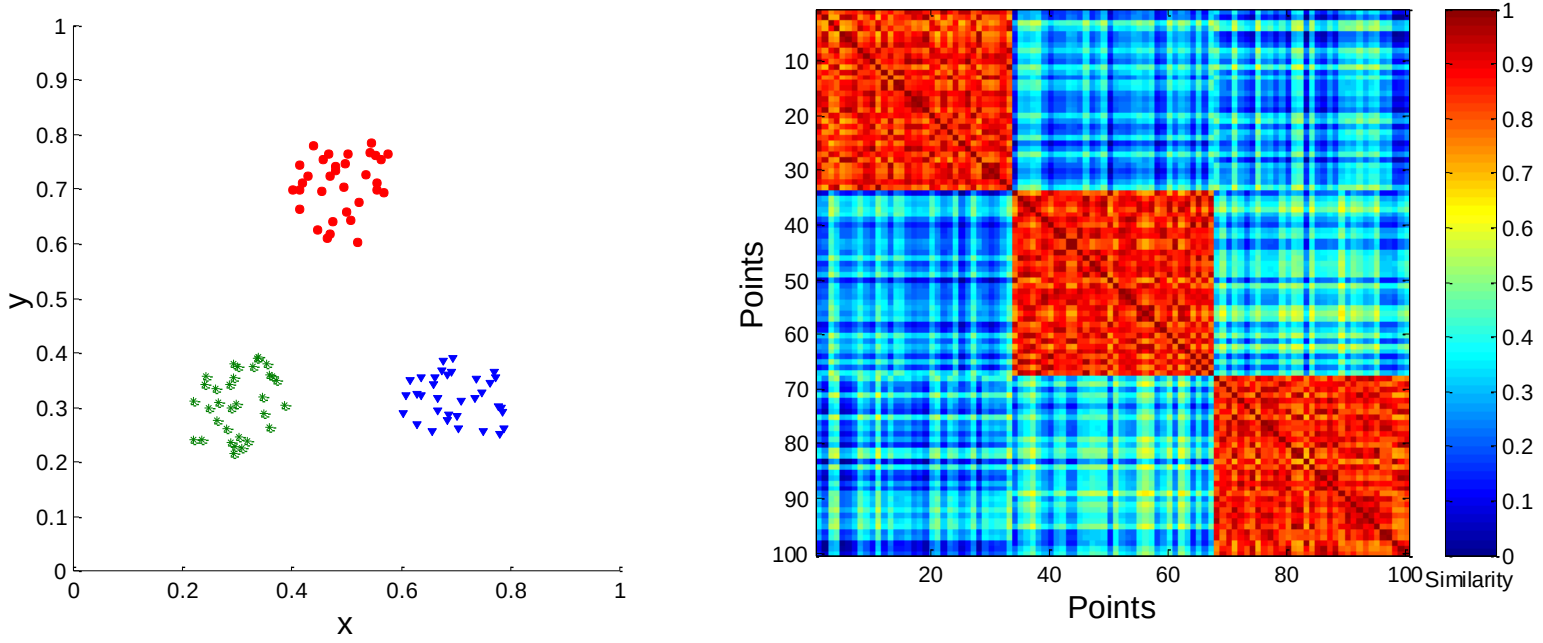
Corr = -0.9235



Corr = -0.5810

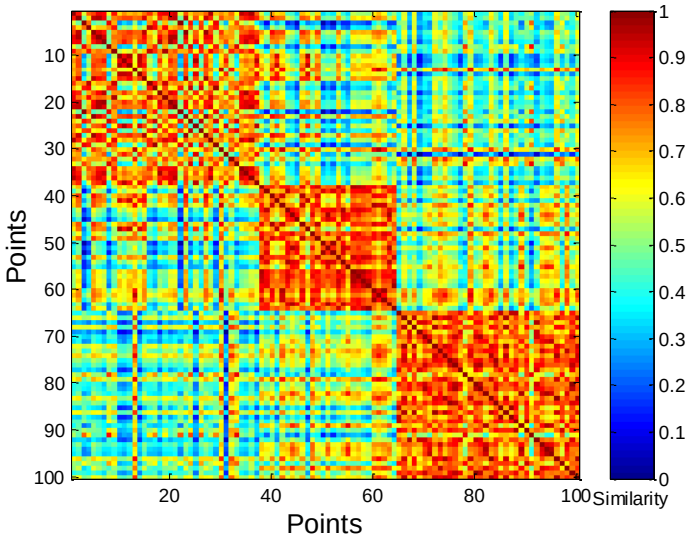
Αξιολόγηση ομάδων μέσω του πίνακα ομοιότητας

