ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ

Αποθήκες Δεδομένων και Εξόρυξη Γνώσης από Δεδομένα

Ανάλυση Συστάδων

Μανώλης Μαραγκουδάκης

Τμήμα Μηχανικών Πληροφοριακών και Επικοινωνιακών Συστημάτων

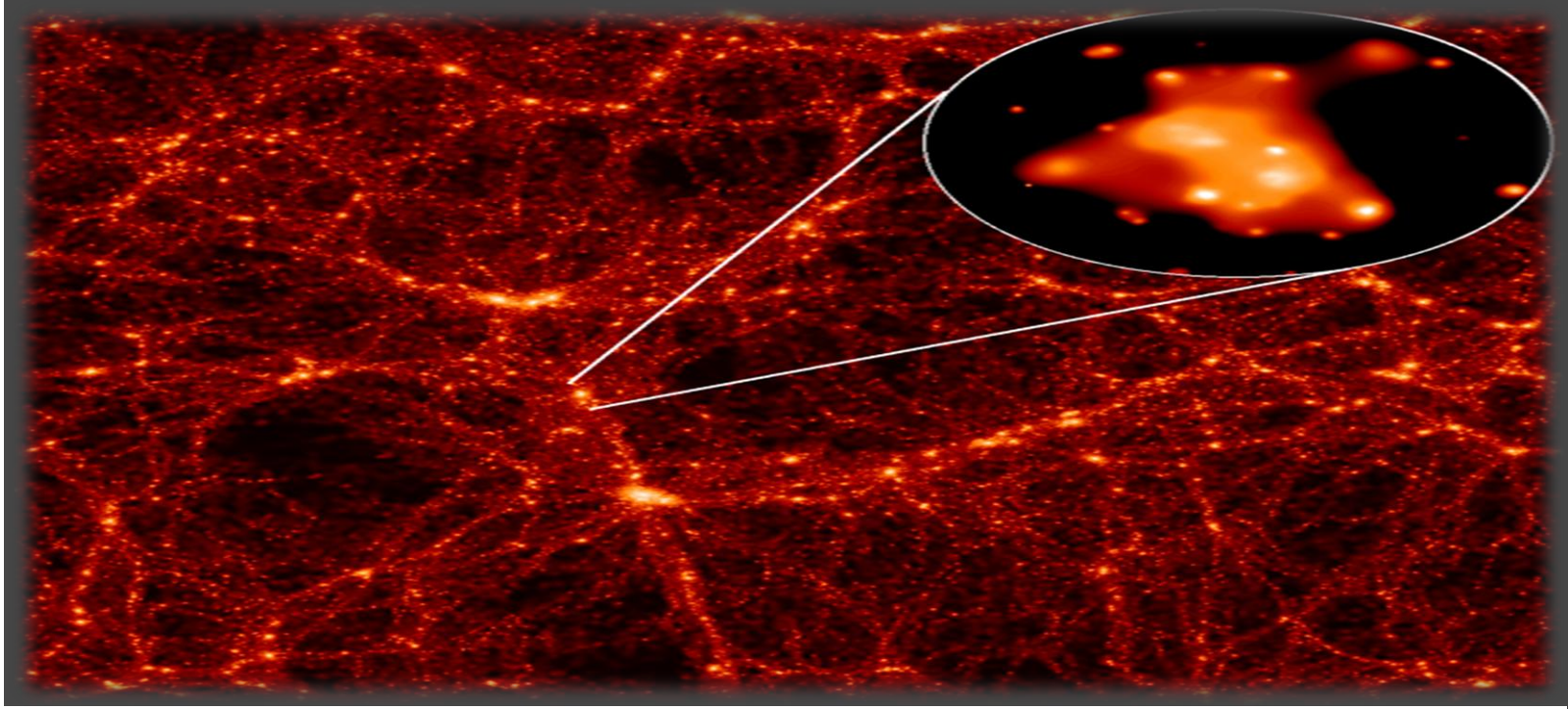
Άδειες Χρήσης

- Το παρόν εκπαιδευτικό υλικό υπόκειται σε άδειες χρήσης Creative Commons.
- Για εκπαιδευτικό υλικό, όπως εικόνες, που υπόκειται σε άλλου τύπου άδειας χρήσης, η άδεια χρήσης αναφέρεται ρητώς.

Χρηματοδότηση

- Το παρόν εκπαιδευτικό υλικό έχει αναπτυχθεί στα πλαίσια του εκπαιδευτικού έργου του διδάσκοντα.
- Το έργο «**Ανοικτά Ακαδημαϊκά Μαθήματα στο Πανεπιστήμιο Αιγαίου**» έχει χρηματοδοτήσει μόνο τη αναδιαμόρφωση του εκπαιδευτικού υλικού.
- Το έργο υλοποιείται στο πλαίσιο του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» και συγχρηματοδοτείται από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο) και από εθνικούς πόρους.

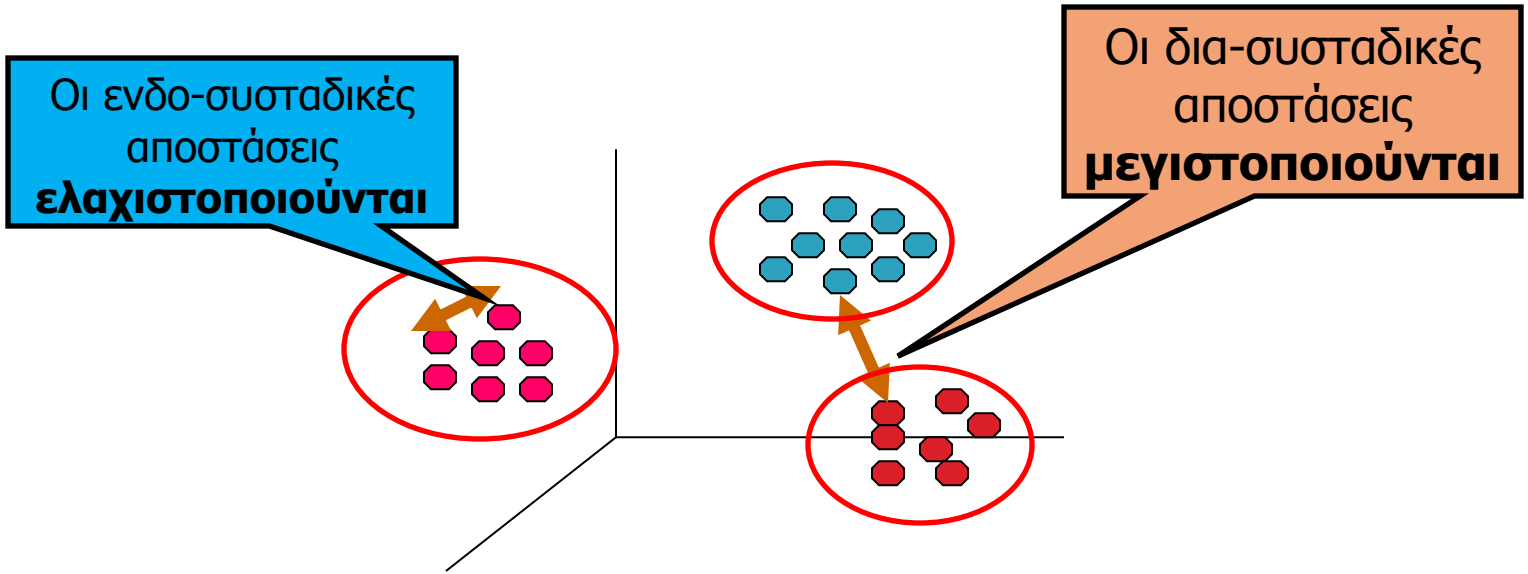
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης



Ενότητα 7: Ανάλυση Συστάδων

Τι είναι η ανάλυση συστάδων

- Η εύρεση ομάδων (συστάδων) αντικειμένων, τέτοιων ώστε τα αντικείμενα μιας ομάδας να είναι όμοια (ή συσχετιζόμενα) και διαφορετικά (ή μη συσχετιζόμενα) από τα αντικείμενα άλλων ομάδων



Εφαρμογές ανάλυσης συστάδων

- **Κατανόηση**
  - Ομαδοποίηση συσχετιζόμενων εγγράφων για πλοήγηση
  - Ομαδοποίηση γονιδίων και πρωτεϊνών με όμοιες λειτουργίες
  - Ομαδοποίηση μετοχών με όμοιες διακυμάνσεις στην τιμή
- **Σύνοψη**
  - Μείωση του μεγάλου όγκου δεδομένων

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

Ομαδοποίηση ατμοσφαιρικής πίεσης στην Αυστραλία

Τι ΔΕΝ είναι ανάλυση συστάδων

- Ταξινόμηση υπό επίβλεψη
  - Έχει πληροφορία για την ετικέτα της κλάσης
- Απλή τμηματοποίηση
  - Διαχωρισμός φοιτητών σε διαφορετικές ομάδες με αλφαβητικό τρόπο
- Αποτελέσματα μιας επερώτησης (query)
  - Οι ομαδοποιήσεις είναι αποτέλεσμα μιας εξωτερικής προδιαγραφής

Ασάφεια στην έννοια της ομάδας

Πόσες ομάδες βλέπετε;

Α. Δυο ομάδες;

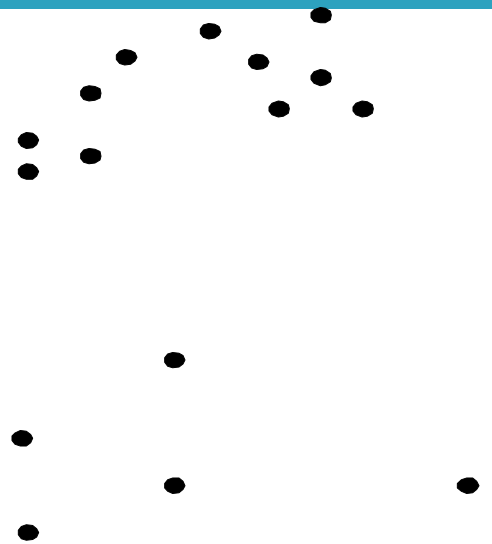
Β. Έξι ομάδες;

Γ. Τέσσερις ομάδες;

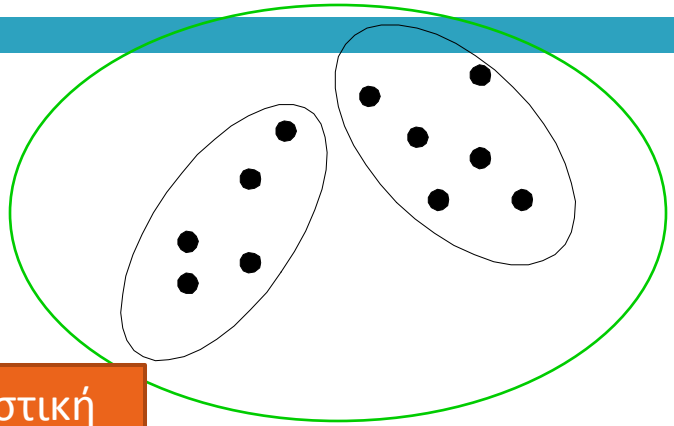
Τύποι ομάδων

- Συσταδοποίηση
  - Ένα σύνολο ομάδων
- Σημαντικός διαχωρισμός μεταξύ ιεραρχικών και διαχωριστικών ομάδων
- Ιεραρχική Συσταδοποίηση (Hierarchical Clustering)
  - Ένα σύνολο από εμφωλευμένες (nested) ομάδες
  - Επιτρέπουμε σε μια συστάδα να έχει υποσυστάδες οργανωμένες σε ένα ιεραρχικό δέντρο
- Διαχωριστική Συσταδοποίηση (Partinional clustering)
  - Ένας διαμερισμός των αντικειμένων σε μη επικαλυπτόμενα - non-overlapping - υποσύνολα (συστάδες) τέτοιος ώστε κάθε αντικείμενο ανήκει σε ακριβώς ένα υποσύνολο

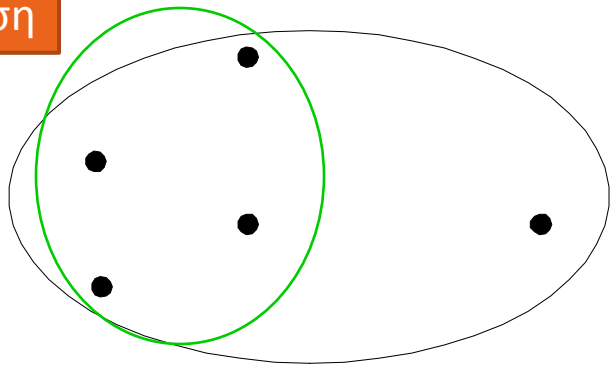
Διαχωριστική Συσταδοποίηση



Αρχικά σημεία



Διαχωριστική συσταδοποίηση

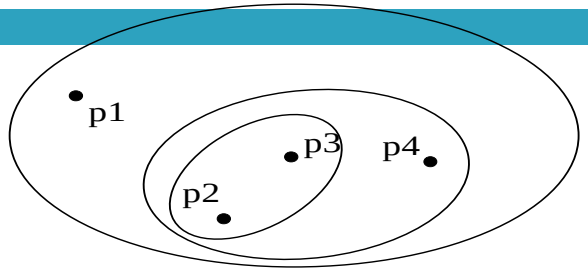


Σημείωση:

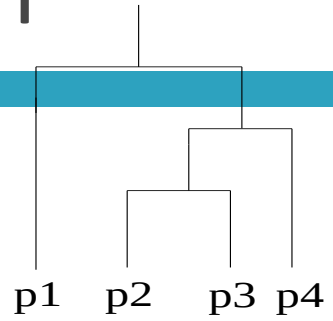
Θόρυβος - outlier

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

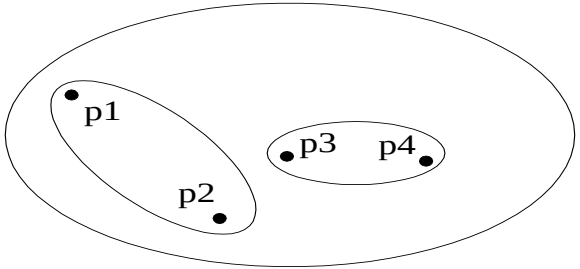
Ιεραρχική συσταδοποίηση



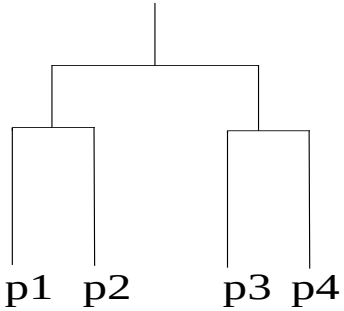
Παραδοσιακή ιεραρχική συσταδοποίηση



Παραδοσιακό δενδρόγραμμα



Μη-παραδοσιακή ιεραρχική συσταδοποίηση



Μη- παραδοσιακό δενδρόγραμμα

Άλλοι διαχωρισμοί μεταξύ συνόλων ομάδων

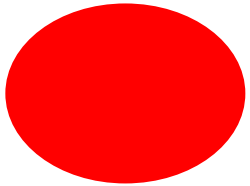
- Επικαλυπτόμενα ή όχι
  - Τα σημεία ανήκουν σε πολλαπλές ομάδες (π.χ. οριακά σημεία)
- Ασαφή ή όχι
  - Στην ασαφή συσταδοποίηση ένα σημείο ανήκει σε κάθε συστάδα με κάποιο βάρος μεταξύ του 0 και του 1
  - Συχνά τα βάρη για κάθε σημείο έχουν άθροισμα 1
  - Η πιθανοτική συσταδοποίηση έχει παρόμοια χαρακτηριστικά
- Μερικά ή όχι
  - Σε μερικές περιπτώσεις ομαδοποιούμε ορισμένα δεδομένα
- Ετερογενή ή όχι
  - πολύ διαφορετικά μεγέθη, σχήματα και πυκνότητες (densities)

Τύποι συστάδων

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

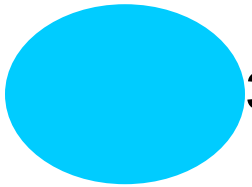
Τύποι: καλώς διαχωρισμένες συστάδες

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

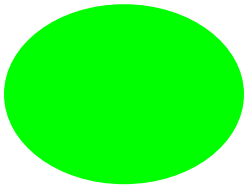


Συχνά υπάρχει η έννοια του κατωφλιού (threshold)

Όχι απαραίτητα κυκλικές (οποιοδήποτε σχήμα)



3 καλώς διαχωρισμένες συστάδες



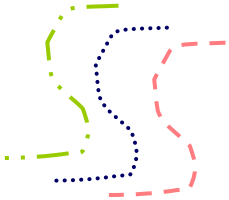
Τύποι: συστάδες βασισμένες σε κέντρο

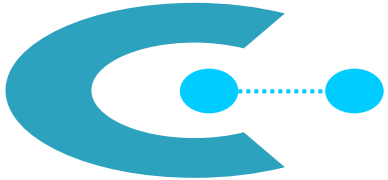
- Μια συστάδα είναι ένα σύνολο από αντικείμενα τέτοιο ώστε ένα αντικείμενο στην ομάδα είναι κοντινότερο σε (ή πιο όμοιο με) το «κέντρο» ή πρότυπο της ομάδας από ότι από το κέντρο οποιασδήποτε άλλης ομάδας.
  - Το κέντρο της ομάδας είναι συχνά
    - centroid, ο μέσος όρος των σημείων της συστάδας, ή
    - medoid, το πιο «αντιπροσωπευτικό» σημείο της συστάδας (πχ όταν υπάρχουν κατηγορικά γνωρίσματα)

4 συστάδες βασισμένες στο κέντρο

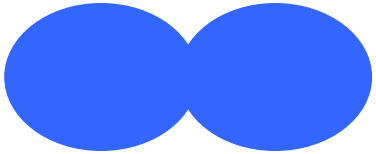
Τύποι: Συνεχείς συστάδες

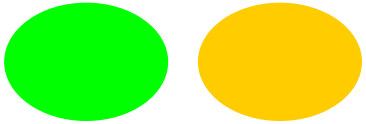
- Συνεχής συστάδα (Πλησιέστερος γείτονας ή μεταβατική)
  - Μια συστάδα είναι ένα σύνολο σημείων τέτοιο ώστε κάθε σημείο είναι **ποιο κοντά σε ένα ή περισσότερα σημεία της συστάδας** από ότι σε οποιοδήποτε σημείο εκτός ομάδας
  - Συχνά σε περιπτώσεις συστάδων με μη κανονικό σχήμα ή με αλληλοπλεκόμενα σχήματα
  - Πρόβλημα με θόρυβο





