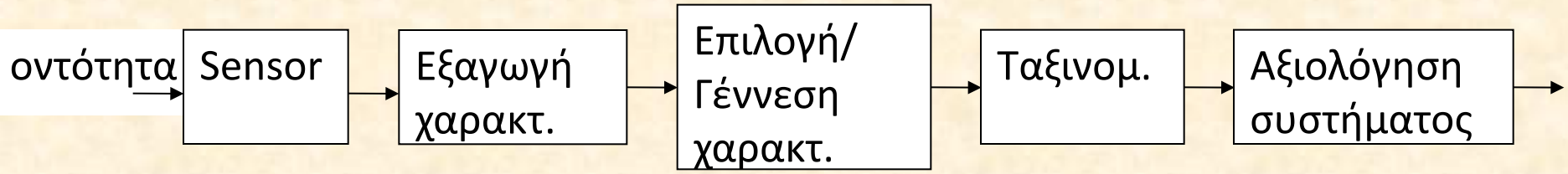
ΑΝΑΓΝΩΡΙΣΗ ΠΡΟΤΥΠΩΝ (PATTERN RECOGNITION)

Σέργιος Θεοδωρίδης

Κωνσταντίνος Κουτρούμπας

Επιλογή / γέννεση χαρακτηριστικών (feature selection / classification)

Τυπική αρχιτεκτονική συστήματος ταξινόμησης



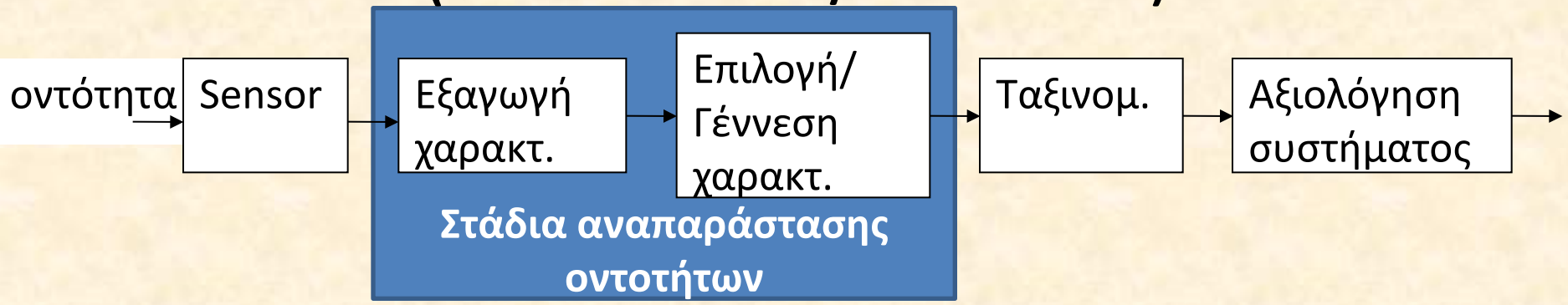
Επιβλεπόμενη ταξινόμηση (Supervised Classification):

Δεδομένα: Ένα πλήθος κλάσεων $\omega_1,\ldots \omega_M$ στις οποίες τα στοιχεία ενός **συνόλου "οντοτήτων"** θα ταξινομηθούν.

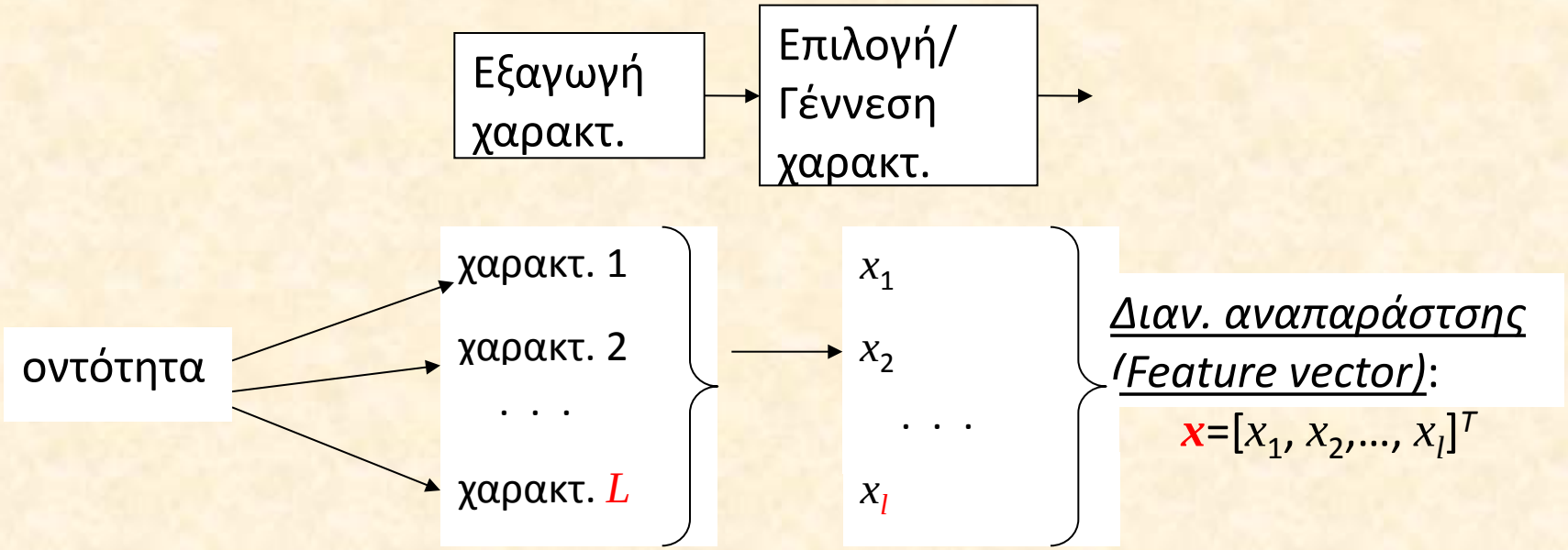
Στόχος:

Δοθείσης μιας οντότητας, **καταχώρησέ** την στην "πιο κατάλληλη" κλάση (με άλλα λόγια, **εκτίμησε** την "πιο κατάλληλη" κλάση γι' αυτή).

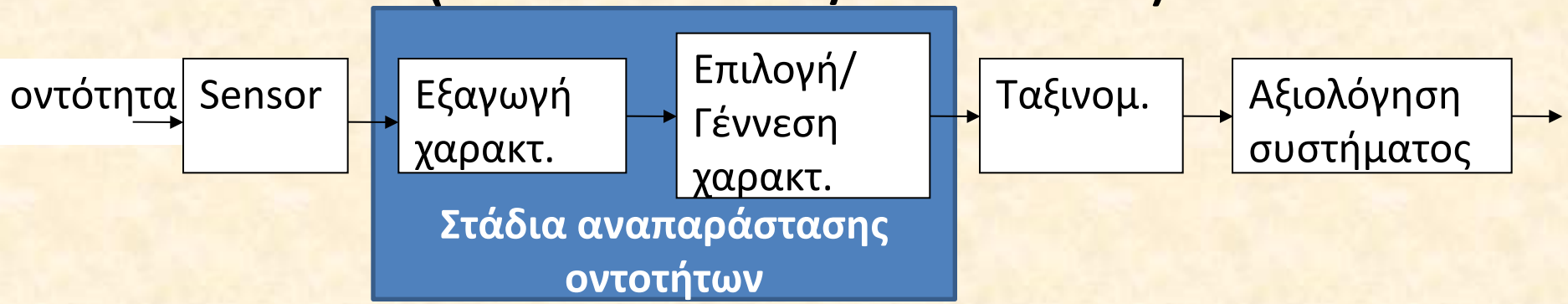
Επιλογή / γέννεση χαρακτηριστικών (feature selection / classification)



Αναπαράσταση οντοτήτων (πώς δουλεύει)



Επιλογή / γέννεση χαρακτηριστικών (feature selection / classification)



Επιλογή χαρακτηριστικών: Ανάμεσα από τα L αρχικά χαρακτηριστικά **επέλεξε** εκείνα που έχουν αυξημένη διακριτική ικανότητα.

Γιατί επιλογή χαρακτηριστικών;

Μεγάλες τιμές του L έχουν τρία μειονεκτήματα:

- Υψηλές υπολογιστικές απαιτήσεις
- Κακές εκτιμήσεις λάθους (peaking phenomenon(*))
- Απόδοση με χαμηλή ικανότητα γενίκευσης

(*) Για δεδομένο N, η πρόσθεση χαρακτ. πάνω από συγκεκριμένο όριο οδηγεί σε αύξηση της πιθανότητας σφάλματος

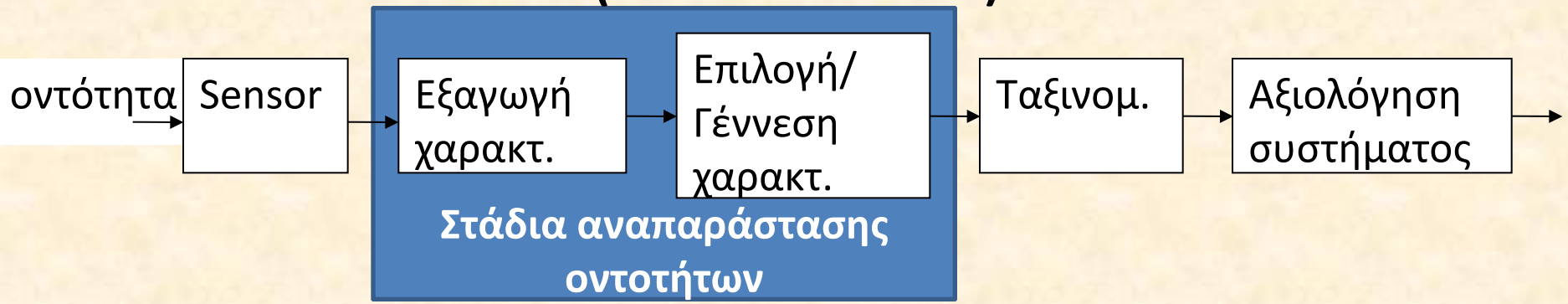
$l<N/3$ είναι μια λογική επιλογή σε αρκετές περιπτώσεις.

Στόχοι της επιλογής χαρακτηριστικών:

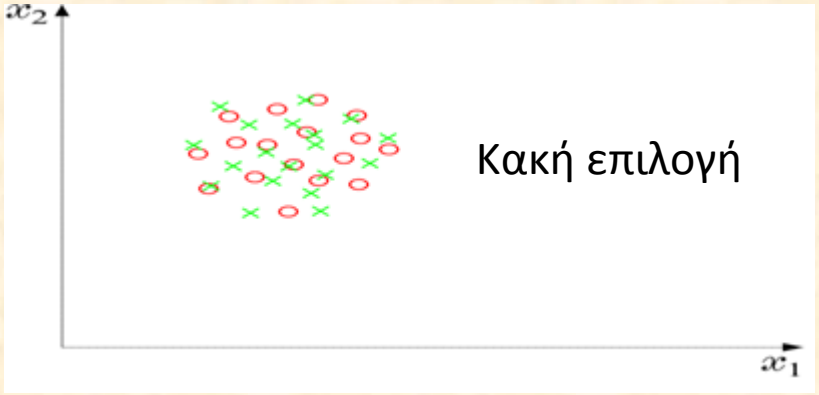
- Ο **προσδιορισμός** του **"βέλτιστου" αριθμού** l (από L) χαρακτηριστικά.
- Η **επιλογή** των **"βέλτιστων"** l χαρακτηριστικών.

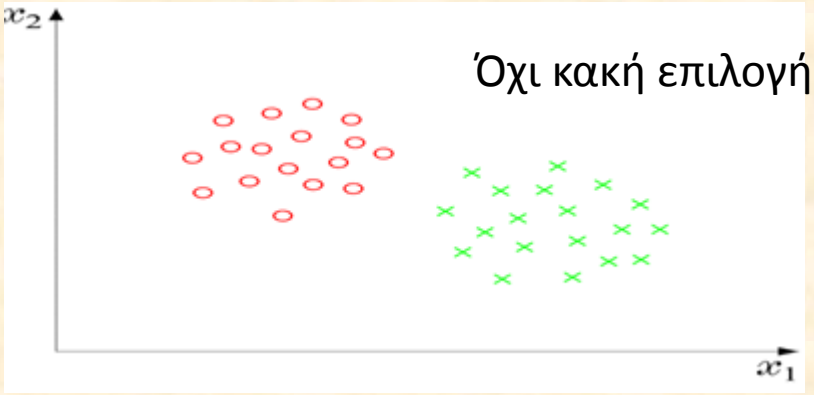
ΣΗΜ.: Σε μερικές περιπτώσεις, οι παραπάνω διαδικασίες γίνονται ταυτόχρονα.

Επιλογή χαρακτηριστικών (feature selection)



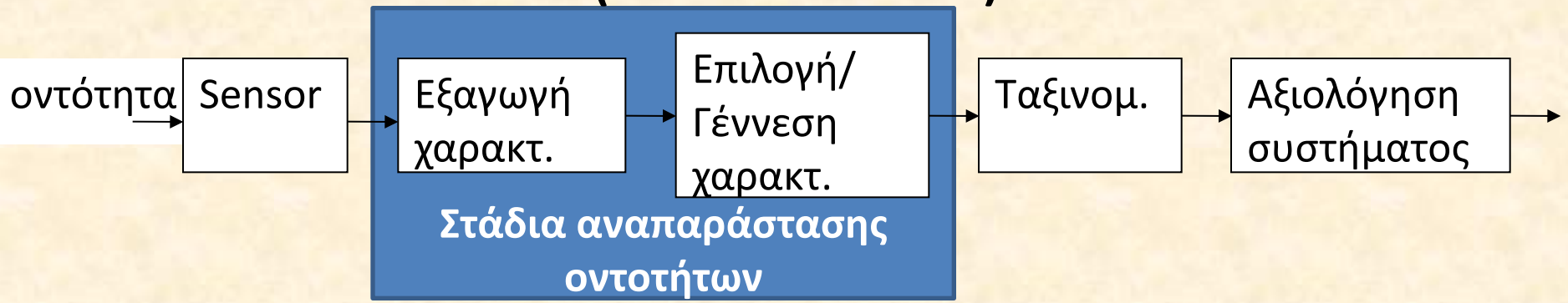
Επιλογή χαρακτηριστικών: Μεταξύ L (αρχικών) features **επέλεξε** εκείνα που έχουν αυξημένη διακριτική ικανότητα για τις κλάσεις.



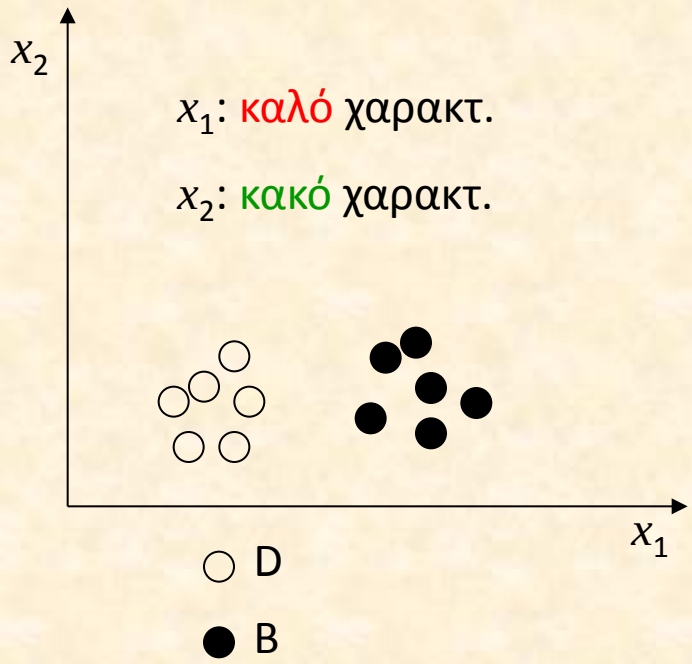


Ο βέλτιστος συνδυασμός χαρακτηριστικών αναπαριστά τις κλάσεις ώστε να έχουμε
-**Μικρή διακύμανση** μέσα στις **κλάσεις** (συμπαγείς κλάσεις)
-**Μεγάλη απόσταση** μεταξύ κλάσεων.

Επιλογή χαρακτηριστικών (feature selection)



Επιλογή χαρακτηριστικών: Μεταξύ L (αρχικών) features **επέλεξε** εκείνα που έχουν αυξημένη διακριτική ικανότητα για τις κλάσεις.



- Μέθοδοι επιλογής εξαρτώμενες από την εφαρμογή

  Βασίζονται στη "φυσική" του υπό μελέτη θέματος (π.χ., η επιλογή συγκεκριμένου μήκους κύματος για το διαχωρισμό δύο υλικών)

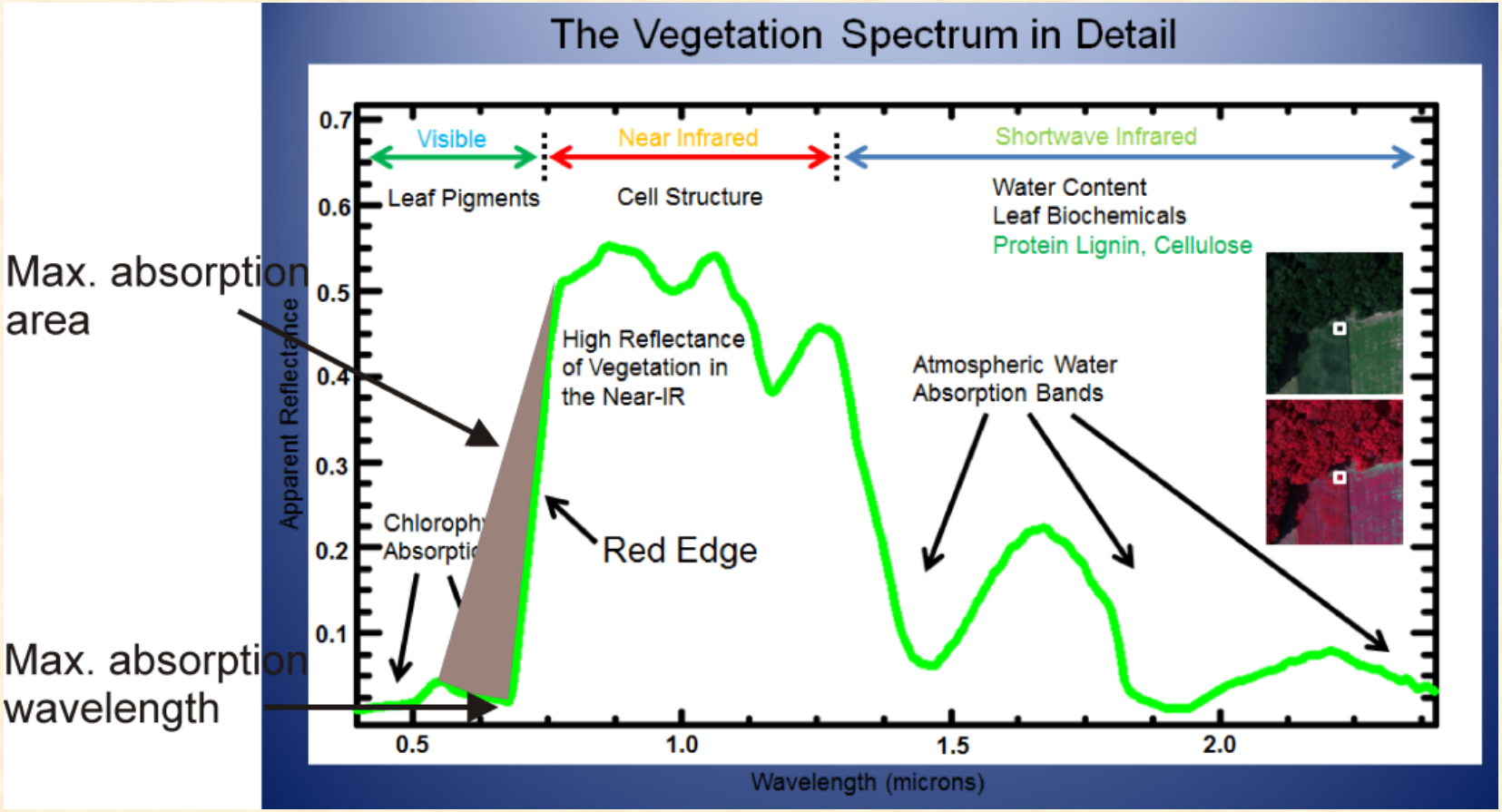
- Μέθοδοι επιλογής μη εξαρτώμενες από την εφαρμογή

  Βασίζονται στη χρήση μαθηματικών εργαλείων (στατιστικοί έλεγχοι, μέθοδοι βελτιστοποίησης)

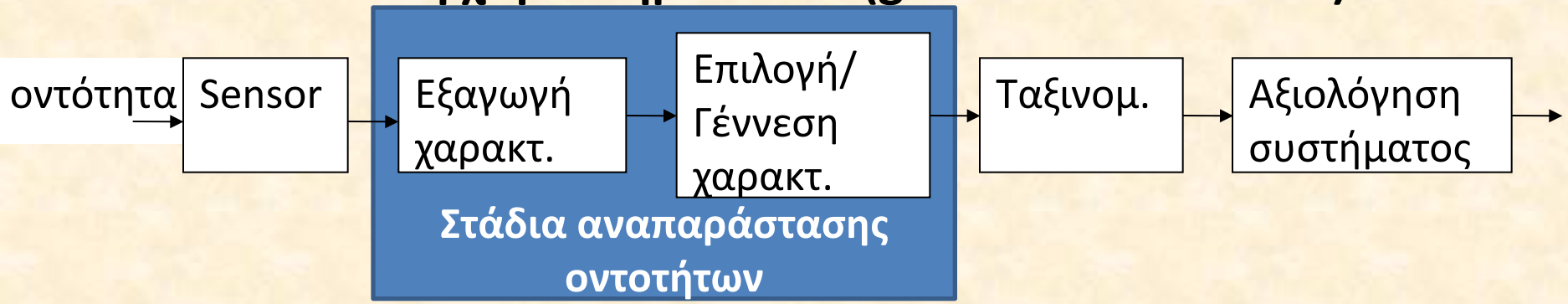
  Τα χαρακτ. Μπορεί να **ελεγχθούν (a)** ανεξάρτητα το ένα από το άλλο ή **(b)** σε ομάδες.

Επιλογή χαρακτηριστικών (feature selection)

Παράδειγμα επιλογής φυσικών χαρακτ.



Γέννεση χαρακτηριστικών (generation selection)



Γέννεση χαρακτηριστικών: «Γέννησε» νέα χαρακτηριστικά από ήδη υπάρχοντα.

Στόχος: Η περαιτέρω **μείωση** του αριθμού των χαρακτηριστικών.

Δύο δημοφιλείς μέθοδοι:

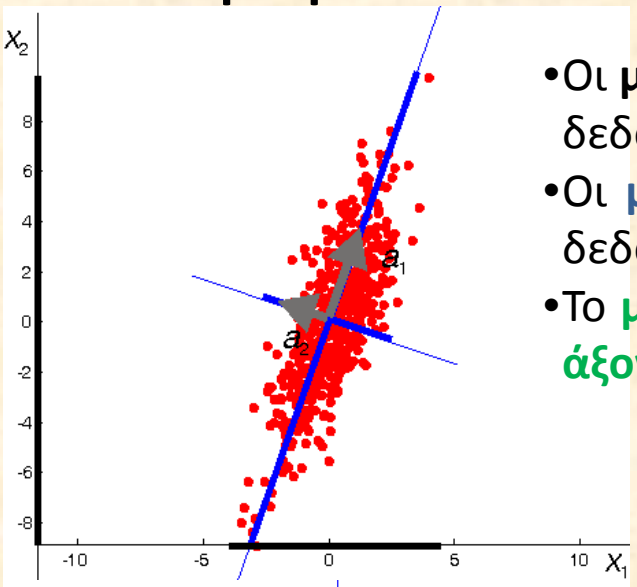
Ανάλυση κυρίων συνιστωσών (Principal component analysis - PCA):

Ο αρχικός χώρος **μετασχηματίζεται** σε νέο ορθογώνιο χώρο όπου τα **χαρακτηριστικά** είναι **μη συσχετισμένα**. Συγκεκριμένα**:** κατά μήκος του (καλούμενου) 1ου κύριου άξονα διατηρείται η μέγιστη δυνατή διακύμανση του συνόλου δεδομένων, κατά μήκος του (καλούμενου) 2ου κύριου άξονα διατηρείται η μέγιστη δυνατή (από την **εναπομείνασα**) διακύμανση κλπ.

