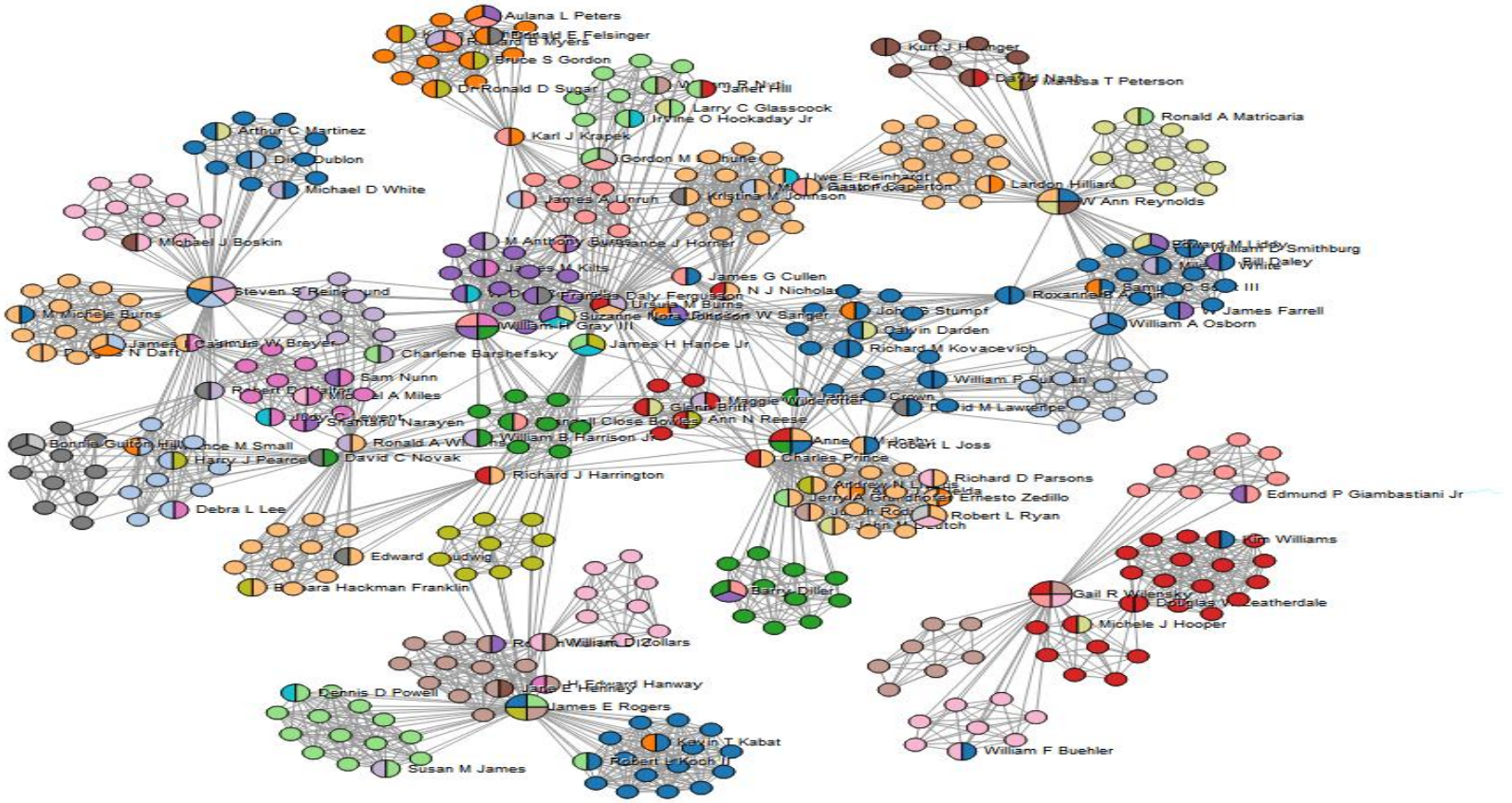
Εξόρυξη Δεδομένων

Ομαδοποίηση:

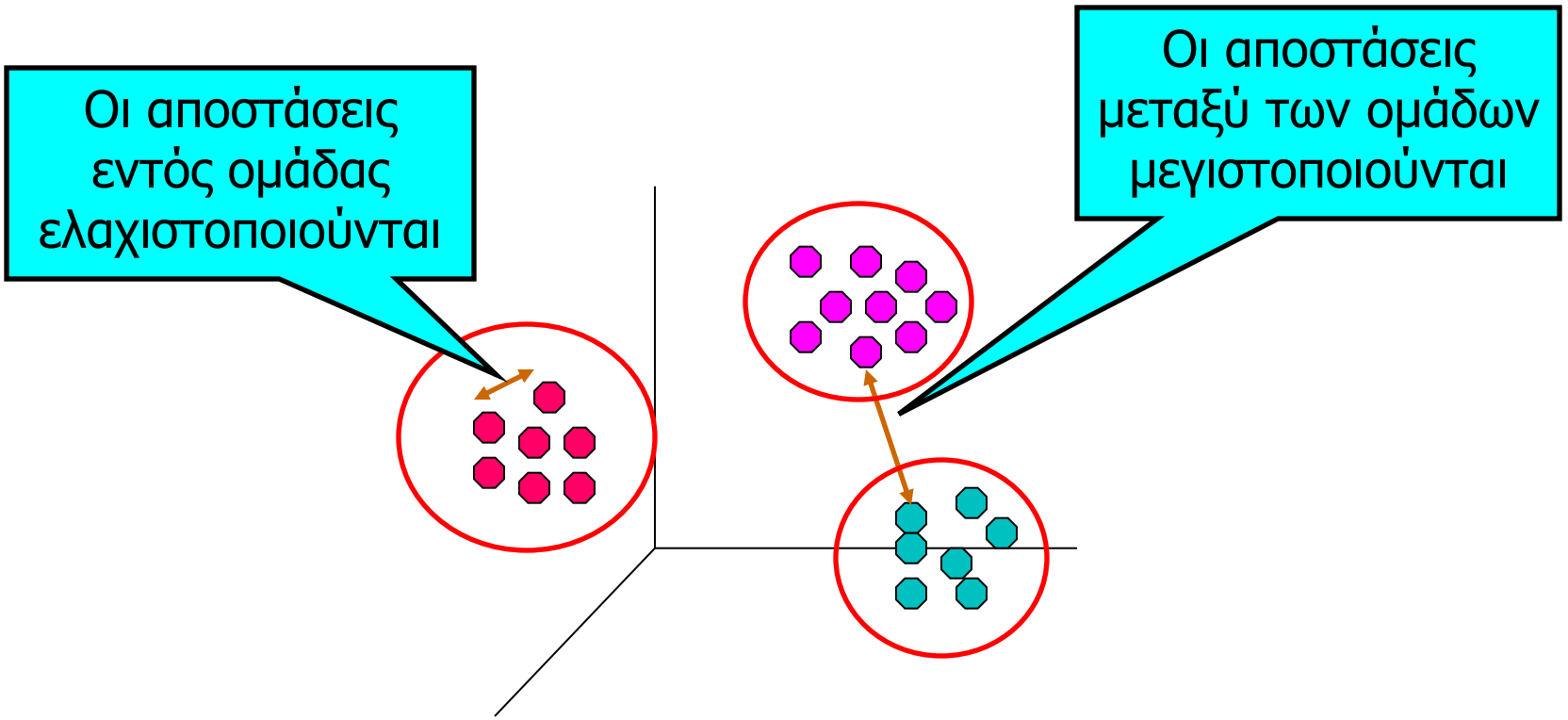
Βασικές Έννοιες και Αλγόριθμοι

(Σημειώσεις μεταφρασμένες από το Κεφάλαιο 8 του βιβλίου των Tan, Steinbach, Kumar)



Τι είναι η Ομαδοποίηση;

- Η εύρεση ομάδων/συστάδων αντικειμένων έτσι ώστε τα αντικείμενα μίας ομάδας να είναι παρόμοια (ή να σχετίζονται) το ένα με το άλλο και να διαφέρουν (ή να μην σχετίζονται) με αντικείμενα από άλλες ομάδες



Εφαρμογές της Ομαδοποίησης

- **Ανακάλυψη-Γνώση**
  - Ομαδοποίηση σχετικών κειμένων για περιήγηση, ομαδοποίηση γονιδίων και πρωτεϊνών που έχουν παρόμοια λειτουργία, ή ομαδοποίηση αποθεμάτων που έχουν παρόμοιες διακυμάνσεις τιμών
- **Σύνοψη-Συγχώνευση**
  - Μείωση του μεγέθους μεγάλων συνόλων 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 βροχοπτώσεων στην Αυστραλία

Τι δεν αποτελεί Ομαδοποίηση;

- Η Εποπτευόμενη Κατηγοριοποίηση
  - Η κατηγοριοποίηση έχει πληροφορίες για τις ετικέτες των κλάσεων
- Η Απλή Τμηματοποίηση
  - Ο διαχωρισμός των φοιτητών αλφαβητικά σε διαφορετικές ομάδες μητρώου με βάση το επώνυμό τους
- Τα αποτελέσματα ενός ερωτήματος
  - Όταν οι ομαδοποιήσεις είναι αποτέλεσμα εξωτερικών προδιαγραφών
- Ο Διαχωρισμός Γράφων
  - Μπορεί να υπάρχει μία αμοιβαία συνάφεια και συγγένεια στα προβλήματα αυτά, αλλά οι περιοχές τους δεν είναι πανομοιότυπες

Η έννοια μίας ομάδας μπορεί να είναι ασαφής

Πόσες Ομάδες έχουμε;

Έξι Ομάδες

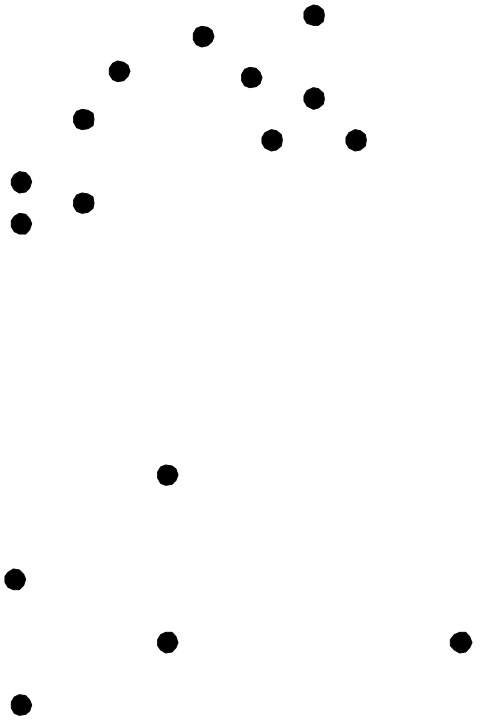
Δύο Ομάδες

Τέσσερις Ομάδες

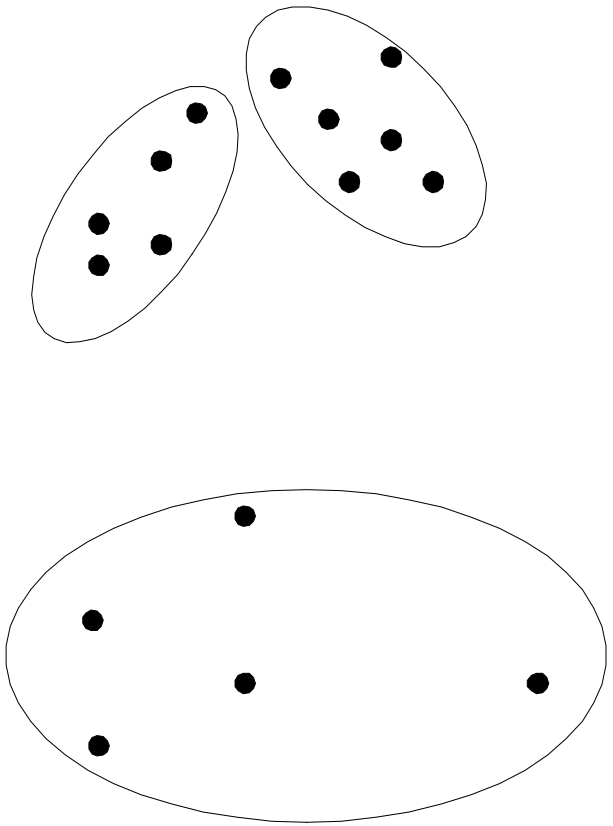
Τύποι Ομαδοποιήσεων

- Η ομαδοποίηση είναι η παραγωγή ενός συνόλου ομάδων
- Υπάρχει σημαντική διάκριση μεταξύ ιεραρχικής και διαχωριστικής παραγωγής συνόλων ομάδων
- Ομαδοποίηση Διαμέρισης (Partitional Clustering)
  - Μια διαμέριση των δεδομένων αντικειμένων σε ξένα μεταξύ τους υποσύνολα (ομάδες) έτσι ώστε κάθε δεδομένο αντικείμενο να βρίσκεται σε ακριβώς ένα υποσύνολο
- Ιεραρχική Ομαδοποίηση (Hierarchical clustering)
  - Ένα σύνολο εμφωλευμένων ομάδων που είναι οργανωμένες σε ένα δέντρο ιεραρχίας

Ομαδοποίηση Διαχωρισμού

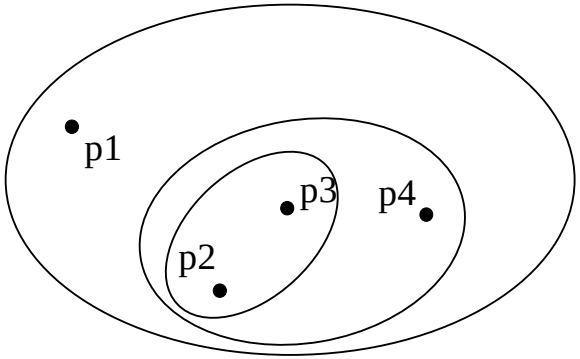


Αρχικά Σημεία

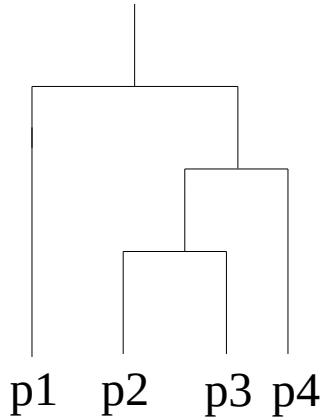


Μία Ομαδοποίηση Διαχωρισμού

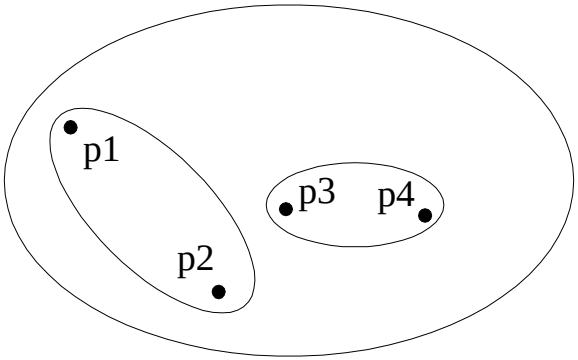
Ιεραρχική Ομαδοποίηση



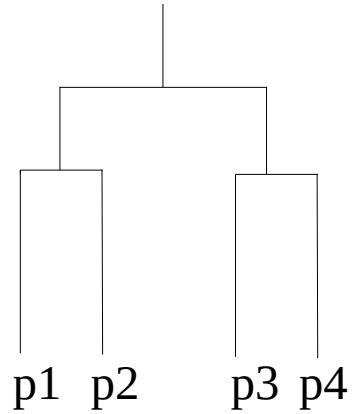
Κλασική Ιεραρχική Ομαδοποίηση



Κλασικό Δενδροδιάγραμμα



Μη-κλασική Ιεραρχική Ομαδοποίηση



Μη-κλασικό Δενδροδιάγραμμα

Άλλες διακρίσεις μεταξύ Ομαδοποιήσεων

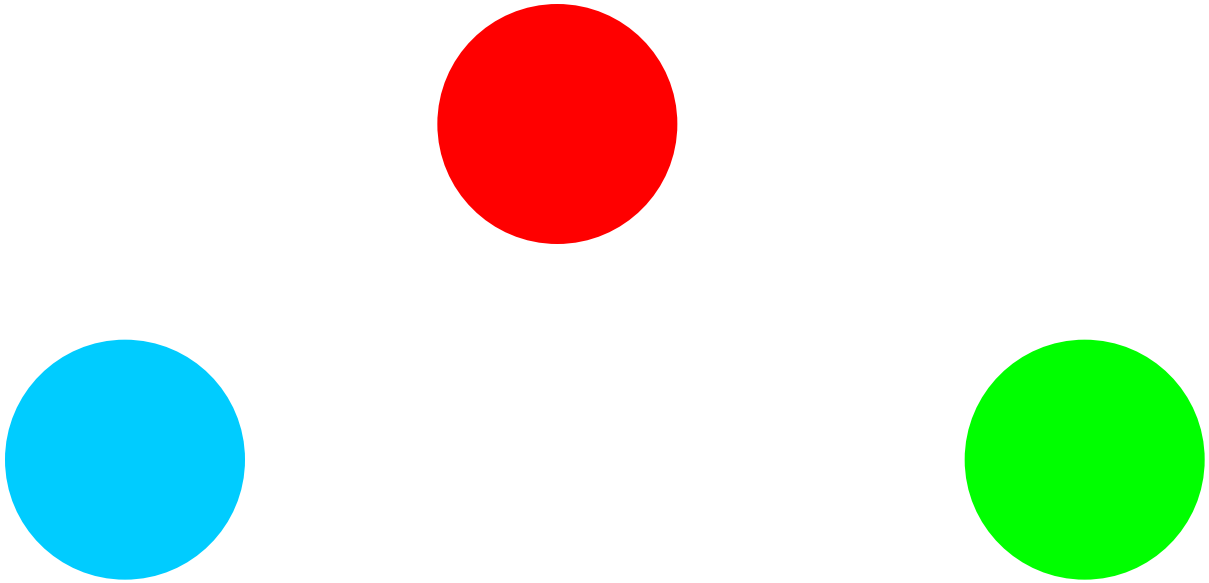
- Αποκλειστικές και Μη-αποκλειστικές
  - Σε μη-αποκλειστικές ομαδοποιήσεις, τα σημεία μπορεί να ανήκουν σε πολλές ομάδες.
  - Μπορεί να παριστάνουν πολλαπλές κλάσεις ή να είναι συνοριακά σημεία
- Ασαφής και Μη-ασαφής
  - Σε μία ασαφή ομαδοποίηση, ένα σημείο ανήκει σε κάθε ομάδα με κάποιο συντελεστή βαρύτητας (βάρος) μεταξύ 0 και 1
  - Τα βάρη πρέπει να έχουν άθροισμα 1
  - Η πιθανοτική ομαδοποίηση έχει παρόμοια χαρακτηριστικά
- Μερική και Πλήρης
  - Σε μερικές περιπτώσεις θέλουμε να ομαδοποιήσουμε μόνο μερικά από τα δεδομένα
- Ετερογενής και Ομογενής
  - Οι ομάδες μπορεί να έχουν πολύ διαφορετικά μεγέθη, σχήματα/μορφές και πυκνότητες

Τύποι Ομάδων

- Καλά διαχωρισμένες ομάδες
- Ομάδες που βασίζονται σε ένα κέντρο
- Συνεκτικές ομάδες
- Ομάδες που βασίζονται στην πυκνότητα
- Ομάδες που βασίζονται σε κάποια Ιδιότητα ή Έννοια
- Ομάδες που περιγράφονται από μία συνάρτηση στόχου (objective function)

Τύποι Ομάδων: Καλά Διαχωρισμένες

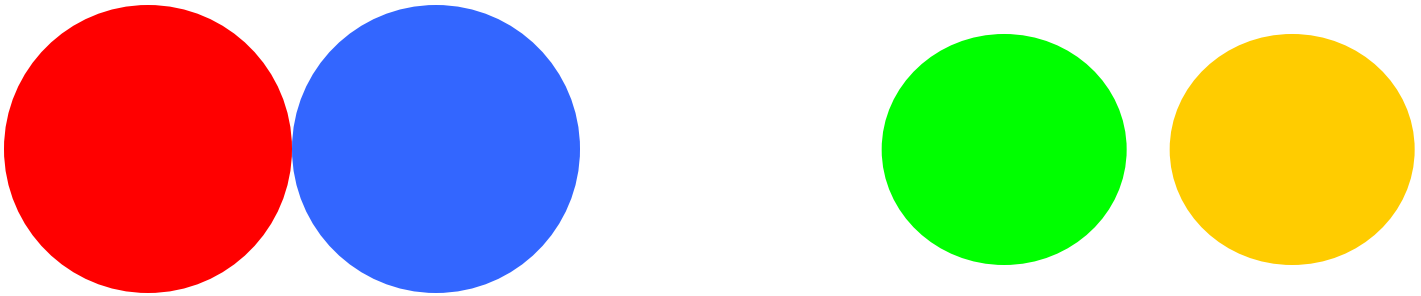
- Καλά Διαχωρισμένες Ομάδες:
  - Μία ομάδα είναι ένα σύνολο σημείων τέτοια ώστε κάθε σημείο της ομάδας να είναι όσο πιο κοντά γίνεται (ή όσο πιο όμοιο γίνεται) με κάθε άλλο σημείο στην ομάδα αυτή παρά με ένα οποιοδήποτε άλλο σημείο που δεν περιέχεται στην ομάδα.



3 καλά διαχωρισμένες ομάδες

Τύποι Ομάδων: που βασίζονται σε ένα κέντρο

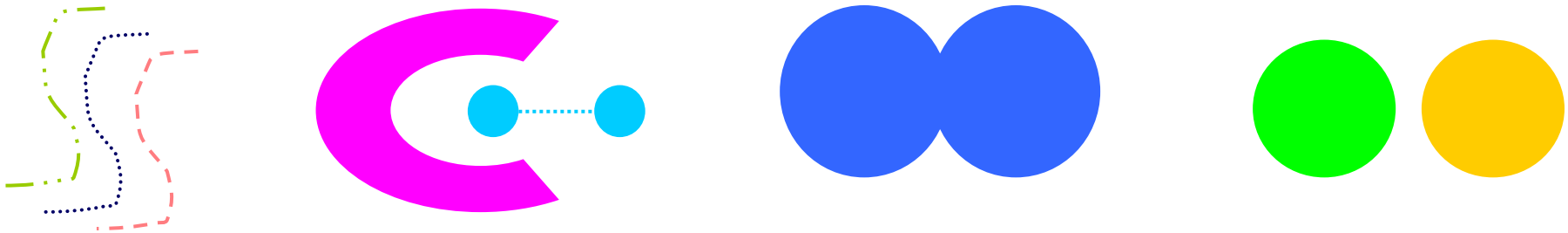
- Ομάδες που βασίζονται σε ένα κέντρο
  - Μία ομάδα είναι ένα σύνολο αντικειμένων τέτοια ώστε κάθε αντικείμενο της ομάδας να είναι όσο πιο κοντά γίνεται (ή όσο πιο όμοιο γίνεται) με το «κέντρο» της ομάδας παρά με το κέντρο οποιασδήποτε άλλης ομάδας
  - Το κέντρο της ομάδας ονομάζεται centroid, όταν προκύπτει από τον μέσο όρο όλων των σημείων της ομάδας, ή medoid, όταν εκλέγεται από το πιο «αντιπροσωπευτικό» σημείο της



4 ομάδες βασισμένες σε κέντρα

Τύποι Ομάδων: Συνεκτικές Ομάδες

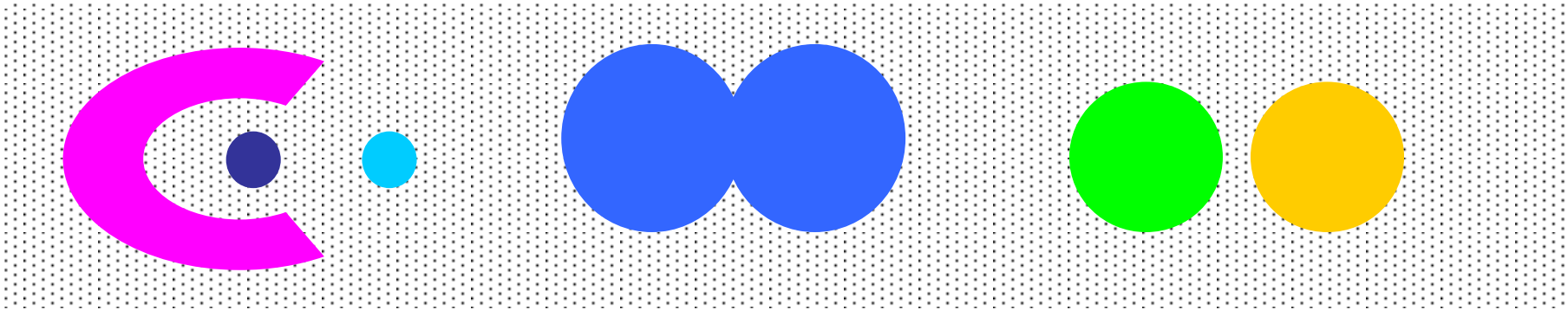
- Συνεκτικές Ομάδες (κοντινότερου γείτονα ή μεταβατικές)
  - Μία ομάδα είναι ένα σύνολο σημείων τέτοια ώστε κάθε σημείο της ομάδας να είναι όσο πιο κοντά γίνεται (ή όσο πιο όμοιο γίνεται) με ένα ή περισσότερα άλλα σημεία στην ομάδα αυτή παρά με ένα οποιοδήποτε άλλο σημείο που δεν περιέχεται στην ομάδα.



8 συνεκτικές ομάδες

Τύποι Ομάδων: που βασίζονται στην πυκνότητα

- Ομάδες που βασίζονται στην πυκνότητα
  - Μία ομάδα είναι μία πυκνή περιοχή σημείων η οποία διαχωρίζεται από περιοχές χαμηλής πυκνότητας και από άλλες περιοχές υψηλής πυκνότητας.
  - Χρησιμοποιείται όταν οι ομάδες είναι ακανόνιστες ή μπερδεμένες, καθώς και όταν υπάρχει θόρυβος στα δεδομένα ή ακραίες-απομονωμένες τιμές (outliers).

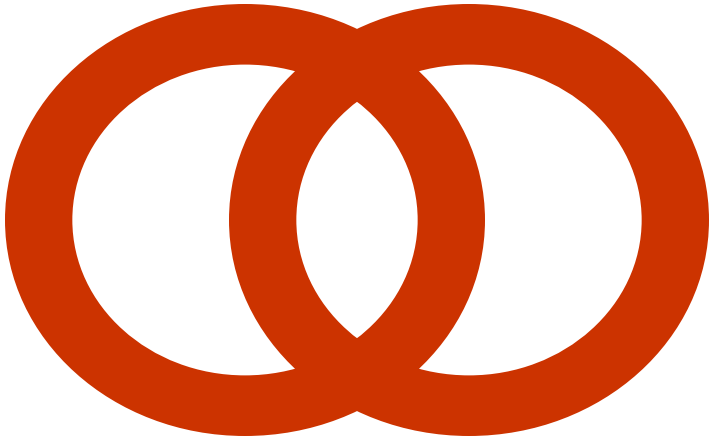


6 ομάδες που βασίζονται στην πυκνότητα

Τύποι Ομάδων: Εννοιολογικές Ομάδες

- Ομάδες που βασίζονται σε κάποια Ιδιότητα ή Έννοια
  - Οι ομάδες που βρίσκονται, μοιράζονται μία κοινή ιδιότητα ή παριστάνουν μία συγκεκριμένη έννοια.

.



2 Επικαλυπτόμενοι Κύκλοι

Τύποι Ομάδων: με Συνάρτηση Στόχου

- Ομάδες που ορίζονται από μία συνάρτηση στόχου
  - Βρίσκονται οι ομάδες που ελαχιστοποιούν ή μεγιστοποιούν μία αντικειμενική συνάρτηση.
  - Καταγράφονται όλοι οι δυνατοί τρόποι διαχωρισμού των σημείων σε ομάδες και αξιολογείται η ποιότητα κάθε υποψήφιου συνόλου ομάδων χρησιμοποιώντας της δεδομένη συνάρτηση στόχου. (NP Hard)
  - Μπορεί να ορίζονται καθολικοί ή τοπικοί στόχοι.
    - Οι ιεραρχικοί αλγόριθμοι ομαδοποίησης έχουν συνήθως τοπικούς στόχους
    - Οι αλγόριθμοι διαμέρισης συνήθως έχουν καθολικούς στόχους
  - Μία παραλλαγή της προσέγγισης με καθολική συνάρτηση στόχου είναι η προσαρμογή των δεδομένων σε ένα παραμετροποιημένο μοντέλο.
    - Οι παράμετροι για το μοντέλο προσδιορίζονται από τα δεδομένα.
    - Μεικτά μοντέλα υποθέτουν ότι τα δεδομένα είναι μία ανάμειξη από δεδομένα στατιστικών κατανομών.