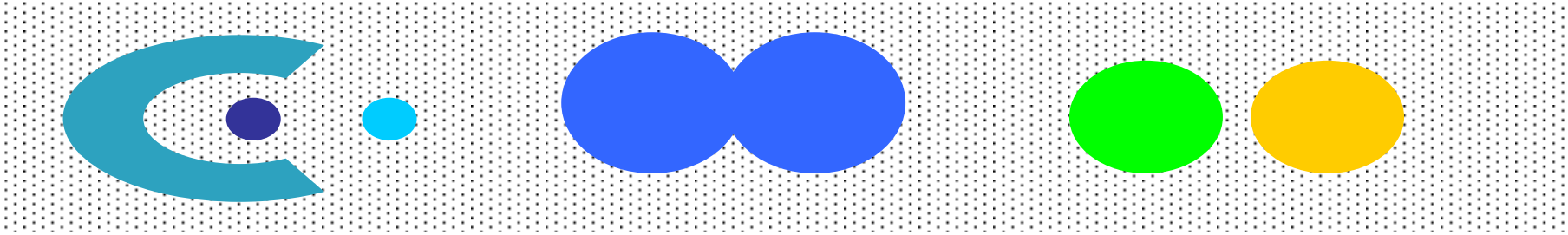
8 συνεχείς συστάδες





Τύποι: βασισμένες στην πυκνότητα

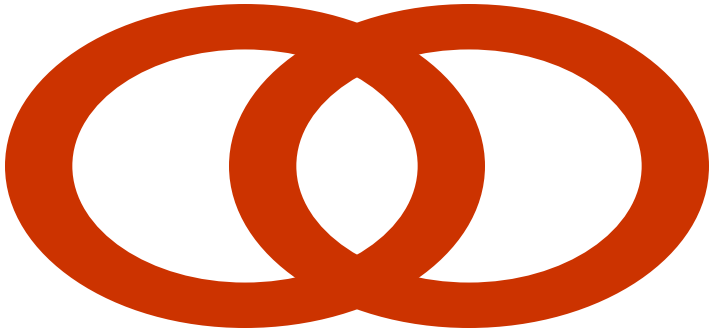
- Με βάση την πυκνότητα
  - Μια συστάδα είναι μια πυκνή περιοχή σημείων, που διαχωρίζεται από περιοχές χαμηλής πυκνότητας σε σχέση με άλλες που έχουν υψηλή πυκνότητα
    - Χρησιμοποιούνται σε περιπτώσεις συστάδων με μη κανονικό σχήμα ή με αλληλοπλεκόμενα σχήματα ή όταν θόρυβος ή outliers



6 συστάδες βασισμένες στην πυκνότητα

Τύποι: εννοιολογική συσταδοποίηση

- Εννοιολογικές συστάδες ή συστάδες κοινής ιδιότητας
  - Εύρεση ομάδων που μοιράζονται κάποια κοινή ιδιότητα ή αναπαριστούν συγκεκριμένη έννοια



2 επικαλυπτόμενοι κύκλοι

Τύποι: αντικειμενική συνάρτηση

- Συστάδες με 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 ολικού

Τύποι: αντικειμενική συνάρτηση

- Χρησιμοποιείται ένας πίνακας εγγύτητας για να ορίσει ένα ζυγισμένο γράφο
  - Κόμβοι: σημεία που θέλουμε να ομαδοποιήσουμε
  - Ακμές: αναπαριστούν την εγγύτητα μεταξύ των σημείων
- Η συσταδοποίηση γίνεται η διαδικασία διαχωρισμού του γράφου σε συνδεδεμένα στοιχεία
- Θέλουμε να ελαχιστοποιήσουμε τα βάρη των ακμών μεταξύ των συστάδων και να μεγιστοποιήσουμε τα βάρη μέσα στις συστάδες

Χαρακτηριστικά 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

Αλγόριθμοι συσταδοποίησης

- K-means και οι παραλλαγές του
- Ιεραρχική συσταδοποίηση
- Συσταδοποίηση βασισμένη στην πυκνότητα

K-means

- Είναι διαχωριστικός αλγόριθμος
- Κάθε συστάδα σχετίζεται με ένα κεντικό σημείο **(centroid)**
- Σε κάθε σημείο ανατίθεται η ομάδα με το πλησιέστερο centroid
- Προσδιορίζεται ο αριθμός των συστάδων, K
- Ο βασικός αλγόριθμος είναι πολύ απλός

K-means

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: Λεπτομέρειες

- Τα αρχικά κεντρικά σημεία επιλέγονται συχνά με τυχαία σειρά
  - Οι συστάδες που παράγονται διαφέρουν η μια από τις άλλες
- Το centroid είναι (συνήθως) το μέσο (Mean) των σημείων της συστάδας
- Η εγγύτητα μετράται με την Ευκλείδεια απόσταση, την ομοιότητα συνημίτονου, τη συσχέτιση, κτλ.
- Για τις παραπάνω απλές μετρικές αποστάσεων, ο K-means θα συγκλίνει με βεβαιότητα
- Η σύγκλιση γίνεται στις πρώτες επαναλήψεις
- Η πολυπλοκότητα είναι της τάξεως: O(n * K * I * d)
  - n = σημεία, K = συστάδες, I = επαναλήψεις, d = ιδιότητες (attributes)

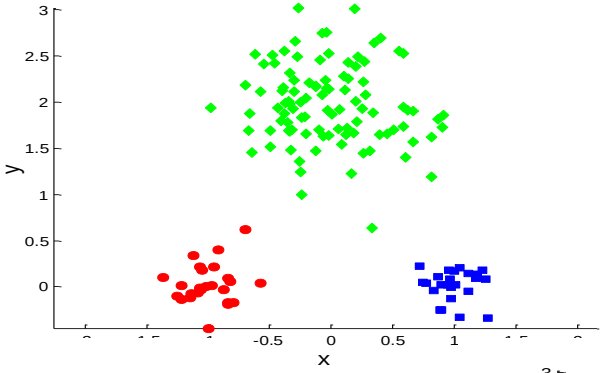
K-means: εκτίμηση ποιότητας

- 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

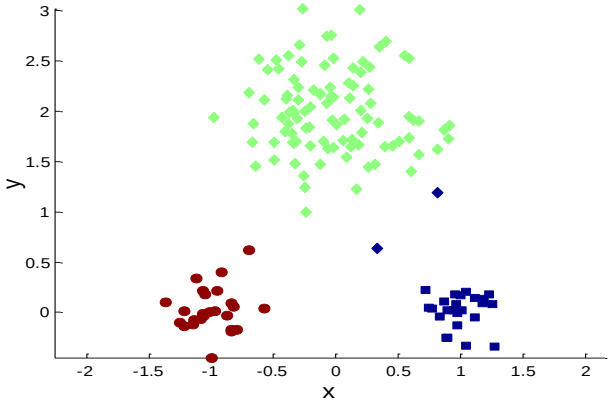
$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$

- 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 $c_i = 1/m_i \, \Sigma \, x \in C_i \, x$
- Δοθέντων δύο συστάδων, μπορούμε να επιλέξουμε αυτήν με το μικρότερο λάθος

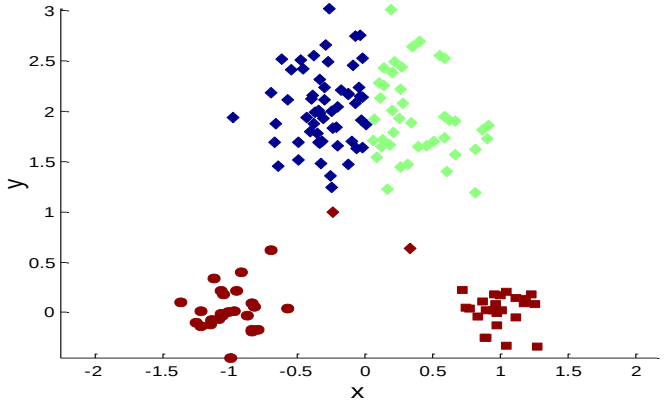
Δυο διαφορετικές συσταδοποιήσεις με k-means



Αρχικά σημεία

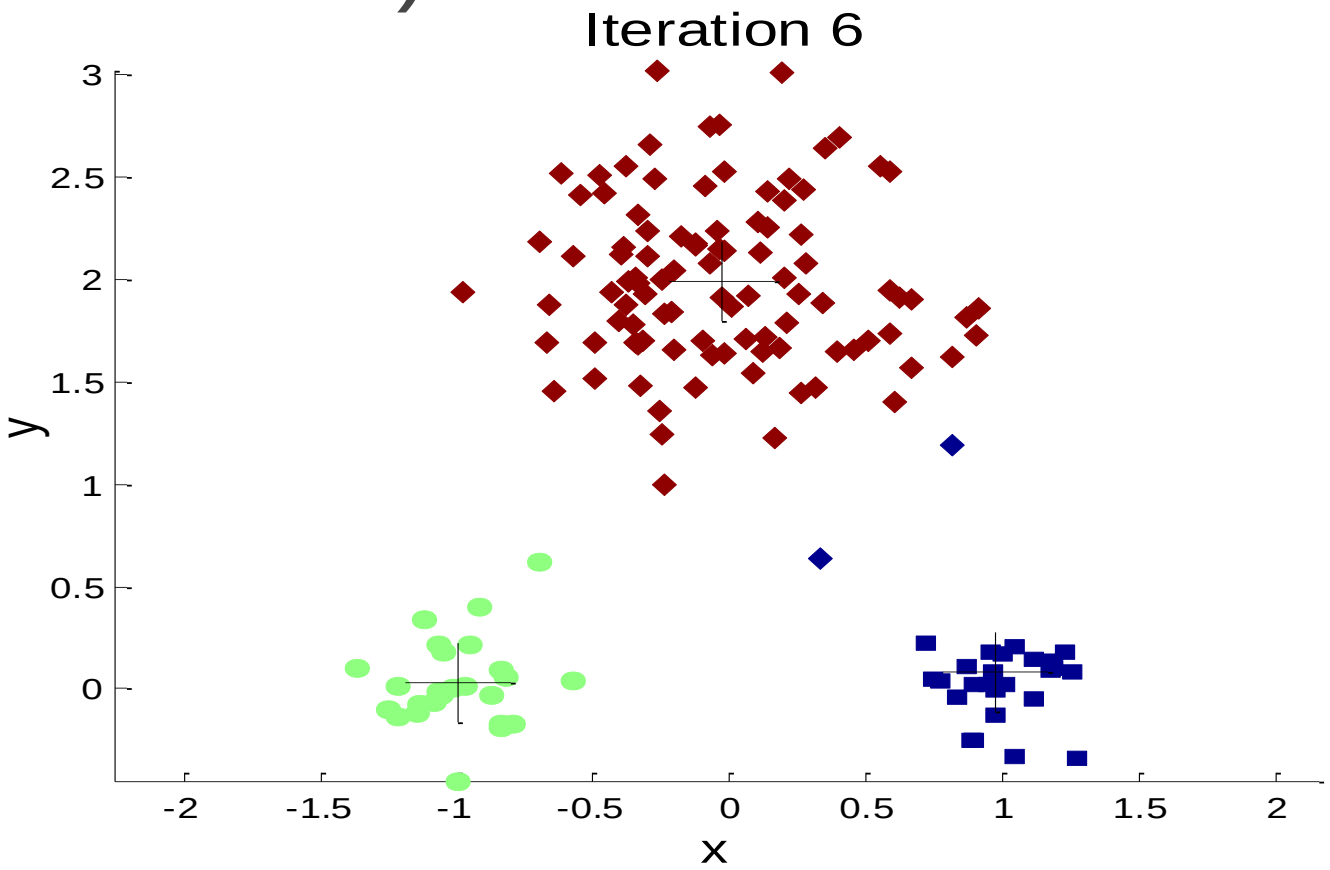


Βέλτιστη συσταδοποίηση

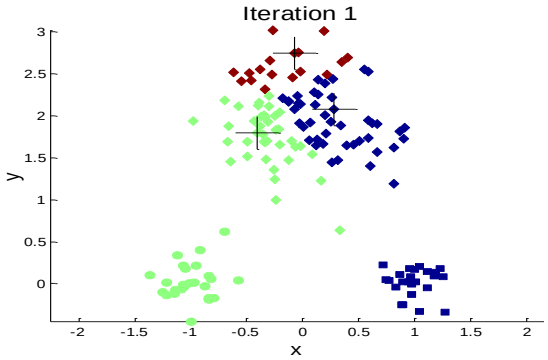


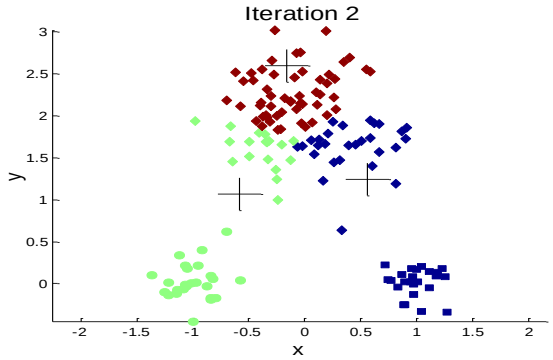
Υπό-βέλτιστη συσταδοποίηση

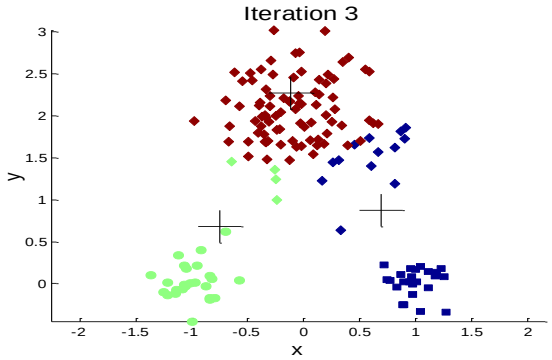
Η σημασία επιλογής αρχικών σημείων (centroids)