Προβάλοντας στο χώρο που ορίζεται από τους «πιο σημαντικούς» κύριους άξονες επιτυγχάνουμε **μείωση** της **διάστασης**.

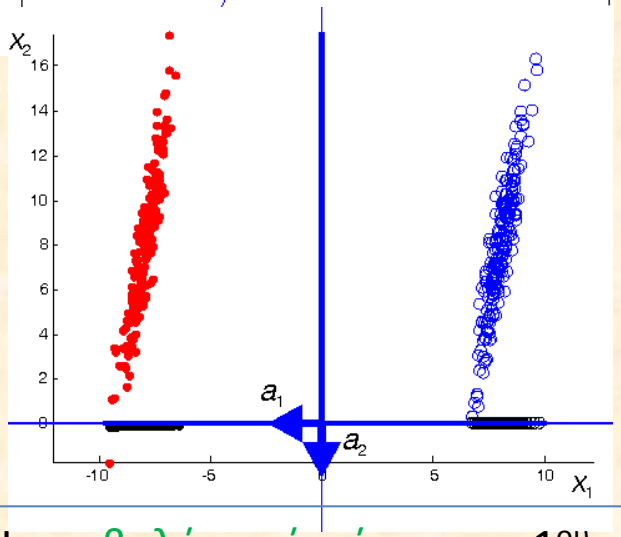
Γέννεση χαρακτηριστικών (generation selection)

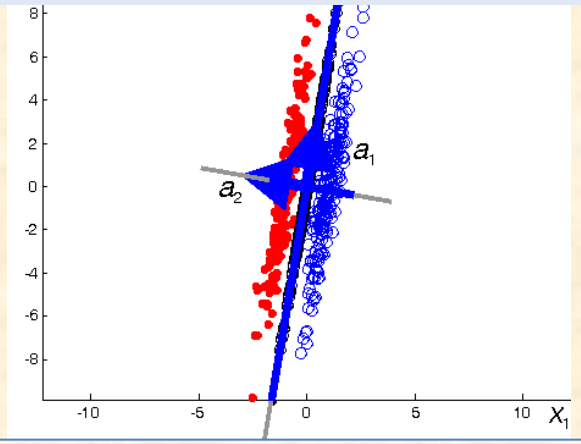
Ανάλυση κύριων συνιστωσών - Principal Component Analysis - PCA)



- Οι **μαύρες γραμμές** δείχνουν το διάστημα τιμών των διανυσμάτων δεδομένων κατά μήκος των **αρχικών αξόνων**.
- Οι **μπλε γραμμές** δείχνουν το διάστημα τιμών των διανυσμάτων δεδομένων κατά μήκος των **κύριων αξόνων**.
- Το **μέγιστο εύρος** τιμών παρατηρείται κατά μήκος του **1ου κύριου άξονα.**

ΣΗΜ: Η **διατήρηση** της μέγιστης δυνατής διασποράς του συνόλου δεδομένων **ΔΕΝ εγγυάται** απαραίτητα και τη **διατήρηση** της διαχωρισιμότητας των κλάσεων.





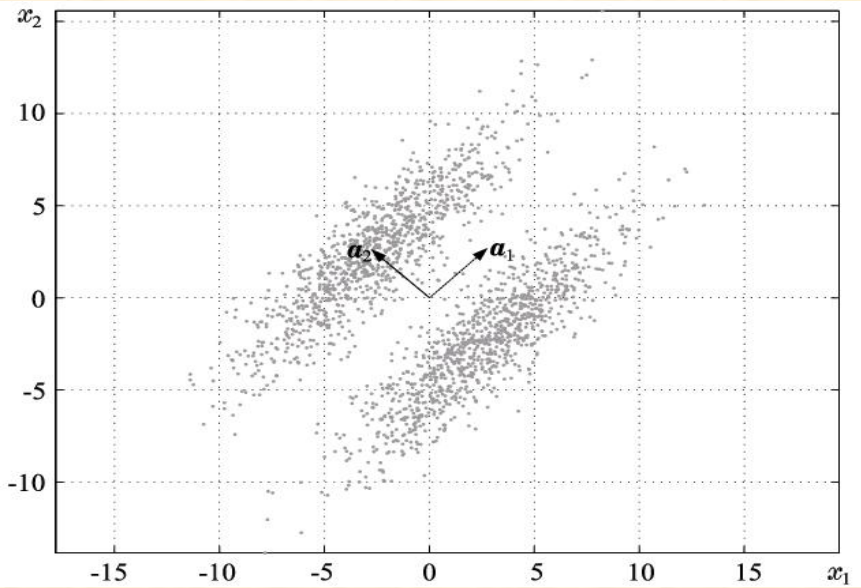
Η προβολή κατά μήκος του 1ου κύρ. άξονα (a_1) διατηρεί τη διαχωρισιμότητα των κλάσεων.

Η προβολή κατά μήκος του 1ου κύρ. άξονα (a_1) **ΔΕΝ** διατηρεί τη διαχωρισιμότητα των κλάσεων.

Γέννεση χαρακτηριστικών (generation selection)

Ανάλυση ανεξάρτητων συνιστωσών - Independent Component Analysis-ICA)

- **Παράγει** στατιστικώς ανεξάρτητα χαρακτηριστικά
- Δείχνει **προτίμηση** σε χαρακτ. των οποίων η κατανομή έχει τη μικρότερη δυνατή ομοιότητα με την Gaussian pdf (αυτά είναι πιο πιθανόν να διατηρήσουν την πληροφορία διαχωρισμού των κλάσεων)



$K_4(y_2) = -1.7$

$K_4(y_1) = 0.1$

- Κατά μήκος της a_2 η στατιστική είναι bimodal.
- Κατά μήκος της a_1 η στατιστική είναι unimodal (μοιάζει με Gaussian).

Ομαδοποίηση (Clustering) – Βασικές έννοιες

Ορισμός ομαδοποίησης

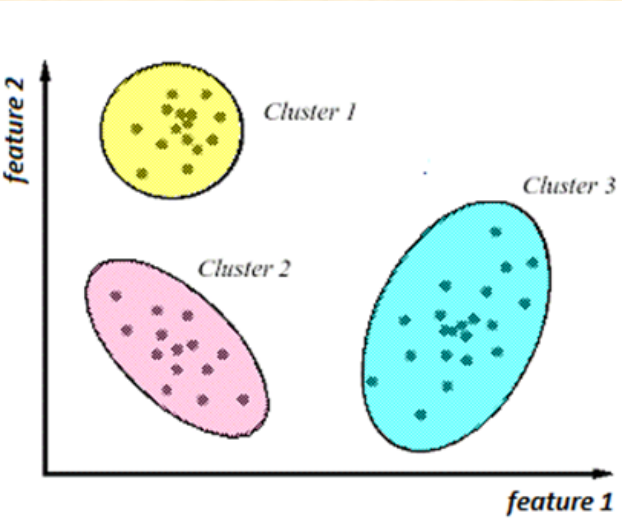
Ομαδοποίηση: Η διαδικασία κατά την οποία "όμοιες" **οντότητες** καταχωρούνται στο ίδιο σύνολο (**"ομάδα"** - **"cluster"**).

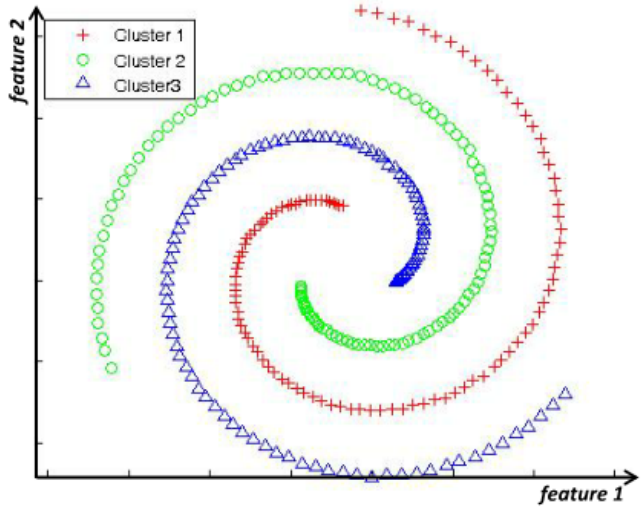
Ορισμός ομάδας: παρατηρήσεις

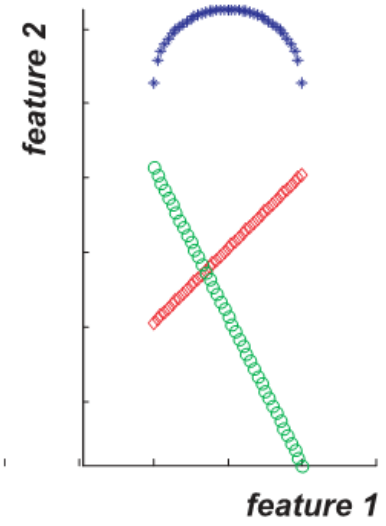
✓ Δεν υπάρχει αυστηρός ορισμός για την έννοια της ομάδας.

✓ Ωστόσο, εννοούμε συνήθως μια **συσσώρευση σημείων** γύρω από :

- ένα συγκεκριμένο σημείο του χώρου χαρακτηριστικών (στην περίπτωση αυτή η ομάδα μοντελοποιείται συνήθως από κανονική κατανομή).
- ένα manifold (π.χ. υπερεπίπεδο, υπερσφαίρα) στο χώρο χαρακτηριστικών.



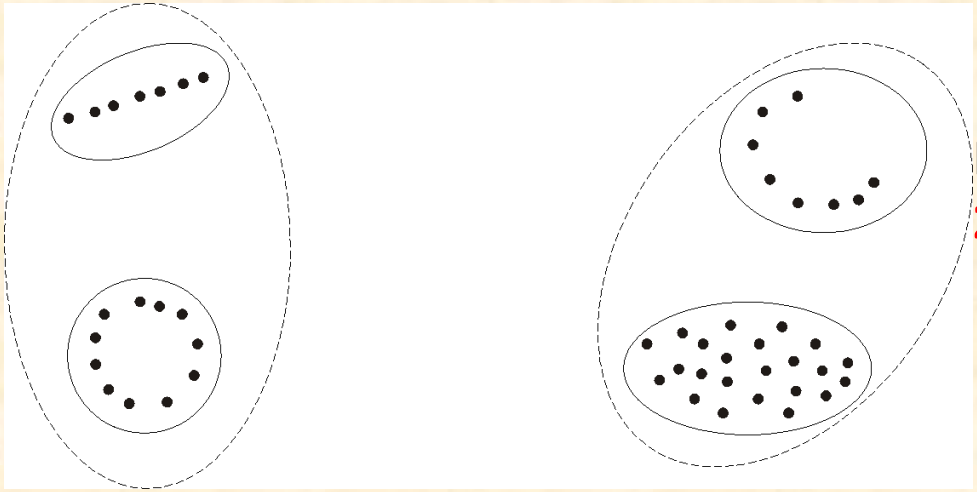




Ομαδοποίηση (Clustering) – Βασικές έννοιες

Ομαδοποίηση είναι η διαδικασία της ταυτοποίησης συσσωρεύσεων σημείων σ' έναν *l*-διάστατο χώρο.

1η Παρατήρηση: Για την ομαδοποίηση, η **υποκειμενικότητα** είναι μια αναπόφευκτη πραγματικότητα..



Ποιος είναι ο «σωστός» αριθμός ομάδων;;
2 ή **4;;**

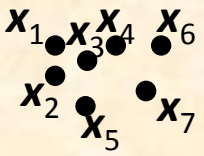
2η παρατήρηση: Υπάρχει τουλάχιστον μία "σφαιρική" παράμετρος σε κάθε αλγόριθμο ομαδοποίησης (π.χ. ο αριθμός των ομάδων) που χρειάζεται να καθοριστεί από το χρήστη (**ελλιπώς ορισμένο** (**ill-posed**) πρόβλημα).

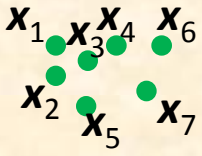
Ομαδοποίηση (Clustering) – Βασικές έννοιες

Ομαδοποίηση είναι η διαδικασία της ταυτοποίησης συσσωρεύσεων σημείων σ' έναν *l*-διάστατο χώρο.

Αναπαράσταση ομάδας

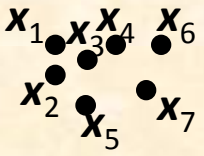
✓ Μέσω όλων των σημείων της (μη-παραμετρική αναπαράσταση)

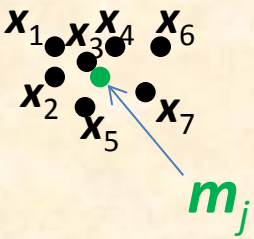




$\{x_1, x_2, x_3, x_4, x_5, x_6, x_7\}$

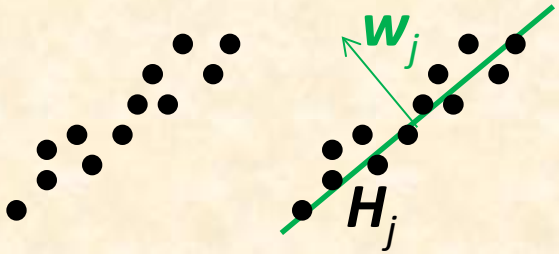
✓ Μέσω ενός συνόλου παραμέτρων (παραμετρική αναπαράσταση)





Παράμετροι:

$m_j = [m_{j1}, m_{j2}, \ldots, m_{jl}]^T$



H_j: $\theta_j^T x_i + \theta_{j0} = 0$

Παράμετροι:

$\theta_j = [\theta_{j1}, \theta_{j2}, \ldots, \theta_{jl}]^T$, θ_{j0}

Ομαδοποίηση (Clustering) – Βασικές έννοιες

Ομαδοποίηση: Η έννοια της ομοιότητας

Μια σημαντική παρατήρηση: Η έννοια της εγγύτητας (proximity) μεταξύ διανυσμάτων (ή μεταξύ ομάδων σε κάποιες περιπτώσεις) πρέπει να ποσοτικοποιηθεί.

Μέτρα ομοιότητας μεταξύ $\boldsymbol{x}_i$ και $\boldsymbol{x}_j$: Όσο υψηλότερη είναι η τιμή τους, τόσο πιο όμοια είναι τα $\boldsymbol{x}_i$ και $\boldsymbol{x}_j$.

Μέτρα ανομοιότητας μεταξύ $\boldsymbol{x}_i$ και $\boldsymbol{x}_j$: Όσο υψηλότερη είναι η τιμή τους, τόσο λιγότερο όμοια είναι τα $\boldsymbol{x}_i$ και $\boldsymbol{x}_j$.

Σε αρκετές περιπτώσεις, χρειάζεται επίσης ο ορισμός ενός κατωφλίου για τον ορισμό των εννοιών “όμοια”, “ανόμοια”.

Παραδείγματα:

(a) Μεταξύ **διανυσμάτων**: τετραγωνική Ευκλείδ. απόσταση $d(\boldsymbol{x}_i,\boldsymbol{x}_j)=\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2$

(b) Μεταξύ συνόλων (**ομάδων**): $d(C_i,C_j) = \min_{\boldsymbol{x}\in Ci,\boldsymbol{y}\in Cj} d(\boldsymbol{x},\boldsymbol{y})$ (ή max ή mean)

ΣΗΜ: Η επιλογή του μέτρου εγγύτητας μπορεί να επηρεάσει το αποτέλεσμα ομαδοποίησης.

Ομαδοποίηση (Clustering) – Βασικές έννοιες

Σύμβαση:

Ομαδοποίηση $\leftrightarrow$ Σύνολο ομάδων

$\mathcal{C}$ = $\{C_1, C_2, \ldots, C_m\}$

Ιεραρχία ομαδοποιήσεων $\leftrightarrow$ Σύνολο ομαδοποιήσεων

$\mathcal{H}$ = $\{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_N\}$

Παράδειγμα:

Σημεία δεδομένων: x_1, x_2, x_3, x_4, x_5

Ομάδες: $C_1=\{x_1, x_3\}$, $C_2=\{x_2, x_4, x_5\}$

Ομαοδοποίηση: $\mathcal{C}=\{C_1, C_2\}$

Ιεραρχία ομαδοποιήσεων: $\mathcal{H}=\{ \{\{x_1\}, \{x_2\}, \{x_3\}, \{x_4\}, \{x_5\}\},$
$\{\{x_1, x_2\}, \{x_3\}, \{x_4\}, \{x_5\}\},$
$\{\{x_1, x_2, x_3\}, \{x_4\}, \{x_5\}\},$
$\{\{x_1, x_2, x_3\}, \{x_4, x_5\}\},$
$\{\{x_1, x_2, x_3, x_4, x_5\}\}$ $\}$

Κατηγοριοποίηση αλγορίθμων ομαδοποίησης

Ιεραρχικοί αλγόριθμοι:

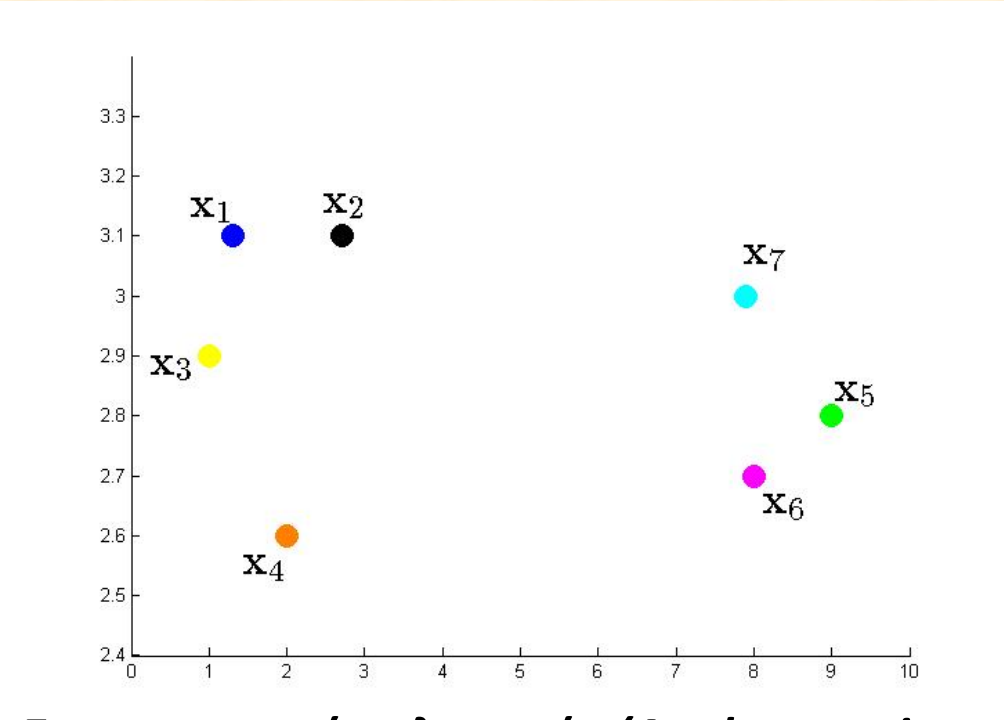
- **Παράγουν** μια ιεραρχία ομαδοποιήσεων των δεδομένων.
- Μπορεί να είναι **είτε** παραμετρικοί **είτε** μη παραμετρικοί.
- Κύριες φιλοσοφίες: συσσωρευτική (agglomerative), διαιρετική (divisive).

Μη ιεραρχικοί αλγόριθμοι:

- **Παράγουν** μια ομαδοποίηση των δεδομένων.
- **Βασίζονται** σε ποικιλία ιδεών που προέρχονται από
  - Το πλαίσιο της βελτιστοποίησης συναρτήσεων κόστους (Cost function optimization - **CFO**) framework (κυρίως παραμετρικοί αλγόριθμοι)
    - Πιθανοτικό πλαίσιο (παραμετρικοί αλγόριθμοι)
    - Μη πιθανοτικό πλαίσιο (παραμετρικοί αλγόριθμοι)
  - Το Πλαίσιο θεωρίας γράφων (π.χ. spectral clustering, ελάχιστο δένδρο κάλυψης) (κυρίως μη παραμετρικοί αλγόριθμοι)
  - Πυκνότητα ομάδων (Cluster-density) (μη παραμετρικοί αλγόριθμοι)

Ιεραρχικοί αλγόριθμοι ομαδοποίησης

$\mathbf{x}_1$ $\mathbf{x}_2$ $\mathbf{x}_3$ $\mathbf{x}_4$ $\mathbf{x}_5$ $\mathbf{x}_6$ $\mathbf{x}_7$



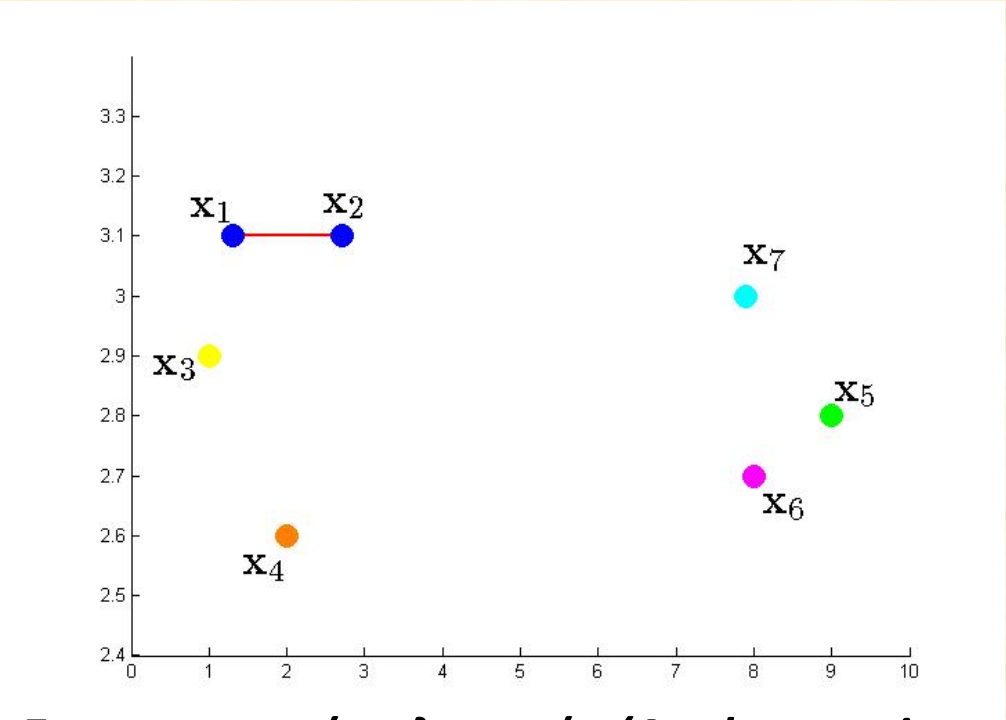
Συσσωρευτική φιλοσοφία (Agglomerative philosophy):

- Στην αρχική ομαδοποίηση όλα τα διανύσματα δεδομένων **ανήκουν** σε διαφορετικές ομάδες.
- Σε κάθε βήμα ορίζεται μια νέα ομαδοποίηση **συνενώνοντας** σε ένα τις δύο όμοιες μεταξύ τους ομάδες.
- Στην τελική ομαδοποίηση όλα τα διανύσματα **ανήκουν** στην ίδια ομάδα.

