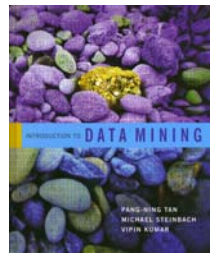
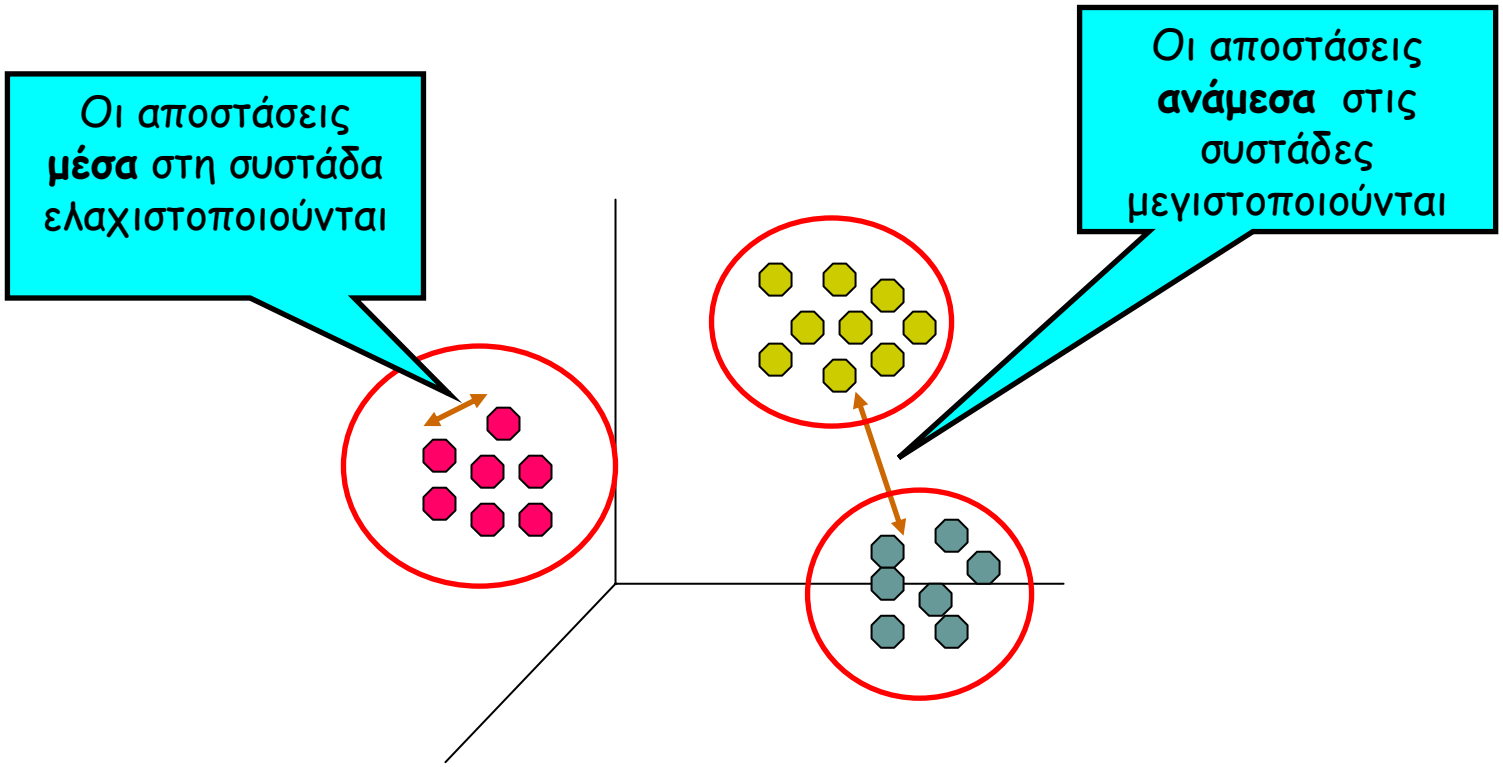
Συσταδοποίηση Ι

Οι διαφάνειες στηρίζονται στο **P.-N. Tan, M.Steinbach, V. Kumar, «Introduction to Data Mining»,** Addison Wesley, 2006



Τι είναι συσταδοποίηση

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



Περιεχόμενα

Είδη Συσταδοποίησης
3 Γνωστούς Αλγορίθμους
- k-means
- Ιεραρχική Συσταδοποίηση
- DBSCAN

Εφαρμογές

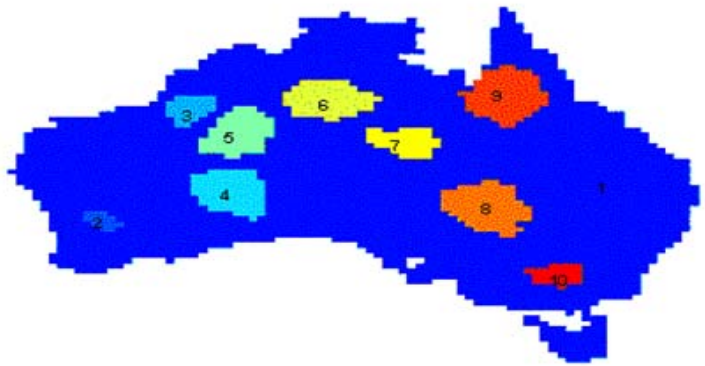
Κατανόηση

Ομαδοποίηση σχετιζόμενων αρχείων για browsing, ομαδοποίηση γονιδίων και πρωτεϊνών που έχουν την ίδια λειτουργία, ή μετοχών με παρόμοια διακύμανση τιμών

Περίληψη

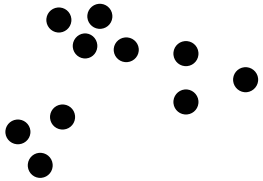
Ελάττωση του μεγέθους μεγάλων συνόλων

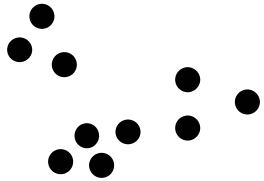
	Discovered Clusters	*Industry Group*
1	Applied-Matl-DOWN,Bay-Network-Down,3-COM-DOWN, Cabletron-Sys-DOWN,CISCO-DOWN,HP-DOWN, DSC-Comm-DOWN,INTEL-DOWN,LSI-Logic-DOWN, Micron-Tech-DOWN,Texas-Inst-Down,Tellabs-Inc-Down, Natl-Semiconduct-DOWN,Oracl-DOWN,SGI-DOWN, Sun-DOWN	Technology1-DOWN
2	Apple-Comp-DOWN,Autodesk-DOWN,DEC-DOWN, 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



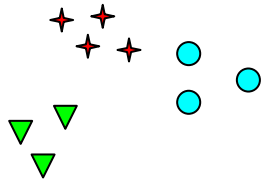
Clustering precipitation in Australia

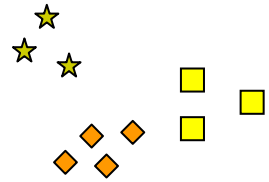
Ασάφεια



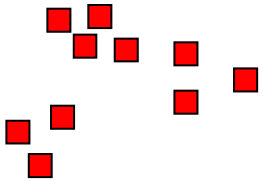


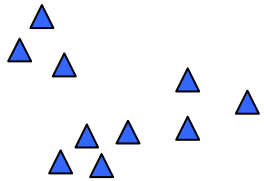
Πόσες Ομάδες?



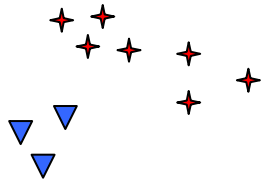


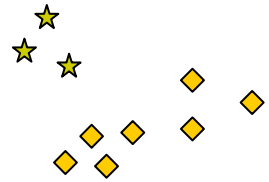
6 ομάδες





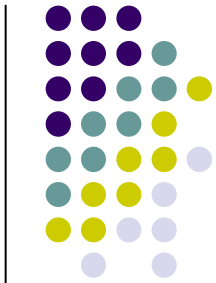
2 ομάδες





4 ομάδες

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)

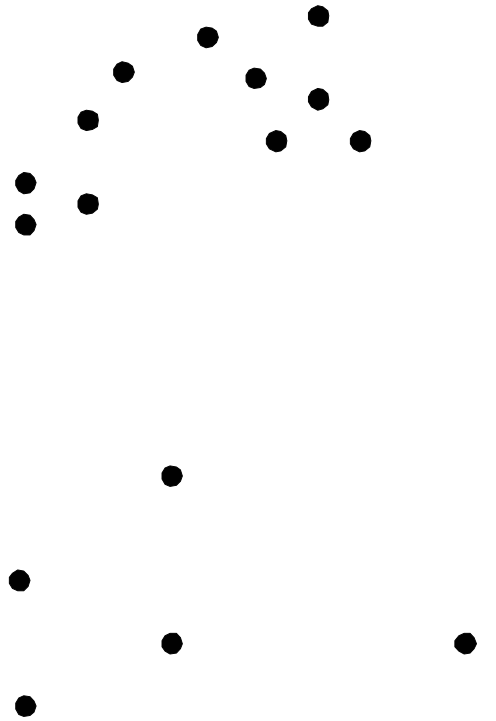
Ένας διαμερισμός των αντικειμένων σε μη επικαλυπτόμενα - non-overlapping - υποσύνολα (συστάδες) τέτοιος ώστε κάθε αντικείμενο ανήκει σε ακριβώς ένα υποσύνολο

Ιεραρχική Συσταδοποίηση (Hierarchical clustering

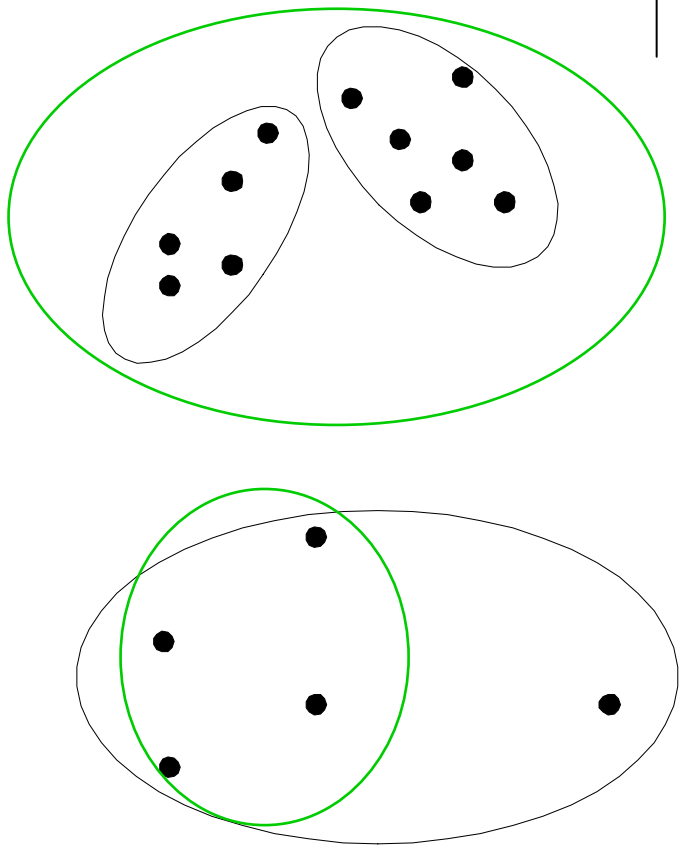
Ένα σύνολο από *εμφωλευμένες* (nested) ομάδες

Επιτρέπουμε σε μια συστάδα να έχει υποσυστάδες οργανωμένες σε ένα ιεραρχικό δέντρο

Διαχωριστική και Ιεραρχική Συσταδοποίηση



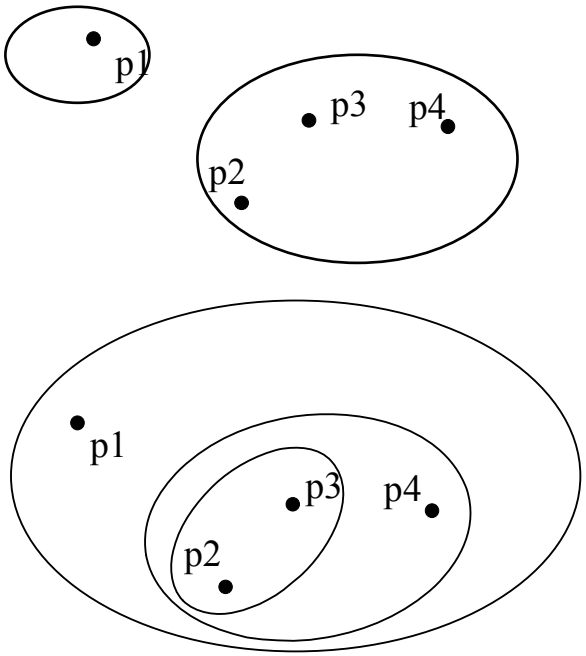
Αρχικά Σημεία



Σημείωση: Θόρυβος - outlier

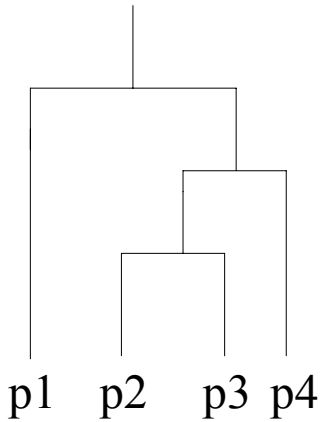
Outlier (ακραίο σημείο) τιμές που είναι εξαιρέσεις ως προς τα συνηθισμένες ή αναμενόμενες τιμές

Διαχωριστική και Ιεραρχική Συσταδοποίηση



Ιεραρχική Συσταδοποίηση

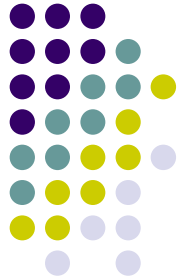
Διαχωριστική Συσταδοποίηση



Παραδοσιακό Δένδρο-γράμμα (Dendrogram)

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

Άλλες διακρίσεις μεταξύ συνόλων συστάδων



Επικαλυπτόμενο ή όχι

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

Μερική - Πλήρης

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

Ετερογενή - Ομογενή

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

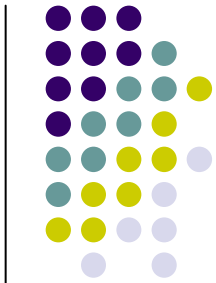
Ασαφή συσταδοποίηση

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

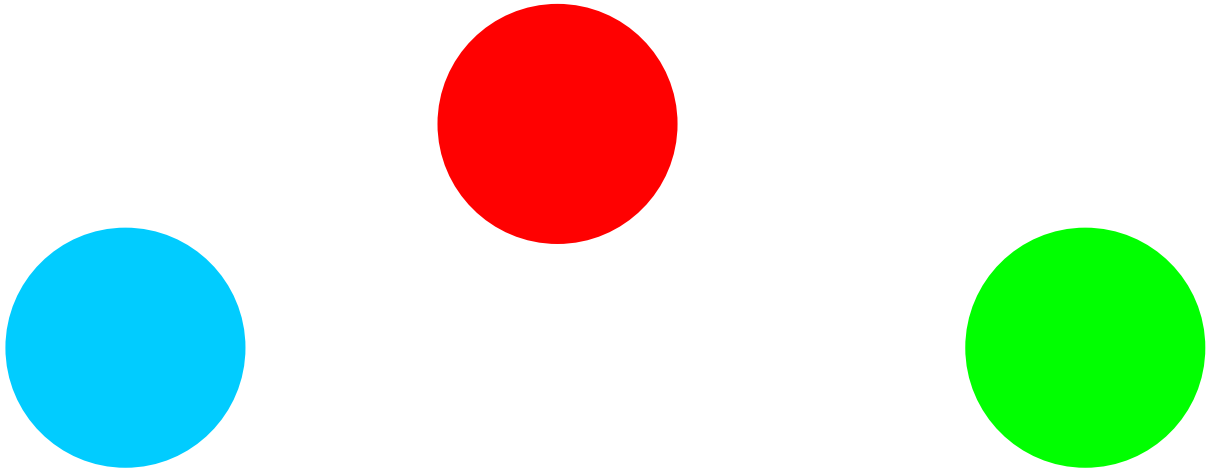
Είδη Συστάδων



- 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)

Καλώς Διαχωρισμένες Συστάδες

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



3 καλώς-διαχωρισμένες συστάδες

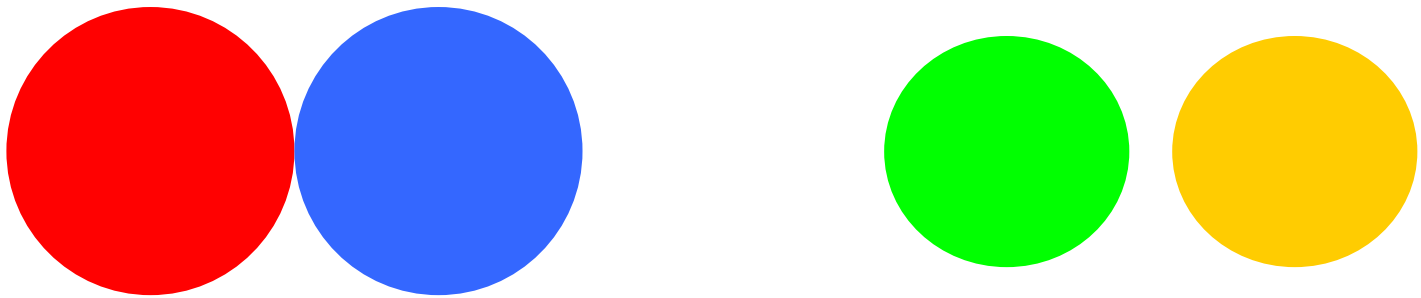
Συχνά υπάρχει η έννοια του κατωφλιού (threshold)

Όχι απαραίτητα κυκλικοί (οποιοδήποτε σχήμα)

Συστάδες βασισμένες σε κέντρο ή πρότυπο

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

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



4 συστάδες βασισμένες σε κέντρο

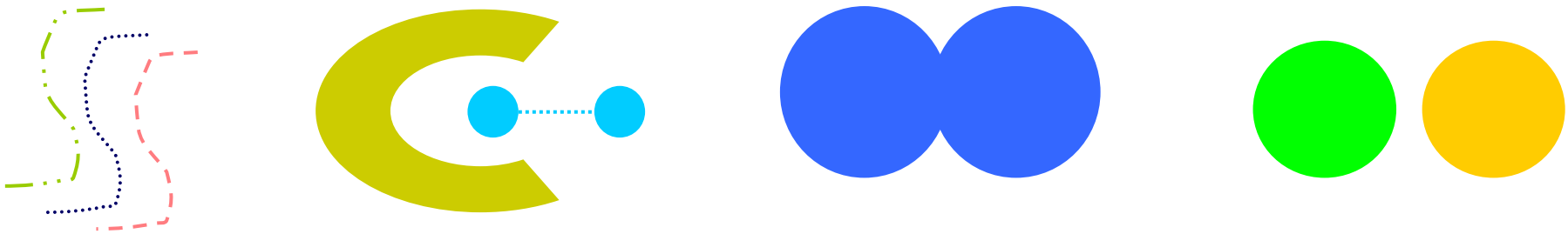
Συνεχής Συστάδες

Συνεχής Συστάδες (Contiguous Cluster) (Κοντινότερος γείτονα ή μεταβατικά)

Μια συστάδα είναι ένα σύνολο σημείων τέτοιο ώστε κάθε σημείο είναι ποιο κοντά σε ένα ή περισσότερα σημεία της συστάδας από ότι σε οποιοδήποτε σημείο εκτός ομάδας

Συχνά σε περιπτώσεις συστάδων με μη κανονικό σχήμα ή με αλληλοπλεκόμενα σχήματα

Πρόβλημα με θόρυβο

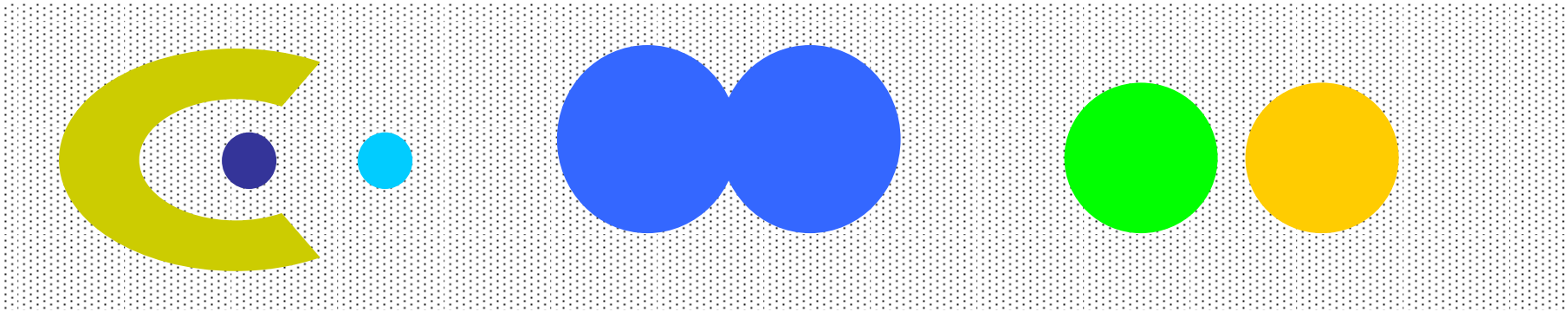


8 contiguous clusters

Συστάδες βασισμένες στην πυκνότητα

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

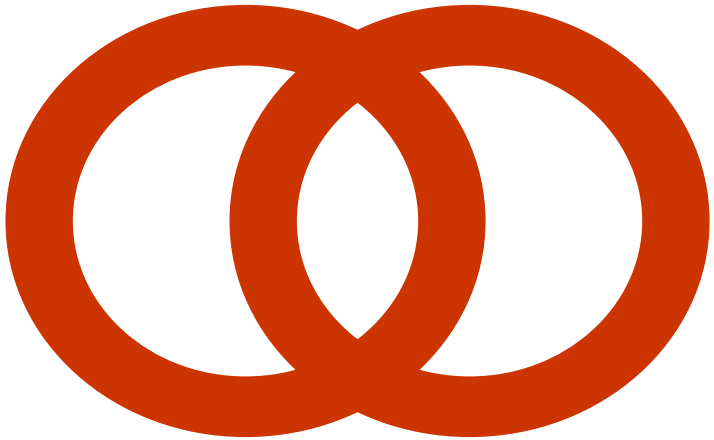
Συχνά σε περιπτώσεις συστάδων με μη κανονικό σχήμα ή με αλληλοπλεκόμενα σχήματα ή όταν θόρυβος ή outliers



6 density-based clusters

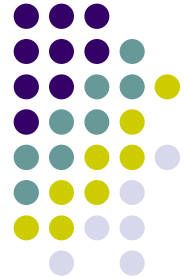
Εννοιολογική συσταδοποίηση

Συστάδες με κοινή ιδιότητα ή εννοιολογικές συστάδες.



