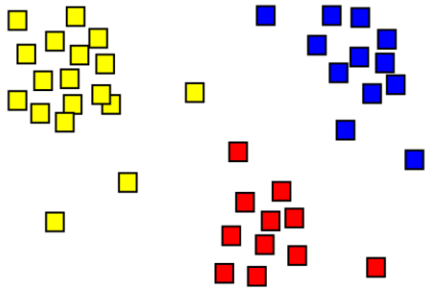
ΣΥΣΤΑΔΟΠΟΙΗΣΗ (CLUSTERING)

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

UNIVERSITY OF PIRAEUS

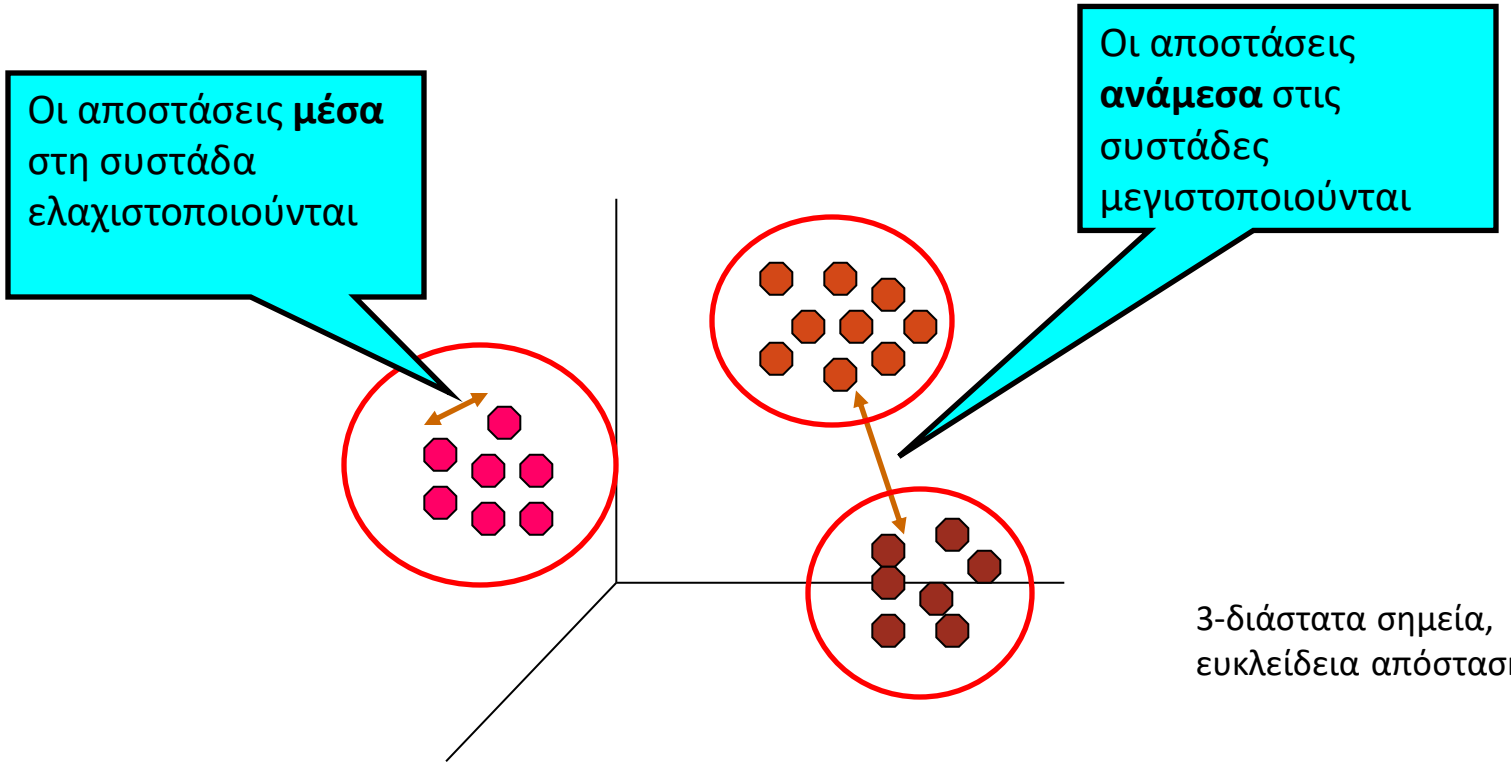
Γεράσιμος Ραζής (razis@uth.gr)



Συστάδες (Clusters)

Τι είναι η Συσταδοποίηση;

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



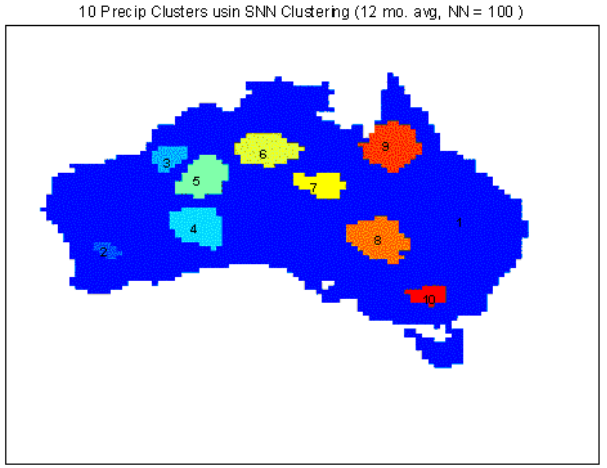
3-διάστατα σημεία, ευκλείδεια απόσταση

Εφαρμογές

Ομαδοποίηση:

- Γονιδίων και πρωτεϊνών που έχουν την ίδια λειτουργία
- Εικόνων
- Χαρακτηριστικά ασθενειών
- Μετοχών με παρόμοια διακύμανση τιμών
- Weblog για εύρεση παρόμοιων προτύπων προσπέλασης
- Σχετιζόμενων αρχείων για browsing
- Κειμένων
- Πελατών με παρόμοια συμπεριφορά

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

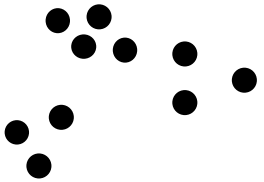


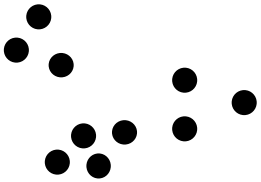
Εφαρμογές

- **Κατανόηση / Stand-alone εφαρμογή/εργαλείο**
  - Οπτικοποίηση
  - Συμπεράσματα για την κατανομή

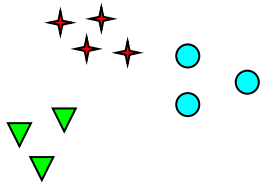
- **Βήμα Προεπεξεργασίας**
  - Περίληψη: Ελάττωση του μεγέθους μεγάλων συνόλων χρήσης αντιπροσωπευτικών σημείων από κάθε συστάδα - πρωτότυπα (prototypes)
  - Συμπίεση
  - Αποδοτική κατασκευή ευρετηρίων, εύρεση κοντινότερου γείτονα, κλπ

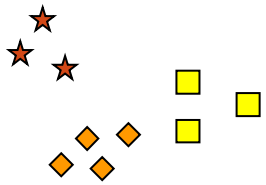
Ασάφεια



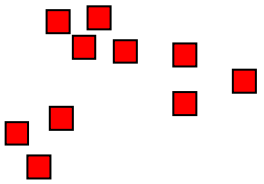


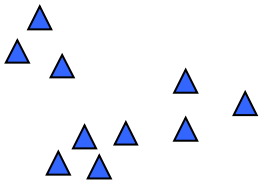
Πόσες Ομάδες?



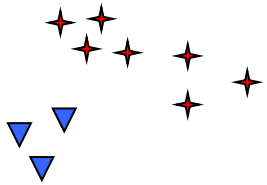


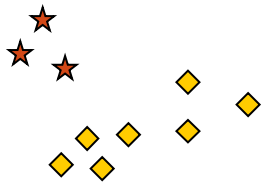
6 ομάδες





2 ομάδες





4 ομάδες

Πότε μια Συσταδοποίηση είναι καλή;

Μια μέθοδος συσταδοποίησης είναι καλή αν παράγει συστάδες καλής ποιότητας:

- μεγάλη ομοιότητα εντός της συστάδας, και
- μικρή ομοιότητα ανάμεσα στις συστάδες

Η ποιότητα εξαρτάται από τη:

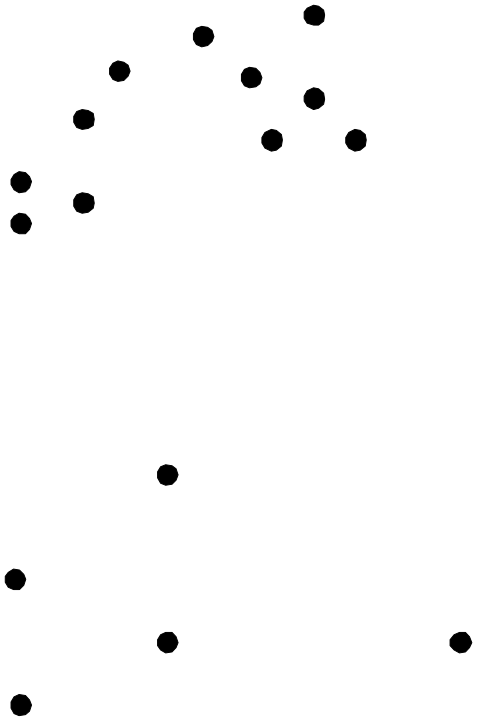
- μέτρηση ομοιότητας, και
- μέθοδο υλοποίησης της συσταδοποίησης

Είδη Συσταδοποίησης

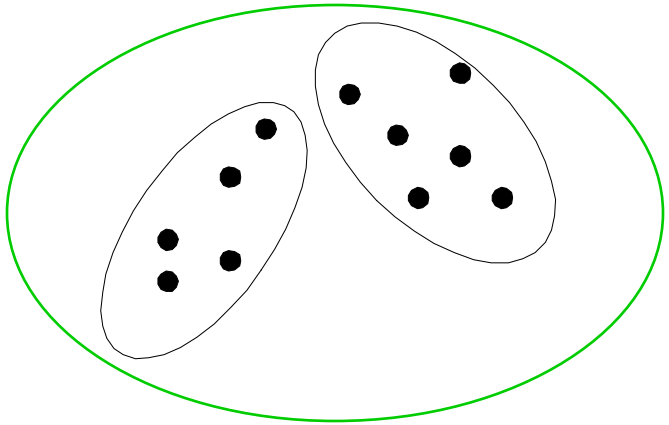
- Μια συσταδοποίηση είναι ένα σύνολο από συστάδες
- Βασική διάκριση ανάμεσα στο *ιεραρχικό* (*hierarchical*) και *διαχωριστικό* (*partitional*) σύνολο από ομάδες

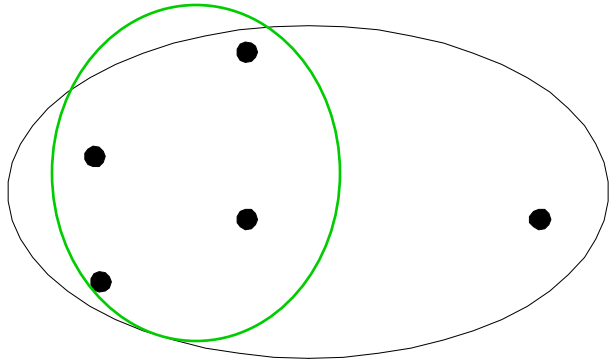
- Ιεραρχική Συσταδοποίηση (Hierarchical clustering)
  - Ένα σύνολο από *εμφωλευμένες* (nested) ομάδες
  - Επιτρέπουμε σε μια συστάδα να έχει υποσυστάδες οργανωμένες σε ένα ιεραρχικό δέντρο
- Διαχωριστική Συσταδοποίηση (Partitional Clustering)
  - Ένας διαμερισμός των αντικειμένων σε μη επικαλυπτόμενα (non-overlapping) υποσύνολα ούτως ώστε κάθε αντικείμενο να ανήκει σε ακριβώς ένα υποσύνολο