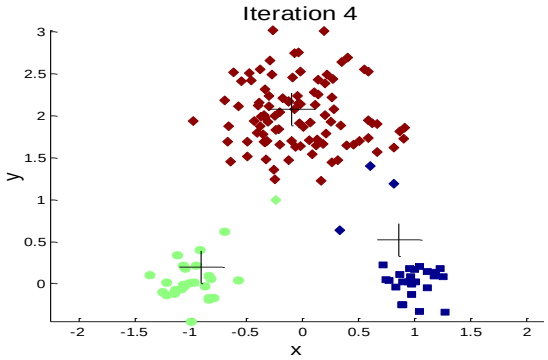


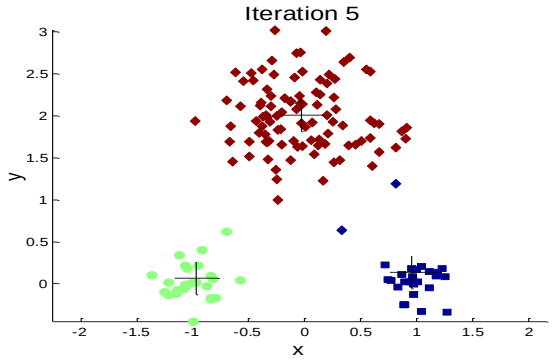
Η σημασία επιλογής αρχικών σημείων

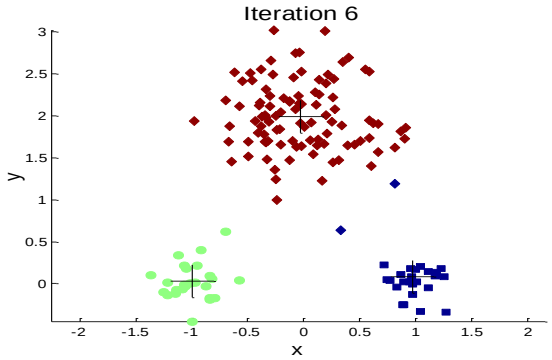




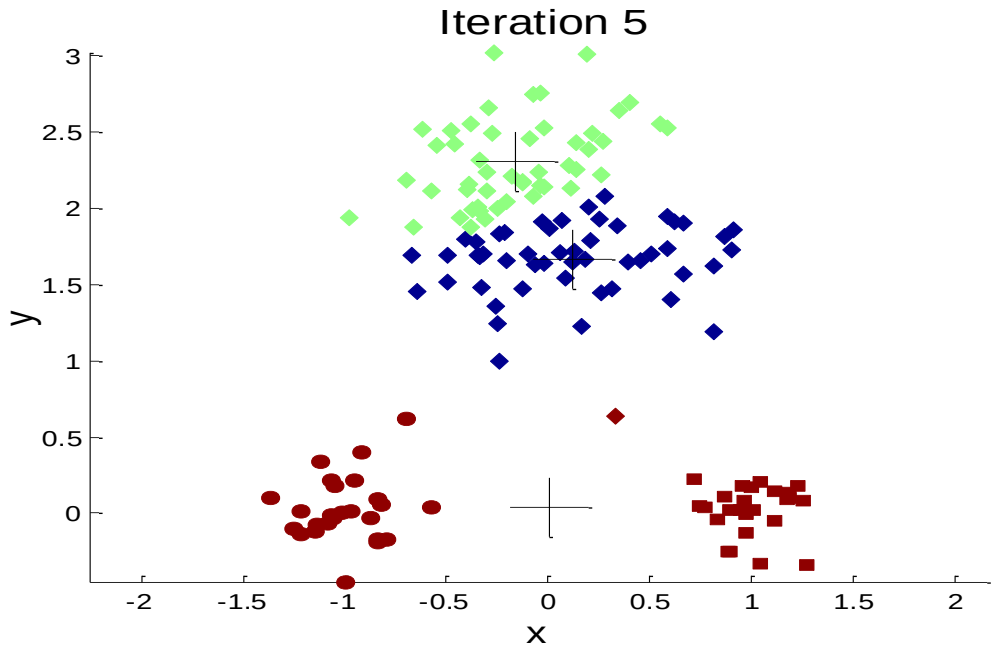




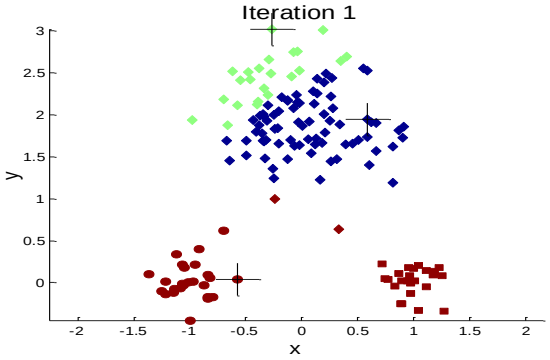


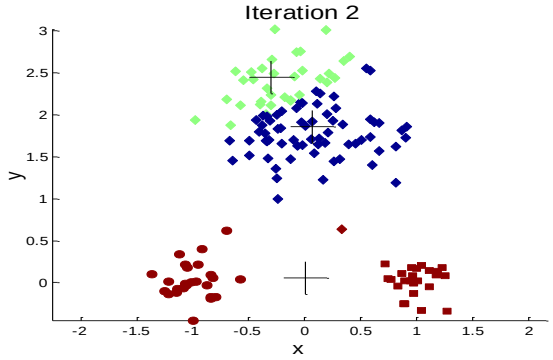


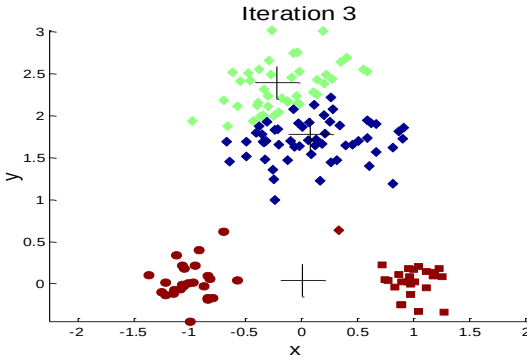
Η σημασία επιλογής αρχικών σημείων

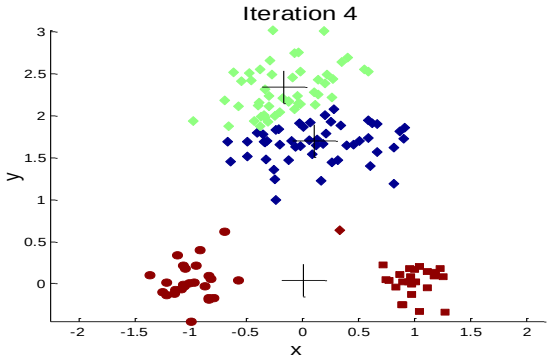


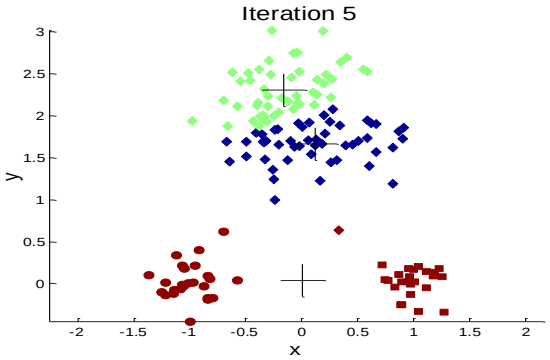
Η σημασία επιλογής αρχικών σημείων











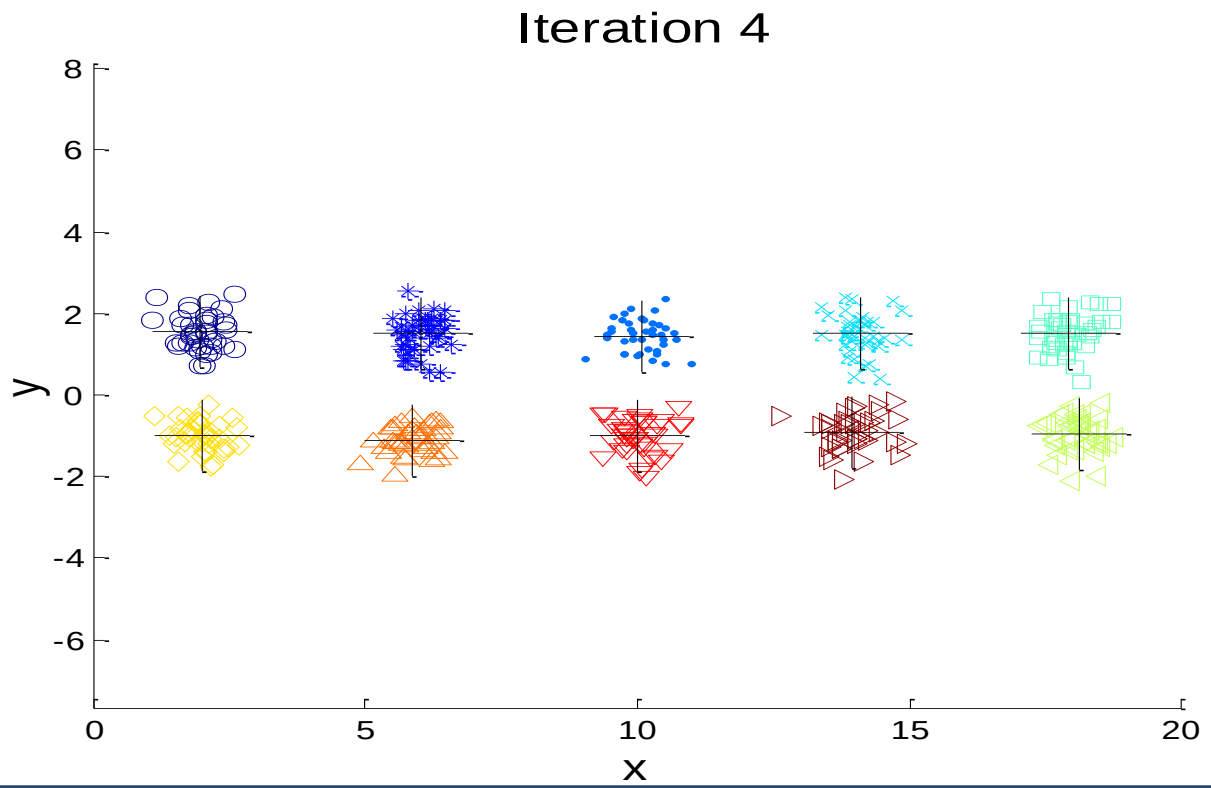
Προβλήματα επιλογής αρχικών σημείων

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

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

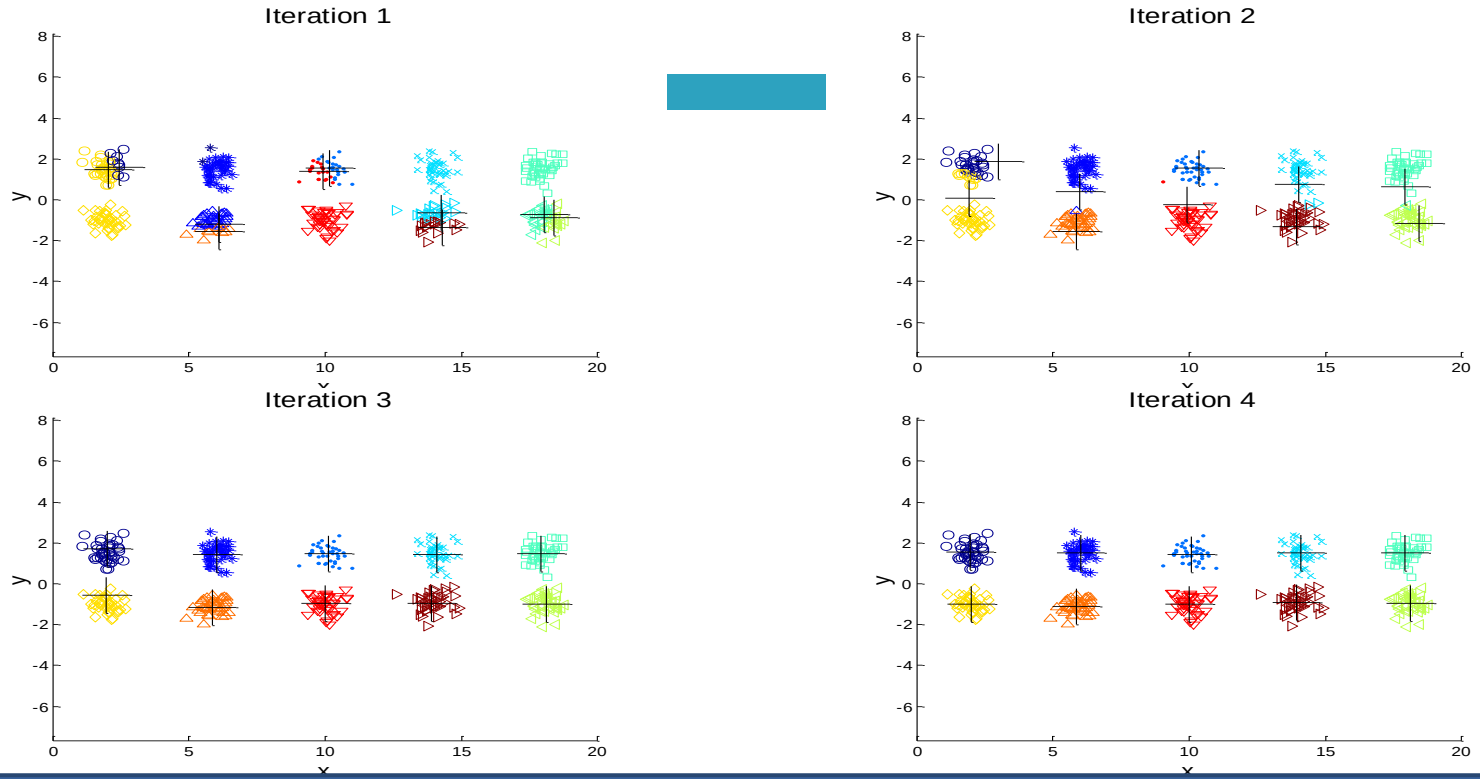
- Π.χ. αν K = 10, τότε P = $10!/10^{10}$ = 0.00036
  - Μερικές φορές τα αρχικά σημεία βελτιώνουν τη θέση τους και άλλες φορές όχι
  - Θα δούμε ένα παράδειγμα με 5 ζευγάρια συστάδων

Παράδειγμα 10 συστάδων



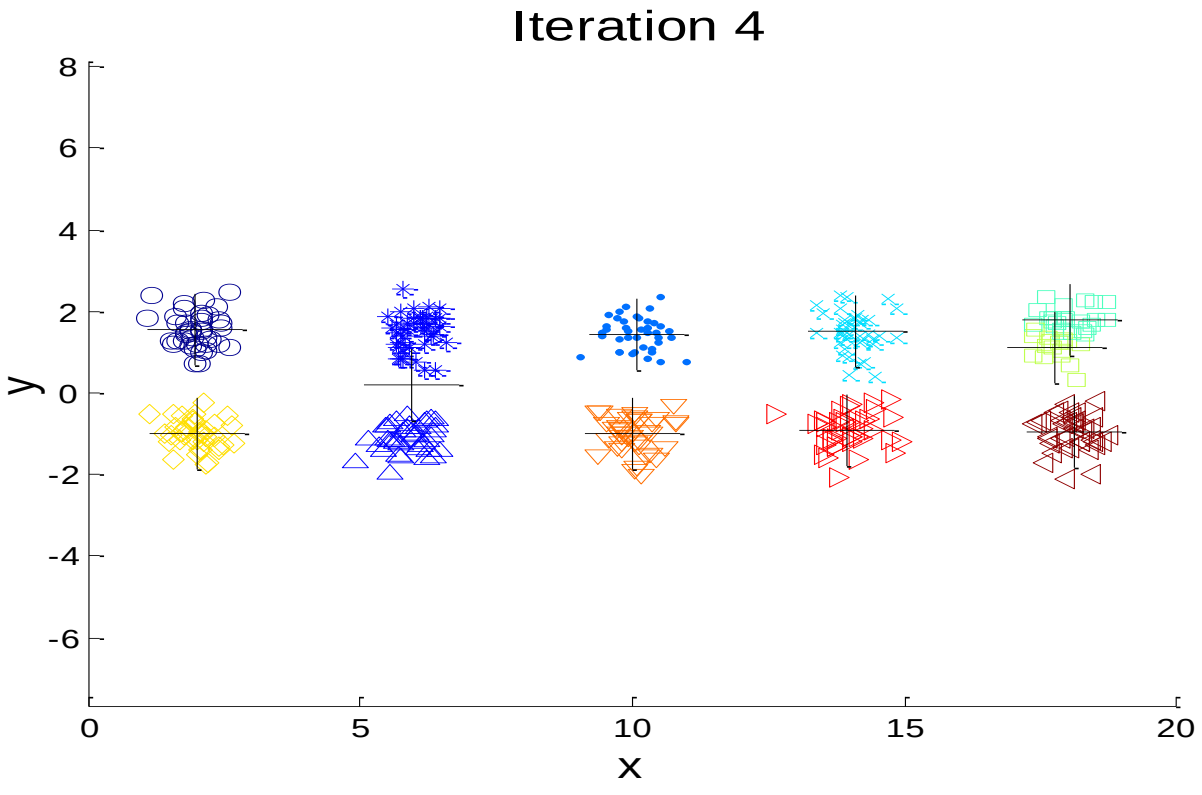
Ξεκινώντας με δύο αρχικά σημεία σε κάθε συστάδα κάθε ζεύγους συστάδων

Παράδειγμα 10 συστάδων



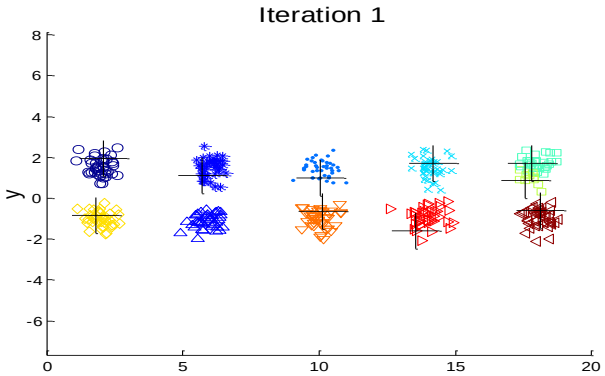
Ξεκινώντας με δύο αρχικά σημεία σε κάθε συστάδα κάθε ζεύγους συστάδων

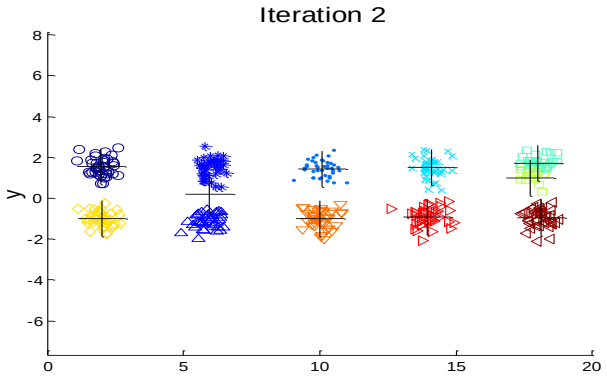
Παράδειγμα 10 συστάδων

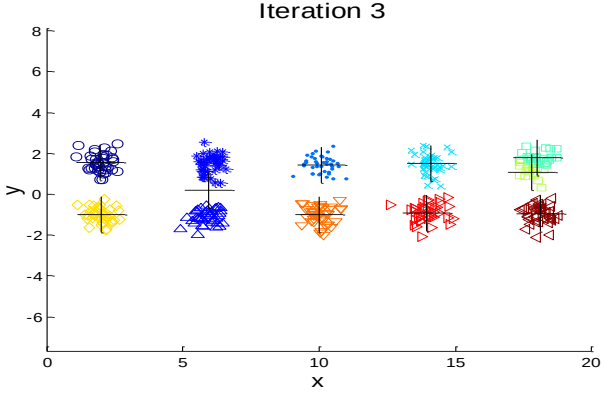


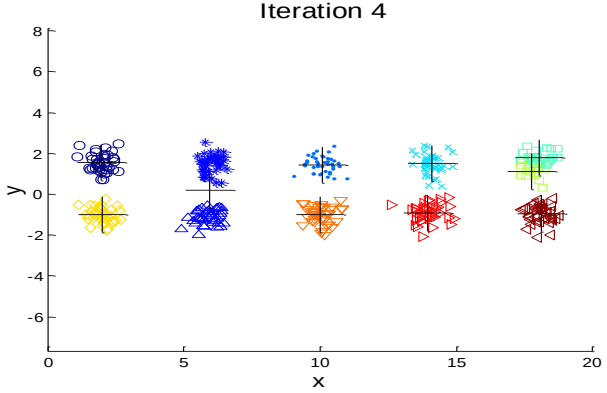
Ξεκινώντας με κάποια ζευγάρια συστάδων να έχουν τρία κεντρικά σημεία και άλλα μόνο ένα

Παράδειγμα 10 συστάδων









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

Προτεινόμενες λύσεις για την επιλογή των κεντρικών σημείων

- Πολλαπλά περάσματα του αλγόριθμου
  - Βοηθά, αλλά η πιθανότητες δεν είναι με το μέρος μας (πολλές περιπτώσεις)
- Δειγματοληψία και ιεραρχική συσταδοποίηση
- Επιλογή περισσότερων από k κεντρικών σημείων και επιλογή από αυτά
  - Επιλέγουμε τα πιο ευρέως διαχωρίσιμα
- 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 σε outliers

Αντιμετωπίζοντας τις άδεις 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 (Sum of squared error)
  - 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

Σταδιακή ενημέρωση των 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 Μετά- Επεξεργασία

- Προ-επεξεργασία
  - Κανονικοποίηση των δεδομένων
  - Απομάκρυνση των outliers
- Μετά-επεξεργασία
  - 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)

Διχοτόμηση του 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

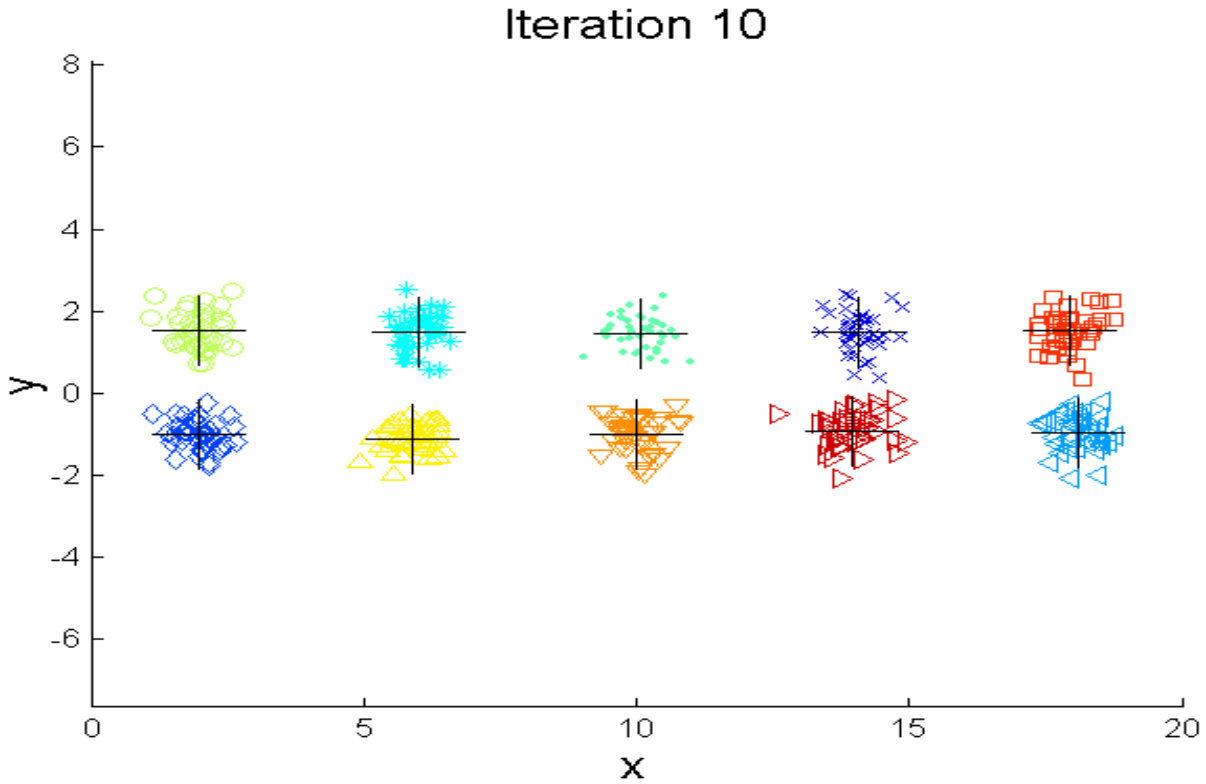
6: Πρόσθεσε στη λίστα από τις δυο συστάδες που προέκυψαν από τη διχοτόμηση αυτήν με το μικρότερο SSE

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

Διχοτόμηση του K-means

- Ποια συστάδα να διασπάσουμε;
  - Τη μεγαλύτερη
  - Αυτή με το μεγαλύτερο SSE
  - Συνδυασμό των παραπάνω
- Μπορεί να χρησιμοποιηθεί και ως ιεραρχικός