Ιεραρχικοί αλγόριθμοι ομαδοποίησης

$\mathbf{x}_1$ $\mathbf{x}_2$ $\mathbf{x}_3$ $\mathbf{x}_4$ $\mathbf{x}_5$ $\mathbf{x}_6$ $\mathbf{x}_7$

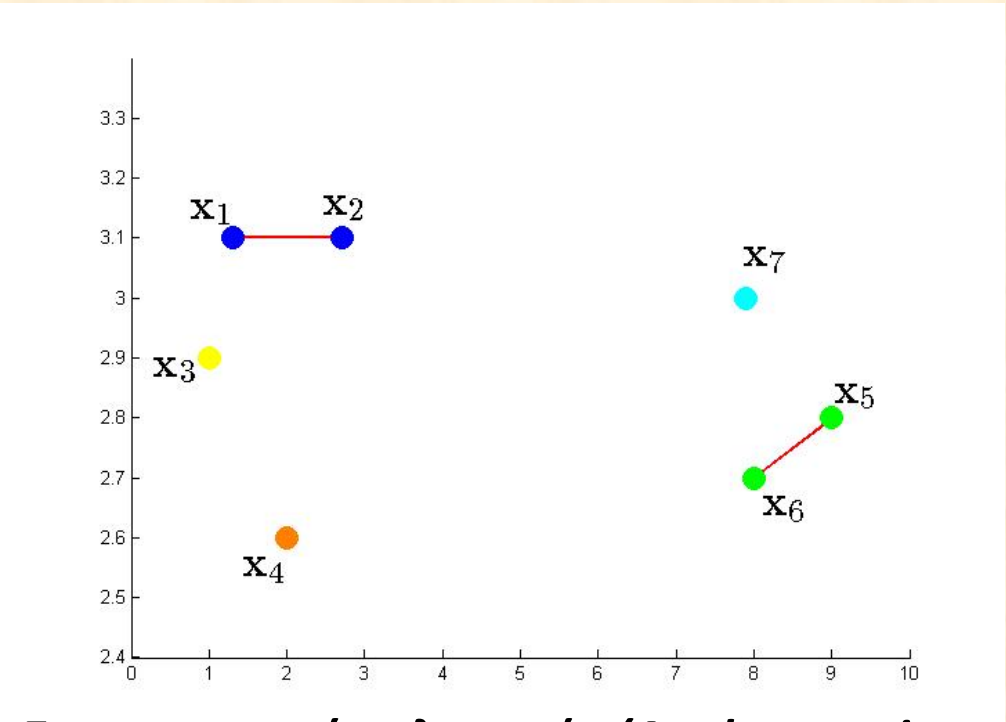
1.4



Συσσωρευτική φιλοσοφία (Agglomerative philosophy):

- Στην αρχική ομαδοποίηση όλα τα διανύσματα δεδομένων **ανήκουν** σε διαφορετικές ομάδες.
- Σε κάθε βήμα ορίζεται μια νέα ομαδοποίηση **συνενώνοντας** σε ένα τις δύο όμοιες μεταξύ τους ομάδες.
- Στην τελική ομαδοποίηση όλα τα διανύσματα **ανήκουν** στην ίδια ομάδα.

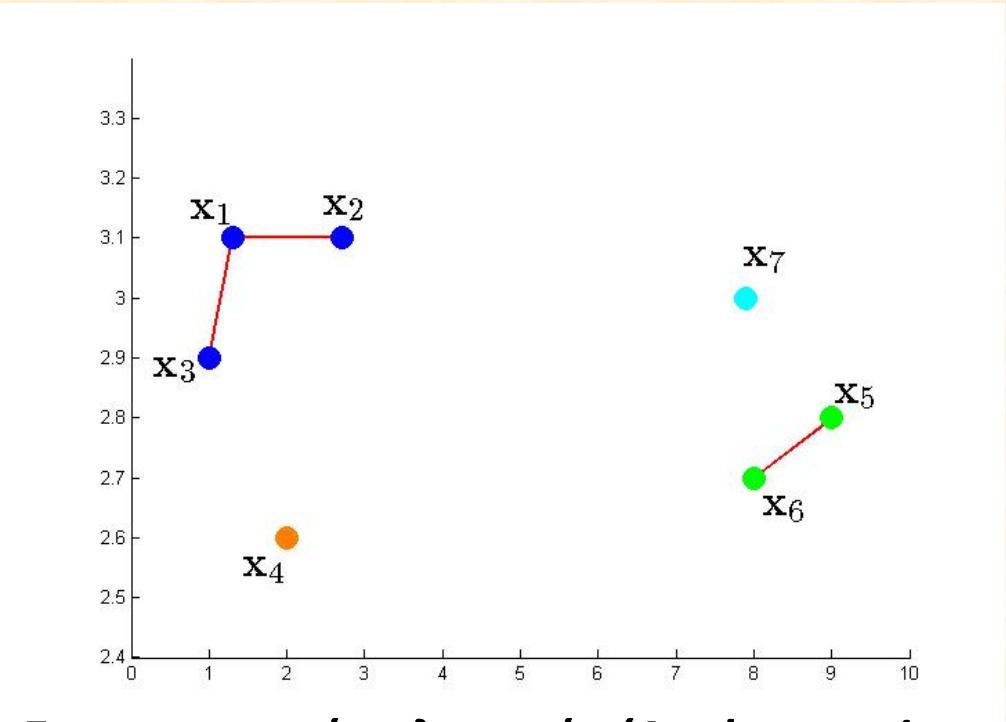
Ιεραρχικοί αλγόριθμοι ομαδοποίησης

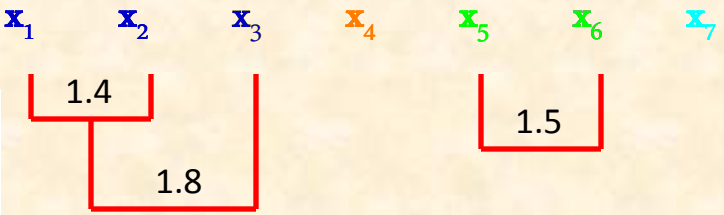


<u>Συσσωρευτική φιλοσοφία (Agglomerative philosophy):</u>

- Στην αρχική ομαδοποίηση όλα τα διανύσματα δεδομένων **ανήκουν** σε διαφορετικές ομάδες.
- Σε κάθε βήμα ορίζεται μια νέα ομαδοποίηση **συνενώνοντας** σε ένα τις δύο όμοιες μεταξύ τους ομάδες.
- Στην τελική ομαδοποίηση όλα τα διανύσματα **ανήκουν** στην ίδια ομάδα.

Ιεραρχικοί αλγόριθμοι ομαδοποίησης

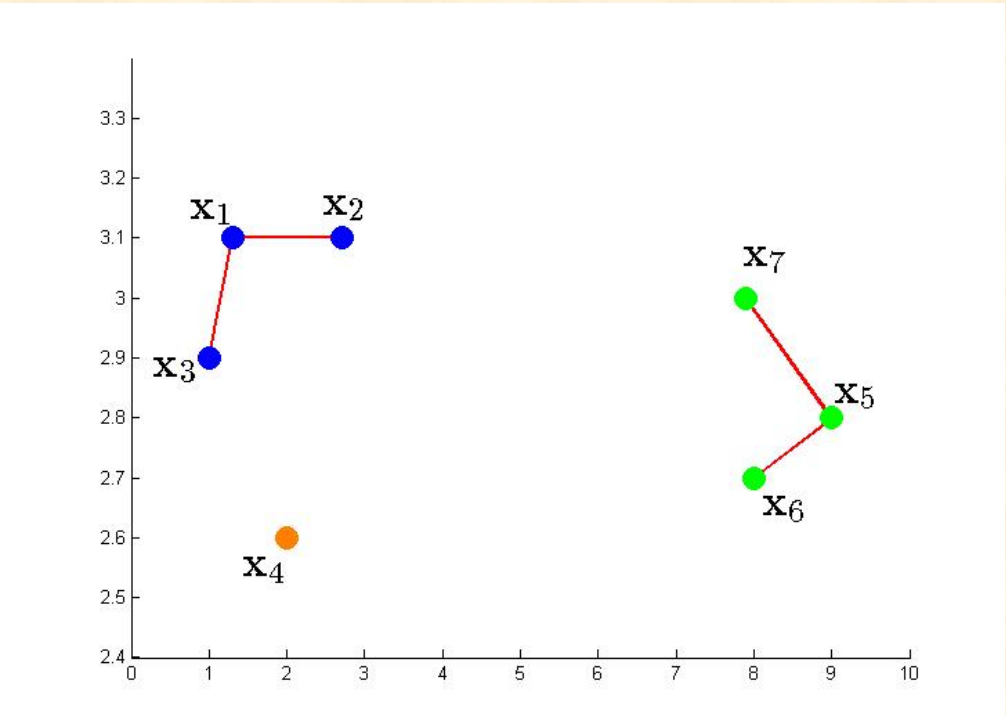


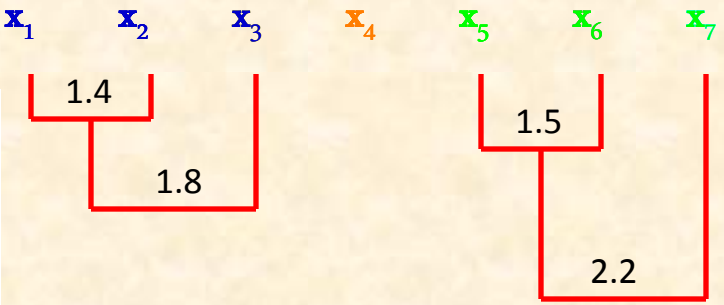


<u>Συσσωρευτική φιλοσοφία (Agglomerative philosophy):</u>

- Στην αρχική ομαδοποίηση όλα τα διανύσματα δεδομένων **ανήκουν** σε διαφορετικές ομάδες.
- Σε κάθε βήμα ορίζεται μια νέα ομαδοποίηση **συνενώνοντας** σε ένα τις δύο όμοιες μεταξύ τους ομάδες.
- Στην τελική ομαδοποίηση όλα τα διανύσματα **ανήκουν** στην ίδια ομάδα.

Ιεραρχικοί αλγόριθμοι ομαδοποίησης

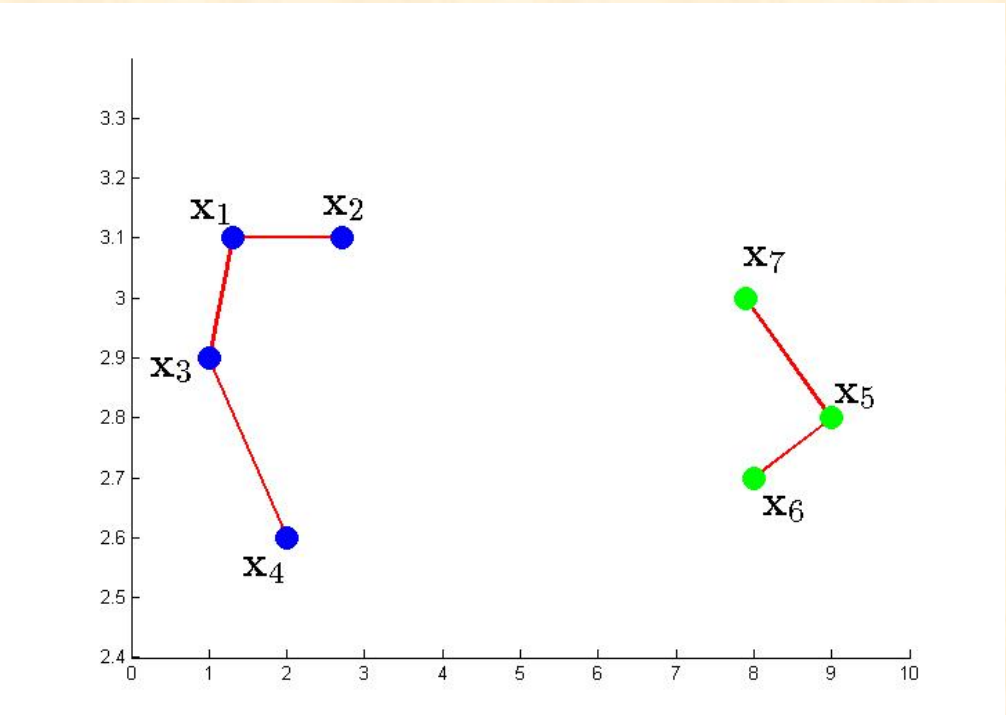


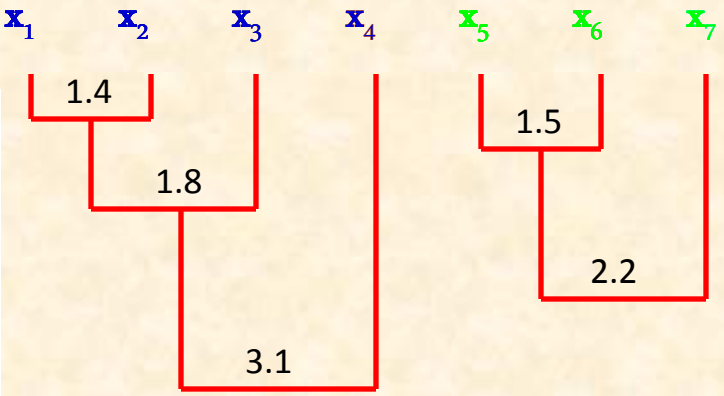


Συσσωρευτική φιλοσοφία (Agglomerative philosophy):

- Στην αρχική ομαδοποίηση όλα τα διανύσματα δεδομένων **ανήκουν** σε διαφορετικές ομάδες.
- Σε κάθε βήμα ορίζεται μια νέα ομαδοποίηση **συνενώνοντας** σε ένα τις δύο όμοιες μεταξύ τους ομάδες.
- Στην τελική ομαδοποίηση όλα τα διανύσματα **ανήκουν** στην ίδια ομάδα.

Ιεραρχικοί αλγόριθμοι ομαδοποίησης

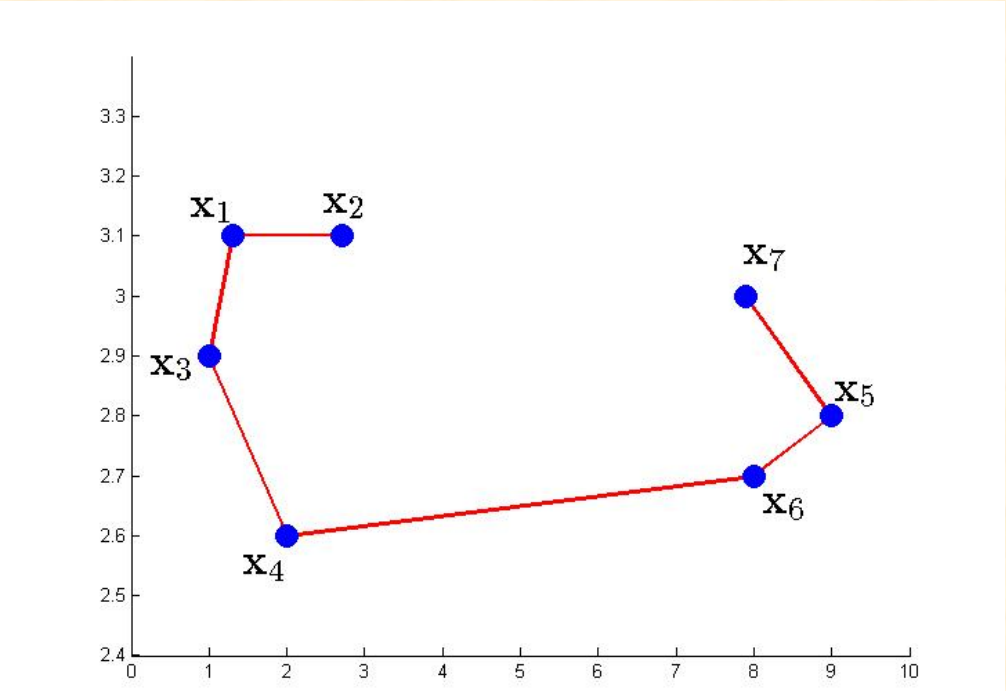


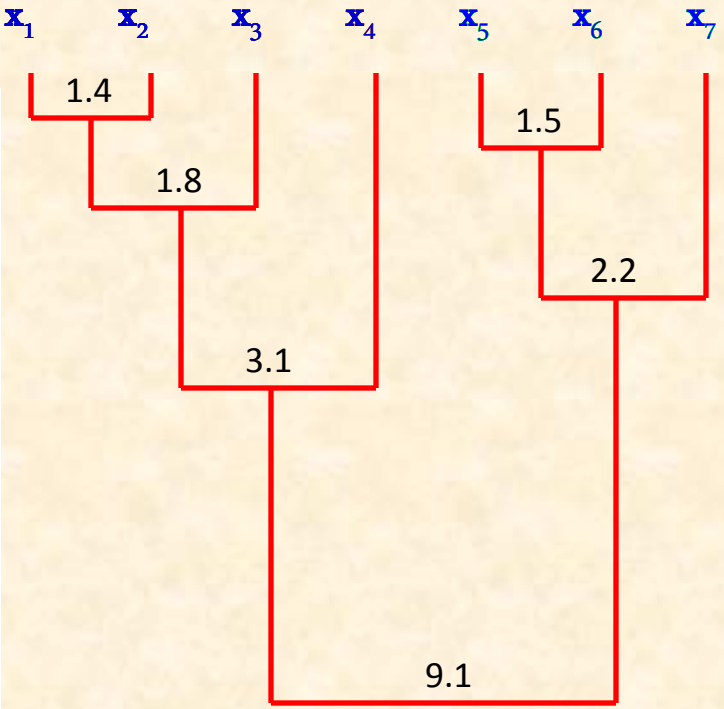


<u>Συσσωρευτική φιλοσοφία (Agglomerative philosophy):</u>

- Στην αρχική ομαδοποίηση όλα τα διανύσματα δεδομένων **ανήκουν** σε διαφορετικές ομάδες.
- Σε κάθε βήμα ορίζεται μια νέα ομαδοποίηση **συνενώνοντας** σε ένα τις δύο όμοιες μεταξύ τους ομάδες.
- Στην τελική ομαδοποίηση όλα τα διανύσματα **ανήκουν** στην ίδια ομάδα.

Ιεραρχικοί αλγόριθμοι ομαδοποίησης

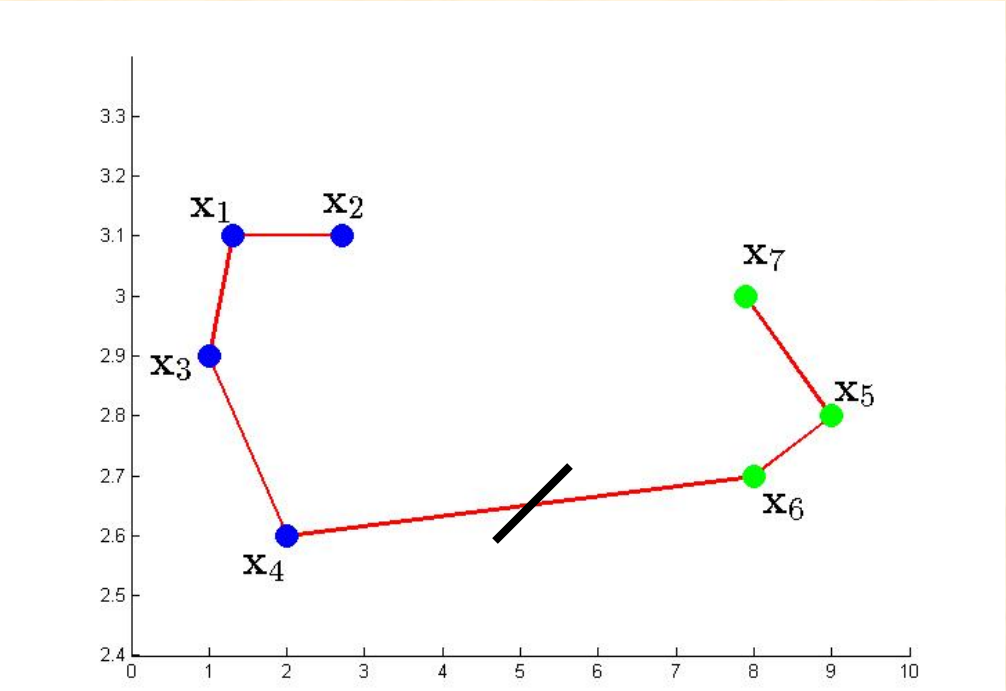


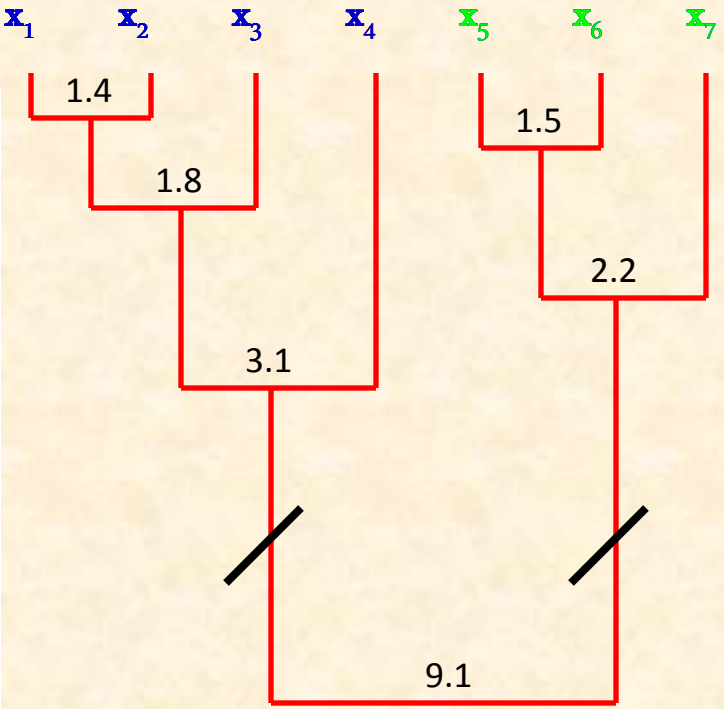


Συσσωρευτική φιλοσοφία (Agglomerative philosophy):

- Στην αρχική ομαδοποίηση όλα τα διανύσματα δεδομένων **ανήκουν** σε διαφορετικές ομάδες.
- Σε κάθε βήμα ορίζεται μια νέα ομαδοποίηση **συνενώνοντας** σε ένα τις δύο όμοιες μεταξύ τους ομάδες.
- Στην τελική ομαδοποίηση όλα τα διανύσματα **ανήκουν** στην ίδια ομάδα.

Ιεραρχικοί αλγόριθμοι ομαδοποίησης

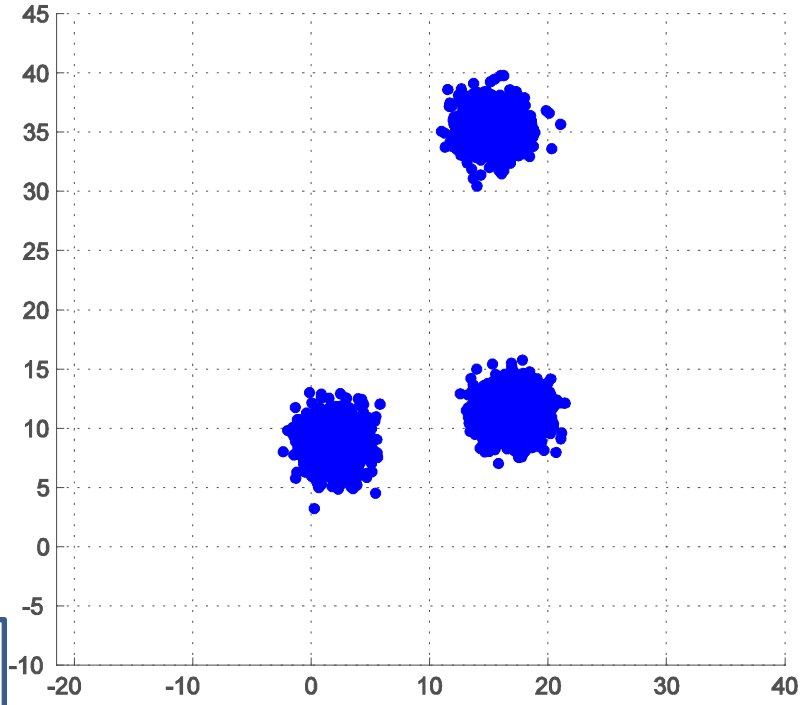




<u>Συσσωρευτική φιλοσοφία (Agglomerative philosophy):</u>

- Στην αρχική ομαδοποίηση όλα τα διανύσματα δεδομένων **ανήκουν** σε διαφορετικές ομάδες.
- Σε κάθε βήμα ορίζεται μια νέα ομαδοποίηση **συνενώνοντας** σε ένα τις δύο όμοιες μεταξύ τους ομάδες.
- Στην τελική ομαδοποίηση όλα τα διανύσματα **ανήκουν** στην ίδια ομάδα.

