Συσταδοποίηση (clustering)

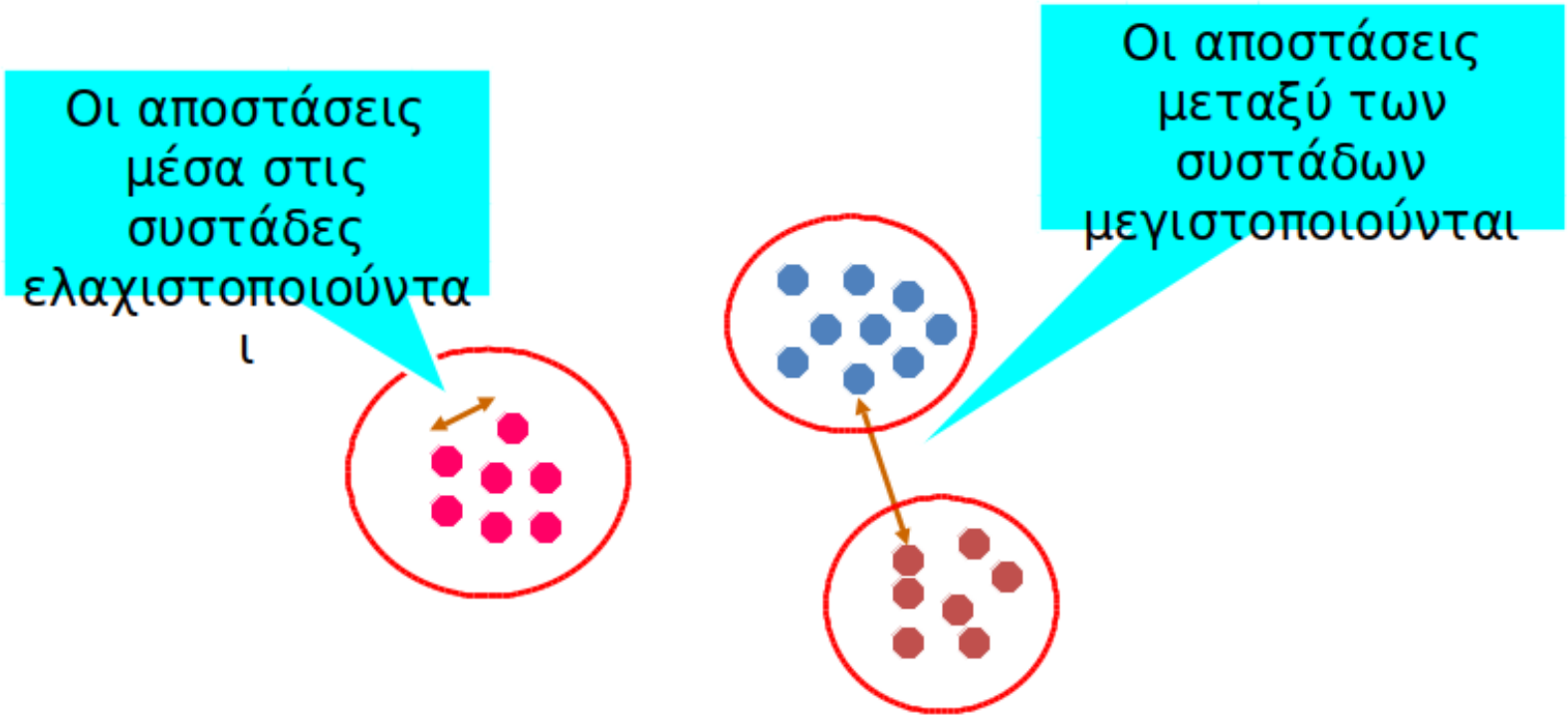
Περιεχόμενα σημερινής διάλεξης

- Γενικά για συσταδοποίηση
- Αλγόριθμος K-means
- Αλγόριθμος ιεραρχικής συσταδοποίησης
- Μέτρα απόστασης - ομοιότητας

Συσταδοποίηση clustering

Τί είναι η συσταδοποίηση;

- Εύρεση συστάδων (ομάδων) αντικειμένων έτσι ώστε τα αντικείμενα σε μια ομάδα να είναι παρόμοια το ένα με το άλλο, και διαφορετικά από τα αντικείμενα άλλων ομάδων.



Συσταδοποίηση clustering

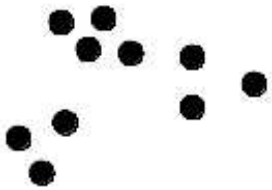
Πως γίνεται η συσταδοποίηση;

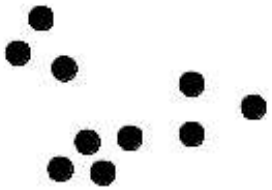
- Δίδονται:
  Ένα σύνολο από σημεία που το καθένα έχει κάποια γνωρίσματα
  Μια μέθοδος μέτρησης της απόστασης/ομοιότητας μεταξύ τους
- Ζητείται η εύρεση **συστάδων** (ομάδων) τέτοιων ώστε:
  Τα σημεία σε μια συστάδα είναι πιο όμοια μεταξύ τους
  Τα σημεία σε διαφορετικές συστάδες είναι λιγότερο όμοια μεταξύ τους.
- Υπόθεση: Οι συστάδες δεν είναι γνωστές εκ των προτέρων.

Συσταδοποίηση clustering

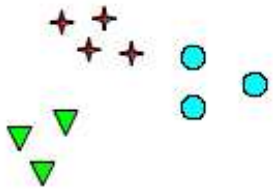
Τί είναι η συσταδοποίηση;

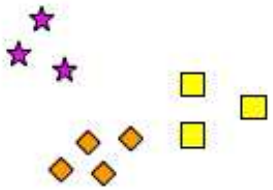
Η έννοια της συστάδας είναι ασαφής.



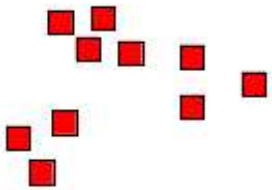


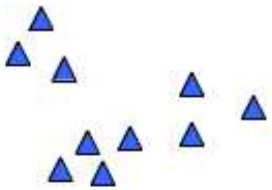
Πόσες συστάδες;



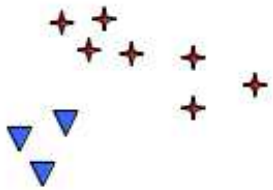


Έξι Συστάδες





Δύο Συστάδες



Τέσσερεις Συστάδες

Συσταδοποίηση clustering

Πότε μια συσταδοποίηση είναι καλή;

Πότε μια συσταδοποίηση είναι καλή;

- Μεγάλη ομοιότητα εντός των συστάδων
- Μικρή ομοιότητα ανάμεσα στις συστάδες

Η ποιότητα εξαρτάται από την

- Μέτρηση ομοιότητας
- Μέθοδο εύρεσης της συσταδοποίησης

Συσταδοποίηση clustering

Που χρησιμοποιείται η συσταδοποίηση;

Που χρησιμοποιείται η συσταδοποίηση;

- Κατανόηση δεδομένων
  οπτικοποίηση, συμπεράσματα για την κατανομή τους
- Σύνοψη/συμπίεση δεδομένων
  Μείωση μεγέθους μεγάλων συνόλων δεδομένων, χρήση αντιπροσωπευτικών σημείων από κάθε συστάδα
- Αποδοτική κατασκευή ευρετηρίων - εύρεση κοντινότερου γείτονα

Που χρησιμοποιείται η συσταδοποίηση;

- Market segmentation
  Στόχος: Χωρισμός των καταναλωτών σε ομάδες έτσι ώστε τα μέλη κάθε ομάδας να είναι ο στόχος συγκεκριμένης πολιτικής marketing.
  Προσέγγιση: Συγκέντρωση διαφορετικών γνωρισμάτων για τους καταναλωτές.
  Ορισμός ‘‘ομοιότητας’’ ανάμεσα στους πελάτες.
  Δημιουργία ομάδων με όμοιους πελάτες.
  Μέτρηση της ποιότητας της ομαδοποίησης (π.χ. παρατηρώντας τις αγοραστικές συνήθειες στην ίδια ομάδα και ανάμεσα σε διαφορετικές ομάδες)

Που χρησιμοποιείται η συσταδοποίηση;

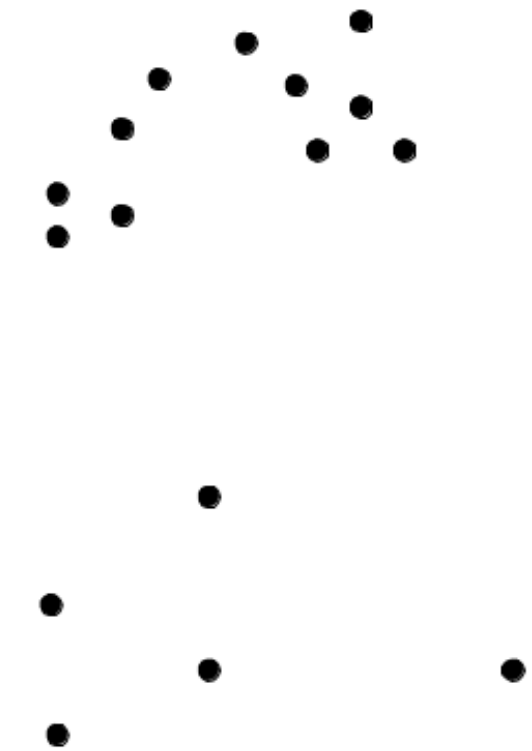
- Συσταδοποίηση εγγράφων
  Στόχος: Εύρεση ομάδων από έγγραφα που είναι όμοια μεταξύ τους με βάση τους σημαντικούς όρους που εμφανίζονται σ᾽ αυτά.
  Προσέγγιση: Εύρεση για κάθε έγγραφο των όρων που εμφανίζονται συχνά σ᾽ αυτό.
  Μέτρηση ομοιότητας με βάση τη συχνότητα των διαφορετικών όρων και χρήση της για την δημιουργία συστάδων.
  Όφελος: Μέθοδοι ανάκτησης πληροφορίας μπορούν να χρησιμοποιήσουν τις συστάδες για να συσχετίσουν ένα καινούργιο έγγραφο ή έναν όρο αναζήτησης με τα έγγραφα κάθε συστάδας.

Συσταδοποίηση clustering

Κατηγορίες συσταδοποίησης

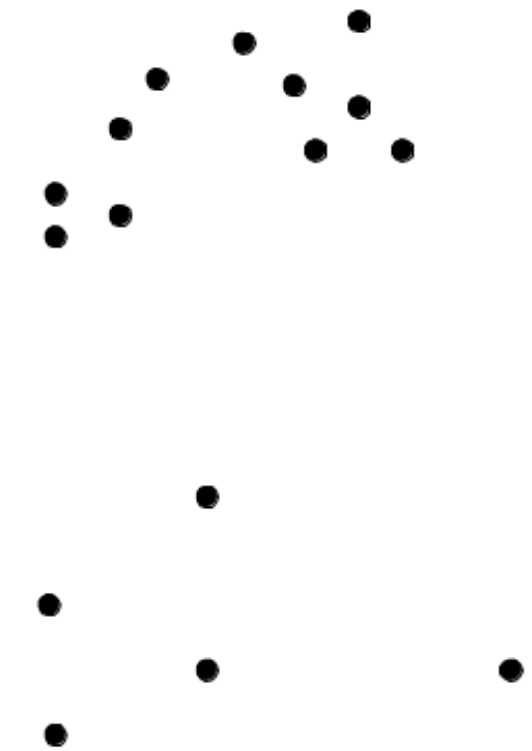
- Διαμεριστική ή διαχωριστική συσταδοποίηση (partitional clustering
  Μια διαμέριση (partition) του συνόλου των αντικειμένων σε μη επικαλυπτόμενα (ξένα) υποσύνολα (συστάδες) έτσι ώστε κάθε αντικείμενο ανήκει σε ένα ακριβώς υποσύνολο.
- Ιεραρχική συσταδοποίηση (hierarchical clustering)
  Ένα σύνολο από εμφωλευμένες (nested) συστάδες οργανωμένες σε ένα ιεραρχικό δένδρο.

Η διαχωριστική συσταδοποίηση με μια ματιά

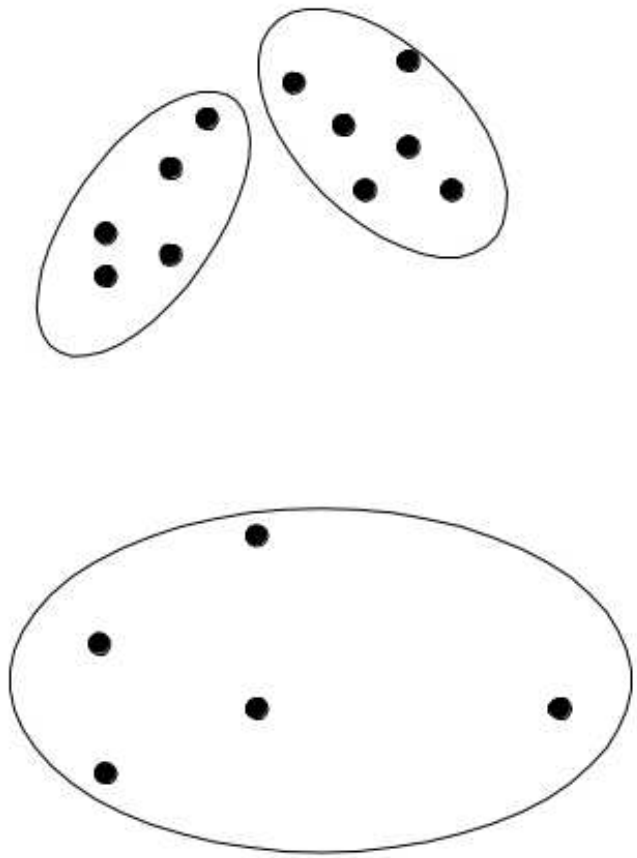


Αρχικά Σημεία

Η διαχωριστική συσταδοποίηση με μια ματιά

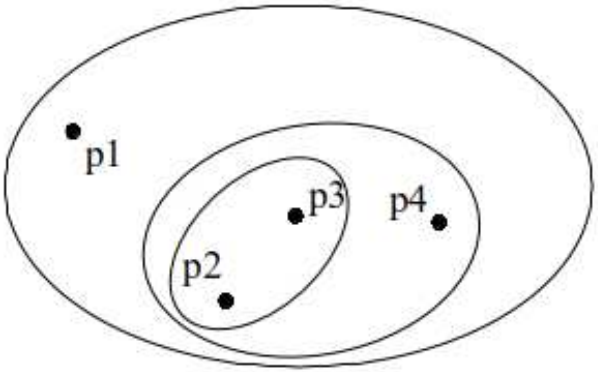


Αρχικά Σημεία

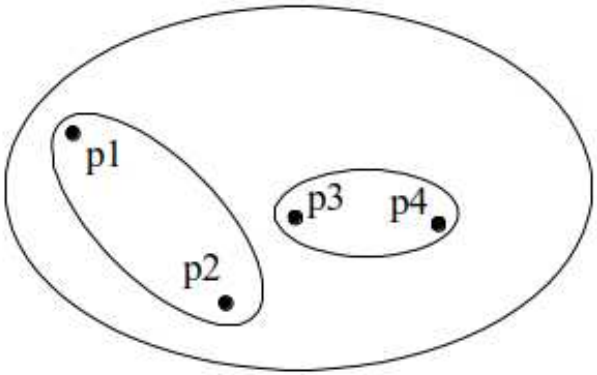


Διαμεριστική Συσταδοποίηση

Η ιεραρχική συσταδοποίηση με μια ματιά

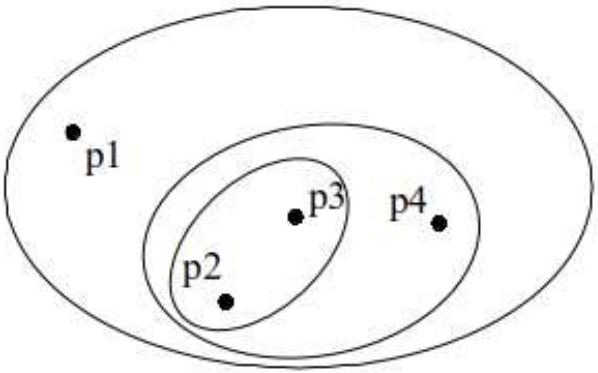


Παραδοσιακή Ιεραρχική Συσταδοποίηση

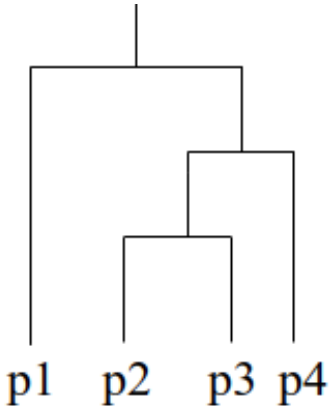


Μη Παραδοσιακή Ιεραρχική Συσταδοποίηση

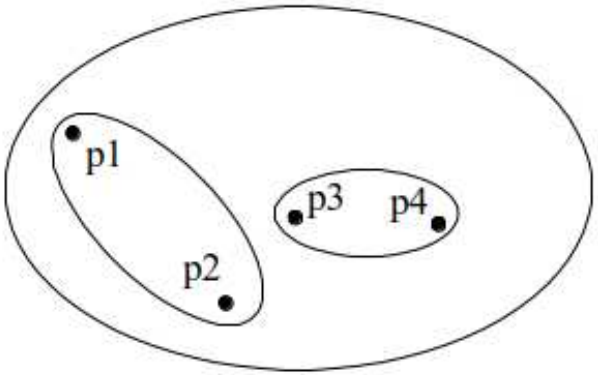
Η ιεραρχική συσταδοποίηση με μια ματιά



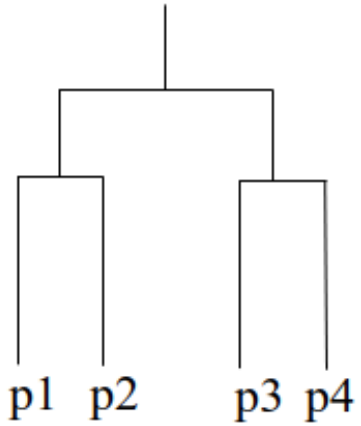
Παραδοσιακή Ιεραρχική Συσταδοποίηση



Παραδοσιακό Δενδρόγραμμα



Μη Παραδοσιακή Ιεραρχική Συσταδοποίηση



Μη-Παραδοσιακό Δενδρόγραμμα

Τα φύλλα είναι απλά σημεία

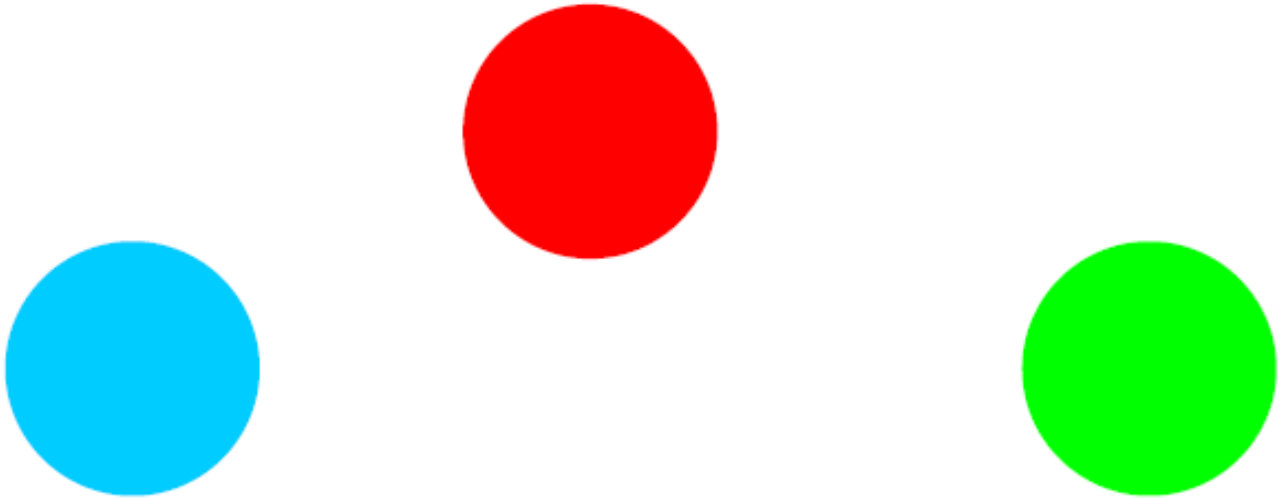
Τύποι συστάδων

Ανάλογα με την μορφή των δεδομένων (π.χ. διακριτά αριθμητικά, διακριτά μη αριθμητικά, συνεχή, κλπ.) υπάρχουν διάφοροι τύποι συστάδων.

Συσταδοποίηση clustering

Τύποι συστάδων: Καλώς διαχωρισμένες συστάδες

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



3 καλώς-διαχωρισμένες συστάδες

Συχνά υπάρχει η έννοια του κατωφλιού (threshold)

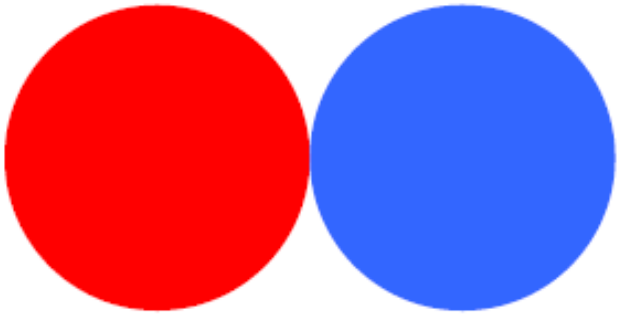
Όχι απαραίτητα κυκλικοί (οποιοδήποτε σχήμα)

Συσταδοποίηση clustering

Τύποι συστάδων: Συστάδες βασισμένες σε κέντρο

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

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



4 συστάδες βασισμένες σε κέντρο

Τείνουν στο να είναι κυκλικοί

Συσταδοποίηση clustering

Τύποι συστάδων: Συνεχείς συστάδες

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

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

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

Πρόβλημα με θόρυβο

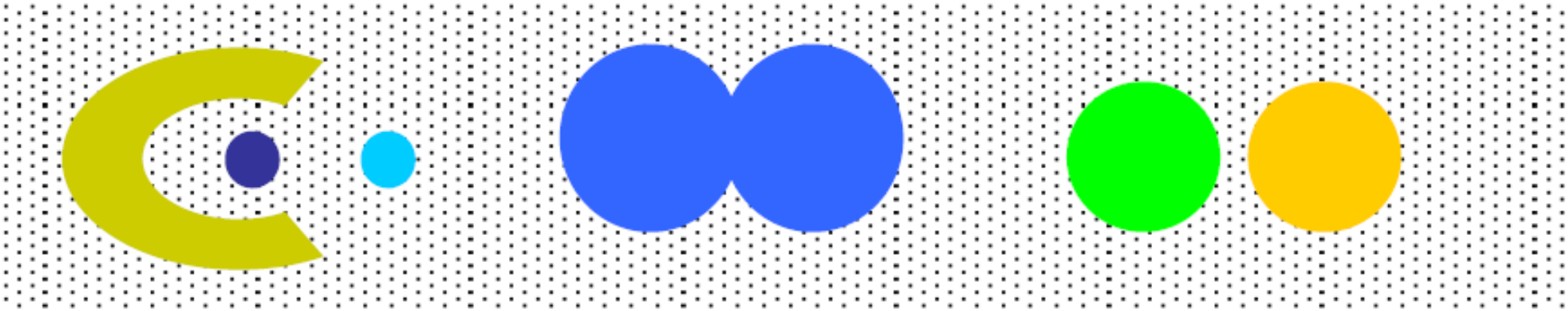
8 συνεχείς συστάδες

Συσταδοποίηση clustering

Τύποι συστάδων: Συστάδες βασισμένες στην πυκνότητα

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

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

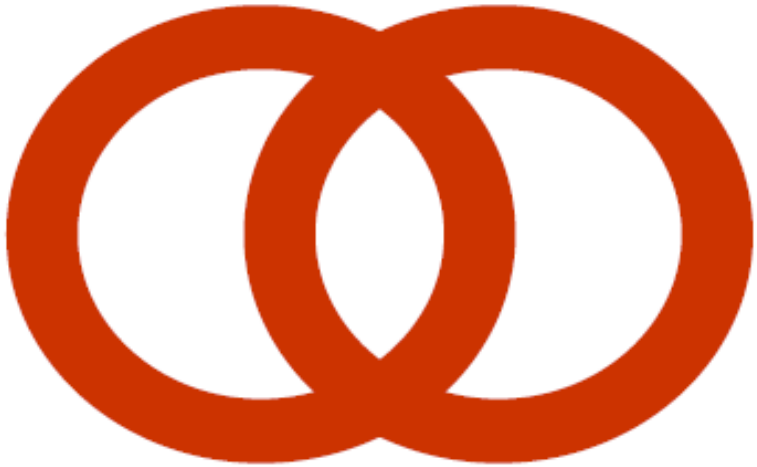


6 συστάδες βασισμένες στην πυκνότητα

Συσταδοποίηση clustering

Τύποι συστάδων: Εννοιολογική συσταδοποίηση

Συστάδες με κοινή ιδιότητα ή εννοιολογικές συστάδες.