Διχοτόμηση του K-means: Παράδειγμα

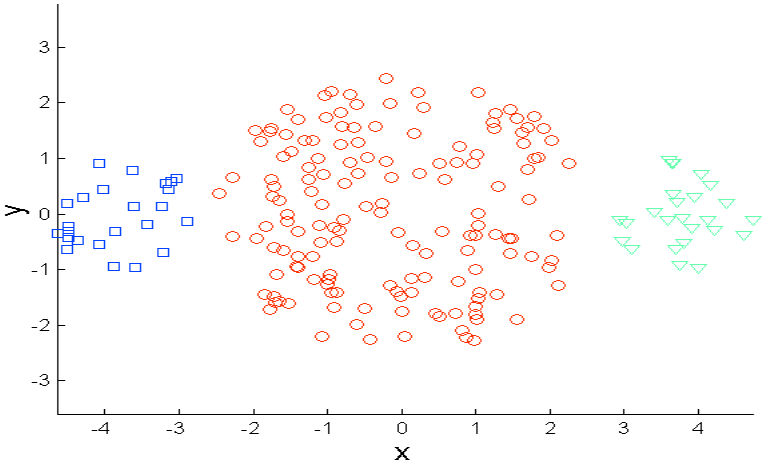


Περιορισμοί του K-means

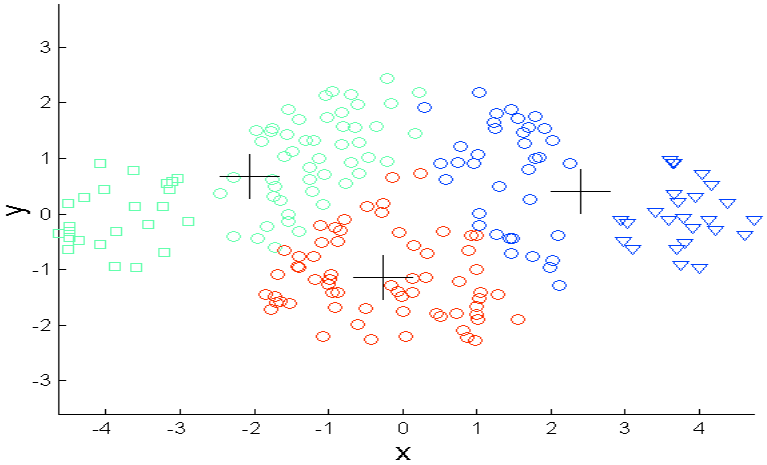
- Ο K-means έχει προβλήματα όταν οι συστάδες έχουν διαφορετικά
  - Διαφορετικά Μεγέθη
  - Διαφορετικές Πυκνότητες
  - Μη-σφαιροειδή σχήματα

- Έχει προβλήματα όταν τα δεδομένα έχουν outliers

Περιορισμοί K-means:διαφορετικά μεγέθη



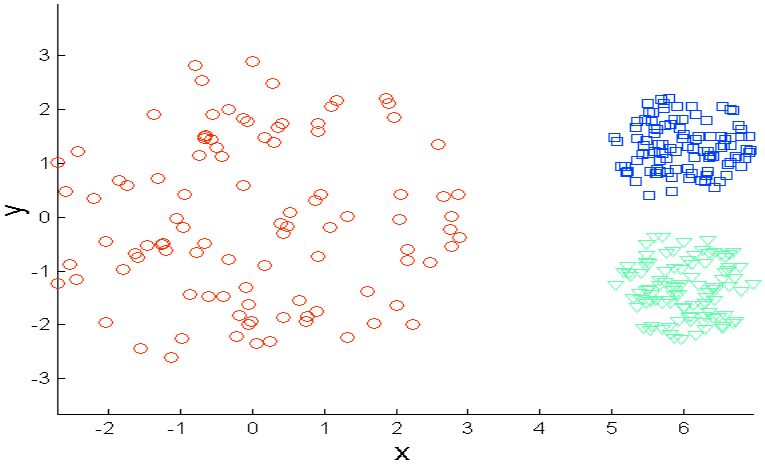
Αρχικά σημεία



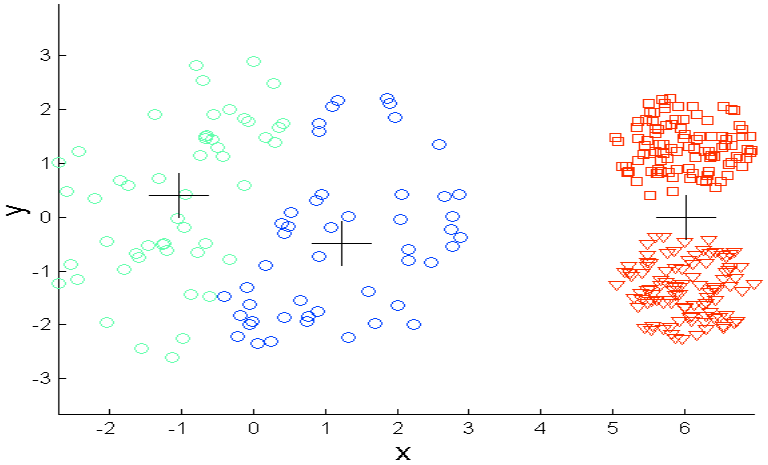
K-means (3 συστάδες)

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

Περιορισμοί K-means:διαφορετικές πυκνότητες



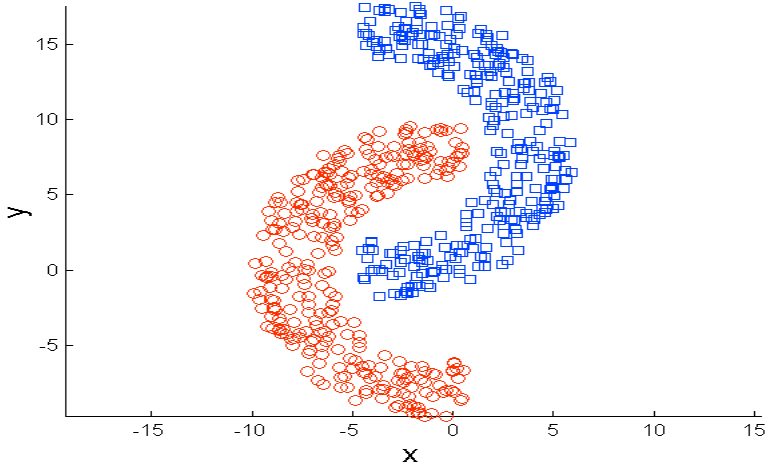
Αρχικά σημεία



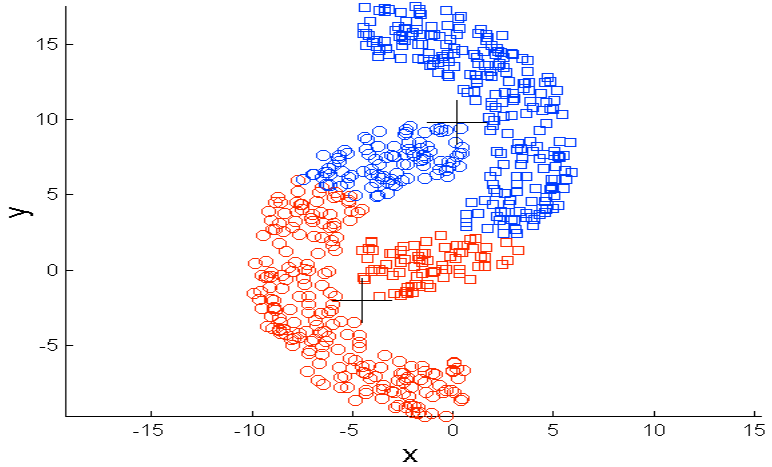
K-means (3 συστάδες)

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

Περιορισμοί K-means:μη-σφαιροειδή σχήματα



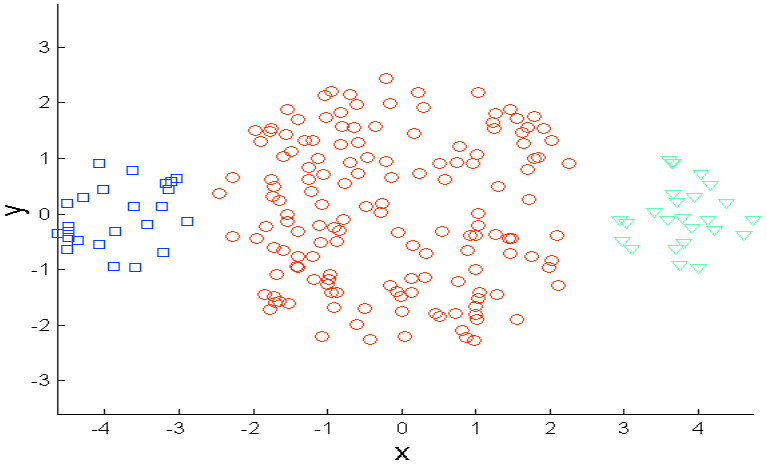
Αρχικά σημεία



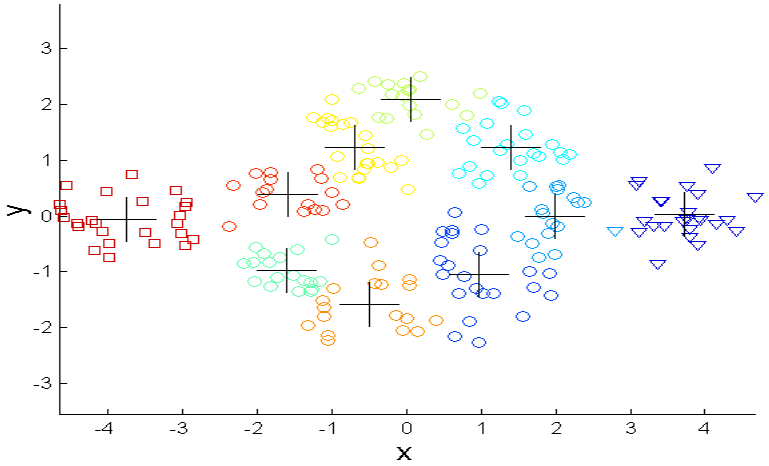
K-means (2 συστάδες)

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

Παρακάμπτοντας τα προβλήματα…



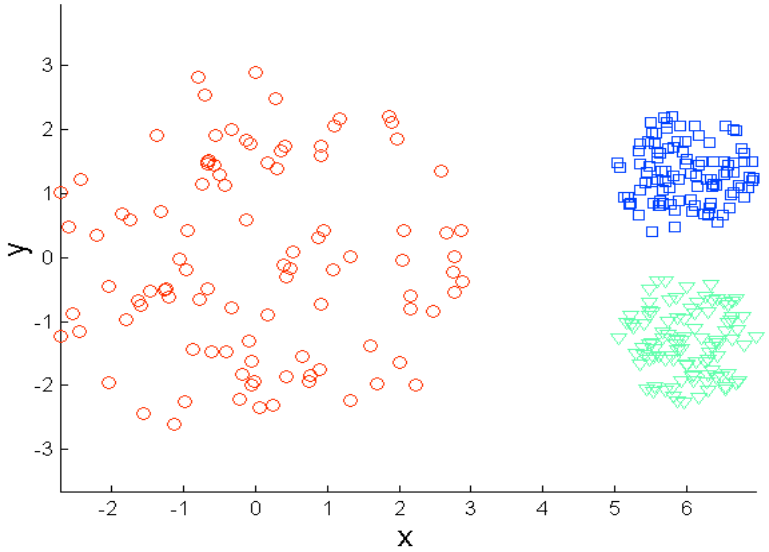
Αρχικά σημεία



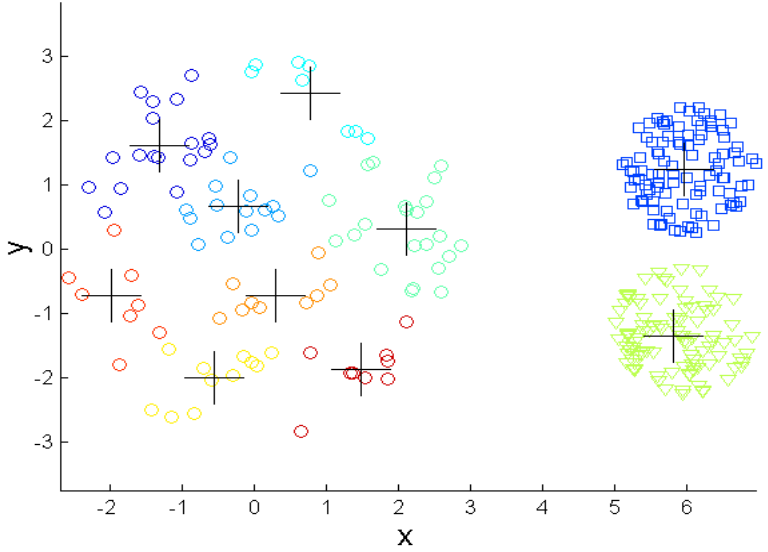
Συστάδες K-means

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

Παρακάμπτοντας τα προβλήματα… (Παράδειγμα)

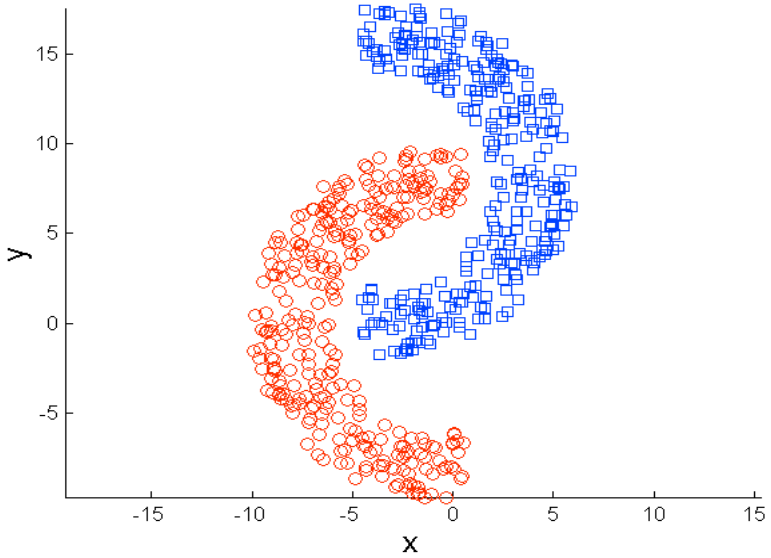


Αρχικά σημεία

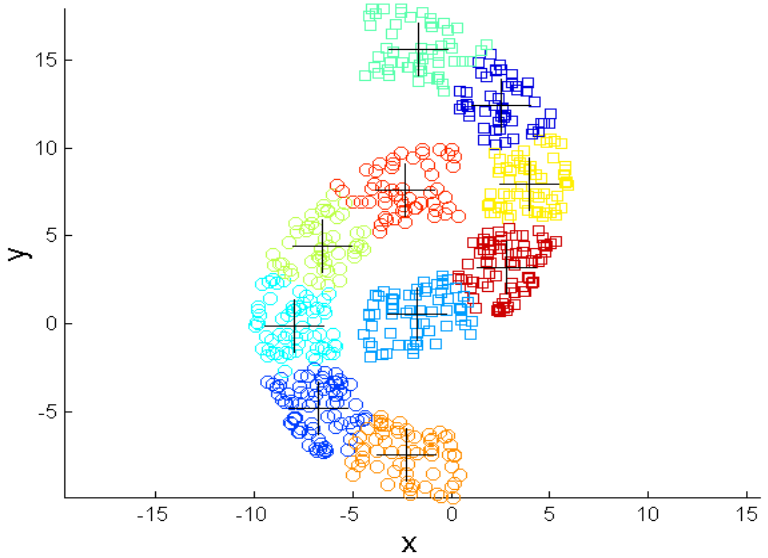


K-means συστάδες

Παρακάμπτοντας τα προβλήματα… (Παράδειγμα)



Αρχικά σημεία



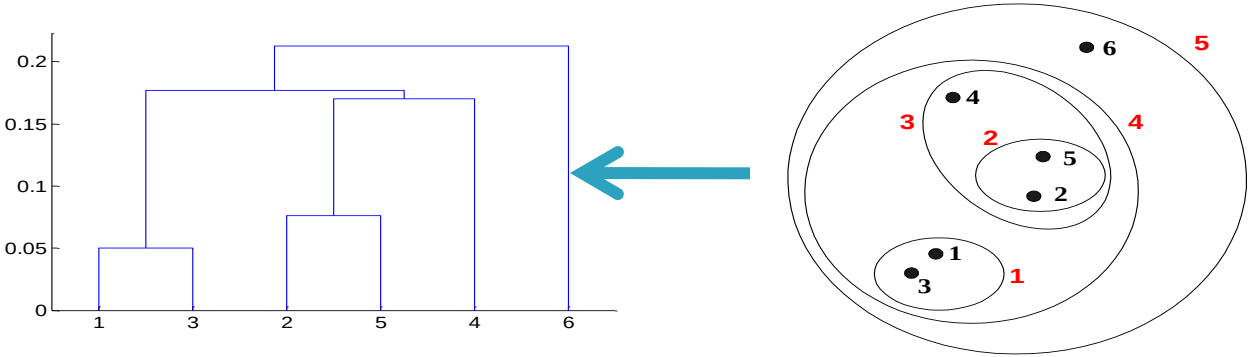
K-means συστάδες

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

- Παράγει ένα σύνολο εμφωλιασμένων συστάδων που οργανώνονται σαν ένα ιεραρχικό δέντρο
- Οπτικοποιούνται σαν ένα **δενδρό-γραμμα**
  - Ένα διάγραμμα που μοιάζει με δέντρο και και καταγράφει τις ακολουθίες από συγχωνεύσεις (merges) και διαχωρισμούς (splits)



Πλεονεκτήματα ιεραρχικής συσταδοποίησης

- 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 εύλογες ταξινομήσεις
  - Π.χ. από τη βιολογία
    - Ζωικό βασίλειο, φυλογένεση, …

Ιεραρχική Συσταδοποίηση

- 2 βασικά είδη
  - Συγκεντρωτικός (Agglomerative):
    - Ξεκινά με κάθε σημείο ως ξεχωριστή συστάδα
    - Σε κάθε βήμα, συγχωνεύει το ποιο κοντινό ζευγάρι συστάδων μέχρι να μείνει μόνο μία (ή k) συστάδες
  - Διαιρετικός (Divisive):
    - Ξεκινά με μια, περιεκτική συστάδα
    - Σε κάθε βήμα, διαιρείται η συστάδα έως κάθε συστάδα να περιέχει ένα σημείο (ή να υπάρχουν k συστάδες)
- Οι παραδοσιακοί αλγόριθμοι χρησιμοποιούν ένα πίνακα ομοιότητας ή απόστασης
  - διαχωρισμός ή συγχώνευση **μιας (1)** ομάδας τη φορά

