Συσταδοποίηση Ι

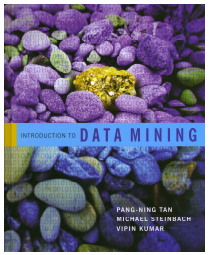
Εισαγωγή

Ο αλγόριθμος k-means

Αποστάσεις

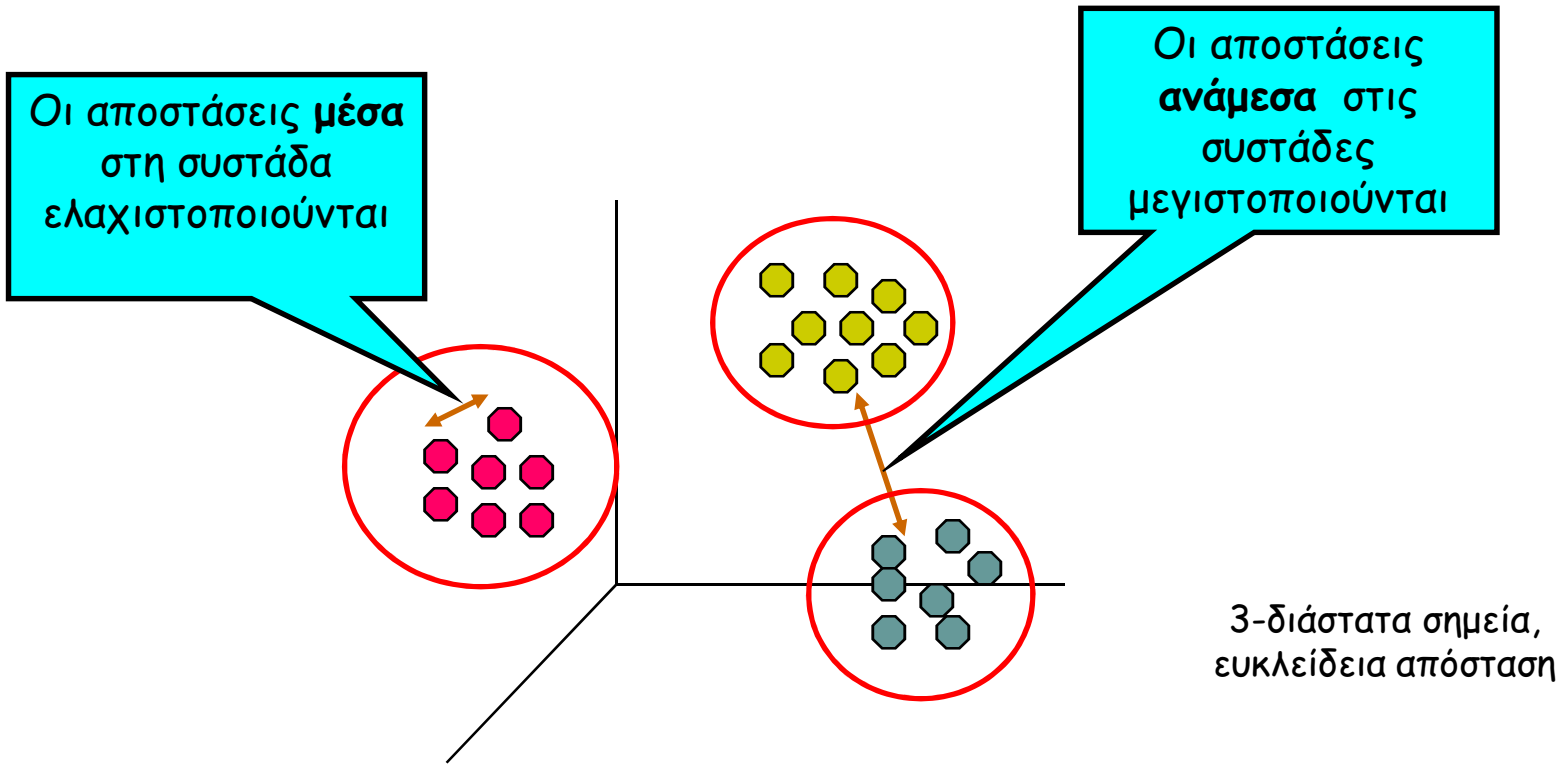
Ιεραρχική Συσταδοποίηση

Μέρος των διαφανειών είναι από το **P.-N. Tan, M.Steinbach, V. Kumar, «Introduction to Data Mining»**, Addison Wesley, 2006



Τι είναι συσταδοποίηση

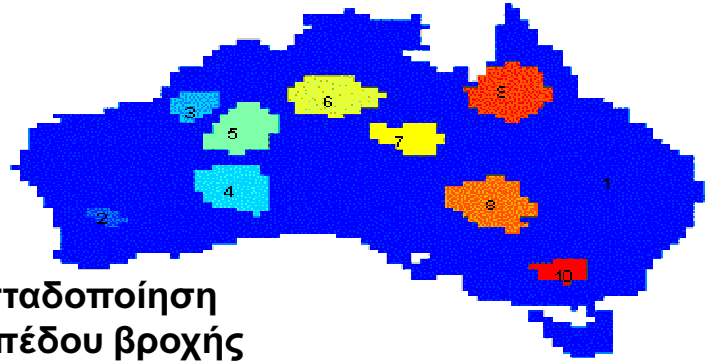
Εύρεση συστάδων (ομάδων) αντικειμένων έτσι ώστε τα αντικείμενα σε κάθε συστάδα να είναι όμοια (ή να σχετίζονται) και διαφορετικά (ή μη σχετιζόμενα) από τα αντικείμενα των άλλων συστάδων



Εφαρμογές

- ομαδοποίηση γονιδίων και πρωτεϊνών που έχουν την ίδια λειτουργία,
- εικόνες,
- χαρακτηριστικά ασθενειών
- μετοχών με παρόμοια διακύμανση τιμών,
- ομαδοποίηση weblog για εύρεση παρόμοιων προτύπων προσπέλασης,
- ομαδοποίηση σχετιζόμενων αρχείων για browsing,
- ομαδοποίηση κειμένων
- πελάτες με παρόμοια συμπεριφορά

	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



Συσταδοποίηση επιπέδου βροχής (precipitation) στην Αυστραλία!

Εφαρμογές



Κατανόηση – Stand-alone εφαρμογή/εργαλείο
οπτικοποίηση, συμπεράσματα για την κατανομή

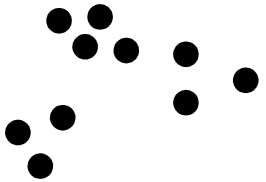
Βήμα Προεπεξεργασίας

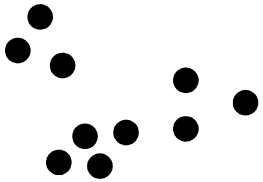
- Περίληψη: Ελάττωση του μεγέθους μεγάλων συνόλων χρήση αντιπροσωπευτικών σημείων από κάθε συστάδα – πρωτότυπα (prototypes),

- Συμπίεση ή

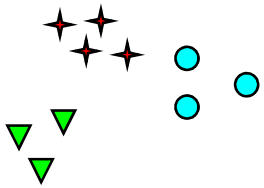
- Αποδοτική κατασκευή ευρετηρίων – εύρεση κοντινότερου γείτονα κλπ

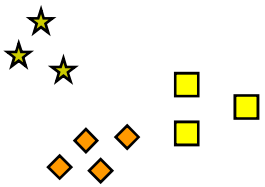
Ασάφεια



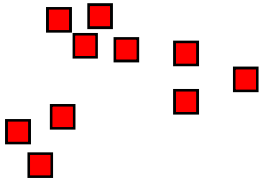


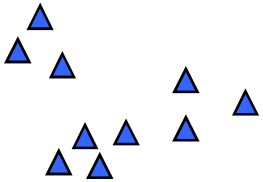
Πόσες Ομάδες?



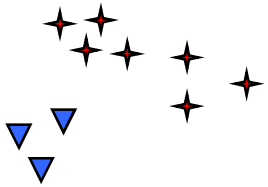


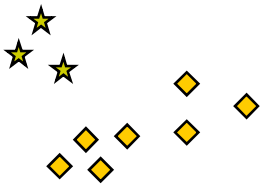
6 ομάδες





2 ομάδες





4 ομάδες



Πότε μια συσταδοποίηση είναι καλή;

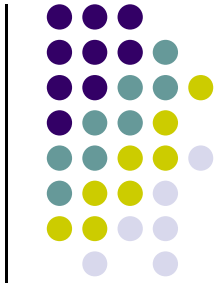
Μια μέθοδος συσταδοποίησης είναι καλή αν παράγει συστάδες καλής ποιότητας

- Μεγάλη ομοιότητα εντός της συστάδας και
- Μικρή ομοιότητα ανάμεσα στις συστάδες

Η ποιότητα εξαρτάται από τη

- Μέτρηση ομοιότητας και
- Μέθοδο υλοποίησης της συσταδοποίησης

Είδη συσταδοποίησης



Μια συσταδοποίηση είναι ένα σύνολο από συστάδες

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

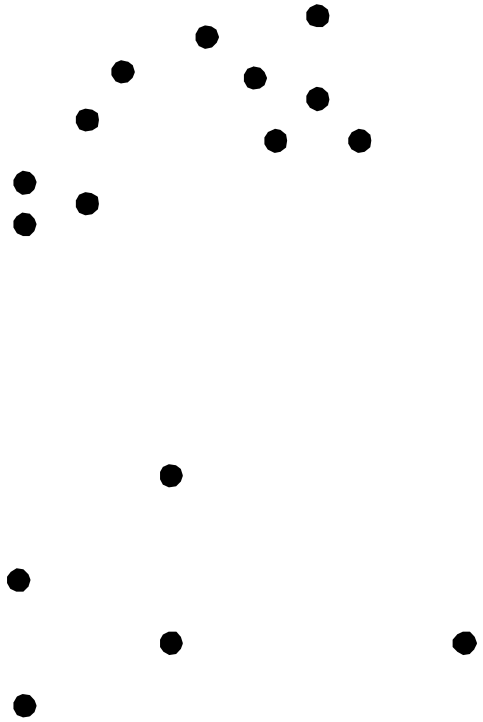
Διαχωριστική Συσταδοποίηση (Partitional Clustering)

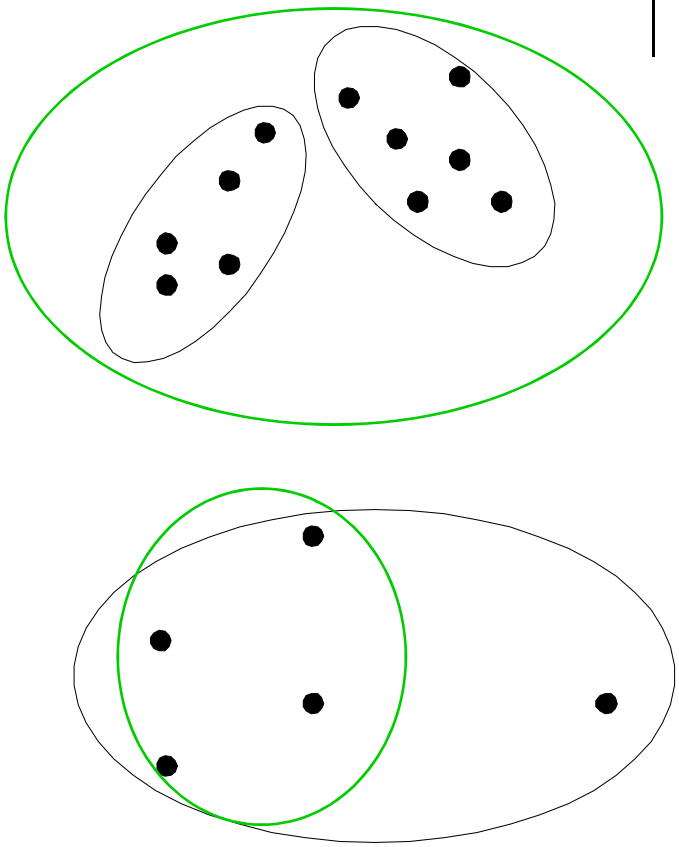
Ένας διαμερισμός των αντικειμένων σε μη επικαλυπτόμενα - non-overlapping - υποσύνολα (συστάδες) τέτοιος ώστε κάθε αντικείμενο ανήκει σε ακριβώς ένα υποσύνολο

Ιεραρχική Συσταδοποίηση (Hierarchical clustering

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

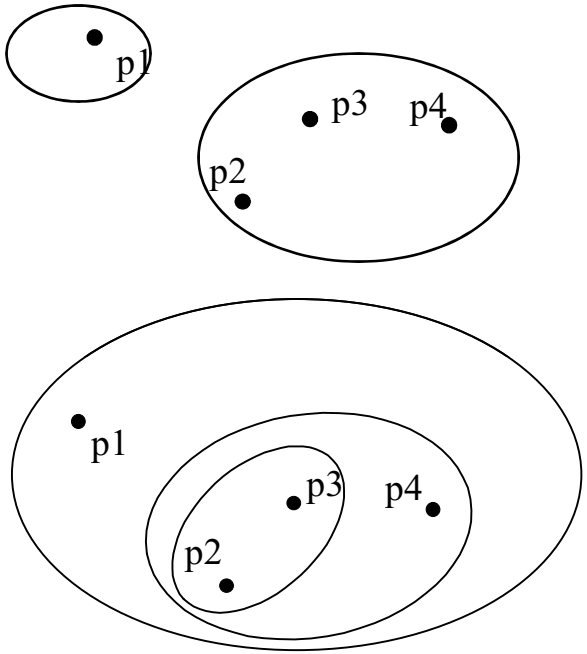
Διαχωριστική και Ιεραρχική Συσταδοποίηση





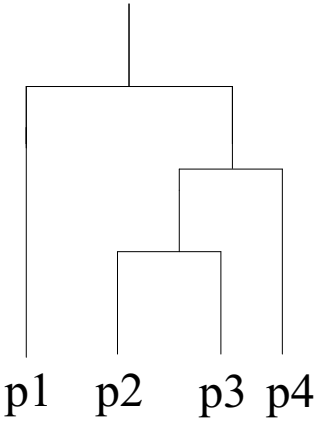
Αρχικά Σημεία

Διαχωριστική και Ιεραρχική Συσταδοποίηση



Ιεραρχική Συσταδοποίηση

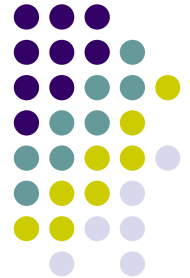
Διαχωριστική Συσταδοποίηση



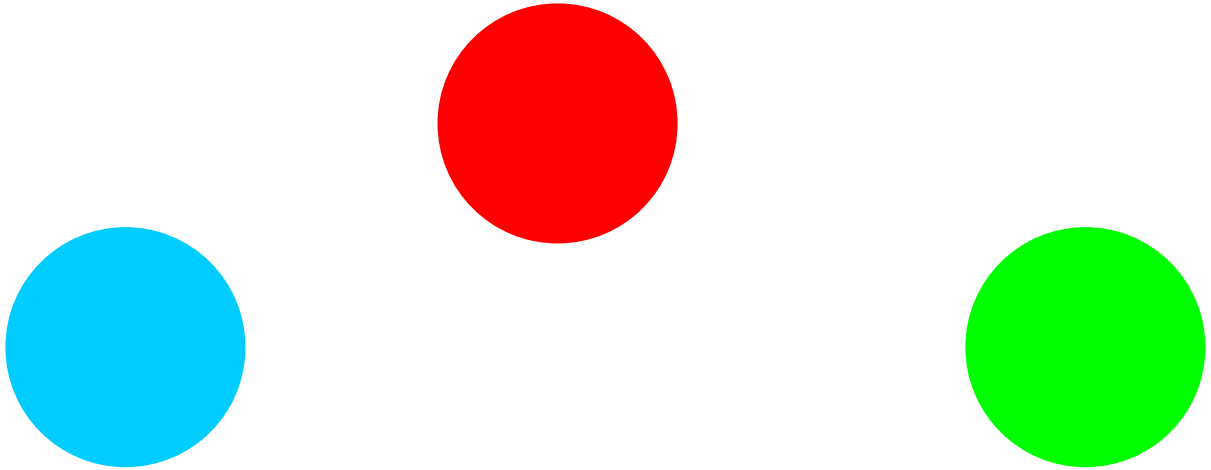
Παραδοσιακό Δένδρο-γράμμα (Dendrogram)

- Φύλλα: απλά σημεία ή απλές συστάδες
- Ως ακολουθία διαχωριστικών
- Να «κόψουμε» το 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)

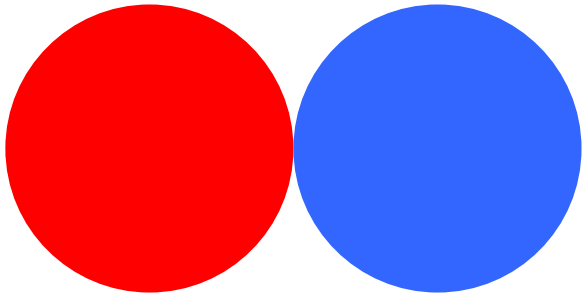
Όχι απαραίτητα κυκλικοί (οποιοδήποτε σχήμα)

Τύποι συστάδων: Συστάδες βασισμένες σε κέντρο ή πρότυπο

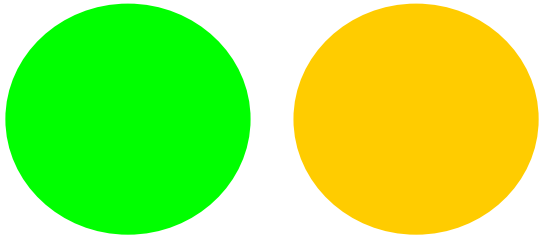
Μια συστάδα είναι ένα σύνολο από αντικείμενα τέτοιο ώστε ένα αντικείμενο στην συστάδα είναι κοντινότερο σε (ή πιο όμοιο με) το «κέντρο» ή πρότυπο της συστάδας από ότι από το κέντρο οποιασδήποτε άλλης συστάδας.

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)

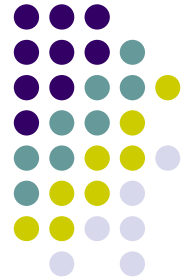


4 συστάδες βασισμένες σε κέντρο



Τείνουν στο να είναι κυκλικοί

Τύποι συστάδων: Συνεχής Συστάδες

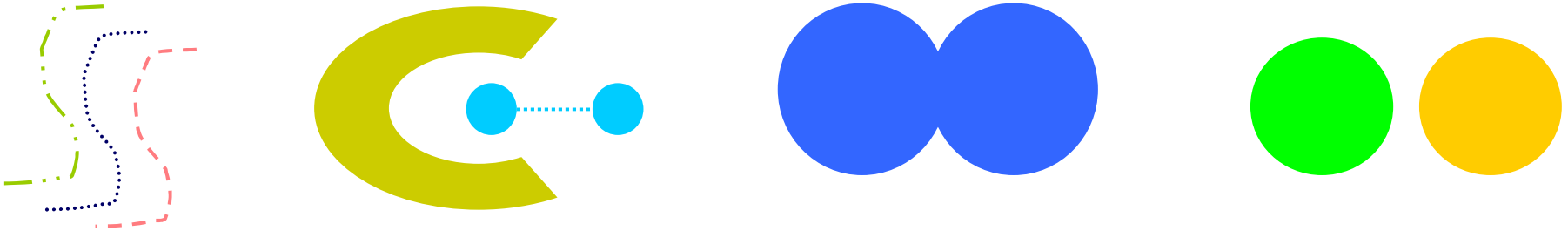


Συνεχής Συστάδες (Contiguous Cluster) (Κοντινότερος γείτονα ή μεταβατικά) – Βάσει γειτνίασης

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

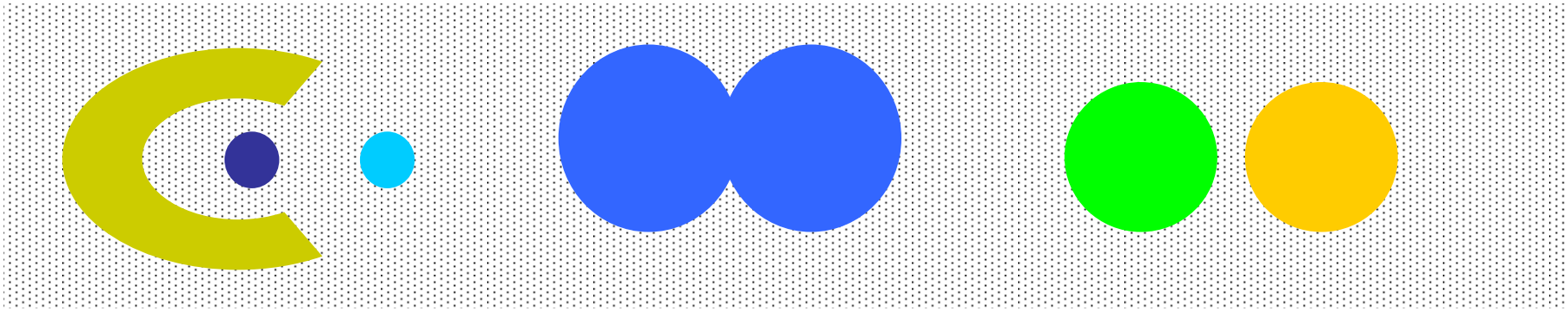


8 συνεχείς συστάδες

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

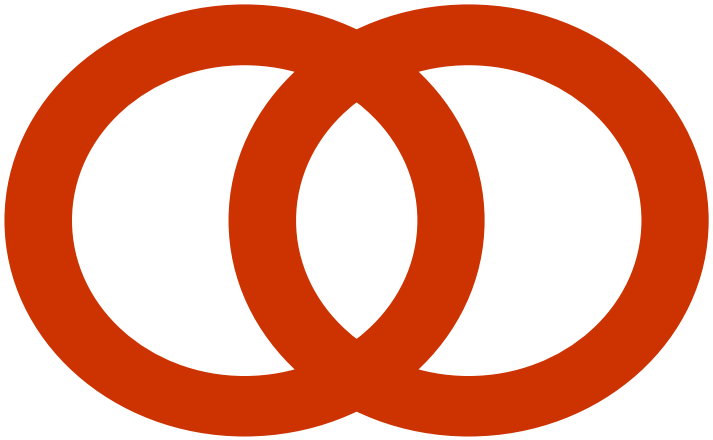


6 συστάδες βασισμένες στην πυκνότητα

Τύποι συστάδων: Εννοιολογική συσταδοποίηση

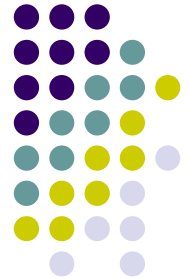


Συστάδες με κοινή ιδιότητα ή εννοιολογικές συστάδες.



2 αλληλοκαλυπτόμενοι κύκλοι

Τύποι συστάδων: Συστάδες βασισμένες σε μια Αντικειμενική Συνάρτηση



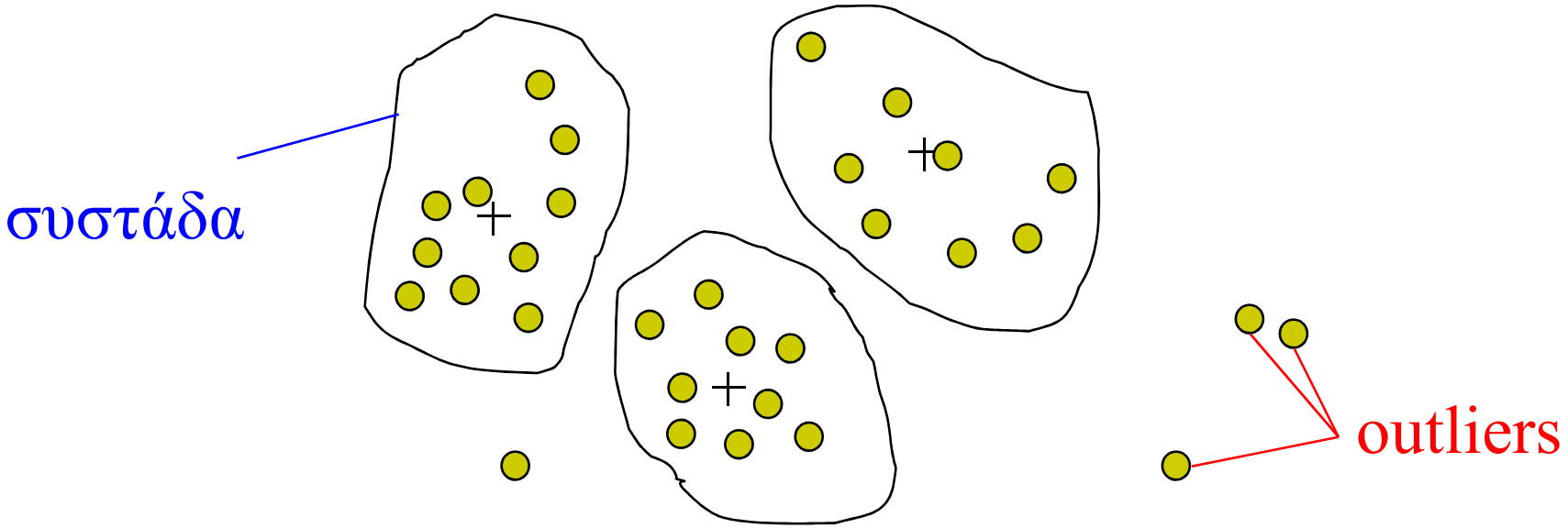
Εύρεση συστάδων που ελαχιστοποιούν ή μεγιστοποιούν μια αντικειμενική συνάρτηση

Απαρίθμηση όλων των δυνατών τρόπων χωρισμού των σημείων σε συστάδες και υπολογισμού του «πόσο καλό» (“goodness”) είναι κάθε πιθανό σύνολο από συστάδες χρησιμοποιώντας τη δοθείσα αντικειμενική συνάρτηση (NP-hard)

- Οι στόχοι (objectives) μπορεί να είναι ολικοί (global) ή τοπικοί (local)
- Οι ιεραρχικοί συνήθως τοπικού
- Οι διαχωριστικοί ολικές

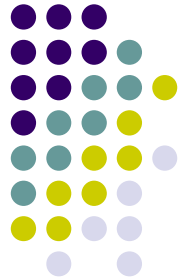


Αντιμετώπιση θορύβου και outliers



Outlier (ακραίο σημείο) τιμές που είναι εξαιρέσεις ως προς τα συνηθισμένες ή αναμενόμενες τιμές

Άλλες διακρίσεις μεταξύ συνόλων συστάδων



Επικαλυπτόμενο ή όχι

Ένα σημείο ανήκει σε περισσότερες από μια συστάδες (πχ οριακά σημεία)

Ασαφή συσταδοποίηση

Στην ασαφή συσταδοποίηση ένα σημείο ανήκει σε κάθε συστάδα με κάποιο βάρος μεταξύ του 0 και του 1
Συχνά τα βάρη για κάθε σημείο έχουν άθροισμα 1
Η πιθανοτική συσταδοποίηση έχει παρόμοια χαρακτηριστικά

Μερική - Πλήρης

Σε ορισμένες περιπτώσεις θέλουμε να ομαδοποιήσουμε μόνο κάποια από τα δεδομένα (άλλα θόρυβος, ή μη ενδιαφέρουσα πληροφορία)

Ετερογενή - Ομογενή

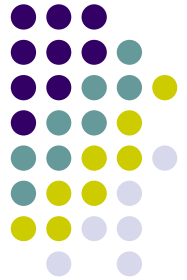
Συστάδες με πολύ διαφορετικά μεγέθη, σχήματα και πυκνότητες (densities)

Αλγόριθμοι Συσταδοποίησης



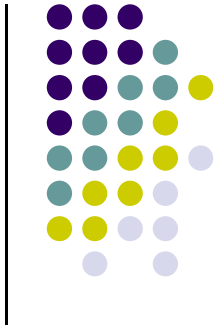
Θα δούμε ανάμεσα σε άλλους τους:

- K-means και παραλλαγές
- Ιεραρχική Συσταδοποίηση
- Συσταδοποίηση με βάση την Πυκνότητα (DBSCAN)
- BIRCH (δεδομένα στο δίσκο!)