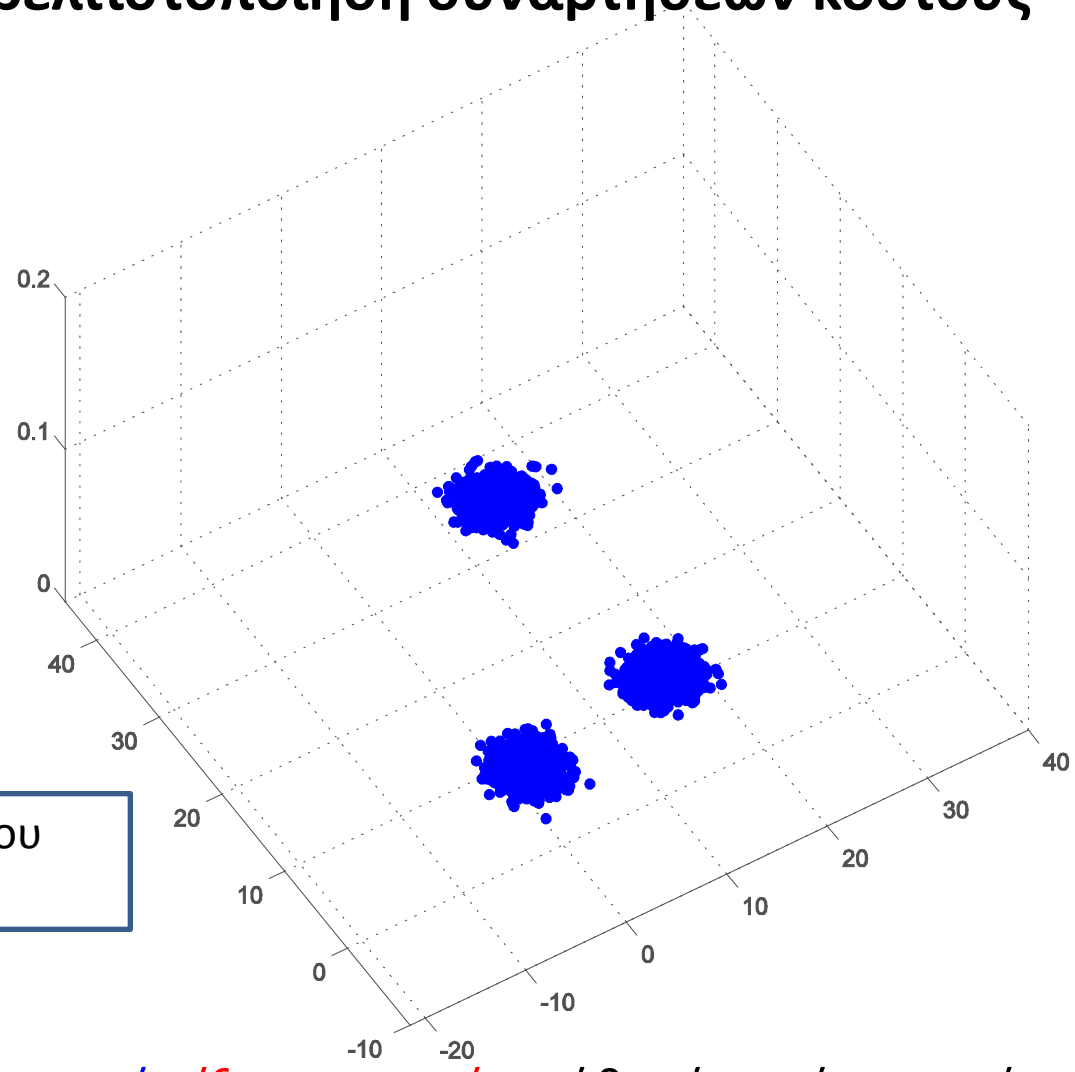
Πιθανοτικοί αλγόριθμοι ομαδοποίησης βασιζόμενοι στην βελτιστοποίηση συναρτήσεων κόστους



Προαπαίτηση: Γνώση του αριθμού των ομάδων.

- **Υιοθέτησε** μια παραμετρική μίξη κατανομών, κάθε μία από τις οποίες αντιστοιχεί σε μια ομάδα (π.χ., μίξη Gaussians), και οι παράμετροί της αρχικοποιούνται τυχαία.
- **Μετακίνησε** προοδευτικά κάθε κατανομή πάνω από μία ομάδα, μέσω της **βελτιστοποίησης** ενός κριτηρίου.

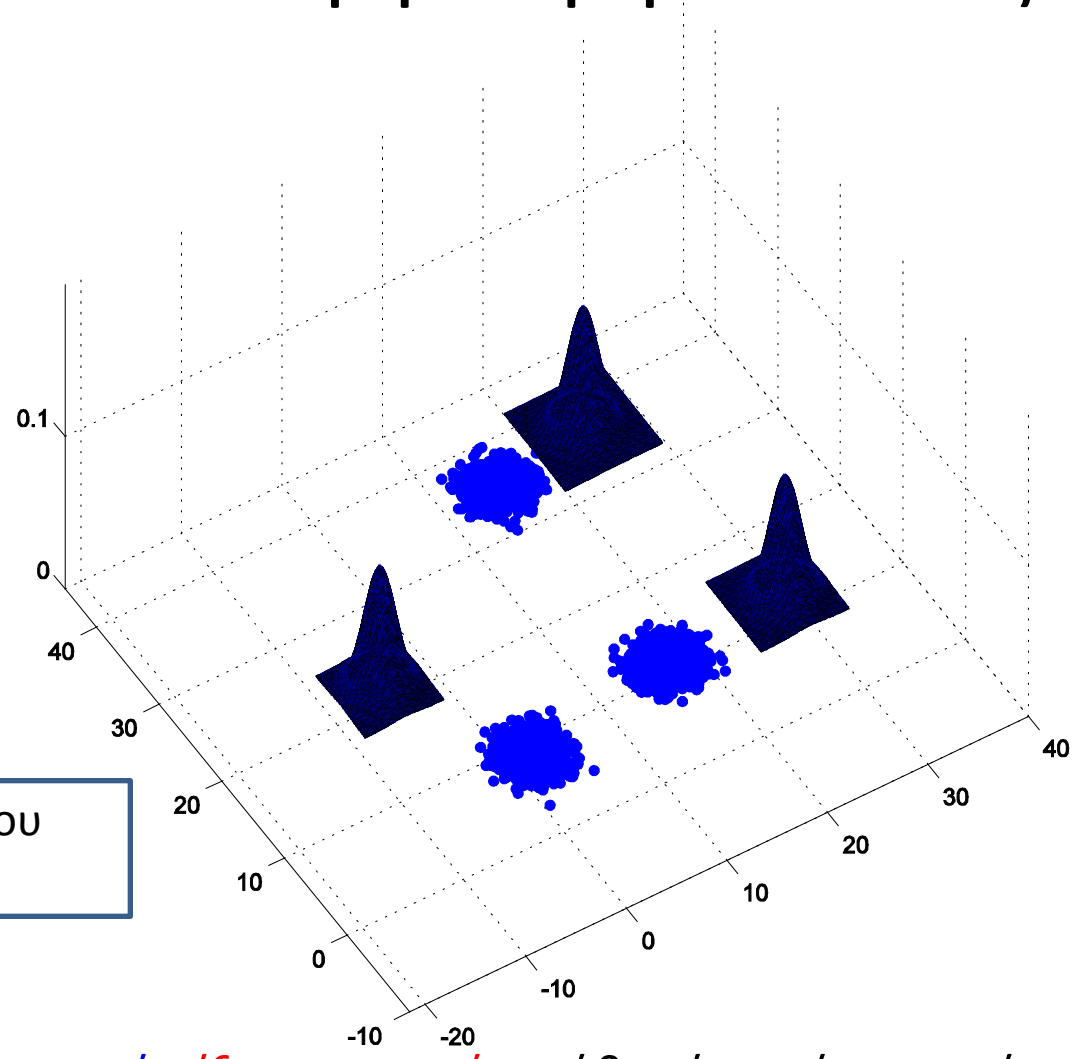
Πιθανοτικοί αλγόριθμοι ομαδοποίησης βασιζόμενοι στην βελτιστοποίηση συναρτήσεων κόστους



Προαπαίτηση: Γνώση του αριθμού των ομάδων.

- **Υιοθέτησε** μια παραμετρική μίξη κατανομών, κάθε μία από τις οποίες αντιστοιχεί σε μια ομάδα (π.χ., μίξη Gaussians), και οι παράμετροί της αρχικοποιούνται τυχαία.
- **Μετακίνησε** προοδευτικά κάθε κατανομή πάνω από μία ομάδα, μέσω της **βελτιστοποίησης** ενός κριτηρίου.

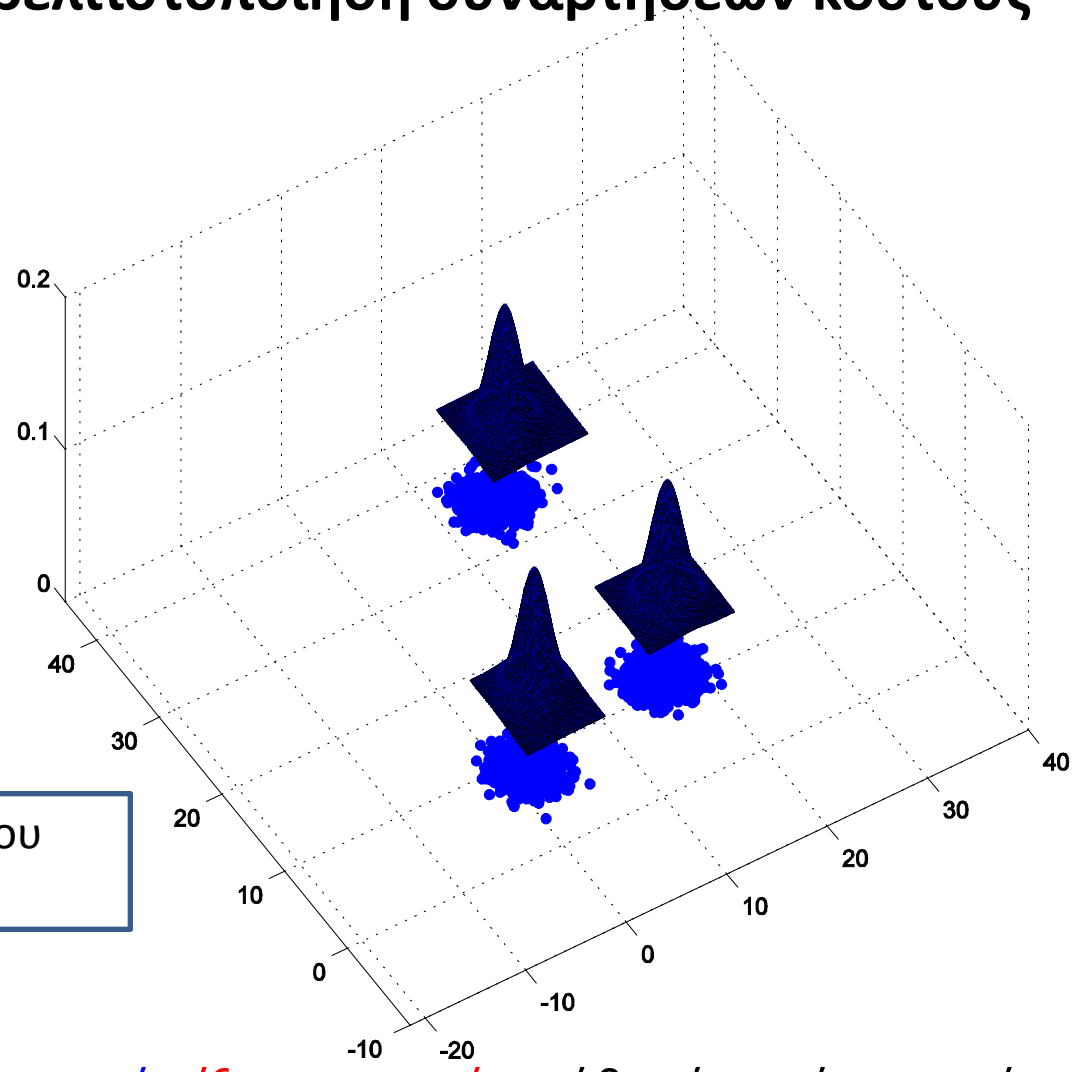
Πιθανοτικοί αλγόριθμοι ομαδοποίησης βασιζόμενοι στην βελτιστοποίηση συναρτήσεων κόστους



Προαπαίτηση: Γνώση του αριθμού των ομάδων.

- **Υιοθέτησε** μια παραμετρική μίξη κατανομών, κάθε μία από τις οποίες αντιστοιχεί σε μια ομάδα (π.χ., μίξη Gaussians), και οι παράμετροί της αρχικοποιούνται τυχαία.
- **Μετακίνησε** προοδευτικά κάθε κατανομή πάνω από μία ομάδα, μέσω της **βελτιστοποίησης** ενός κριτηρίου.

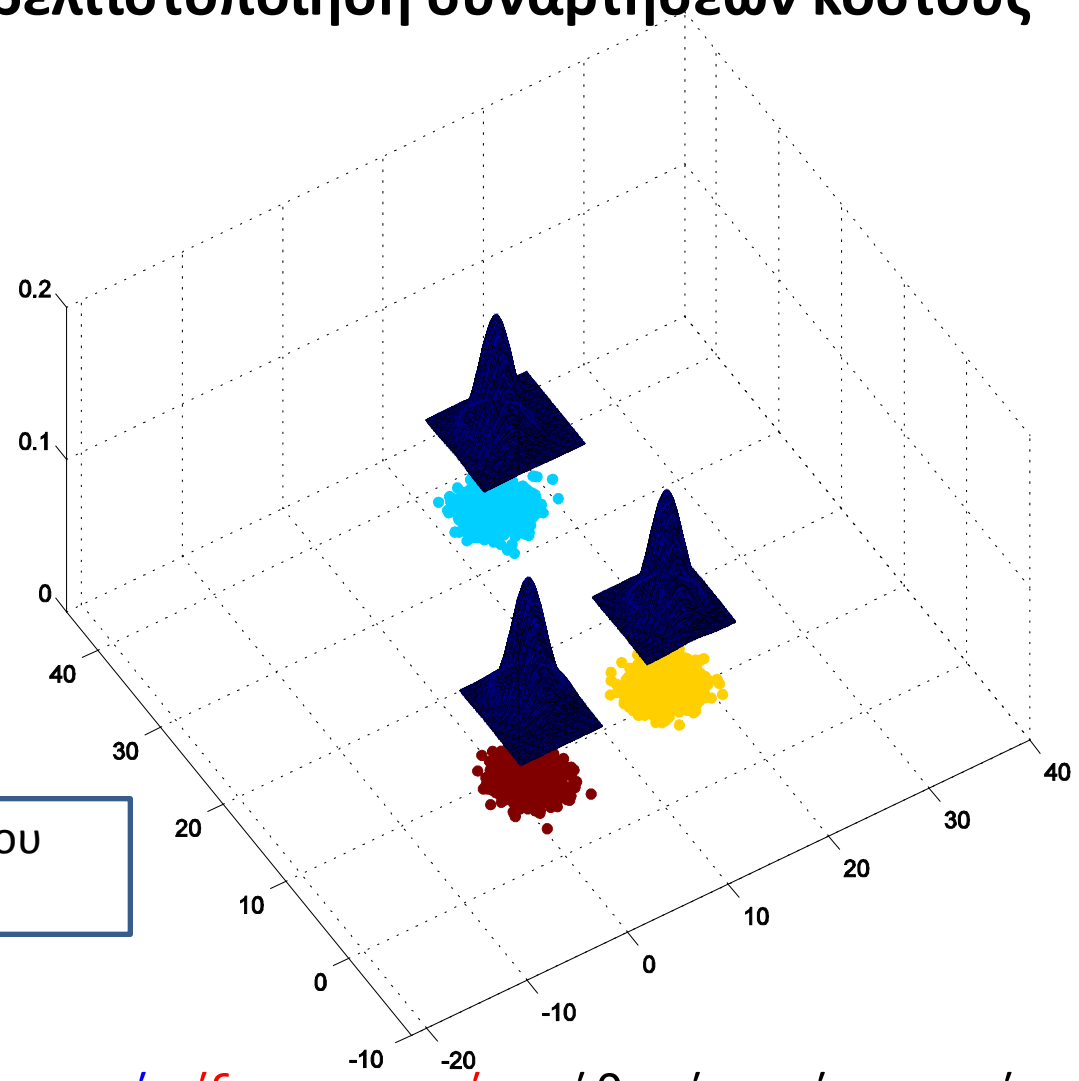
Πιθανοτικοί αλγόριθμοι ομαδοποίησης βασιζόμενοι στην βελτιστοποίηση συναρτήσεων κόστους



Προαπαίτηση: Γνώση του αριθμού των ομάδων.

- **Υιοθέτησε** μια παραμετρική μίξη κατανομών, κάθε μία από τις οποίες αντιστοιχεί σε μια ομάδα (π.χ., μίξη Gaussians), και οι παράμετροί της αρχικοποιούνται τυχαία.
- **Μετακίνησε** προοδευτικά κάθε κατανομή πάνω από μία ομάδα, μέσω της **βελτιστοποίησης** ενός κριτηρίου.

Πιθανοτικοί αλγόριθμοι ομαδοποίησης βασιζόμενοι στην βελτιστοποίηση συναρτήσεων κόστους



Προαπαίτηση: Γνώση του αριθμού των ομάδων.

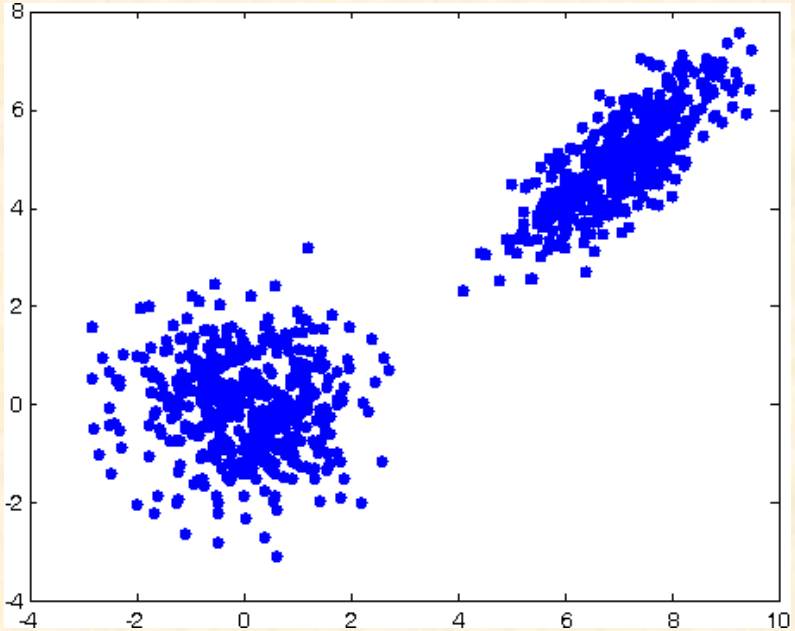
- **Υιοθέτησε** μια παραμετρική μίξη κατανομών, κάθε μία από τις οποίες αντιστοιχεί σε μια ομάδα (π.χ., μίξη Gaussians), και οι παράμετροί της αρχικοποιούνται τυχαία.
- **Μετακίνησε** προοδευτικά κάθε κατανομή πάνω από μία ομάδα, μέσω της **βελτιστοποίησης** ενός κριτηρίου.

Πιθανοτικοί αλγόριθμοι ομαδοποίησης βασιζόμενοι στην βελτιστοποίηση συναρτήσεων κόστους

Ο αλγόριθμος Αναμονής-Μεγιστοποίησης (Expectation – Maximization - EM) (μοντέλα μίξης - mixture models)

Μοντέλο μίξης: Σταθμισμένο άθροισμα pdfs γνωστής παραμετρικής μορφής

$$p(x)=\sum_{j=1}^{m} P_j\, p(x \mid j), \quad \sum_{j=1}^{m} P_j = 1, \quad \int_{-\infty}^{+\infty} p(x \mid j) = 1$$



Παρατηρήσεις:

- Απαιτείται εκ των προτέρων γνώση του αριθμού των ομάδων, *m*.
- Κάθε μία pdf αντιστοιχεί και σε μία φυσική ομάδα.

Συμβολισμός:

$p(\boldsymbol{x}|j) = p(\boldsymbol{x}|j;\boldsymbol{\theta}_j)$

$\boldsymbol{\Theta}=(\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_m)$, $\boldsymbol{P}=[P_1,\ldots,P_m]^T$

$X=\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N\}$: ελλιπή (παρατηρούμενα) δεδομένα.

$X^c=\{(\boldsymbol{x}_1,j_1),\ldots,(\boldsymbol{x}_N,j_N)\}$: πλήρη δεδομένα

Πιθανοτικοί αλγόριθμοι ομαδοποίησης βασιζόμενοι στην βελτιστοποίηση συναρτήσεων κόστους

Ο ΕΜ αλγόριθμος (μίξη μοντέλων)

Στόχος:

Εκτίμηση των ***Θ*** και ***P*** μέσω μεγιστοποίησης της αναμενόμενης τιμής της συνάρτησης log-likelihood **δεσμευμένης** στα παρατηρούμενα δεδομένα.

$$\ln p(X;\Theta,P)=\sum_{n=1}^{N}\sum_{j=1}^{m}P(j\mid \boldsymbol{x}_n;\boldsymbol{\theta}_j)\ln p(\boldsymbol{x}_n,j;\boldsymbol{\theta}_j)=$$

$$\sum_{n=1}^{N}\sum_{j=1}^{m}P(j\mid \boldsymbol{x}_n;\boldsymbol{\theta}_j)\ln\left(p(\boldsymbol{x}_n\mid j;\boldsymbol{\theta}_j)P_j\right)$$

Μίξη κανονικών κατανομών

$$p(\boldsymbol{x})=\sum_{j=1}^{m}P_j\,p(\boldsymbol{x}\mid j;\boldsymbol{\mu}_j,\Sigma_j),$$

$$p(\boldsymbol{x}\mid j;\boldsymbol{\mu}_j,\Sigma_j)=\frac{1}{(2\pi)^{l/2}\mid\Sigma_j\mid^{1/2}}\exp\left(-0.5\cdot(\boldsymbol{x}-\boldsymbol{\mu}_j)^T\Sigma_j^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_j)\right)$$

Πιθανοτικοί αλγόριθμοι ομαδοποίησης βασιζόμενοι στην βελτιστοποίηση συναρτήσεων κόστους

Ο **ΕΜ** αλγόριθμος για την περίπτωση της **μίξης κανονικών κατανομών**.