Τύποι Ομάδων: με Συνάρτηση Στόχου

- Αποτυπώνεται το πρόβλημα της ομαδοποίησης σε ένα διαφορετικό πεδίο και επιλύνεται ένα σχετικό πρόβλημα στο πεδίο αυτό
  - Ο πίνακας εγγύτητας (proximity matrix) ορίζει έναν γράφο με βάρη, όπου οι κόμβοι του είναι τα σημεία που πρέπει να ομαδοποιηθούν και οι ακμές του με βάρη παριστάνουν την εγγύτητα μεταξύ των σημείων
  - Τότε η ομαδοποίηση είναι ισοδύναμη με τον διαχωρισμό του γράφου σε συνεκτικά τμήματα, όπου το καθένα ορίζει και μία ομάδα.
  - Πρέπει να ελαχιστοποιηθεί το βάρος ακμών ανάμεσα στις ομάδες και να μεγιστοποιηθεί μέσα στις ομάδες

Τα χαρακτηριστικά των αρχικών δεδομένων είναι σημαντικά

- Μέτρο εγγύτητας ή πυκνότητας
  - Πρόκειται για ένα παράγωγο μέτρο, αλλά είναι βασικό στην ομαδοποίηση
- Μέτρο αραιότητας
  - Καθορίζει τον τύπο της ομοιότητας
  - Συνεισφέρει στην αποδοτικότητα
- Τύπος χαρακτηριστικών
  - Καθορίζει τον τύπο της ομοιότητας
- Τύπος δεδομένων
  - Καθορίζει τον τύπο της ομοιότητας
  - Άλλα χαρακτηριστικά, π.χ. αυτοσυσχέτιση
- Πλήθος Διαστάσεων
- Θόρυβος και Outliers
- Τύπος της Κατανομής

Αλγόριθμοι Ομαδοποίησης

- K-means και οι παραλλαγές του
- Ιεραρχική Ομαδοποίηση (Hierarchical clustering)
- Ομαδοποίηση που βασίζεται στην πυκνότητα (Density-based clustering)

Ο Αλγόριθμος K-means

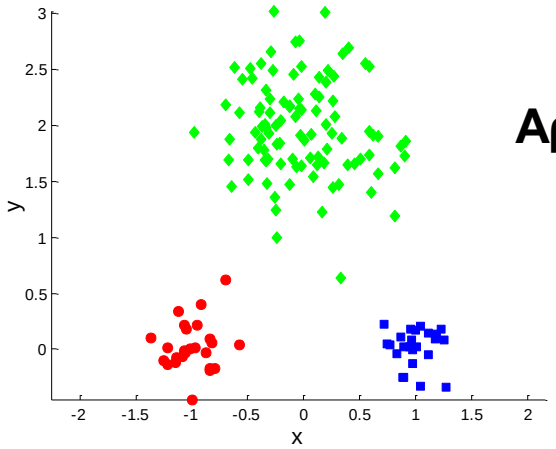
- Είναι μέθοδος ομαδοποίησης διαμέρισης
- Κάθε ομάδα συνδέεται με ένα centroid (κεντρικό σημείο)
- Κάθε σημείο αντιστοιχείται στην ομάδα που έχει το κοντινότερο centroid
- Το πλήθος των ομάδων (K) πρέπει να προκαθοριστεί
- Ο βασικός αλγόριθμος είναι πολύ απλός:

1: Select K points as the initial centroids.
2: **repeat**
3: Form K clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: **until** The centroids don't change

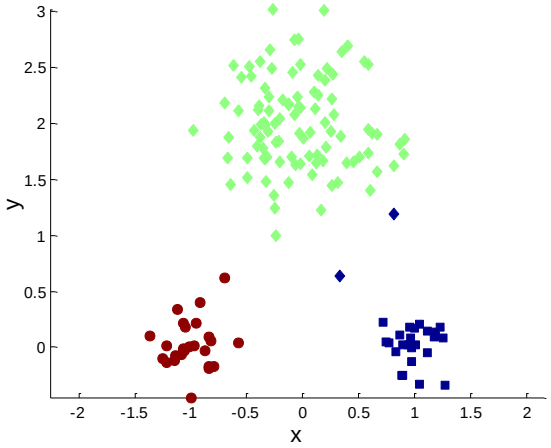
Αλγόριθμος K-means – Λεπτομέρειες

- Τα αρχικά centroids επιλέγονται συνήθως τυχαία.
  - Οι ομάδες που παράγονται συνήθως ποικίλουν μεταξύ των εκτελέσεων του αλγορίθμου για το ίδιο σύνολο δεδομένων.
- Το centroid είναι (τυπικά) ο μέσος όρος των σημείων στην ομάδα. Π.χ. για τα σημεία (1,1), (2,3) και (6,2) το centroid είναι: ((1+2+6)/3,(1+3+2)/3)=(3,2)
- Η «εγγύτητα» μετριέται από την Ευκλείδεια απόσταση, ή την cosine similarity, ή το correlation, κλπ.
- Ο K-means συγκλίνει όταν εφαρμόζονται κοινά μέτρα ομοιότητας όπως αυτά που αναφέρονται πιο πάνω.
- Τις περισσότερες φορές η σύγκλιση συμβαίνει στις πρώτες επαναλήψεις.
  - Συχνά η συνθήκη τερματισμού καθορίζεται από το «αν σχετικά λίγα σημεία αλλάζουν ομάδες»
- Η χρονική πολυπλοκότητά του είναι O(n * K * I * d)
  - n = πλήθος σημείων, K = πλήθος ομάδων,
    I = αριθμός επαναλήψεων, d = πλήθος χαρακτηριστικών/διαστάσεων

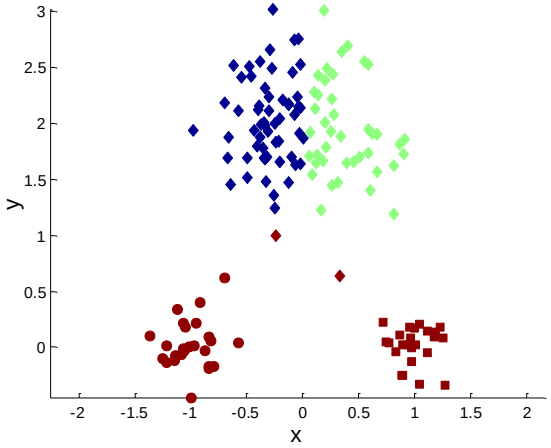
Δύο διαφορετικές ομαδοποιήσεις του K-means



Αρχικά Σημεία

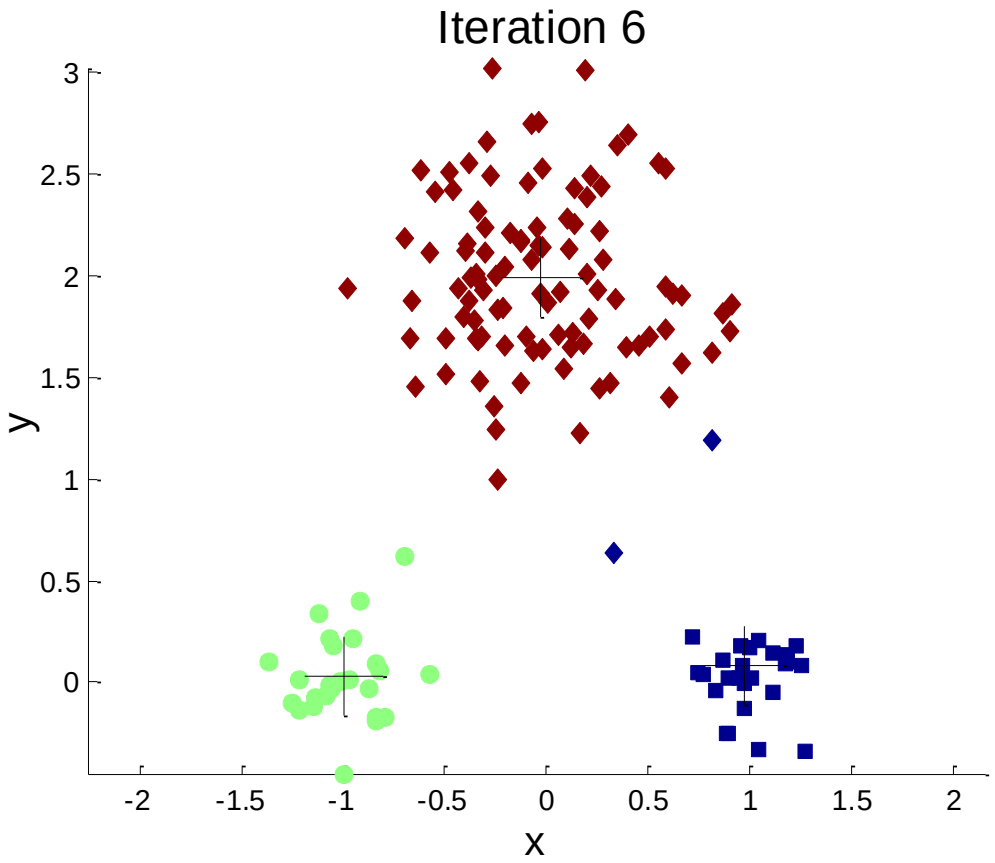


Βέλτιστη Ομαδοποίηση

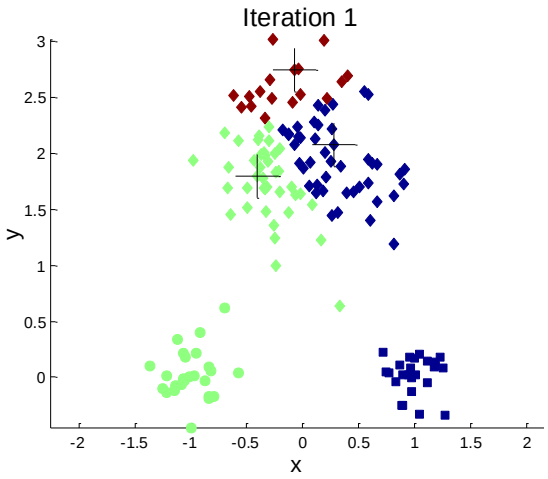


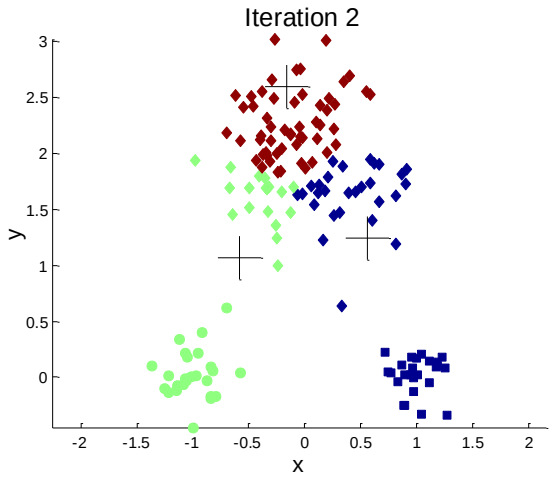
Μη-βέλτιστη Ομαδοποίηση

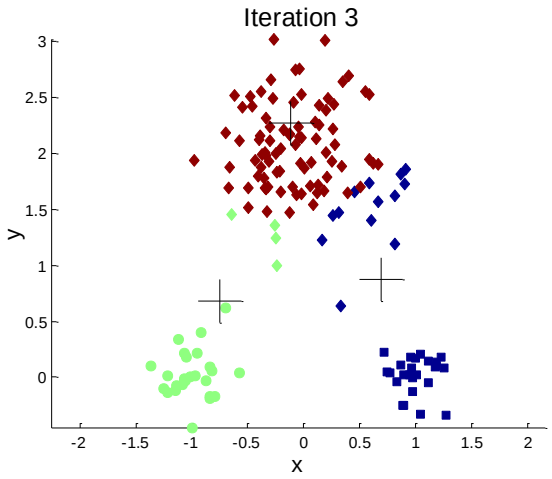
Η σημασία της επιλογής των αρχικών Centroids

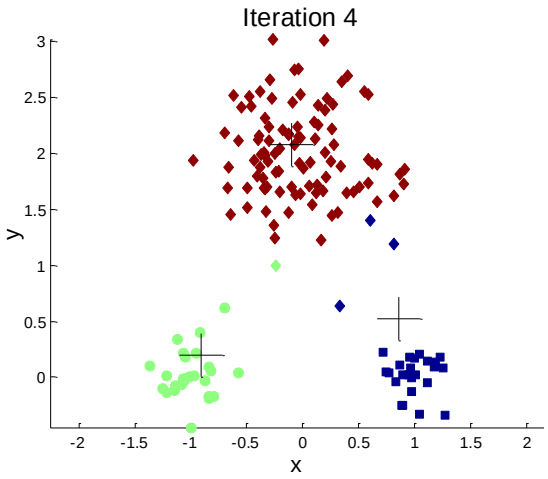


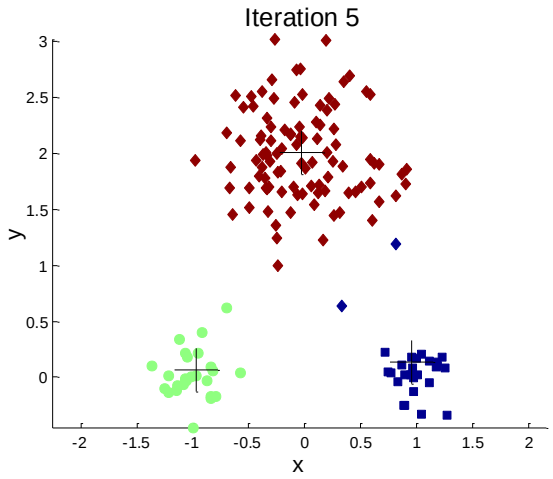
Η σημασία της επιλογής των αρχικών Centroids

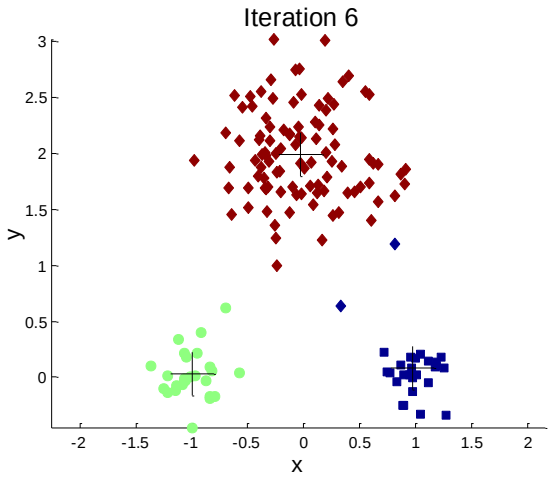












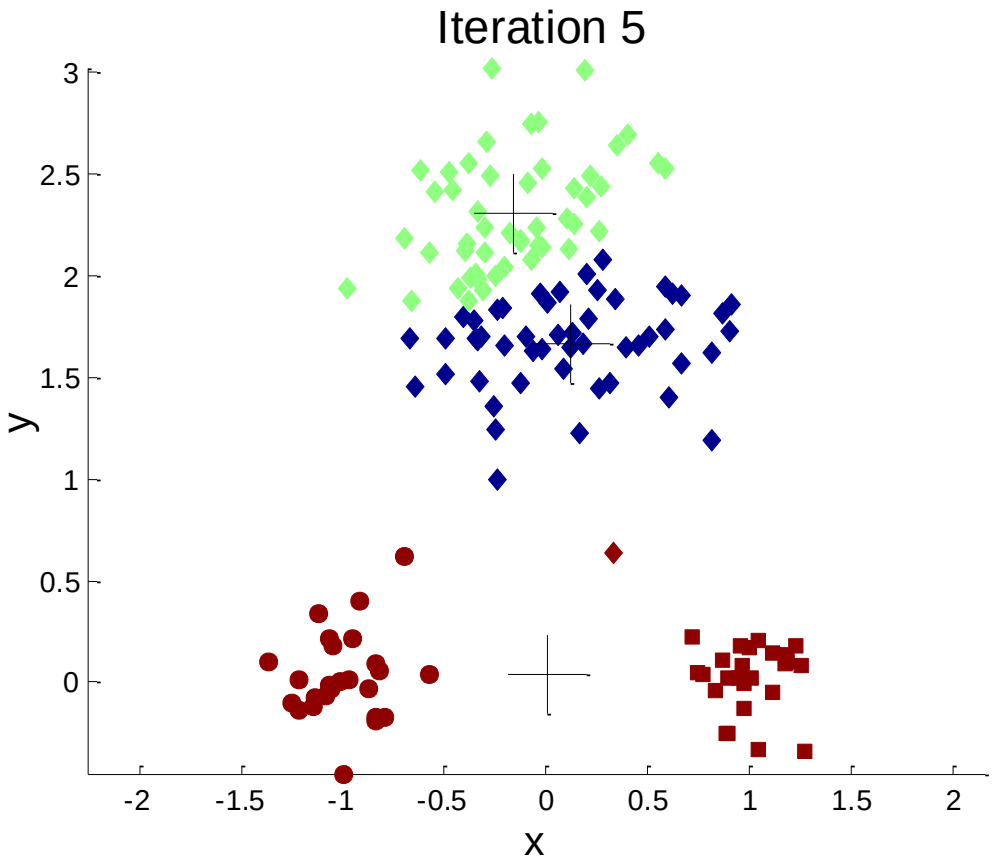
Αξιολόγηση των Ομάδων του K-means

- Το πιο κοινό μέτρο είναι το Sum of Squared Error (SSE)
  - Για κάθε σημείο, το σφάλμα καθορίζεται από την απόσταση προς την κοντινότερη ομάδα
  - Το SSE, υπολογίζεται από το άθροισμα των τετραγώνων αυτών των σφαλμάτων:

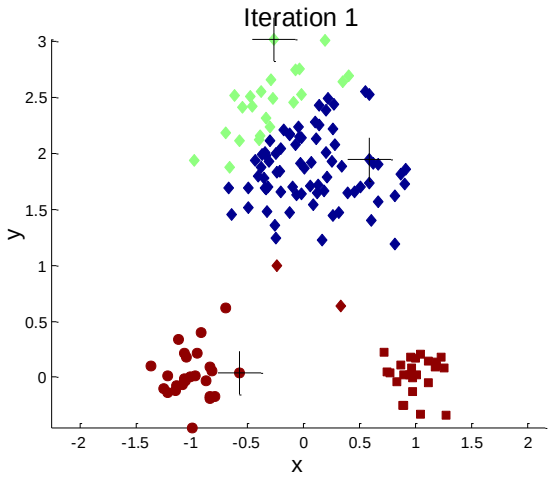
$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$

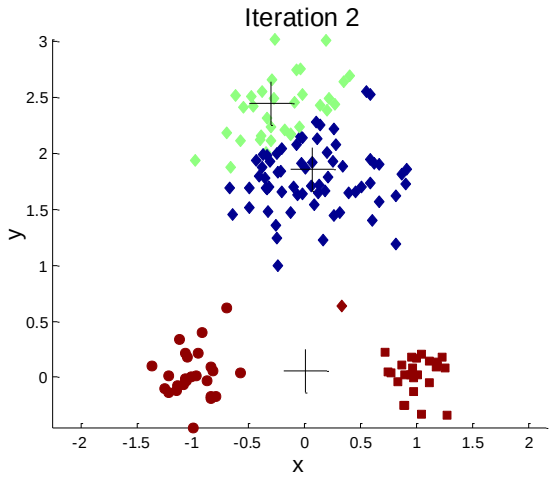
  - x είναι ένα σημείο στην ομάδα C_i και m_i είναι το αντιπροσωπευτικό σημείο της ομάδας C_i
    - το m_i αντιστοιχεί στο κέντρο (mean) της ομάδας
  - Μεταξύ δύο ομάδων, μπορούμε να επιλέξουμε εκείνη που έχει το μικρότερο σφάλμα
  - Ένας εύκολος τρόπος για να μειωθεί το SSE είναι να αυξηθεί το K, το πλήθος των ομάδων
    - Μία καλή ομαδοποίηση με μικρό K μπορεί να έχει πιο μικρό SSE από μία κακή ομαδοποίηση με πιο μεγάλο K

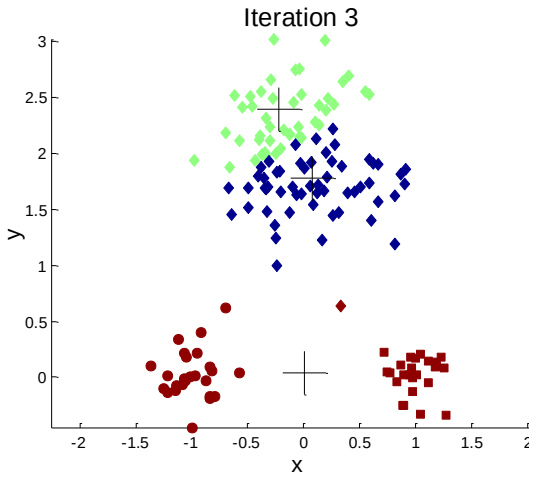
Η σημασία της επιλογής των αρχικών Centroids

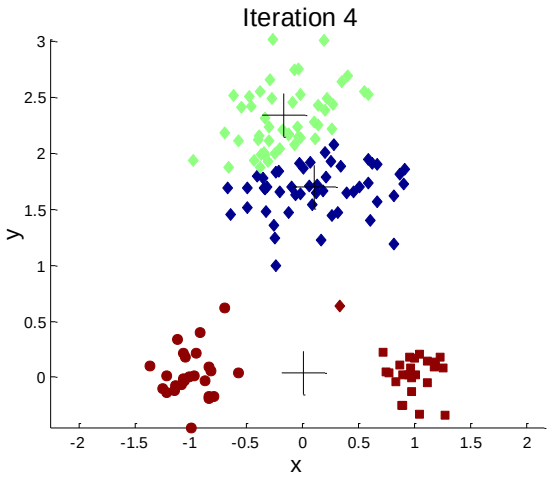


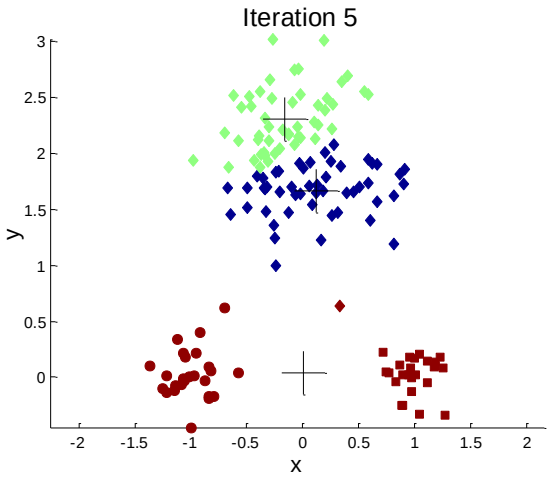
Η σημασία της επιλογής των αρχικών Centroids











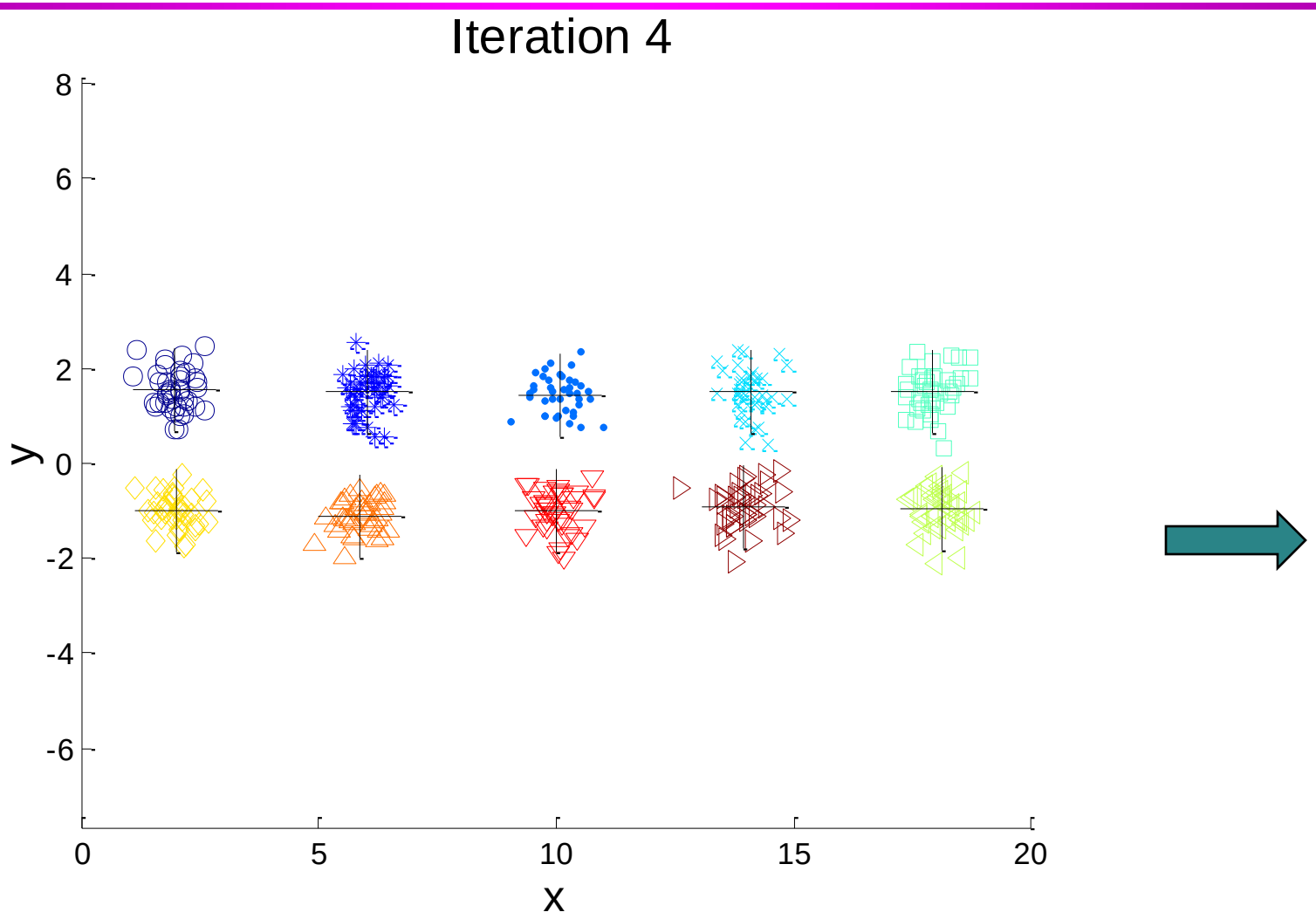
Προβλήματα με την επιλογή των αρχικών σημείων

- Αν υπάρχουν K «πραγματικές» ομάδες τότε η πιθανότητα επιλογής ενός centroid από κάθε ομάδα είναι μικρή.
  - Η πιθανότητα είναι σχετικά μικρή όταν το K είναι μεγάλο
  - Αν οι ομάδες είναι του ιδίου μεγέθους n τότε:

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

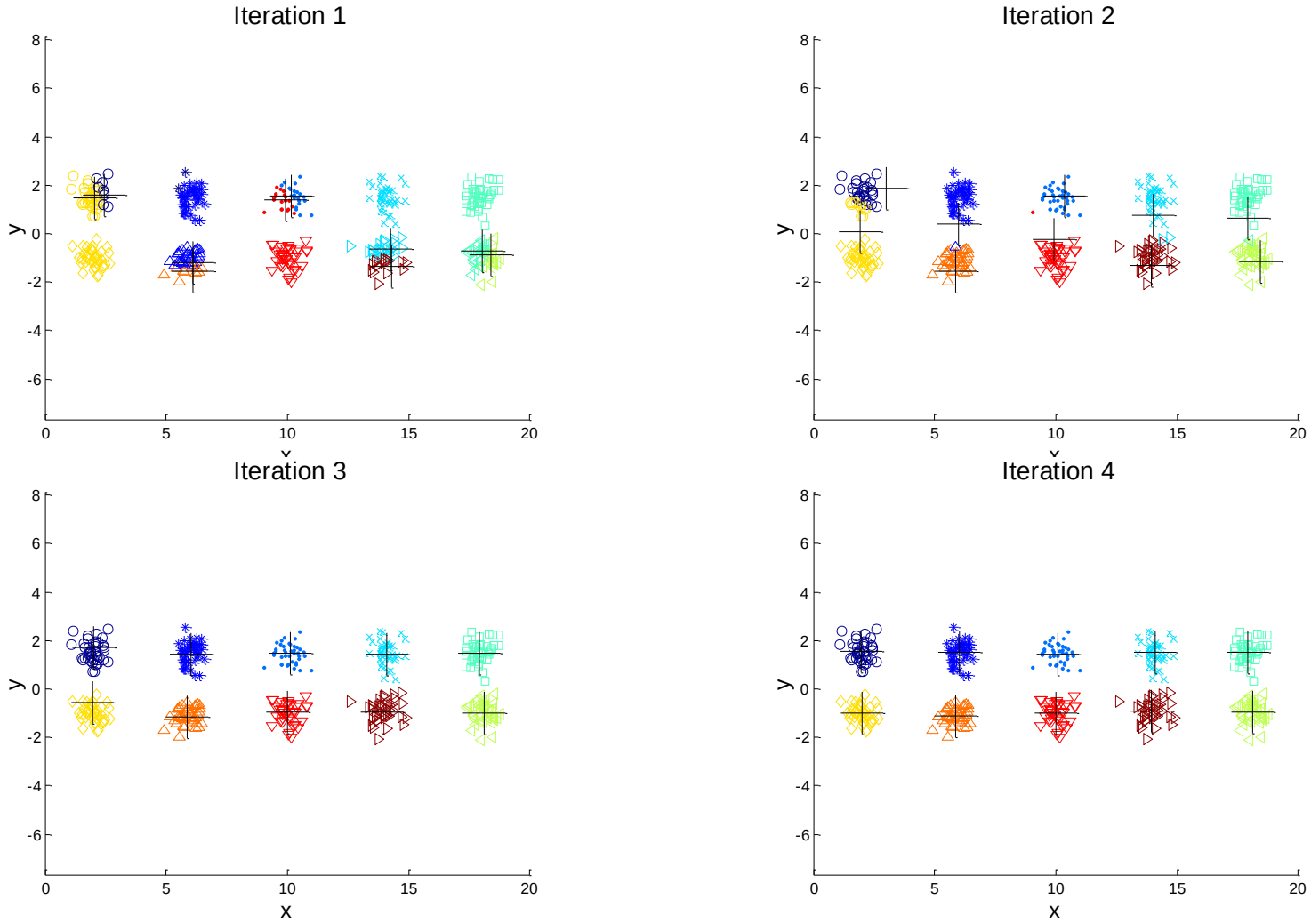
  - Για παράδειγμα αν K = 10, τότε η πιθανότητα είναι ίση με $10!/10^{10} = 0.00036$
  - Μερικές φορές τα αρχικά centroids αναπροσαρμόζονται με τον σωστό τρόπο, και άλλες φορές δεν το κάνουν αυτό
  - Ας θεωρήσουμε ένα παράδειγμα με 5 ζεύγη ομάδων

Παράδειγμα 10 Ομάδων



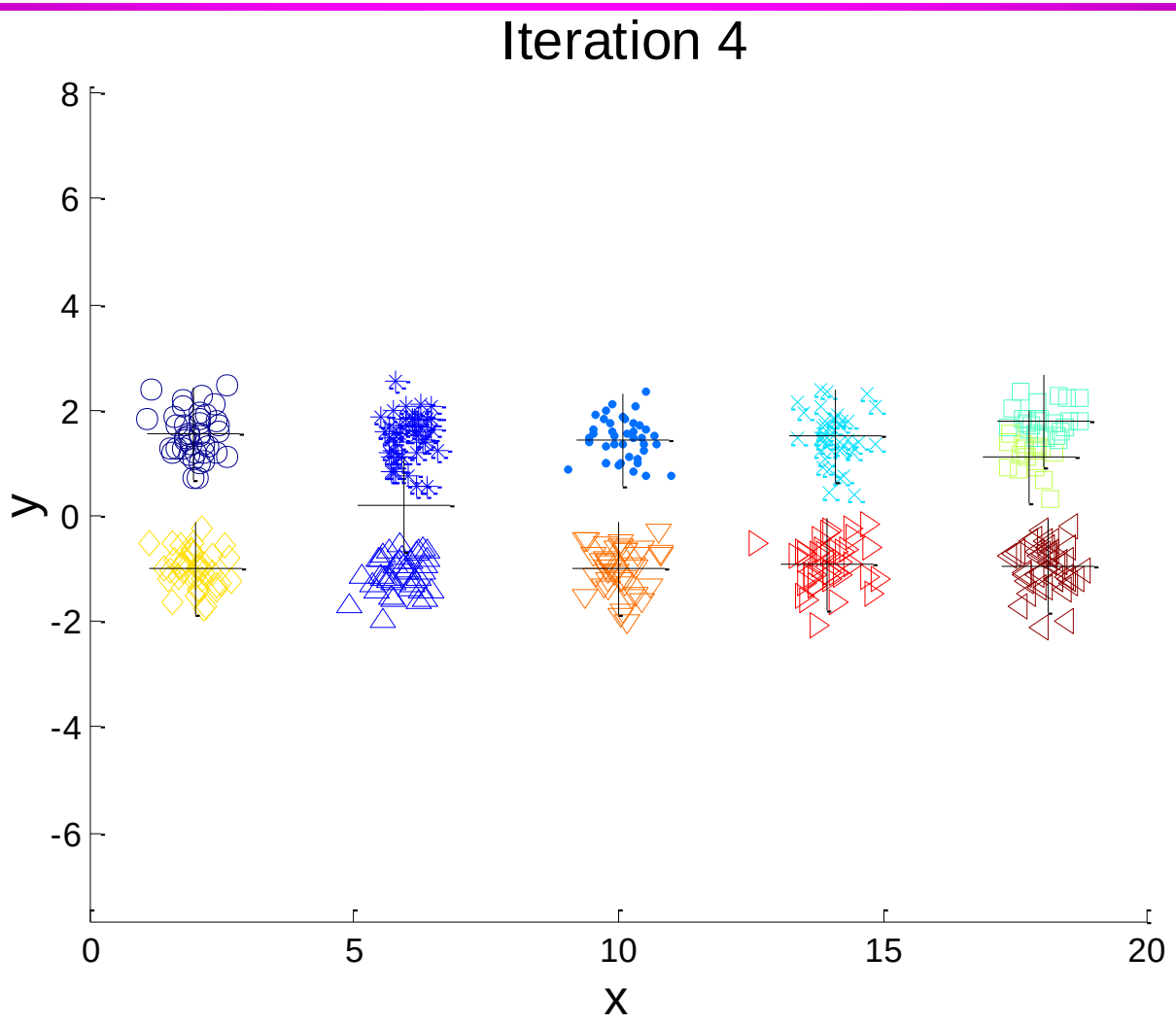
Ξεκινώντας με δύο αρχικά centroids σε μία ομάδα για κάθε ζεύγος ομάδων

Παράδειγμα 10 Ομάδων



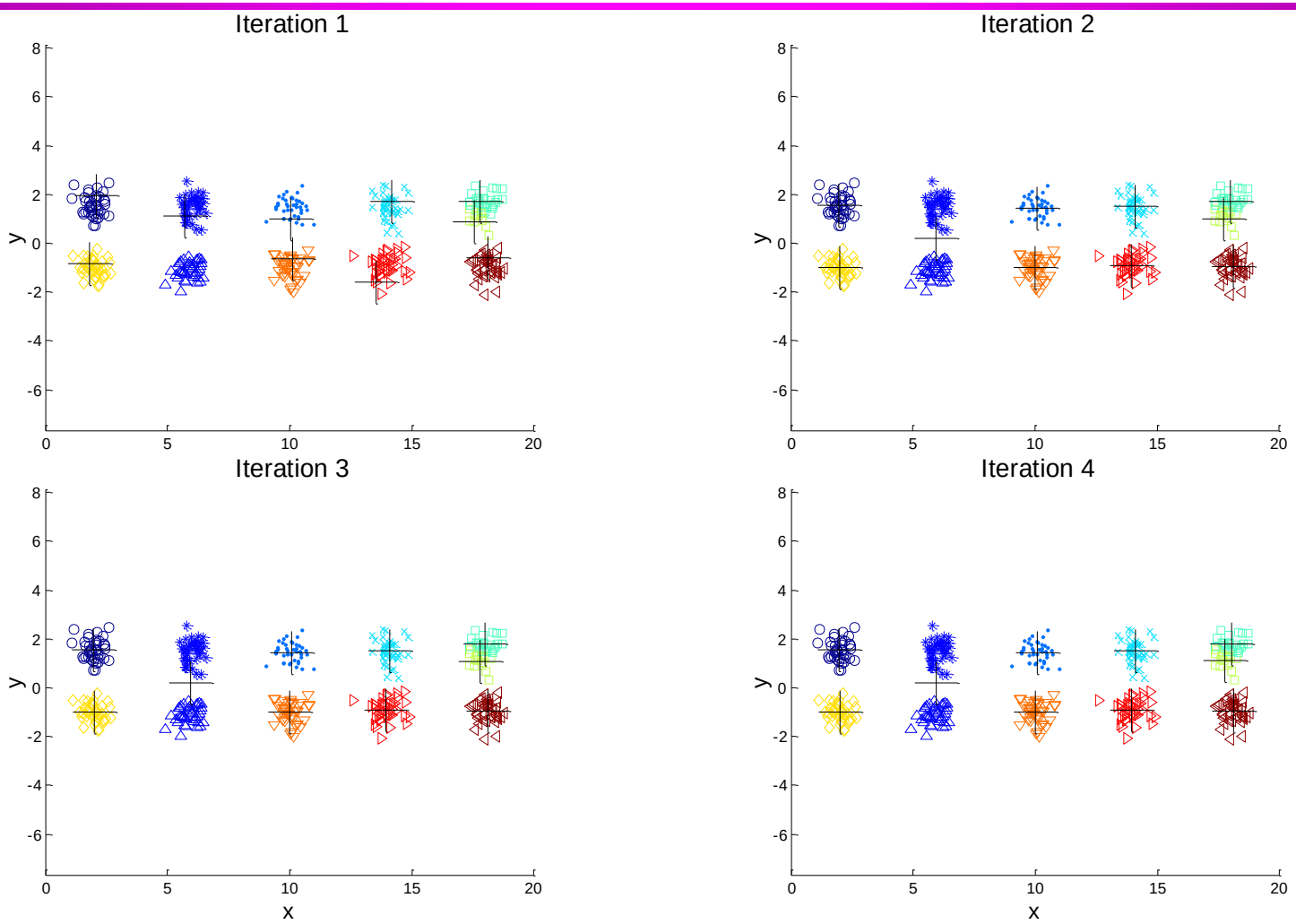
Ξεκινώντας με δύο αρχικά centroids σε μία ομάδα για κάθε ζεύγος ομάδων

Παράδειγμα 10 Ομάδων



Ξεκινώντας με μερικά ζεύγη ομάδων να έχουν 3 αρχικά centroids, ενώ άλλα να έχουν μόνο 1.

Παράδειγμα 10 Ομάδων



Ξεκινώντας με μερικά ζεύγη ομάδων να έχουν 3 αρχικά centroids, ενώ άλλα να έχουν μόνο 1.

Λύσεις για το πρόβλημα των αρχικών centroids

- Πολλαπλές εκτελέσεις του αλγορίθμου
  - Βοηθάει, αλλά οι πιθανότητες δεν είναι με το μέρος μας
- Δειγματοληψία και χρήση ιεραρχικής ομαδοποίησης για να προσδιοριστούν τα αρχικά centroids
- Επιλέγουμε περισσότερα από k αρχικά centroids και μετά εκλέγουμε μερικά μεταξύ αυτών των αρχικών centroids με κάποιο κριτήριο
  - Επιλέγουμε τα πιο διαχωρισμένα
- Μεταγενέστερη επεξεργασία (post-processing)
- Αλγόριθμος Bisecting K-means
  - Δεν είναι τόσο ευαίσθητος σε θέματα αρχικοποίησης

Χειρισμός των Κενών Ομάδων

- Ο βασικός αλγόριθμος K-means μπορεί να παράγει και κενές ομάδες
- Διάφορες στρατηγικές
  - Επιλέγεται ένα σημείο που συμβάλλει το περισσότερο στο μέτρο SSE (και τοποθετείται στην κενή ομάδα)
  - Επιλέγεται ένα σημείο από την ομάδα που έχει το υψηλότερο SSE (και τοποθετείται στην κενή ομάδα)
  - Εάν υπάρχουν πολλές κενές ομάδες, η παραπάνω διαδικασία μπορεί να επαναληφθεί πολλές φορές.

Ενημερώνοντας τα Κέντρα Αυξητικά

- Στον βασικό αλγόριθμο K-means, τα centroids ενημερώνονται όταν όλα τα σημεία έχουν ανατεθεί σε κάποιο centroid
- Μία εναλλακτική λύση είναι να ενημερώνονται τα centroids μετά από κάθε ανάθεση (αυξητική προσέγγιση)
  - Κάθε ανάθεση ενημερώνει από 0 έως 2 centroids
  - Είναι υπολογιστικά ακριβή λύση
  - Εισάγει μία εξάρτηση διάταξης
  - Ποτέ δεν παράγει κενή ομάδα
  - Μπορεί να χρησιμοποιήσει «βάρη» για να μεταβάλει την επίδραση

Pre-processing και Post-processing

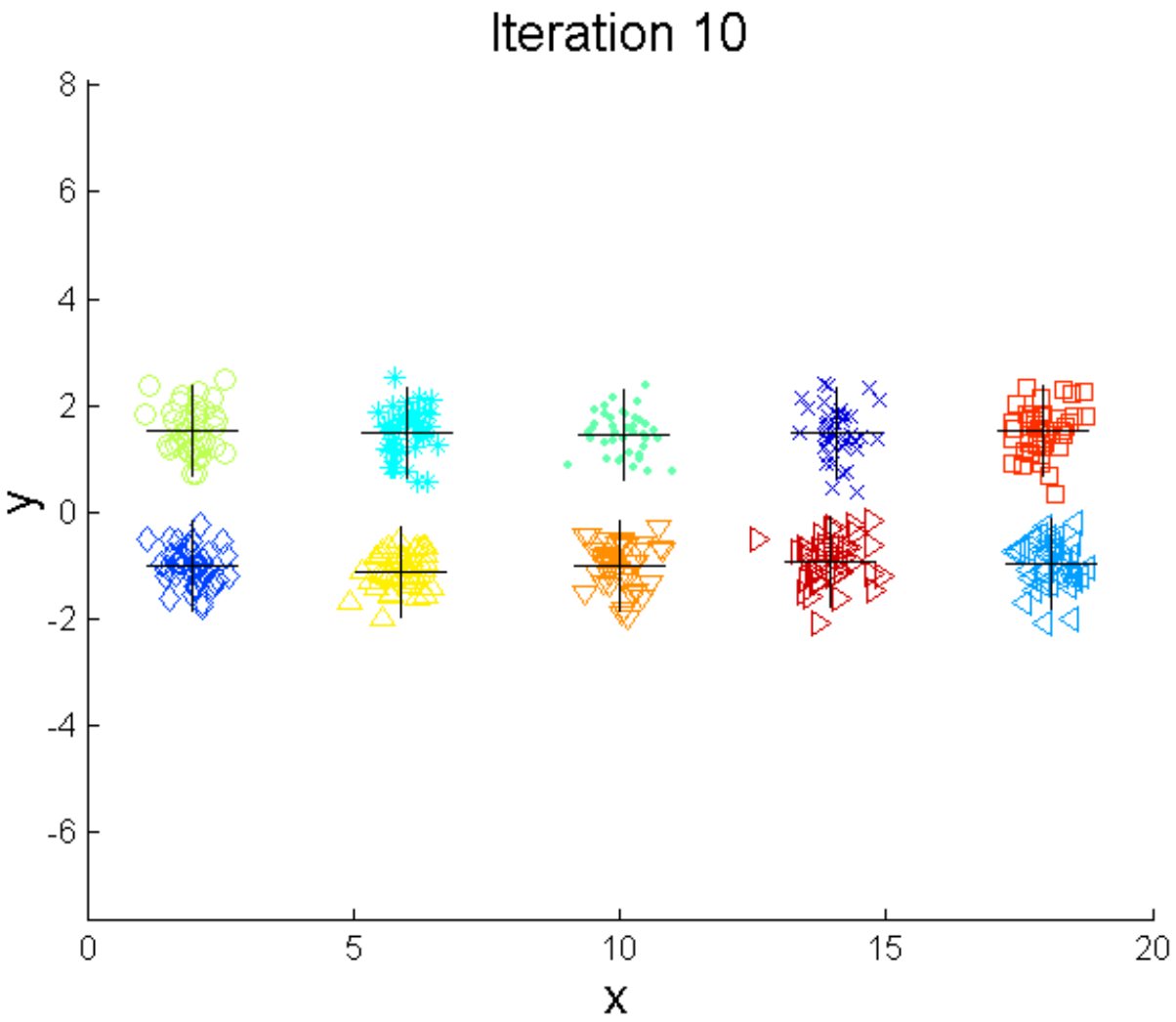
- Pre-processing
  - Τα δεδομένα κανονικοποιούνται
  - Εξαλείφονται οι outliers
- Post-processing
  - Εξαλείφονται οι μικρές ομάδες οι οποίες μπορεί να αντιπροσωπεύουν outliers
  - Διαχωρίζονται οι «χαλαρές» ομάδες, δηλαδή οι ομάδες με σχετικά υψηλό SSE
  - Συγχωνεύονται οι «κοντινές» ομάδες οι οποίες έχουν σχετικά χαμηλό SSE
  - Αυτά τα βήματα μπορούν να χρησιμοποιηθούν και κατά τη διάρκεια της διαδικασίας της ομαδοποίησης
    - Αλγόριθμος ISODATA

Αλγόριθμος Bisecting K-means

- Ο αλγόριθμος Bisecting K-means
  - Είναι μία παραλλαγή του K-means ο οποίος μπορεί να παράγει μία ομαδοποίηση διαμέρισης ή μία ιεραρχική ομαδοποίηση

1: Initialize the list of clusters to contain the cluster containing all points.
2: **repeat**
3: Select a cluster from the list of clusters
4: **for** $i = 1$ to *number_of_iterations* **do**
5: Bisect the selected cluster using basic K-means
6: **end for**
7: Add the two clusters from the bisection with the lowest SSE to the list of clusters.
8: **until** Until the list of clusters contains K clusters

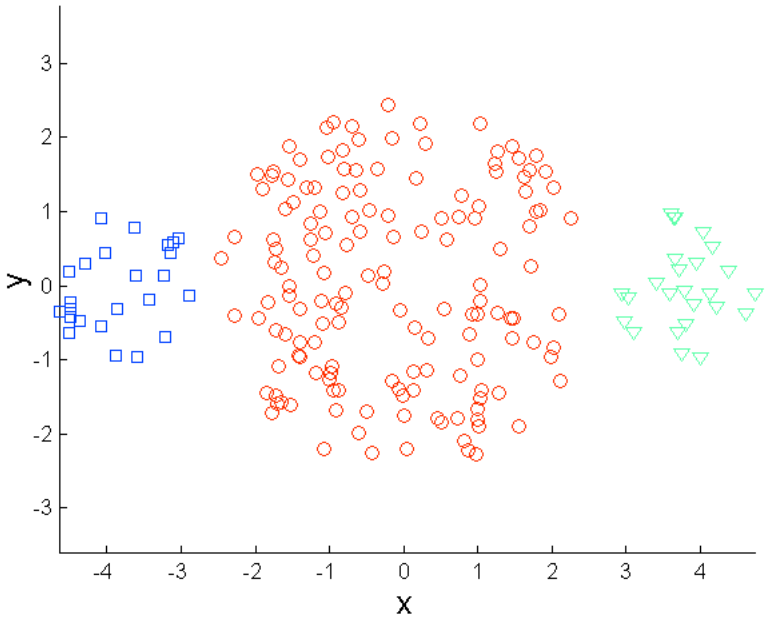
Παράδειγμα του Bisecting K-means



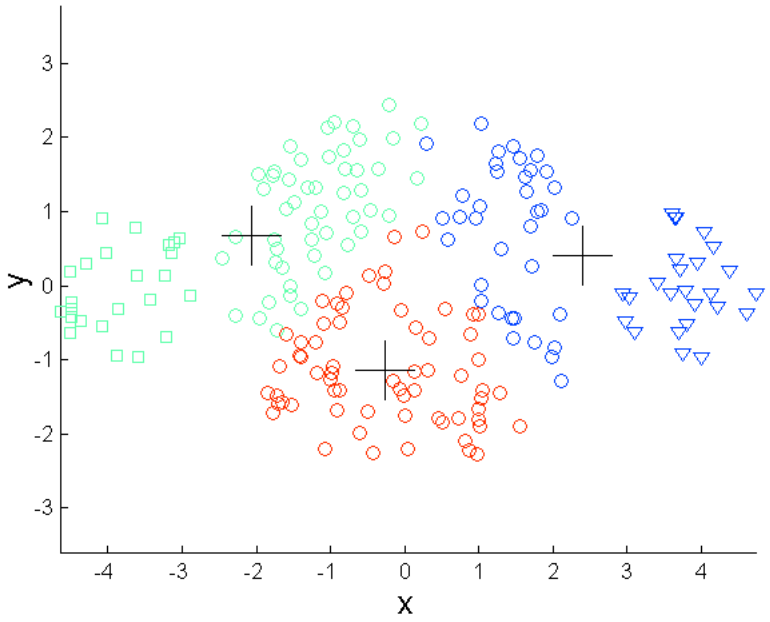
Περιορισμοί του K-means

- Ο αλγόριθμος K-means παρουσιάζει προβλήματα όταν οι ομάδες διαφέρουν σε:
  - Μεγέθη
  - Πυκνότητες
  - Ή έχουν μη-σφαιρικά σχήματα
- Ο αλγόριθμος K-means παρουσιάζει προβλήματα όταν τα δεδομένα περιλαμβάνουν outliers.

Περιορισμοί του K-means: Διαφορετικά Μεγέθη

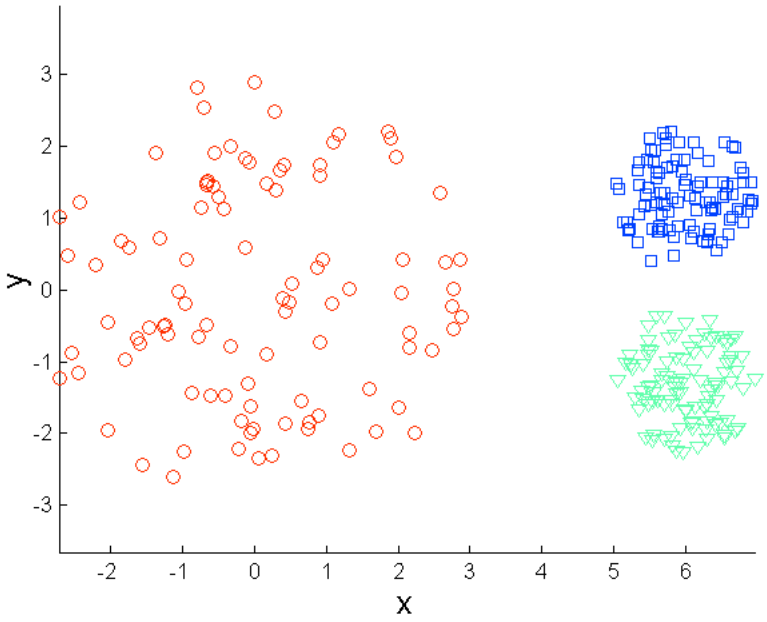


Αρχικά Σημεία

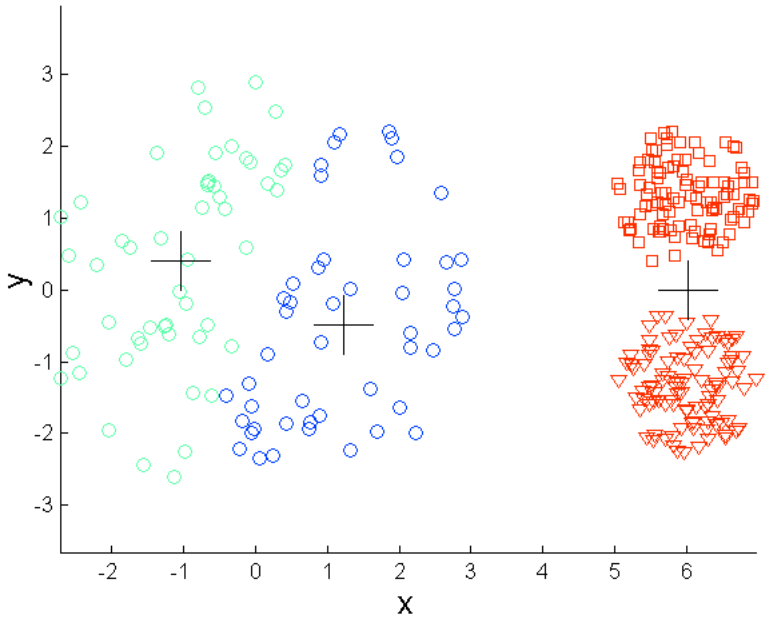


K-means (3 Ομάδες)

Περιορισμοί του K-means: Διαφορετικές Πυκνότητες

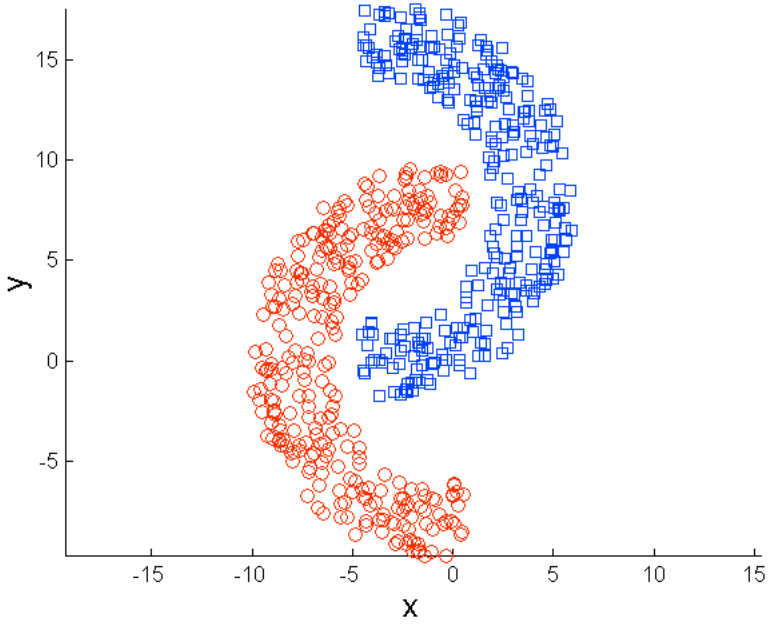


Αρχικά Σημεία

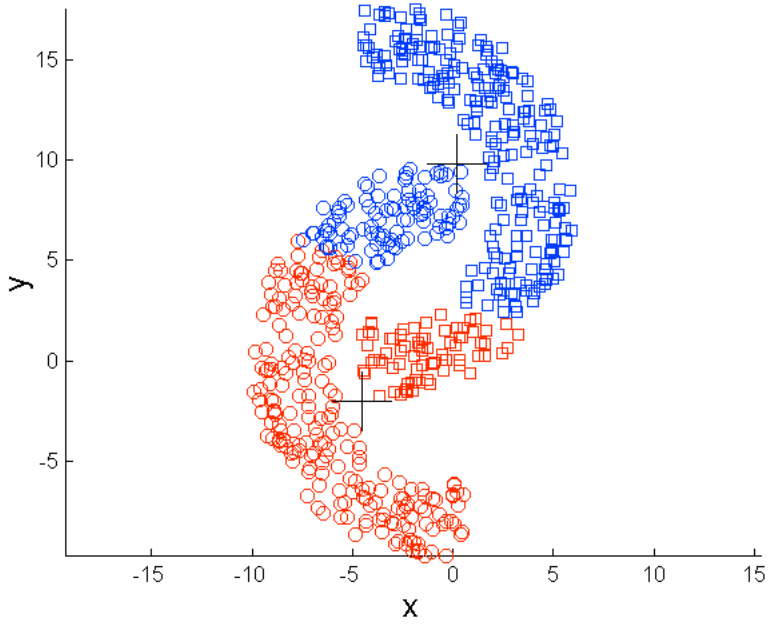


K-means (3 Ομάδες)