2 αλληλοκαλυπτόμενοι κύκλοι

Συσταδοποίηση clustering

Τύποι συστάδων: Άλλες διακρίσεις

Επικαλυπτόμενο ή όχι

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

Ασαφή συσταδοποίηση

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

Μερική - Πλήρης

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

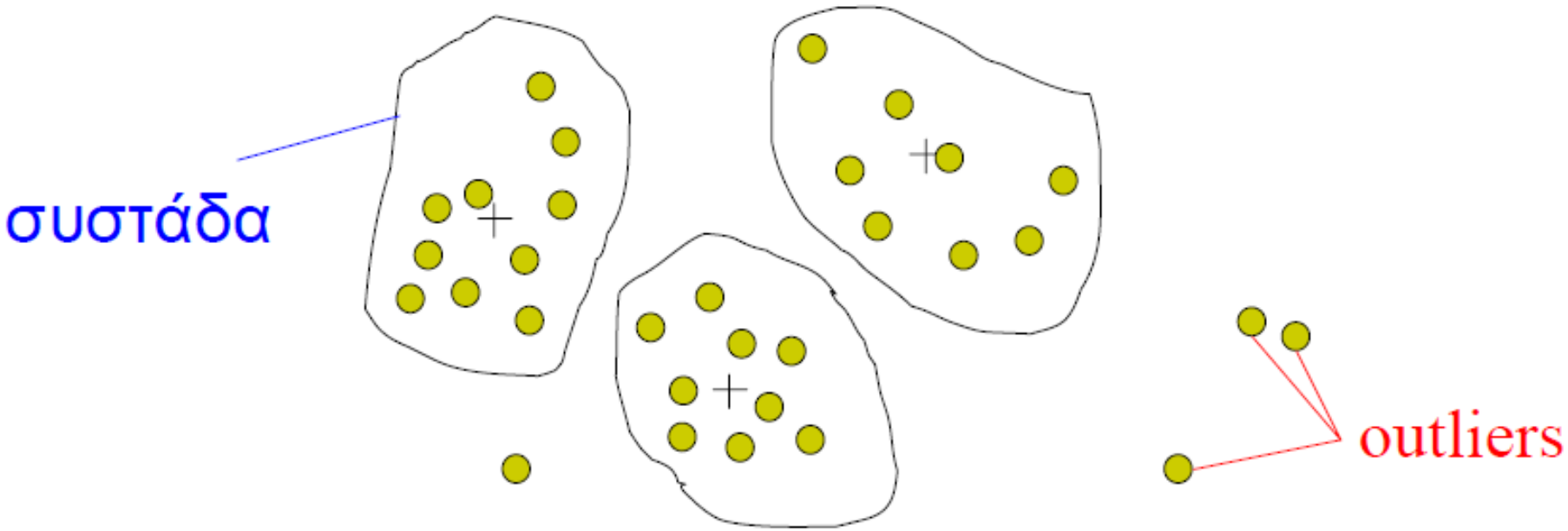
Ετερογενή - Ομογενή

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

Συσταδοποίηση clustering

Γενικές απαιτήσεις

Αντιμετώπιση θορύβου και outliers



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

Σημαντικοί αλγόριθμοι συσταδοποίησης

- K-means και παραλλαγές
- Συσσωρευτική ιεραρχική συσταδοποίηση

Ο **αλγόριθμος** K-**means**

Ο αλγόριθμος K-means

Ο αλγόριθμος K-means είναι διαχωριστικός αλγόριθμος
Κάθε συστάδα συσχετίζεται με ένα κεντρικό σημείο (κεντροειδές - centroid)
Κάθε σημείο τοποθετείται στη συστάδα με το κοντινότερο κεντρικό σημείο.
Ο αριθμός K των συστάδων είναι είσοδος στον αλγόριθμο.

Ο αλγόριθμος K-means

Η βασική περιγραφή του K-means είναι πολύ απλή.

Ο αλγόριθμος K-means

Είσοδος: Ο αριθμός K των συστάδων
Έξοδος: K συστάδες

1. Επιλέγονται K σημεία ως τα αρχικά κέντρα των συστάδων
2. Όσο τα κέντρα των συστάδων αλλάζουν
   1. Ανάθεση κάθε σημείου στο κοντινότερο από τα K κεντρικά σημεία.
   2. Επανα-υπολογισμός του κέντρου κάθε συστάδας.

Ο αλγόριθμος K-means

Παρατηρήσεις

1. Τα αρχικά κεντρικά σημεία συνήθως επιλέγονται τυχαία.
2. Οι συστάδες που παράγονται διαφέρουν από την μια εκτέλεση του αλγορίθμου στην άλλη.
3. Η ομοιότητα/εγγύτητα των σημείων υπολογίζεται με κάποια απόσταση που εξαρτάται από το είδος των σημείων/δεδομένων. Επειδή η απόσταση υπολογίζεται συχνά πρέπει ο υπολογισμός της να είναι σχετικά απλός.
4. Τα κεντρικό σημεία είναι συνήθως το κεντροειδές ή το μέσο των σημείων της συστάδας. Μπορεί να μην ανήκει στα δεδομένα εισόδου!

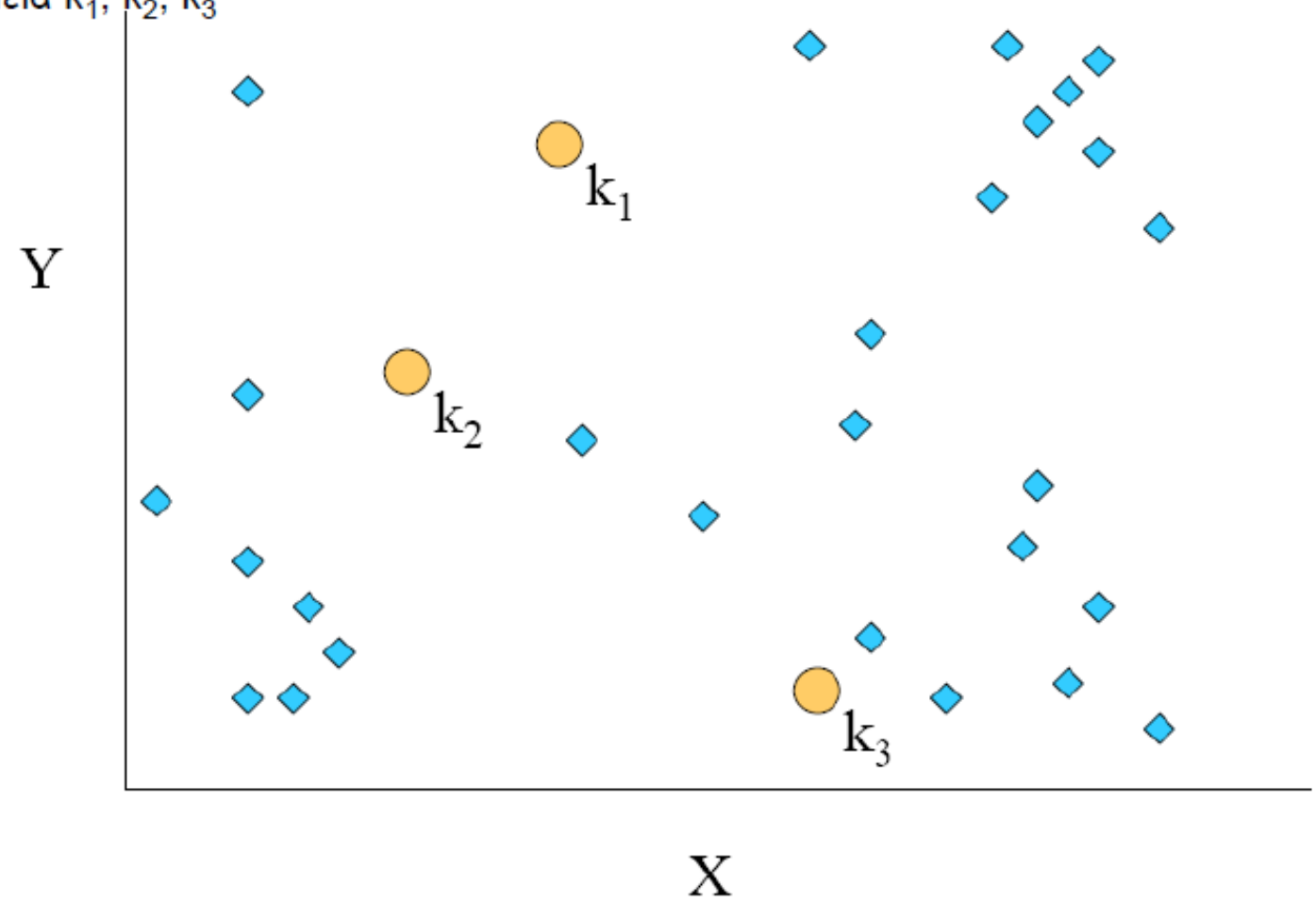
Ο αλγόριθμος K-means

Ένα παράδειγμα

Αρχική κατάσταση,

K = 3 συστάδες

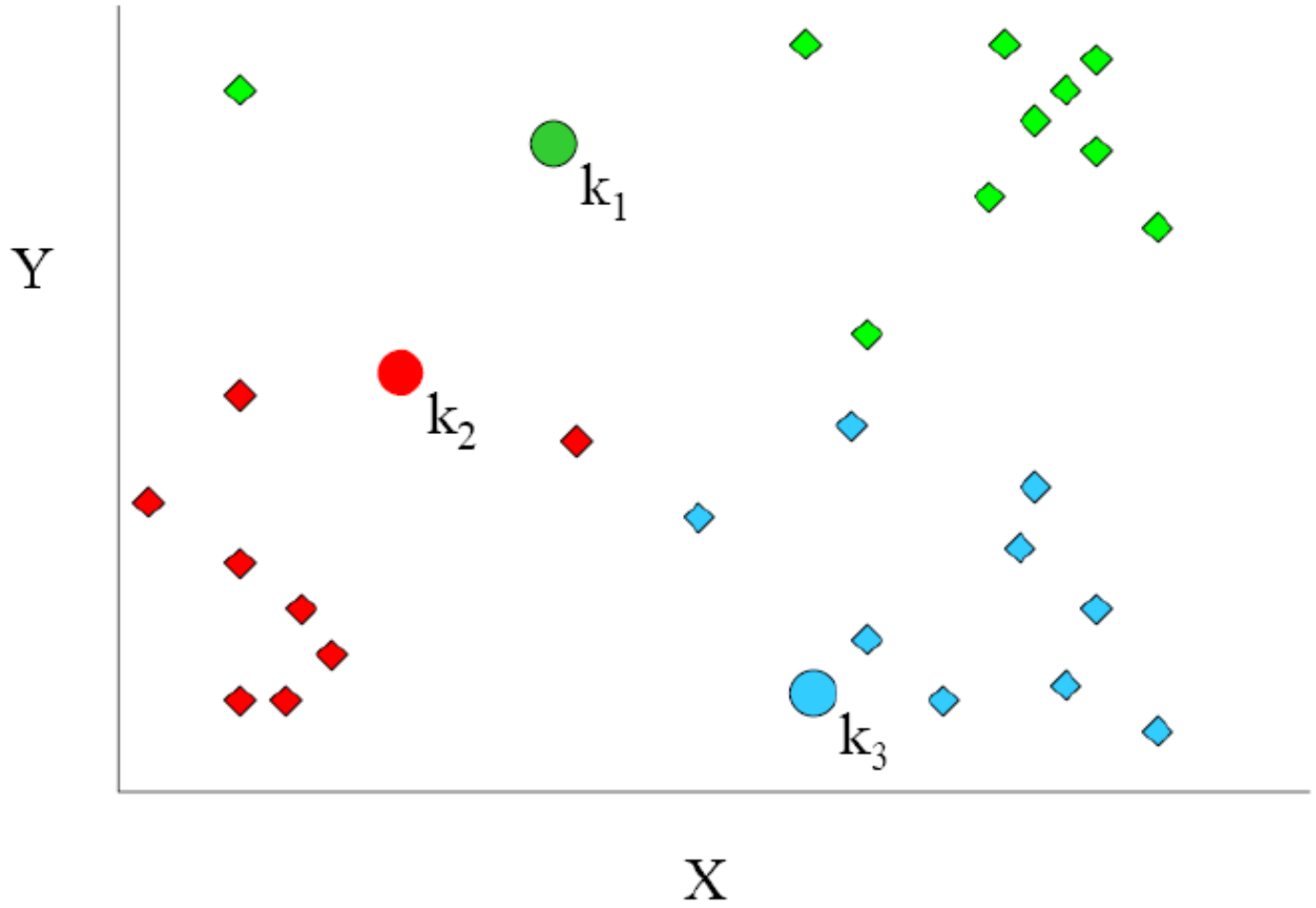
Αρχικά σημεία k_1, k_2, k_3



Ο αλγόριθμος K-means

Ένα παράδειγμα

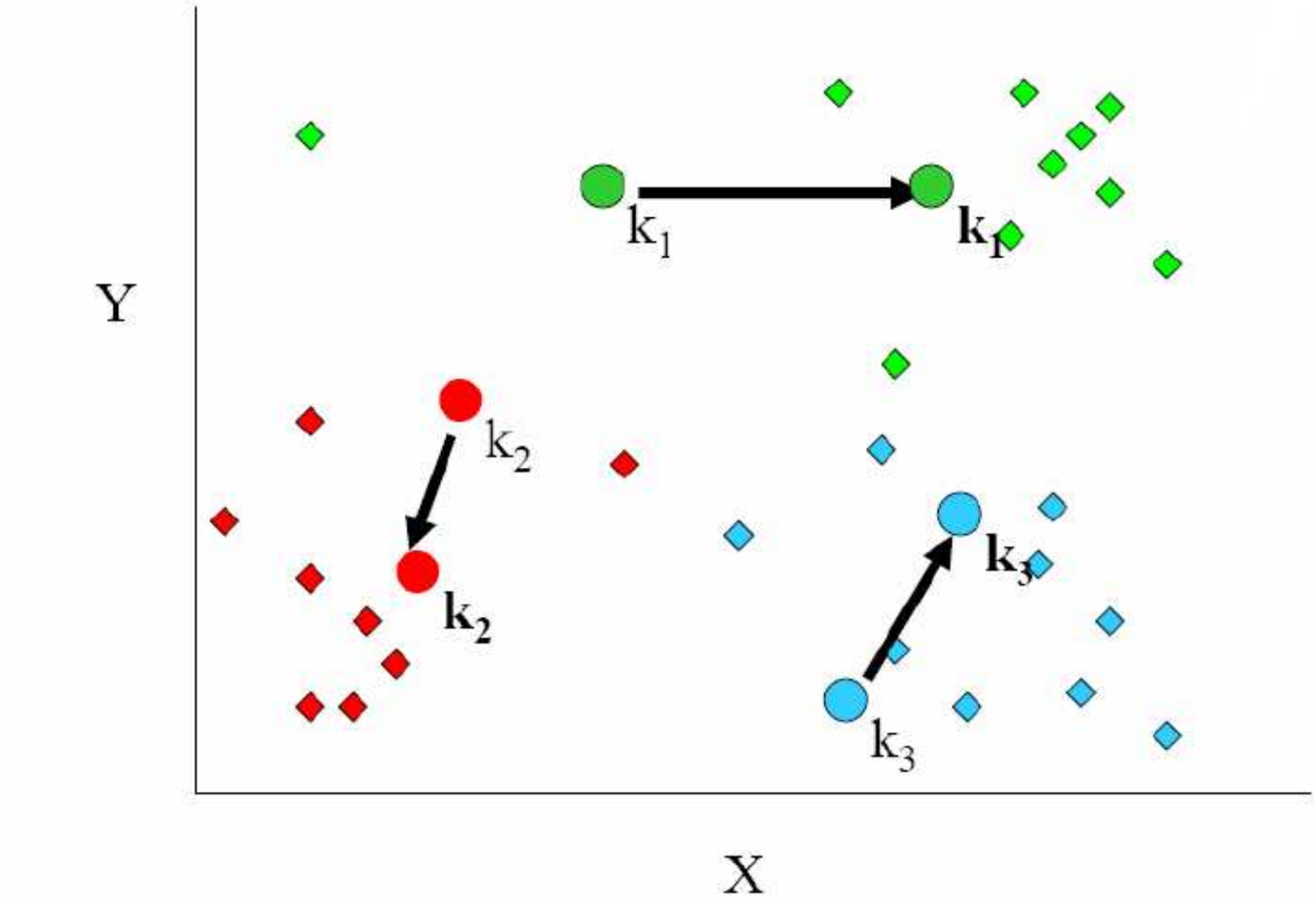
Τα σημεία ανατίθενται στο πιο γειτονικό από τα 3 αρχικά σημεία