K-means

K-means: Γενικά



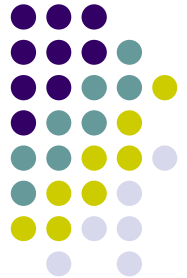
Διαχωριστικός αλγόριθμος

(βασισμένος σε πρότυπο) Κάθε συστάδα συσχετίζεται με ένα κεντρικό σημείο (centroid)

Κάθε σημείο ανατίθεται στη συστάδα με το κοντινότερο κεντρικό σημείο

Ο αριθμός των ομάδων, Κ, είναι είσοδος στον αλγόριθμο

K-means: Βασικός Αλγόριθμος



Βασικός αλγόριθμος

1: Επιλογή *K* σημείων ως τα αρχικά κεντρικά σημεία

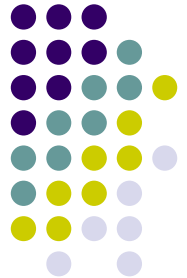
2: **Repeat**

3: Ανάθεση όλων των αρχικών σημείων στο *κοντινότερο* τους από τα *K* κεντρικά σημεία

4: Επανα-υπολογισμός του *κεντρικού σημείου* κάθε συστάδας

5: **Until** τα κεντρικά σημεία να μην αλλάζουν

K-means: Βασικός Αλγόριθμος

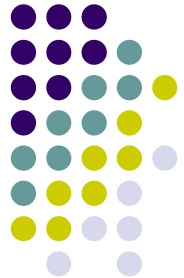


Παράδειγμα

2 **4** 10 12 **3** 20 30 11 15

Έστω k = 2, και αρχικά επιλέγουμε το 3 και το 4

K-means: Βασικός Αλγόριθμος



Παρατηρήσεις

1. Τα αρχικά κεντρικά σημεία συνήθως επιλέγονται τυχαία

Οι συστάδες που παράγονται διαφέρουν από το ένα τρέξιμο του αλγορίθμου στο άλλο

K-means: Βασικός Αλγόριθμος



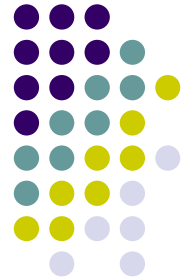
Παρατηρήσεις (συνέχεια)

2. Η εγγύτητα των σημείων υπολογίζεται με βάση κάποια απόσταση που εξαρτάται από το είδος των σημείων, στα παραδείγματα θα θεωρήσουμε την *Ευκλείδεια απόσταση*

- Επειδή η απόσταση υπολογίζεται συχνά ο υπολογισμός της πρέπει να είναι σχετικά *απλός*

3. Το κεντρικό σημείο είναι (συνήθως) το μέσο (mean) των σημείων της συστάδας (το οποίο μπορεί να *μην είναι ένα από τα δεδομένα εισόδου*)

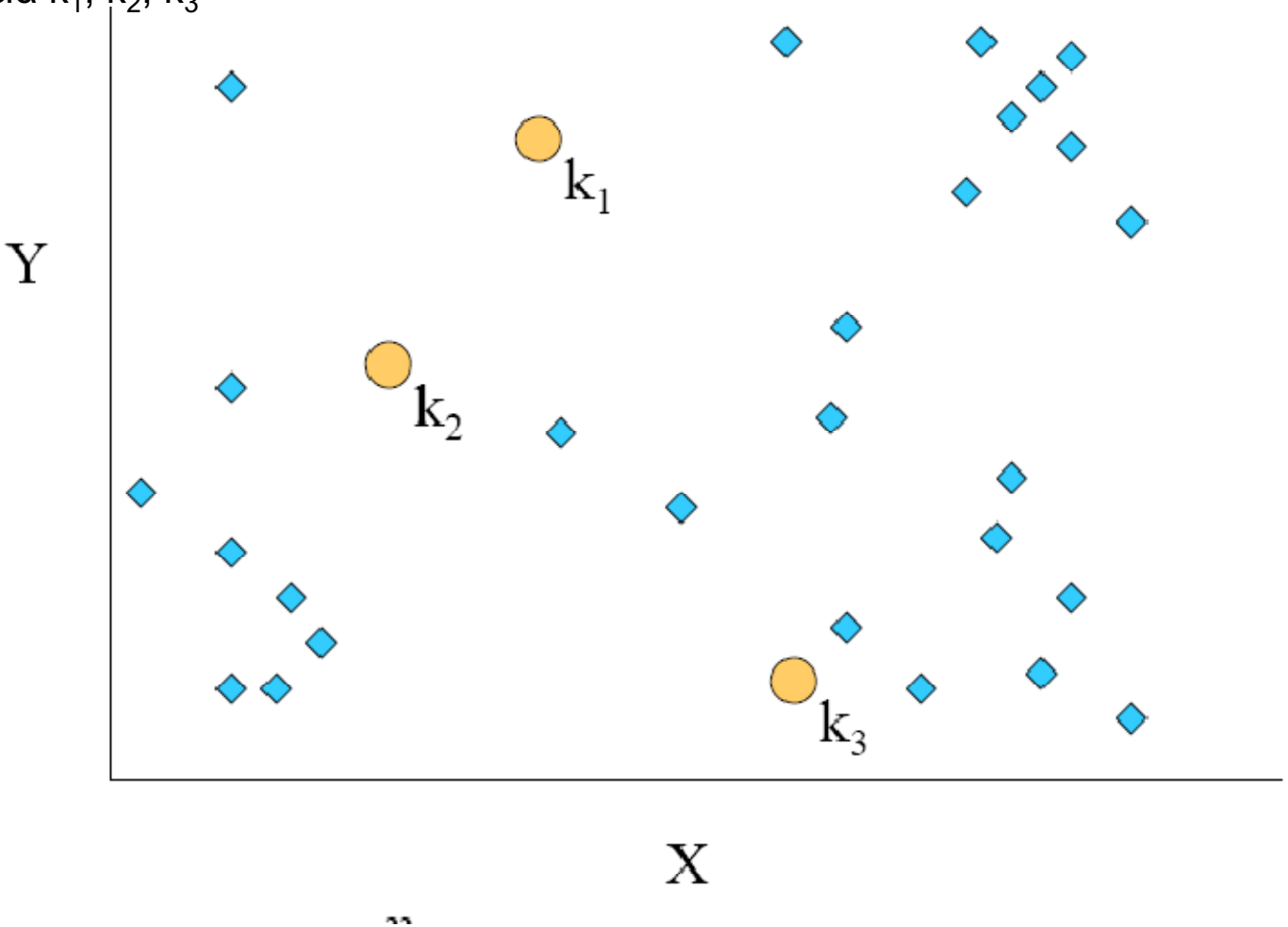
K-means: Βασικός Αλγόριθμος



Αρχική κατάσταση,

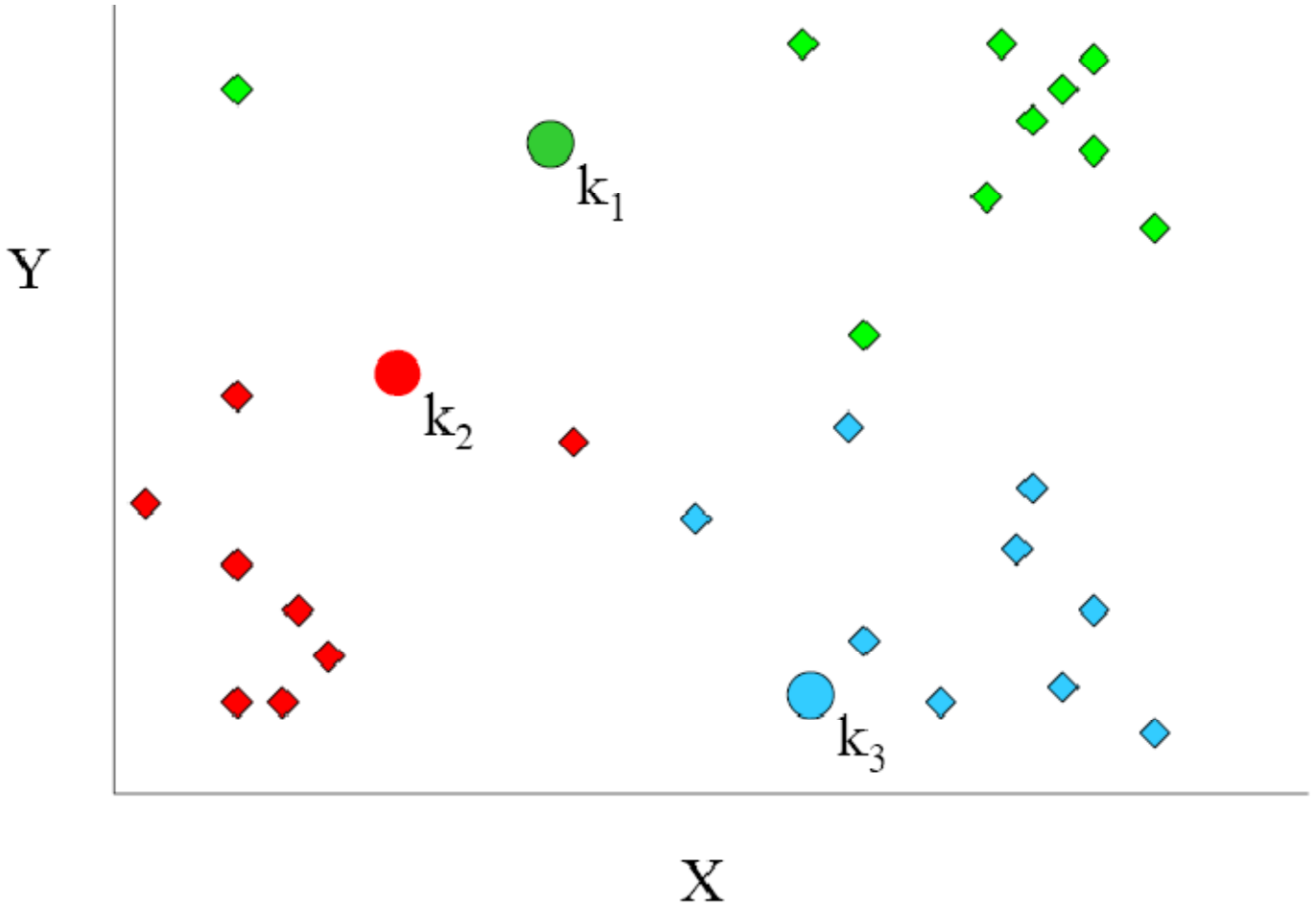
K = 3 συστάδες

Αρχικά σημεία k_1, k_2, k_3



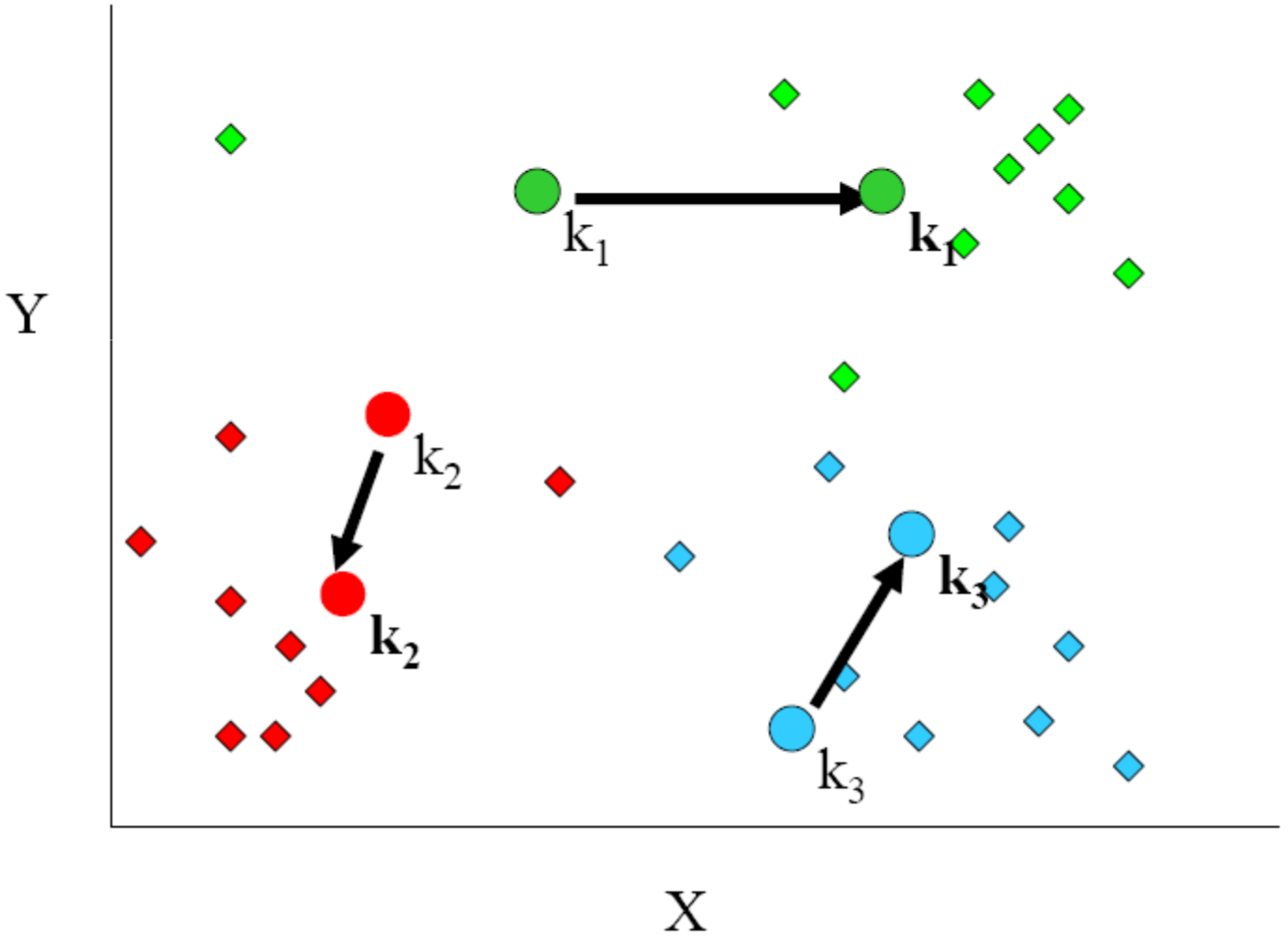
K-means: Βασικός Αλγόριθμος

Τα σημεία ανατίθενται στο πιο γειτονικό από τα 3 αρχικά σημεία



K-means: Βασικός Αλγόριθμος

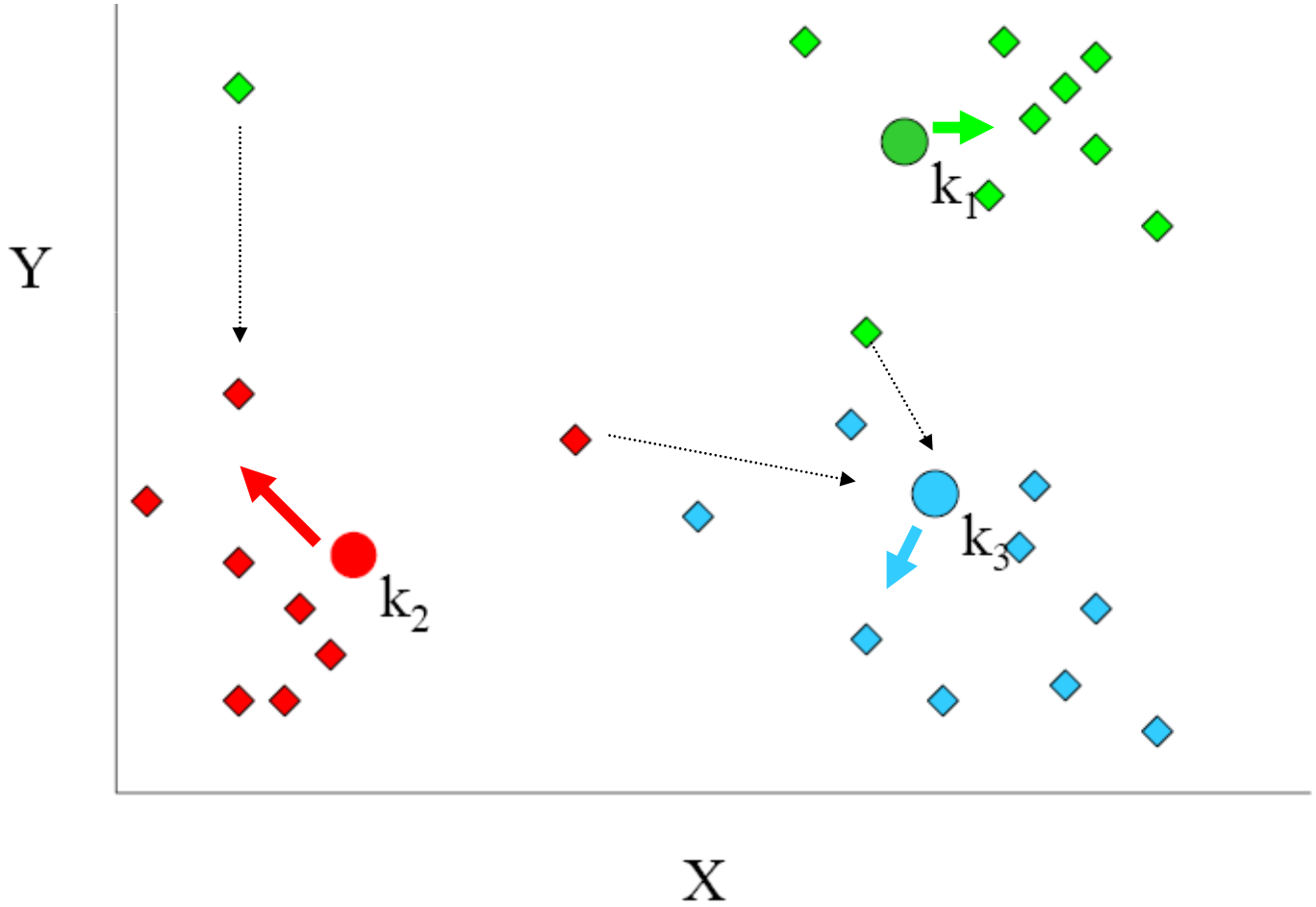
Επανα-υπολογισμός του κέντρου (κέντρου βάρους) κάθε σημείου



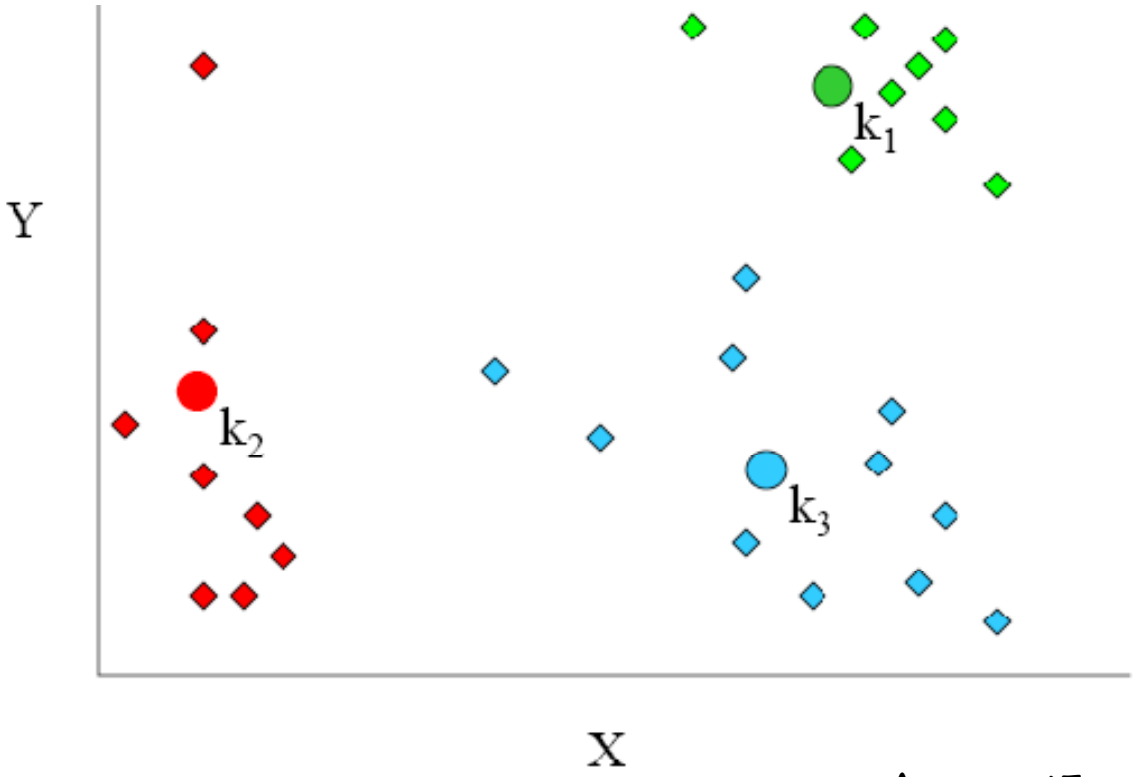
K-means: Βασικός Αλγόριθμος

Νέα ανάθεση των σημείων

Νέα κέντρα βάρους

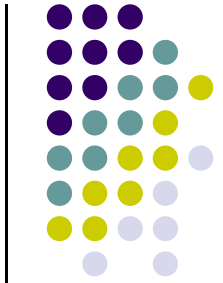


K-means: Βασικός Αλγόριθμος



Δεν αλλάζει τίποτα -> ΤΕΛΟΣ

K-means: Βασικός Αλγόριθμος

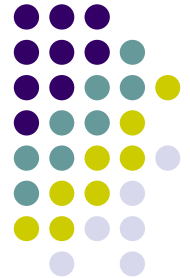


Παρατηρήσεις (συνέχεια)

- Χώρος: αποθηκεύουμε μόνα τα κέντρα

- Η πολυπλοκότητα είναι O(I * n * K * d)
  n = αριθμός σημείων,
  K = αριθμός συστάδων,
  I = αριθμός επαναλήψεων,
  d = αριθμός γνωρισμάτων (διάσταση)

K-means: Βασικός Αλγόριθμος



Παρατηρήσεις (συνέχεια)

- Για συνηθισμένα μέτρα ομοιότητας, ο αλγόριθμος συγκλίνει
  Η σύγκλιση συμβαίνει συνήθως τις αρχικές πρώτες επαναλήψεις

- Συχνά η τελική συνθήκη αλλάζει σε

Until

- σχετικά λίγα σημεία να αλλάζουν συστάδα – ή

- η απόσταση μεταξύ των νέων κεντρικών σημείων από τα παλιά
  να είναι μικρή

K-means: Εκτίμηση ποιότητας



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

Η πιο συνηθισμένη μέτρηση είναι το *άθροισμα των τετράγωνων του λάθους* (Sum of Squared Error (SSE))

Για κάθε σημείο, το λάθος είναι η απόστασή του από την κοντινότερη συστάδα

Για να πάρουμε το SSE, παίρνουμε το τετράγωνο αυτών των λαθών και τα προσθέτουμε

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

Όπου dist Ευκλείδεια απόσταση, *x* είναι ένα σημείο στη συστάδα C_i και m_i είναι ο αντιπρόσωπος (κεντρικό σημείο) της συστάδας C_i

- Μπορούμε να δείξουμε ότι το σημείο που ελαχιστοποιεί το SSE για τη συστάδα είναι ο μέσος όρος $c_i = 1/m_i \Sigma_{x \in Ci} x$

- Δοθέντων δύο συστάδων, μπορούμε να επιλέξουμε αυτήν με το μικρότερο λάθος

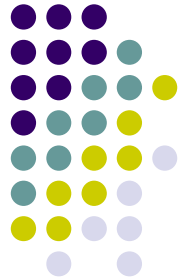
K-means: Εκτίμηση ποιότητας



Ένας τρόπος να βελτιώσουμε τη συσταδοποίηση (ελάττωση του SSE) είναι να μεγαλώσουμε το K

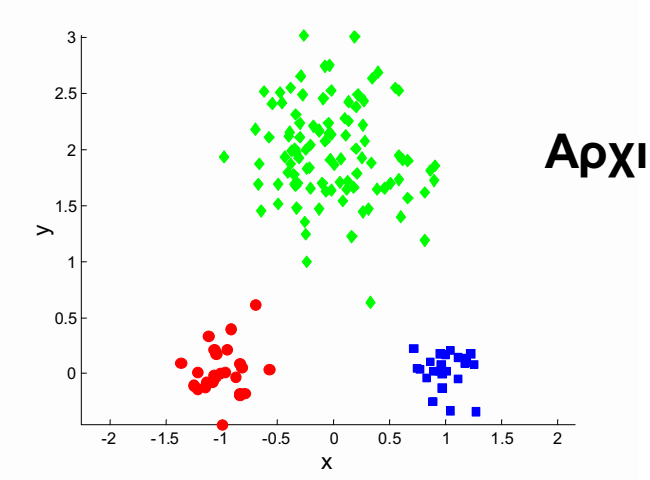
Αλλά γενικά μια καλή συσταδοποίηση με μικρό K μπορεί να έχει μικρότερο SSE από μια κακή συσταδοποίηση με μεγάλο K