Διαχωριστική και Ιεραρχική Συσταδοποίηση



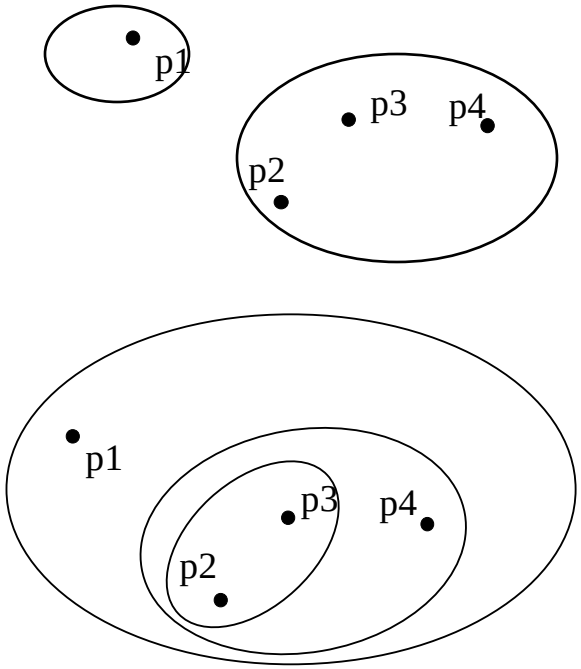
Αρχικά Σημεία



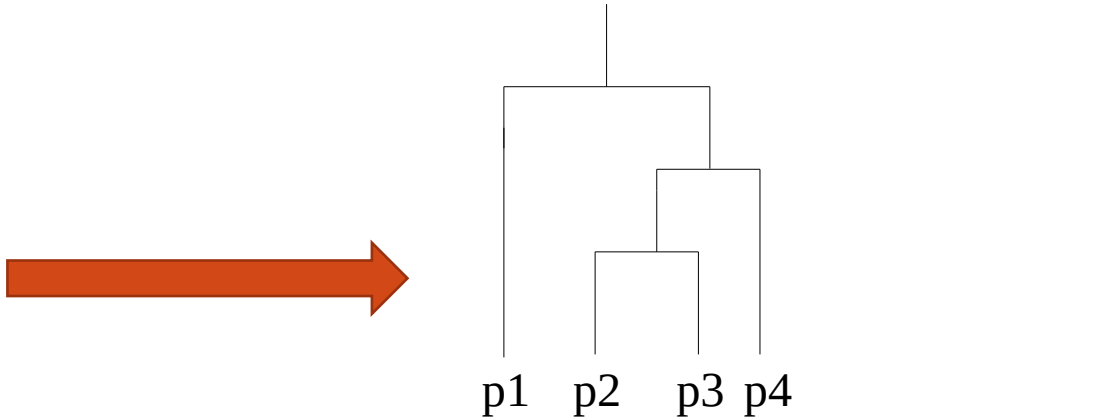


Διαχωριστική και Ιεραρχική Συσταδοποίηση

Διαχωριστική Συσταδοποίηση



Ιεραρχική Συσταδοποίηση



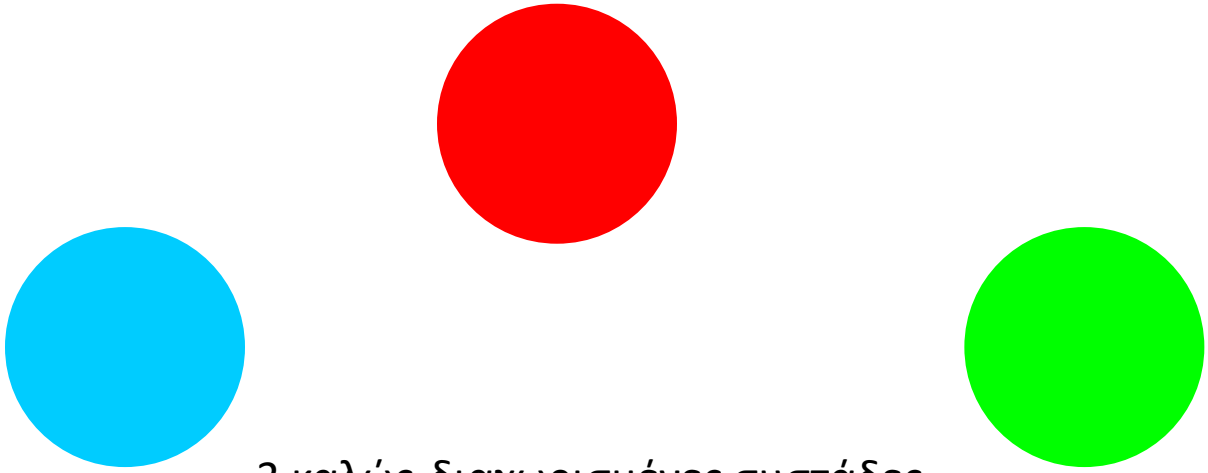
Παραδοσιακό Δενδρόγραμμα (Dendrogram)

- Φύλλα: απλά σημεία ή απλές συστάδες
- Μπορούμε να «κόψουμε» το δέντρο

Τύποι Συστάδων: Καλώς Διαχωρισμένες

Μια *συστάδα* είναι ένα σύνολο από σημεία τέτοια ώστε:

- κάθε σημείο μιας συστάδας είναι κοντινότερο σε (ή πιο όμοιο με) όλα τα άλλα σημεία της συστάδας από ότι σε οποιοδήποτε άλλο σημείο που δεν ανήκει στη συστάδα



3 καλώς-διαχωρισμένες συστάδες

- Συχνά υπάρχει η έννοια του κατωφλιού (threshold)
- Όχι απαραίτητα κυκλικοί (οποιοδήποτε σχήμα)

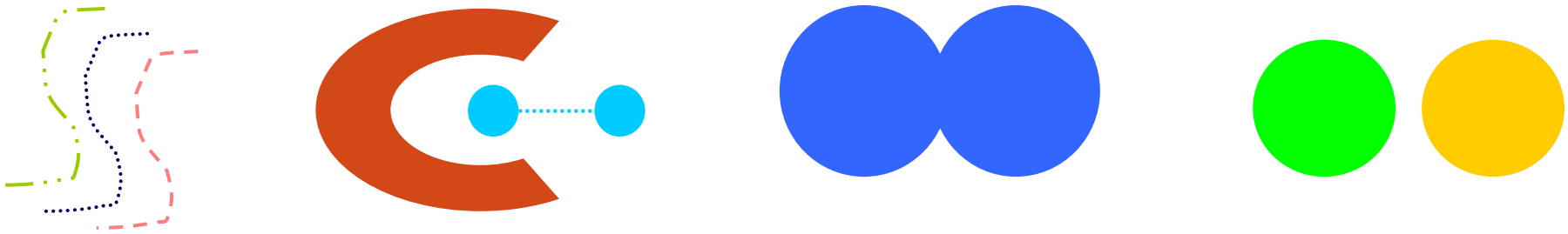
Τύποι Συστάδων: Βασισμένες σε κέντρο ή πρότυπο

- 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
- Κέντρο συστάδας:
  - Centroid: ο μέσος όρος των σημείων της συστάδας, ή
  - Medoid: το πιο «αντιπροσωπευτικό» σημείο της συστάδας (π.χ. για κατηγορικά γνωρίσματα)

4 συστάδες βασισμένες σε κέντρο
(Τείνουν στο να είναι κυκλικές)

Τύποι Συστάδων: Συνεχής (Contiguous Cluster)

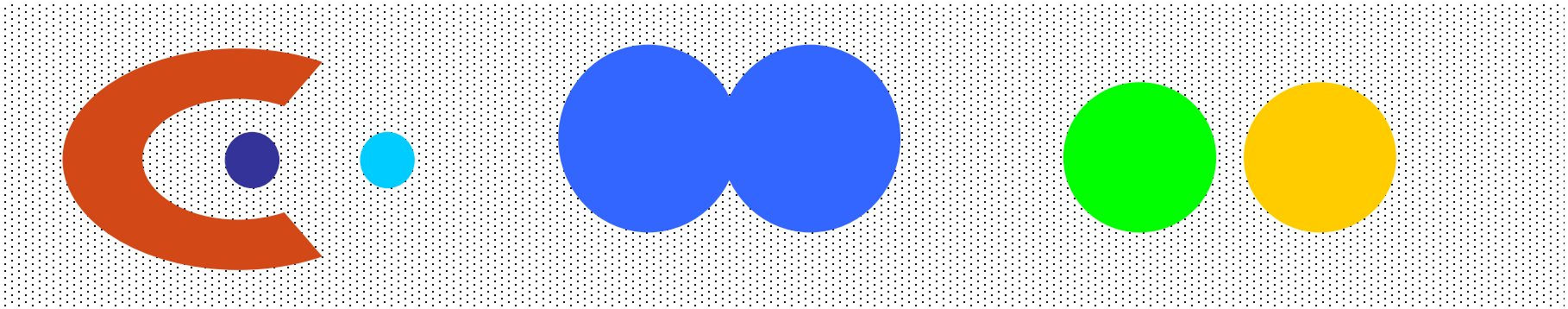
- Μια συστάδα είναι ένα σύνολο σημείων τέτοιο ώστε κάθε σημείο είναι πιο κοντά σε ένα ή περισσότερα σημεία της συστάδας από ό,τι σε οποιοδήποτε άλλο σημείο εκτός συστάδας
- Συχνά σε περιπτώσεις συστάδων με μη κανονικό σχήμα ή με αλληλοπλεκόμενα σχήματα
  - Ή όταν έχουμε γραφήματα και θέλουμε να βρούμε συνεκτικά υπογραφήματα
- Πρόβλημα με θόρυβο



8 συνεχείς συστάδες

Τύποι Συστάδων: Βασισμένες στην πυκνότητα

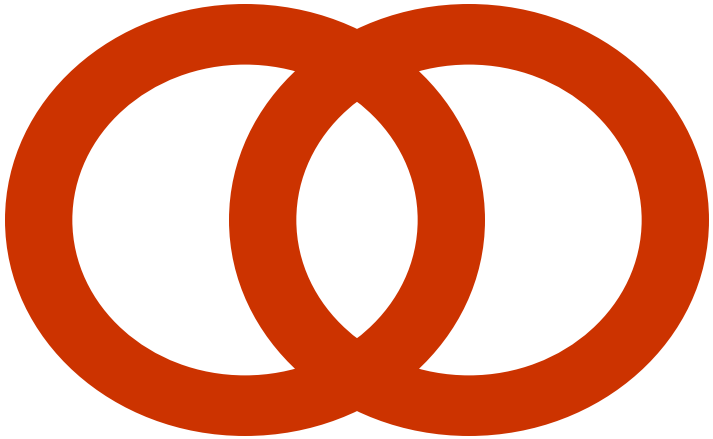
- Μια συστάδα είναι μια πυκνή περιοχή από σημεία την οποία χωρίζουν από άλλες περιοχές μεγάλης πυκνότητας περιοχές χαμηλής πυκνότητας
- Συχνό σε περιπτώσεις συστάδων με:
  - μη κανονικό σχήμα,
  - με αλληλοπλεκόμενα σχήματα,
  - ύπαρξη θορύβου ή outliers