Περιορισμοί του K-means: Μη-σφαιρικά Σχήματα

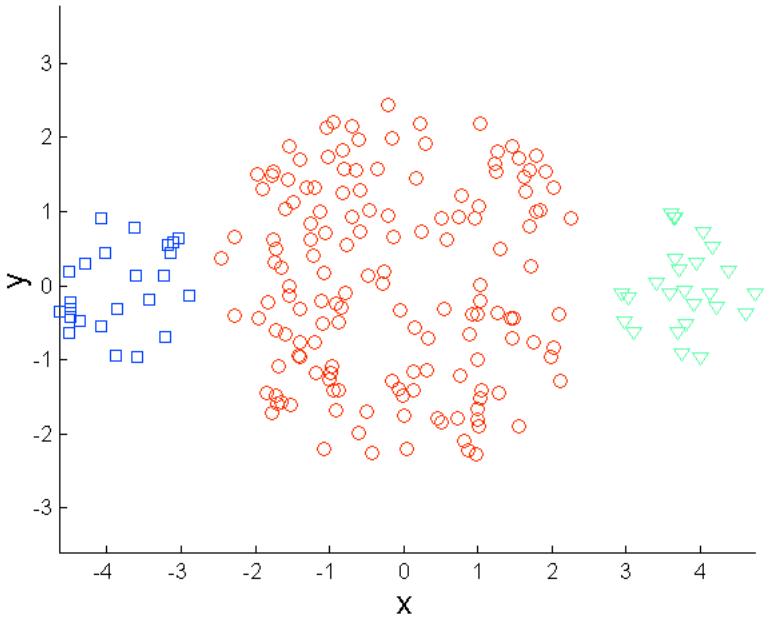


Αρχικά Σημεία

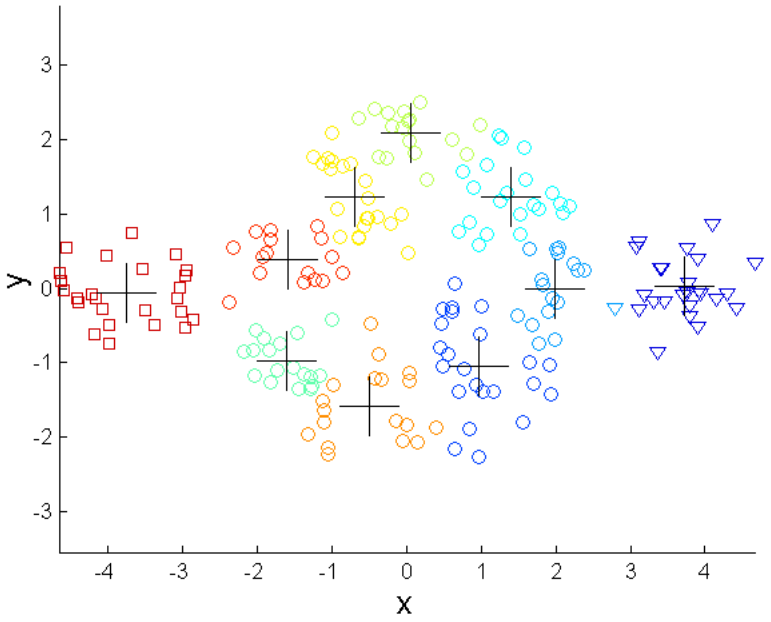


K-means (2 Ομάδες)

Ξεπερνώντας τους περιορισμούς του K-means



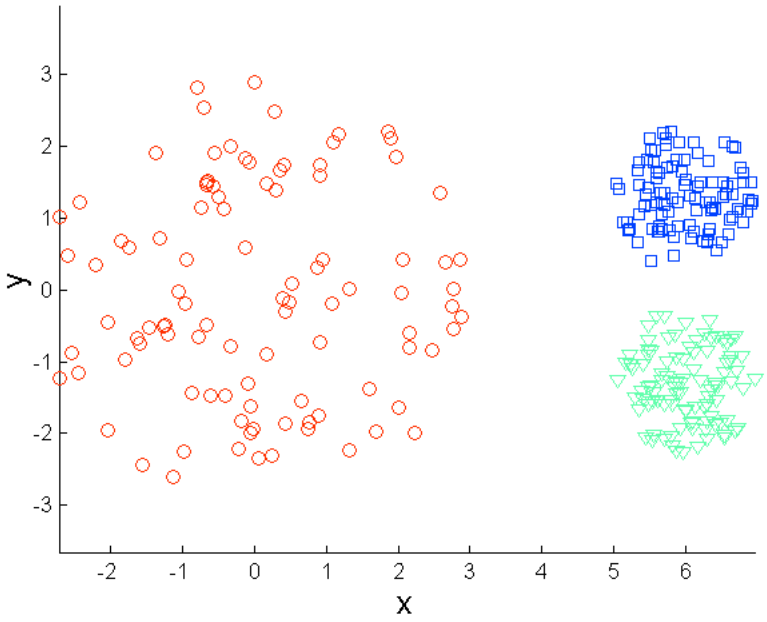
Αρχικά Σημεία



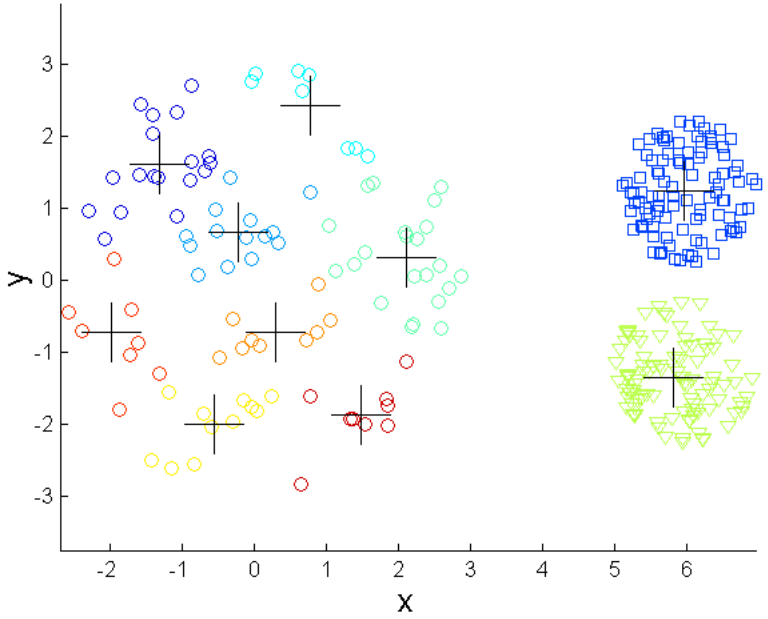
Ομάδες του K-means

Μία λύση είναι να χρησιμοποιήσουμε πολλές ομάδες.
Βρίσκει τμήματα των ομάδων αλλά πρέπει να τα βάλλει μαζί.

Ξεπερνώντας τους περιορισμούς του K-means

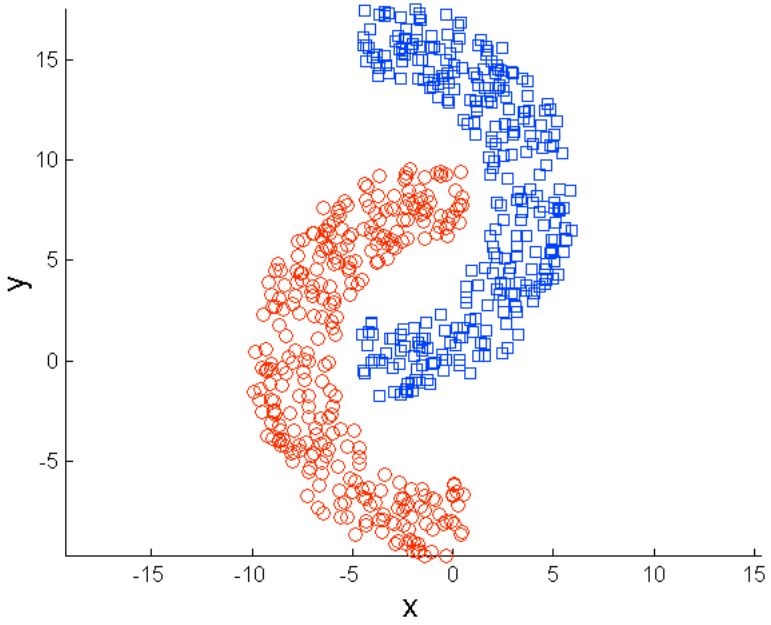


Αρχικά Σημεία

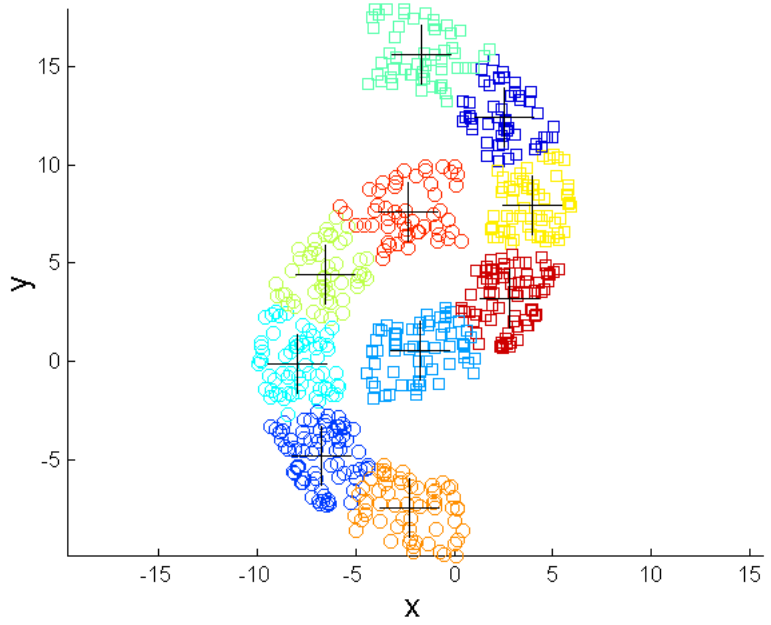


Ομάδες του K-means

Ξεπερνώντας τους περιορισμούς του K-means



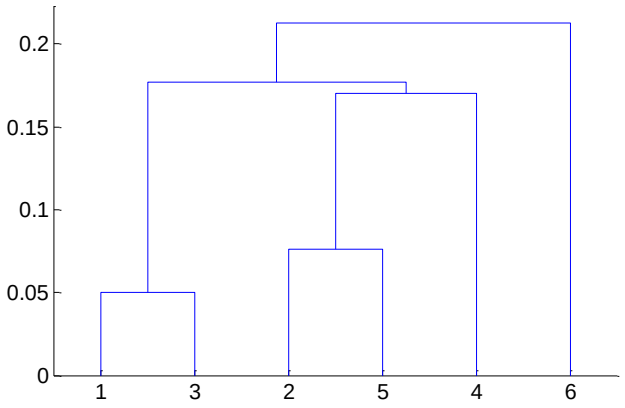
Αρχικά Σημεία

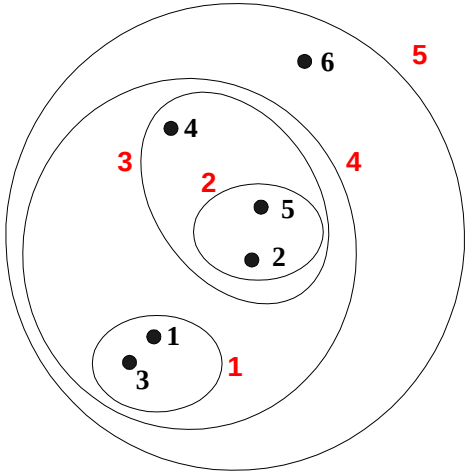


Ομάδες του K-means

Ιεραρχική Ομαδοποίηση (Hierarchical Clustering)

- Παράγει ένα σύνολο εμφωλευμένων ομάδων που είναι οργανωμένες σε ένα ιεραρχικό δέντρο
- Μπορεί να απεικονιστεί ως ένα δενδροδιάγραμμα
  - Ένα διάγραμμα που μοιάζει με δέντρο και καταγράφει την ακολουθία συγχώνευσης ή διαχωρισμού





Πλεονεκτήματα της Ιεραρχικής Ομαδοποίησης

- Δεν χρειάζεται να θεωρήσουμε κάποιο συγκεκριμένο αριθμό ομάδων
  - Οποιοσδήποτε επιθυμητός αριθμός ομάδων μπορεί να προκύψει με «αποκοπή» στο σωστό επίπεδο του δενδροδιαγράμματος
- Οι ομάδες μπορεί να αντιστοιχούν σε εννοιολογικές τάξεις
  - Παράδειγμα στις βιολογικές επιστήμες (π.χ., βασίλειο των ζώων, ανακατασκευή φυλογενετικής, κλπ.)

Ιεραρχική Ομαδοποίηση

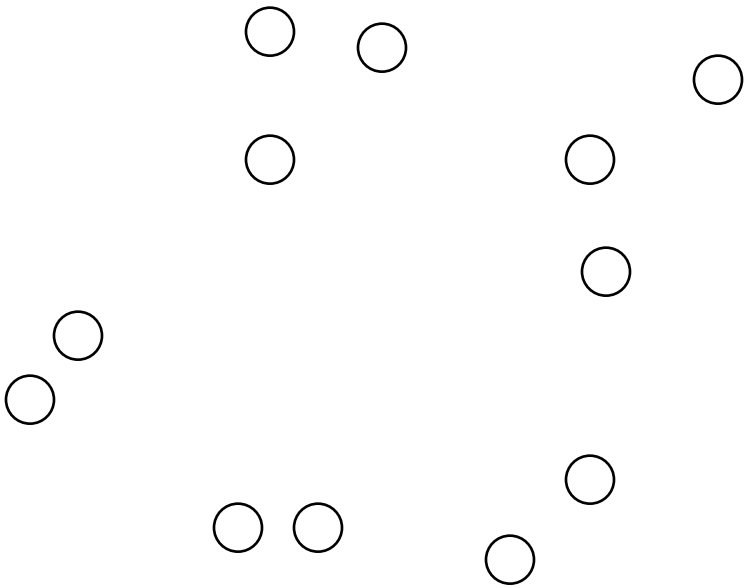
- Υπάρχουν δύο βασικοί τύποι ιεραρχικής ομαδοποίησης
  - Ομαδοποίηση Επικόλλησης (Agglomerative):
    - Ξεκινά με τα σημεία να αποτελούν ανεξάρτητες ομάδες
    - Σε κάθε βήμα συγχωνεύεται το πλησιέστερο ζεύγος ομάδων μέχρι να μείνει μόνο μία ομάδα (ή k ομάδες)
  - Ομαδοποίηση Διαίρεσης (Divisive):
    - Ξεκινά με μία ομάδα που περιέχει όλα τα σημεία
    - Σε κάθε βήμα διαιρείται μία ομάδα μέχρι κάθε ομάδα να περιέχει ένα μόνο σημείο (ή αν δημιουργηθούν k ομάδες)
- Οι κλασικοί ιεραρχικοί αλγόριθμοι χρησιμοποιούν έναν πίνακα ομοιότητας ή αποστάσεων (similarity-distance matrix)
  - Συγχωνεύεται ή διασπάται μία ομάδα κάθε φορά

Αλγόριθμος Ομαδοποίησης Επικόλλησης

- Είναι η πιο δημοφιλής τεχνική ιεραρχικής ομαδοποίησης
- Ο βασικός αλγόριθμος είναι απλός:
  1. Compute the proximity matrix
  2. Let each data point be a cluster
  3. **Repeat**
  4. Merge the two closest clusters
  5. Update the proximity matrix
  6. **Until** only a single cluster remains
- Η βασική λειτουργία του είναι ο υπολογισμός της εγγύτητας μεταξύ δύο ομάδων
  - Οι διαφορετικές προσεγγίσεις για τον ορισμό της απόστασης μεταξύ των ομάδων διακρίνουν και τους διαφορετικούς αλγόριθμους

Αρχική Κατάσταση

- Ξεκινά με ομάδες μεμονωμένων σημείων και έναν πίνακα εγγύτητας:

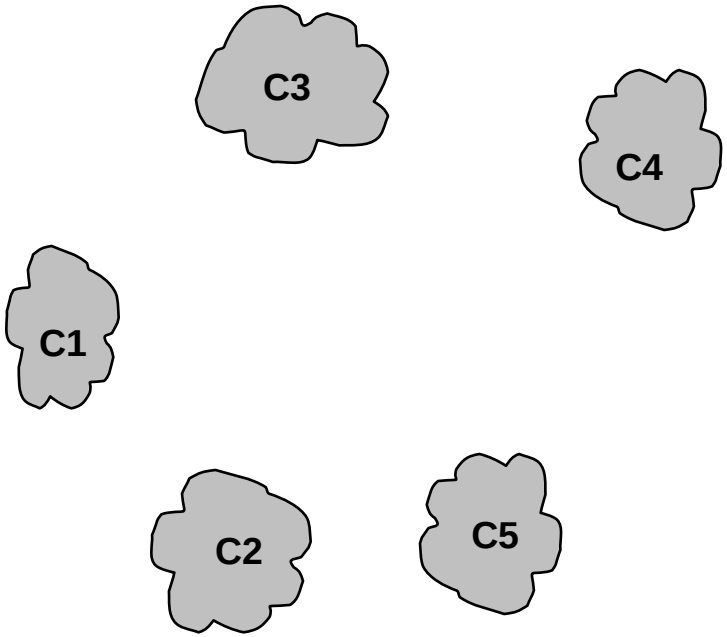


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

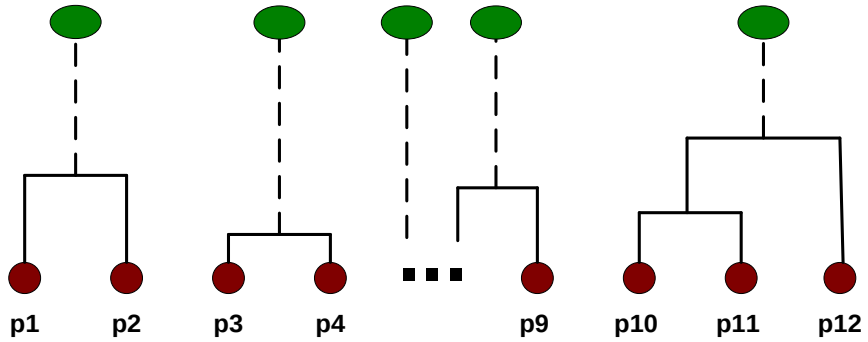
Ενδιάμεση Κατάσταση

- Μετά από κάποια βήματα συγχώνευσης έχουμε μερικές ομάδες:



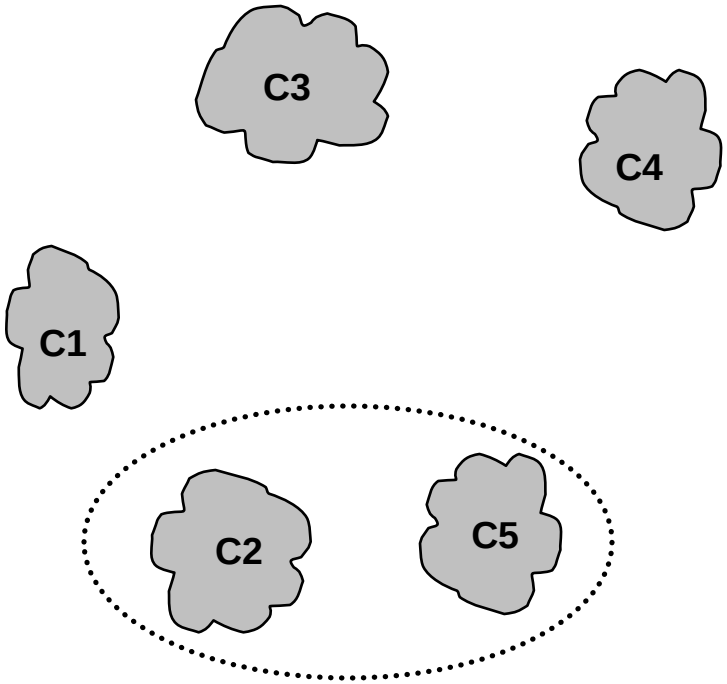
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Proximity Matrix



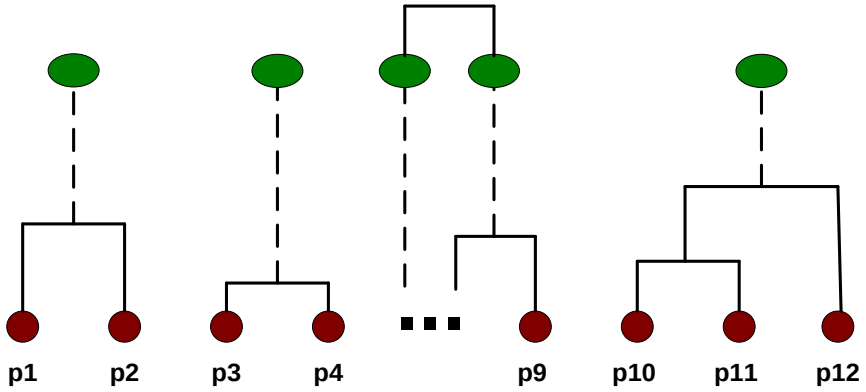
Ενδιάμεση Κατάσταση

- Θέλουμε να συγχωνεύσουμε τις δύο πλησιέστερες ομάδες (C2 και C5) και να ενημερώσουμε τον πίνακα εγγύτητας.



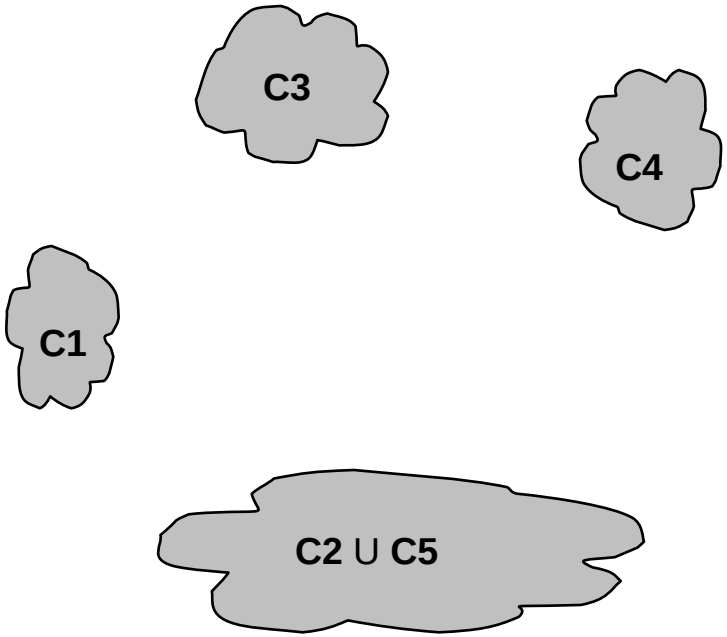
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Proximity Matrix



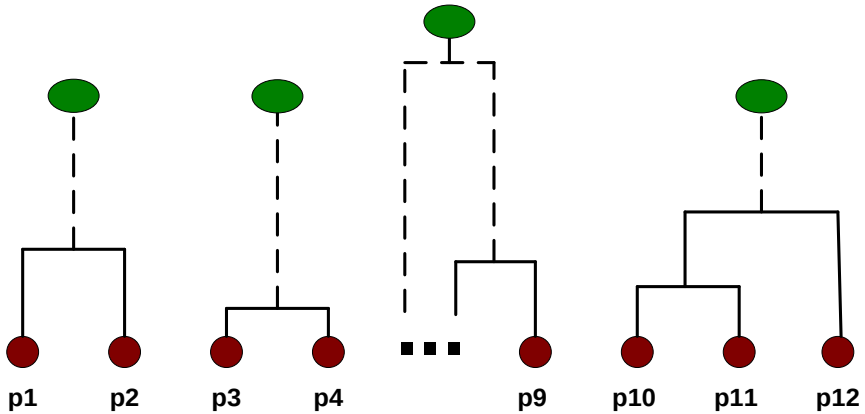
Μετά τη συγχώνευση

- Το ερώτημα είναι: «Πώς θα ενημερώσουμε τον πίνακα εγγύτητας;»

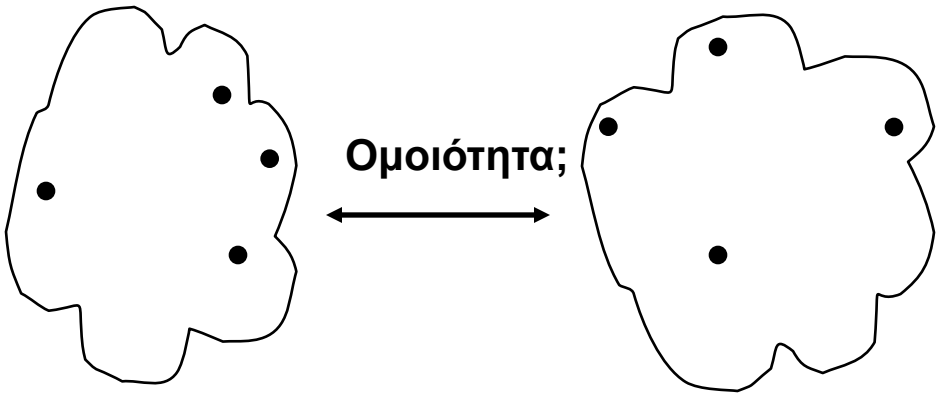


	C1	C2 U C5	C3	C4
C1		?		
C2 U C5	?	?	?	?
C3		?		
C4		?		

Proximity Matrix



Πώς θα οριστεί η ομοιότητα μεταξύ ομάδων



- MIN
- MAX
- Group Average
- Απόσταση μεταξύ Centroids
- Άλλες μέθοδοι που καθοδηγούνται από μία συνάρτηση στόχου
  - Η μέθοδος Ward χρησιμοποιεί το τετραγωνικό σφάλμα

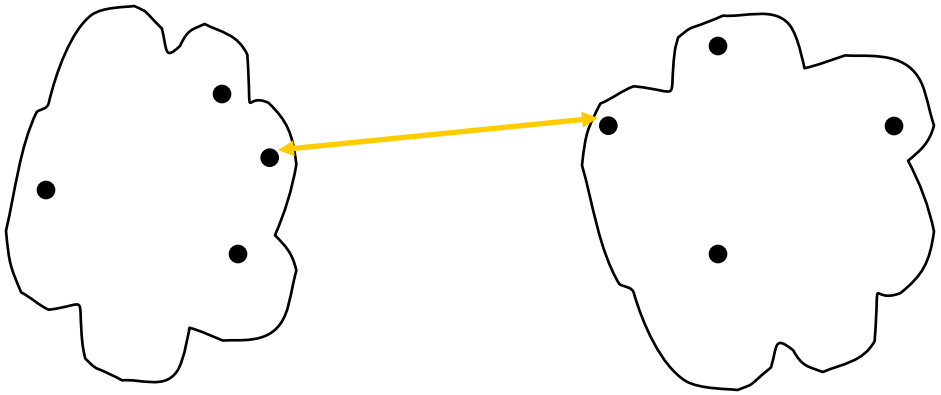
	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						

.

.