K-means: Βασικός Αλγόριθμος

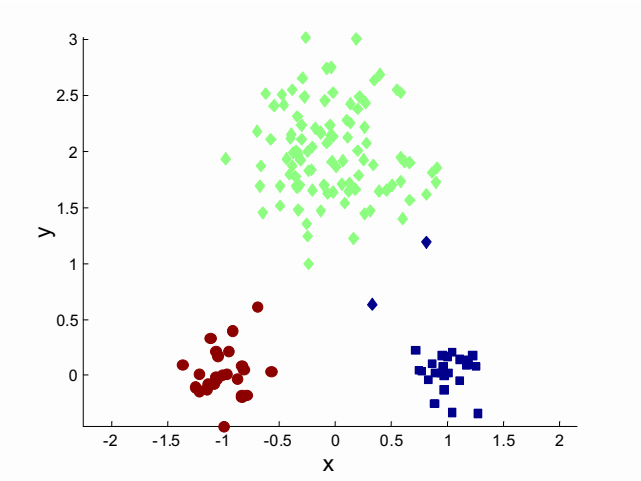


- Το αποτέλεσμα εξαρτάται από την επιλογή των αρχικών σημείων

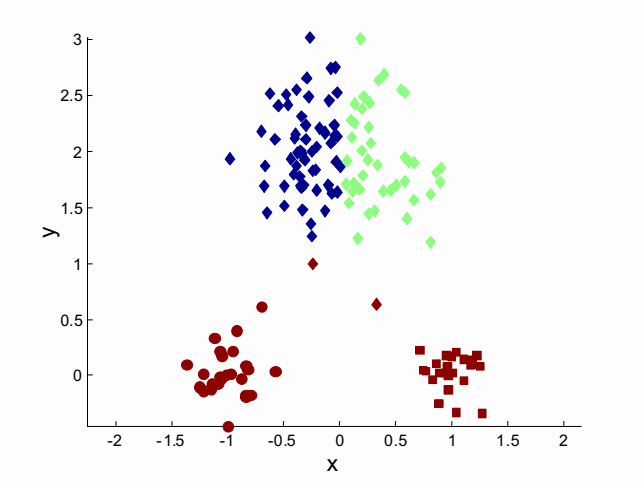
K-means: Παράδειγμα



Αρχικά σημεία

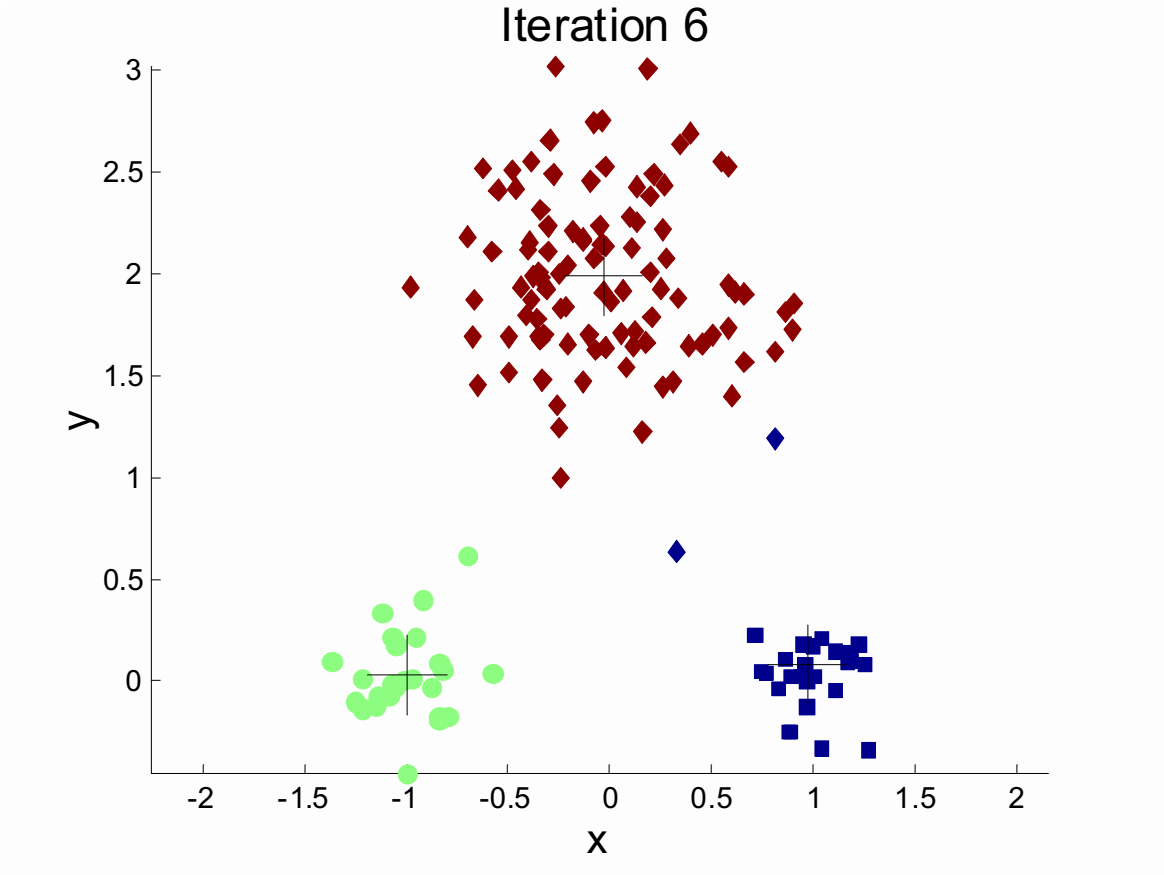


Βέλτιστη συσταδοποίηση

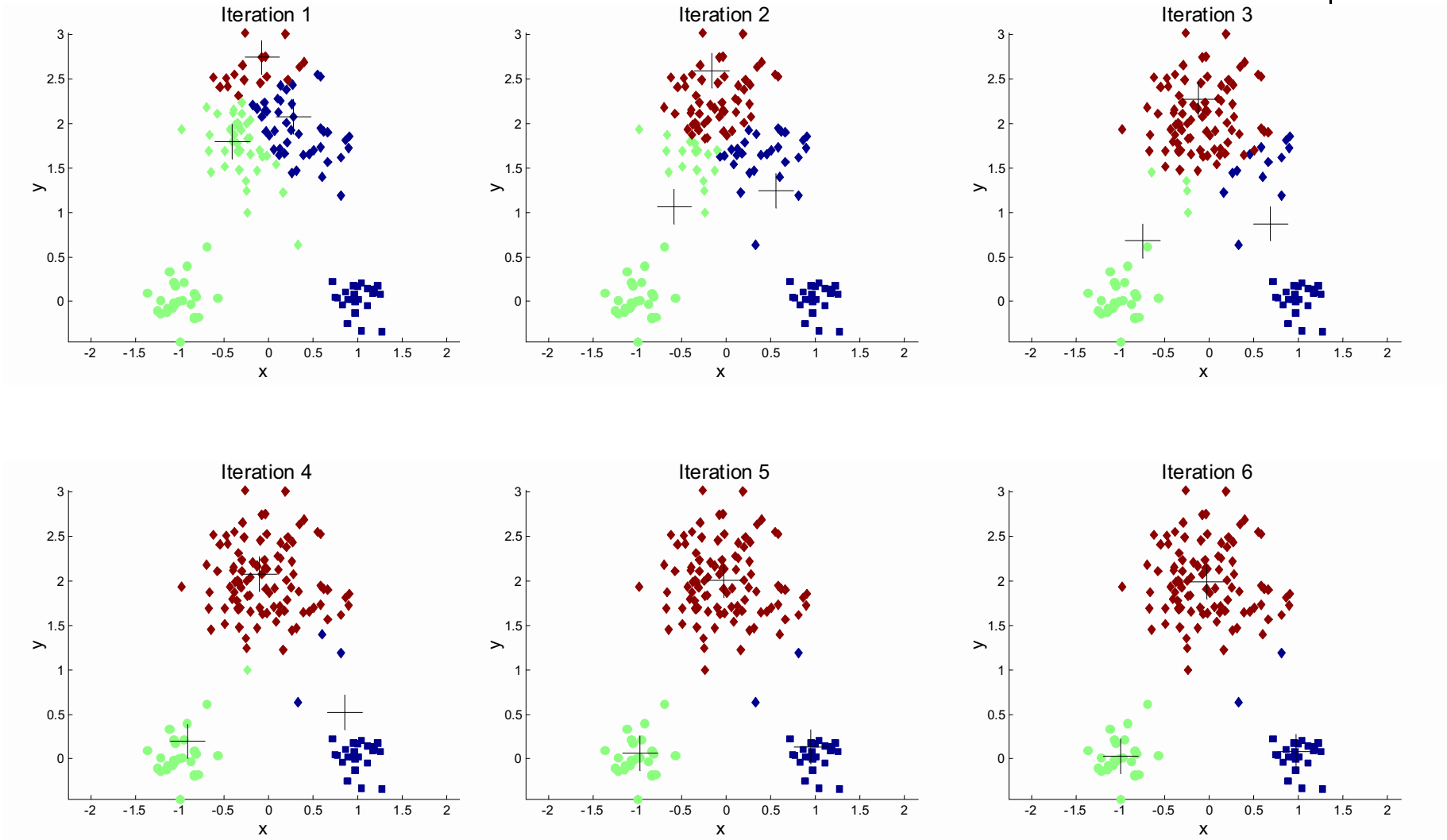


Υπό-βέλτιστη συσταδοποίση

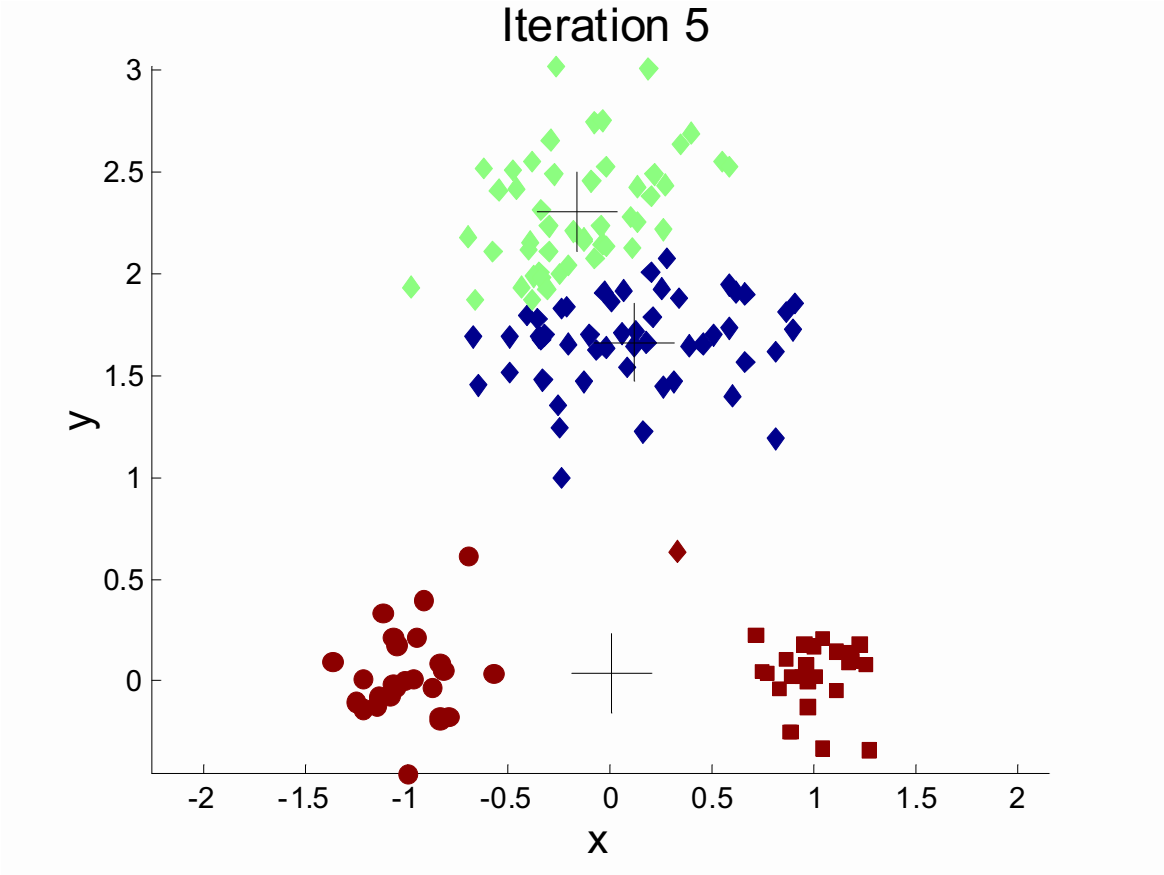
K-means: Επιλογή αρχικών σημείων



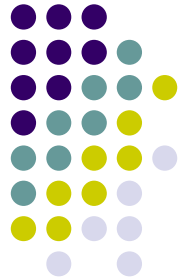
K-means: Επιλογή αρχικών σημείων

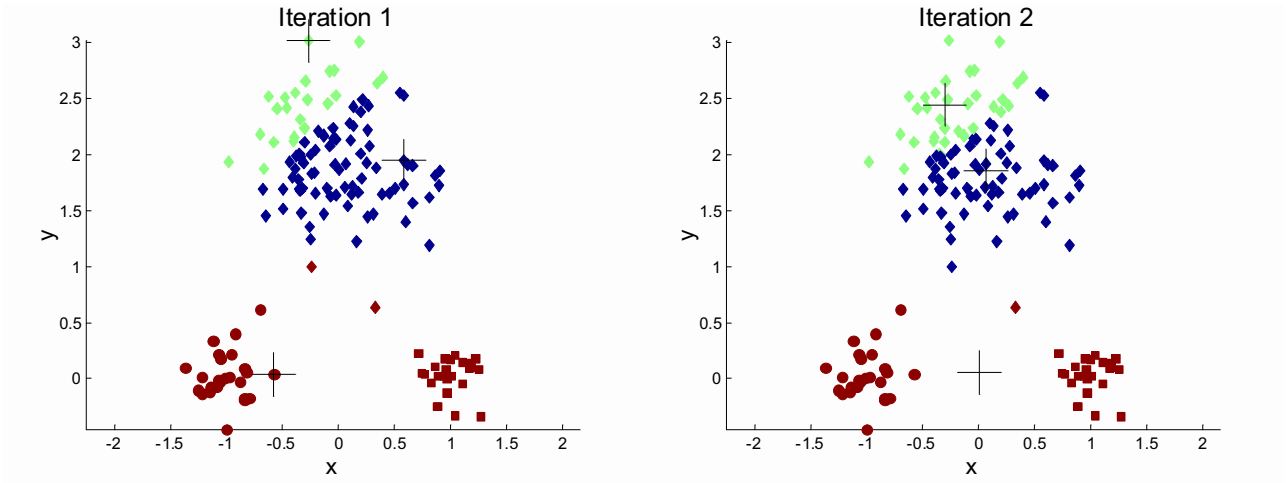


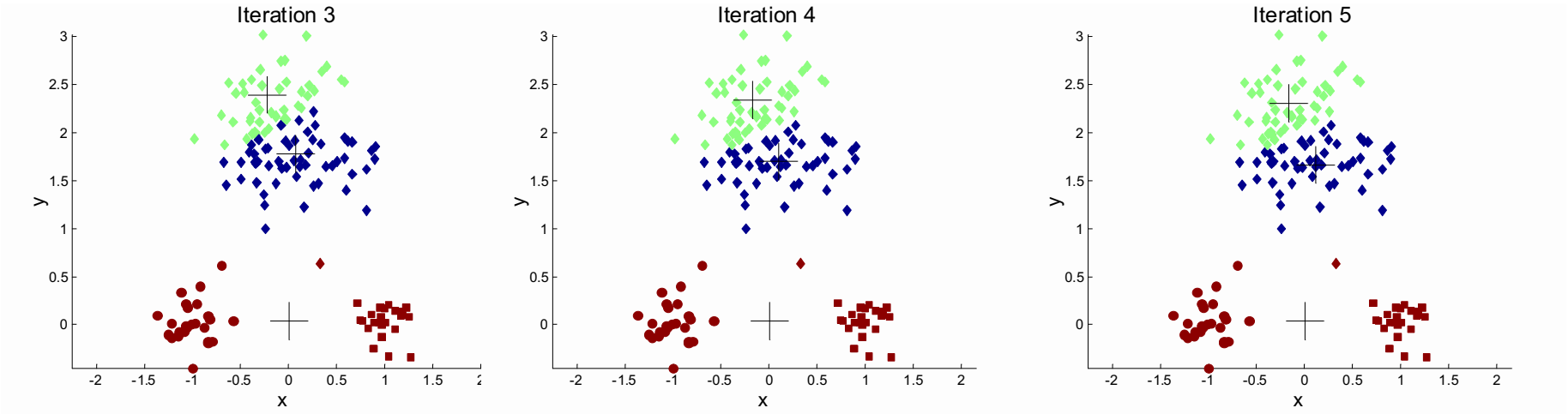
K-means: Επιλογή αρχικών σημείων



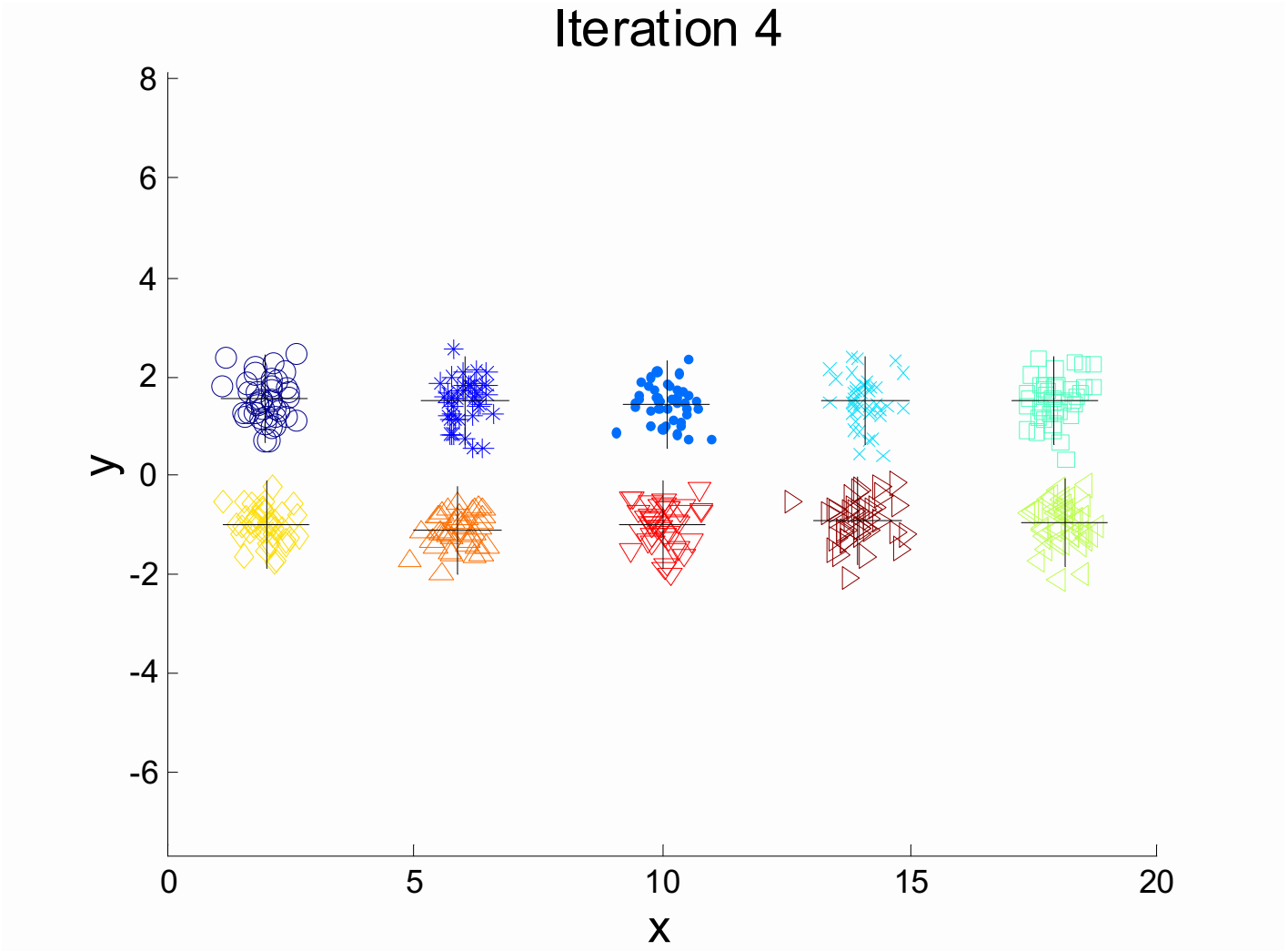
K-means: Επιλογή αρχικών σημείων





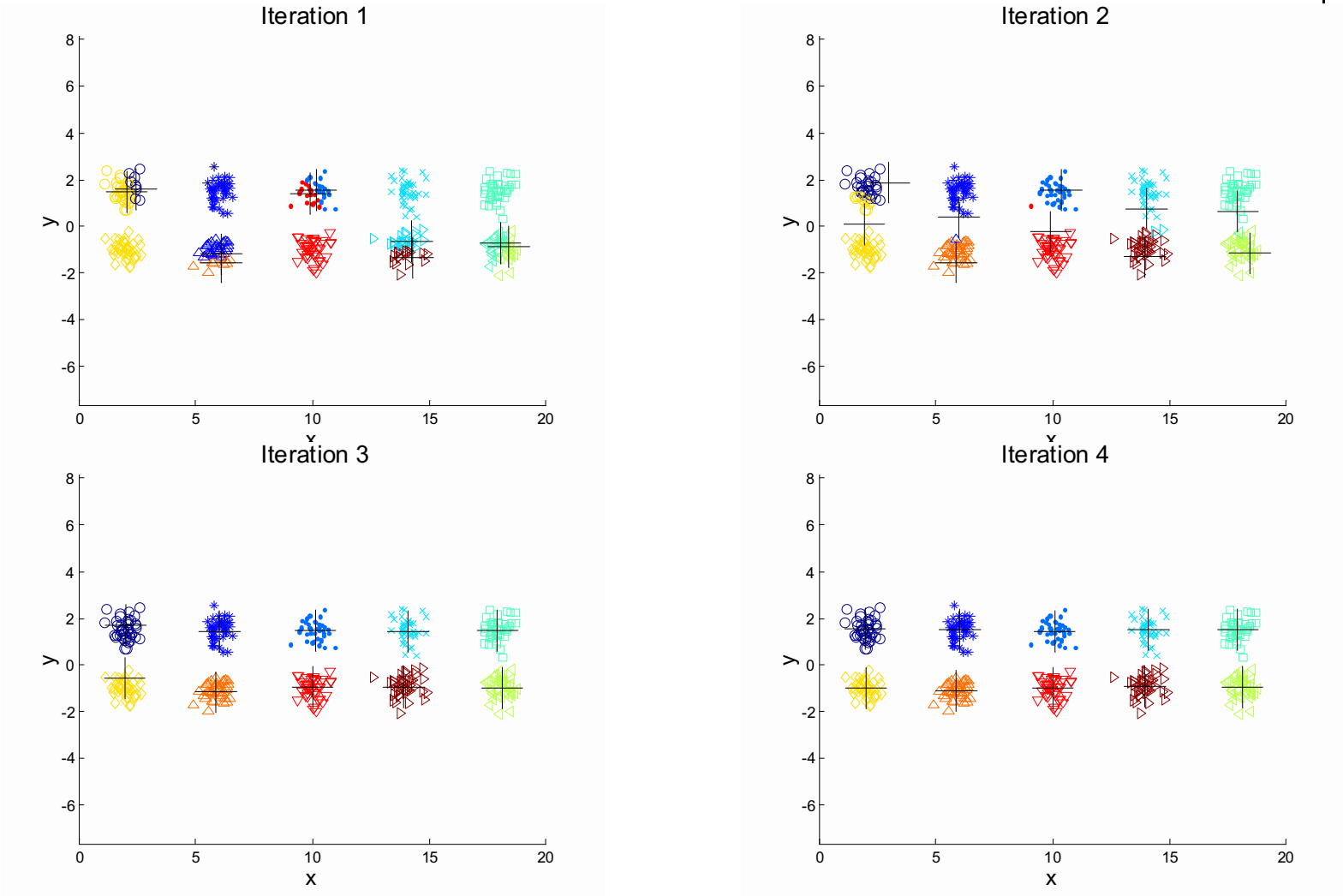


Παράδειγμα 10 συστάδων



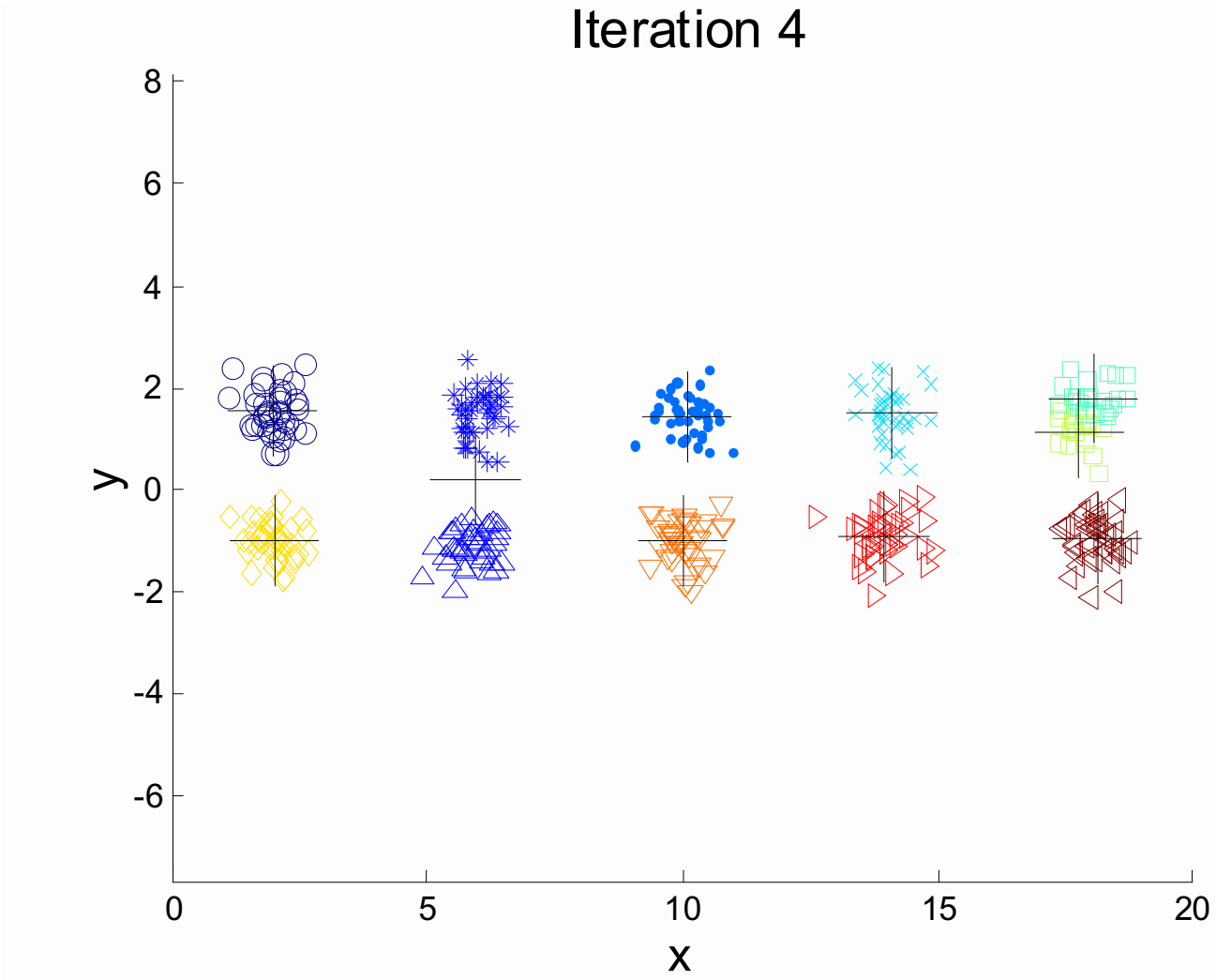
Ξεκινώντας με δύο αρχικά σημεία σε κάθε συστάδα κάθε ζεύγους συστάδων

Παράδειγμα 10 συστάδων



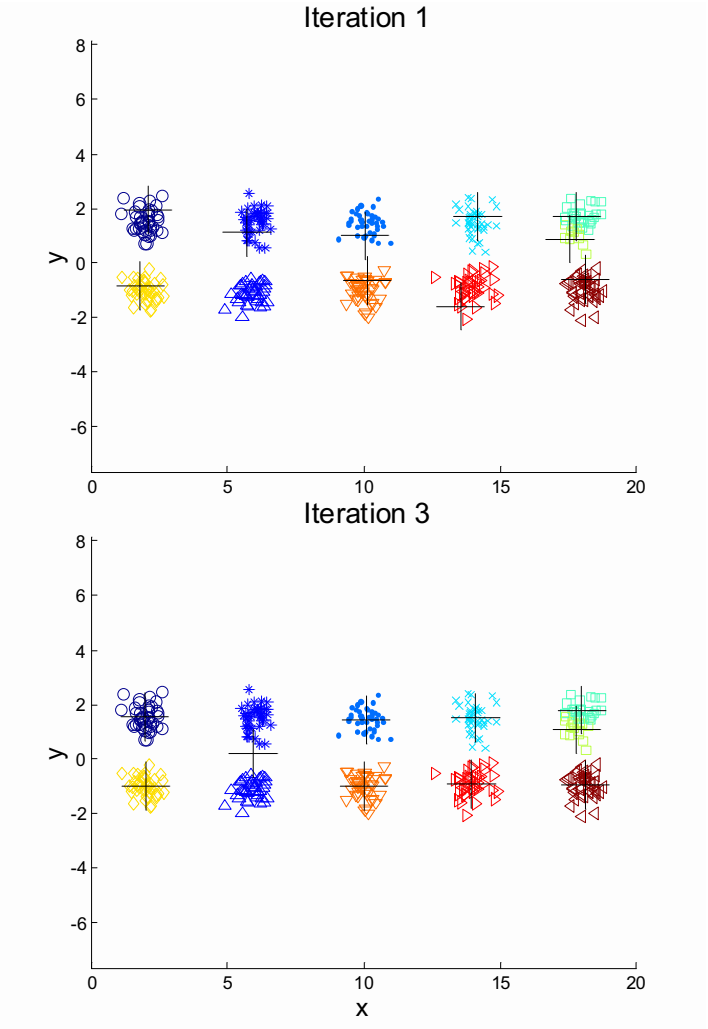
Ξεκινώντας με δύο αρχικά σημεία σε κάθε συστάδα κάθε ζεύγους συστάδων

