Τμήμα Πληροφορικής με Εφαρμογές στη Βιοϊατρική
Σχολή Θετικών Επιστημών
Πανεπιστήμιο Θεσσαλίας

ΑΝΑΓΝΩΡΙΣΗ ΠΡΟΤΥΠΩΝ

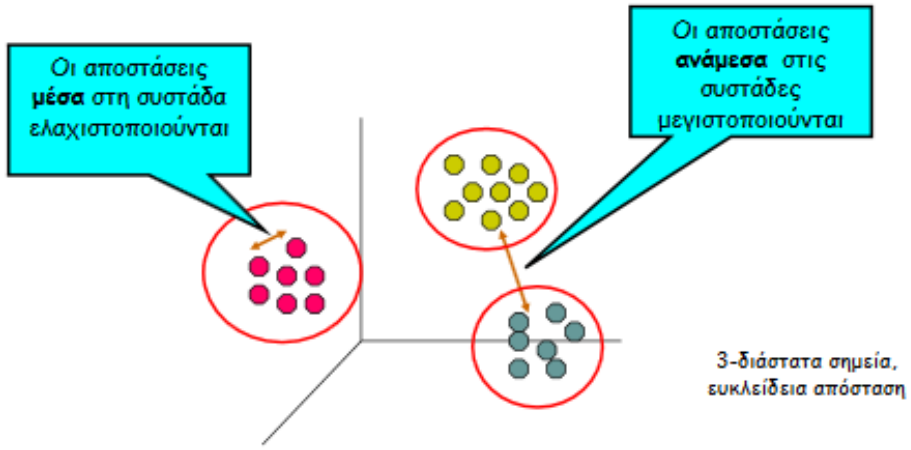
Ανάλυση Δεδομένων

Αριστείδης Γ. Βραχάτης, Dipl-Ing, M.Sc, PhD
Adjunct Lecturer

Συσταδοποίηση - Clustering

- Είναι η διαδικασία της κατηγοριοποίησης των δεδομένων σε σύνολα ομοειδών αντικειμένων καλούμενα ομάδες (clusters)

- Στόχος
  - Να παράγει ένα σύνολο από ομάδες με υψηλή εντός των ομάδων ομοιότητα (intra-cluster similarity), ενώ παράλληλα να διατηρείται χαμηλή η ομοιότητα μεταξύ των διαφόρων ομάδων (inter-cluster similarity)



- Εφαρμογές
  - Ευρύ φάσμα εφαρμογών, από τις κοινωνικές επιστήμες, την οικονομία, την αναγνώριση προτύπων έως την βιοπληροφορική, την αστροφυσική και σεισμολογία

Ομαδοποίηση - Clustering

- Well Separated
  - μία συστάδα είναι το σύνολο των αντικειμένων όπου κάθε αντικείμενο είναι πιο κοντά σε κάθε άλλο αντικείμενο της συστάδας, από ότι σε κάποιο άλλο αντικείμενο.

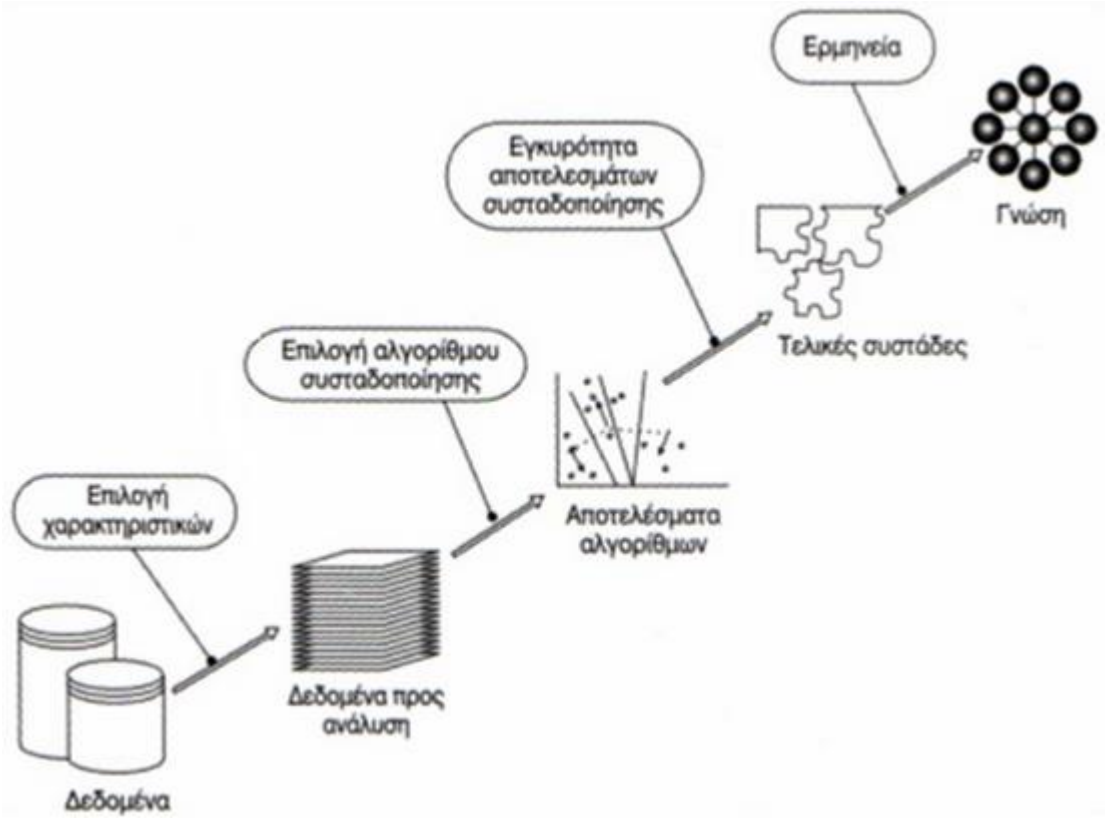
- Prototype Based
  - μία συστάδα είναι τα αντικείμενα που είναι πιο κοντά σε ένα πρωτότυπο (prototype) από ότι κάποιο άλλο αντικείμενο. Συνήθως σαν πρωτότυπο επιλέγεται το μέσο των σημείων μίας συστάδας.

- Graph Based
  - μία συνεκτική συνιστώσα ή μία κλίκα του γραφήματος.

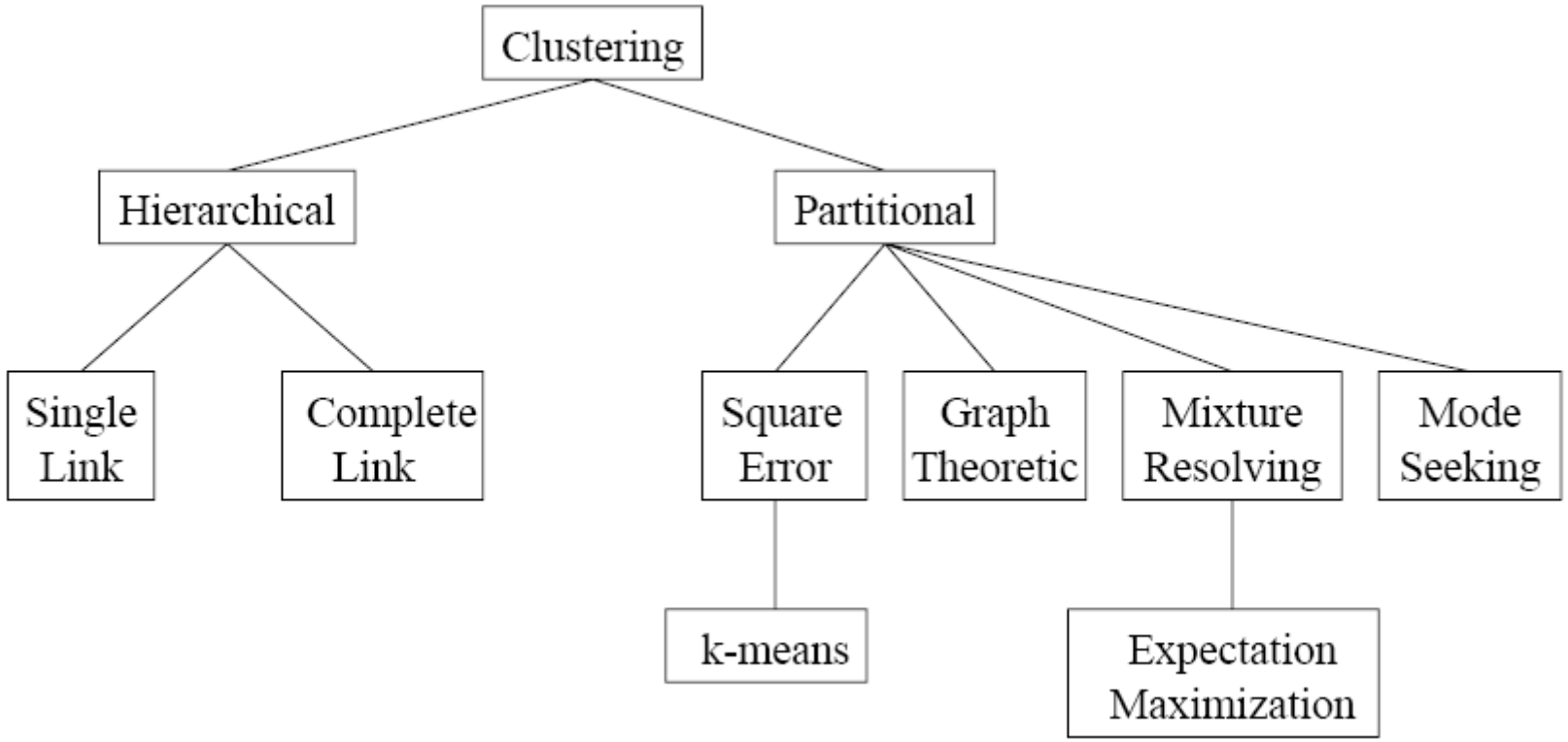
- Density Based
  - μία πυκνή περιοχή αντικειμένων που περιβάλλεται από μία αραιή

- Shared Property (conceptual clusters)
  - σύνολο αντικειμένων που μοιράζονται μία ιδιότητα – έχει εφαρμογή κυρίως σε κατηγορικά αντικείμενα

Βήματα Διαδικασίας Συσταδοποίησης



Κατηγοριοποίηση των Αλγορίθμων Ομαδοποίησης



Είδη Ομαδοποίησης

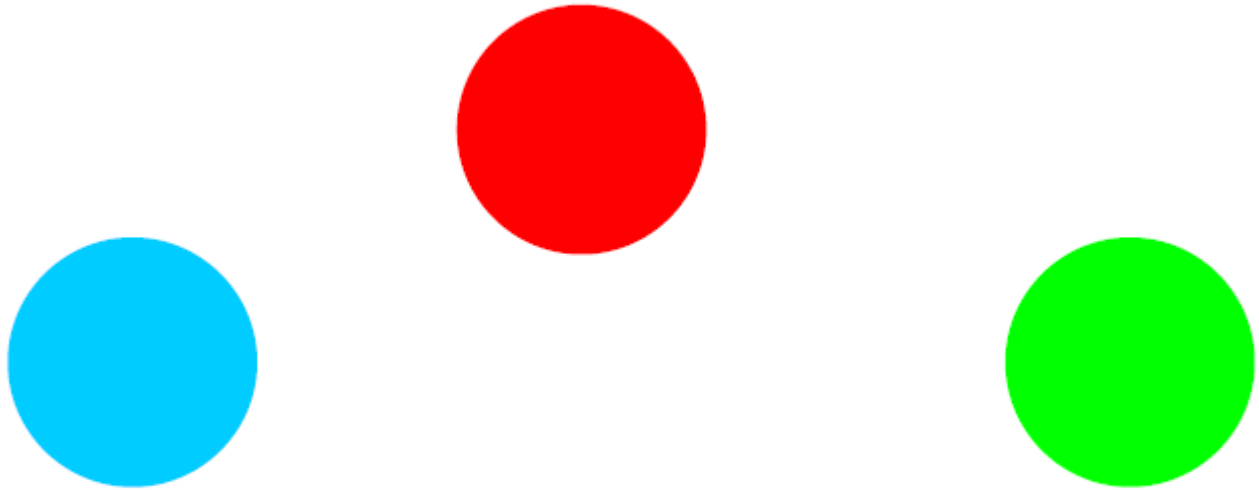
- Βασική διάκριση ανάμεσα στο ιεραρχικό (hierarchical) και διαχωριστικό (partitional) σύνολο από ομάδες

- Διαχωριστική Συσταδοποίηση (Partitional Clustering)
  - Ένας διαμερισμός των αντικειμένων σε μη επικαλυπτόμενα -non-overlapping - υποσύνολα (συστάδες) τέτοιος ώστε κάθε αντικείμενο ανήκει σε ακριβώς ένα υποσύνολο

- Ιεραρχική Συσταδοποίηση (Hierarchical clustering)
  - Ένα σύνολο από εμφωλευμένες (nested) ομάδες Επιτρέπουμε σε μια συστάδα να έχει υπο-συστάδες οργανωμένες σε ένα ιεραρχικό δέντρο

Τύποι συστάδων: Καλώς Διαχωρισμένες Συστάδες

Μια συστάδα είναι ένα σύνολο από σημεία τέτοια ώστε κάθε σημείο μιας συστάδας είναι κοντινότερο σε (ή πιο όμοιο με) όλα τα άλλα σημεία της συστάδας από ότι σε οποιοδήποτε άλλο σημείο που δεν ανήκει στη συστάδα.



3 καλώς-διαχωρισμένες συστάδες

Συχνά υπάρχει η έννοια του κατωφλιού (threshold)

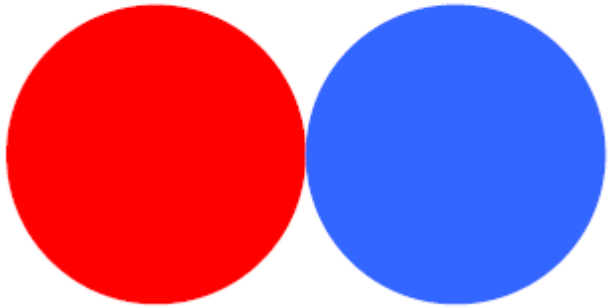
Όχι απαραίτητα κυκλικοί (οποιοδήποτε σχήμα)

Τύποι συστάδων: Συστάδες βασισμένες σε κέντρο ή πρότυπο

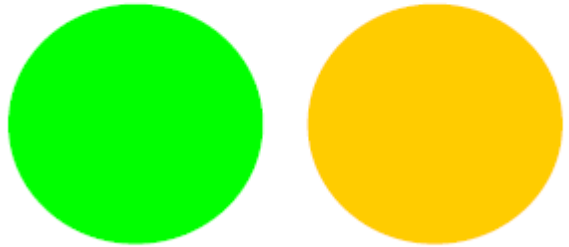
Μια συστάδα είναι ένα σύνολο από αντικείμενα τέτοιο ώστε ένα αντικείμενο στην συστάδα είναι κοντινότερο σε (ή πιο όμοιο με) το «κέντρο» ή πρότυπο της συστάδας από ότι από το κέντρο οποιασδήποτε άλλης συστάδας.