2 Overlapping Circles

Συστάδες βασισμένες σε μια Αντικειμενική Συνάρτηση

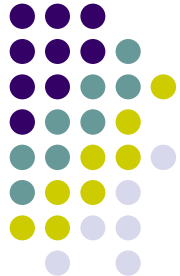


Εύρεση συστάδων που ελαχιστοποιούν ή μεγιστοποιούν μια αντικειμενική συνάρτηση

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)

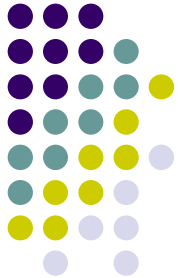
Οι στόχοι (objectives) μπορεί να είναι ολικοί (global) ή τοπικοί (local)
Οι ιεραρχικοί συνήθως τοπικού
Οι διαχωριστικοί ολικές

Χαρακτηριστικά των Δεδομένων Εισόδου



- Πυκνότητα
- Είδος γνωρισμάτων
  - Καθορίζει τον τύπο της ομοιότητας
- Είδος δεδομένων
  - Καθορίζει τον τύπο της ομοιότητας
  - Άλλα χαρακτηριστικά, όπως η αυτοσυσχέτιση (autocorrelation)
- Διάσταση
- Θόρυβος και Outliers
- Είδος κατανομών

Κριτήρια Ομοιότητας -Απόσταση



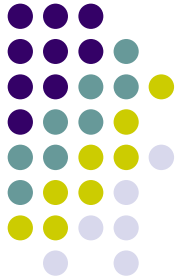
Τι είναι τα δεδομένα και η απόσταση τους;

Interval-scaled (συνεχείς τιμές σε κάποιο διάστημα)
Πχ ύψος και βάρος, θερμοκρασία, συντεταγμένες, κλπ

Η μονάδα μέτρησης επηρεάζει => standardization
Ίδιο βάρος σε κάθε μεταβλητή (αλλά εξαρτάται και από την εφαρμογή, πχ ύψος πιο σημαντικό στην καλαθοσφαίριση)

Mean absolute deviation (μέση απόλυτη απόκλιση) για τον υπολογισμό του z-score

Κριτήρια Ομοιότητας -Απόσταση

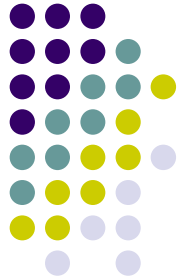


Interval Scaled (συνέχεια)

Ευκλείδεια
City-block ή Manhattan
Minkowski

Εκδοχές τους με βάρη

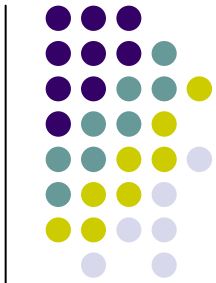
Κριτήρια Ομοιότητας



Συναρτήσεις απόστασης (distance functions)

$d(i, j) \geq 0$
$d(i, i) = 0$ (ανακλαστική)
$d(i, j) = d(j, i)$ (συμμετρική)
$d(i, j) \leq d(i, h) + d(h, j)$ (τριγωνική ανισότητα)

Κριτήρια Ομοιότητας



Δυαδικές μεταβλητές

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

Μη συμμετρικές (η συμφωνία στο 1 πιο σημαντική)
Non-invariant (Jaccard)

Κριτήρια Ομοιότητας

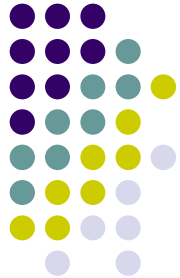
Κατηγορικές μεταβλητές

Πχ μετράμε απλώς τις κοινές τιμές

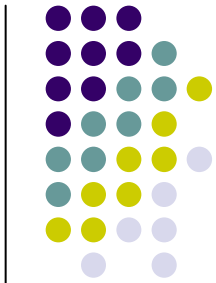
Ordinal

Μεικτές τιμές

Αλγόριθμοι Συσταδοποίησης

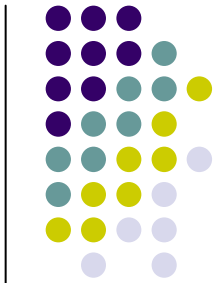


- K-means και παραλλαγές
- Ιεραρχική Συσταδοποίηση
- Συσταδοποίηση με βάση την Πυκνότητα



K-means

K-means: Γενικά



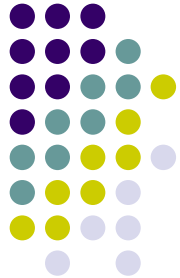
Διαχωριστικός αλγόριθμος

Κάθε συστάδα συσχετίζεται με ένα κεντρικό σημείο (centroid)

Κάθε σημείο ανατίθεται στη συστάδα με το κοντινότερο κεντρικό σημείο

Καθορίζεται ο αριθμός των ομάδων, **K**

K-means: Βασικός Αλγόριθμος



Βασικός αλγόριθμος

1: Επιλογή K σημείων ως τα αρχικά κεντρικά σημεία

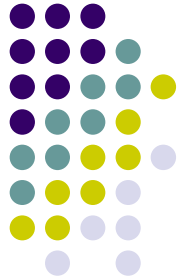
2: **Repeat**

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

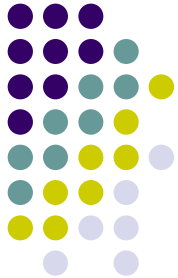
K-means: Βασικός Αλγόριθμος



Παρατηρήσεις

- Τα αρχικά κεντρικά σημεία συνήθως επιλέγονται τυχαία
- Οι συστάδες που παράγονται διαφέρουν από το ένα τρέξιμο του αλγορίθμου στο άλλο
- Το κεντρικό σημείο είναι (συνήθως) το μέσο (mean) των σημείων της συστάδας

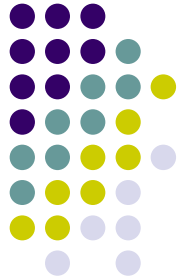
K-means: Βασικός Αλγόριθμος



Παρατηρήσεις (συνέχεια)

- Η εγγύτητα των σημείων υπολογίζεται με βάση την Ευκλείδεια απόσταση, την ομοιότητα συνημίτονου (cosine similarity), τη συσχέτιση correlation (κλπ)
- Επειδή υπολογίζεται συχνά πρέπει να είναι σχετικά *απλή*

K-means: Βασικός Αλγόριθμος



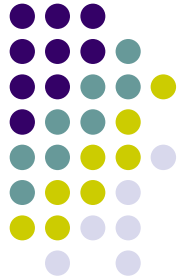
Παρατηρήσεις (συνέχεια)

- Για αυτά τα 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

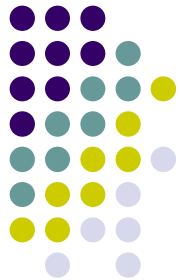
K-means: Βασικός Αλγόριθμος



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: Εκτίμηση ποιότητας



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

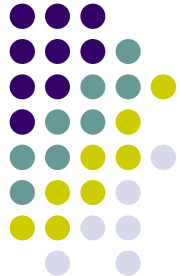
$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$

όπου 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 $c_i = 1/m_i \sum_{x \in C_i} x$

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

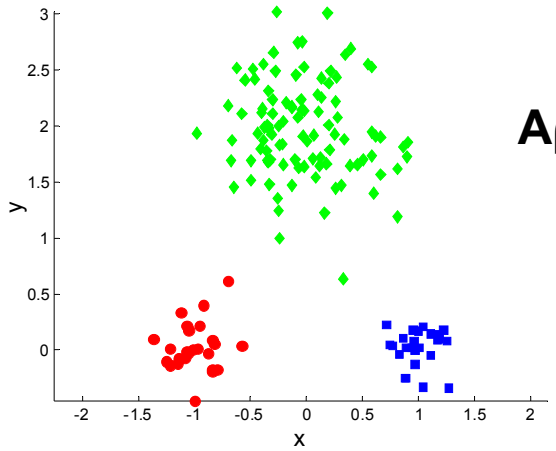
K-means: Εκτίμηση ποιότητας



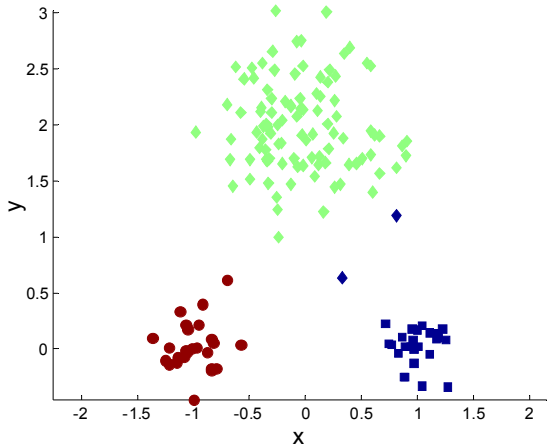
Ένας τρόπος να βελτιώσουμε τη συσταδοποίηση (ελάττωση του SSE) είναι να μεγαλώσουμε το K