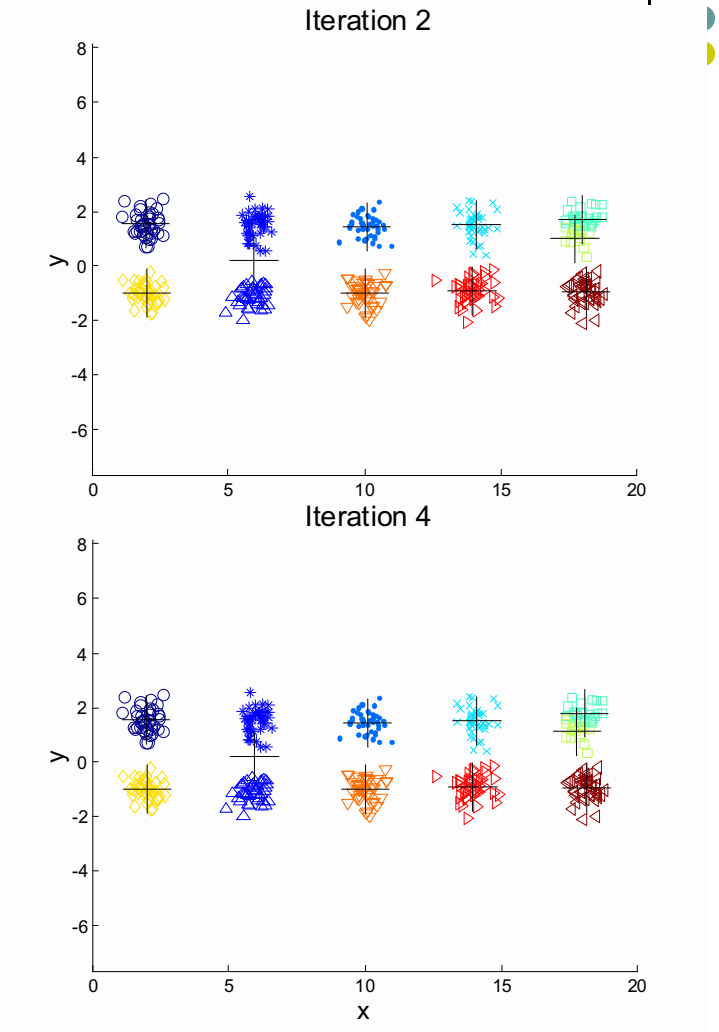
Παράδειγμα 10 συστάδων



Ξεκινώντας με κάποια ζευγάρια συστάδων να έχουν τρία κεντρικά σημεία και άλλα μόνο ένα

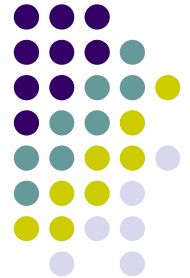
Παράδειγμα 10 συστάδων





Ξεκινώντας με κάποια ζευγάρια συστάδων να έχουν τρία κεντρικά σημεία και άλλα μόνο ένα

K-means: Επιλογή αρχικών σημείων

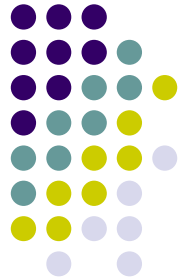


Αν υπάρχουν Κ «πραγματικές συστάδες» η πιθανότητα να επιλέξουμε ένα κέντρο από κάθε συστάδα είναι μικρή, συγκεκριμένα αν όλες οι συστάδες έχουν το ίδιο μέγεθος n, τότε:

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

Για παράδειγμα, αν Κ = 10, η πιθανότητα είναι = 10!/1010 = 0.00036

K-means: Λύσεις για την επιλογή αρχικών σημείων



Πολλαπλά τρεξίματα

Βοηθά, αλλά πολλές περιπτώσεις

Δειγματοληψία και χρήση κάποιας ιεραρχικής τεχνικής

Επιλογή παραπάνω από k αρχικών σημείων και μετά επιλογή k από αυτά τα αρχικά κεντρικά σημεία (πχ τα πιο απομακρυσμένα μεταξύ τους)

Σταδιακή επιλογή

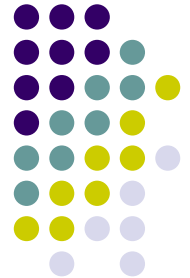
Επιλογή του πρώτου σημείου τυχαία ή ως το μέσο όλων των σημείων

Για καθένα από τα υπόλοιπα αρχικά σημεία

επέλεξε αυτό που είναι πιο μακριά από τα μέχρι τώρα επιλεγμένα αρχικά σημεία

- Μπορεί να οδηγήσει στην επιλογή outliers
- Ο υπολογισμός του πιο απομακρυσμένου σημείου είναι δαπανηρός
- Συχνά εφαρμόζεται σε δείγματα

K-means: Άδειες συστάδες



Ο βασικός αλγόριθμος μπορεί να οδηγήσει σε άδειες αρχικές συστάδες

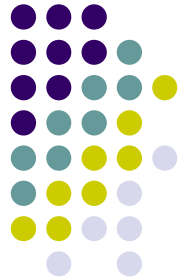
Πολλές στρατηγικές

Επιλογή του σημείου που είναι πιο μακριά από όλα τα τωρινά κέντρα = επιλογή του σημείου που συμβάλει περισσότερο στο SSE

Ένα σημείο από τη συστάδα με το υψηλότερο SSE – θα οδηγήσει σε «σπάσιμο» της άρα σε μείωση του λάθους

Αν πολλές *άδειες συστάδες*, τα παραπάνω βήματα μπορεί να επαναληφτούν πολλές φορές

K-means: Σταδιακή ενημέρωση κεντρικών σημείων

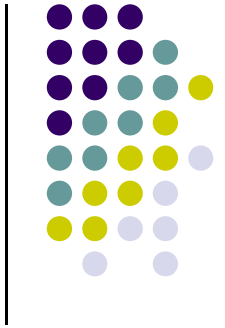


Στο βασικό K-means, το κέντρα ενημερώνεται αφού όλο τα σημεία έχουν ανατεθεί στο κέντρο

Μια παραλλαγή είναι να ενημερώνονται τα κέντρα μετά από κάθε ανάθεση (incremental approach)

- Κάθε ανάθεση ενημερώνει 0 ή 2 κέντρα
- Πιο δαπανηρό
- Έχει σημασία η σειρά εισαγωγής/εξέτασης των σημείων
- Δεν υπάρχουν άδειες συστάδες
- Μπορεί να χρησιμοποιηθούν βάρη – αν υπάρχει κάποια τυχαία αντικειμενική συνάρτηση – έλεγχος τι συμφέρει κάθε φορά

Προ και Μετα Επεξεργασία



Ολικό SSE και SSE Συστάδας

Προ-επεξεργασία

- Κανονικοποίηση των δεδομένων
- Απομάκρυνση outliers

Post-processing

Split-Merge (διατηρώντας το ίδιο K)

- Διαχωρισμός (split) συστάδων με το σχετικά μεγαλύτερο SSE
- Δημιουργία μια νέας συστάδας: πχ επιλέγοντας το σημείο που είναι πιο μακριά από όλα τα κέντρα ή τυχαία επιλογή σημείου ή επιλογή του σημείου με το μεγαλύτερο SSE

- Συνένωση (merge) συστάδων που είναι σχετικά κοντινές (τα κέντρα τους έχουν την μικρότερη απόσταση) ή τις δυο συστάδες που οδηγούν στην μικρότερη αύξηση του SSE
- Διαγραφή συστάδας και ανακατανομή των σημείων της σε άλλες συστάδες (αυτό που οδηγεί στην μικρότερη αύξηση του SSE)

K-means με διχοτόμηση (bisecting k-means)



Παραλλαγή που μπορεί να παράγει μια διαχωριστική ή ιεραρχική συσταδοποίηση

1: Αρχικοποίηση της λίστας των συστάδων ώστε να περιέχει μια συστάδα που περιέχει όλα τα σημεία

2: **Repeat**

3: Επιλογή μιας συστάδας από τη λίστα των συστάδων

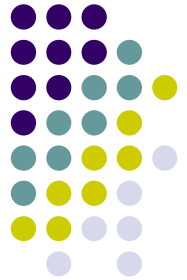
4: **for** i = 1 to number_of_trials **do**

5: διχοτόμησε την επιλεγμένη συστάδα χρησιμοποιώντας το βασικό k-means

6: Πρόσθεσε στη λίστα από τις δυο συστάδες που προέκυψαν από τη διχοτόμηση αυτήν με το μικρότερο SSE

5: **Until** η λίστα των συστάδων να έχει K συστάδες

K-means με διχοτόμηση (bisecting k-means)

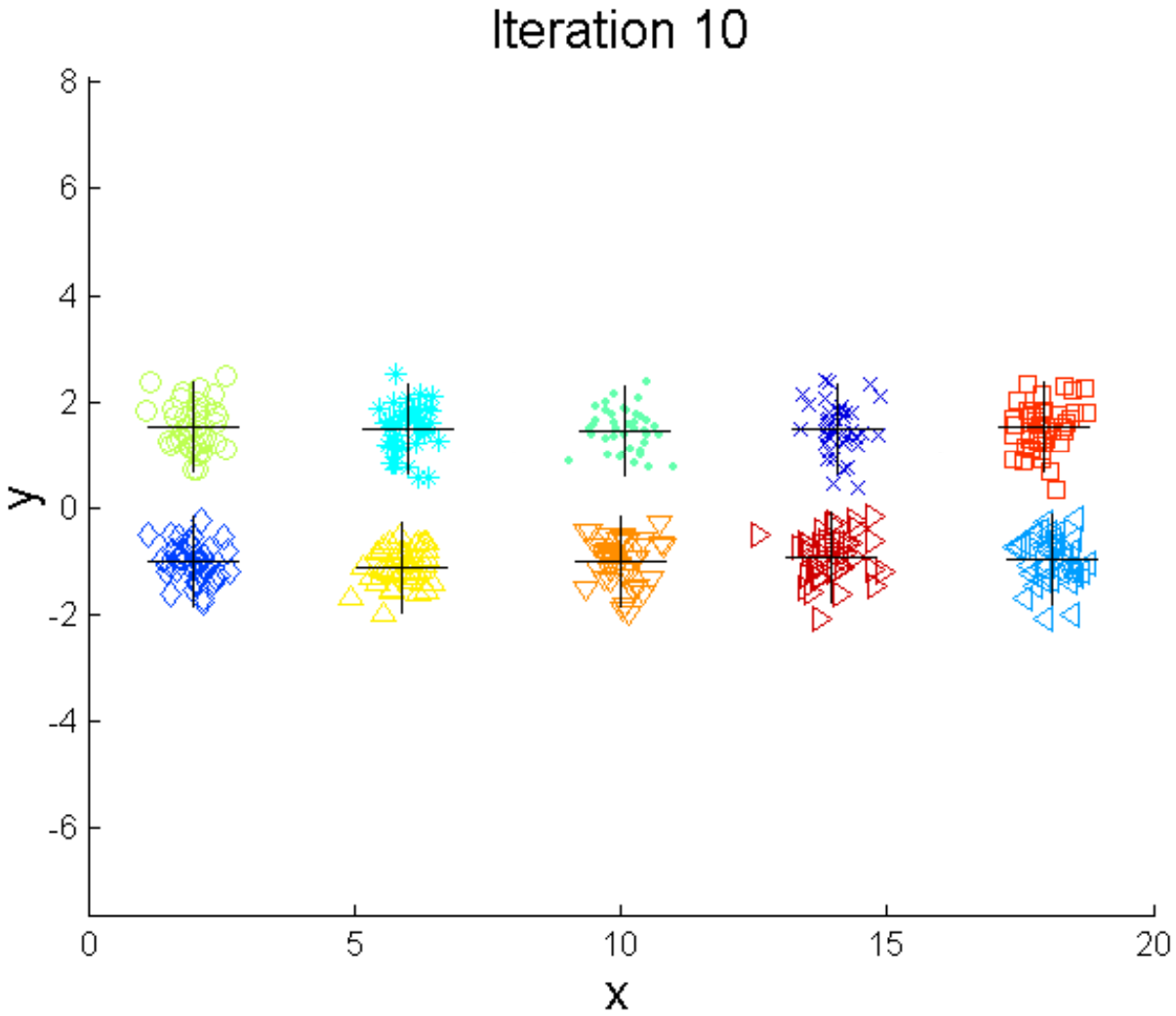


Ποια συστάδα να διασπάσουμε;

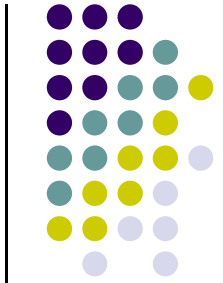
- Τη μεγαλύτερη
- Αυτή με το μεγαλύτερο SSE
- Συνδυασμό των παραπάνω

Μπορεί να χρησιμοποιηθεί και ως ιεραρχικός

K-means με διχοτόμηση



K-means: Περιορισμοί

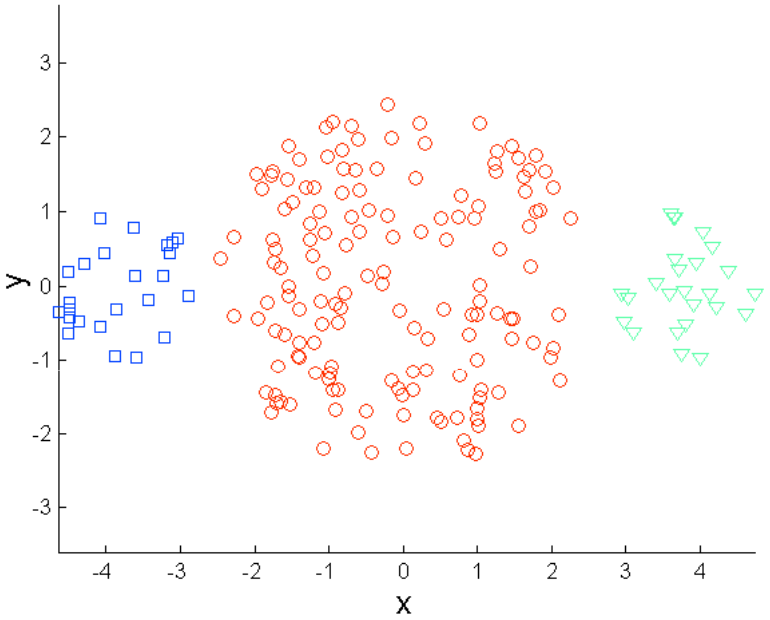


Ο K-means έχει προβλήματα όταν οι συστάδες έχουν

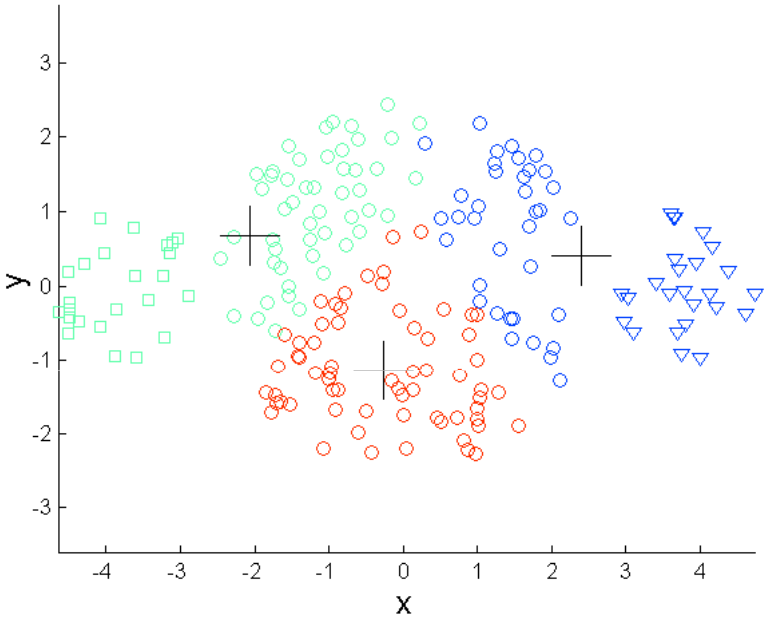
Διαφορετικά Μεγέθη
Διαφορετικές Πυκνότητες
Non-globular shapes

Έχει προβλήματα όταν τα δεδομένα έχουν outliers

K-means: Περιορισμοί – διαφορετικά μεγέθη



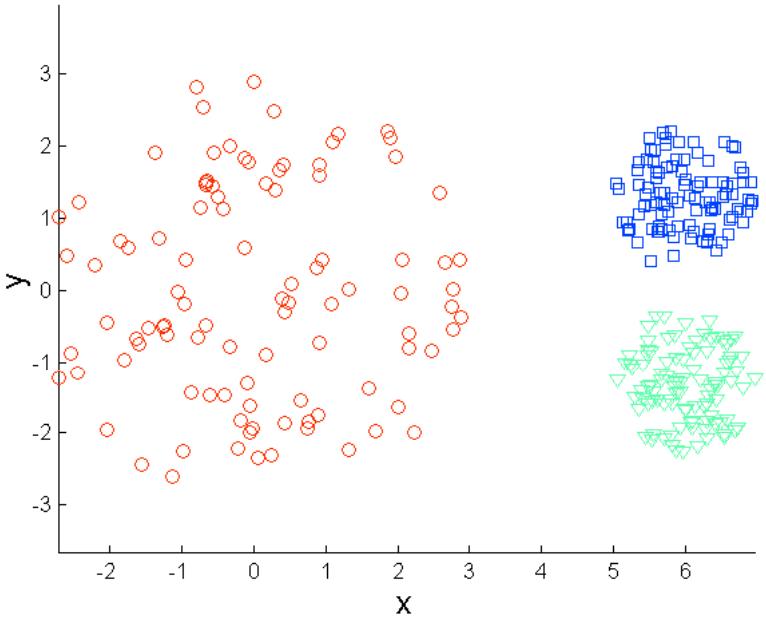
Αρχικά σημεία



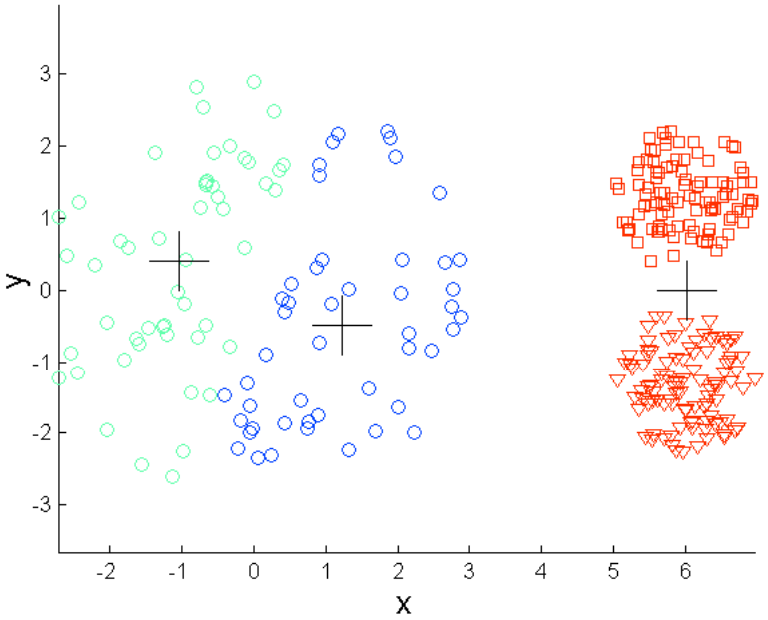
K-means (3 συστάδες)

Δεν μπορεί να βρει το μεγάλο κόκκινο, γιατί είναι πολύ μεγαλύτερος από τους άλλους

K-means: Περιορισμοί – διαφορετικές πυκνότητες



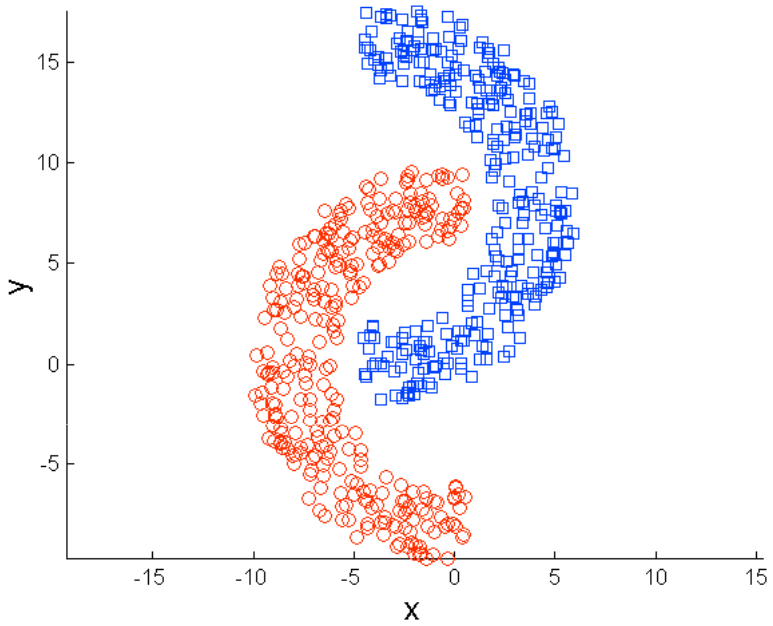
Αρχικά σημεία

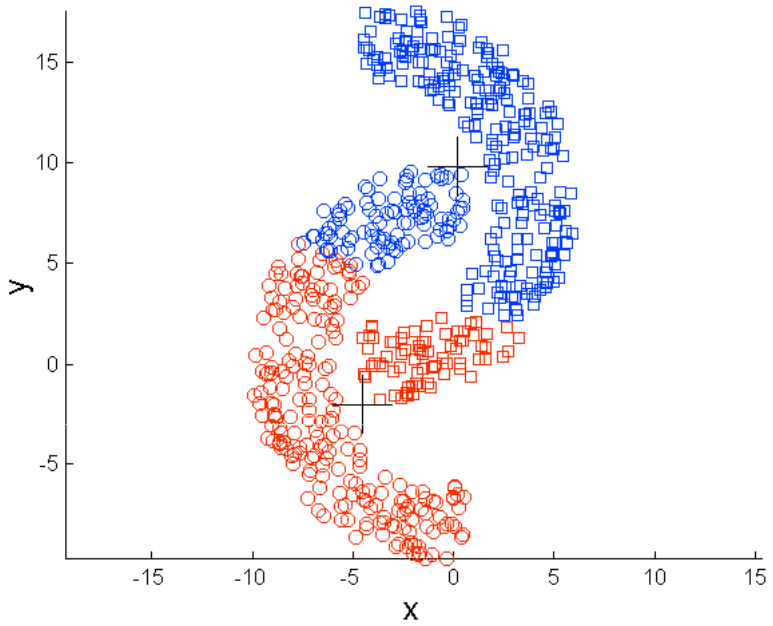


K-means (3 συστάδες)

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

K-means: Περιορισμοί – μη κυκλικά σχήματα



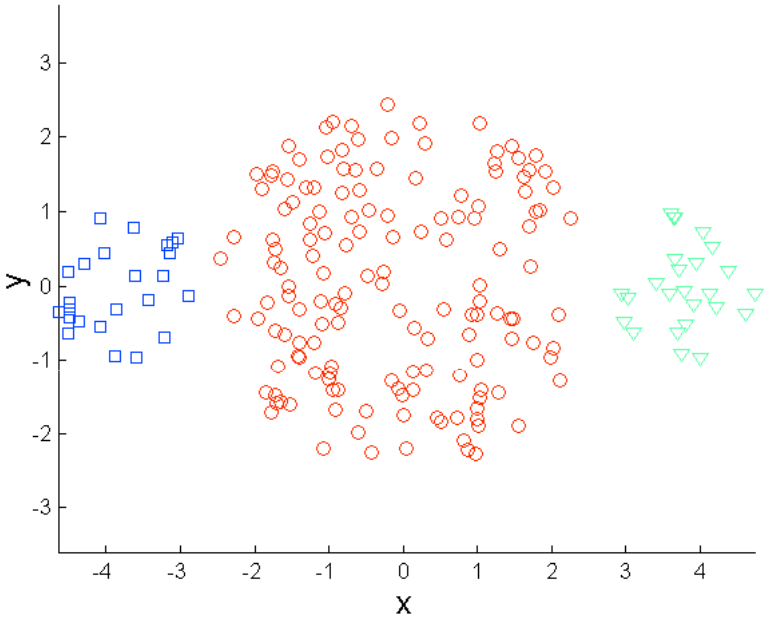


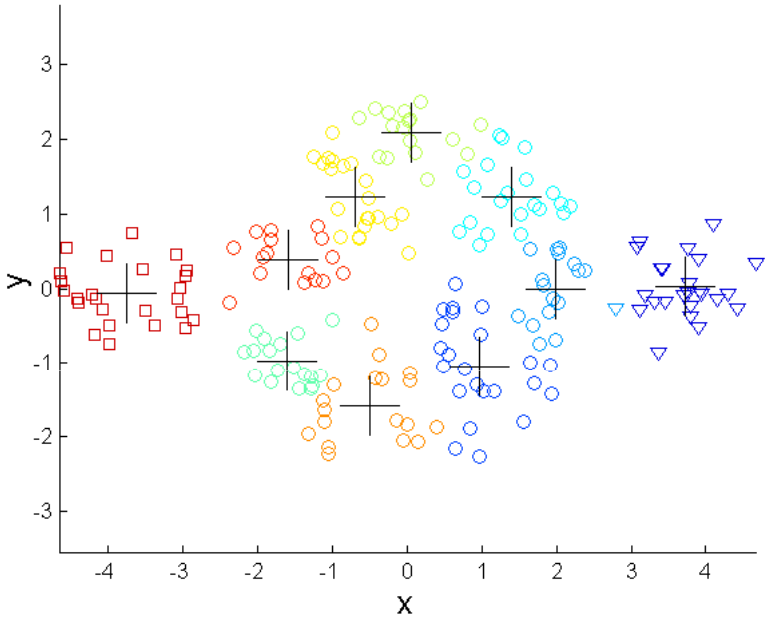
Αρχικά σημεία

K-means (2 συστάδες)

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

K-means: Περιορισμοί



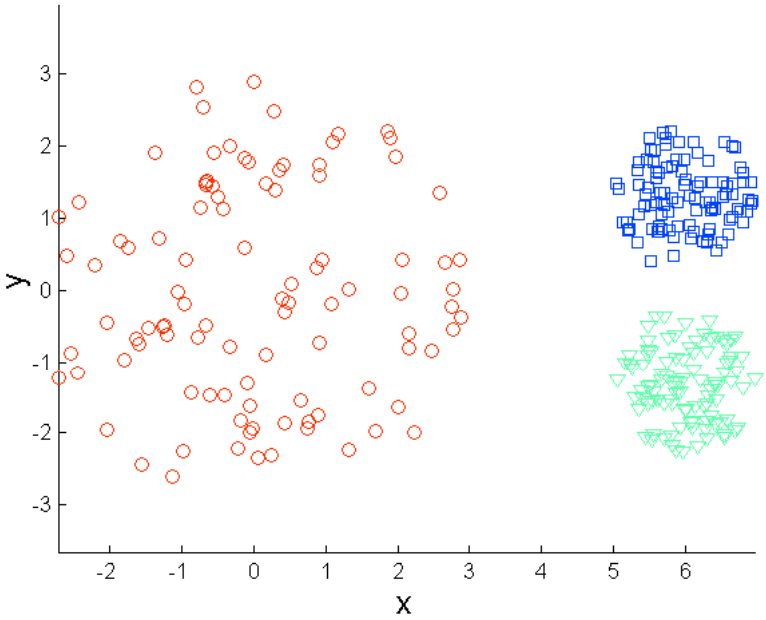


Αρχικά Σημεία

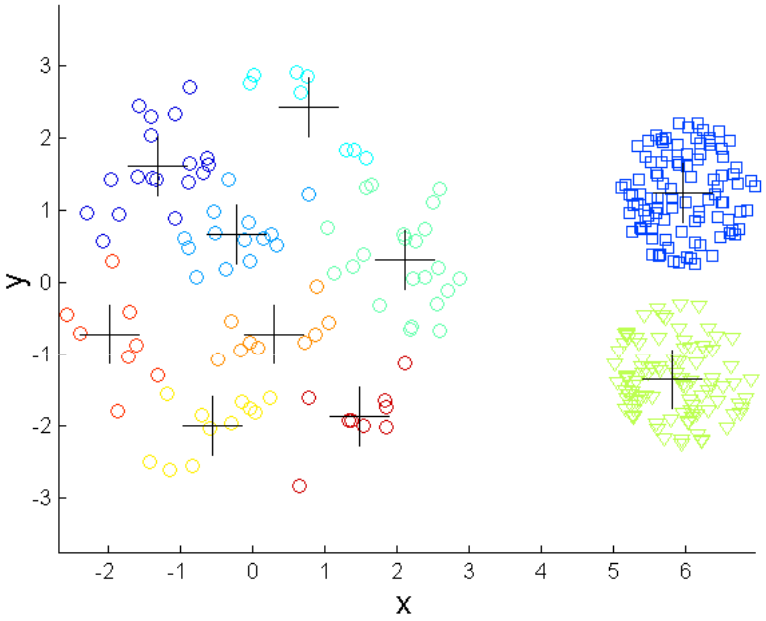
K-means Συστάδες

Μια λύση είναι να χρησιμοποιηθούν πολλές συστάδες
Βρίσκει τμήματα των συστάδων, αλλά πρέπει να τα συγκεντρώσουμε