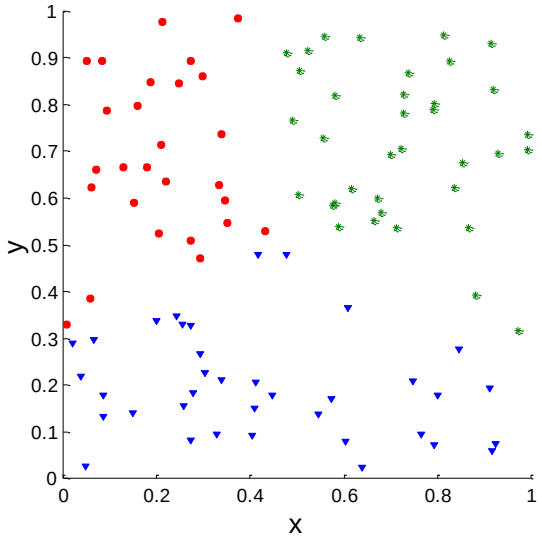
- Ταξινομήστε τον πίνακα ομοιότητας με βάση τις ετικέτες των ομάδων και οπτικοποιήστε:



Αξιολόγηση ομάδων μέσω του πίνακα ομοιότητας

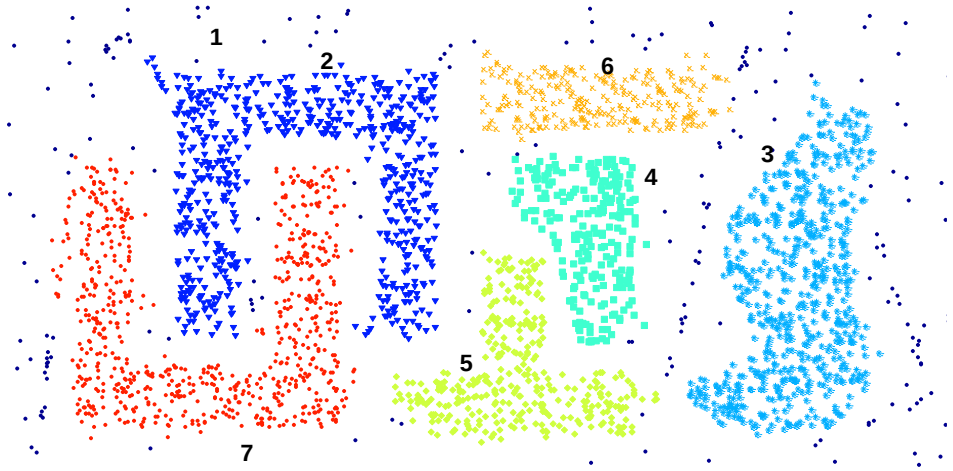
- Οι ομάδες σε τυχαία δεδομένα δεν είναι ξεκάθαρες