Το κέντρο της ομάδας είναι συχνά

- centroid, ο μέσος όρος των σημείων της συστάδας, ή
- a medoid, το πιο «αντιπροσωπευτικό» σημείο της συστάδας (πχ όταν κατηγορικά γνωρίσματα)



4 συστάδες βασισμένες σε κέντρο



Τείνουν στο να είναι κυκλικοί

Τύποι συστάδων: Συνεχής Συστάδες

Συνεχής Συστάδες (Contiguous Cluster) (Κοντινότερος γείτονα ή μεταβατικά) – Βάσει γειτνίασης

Μια συστάδα είναι ένα σύνολο σημείων τέτοιο ώστε κάθε σημείο είναι πιο κοντά σε ένα ή περισσότερα σημεία της συστάδας από ό,τι σε οποιοδήποτε άλλο σημείο εκτός συστάδας

Συχνά σε περιπτώσεις συστάδων με μη κανονικό σχήμα ή με αλληλοπλεκόμενα σχήματα – ή όταν έχουμε γραφήματα και θέλουμε να βρούμε συνεκτικά υπογραφήματα

Πρόβλημα με θόρυβο

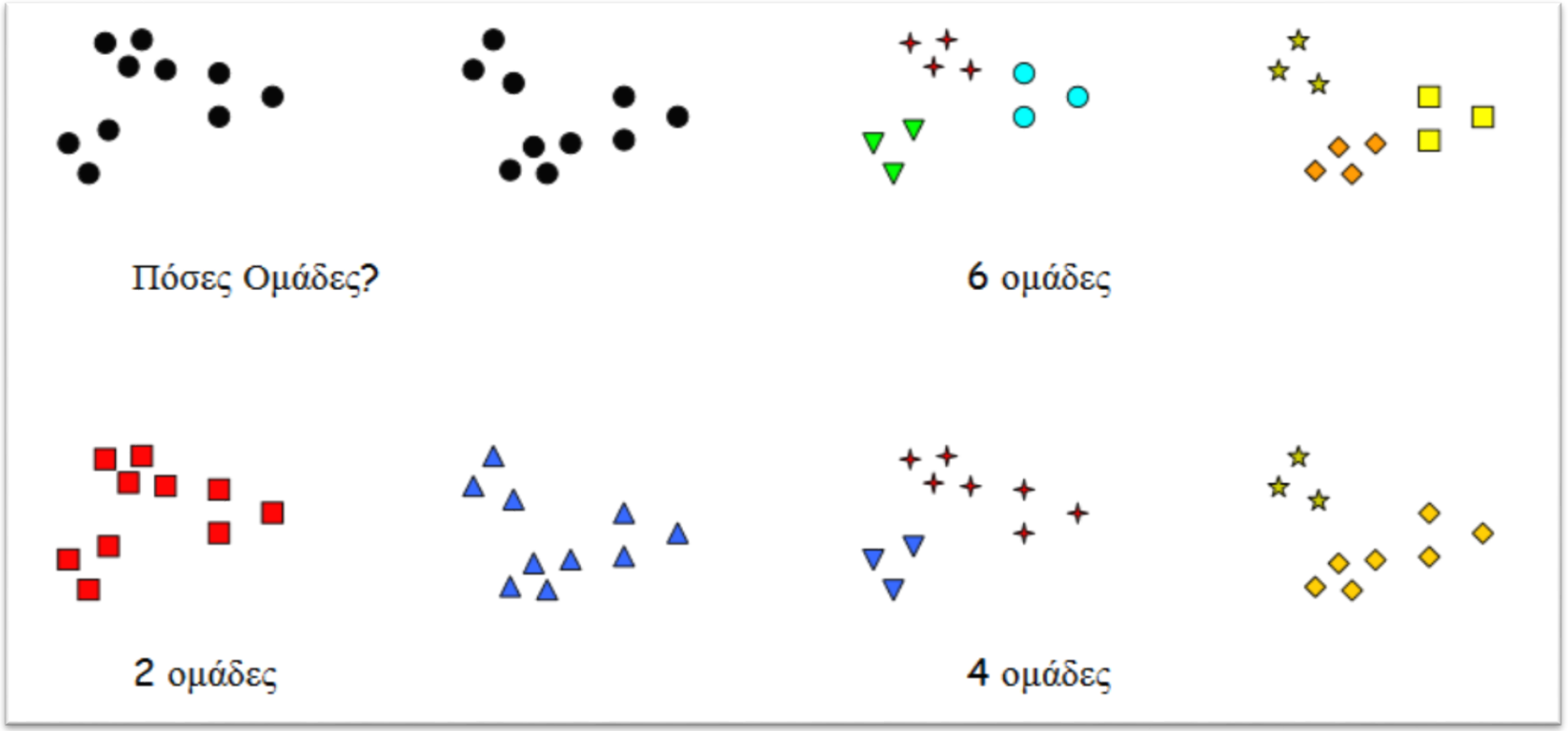
8 συνεχείς συστάδες

Τύποι συστάδων: Συστάδες βασισμένες στην πυκνότητα

– Μια συστάδα είναι μια πυκνή περιοχή από σημεία την οποία χωρίζουν από άλλες περιοχές μεγάλης πυκνότητας περιοχές χαμηλής πυκνότητας

– Συχνά σε περιπτώσεις συστάδων με μη κανονικό σχήμα ή με αλληλοπλεκόμενα σχήματα ή όταν θόρυβος ή outliers

Ασαφεια



Ο αλγόριθμος K-μέσων (k-means)

– Θέλουμε να βρούμε εκείνο το σύνολο k σημείων στον d-διάστατο χώρο, το οποίο ελαχιστοποιεί την μέση απόσταση ελαχίστων τετραγώνων κάθε σημείου από το κοντινότερό του κέντρο

$$d_E(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

– Η συνάρτηση βαθμολόγησης που βασίζεται στο άθροισμα των τετραγώνων των σφαλμάτων (SSE) ορίζεται ως:

$$SSE(\mathrm{C}) = \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in C_i} \| \mathbf{x}_j - \boldsymbol{\mu}_i \|^2$$

– Ο στόχος μας είναι να βρούμε εκείνη τη συσταδοποίηση που ελαχιστοποιεί τη βαθμολογία SSE:

$$\mathrm{C}^* = \arg\min_{\mathrm{C}} \{SSE(\mathrm{C})\}$$

– Ο αλγόριθμος K μέσων χρησιμοποιεί μια άπληστη επαναληπτική τεχνική για να βρει μια συσταδοποίηση που ελαχιστοποιεί την αντικειμενική συνάρτηση SSE.

Ο αλγόριθμος K-μέσων (k-means)

- Ο αλγόριθμος K μέσων καθορίζει τις αρχικές τιμές των μέσων για τις συστάδες παράγοντας με τυχαίο τρόπο k σημεία στον χώρο δεδομένων. Κάθε επανάληψη του αλγορίθμου K μέσων αποτελείται από δύο βήματα: (1) την αντιστοίχιση σε συστάδες και (2) την ενημέρωση των κέντρων βάρους.

- Με την προϋπόθεση ότι δίνονται οι μέσοι των k συστάδων, κάθε σημείο $x_j \in D$ αντιστοιχίζεται στον πλησιέστερο μέσο κατά τη διάρκεια του πρώτου βήματος του αλγορίθμου· αυτό προκαλεί μια συσταδοποίηση, με κάθε συστάδα C_i να περιλαμβάνει σημεία που βρίσκονται πιο κοντά στον μέσο μ_i σε σύγκριση με τον μέσο οποιασδήποτε άλλης συστάδας. Δηλαδή, κάθε σημείο x_j αντιστοιχίζεται στη συστάδα C_{j*}, όπου

$$j^* = \arg\min_{k}^{i=1} \left\{ \| \mathbf{x}_j - \boldsymbol{\mu}_i \|^2 \right\}$$

- Για ένα καθορισμένο σύνολο συστάδων C_i, $i = 1,\ldots, k$, στο δεύτερο βήμα του αλγορίθμου (ενημέρωση των κέντρων βάρους) υπολογίζονται νέες μέσες τιμές για κάθε συστάδα από τα σημεία του συνόλου C_i.

- Τα βήματα της αντιστοίχισης σε συστάδες και της ενημέρωσης των κέντρων βάρους εκτελούνται επαναληπτικά μέχρι να καταλήξουμε σε ένα σταθερό σημείο ή σε τοπικά ελάχιστα.

Ο αλγόριθμος K-μέσων (k-means)

- ΣΚΟΠΟΣ : Εύρεση των κέντρων των ομάδων
- ΜΕΘΟΔΟΣ : Ελαχιστοποίηση του σφάλματος, J

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} (||x_i^{(j)} - c_j||)^2$$

- ΒΗΜΑΤΑ
  I. Ορισμός K κέντρων συστάδων με τυχαίο τρόπο
  II. Εισαγωγή αντικειμένου στη συστάδα με το πιο κοντινό κέντρο
  III. Ανανέωση του κέντρου της συστάδας
  IV. Επανάληψη των βημάτων 2,3 μέχρι τη σύγκλιση (αλλαγή στις συστάδες μικρότερη από ένα κατώφλι)

- Ουσιαστικά, ο αλγόριθμος προσπαθεί επαναληπτικά να «μειώσει» την απόσταση όλων των σημείων από ένα σημείο της συστάδας

Ο αλγόριθμος K-μέσων (k-means)

K-MEANS (D, k, ϵ **)**:

1 $t \leftarrow 0$

2 Καθορισμός αρχικής τιμής για k κέντρα βάρους με τυχαίο τρόπο: $\boldsymbol{\mu}_1^t, \boldsymbol{\mu}_2^t, \ldots, \boldsymbol{\mu}_k^t \in \mathbb{R}^d$

3 **repeat**

4 $\quad t \leftarrow t + 1$

5 $\quad C_j \leftarrow \emptyset$ για όλα τα $j = 1, \ldots, k$

$\quad$ // Βήμα αντιστοίχισης σε συστάδες

6 $\quad$ **foreach** $\mathbf{x}_j \in \mathbf{D}$ **do**

7 $\quad\quad j^* \leftarrow \arg\min_i \left\{ \left\| \mathbf{x}_j - \boldsymbol{\mu}_i^{t-1} \right\|^2 \right\}$ // Αντιστοίχιση του $\mathbf{x}_j$ στο πλησιέστερο κέντρο βάρους

8 $\quad\quad C_{j^*} \leftarrow C_{j^*} \cup \{\mathbf{x}_j\}$

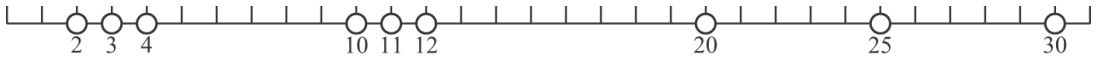
$\quad$ // Βήμα ενημέρωσης των κέντρων βάρους

9 $\quad$ **foreach** $i = 1$ *to* k **do**

10 $\quad\quad \boldsymbol{\mu}_i^t \leftarrow \frac{1}{|C_i|} \sum_{\mathbf{x}_j \in C_i} \mathbf{x}_j$

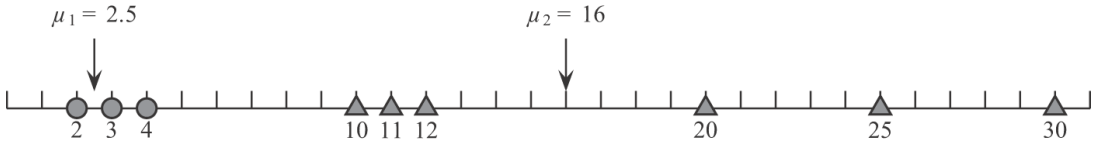
11 **until** $\sum_{i=1}^{k} \left\| \boldsymbol{\mu}_i^t - \boldsymbol{\mu}_i^{t-1} \right\|^2 \leq \epsilon$

Ο αλγόριθμος Κ μέσων στη μία διάσταση

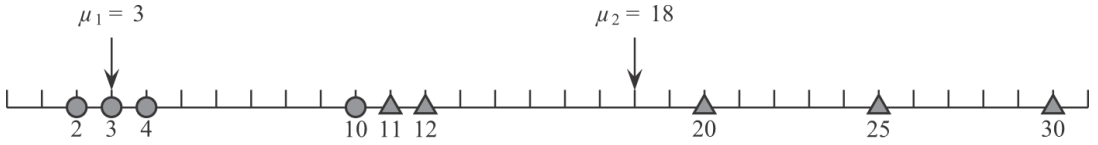


(α) Αρχικό σύνολο δεδομένων

(β) Επανάληψη: $t = 1$



(γ) Επανάληψη: $t = 2$



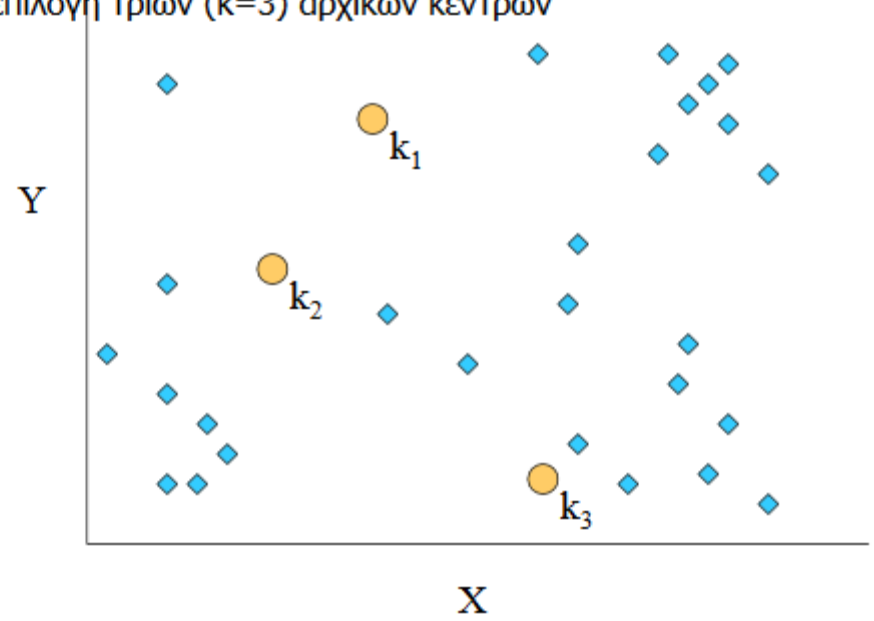
(δ) Επανάληψη: $t = 3$

(ε) Επανάληψη: $t = 4$

(στ) Επανάληψη: $t = 5$ (σύγκλιση)

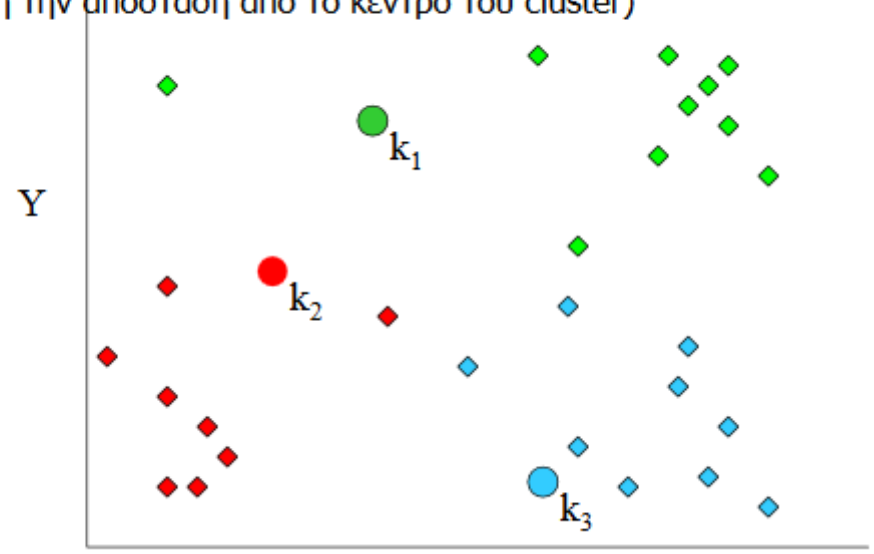
K-means σε 2 διαστάσεις

- Τυχαία επιλογή τριών (k=3) αρχικών κέντρων



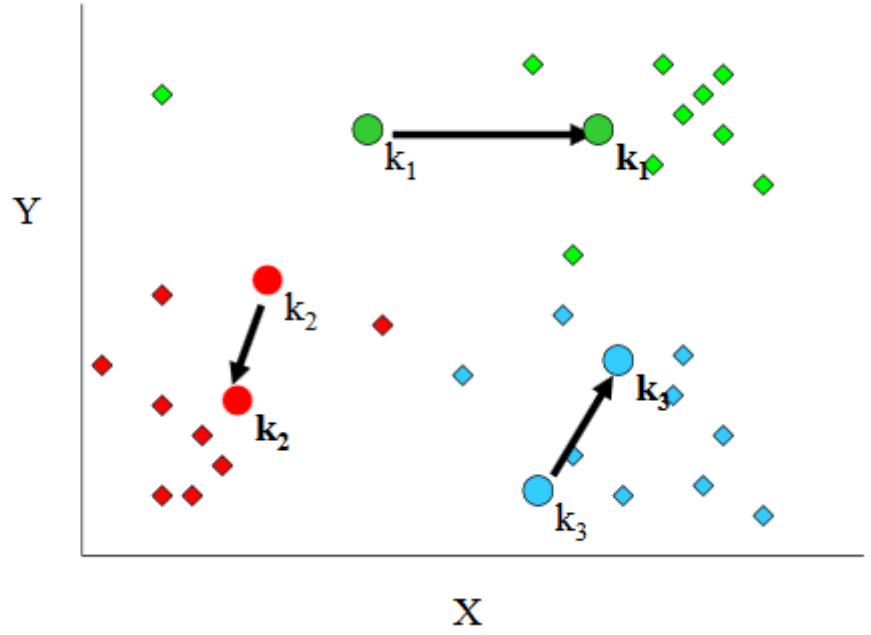
K-means σε 2 διαστάσεις

- Εκχώρηση κάθε στοιχείου στο πλησιέστερό του cluster
(με βάση την απόσταση από το κέντρο του cluster)