Συγκεντρωτική Συσταδοποίηση

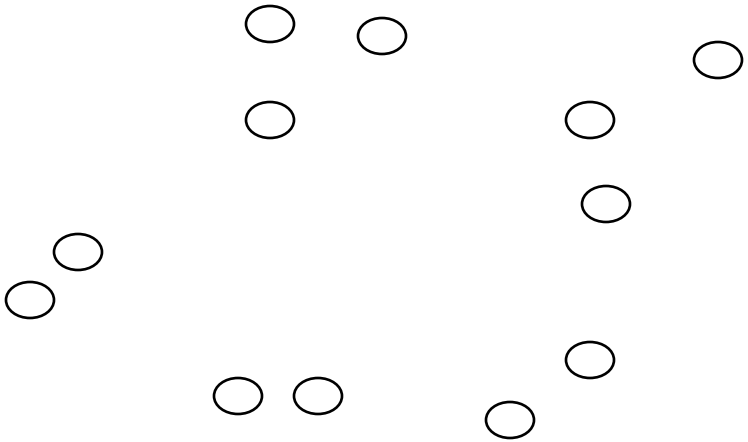
- Περισσότερο δημοφιλής
- Ο βασικός αλγόριθμος είναι σαφής:
  1. Υπολογισμός του πίνακα εγγύτητας
  2. Έστω κάθε σημείο μια συστάδα
  3. **Repeat**
  4. Συγχώνευση των δυο πλησιέστερων συστάδων
  5. Ανανέωση του πίνακα εγγύτητας
  6. **Until** να μείνει μόνο μια συστάδα
- Το σημείο-κλειδί είναι **ο υπολογισμός της εγγύτητας** δυο συστάδων
  - Ανάλογα με τον τρόπο υπολογισμού παρουσιάζονται και διαφορετικοί αλγόριθμοι

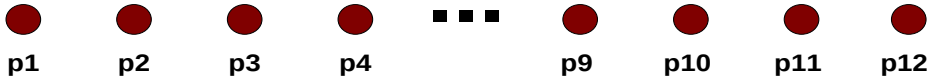
Αρχική Περίπτωση

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

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

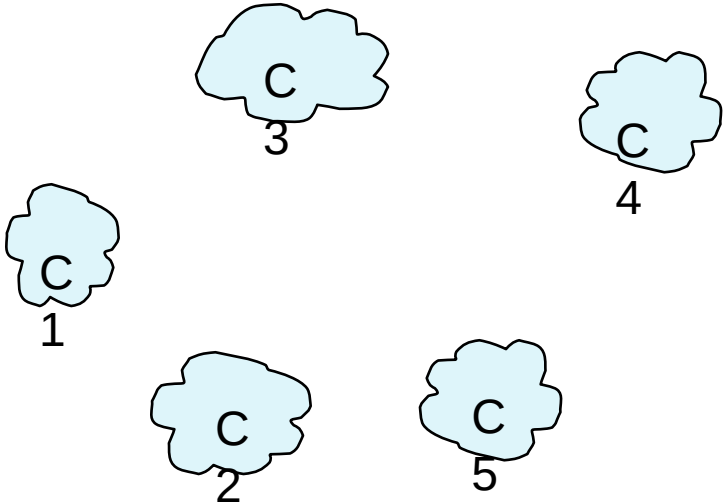
Πίνακας εγγύτητας ή γειτνίασης





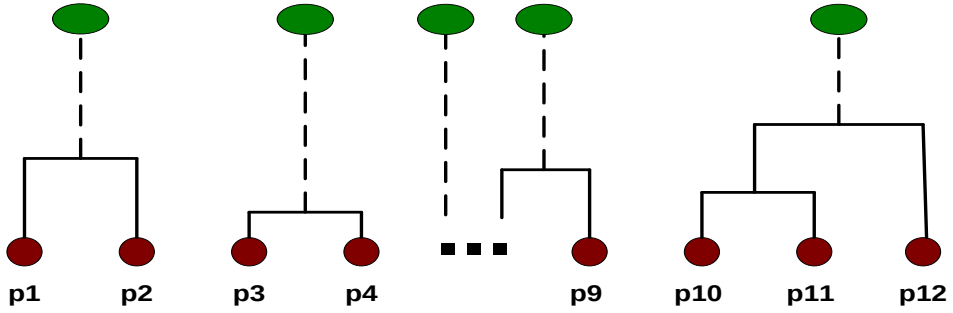
Ενδιάμεση περίπτωση

Έπειτα από ορισμένες συγχωνεύσεις παράγονται κάποιες συστάδες



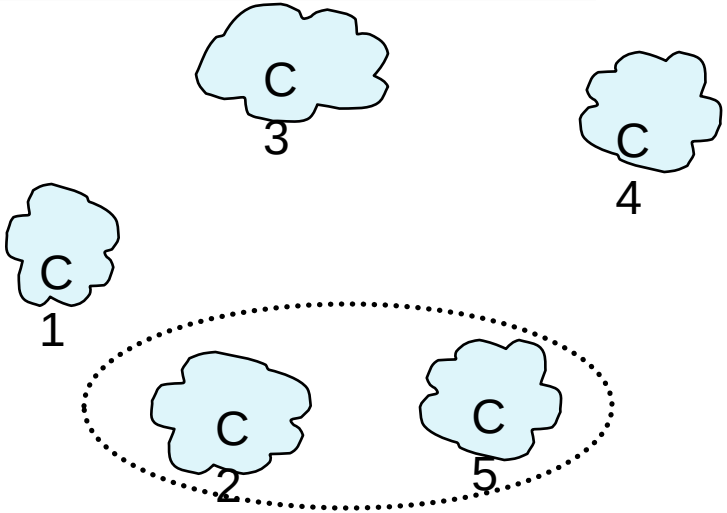
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

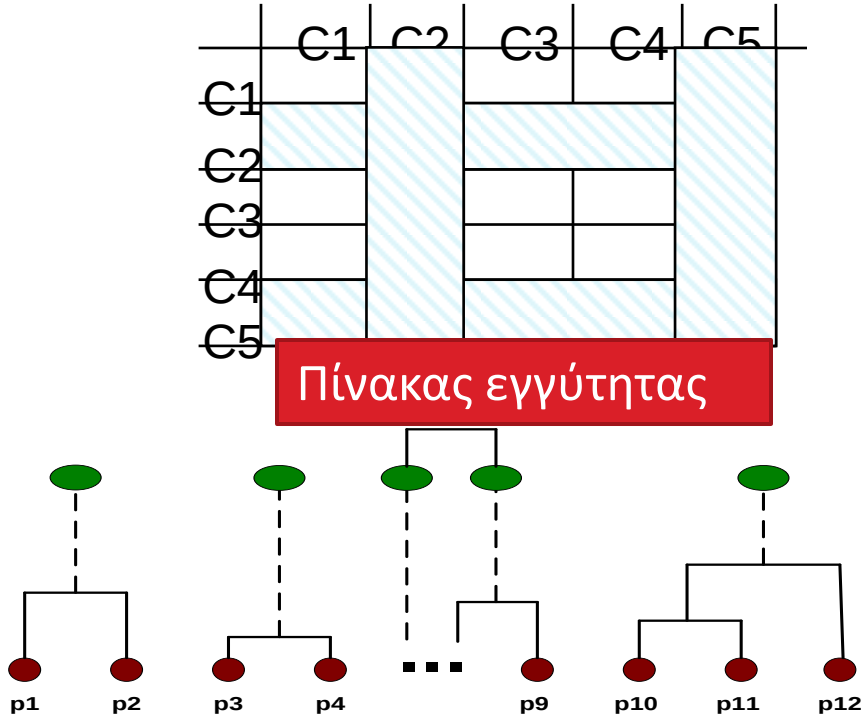
Πίνακας εγγύτητας



Ενδιάμεση περίπτωση

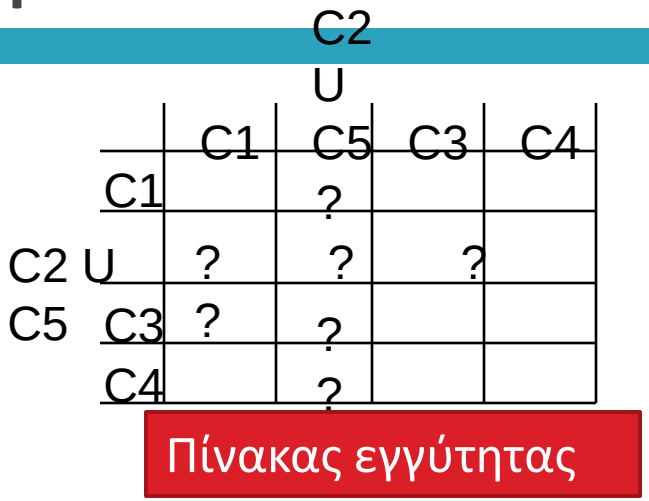
Θέλουμε να συγχωνεύσουμε τις δύο κοντινότερες συστάδες (C2 και C5) και να ενημερώσουμε τον πίνακα εγγύτητας.



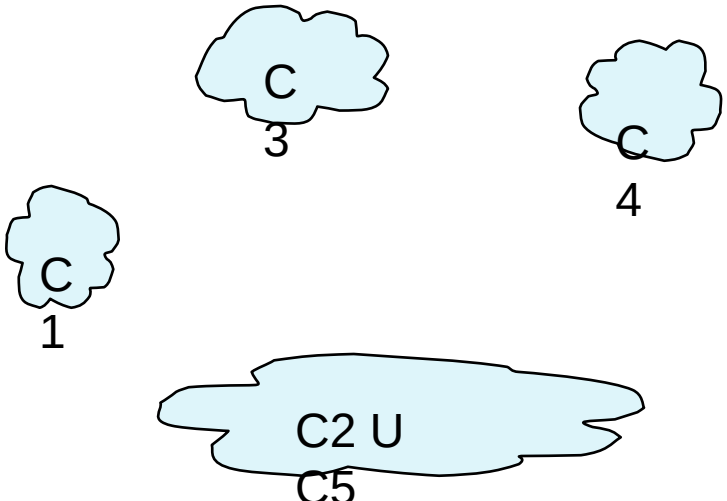


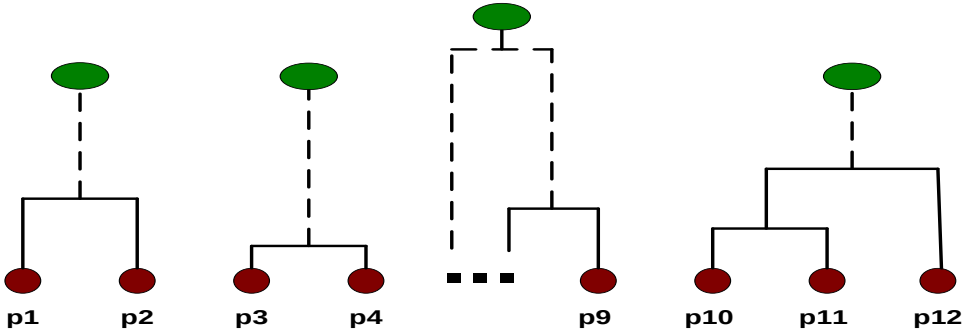
Μετά τη συγχώνευση

η ερώτηση είναι: Πως ενημερώνουμε τον πίνακα εγγύτητας;

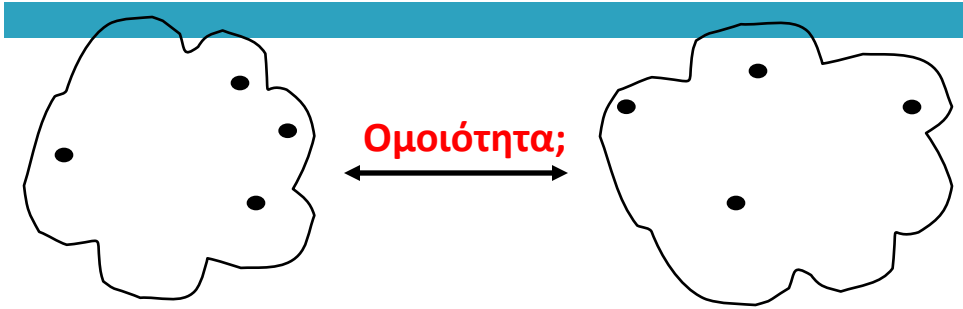


Πίνακας εγγύτητας





Ομοιότητα συστάδων

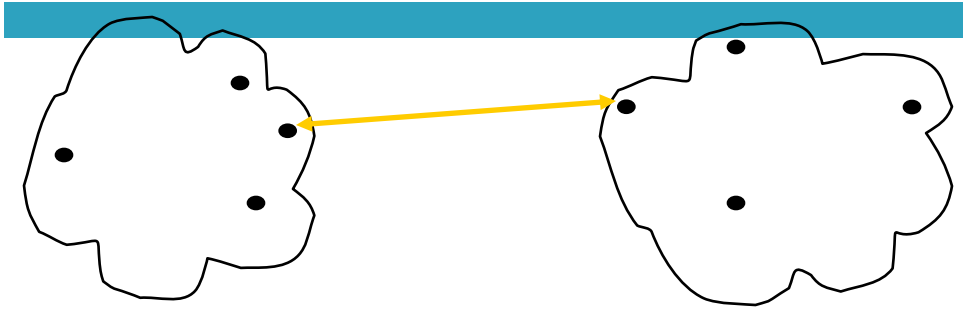


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Πίνακας Εγγύτητας

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

Ομοιότητα συστάδων

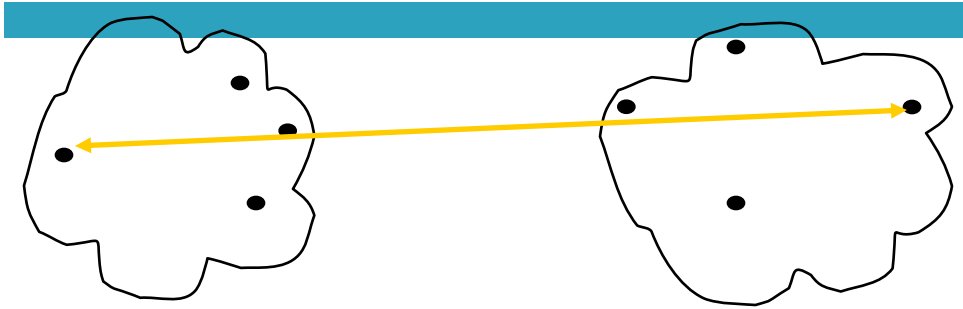


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Πίνακας Εγγύτητας

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

Ομοιότητα συστάδων

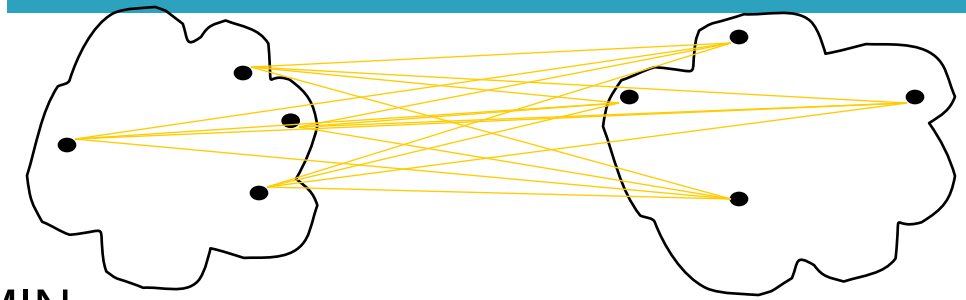


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Πίνακας Εγγύτητας

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

Ομοιότητα συστάδων

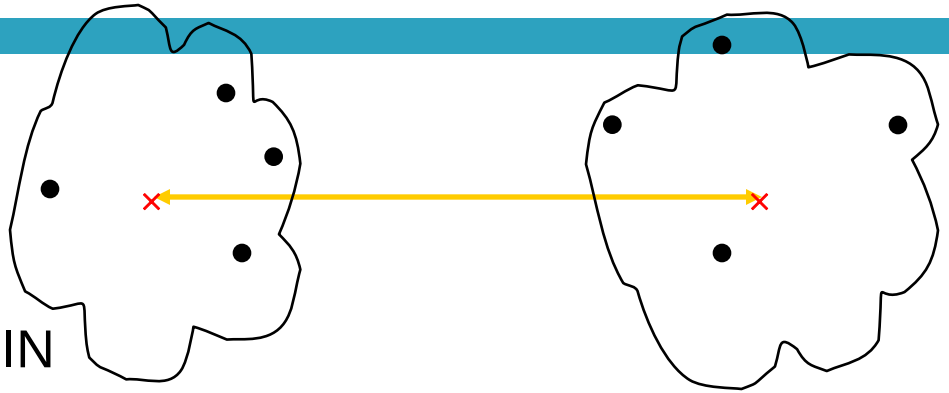


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Πίνακας Εγγύτητας

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

Ομοιότητα συστάδων



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

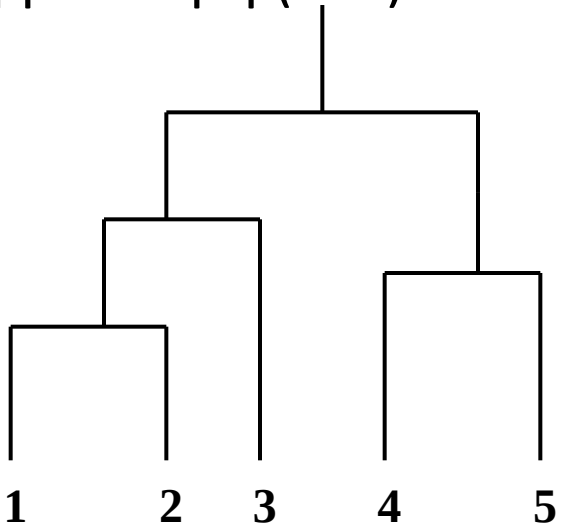
Πίνακας Εγγύτητας

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

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



Παράδειγμα

1 (0.4, 0.53)

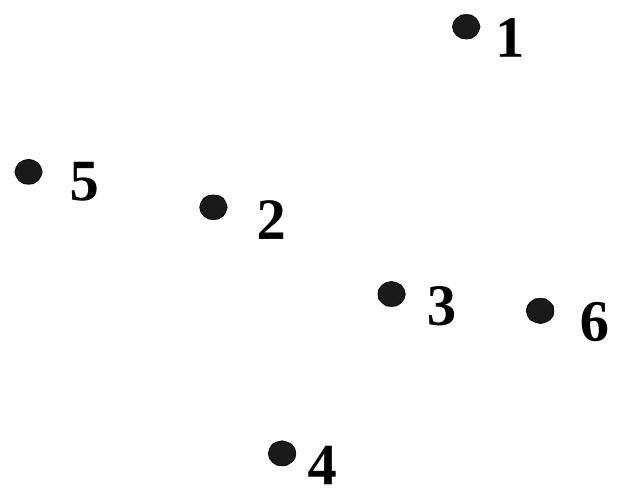
2 (0.22, 0.38)

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

Παράδειγμα

1 (0.4, 0.53)

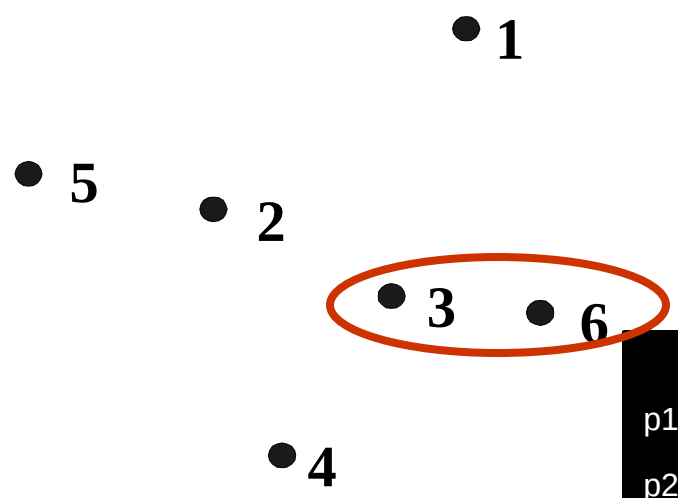
2 (0.22, 0.38)