6 συστάδες βασισμένες στην πυκνότητα

Τύποι Συστάδων: Συσταδοποίηση βάσει ιδιοτήτων ή εννοιών

Συστάδες με κοινή ιδιότητα ή εννοιολογικές συστάδες



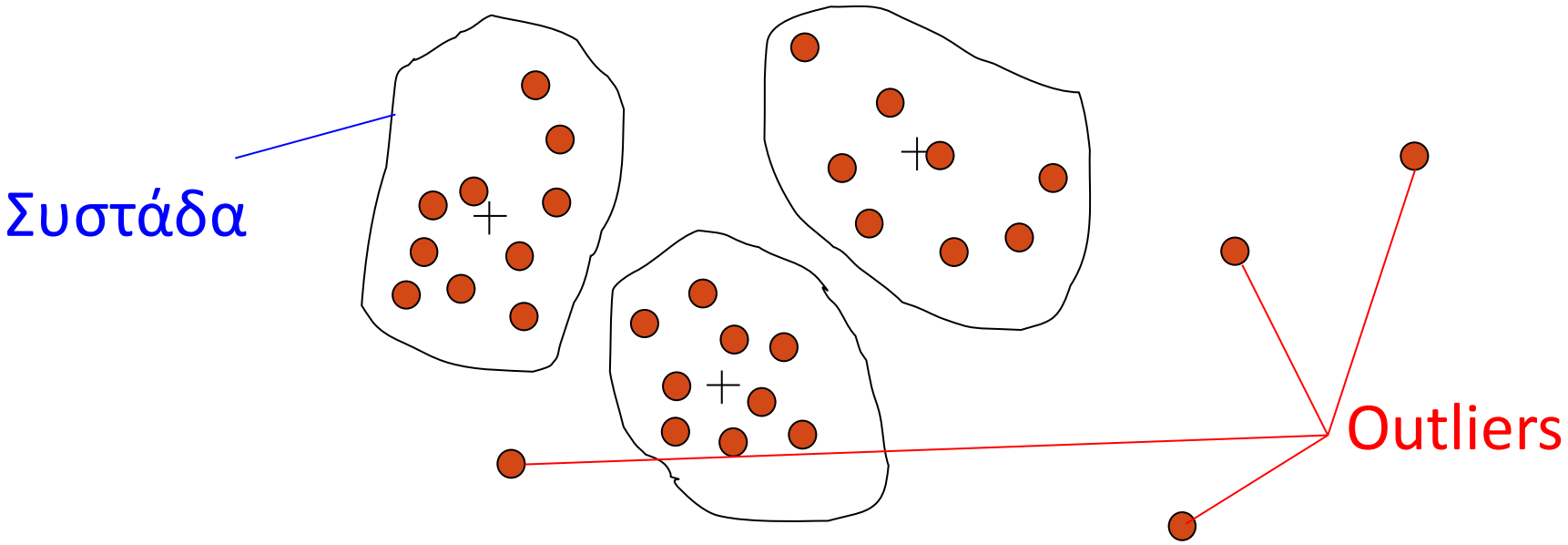
2 αλληλοκαλυπτόμενοι κύκλοι

Τύποι Συστάδων: Βασισμένες σε Αντικειμενική Συνάρτηση

- Εύρεση συστάδων που ελαχιστοποιούν ή μεγιστοποιούν μια *Αντικειμενική Συνάρτηση (Objective Function)*
- Απαρίθμηση όλων των δυνατών τρόπων χωρισμού των σημείων σε συστάδες και υπολογισμού του «πόσο καλό» είναι κάθε πιθανό σύνολο χρησιμοποιώντας τη δοθείσα αντικειμενική συνάρτηση
  - Οι στόχοι (objectives) μπορεί να είναι ολικοί (global) ή τοπικοί (local)
    - Η Ιεραρχική Συσταδοποίηση είναι συνήθως τοπικού στόχου
    - Η Διαχωριστική Συσταδοποίηση είναι συνήθως ολικού στόχου

Γενικές Απαιτήσεις

Αντιμετώπιση *θορύβου* και *outliers*



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

Άλλες διακρίσεις μεταξύ συνόλων Συστάδων

- Επικαλυπτόμενο σημείο ή όχι
  - Ένα σημείο ανήκει σε περισσότερες από μια συστάδες (π.χ. οριακά σημεία)
- Ασαφής συσταδοποίησης
  - Ένα σημείο ανήκει σε κάθε συστάδα με κάποιο βάρος μεταξύ 0 και 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 θεωρούνται θόρυβος, ή μη ενδιαφέρουσα πληροφορία
- Ετερογενή - Ομογενή
  - Συστάδες με πολύ διαφορετικά μεγέθη, σχήματα και πυκνότητες (densities)

Αλγόριθμοι Συσταδοποίησης

Θα δούμε τους:

- K-means (και παραλλαγές)
- Ιεραρχική Συσταδοποίηση
- Συσταδοποίηση με βάση την Πυκνότητα (DBSCAN)

Όμως πρώτα…
μία παρένθεση!!

- Μέση Τιμή
- Ομοιότητα και Απόσταση

Μέση Τιμή

Γενική Τάση

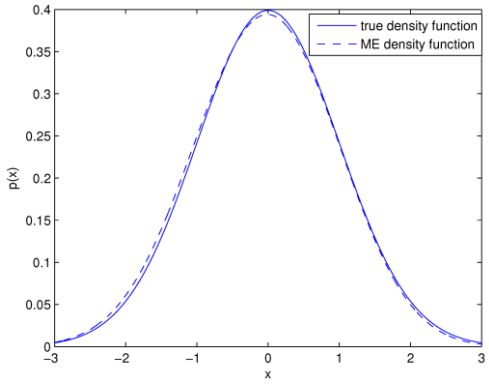
- Αριθμητικός Μέσος / Μέση Τιμή (Mean) $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
  - Αριθμητικός μέσος με βάρος (Weighted arithmetic mean) $\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$
  - Trimmed mean: κόβουμε τις ακραίες τιμές (π.χ. τα μεγαλύτερα και μικρότερα (p/2)%)
- Μέσο - μεσαία τιμή (median) - διάμεσος
  - Μεσαία τιμή αν μονός αριθμός τιμών, αλλιώς ο μέσος όρος των δυο μεσαίων τιμών
  - Το μέσο συμπεριφέρεται καλύτερα όταν δεδομένα με μη ομοιόμορφη κατανομή (skewed)

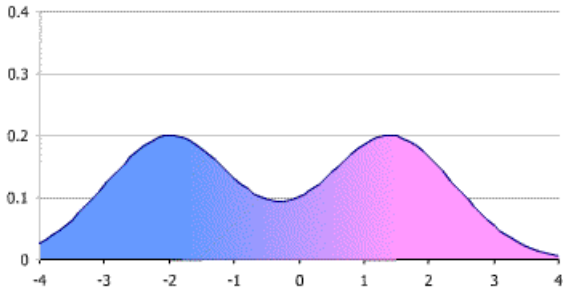
Άσκηση: {1 2 3 4 5 90}

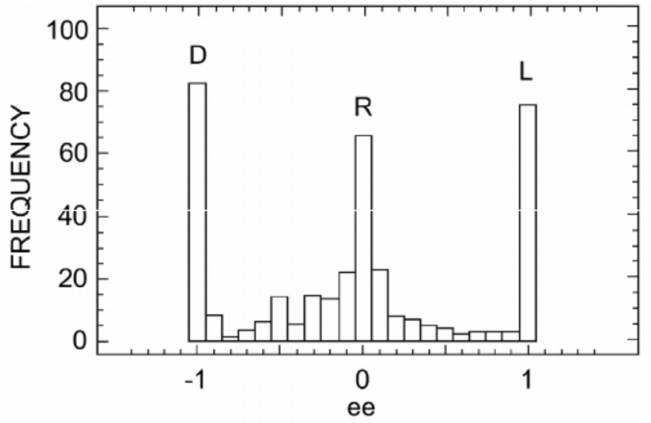
- Μέσο;
- Μέση τιμή;
- Trimmed 40%;

Γενική Τάση

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



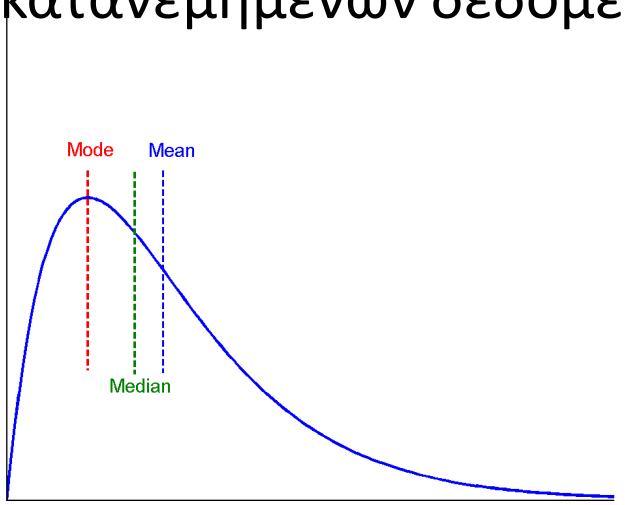


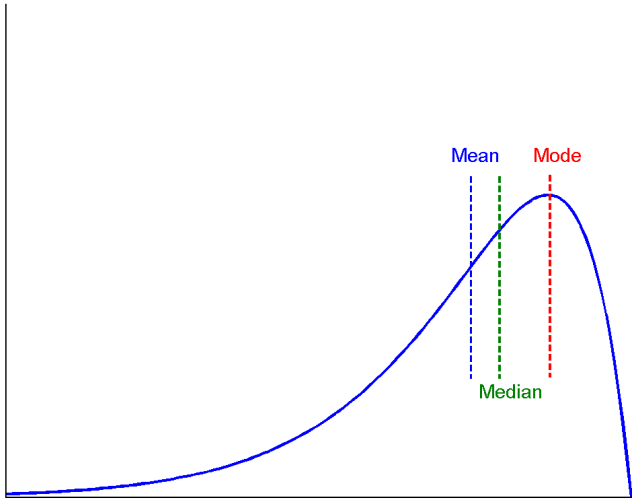


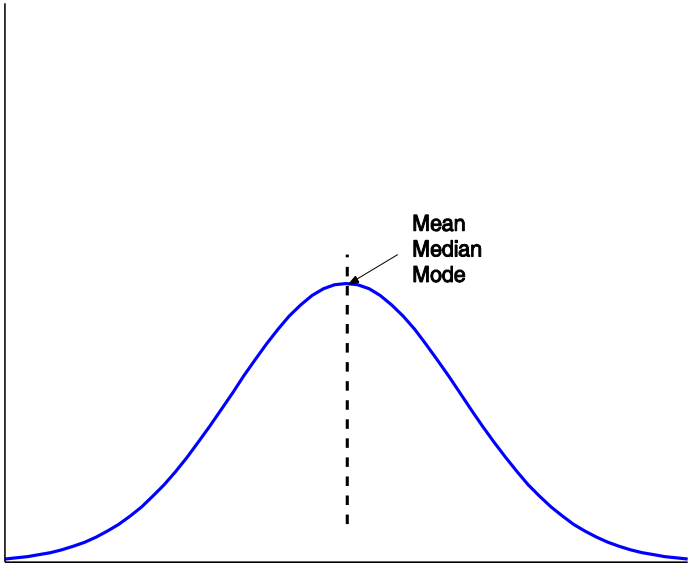
- Midrange (μέσο διαστήματος)
  - (min()+max()) / 2

Γενική Τάση

Οι mode, mean, και median ουδετέρων, θετικά, και αρνητικά κατανεμημένων δεδομένων