- Αρχικοποίηση: $\boldsymbol{\mu}_j=\boldsymbol{\mu}_j^{(0)}$, $\Sigma_j=\Sigma_j^{(0)}$, $\boldsymbol{P}=\boldsymbol{P}^{(0)}$
- t=0
- Επανέλαβε

Expectation step

$$P\left(j \mid \boldsymbol{x}_n;\Theta^{(t)},P^{(t)}\right)=\frac{p\left(\boldsymbol{x}_n \mid j;\boldsymbol{\theta}_k^{(t)}\right)P_j^{(t)}}{\sum_{q=1}^{m} p\left(\boldsymbol{x}_n \mid q;\boldsymbol{\theta}_q^{(t)}\right)P_q^{(t)}} \equiv \gamma_{jn}^{(t)}$$

Maximization step

➢ $t=t+1$

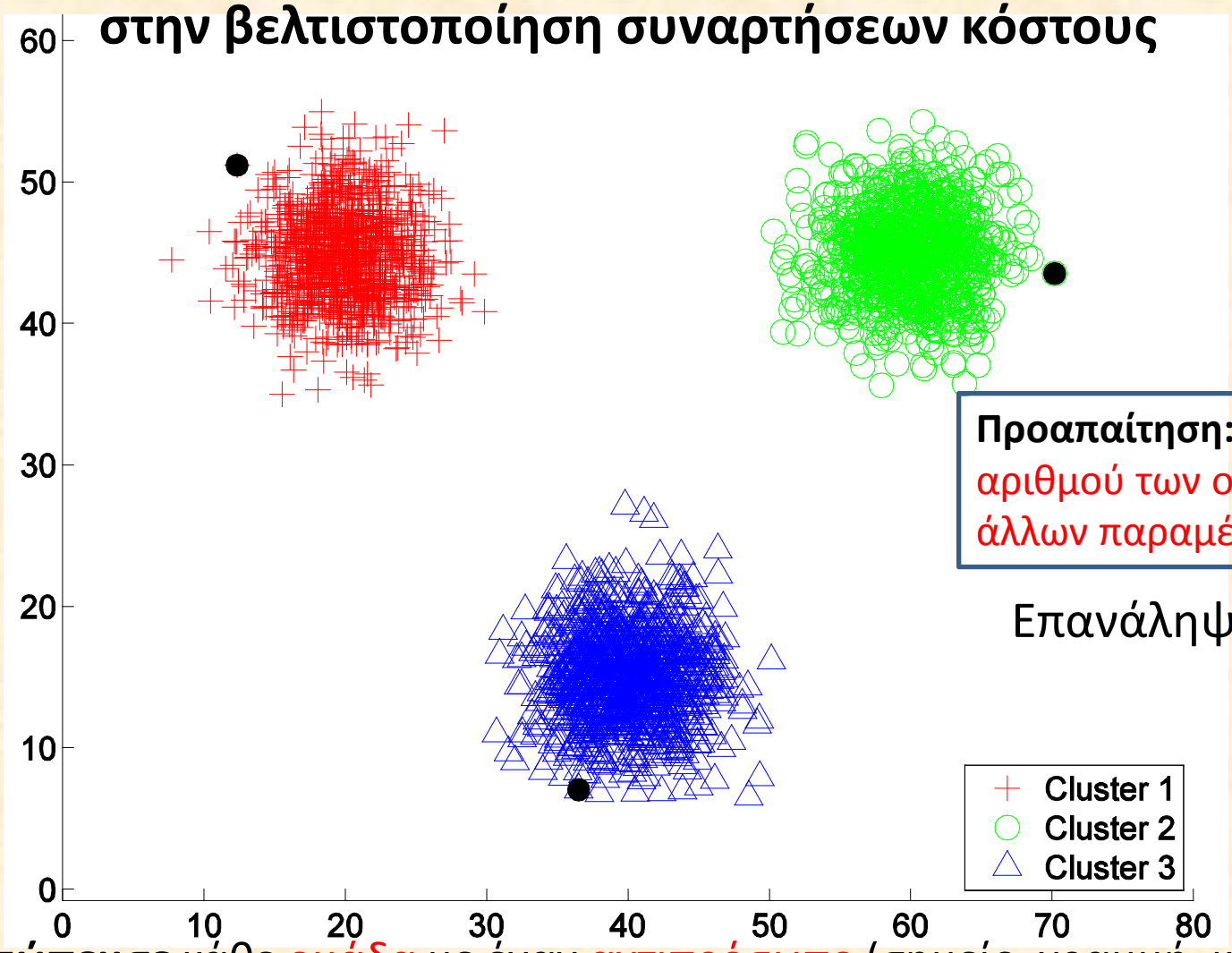
$$\boldsymbol{\mu}_j^{(t+1)}=\sum_{n=1}^{N}\gamma_{jn}^{(t)}\boldsymbol{x}_n \Big/ \sum_{n=1}^{N}\gamma_{jn}^{(t)}$$

$$\Sigma_j^{(t+1)}=\sum_{n=1}^{N}\gamma_{jn}^{(t)}(\boldsymbol{x}_n-\boldsymbol{\mu}_j^{(t+1)})(\boldsymbol{x}_n-\boldsymbol{\mu}_j^{(t+1)})^T \Big/ \sum_{n=1}^{N}\gamma_{jn}^{(t)}$$

$$P_j^{(t+1)}=\frac{1}{N}\sum_{n=1}^{N}\gamma_{jn}^{(t)}$$

- Έως ότου ικανοποιηθεί ένα κατάλληλο κριτήριο

Μη πιθανοτικοί αλγόριθμοι ομαδοποίησης βασιζόμενοι στην βελτιστοποίηση συναρτήσεων κόστους

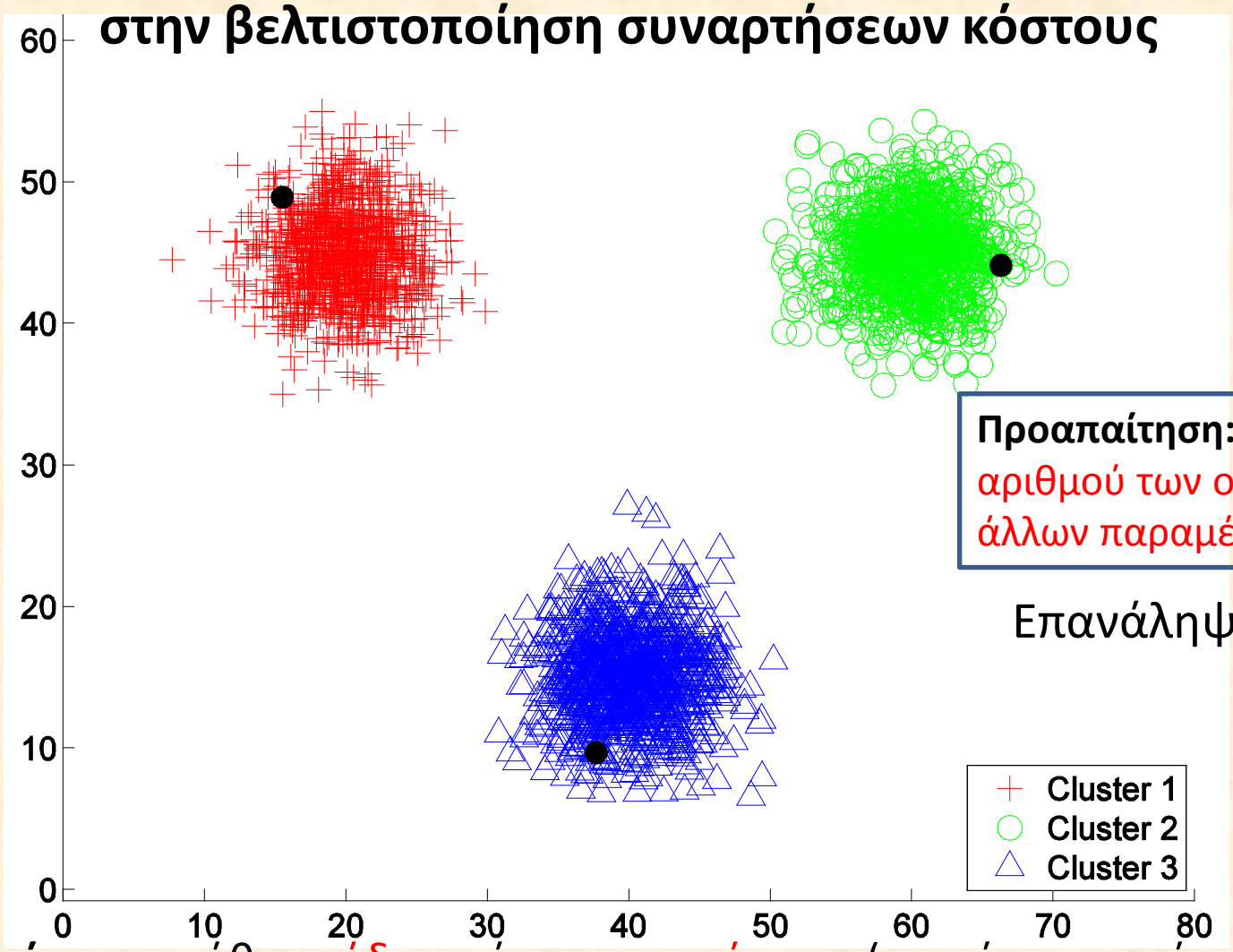


Προαπαίτηση: Γνώση του αριθμού των ομάδων ή άλλων παραμέτρων.

Επανάληψη 1

- **Αντιπροσώπευσε** κάθε ομάδα με έναν αντιπρόσωπο (σημείο, γραμμή, κλπ)
- **Αρχικοποίησε** τους αντιπροσώπους σε τυχαίες θέσεις
- **Μετακίνησε** προοδευτικά τους αντιπροσώπους προς τις φυσικές ομάδες που σχηματίζουν τα δεδομένα, **βελτιστοποιώντας** ένα κριτήριο.

Μη πιθανοτικοί αλγόριθμοι ομαδοποίησης βασιζόμενοι στην βελτιστοποίηση συναρτήσεων κόστους

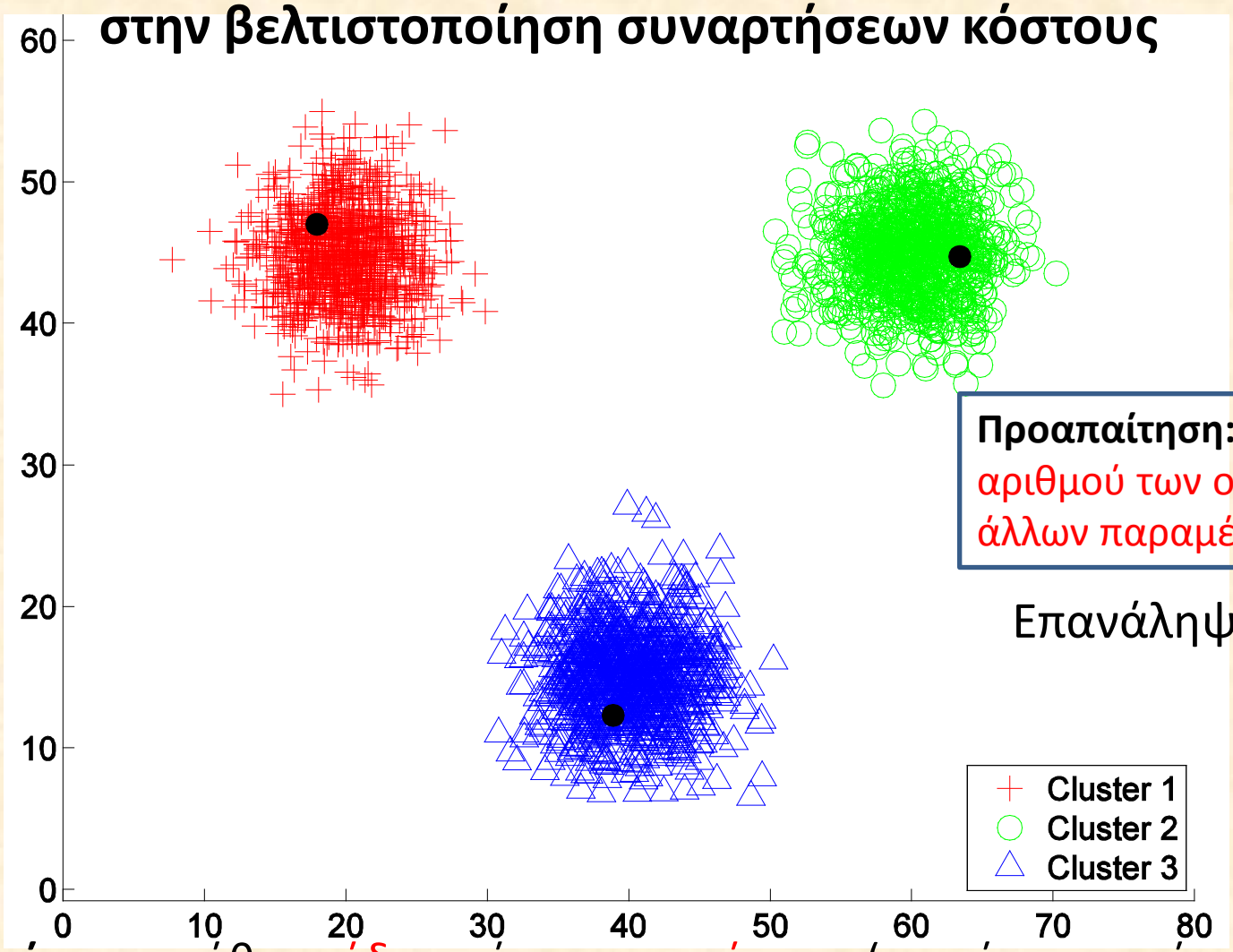


Προαπαίτηση: Γνώση του αριθμού των ομάδων ή άλλων παραμέτρων.

Επανάληψη 2

- **Αντιπροσώπευσε** κάθε ομάδα με έναν αντιπρόσωπο (σημείο, γραμμή, κλπ)
- **Αρχικοποίησε** τους αντιπροσώπους σε τυχαίες θέσεις
- **Μετακίνησε** προοδευτικά τους αντιπροσώπους προς τις φυσικές ομάδες που σχηματίζουν τα δεδομένα, **βελτιστοποιώντας** ένα κριτήριο.

Μη πιθανοτικοί αλγόριθμοι ομαδοποίησης βασιζόμενοι στην βελτιστοποίηση συναρτήσεων κόστους

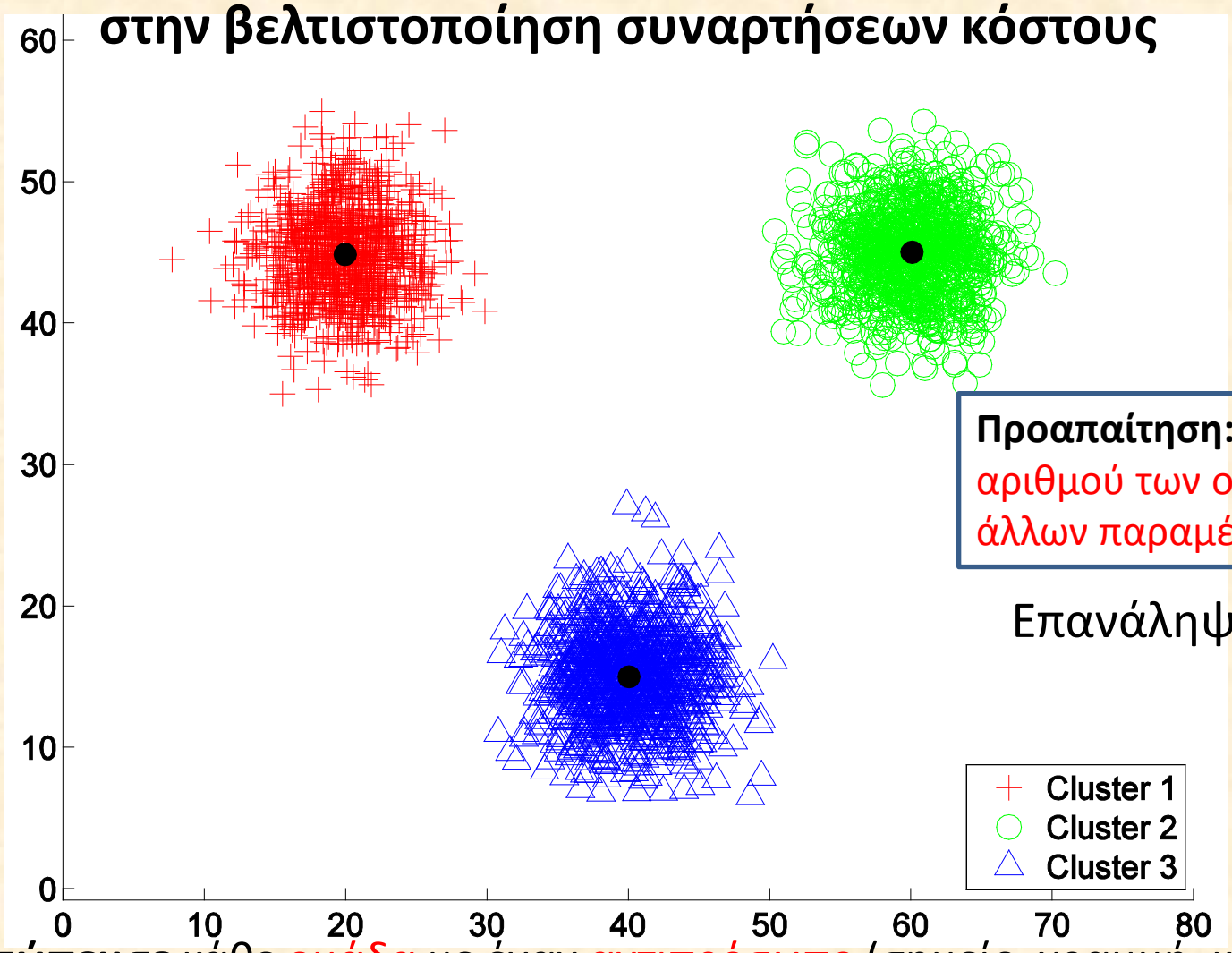


Προαπαίτηση: Γνώση του αριθμού των ομάδων ή άλλων παραμέτρων.

Επανάληψη 3

- **Αντιπροσώπευσε** κάθε ομάδα με έναν αντιπρόσωπο (σημείο, γραμμή, κλπ)
- **Αρχικοποίησε** τους αντιπροσώπους σε τυχαίες θέσεις
- **Μετακίνησε** προοδευτικά τους αντιπροσώπους προς τις φυσικές ομάδες που σχηματίζουν τα δεδομένα, **βελτιστοποιώντας** ένα κριτήριο.

Μη πιθανοτικοί αλγόριθμοι ομαδοποίησης βασιζόμενοι στην βελτιστοποίηση συναρτήσεων κόστους



Προαπαίτηση: Γνώση του αριθμού των ομάδων ή άλλων παραμέτρων.

Επανάληψη 4 (τελική)

- **Αντιπροσώπευσε** κάθε ομάδα με έναν αντιπρόσωπο (σημείο, γραμμή, κλπ)
- **Αρχικοποίησε** τους αντιπροσώπους σε τυχαίες θέσεις
- **Μετακίνησε** προοδευτικά τους αντιπροσώπους προς τις φυσικές ομάδες που σχηματίζουν τα δεδομένα, **βελτιστοποιώντας** ένα κριτήριο.

Μη πιθανοτικοί αλγόριθμοι ομαδοποίησης βασιζόμενοι στην βελτιστοποίηση συναρτήσεων κόστους – Ο αλγ. k-means

- Πραγματοποιεί επεξεργασία κατά δέσμες και επιστρέφει μια ομαδοποίηση
- Είναι αλγόριθμος αυστηρής (hard) ομαδοποίησης, που χρησιμοποιεί σημεία αντιπροσώπους (θ_j) για την αντιπροσώπευση των ομάδων (C_j).
- Προκύπτει από τη βελτιστοποίηση της ακόλουθης συνάρτησης κόστους

$$J(\Theta, U) = \sum_{i=1}^{N} \sum_{j=1}^{m} u_{ij} \| \boldsymbol{x}_i - \boldsymbol{\theta}_j \|^2$$

N: αριθμός σημείων
m: αριθμός ομάδων

όπου $\theta = \{\theta_1, \ldots, \theta_m\}$ και $U=[u_{ij}]$, με $u_{ij} = \begin{cases} 1, & \alpha\nu\ \boldsymbol{x}_i \in C_j \\ 0, & \delta\iota\alpha\varphi o\rho\varepsilon\tau\iota\kappa\alpha \end{cases}$

- Είναι επαναληπτικός αλγόριθμος.
- Αρχικά τοποθετεί τους αντιπροσώπους θ_j σε τυχαίες θέσεις στο χώρο.
- Προοδευτικά μετακινεί τους αντιπροσώπους προς τα κέντρα των φυσικών ομάδων που σχηματίζουν τα δεδομένα.
- **Τερματίζει** όταν τα θ_j's δεν αλλάζουν μεταξύ δύο διαδοχικών επαναλήψεων.
- Η υπολογιστική πολυπλοκότητα είναι $O(kN)$ ($k =$ αριθμός επαναλήψεων).
- Απαιτεί εκ των προτέρων γνώση του αριθμού των ομάδων *m*.

Μη πιθανοτικοί αλγόριθμοι ομαδοποίησης βασιζόμενοι στην βελτιστοποίηση συναρτήσεων κόστους – Ο αλγ. k-means

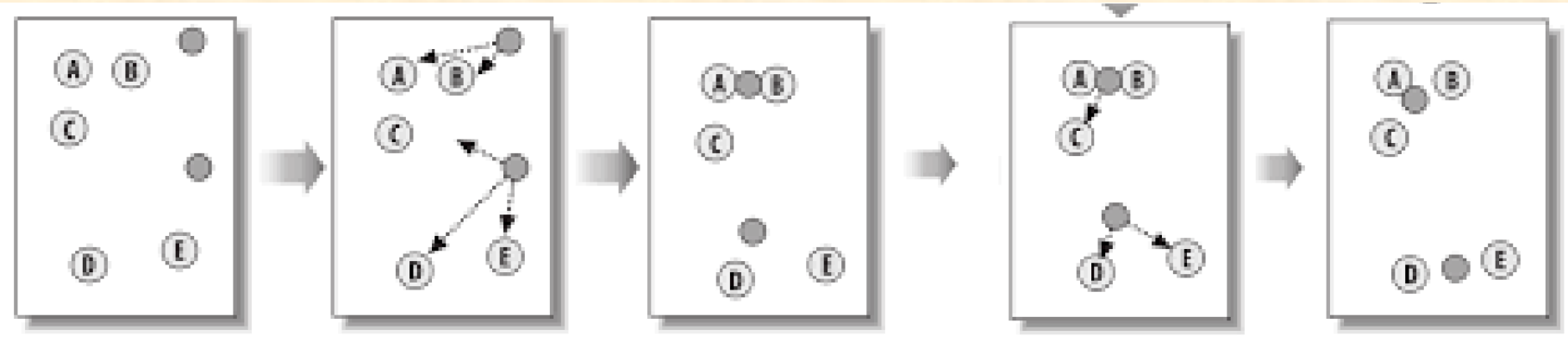
Ο αλγόριθμος k-means

Επέλεξε αυθαίρετες αρχικές εκτιμήσεις $\theta_j(0)$ για τα θ_j', $j=1,\ldots,m$.