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

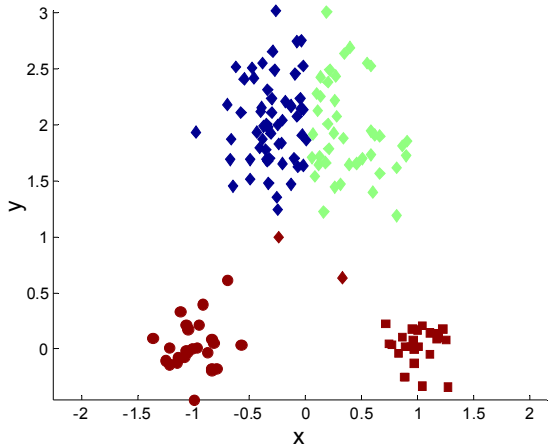
K-means: Παράδειγμα



Αρχικά σημεία

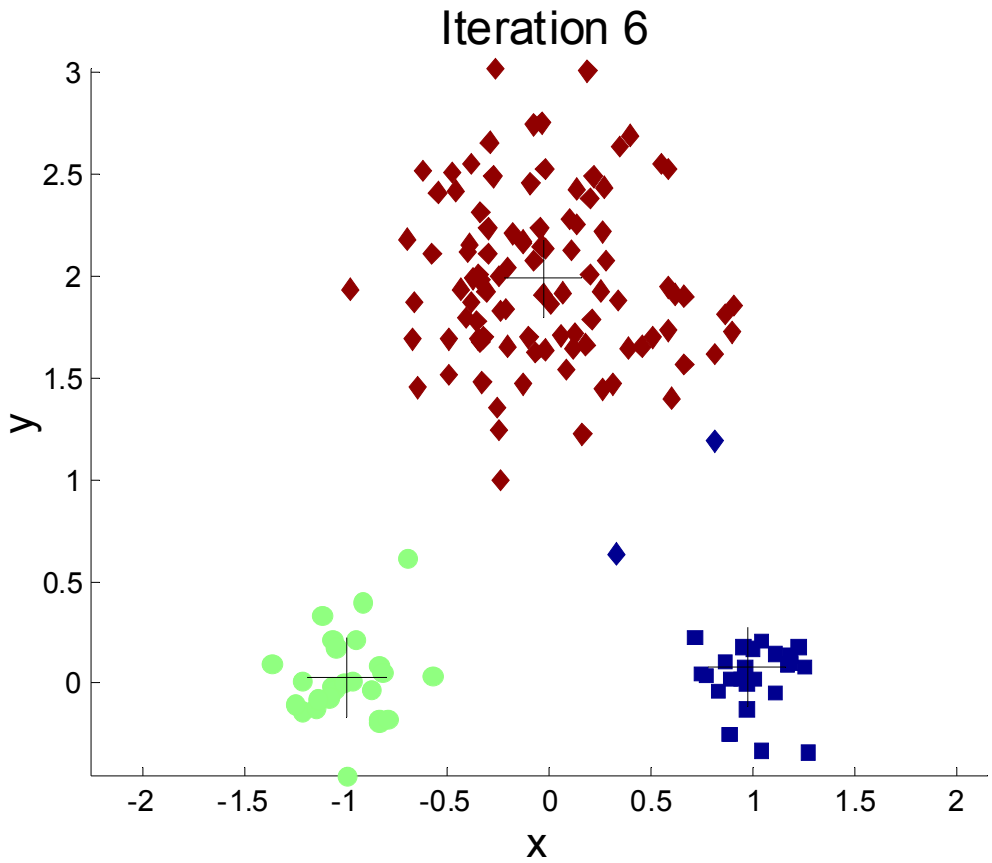


Βέλτιστη συσταδοποίηση

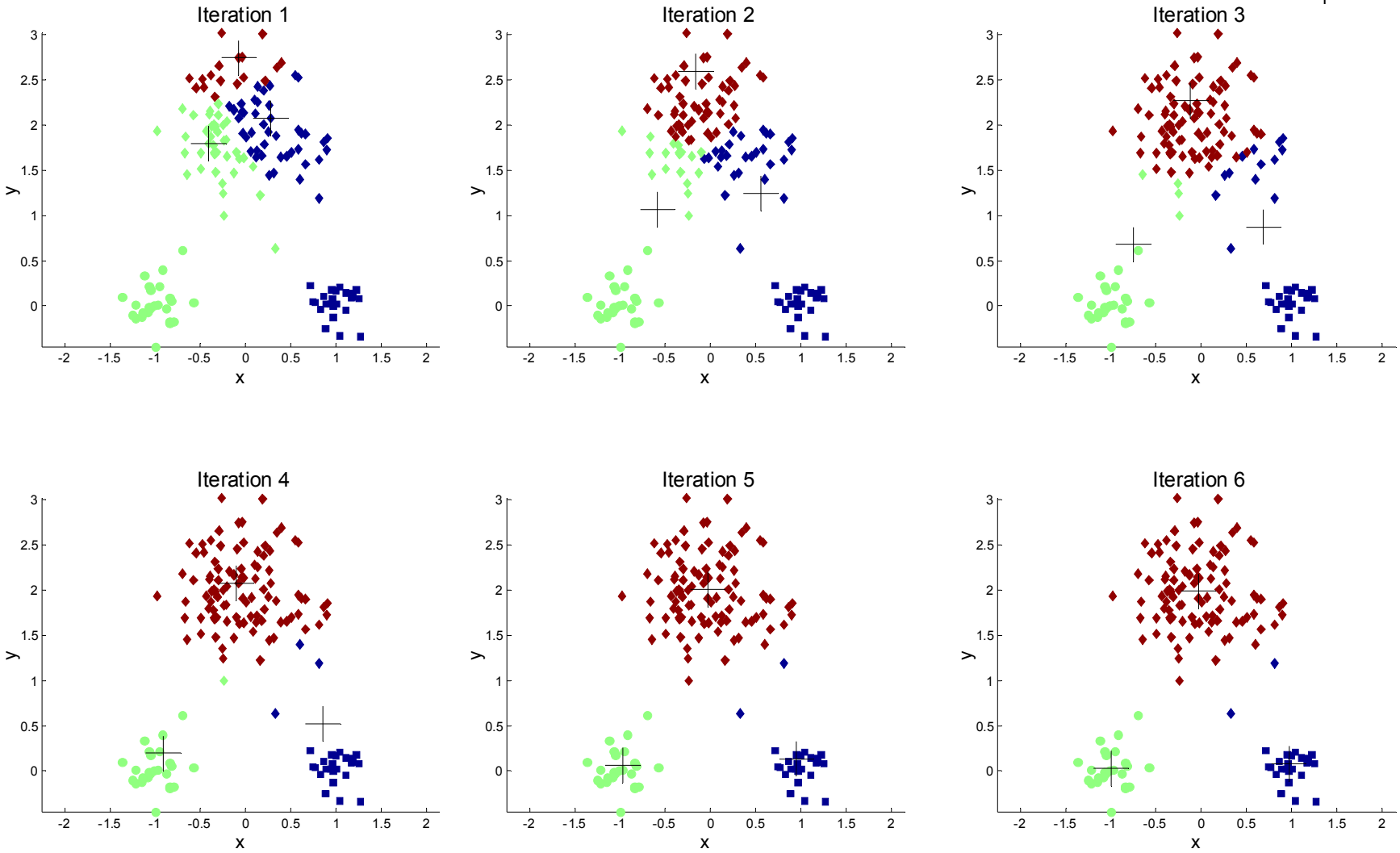


Υπό-βέλτιστη συσταδοποίση

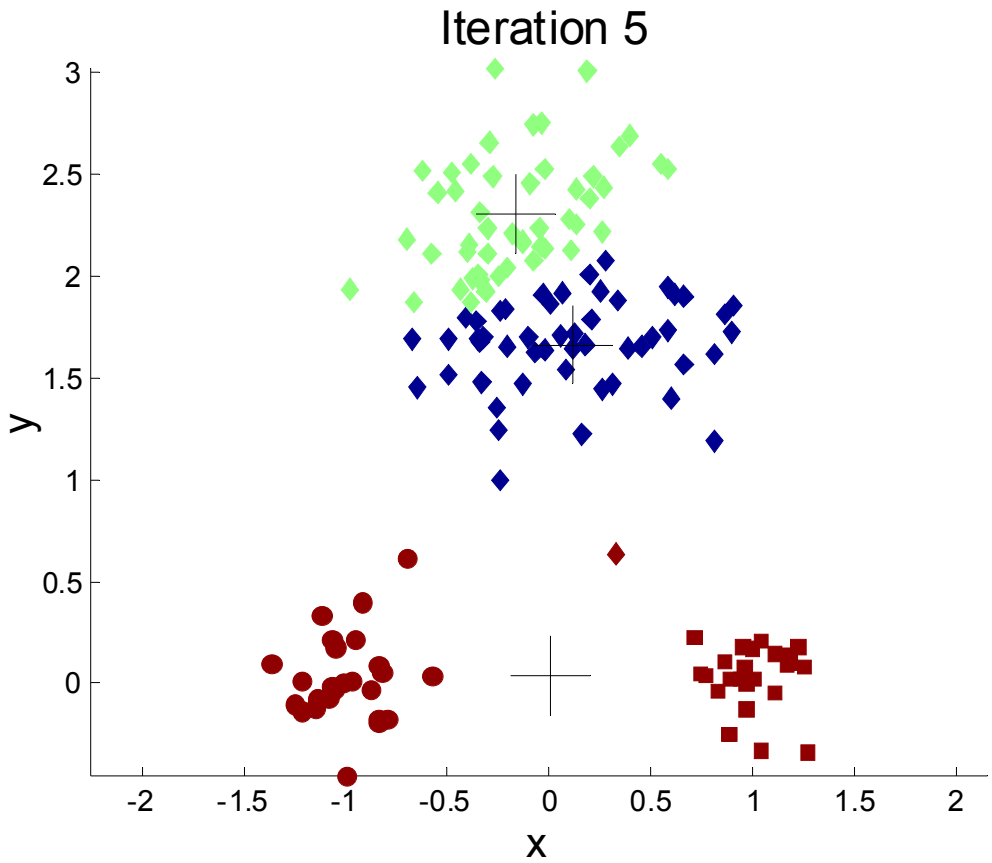
K-means: Επιλογή αρχικών σημείων



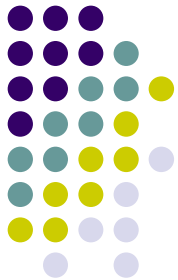
K-means: Επιλογή αρχικών σημείων

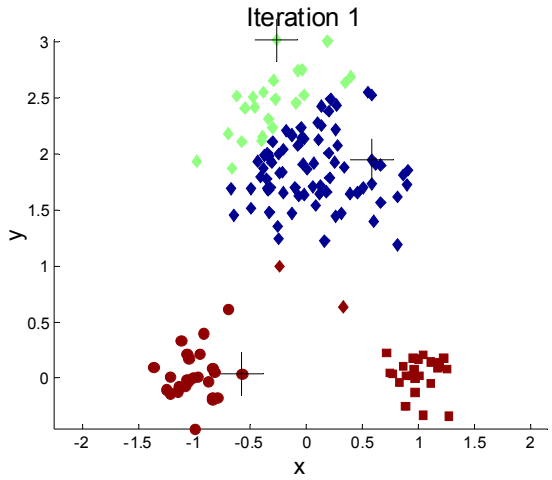


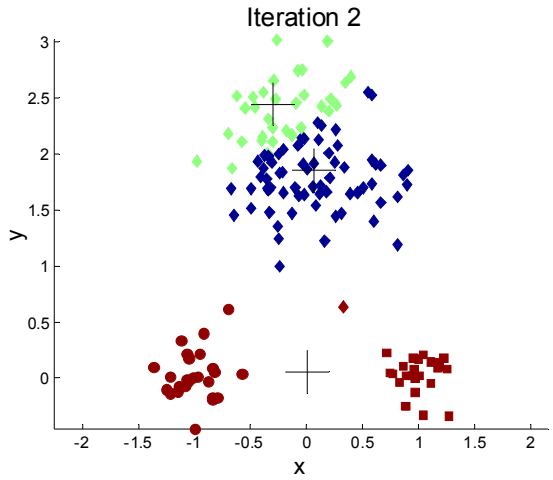
K-means: Επιλογή αρχικών σημείων

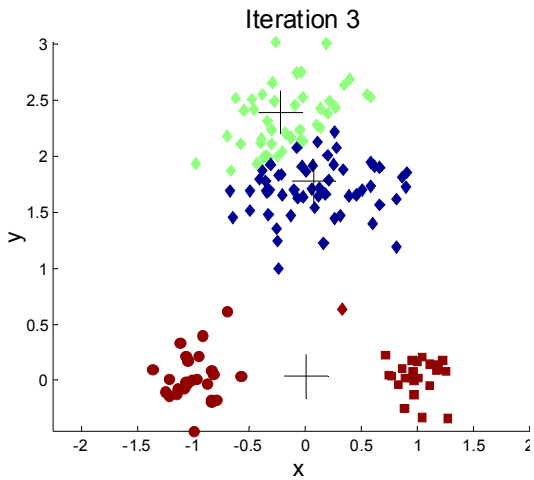


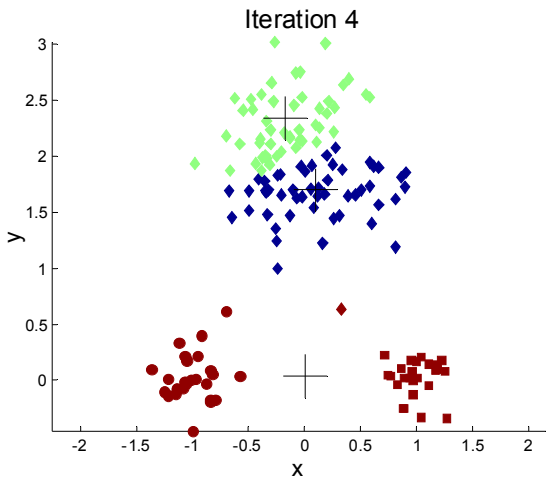
K-means: Επιλογή αρχικών σημείων

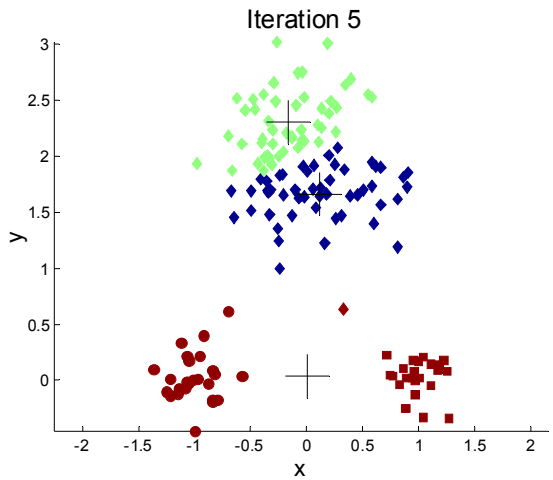




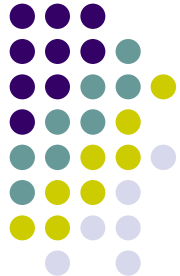








K-means: Επιλογή αρχικών σημείων



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

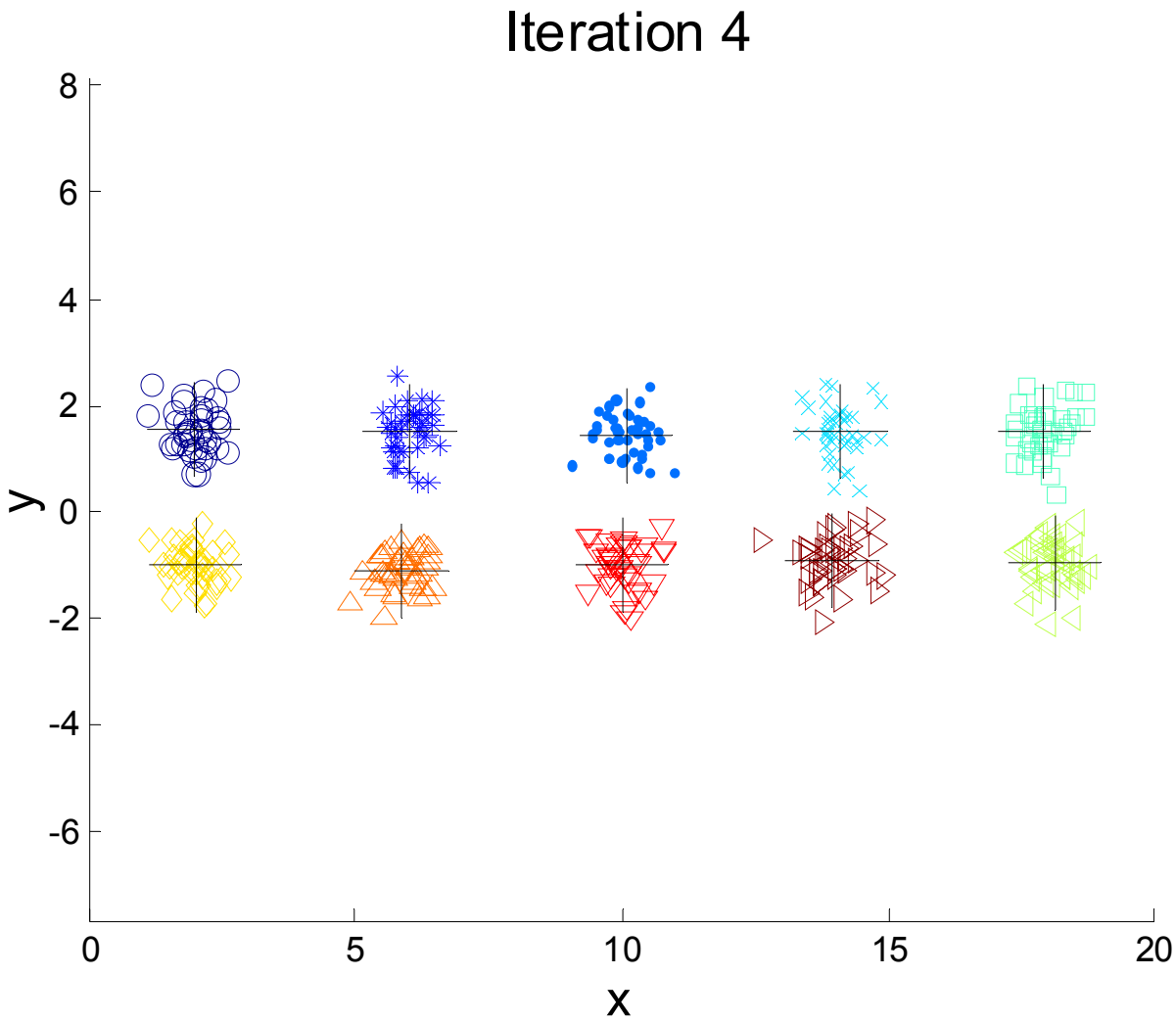
$$P = \frac{\text{number of ways to select one centroid from each cluster}}{\text{number of ways to select } K \text{ centroids}} = \frac{K! n^K}{(Kn)^K} = \frac{K!}{K^K}$$

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

Μερικές φορές τα αρχικά σημεία βελτιώνουν τη θέση τους και άλλες φορές όχι

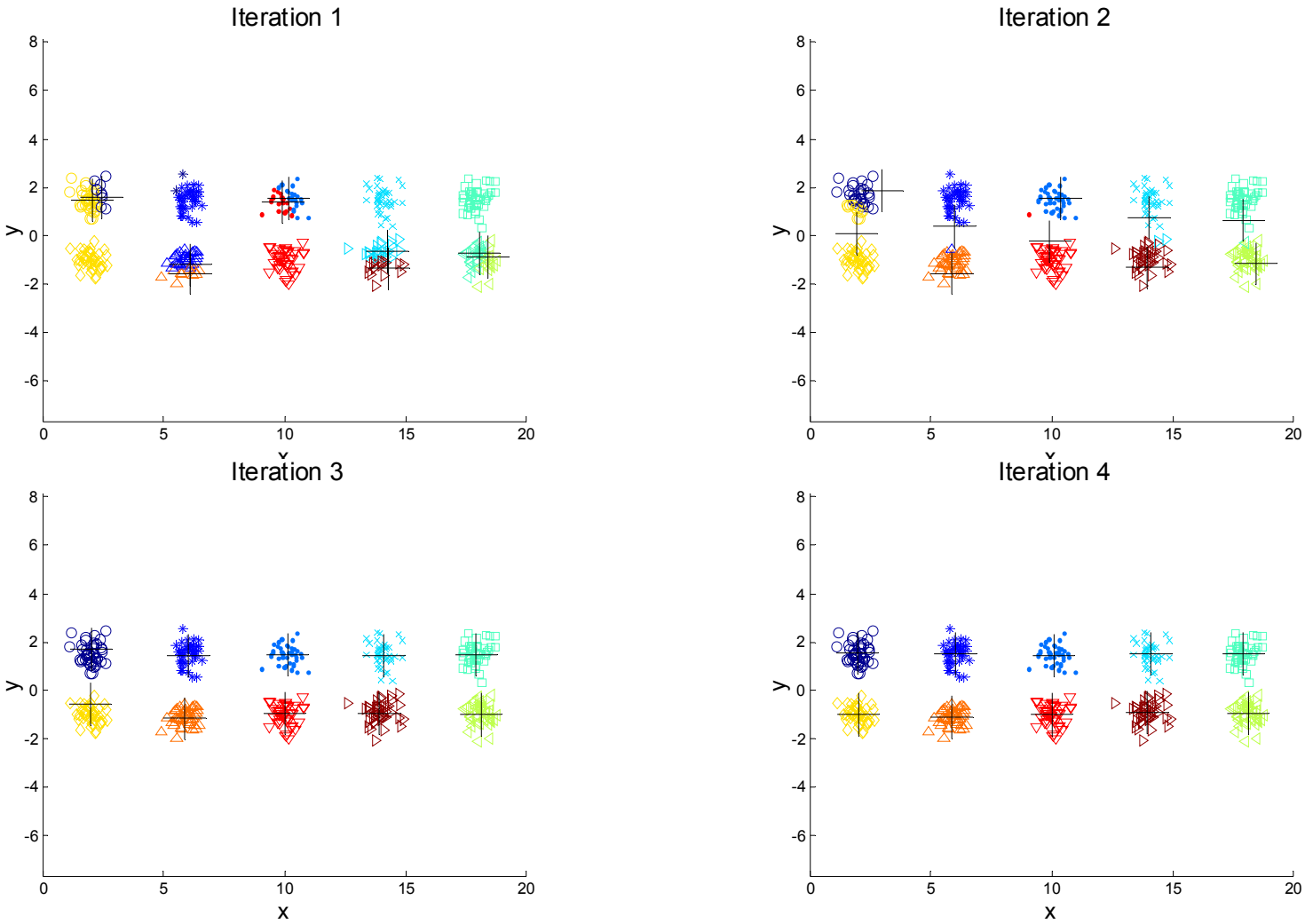
Θα δούμε ένα παράδειγμα με 5 ζευγάρια συστάδων

Παράδειγμα 10 συστάδων



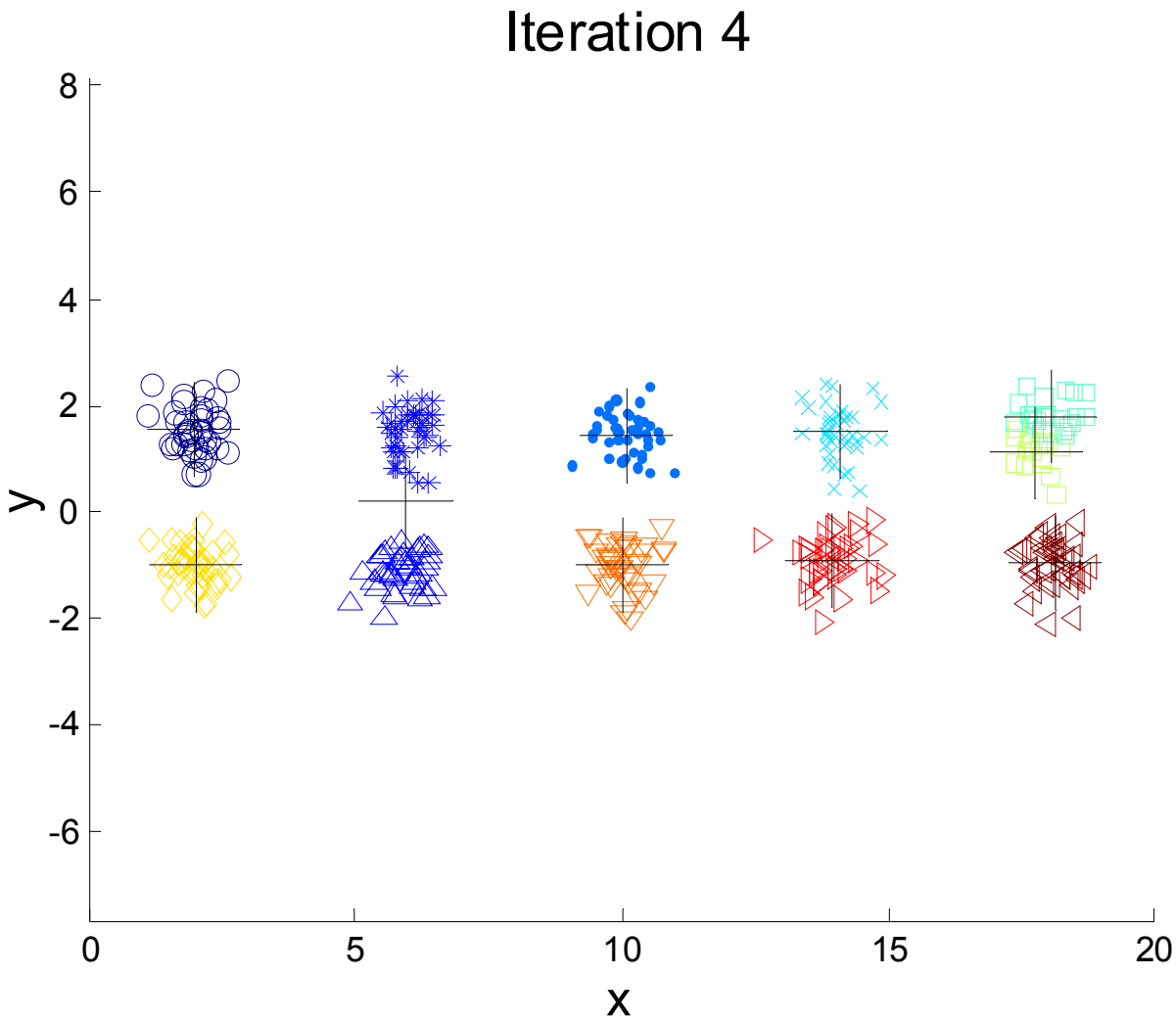
Ξεκινώντας με δύο αρχικά σημεία σε κάθε συστάδα κάθε ζεύγους συστάδων

Παράδειγμα 10 συστάδων



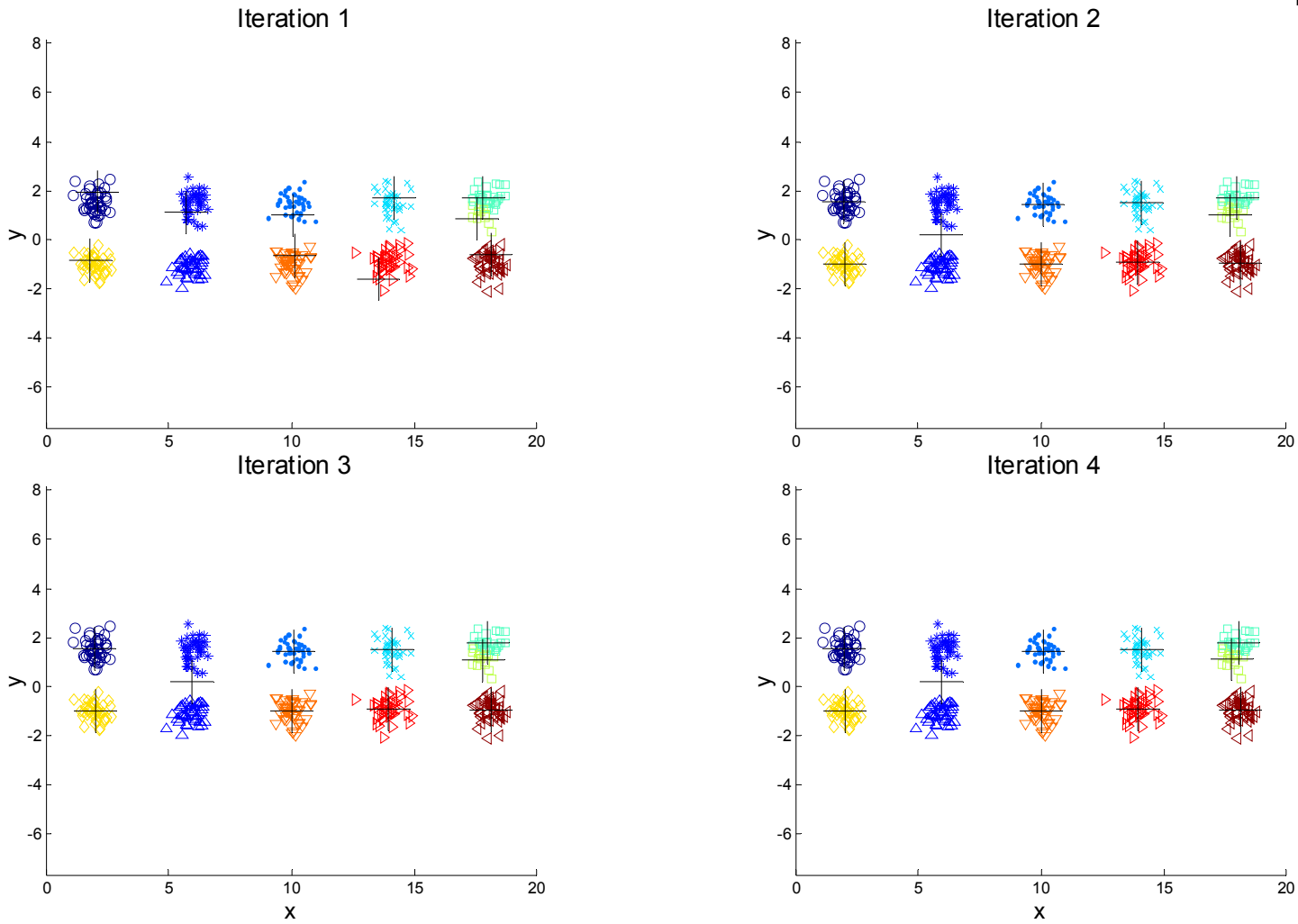
Ξεκινώντας με δύο αρχικά σημεία σε κάθε συστάδα κάθε ζεύγους συστάδων