Ο αλγόριθμος K-means

Ένα παράδειγμα

Επανα-υπολογισμός του κέντρου
(κέντρου βάρους) κάθε σημείου

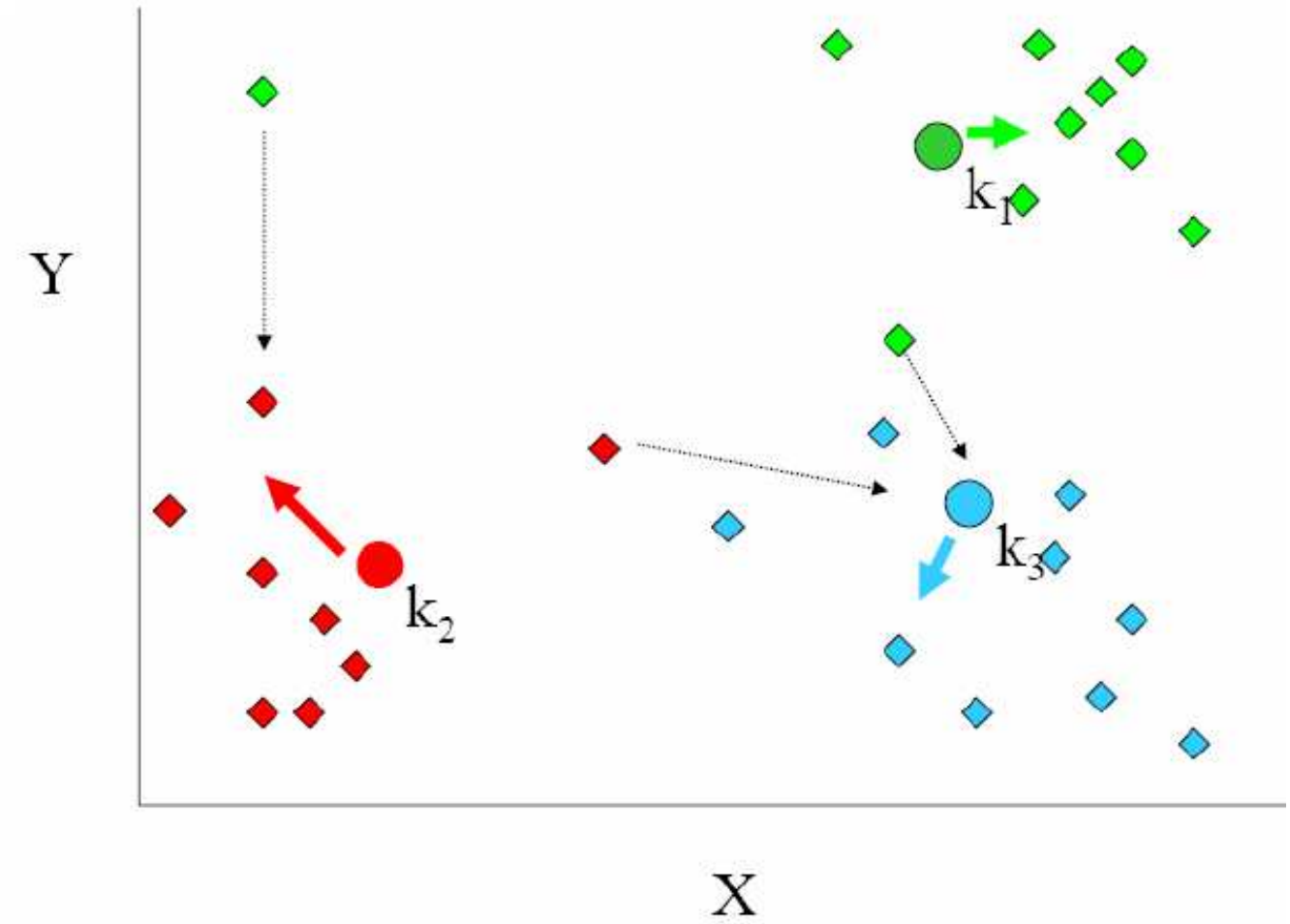


Ο αλγόριθμος K-means

Ένα παράδειγμα

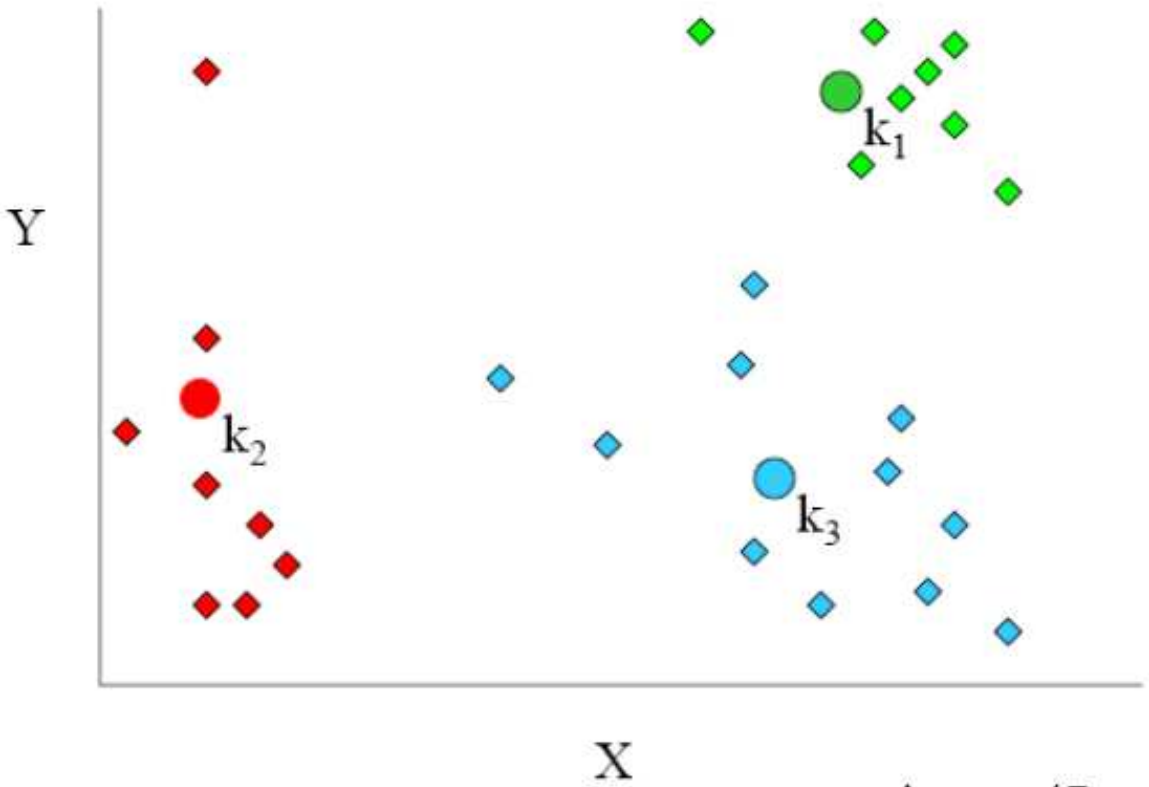
Νέα ανάθεση των σημείων

Νέα κέντρα βάρους



Ο αλγόριθμος K-means

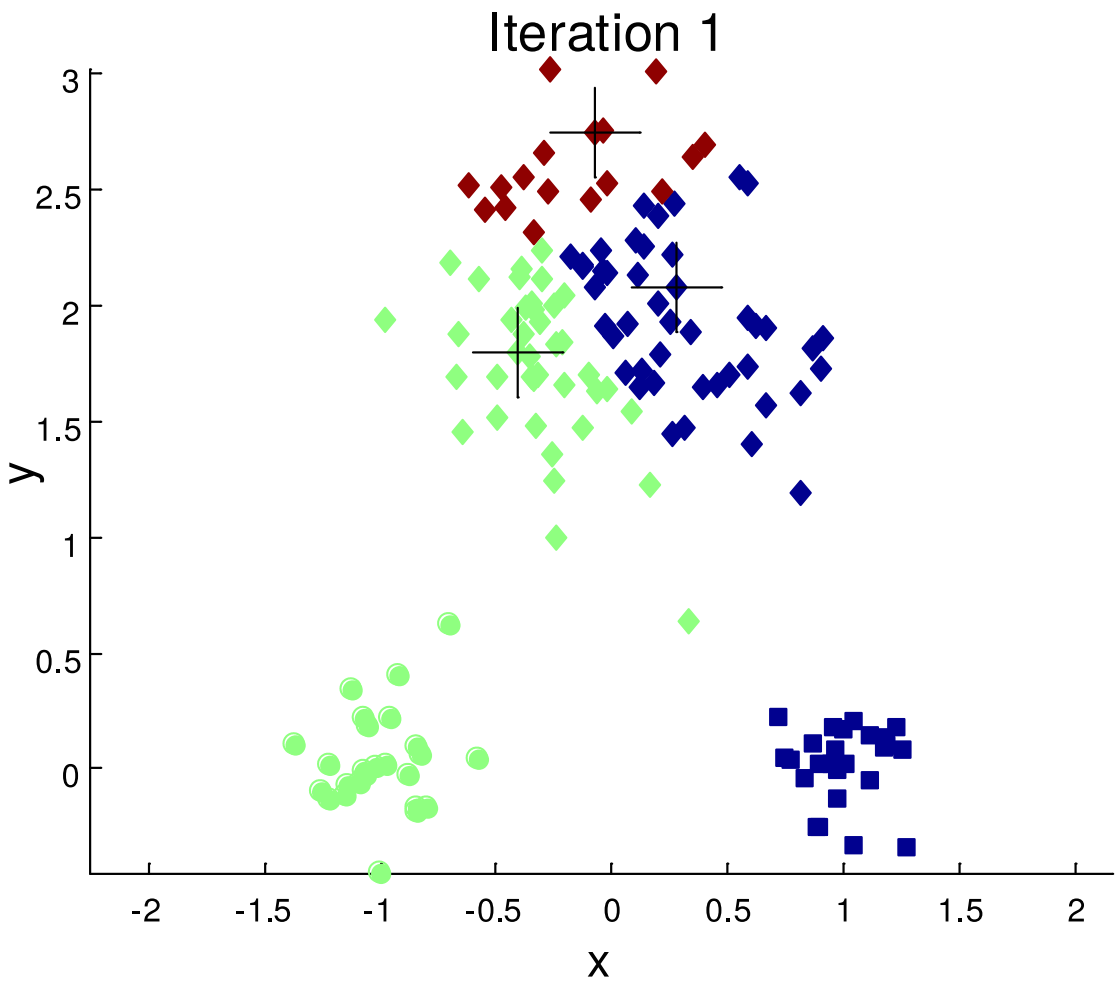
Ένα παράδειγμα



Δεν αλλάζει τίποτα -> ΤΕΛΟΣ

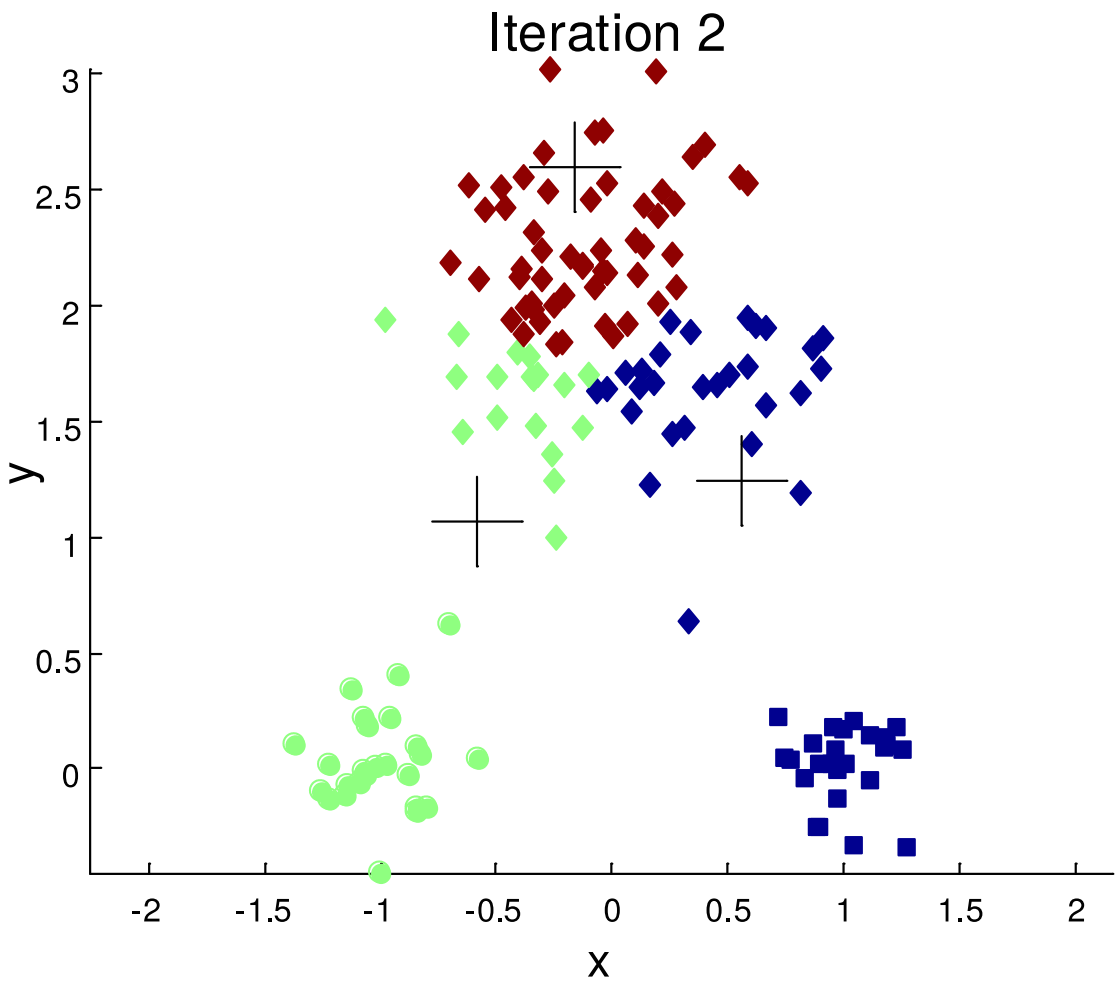
Ο αλγόριθμος K-means

Ένα άλλο παράδειγμα



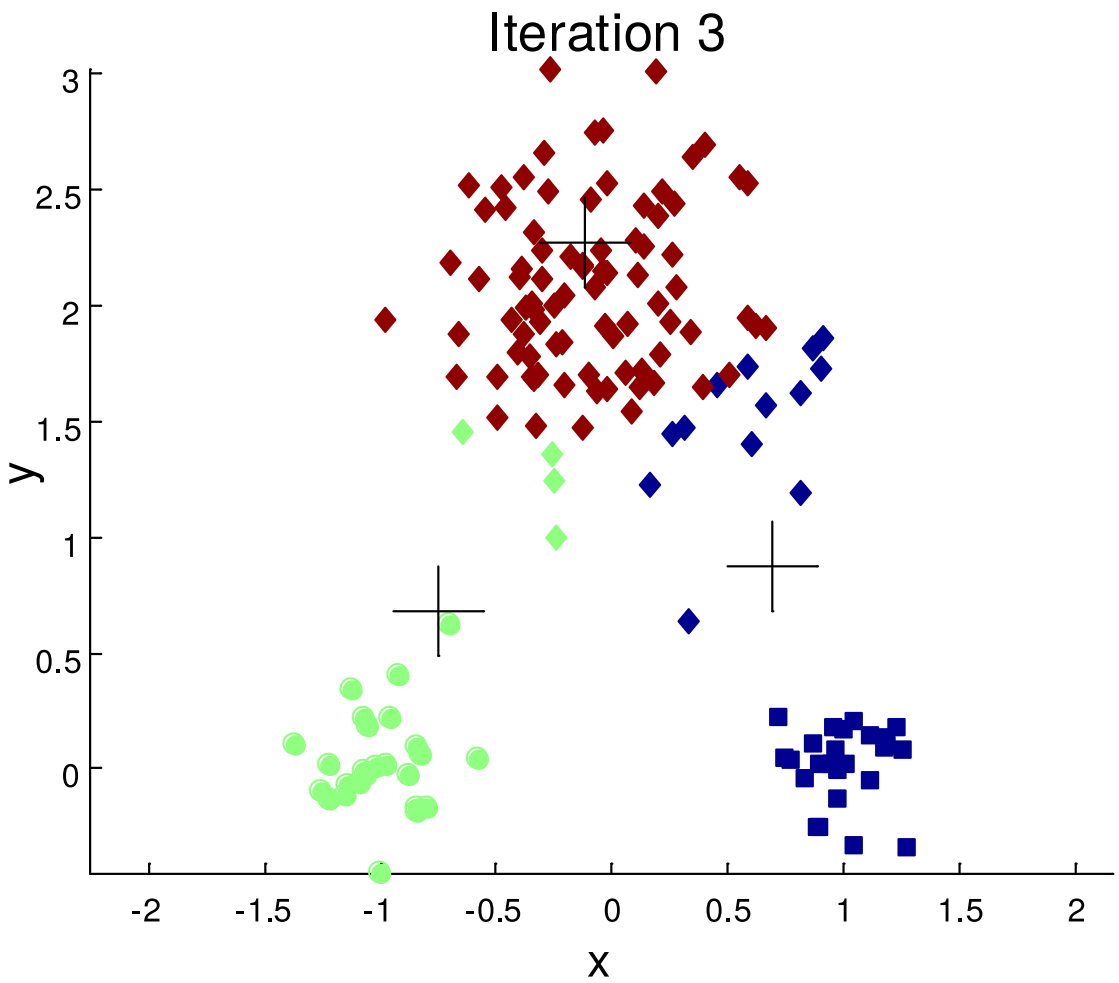
Ο αλγόριθμος K-means

Ένα άλλο παράδειγμα



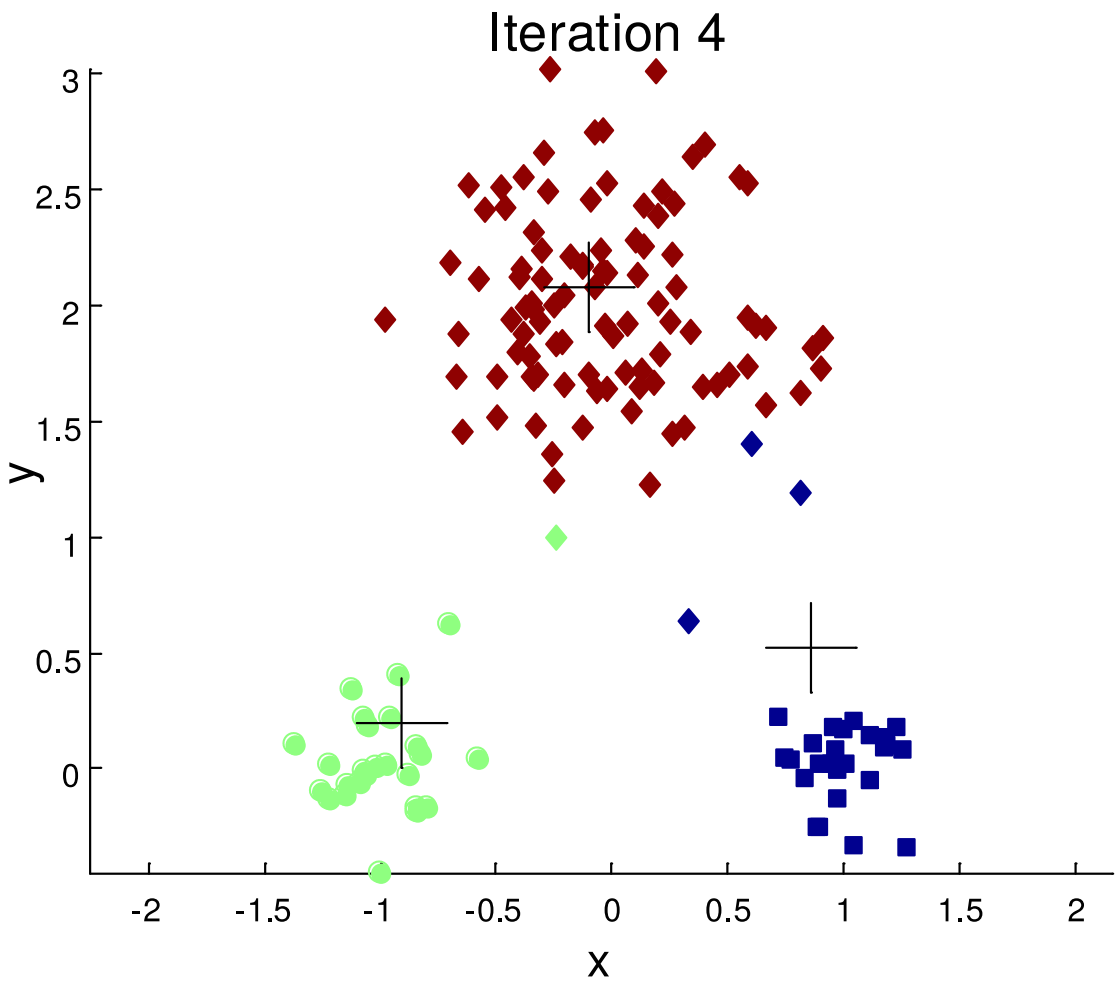
Ο αλγόριθμος K-means

Ένα άλλο παράδειγμα



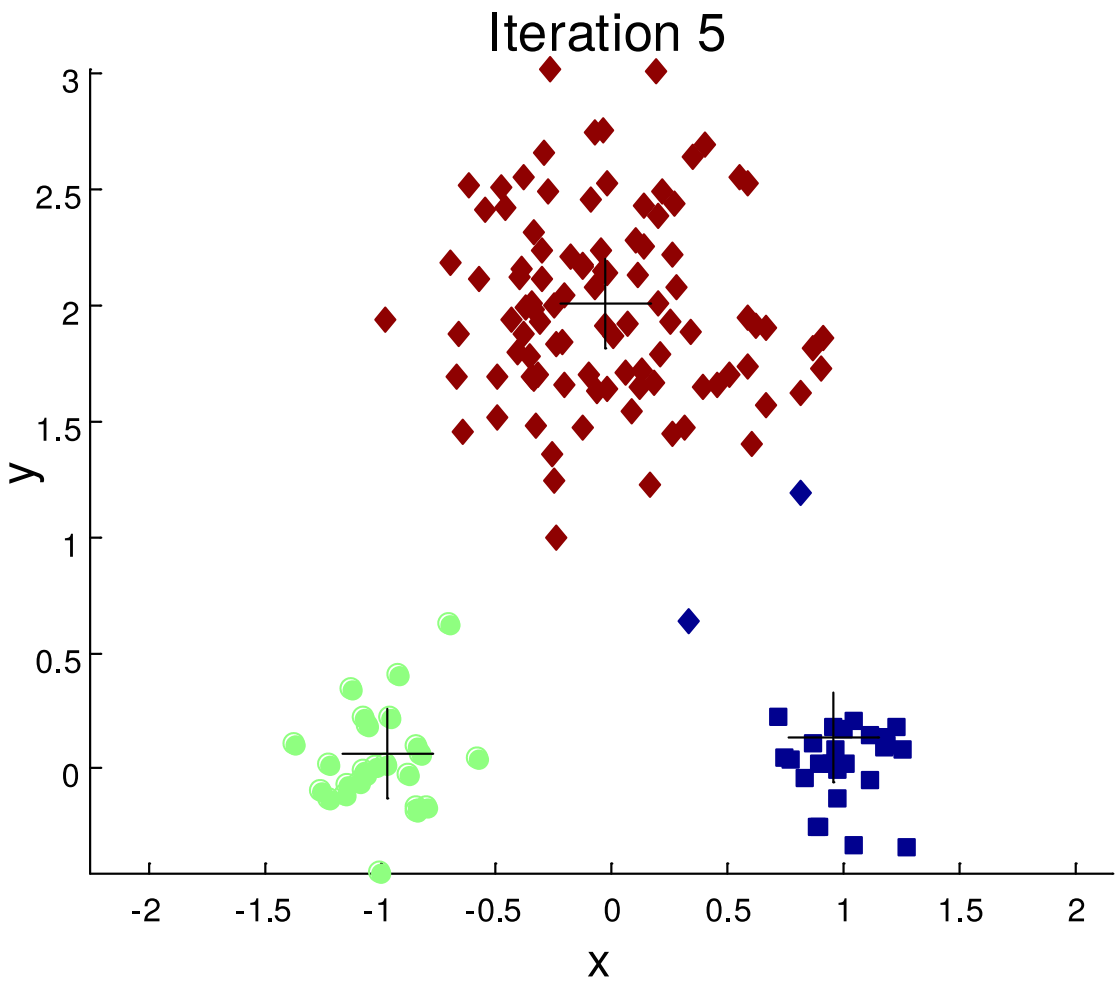
Ο αλγόριθμος K-means

Ένα άλλο παράδειγμα



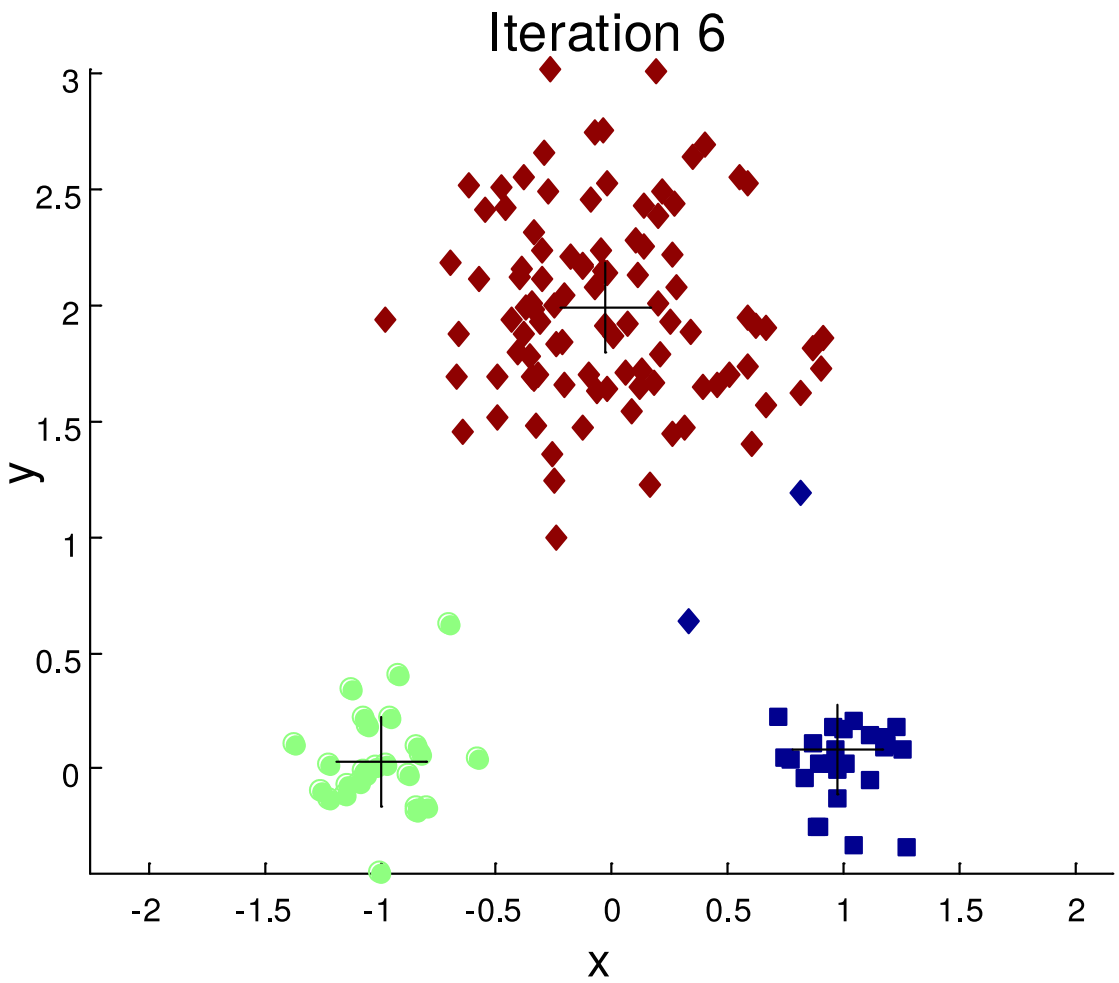
Ο αλγόριθμος K-means

Ένα άλλο παράδειγμα



Ο αλγόριθμος K-means

Ένα άλλο παράδειγμα



Ο αλγόριθμος K-means

Παρατηρήσεις (συνέχεια)

1. Απαιτήσεις σε μνήμη: Αποθηκεύουμε μόνο τα κέντρα
2. Η πολυπλοκότητα είναι $O(I \cdot n \cdot K * d)$
   όπου
   n: αριθμός των σημείων
   K: αριθμός των συστάδων
   I: αριθμός των επαναλήψεων
   d: αριθμός των γνωρισμάτων/χαρακτηριστικών (διάσταση)
3. Για συνηθισμένα μέτρα ομοιότητας ο αλγόριθμος συγκλίνει (ολοκληρώνεται). Η σύγκλιση γίνεται συνήθως στις πρώτες επαναλήψεις.
4. Συνήθως η συνθήκη εξόδου του αλγορίθμου είναι
   είτε σχετικά λίγα σημεία άλλαξαν συστάδα
   είτε η απόσταση μεταξύ των νέων κεντρικών σημείων από τα παλιά είναι μικρή.

Ο αλγόριθμος K-means

Παρατηρήσεις (συνέχεια)

1. Ένα μέτρο της ποιότητας του αποτελέσματος του K-means είναι το **άθροισμα των τετραγώνων των σφαλμάτων** (sum of squared error (SSE)) το οποίο ορίζεται ως εξής:

$$SSE = \sum_{i=1}^{K} \sum_{\text{Για κάθε } x \in C_i} d^2(m_i, x)$$

όπου $C_1, C_2, \ldots, C_k$ είναι οι K συστάδες και $m_1, m_2, \ldots, m_k$ είναι τα κέντρα των K συστάδων αντίστοιχα και $d(x, y)$ είναι η απόσταση ανάμεσα στα x και y.
Με άλλα λόγια, αθροίζουμε τα τετράγωνα των αποστάσεων κάθε σημείου των δεδομένων μας από το κεντρικό σημείο της συστάδας που ανήκει.

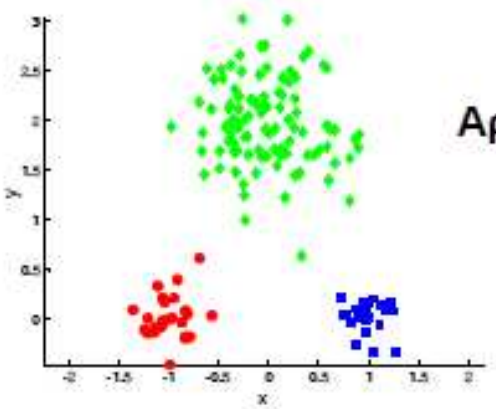
Ο αλγόριθμος K-means

Παρατηρήσεις (συνέχεια)

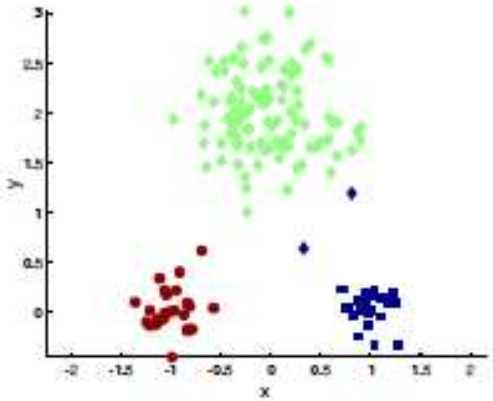
1. Το μέτρο SSE είναι καλός δείκτης της ποιότητας του K-means διότι ουσιαστικά ο αλγόριθμος προσπαθεί επαναληπτικά να μειώσει την απόσταση όλων σημείων από ένα σημείο της συστάδας.
2. Αν έχουμε δύο διαφορετικές εξόδους του K-means προτιμάμε την έξοδο που έχει το μικρότερο SSE.
3. Ένας τρόπος να μειωθεί το SSE είναι αυξάνοντας το K δηλαδή των αριθμών των συστάδων.