Γενική Τάση

- Distributed measure (κατανεμημένη μέτρηση)
  - Μπορεί να υπολογιστεί αν χωρίσουμε τα αρχικά δεδομένα σε μικρότερα υποσύνολα, υπολογίσουμε την τιμή σε κάθε υποσύνολο και τις συγχωνεύουμε
    - Π.χ. sum(), count(), max(), min()
- Algebraic measure (αλγεβρική μέτρηση)
  - Μπορεί να υπολογιστεί αν εφαρμόσουμε μια αλγεβρική (πολυωνυμική) συνάρτηση σε μία ή περισσότερες κατανεμημένες μετρήσεις
    - Π.χ. avg()= sum() / count()
- Holistic measure (ολιστική μέτρηση)
  - Πρέπει να υπολογιστεί στο σύνολο των δεδομένων

Διασπορά

- Variance σ^2 (Διακύμανση)
  - Μετρά «πόσο μακριά» ένα σύνολο αριθμών απλώνεται από τη μέση τιμή του

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

Mean (μέση τιμή)

- Standard deviation σ (Τυπική Απόκλιση)
  - Υπολογίζει το ποσό της μεταβολής ή της διασποράς ενός συνόλου τιμών δεδομένων

27

Απόσταση & Ομοιότητα

Κριτήρια Ομοιότητας - Απόσταση

- Ομοιότητα (Similarity)
  - Μια αριθμητική μέτρηση για το πόσο όμοια είναι δυο αντικείμενα
  - Μεγαλύτερη όσο πιο όμοια είναι τα αντικείμενα μεταξύ τους
  - Συχνά τιμές στο [0, 1]
- Μη Ομοιότητα (Dissimilarity)
  - Μια αριθμητική μέτρηση για το πόσο διαφορετικά είναι δυο αντικείμενα
  - Μικρότερη όσο ποιο όμοια είναι τα αντικείμενα μεταξύ τους
  - Η ελάχιστη τιμή είναι συνήθως 0, αλλά το άνω όριο διαφέρει

Κριτήρια Ομοιότητας - Απόσταση

- Η ομοιότητα-μη ομοιότητα μεταξύ δύο αντικειμένων μετριέται συνήθως βάση μιας *συνάρτησης απόστασης* ανάμεσα στα αντικείμενα
- Εξαρτάται από το είδος των δεδομένων, δηλαδή από το είδος των γνωρισμάτων τους
- Δύο μεγάλες κατηγορίες:
  - Ευκλείδειες: Βασισμένες στη θέση των σημείων, αποστάσεις των σημείων στο χώρο
  - Μη Ευκλείδειες: Βασισμένες σε άλλες ιδιότητες των σημείων πλην της θέσης τους

Ευκλείδειες Μετρικές Απόστασης

Έστω δυο μεταβλητές *i* και *j* με *n* γνωρίσματα x_{ik} και x_{jk}

- Ο πιο συνηθισμένος τρόπος → Ευκλείδεια απόσταση

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

- *Βάρη* για Ευκλείδεια απόσταση

$$d(i,j)=\sqrt{(w_1|x_{i_1}-x_{j_1}|^2+w_2|x_{i_2}-x_{j_2}|^2+...+w_n|x_{i_n}-x_{j_n}|^2)}$$

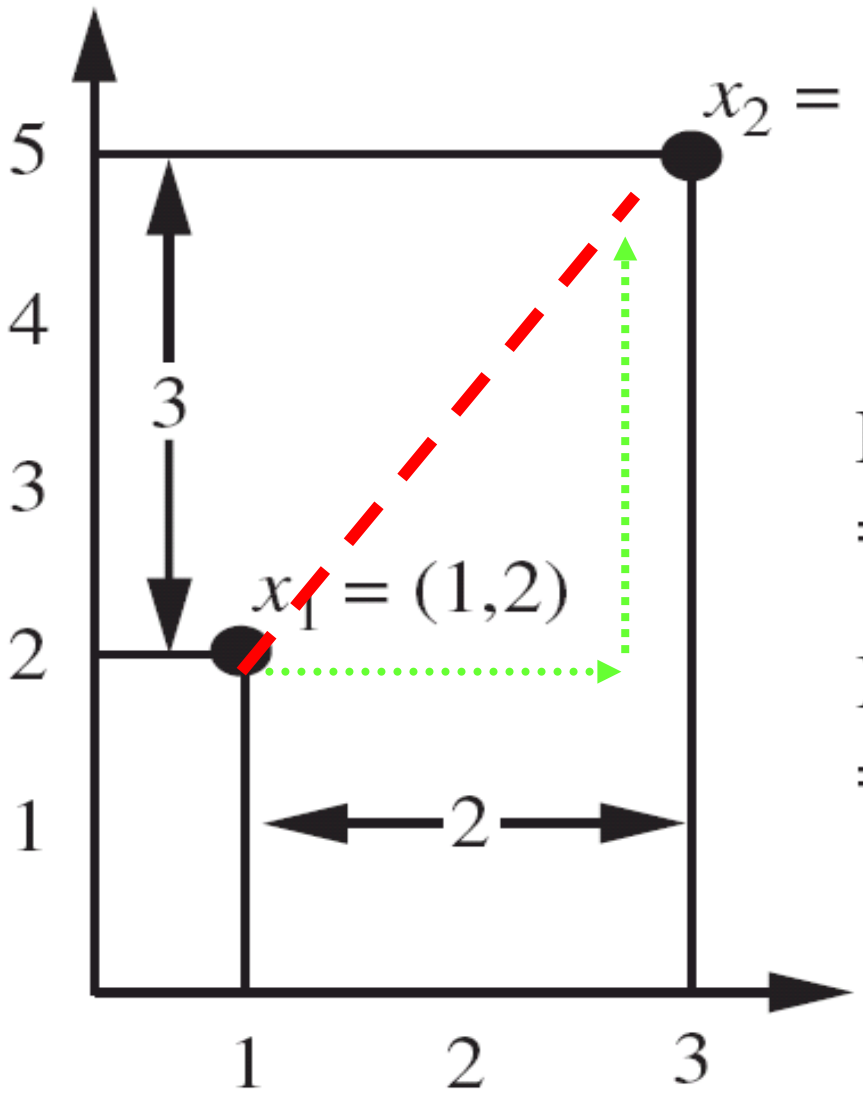
- Manhattan ή Taxicab ή City Block απόσταση

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

- Minkowski απόσταση

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

Παράδειγμα Απόστασης

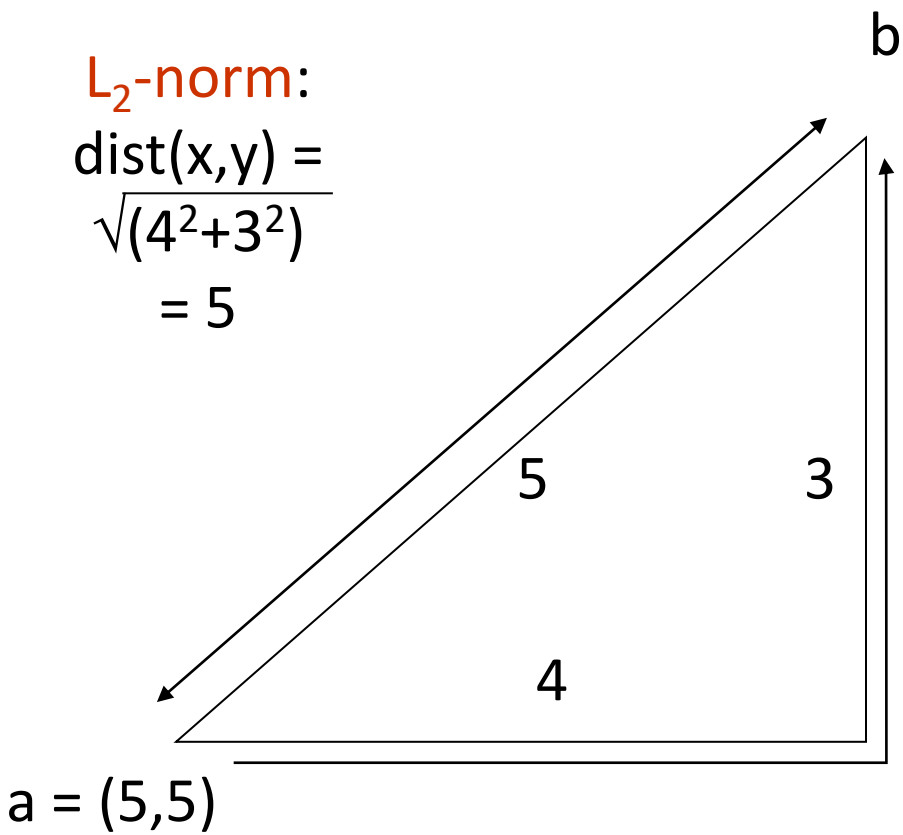


Euclidean distance
$= (2^2 + 3^2)^{1/2} = 3.61$

Manhattan distance
$= 2 + 3 = 5$

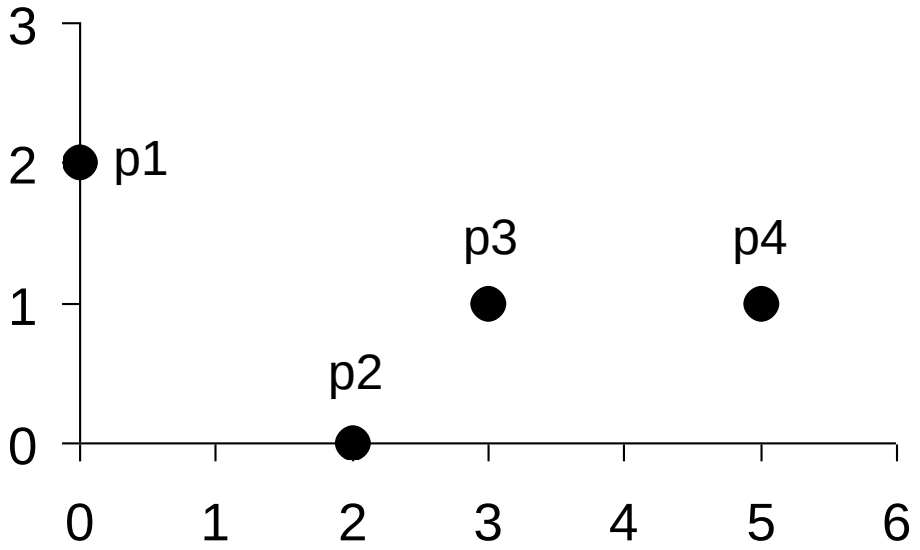
Παράδειγμα Απόστασης

L_2-norm:
dist(x,y) =
$\sqrt{(4^2+3^2)}$
= 5



L_1-norm:
dist(x,y) =
4+3 = 7

Παράδειγμα Απόστασης



Πίνακας Δεδομένων

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

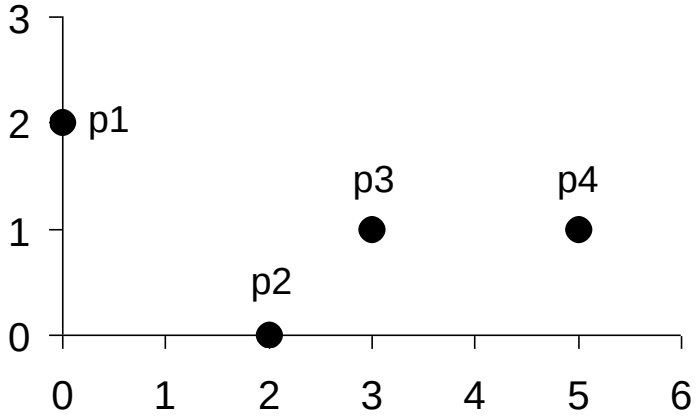
Πίνακας Απόστασης

	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

Ορισμός Απόστασης

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

Παράδειγμα Απόστασης



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

Μη Ευκλείδειες Μετρικές Απόστασης

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

Δυαδικές Μεταβλητές

Συχνά έχουμε δεδομένα με μόνο δυαδικά γνωρίσματα (δυαδικά διανύσματα)