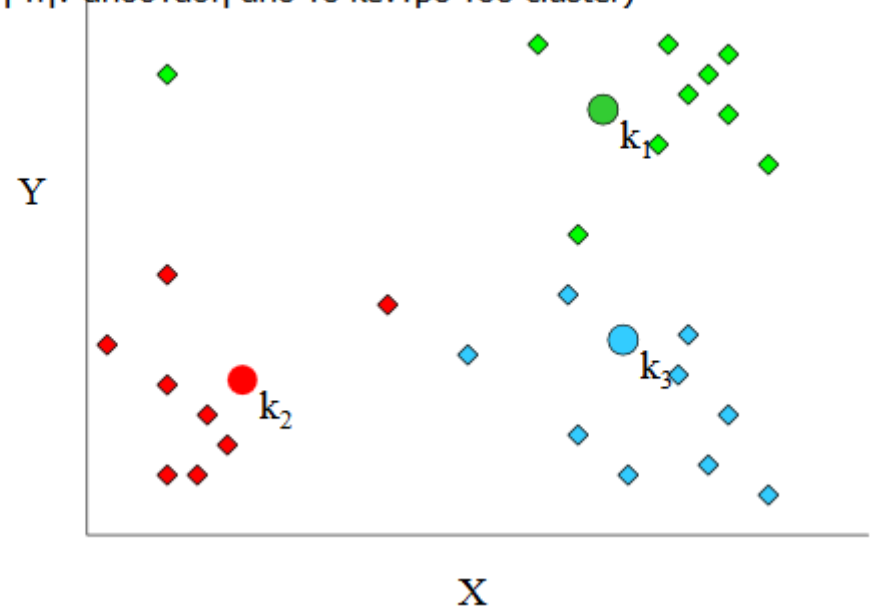
K-means σε 2 διαστάσεις

- Επανυπολογισμός του νέου κέντρου βάρους του κάθε cluster

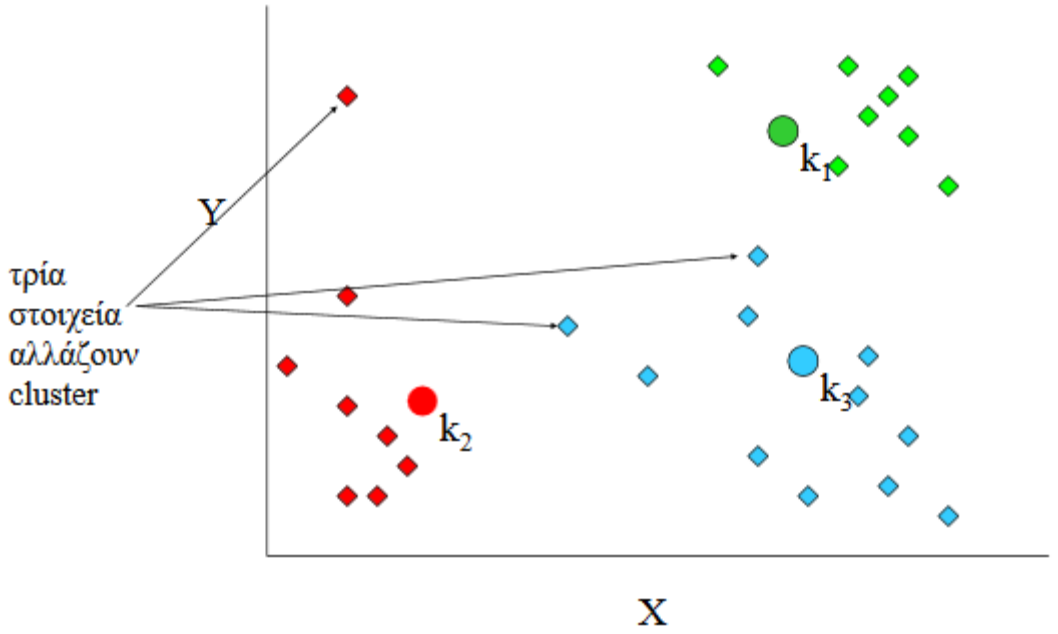


K-means σε 2 διαστάσεις

- Εκχώρηση κάθε στοιχείου στο πλησιέστερό του cluster
  (με βάση την απόσταση από το κέντρο του cluster)

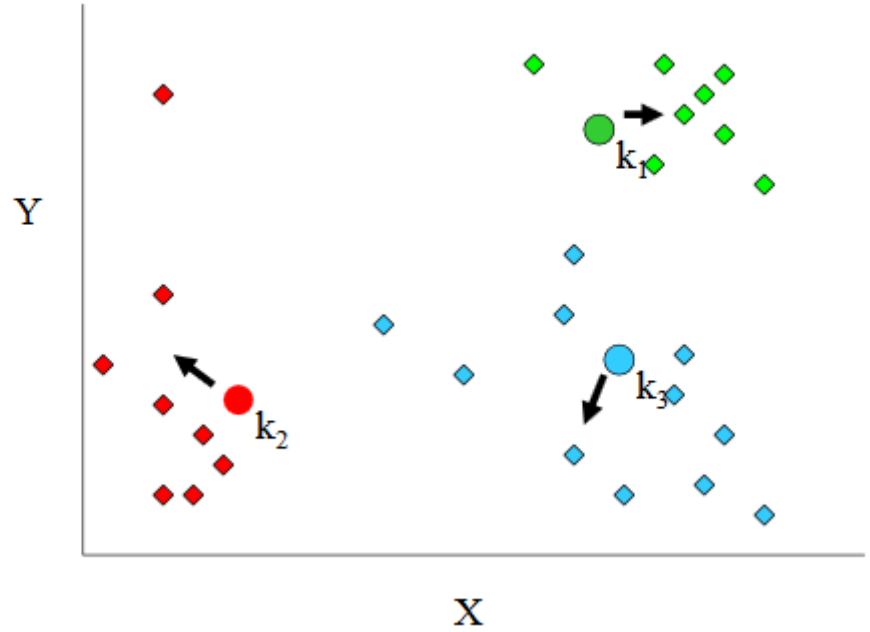


K-means σε 2 διαστάσεις

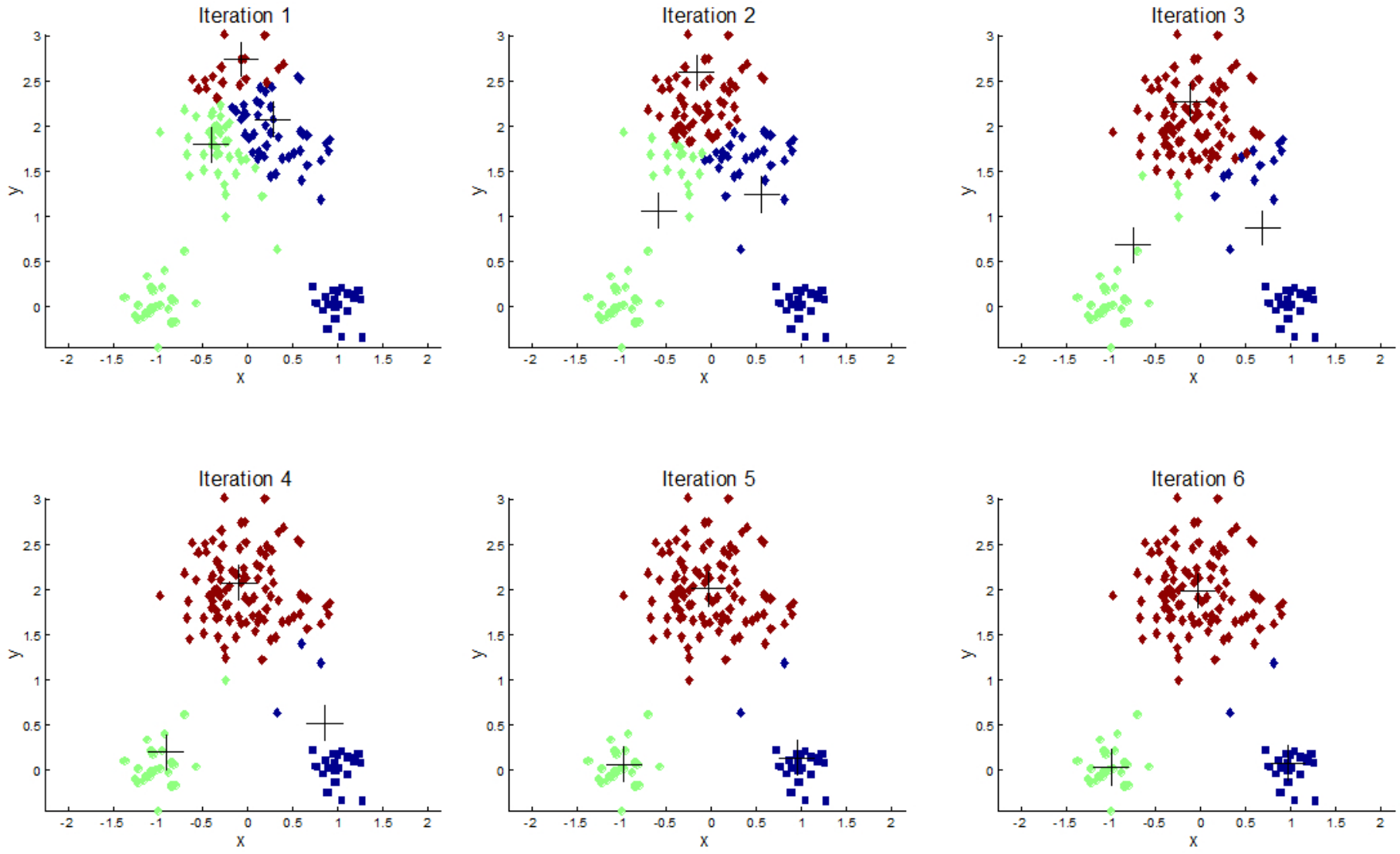


K-means σε 2 διαστάσεις

- Επανυπολογισμός του νέου κέντρου βάρους του κάθε cluster



Αλγόριθμος k-means - ΒΗΜΑΤΑ



Παραδείγματα

Ασκηση 1

– Δίνεται: {2, 4, 10, 12, 3, 20, 30, 11, 25}, k=2 και Τυχαία επιλέγουμε, έστω κέντρα m1=3, m2=4

Άσκηση 2

– Δίνεται:

X1	X2
1	1
2	1
4	4
1	2
4	5
5	4
1	3
8	3

και τυχαία κέντρα k1 = {1, 0}, k2 = {1, 3}

Επιλογή k

- Χρήση άλλης μεθόδου ομαδοποίησης
- Εφαρμογή του αλγορίθμου για διάφορες τιμές του k
- Χρήση πρότερης γνώσης για το είδος των δεδομένων
  - Κακοήθης - Καλοήθης

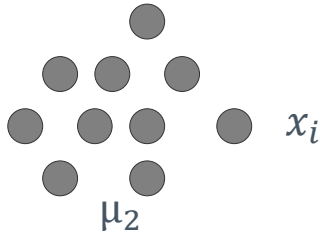
K-means - Συμπέρασμα

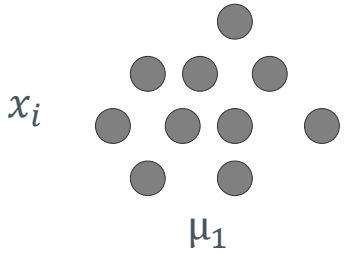
- Πλεονεκτήματα
  - Απλός, κατανοητός
  - Τα αντικείμενα ανατίθενται αυτόματα σε κάποιο cluster
  - Ταχύτητα σύγκλισης

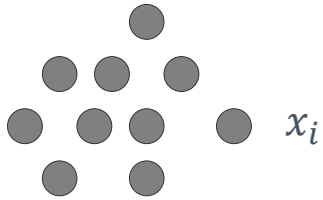
- Μειονεκτήματα
  - Πρέπει να οριστεί ο αριθμός των clusters
  - Όλα τα αντικείμενα πρέπει υποχρεωτικά να ανήκουν σε κάποιο cluster
  - Δε δουλεύει για μη αριθμητικά δεδομένα
  - Μη-ντετερμινιστικός

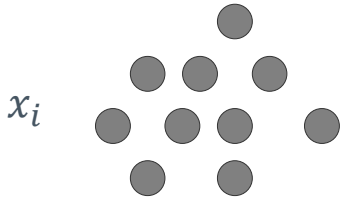
Ερωτηση

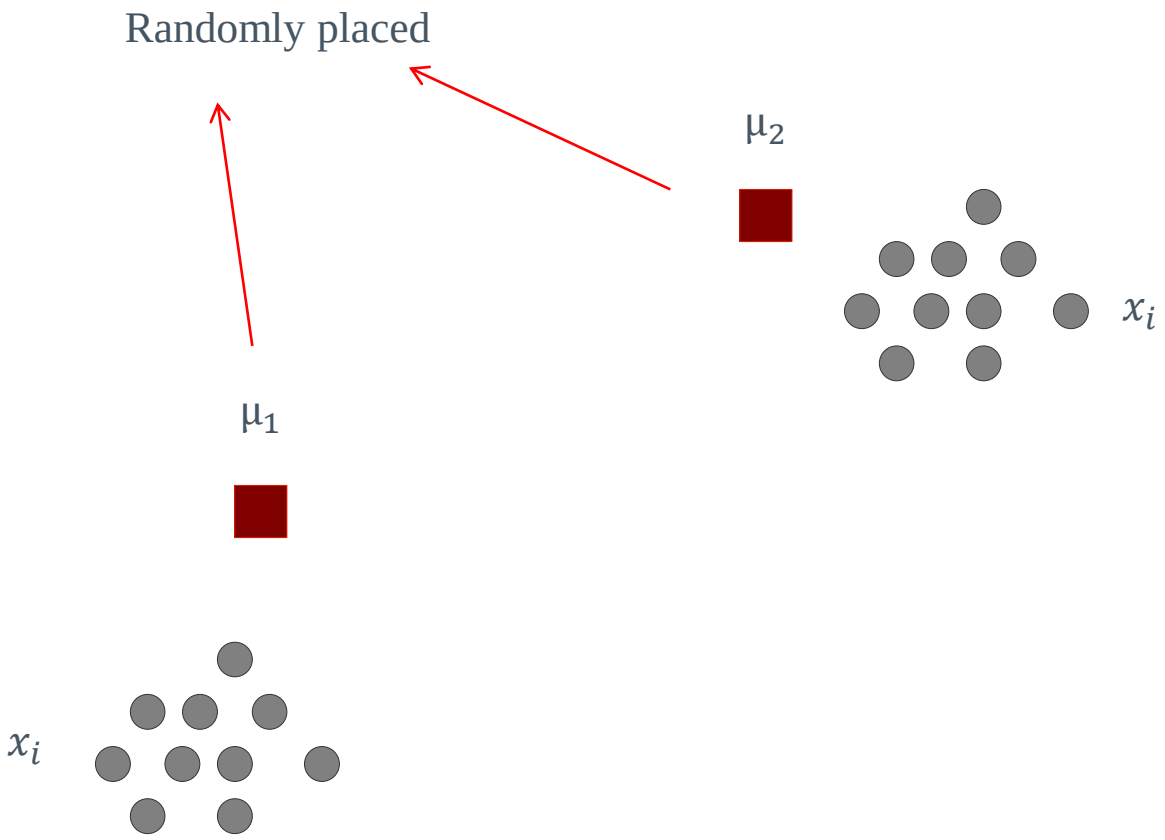
– Ποια αρχικοποίηση θα μπέρδευε τον K-means ? (εστω k = 2)