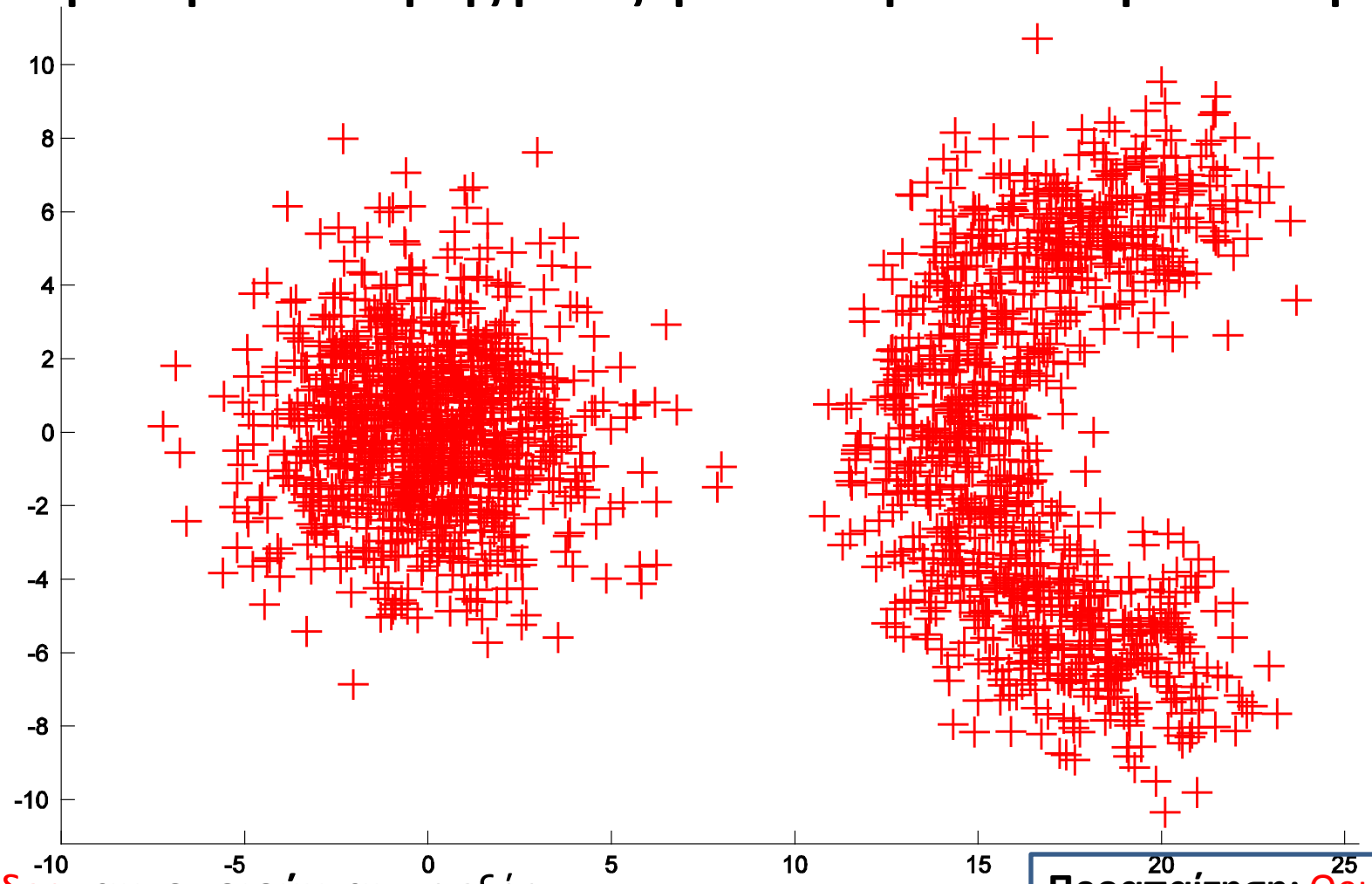
Επανέλαβε

- Για $i=1$ έως N *Προσδιορισμός ομάδων*
  - Προσδιόρισε τον πλησιέστερο αντιπρόσωπο, έστω θ_j, για το x_i
  - Θέσε $u_{ij}=1$ και $u_{iq}=0$, για $q=1,\ldots,m$, $q\neq j$.
- Τέλος {Για}
- Για $j=1$ έως m *Ενημέρωση παραμέτρων*
  - Επανεκτίμησε το θ_j ως το μέσο διάνυσμα των $x_i \in X$ με $u_{ij}=1$.
- Τέλος {Για}

Έως ότου οι τιμές των θ_j παραμείνουν ίδιες μεταξύ δύο διαδοχικών επαναλήψεων



Αλγόριθμοι ομαδοποίησης βασιζόμενοι στην πυκνότητα των ομάδων

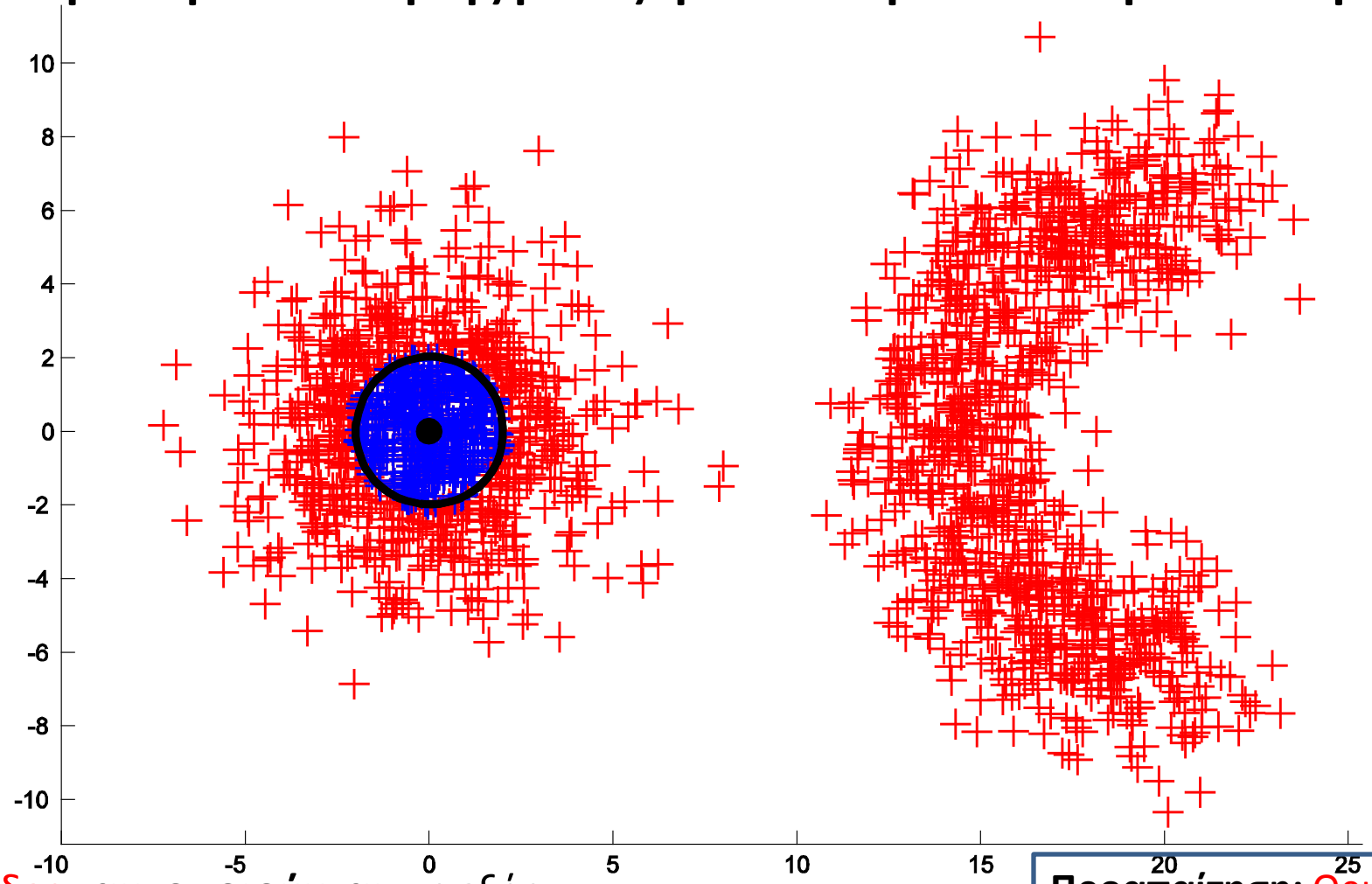


Οι ομάδες **ταυτοποιούνται** ως εξής:

- **Όρισε μια** νέα ομάδα C **επιλέγοντας** ένα σημ. δεδομένων x.
- **Καταχώρησε** όλα τα σημεία που βρίσκονται στη γειτονιά του x στην ίδια ομάδα.
- **Επανέλαβε** αναδρομικά το προηγούμενο βήμα **έως ότου** όλα τα γειτονικά σημεία ΟΛΩΝ των $x \in C$ **καταχωρηθούν** στο C.

Προαπαίτηση: Ορισμός μεγέθους γειτονιάς

Αλγόριθμοι ομαδοποίησης βασιζόμενοι στην πυκνότητα των ομάδων

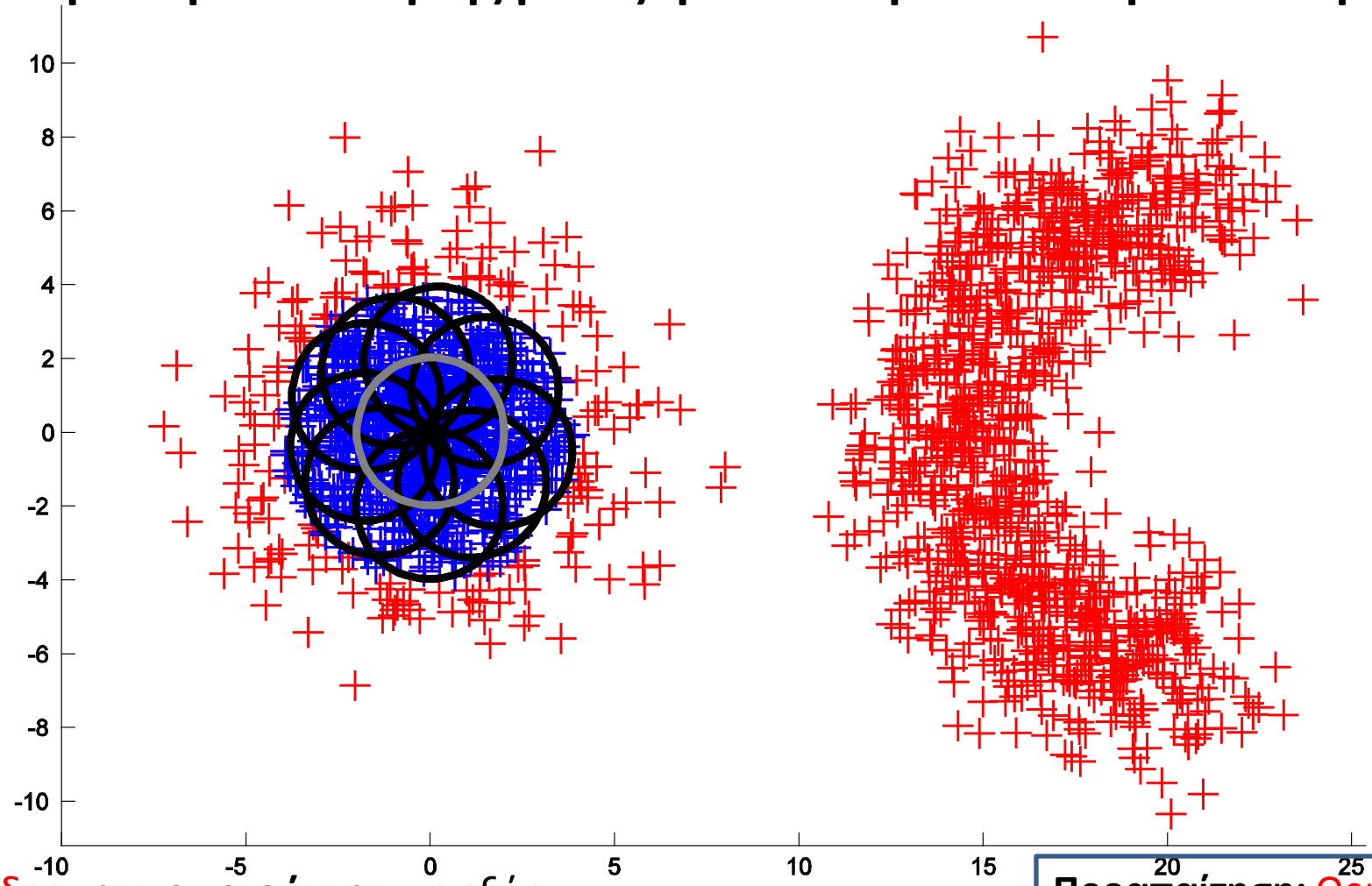


Οι ομάδες **ταυτοποιούνται** ως εξής:

- **Όρισε μια** νέα ομάδα C **επιλέγοντας** ένα σημ. δεδομένων x.
- **Καταχώρησε** όλα τα σημεία που βρίσκονται στη γειτονιά του x στην ίδια ομάδα.
- **Επανέλαβε** αναδρομικά το προηγούμενο βήμα **έως ότου** όλα τα γειτονικά σημεία ΟΛΩΝ των $x \in C$ **καταχωρηθούν** στο C.

Προαπαίτηση: Ορισμός μεγέθους γειτονιάς

Αλγόριθμοι ομαδοποίησης βασιζόμενοι στην πυκνότητα των ομάδων

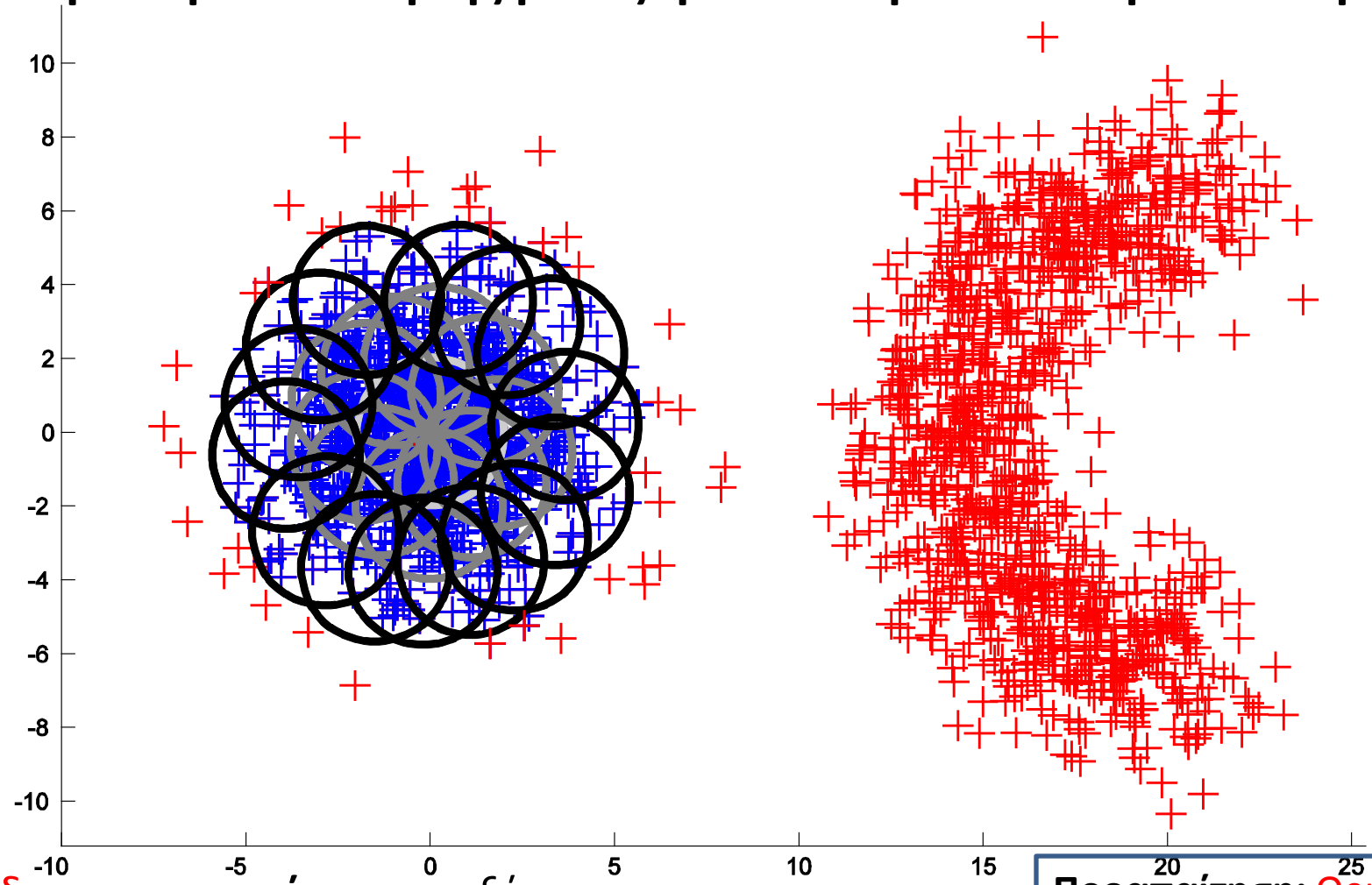


Οι ομάδες **ταυτοποιούνται** ως εξής:

- **Όρισε μια** νέα ομάδα C **επιλέγοντας** ένα σημ. δεδομένων x.
- **Καταχώρησε** όλα τα σημεία που βρίσκονται στη γειτονιά του x στην ίδια ομάδα.
- **Επανέλαβε** αναδρομικά το προηγούμενο βήμα **έως ότου** όλα τα γειτονικά σημεία ΟΛΩΝ των $x \in C$ **καταχωρηθούν** στο C.

Προαπαίτηση: Ορισμός μεγέθους γειτονιάς

Αλγόριθμοι ομαδοποίησης βασιζόμενοι στην πυκνότητα των ομάδων

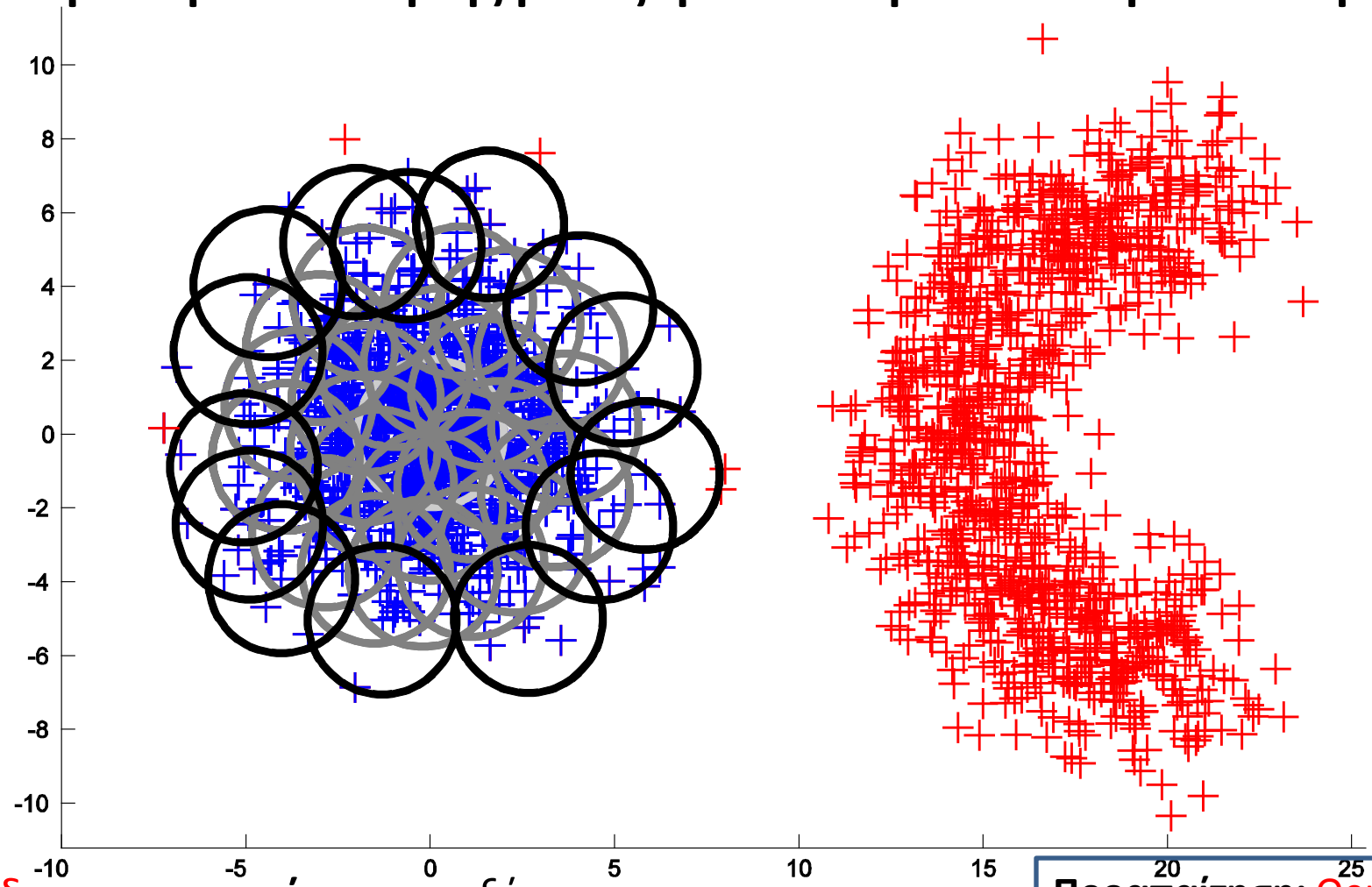


Οι ομάδες **ταυτοποιούνται** ως εξής:

- **Όρισε μια** νέα ομάδα C **επιλέγοντας** ένα σημ. δεδομένων x.
- **Καταχώρησε** όλα τα σημεία που βρίσκονται στη γειτονιά του x στην ίδια ομάδα.
- **Επανέλαβε** αναδρομικά το προηγούμενο βήμα **έως ότου** όλα τα γειτονικά σημεία ΟΛΩΝ των $x \in C$ **καταχωρηθούν** στο C.

Προαπαίτηση: Ορισμός μεγέθους γειτονιάς

Αλγόριθμοι ομαδοποίησης βασιζόμενοι στην πυκνότητα των ομάδων

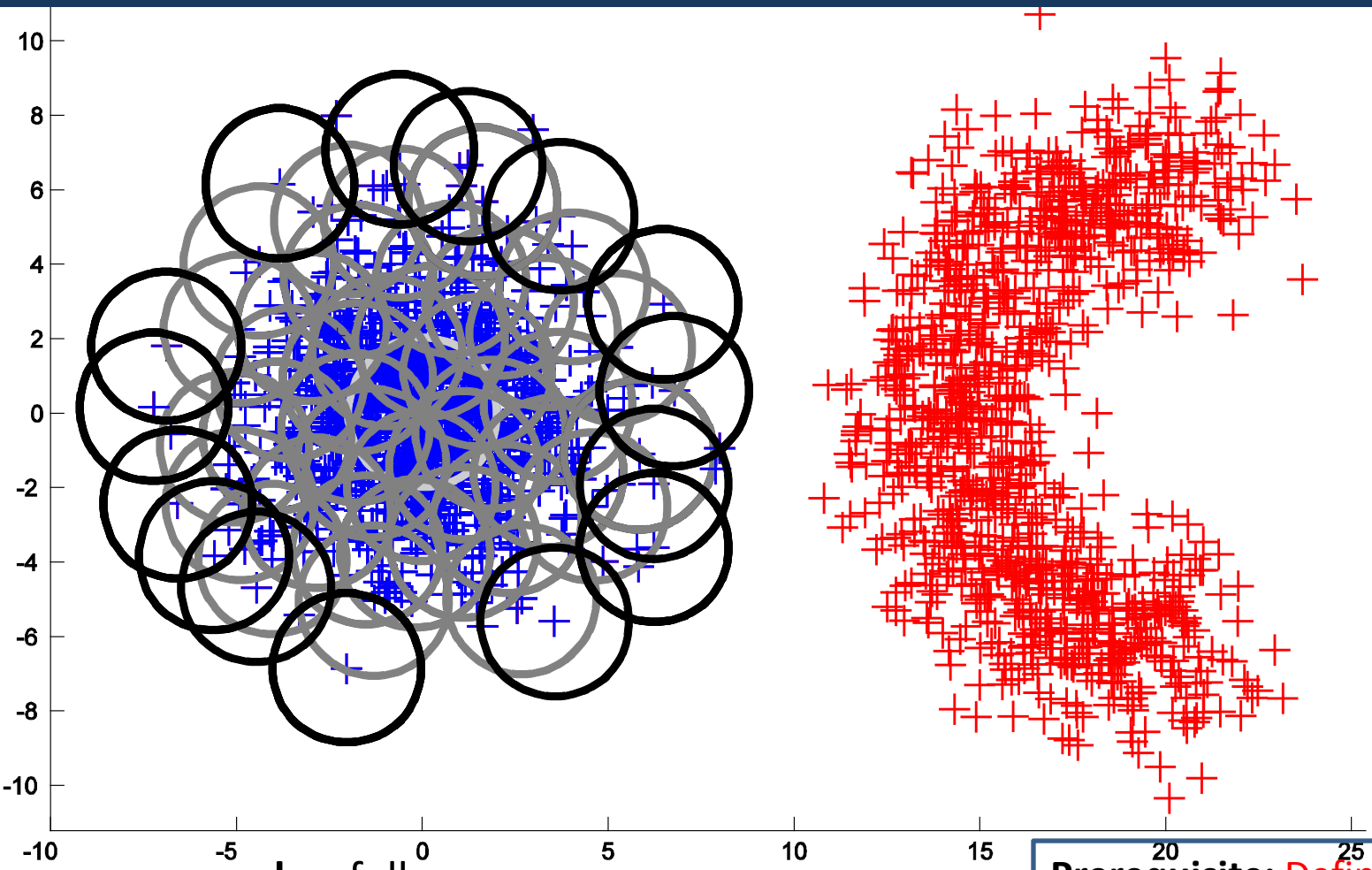


Οι ομάδες **ταυτοποιούνται** ως εξής:

- **Όρισε μια** νέα ομάδα C **επιλέγοντας** ένα σημ. δεδομένων x.
- **Καταχώρησε** όλα τα σημεία που βρίσκονται στη γειτονιά του x στην ίδια ομάδα.
- **Επανέλαβε** αναδρομικά το προηγούμενο βήμα **έως ότου** όλα τα γειτονικά σημεία ΟΛΩΝ των $x \in C$ **καταχωρηθούν** στο C.