Proximity Matrix

Πώς θα οριστεί η ομοιότητα μεταξύ ομάδων



- MIN
- MAX
- Group Average
- Απόσταση μεταξύ Centroids
- Άλλες μέθοδοι που καθοδηγούνται από μία συνάρτηση στόχου
  - Η μέθοδος Ward χρησιμοποιεί το τετραγωνικό σφάλμα

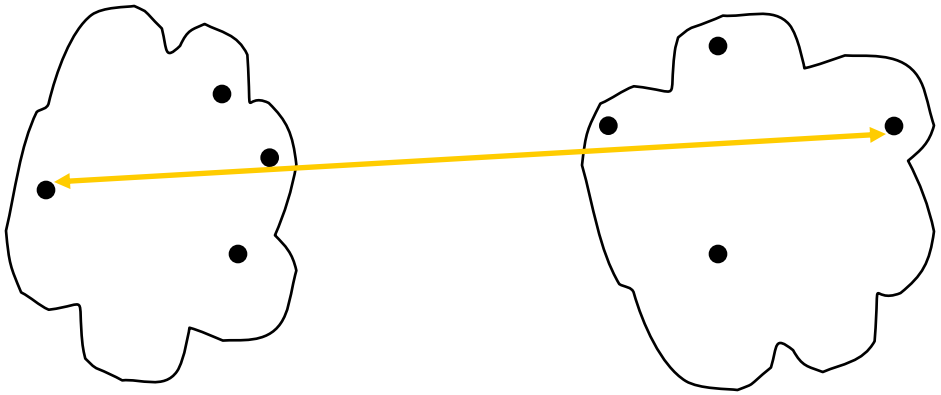
	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						

.

.

Proximity Matrix

Πώς θα οριστεί η ομοιότητα μεταξύ ομάδων



- MIN
- MAX
- Group Average
- Απόσταση μεταξύ Centroids
- Άλλες μέθοδοι που καθοδηγούνται από μία συνάρτηση στόχου
  - Η μέθοδος Ward χρησιμοποιεί το τετραγωνικό σφάλμα

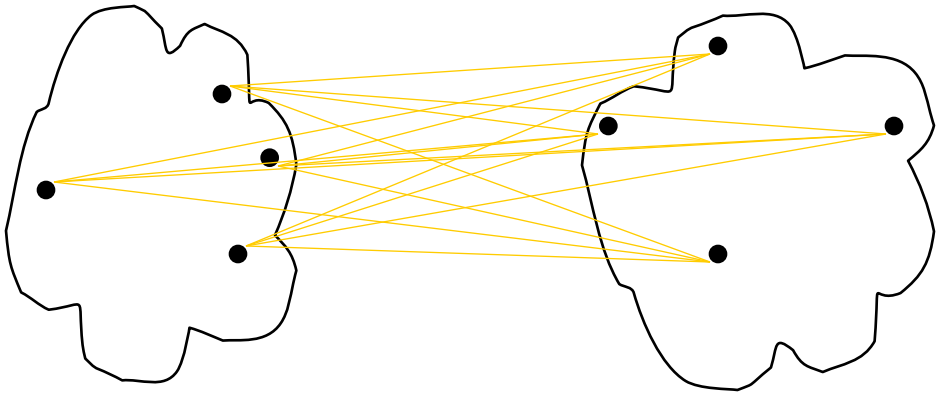
	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						

.

.

Proximity Matrix

Πώς θα οριστεί η ομοιότητα μεταξύ ομάδων



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

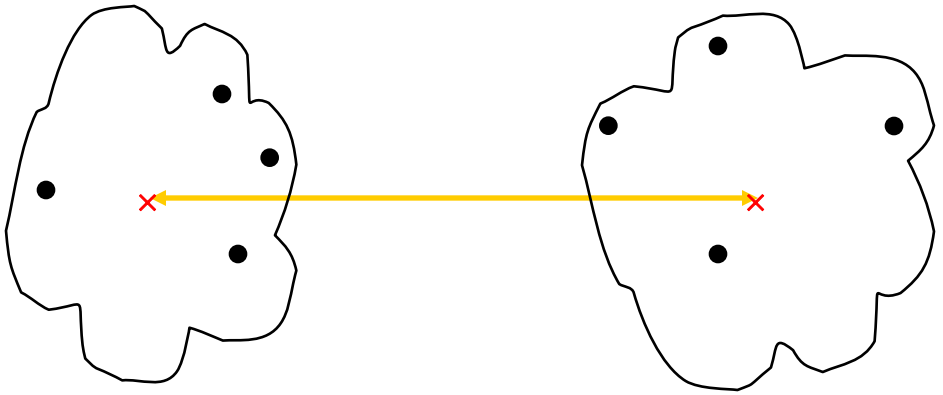
.

.

Proximity Matrix

- MIN
- MAX
- Group Average
- Απόσταση μεταξύ Centroids
- Άλλες μέθοδοι που καθοδηγούνται από μία συνάρτηση στόχου
  - Η μέθοδος Ward χρησιμοποιεί το τετραγωνικό σφάλμα

Πώς θα οριστεί η ομοιότητα μεταξύ ομάδων



- MIN
- MAX
- Group Average
- Απόσταση μεταξύ Centroids
- Άλλες μέθοδοι που καθοδηγούνται από μία συνάρτηση στόχου
  - Η μέθοδος του Ward χρησιμοποιεί το τετραγωνικό σφάλμα

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

.

.

Proximity Matrix

Παράδειγμα

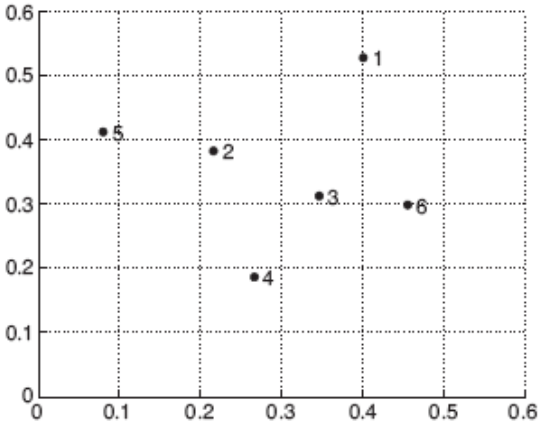


Figure 8.15. Set of 6 two-dimensional points.

Point	x Coordinate	y Coordinate
p1	0.40	0.53
p2	0.22	0.38
p3	0.35	0.32
p4	0.26	0.19
p5	0.08	0.41
p6	0.45	0.30

Table 8.3. xy coordinates of 6 points.

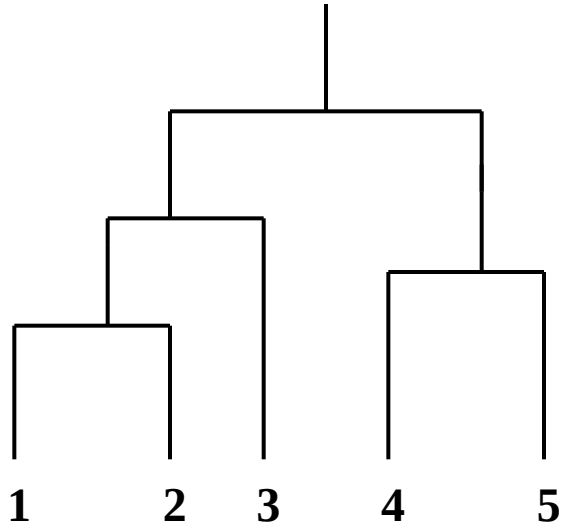
	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

Table 8.4. Euclidean distance matrix for 6 points.

Ομοιότητα Ομάδων: MIN (απλό link)

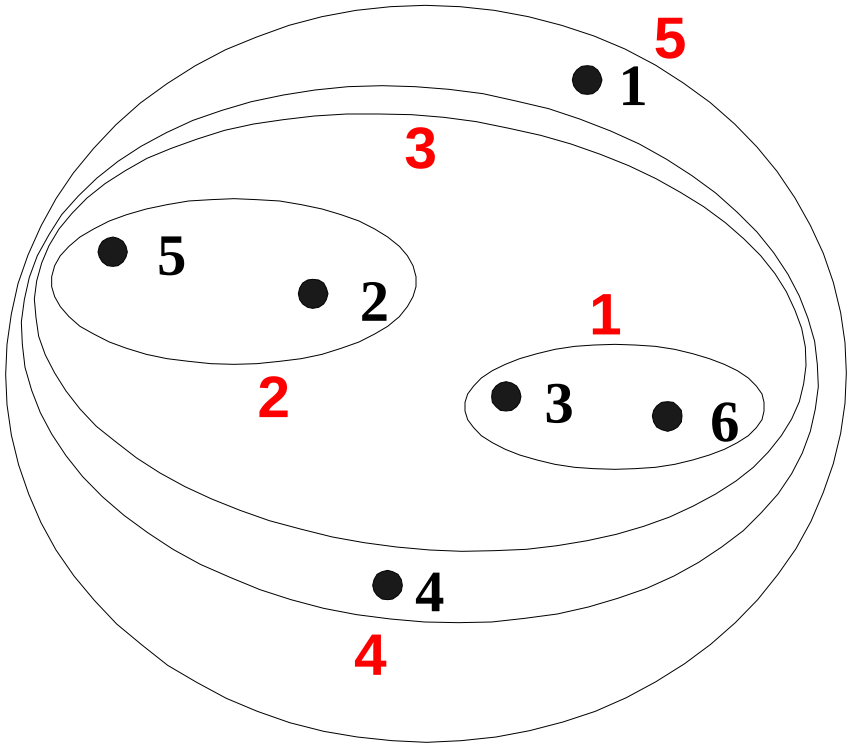
- Η ομοιότητα δύο ομάδων βασίζεται στα δύο πιο όμοια (πλησιέστερα) σημεία των διαφορετικών αυτών ομάδων
  - Προσδιορίζεται μόνο από ένα ζεύγος σημείων, δηλαδή από ένα απλό link του πίνακα εγγύτητας.

	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

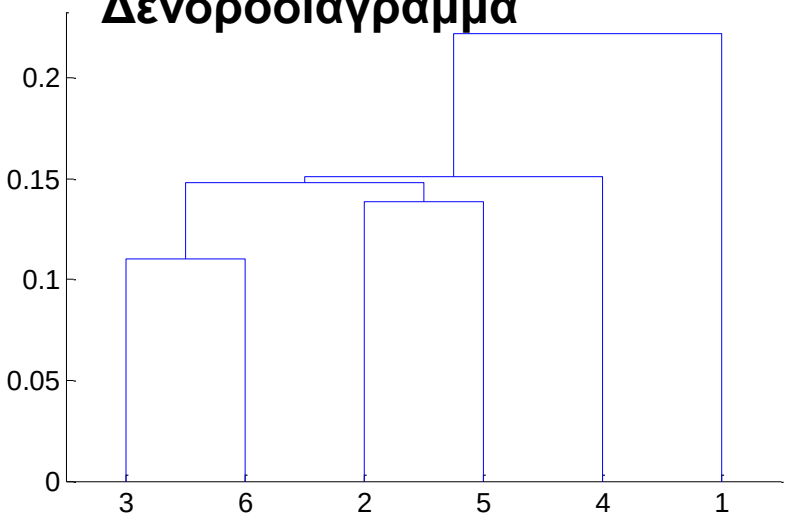


Ιεραρχική Ομαδοποίηση: MIN

Εμφωλευμένες Ομάδες



Δενδροδιάγραμμα



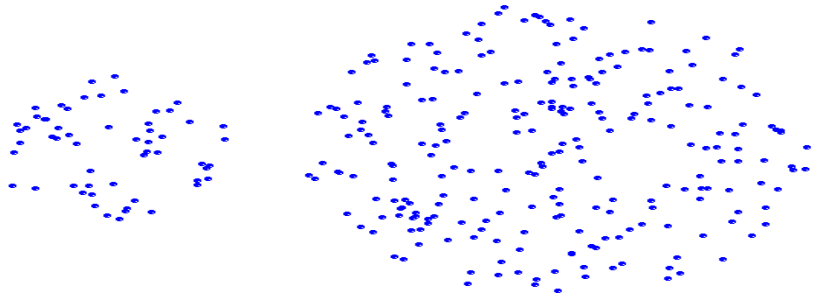
	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

Dist({3,6},{2,5}) = min(dist(3,2), dist(6,2),

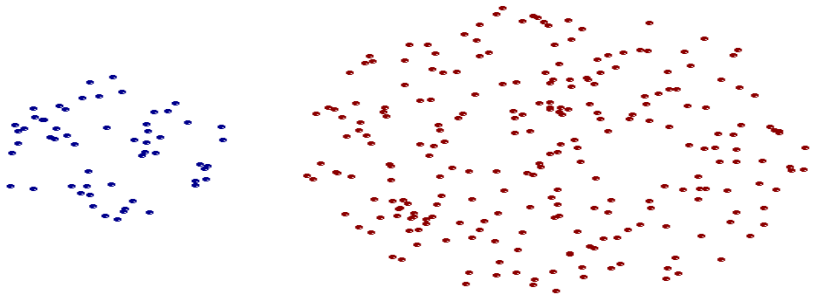
dist(3,5), dist(6,5))= min(0.15, 0.25, 0.28, 0.39)

=0.15

Πλεονεκτήματα του MIN



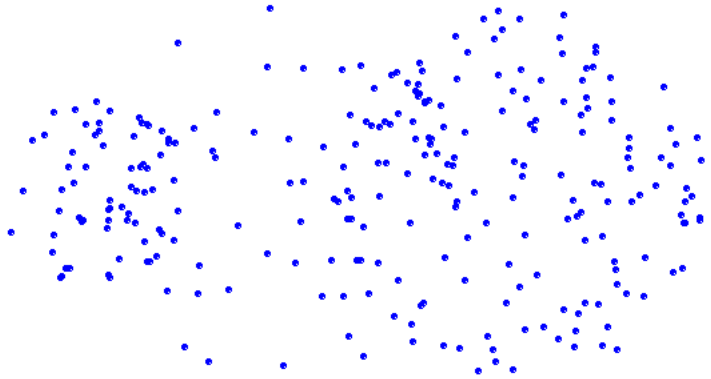
Αρχικά Σημεία



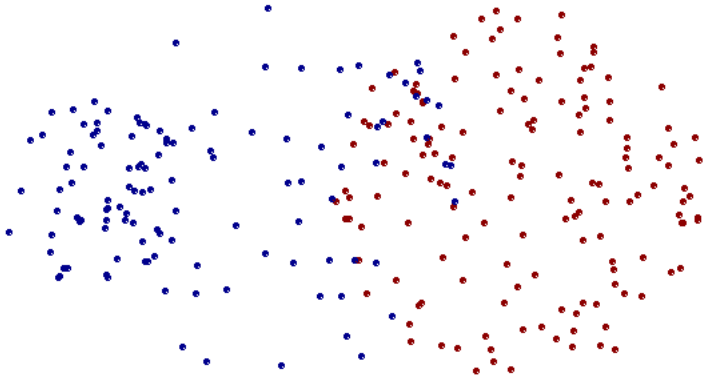
Δύο Ομάδες

- Μπορεί να χειριστεί μη-ελλειπτικά σχήματα

Μειονεκτήματα του MIN



Αρχικά Σημεία



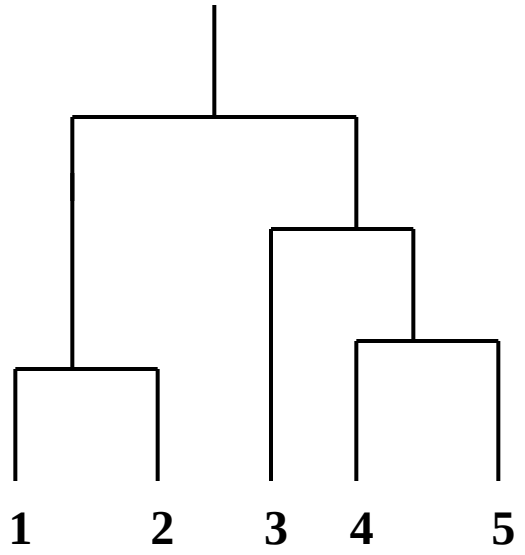
Δύο Ομάδες

- Είναι ευαίσθητο σε θόρυβο και outliers

Ομοιότητα Ομάδων: MAX (πλήρης σύνδεση)

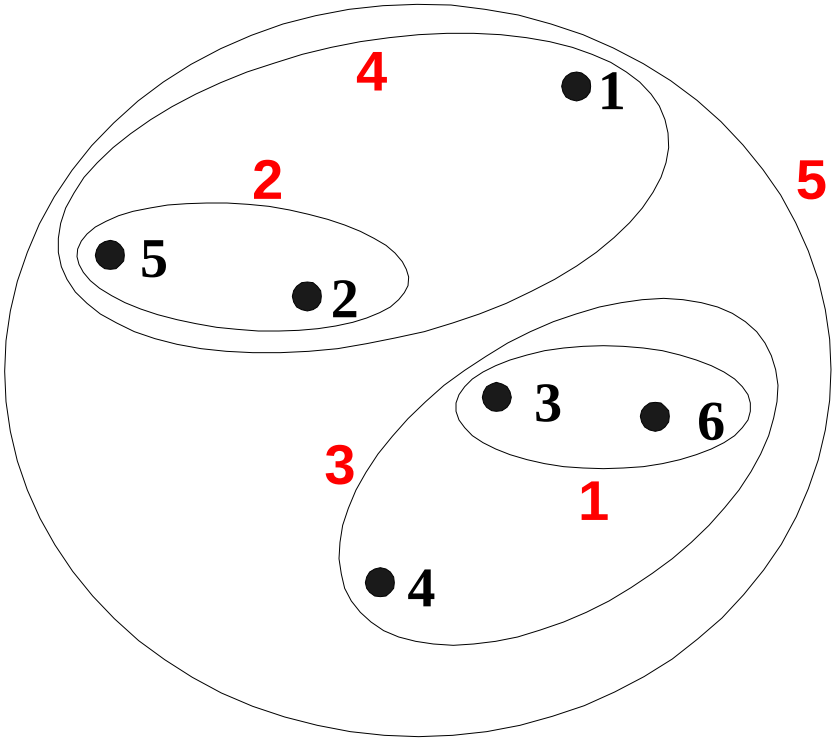
- Η ομοιότητα δύο ομάδων βασίζεται στα δύο λιγότερο όμοια (τα πιο απομακρυσμένα) σημεία των διαφορετικών αυτών ομάδων
  - Προσδιορίζεται από όλα τα ζεύγη σημείων μεταξύ των δύο ομάδων

	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

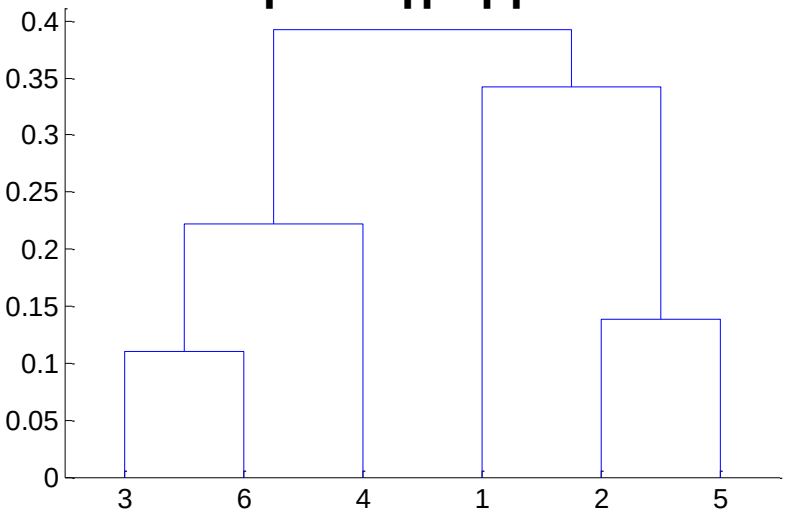


Ιεραρχική Ομαδοποίηση: MAX

Εμφωλευμένες Ομάδες



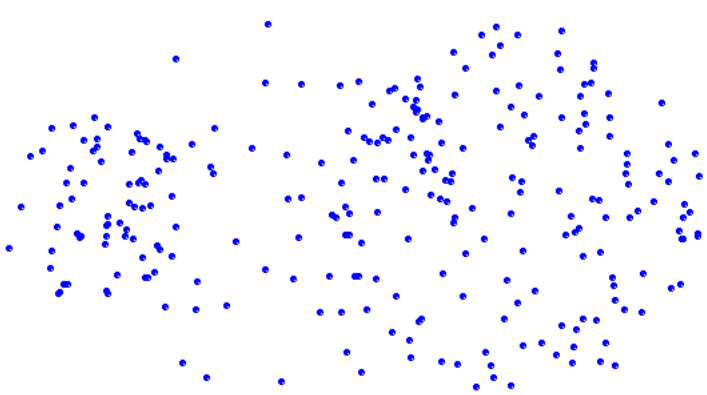
Δενδροδιάγραμμα



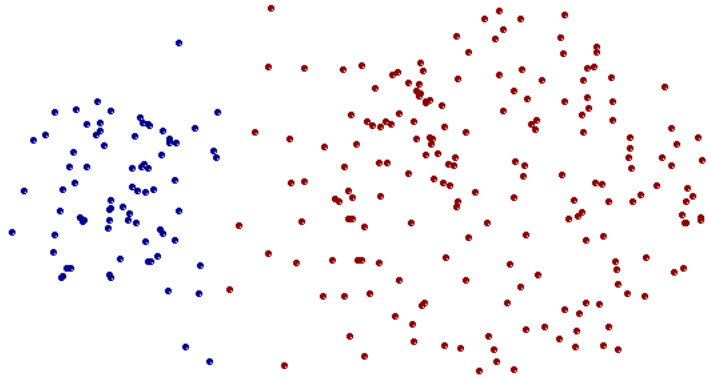
Dist({3,6},{4}) = max(dist(3,4), dist(6,4))= max(0.15, 0.22) =0.22
Dist({3,6},{1}) = max(dist(3,1), dist(6,1))= max(0.22, 0.23) =0.23

	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

Πλεονεκτήματα του MAX



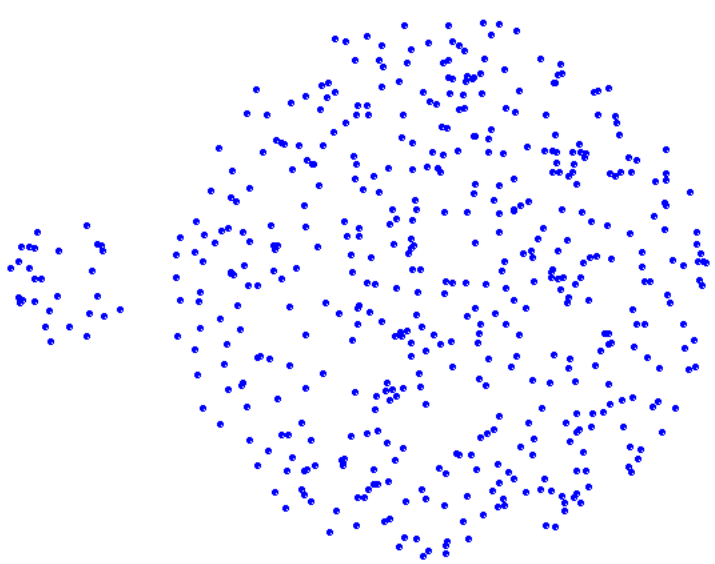
Αρχικά Σημεία



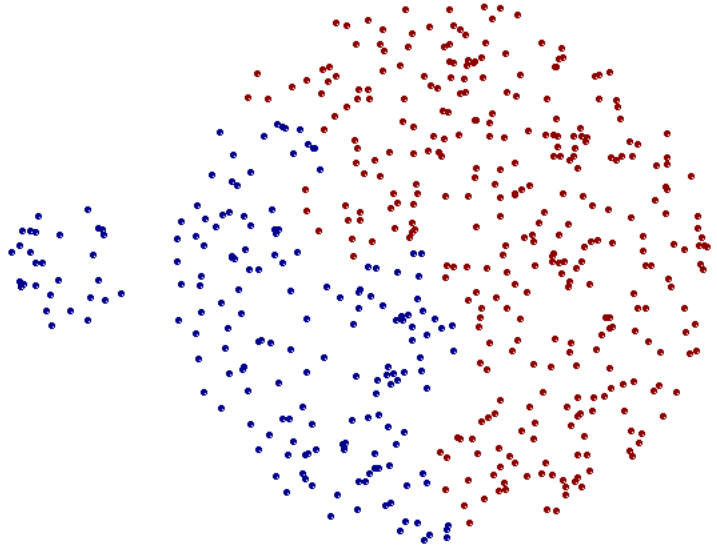
Δύο Ομάδες

- Λιγότερο ευαίσθητο στο θόρυβο και τους outliers

Μειονεκτήματα του MAX



Αρχικά Σημεία



Δύο Ομάδες

- Τείνει να διασπάσει τις μεγάλες ομάδες
- Μεροληπτεί υπέρ των σφαιρικών ομάδων

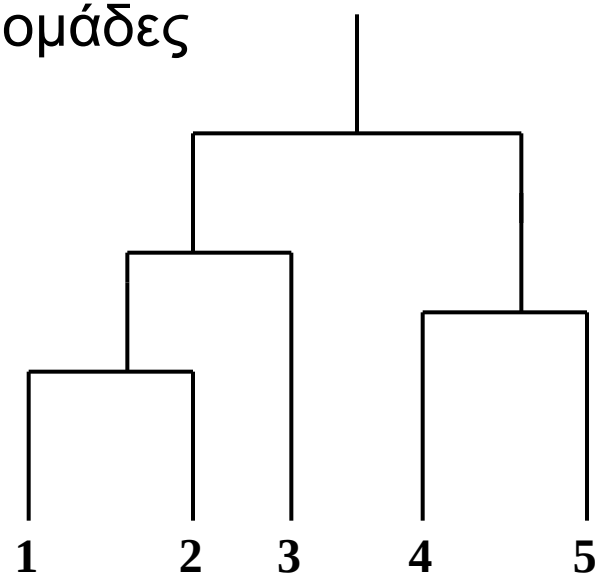
Ομοιότητα Ομάδων: Group Average

- Η εγγύτητα δύο ομάδων είναι ο μέσος όρος των τιμών εγγύτητας από όλα τα ζεύγη σημείων μεταξύ των δύο ομάδων.