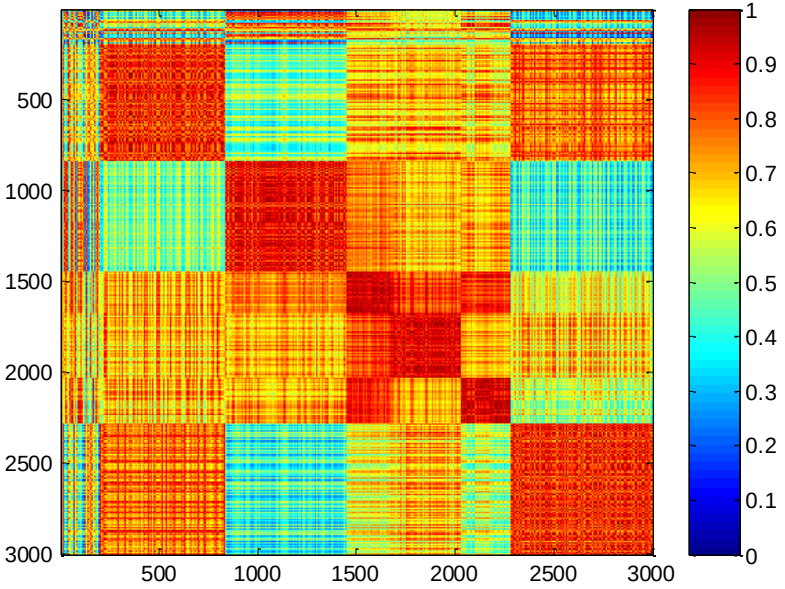




K-means

Αξιολόγηση ομάδων μέσω του πίνακα ομοιότητας

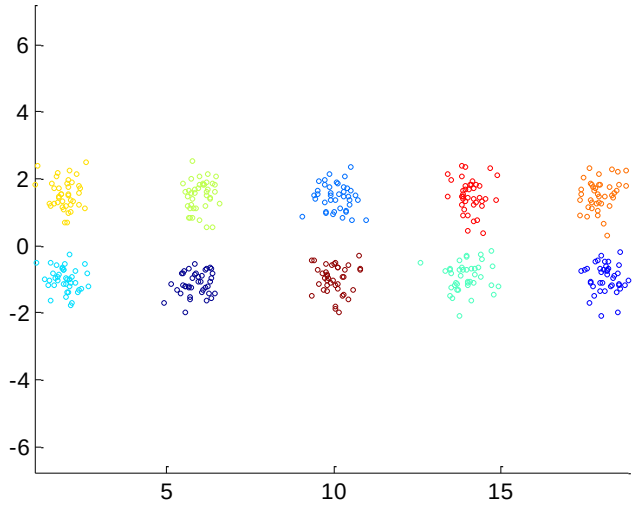


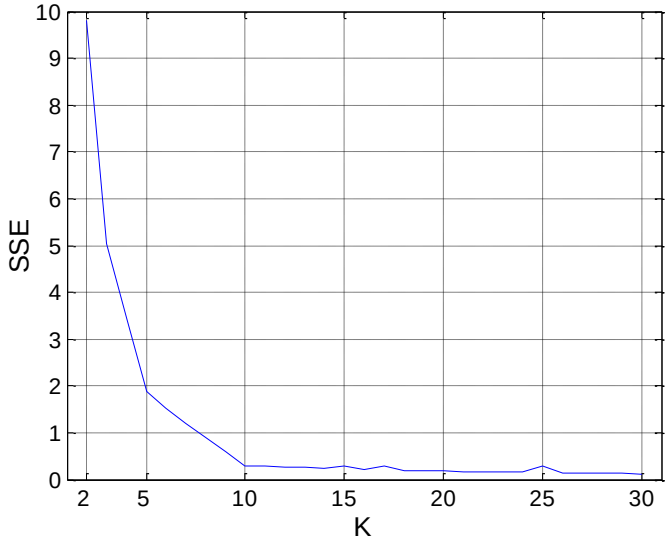


DBSCAN

Αξιολόγηση ομάδων μέσω του SSE

- Η ομαδοποίηση σε πιο σύνθετα σετ δεδομένων δεν είναι τόσο ξεκάθαρη
- Το SSE είναι καλή μετρική για τη σύγκριση δυο ομαδοποιήσεων ή δυο ομάδων
- Μπορεί να χρησιμοποιηθεί και για τον υπολογισμό του αριθμού των ομάδων