Αδυναμίες του K-means

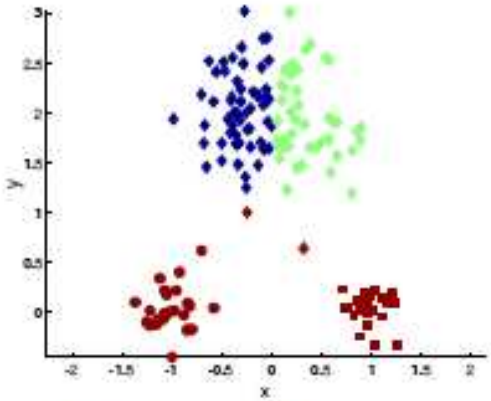
Το αποτέλεσμα εξαρτάται από την επιλογή των αρχικών σημείων



Αρχικά σημεία



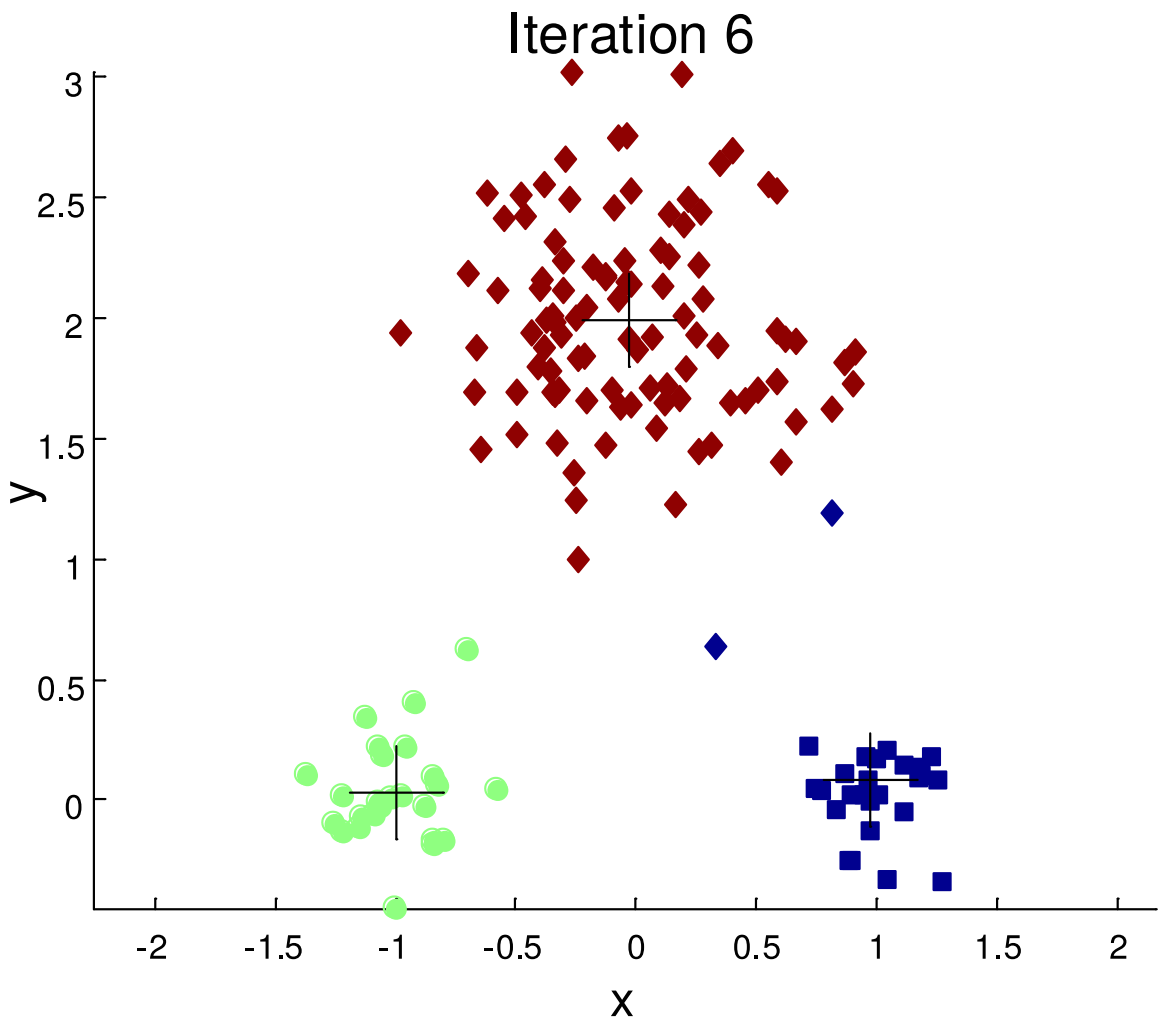
Βέλτιστη συσταδοποίηση



Υπό-βέλτιστη συσταδοποίση

Αδυναμίες του K-means

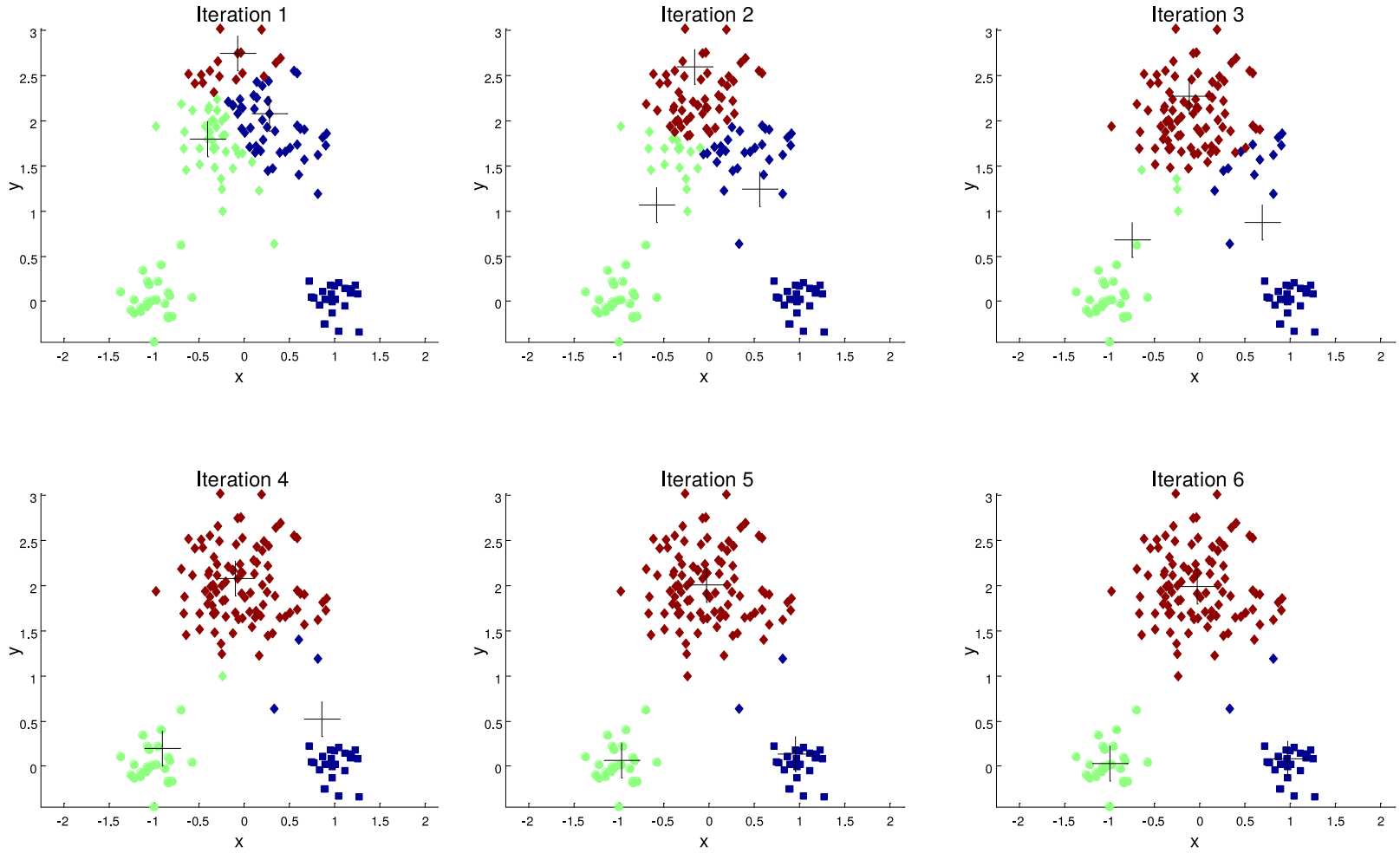
Το αποτέλεσμα εξαρτάται από την επιλογή των αρχικών σημείων



(Καλό σενάριο)

Αδυναμίες του K-means

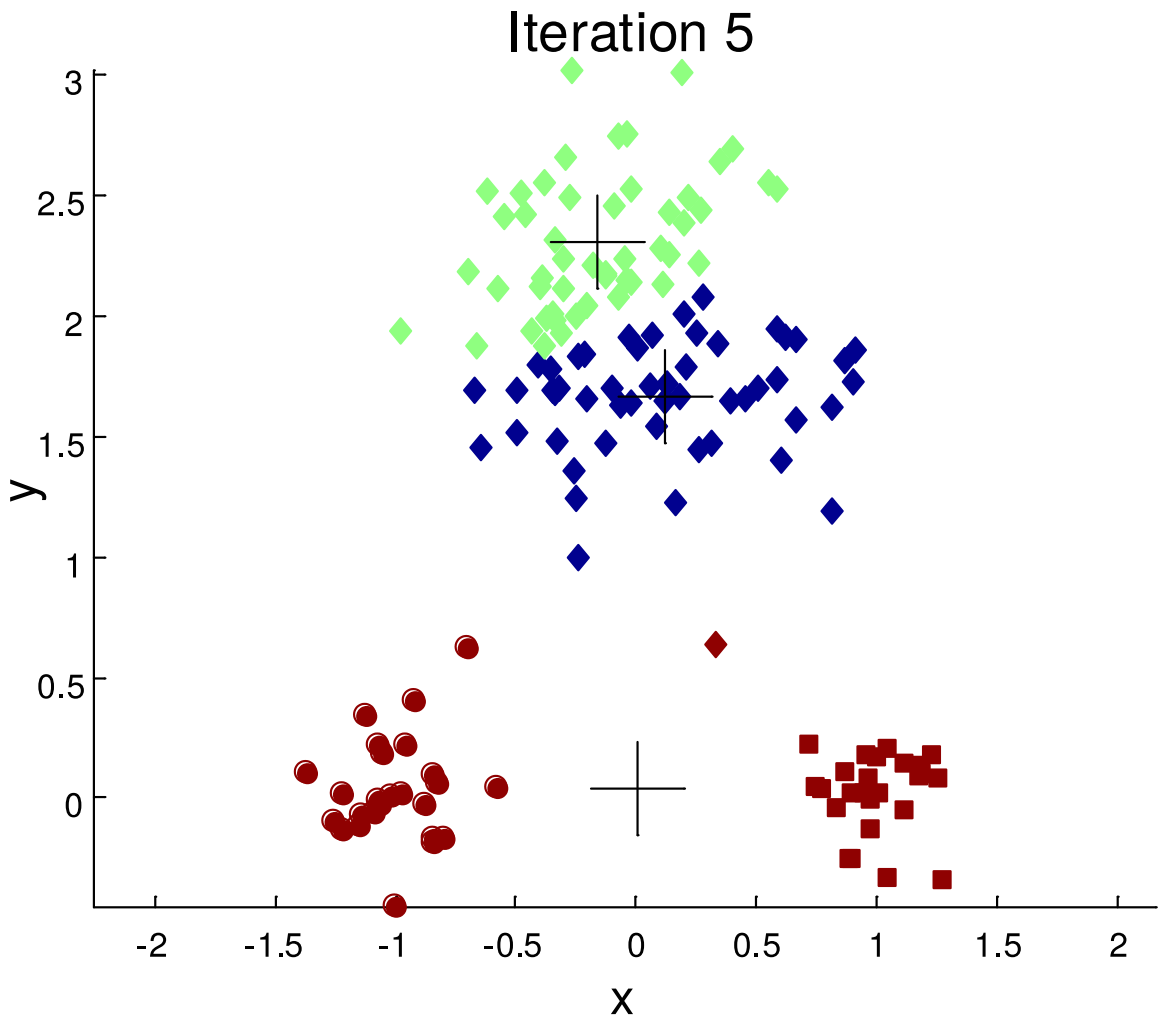
Το αποτέλεσμα εξαρτάται από την επιλογή των αρχικών σημείων



(Καλό σενάριο)

Αδυναμίες του K-means

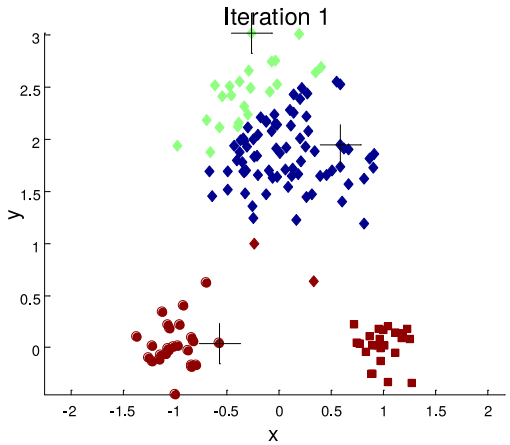
Το αποτέλεσμα εξαρτάται από την επιλογή των αρχικών σημείων

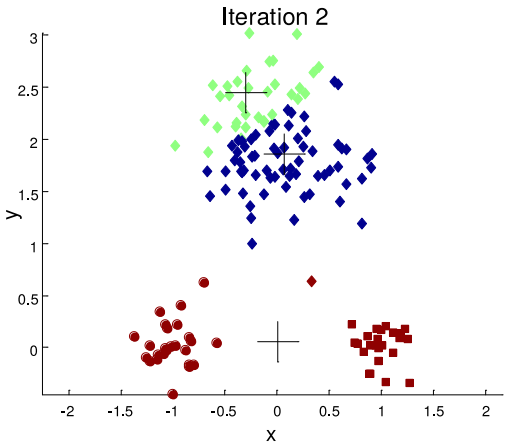


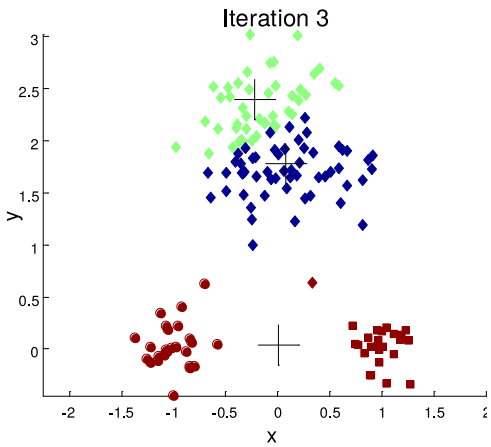
(Κακό σενάριο)

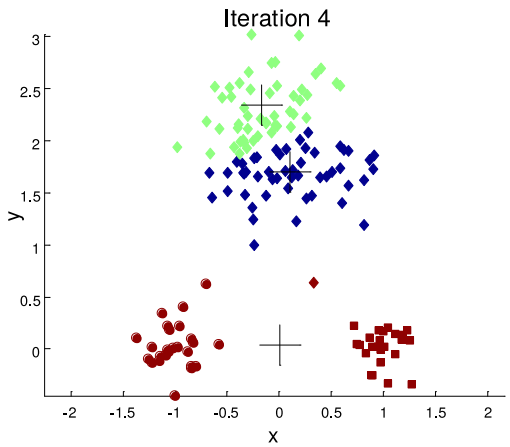
Αδυναμίες του K-means

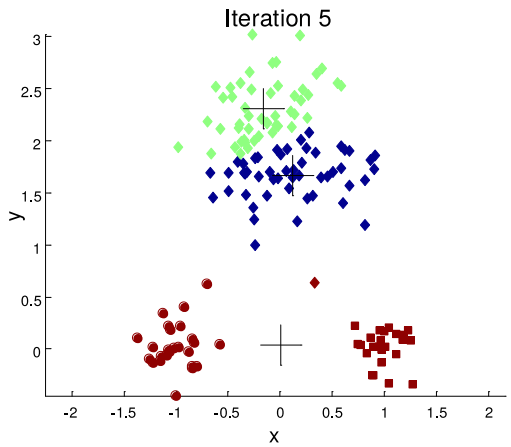
Το αποτέλεσμα εξαρτάται από την επιλογή των αρχικών σημείων











(Κακό σενάριο)

Αδυναμίες του K-means

Το αποτέλεσμα εξαρτάται από την επιλογή των αρχικών σημείων

Παρατήρηση

Αν υπάρχουν K ''πραγματικές'' συστάδες, τότε η πιθανότητα να επιλέξουμε τυχαία ένα κέντρο από κάθε πραγματική συστάδα είναι μικρή.
Συγκεκριμένα, αν όλες οι συστάδες έχουν μέγεθος n, η πιθανότητα είναι

$$P = \frac{\text{αριθμός τρόπων να επιλέξουμε ένα κέντρο από κάθε συστάδα}}{\text{αριθμός τρόπων να διαλέξουμε K κέντρα}}$$
$$= \frac{K!n^K}{(Kn)^K} = \frac{K!}{K^K}$$

Για παράδειγμα, αν $K = 10$, η πιθανότητα είναι
$P = 10!/10^{10} = 0.00036$.

Αδυναμίες του K-means

Το αποτέλεσμα εξαρτάται από την επιλογή των αρχικών σημείων

Λύσεις για την επιλογή των αρχικών σημείων

- Πολλές εκτελέσεις του αλγορίθμου. (Βοηθά αλλά πολλές περιπτώσεις)
- Επιλογή πάνω από K αρχικών σημείων και μετά επιλογή K από αυτά (π.χ. τα πιο απομακρυσμένα).
- Σταδιακή επιλογή.
  Επιλογή του πρώτου σημείου τυχαία.
  Για κάθε ένα από τα υπόλοιπα αρχικά σημεία επέλεξε αυτό που είναι πιο μακριά από τα μέχρι τώρα επιλεγμένα.
  (Παρατήρηση: Μπορεί να οδηγήσει στην επιλογή outliers. Δαπανηρός ο υπολογισμός του πιο απομακρυσμένου. Συχνά εφαρμόζεται σε τυχαία δείγματα)

Αδυναμίες του K-means

Μπορεί να προκύψουν άδειες αρχικές συστάδες

Λύσεις

- Επαναλαμβάνουμε ξανά την επιλογή.
- Μπορούμε να επιλέξουμε κάποια από τα σημεία ως αρχικά σημεία.

Μια παραλλαγή του K-means

K-means με διχοτόμηση

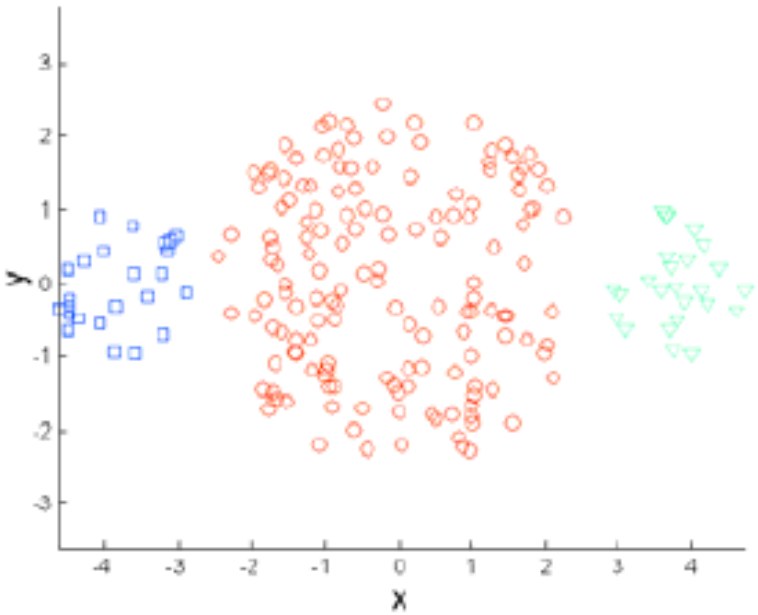
Ο αλγόριθμος K-means με διχοτόμηση

1. Αρχικοποίηση της λίστας των συστάδων ώστε ακριβώς μια συστάδα που περιέχει όλα τα σημεία.
2. Μέχρι να δημιουργηθούν K συστάδες
   1. Επιλογή μιας συστάδας από τη λίστα των συστάδων και διχοτόμηση της επιλεγμένης συστάδας χρησιμοποιώντας τον αλγόριθμο K-means
   2. Προσθήκη στη λίστα των συστάδων εκείνης της συστάδας (από τις δύο που προέκυψαν) με το μικρότερο SSE.

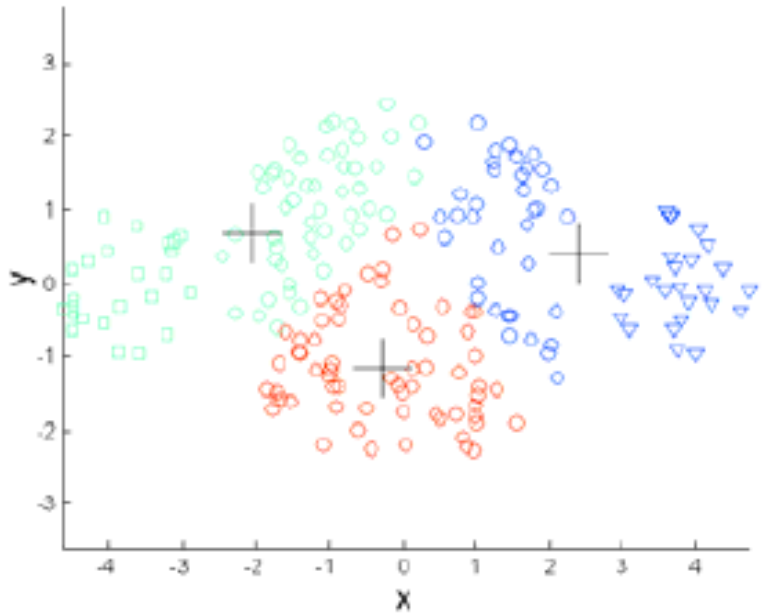
Συνήθως διασπάμε την συστάδα με το μεγαλύτερο SSE ή με το μεγαλύτερο πλήθος σημείων, ή με κάποιο συνδυασμό των δύο.

Αδυναμίες του K-means

Ο K-means έχει αδυναμίες όταν οι συστάδες έχουν διαφορετικά μεγέθη.



Αρχικά σημεία

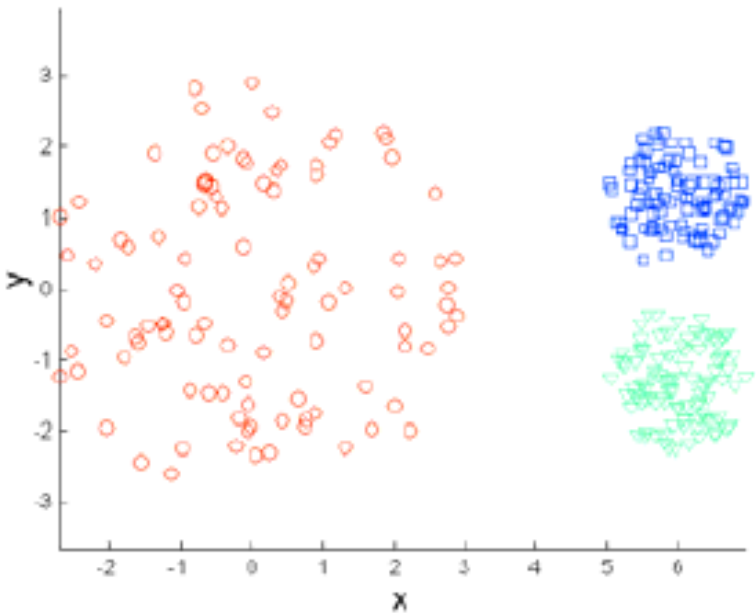


K-means (3 συστάδες)

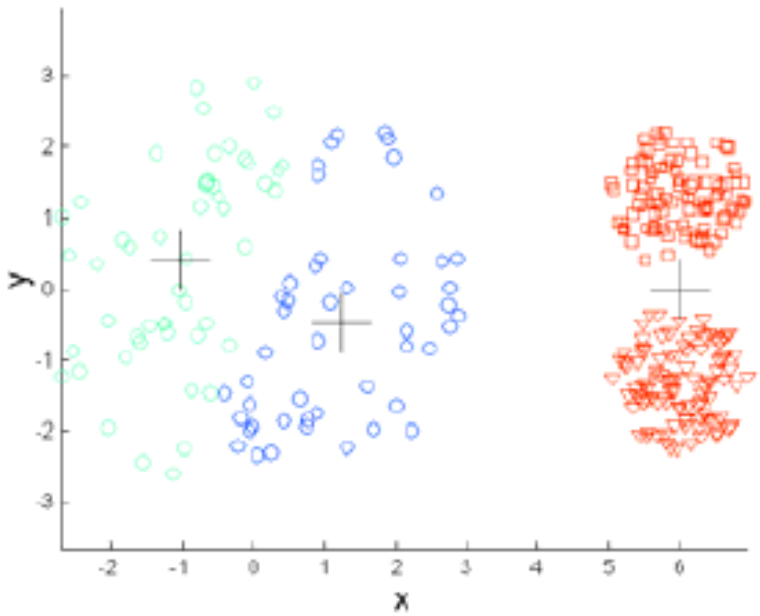
Δεν μπορεί να βρει το μεγάλο κόκκινο, γιατί είναι πολύ μεγαλύτερος από τους άλλους

Αδυναμίες του K-means

Ο K-means έχει αδυναμίες όταν οι συστάδες έχουν διαφορετική πυκνότητα.