K-means: Περιορισμοί

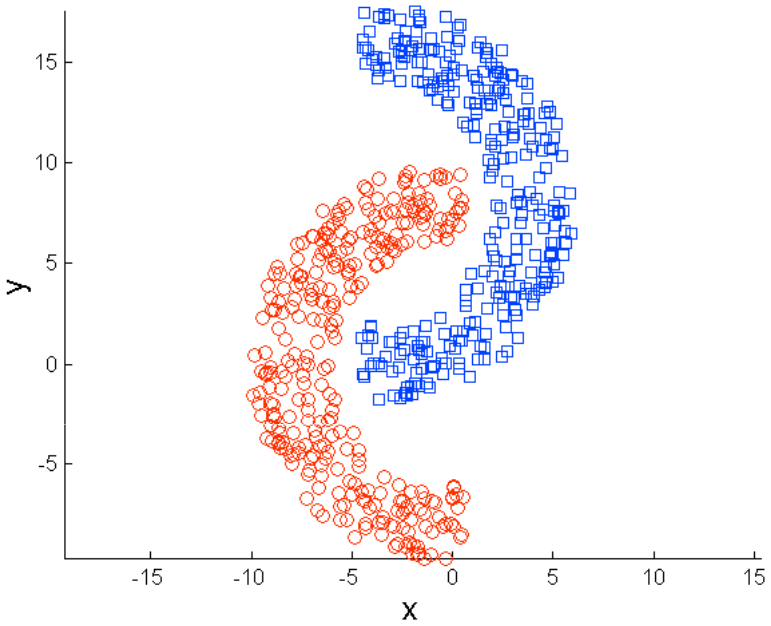


Αρχικά σημεία

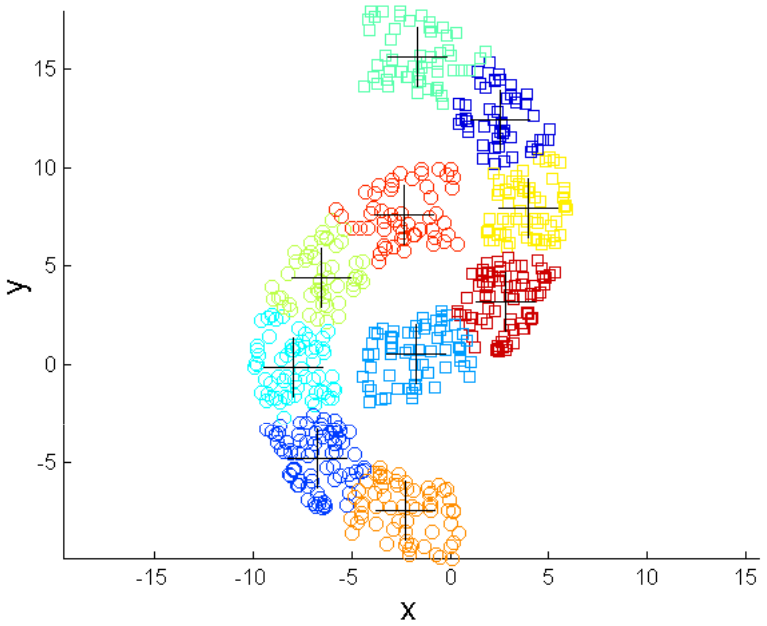


K-means Συστάδες

K-means: Περιορισμοί – διαφορετικά μεγέθη



Αρχικά Σημεία



K-means Συστάδες

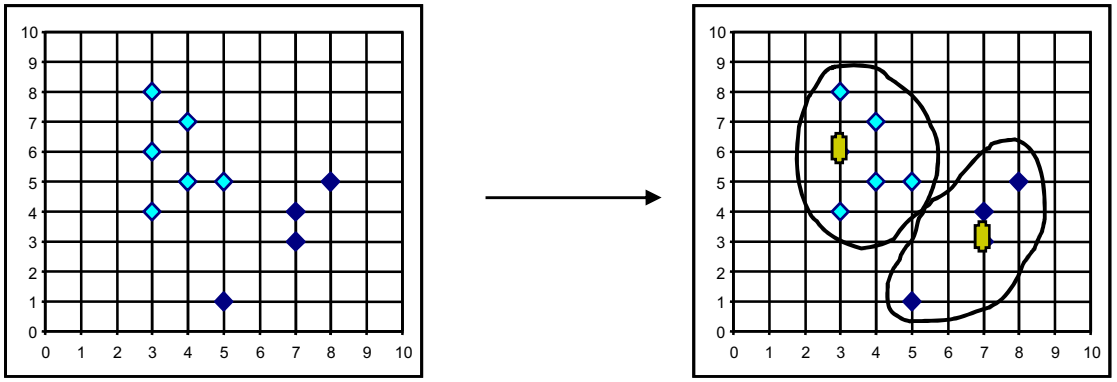
K-medoid

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)



Παρένθεση – Ορισμοί

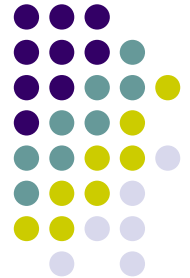


Μια παρένθεση

Μέση τιμή

Απόσταση

Γενική Τάση



- Αριθμητικό Μέσο/Μέση Τιμή- Mean (αλγεβρική μέτρηση) (sample vs. population):
  - Αριθμητικό μέσο με βάρος (Weighted arithmetic mean)
  - Trimmed mean: κόβουμε τις ακραίες τιμές (πχ τα μεγαλύτερα και μικρότερα (p/2)%)
- Μέσο – μεσαία τιμή (median) - **διάμεσο**:
  - Μεσαία τιμή αν μονός αριθμός, ο μέσος όρος των δυο μεσαίων τιμών, αλλιώς

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

Το μέσο συμπεριφέρεται καλύτερα όταν δεδομένα με μη ομοιόμορφη κατανομή (skewed)

Παράδειγμα

1 2 3 4 5 90

Μέσο

Μέση τιμή

Trimmed 40%

Γενική Τάση

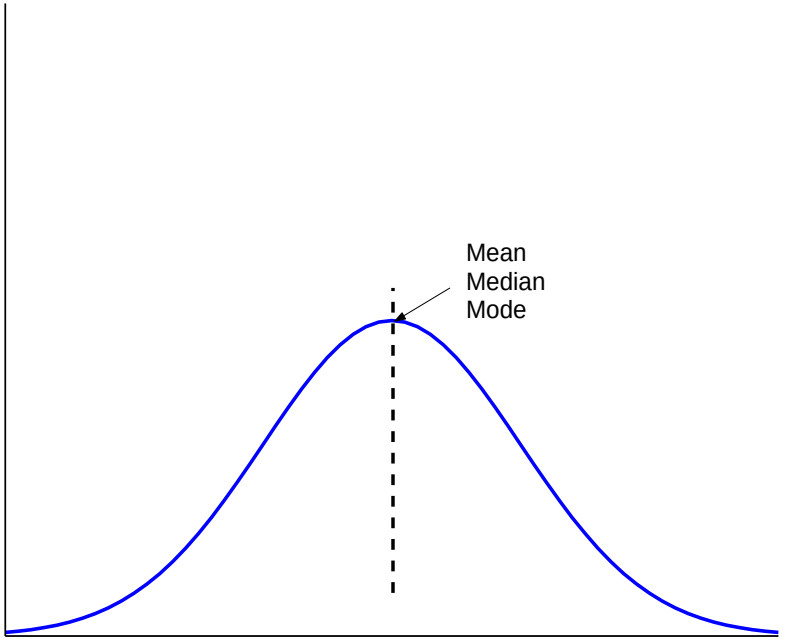


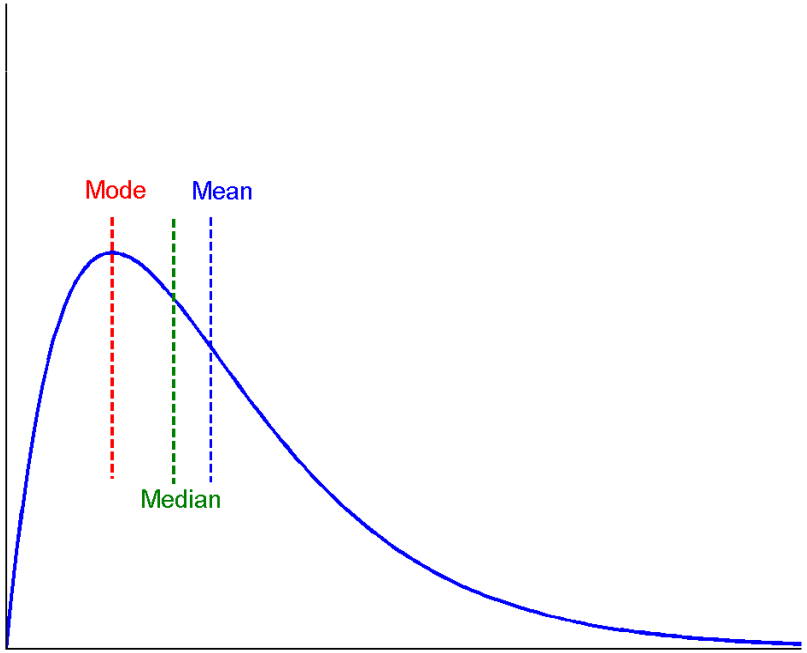
- Mode
  - Η τιμή που εμφανίζεται πιο συχνά στα δεδομένα
  - Unimodal, bimodal, trimodal (πιο συχνά εμφανίζεται μία, δύο ή τρεις διαφορετικές τιμές)

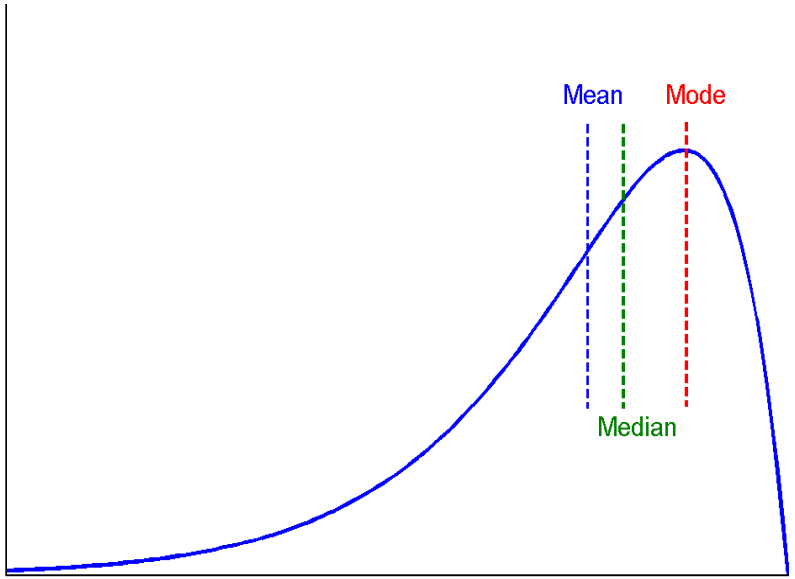
- Midrange (μέσο διαστήματος)
  - (min()+max())/2

Γενική Τάση

Median, mean and mode of symmetric, positively and negatively skewed data







Γενική Τάση

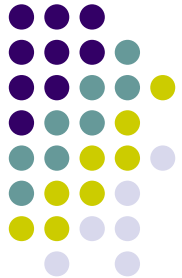


Distributed measure (κατανεμημένη μέτρηση): μπορούν να υπολογιστούν αν χωρίσουμε τα αρχικά δεδομένα σε μικρότερα υποσύνολα, υπολογίσουμε την τιμή σε κάθε υποσύνολο και τις συγχωνεύουμε πχ sum(), count(), max(), min()

Algebraic measure (αλγεβρική μέτρηση): μπορεί να υπολογιστεί αν εφαρμόσουμε μια αλγεβρική (πολυωνυμική) συνάρτηση σε μία ή περισσότερες κατανεμημένες μετρήσεις (πχ avg()= sum()/count())

Holistic measure (ολιστική μέτρηση) πρέπει να υπολογιστεί στο σύνολο των δεδομένων

Διασπορά

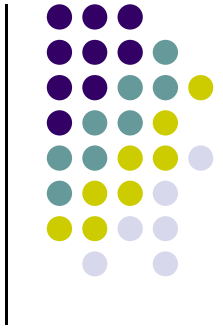


Mean (μέση τιμή)

Variance (σ^2)

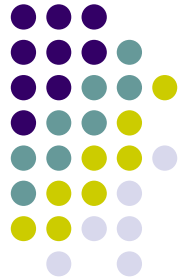
$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2]$$

Standard deviation (σ)



Απόσταση και Ομοιότητα

Κριτήρια Ομοιότητας -Απόσταση



Ομοιότητα

Μια αριθμητική μέτρηση για το πόσο όμοια είναι δυο αντικείμενα

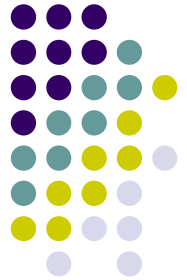
- Μεγαλύτερη όσο πιο όμοια είναι τα αντικείμενα μεταξύ τους
- Συχνά τιμές στο [0, 1]

Μη Ομοιότητα (dissimilarity)

Μια αριθμητική μέτρηση για το πόσο διαφορετικά είναι δυο αντικείμενα

- Μικρότερη όσο ποιο όμοια είναι τα αντικείμενα μεταξύ τους
- Η ελάχιστη τιμή είναι συνήθως 0 (όταν τα ίδια), αλλά το πάνω όρο διαφέρει

Κριτήρια Ομοιότητας

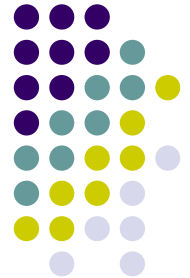


Γνωστές ιδιότητες για την ομοιότητα:

s(p, q) = 1 (ή μέγιστη ομοιότητα) μόνο αν p = q.

s(p, q) = s(q, p) για κάθε p και q (Συμμετρία)

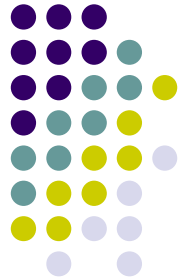
Κριτήρια Ομοιότητας -Απόσταση



Η ομοιότητα-μη ομοιότητα μεταξύ δύο αντικειμένων μετριέται συνήθως βάση μιας συνάρτησης απόστασης ανάμεσα στα αντικείμενα

Εξαρτάται από το είδος των δεδομένων, δηλαδή από το είδος των γνωρισμάτων τους

Κριτήρια Ομοιότητας



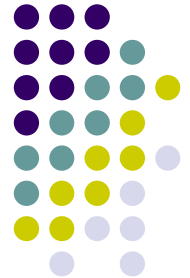
Συναρτήσεις απόστασης (distance functions)

Συχνές ιδιότητες:

1. $d(i, j) \geq 0$
2. $d(i, i) = 0$ (ανακλαστική)
3. $d(i, j) = d(j, i)$ (συμμετρική)
4. $d(i, j) \leq d(i, h) + d(h, j)$ (τριγωνική ανισότητα)

Όταν ισχύουν και οι 4, η συνάρτηση απόστασης ονομάζεται και **μετρική απόστασης (distance metric)**

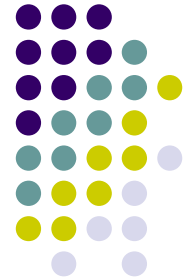
Αποστάσεις



Δύο μεγάλες κατηγορίες:

- *Ευκλείδειες* (βασισμένες στη θέση των σημείων, αποστάσεις των σημείων στο χώρο, έννοια της μέσης τιμής)
- *Μη Ευκλείδειες* (βασισμένες σε άλλες ιδιότητες των σημείων πλην της θέσης τους)

Ορισμός Απόστασης



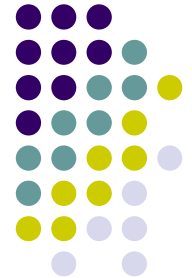
Έστω δυο μεταβλητές i και j με n γγνωρίσματα x_{ik} και x_{jk} $i = (x_{i1}, x_{i2}, \ldots, x_{in})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jn})$

Ο πιο συνηθισμένος τρόπος – Ευκλείδεια απόσταση:

$$d(i,j) = \sqrt{(|x_{i_1} - x_{j_1}|^2 + |x_{i_2} - x_{j_2}|^2 + \ldots + |x_{i_n} - x_{j_n}|^2)}$$

Είναι μετρική απόστασης

Ορισμός Απόστασης



Έστω δυο μεταβλητές i και j με n γνωρίσματα x_{ik} και x_{jk} $i = (x_{i1}, x_{i2}, \ldots, x_{in})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jn})$

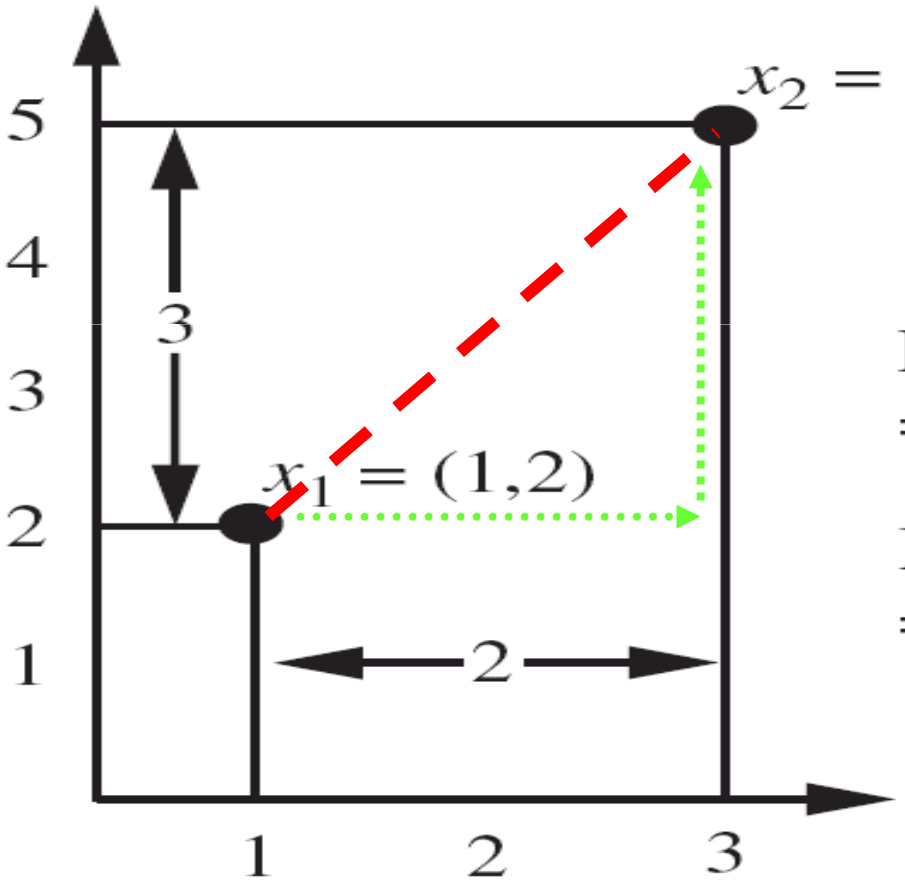
Manhattan ή city-block

$$L_1(i,j) = |x_{i_1} - x_{j_1}| + |x_{i_2} - x_{j_2}| + \ldots + |x_{i_n} - x_{j_n}|$$

Είναι μετρική απόστασης

Ορισμός Απόστασης

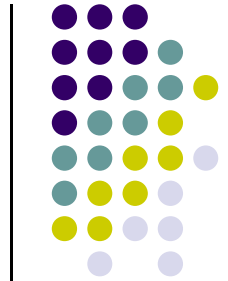
Παράδειγμα



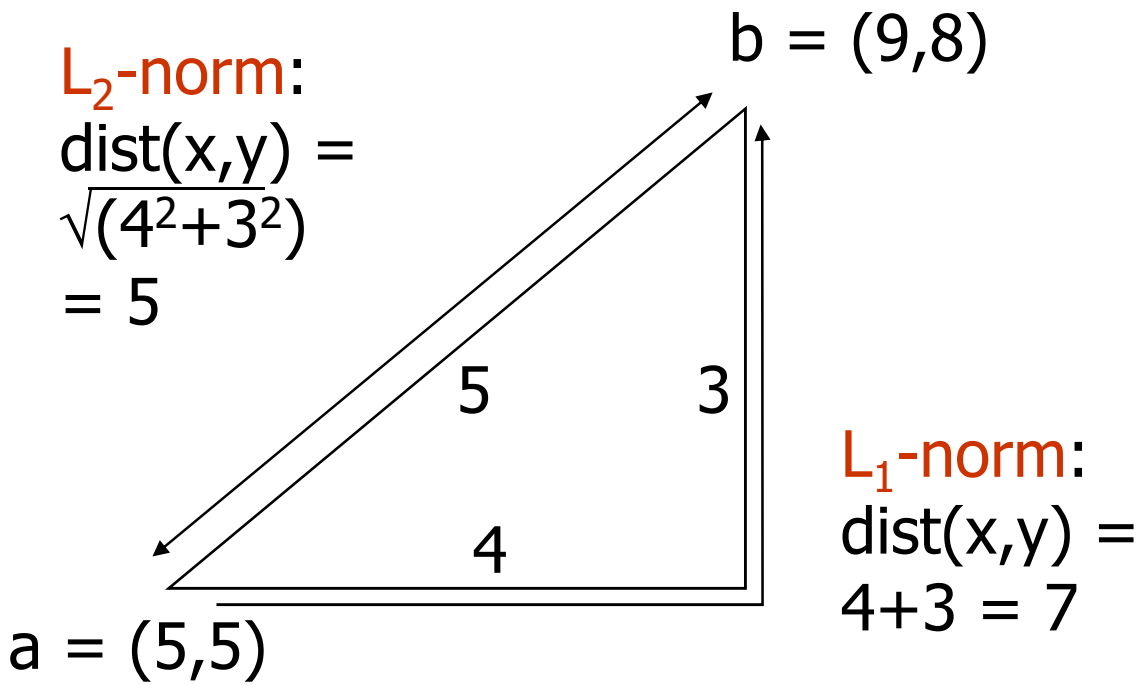
Euclidean distance
$= (2^2 + 3^2)^{1/2} = 3.61$

Manhattan distance
$= 2 + 3 = 5$

Ορισμός Απόστασης



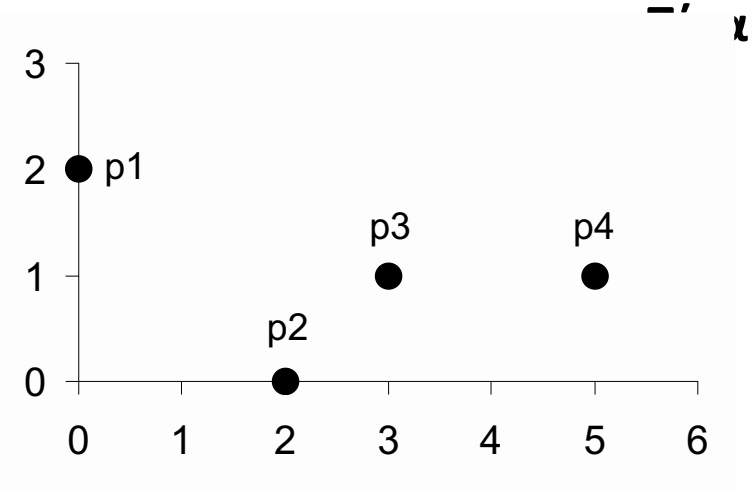
Παράδειγμα



Ορισμός Απόστασης

Παράδειγμα

–’ ικας Δεδομένων

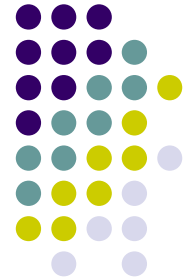


point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

Πίνακας Απόστασης

Ορισμός Απόστασης



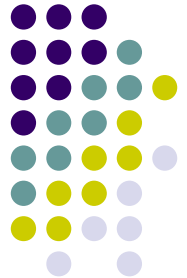
Έστω δυο μεταβλητές i και j με n γνωρίσματα x_{ik} και x_{jk} $i = (x_{i1}, x_{i2}, \ldots, x_{in})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jn})$

Minkowski (p-norm):

$$L_p(i,j) = \left(|x_{i1} - x_{j1}|^p + |x_{i2} - x_{j2}|^p + \ldots + |x_{in} - x_{jn}|^p \right)^{1/p}$$

Είναι μετρική απόστασης

Ορισμός Απόστασης

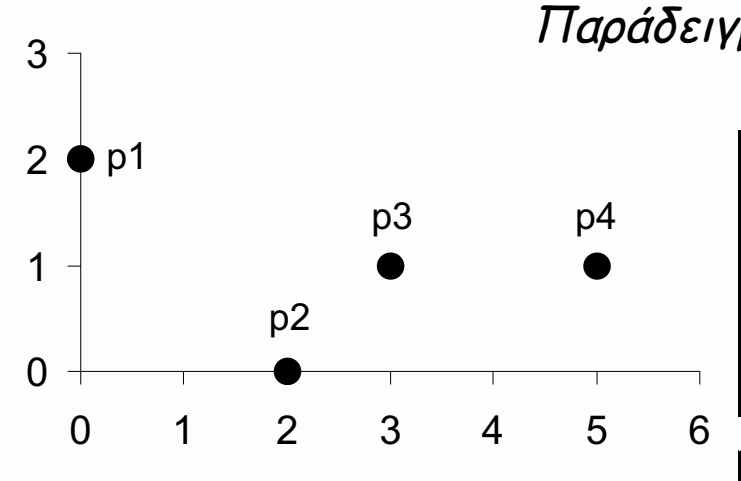


- p = 1. City block (Manhattan, taxicab, L_1 norm).
  - Hamming distance, όταν δυαδικά διανύσματα = αριθμός bits που διαφέρουν
- p = 2. Ευκλείδεια απόσταση
- $p \rightarrow \infty$. "supremum" (L_{max} norm, L_∞ norm) απόσταση.
  - Η μέγιστη απόσταση μεταξύ οποιουδήποτε γνωρίσματος (διάστασης) των δυο διανυσμάτων

το maximum το όριο όταν το p τείνει στο ∞ στην L_p norm: παίρνοντας την p^{th} των διαφορών, αθροίζοντας και παίρνοντας την n^{th} ρίζα.