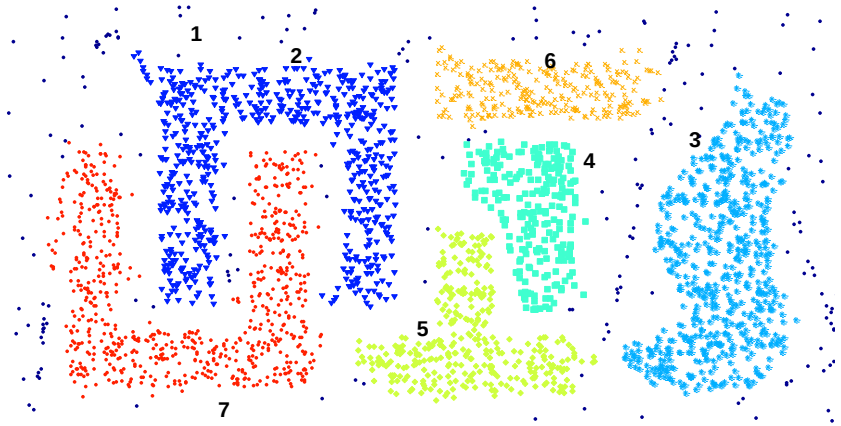


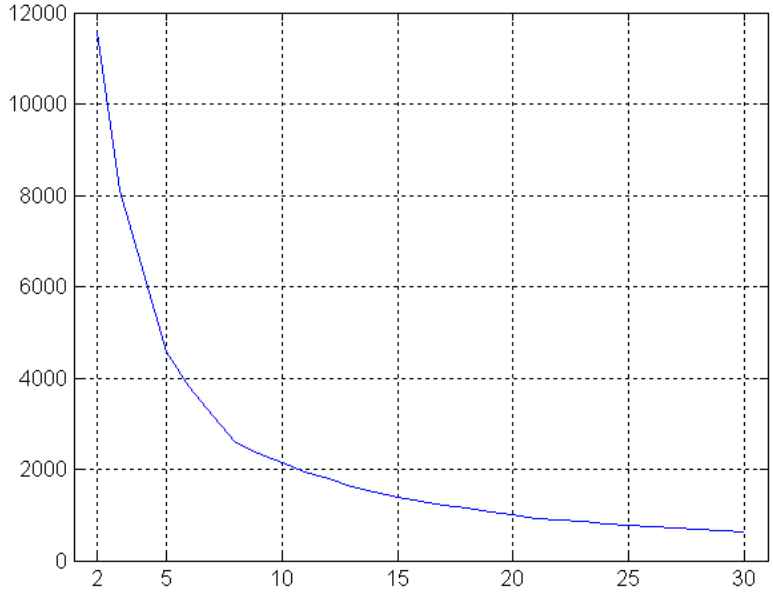


Το SSE των ομάδων με K-means

Αξιολόγηση ομάδων μέσω του SSE

- Η καμπύλη του SSE για ένα πιο σύνθετο σετ δεδομένων

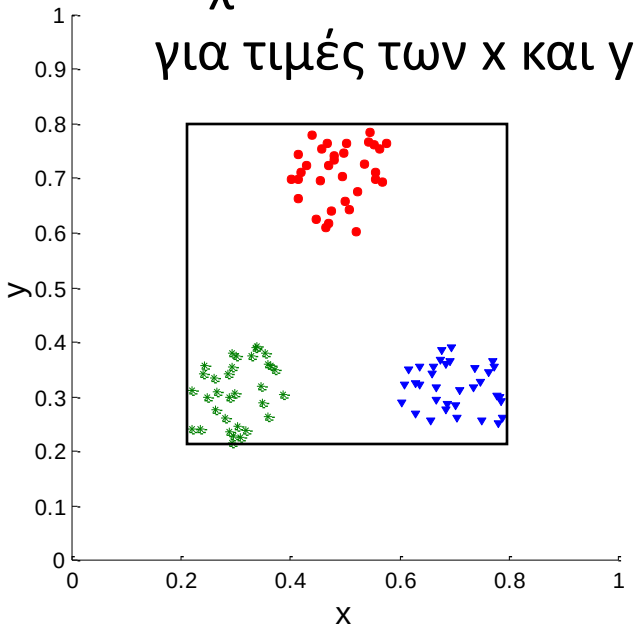


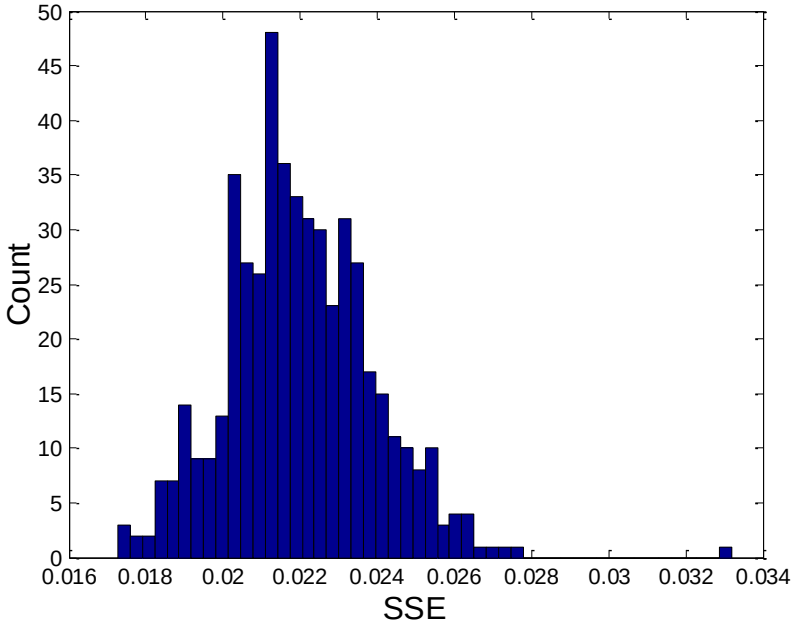


Το SSE των ομάδων με K-means

Στατιστική υποδομή για το SSE

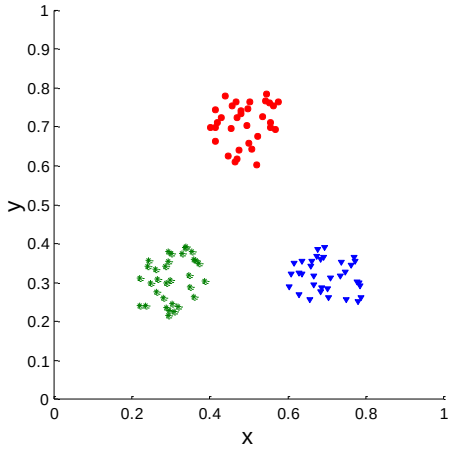
- Παράδειγμα
  - Σύγκριση του SSE= 0.005 (τιμή ομαδοποίησης) σε σχέση με SSE σε τυχαία δεδομένα
  - Το ιστόγραμμα δείχνει το SSE τριών ομάδων σε 500 επαναλήψεις τυχαίων 100άδων σημείων, κατανεμημένων στο εύρος 0.2 – 0.8 για τιμές των x και y