Παράδειγμα 10 συστάδων



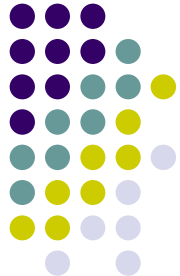
Ξεκινώντας με κάποια ζευγάρια συστάδων να έχουν τρία κεντρικά σημεία και άλλα μόνο ένα

Παράδειγμα 10 συστάδων



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

K-means: Λύσεις για την επιλογή αρχικών σημείων



Πολλαπλά τρεξίματα
- Βοηθά, αλλά πολλές περιπτώσεις

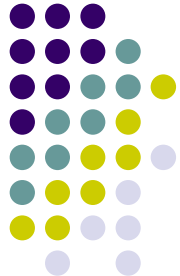
Δειγματοληψία και χρήση κάποιας ιεραρχικής τεχνικής

Επιλογή παραπάνω από k αρχικών σημείων και μετά επιλογή k από αυτά τα αρχικά κεντρικά σημεία

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

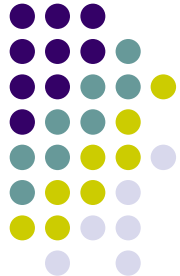
Πολλές στρατηγικές

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

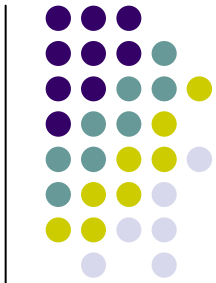


Στο βασικό K-means, το κέντρα ενημερώνεται αφού όλο τα σημεία έχουν ανατεθεί στο κέντρο

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

- Κάθε ανάθεση ενημερώνει 0 ή 2 κέντρα
- Ποιο 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

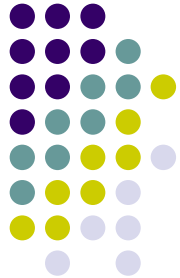
Post-processing

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

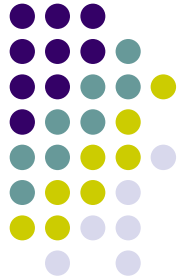
4: **for** i = 1 to number_of_trials **do**

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)

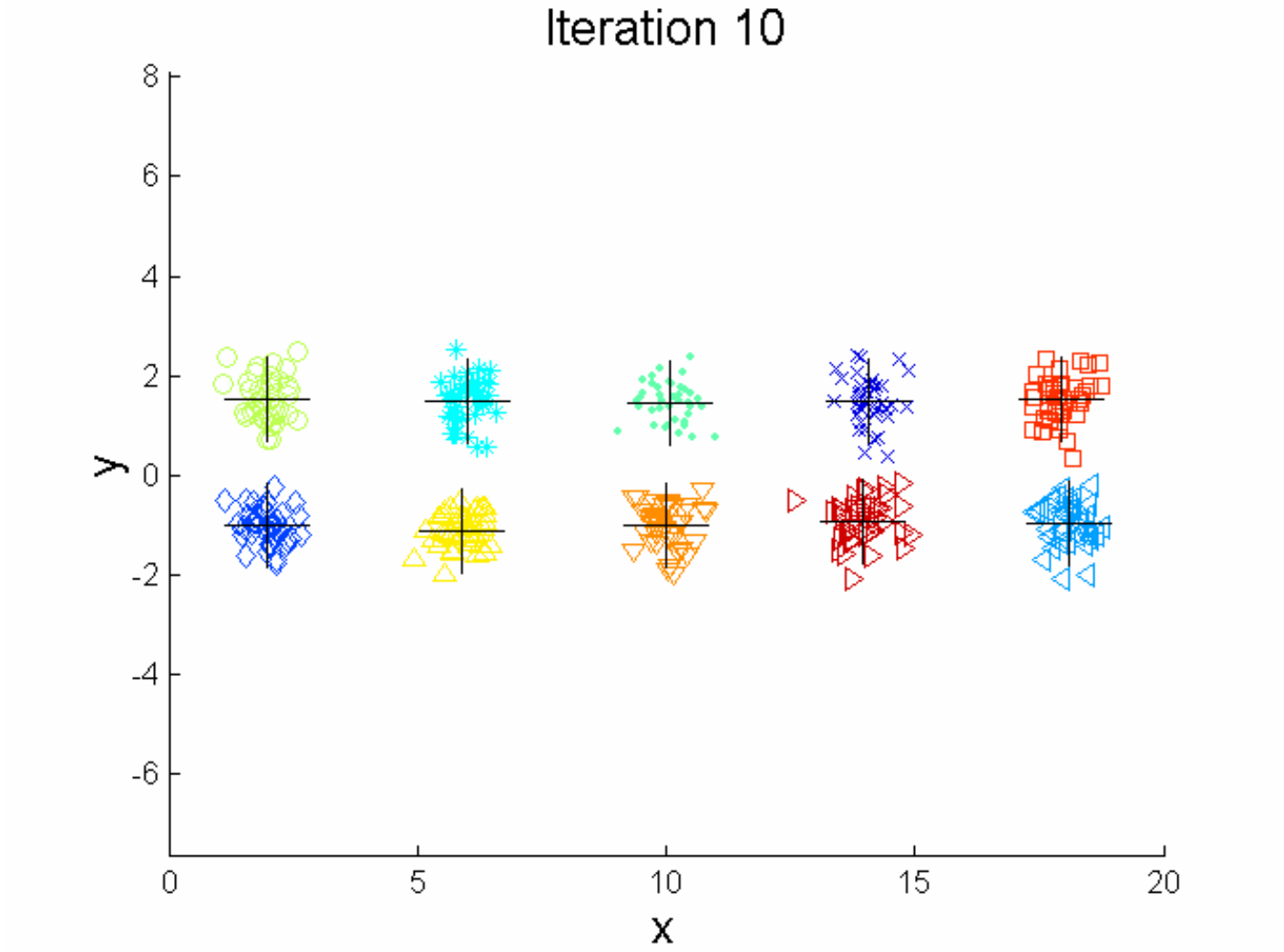


Ποια συστάδα να διασπάσουμε;

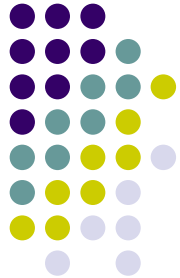
- Τη μεγαλύτερη
- Αυτή με το μεγαλύτερο SSE
- Συνδυασμό των παραπάνω

Μπορεί να χρησιμοποιηθεί και ως ιεραρχικός

K-means με διχοτόμηση



K-means: Περιορισμοί

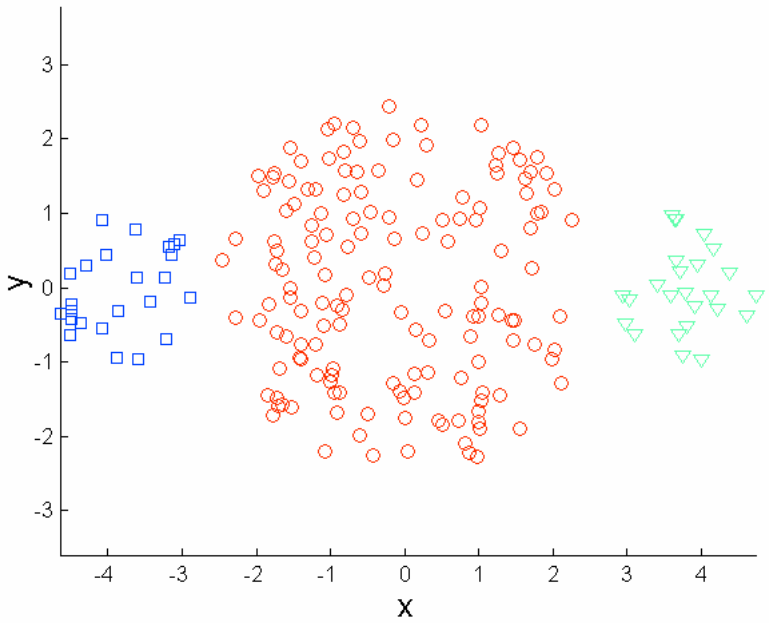


Ο K-means έχει προβλήματα όταν οι συστάδες έχουν διαφορετικά

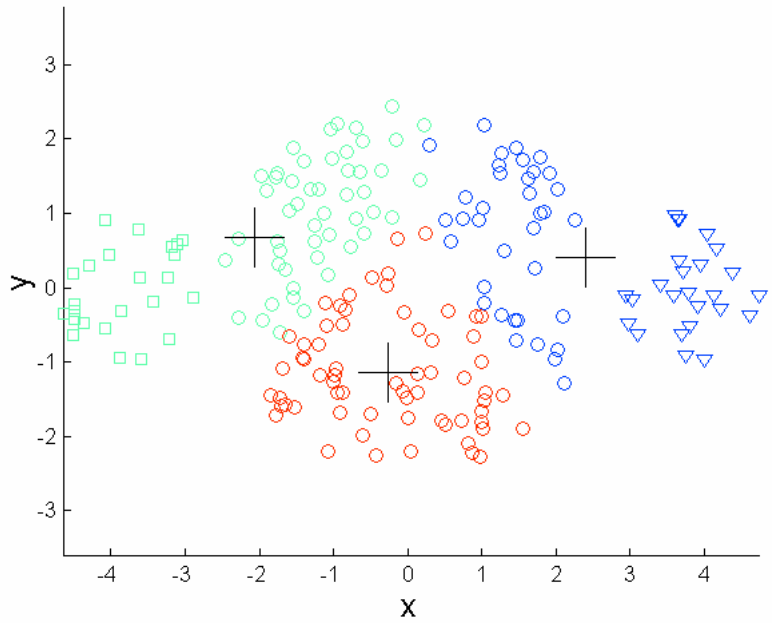
- Διαφορετικά Μεγέθη
- Διαφορετικές Πυκνότητες
- Non-globular shapes

Έχει προβλήματα όταν τα δεδομένα έχουν outliers

K-means: Περιορισμοί – διαφορετικά μεγέθη



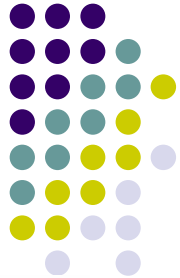
Αρχικά σημεία

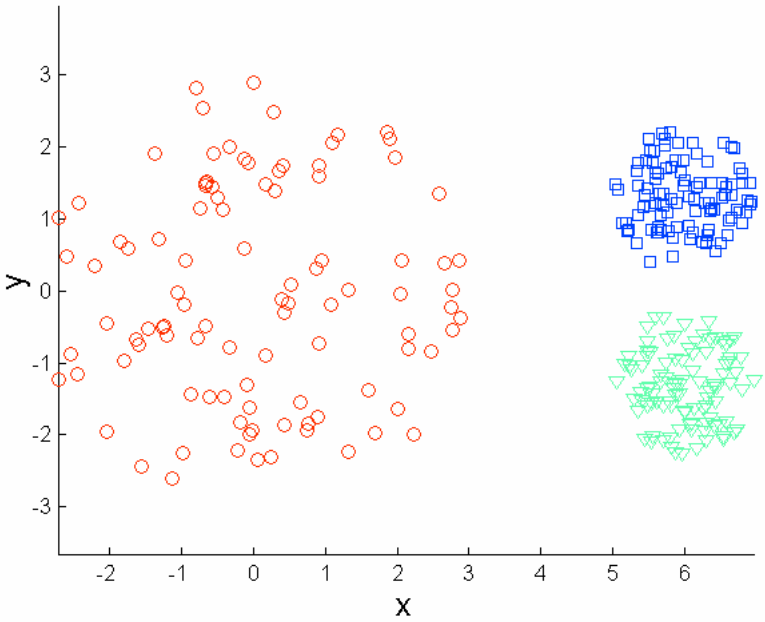


K-means (3 συστάδες)

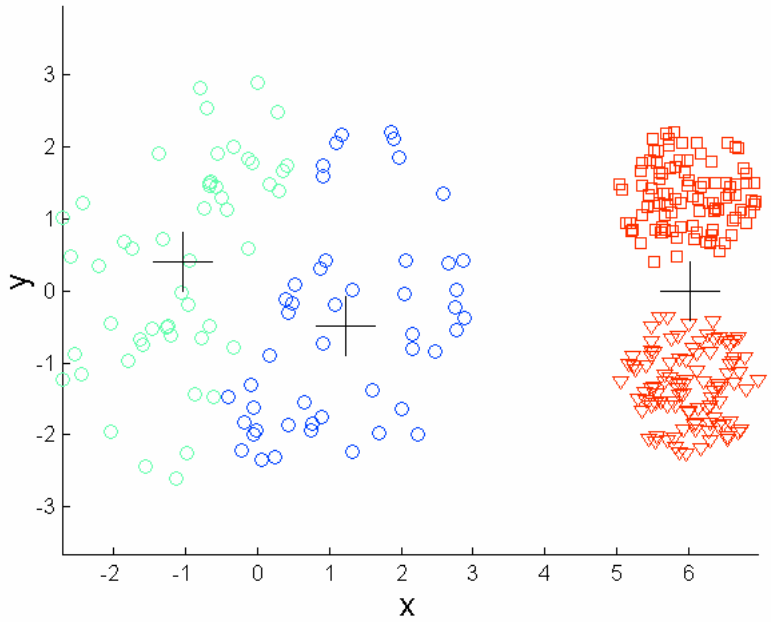
Δεν μπορεί να βρει το μεγάλο κόκκινο, γιατί είναι πολύ μεγαλύτερος από τους άλλους

K-means: Περιορισμοί – διαφορετικές πυκνότητες





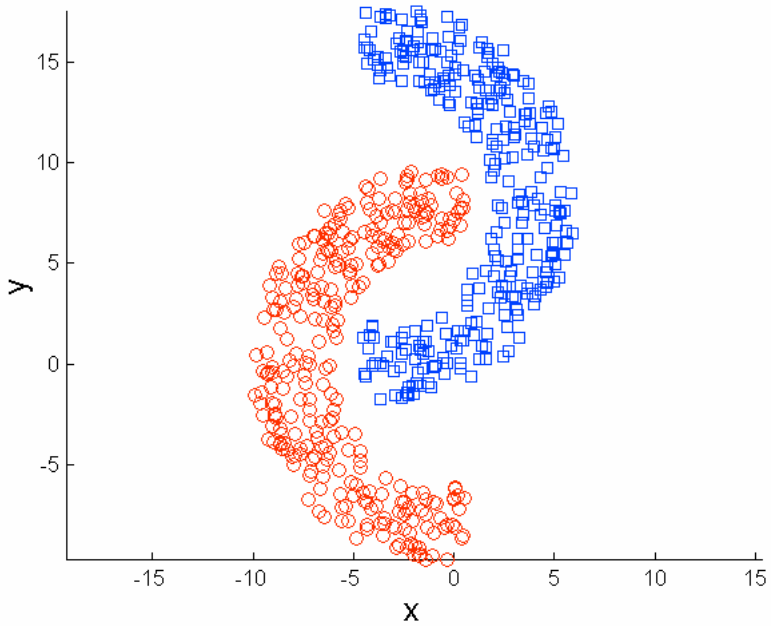
Αρχικά σημεία



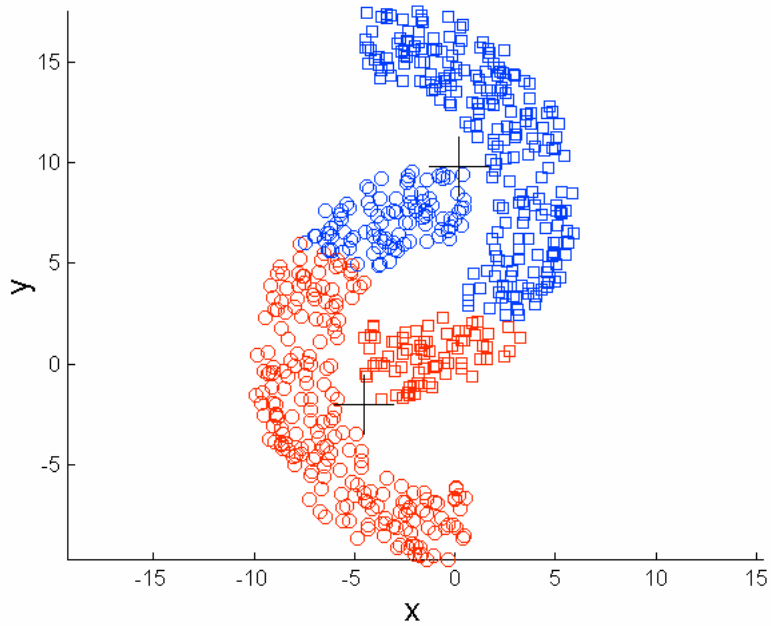
K-means (3 συστάδες)

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

K-means: Περιορισμοί – μη κυκλικά σχήματα



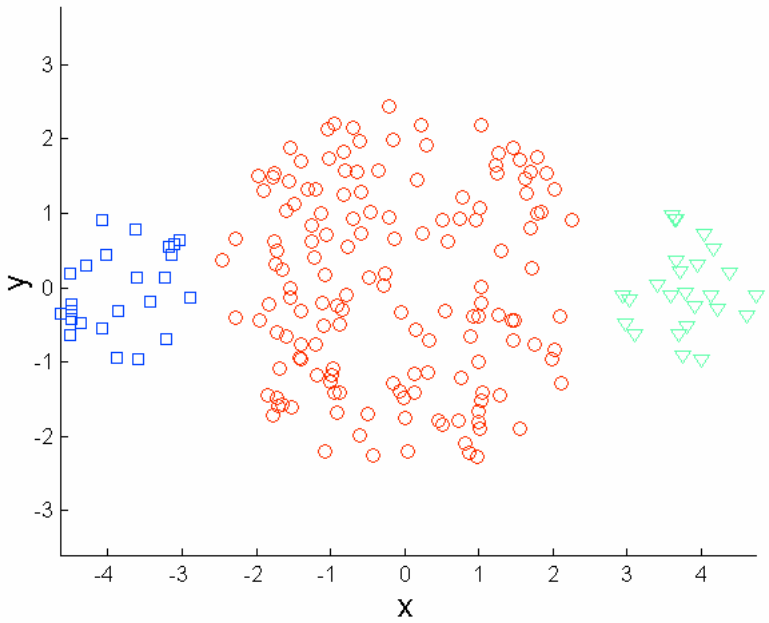
Αρχικά σημεία



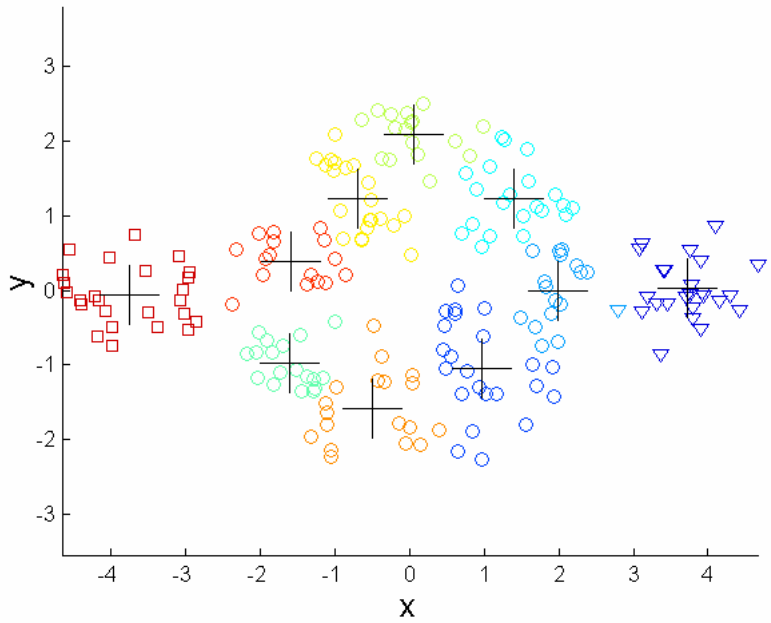
K-means (2 συστάδες)