Αρχικά σημεία

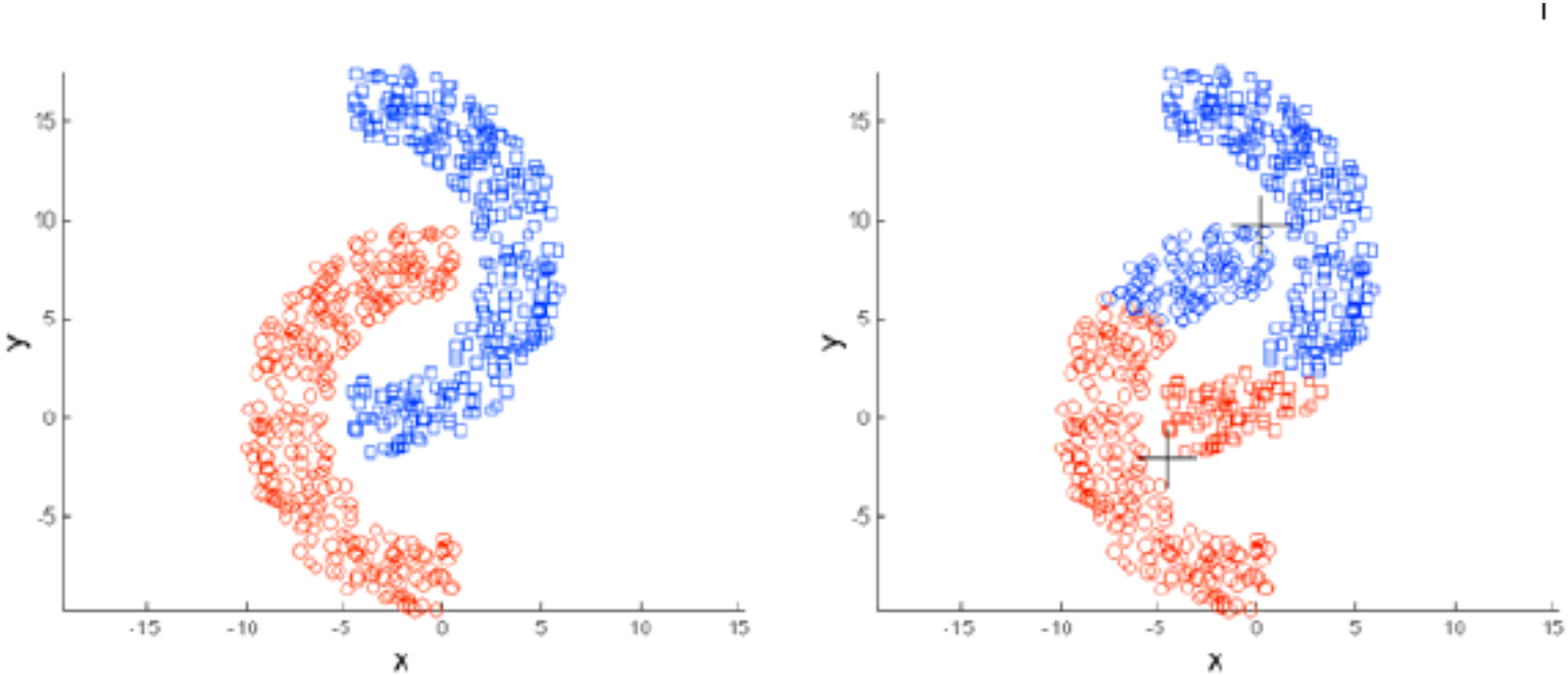


K-means (3 συστάδες)

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

Αδυναμίες του K-means

Ο K-means έχει αδυναμίες όταν οι συστάδες έχουν μη κυκλικό σχήμα.



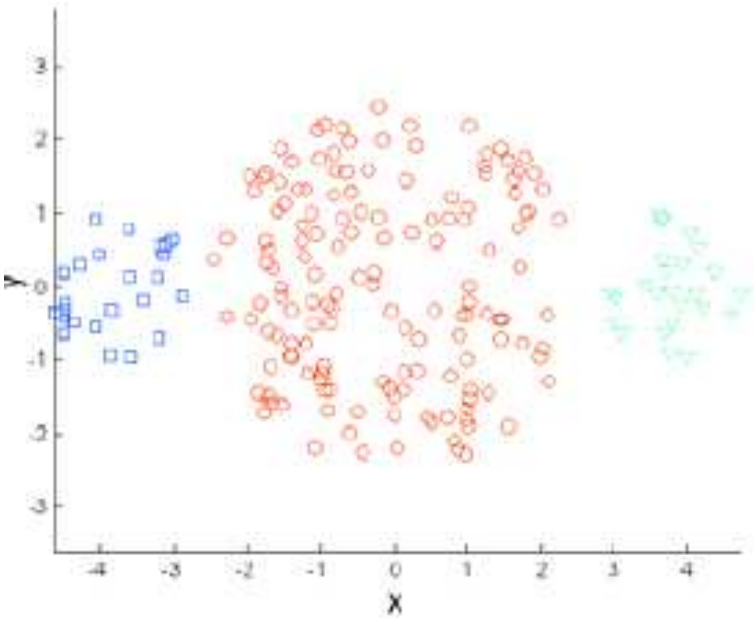
Αρχικά σημεία

K-means (2 συστάδες)

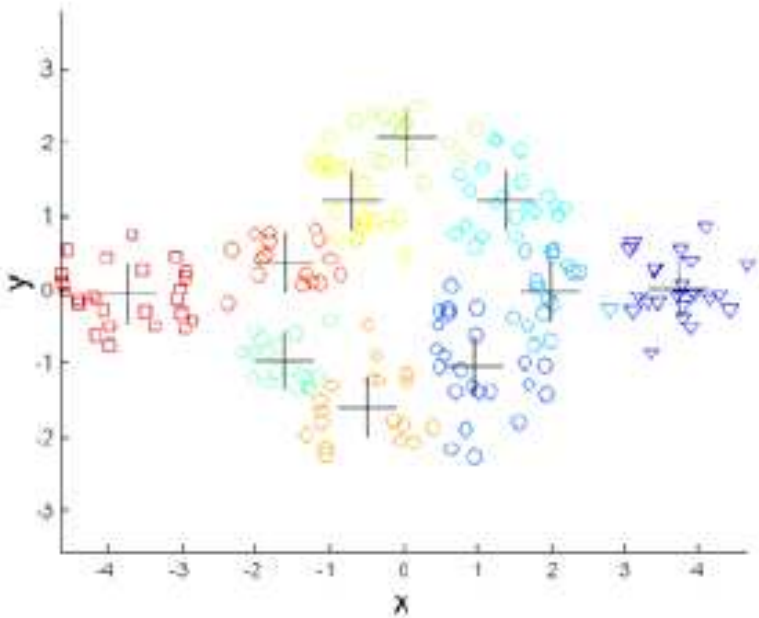
Δεν μπορεί να βρει τις δύο συστάδες γιατί έχουν μη κυκλικά σχήματα

Αδυναμίες του K-means

Ο K-means έχει αδυναμίες όταν οι συστάδες έχουν μη κυκλικό σχήμα.



Αρχικά Σημεία

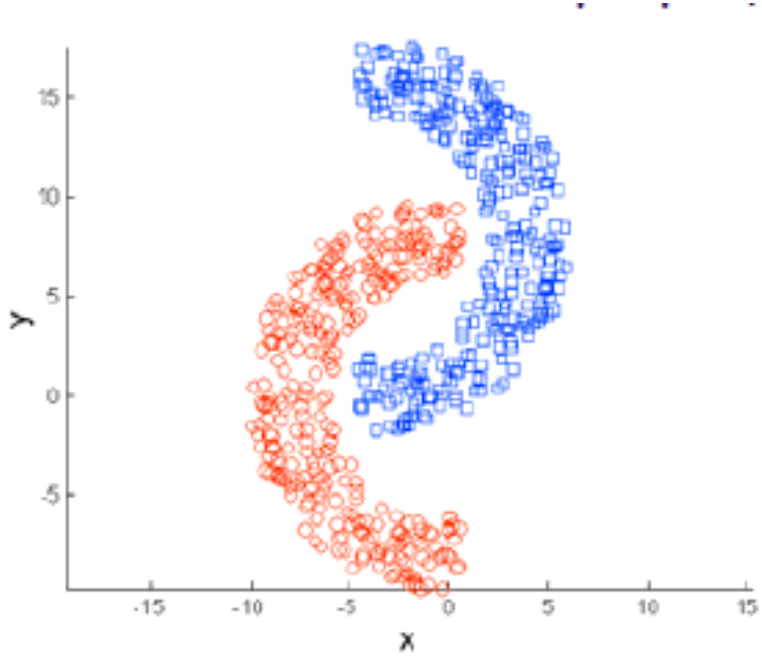


K-means Συστάδες

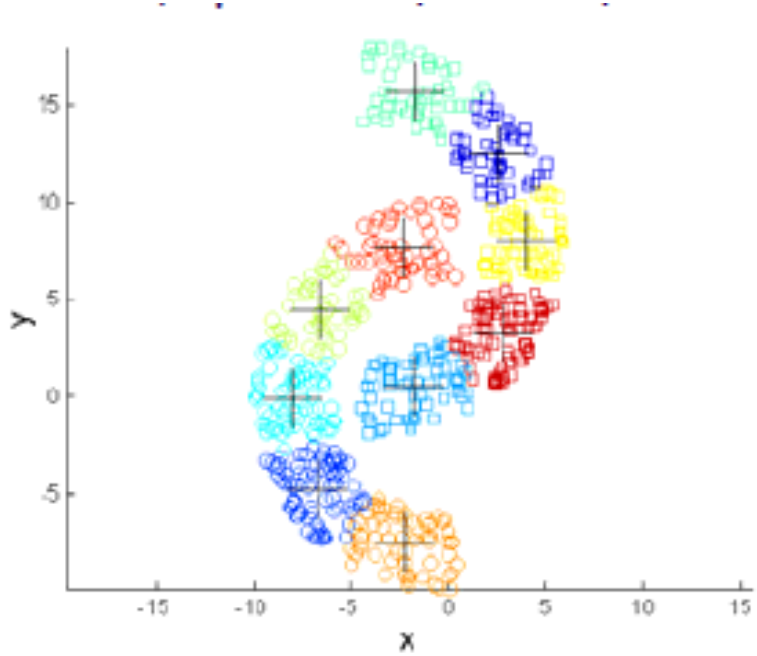
Μια λύση είναι να χρησιμοποιηθούν πολλές συστάδες
Βρίσκει τμήματα των συστάδων, αλλά πρέπει να τα συγκεντρώσουμε

Αδυναμίες του K-means

Ο K-means έχει αδυναμίες όταν οι συστάδες έχουν μη κυκλικό σχήμα.



Αρχικά Σημεία



K-means Συστάδες

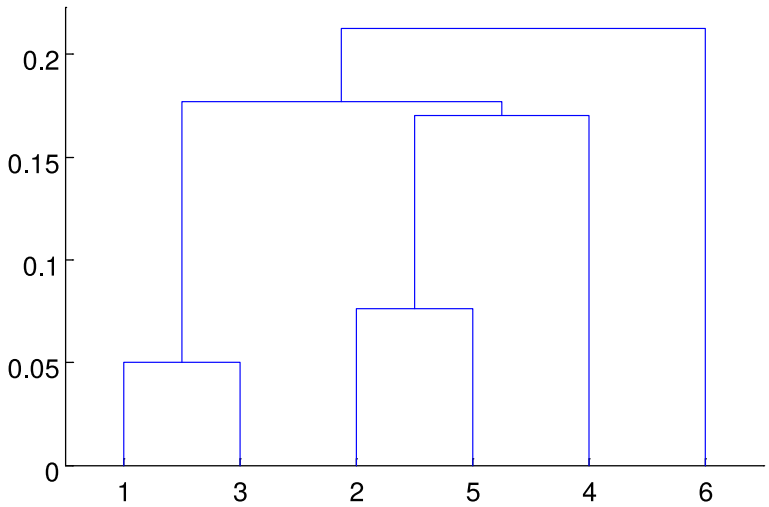
Ο αλγόριθμος K-medoid

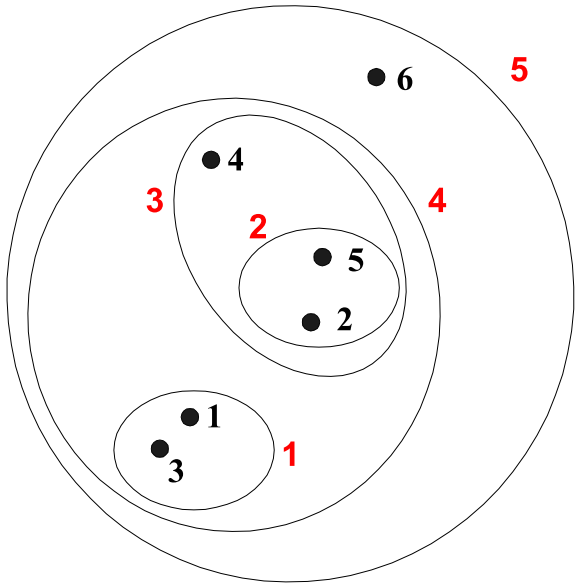
Διαλέγει ένα αντιπροσωπευτικό σημείο από τα δεδομένα και ελαχιστοποιεί την απόσταση από αυτό:
- **Medoid**: το πιο κεντρικο σημείο της συστάδας (αντί να χρησιμοποιεί το mean)
Μειώνει την ευαισθησία σε outliers.
Μπορεί να εφαρμοσθεί σε δεδομένα οποιουδήποτε τύπου (π.χ. και για κατηγορικά δεδομένα).

Ιεραρχική συσταδοποίηση (Hierarchical clustering)

Ιεραρχική συσταδοποίηση

Παράγει ένα σύνολο από εμφωλευμένες συστάδες οργανωμένες σε ένα ιεραρχικό δένδρο.
Μπορεί να παρασταθεί με ένα δενδρόγραμμα.





Ιεραρχική συσταδοποίηση

Δεν χρειάζεται να υποθέσουμε ένα συγκεκριμένο αριθμό συστάδων.
Οποιοσδήποτε επιθυμητός αριθμός από συστάδες μπορεί να επιτευχθεί κόβοντας το δενδρόγραμμα στο κατάλληλο επίπεδο.
Οι συστάδες μπορεί να αντιστοιχούν σε λογικές ταξινομήσεις

Ιεραρχική συσταδοποίηση

Υπάρχουν δύο βασικοί τύποι ιεραρχικής συσταδοποίησης:

- Συσσωρευτικός (Agglomerative)
  Αρχίζει με τα σημεία ως διαφορετικές συστάδες.
  Σε κάθε βήμα συγχωνεύει το πιο **κοντινό ζευγάρι** συστάδων μέχρι να μείνει μόνο μία (ή γενικότερα k) συστάδες
- Διαιρετικός (Divisive)
  Αρχίζει με μόνο μια συστάδα που περιέχει όλα τα σημεία.
  Σε κάθε βήμα διαχωρίζει μια συστάδα, έως ότου κάθε συστάδα να περιέχει μόνο ένα σημείο (ή γενικότερα να δημιουργηθούν k συστάδες).

Ιεραρχική συσταδοποίηση

Οι παραδοσιακοί αλγόριθμοι ιεραρχικής συσταδοποίησης χρησιμοποιούν ένα πίνακα ομοιότητας ή απόστασης.
Κάθε φορά γίνεται διαχωρισμός μιας ομάδας σε δύο (ή συγχώνευση δύο ομάδων σε μια.)

Συσσωρευτική ιεραρχική συσταδοποίηση (ΣΙΣ)

Ο πιο δημοφιλής αλγόριθμος είναι ο συσσωρευτικός που έχει την παρακάτω περιγραφή:

Συσσωρευτικός ιεραρχικός αλγόριθμος

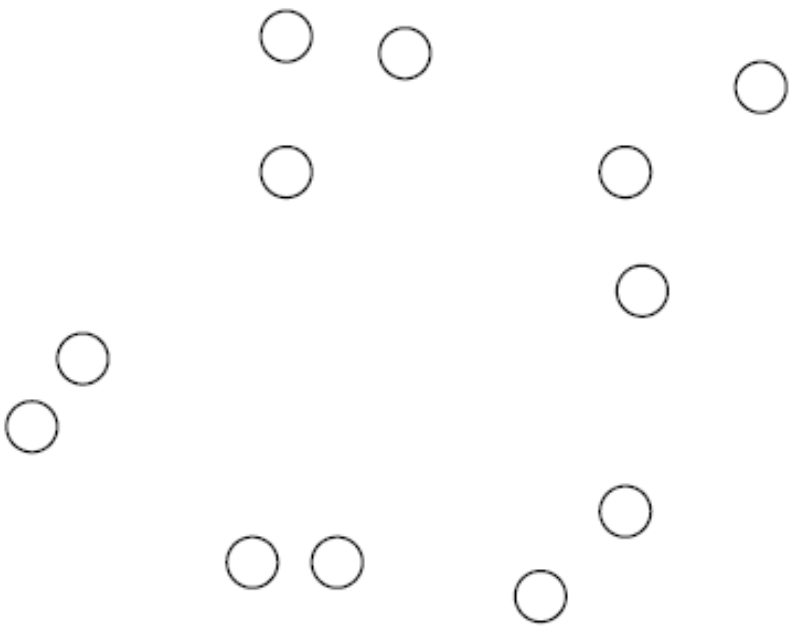
1. Καθε σημείο αποτελεί μια συστάδα
2. Υπολογισμός του πίνακα γειτνίασης των συστάδων
3. Μέχρι να μείνει μόνο μια συστάδα
   1. Συγχώνευση των δύο κοντινότερων συστάδων σε μια.
   2. Ενημέρωση του πίνακα γειτνίασης των συστάδων

Η βασική λειτουργία του αλγορίθμου είναι ο υπολογισμός της γειτνίασης δύο συστάδων.
Υπάρχουν διαφορετικοί αλγόριθμοι με βάση το πως ορίζεται η απόσταση ανάμεσα σε δύο συστάδες

Συσσωρευτική ιεραρχική συσταδοποίηση (ΣΙΣ)

Ένα σχηματικό παράδειγμα

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



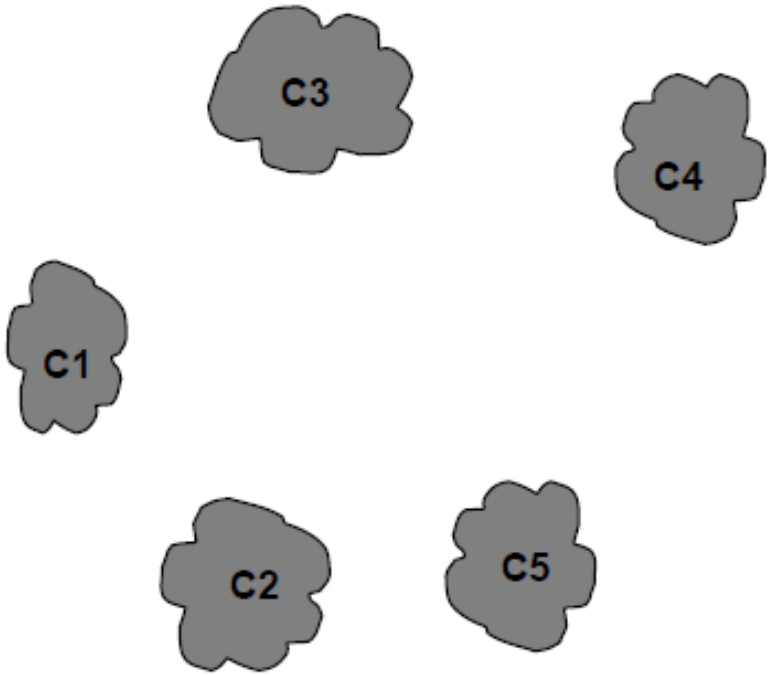
	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Πίνακας Γειτνίασης

Συσσωρευτική ιεραρχική συσταδοποίηση (ΣΙΣ)

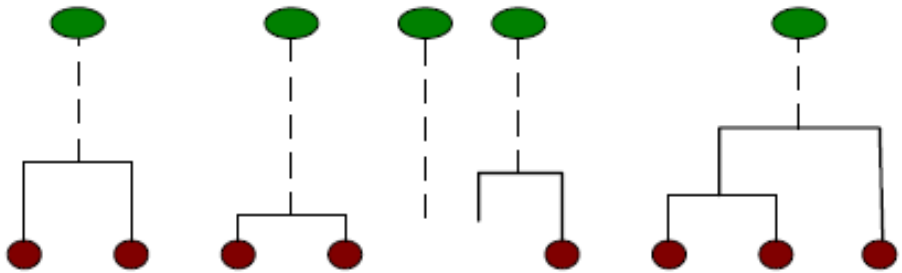
Ένα σχηματικό παράδειγμα

Μετά από κάποιες συγχωνεύσεις, έχουμε κάποιες συστάδες



	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

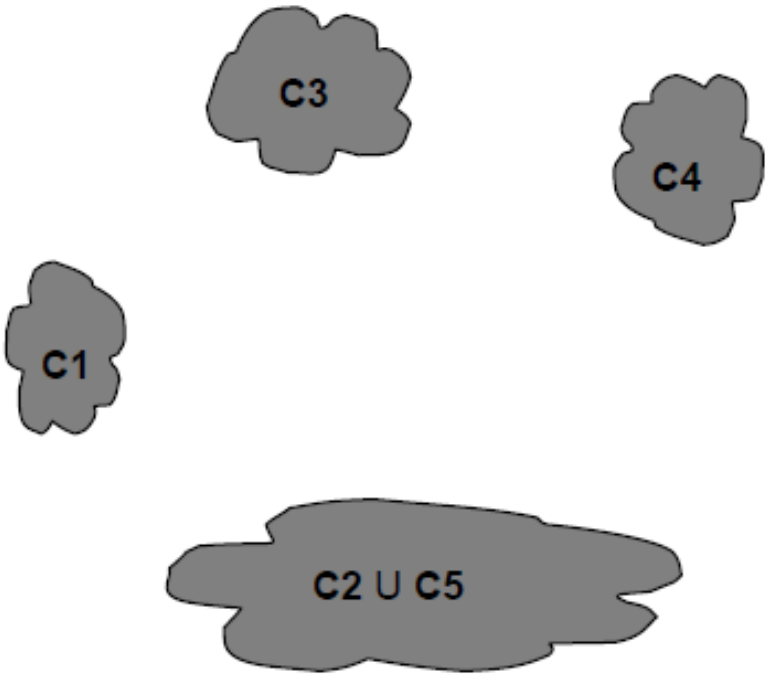
Πίνακας Γειτνίασης



Συσσωρευτική ιεραρχική συσταδοποίηση (ΣΙΣ)

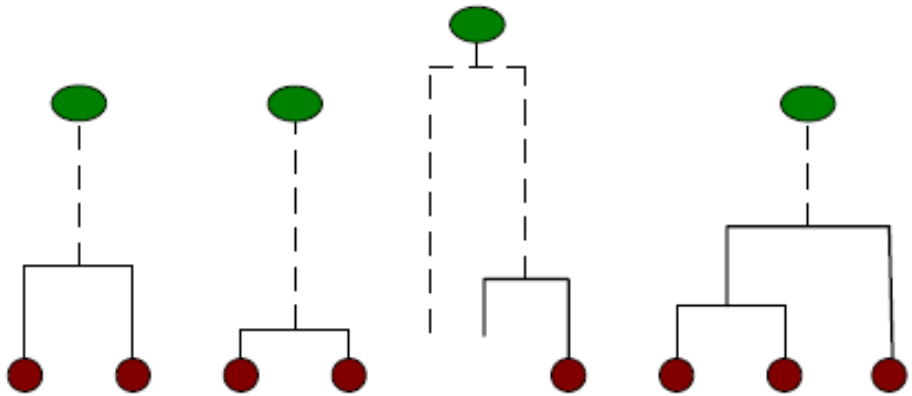
Ένα σχηματικό παράδειγμα

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

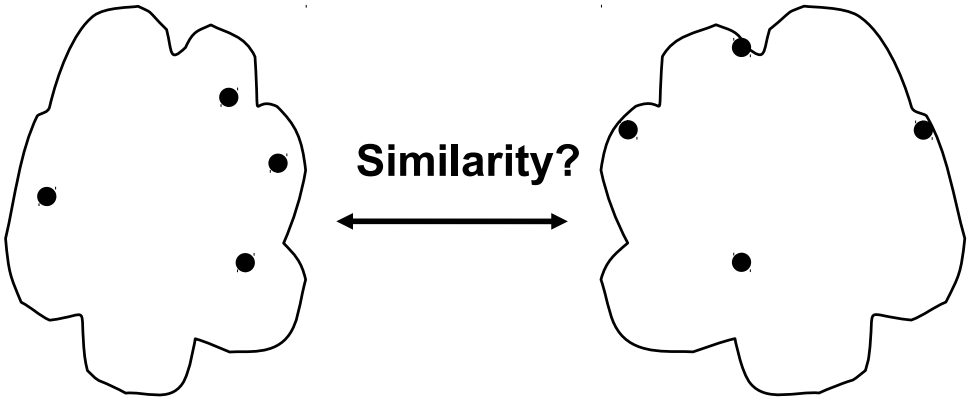


	C1	C2 U C5	C3	C4
C1		?		
C2 U C5	?	?	?	?
C3		?		
C4		?		