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

Καθορίζεται μόνο από μια ακμή - την μικρότερη

Παράδειγμα

1 (0.4, 0.53)

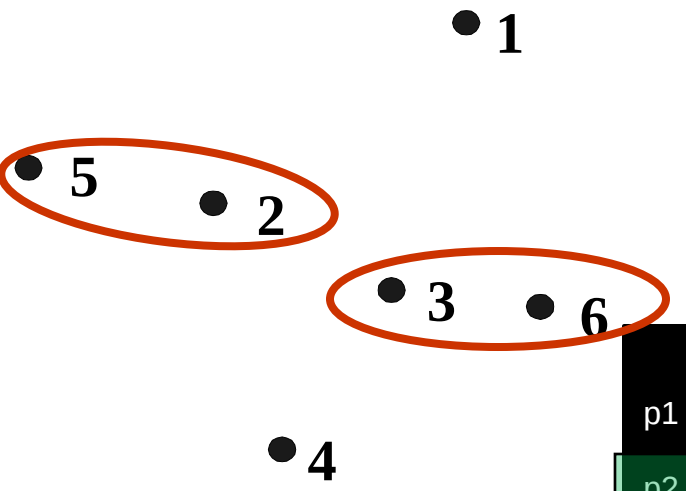
2 (0.22, 0.38)

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

Παράδειγμα

1 (0.4, 0.53)

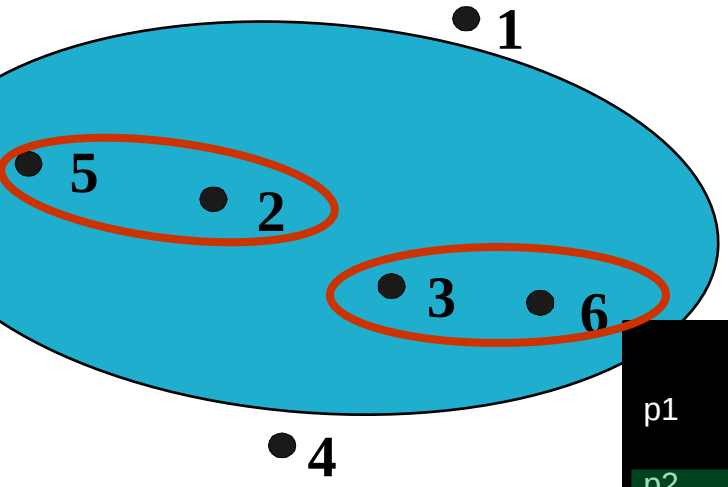
2 (0.22, 0.38)

3 (0.35, 0.32)

4 (0.26, 0.19)

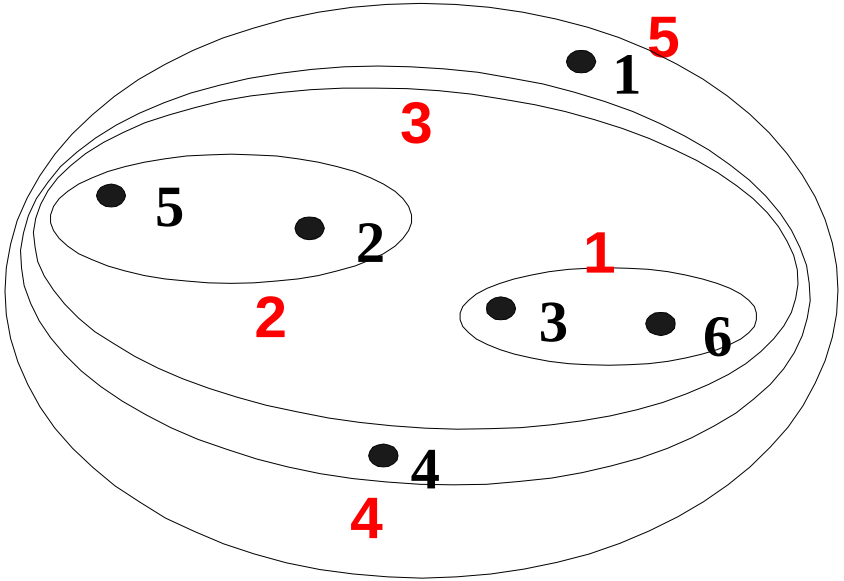
5 (0.08, 0.41)

6 (0.45, 0.30)

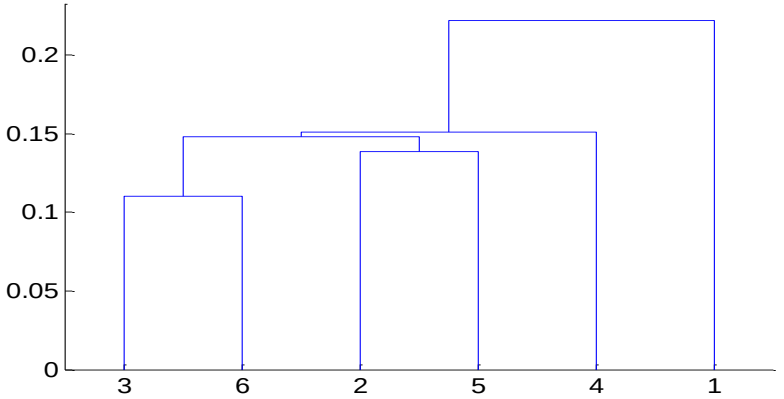


	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

Ορισμός απόστασης μεταξύ συστάδων: MIN



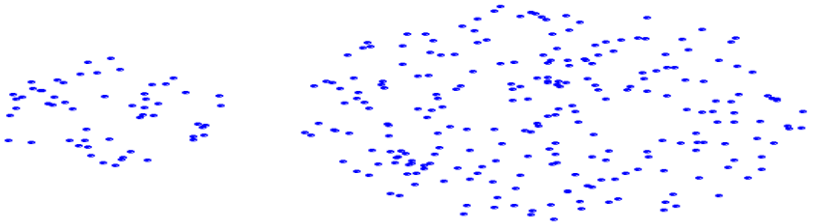
Εμφωλιασμένες συστάδες



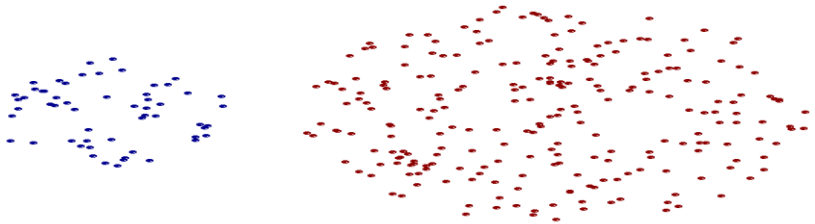
Dendrogram

Το δεντρόγραμμα δίνει και τις αποστάσεις

Πλεονεκτήματα του MIN



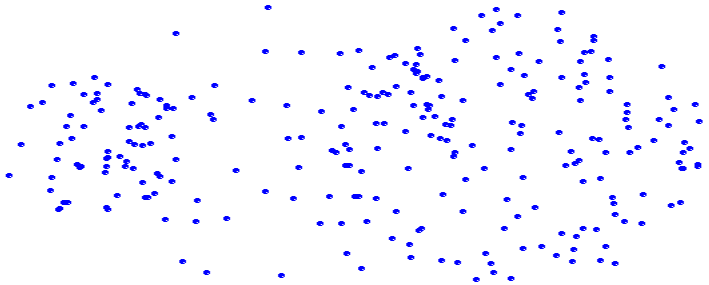
Αρχικά Σημεία



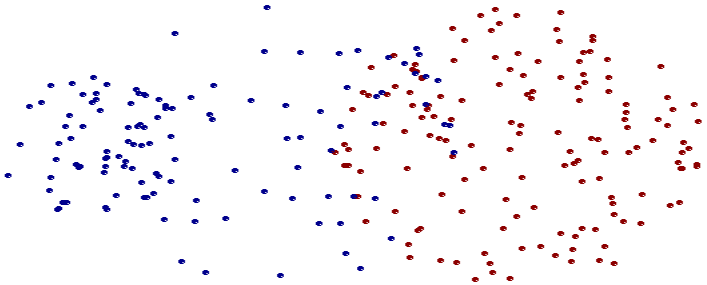
2 Συστάδες

- Μπορεί να αντιμετωπίσει μη-ελλειπτικά δεδομένα

Μειονεκτήματα του MIN



Αρχικά Σημεία



2 Συστάδες

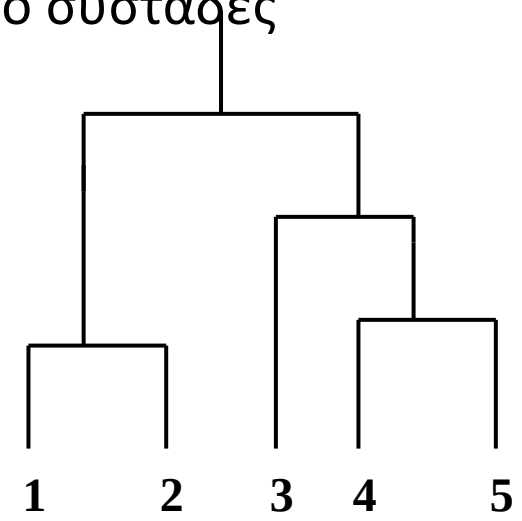
- Ευαίσθητος στο Θόρυβο και στα Ακραία Σημεία

MAX ή πλήρους συνδεσιμότητας (complete linkage)

- Η ομοιότητα μεταξύ δυο συστάδων βασίζεται στα δυο λιγότερο όμοια (πιο 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



Παράδειγμα

1 (0.4, 0.53)

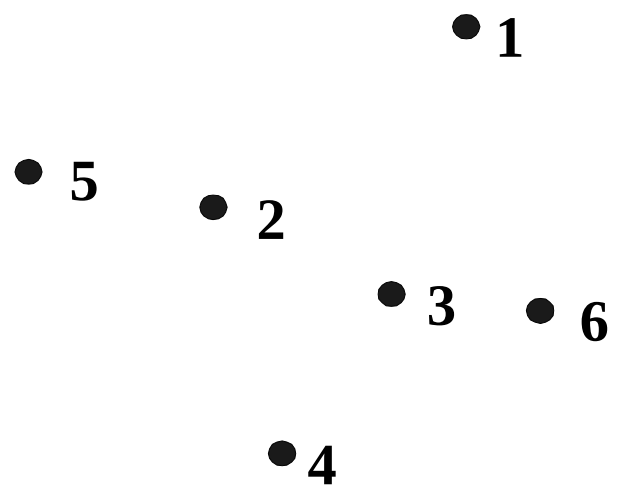
2 (0.22, 0.38)

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

Παράδειγμα

1 (0.4, 0.53)

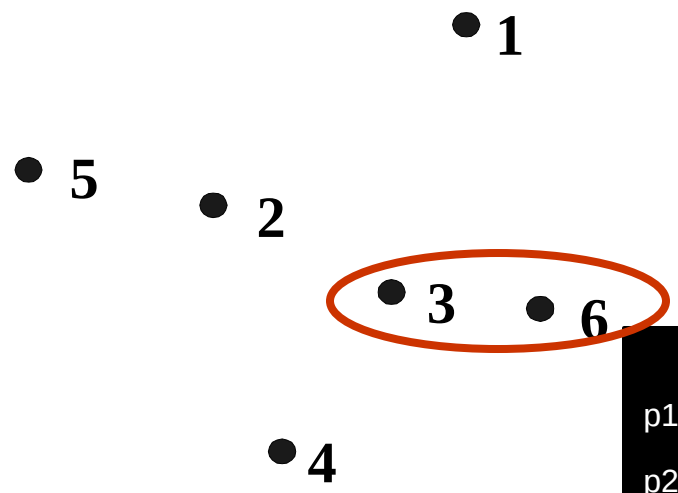
2 (0.22, 0.38)

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**

Παράδειγμα

1 (0.4, 0.53)

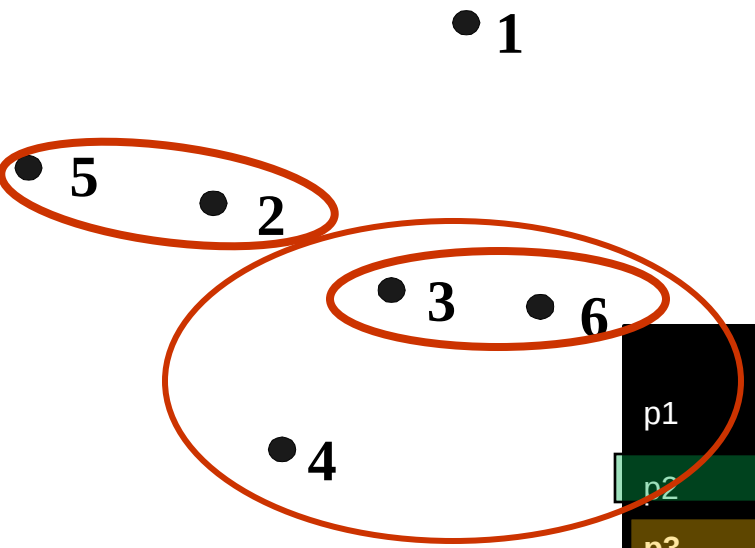
2 (0.22, 0.38)

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

Παράδειγμα

1 (0.4, 0.53)

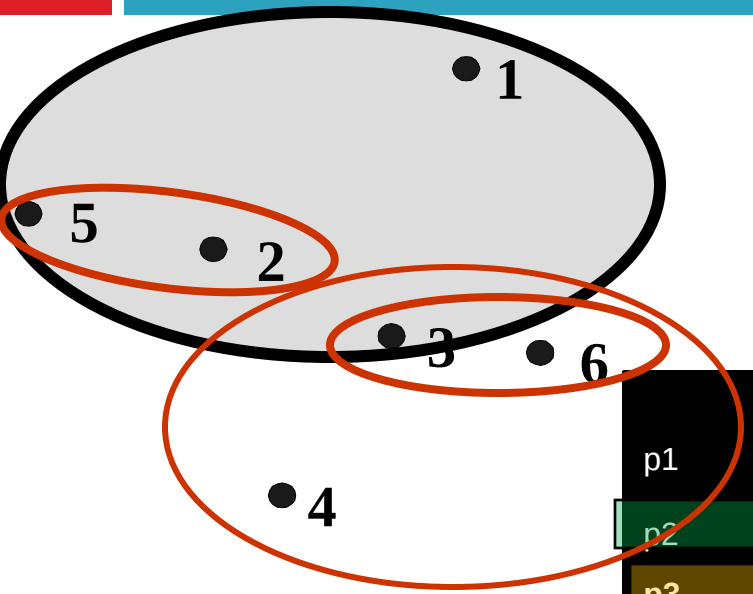
2 (0.22, 0.38)

3 (0.35, 0.32)

4 (0.26, 0.19)

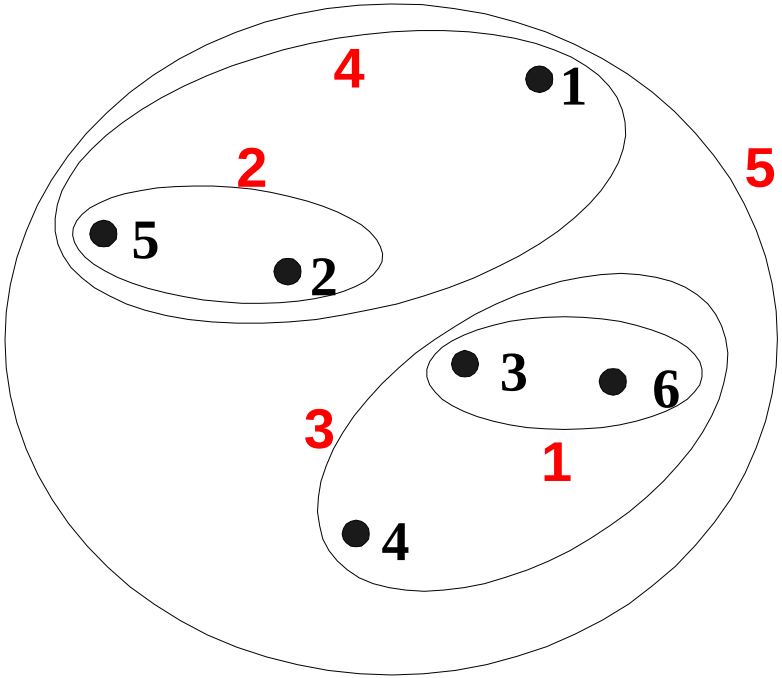
5 (0.08, 0.41)

6 (0.45, 0.30)

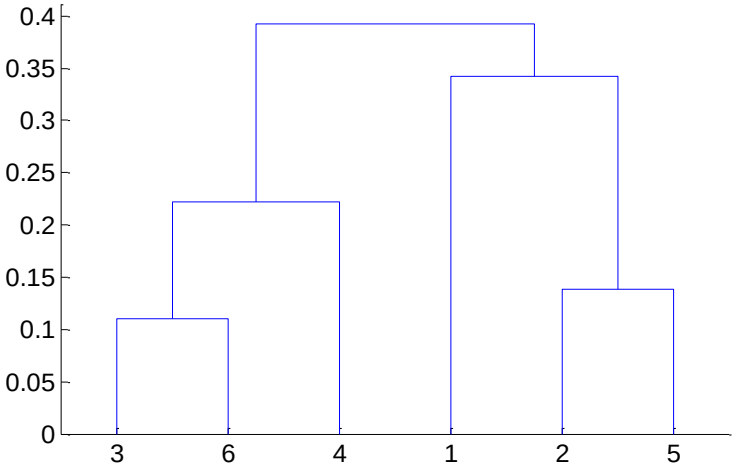


	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	**0.25**
p3	0.22	0.15	0.00	0.15	0.28	**0.11**
p4	**0.37**	0.20	0.15	0.00	0.29	0.22
p5	**0.34**	0.14	**0.28**	**0.29**	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