- Συμμετρικές (τιμές 0 και 1 έχουν την ίδια σημασία)
  - Invariant ομοιότητα
- Μη συμμετρικές (η συμφωνία στο 1 πιο σημαντική – π.χ. σηματοδοτεί την ύπαρξη κάποιας ασθένειας)
  - Non-invariant (Jaccard)

Μη Ευκλείδειες Μετρικές Απόστασης

Μεταξύ δύο αντικειμένων *i* και *j* με δυαδικά γνωρίσματα

- M_{01} = ο αριθμός των γνωρισμάτων που το *i* έχει τιμή 0 και το *j* έχει 1
- M_{10} = ο αριθμός των γνωρισμάτων που το *i* έχει τιμή 1 και το *j* έχει 0
- M_{00} = ο αριθμός των γνωρισμάτων που το *i* έχει τιμή 0 και το *j* έχει 0
- M_{11} = ο αριθμός των γνωρισμάτων που το *i* έχει τιμή 1 και το *j* έχει 1

ΟΜΟΙΟΤΗΤΑ

- Απλό ταίριασμα – συμμετρικές μεταβλητές
  - SMC = αριθμός ταιριασμάτων / αριθμός γνωρισμάτων = $(M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00})$
- J = αριθμός 11 ταιριασμάτων / αριθμό μη μηδενικών γνωρισμάτων = $(M_{11}) / (M_{01} + M_{10} + M_{11})$
- J → Συντελεστής Jaccard (Jaccard Coefficient) - μη συμμετρικές μεταβλητές (διαφορετική σημασία στην τιμή 1 και στην τιμή 0)

Παράδειγμα Μη Ευκλείδειας Μετρικής Απόστασης

p = 1 0 0 0 0 0 0 0 0 0 $\quad M_{01} = 2, \quad M_{00} = 7$

q = 0 0 0 0 0 0 1 0 0 1 $\quad M_{10} = 1, \quad M_{11} = 0$

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

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

Παράδειγμα Μη Ευκλείδειας Μετρικής Απόστασης

Πίνακας συνάφειας (Contingency table) για δυαδικά δεδομένα

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

Μέτρηση απόστασης για συμμετρικές δυαδικές μεταβλητές

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

Μέτρηση απόστασης για μη συμμετρικές δυαδικές μεταβλητές

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

Jaccard coefficient

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

Κατηγορικές Μεταβλητές χωρίς Διάταξη (nominal)

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

1η Μέθοδος: Απλό ταίριασμα

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

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

2η Μέθοδος: Χρήση πολλών δυαδικών μεταβλητών

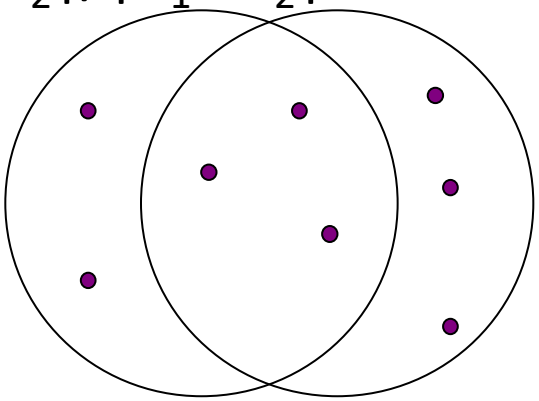
Μία για κάθε μία από τις μ τιμές

Jaccard Ομοιότητα για Σύνολα

- Η Jaccard ομοιότητα για δύο σύνολα είναι το μέγεθος της τομής προς το μέγεθος της ένωσης τους

$$Sim\ (C_1, C_2) = |C_1 \cap C_2| / |C_1 \cup C_2|$$

- Παράδειγμα
  - Μέγεθος τομής = 3, Μέγεθος ένωσης = 8
  - Jaccard ομοιότητα = 3/8



- Δυαδική αναπαράσταση συνόλου (1 το στοιχείο υπάρχει, 0 αλλιώς)
  - p1 = 10111, p2 = 10011
  - Μέγεθος τομής = 3, Μέγεθος ένωσης = 4, Jaccard ομοιότητα (όχι απόσταση) = 3/4
  - Απόσταση: d(x,y) = 1 - (Jaccard ομοιότητα) = 1/4

Ομοιότητα Συνημίτονου (Cosine Similarity)

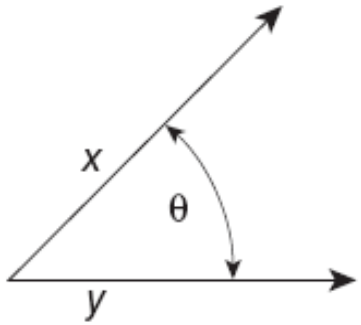
Θέλουμε μια απόσταση που να αγνοεί τα 0 (όπως η Jaccard) αλλά να δουλεύει και για μη δυαδικά δεδομένα

Επίσης, αγνοεί το *μήκος* των διανυσμάτων

- Αν d_1 and d_2 είναι διανύσματα κειμένου:
  $\cos(d_1, d_2) = (d_1 \bullet d_2) / ||d_1|| \, ||d_2||$,
  όπου • εσωτερικό γινόμενο $||d||$ το μήκος του d

- Παράδειγμα:
  d_1 = **3 2 0 5 0 0 0 2 0 0** , d_2 = **1 0 0 0 0 0 0 1 0 2**
- $d_1 \bullet d_2 = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5$
- $||d_1|| = (3*3+2*2+0*0+5*5+0*0+0*0+0*0+2*2+0*0+0*0)^{0.5} = (42)^{0.5} = 6.481$
- $||d_2|| = (1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2)^{0.5} = (6)^{0.5} = 2.245$
- $\cos(d_1, d_2) = .3150$

Ομοιότητα Συνημίτονου (Cosine Similarity)



Γεωμετρική ερμηνεία

- Ομοιότητα 1, όταν η γωνία 0
  - τα x και y ίδια (αν εξαιρέσουμε το μήκος τους)
- Ομοιότητα 0, όταν η γωνία 90 (κανένας κοινός όρος)

Απόσταση Edit

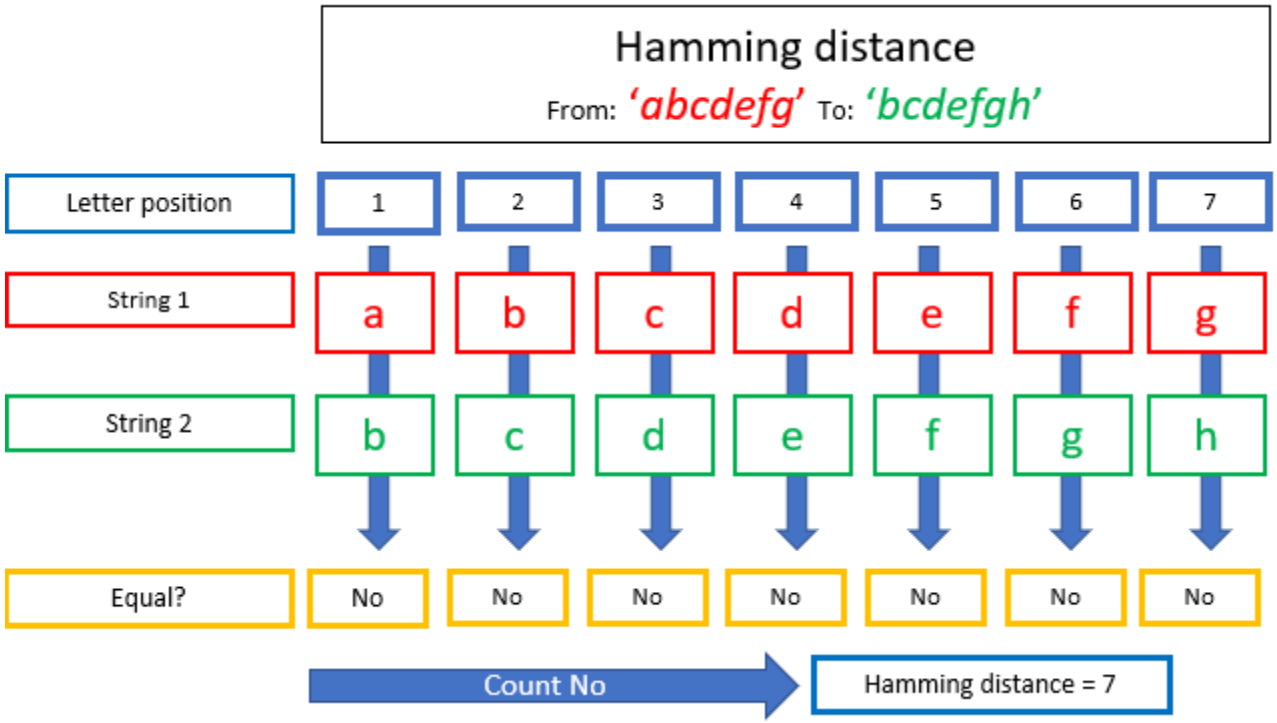
Για δύο συμβολοσειρές (strings) ορίζεται ως ο ελάχιστος αριθμός εισαγωγών/διαγραφών χαρακτήρων που χρειάζονται για να πάμε από τη μία στην άλλη

Παράδειγμα

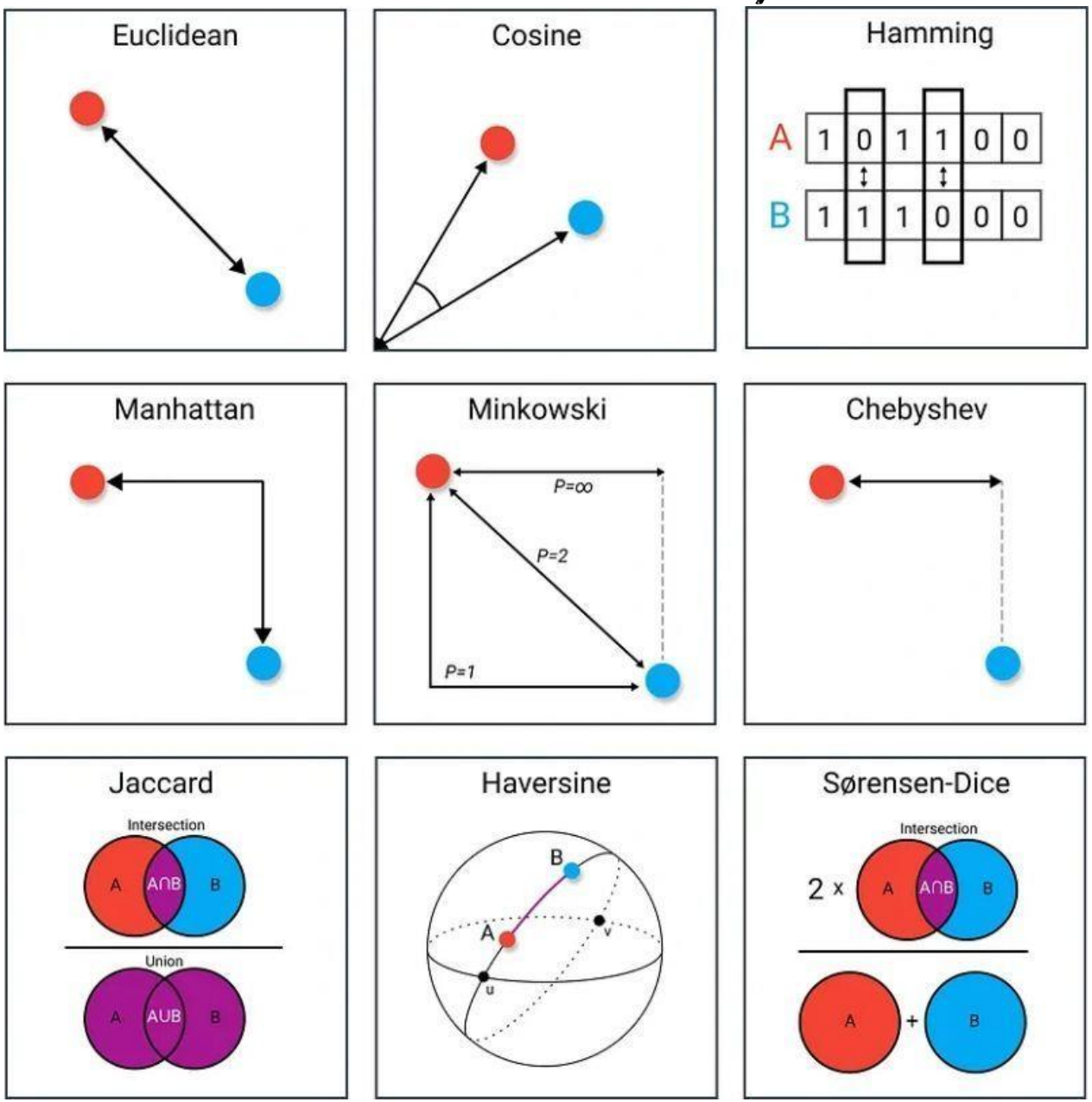
- *x* = *abcde*, *y* = *bcduve*
- Μετατροπή του *x* σε *y* διαγράφοντας το *a*, και έπειτα εισάγοντας τα *u*, *v* μετά το *d*
  - Edit Απόσταση = 3