Πίνακας Γειτνίασης

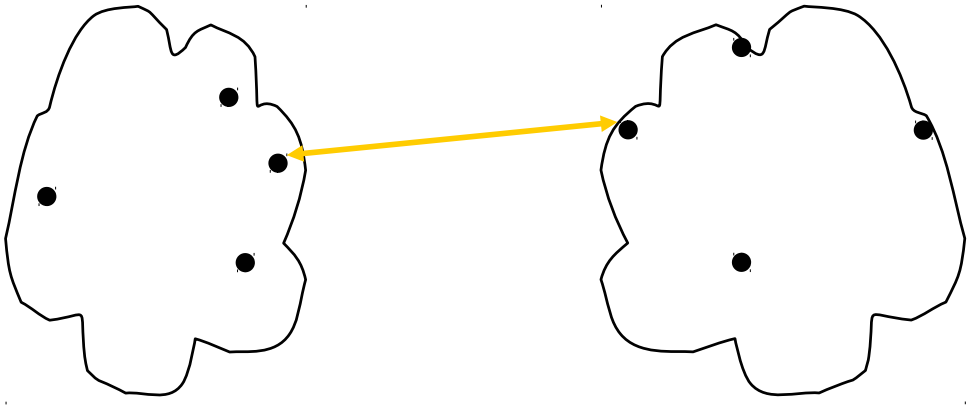


Τρόποι ορισμού απόστασης ανάμεσα σε συστάδες



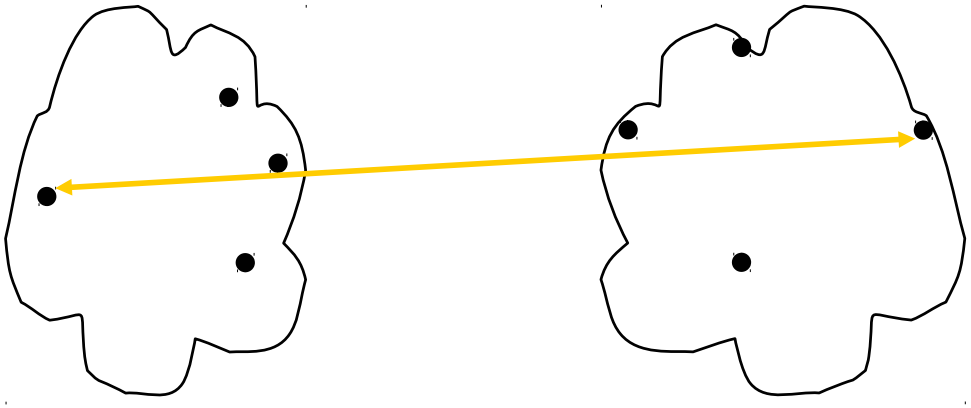
1. MIN ή Single Link
2. MAX ή Complete Linkage
3. Μέση απόσταση
4. Απόσταση ανάμεσα στα κέντρα τους
5. Μέθοδος του Ward

Τρόποι ορισμού απόστασης ανάμεσα σε συστάδες



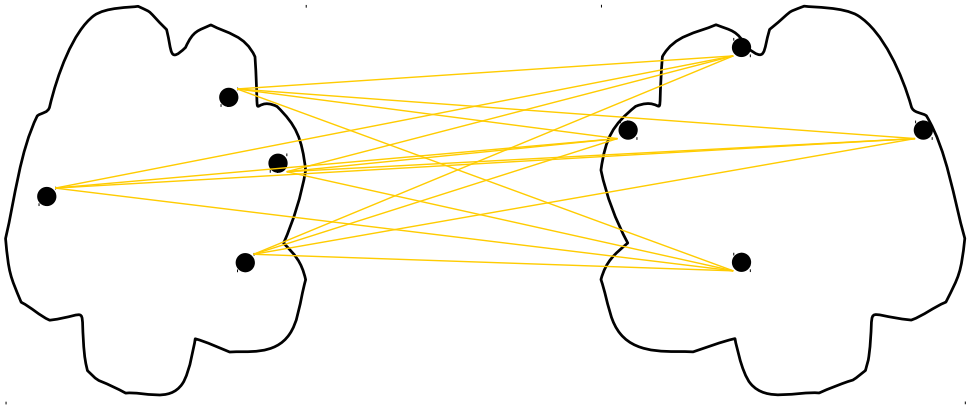
1. MIN ή Single Link
2. MAX ή Complete Linkage
3. Μέση απόσταση
4. Απόσταση ανάμεσα στα κέντρα τους
5. Μέθοδος του Ward

Τρόποι ορισμού απόστασης ανάμεσα σε συστάδες



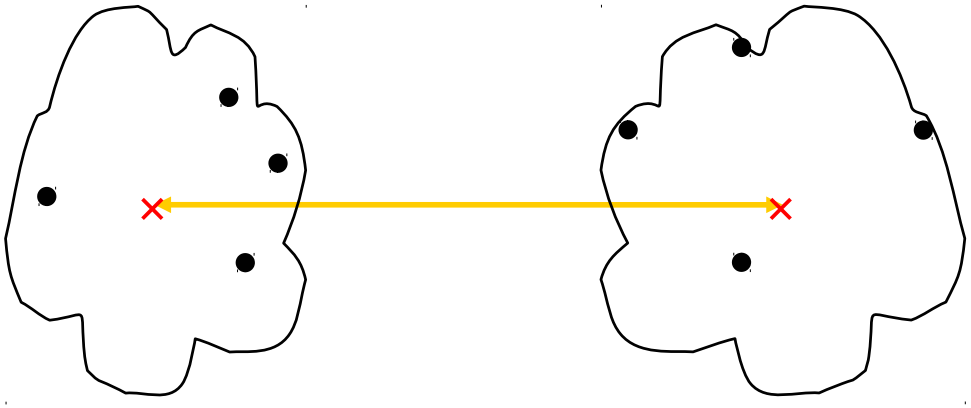
1. MIN ή Single Link
2. MAX ή Complete Linkage
3. Μέση απόσταση
4. Απόσταση ανάμεσα στα κέντρα τους
5. Μέθοδος του Ward

Τρόποι ορισμού απόστασης ανάμεσα σε συστάδες



1. MIN ή Single Link
2. MAX ή Complete Linkage
3. Μέση απόσταση
4. Απόσταση ανάμεσα στα κέντρα τους
5. Μέθοδος του Ward

Τρόποι ορισμού απόστασης ανάμεσα σε συστάδες



1. MIN ή Single Link
2. MAX ή Complete Linkage
3. Μέση απόσταση
4. Απόσταση ανάμεσα στα κέντρα τους
5. Μέθοδος του Ward

Συσσωρευτική ιεραρχική συσταδοποίηση (ΣΙΣ)

Χρησιμοποιώντας την απόσταση MIN

MIN ή μοναδικής ακμής ή απλού συνδέσμου (single link)

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

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

Ονομάζεται και μέθοδος συσταδοποίησης κοντινότερου γείτονα

Συσσωρευτική ιεραρχική συσταδοποίηση (ΣΙΣ)

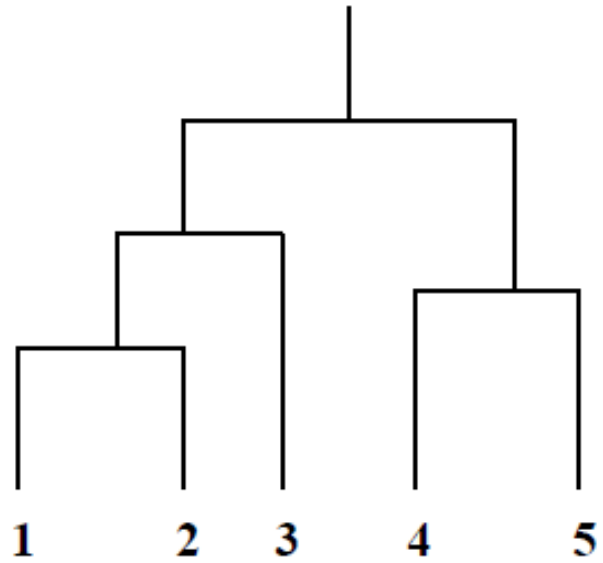
Χρησιμοποιώντας την απόσταση MIN - **Ομοιότητα**

MIN ή μοναδικής ακμής ή απλού συνδέσμου (single link)

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

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

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



Προσοχή: ομοιότητα –> τα ποιο όμοια

Συσσωρευτική ιεραρχική συσταδοποίηση (ΣΙΣ)

Χρησιμοποιώντας την απόσταση MIN - **Απόσταση**

1 (0.4, 0.53)

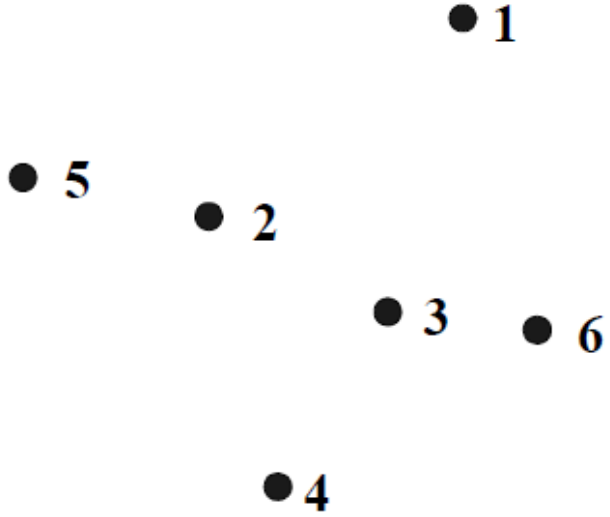
2 (0.22, 0.38)

3 (0.35, 0.32)

4 (0.26, 0.19)

5 (0.08, 0.41)

6 (0.45, 0.30)



Πίνακας απόστασης (Ευκλείδεια)

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

Συσσωρευτική ιεραρχική συσταδοποίηση (ΣΙΣ)

Χρησιμοποιώντας την απόσταση MIN - **Απόσταση**

1 (0.4, 0.53)

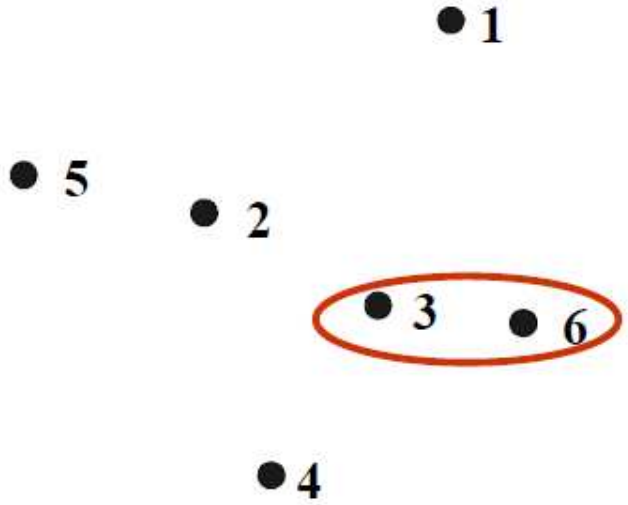
2 (0.22, 0.38)

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**

Συσσωρευτική ιεραρχική συσταδοποίηση (ΣΙΣ)

Χρησιμοποιώντας την απόσταση MIN - **Απόσταση**

1 (0.4, 0.53)

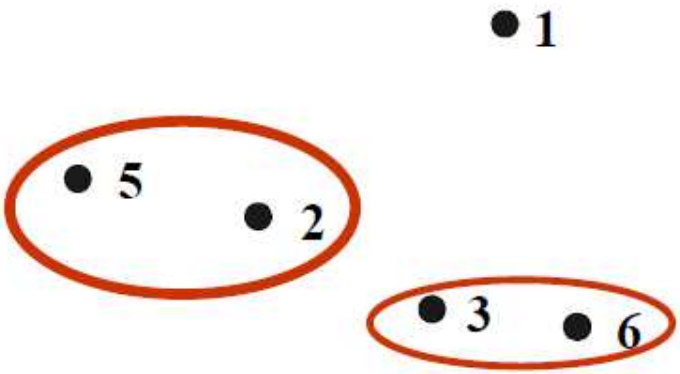
2 (0.22, 0.38)

3 (0.35, 0.32)

4 (0.26, 0.19)

5 (0.08, 0.41)

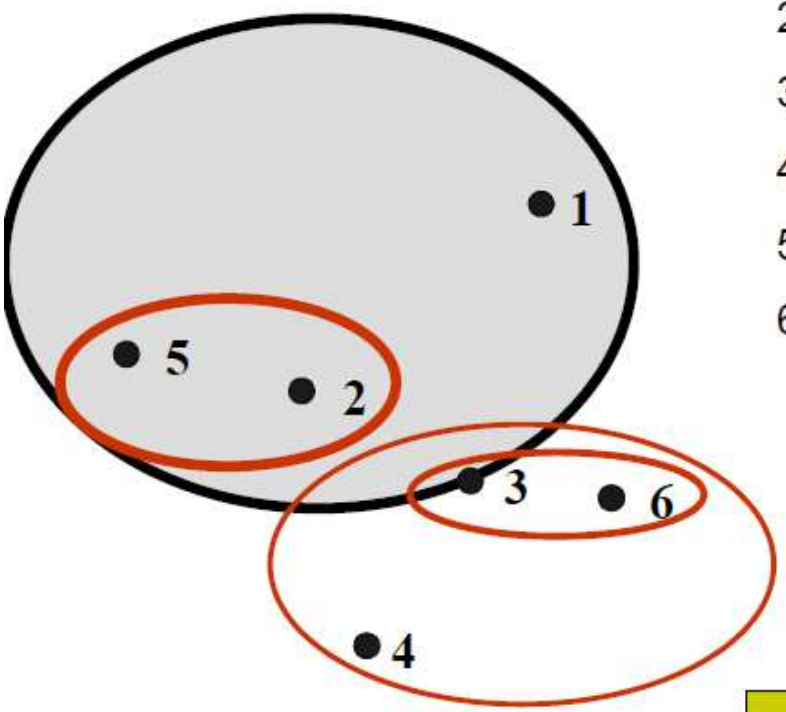
6 (0.45, 0.30)



	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**

Συσσωρευτική ιεραρχική συσταδοποίηση (ΣΙΣ)

Χρησιμοποιώντας την απόσταση MIN - **Απόσταση**



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)

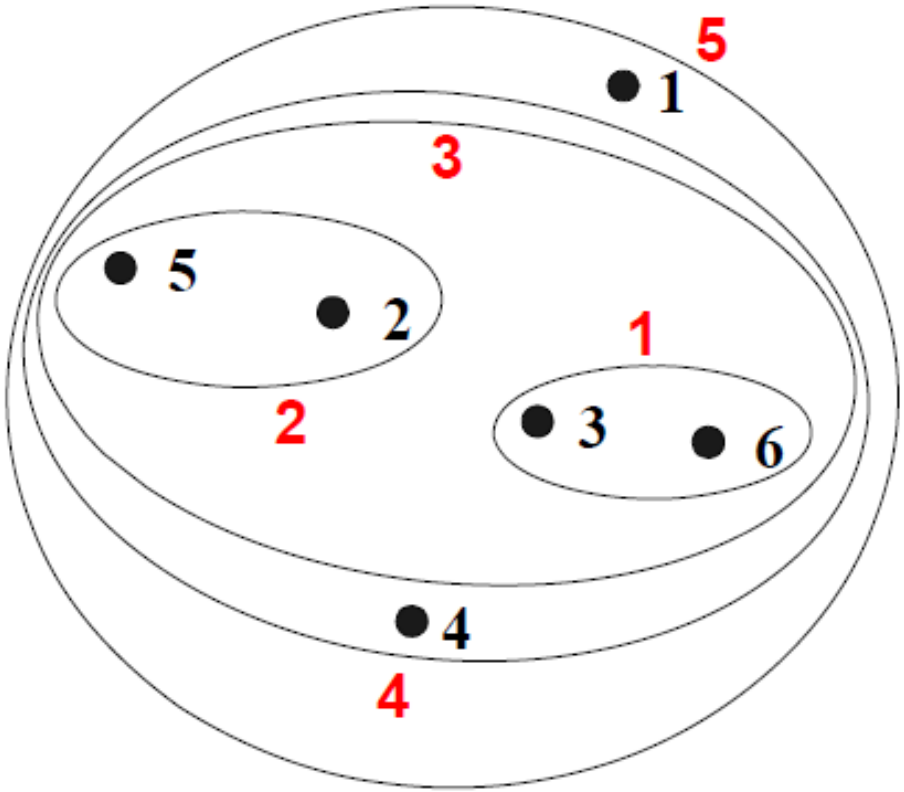
6 (0.45, 0.30)

	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

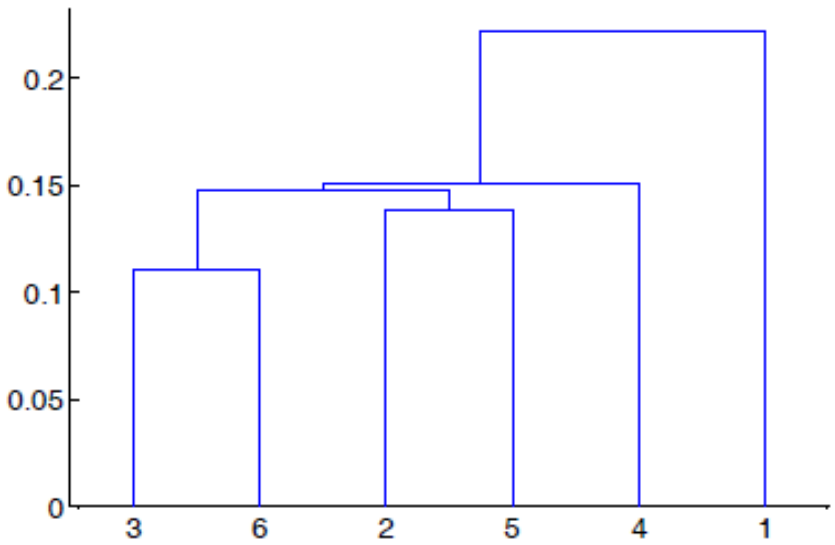
Συσσωρευτική ιεραρχική συσταδοποίηση (ΣΙΣ)

Χρησιμοποιώντας την απόσταση MIN - **Απόσταση**

Τελικά:



Φωλιασμένες Συστάδες



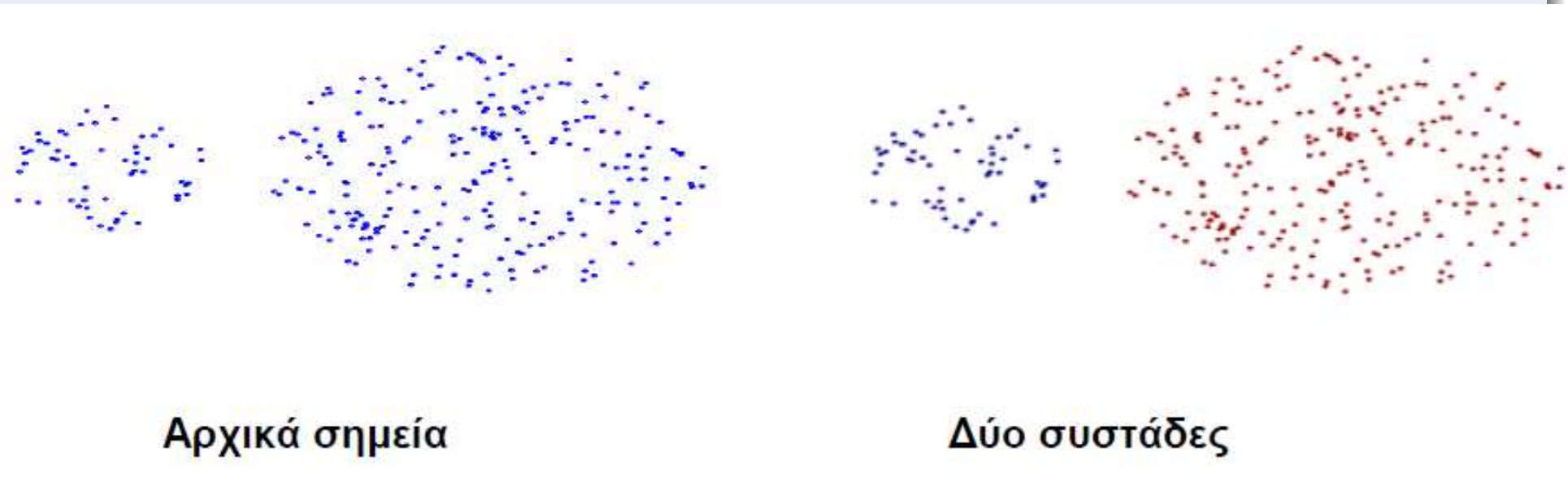
Δεντρόγραμμα

Το δεντρόγραμμα (y-άξονας) δίνει και τις αποστάσεις

Συσσωρευτική ιεραρχική συσταδοποίηση (ΣΙΣ)

Χρησιμοποιώντας την απόσταση MIN

Πλεονεκτήματα



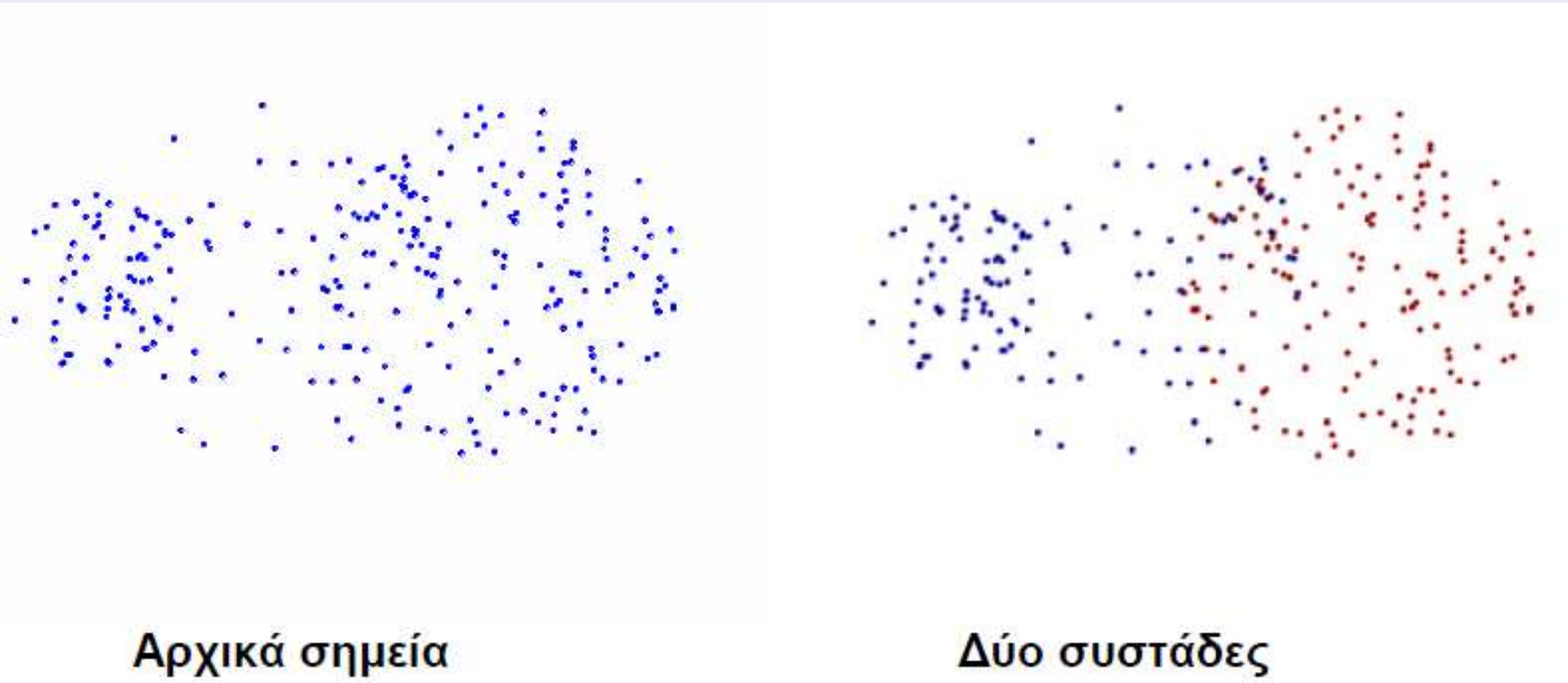
Αναγνωρίζει τις συνεκτικές συστάδες.

Μπορεί να χειριστεί καλά και συστάδες με μη κυκλικό σχήμα.

Συσσωρευτική ιεραρχική συσταδοποίηση (ΣΙΣ)

Χρησιμοποιώντας την απόσταση MIN

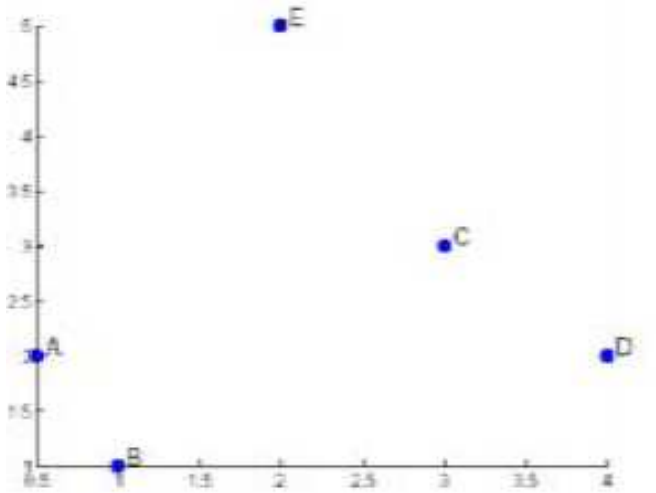
Μειονεκτήματα



Είναι ευαίσθητο σε θόρυβο και outliers.

Συσσωρευτική ιεραρχική συσταδοποίηση (ΣΙΣ)

Χρησιμοποιώντας την απόσταση MAX



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

Συσσωρευτική ιεραρχική συσταδοποίηση (ΣΙΣ)

Χρησιμοποιώντας την απόσταση MAX - **Ομοιότητα**

MAX ή πλήρους συνδεσιμότητας (complete linkage)

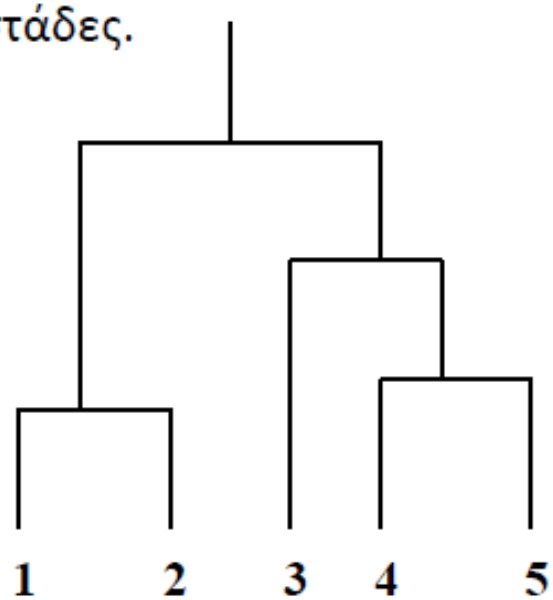
- Αναζητά κλίκες

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

Καθορίζεται από όλα τα ζεύγη τιμών στις δύο συστάδες.