Ορισμός Απόστασης

Παράδειγμα



point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

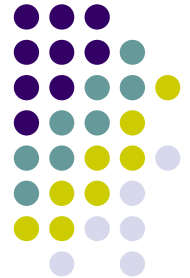
L1	p1	p2	p3	p4
p1	0	4	4	6
p2	4	0	2	4
p3	4	2	0	2
p4	6	4	2	0

L2	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

L_∞	p1	p2	p3	p4
p1	0	2	3	5
p2	2	0	1	3
p3	3	1	0	2
p4	5	3	2	0

Πίνακες Απόστασης

Ορισμός Απόστασης



Συχνά,

Βάρη πχ για Ευκλείδεια απόσταση:

$$d(i,j)=\sqrt{(w_1|x_{i1}-x_{j1}|^2+w_2|x_{i2}-x_{j2}|^2+...+w_n|x_{in}-x_{jn}|^2)}$$

Μη Ευκλείδειες αποστάσεις



- Jaccard distance
- Cosine distance
- Edit distance
- Hamming Distance

Ορισμός Απόστασης



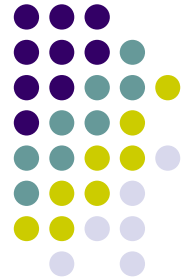
Δυαδικές Μεταβλητές

Συχνά δεδομένα με μόνο δυαδικά γνωρίσματα (δυαδικά διανύσματα)

- Συμμετρικές (τιμές 0 και 1 έχουν την ίδια σημασία)
  Invariant ομοιότητα

- Μη συμμετρικές (η συμφωνία στο 1 πιο σημαντική – πχ όταν το 1 σηματοδοτεί την ύπαρξη κάποιας ασθένειας)
  Non-invariant (Jaccard)

Ορισμός Απόστασης



Μεταξύ δύο αντικειμένων i και j με δυαδικά γνωρίσματα

M_{01} = ο αριθμός των γνωρισμάτων που το i έχει τιμή 0 και το j έχει 1

M_{10} = ο αριθμός των γνωρισμάτων που το i έχει τιμή 1 και το j έχει 0

M_{00} = ο αριθμός των γνωρισμάτων που το i έχει τιμή 0 και το j έχει 0

M_{11} = ο αριθμός των γνωρισμάτων που το i έχει τιμή 1 και το j έχει 1

ΟΜΟΙΟΤΗΤΑ

Απλό ταίριασμα – συμμετρικές μεταβλητές

SMC = αριθμός ταιριασμάτων / αριθμός γνωρισμάτων

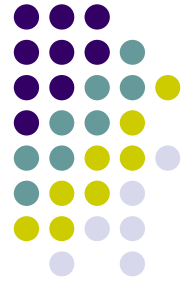
$= (M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00})$

J = αριθμός 11 ταιριασμάτων / αριθμό μη μηδενικών γνωρισμάτων

$= (M_{11}) / (M_{01} + M_{10} + M_{11})$

J - Συντελεστής Jaccard – Jaccard Coefficient – μη συμμετρικές μεταβλητές (διαφορετική σημασία στην τιμή 1 και στην τιμή 0)

Ορισμός Απόστασης



Παράδειγμα

p = 1 0 0 0 0 0 0 0 0 0

q = 0 0 0 0 0 0 1 0 0 1

$M_{01} = 2$

$M_{10} = 1$

$M_{00} = 7$

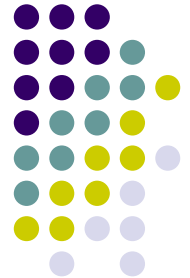
$M_{11} = 0$

SMC = $(M_{11} + M_{00})/(M_{01} + M_{10} + M_{11} + M_{00})$ = (0+7) / (2+1+0+7) = 0.7

J = (M_{11}) / $(M_{01} + M_{10} + M_{11})$ = 0 / (2 + 1 + 0) = 0

J = #1(p BAND q) / #1(p BOR Q)

Ορισμός Απόστασης



Contigency πίνακας για δυαδικά δεδομένα

		Αντικείμενο *j*	
		1	0
Αντικείμενο *i*	1	M_{11}	M_{10}
	0	M_{01}	M_{00}

Μέτρηση απόστασης για συμμετρικές δυαδικές μεταβλητές

1 – συμμετρική-ομοιότητα

$$d(i,j) = \frac{M_{10} + M_{01}}{M_{11} + M_{10} + M_{01} + M_{00}}$$

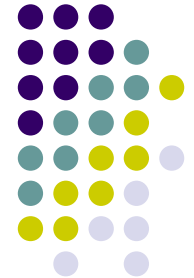
Μέτρηση απόστασης για μη συμμετρικές δυαδικές μεταβλητές

$$d(i,j) = \frac{M_{10} + M_{01}}{M_{11} + M_{10} + M_{01}}$$

Jaccard coefficient

$$sim_{Jaccard}(i,j) = \frac{M_{11}}{M_{11} + M_{10} + M_{01}}$$

Ορισμός Απόστασης



Παράδειγμα

τα γνωρίσματα μη συμμετρικά
Έστω Y-P να αντιστοιχούν στο 1 και το N στο 0

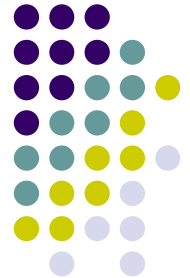
Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

$$d(jack, mary) = \frac{0+1}{2+0+1} = 0.33$$

$$d(jack, jim) = \frac{1+1}{1+1+1} = 0.67$$

$$d(jim, mary) = \frac{1+2}{1+1+2} = 0.75$$

Ορισμός Απόστασης



Κατηγορικές Μεταβλητές χωρίς Διάταξη (nominal)

Γενίκευση των δυαδικών μεταβλητών (γνωρισμάτων) όπου μπορούν να πάρουν παραπάνω από 2 τιμές, πχ κόκκινο, πράσινο, κίτρινο

1η Μέθοδος: Απλό ταίριασμα

m: # ταιριάσματα, *p*: συνολικός # μεταβλητών

$$d(i, j) = \frac{p - m}{p}$$

2η Μέθοδος: Χρήση πολλών δυαδικών μεταβλητών

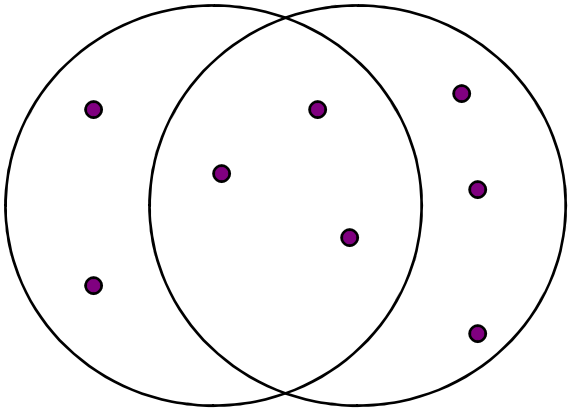
Μία για κάθε μία από τις μ τιμές

Jaccard ομοιότητα για Σύνολα



- Η Jaccard ομοιότητα για δύο σύνολα είναι το μέγεθος της τομής προς το μέγεθος της ένωσης τους
  - Sim $(C_1, C_2) = |C_1 \cap C_2| / |C_1 \cup C_2|$.

Παράδειγμα



3 τομή.
8 ένωση.
Jaccard ομοιότητα
= 3/8

Jaccard ομοιότητα για Σύνολα



Δυαδική αναπαράσταση συνόλου (1 το στοιχείο υπάρχει, 0 Αλλιώς)

p_1 = 10111; p_2 = 10011.

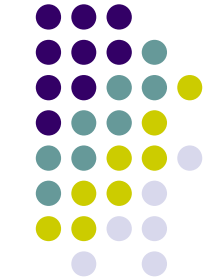
- Μέγεθος τομής = 3; Μέγεθος ένωσης = 4, ομοιότητα Jaccard (όχι απόσταση) = 3/4.
- d(x,y) = 1 – (ομοιότητα Jaccard) = 1/4.

Jaccard ομοιότητα για Σύνολα



- $d(x,x) = 0$ γιατί $x \cap x = x \cup x$.
- $d(x,y) = d(y,x)$ γιατί η ένωση και η τομή είναι συμμετρικές
- $d(x,y) \geq 0$ γιατί $|x \cap y| \leq |x \cup y|$.
- $d(x,y) \leq d(x,z) + d(z,y)$?.

Ορισμός Απόστασης



Ομοιότητα συνημίτονου (cosine similarity)

- Αν d_1 and d_2 είναι διανύσματα κειμένου

$$\cos(d_1, d_2) = (d_1 \bullet d_2) / ||d_1|| \, ||d_2|| ,$$

όπου • εσωτερικό γινόμενο $||d||$ το μήκος του d.

- Παράδειγμα:

d_1 = 3 2 0 5 0 0 0 2 0 0
d_2 = 1 0 0 0 0 0 0 1 0 2

Θέλουμε μια απόσταση που να αγνοεί τα 0 (όπως η Jaccard) αλλά να δουλεύει και για μη δυαδικά δεδομένα

Επίσης, αγνοεί το *μήκος* των διανυσμάτων

$d_1 \bullet d_2$= 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5

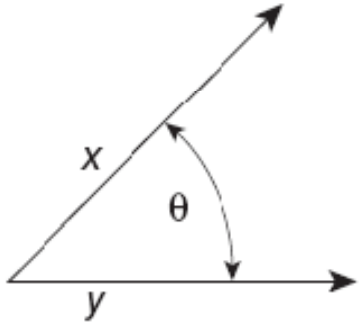
$||d_1||$ = (3*3+2*2+0*0+5*5+0*0+0*0+0*0+2*2+0*0+0*0)$^{0.5}$ = (42) $^{0.5}$ = 6.481

$||d_2||$ = (1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2) $^{0.5}$ = (6) $^{0.5}$ = 2.245

$\cos(d_1, d_2)$ = .3150

Ορισμός Απόστασης

Ομοιότητα συνημίτονου (cosine similarity)

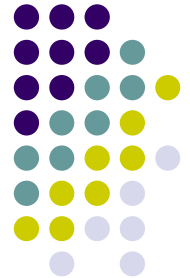


Γεωμετρική ερμηνεία

Ομοιότητα 1, όταν η γωνία 0 – που σημαίνει ότι τα x και y ίδια (αν εξαιρέσουμε το μήκος τους)

Ομοιότητα 0, όταν η γωνία 90 (κανένας κοινός όρος)

Απόσταση Edit



- Για δύο συμβολοσειρές (strings) ο ελάχιστος αριθμός εισαγωγών/διαγραφών χαρακτήρων που χρειάζονται για να πάμε από τη μία στην άλλη

- *x* = *abcde* ; *y* = *bcduve*.
- Turn *x* into *y* by deleting *a*, then inserting *u* and *v* after *d*.
  - Edit distance = 3..

Παρένθεση –Ορισμοί



Κλείνει η παρένθεση

Γενικές Απαιτήσεις



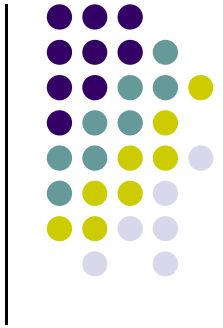
- Scalability – στον αριθμό σημείων και διαστάσεων
- Να υποστηρίζει διαφορετικούς τύπους δεδομένων
- Να υποστηρίζει συστάδες με διαφορετικά σχήματα (συνήθως, «σφαίρες»)
- Να είναι εύκολο να δώσουμε τιμές στις παραμέτρους εισόδου (αριθμό συστάδων, μέγεθος κλπ)
- Να μην εξαρτάται από τη *σειρά* επεξεργασίας των σημείων εισόδου

Γενικές Απαιτήσεις



- Δυναμικά μεταβαλλόμενα δεδομένα
  - Αλλαγή συστάδων με το πέρασμα του χρόνου
- Απόδοση (scaling)
  - Disk-resident vs Main memory

Είδη Συστάδων



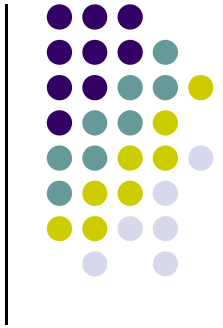
- Καλώς διαχωρισμένες συστάδες
- Συστάδες βασισμένες σε κέντρο
- Συνεχής (contiguous) συστάδες
- Συστάδες Βασισμένες σε πυκνότητα
- Βασισμένα σε ιδιότητες ή έννοιες
- Περιγράφονται από μια αντικειμενική συνάρτηση (Objective Function)

Αλγόριθμοι Συσταδοποίησης



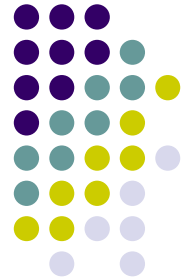
Θα δούμε ανάμεσα σε άλλους τους:

- K-means και παραλλαγές
- Ιεραρχική Συσταδοποίηση
- Συσταδοποίηση με βάση την Πυκνότητα (DBSCAN)
- BIRCH (δεδομένα στο δίσκο!)



Ιεραρχική Συσταδοποίηση

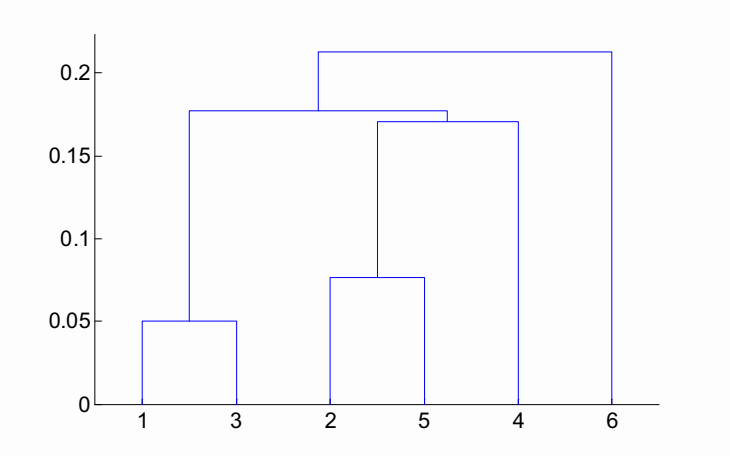
Ιεραρχική Συσταδοποίηση: Βασικά

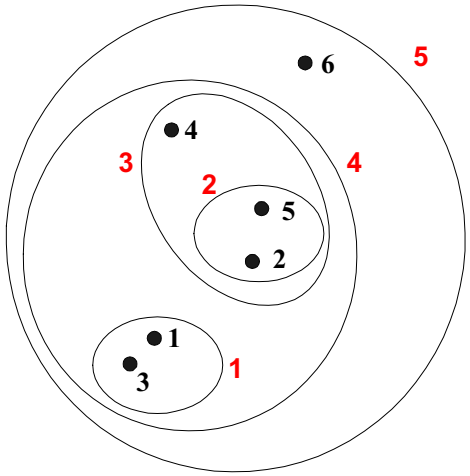


Παράγει ένα σύνολο από εμφωλευμένες συστάδες οργανωμένες σε ένα ιεραρχικό δέντρο

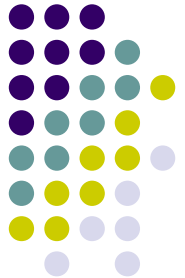
Μπορεί να παρασταθεί με ένα δένδρο-γραμμα

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





Ιεραρχική Συσταδοποίηση: Πλεονεκτήματα



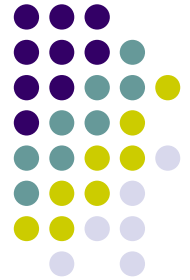
- Δε χρειάζεται να υποθέσουμε ένα συγκεκριμένο αριθμό από συστάδες

  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, …)

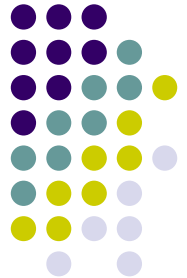
Ιεραρχική Συσταδοποίηση



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)

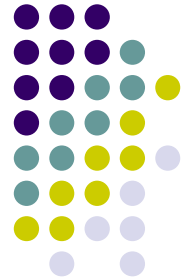
Ιεραρχική Συσταδοποίηση



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

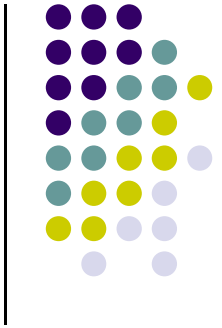
Βασικός Αλγόριθμος

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

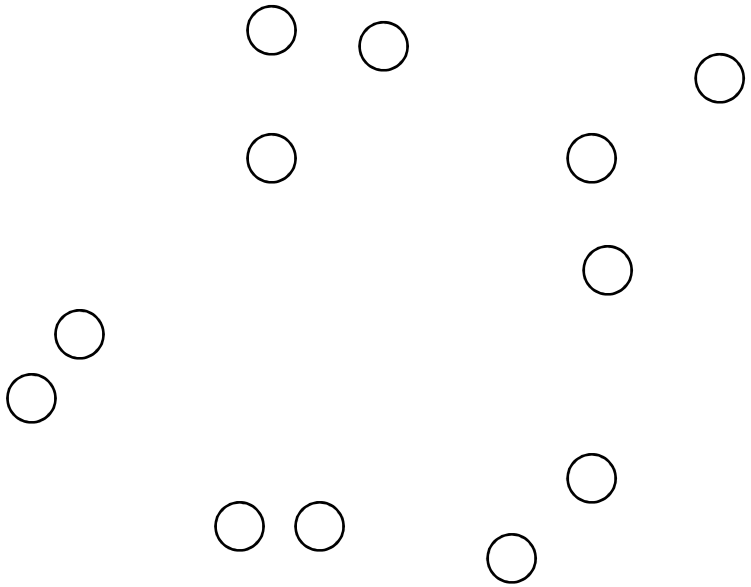
Βασική λειτουργία είναι ο υπολογισμός της γειτνίασης δυο συστάδων

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)



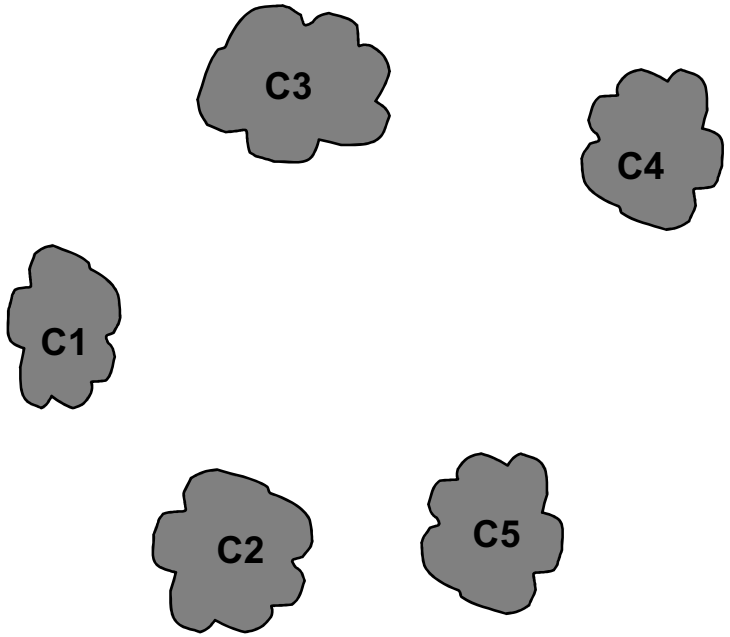
	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Πίνακας Γειτνίασης



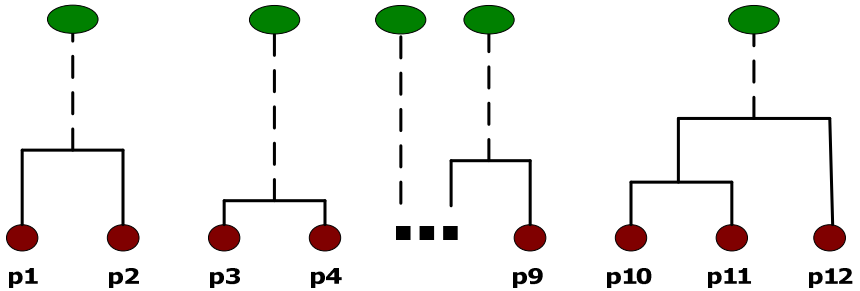
Συσσωρευτική Ιεραρχική Συσταδοποίηση

Μετά από κάποιες συγχωνεύσεις, έχουμε κάποιες συστάδες



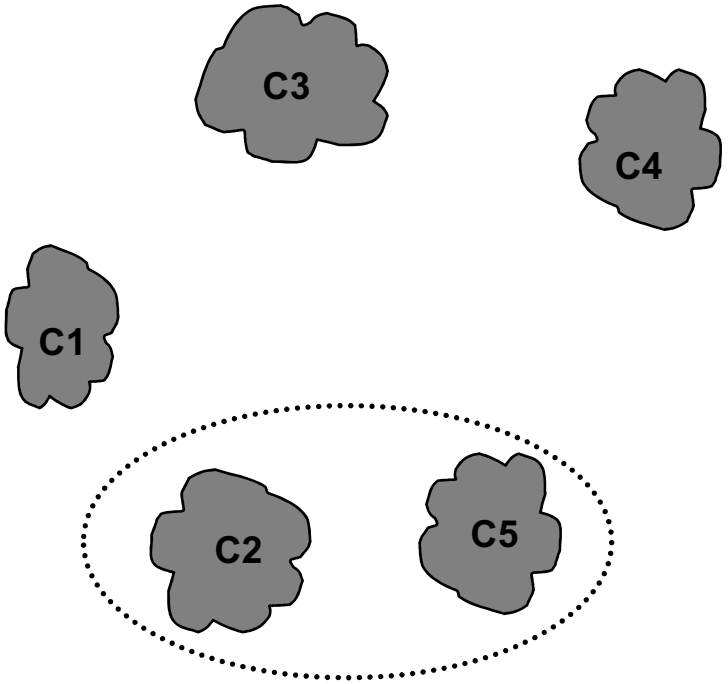
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Πίνακας Γειτνίασης



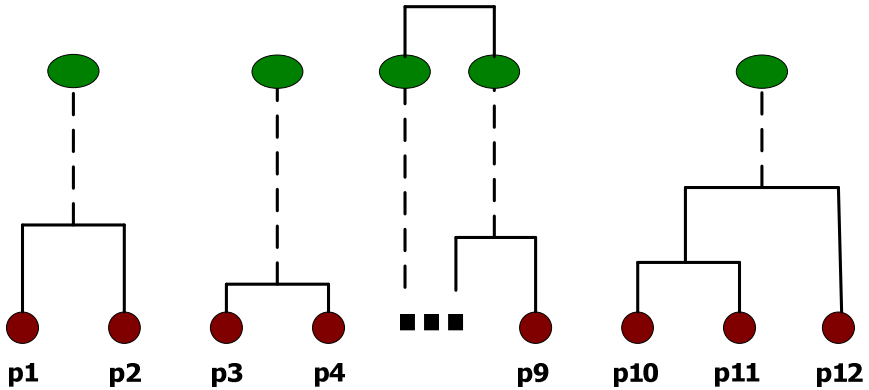
Συσσωρευτική Ιεραρχική Συσταδοποίηση

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



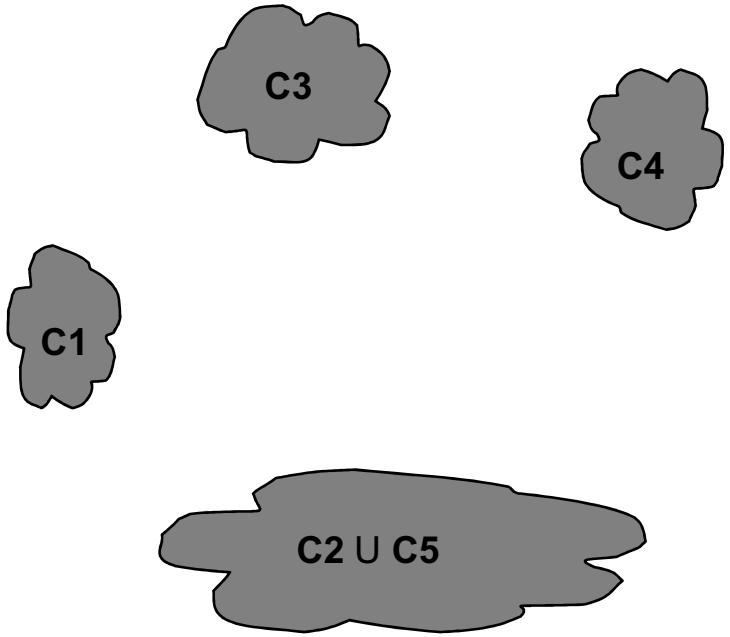
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Πίνακας Γειτνίασης



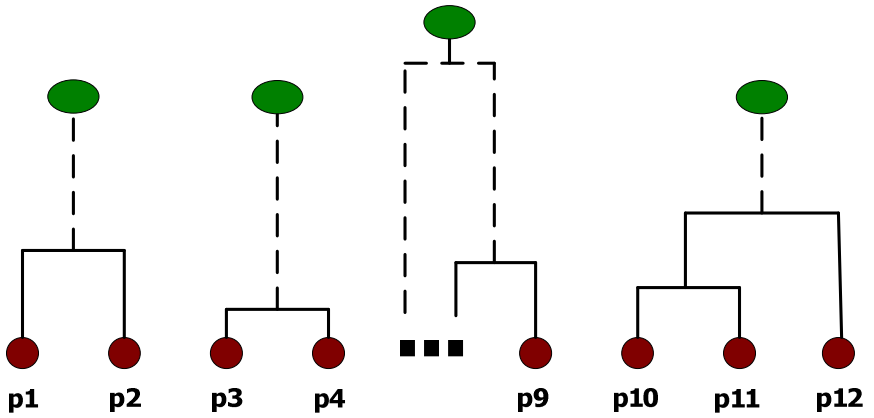
Συσσωρευτική Ιεραρχική Συσταδοποίηση

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

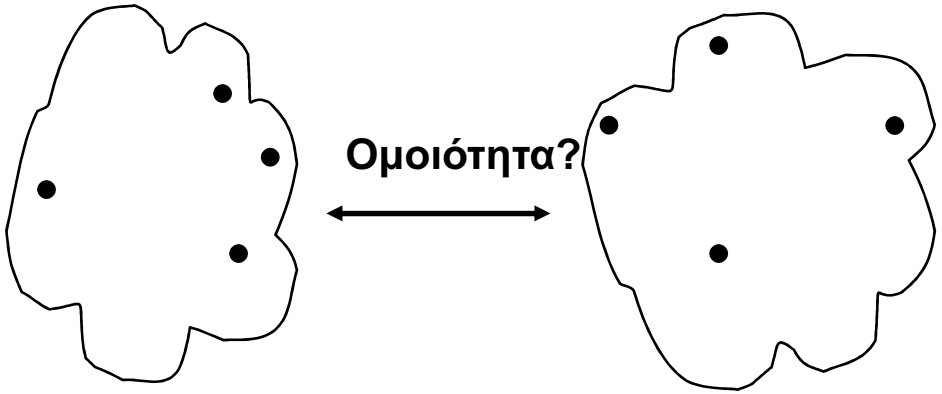


	C1	C2 U C5	C3	C4
C1		?		
C2 U C5	?	?	?	?
C3		?		
C4		?		