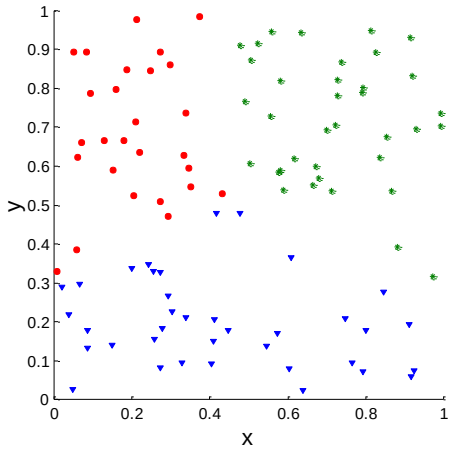


Στατιστική υποδομή για τη συσχέτιση

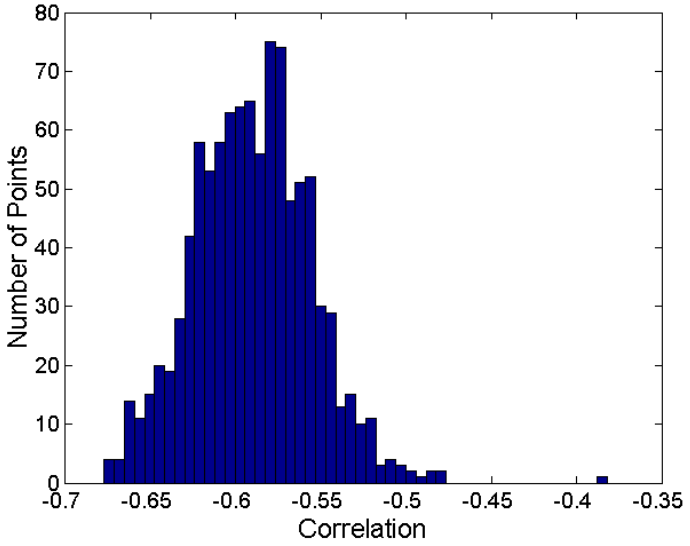
- Συσχέτιση των πινάκων γεγονότων και γειτνίασης για την ομαδοποίηση με K-means των παρακάτω σετ δεδομένων



Corr = -0.9235

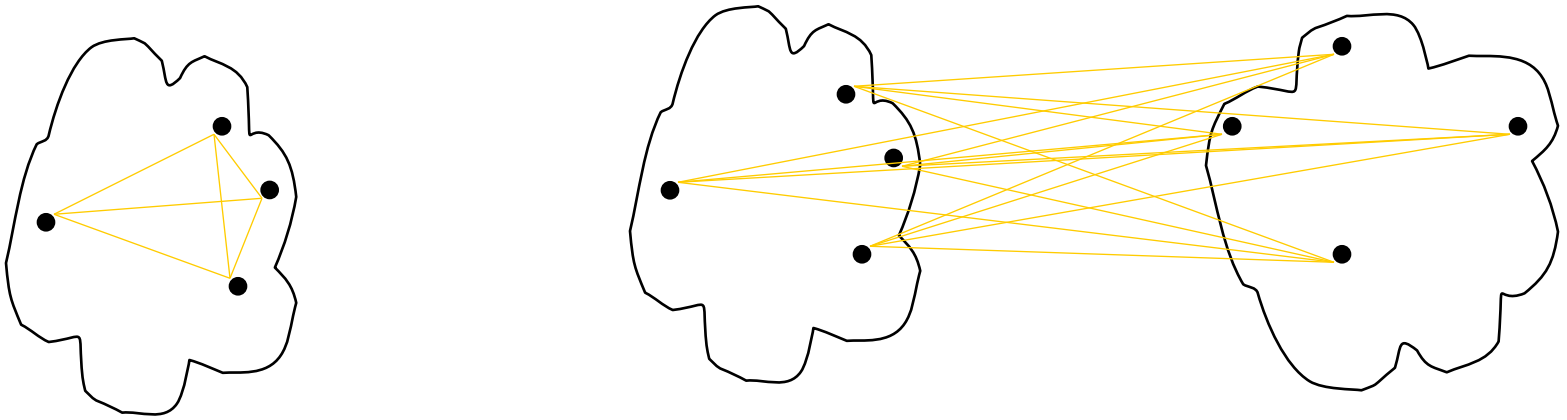


Corr = -0.5810



Cohesion και Separation

- Ένας γράφος γειτνίασης μπορεί να χρησιμοποιηθεί για τον υπολογισμό του cohesion και του separation.
  - Το cohesion μιας ομάδας είναι το άθροισμα των βαρών όλων των συνδέσεων σε μια ομάδα
  - Το separation μιας ομάδας είναι το άθροισμα των βαρών όλων των κόμβων μιας ομάδας με τους κόμβους εκτός της ομάδας



cohesion **separation**

Cohesion και Separation

- **Cohesion Ομάδας:** Μετρά πόσο στενά συνδεδεμένα είναι τα αντικείμενα μιας ομάδας (SSE)
- **Separation Ομάδας:** Μετρά πόσο διακριτές ή καλά διαχωρισμένες είναι οι ομάδες μεταξύ τους (τετραγωνικό σφάλμα)
- Το Cohesion μετράται ως το SSE μέσα στην ομάδα

$$\boldsymbol{WSS} = \sum_{\boldsymbol{i}} \sum_{\boldsymbol{x} \in \boldsymbol{C_i}} (\boldsymbol{x} - \boldsymbol{m_i})^2$$