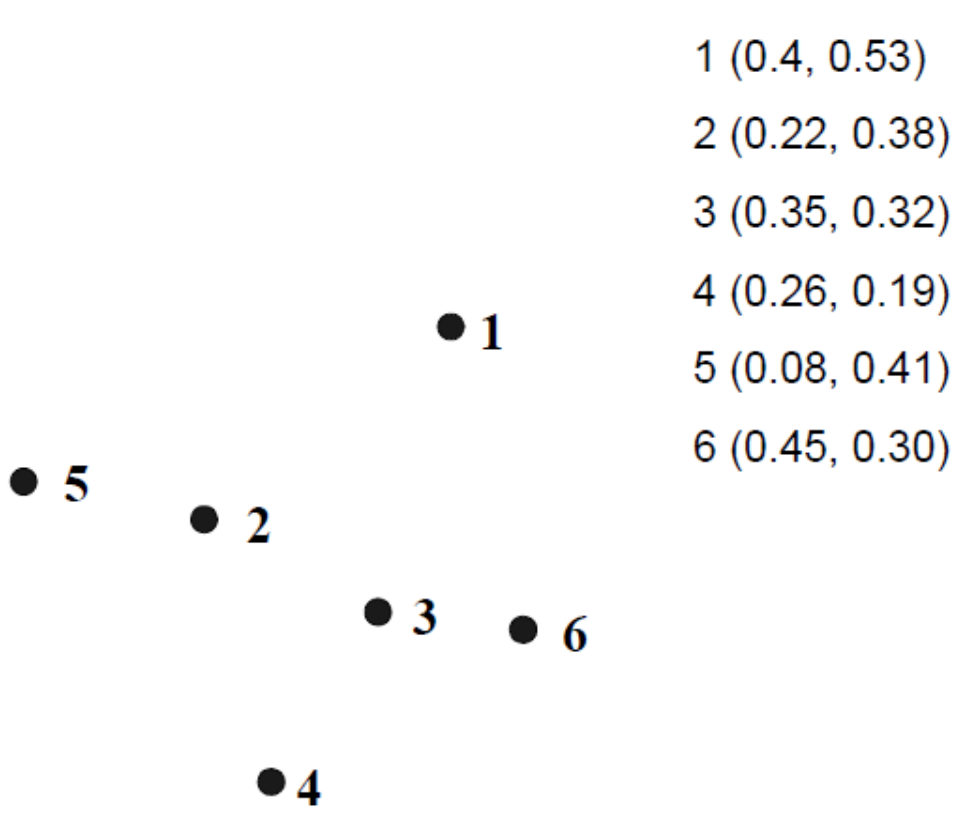
	I1	I2	I3	I4	I5
I1	1.00	0.90	0.10	0.65	0.20
I2	0.90	1.00	0.70	0.60	0.50
I3	0.10	0.70	1.00	0.40	0.30
I4	0.65	0.60	0.40	1.00	0.80
I5	0.20	0.50	0.30	0.80	1.00

ομοιότητα



Συσσωρευτική ιεραρχική συσταδοποίηση (ΣΙΣ)

Χρησιμοποιώντας την απόσταση MAX - **Απόσταση**



	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

Συσσωρευτική ιεραρχική συσταδοποίηση (ΣΙΣ)

Χρησιμοποιώντας την απόσταση MAX - **Απόσταση**

1 (0.4, 0.53)

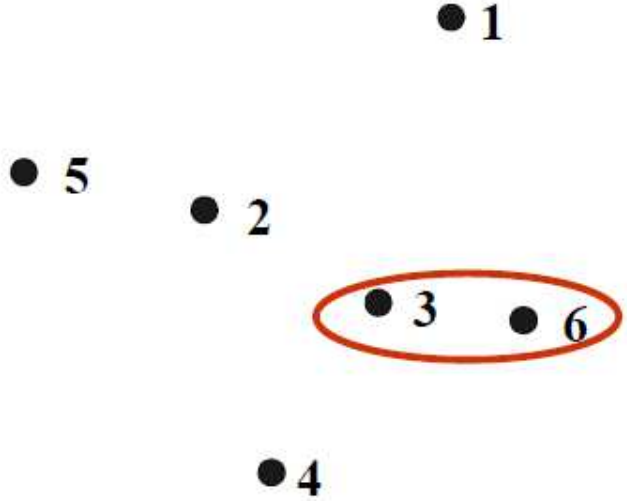
2 (0.22, 0.38)

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

Συσσωρευτική ιεραρχική συσταδοποίηση (ΣΙΣ)

Χρησιμοποιώντας την απόσταση MAX - **Απόσταση**

1 (0.4, 0.53)

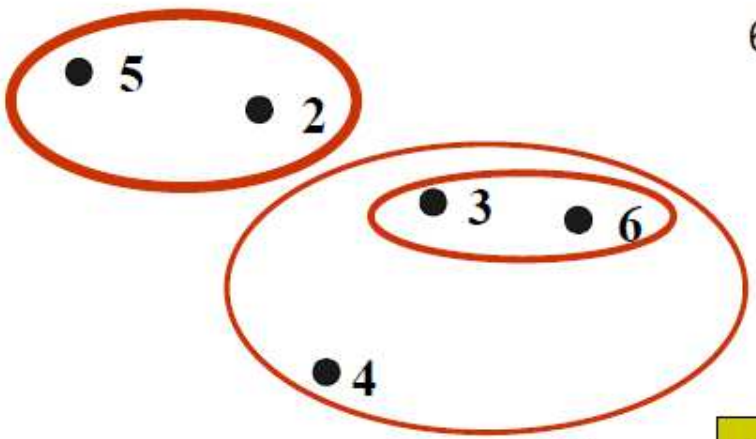
2 (0.22, 0.38)

3 (0.35, 0.32)

4 (0.26, 0.19)

5 (0.08, 0.41)

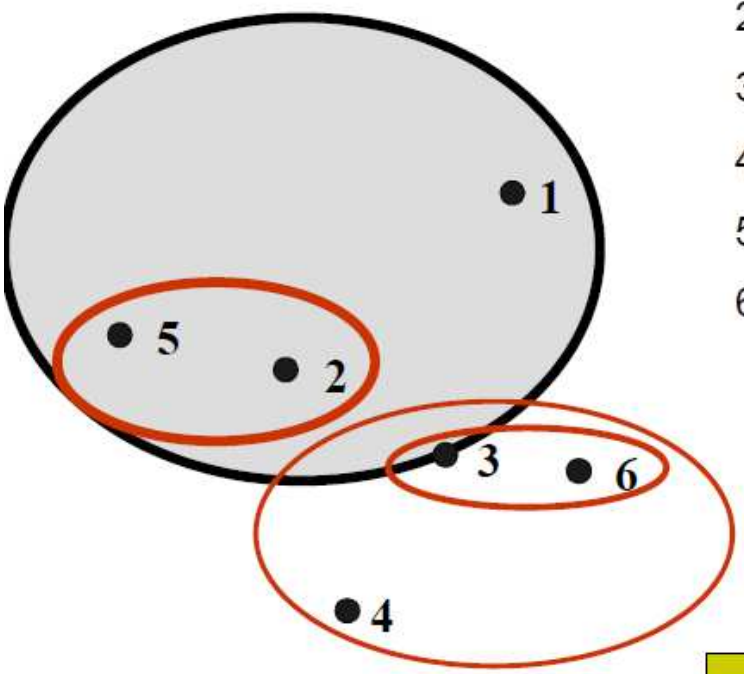
6 (0.45, 0.30)



	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

Συσσωρευτική ιεραρχική συσταδοποίηση (ΣΙΣ)

Χρησιμοποιώντας την απόσταση MAX - **Απόσταση**



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)

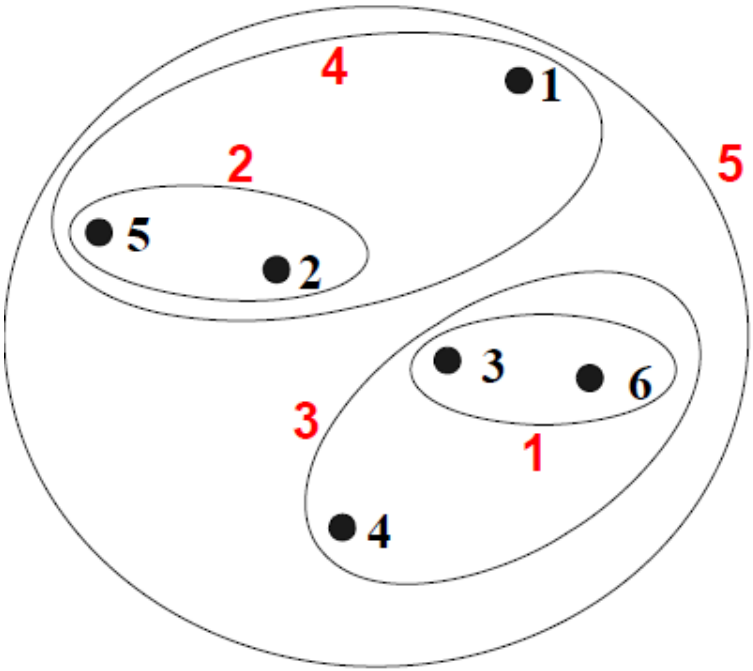
6 (0.45, 0.30)

	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**

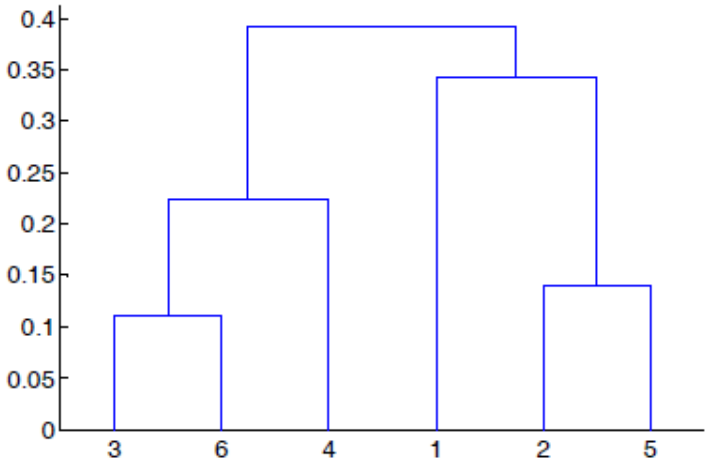
Συσσωρευτική ιεραρχική συσταδοποίηση (ΣΙΣ)

Χρησιμοποιώντας την απόσταση MAX - **Απόσταση**

Τελικά:



Φωλιασμένες Συστάδες

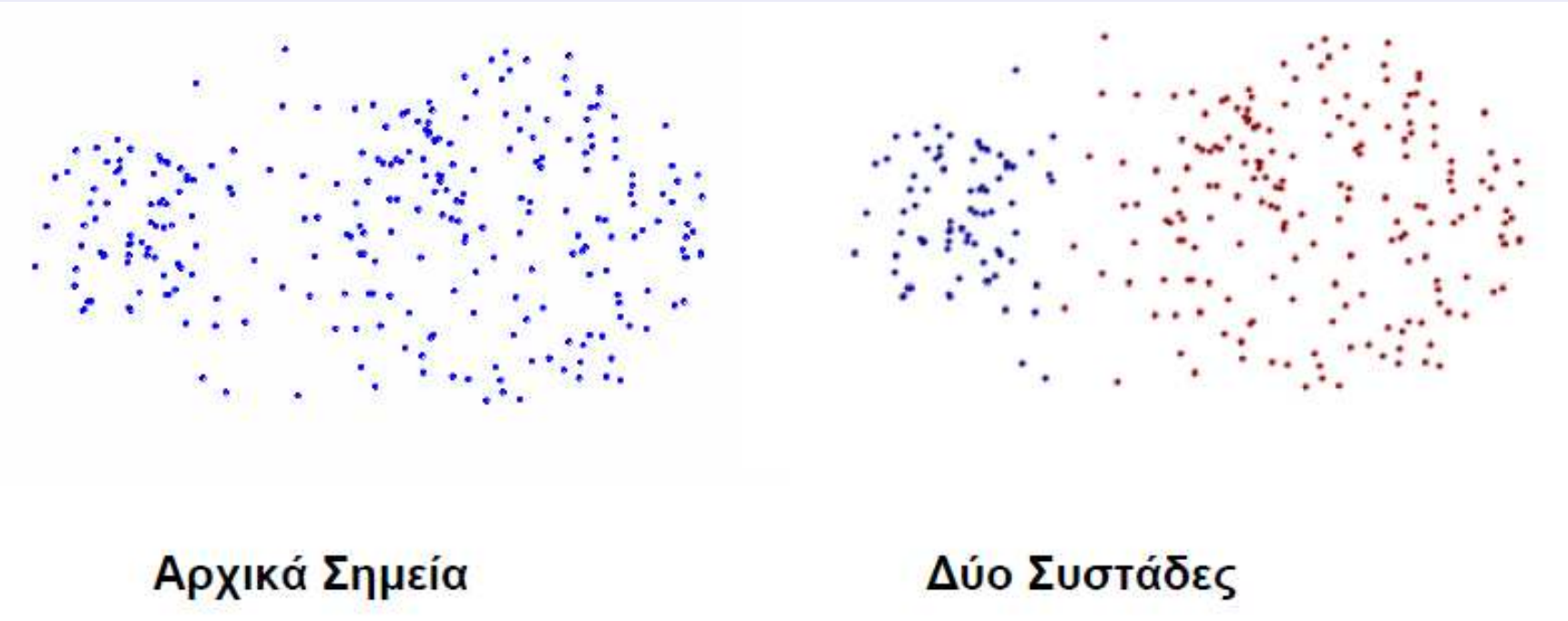


Δεντρόγραμμα

Συσσωρευτική ιεραρχική συσταδοποίηση (ΣΙΣ)

Χρησιμοποιώντας την απόσταση MAX

Πλεονεκτήματα

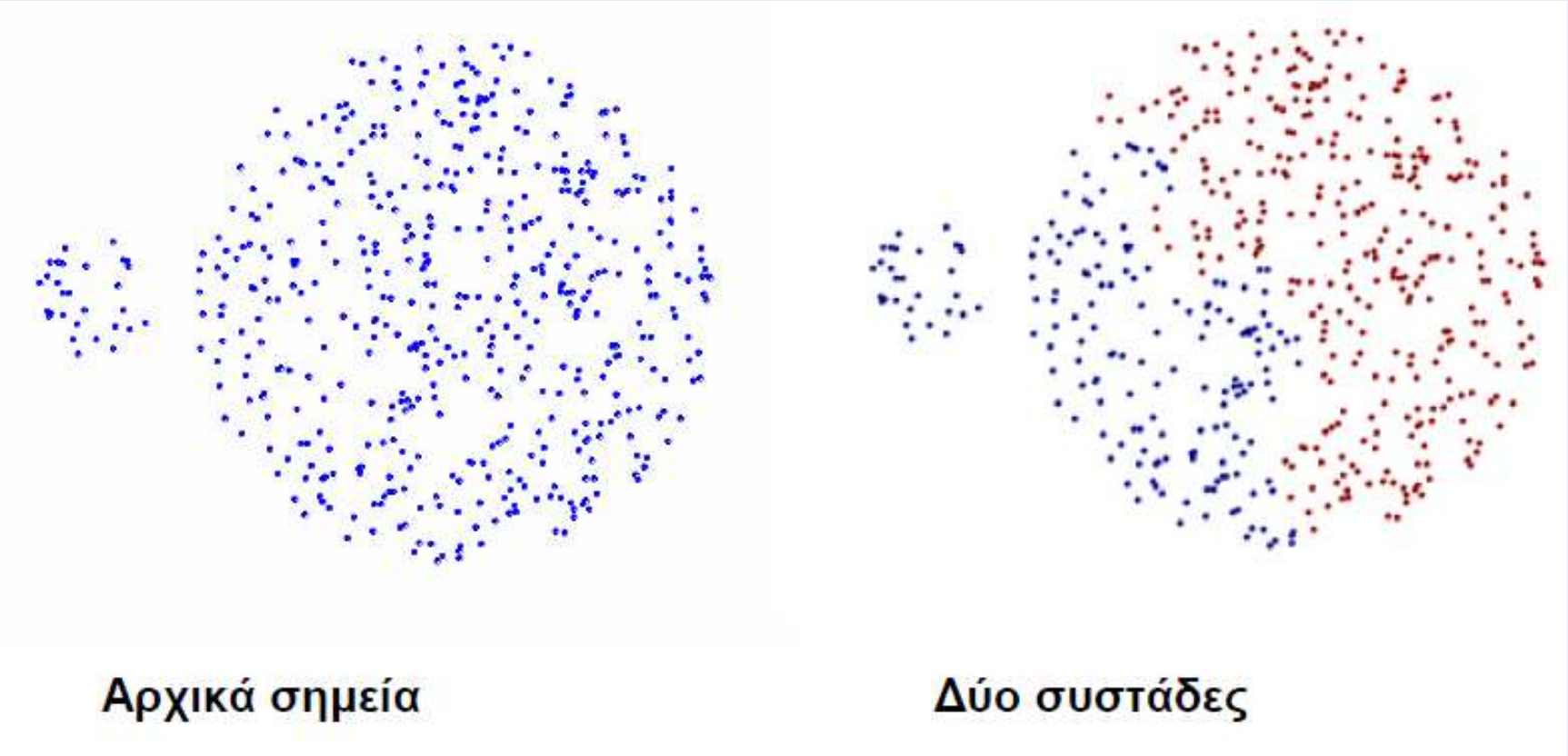


Λιγότερο ευαίσθητο σε θόρυβο και outliers.

Συσσωρευτική ιεραρχική συσταδοποίηση (ΣΙΣ)

Χρησιμοποιώντας την απόσταση MAX

Μειονεκτήματα

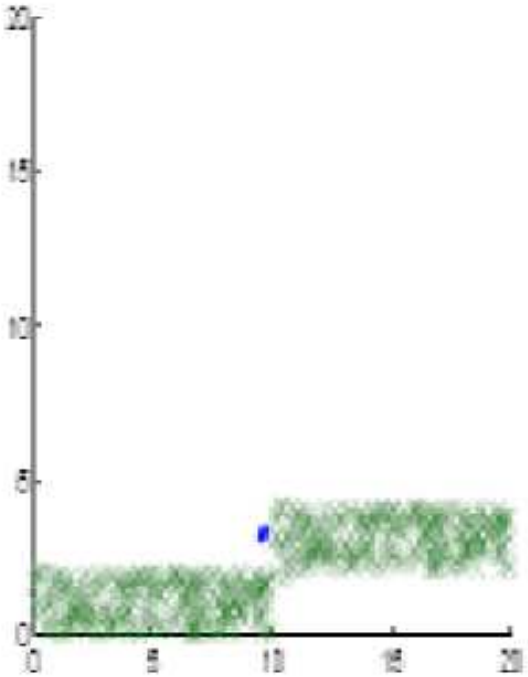


Τείνει να διασπά τις μεγάλες συστάδες

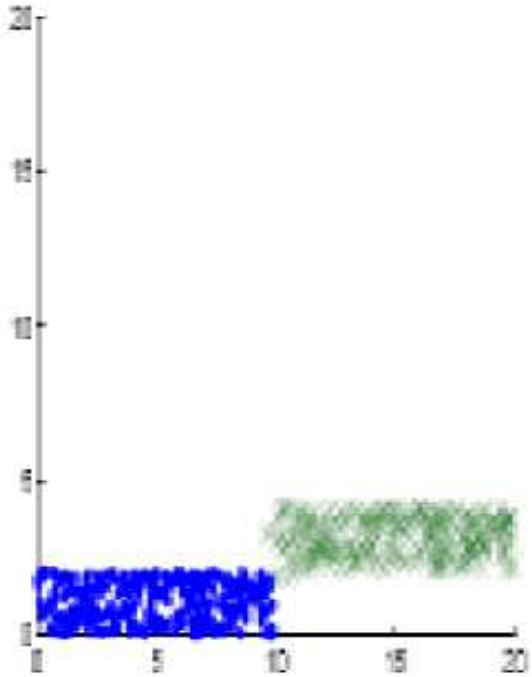
Οδηγεί συνήθως σε κυκλικά σχήματα

Συσσωρευτική ιεραρχική συσταδοποίηση (ΣΙΣ)

Σύγκριση MIN και MAX



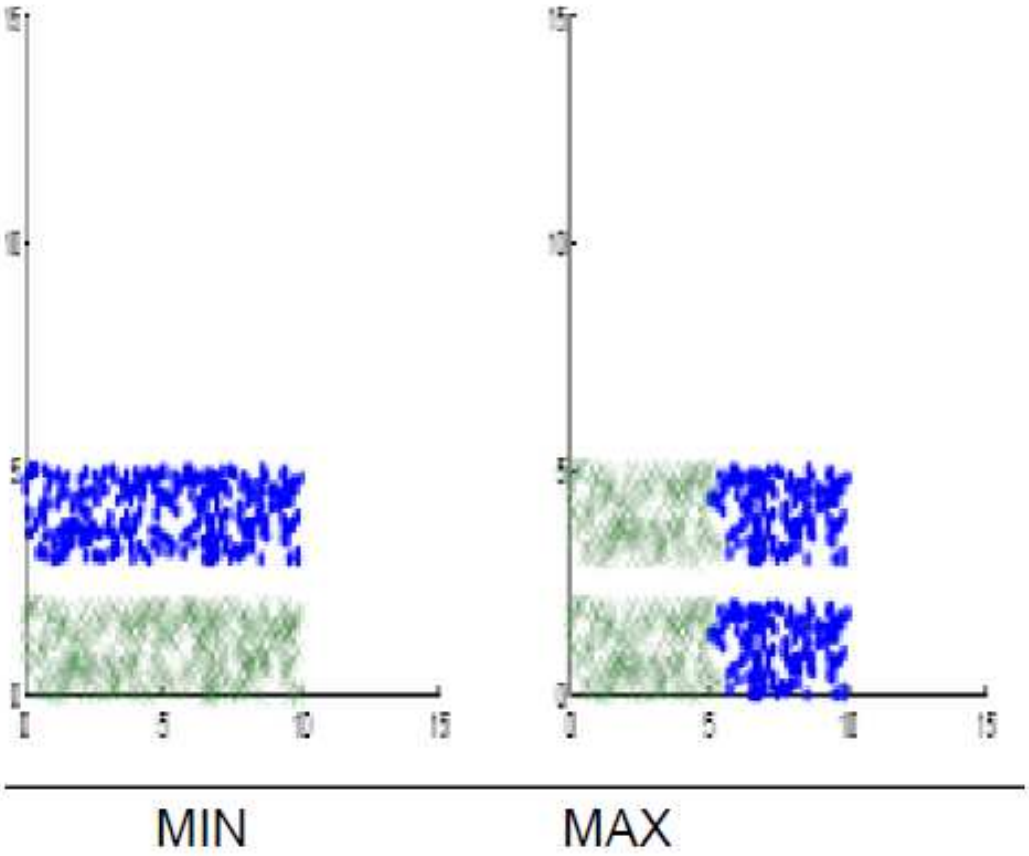
MIN



MAX

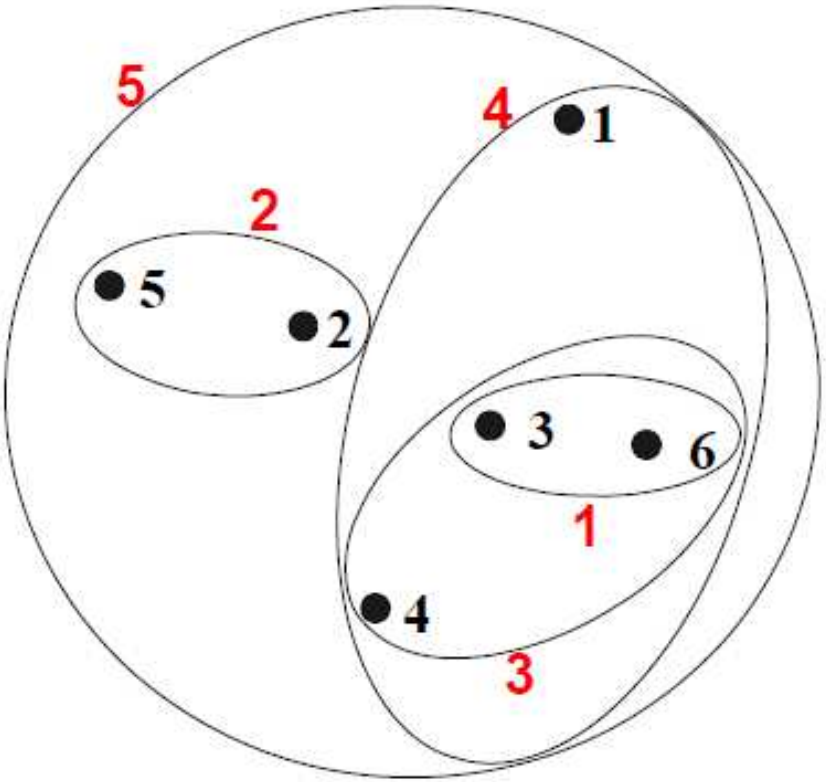
Συσσωρευτική ιεραρχική συσταδοποίηση (ΣΙΣ)

Σύγκριση MIN και MAX

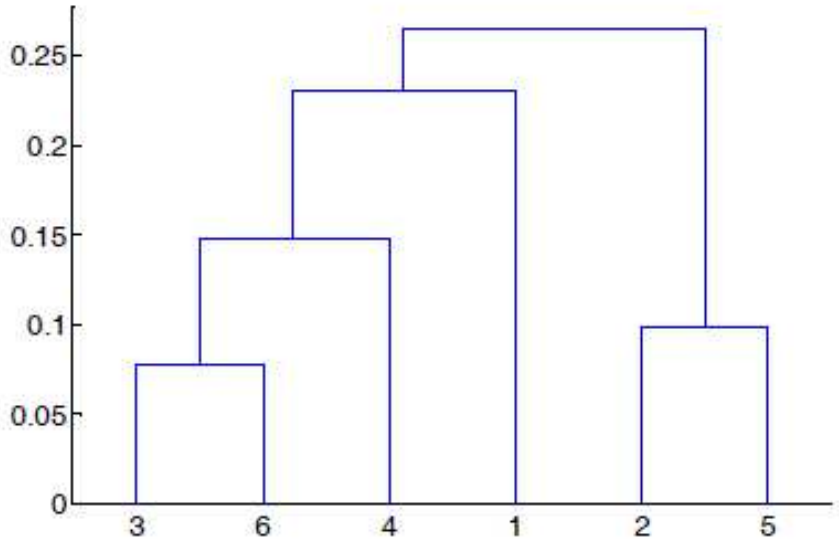


Συσσωρευτική ιεραρχική συσταδοποίηση (ΣΙΣ)

Χρησιμοποιώντας την μέση απόσταση



Φωλιασμένες Συστάδες



Dendrogram

Πλεονεκτήματα: Μικρότερη ευαισθησία σε θόρυβο και outliers

Μειονεκτήματα: Ευνοεί κυκλικές συστάδες.