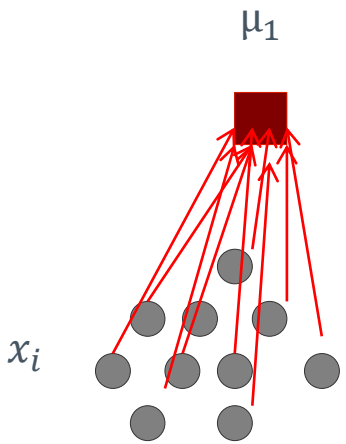


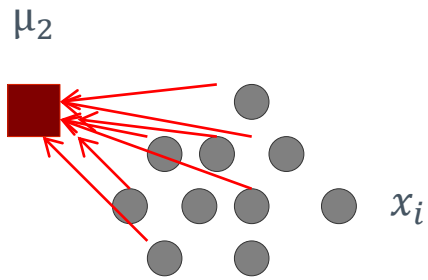

x_i


x_i

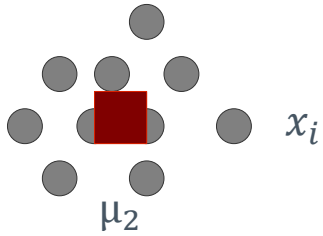

Randomly placed
μ_2
x_i
μ_1
x_i

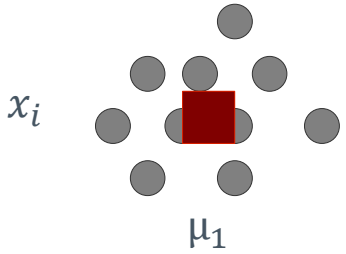
Find optimal allocation of points





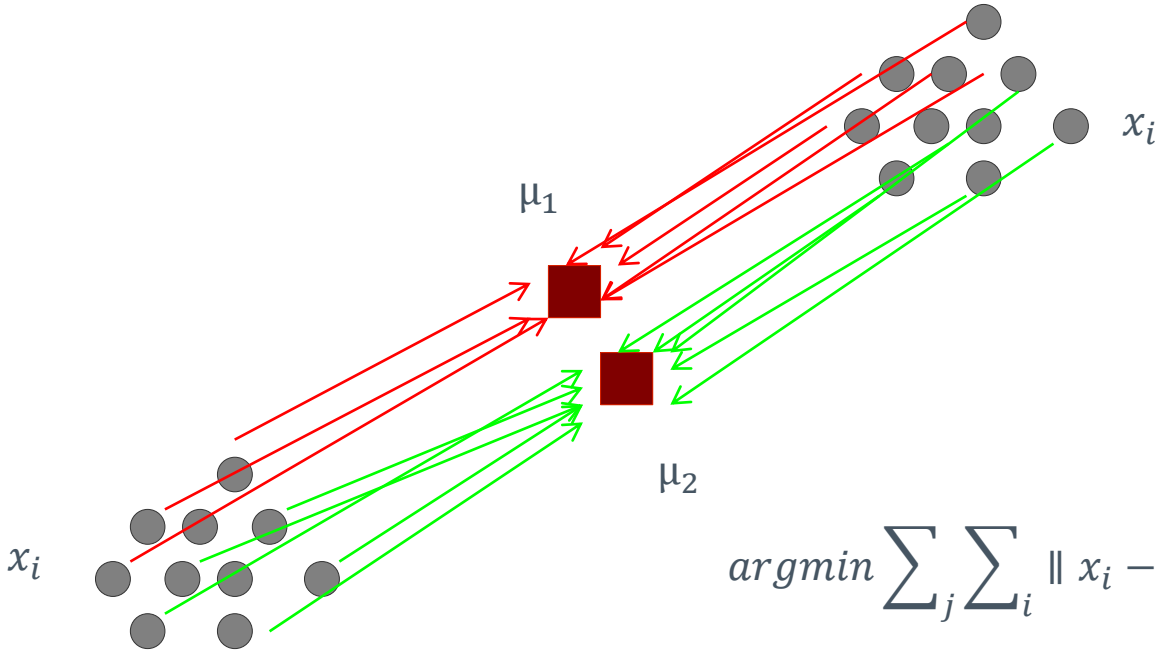
Reassign and repeat





$$argmin \sum_j \sum_i \| x_i - \mu_j \|$$

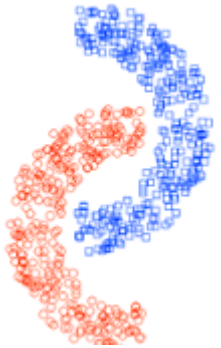
Problems with local optimum of the optimization function

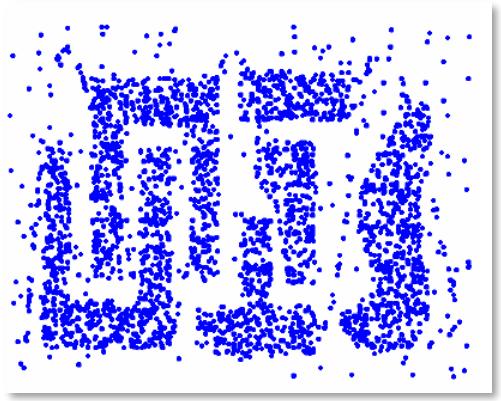


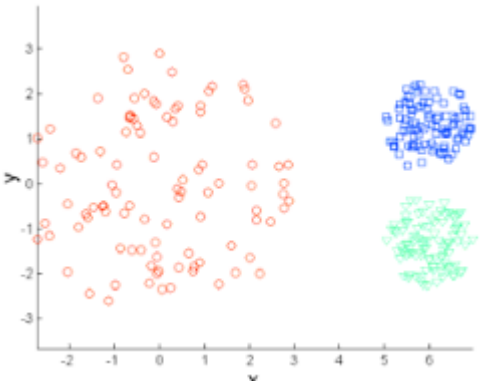
$$argmin \sum_j \sum_i \| x_i - \mu_j \|$$

Ερωτηση

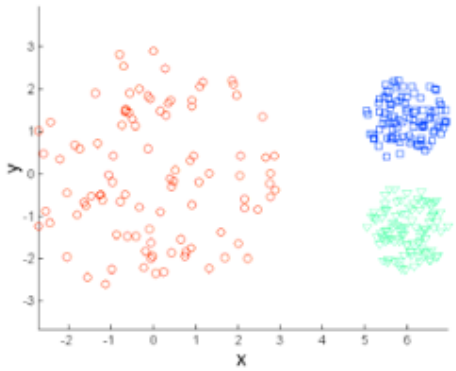
– Θα λειτουργούσε καλά ο k-means (εστω ότι δίναμε σωστό K)



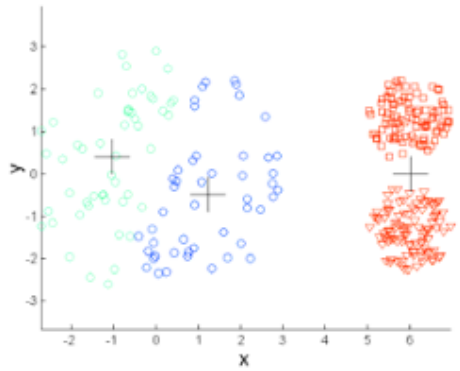




K-means: Περιορισμοί – Διαφορετικές Πυκνότητες



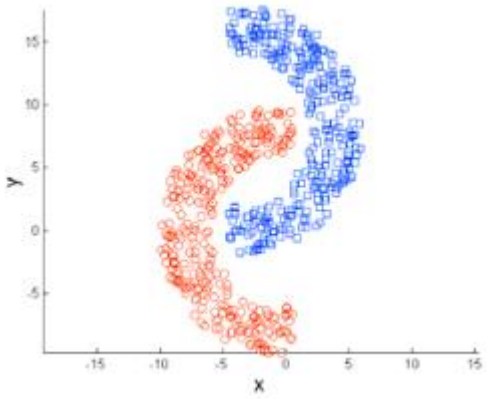
Αρχικά σημεία

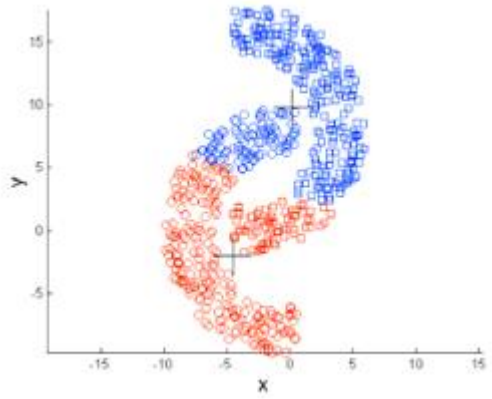


K-means (3 συστάδες)

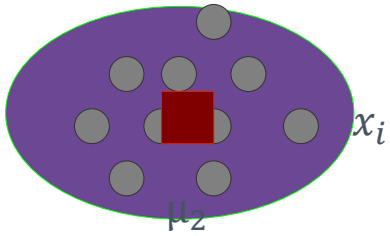
Δεν μπορεί να διαχωρίσει τους δυο μικρούς γιατί είναι πολύ πυκνοί σε σχέση με τον ένα μεγάλο

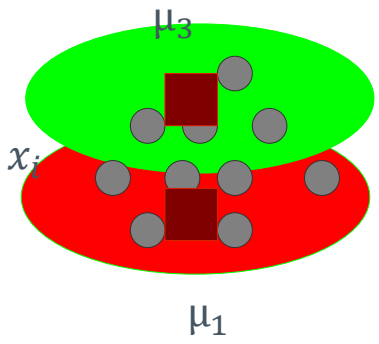
Δεν μπορεί να βρει τις δύο συστάδες γιατί έχουν μη κυκλικά σχήματα





Problems with wrong number of clusters

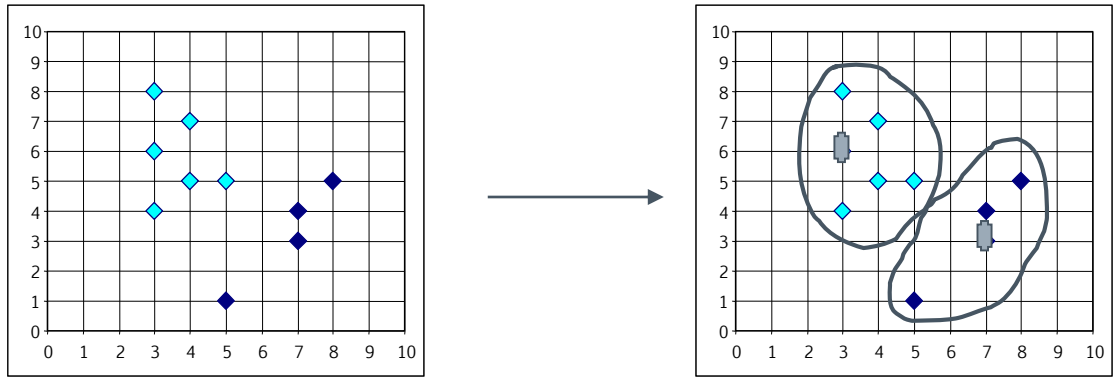




$$argmin \sum_j \sum_i \| x_i - \mu_j \|$$

K-medoid

- Συνήθως συνεχή d-διάστατο χώρο
- Διαλέγει ένα αντιπροσωπευτικό σημείο από τα δεδομένα και ελαχιστοποιεί την απόσταση από αυτό – Medoid: το πιο κεντρικό σημείο της συστάδας (αντί να χρησιμοποιεί το mean)
- Μειώνει την ευαισθησία σε outliers
- Μπορεί να εφαρμοστεί σε δεδομένα οποιουδήποτε τύπου (πχ και για κατηγορικά δεδομένα)

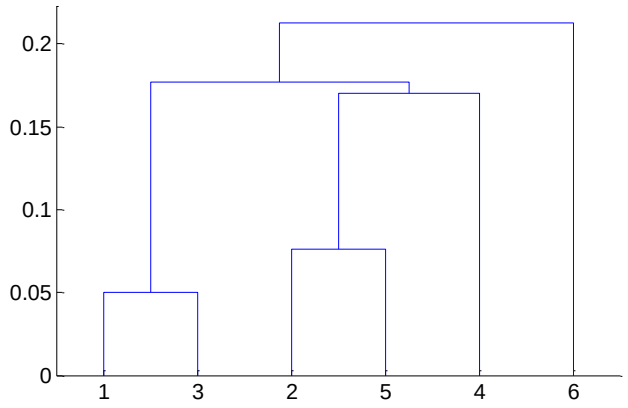


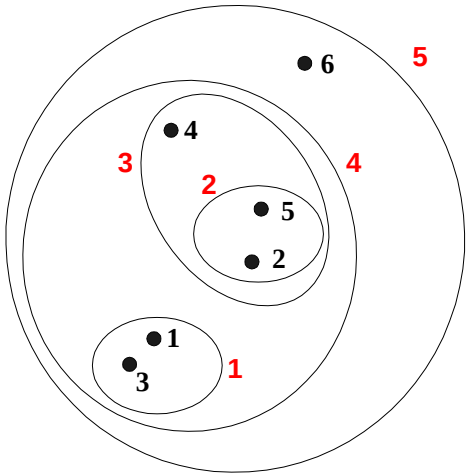
Ιεραρχική Συσταδοποίηση: Βασικά

Παράγει ένα σύνολο από εμφωλευμένες συστάδες οργανωμένες σε ένα ιεραρχικό δέντρο

Μπορεί να παρασταθεί με ένα δένδρο-γραμμα

Ένα διάγραμμα που μοιάζει με δένδρο και καταγράφει τις ακολουθίες από συγχωνεύσεις (merges) και διαχωρισμούς (splits)





Ιεραρχική Συσταδοποίηση: Πλεονεκτήματα

- Δε χρειάζεται να υποθέσουμε ένα συγκεκριμένο αριθμό από συστάδες

  Οποιοσδήποτε επιθυμητός αριθμός από συστάδες μπορεί να επιτευχθεί κόβοντας το δενδρόγραμμα στο κατάλληλο επίπεδο

- Μπορεί να αντιστοιχούν σε λογικές ταξινομήσεις

Για παράδειγμα στις βιολογικές επιστήμες (ζωικό βασίλειο, phylogeny reconstruction, …)

Ιεραρχική Συσταδοποίηση

Δυο βασικοί τύποι ιεραρχικής συσταδοποίησης

- Συσσωρευτικός (Agglomerative):
  - Αρχίζει με τα σημεία ως ξεχωριστές συστάδες
  - Σε κάθε βήμα, συγχωνεύει το πιο κοντινό ζευγάρι συστάδων μέχρι να μείνει μόνο μία (ή k) συστάδες

- Διαιρετικός (Divisive):
  - Αρχίζει με μία συστάδα που περιέχει όλα τα σημεία
  - Σε κάθε βήμα, διαχωρίζει μία συστάδα, έως κάθε συστάδα να περιέχει μόνο ένα σημείο (ή να δημιουργηθούν k συστάδες)

Ιεραρχική Συσταδοποίηση

Οι παραδοσιακοί αλγόριθμοι

- χρησιμοποιούν έναν **πίνακα** ομοιότητα ή απόστασης
  - διαχωρισμός ή συγχώνευση *μιας* ομάδας τη φορά

Συσσωρευτική Ιεραρχική Συσταδοποίηση (ΣΙΣ)