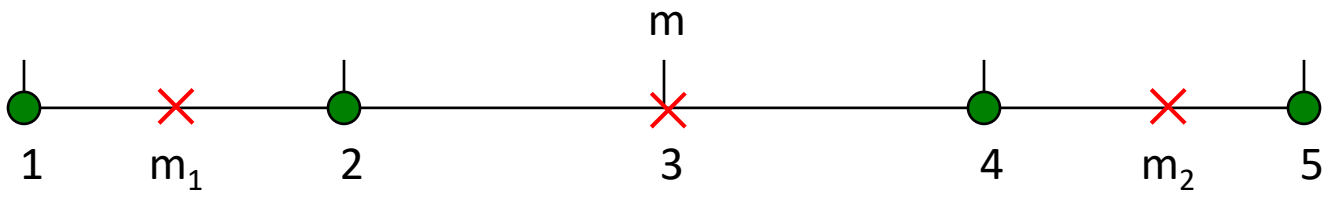
- Το Separation μετράται ως το άθροισμα των τετραγώνων ανάμεσα στις ομάδες

$$BSS = \sum_{i} |C_i| (m - m_i)^2$$

όπου $|C_i|$ είναι το μέγεθος της ομάδας i

Cohesion και Separation: παράδειγμα

- BSS + WSS = σταθερό



K=1 cluster:

$$WSS = (1-3)^2 + (2-3)^2 + (4-3)^2 + (5-3)^2 = 10$$

$$BSS = 4 \times (3-3)^2 = 0$$

$$Total = 10 + 0 = 10$$

K=2 clusters:

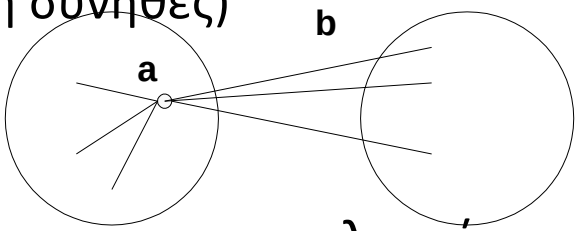
$$WSS = (1-1.5)^2 + (2-1.5)^2 + (4-4.5)^2 + (5-4.5)^2 = 1$$

$$BSS = 2 \times (3-1.5)^2 + 2 \times (4.5-3)^2 = 9$$

$$Total = 1 + 9 = 10$$

Συντελεστής Silhouette

- Ο συντελεστής Silhouette συνδυάζει τη λογική του cohesion και του separation, αλλά για μεμονωμένα σημεία, καθώς και για ομάδες και ομαδοποιήσεις
- Για ένα σημείο, *i:*
  - Υπολόγισε το ***a*** = η μέση απόσταση του *i* από τα υπόλοιπα σημεία στην ομάδα
  - Υπολόγισε το ***b*** = min (μέσης απόστασης του *i* από τα σημεία στις άλλες ομάδες)
  - Ο συντελεστής silhouette για το σημείο αυτό είναι:
    $s = 1 - a/b$ if $a < b$, (ή $s = b/a - 1$ αν $a \geq b$, μη σύνηθες)
  - Τυπικά, ανάμεσα στο 0 και το 1
  - Όσο πιο κοντά στο 1, τόσο καλύτερα



- Μπορεί να γίνει υπολογισμός του μέσου πλάτους του συντελεστή για ομάδες και ομαδοποιήσεις

Ανακεφαλαίωση

- Μη επιβλεπόμενη μάθηση
- Διάφοροι τύποι ομαδοποίησης
  - Διαχωρισμού
  - Ιεραρχική
  - Πυκνωτική
- K-Means και παράγωγα
- Αξιολόγηση ομάδων με βάση εξωτερικούς, εσωτερικούς και σχετικούς δείκτες
- Πηγές:
  - Introduction to Data Mining, Tan, Steinbach, Kumar.
  - Pattern recognition, Theodoridis & Koutroumbas.