Μέτρα ομοιότητας και απόστασης

Μέτρα ομοιότητας και απόστασης

Απόσταση

Μια συνάρτηση $d(p, q)$ ονομάζεται **μετρική** ή **απόσταση** αν έχει τις παρακάτω ιδιότητες:

- $d(p, q) \geq 0$ (μη αρνητική)
- $d(p, q) = 0 \Leftrightarrow p = q$ (ανακλαστική)
- $d(p, q) = d(q, p)$ (συμμετρική)
- $d(p, q) \leq d(p, r) + d(r, p)$ (τριγωνική ανισότητα)

Μέτρα ομοιότητας και απόστασης

Ευκλείδεια απόσταση

Αν p και q είναι δύο αντικείμενα με n γνωρίσματα που έχουν αριθμητικές τιμές: $p = (x_1, x_2, \ldots, x_n)$ και $q = (y_1, y_2, \ldots, y_n)$ τότε η Ευκλείδεια απόσταση των p, q ορίζεται

$$d(p, q) = L_2(p, q) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2}$$

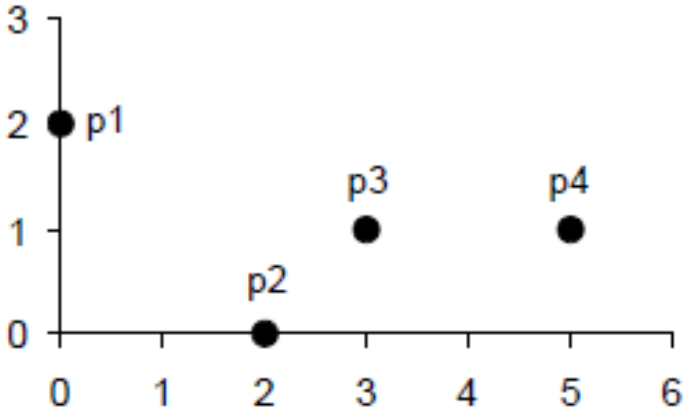
Άσκηση

Να βρεθούν ανα δύο οι Ευκλείδειες αποστάσεις των σημείων $p = (1, -1, 2)$, $q = (1, 0, 1)$ και $r = (-1, 1, -1)$. Ποια σημεία απέχουν την μικρότερη απόσταση;

Μέτρα ομοιότητας και απόστασης

Ευκλείδεια απόσταση

Παράδειγμα



Πίνακας Δεδομένων

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

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

Πίνακας Απόστασης

Μέτρα ομοιότητας και απόστασης

Απόσταση Manhattan

Αν p και q είναι δύο αντικείμενα με n γνωρίσματα που έχουν αριθμητικές τιμές: $p = (x_1, x_2, \ldots, x_n)$ και $q = (y_1, y_2, \ldots, y_n)$ τότε η απόσταση Manhattan των p, q ορίζεται

$$d(p, q) = L_1(p, q) = |x_1 - y_1| + |x_2 - y_2| + \cdots + |x_n - y_n|$$

Άσκηση

Να βρεθούν ανα δύο οι αποστάσεις Manhattan των σημείων $p = (1, -1, 2)$, $q = (1, 0, 1)$ και $r = (-1, 1, -1)$. Ποια σημεία απέχουν την μικρότερη απόσταση;

Μέτρα ομοιότητας και απόστασης

Γενίκευση: Απόσταση Minkowski

Αν p και q είναι δύο αντικείμενα με n γνωρίσματα που έχουν αριθμητικές τιμές: $p = (x_1, x_2, \ldots, x_n)$ και $q = (y_1, y_2, \ldots, y_n)$ τότε η L_p απόσταση Minkowski των p, q ορίζεται

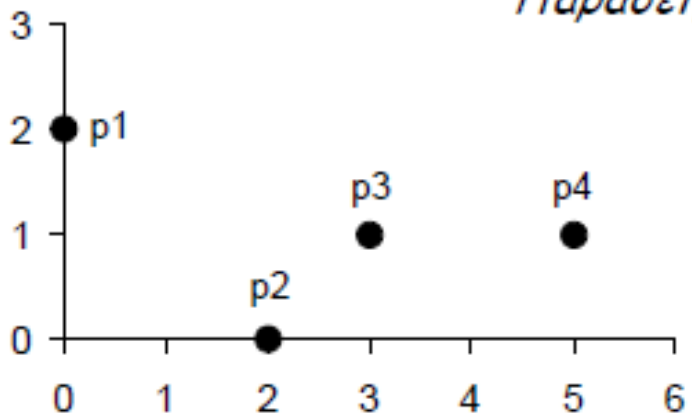
$$d(p,q) = L_p(p,q) = \sqrt[p]{(x_1 - y_1)^p + (x_2 - y_2)^p + \cdots + (x_n - y_n)^p}$$

Για $p = 2$ προκύπτει η Ευκλείδεια απόσταση.
Για $p = 1$ προκύπτει η απόσταση Manhattan.
Για $p \to \infty$ προκύπτει ότι
$d(p,q) = \max\{|x_1 - y_1|, |x_2 - y_2|, \ldots, |x_n - y_n|\}$.

Μέτρα ομοιότητας και απόστασης

Απόσταση Minkowksi

Παράδειγμα



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

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

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

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

Πίνακες Απόστασης

Μέτρα ομοιότητας και απόστασης

Ευκλείδεια απόσταση με βάρη

Για να δώσουμε έμφαση σε κάποια γνωρίσματα ή να τα αξιολογήσουμε χρησιμοποιούμε βάρη.
Αν p και q είναι δύο αντικείμενα με n γνωρίσματα που έχουν αριθμητικές τιμές: $p = (x_1, x_2, \ldots, x_n)$ και $q = (y_1, y_2, \ldots, y_n)$ τότε η Ευκλείδεια απόσταση με βάρη των p, q ορίζεται

$$d(p, q) = \sqrt{w_1(x_1 - y_1)^2 + w_2(x_2 - y_2)^2 + \cdots + w_n(x_n - y_n)^2}$$

Συνήθως, $w_1 + w_2 + \cdots + w_n = 1$ και $w_1, w_2, \ldots, w_n \geq 0$.

Ομοιότητα (similarity)

Μια αριθμητική μέτρηση $s(p, q)$ για το πόσο όμοια είναι δύο αντικείμενα p, q
Μεγαλύτερη όσο πιο όμοια είναι τα αντικείμενα μεταξύ τους
Συχνά τιμές στο $[0, 1]$.

Ιδιότητες

- $s(p, q) = 1 \Leftrightarrow p = q$.
- $s(p, q) = s(q, p)$ (συμμετρία)

Η ομοιότητα μεταξύ δύο αντικειμένων μετριέται συνήθως με τη βοήθεια μιας συνάρτησης απόστασης ανάμεσα στα αντικείμενα.
Εξαρτάται από το είδος των δεδομένων, δηλαδή από το είδος των χαρακτηριστικών/γνωρισμάτων τους.

Μέτρα ομοιότητας και απόστασης

Δυαδικά δεδομένα

Συχνά έχουμε δεδομένα με μόνο δυαδικά γνωρίσματα (π.χ. Φύλο (Άνδρας ή Γυναίκα)) τα οποία συνήθως κωδικοποιούνται με 0 ή 1.
Αν τα γνωρίσματα είναι συμμετρικά (π.χ. Φύλο) τότε είναι αυθαίρετη η επιλογή του 0 ή του 1.
Αν τα γνωρίσματα δεν είναι συμμετρικά (π.χ. Έχει νοσήσει από μια ασθένεια ή όχι, Έχει κάποιο χαρακτηριστικό ή όχι) τότε συνήθως το 1 κωδικοποιεί την ύπαρξη του χαρακτηριστικού ή την τιμή που θεωρείται πιο σημαντική.

Μέτρα ομοιότητας και απόστασης

Δυαδικά δεδομένα

Μεταξύ δύο αντικειμένων p, q με δυαδικά γνωρίσματα ορίζουμε τις συναρτήσεις:
M_{01} = ο αριθμός των γνωρισμάτων όπου το p έχει τιμή 0 και το q έχει 1.
M_{10} = ο αριθμός των γνωρισμάτων όπου το p έχει τιμή 1 και το q έχει 0.
M_{00} = ο αριθμός των γνωρισμάτων όπου το p έχει τιμή 0 και το q έχει 0.
M_{11} = ο αριθμός των γνωρισμάτων όπου το p έχει τιμή 1 και το q έχει 1.
Προφανώς, αν τα αντικείμενα έχουν n γνωρίσματα τότε
$M_{01} + M_{10} + M_{00} + M_{11} = n$.

Απλό ταιριασμα (για συμμετρικά γνωρίσματα)

$$SMC = \frac{\text{αριθμός ταιριασμάτων}}{\text{αριθμός γνωρισμάτων}} = \frac{M_{00} + M_{11}}{M_{01} + M_{10} + M_{00} + M_{11}}$$

Συντελεστής Jaccard (για μή συμμετρικά γνωρίσματα)

$$J = \frac{\text{αριθμός ταιριασμάτων } 11}{\text{αριθμός μη μηδενικών ταιριασμάτων}} = \frac{M_{11}}{M_{01} + M_{10} + M_{11}}$$

Μέτρα ομοιότητας και απόστασης

Μέτρα ομοιότητας για δυαδικά δεδομένα

Παράδειγμα

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

$M_{01} = 2$
$M_{10} = 1$
$M_{00} = 7$
$M_{11} = 0$

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

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

Μέτρα ομοιότητας και απόστασης

Αποστάσεις για δυαδικά δεδομένα

Απόσταση όταν έχουμε συμμετρικά γνωρίσματα

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

Απόσταση όταν έχουμε μη συμμετρικά γνωρίσματα

$$d(p,q) = \frac{M_{10} + M_{01}}{M_{11} + M_{10} + M_{01}}$$

Απόσταση Jaccard

$$d(p,q) = \frac{M_{11}}{M_{11} + M_{10} + M_{01}}$$

Μέτρα ομοιότητας και απόστασης

Αποστάσεις για δυαδικά δεδομένα

Παράδειγμα

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

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

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

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

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

Μέτρα ομοιότητας και απόστασης

Κατηγορικά δεδομένα χωρίς διάταξη (nominal)

Αποτελούν γενίκευση των δυαδικών γνωρισμάτων όπου μπορούν να πάρουν πάνω από δύο τιμές, π.χ. κόκκινο, πράσινο, κίτρινο

1η μέθοδος: Απλό ταίριασμα.

Αν m είναι ο αριθμός των ταιριασμάτων και n ο συνολικός αριθμός των μεταβλητών.

$$d(p, q) = \frac{n - m}{n}$$

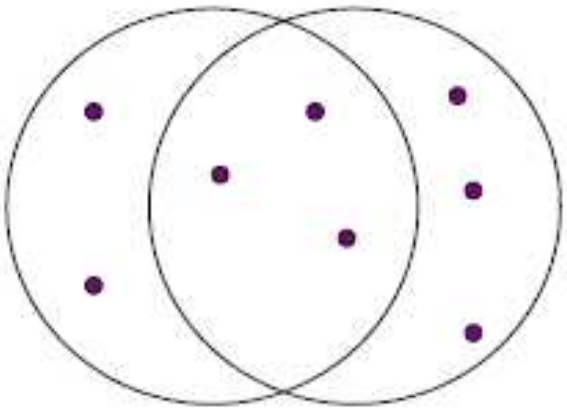
2η μέθοδος: Χρήση πολλών δυαδικών μεταβλητών

Μια μεταβλητή για κάθε μια από τις πιθανές τιμές ενός γνωρίσματος

Π.χ. Είναι κόκκινο; Είναι πράσινο; Είναι κίτρινο; κ.ο.κ.

Jaccard ομοιότητα για δύο σύνολα A, B

$$s(A, B) = \frac{|A \cap B|}{A \cup B}$$



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

Παρατήρηση: Ένα σύνολο μπορεί να αναπαρασταθεί ως μια δυαδική λέξη.

Μέτρα ομοιότητας και απόστασης

Αποστάσεις για σύνολα

Με βάση την Jaccard ομοιότητα μπορούμε να ορίσουμε την απόσταση δύο συνόλων A, B:

$$d(A, B) = 1 - s(A, B)$$

Εύκολα προκύπτει ότι η συνάρτηση $d(A, B)$ είναι μετρική.

Μέτρα ομοιότητας και απόστασης

Ομοιότητα συνημιτόνου

Εφαρμόζεται για δεδομένα με αριθμητικά γνωρίσματα. Δεν εξαρτάται από τον αριθμό των γνωρισμάτων. Αγνοεί τα 0 (όπως και η Jaccard). Αν p και q είναι δύο αντικείμενα με n γνωρίσματα που έχουν αριθμητικές τιμές: $p = (x_1, x_2, \ldots, x_n)$ και $q = (y_1, y_2, \ldots, y_n)$ τότε

$$\cos(p, q) = \frac{< p, q >}{|p||q|}$$

όπου $< p, q >$ είναι το εσωτερικό γινόμενο των p, q δηλαδή

$$< p, q >= x_1 y_1 + x_2 y_2 + \cdots + x_n y_n$$

και $|p|$, $|q|$ είναι τα μήκη των p, q που ορίζονται ως

$$|p| = \sqrt{< p, p >} \text{ και } |q| = \sqrt{< q, q >}$$

Μέτρα ομοιότητας και απόστασης

Ομοιότητα συνημιτόνου

Παράδειγμα

$d_1 = 3\ 2\ 0\ 5\ 0\ 0\ 0\ 2\ 0\ 0$

$d_2 = 1\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 2$

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

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

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

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

Παράδειγμα

Έστω ότι έχουμε μια συλλογή από n έγγραφα κειμένου D_1, D_2, ..., D_n στην οποία πρόκειται να κάνουμε αρκετές αναζητήσεις με λέξεις κλειδιά (keywords) ή όρους (terms). Ένας κλασσικός τρόπος για να αυξήσουμε την ταχύτητα των αναζητήσεων είναι να κάνουμε προεπεξεργασία των εγγράφων και να φτιάξουμε μια μήτρα όρων – εγγράφων (term-by-document) D η οποία έχει διαστάσεις $m \times n$, όπου m είναι ο αριθμός των διαφορετικών λέξεων κλειδιών ή όρων, στην οποία το στοιχείο της i-γραμμής και j-στήλης είναι ο αριθμός των εμφανίσεων του όρου t_i στο έγγραφο D_j.

Παράδειγμα (συνέχεια)

Έτσι, για παράδειγμα σε μια συλλογή με 9 έγγραφα D_1, D_2, ..., D_9 για την οποία μας ενδιαφέρει η αναζήτηση με λέξεις κλειδιά που μπορούν να επιλεγούν μεταξύ 5 όρων t_1, t_2, t_3, t_4, t_5 δημιουργήθηκε η παρακάτω μήτρα όρων - εγγράφων:

$D =$

	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8	D_9
t_1	1	1	2	0	1	0	1	0	1
t_2	0	1	0	1	0	1	1	0	0
t_3	0	2	0	2	0	1	0	1	1
t_4	1	0	1	0	1	0	2	1	0
t_5	1	2	1	0	0	1	0	0	1

(Σύμφωνα με την μήτρα D, το έγγραφο D_3 περιέχει δύο εμφανίσεις του όρου t_1, καμία εμφάνιση των όρων t_2, t_3 και από μια εμφάνιση των όρων t_4, t_5.)

Παράδειγμα (συνέχεια)

Εξετάζοντας μια-μία όλες τις στήλες της μήτρας D παρατηρούμε ότι κανένα έγγραφο δεν περιέχει ακριβώς αυτούς τους όρους.
Σύμφωνα με την αναπαράσταση αυτή καθένα από τα έγγραφα D_1, D_2, ..., D_9 αντιστοιχεί σε ένα διάνυσμα που ανήκει σε ένα χώρο 5 διαστάσεων με άξονες τα t_1, t_2, ..., t_5, ή ισοδύναμα ένα διάνυσμα του $\mathbb{R}^5$. Προκειμένου να επιλέξουμε ποιό από τα D_1, D_2, ..., D_9 ταιριάζει καλύτερα στην αναζήτησή μας, θεωρουμε επίσης ότι η αναζήτησή μας αντιστοιχεί στο διάνυσμα

$$\mathbf{q} = (0, 1, 0, 1, 1)$$

(Στην θέση i του $\mathbf{q}$ εμφανίζεται 1 αν η αναζήτησή μας περιέχει τον όρο t_i, αλλιώς εμφανίζεται 0.)

Παράδειγμα (συνέχεια)

Υπολογίζουμε την ομοιότητα ανάμεσα στο q και τα έγγραφα D_1, D_2, ..., D_9 με βάση το συνημίτονο της γωνίας που σχηματίζει διάνυσμα q με καθένα από τα διανύσματα-στήλες που αντιστοιχούν σε κάθε έγγραφο D_i, $i = 1, 2, \ldots, 9$.

Παράδειγμα (συνέχεια)

Έτσι, αν θ_1, θ_2, ..., θ_9 είναι οι γωνίες μεταξύ του q και των D_1, D_2, ..., D_9 αντίστοιχα, τότε

$$\cos\theta_1 = \frac{\langle \mathbf{q}, D_1\rangle}{\|\mathbf{q}\|\,\|D_1\|} = \frac{\langle (0,1,0,1,1),(1,0,0,1,1)\rangle}{\|0,1,0,1,1\|\,\|1,0,0,1,1\|} = \frac{2}{3} \simeq 0.666666$$

$$\cos\theta_2 = \frac{\langle \mathbf{q}, D_2\rangle}{\|\mathbf{q}\|\,\|D_2\|} = \frac{\langle (0,1,0,1,1),(1,1,2,0,2)\rangle}{\|0,1,0,1,1\|\,\|1,1,2,0,2\|} = \frac{3}{\sqrt{30}} \simeq 0.547723$$

$$\cos\theta_3 = \frac{\langle \mathbf{q}, D_3\rangle}{\|\mathbf{q}\|\,\|D_3\|} = \frac{\langle (0,1,0,1,1),(2,0,0,1,1)\rangle}{\|0,1,0,1,1\|\,\|2,0,0,1,1\|} = \frac{2}{\sqrt{18}} \simeq 0.471405$$

$$\cos\theta_4 = \frac{\langle \mathbf{q}, D_4\rangle}{\|\mathbf{q}\|\,\|D_4\|} = \frac{\langle (0,1,0,1,1),(0,1,2,0,0)\rangle}{\|0,1,0,1,1\|\,\|0,1,2,0,0\|} = \frac{1}{\sqrt{15}} \simeq 0.258199$$

Παράδειγμα (συνέχεια)

$$\cos\theta_5 = \frac{\langle \mathbf{q}, D_5\rangle}{\|\mathbf{q}\|\,\|D_5\|} = \frac{\langle (0,1,0,1,1),(1,0,0,1,0)\rangle}{\|0,1,0,1,1\|\,\|1,0,0,1,0\|} = \frac{1}{\sqrt{6}} \simeq 0.408248$$

$$\cos\theta_6 = \frac{\langle \mathbf{q}, D_6\rangle}{\|\mathbf{q}\|\,\|D_6\|} = \frac{\langle (0,1,0,1,1),(0,1,1,0,1)\rangle}{\|0,1,0,1,1\|\,\|0,1,1,0,1\|} = \frac{2}{3} \simeq 0.666666$$

$$\cos\theta_7 = \frac{\langle \mathbf{q}, D_7\rangle}{\|\mathbf{q}\|\,\|D_7\|} = \frac{\langle (0,1,0,1,1),(1,1,0,2,0)\rangle}{\|0,1,0,1,1\|\,\|1,1,0,2,0\|} = \frac{1}{\sqrt{2}} \simeq 0.707107$$

$$\cos\theta_8 = \frac{\langle \mathbf{q}, D_8\rangle}{\|\mathbf{q}\|\,\|D_8\|} = \frac{\langle (0,1,0,1,1),(0,0,1,1,0)\rangle}{\|0,1,0,1,1\|\,\|0,0,1,1,0\|} = \frac{1}{\sqrt{6}} \simeq 0.408248$$

$$\cos\theta_9 = \frac{\langle \mathbf{q}, D_9\rangle}{\|\mathbf{q}\|\,\|D_9\|} = \frac{\langle (0,1,0,1,1),(1,0,1,0,1)\rangle}{\|0,1,0,1,1\|\,\|1,0,1,0,1\|} = \frac{1}{3} \simeq 0.333333$$

Παράδειγμα (συνέχεια)

Με βάση τους υπολογισμούς μας, η μικρότερη γωνία εμφανίζεται μεταξύ του $\mathbf{q}$ και του D_7 (διότι όσο μεγαλύτερο το συνημίτονο της γωνίας θ, τόσο μικρότερη η γωνία θ). Επομένως, το έγγραφο D_7 ταιριάζει καλύτερα στην αναζήτηση των όρων t_2, t_4 και t_5 σε σχέση με τα υπόλοιπα έγγραφα της συλλογής μας.
Επιπλέον, με βάση τα συνημίτονα που υπολογίσαμε μπορούμε να διάταξουμε σε φθίνουσα σειρά (ως προς το ταίριασμα με την αναζήτησή μας) τα έγγραφα της συλλογής ως εξής:

$$D_7, D_1, D_6, D_2, D_3, D_5, D_8, D_9, D_4.$$

Παράδειγμα (συνέχεια)

Παρατήρηση Αν δύο έγγραφα περιέχουν ακριβώς τις ίδιες εμφανίσεις που αφορούν όρους της αναζήτησής μας $\mathbf{q}$, αλλά περιέχουν επιπλέον όρους που δεν περιέχονται στην αναζήτησή μας (π.χ. τα D_1, D_3, όπου το D_1 περιέχει μια εμφάνιση του όρου t_1 που δεν εμφανίζεται στην αναζήτηση $\mathbf{q}$, ενώ το D_3 περιέχει δύο εμφανίσεις του όρου t_1), τότε προτιμάται αυτό που περιέχει λιγότερους επιπλέον όρους (αφού με βάση τον τύπο του συνημιτόνου, τα εσωτερικά γινόμενα με το q στον αριθμητή είναι ίσα για τα δύο έγγραφα (π.χ. $\langle \mathbf{q}, D_1 \rangle = \langle \mathbf{q}, D_3 \rangle = 2$), ενώ το έγγραφο που περιέχει περισσότερους επιπλέον όρους οι οποίοι δεν περιλαμβάνονται στην αναζήτηση θα έχει μεγαλύτερη νόρμα (π.χ. $\|D_3\| = \sqrt{6} > \sqrt{3} = \|D_1\|$), επομένως μεγαλύτερο παρανομαστή, οπότε και μικρότερο συνημίτονο (πράγματι, $\cos\theta_1 = \frac{2}{\sqrt{9}} > \frac{2}{\sqrt{18}} = \cos\theta_3$.)).

Απόσταση Edit (Βιοπληροφορική)

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

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

Πηγές διαφανειών

- P.-N. Tan, M. Steinbach, A. Karpatne, V. Kumar, Introduction to Data Mining, 2nd edition
- Μαθήματα εξόρυξης δεδομένων, ΕΑΠ.