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

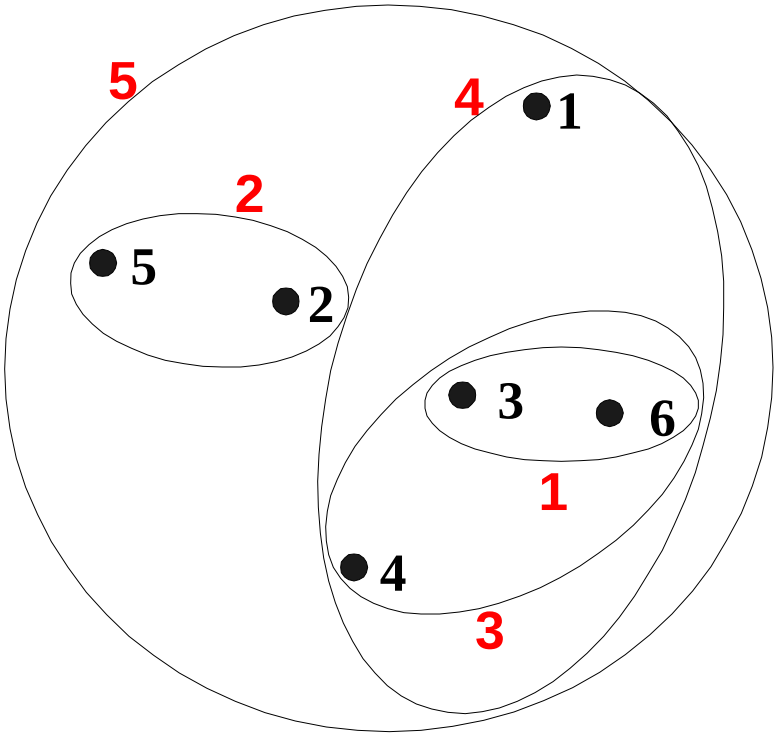
- Χρησιμοποιείται η μέση συνδεσιμότητα για κλιμάκωση καθώς η συνολική εγγύτητα ευνοεί τις μεγάλες ομάδες

	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

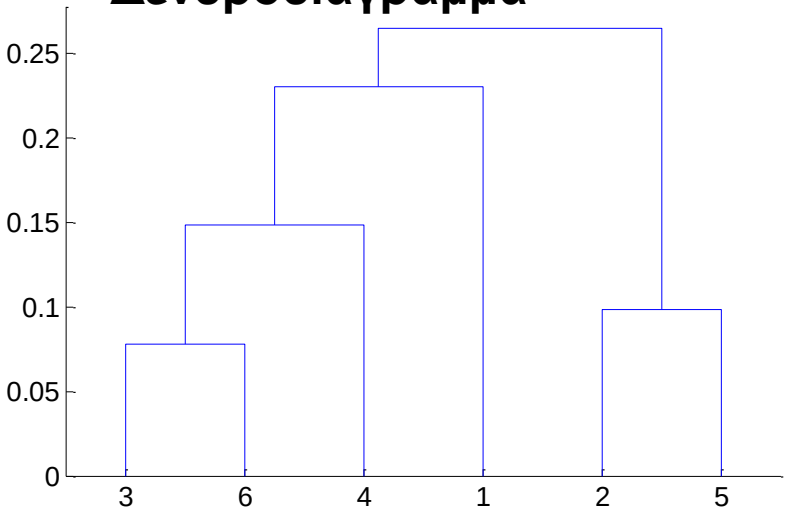


Ιεραρχική Ομαδοποίηση: Group Average

Εμφωλευμένες Ομάδες



Δενδροδιάγραμμα



	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

Dist({3,6,4},{1}) = (0.22 + 0.23 + 0.37) / (3*1) =0.28

Dist({3,6,4},{2,5}) = (0.15+0.28+0.25+0.39+0.20+0.29)/ (3*2) =0.26

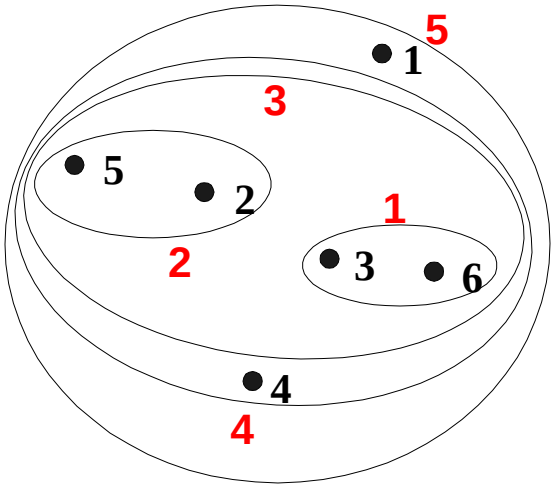
Ιεραρχική Ομαδοποίηση: Group Average

- Βρίσκεται κάπου μεταξύ απλής και πλήρης σύνδεσης
- Πλεονεκτήματα
  - Λιγότερο ευαίσθητο σε θόρυβο και outliers
- Μειονεκτήματα
  - Μεροληπτεί υπέρ των σφαιρικών ομάδων

Ομοιότητα Ομάδων: Μέθοδος Ward

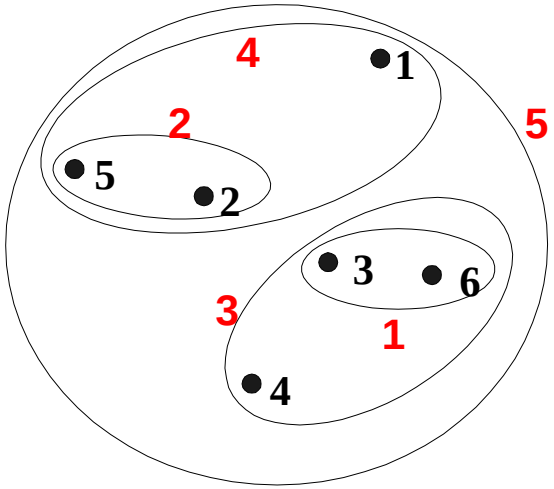
- Η ομοιότητα δύο ομάδων βασίζεται στην αύξηση του τετραγωνικού σφάλματος όταν συγχωνεύονται δύο ομάδες
  - Παρόμοια με το group average όταν η απόσταση μεταξύ των σημείων είναι το τετράγωνο της κανονικής απόστασης
- Λιγότερο ευαίσθητη σε θόρυβο και outliers
- Μεροληπτεί υπέρ των σφαιρικών ομάδων
- Ιεραρχική μέθοδος ανάλογη του K-means
  - Μπορεί να χρησιμοποιηθεί για την αρχικοποίηση του αλγορίθμου K-means

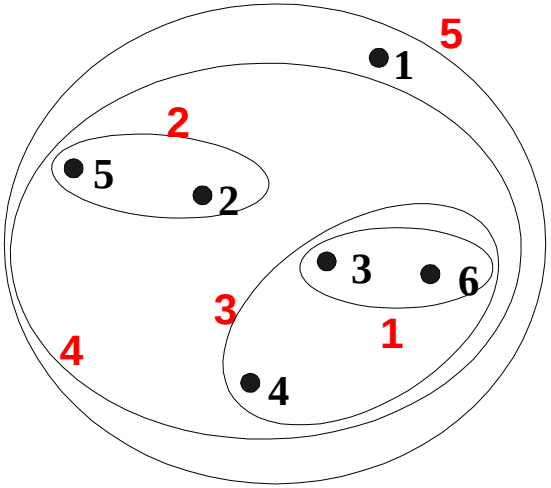
Ιεραρχική Ομαδοποίηση: Σύγκριση



MIN

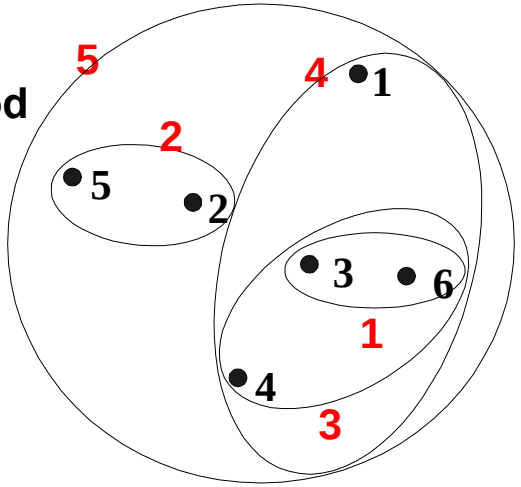
MAX





Group Average

Ward's Method



Ιεραρχική Ομαδοποίηση: Απαιτήσεις Χρόνου και Χώρου

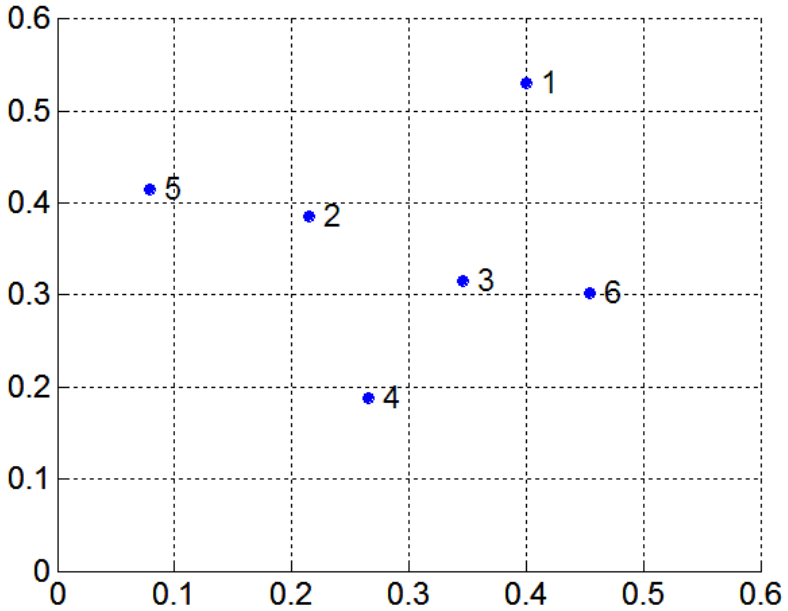
- Χώρος: $O(N^2)$ καθώς χρησιμοποιείται ο πίνακας εγγύτητας
  - N είναι το πλήθος των σημείων
- Χρόνος: $O(N^3)$ σε πολλές περιπτώσεις
  - Υπάρχουν N βήματα και σε κάθε βήμα ο πίνακας εγγύτητας (μεγέθους N^2) πρέπει να διαβαστεί και να ενημερωθεί
  - Η χρονική πολυπλοκότητα μπορεί να μειωθεί σε $O(N^2 \log(N))$ με κάποιες στρατηγικές υλοποίησης

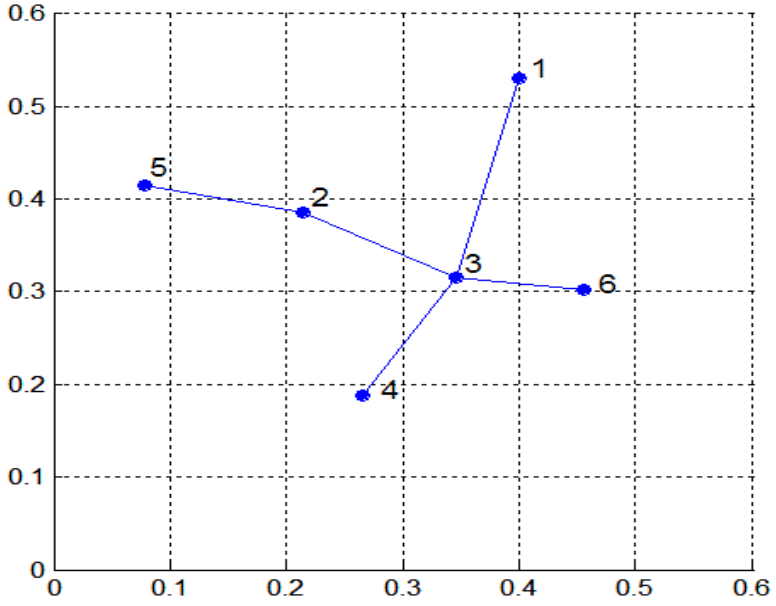
Ιεραρχική Ομαδοποίηση: Προβλήματα και Περιορισμοί

- Μόλις αποφασιστεί η συγχώνευση δύο ομάδων δεν μπορεί να αναστραφεί
- Καμία συνάρτηση στόχου δεν ελαχιστοποιείται άμεσα
- Οι διάφορες μορφές της μεθόδου έχουν προβλήματα με ένα ή περισσότερα από τα παρακάτω:
  - Είναι ευαίσθητη στο θόρυβο και στους outliers
  - Παρουσιάζει δυσκολία στο χειρισμό ομάδων διαφορετικού μεγέθους και κυρτών σχημάτων
  - Διασπά τις μεγάλες ομάδες

MST: Divisive Hierarchical Clustering

- Κατασκευάζεται το MST (Minimum Spanning Tree)
  - Αρχικά το δέντρο περιλαμβάνει ένα οποιοδήποτε σημείο
  - Κατόπιν (και επαναληπτικά μέχρι να τελειώσουν τα σημεία) αναζητείται το πλησιέστερο ζεύγος σημείων (p, q) έτσι ώστε το ένα σημείο (p) να βρίσκεται στο τρέχον δέντρο αλλά το άλλο (q) όχι
  - Προστίθεται το q στο δέντρο με την ακμή μεταξύ p και q





MST: Divisive Hierarchical Clustering

- Χρησιμοποιείται το MST για την κατασκευή της ιεραρχίας των ομάδων:

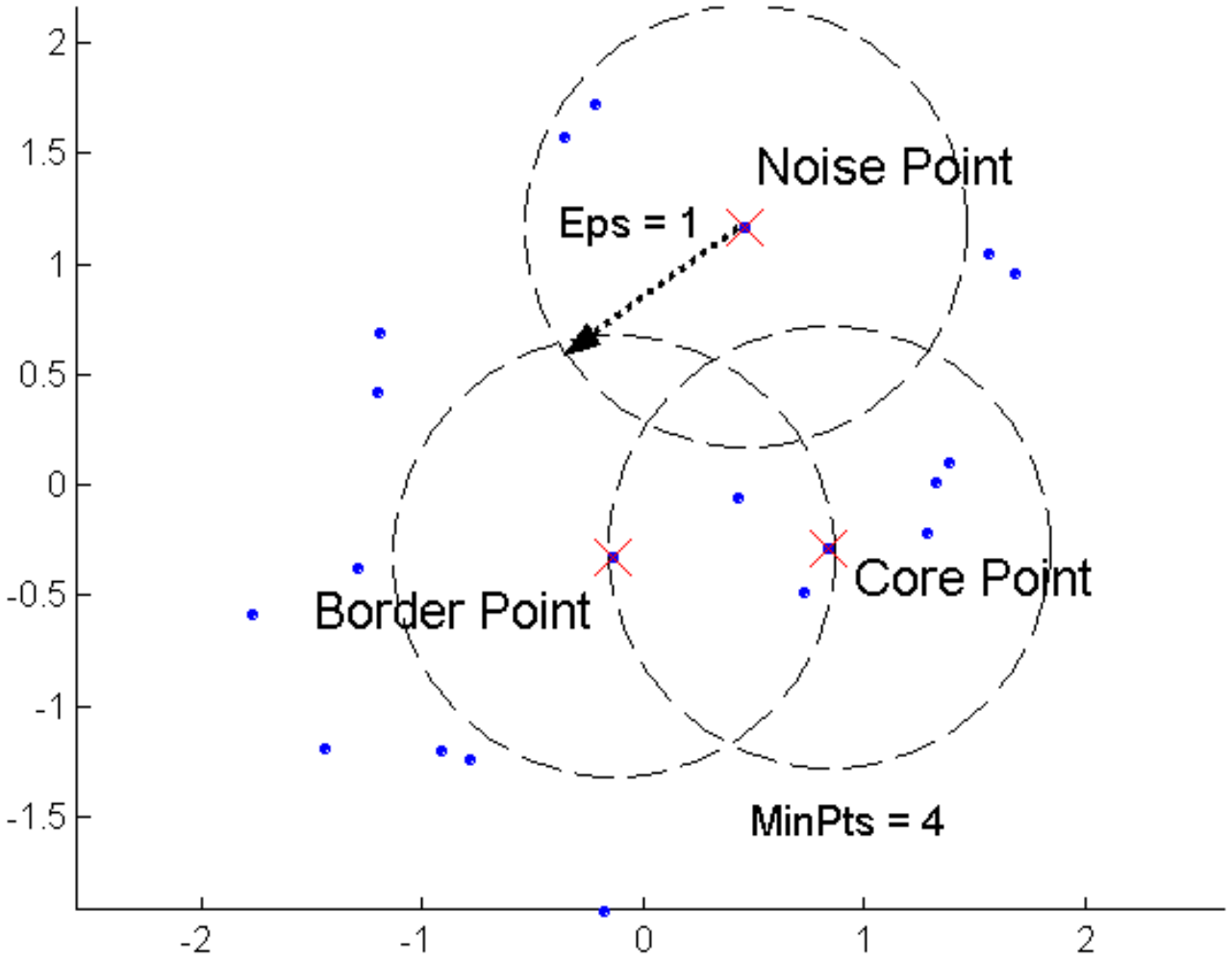
Algorithm 7.5 MST Divisive Hierarchical Clustering Algorithm

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

Αλγόριθμος DBSCAN

- Ο DBSCAN είναι ένας αλγόριθμος που βασίζεται στην πυκνότητα.
  - Πυκνότητα = πλήθος σημείων μέσα σε μία προκαθορισμένη ακτίνα (Eps)
  - Ένα σημείο θα είναι σημείο πυρήνας (core point) αν έχει πάνω από ένα προκαθορισμένο πλήθος σημείων (MinPts) μέσα στην ακτίνα Eps
    - Αυτά είναι σημεία που βρίσκονται στο εσωτερικό μίας ομάδας
  - Ένα συνοριακό σημείο (border point) έχει λιγότερα από MinPts σημεία στην ακτίνα Eps, αλλά βρίσκεται μέσα στην γειτονιά ενός σημείου πυρήνα
  - Ένα σημείο θορύβου (noise point) είναι οποιοδήποτε σημείο δεν είναι ούτε σημείο πυρήνα ούτε συνοριακό σημείο.

DBSCAN: Core, Border, και Noise Points



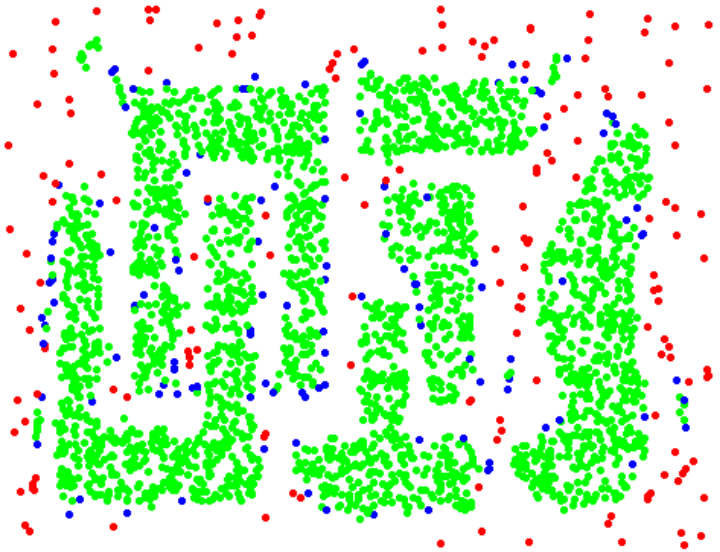
DBSCAN Algorithm

- Εξαλείφονται τα σημεία θορύβου
- Γίνεται ομαδοποίηση στα υπόλοιπα σημεία

current_cluster_label ← 1
for all core points **do**
  if the core point has no cluster label **then**
    current_cluster_label ← *current_cluster_label* + 1
    Label the current core point with cluster label *current_cluster_label*
  end if
  for all points in the *Eps*-neighborhood, except i^{th} the point itself **do**
    if the point does not have a cluster label **then**
      Label the point with cluster label *current_cluster_label*
    end if
  end for
end for

DBSCAN: Core, Border και Noise Points

Αρχικά Σημεία

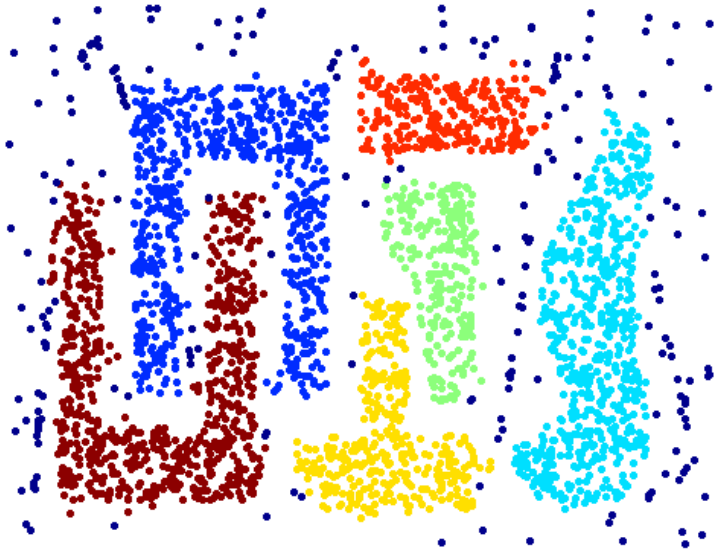


Τύποι σημείων: core, border και noise

Eps = 10, MinPts = 4

Όταν ο DBSCAN λειτουργεί καλά

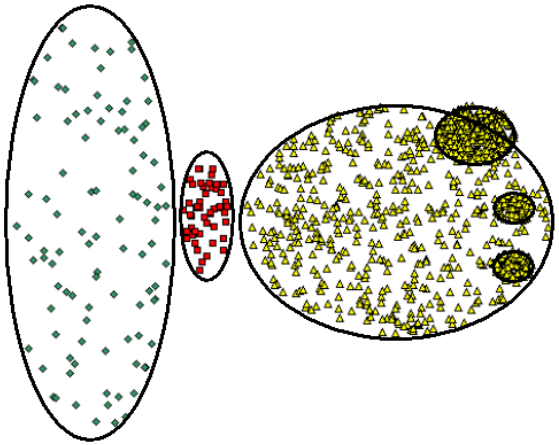
Αρχικά Σημεία



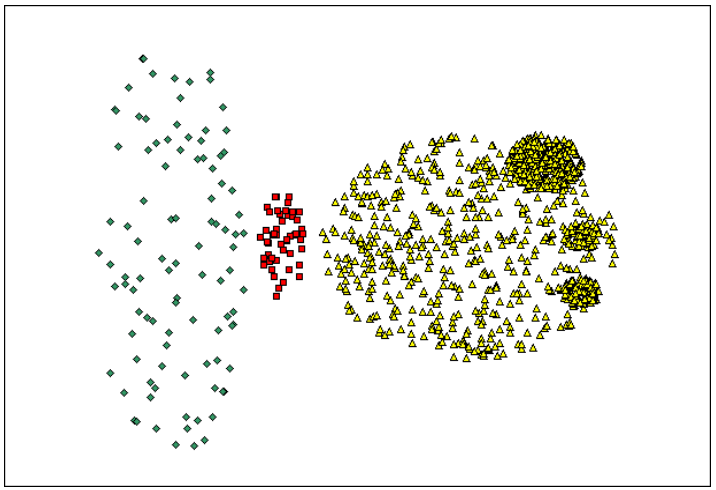
Ομάδες

- Είναι ανθεκτικός στο θόρυβο
- Μπορεί να χειριστεί ομάδες διαφορετικών σχημάτων και μεγεθών

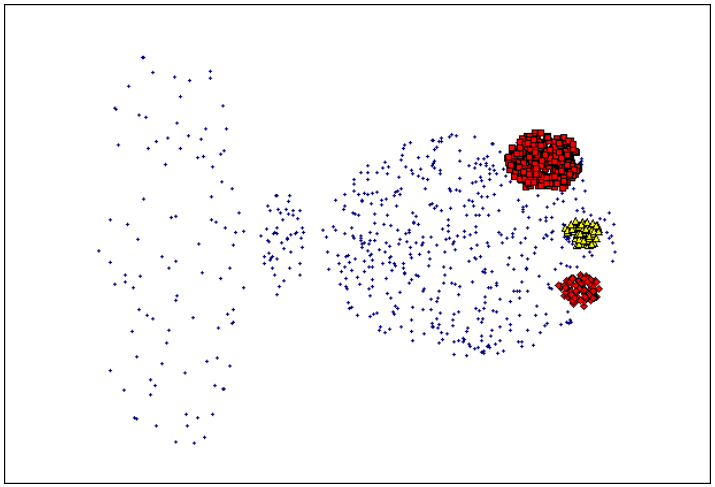
Όταν ο DBSCAN ΔΕΝ λειτουργεί καλά



Αρχικά Σημεία



(MinPts=4, Eps=9.75).

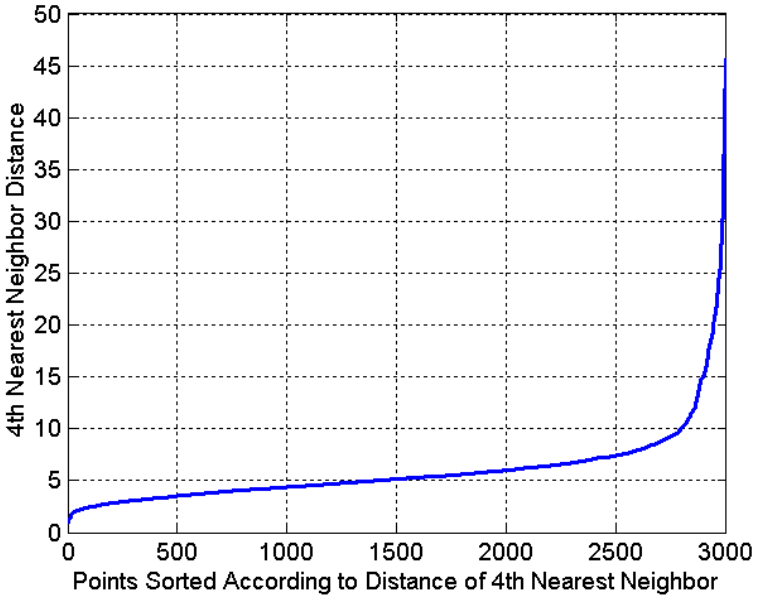


(MinPts=4, Eps=9.92)

- **Πολύ διαφορετικές πυκνότητες**
- **Δεδομένα με πολλές διαστάσεις**

DBSCAN: Προσδιορίζοντας τα EPS και MinPts

- Η ιδέα στηρίζεται στο ότι για τα σημεία που βρίσκονται σε μία ομάδα, οι k–στοι κοντινότεροι γείτονές τους βρίσκονται περίπου στην ίδια απόσταση
- Τα σημεία θορύβου έχουν τους k–στους κοντινότερους γείτονες σε πιο μακρινή απόσταση
- Έτσι απεικονίζονται και ταξινομούνται οι αποστάσεις κάθε σημείου από τους k–στους κοντινότερους γείτονές τους:

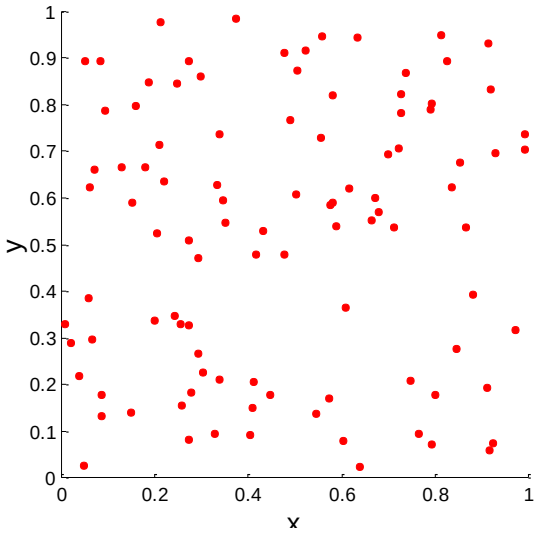


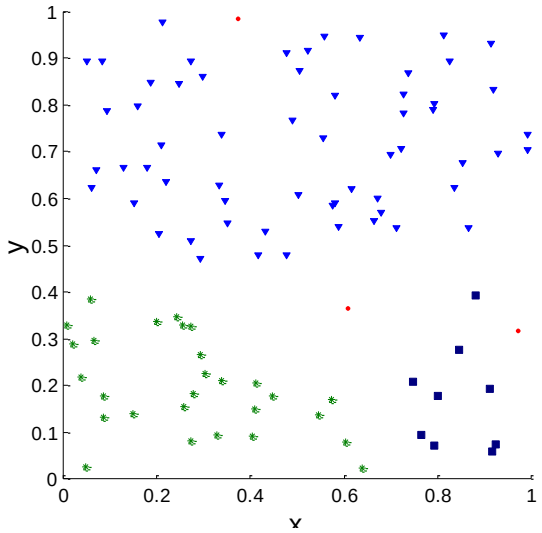
Αξιολόγηση της Ομαδοποίησης

- Για την καθοδηγούμενη ομαδοποίηση έχουμε μία ποικιλία μέτρων για να αξιολογήσουμε πόσο καλό είναι το μοντέλο μας:
  - accuracy, precision, recall
- Για την ανάλυση των ομάδων η ανάλογη ερώτηση είναι πως θα αξιολογήσουμε το πόσο καλή είναι η ομαδοποίηση των ομάδων που παράγονται;
- Αλλά «οι ομάδες είναι στο μάτι του θεατή»!
- Τότε γιατί θέλουμε να τις αξιολογήσουμε;
  - Για να αποφύγουμε την εύρεση προτύπων σε θόρυβο
  - Για να συγκρίνουμε αλγορίθμους ομαδοποίησης
  - Για να συγκρίνουμε δύο σύνολα από ομάδες
  - Για να συγκρίνουμε δύο ομάδες

Ομάδες που βρέθηκαν σε τυχαία δεδομένα

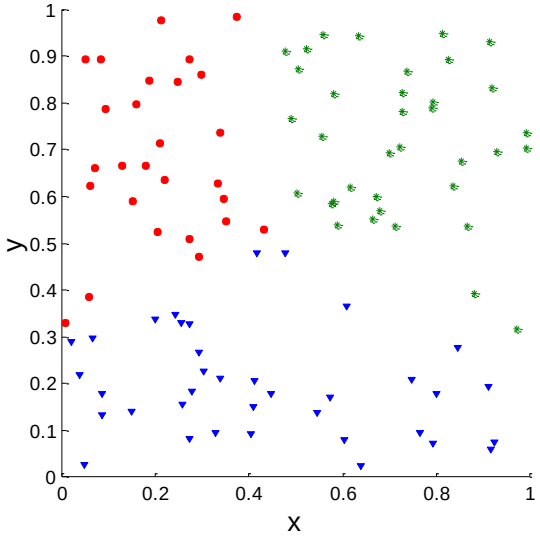
Τυχαία Σημεία

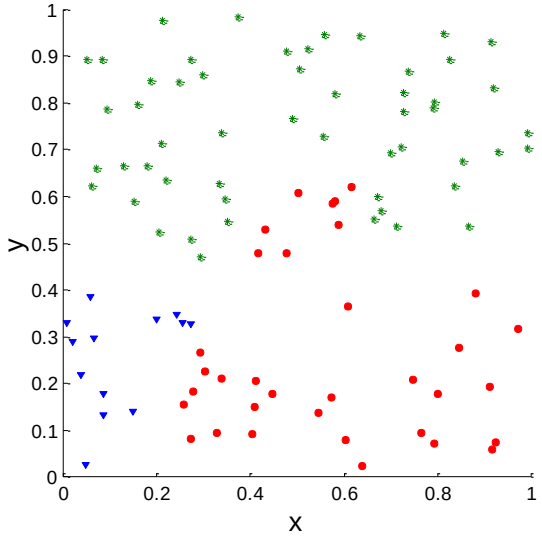




DBSCAN

K-means





Complete Link

Διάφορες πτυχές της αξιολόγησης των ομάδων

1. Προσδιορίζουμε την τάση ομαδοποίησης ενός συνόλου δεδομένων, δηλαδή διακρίνουμε αν υπάρχει πραγματικά στα δεδομένα μη-τυχαία δομή.
2. Συγκρίνουμε τα αποτελέσματα της ανάλυσης των ομάδων με εξωτερικά γνωστά αποτελέσματα, π.χ. με γνωστές τις ετικέτες των ομάδων.
3. Αξιολογούμε πόσο καλά τα αποτελέσματα της ανάλυσης των ομάδων προσαρμόζονται στα δεδομένα *χωρίς* να αναφερθούμε σε εξωτερικές πληροφορίες.
   - Χρησιμοποιούμε μόνο τα δεδομένα
4. Συγκρίνουμε τα αποτελέσματα δύο διαφορετικών συνόλων ανάλυσης ομάδων για να προσδιορίσουμε ποια είναι καλύτερη.
5. Προσδιορίζουμε το «σωστό» πλήθος των ομάδων.