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

K-means: Περιορισμοί



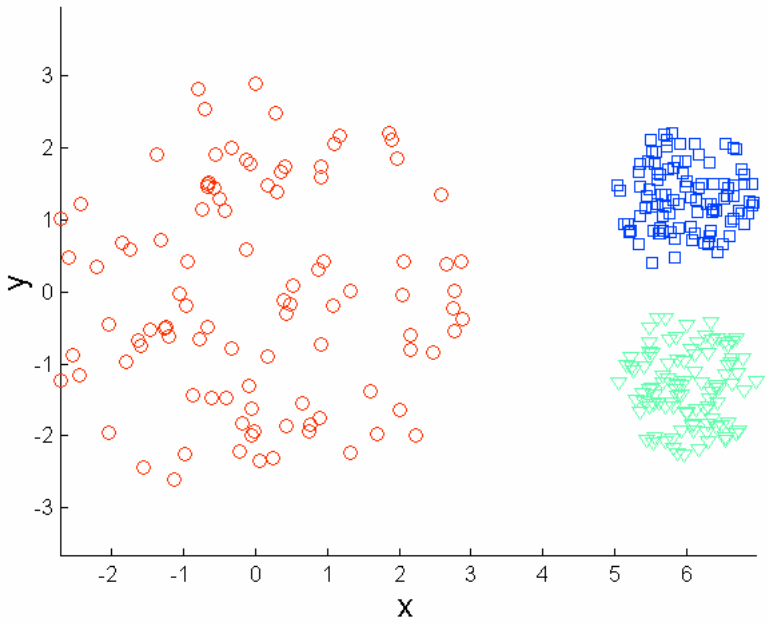
Original Points



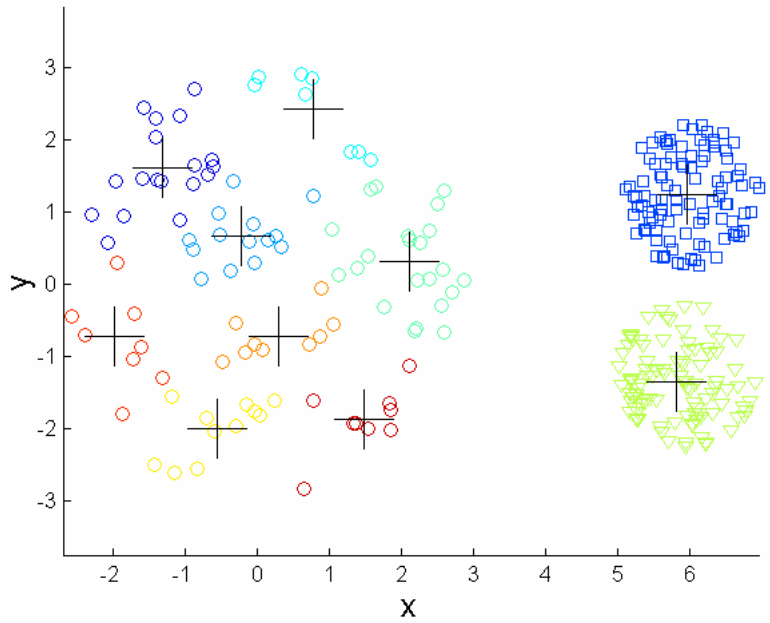
K-means Clusters

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

K-means: Περιορισμοί

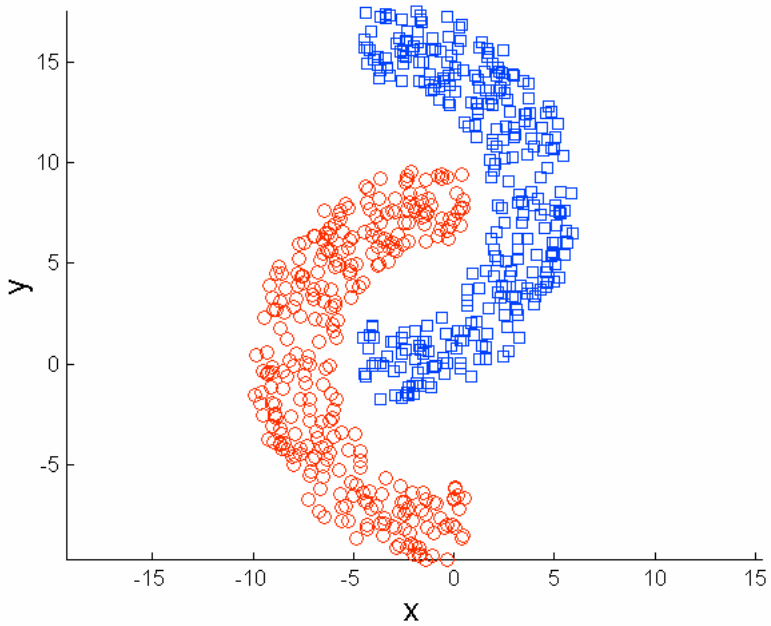


Αρχικά σημεία

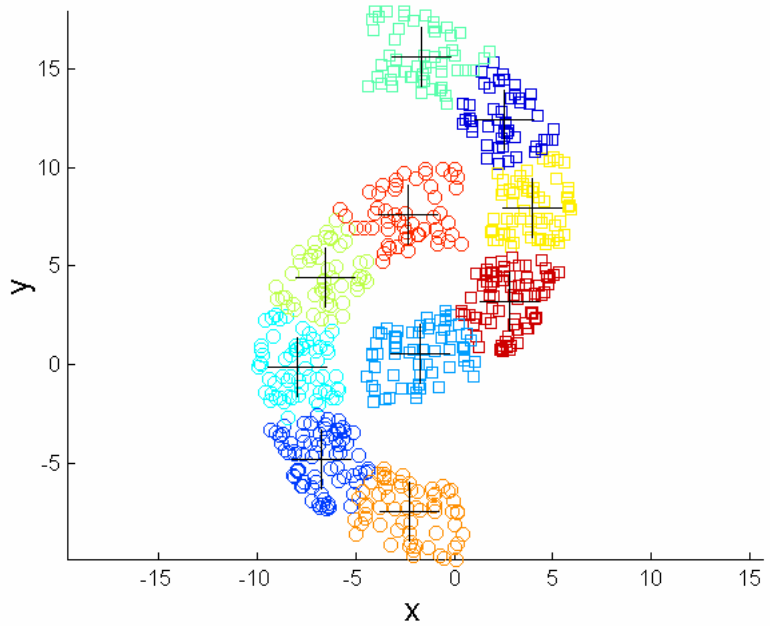


K-means Συστάδες

K-means: Περιορισμοί – διαφορετικά μεγέθη

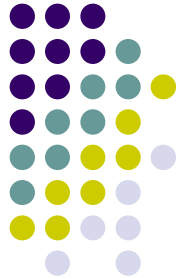


Αρχικά Σημεία



K-means Συστάδες

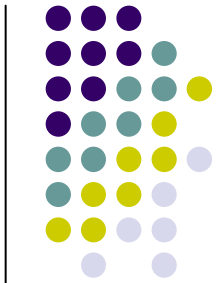
K-medoid



Συνήθως συνεχή d-διάστατο χώρο

Διαλέγει ένα αντιπροσωπευτικό σημείο από τα δεδομένα και ελαχιστοποιεί την απόσταση από αυτό

Μπορεί να εφαρμοστεί σε δεδομένα οποιουδήποτε τύπου (πχ και για κατηγορικά δεδομένα)



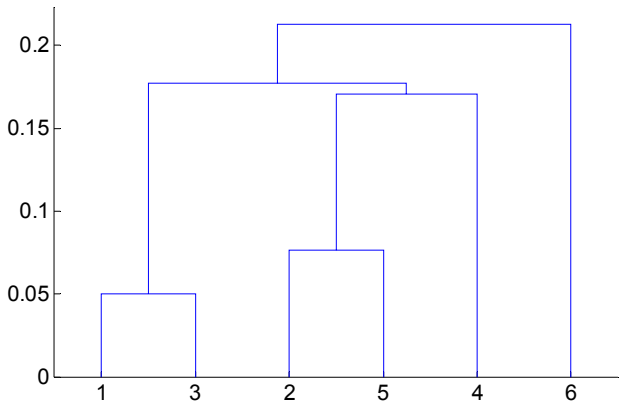
Ιεραρχική Συσταδοποίηση

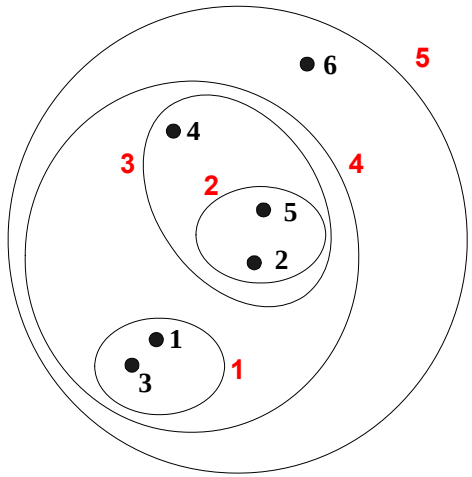
Ιεραρχική Συσταδοποίηση: Βασικά

Παράγει ένα σύνολο από εμφωλευμένες συστάδες οργανωμένες σε ένα ιεραρχικό δέντρο

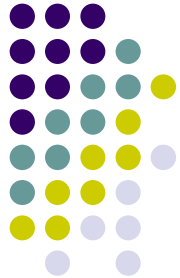
Μπορεί να παρασταθεί με ένα δένδρο-γραμμα

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





Ιεραρχική Συσταδοποίηση: Πλεονεκτήματα

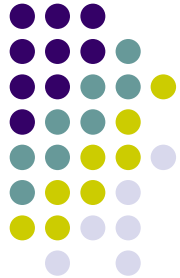


- Δε χρειάζεται να υποθέσουμε ένα συγκεκριμένο αριθμό από συστάδες

  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, …)



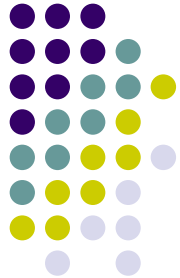
Ιεραρχική Συσταδοποίηση

Δυο βασικοί τύποι ιεραρχικής συσταδοποίησης

- Συσσωρευτικός (Agglomerative):
  - Αρχίζει με τα σημεία ως ξεχωριστές συστάδες
  - Σε κάθε βήμα, συγχωνεύει το ποιο 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)

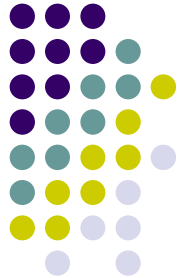
Ιεραρχική Συσταδοποίηση



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

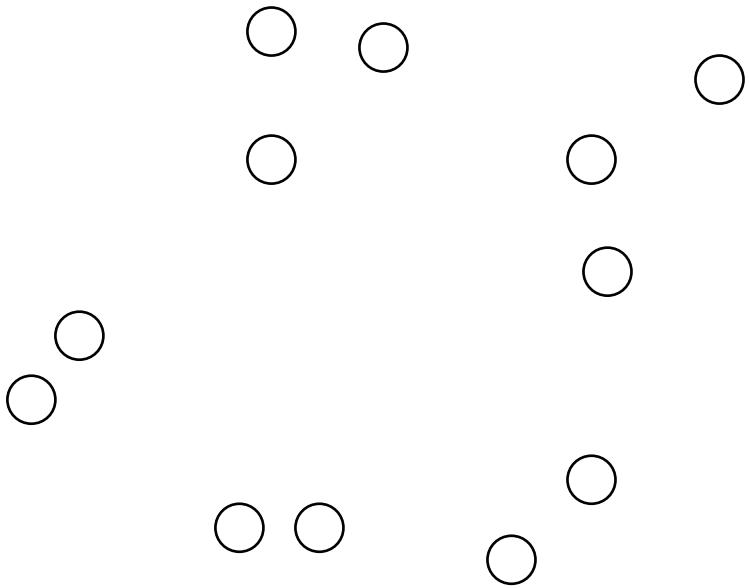
1: Υπολογισμός του Πίνακα Γειτνίασης
2: Έστω κάθε σημείο αποτελεί και μια συστάδα
3: **Repeat**
4: Συγχώνευση των δύο κοντινότερων συστάδων
5: Ενημέρωση του Πίνακα Γειτνίασης
6: **Until** να μείνει μία μόνο συστάδα

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)

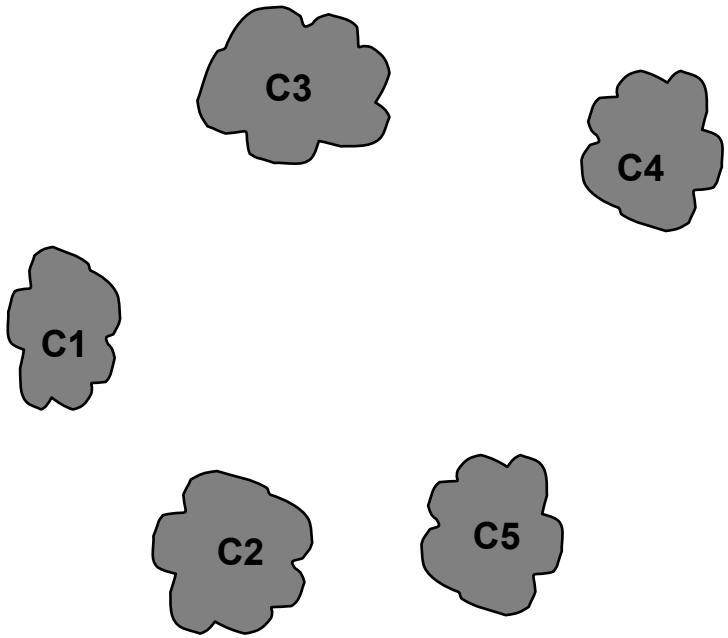


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Πίνακας Γειτνίασης

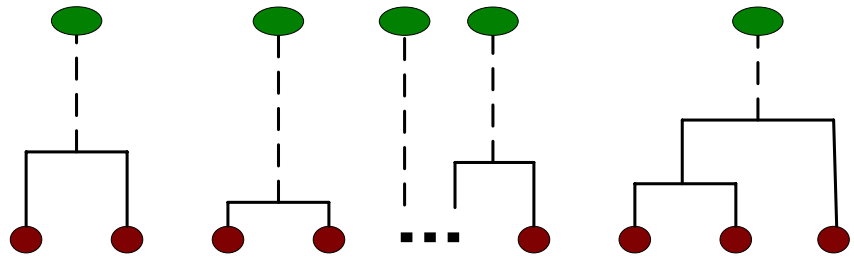
Συσσωρευτική Ιεραρχική Συσταδοποίηση

Μετά από κάποιες συγχωνεύσεις, έχουμε κάποιες συστάδες



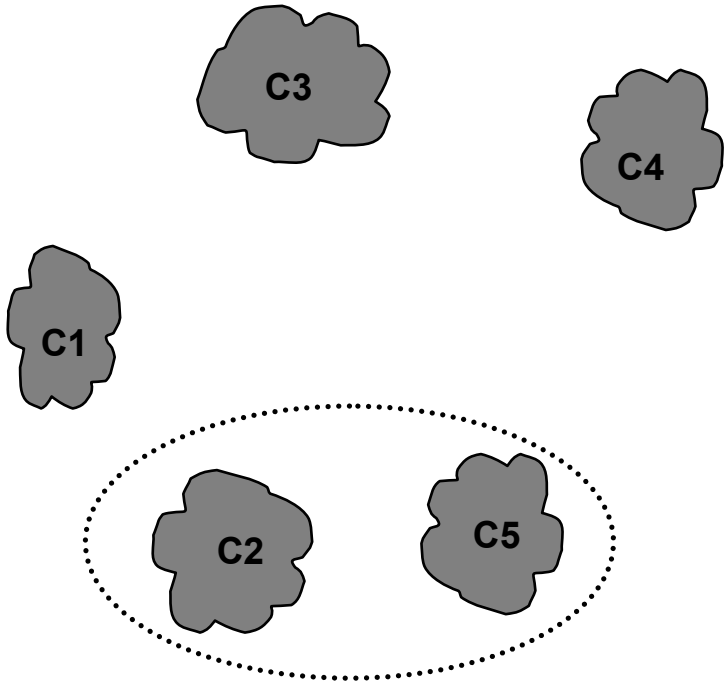
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Πίνακας Γειτνίασης



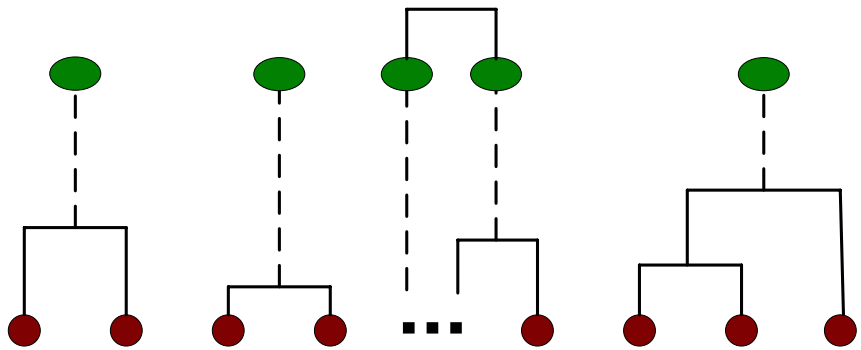
Συσσωρευτική Ιεραρχική Συσταδοποίηση

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



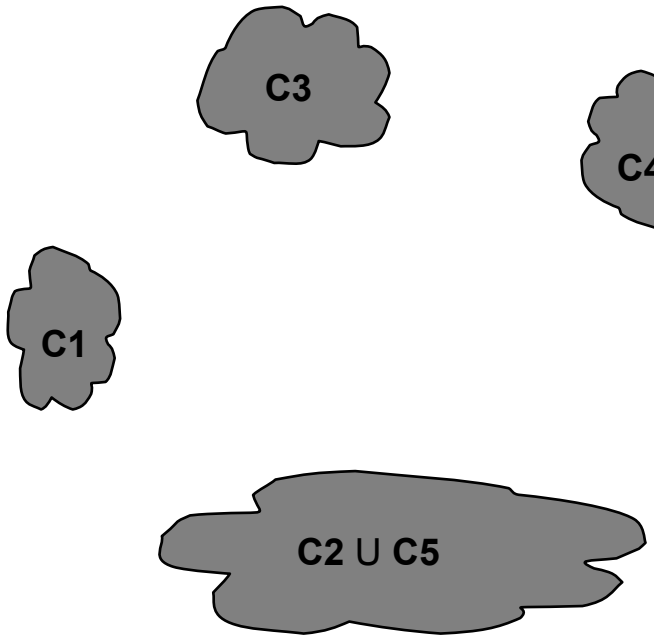
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Πίνακας Γειτνίασης



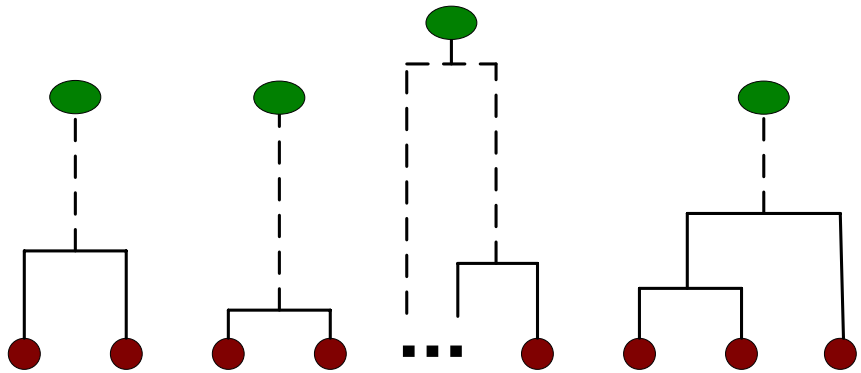
Συσσωρευτική Ιεραρχική Συσταδοποίηση

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

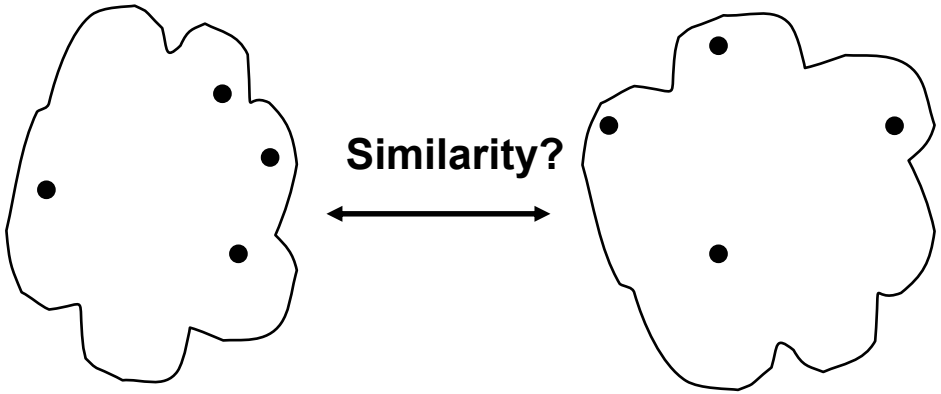


	C1	C2 U C5	C3	C4
C1		?		
C2 U C5	?	?	?	?
C3		?		
C4		?		