Προαπαίτηση: Ορισμός μεγέθους γειτονιάς

Clustering – Density-based algorithms

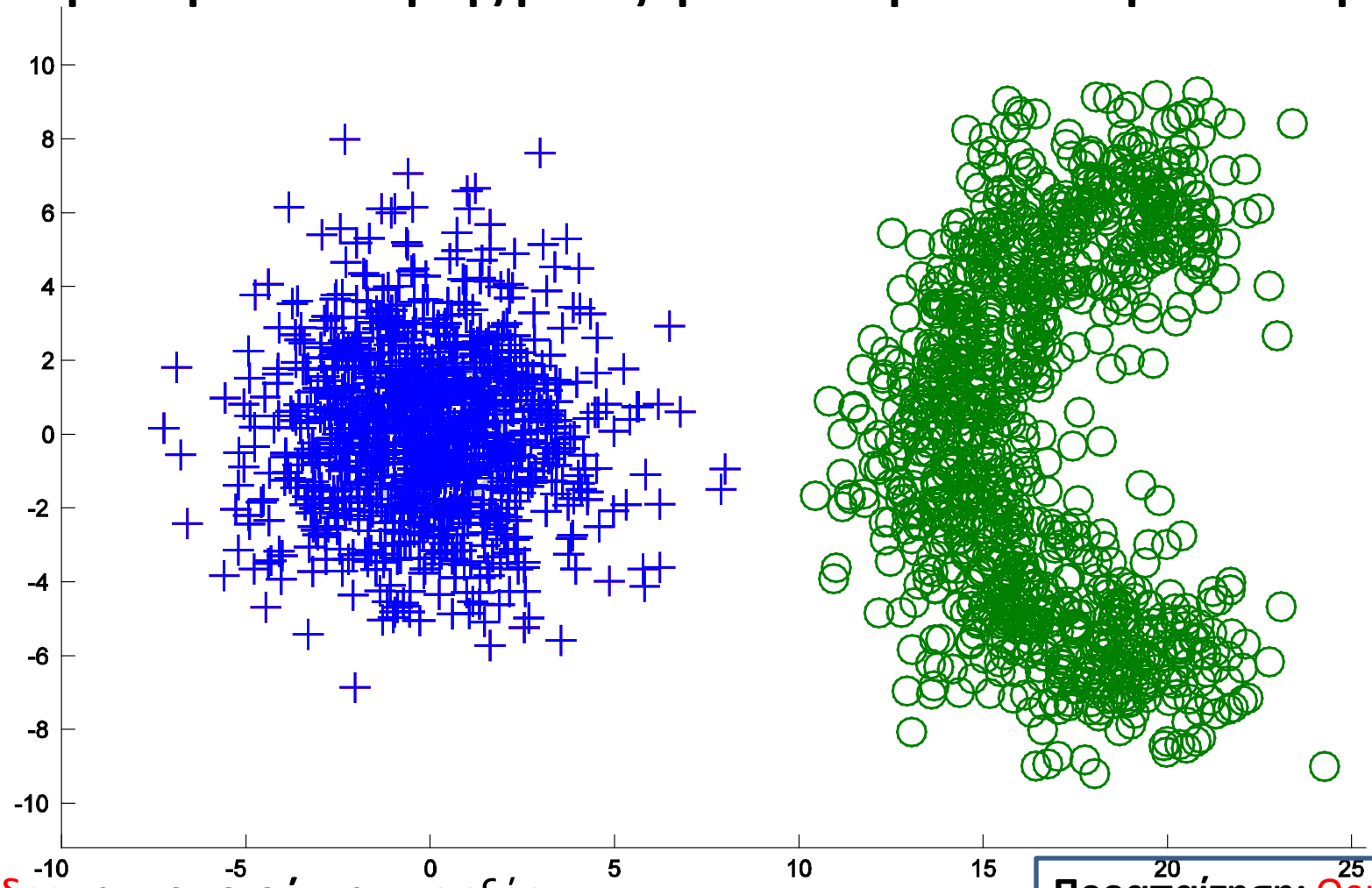


Clusters are **recovered** as follows:

- **Start** a new cluster C by **choosing** a data point x.
- **Assign** all the data points that lie in the neighborhood of x to the same cluster.
- **Repeat** recursively the previous step **until** all neighboring points of ALL $x \in C$ are **assigned** to C.

Prerequisite: Definition of the neighborhood size

Αλγόριθμοι ομαδοποίησης βασιζόμενοι στην πυκνότητα των ομάδων

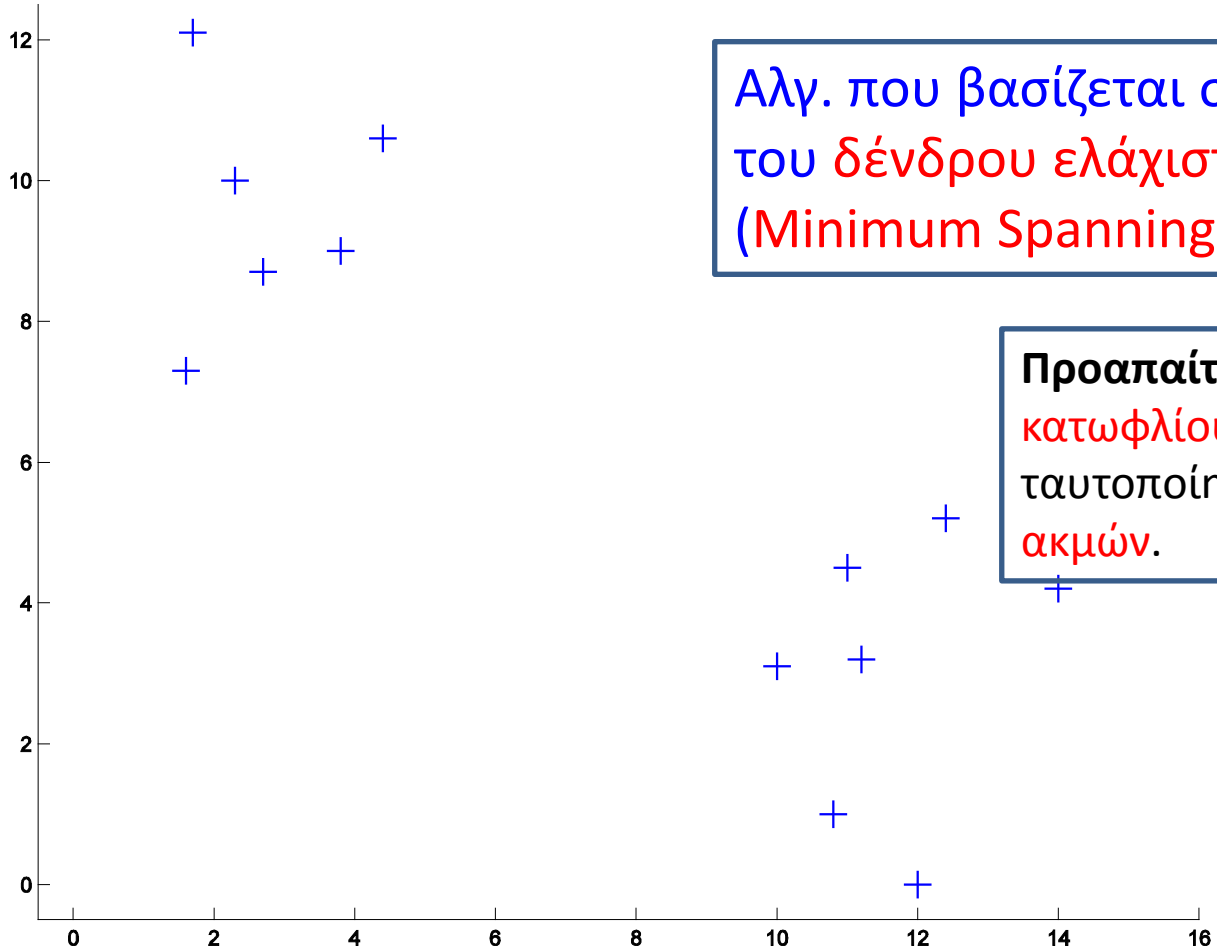


Οι ομάδες **ταυτοποιούνται** ως εξής:

- **Όρισε μια** νέα ομάδα C **επιλέγοντας** ένα σημ. δεδομένων x.
- **Καταχώρησε** όλα τα σημεία που βρίσκονται στη γειτονιά του x στην ίδια ομάδα.
- **Επανέλαβε** αναδρομικά το προηγούμενο βήμα **έως ότου** όλα τα γειτονικά σημεία ΟΛΩΝ των $x \in C$ **καταχωρηθούν** στο C.

Προαπαίτηση: Ορισμός μεγέθους γειτονιάς

Αλγόριθμοι ομαδοποίησης που βασίζονται σε έννοιες γράφων

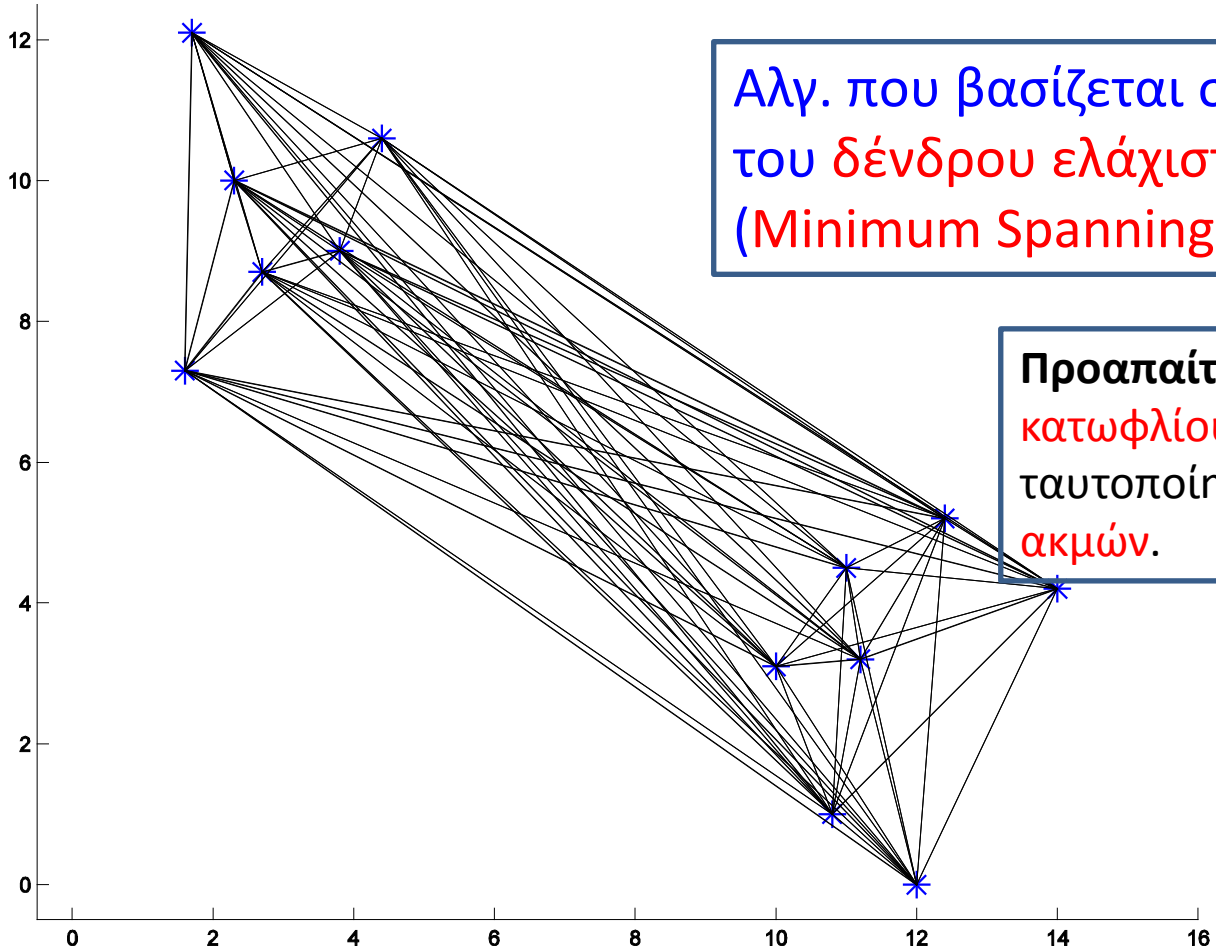


Αλγ. που βασίζεται στην έννοια του δένδρου ελάχιστης κάλυψης (Minimum Spanning Tree - MST).

Προαπαίτηση: Ορισμός κατωφλίου για την ταυτοποίηση "μεγάλων" ακμών.

- **Όρισε** τον **πλήρη** γράφο με κορυφές τα σημεία δεδομένων και ακμές τα τμήματα που συνδέουν κάθε ζεύγος κορυφών.
- **Στάθμισε** κάθε ακμή με την απόσταση μεταξύ των δύο κορυφών που αυτή συνδέει.
- **Όρισε** το MST του γράφου **και αφαίρεσε** τις **"ασυνήθιστα μεγάλες"** ακμές.
- Οι υπό-γράφοι που δημιουργούνται **αντιστοιχούν** στις ομάδες.

Αλγόριθμοι ομαδοποίησης που βασίζονται σε έννοιες γράφων

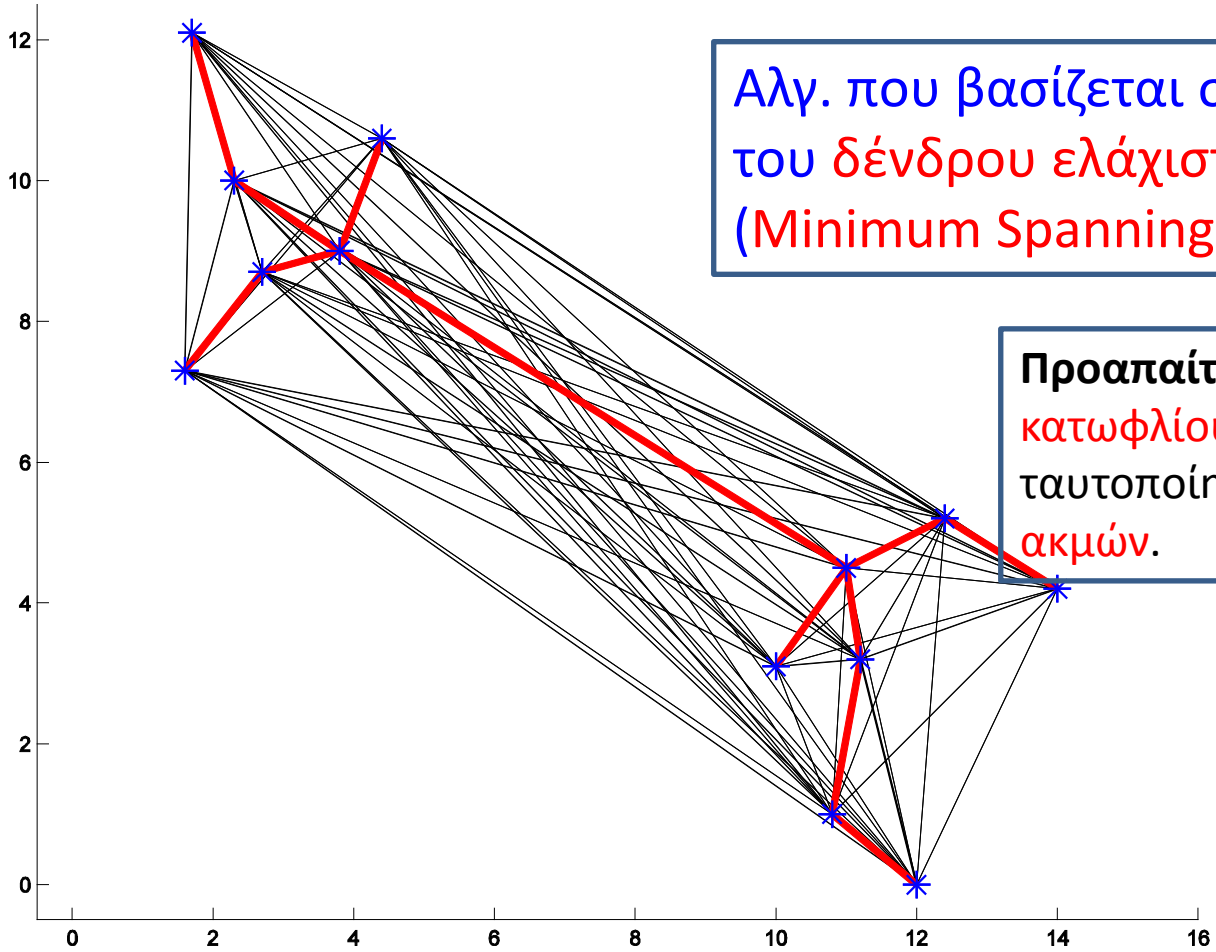


Αλγ. που βασίζεται στην έννοια του δένδρου ελάχιστης κάλυψης (Minimum Spanning Tree - MST).

Προαπαίτηση: Ορισμός κατωφλίου για την ταυτοποίηση "μεγάλων" ακμών.

- **Όρισε** τον **πλήρη** γράφο με κορυφές τα σημεία δεδομένων και ακμές τα τμήματα που συνδέουν κάθε ζεύγος κορυφών.
- **Στάθμισε** κάθε ακμή με την απόσταση μεταξύ των δύο κορυφών που αυτή συνδέει.
- **Όρισε** το MST του γράφου **και αφαίρεσε** τις **"ασυνήθιστα μεγάλες"** ακμές.
- Οι υπό-γράφοι που δημιουργούνται **αντιστοιχούν** στις ομάδες.

Αλγόριθμοι ομαδοποίησης που βασίζονται σε έννοιες γράφων

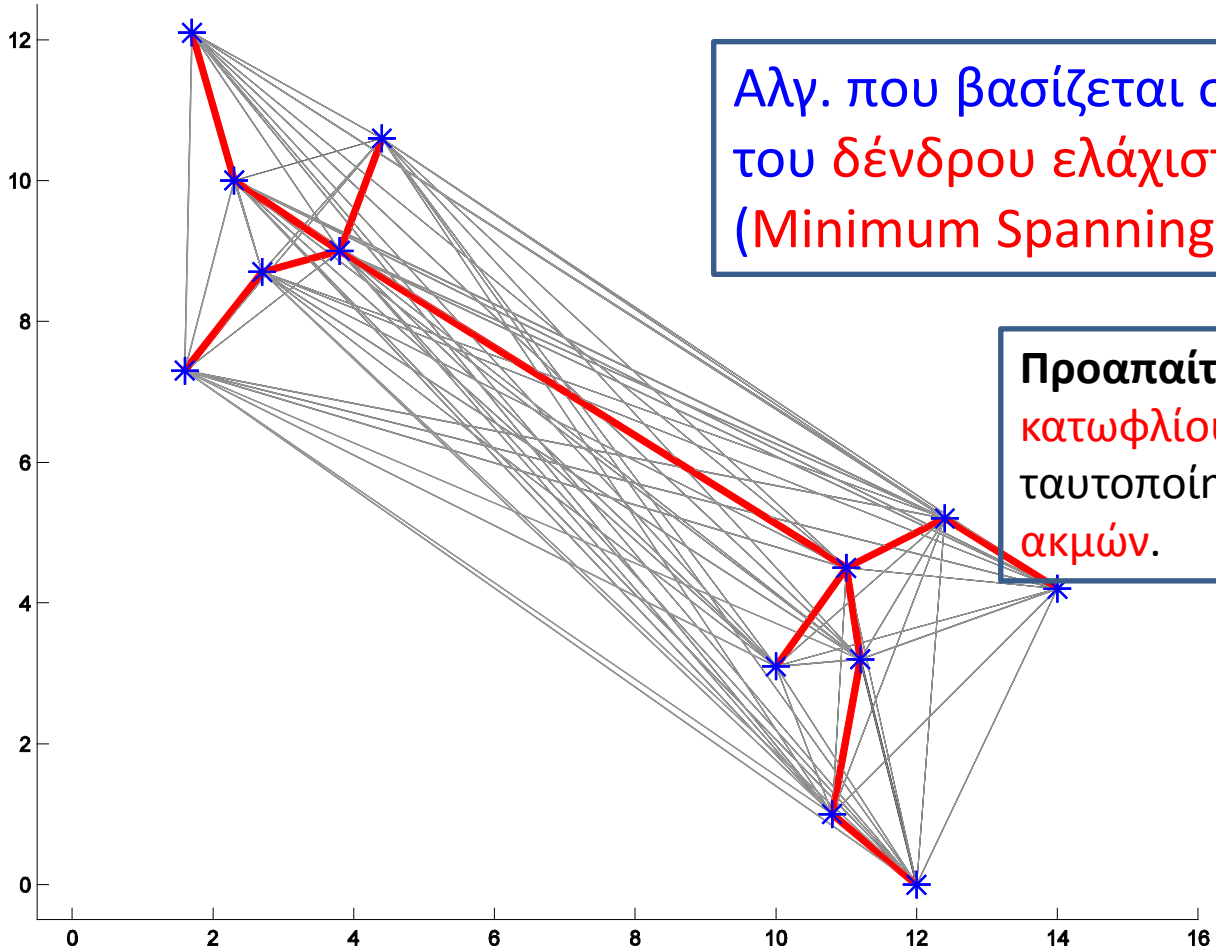


Αλγ. που βασίζεται στην έννοια του δένδρου ελάχιστης κάλυψης (Minimum Spanning Tree - MST).

Προαπαίτηση: Ορισμός κατωφλίου για την ταυτοποίηση "μεγάλων" ακμών.

- **Όρισε** τον **πλήρη** γράφο με κορυφές τα σημεία δεδομένων και ακμές τα τμήματα που συνδέουν κάθε ζεύγος κορυφών.
- **Στάθμισε** κάθε ακμή με την απόσταση μεταξύ των δύο κορυφών που αυτή συνδέει.
- **Όρισε** το MST του γράφου **και αφαίρεσε** τις **"ασυνήθιστα μεγάλες"** ακμές.
- Οι υπό-γράφοι που δημιουργούνται **αντιστοιχούν** στις ομάδες.