Η πιο δημοφιλής τεχνική συσταδοποίησης

Βασικός Αλγόριθμος

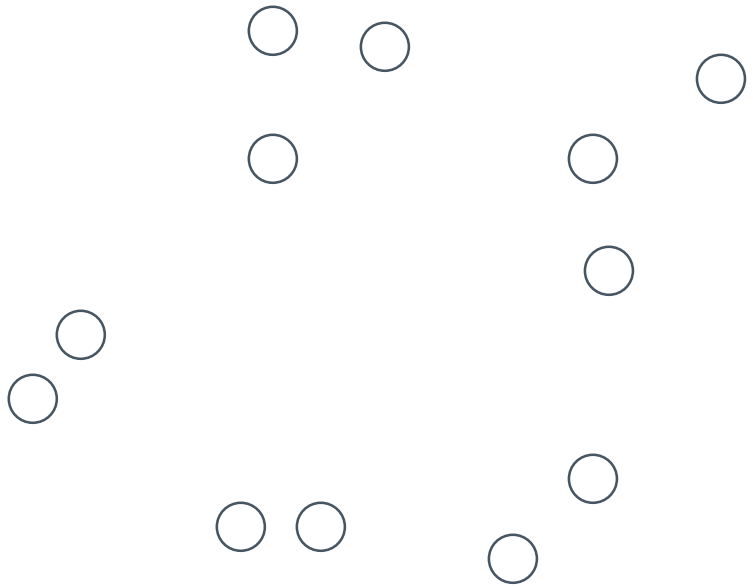
1: Υπολογισμός του Πίνακα Γειτνίασης
2: Έστω κάθε σημείο αποτελεί και μια συστάδα
3: **Repeat**
4: Συγχώνευση των δύο κοντινότερων συστάδων
5: Ενημέρωση του Πίνακα Γειτνίασης
6: **Until** να μείνει μία μόνο συστάδα

Βασική λειτουργία είναι ο υπολογισμός της γειτνίασης δυο συστάδων

Διαφορετικοί αλγόριθμοι με βάση το πως ορίζεται η απόσταση ανάμεσα σε δύο συστάδες

Συσσωρευτική Ιεραρχική Συσταδοποίηση

Αρχικά: Κάθε σημείο και συστάδα και ένας Πίνακας Γειτνίασης (proximity matrix)

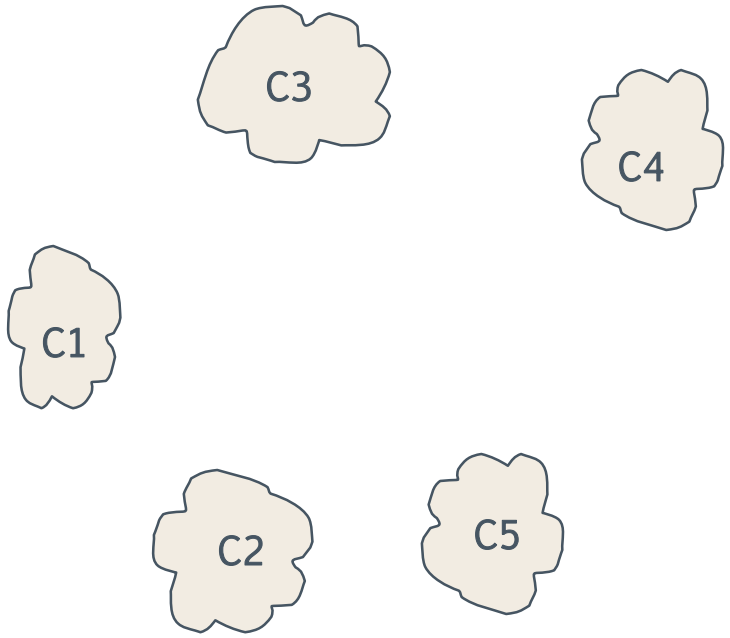


	p1	p2	p3	p4	p5	. .
p1						.
p2						
p3						
p4						
p5						
.						
.						
.						

Πίνακας Γειτνίασης

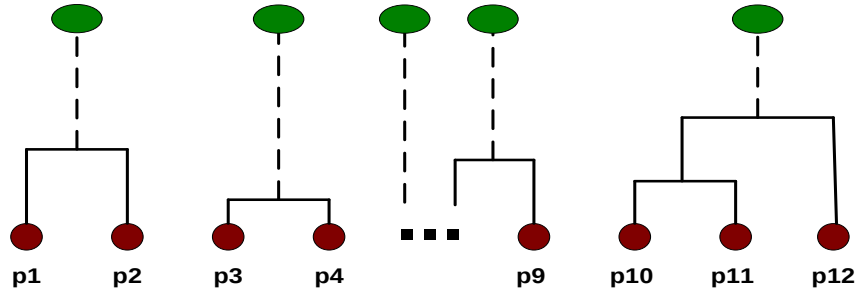
Συσσωρευτική Ιεραρχική Συσταδοποίηση

Μετά από κάποιες συγχωνεύσεις, έχουμε κάποιες συστάδες



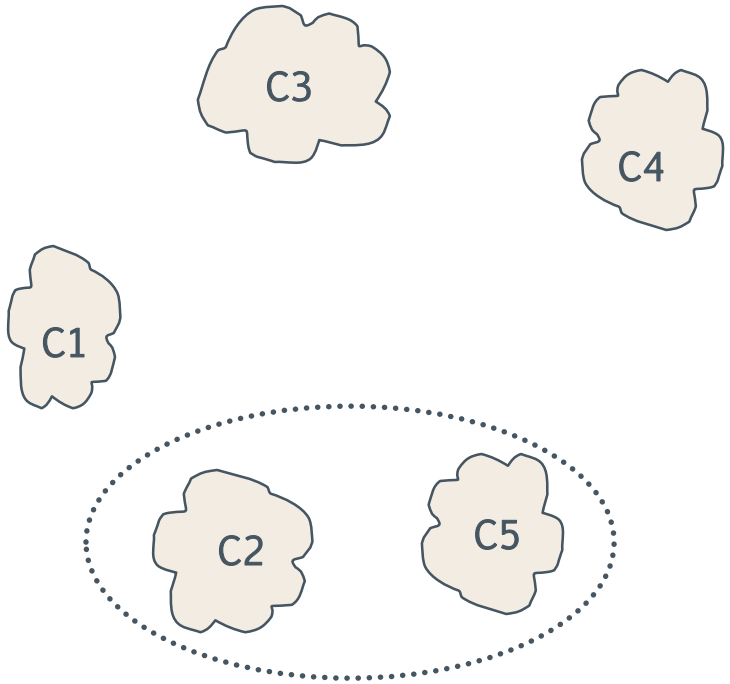
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Πίνακας Γειτνίασης



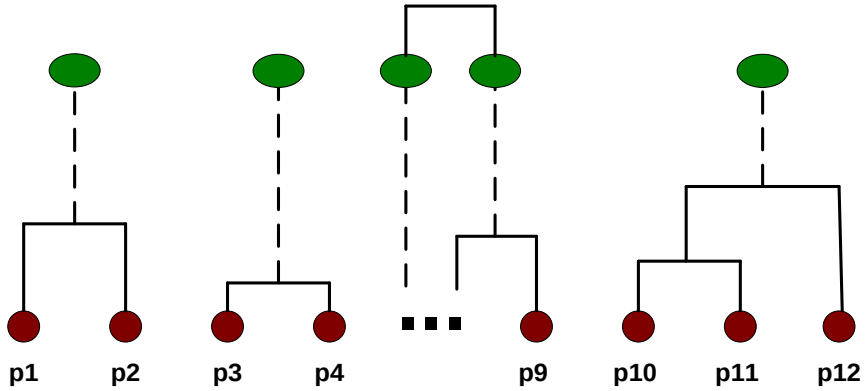
Συσσωρευτική Ιεραρχική Συσταδοποίηση

Θέλουμε να συγχωνεύσουμε τις δύο κοντινότερες συστάδες (C2 και C5) και να ενημερώσουμε τον πίνακα γειτνίασης.



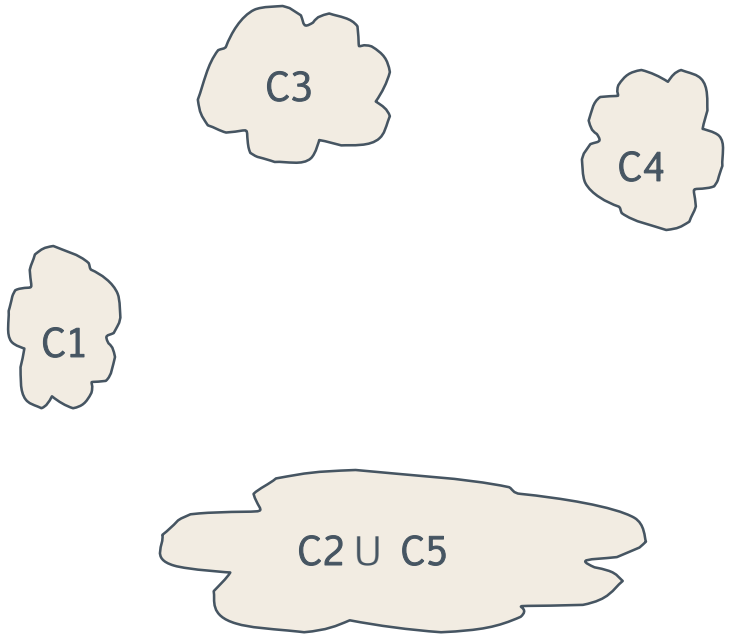
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Πίνακας Γειτνίασης



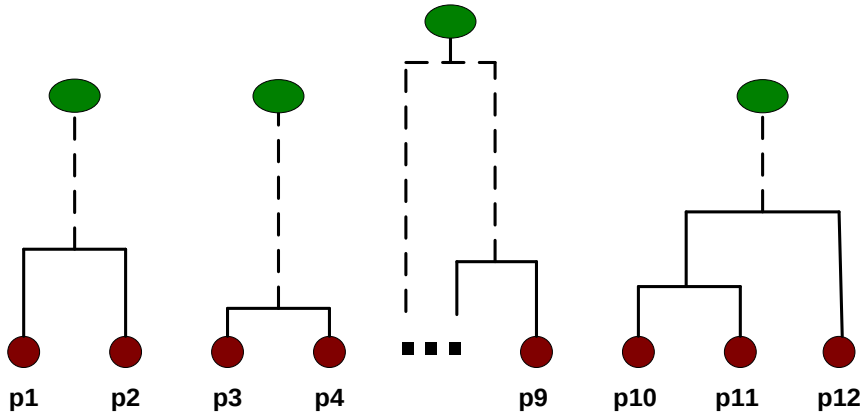
Συσσωρευτική Ιεραρχική Συσταδοποίηση

Μετά τη συγχώνευση η ερώτηση είναι:
Πως ενημερώνουμε τον πίνακα γειτνίασης

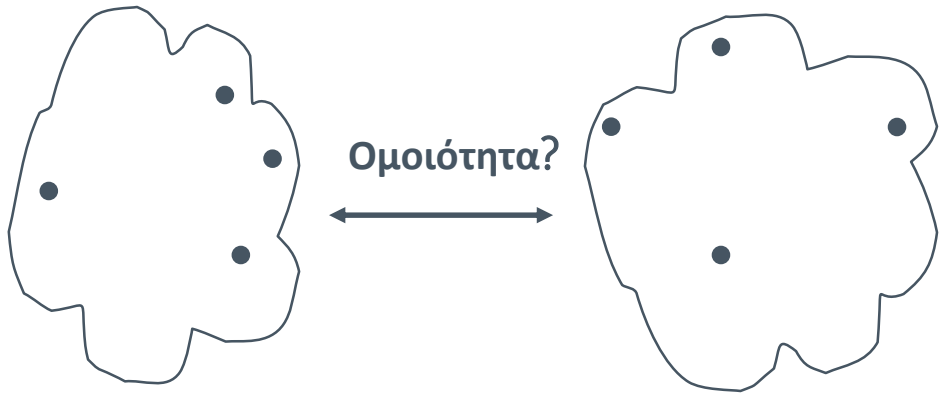


	C1	C2 U C5	C3	C4
C1		?		
C2 U C5	? ?		?	?
C3		?		
C4		?		

Πίνακας Γειτνίασης



ΣΙΣ: Ορισμός απόστασης μεταξύ συστάδων

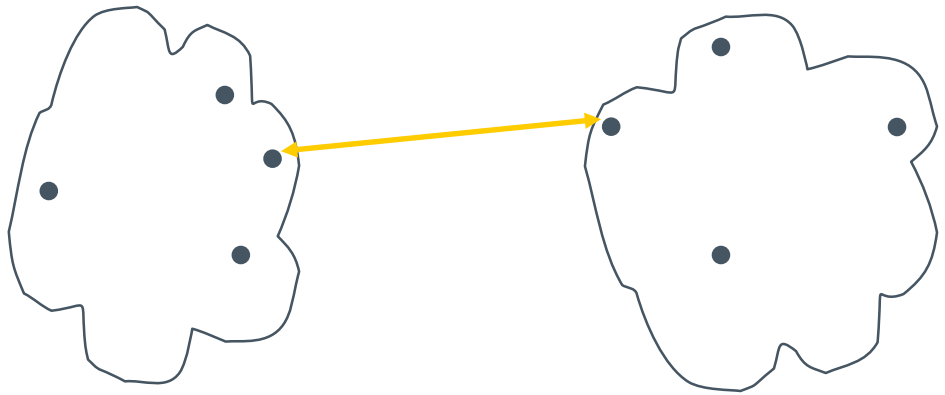


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Πίνακας Γειτνίασης

- MIN
- MAX
- Μέσος όρος της συστάδας
- Η απόσταση μεταξύ των κεντρικών σημείων
- Άλλες μέθοδοι βασισμένες σε μια αντικειμενική συνάρτηση
  - Η μέθοδος του Ward χρησιμοποιεί τετραγωνικά λάθη

ΣΙΣ: Ορισμός απόστασης μεταξύ συστάδων



- **MIN**
- MAX
- Μέσος όρος της ομάδας
- Η απόσταση μεταξύ των κεντρικών σημείων
- Άλλες μέθοδοι βασισμένες σε μια αντικειμενική συνάρτηση
  - Η μέθοδος του Ward χρησιμοποιεί τετραγωνικά λάθη

	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Πίνακας Γειτνίασης

ΣΙΣ: Ορισμός απόστασης μεταξύ συστάδων: MIN

MIN ή μοναδικής ακμής ή απλού συνδέσμου (single link)

Η ομοιότητα μεταξύ δυο συστάδων βασίζεται στα δυο πιο όμοια (πιο γειτονικά) σημεία στις διαφορετικές συστάδες (με όρους γραφημάτων – shortest edge)

Καθορίζεται από ένα ζεύγος τιμών, δηλαδή μια ακμή (link) του γραφήματος γειτνίασης.

Ονομάζεται και μέθοδος συσταδοποίησης κοντινότερου γείτονα

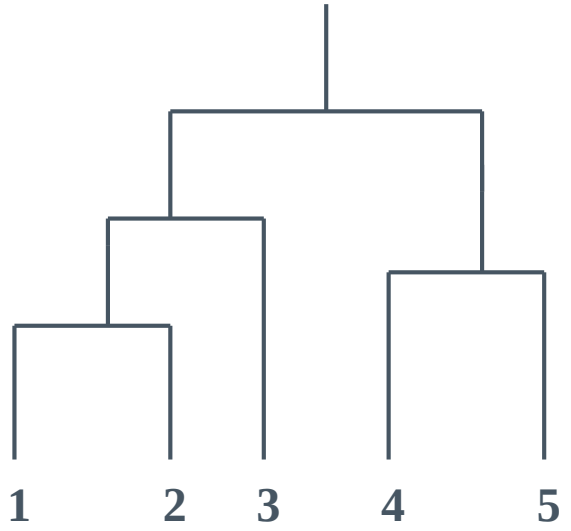
ΣΙΣ: Ορισμός απόστασης μεταξύ συστάδων: MIN

MIN ή μοναδικής ακμής ή απλού συνδέσμου (single link)

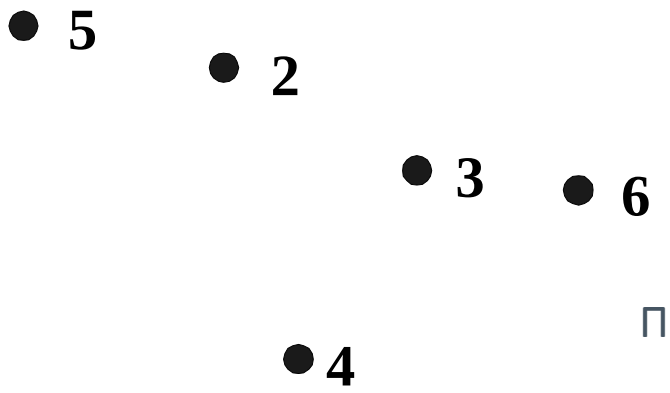
Η ομοιότητα μεταξύ δυο συστάδων βασίζεται στα δυο πιο όμοια (πιο γειτονικά) σημεία στις διαφορετικές συστάδες (με όρους γραφημάτων – shortest edge)

Καθορίζεται από ένα ζεύγος τιμών, δηλαδή μια ακμή (link) του γραφήματος γειτνίασης.