Ορισμός απόστασης μεταξύ συστάδων: MAX

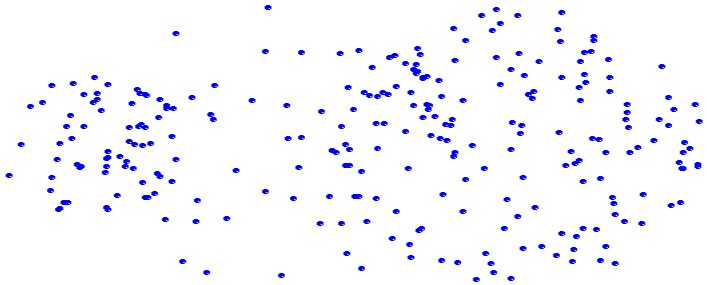


Εμφωλιασμένες συστάδες

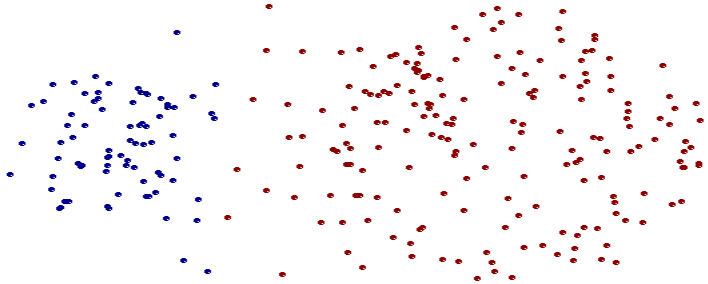


Dendrogram

Πλεονεκτήματα του MAX



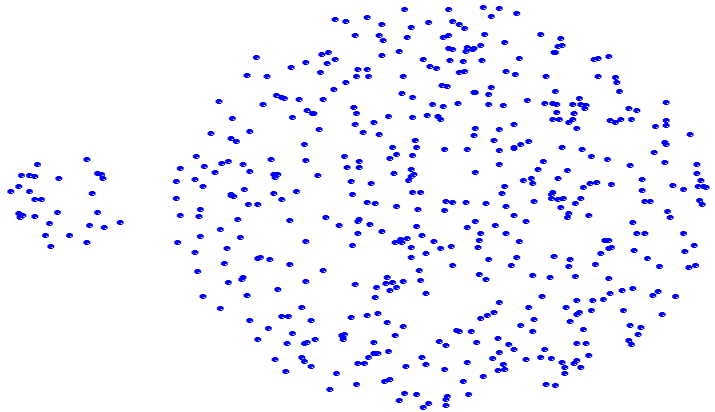
Αρχικά σημεία



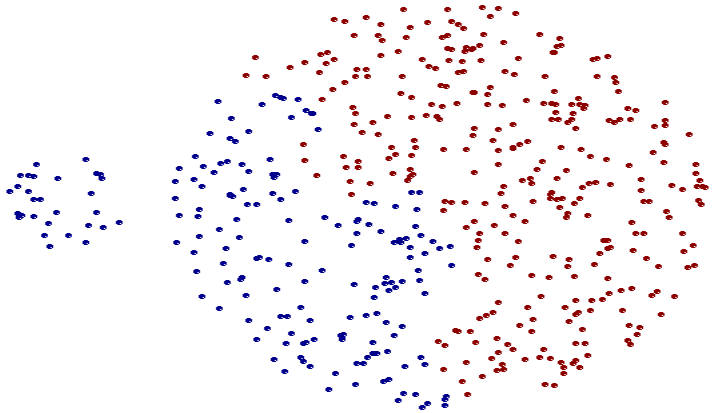
2 Συστάδες

- Λιγότερο ευαίσθητος στο θόρυβο

Μειονεκτήματα του MAX



Αρχικά Σημεία



2 Συστάδες

- Τείνει να διασπά μεγάλες συστάδες
- Προτιμά κυκλικές συστάδες

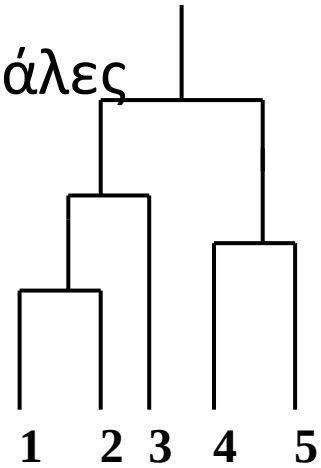
Μέσος όρος της ομάδας

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

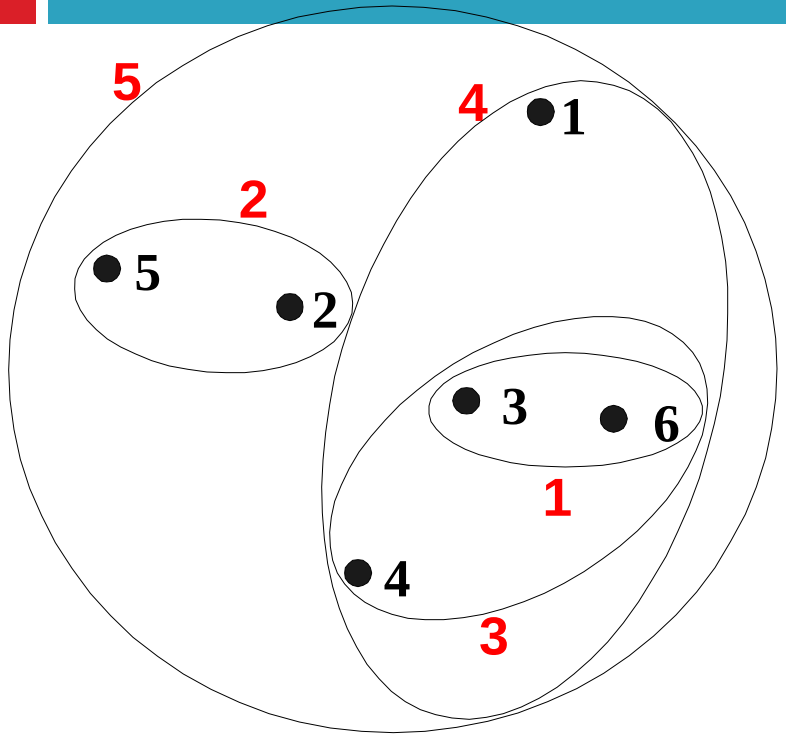
$$\text{proximity}(\text{Cluster}_i, \text{Cluster}_j) = \frac{\sum_{\substack{p_i \in \text{Cluster}_i \\ p_j \in \text{Cluster}_j}} \text{proximity}(p_i, p_j)}{|\text{Cluster}_i| * |\text{Cluster}_j|}$$

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

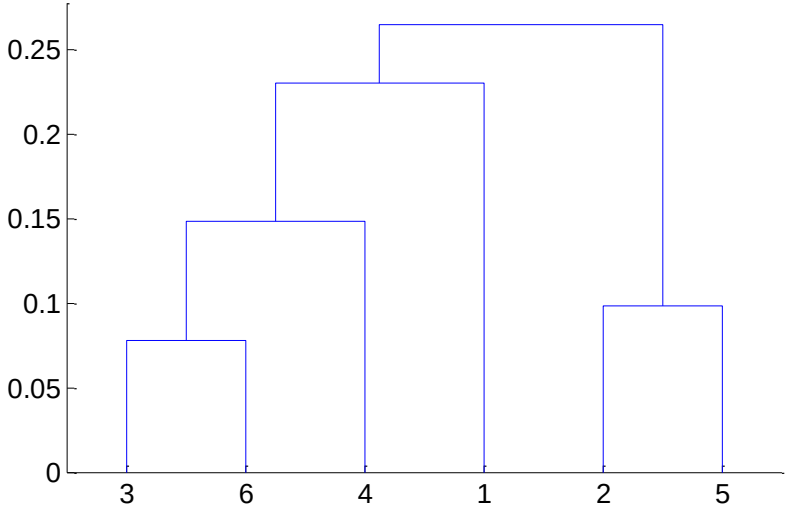
	I1	I2	I3	I4	I5
I1	1.00	0.90	0.10	0.65	0.20
I2	0.90	1.00	0.70	0.60	0.50
I3	0.10	0.70	1.00	0.40	0.30
I4	0.65	0.60	0.40	1.00	0.80
I5	0.20	0.50	0.30	0.80	1.00



Παράδειγμα



Εμφωλιασμένες συστάδες

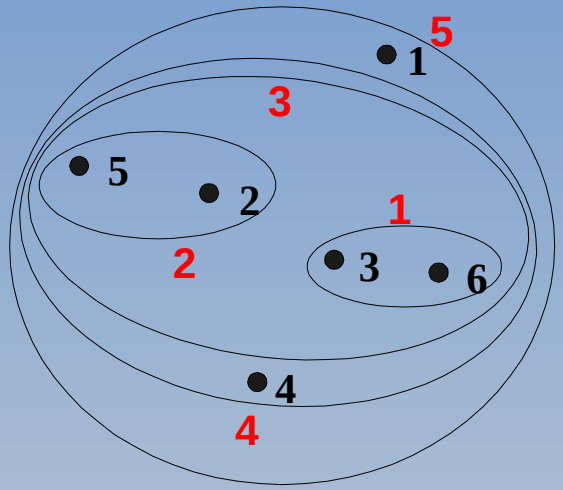


Dendrogram

Μέθοδος του Ward

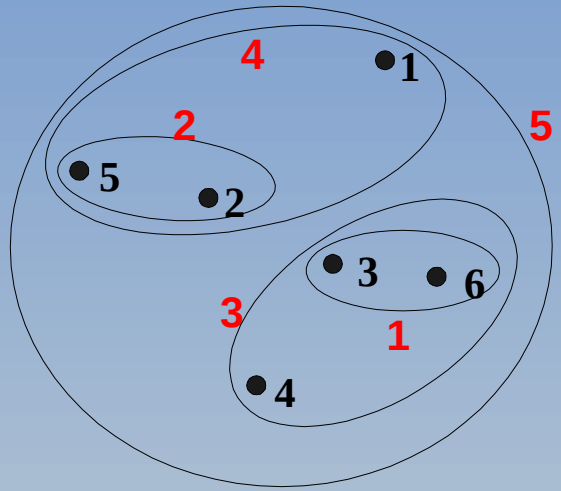
- Βασισμένο στην αύξηση του SSE όταν συγχωνεύονται οι δύο συστάδες
- Ιεραρχικό ανάλογο του k-means
- Μπορεί να χρησιμοποιηθεί για την αρχικοποίηση του k-means

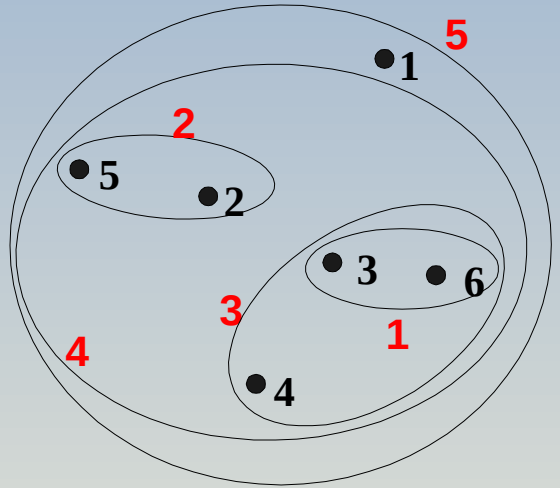
ΣΥΓΚΡΙΣΗ



MIN

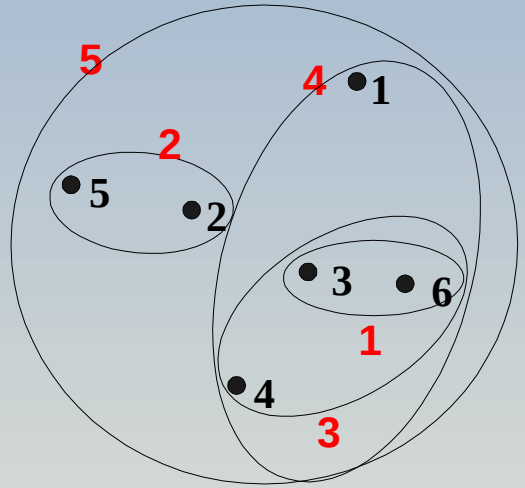
MAX





Μέσος όρος ομάδας

Μέθοδος του Ward



Συσσωρευτική ιεραρχική συσταδοποίηση: πολυπλοκότητα και περιορισμοί

- $O(m^2)$ 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 $O(m^2)$
- Περιορισμοί
  - Οι αποφάσεις είναι τελικές – αφού δυο συστάδες συγχωνευτούν αυτό δεν μπορεί να αλλάξει
  - Δεν ελαχιστοποιούν άμεσα κάποια αντικειμενική συνάρτηση

MST: Διαιρετική ιεραρχική συσταδοποίηση

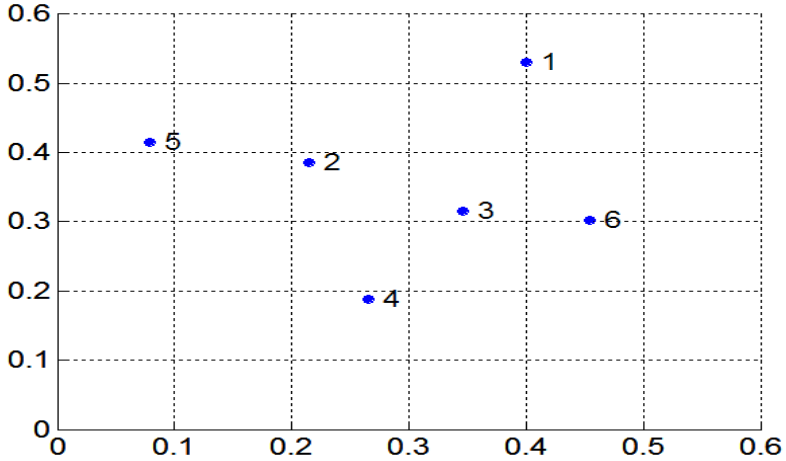
- Χρησιμοποιείστε τον MST για ιεραρχίες συστάδων

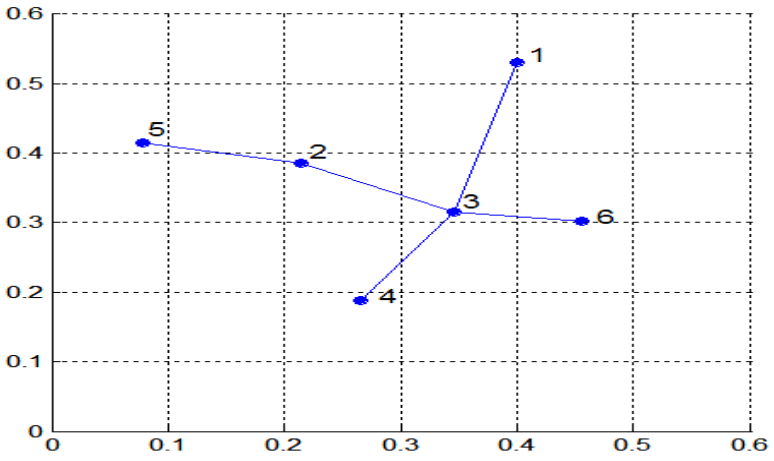
Algorithm 7.5 MST Divisive Hierarchical Clustering Algorithm

1: Compute a minimum spanning tree for the proximity graph.
2: **repeat**
3: Create a new cluster by breaking the link corresponding to the largest distance (smallest similarity).
4: **until** Only singleton clusters remain

MST: Divisive Hierarchical Clustering

- MST (Minimum Spanning Tree)
  - Ξεκίνα με ένα δέντρο που αποτελείται από οποιοδήποτε σημείο
  - Σε διαδοχικά βήματα, βρες το κοντινότερο ζεύγος σημείων (p, q) τέτοιο ώστε το (p) να είναι στο τρέχον δέντρο και το (q) όχι
  - Πρόσθεσε το q στο δέντρο και μια ακμή μεταξύ p και q

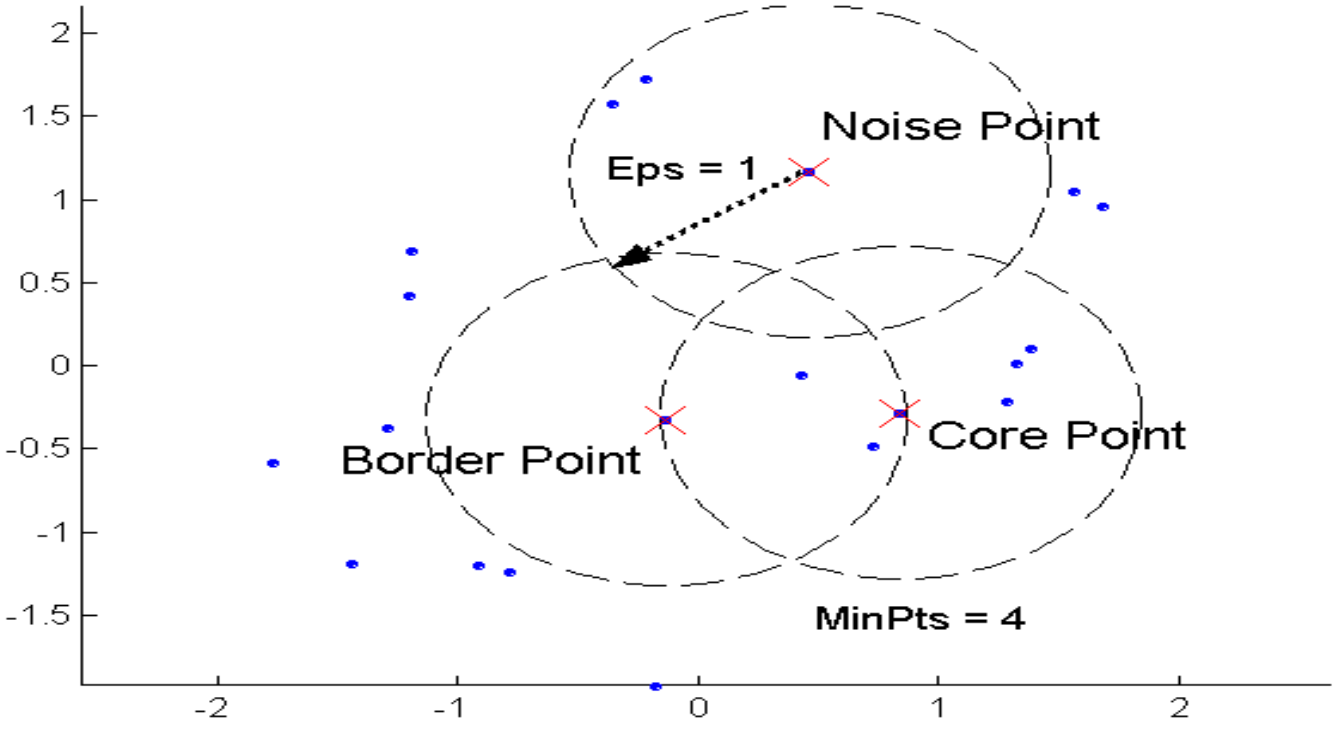




Ο αλγόριθμος DBSCAN

- Ο DBSCAN βασίζεται στην πυκνότητα.
  - Πυκνότητα= αριθμός των σημείων μέσα σε συγκεκριμένη ακτίνα (Eps)
- Κατηγορίες σημείων
  - **Βασικό** (core point) εάν υπάρχουν περισσότερα από ένα προκαθορισμένο αριθμό (MinPts) σημεία σε ακτίνα Eps
    - Είναι εσωτερικά σε μια συστάδα
  - **Οριακό** (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: Core, Border, και Noise