Για τα 2, 3, και 4, μπορούμε επιπλέον να διαχωρίσουμε το αν θέλουμε να αξιολογήσουμε ολόκληρη την ομαδοποίηση ή μεμονωμένες ομάδες.

Μέτρα Αξιολόγησης των Ομάδων

- Τα αριθμητικά μέτρα που εφαρμόζονται για να κρίνουν διάφορες πτυχές της αξιολόγησης των ομάδων, ταξινομούνται στις παρακάτω τρεις κατηγορίες:
  - External Index: Χρησιμοποιείται για τη μέτρηση του βαθμού στον οποίο οι ετικέτες της ομαδοποίησης ταιριάζουν με τις ετικέτες των ομάδων που παρέχονται από εξωτερική πηγή.
    - Entropy, Correlation
  - Internal Index: Χρησιμοποιείται για τη μέτρηση της ποιότητας μίας ομαδοποίησης *χωρίς* αναφορά σε εξωτερικές πληροφορίες.
    - Sum of Squared Error (SSE), Cohesion and Separation
  - Relative Index: Χρησιμοποιείται για τη σύγκριση δύο διαφορετικών ομαδοποιήσεων ή ομάδων.
    - Συχνά external ή internal index χρησιμοποιείται για αυτή τη σύγκριση, π.χ. SSE ή entropy
- Μερικές φορές αναφέρονται ως κριτήρια αντί δείκτες
  - Ωστόσο, μερικές φορές το κριτήριο είναι η γενική στρατηγική ενώ ο δείκτης είναι το αριθμητικό μέτρο με το οποίο εφαρμόζεται το κριτήριο.

Αξιολογώντας την Ομαδοποίηση μέσω Συσχέτισης

- Χρησιμοποιούνται δύο πίνακες:
  - Proximity Matrix (πίνακας εγγύτητας)
  - "Incidence" Matrix (πίνακας γειτνίασης)
    - Έχει μία γραμμή και μία στήλη για κάθε σημείο
    - Μία τιμή του είναι 1 εάν το αντίστοιχο ζεύγος σημείων ανήκει στην ίδια ομάδα
    - Μία τιμή του είναι 0 εάν το αντίστοιχο ζεύγος σημείων ανήκει σε διαφορετικές ομάδες
- Υπολογίζεται η συσχέτιση μεταξύ των δύο πινάκων
  - Από τη στιγμή που οι πίνακες είναι συμμετρικοί, μόνο η συσχέτιση μεταξύ n(n-1) / 2 στοιχείων τους χρειάζεται να υπολογιστεί.
- Η υψηλή συσχέτιση δείχνει ότι τα σημεία ανήκουν στην ίδια ομάδα και είναι κοντά το ένα στο άλλο.
- Δεν είναι καλό μέτρο για κάποιες ομάδες που βασίζονται στην πυκνότητα ή στη συγγένεια.

Συσχέτιση (Correlation)

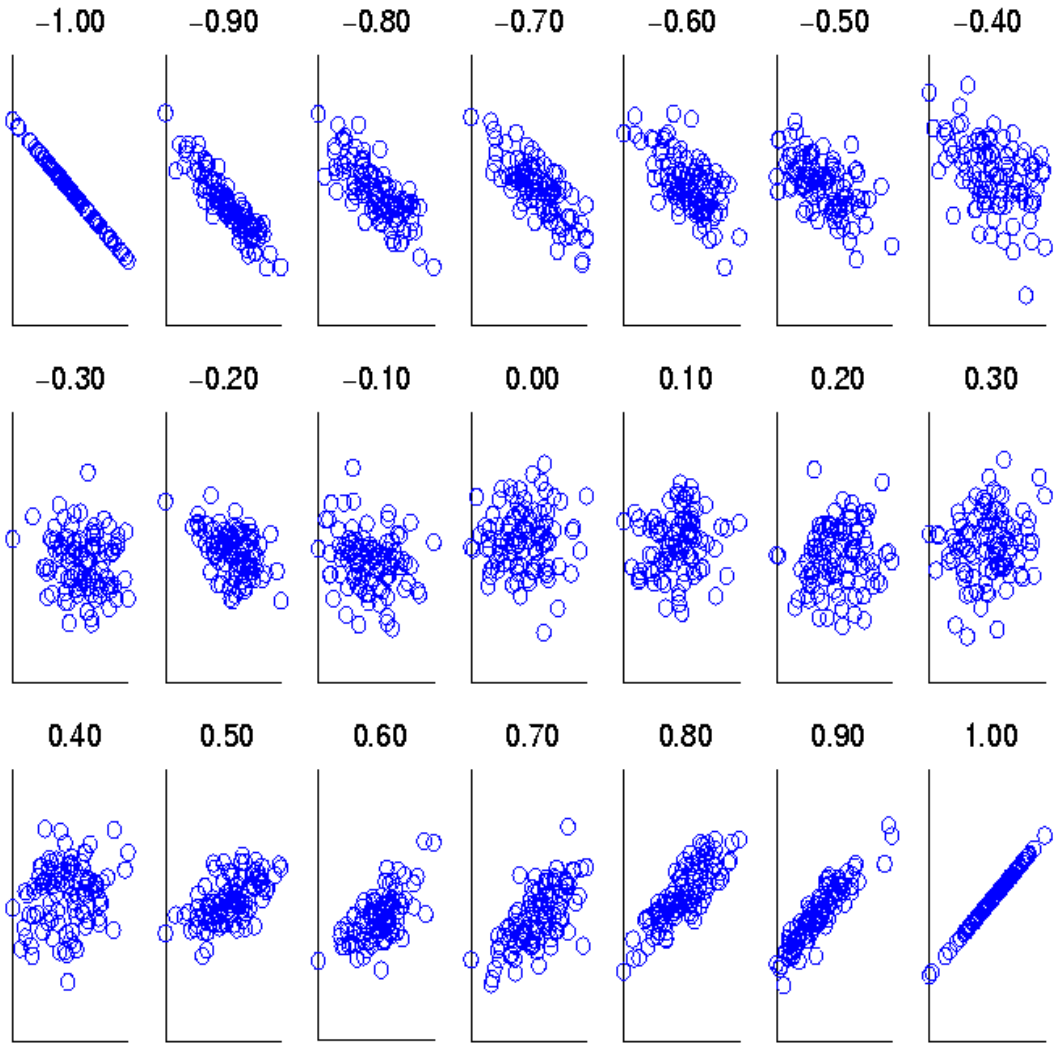
- Η συσχέτιση μετρά την γραμμική σχέση μεταξύ αντικειμένων
- Για να υπολογίσουμε τη συσχέτιση, κανονικοποιούμε τα δεδομένα, p και q, και κατόπιν παίρνουμε το εσωτερικό τους γινόμενο:

$$p'_k = (p_k - mean(p)) / std(p)$$

$$q'_k = (q_k - mean(q)) / std(q)$$

$$correlation(p, q) = p' \bullet q'$$

Οπτική Αξιολόγηση της Συσχέτισης



Τα διαγράμματα διασποράς απεικονίζουν την ομοιότητα από –1 ως 1.

Άσκηση

24. Given the set of cluster labels and similarity matrix shown in Tables 8.4 and 8.5, respectively, compute the correlation between the similarity matrix and the ideal similarity matrix, i.e., the matrix whose ij^{th} entry is 1 if two objects belong to the same cluster, and 0 otherwise.

Table 8.4. Table of cluster labels for Exercise 24.

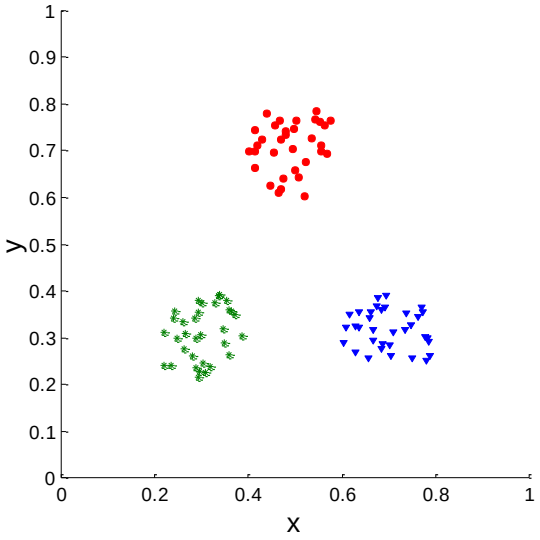
Point	Cluster Label
P1	1
P2	1
P3	2
P4	2

Table 8.5. Similarity matrix for Exercise 24.

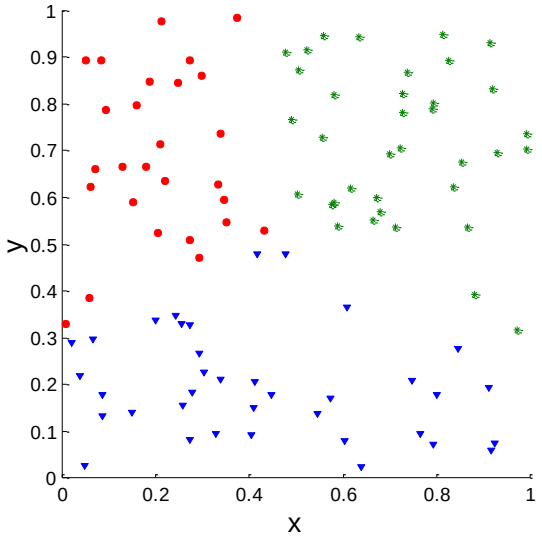
Point	P1	P2	P3	P4
P1	1	0.8	0.65	0.55
P2	0.8	1	0.7	0.6
P3	0.65	0.7	1	0.9
P4	0.55	0.6	0.9	1

Αξιολογώντας την ομαδοποίηση με τη συσχέτιση

- Η συσχέτιση των πινάκων εγγύτητας και γειτνίασης για τις ομαδοποιήσεις του K-means στα ακόλουθα παραδείγματα είναι:



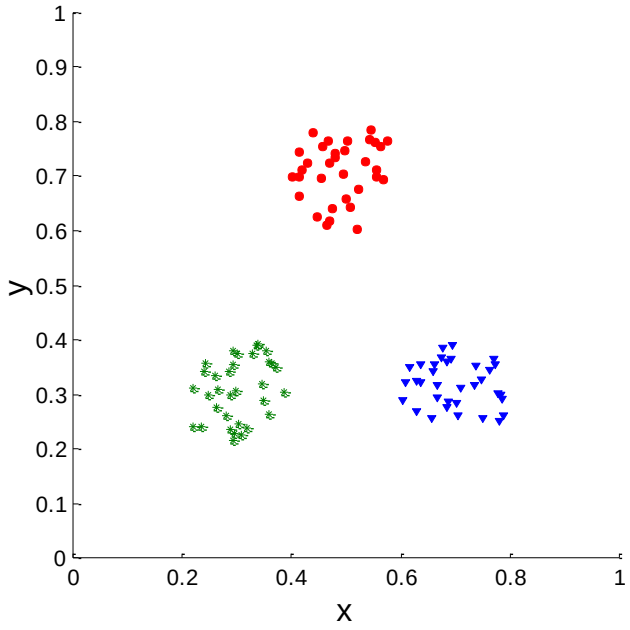
Corr = -0.9235

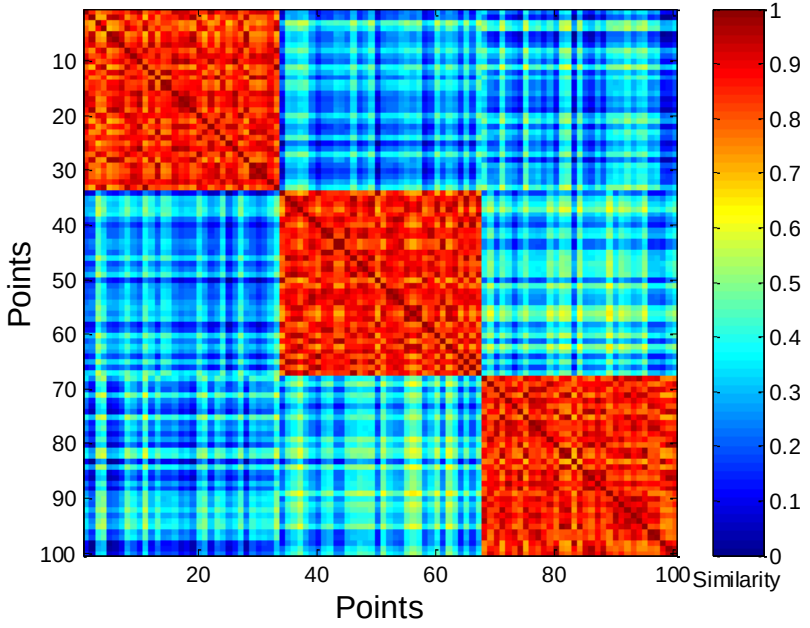


Corr = -0.5810

Χρησιμοποιώντας τον πίνακα ομοιότητας για την αξιολόγηση των ομάδων

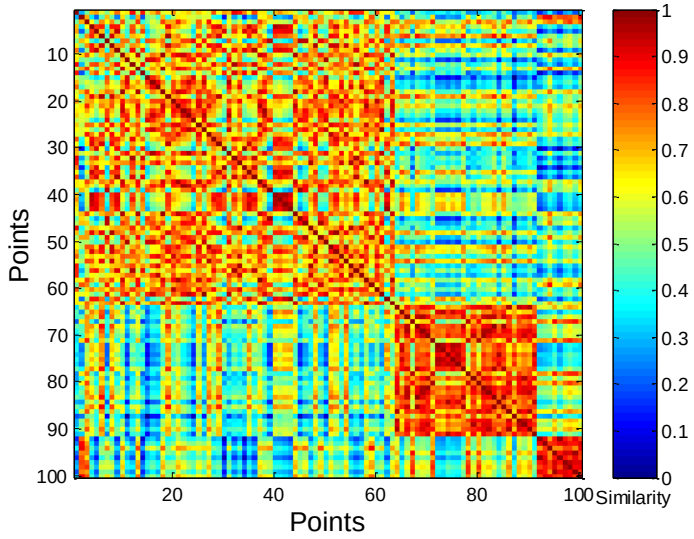
- Ταξινομούμε τον πίνακα ομοιότητας ως προς τις ετικέτες των ομάδων και τον μελετάμε οπτικά.

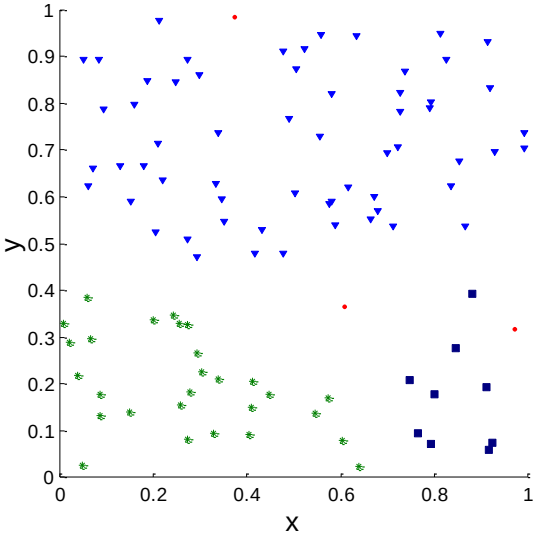




Χρησιμοποιώντας τον πίνακα ομοιότητας για την αξιολόγηση των ομάδων

- Οι ομάδες από τυχαία δεδομένα δεν είναι και τόσο ευκρινείς

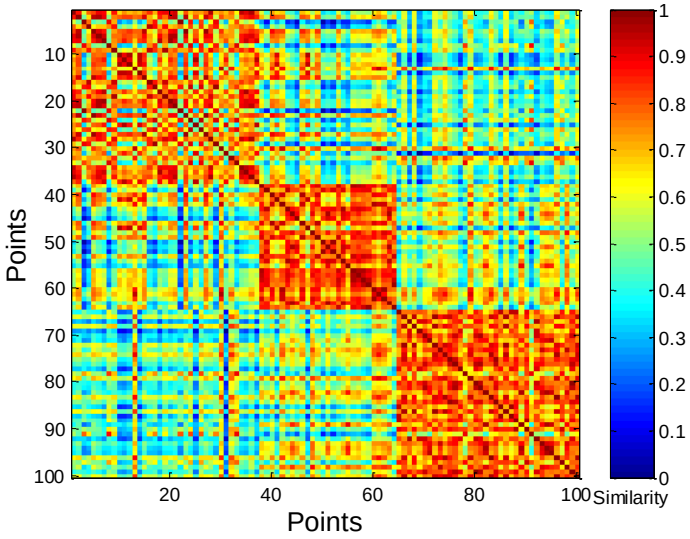


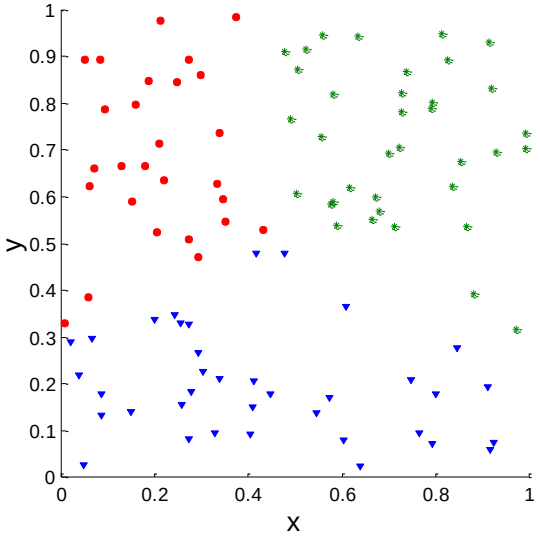


DBSCAN

Χρησιμοποιώντας τον πίνακα ομοιότητας για την αξιολόγηση των ομάδων

- Οι ομάδες από τυχαία δεδομένα δεν είναι και τόσο ευκρινείς

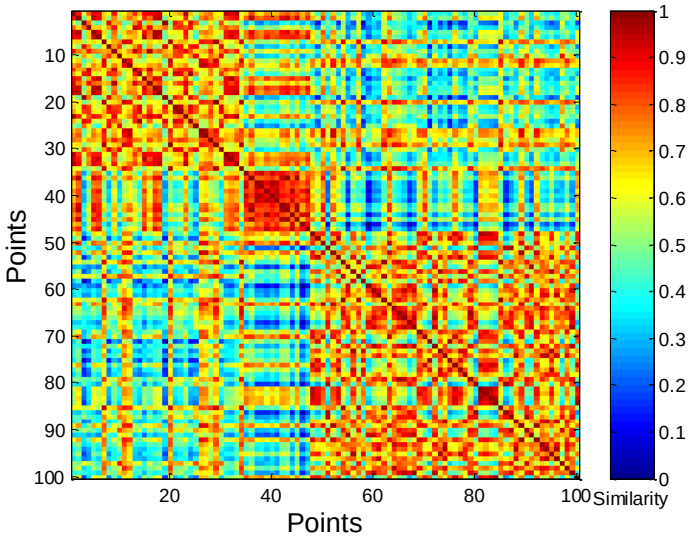


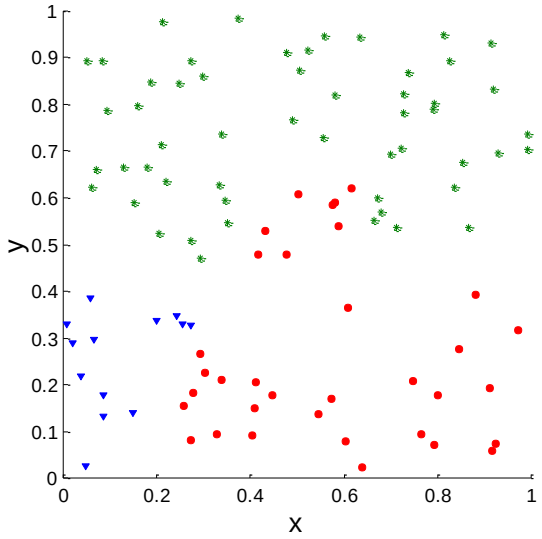


K-means

Χρησιμοποιώντας τον πίνακα ομοιότητας για την αξιολόγηση των ομάδων

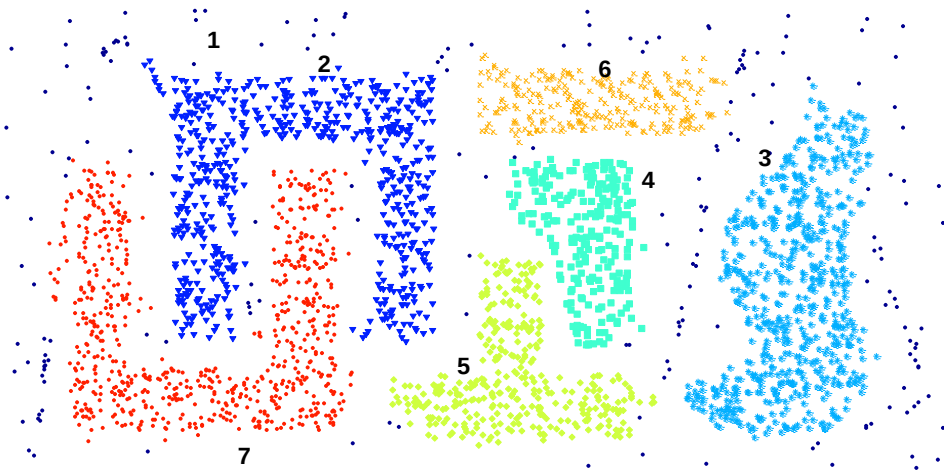
- Οι ομάδες από τυχαία δεδομένα δεν είναι και τόσο ευκρινείς

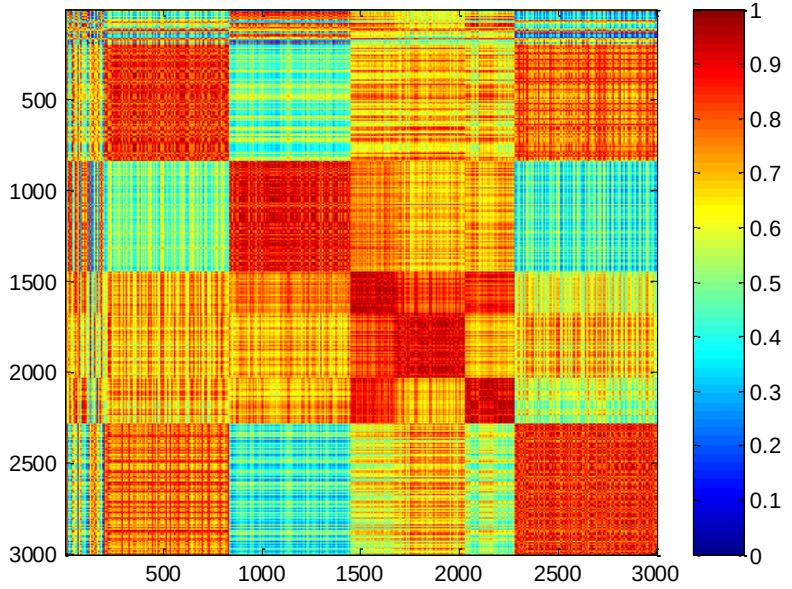




Complete Link

Χρησιμοποιώντας τον πίνακα ομοιότητας για την αξιολόγηση των ομάδων

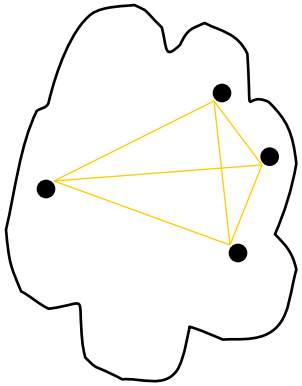




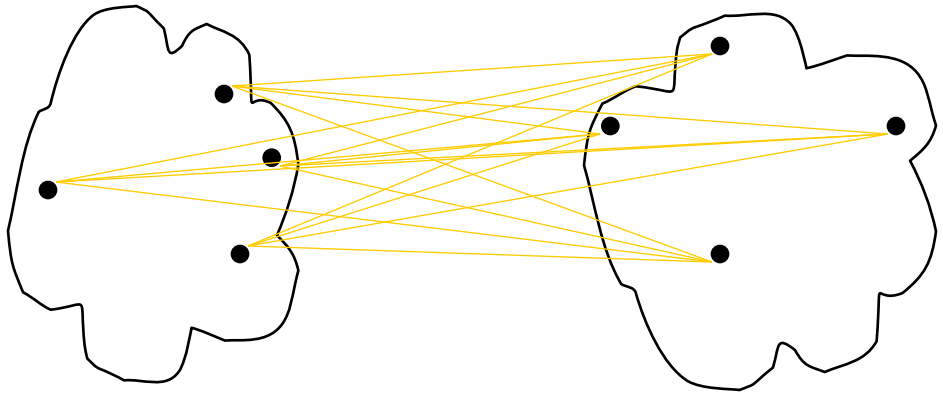
DBSCAN

Internal Measures: Cohesion και Separation

- Μία ομαδοποίηση βασισμένη σε γράφους μπορεί να αξιολογηθεί από τα μέτρα συνοχής (cohesion) και διαχωρισμού (separation).
  - Το μέτρο συνοχής μίας ομάδας (cluster cohesion) είναι το άθροισμα των βαρών όλων των συνδέσμων (ακμών) μέσα στην ομάδα.
  - Το μέτρο διαχωρισμού μίας ομάδας (cluster separation) είναι το άθροισμα όλων των βαρών των συνδέσμων μεταξύ κόμβων της ομάδας με κόμβους έξω από την ομάδα.