	I1	I2	I3	I4	I5
I1	1,00	0,90	0,10	0,65	0,20
I2	0,90	1,00	0,70	0,60	0,50
I3	0,10	0,70	1,00	0,40	0,30
I4	0,65	0,60	0,40	1,00	0,80
I5	0,20	0,50	0,30	0,80	1,00



Προσοχή: ομοιότητα –> τα ποιο όμοια



1 (0.4, 0.53)

2 (0.22, 0.38)

3 (0.35, 0.32)

4 (0.26, 0.19)

5 (0.08, 0.41)

6 (0.45, 0.30)

Πίνακας απόστασης (Ευκλείδεια)

	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

1 (0.4, 0.53)

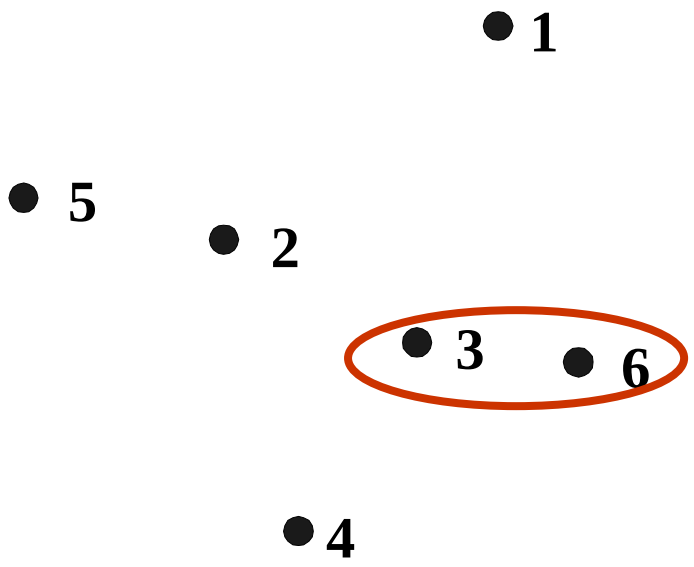
2 (0.22, 0.38)

3 (0.35, 0.32)

4 (0.26, 0.19)

5 (0.08, 0.41)

6 (0.45, 0.30)



Καθορίζεται μόνο από μια ακμή – την μικρότερη

	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

1 (0.4, 0.53)

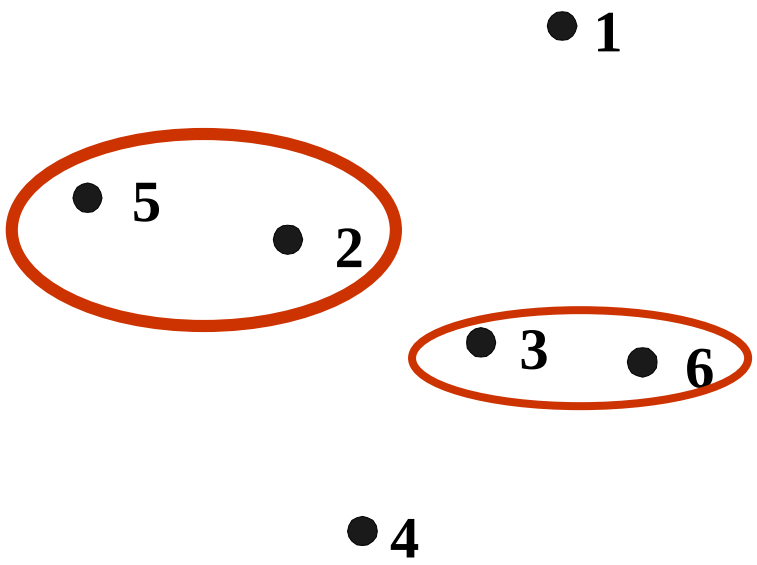
2 (0.22, 0.38)

3 (0.35, 0.32)

4 (0.26, 0.19)

5 (0.08, 0.41)

6 (0.45, 0.30)



	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

1 (0.4, 0.53)

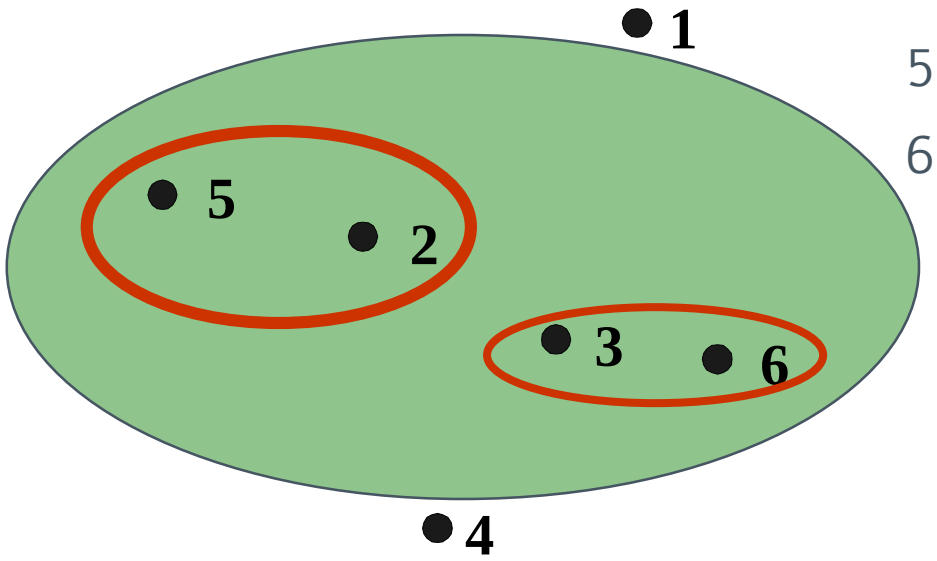
2 (0.22, 0.38)

3 (0.35, 0.32)

4 (0.26, 0.19)

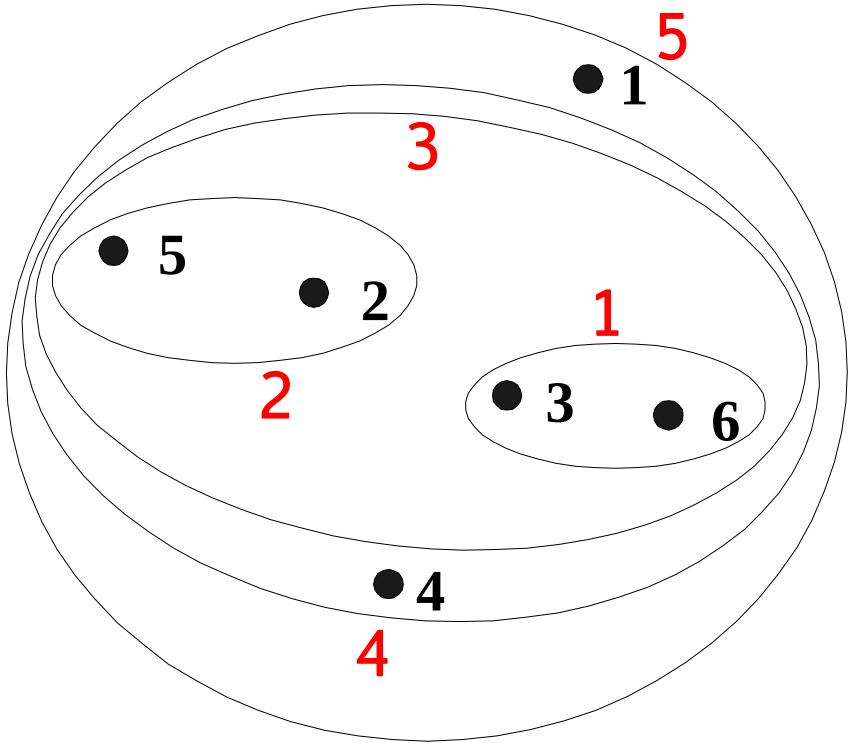
5 (0.08, 0.41)

6 (0.45, 0.30)

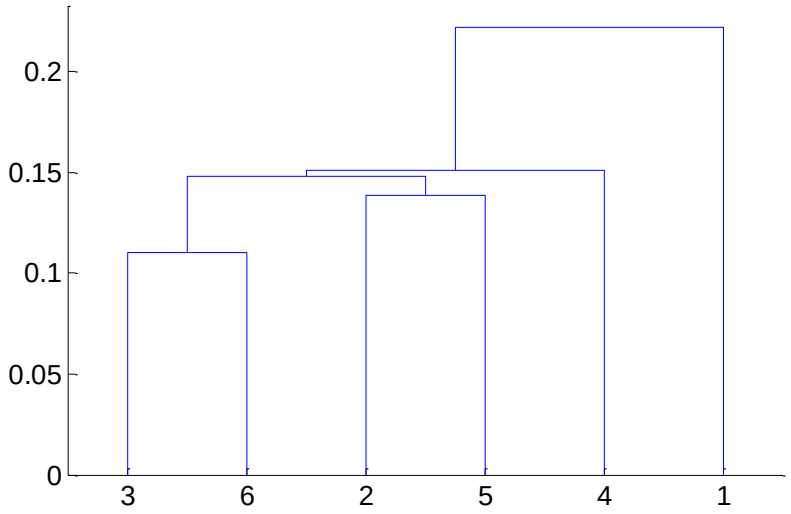


	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	**0.24**	**0.00**	**0.15**	**0.20**	0.14	**0.25**
p3	**0.22**	0.15	**0.00**	**0.15**	**0.28**	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	**0.00**	0.39
p6	0.23	0.25	0.11	0.22	0.39	**0.00**

ΣΙΣ: Ορισμός απόστασης μεταξύ συστάδων: MIN



Φωλιασμένες Συστάδες

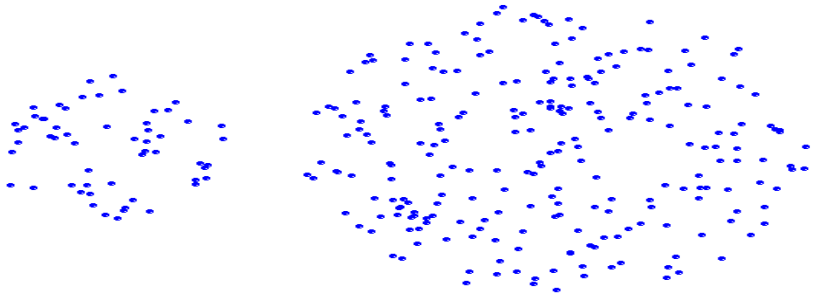


Δεντρόγραμμα

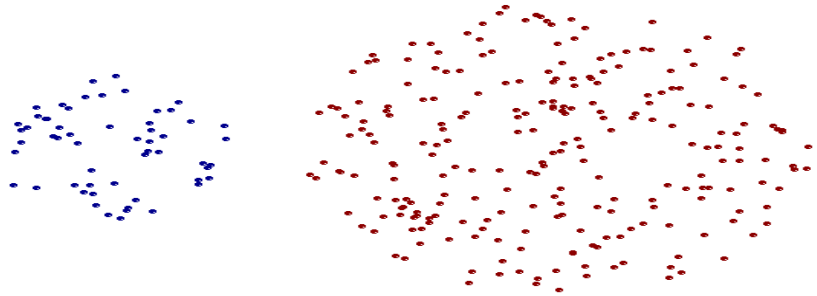
Το δεντρόγραμμα (y-άξονας) δίνει και τις αποστάσεις

ΣΙΣ: Ορισμός απόστασης μεταξύ συστάδων: MIN

Προτερήματα



Αρχικά σημεία



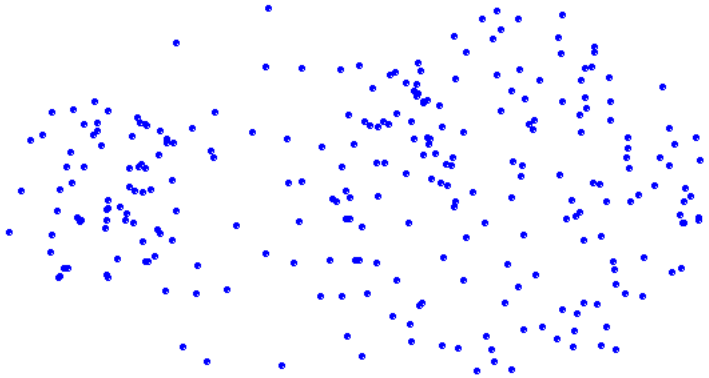
Δύο συστάδες

Contiguity-based (συνεχόμενες συστάδες)

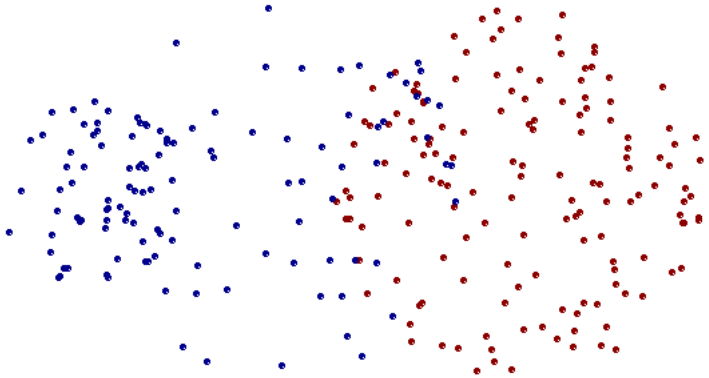
Μπορεί να χειριστεί μη ελλειπτικά (non-elliptical) σχήματα

ΣΙΣ: Ορισμός απόστασης μεταξύ συστάδων: MIN

Μειονεκτήματα



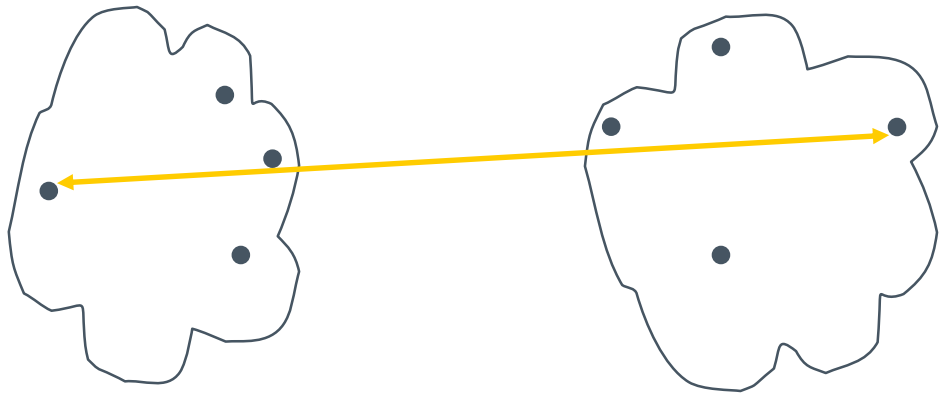
Αρχικά σημεία



Δύο συστάδες

- **Ευαίσθητο σε θόρυβο και outliers**

ΣΙΣ: Ορισμός απόστασης μεταξύ συστάδων

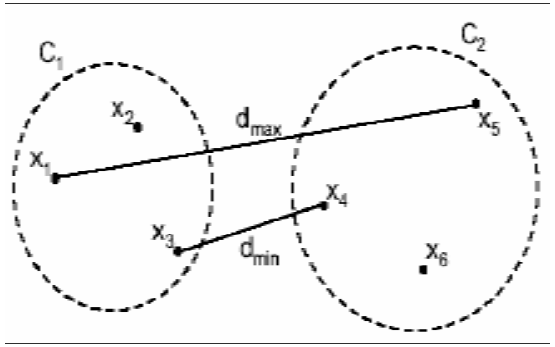


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						

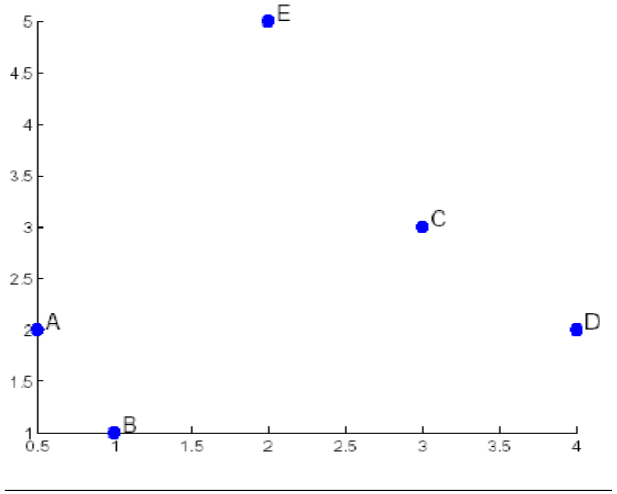
- MIN
- **MAX**
- Μέσος όρος της ομάδας
- Η απόσταση μεταξύ των κεντρικών σημείων
- Άλλες μέθοδοι βασισμένες σε μια αντικειμενική συνάρτηση
  - Η μέθοδος του Ward χρησιμοποιεί τετραγωνικά λάθη