Απόσταση Hamming

- Για δύο συμβολοσειρές (strings) ίσου μήκους ορίζεται ως ο αριθμός θέσεων στις οποίες τα αντίστοιχα σύμβολα είναι διαφορετικά
- Μετρά τον ελάχιστο αριθμό αντικαταστάσεων που χρειάζονται ώστε να μετατραπεί η μία συμβολοσειρά στην άλλη



Αποστάσεις



https://towardsdatascience.com/9-distance-measures-in-data-science

k-means (και παραλλαγές)

Διαίρεση *n* παρατηρήσεων σε *k* συστάδες

Εισαγωγή

- Διαχωριστικός αλγόριθμος (βασισμένος σε πρότυπο)
- Κάθε συστάδα συσχετίζεται με ένα κεντρικό σημείο (centroid)
- Κάθε σημείο ανατίθεται στη συστάδα με το κοντινότερο κεντρικό σημείο
- Ο αριθμός των ομάδων *k* είναι είσοδος στον αλγόριθμο

Αλγόριθμος k-means

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

2: **Repeat**

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

4: Επανυπολογισμός του κεντρικού σημείου κάθε συστάδας

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

- Το αποτέλεσμα εξαρτάται σε μεγάλο βαθμό από την επιλογή των αρχικών σημείων
- Ουσιαστικά, ο αλγόριθμος προσπαθεί επαναληπτικά να «μειώσει» την απόσταση όλων των σημείων από ένα σημείο της συστάδας

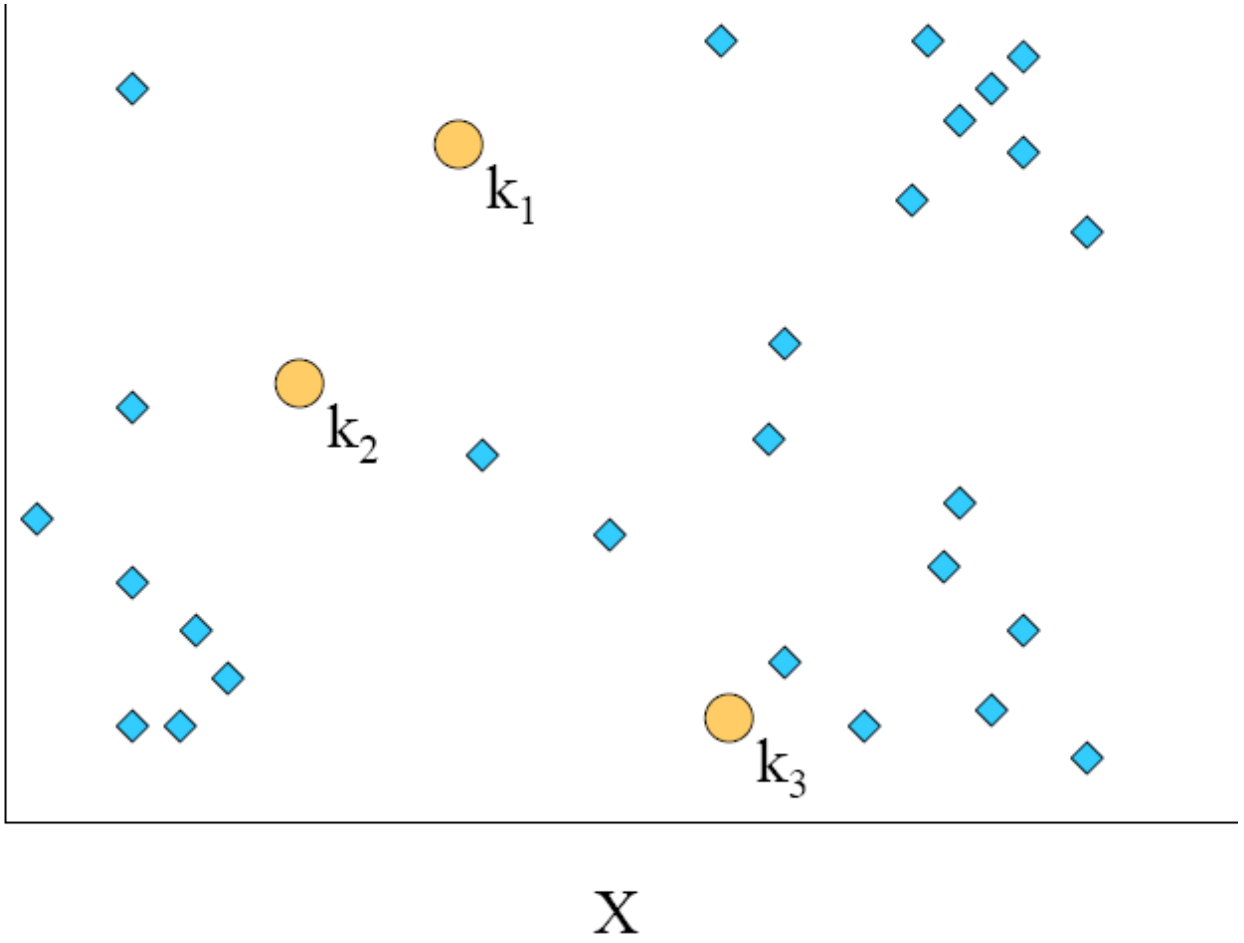
Αλγόριθμος k-means

Παρατηρήσεις

1. Τα αρχικά κεντρικά σημεία συνήθως επιλέγονται τυχαία
   - Οι συστάδες που παράγονται διαφέρουν από τη μία διάτρεξη του αλγορίθμου στην άλλη
2. Η εγγύτητα των σημείων υπολογίζεται με βάση κάποια απόσταση που εξαρτάται από το είδος των σημείων
   - Στα παραδείγματα θα θεωρήσουμε την *Ευκλείδεια απόσταση*
   - Επειδή η απόσταση υπολογίζεται συχνά ο υπολογισμός της πρέπει να είναι σχετικά απλός
3. Το κεντρικό σημείο είναι (συνήθως) το μέσο (mean) των σημείων της συστάδας
   - Μπορεί να *μην είναι ένα από τα δεδομένα εισόδου*