Πίνακας Γειτνίασης



ΣΙΣ: Ορισμός απόστασης μεταξύ συστάδων

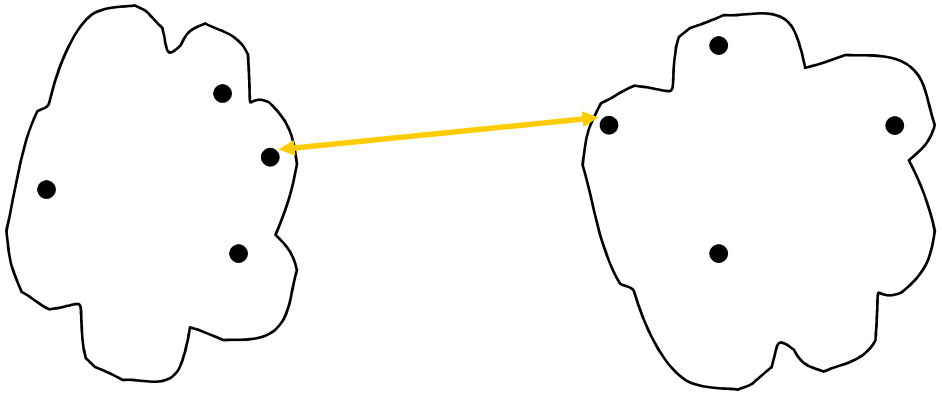


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

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

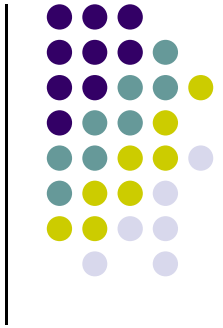


- **MIN**
- MAX
- Μέσος όρος της ομάδας
- Η απόσταση μεταξύ των κεντρικών σημείων
- Άλλες μέθοδοι βασισμένες σε μια αντικειμενική συνάρτηση
  - Η μέθοδος του Ward χρησιμοποιεί τετραγωνικά λάθη

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

Πίνακας Γειτνίασης

ΣΙΣ: Ορισμός απόστασης μεταξύ συστάδων: MIN



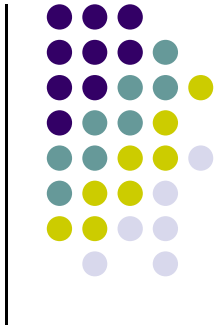
MIN ή μοναδικής ακμής ή απλού συνδέσμου (single link)

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



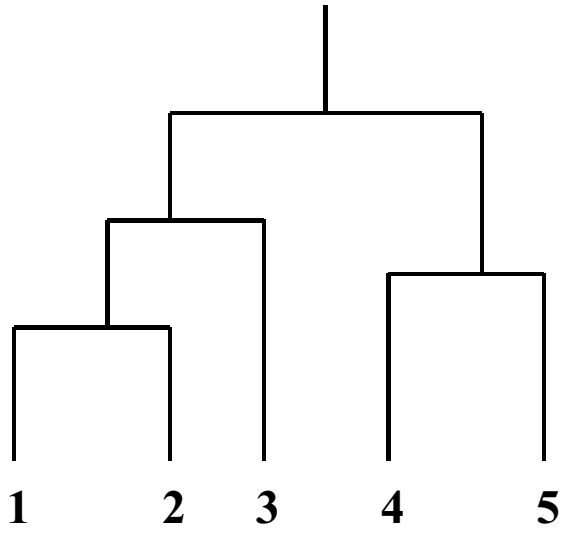
MIN ή μοναδικής ακμής ή απλού συνδέσμου (single link)

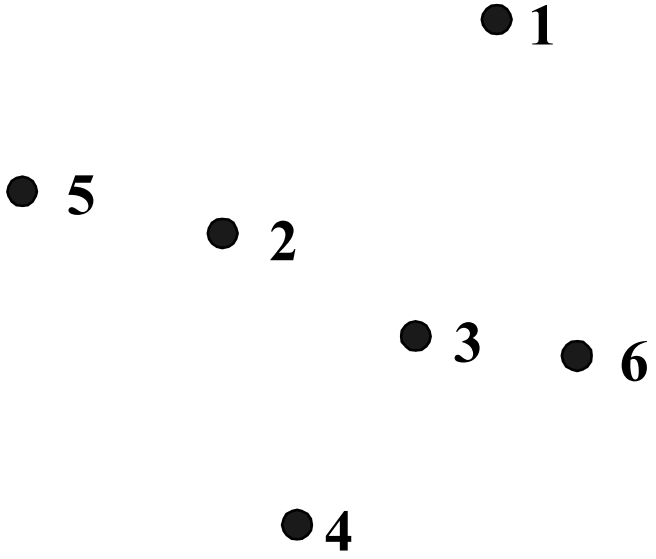
Η ομοιότητα μεταξύ δυο συστάδων βασίζεται στα δυο πιο όμοια (πιο γειτονικά) σημεία στις διαφορετικές συστάδες (με όρους γραφημάτων – shortest edge)

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)

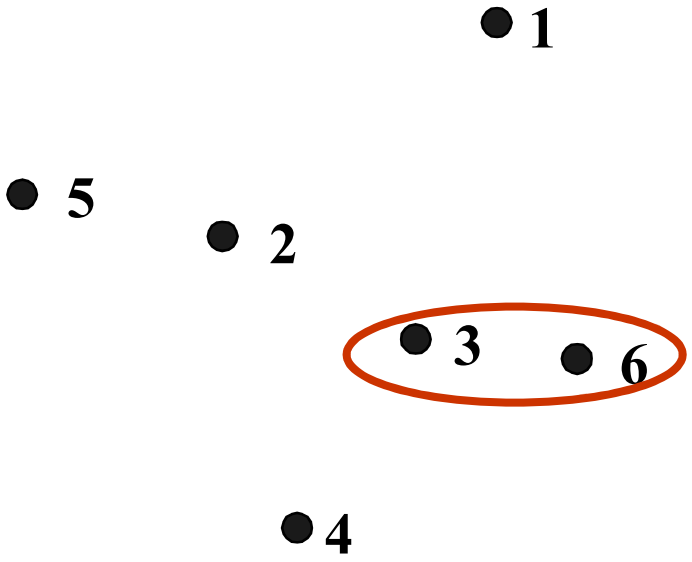
2 (0.22, 0.38)

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)

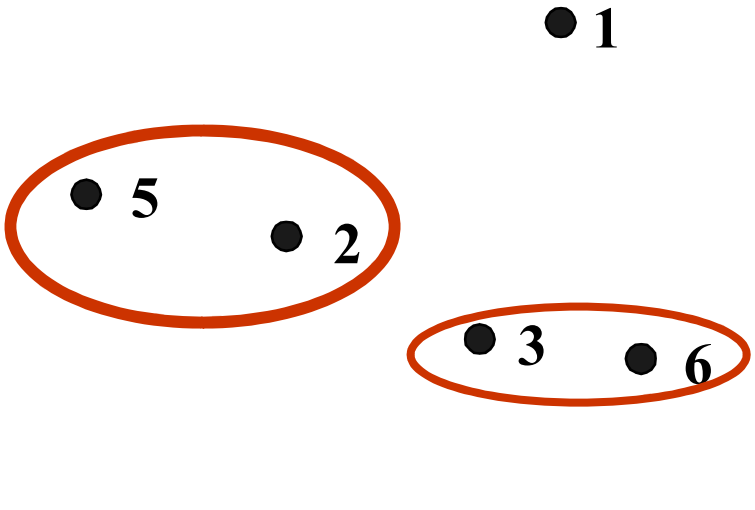
2 (0.22, 0.38)

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)

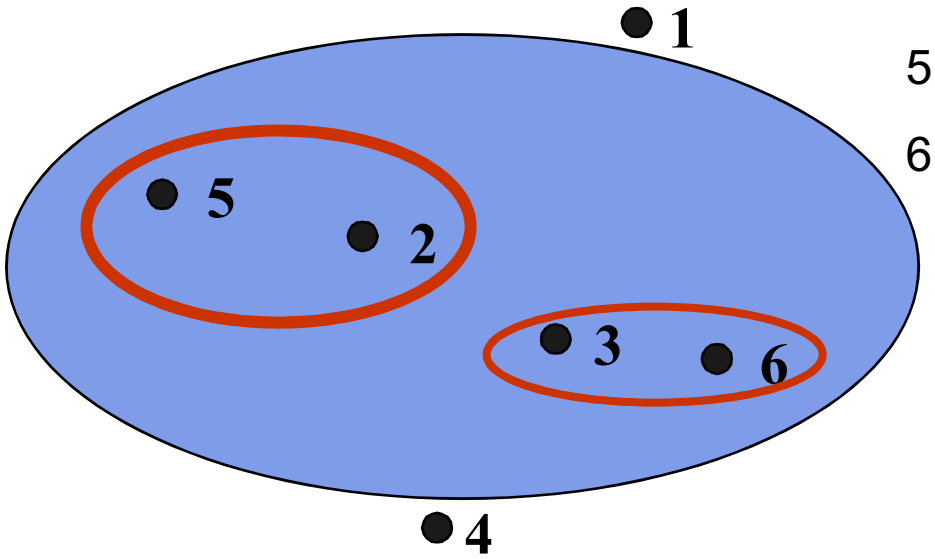
2 (0.22, 0.38)

3 (0.35, 0.32)

4 (0.26, 0.19)

5 (0.08, 0.41)

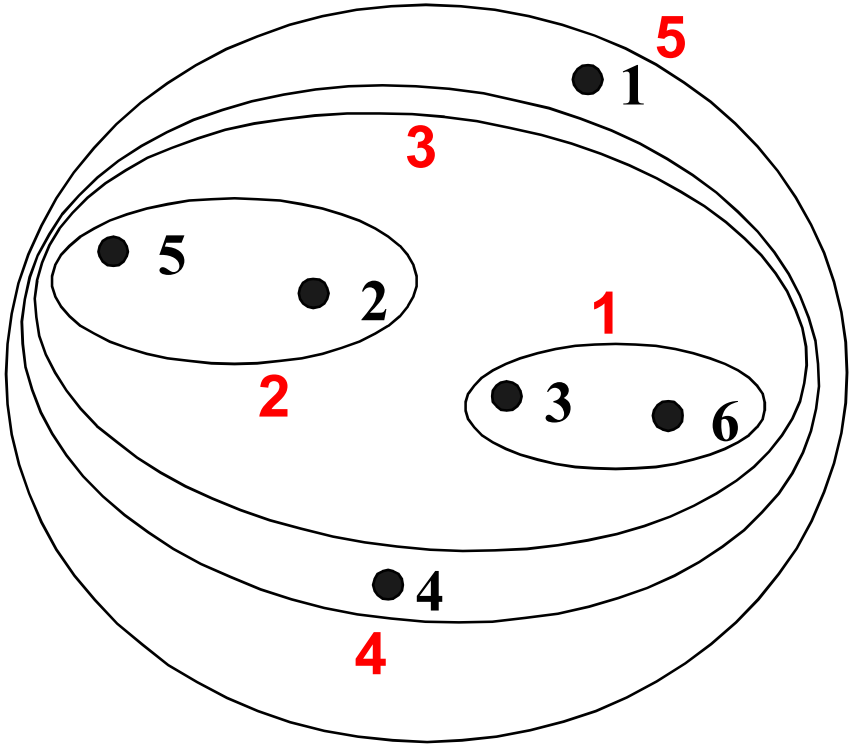
6 (0.45, 0.30)



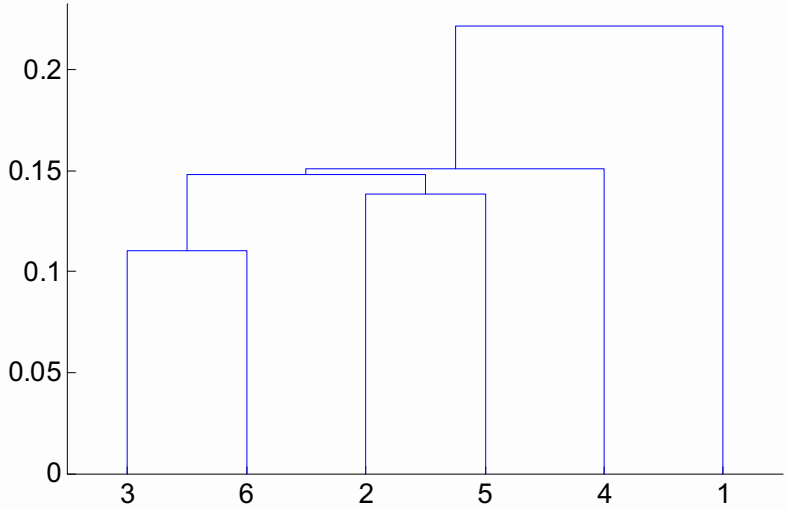
Αρκεί να «δω» μια ακμή

	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



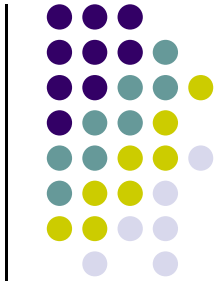
Φωλιασμένες Συστάδες



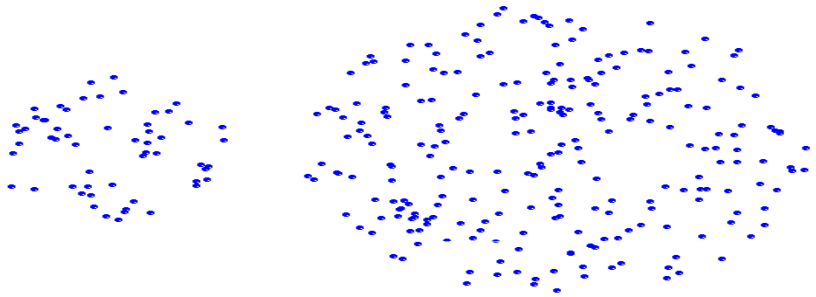
Δεντρόγραμμα

Το δεντρόγραμμα (y-άξονας) δίνει και τις αποστάσεις

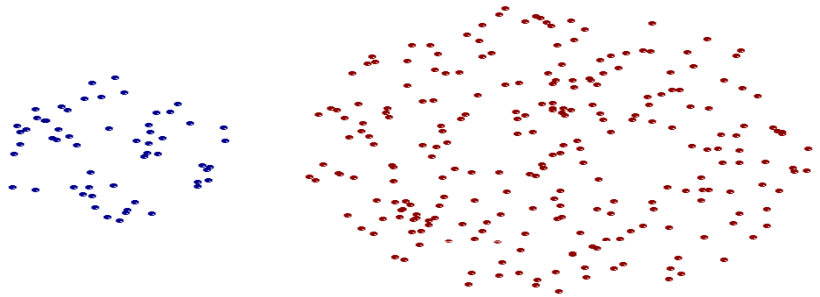
ΣΙΣ: Ορισμός απόστασης μεταξύ συστάδων: MIN



Προτερήματα



Αρχικά σημεία

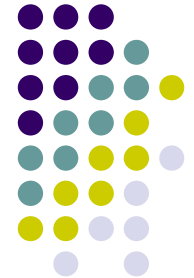


Δύο συστάδες

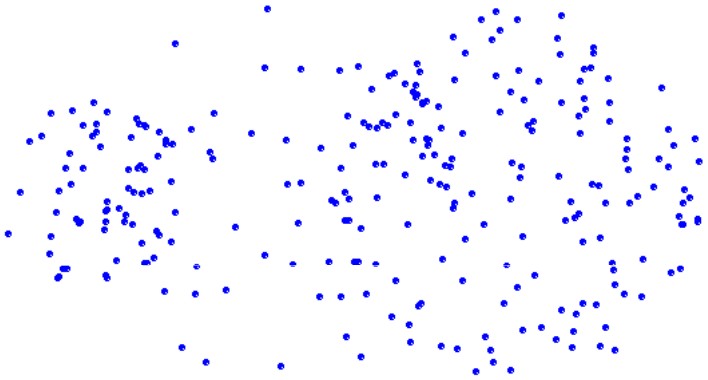
Contiguity-based (συνεχόμενες συστάδες)

Μπορεί να χειριστεί μη ελλειπτικά (non-elliptical) σχήματα

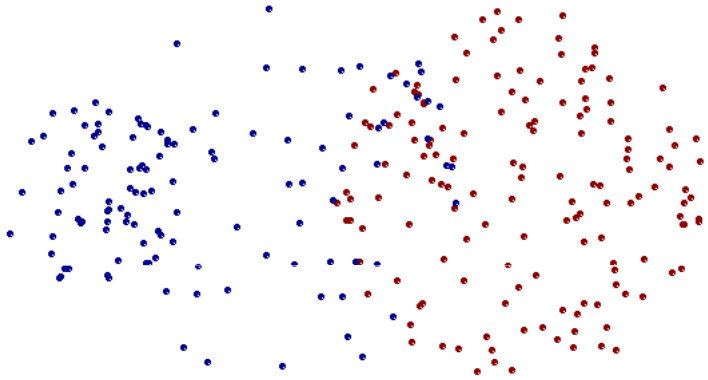
ΣΙΣ: Ορισμός απόστασης μεταξύ συστάδων: MIN



Μειονεκτήματα



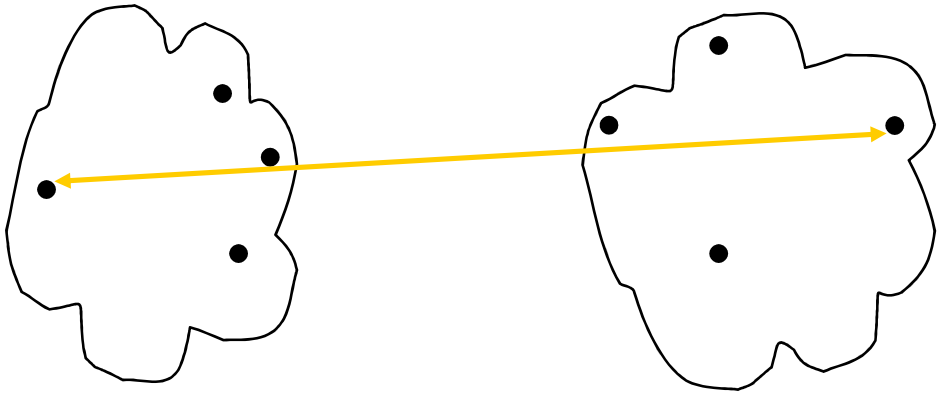
Αρχικά σημεία



Δύο συστάδες

- **Ευαίσθητο σε θόρυβο και outliers**

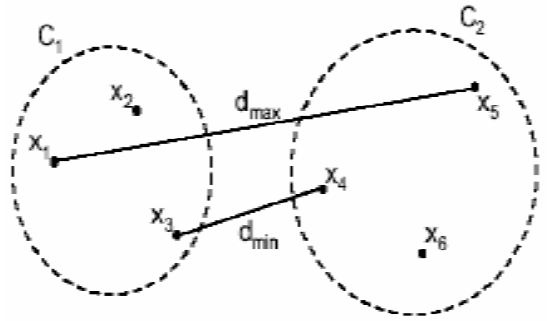
ΣΙΣ: Ορισμός απόστασης μεταξύ συστάδων



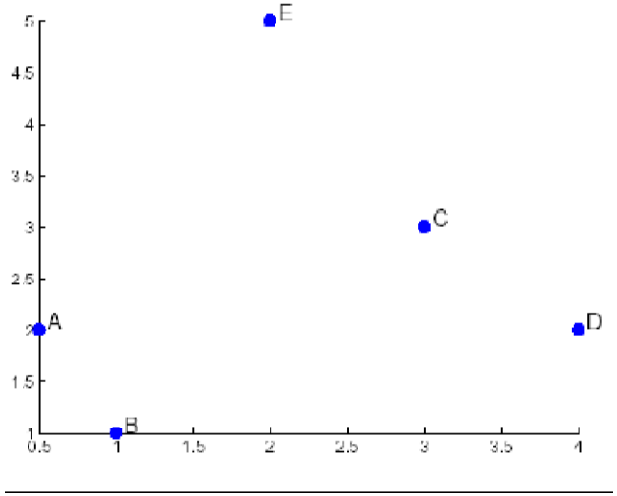
	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						

Πίνακας Γειτνίασης

- 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

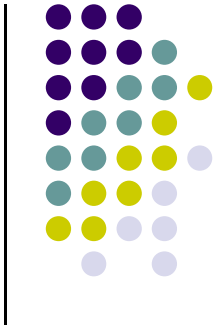


ΣΙΣ: Ορισμός απόστασης μεταξύ συστάδων



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



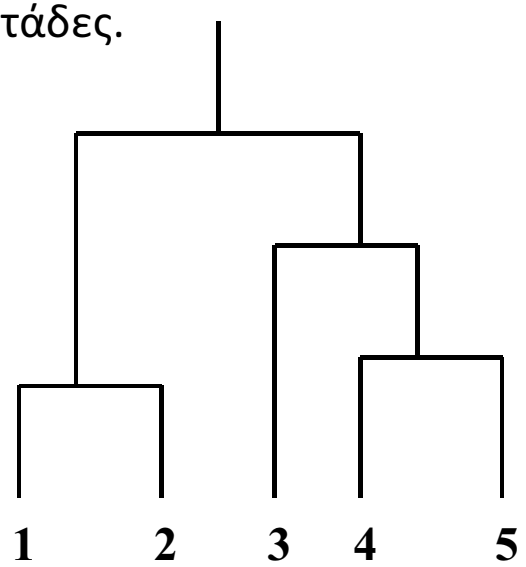
MAX ή πλήρους συνδεσιμότητας (complete linkage)
- Αναζητά κλίκες

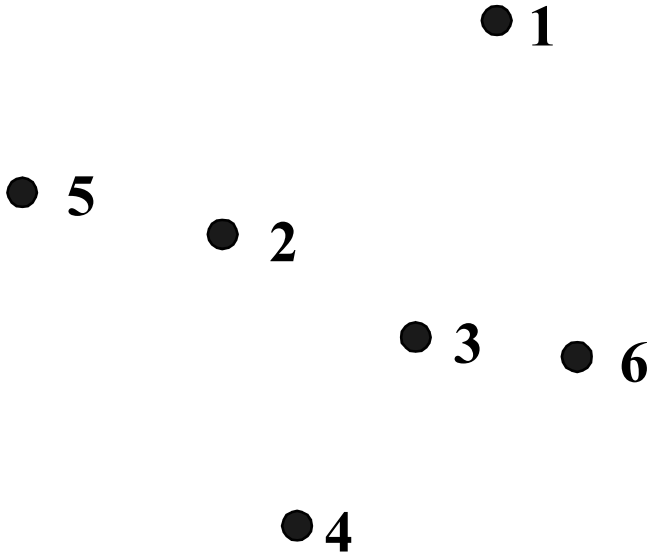
Η ομοιότητα μεταξύ δυο συστάδων βασίζεται στα δυο λιγότερο όμοια (πιο μακρινά) σημεία στις διαφορετικές συστάδες (longest edge) – δηλαδή, οι συστάδες με την μικρότερη τέτοια απόσταση

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)

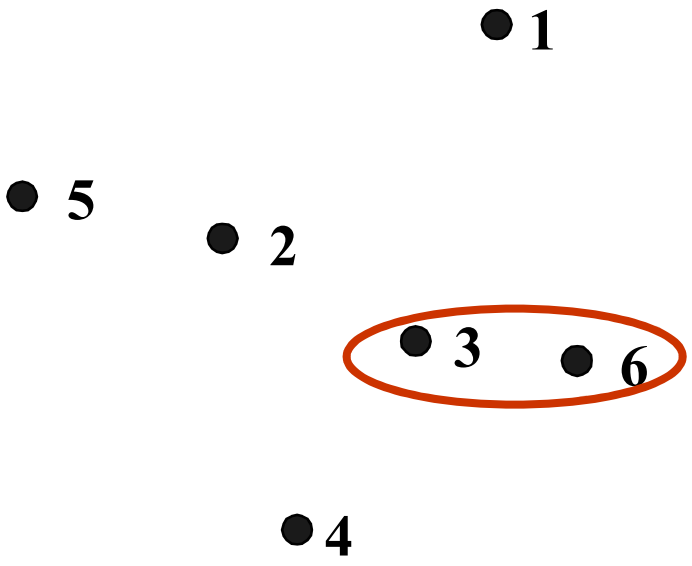
2 (0.22, 0.38)

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)

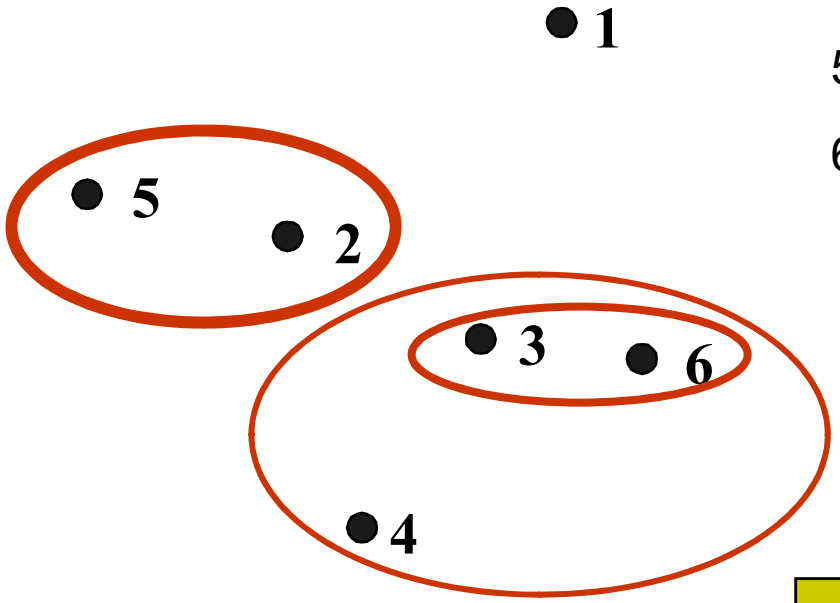
2 (0.22, 0.38)

3 (0.35, 0.32)

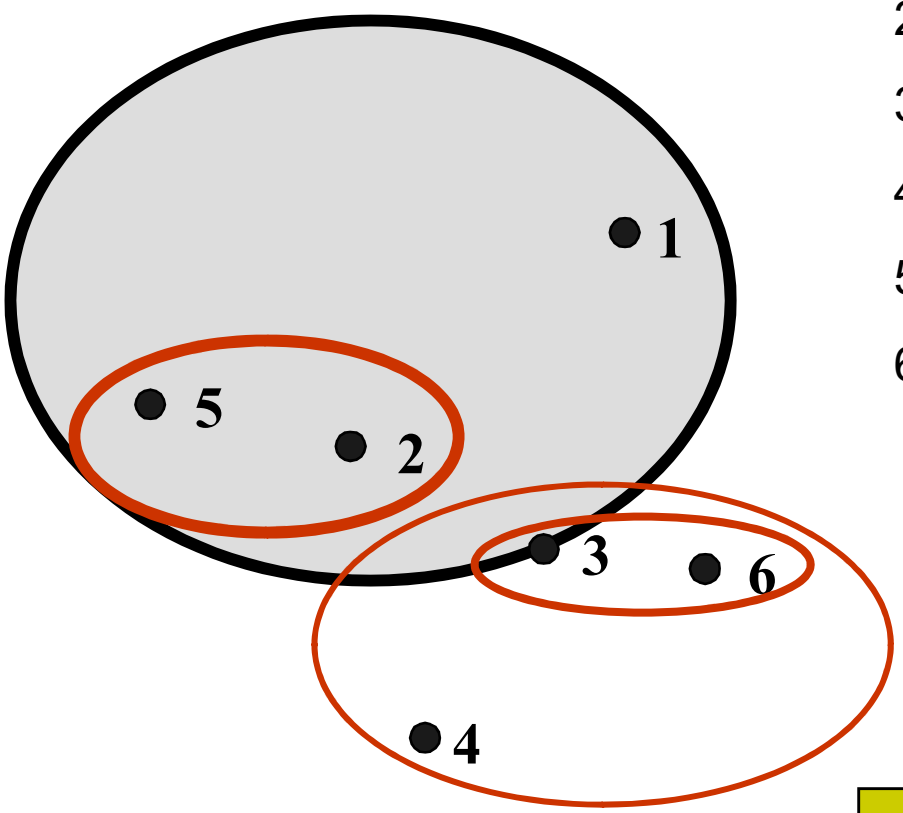
4 (0.26, 0.19)

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**



1 (0.4, 0.53)

2 (0.22, 0.38)

3 (0.35, 0.32)