· **Πίνακας Γειτνίασης**

·



ΣΙΣ: Ορισμός απόστασης μεταξύ συστάδων



0	**1.1180**	**2.6926**	**3.5**	**3.3541**
1.1180	**0**	**2.8282**	**3.1623**	**4.1231**
2.6926	**2.8284**	**0**	**1.4142**	**2.2361**
3.5	**3.1623**	**1.4142**	**0**	**3.6056**
3.3541	**4.1231**	**2.2361**	**3.6056**	**0**

ΣΙΣ: Ορισμός απόστασης μεταξύ συστάδων: MAX

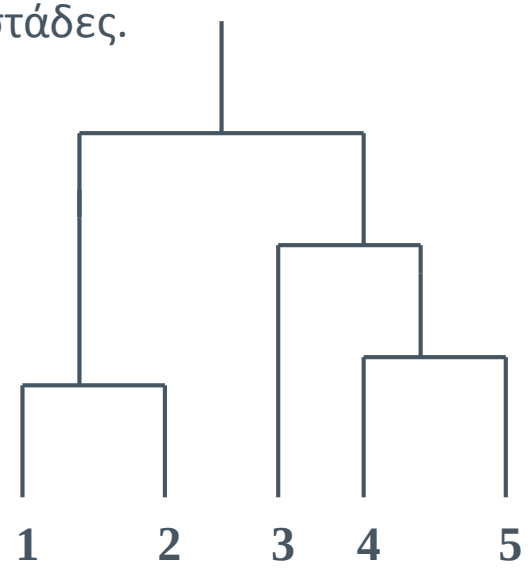
MAX ή πλήρους συνδεσιμότητας (complete linkage)
- Αναζητά κλίκες

Η ομοιότητα μεταξύ δυο συστάδων βασίζεται στα δυο λιγότερο όμοια (πιο μακρινά) σημεία στις διαφορετικές συστάδες (longest edge) – δηλαδή, οι συστάδες με την μικρότερη τέτοια απόσταση

Καθορίζεται από όλα τα ζεύγη τιμών στις δύο συστάδες.

	I1	I2	I3	I4	I5
I1	1.00	0.90	0.10	0.65	0.20
I2	0.90	1.00	0.70	0.60	0.50
I3	0.10	0.70	1.00	0.40	0.30
I4	0.65	0.60	0.40	1.00	0.80
I5	0.20	0.50	0.30	0.80	1.00

ομοιότητα



1 (0.4, 0.53)

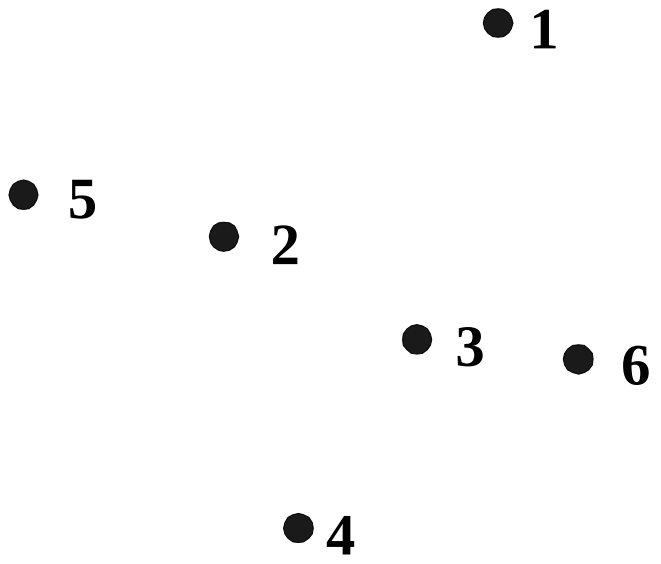
2 (0.22, 0.38)

3 (0.35, 0.32)

4 (0.26, 0.19)

5 (0.08, 0.41)

6 (0.45, 0.30)



	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

1 (0.4, 0.53)

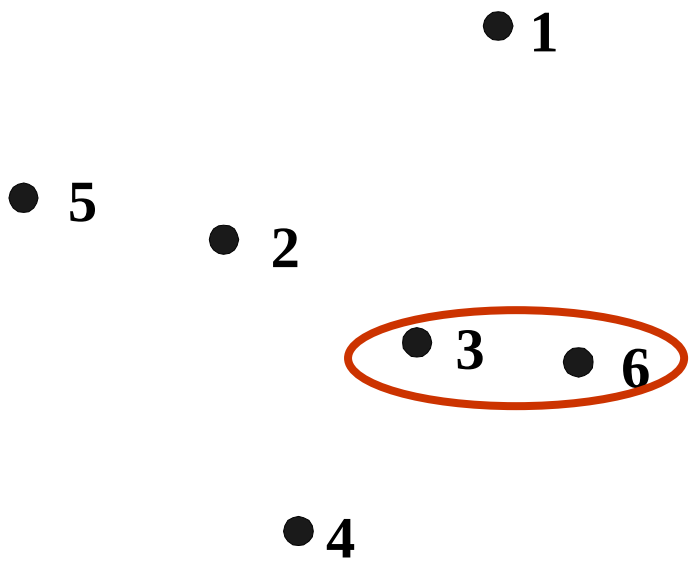
2 (0.22, 0.38)

3 (0.35, 0.32)

4 (0.26, 0.19)

5 (0.08, 0.41)

6 (0.45, 0.30)



	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

1 (0.4, 0.53)

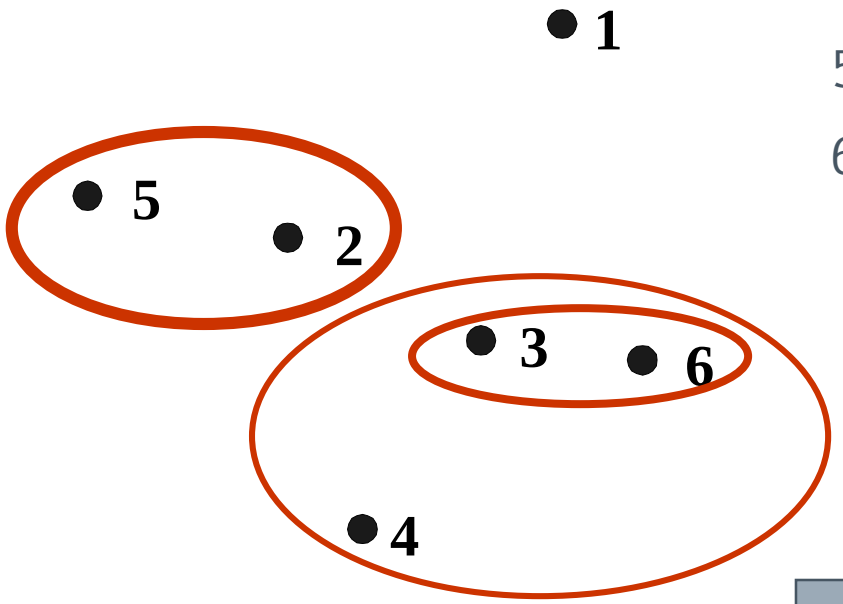
2 (0.22, 0.38)

3 (0.35, 0.32)

4 (0.26, 0.19)

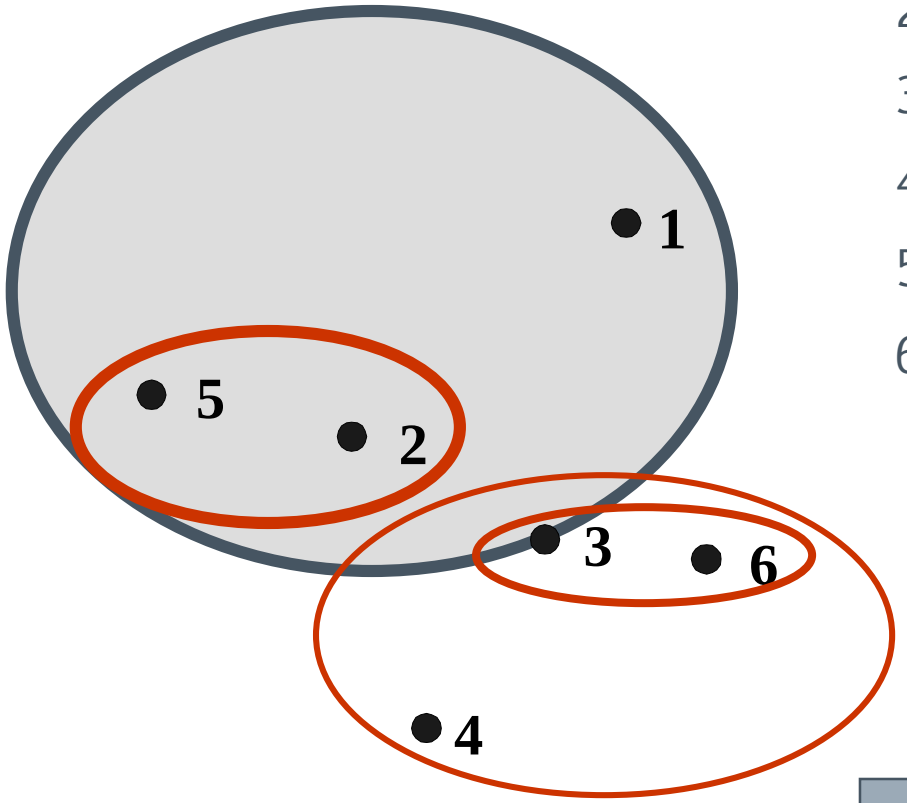
5 (0.08, 0.41)

6 (0.45, 0.30)



	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	**0.00**	0.15	0.20	0.14	**0.25**
p3	**0.22**	0.15	**0.00**	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	**0.34**	0.14	**0.28**	**0.29**	**0.00**	0.39
p6	**0.23**	**0.25**	0.11	**0.22**	**0.39**	**0.00**

π



1 (0.4, 0.53)

2 (0.22, 0.38)

3 (0.35, 0.32)

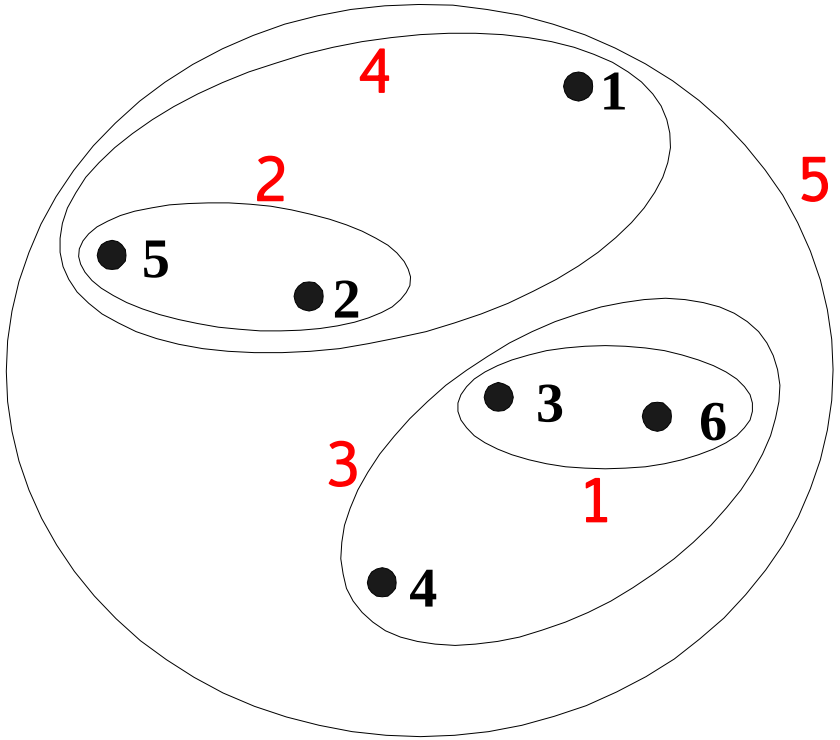
4 (0.26, 0.19)

5 (0.08, 0.41)

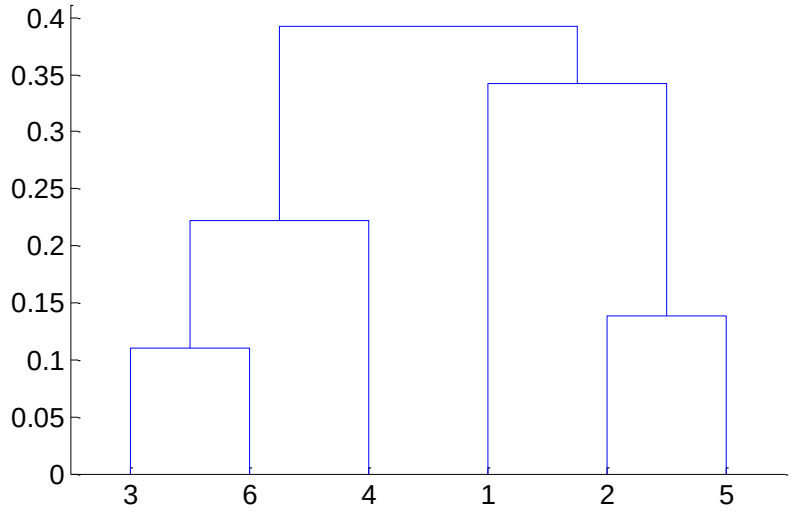
6 (0.45, 0.30)

	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	**0.00**	0.15	0.20	0.14	**0.25**
p3	0.22	0.15	**0.00**	0.15	0.28	0.11
p4	**0.37**	0.20	0.15	0.00	0.29	0.22
p5	**0.34**	0.14	**0.28**	**0.29**	**0.00**	0.39
p6	0.23	**0.25**	0.11	**0.22**	**0.39**	**0.00**