Πίνακας Γειτνίασης



ΣΙΣ: Ορισμός απόστασης μεταξύ συστάδων

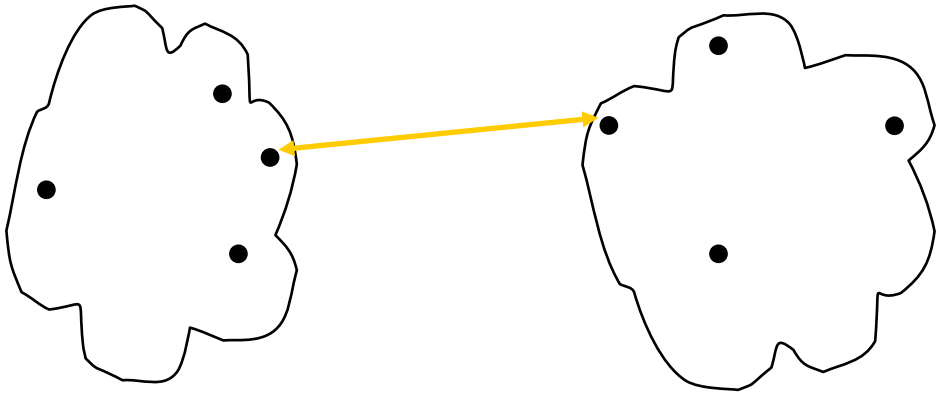


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Πίνακας Γειτνίασης

- MIN
- MAX
- Μέσος όρος της 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



- **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

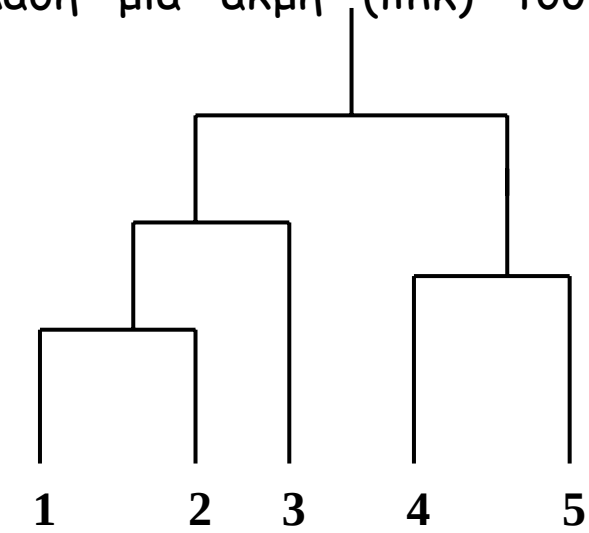
ΣΙΣ: Ορισμός απόστασης μεταξύ συστάδων: MIN

MIN ή μοναδικής ακμής (single link)

Η ομοιότητα μεταξύ δυο συστάδων βασίζεται στα δυο ποιο όμοια (ποιο 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



ομοιότητα

1 (0.4, 0.53)

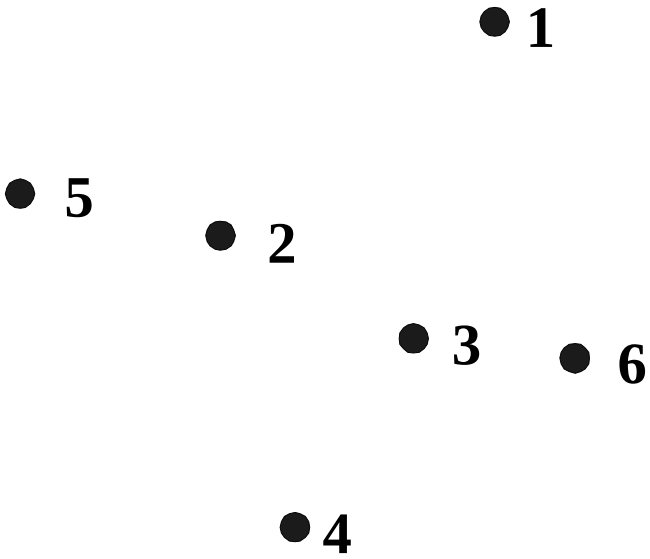
2 (0.22, 0.38)

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)

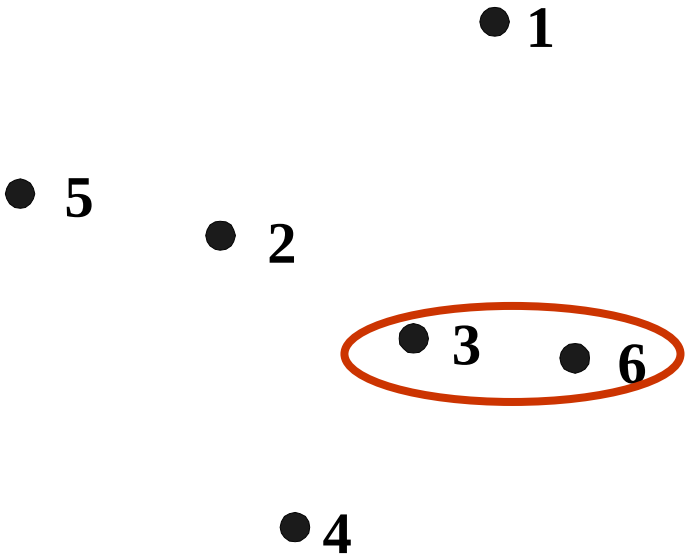
2 (0.22, 0.38)

3 (0.35, 0.32)

4 (0.26, 0.19)

5 (0.08, 0.41)

6 (0.45, 0.30)



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)

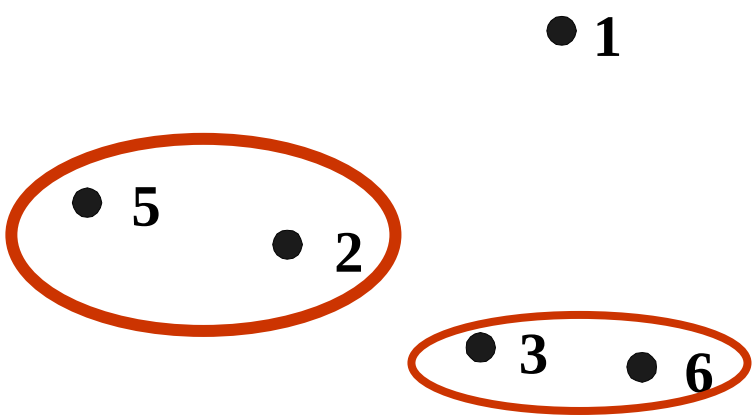
2 (0.22, 0.38)

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)

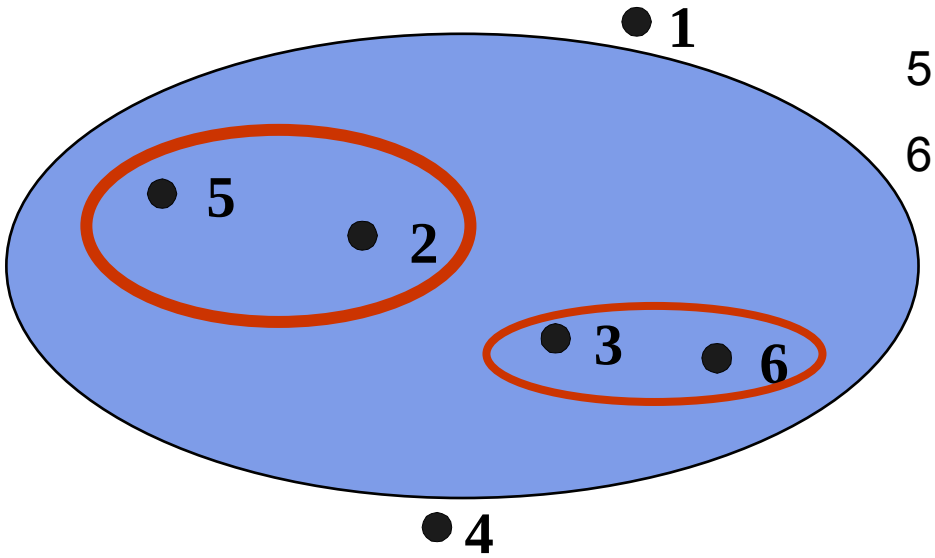
2 (0.22, 0.38)

3 (0.35, 0.32)

4 (0.26, 0.19)

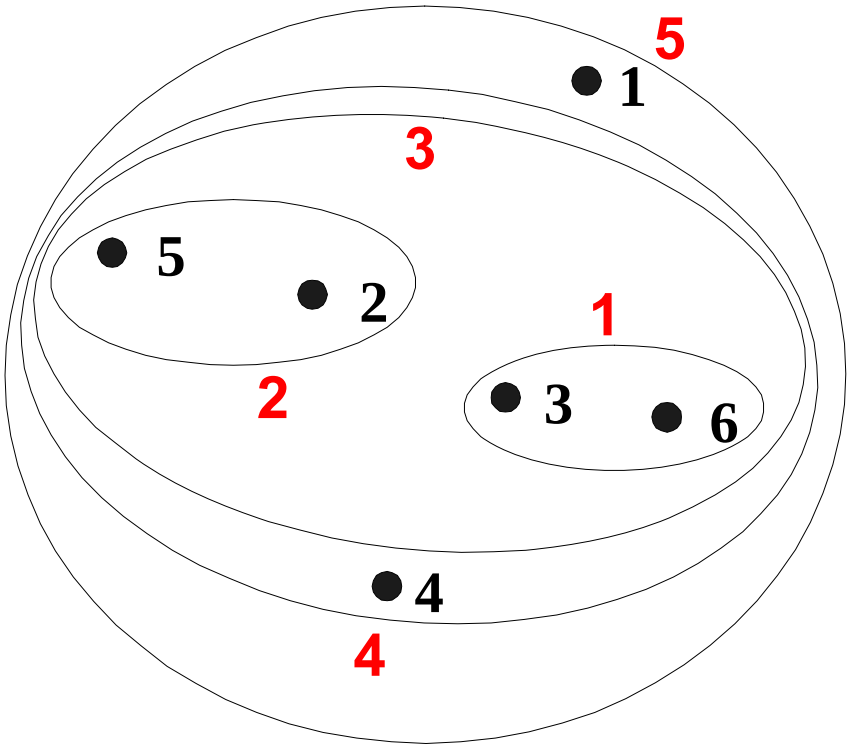
5 (0.08, 0.41)

6 (0.45, 0.30)

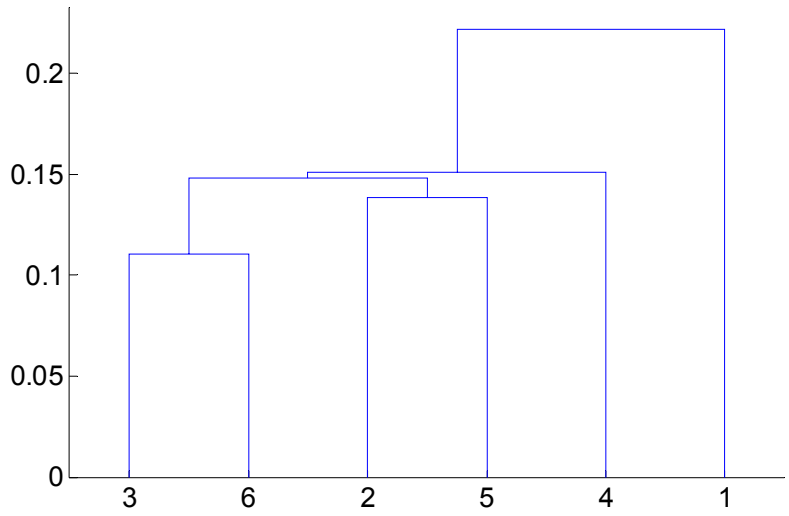


	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	**0.24**	**0.00**	**0.15**	**0.20**	0.14	**0.25**
p3	**0.22**	**0.15**	**0.00**	**0.15**	**0.28**	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	**0.00**	0.39
p6	0.23	0.25	0.11	0.22	0.39	**0.00**

ΣΙΣ: Ορισμός απόστασης μεταξύ συστάδων: MIN



Nested Clusters

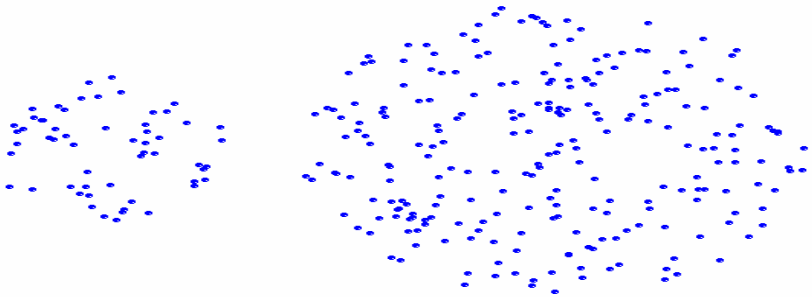


Dendrogram

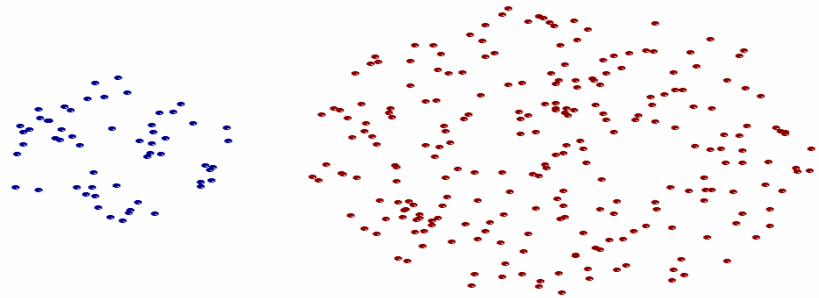
Το δεντρόγραμμα δίνει και τις αποστάσεις

ΣΙΣ: Ορισμός απόστασης μεταξύ συστάδων: MIN

Προτερήματα



Αρχικά σημεία

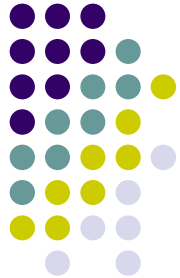


Δύο συστάδες

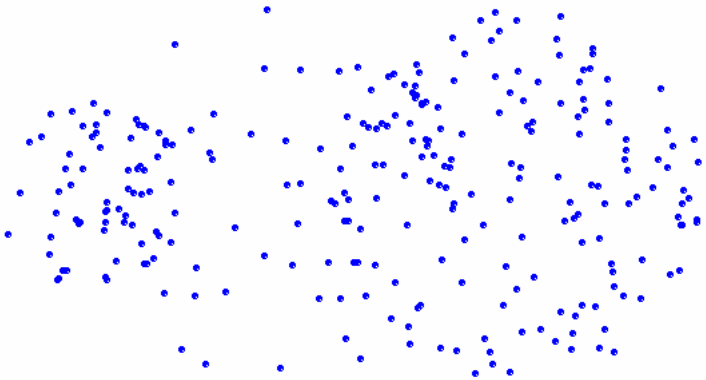
Contiguity-based (συνεχόμενες συστάδες)

Μπορεί να χειριστεί μη ελλειπτικά (non-elliptical) σχήματα

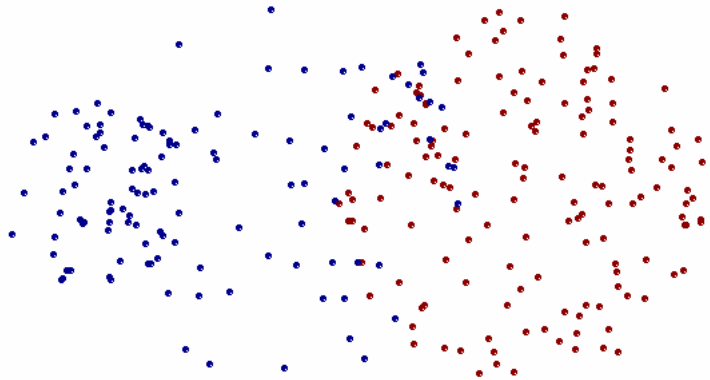
ΣΙΣ: Ορισμός απόστασης μεταξύ συστάδων: MIN



Μειονεκτήματα



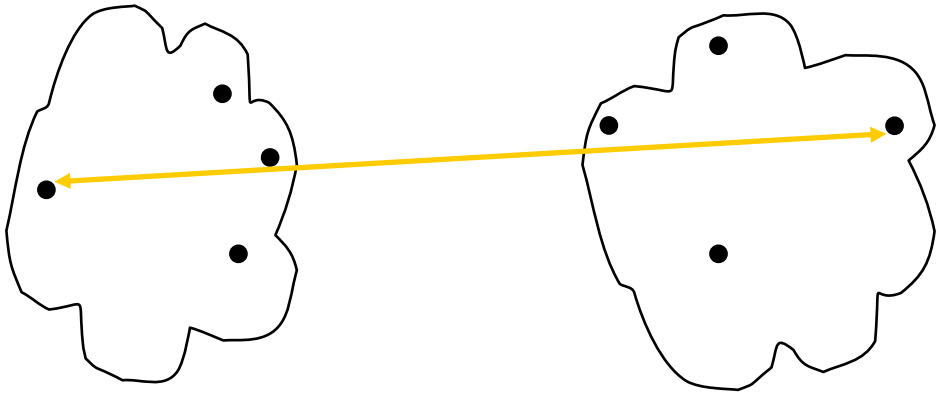
Αρχικά σημεία



Δύο συστάδες

- **Ευαίσθητο σε θόρυβο και outliers**

ΣΙΣ: Ορισμός απόστασης μεταξύ συστάδων



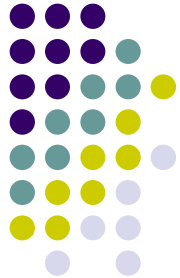
	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						

- 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: MAX



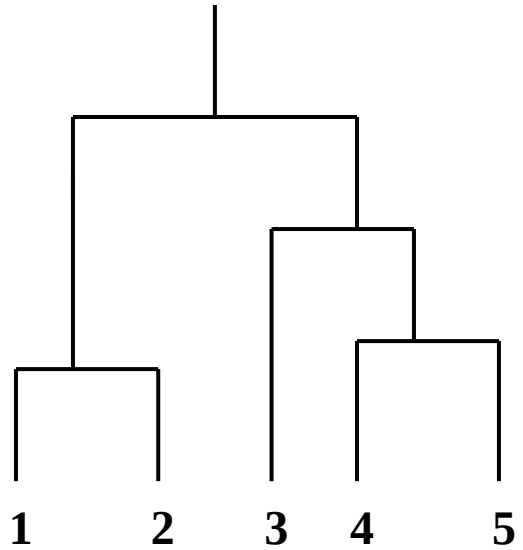
MAX ή πλήρους συνδεσιμότητας (complete linkage)

Η ομοιότητα μεταξύ δυο συστάδων βασίζεται στα δυο λιγότερο όμοια (ποιο μακρινά) σημεία στις διαφορετικές συστάδες (longest edge)

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)