Αλγόριθμοι ομαδοποίησης που βασίζονται σε έννοιες γράφων

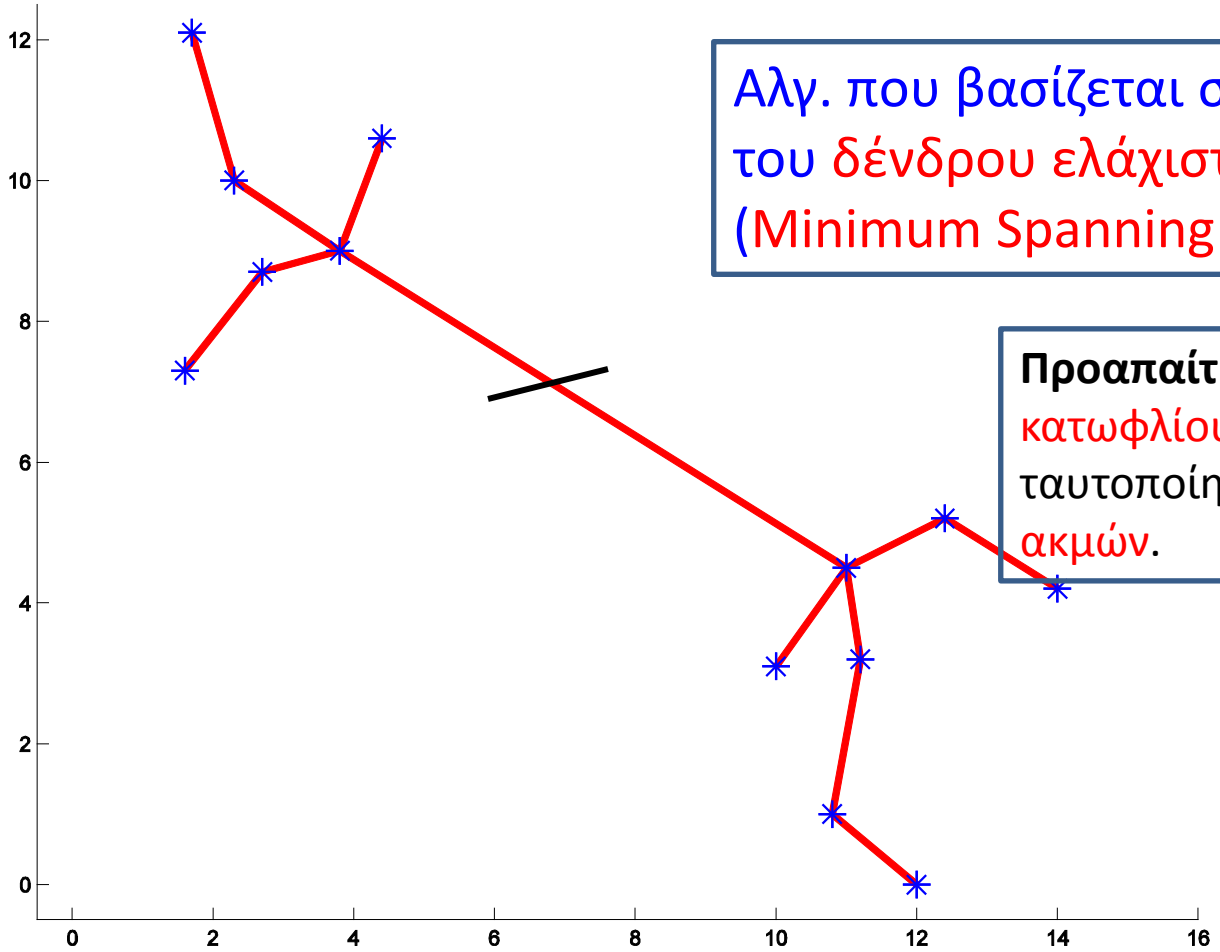


Αλγ. που βασίζεται στην έννοια του δένδρου ελάχιστης κάλυψης (Minimum Spanning Tree - MST).

Προαπαίτηση: Ορισμός κατωφλίου για την ταυτοποίηση "μεγάλων" ακμών.

- **Όρισε** τον **πλήρη** γράφο με κορυφές τα σημεία δεδομένων και ακμές τα τμήματα που συνδέουν κάθε ζεύγος κορυφών.
- **Στάθμισε** κάθε ακμή με την απόσταση μεταξύ των δύο κορυφών που αυτή συνδέει.
- **Όρισε** το MST του γράφου **και αφαίρεσε** τις **"ασυνήθιστα μεγάλες"** ακμές.
- Οι υπό-γράφοι που δημιουργούνται **αντιστοιχούν** στις ομάδες.

Αλγόριθμοι ομαδοποίησης που βασίζονται σε έννοιες γράφων

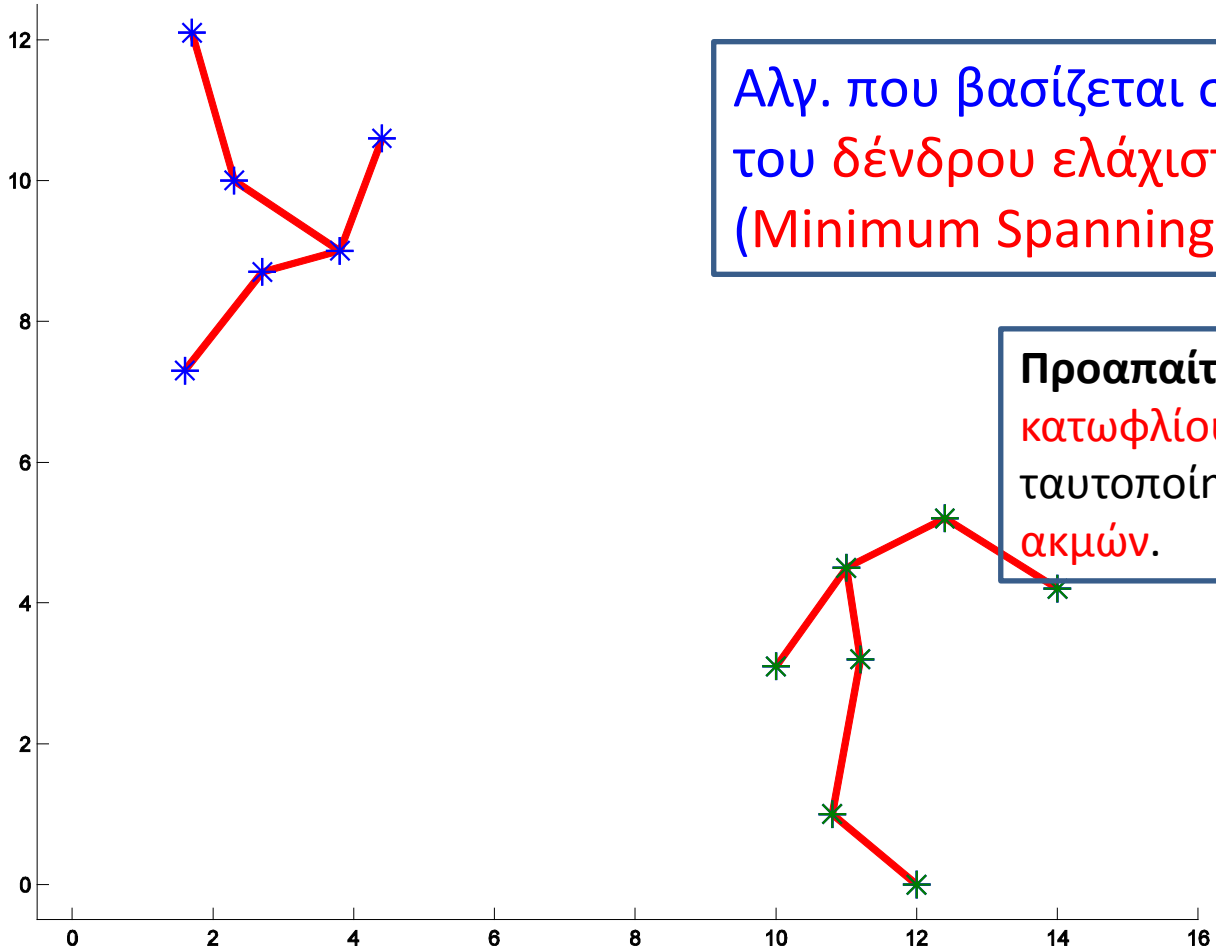


Αλγ. που βασίζεται στην έννοια του δένδρου ελάχιστης κάλυψης (Minimum Spanning Tree - MST).

Προαπαίτηση: Ορισμός κατωφλίου για την ταυτοποίηση "μεγάλων" ακμών.

- **Όρισε** τον **πλήρη** γράφο με κορυφές τα σημεία δεδομένων και ακμές τα τμήματα που συνδέουν κάθε ζεύγος κορυφών.
- **Στάθμισε** κάθε ακμή με την απόσταση μεταξύ των δύο κορυφών που αυτή συνδέει.
- **Όρισε** το MST του γράφου **και αφαίρεσε** τις **"ασυνήθιστα μεγάλες"** ακμές.
- Οι υπό-γράφοι που δημιουργούνται **αντιστοιχούν** στις ομάδες.

Αλγόριθμοι ομαδοποίησης που βασίζονται σε έννοιες γράφων

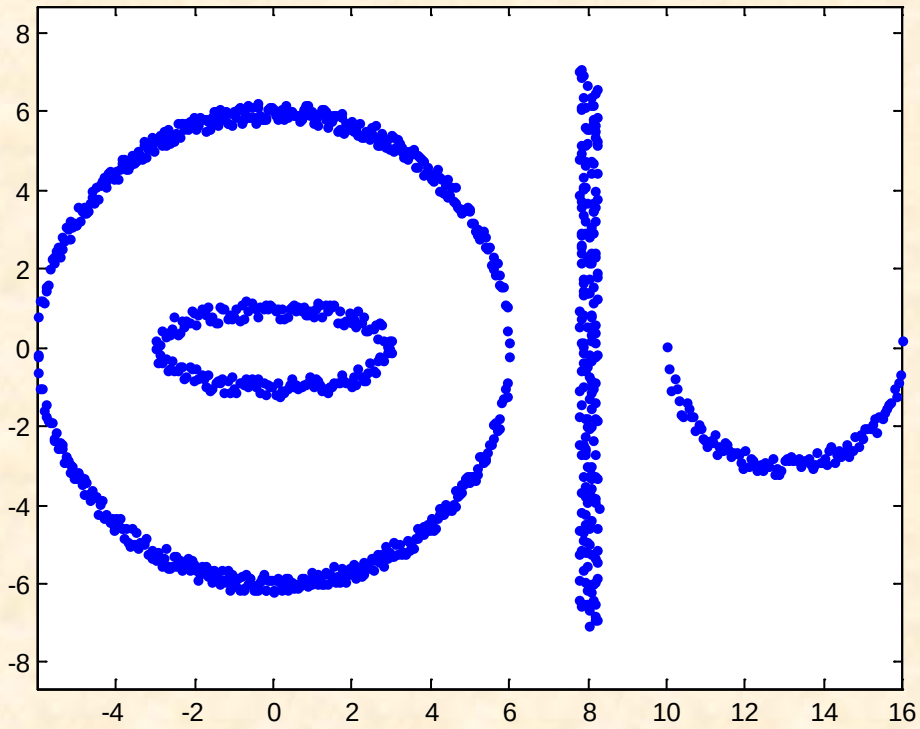


Αλγ. που βασίζεται στην έννοια του δένδρου ελάχιστης κάλυψης (Minimum Spanning Tree - MST).

Προαπαίτηση: Ορισμός κατωφλίου για την ταυτοποίηση "μεγάλων" ακμών.

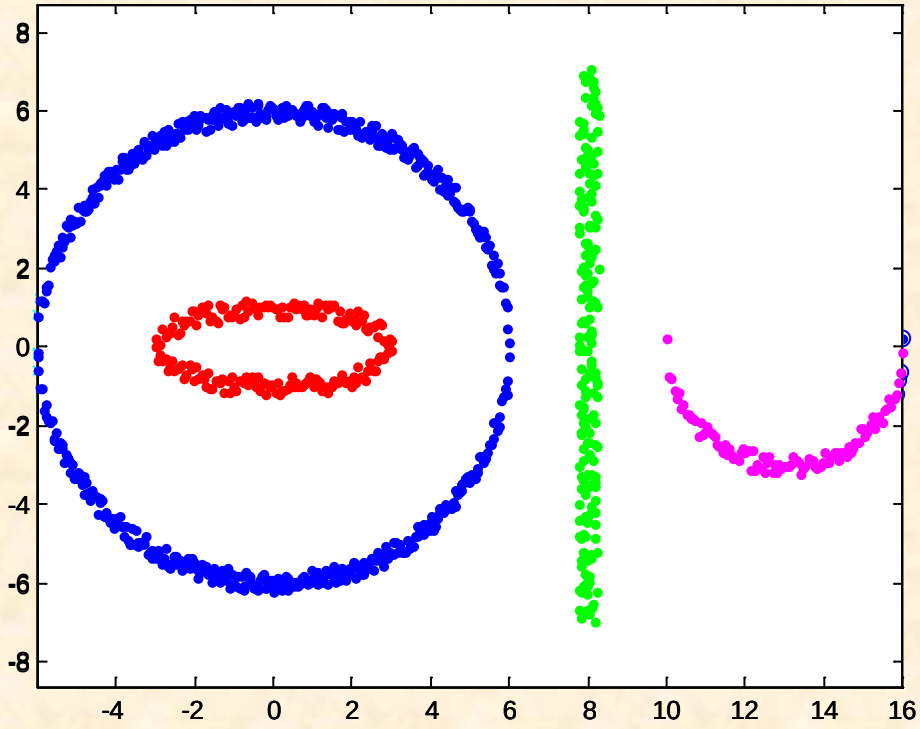
- **Όρισε** τον **πλήρη** γράφο με κορυφές τα σημεία δεδομένων και ακμές τα τμήματα που συνδέουν κάθε ζεύγος κορυφών.
- **Στάθμισε** κάθε ακμή με την απόσταση μεταξύ των δύο κορυφών που αυτή συνδέει.
- **Όρισε** το MST του γράφου **και αφαίρεσε** τις **"ασυνήθιστα μεγάλες"** ακμές.
- Οι υπό-γράφοι που δημιουργούνται **αντιστοιχούν** στις ομάδες.

Παραμετρικοί έναντι μη παραμετρικών αλγορίθμων ομαδοποίησης



Η περίπτωση ομάδων που δεν έχουν την ίδια δομή και δεν τέμνονται μεταξύ τους.

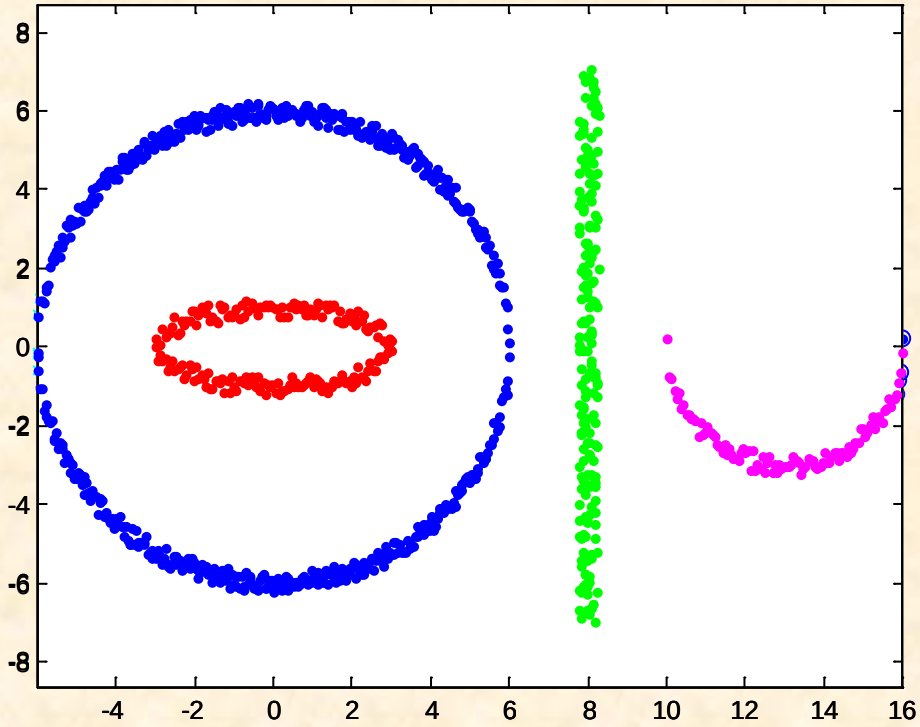
Παραμετρικοί έναντι μη παραμετρικών αλγορίθμων ομαδοποίησης



Παραμετρικοί αλγόριθμοι

Στην περίπτωση ομάδων που δεν έχουν την ίδια δομή και δεν τέμνονται μεταξύ τους, οι παραμετρικοί αλγόριθμοι ομαδοποίησης είναι πιο πιθανό να **αποτύχουν**, σε σχέση με τους μη-παραμετρικούς αλγόριθμους.

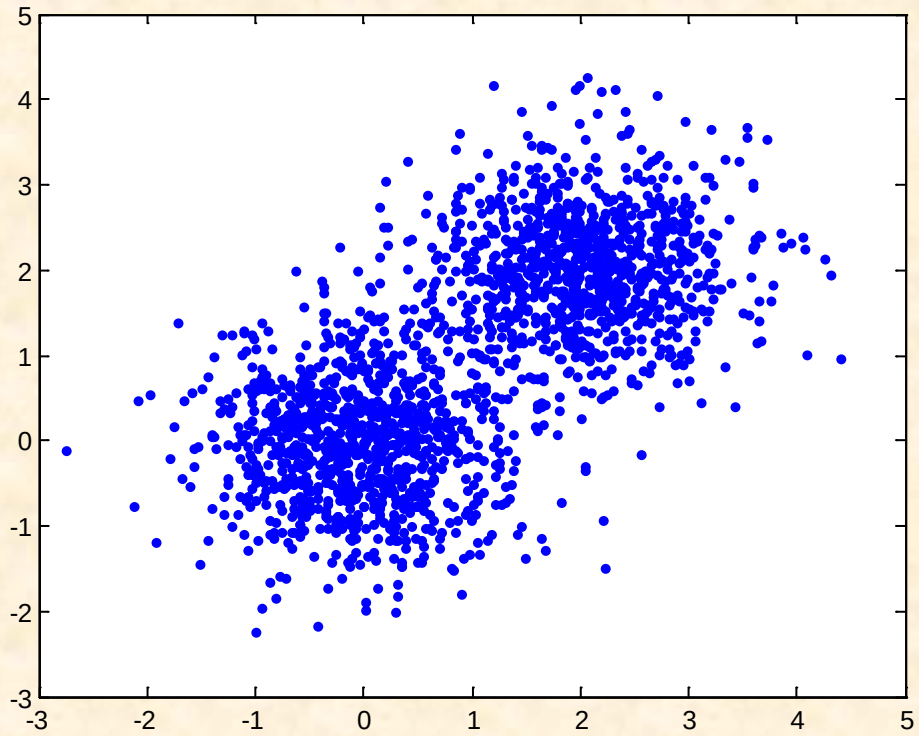
Παραμετρικοί έναντι μη παραμετρικών αλγορίθμων ομαδοποίησης



Μη παραμετρικοί αλγόριθμοι

Στην περίπτωση ομάδων που δεν έχουν την ίδια δομή και δεν τέμνονται μεταξύ τους, οι παραμετρικοί αλγόριθμοι ομαδοποίησης είναι πιο πιθανό να **αποτύχουν**, σε σχέση με τους μη-παραμετρικούς αλγόριθμους.

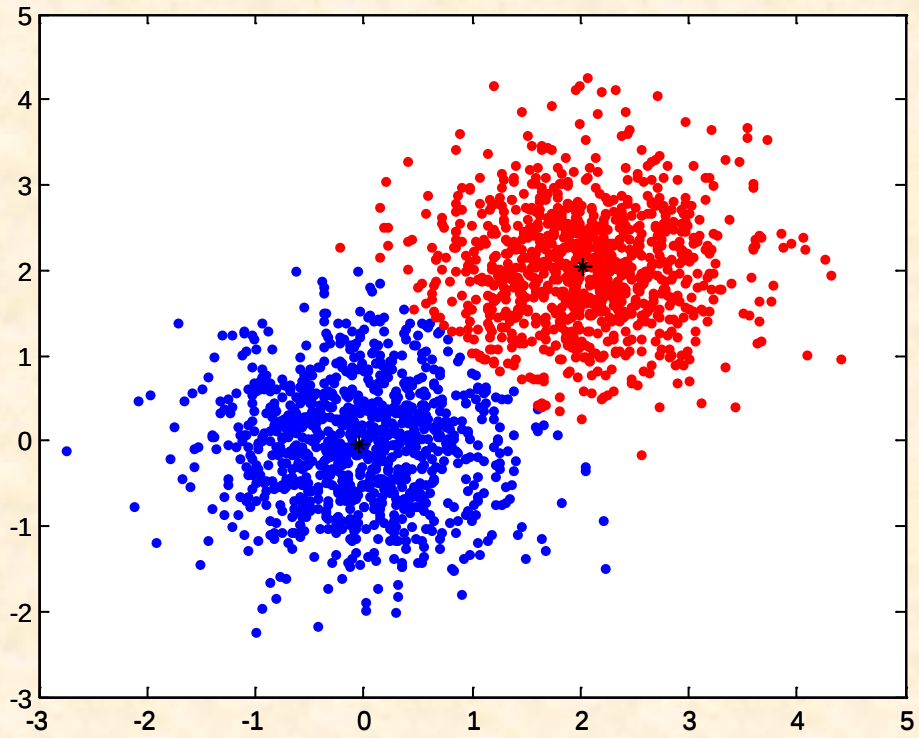
Παραμετρικοί έναντι μη παραμετρικών αλγορίθμων ομαδοποίησης



Μη παραμετρικοί αλγόριθμοι

Όταν οι ομάδες έχουν την ίδια δομή και παρουσιάζουν επικάλυψη, οι μη παραμετρικοί αλγόριθμοι είναι πιο πιθανό να **αποτύχουν** σε σχέση με τους παραμετρικούς αλγόριθμους.

Παραμετρικοί έναντι μη παραμετρικών αλγορίθμων ομαδοποίησης



Παραμετρικοί αλγόριθμοι

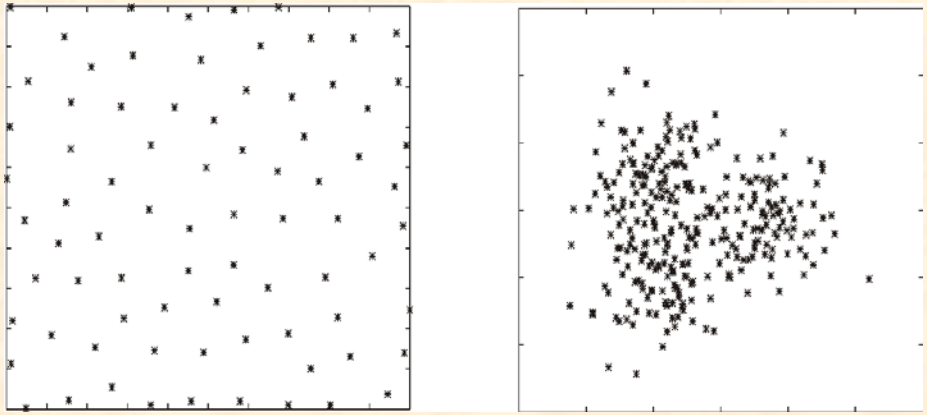
Όταν οι ομάδες έχουν την ίδια δομή και παρουσιάζουν επικάλυψη, οι μη παραμετρικοί αλγόριθμοι είναι πιο πιθανό να **αποτύχουν** σε σχέση με τους παραμετρικούς αλγόριθμους.

Αξιολόγηση αποτελεσμάτων ομαδοποίησης

- Αυτή πραγματοποιείται από έναν ειδικό στο πεδίο της εφαρμογής.
- Εμπεριέχει πάντα την υποκειμενικότητα του ειδικού.
- Ιαν τα αποτελέσματα δεν είναι ικανοποιητικά, η όλη διαδικασία ομαδοποίησης μπορεί να επαναληφθεί με
  a) διαφορετικό μέτρο εγγύτητας ή/και
  b) διαφορετική αναπαράσταση δεδομένων ή/και
  c) διαφορετικό αλγόριθμο ομαδοποίησης

Μια τελευταία σημείωση: **Δεν είναι όλα τα προβλήματα** κατάλληλα για επεξεργασία με τεχνικές ομαδοποίησης.
(Για παράδειγμα, τα διανύσματα σε δεδομένο πρόβλημα, μπορεί να μην σχηματίζουν καθόλου φυσικές ομάδες).

Δεδομένα που παρουσιάζουν κανονικότητες (Regularly spaced Data)



Δεδομένα που σχηματίζουν μία φυσική ομάδα.