ΣΙΣ: Ορισμός απόστασης μεταξύ συστάδων: MAX



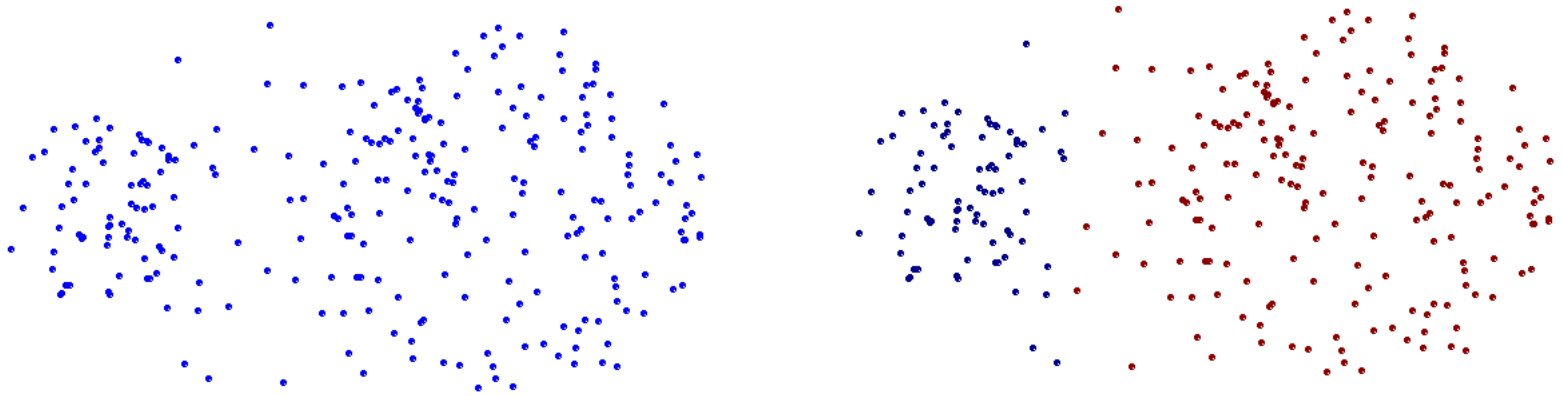
Φωλιασμένες Συστάδες



Δεντρόγραμμα

ΣΙΣ: Ορισμός απόστασης μεταξύ συστάδων: MAX

Πλεονεκτήματα

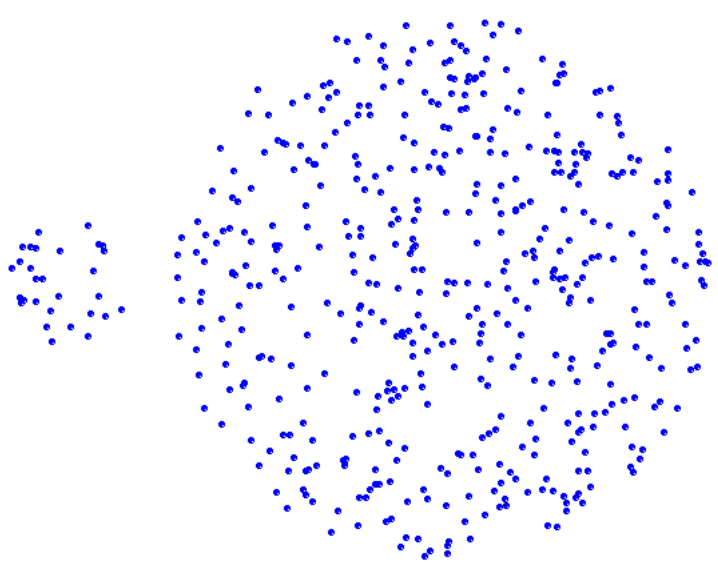


Αρχικά Σημεία **Δύο Συστάδες**

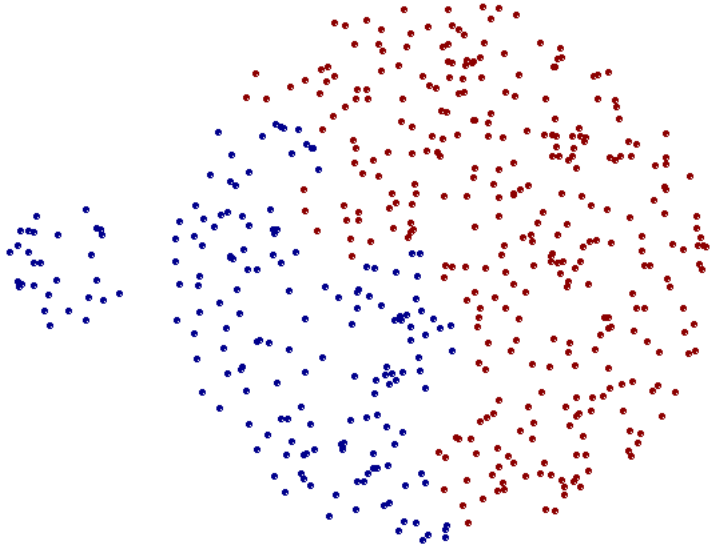
- **λιγότερη εξάρτηση σε θόρυβο και outliers**

ΣΙΣ: Ορισμός απόστασης μεταξύ συστάδων: MAX

Μειονεκτήματα



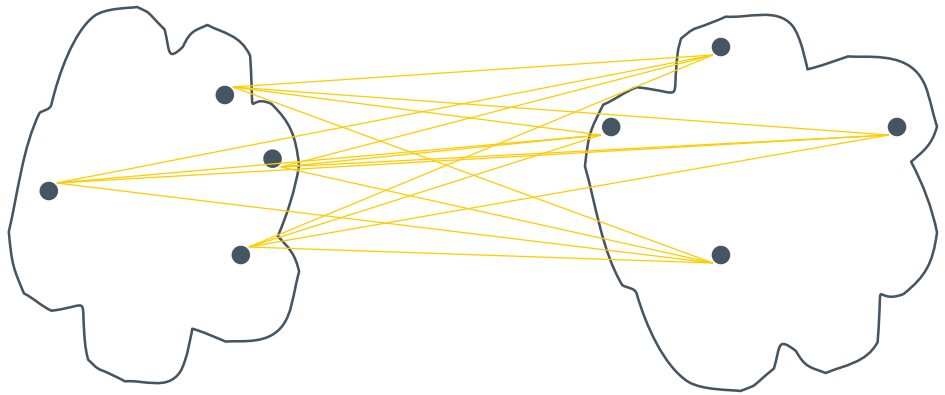
Αρχικά σημεία



Δύο συστάδες

- **Τείνει να διασπά μεγάλες συστάδες**
- **Οδηγεί συνήθως σε κυκλικά σχήματα**

ΣΙΣ: Ορισμός απόστασης μεταξύ συστάδων



	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						

Πίνακας Γειτνίασης

- MIN
- MAX
- **Μέσος όρος της ομάδας (group average)**
- Η απόσταση μεταξύ των κεντρικών σημείων
- Άλλες μέθοδοι βασισμένες σε μια αντικειμενική συνάρτηση
  - Η μέθοδος του Ward χρησιμοποιεί τετραγωνικά λάθη

ΣΙΣ: Ορισμός απόστασης μεταξύ συστάδων: Μέσο Ομάδας

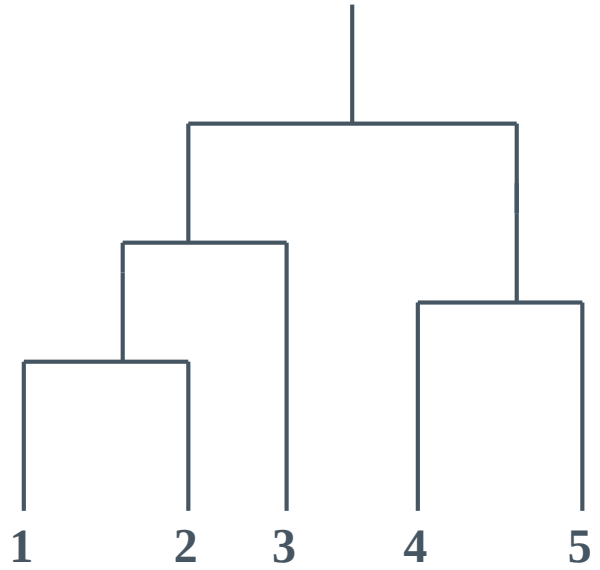
- Κοντινότητα δύο συστάδων είναι η μέση τιμή της ανα-δύο κοντινότητας (average of pairwise proximity) μεταξύ των σημείων των δύο συστάδων.

$$\text{proximity}(\text{Cluster}_i, \text{Cluster}_j) = \frac{\sum_{\substack{p_i \in \text{Cluster}_i \\ p_j \in \text{Cluster}_j}} \text{proximity}(p_i, p_j)}{|\text{Cluster}_i| * |\text{Cluster}_j|}$$

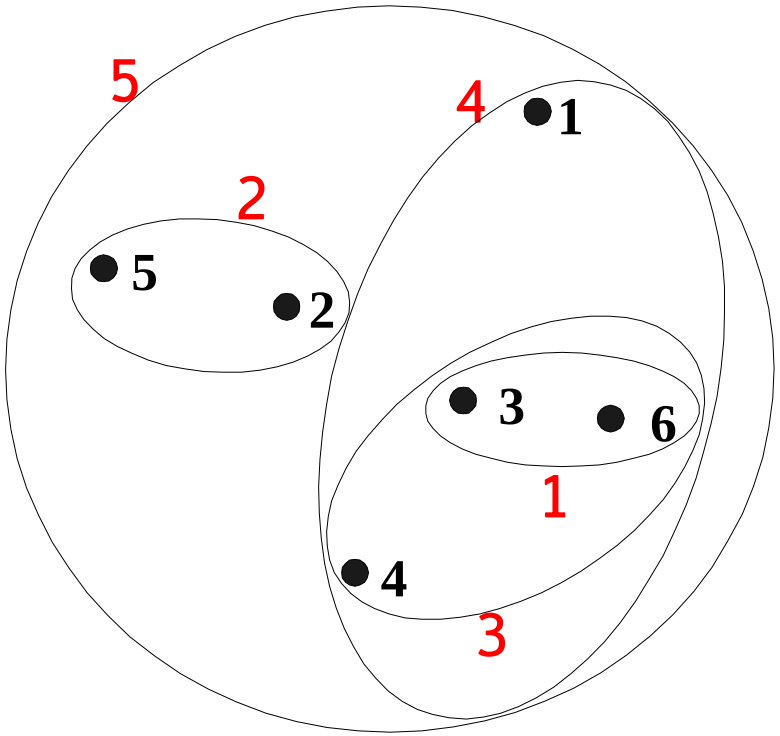
- Χρήση μέσης γιατί η ολική θα έδινε προτίμηση στις μεγάλες συστάδες

ομοιότητα

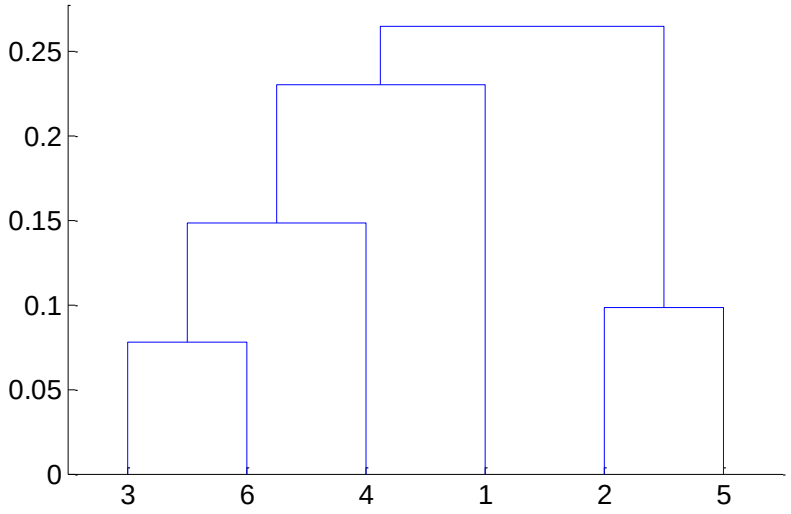
	I1	I2	I3	I4	I5
I1	1.00	0.90	0.10	0.65	0.20
I2	0.90	1.00	0.70	0.60	0.50
I3	0.10	0.70	1.00	0.40	0.30
I4	0.65	0.60	0.40	1.00	0.80
I5	0.20	0.50	0.30	0.80	1.00



ΣΙΣ: Ορισμός απόστασης μεταξύ συστάδων: Μέσο Ομάδας



Φωλιασμένες Συστάδες

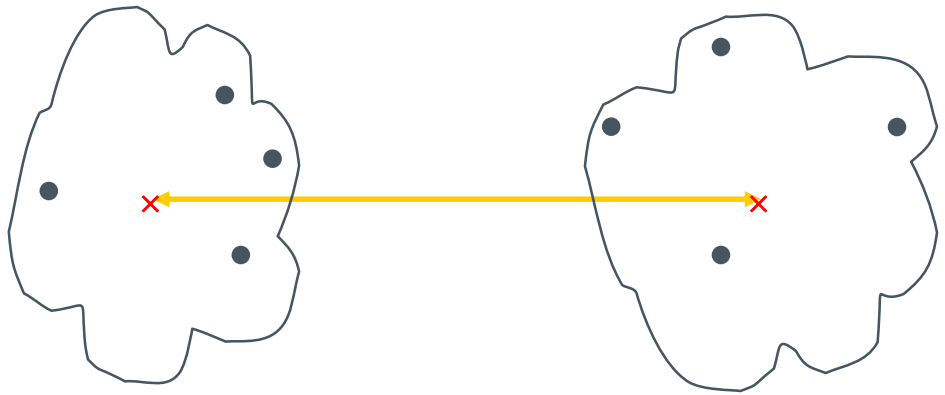


Dendrogram

ΣΙΣ: Ορισμός απόστασης μεταξύ συστάδων: Μέσο Ομάδας

- Ανάμεσα σε MIN-MAX
- Πλεονεκτήματα: μικρότερη ευαισθησία σε θόρυβο και outliers
- Μειονεκτήματα: Ευνοεί κυκλικές συστάδες

ΣΙΣ: Ορισμός απόστασης μεταξύ συστάδων



	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						

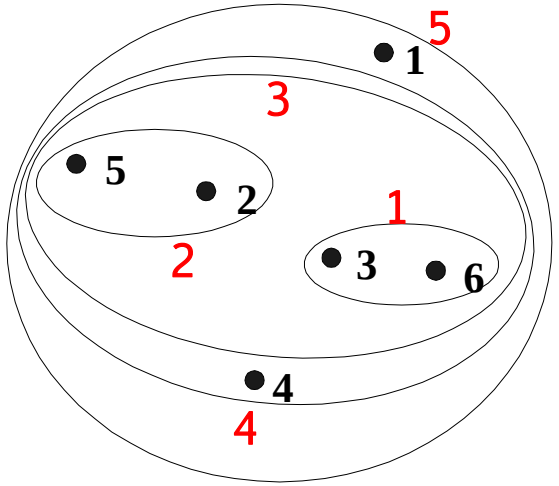
- MIN
- MAX
- Μέσος όρος της ομάδας
- **Η απόσταση μεταξύ των κεντρικών σημείων**
- Άλλες μέθοδοι βασισμένες σε μια αντικειμενική συνάρτηση
  - Η μέθοδος του Ward χρησιμοποιεί τετραγωνικά λάθη

· **Πίνακας Γειτνίασης**

Πρόβλημα: μη μονότονη αύξηση της απόστασης

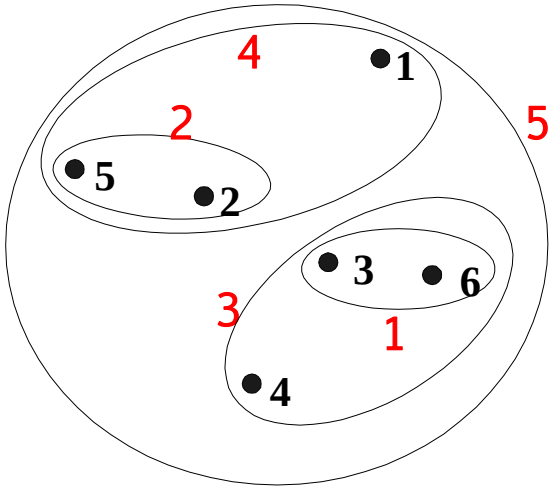
Δηλαδή, δυο συστάδες που συγχωνεύονται μπορεί να έχουν μικρότερη απόσταση από συστάδες που έχουν συγχωνευτεί σε προηγούμενα βήματα

ΣΙΣ: Ορισμός απόστασης μεταξύ συστάδων: Σύγκριση

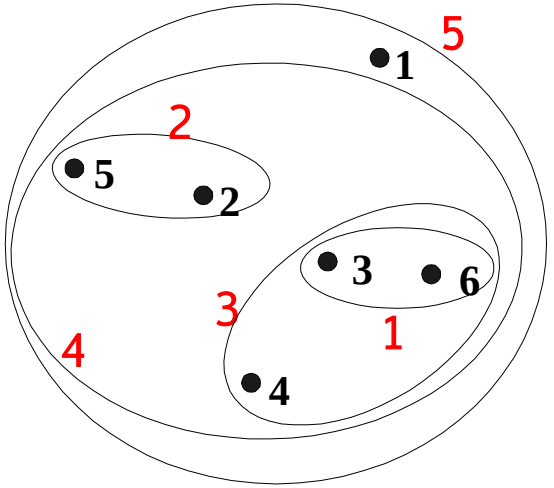


MIN

MAX



Μέθοδος του Ward



Μέσο Ομάδας