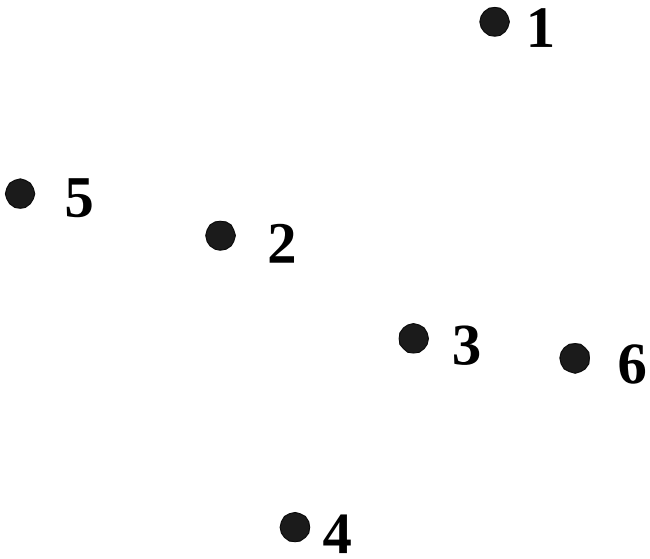
2 (0.22, 0.38)

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)

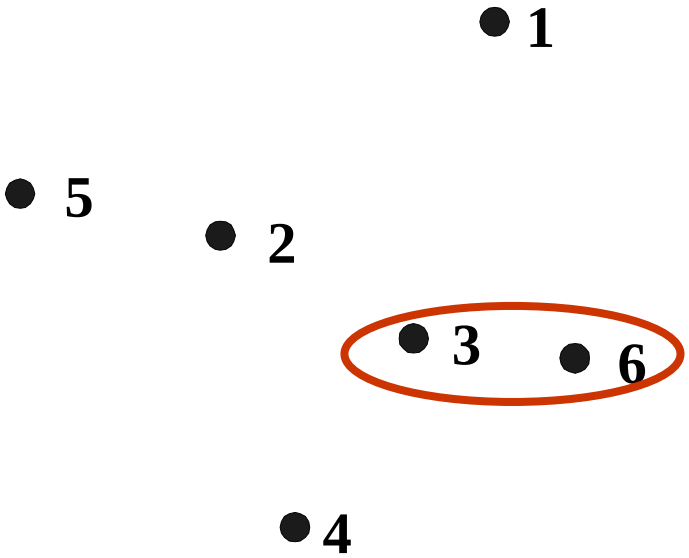
2 (0.22, 0.38)

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)

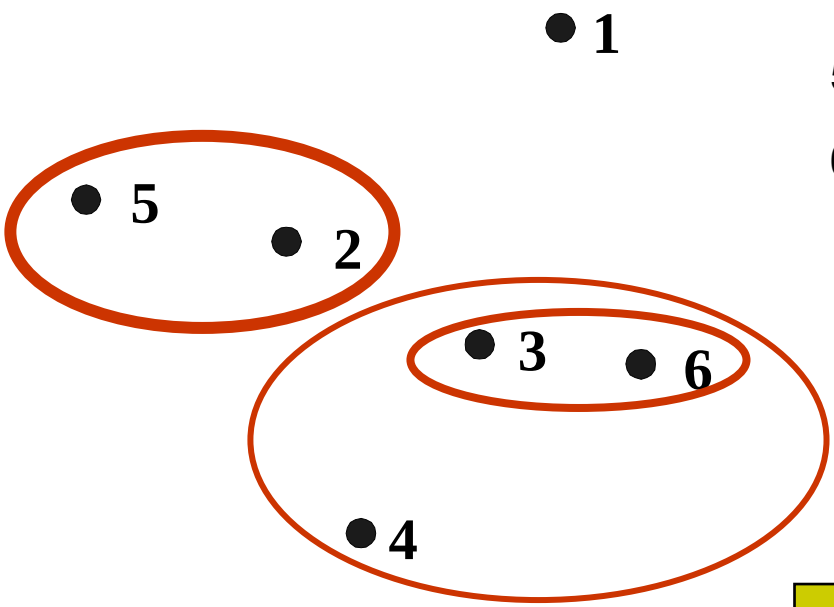
2 (0.22, 0.38)

3 (0.35, 0.32)

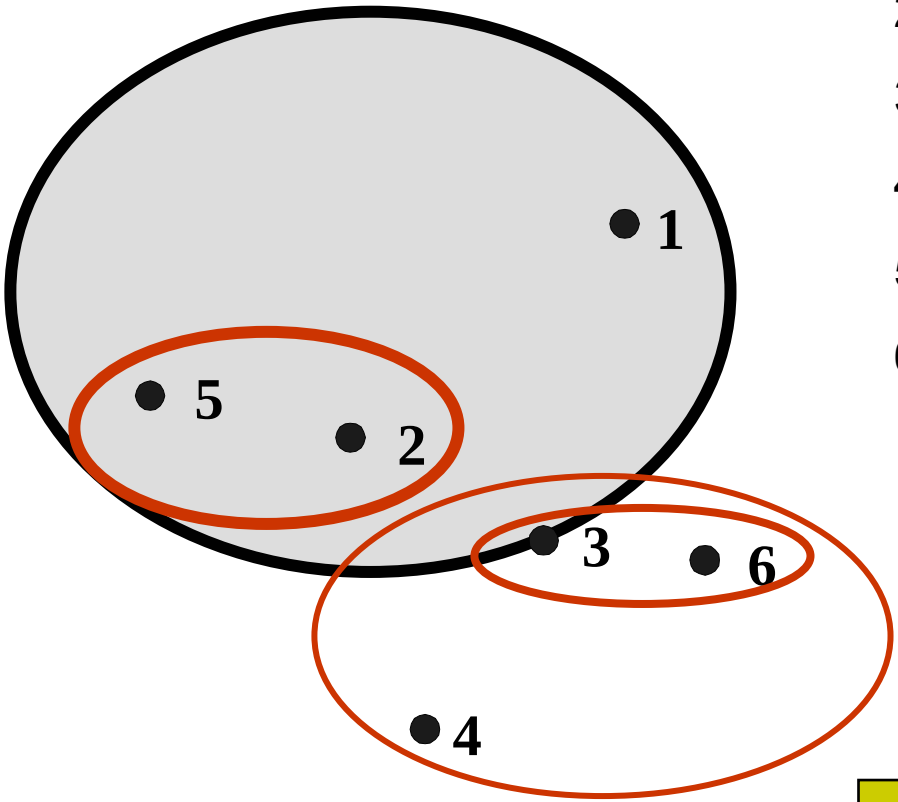
4 (0.26, 0.19)

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)

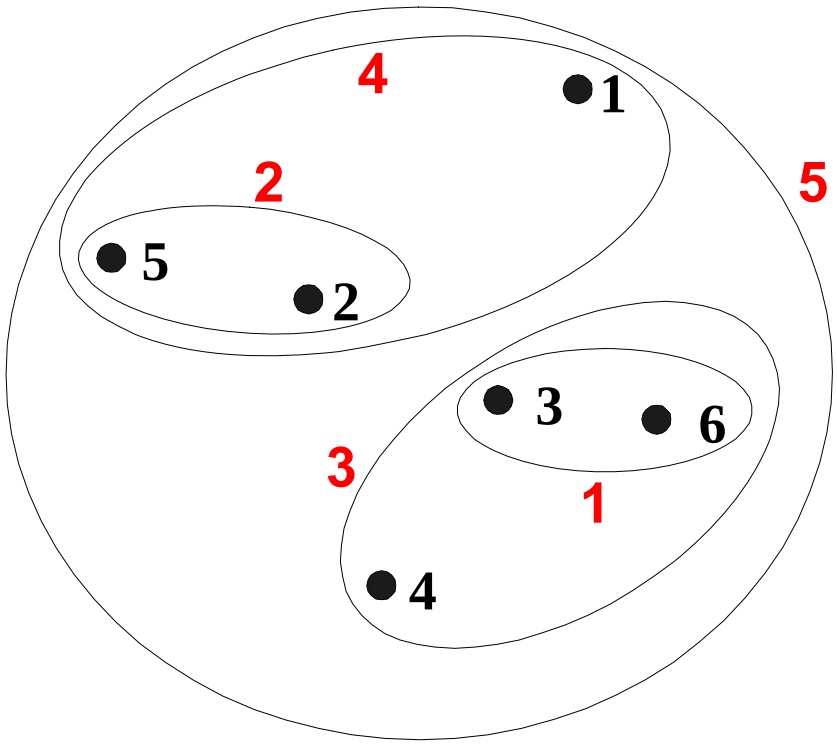
4 (0.26, 0.19)

5 (0.08, 0.41)

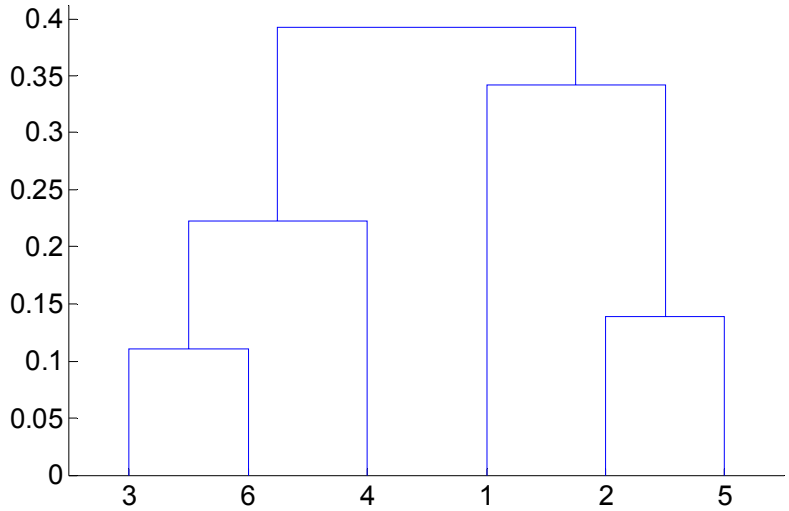
6 (0.45, 0.30)

	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



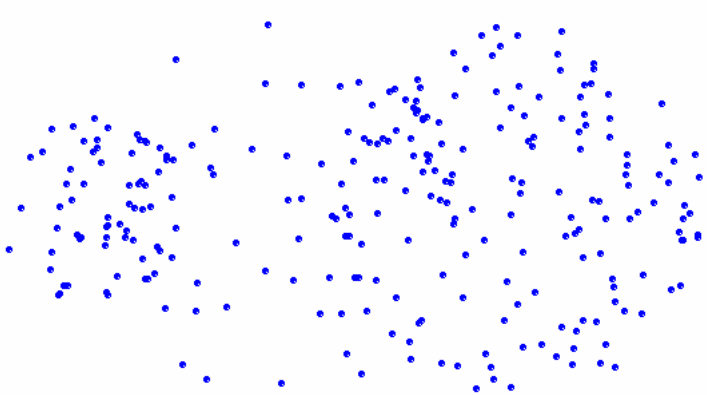
Nested Clusters



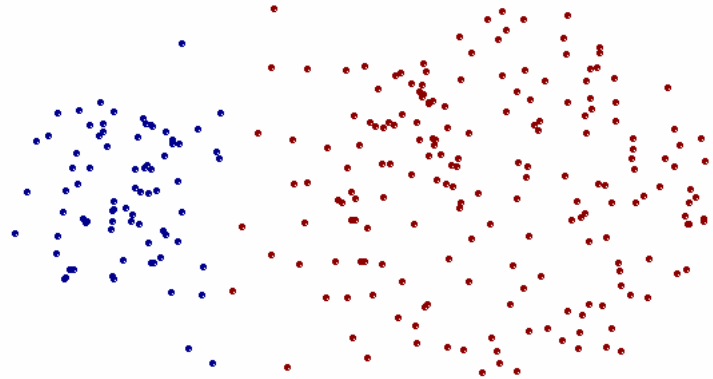
Dendrogram

ΣΙΣ: Ορισμός απόστασης μεταξύ συστάδων: MAX

Πλεονεκτήματα



Original Points

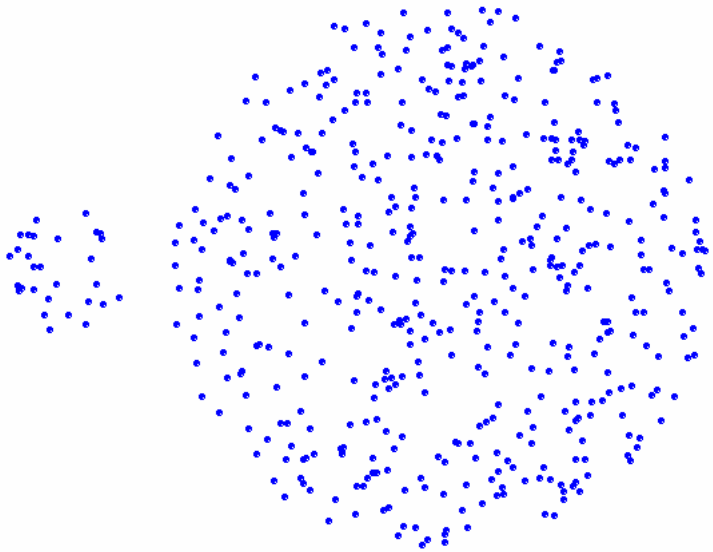


Two Clusters

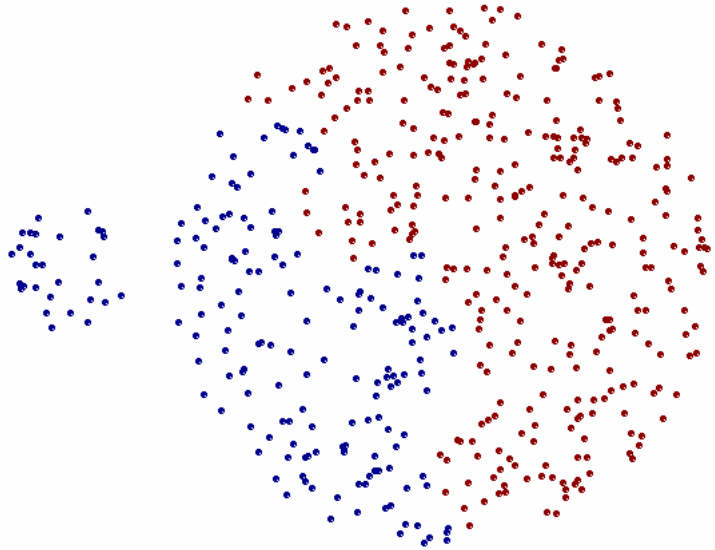
- **λιγότερη εξάρτηση σε θόρυβο και outliers**

ΣΙΣ: Ορισμός απόστασης μεταξύ συστάδων: MAX

Μειονεκτήματα



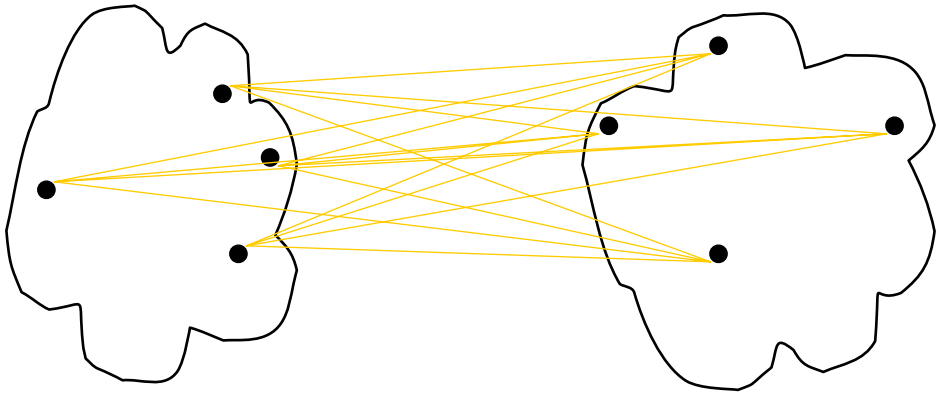
Αρχικά σημεία



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

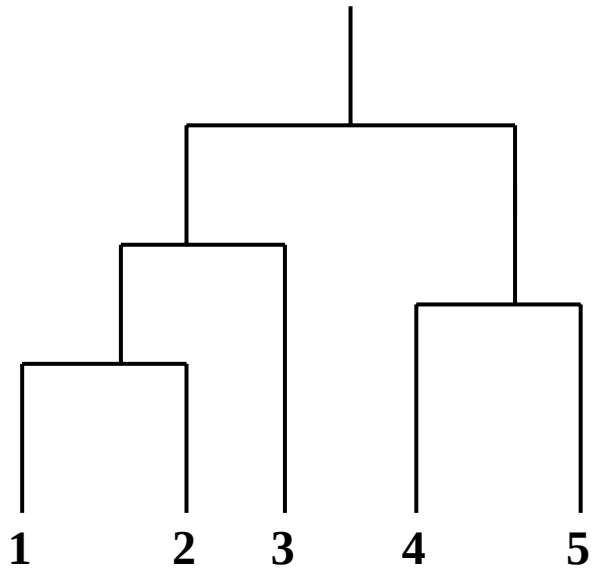
- Κοντινότητα δύο συστάδων είναι η μέση τιμή της ανα-δύο κοντινότητας (average of pairwise proximity) μεταξύ των σημείων των δύο συστάδων.

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

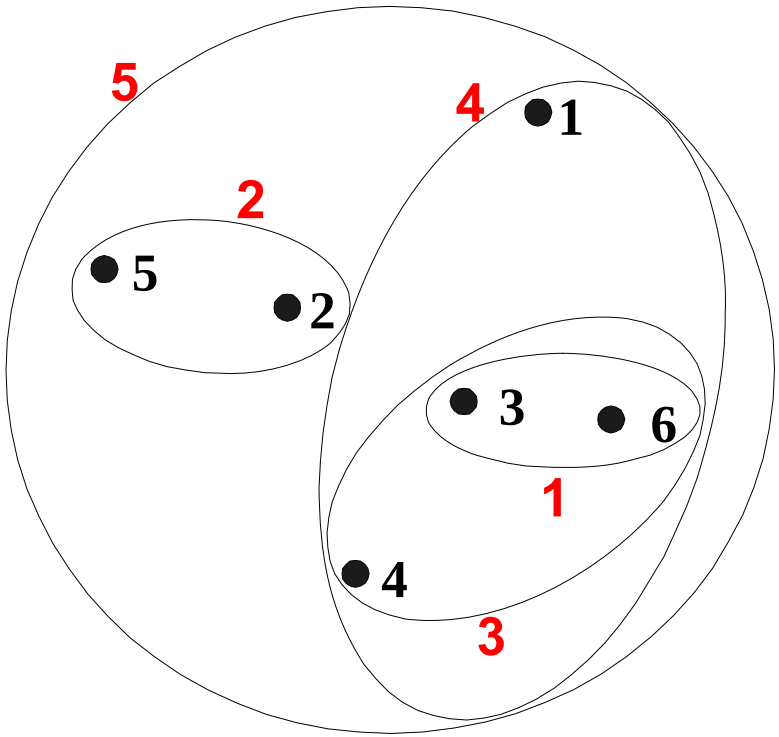
- Χρήση μέσης γιατί η ολική θα έδινα προτίμηση στις μεγάλες συστάδες

ομοιότητα

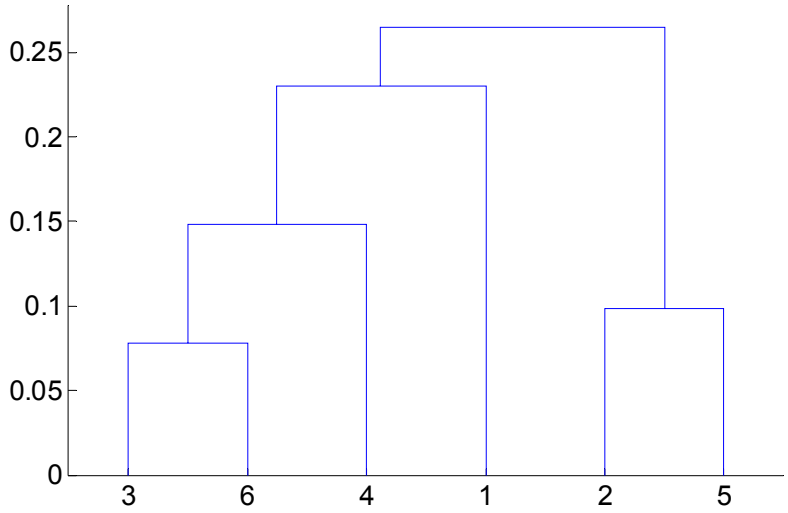
	I1	I2	I3	I4	I5
I1	1.00	0.90	0.10	0.65	0.20
I2	0.90	1.00	0.70	0.60	0.50
I3	0.10	0.70	1.00	0.40	0.30
I4	0.65	0.60	0.40	1.00	0.80
I5	0.20	0.50	0.30	0.80	1.00