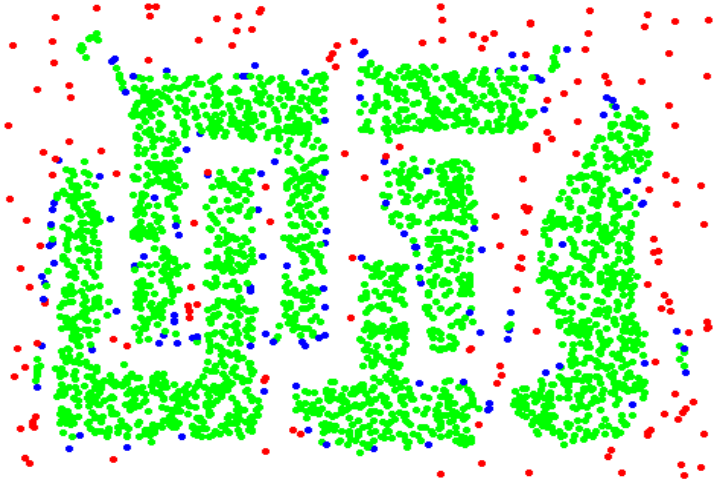
Ο γενικός αλγόριθμος DBSCAN

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: Γενικά



Αρχικά σημεία

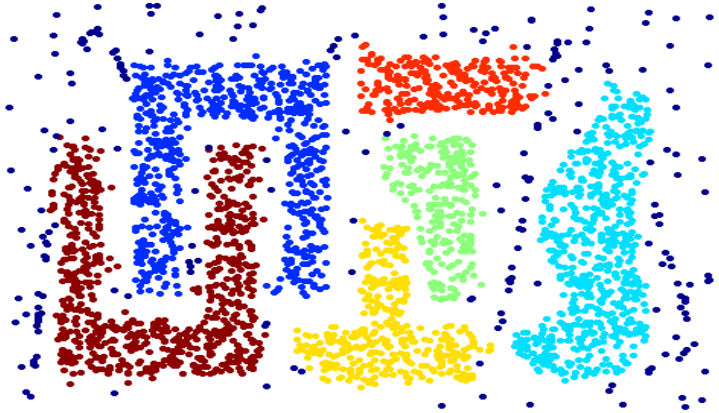
Τύποι σημείων:

Eps = 10, MinPts = 4

core, border & noise

Πότε αποδίδει καλά ο DBSCAN

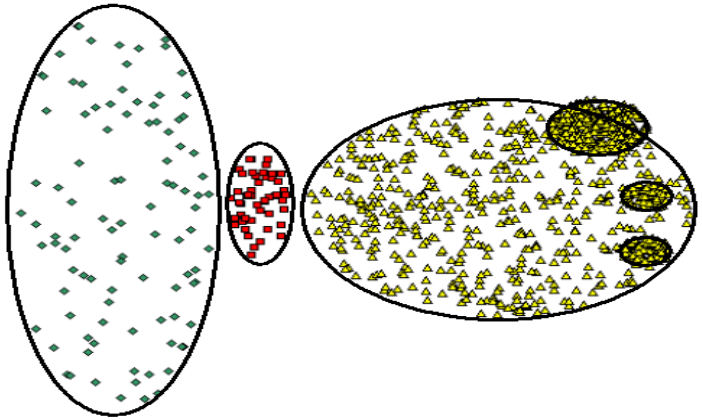
Αρχικά σημεία



Συστάδες

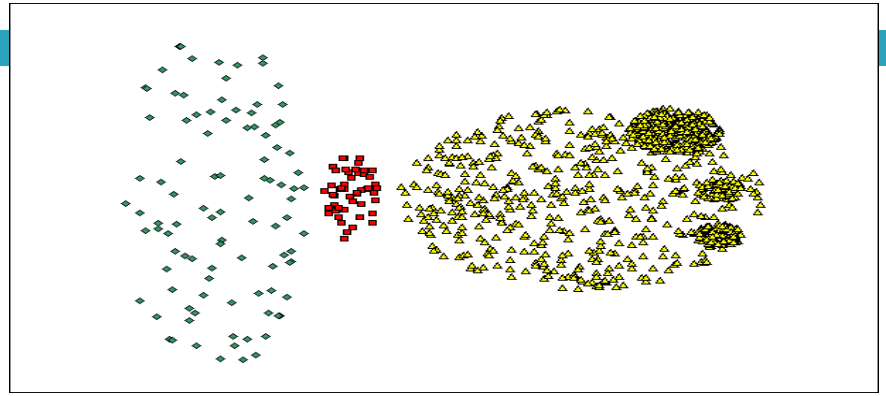
- Ανθεκτικός στο θόρυβο
- Μπορεί να χειριστεί συστάδες διαφορετικών σχημάτων και μεγεθών

Πότε **ΔΕΝ** αποδίδει καλά ο DBSCAN

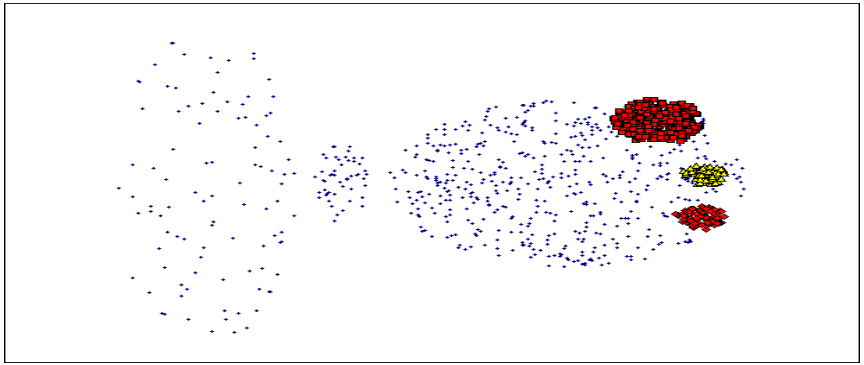


Αρχικά σημεία

- Διαφορετικές πυκνότητες
- Δεδομένα μεγάλης διαστατικότητας



(MinPts=4, Eps=9.75).



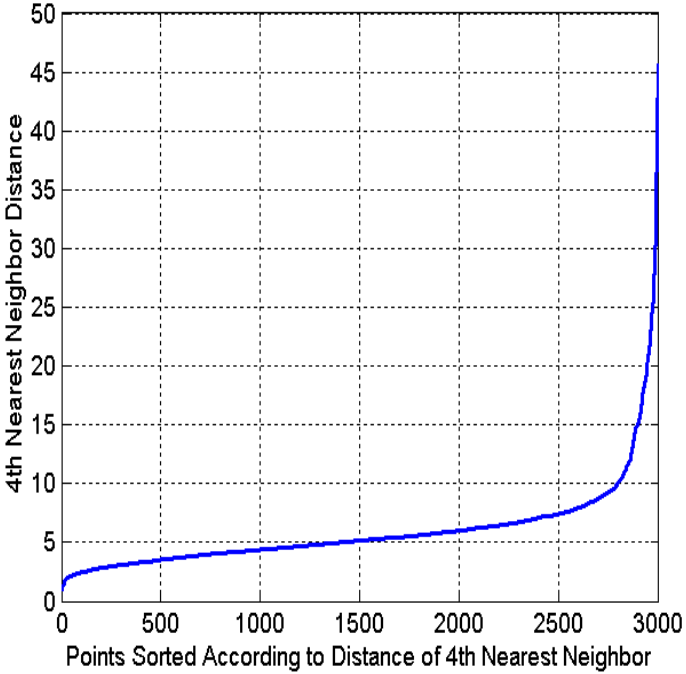
(MinPts=4, Eps=9.92)

Πολυπλοκότητα του DBSCAN

- 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: καθορισμός Eps και MinPts

- Κοιτάμε την απόσταση ενός σημείου από τον k-οστό κοντινότερο γείτονα του -> k-dist
  - 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 φθίνουσα διάταξη
  - Περιμένουμε ξαφνική αλλαγή στο k-dist που αντιστοιχεί στο Eps



Συσταδοποίηση με Mixture Models

- Είναι συχνά βολικό να υποθέτουμε ότι τα δεδομένα έχουν προέλθει από ένα στατιστικό μοντέλο.
- Mixture models→τα δεδομένα να έχουν προέλθει από ένα αριθμό κατανομών
  - Κάθε κατανομή = ένα cluster
- Προαπαιτούμενα:
  - Maximum Likelihood εκτίμηση (μέγιστη πιθανοφάνεια)

Mixture Models

- Οι κατανομές μπορεί να είναι οτιδήποτε αλλά συνήθως η **κανονική** (**Gauss**) κατανομή προτιμάται γιατί
  - Εύκολα κατανοήσιμη
  - Μαθηματικά εύκολη στη χρήση
- Τι είναι τα Mixture models
  - Δοσμένου ενός συνόλου κατανομών (κατά κανόνα ίδιου τύπου) με διαφορετικές παραμέτρους
    - Επέλεξε μια τυχαία και δημιούργησε ένα αντικείμενο (δεδομένο) από αυτή.
    - Επανέλαβε m φορές, όπου m ο αριθμός των αντικειμένων

Ορολογία

- Έστω
  - k κατανομές
  - m αντικείμενα, X={x1,…,xm}
  - Η j-οστή κατανομή έχει παραμέτους θj
  - Θ= το σύνολο των παραμέτρων Θ={θ1,…,θk}
- Η πιθανότητα ενός αντικειμένου x είναι

$$p(x \mid \Theta) = \sum_{i=1}^{k} w_i p_i(x \mid \theta_i)$$

  - Με w_i το βάρος που δηλώνει πόσο πιθανή είναι η επιλογή μιας κατανομής

Ορολογία

- Αν τα αντικείμενα έχουν προέλθει ανεξάρτητα

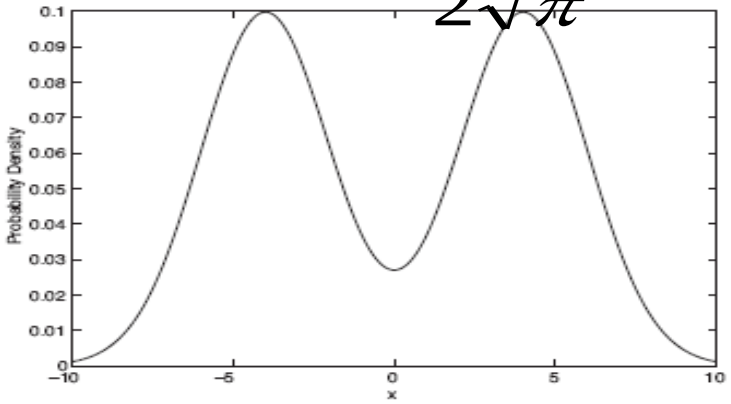
$$p(X \mid \Theta) = \prod_{l=1}^{m} \mathrm{p_i}(\mathrm{x} \mid \Theta) = \prod_{l=1}^{m} \sum_{\mathrm{i}=1}^{\mathrm{k}} \mathrm{w_i p_i}(\mathrm{x} \mid \theta_\mathrm{i})$$

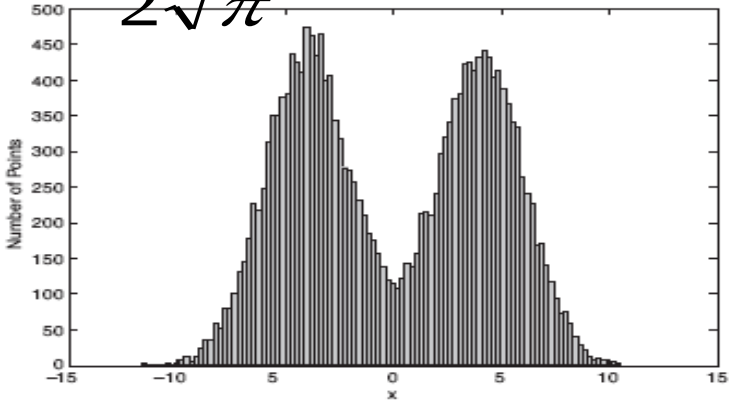
- Τα mixture models μπορούν να βρουν τις παραμέτρους των κατανομών από τα δεδομένα, αλλά δεν παράγουν άμεση αντιστοίχηση με cluster
  - Δηλώνουν πιθανότητα ένα συγκεκριμένο αντικείμενο να ανήκει σε ένα συγκεκριμένο cluster

Παράδειγμα

- Gauss: $p(x \mid \theta) = \frac{1}{\sqrt{2\pi\sigma}} \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}$
- Αν οι παρακάτω κατανομές έχουν κοινό σ=2 και μ1=-4,μ2=4, ενώ τα βάρη επιλογής τους είναι ίδια (0.5) τότε:

$$p(x \mid \Theta) = \frac{1}{2\sqrt{\pi}} \mathrm{e}^{-\frac{(x-4)^2}{2*4^2}} + \frac{1}{2\sqrt{\pi}} \mathrm{e}^{-\frac{(x+4)^2}{2*4^2}}$$





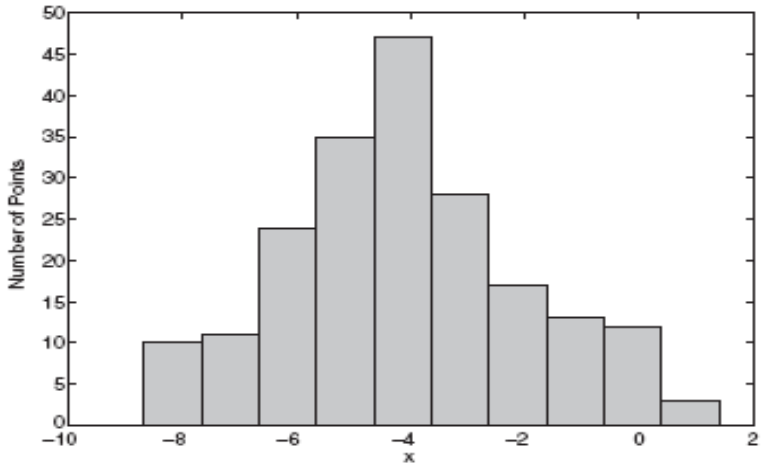
Maximum Likelihood

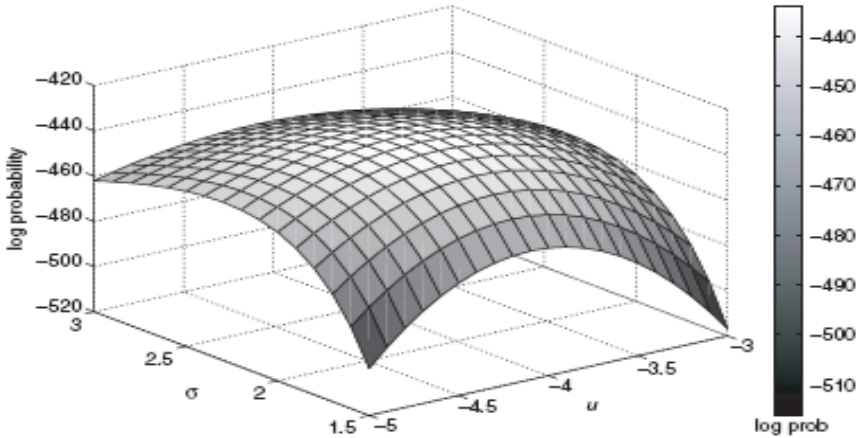
- Ζητούμενο
  - Να βρούμε τις παραμέτρους ενός στατιστικού μοντέλου
- Πως;
  - Είδαμε πως: $p(X \mid \Theta) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$
  - Επειδή είναι πολύ μικρή, συνήθως παίρνουμε λογάριθμο:

$$\log p(X \mid \Theta) = -\sum_{i=1}^{m} \frac{(x_i - \mu)^2}{2\sigma^2} - 0.5m \log 2\pi - m \log \sigma$$

Παράδειγμα

- Έστω 200 σημεία όπως στην εικόνα (μ=-4,σ=2)
- Η δεξιά εικόνα δείχνει το Log-Likelihood για διάφορα μ,σ
  - Εκεί που γίνεται μέγιστο το Log Likelihood, βρίσκεται το πιο πιθανό μ,σ (-4.1,2.1)





Ο αλγόριθμος ΕΜ

- Στη θεωρία
  - Αν ξέρουμε ποια δεδομένα έχουν προκύψει από ποιες κατανομές, μπορούμε να χρησιμοποιήσουμε Log-Likelihood για να βρούμε τις παραμέτρους τους
- Στην πράξη
  - Δεν το γνωρίζουμε…

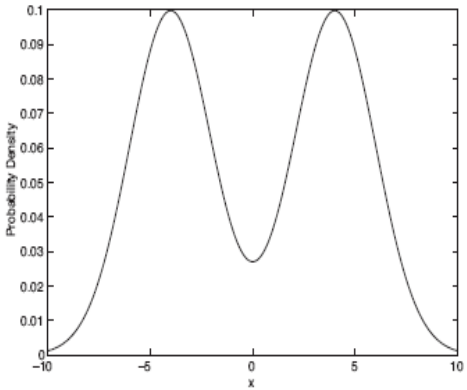
EM

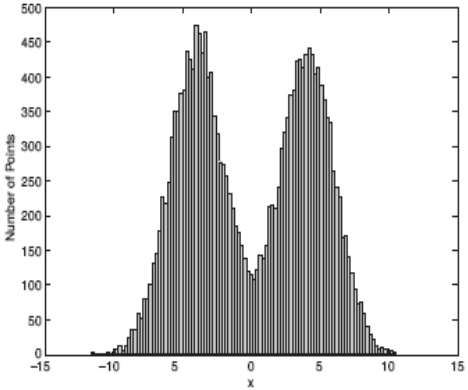
Αλγόριθμος

1. Επιλέγουμε μια αρχική εκτίμηση των παραμέτρων
2. Επανέλαβε
3. Expectation: Για κάθε αντικείμενο, υπολόγισε την πιθανότητα να ανήκει σε κάθε κατανομή
4. Maximization: Με βάση τις πιθανότητες από το expectation βήμα, βρες τις νέες παραμέτρους που μεγιστοποιούν την πιθανοφάνεια
5. Μέχρι να μην αλλάζουν οι παράμετροι

Παράδειγμα

- Έστω δυο οι κατανομές, d1,d2 και 20000 σημεία
  - Έστω σ κοινό =2 (για ευκολία)
  - Ψάχνουμε τα μ1,μ2
- Αρχικά:
  - Έστω μ1=-2, μ2=3
- Expectation:





$$p(\text{κατανομή j}\,|\,x_i,\Theta)=\frac{0.5p(x_i\,|\,\theta_j)}{0.5p(x_i\,|\,\theta_1)+0.5p(x_i\,|\,\theta_2)}$$

Παράδειγμα (συνέχεια)

- Έστω ότι ένα σημείο x=0 έρχεται στο βήμα 2.
- P(x=0|θ1)=0.12 (από τη βασική εξίσωση της κατανομής Gauss) ενώ p(x=0|θ2)=0.06
- Άρα p(κατανομή 1|x=0,Θ)=0.12/(0.12+0.06)=0.66
- Ενώ p(κατανομή 2|x=0,Θ)=0.06/(0.12+0.06)=0.33
- Δηλ το x=0 είναι 2 φορές πιο πιθανό να ανήκει στην κατανομή 1. (με βάση τις αρχικές εκτιμήσεις)
- Το ίδιο γίνεται για όλα τα 20000 σημεία

Παράδειγμα (συνέχεια)

- Maximization:
  - Υπολογισμός μ1,μ2 (νέων)

$$\mu 1 = \sum_{i=1}^{20000} x_i \frac{p(\text{κατανομή } 1 \mid x_i, \Theta)}{\sum_{i=1}^{20000} p(\text{κατανομή } 1 \mid x_i, \Theta)}$$

$$\mu 2 = \sum_{i=1}^{20000} x_i \frac{p(\text{κατανομή } 2 \mid x_i, \Theta)}{\sum_{i=1}^{20000} p(\text{κατανομή } 2 \mid x_i, \Theta)}$$

Παράδειγμα (συνέχεια)

- Μετά από μερικές επαναλήψεις:

Επανάληψη	μ1	μ2
0	-2.00	3.00
1	-3.74	4.10
2	-3.94	4.07
3	-3.97	4.04
4	-3.98	4.03
5	-3.99	4.02
6	-3.99	4.02