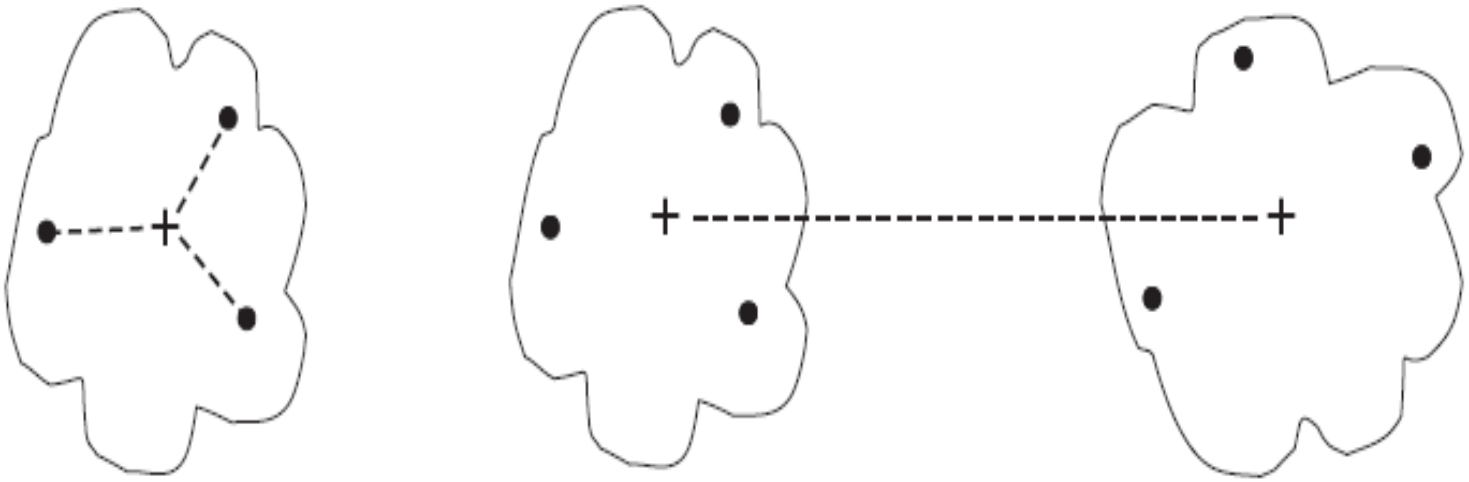
cohesion



separation

Cohesion και Separation (Central-based clusters)

Μία ομαδοποίηση που βασίζεται σε κέντρα μπορεί να αξιολογηθεί από τα μέτρα cohesion και separation.



(a) Cohesion.

(b) Separation.

Cohesion και Separation (Central-based clustering)

- Cluster Cohesion: Μετρά πόσο στενά συσχετίζονται τα αντικείμενα μίας ομάδας
  - Η συνοχή μετριέται από το εσωτερικό άθροισμα τετραγώνων των αποστάσεων εντός της ομάδας (SSE):

$$WSS = \sum_{i} \sum_{x \in C_i} (x - m_i)^2$$

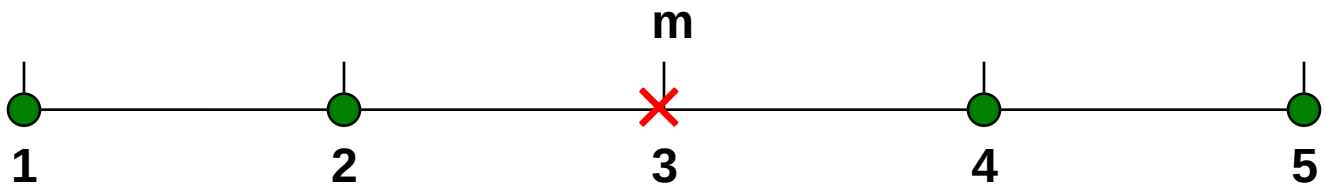
- Cluster Separation: Μετρά πόσο διακεκριμένη ή πόσο καλά διαχωρισμένη είναι μία ομάδα σε σχέση με τις υπόλοιπες ομάδες
  - Ο διαχωρισμός μετριέται από το άθροισμα τετραγώνων των αποστάσεων μεταξύ των ομάδων:

$$BSS = \sum_{i} |C_i| (m - m_i)^2$$

    - όπου $|C_i|$ είναι το μέγεθος της ομάδας i

Cohesion και Separation (Παράδειγμα)

- Παράδειγμα: WSS + BSS = Total SSE (σταθερό)

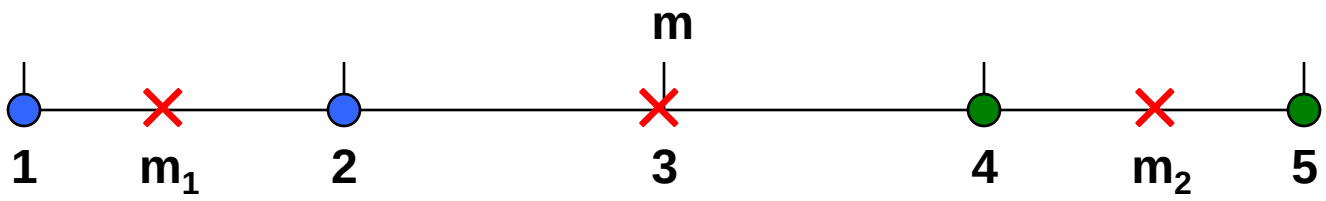


K=1 cluster:

$$WSS = (1-3)^2 + (2-3)^2 + (4-3)^2 + (5-3)^2 = 10$$

$$BSS = 4 \times (3-3)^2 = 0$$

$$Total = 10 + 0 = 10$$



K=2 clusters:

$$WSS = (1-1.5)^2 + (2-1.5)^2 + (4-4.5)^2 + (5-4.5)^2 = 1$$

$$BSS = 2 \times (3-1.5)^2 + 2 \times (3-4.5)^2 = 9$$

$$Total = 1 + 9 = 10$$

Internal Measures: Silhouette Coefficient

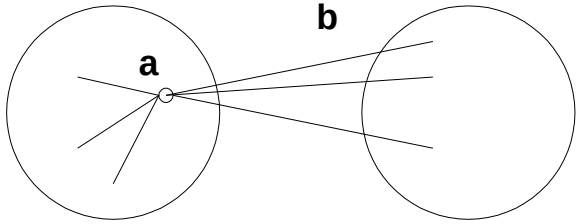
- Ο συντελεστής Silhouette συνδυάζει τις ιδέες των cohesion και separation, αλλά για ανεξάρτητα σημεία.

- Για ένα ανεξάρτητο σημείο, *i*
  - Υπολόγισε το ***a*** = μέση απόσταση του *i* από τα σημεία της ομάδας του
  - Υπολόγισε το ***b*** = min (μέση απόσταση του *i* από σημεία άλλης ομάδας)
  - Ο συντελεστής silhouette για το σημείο τότε ορίζεται από την σχέση:

    $s = 1 - a/b$ όταν $a < b$,

    ή $s = b/a - 1$ όταν $a \geq b$ (που δεν είναι η συνήθης περίπτωση)

  - Τυπικές τιμές από 0 έως 1.
  - Όσο πιο κοντά στο 1 είναι τόσο το καλύτερο



- Μπορεί να υπολογίσει το μέσο πλάτος της μορφής μίας ομάδας ή μίας ομαδοποίησης

Silhouette Coefficient (Παράδειγμα)

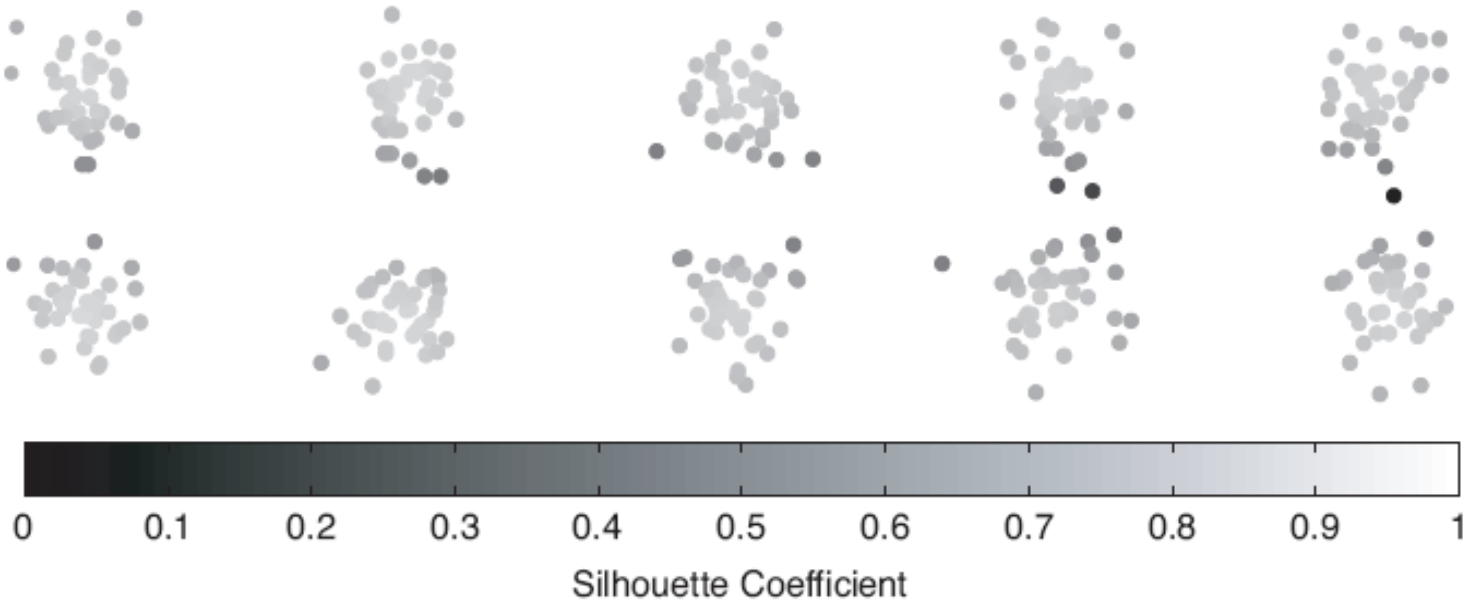
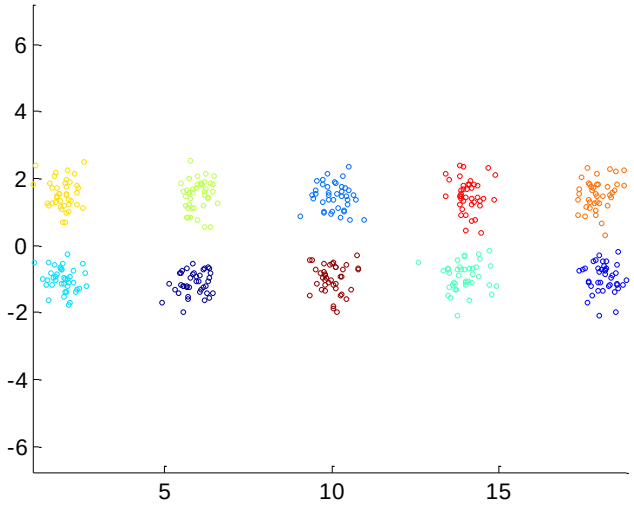
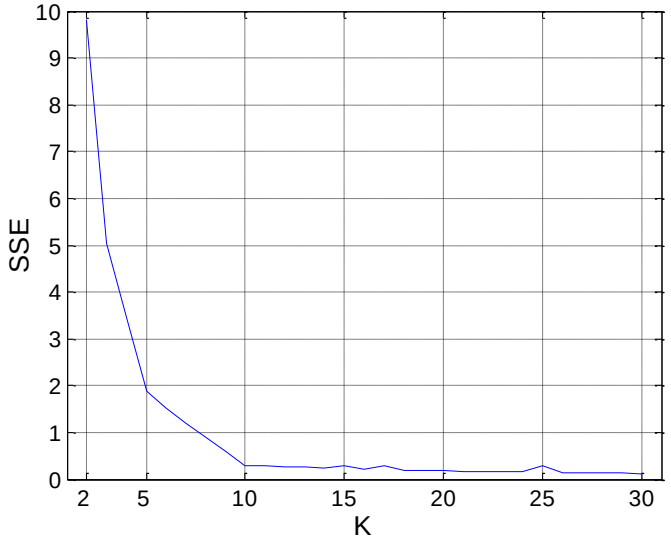


Figure 8.29. Silhouette coefficients for points in ten clusters.

Internal Measures: SSE

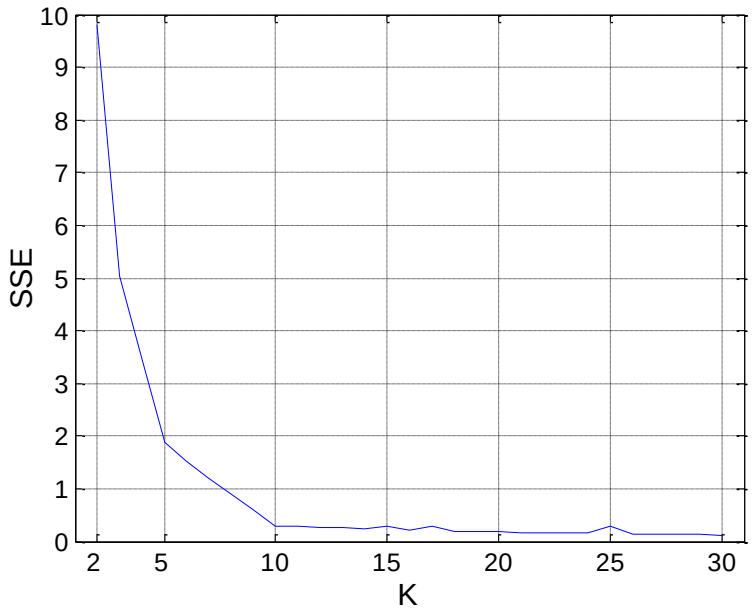
- Οι ομάδες με πιο περίπλοκες μορφές δεν αναγνωρίζονται καλά
- Internal Index: Χρησιμοποιείται για τη μέτρηση της ποιότητας της δομής της ομάδας χωρίς να ληφθεί υπόψη καμία εξωτερική πληροφορία, π.χ. SSE
- Το μέτρο SSE είναι καλό για τη σύγκριση δύο ομαδοποιήσεων ή δύο ομάδων (μέσο SSE)
- Μπορεί επίσης να χρησιμοποιηθεί και για την εκτίμηση του πλήθους των ομάδων

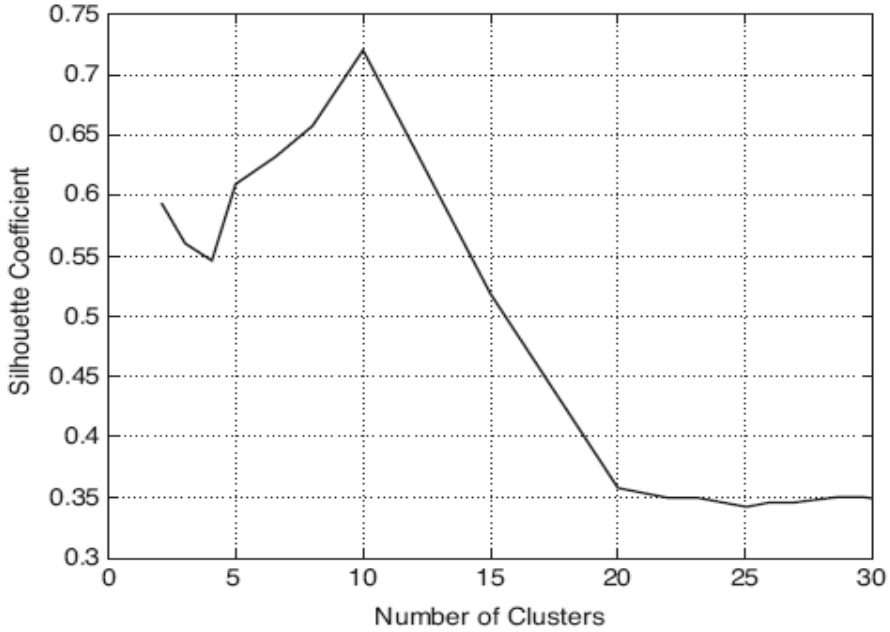




Internal Measures: SSE και Silhouette

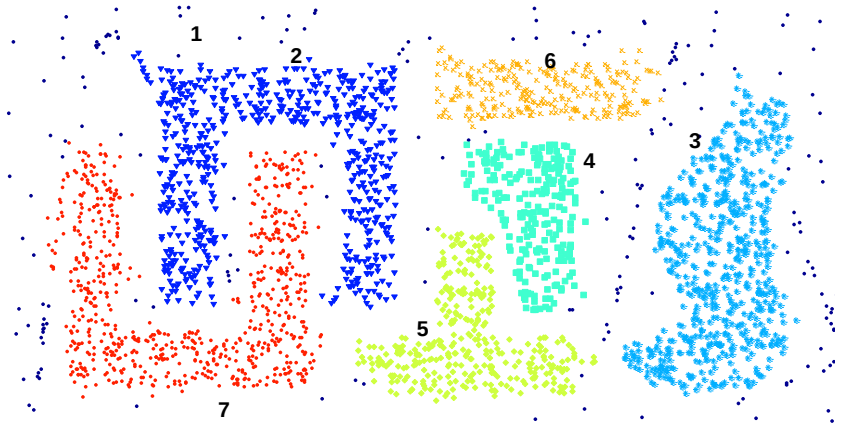
- Τα μέτρα SSE και Average Silhouette Coefficient μπορούν να εκτιμήσουν το πλήθος των ομάδων

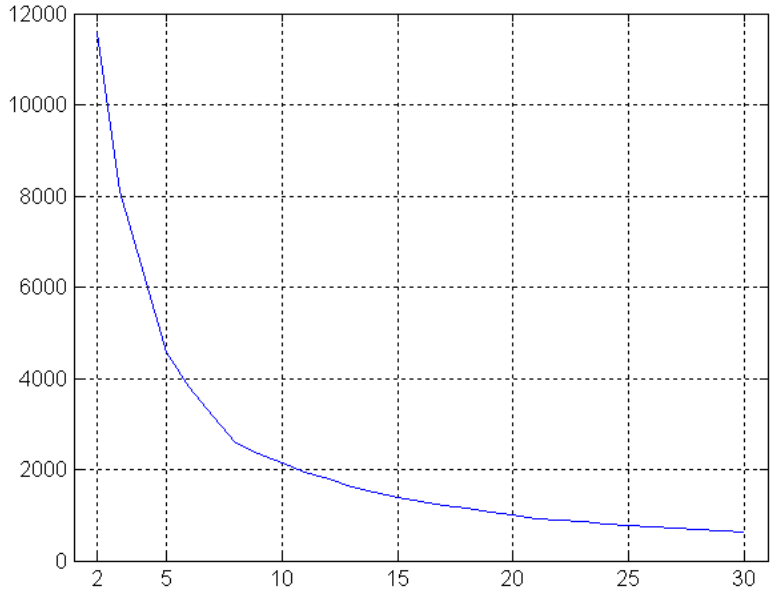




Internal Measures: SSE

- Η καμπύλη του SSE βοηθά σε πιο περίπλοκα σύνολα δεδομένων





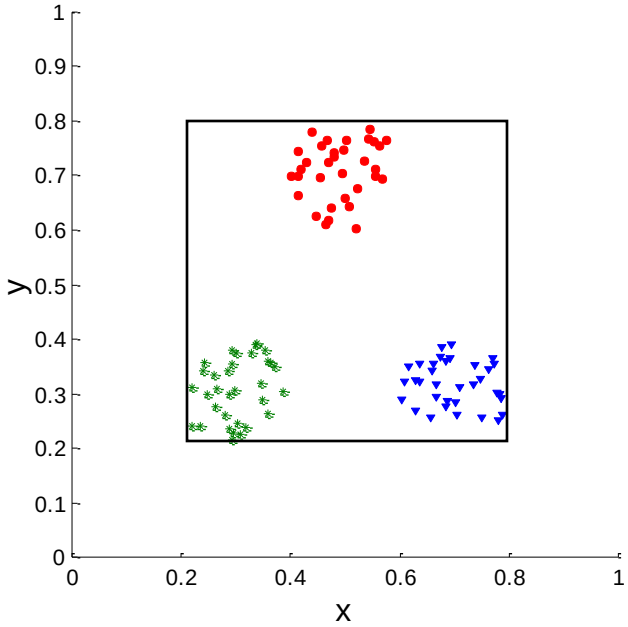
Το SSE των ομάδων που βρέθηκαν με τη χρήση του K-means

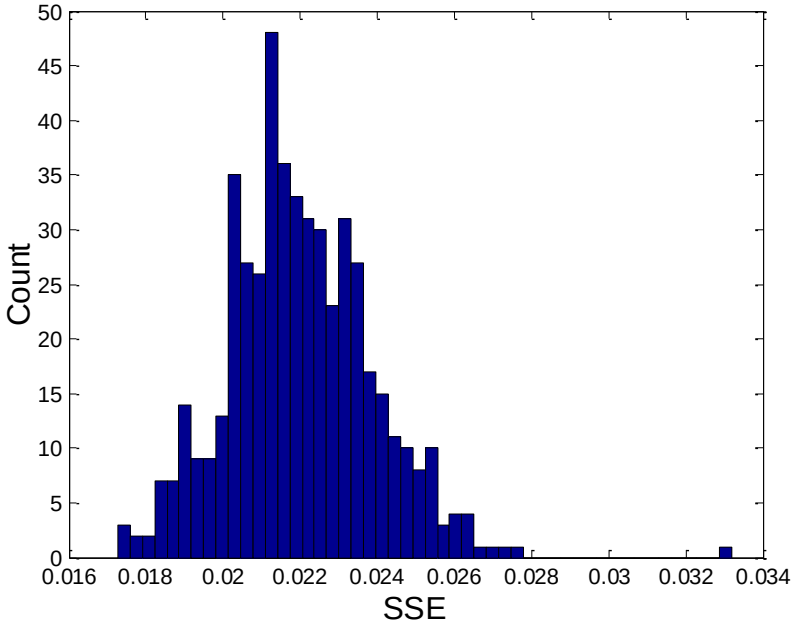
Πλαίσιο Αξιολόγησης Ομάδων

- Χρειάζεται ένα πλαίσιο για την ερμηνεία του κάθε μέτρου
  - Για παράδειγμα, αν το μέτρο αξιολόγησης πάρει την τιμή 10, τότε αυτό είναι καλό, κακό ή ουδέτερο;
- Η στατιστική παρέχει ένα πλαίσιο αξιολόγησης ομάδων
  - Όσο πιο «άτυπο» είναι το αποτέλεσμα μίας ομαδοποίησης τόσο πιο πιθανό είναι να αντιπροσωπεύει μία έγκυρη δομή στα δεδομένα.
  - Μπορεί να συγκρίνει τις τιμές ενός μέτρου όπως προκύπτουν από τυχαία δεδομένα ή τυχαίες ομαδοποιήσεις με αυτές ενός αποτελέσματος ομαδοποίησης που μας ενδιαφέρει.
    - Αν οι τιμές του μέτρου είναι μη-τυπικές, τότε τα αποτελέσματα της ομαδοποίησης είναι έγκυρα.
  - Αυτές οι προσεγγίσεις είναι πιο περίπλοκες και πιο δύσκολες να κατανοηθούν.
- Για τη σύγκριση των αποτελεσμάτων δύο διαφορετικών συνόλων ανάλυσης ομάδων, ένα πλαίσιο δεν είναι τόσο απαραίτητο.
  - Ωστόσο, υπάρχει το ερώτημα εάν η διαφορά μεταξύ δύο τιμών ενός μέτρου είναι σημαντική.

Στατιστικό Πλαίσιο για το SSE

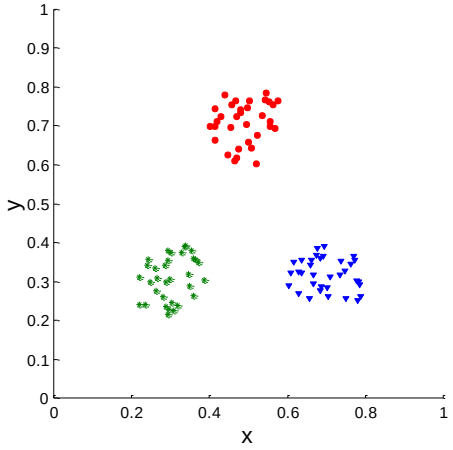
- Παράδειγμα
  - Να συγκριθεί το SSE με τιμή 0.005 έναντι ομαδοποιήσεων με τρεις ομάδες σε τυχαία δεδομένα
  - Το ιστόγραμμα απεικονίζει ότι το SSE από τρεις ομάδες σε 500 σύνολα τυχαίων σημείων μεγέθους 100 κατανέμονται στο διάστημα 0.2 – 0.8 για τις τιμές του x και του y

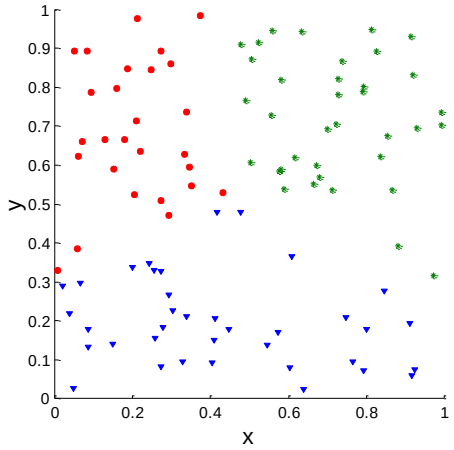


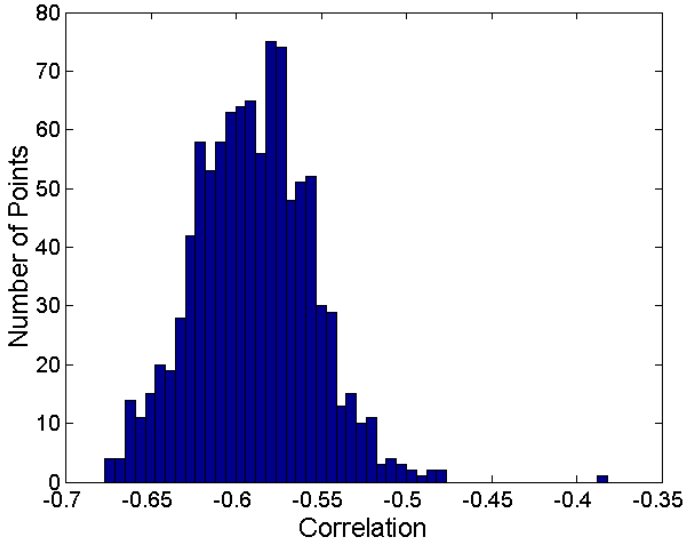


Στατιστικό Πλαίσιο για τη Συσχέτιση

- Η συσχέτιση των πινάκων γειτνίασης και εγγύτητας για τις ομαδοποιήσεις του K-means των παρακάτω δύο συνόλων δεδομένων είναι:







Corr = -0.9235

Corr = -0.5810

Εξωτερικά Μέτρα Αξιολόγησης Ομάδων: Entropy και Purity

Table 5.9. K-means Clustering Results for LA Document Data Set

Cluster	Entertainment	Financial	Foreign	Metro	National	Sports	Entropy	Purity
1	3	5	40	506	96	27	1.2270	0.7474
2	4	7	280	29	39	2	1.1472	0.7756
3	1	1	1	7	4	671	0.1813	0.9796
4	10	162	3	119	73	2	1.7487	0.4390
5	331	22	5	70	13	23	1.3976	0.7134
6	5	358	12	212	48	13	1.5523	0.5525
Total	354	555	341	943	273	738	1.1450	0.7203

entropy For each cluster, the class distribution of the data is calculated first, i.e., for cluster j we compute p_{ij}, the 'probability' that a member of cluster j belongs to class i as follows: $p_{ij} = m_{ij}/m_j$, where m_j is the number of values in cluster j and m_{ij} is the number of values of class i in cluster j. Then using this class distribution, the entropy of each cluster j is calculated using the standard formula $e_j = \sum_{i=1}^{L} p_{ij} \log_2 p_{ij}$, where the L is the number of classes. The total entropy for a set of clusters is calculated as the sum of the entropies of each cluster weighted by the size of each cluster, i.e., $e = \sum_{i=1}^{K} \frac{m_i}{m} e_j$, where m_j is the size of cluster j, K is the number of clusters, and m is the total number of data points.

purity Using the terminology derived for entropy, the purity of cluster j, is given by $purity_j = \max p_{ij}$ and the overall purity of a clustering by $purity = \sum_{i=1}^{K} \frac{m_i}{m} purity_j$.

Τελικά Σχόλια για την Αξιολόγηση Ομάδων

"Η αξιολόγηση των δομών των ομάδων είναι το πιο δύσκολο και το πιο ενοχλητικό κομμάτι της ανάλυσης ομάδων.

Χωρίς επίπονη προσπάθεια σε αυτή την κατεύθυνση, η ανάλυση ομάδων θα παραμένει μία κρυφή τέχνη προσιτή μόνο στους αληθινούς πιστούς που έχουν εμπειρία και μεγάλο κουράγιο."

Algorithms for Clustering Data, Jain and Dubes