ΣΙΣ: Ορισμός απόστασης μεταξύ συστάδων: Μέσο Ομάδας

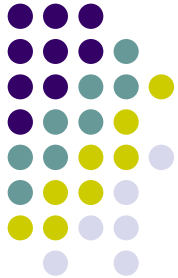


Nested Clusters



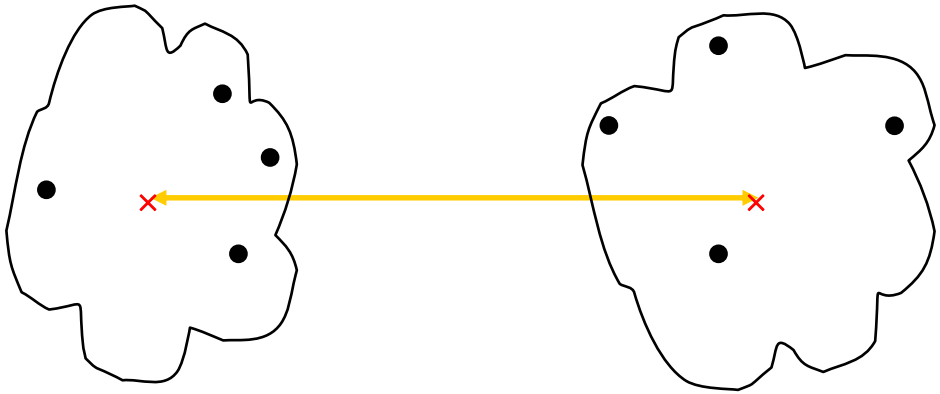
Dendrogram

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

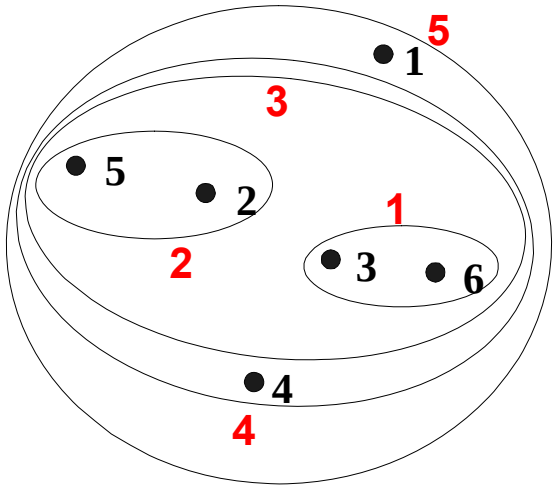
Πρόβλημα: μη μονότονη αύξηση της απόστασης

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

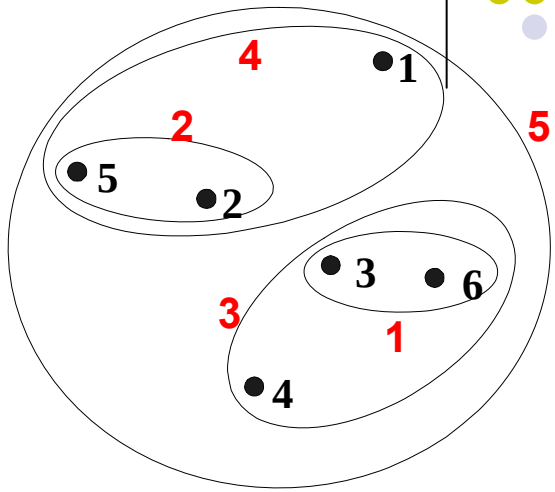
- Βασισμένο στην αύξηση του SSE όταν συγχωνεύονται οι δύο συστάδες
- Ιεραρχικό ανάλογο του k-means
- Μπορεί να χρησιμοποιηθεί για την αρχικοποίηση του k-means

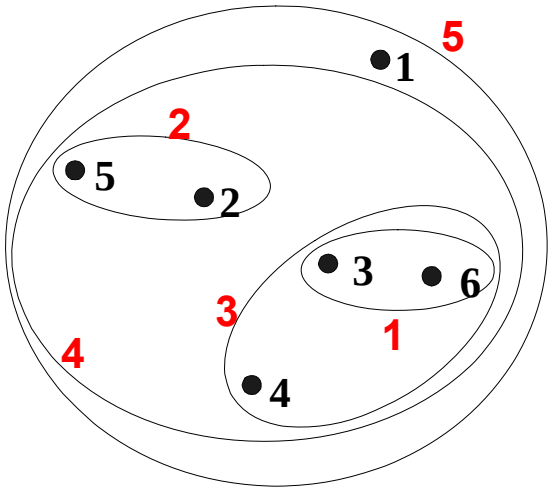
ΣΙΣ: Ορισμός απόστασης μεταξύ συστάδων: Σύγκριση



MIN

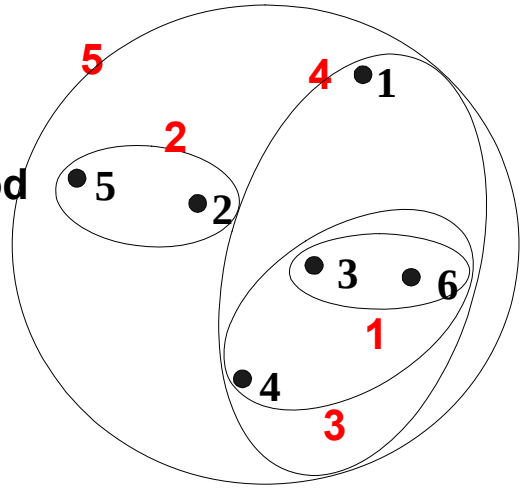
MAX



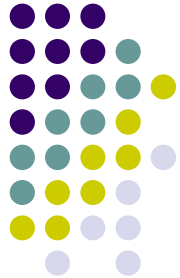


Group Average

Ward's Method



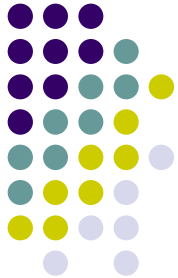
ΣΙΣ: Πολυπλοκότητα Χρόνου και Χώρου



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

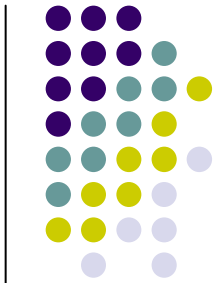
- $O(m^3)$
  - Ξεκινάμε με m συστάδες και μειώνουμε 1 τη φορά
  - Αν γραμμική αναζήτηση του πίνακα $O(m^2)$
  - Καλύτερος χρόνος αν διατηρούμε κάποια ταξινόμηση των αποστάσεων πχ heap

ΣΙΣ: Περιορισμοί και Προβλήματα



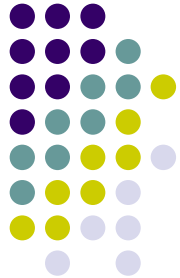
Οι αποφάσεις είναι τελικές – αφού δυο συστάδες συγχωνευτούν αυτό δεν μπορεί να αλλάξει

Δεν ελαχιστοποιούν άμεσα κάποια αντικειμενική συνάρτηση



DBSCAN

DBSCAN: Γενικά



Ο DBSCAN είναι ένας αλγόριθμος βασισμένος στην πυκνότητα
Πυκνότητα = αριθμός σημείων μέσα σε ποια προκαθορισμένη ακτίνα (Eps)

Τα σημεία διαχωρίζονται σε:

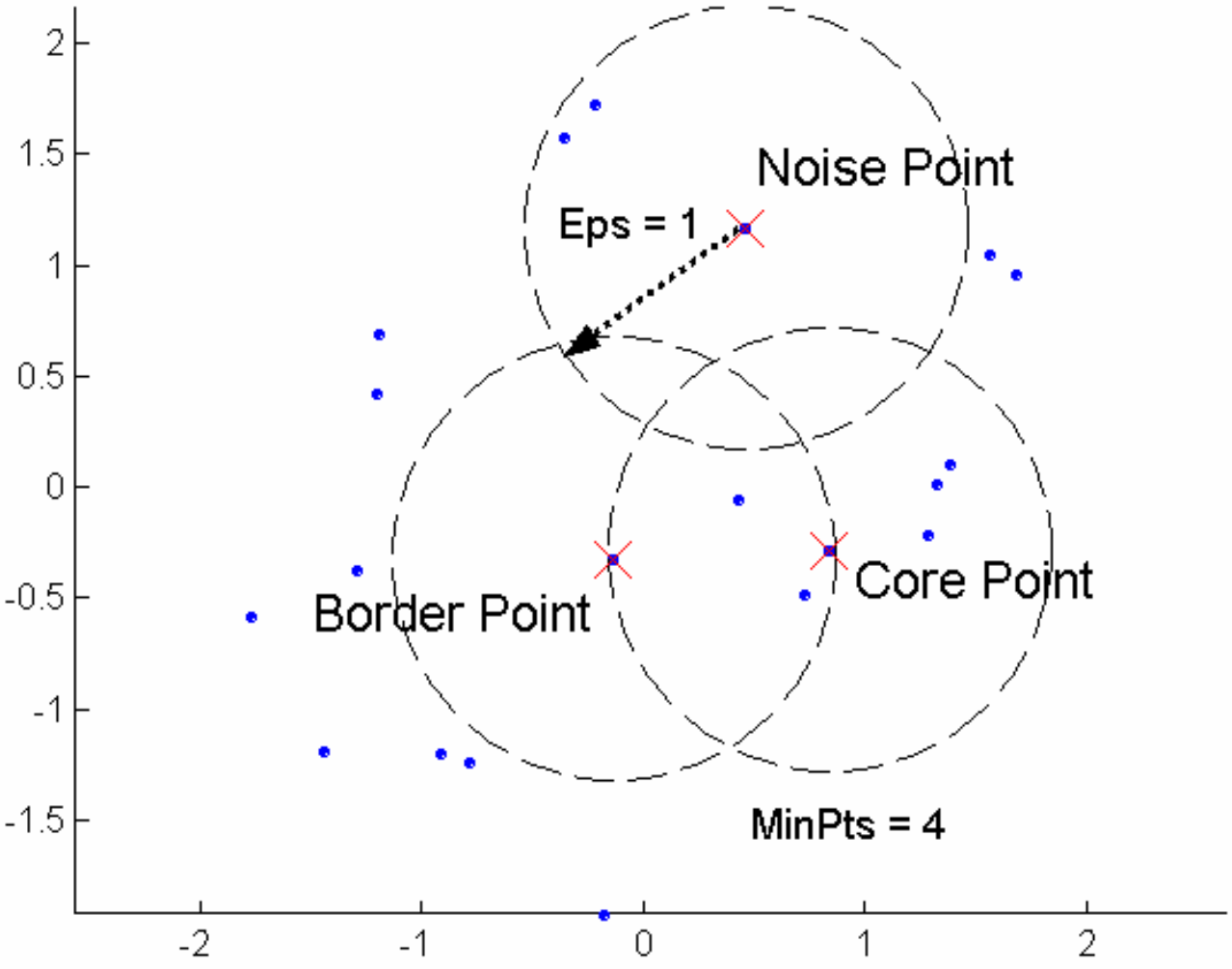
- Βασικά (core): ένα σημείο για το οποίο υπάρχουν περισσότερα από ένα προκαθορισμένο αριθμό (***MinPts***) σημεία σε ακτίνα ***Eps***

Αυτά είναι τα σημεία που είναι στο εσωτερικό μιας συστάδας

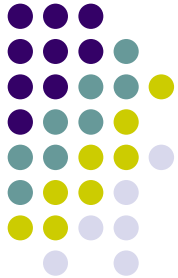
- Οριακά (border): ένα σημείο για το οποίο υπάρχουν λιγότερα από ένα προκαθορισμένο αριθμό (MinPts) σημεία σε ακτίνα Eps, αλλά είναι στη γειτονιά ενός βασικού σημείου

- Θορύβου (noise): ένα σημείο που δεν είναι ούτε βασικό ούτε οριακό

DBSCAN: Γενικά



DBSCAN: Αλγόριθμος



Βασικός Αλγόριθμος

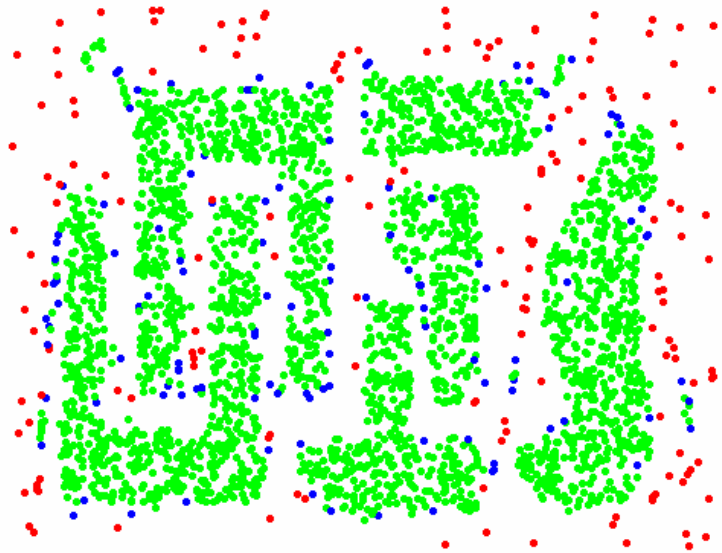
1: Χαρακτήρισε κάθε σημείο ως βασικό, οριακό ή θόρυβο
2: Διέγραψε τα σημεία θορύβου

3: Τοποθέτησε μια ακμή μεταξύ όλων των βασικών σημείων που είναι σε απόσταση έως *Eps* μεταξύ τους
4: Κάνε κάθε ομάδα συνδεδεμένων βασικών σημείων μια διαφορετική συστάδα

5: Ανάθεσε κάθε οριακό σημεία σε μία από τις συστάδες των συσχετιζόμενων του βασικών σημείων

DBSCAN: Αλγόριθμος

Αρχικά σημεία

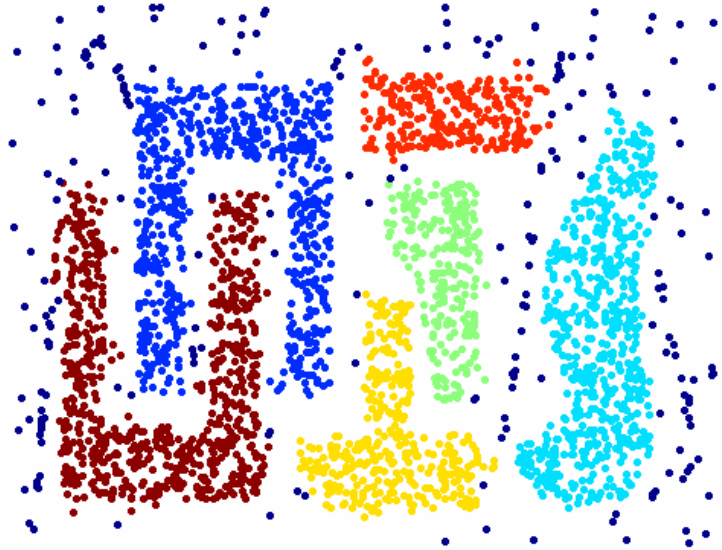


Τύποι σημείων: core, border και noise

Eps = 10, MinPts = 4

DBSCAN: Πλεονεκτήματα

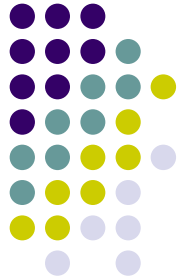
Αρχικά Σημεία



Συστάδες

- Δεν επηρεάζεται από το θόρυβο
- Μπορεί να χειριστεί συστάδες με διαφορετικά σχήματα και μεγέθη

DBSCAN: Πολυπλοκότητα



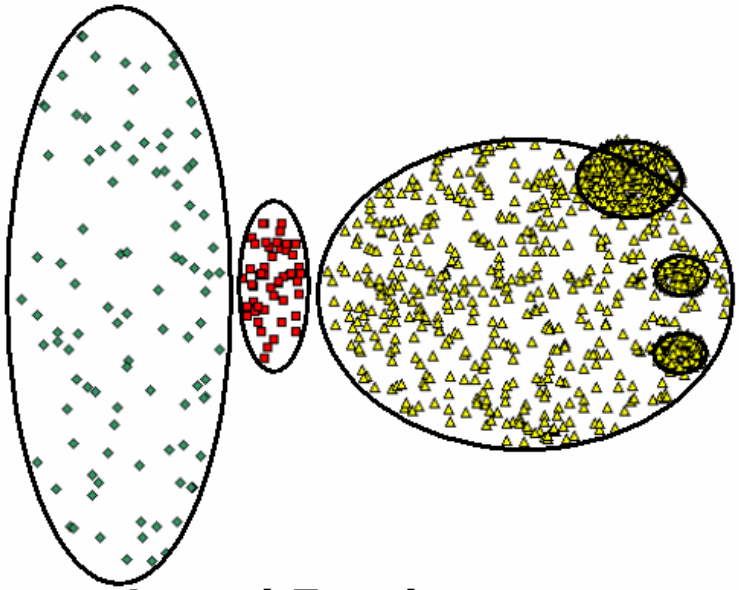
O(m x χρόνος εντοπισμού σημείων σε eps-γειτονιά)

O(m^2)

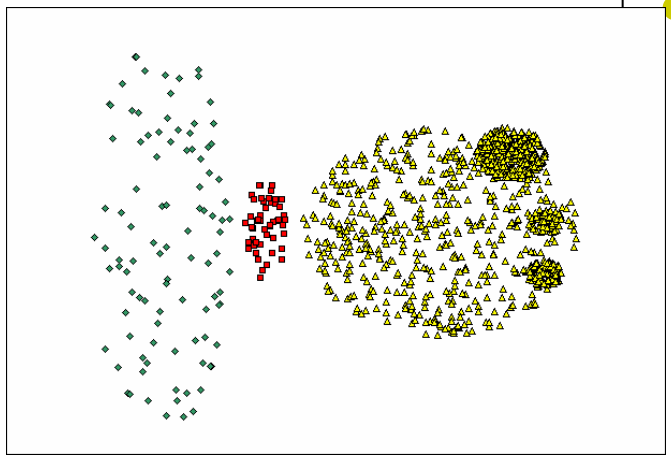
Για μικρό αριθμό διαστάσεων, υπάρχουν δομές που υποστηρίζουν την πράξη σε *O*(m logm)

O(m) χώρος (κρατάμε μόνο ένα label)

DBSCAN: Περιορισμοί

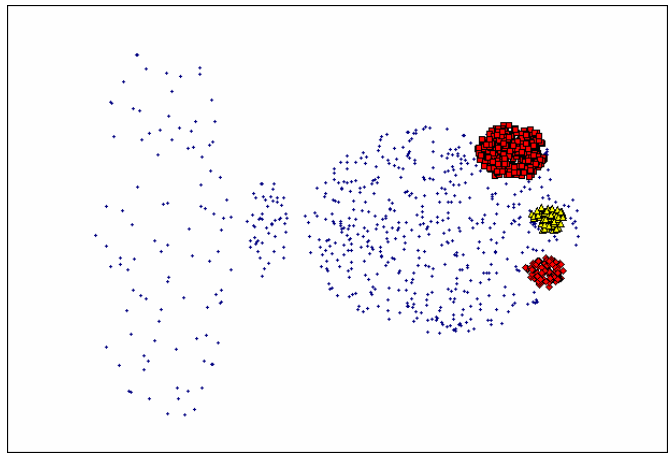


Αρχικά Σημεία



(MinPts=4, Eps=9.75).

- **Διαφορετικές πυκνότητες**
- **Πολυ-διάστατα δεδομένα – δύσκολος ορισμός πυκνότητας και δαπανηρός υπολογισμός γειτόνων**



(MinPts=4, Eps=9.92)

DBSCAN: Καθορισμός των MinPts και Eps

Η ιδέα είναι να κοιτάξουμε την απόσταση ενός σημείου από τον k-οστό κοντινότερο γείτονα του -> k-dist
Γενικά, για τα σημεία που ανήκουν στην ίδια ομάδα, η τιμή του k-dist θα είναι μικρή (αν το k δεν είναι μεγαλύτερο από το μέγεθος της συστάδας)
Θα θέλαμε για τα σημεία μιας συστάδας, να έχουν περίπου την ίδια k-dist
Τα σημεία θορύβου έχουν μεγαλύτερες k-dist

Υπολογίζουμε την k-dist για όλα τα σημεία, για κάποιο k
Ταξινομούμε τις αποστάσεις με φθήνουσα διάταξη
Περιμένουμε ξαφνική αλλαγή στο k-dist που αντιστοιχεί στο Eps
Οπότε k = MinPts και Eps = k-dist