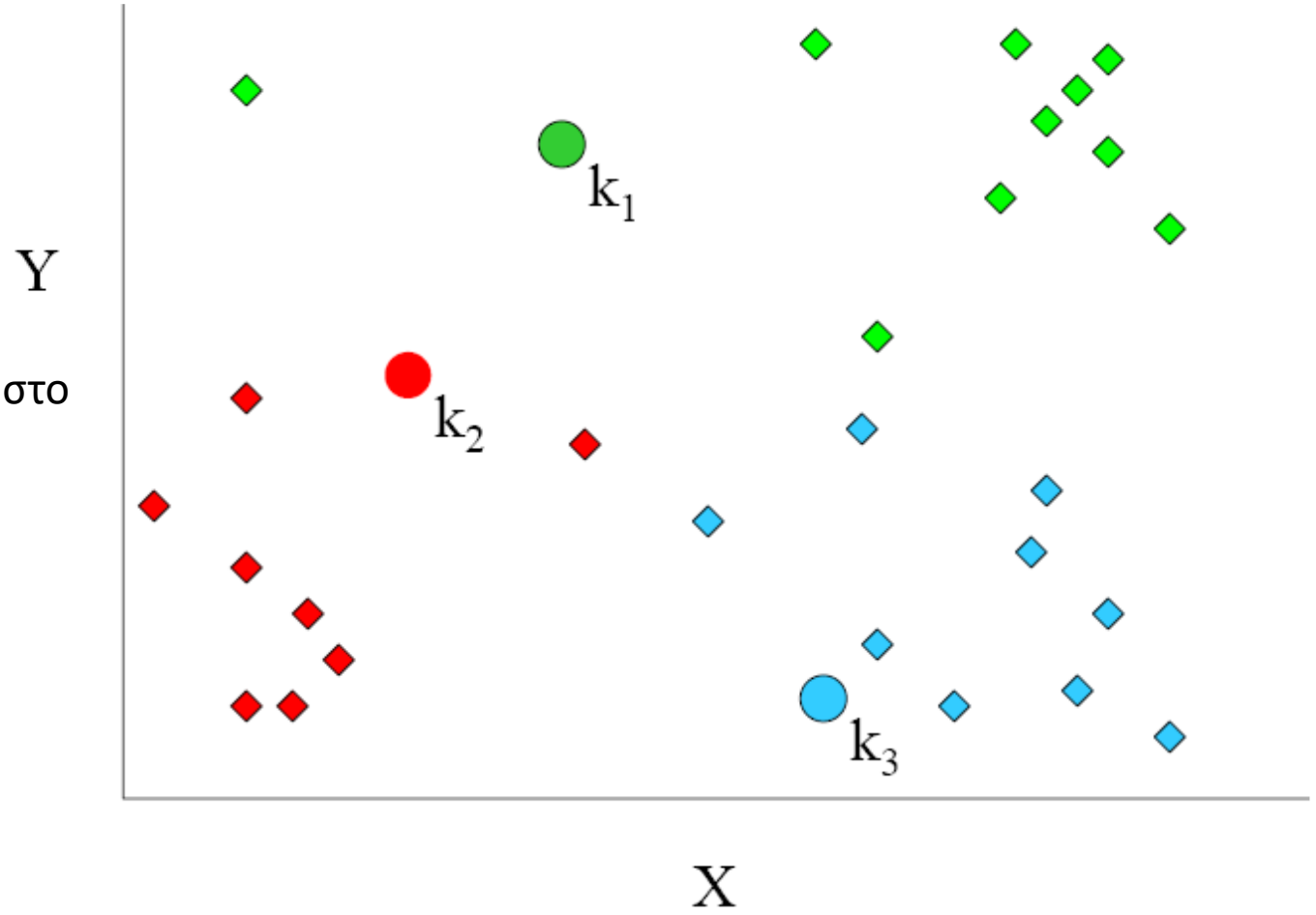
Αλγόριθμος k-means

- Αρχική κατάσταση
- K = 3 συστάδες
- Αρχικά σημεία k_1, k_2, k_3



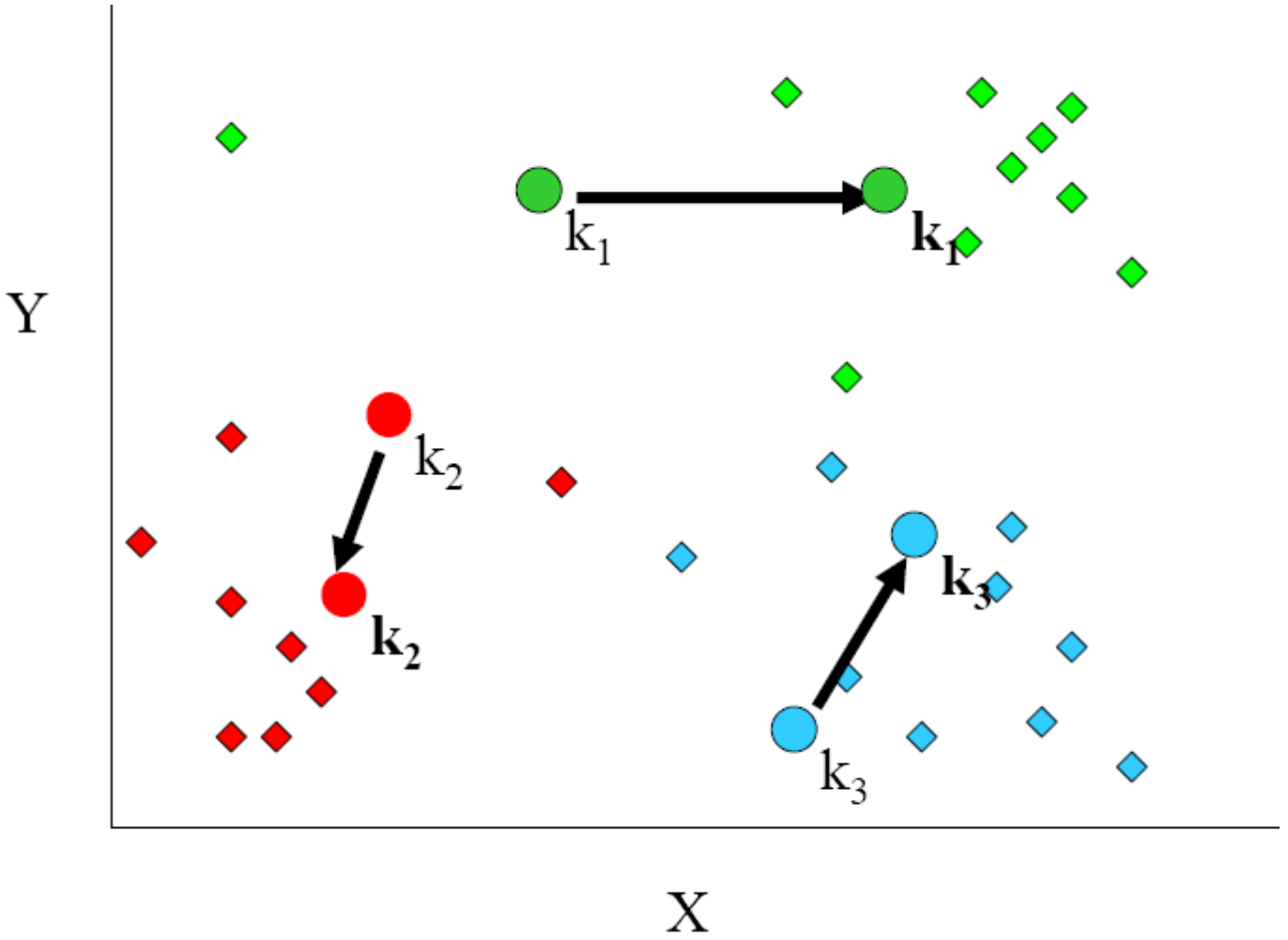
Αλγόριθμος k-means

Τα σημεία ανατίθενται στο πιο γειτονικό από τα 3 αρχικά σημεία



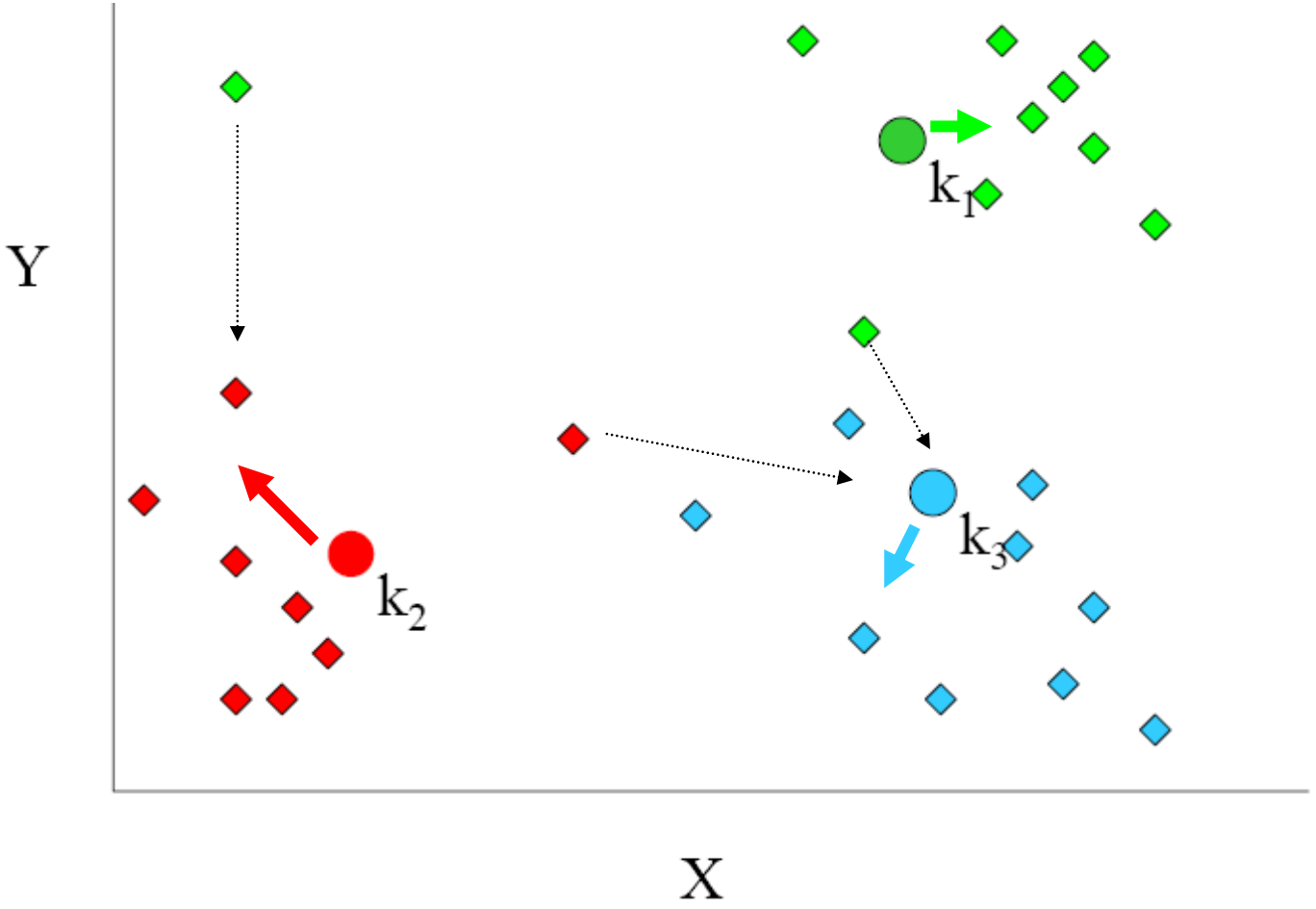
Αλγόριθμος k-means

Επανυπολογισμός του κέντρου (βάρους) κάθε σημείου

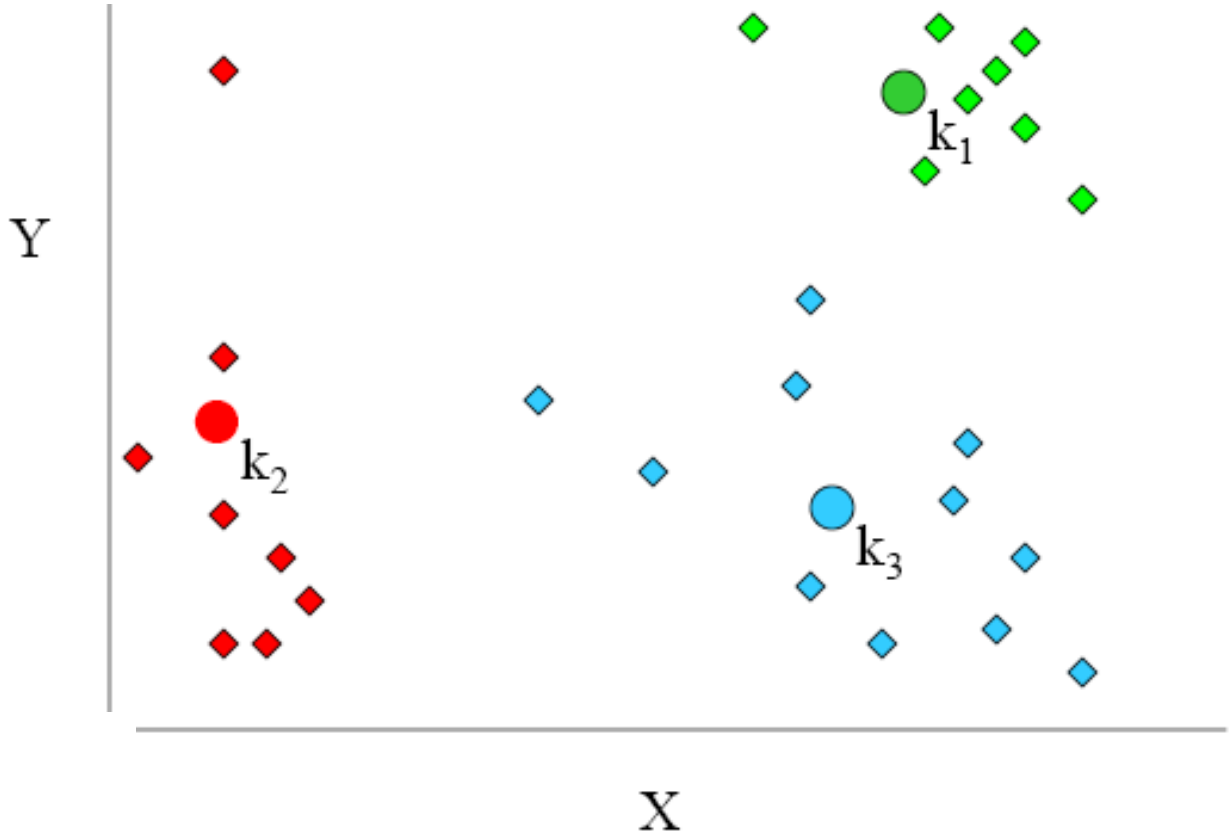


Αλγόριθμος k-means

- Νέα ανάθεση των σημείων
- Νέα κέντρα βάρους



Αλγόριθμος k-means



Δεν αλλάζει τίποτα → ΤΕΛΟΣ!