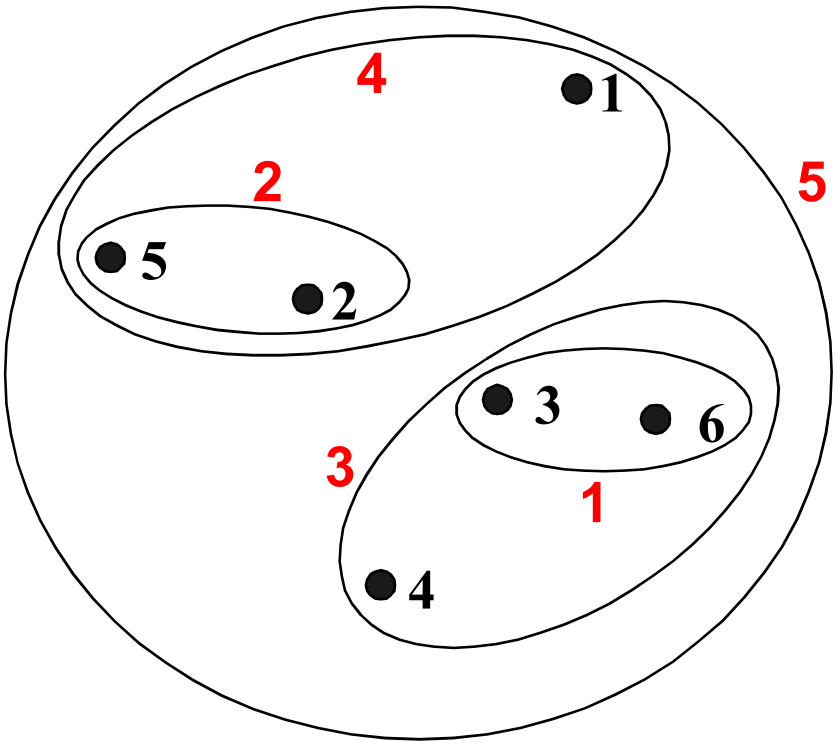
4 (0.26, 0.19)

5 (0.08, 0.41)

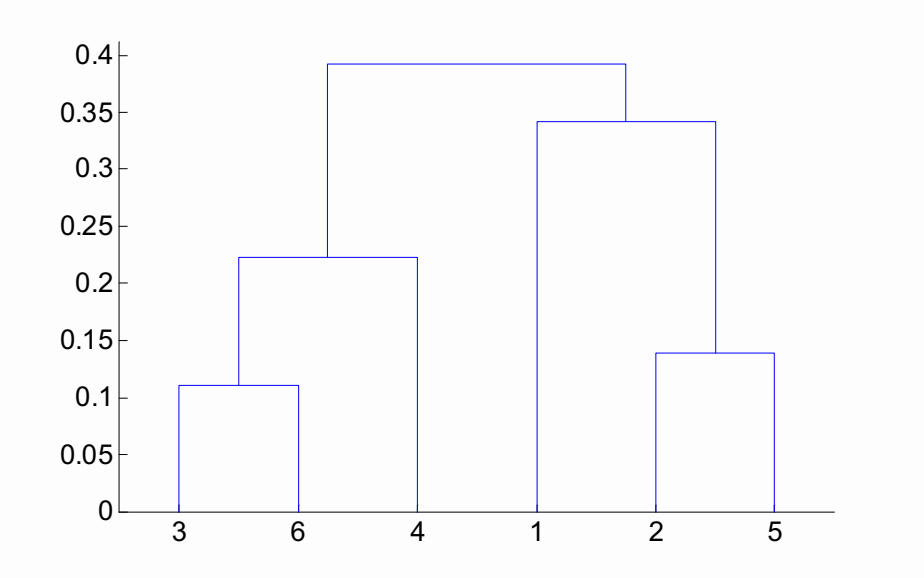
6 (0.45, 0.30)

	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

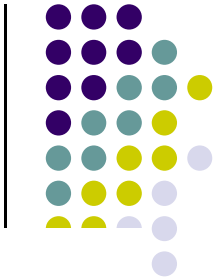


Φωλιασμένες Συστάδες

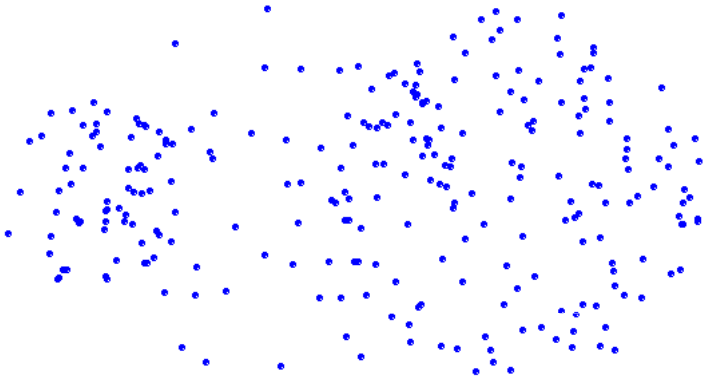


Δεντρόγραμμα

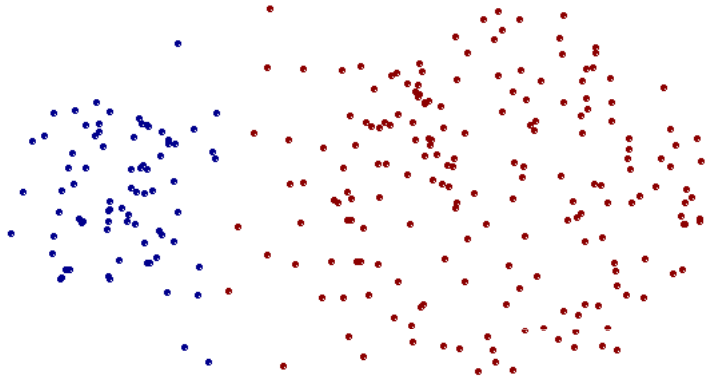
ΣΙΣ: Ορισμός απόστασης μεταξύ συστάδων: MAX



Πλεονεκτήματα



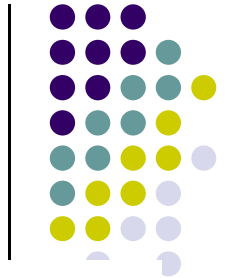
Αρχικά Σημεία



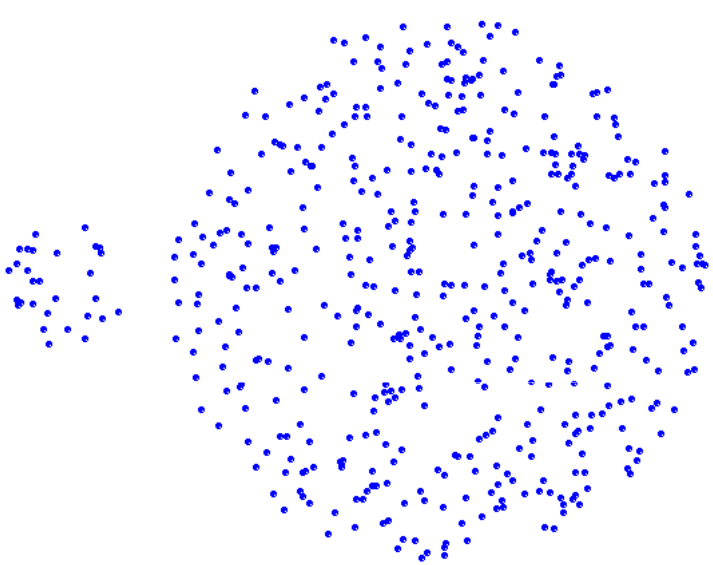
Δύο Συστάδες

- **λιγότερη εξάρτηση σε θόρυβο και outliers**

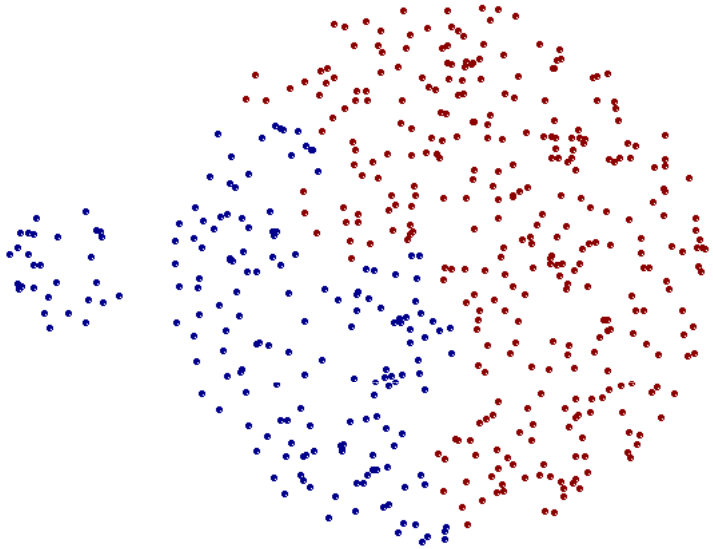
ΣΙΣ: Ορισμός απόστασης μεταξύ συστάδων: MAX



Μειονεκτήματα



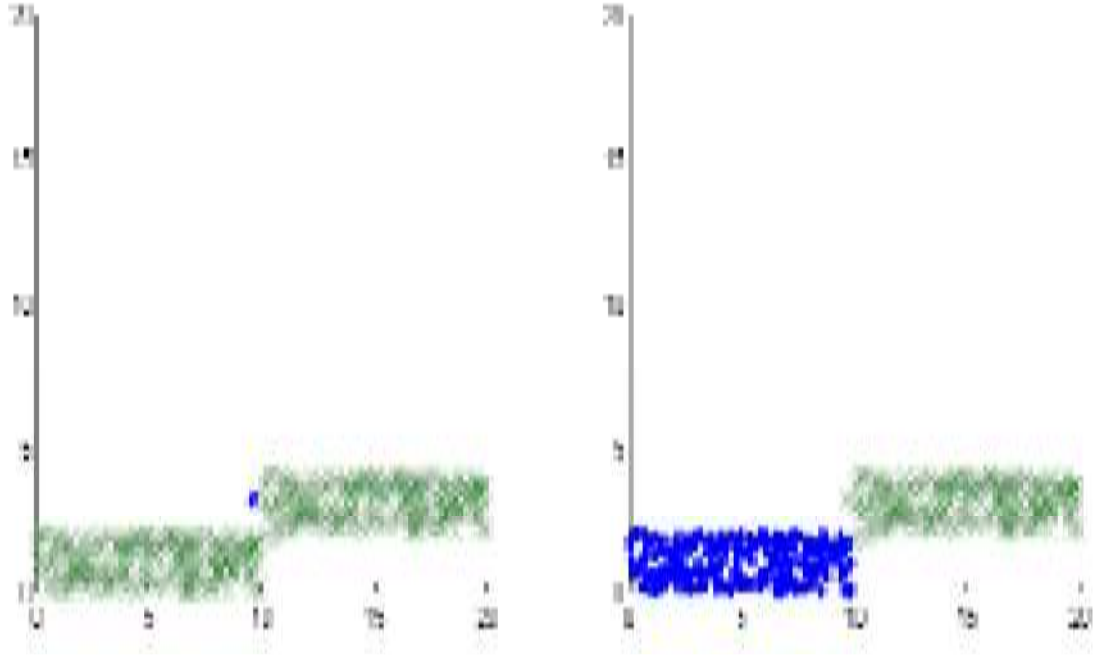
Αρχικά σημεία



Δύο συστάδες

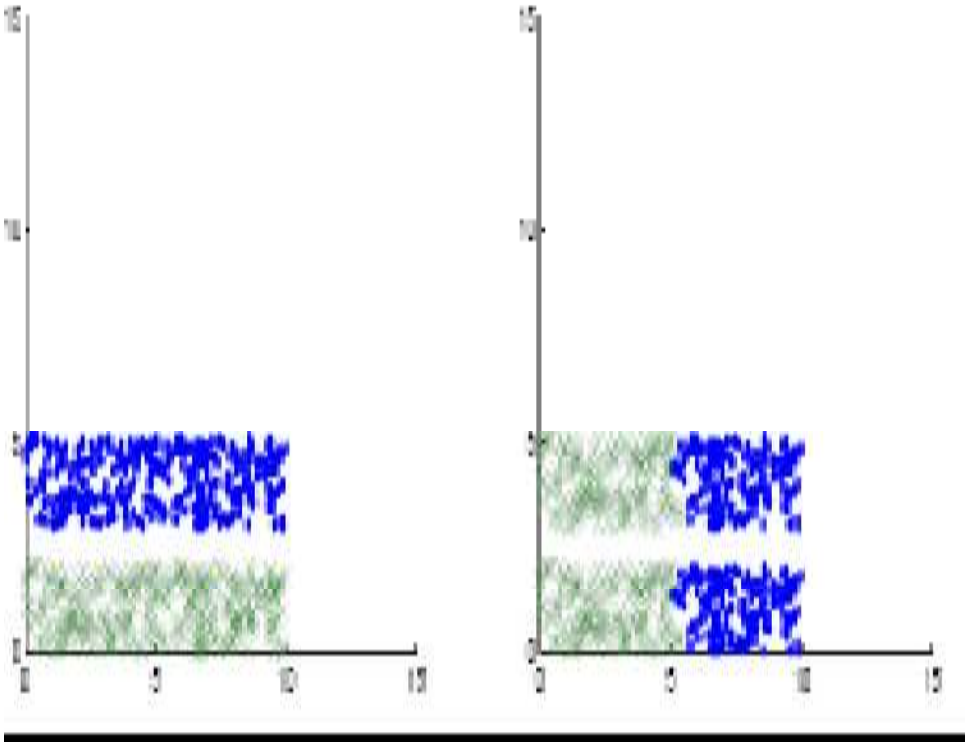
- **Τείνει να διασπά μεγάλες συστάδες**
- **Οδηγεί συνήθως σε κυκλικά σχήματα**

ΣΙΣ: Ορισμός απόστασης μεταξύ συστάδων: MAX

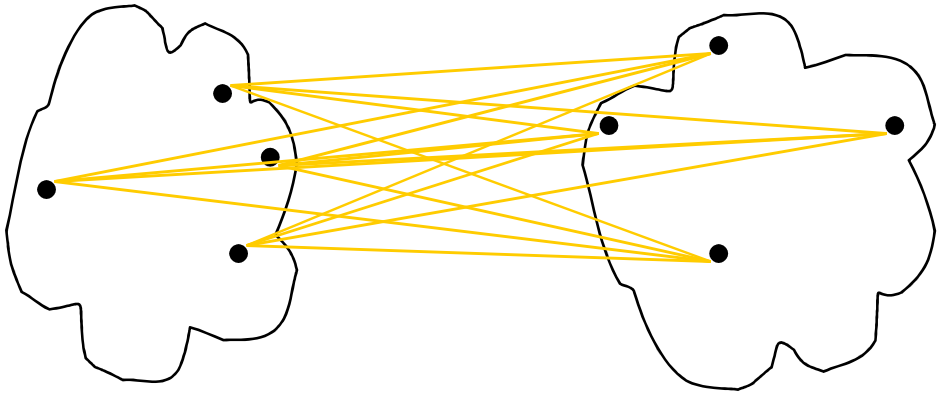


MIN MAX

ΣΙΣ: Ορισμός απόστασης μεταξύ συστάδων: MAX



ΣΙΣ: Ορισμός απόστασης μεταξύ συστάδων

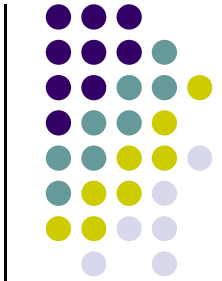


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						

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



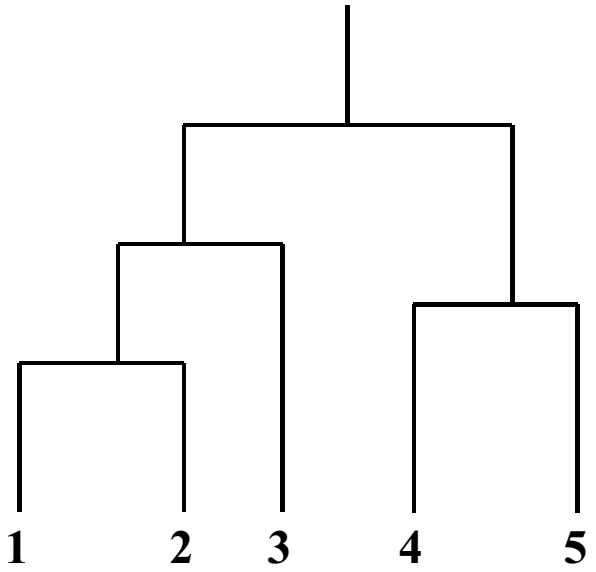
- Κοντινότητα δύο συστάδων είναι η μέση τιμή της ανα-δύο κοντινότητας (average of pairwise proximity) μεταξύ των σημείων των δύο συστάδων.

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

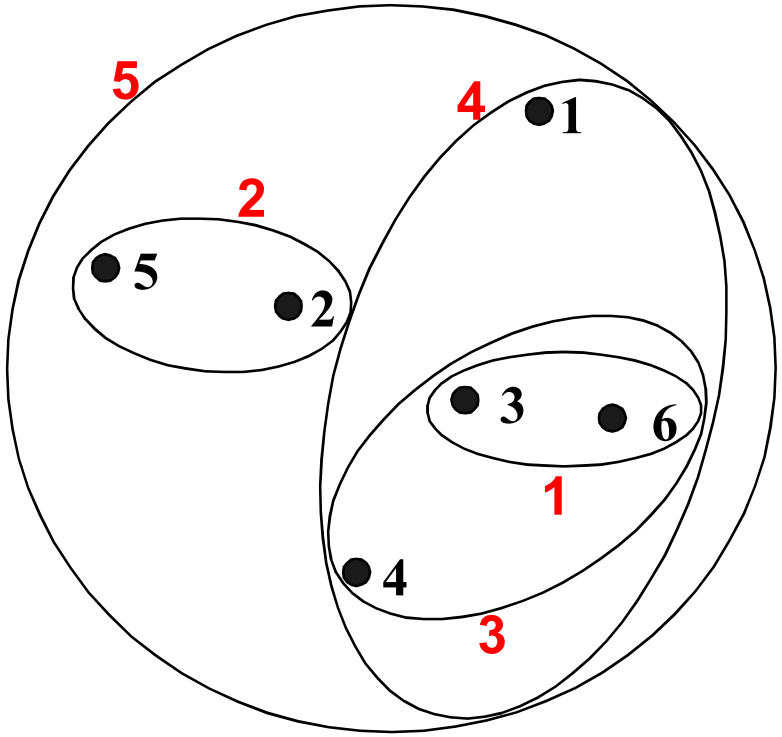
- Χρήση μέσης γιατί η ολική θα έδινε προτίμηση στις μεγάλες συστάδες

ομοιότητα

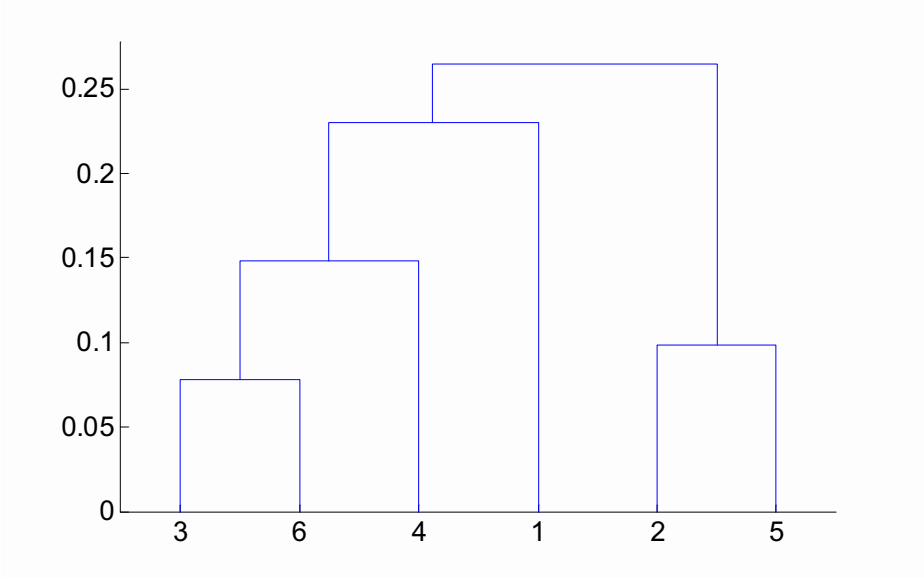
	I1	I2	I3	I4	I5
I1	1.00	0.90	0.10	0.65	0.20
I2	0.90	1.00	0.70	0.60	0.50
I3	0.10	0.70	1.00	0.40	0.30
I4	0.65	0.60	0.40	1.00	0.80
I5	0.20	0.50	0.30	0.80	1.00



ΣΙΣ: Ορισμός απόστασης μεταξύ συστάδων: Μέσο Ομάδας

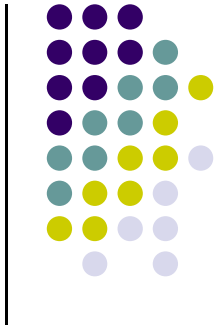


Φωλιασμένες Συστάδες



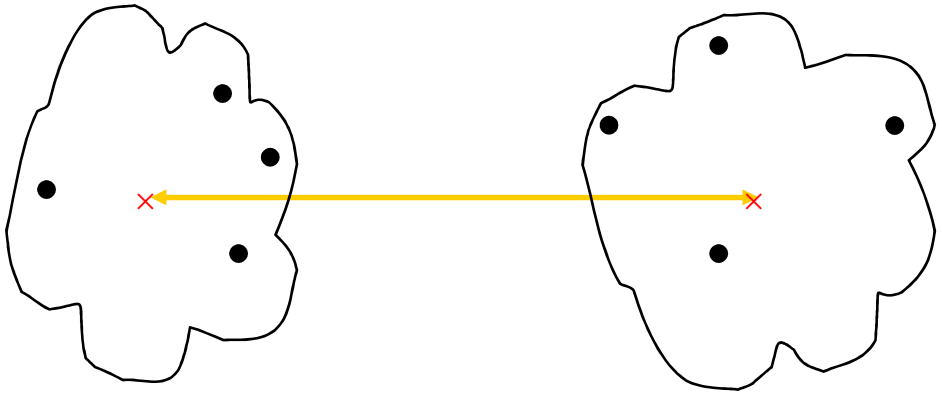
Dendrogram

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						
.						

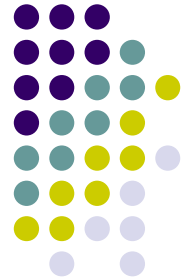
· **Πίνακας Γειτνίασης**

.

- 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

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: Μέθοδος του Ward

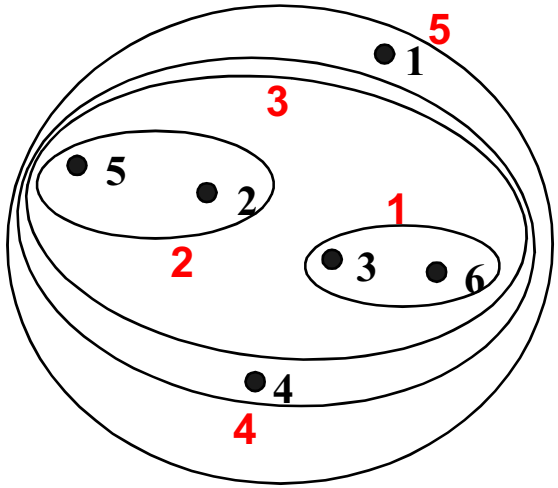
- Βασισμένο στην αύξηση του SSE όταν συγχωνεύονται οι δύο συστάδες
- Ιεραρχικό ανάλογο του k-means
- Μπορεί να χρησιμοποιηθεί για την αρχικοποίηση του k-means

$$D_w(C_i, C_j) = \sum_{x \in C_i} (x - r_i)^2 + \sum_{x \in C_j} (x - r_j)^2 - \sum_{x \in C_{ij}} (x - r_{ij})^2$$

Ward's distance μεταξύ συστάδων C_i and C_j **είναι η διαφορά μεταξύ της ολικού λάθους των δύο συστάδων και του ολκού λάθους αν ενώσουμε τις 2 συστάδες σε μία συστάδα έστω C_{ij}**

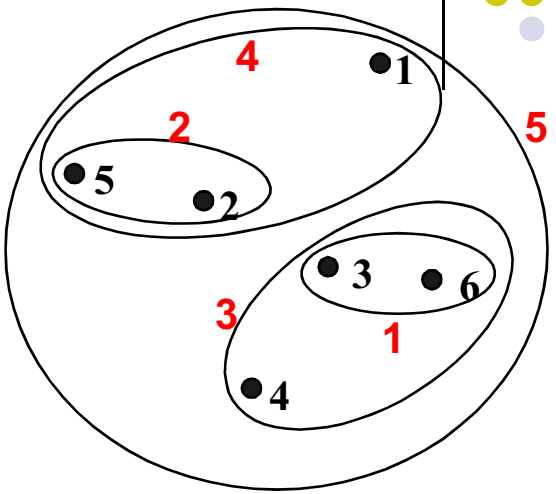
r_i: centroid of C_i
r_j: centroid of C_j
r_{ij}: centroid of C_{ij}

ΣΙΣ: Ορισμός απόστασης μεταξύ συστάδων: Σύγκριση

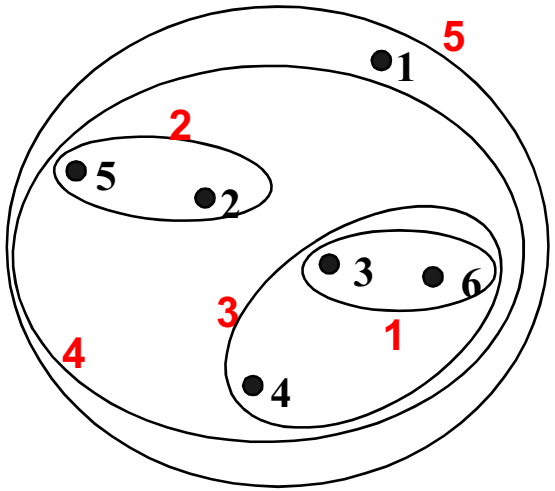


MIN

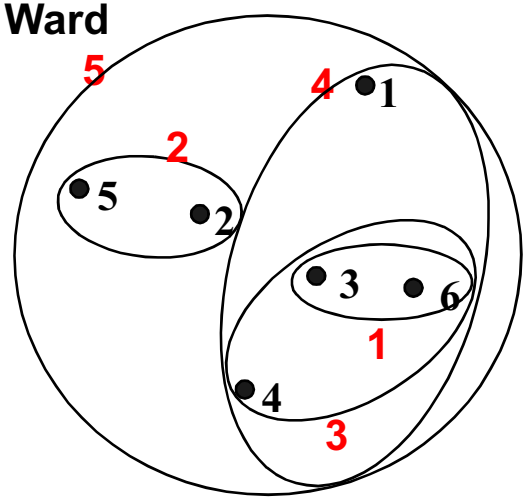
MAX



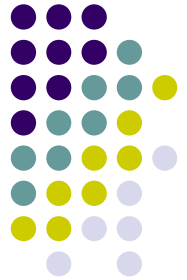
Μέθοδος του Ward



Μέσο Ομάδας



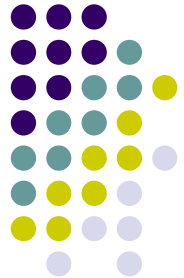
ΣΙΣ: Πολυπλοκότητα Χρόνου και Χώρου



- $O(m^2)$ χώρος για την αποθήκευση του πίνακα γειτνίασης
  - m αριθμός σημείων.

- $O(m^3)$
  - Ξεκινάμε με m συστάδες και μειώνουμε 1 τη φορά
  - Αν γραμμική αναζήτηση του πίνακα $O(m^2)$
  - Καλύτερος χρόνος αν διατηρούμε κάποια ταξινόμηση των αποστάσεων πχ heap

ΣΙΣ: Περιορισμοί και Προβλήματα



Οι αποφάσεις είναι τελικές – αφού δυο συστάδες συγχωνευτούν αυτό δεν μπορεί να αλλάξει

Δεν ελαχιστοποιούν άμεσα κάποια αντικειμενική συνάρτηση



Μια διαιρετική παραλλαγή του MIN βασίζεται σε spanning tree (σκελετικά δέντρα)

1. Χρησιμοποίησε τον πίνακα απόστασης και κατασκεύασε ένα ελάχιστο σκελετικό δέντρο

2. Δημιούργησε μια νέα συστάδα «σπάζοντας» το δέντρο στην ακμή με τη μεγαλύτερη απόσταση (μικρότερη ομοιότητα)