Αλγόριθμος k-means

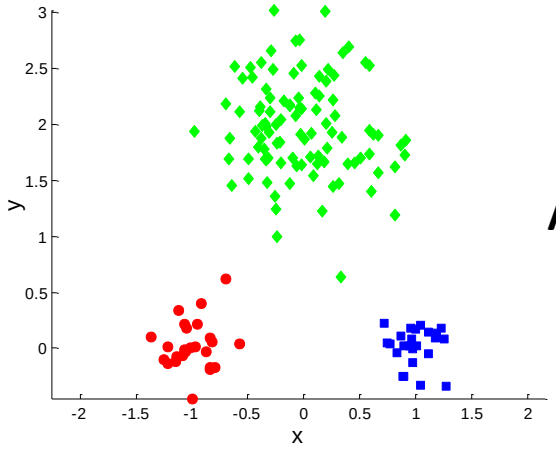
Παρατηρήσεις (συνέχεια)

1. Χώρος: αποθηκεύουμε μόνα τα κέντρα
2. Η πολυπλοκότητα είναι O(I * n * k * d)
   - *n* = αριθμός σημείων
   - *k* = αριθμός συστάδων
   - *I* = αριθμός επαναλήψεων
   - *d* = αριθμός γνωρισμάτων (διάσταση)
3. Για συνηθισμένα μέτρα ομοιότητας, ο αλγόριθμος συγκλίνει
   - Η σύγκλιση συμβαίνει συνήθως στις πρώτες επαναλήψεις
4. Συχνά η τελική συνθήκη αλλάζει σε «Until …
   - … σχετικά λίγα σημεία να αλλάζουν συστάδα» ή
   - … η απόσταση μεταξύ των νέων κεντρικών σημείων από τα παλιά να είναι μικρή»

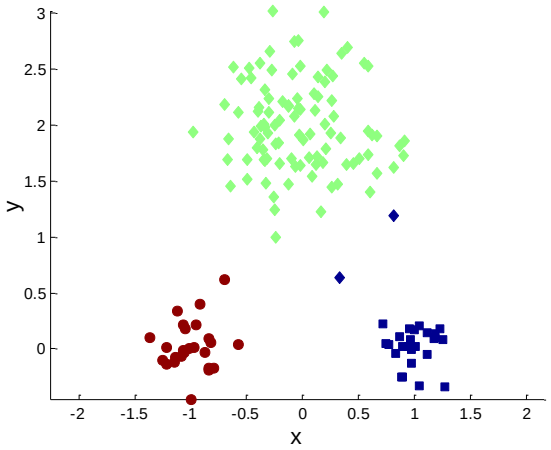
Εκτίμηση ποιότητας

- Η πιο συνηθισμένη μέτρηση είναι το *Άθροισμα του Τετραγωνικού Σφάλματος (ΑΤΣ)* (*Sum of Squared Error (SSE)*)
  - Για κάθε σημείο, το λάθος είναι η απόστασή του από την κοντινότερη συστάδα
  - Για να πάρουμε το SSE, παίρνουμε το τετράγωνο αυτών των λαθών και τα προσθέτουμε $$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$
  - Όπου *dist* ή Ευκλείδεια απόσταση, *x* είναι ένα σημείο στη συστάδα C_i, και m_i είναι ο αντιπρόσωπος (κεντρικό σημείο) της συστάδας C_i
- Δοθέντων δύο συστάδων, μπορούμε να επιλέξουμε αυτήν με το μικρότερο λάθος
- Ένας τρόπος να βελτιώσουμε τη συσταδοποίηση (ελάττωση του SSE) είναι να μεγαλώσουμε το *k*
  - Αλλά γενικά μια καλή συσταδοποίηση με μικρό *k* μπορεί να έχει μικρότερο SSE από μια κακή συσταδοποίηση με μεγάλο *k*

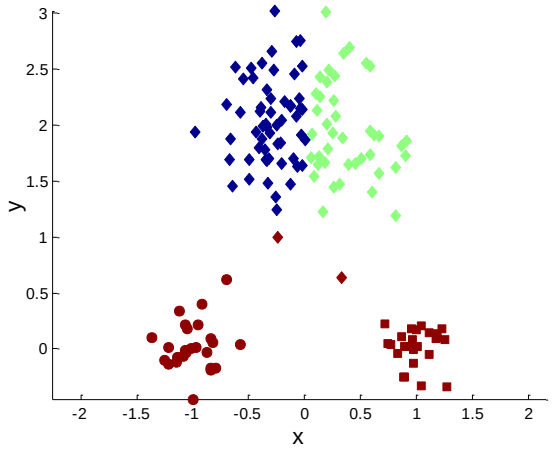
Παράδειγμα 1



Αρχικά σημεία

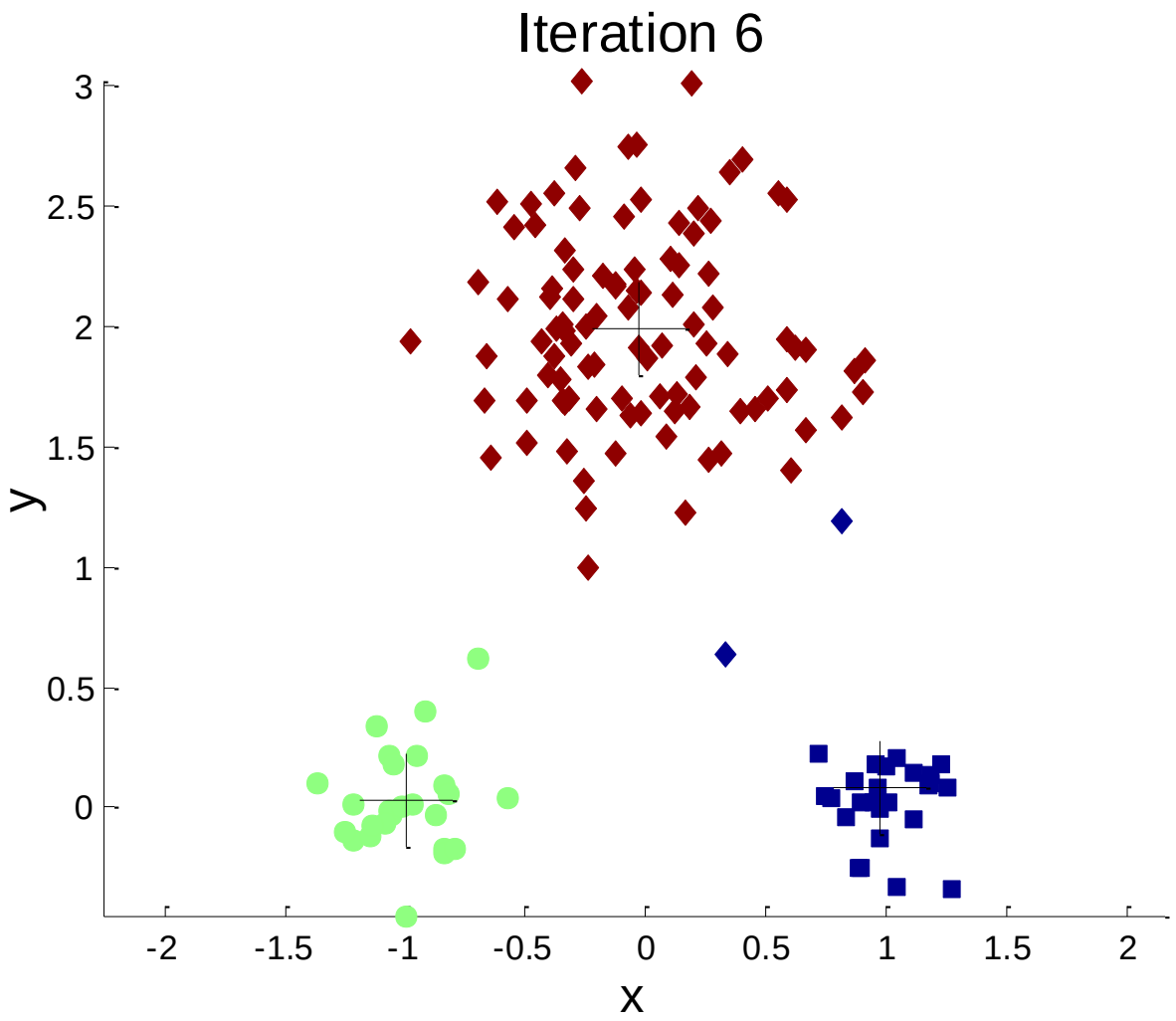


Βέλτιστη συσταδοποίηση

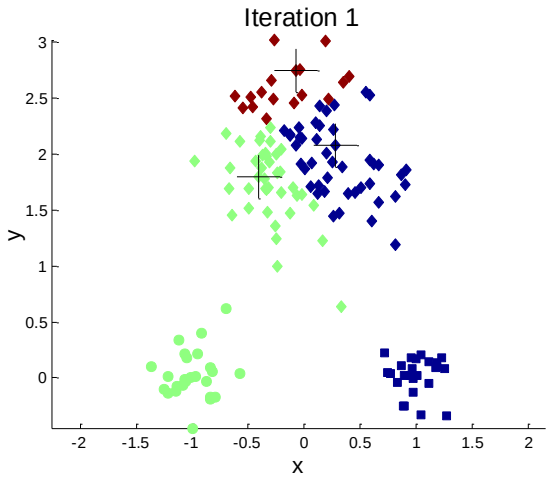


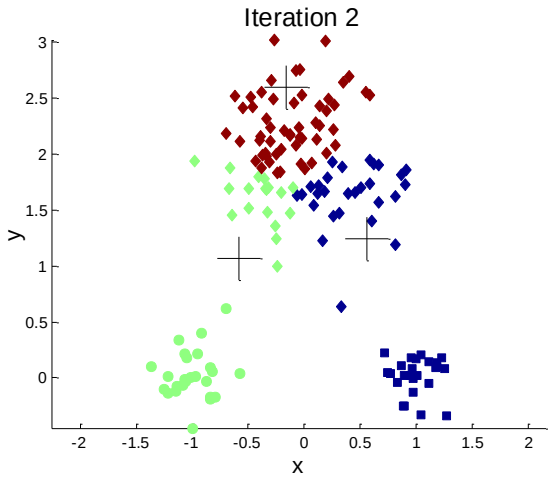
Υπό-βέλτιστη συσταδοποίση

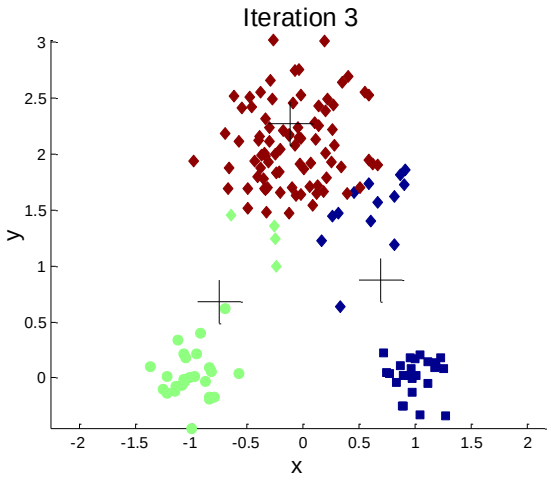
Παράδειγμα 2

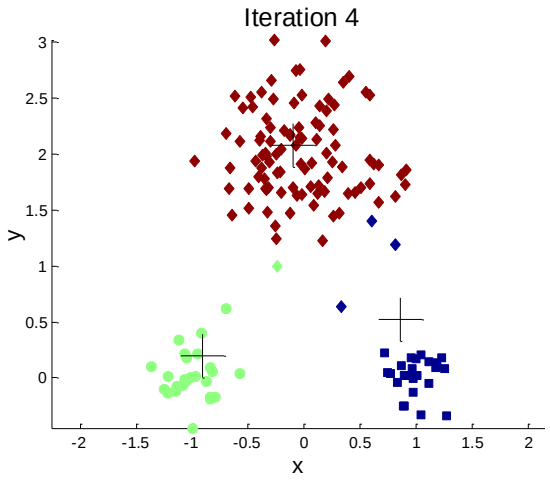


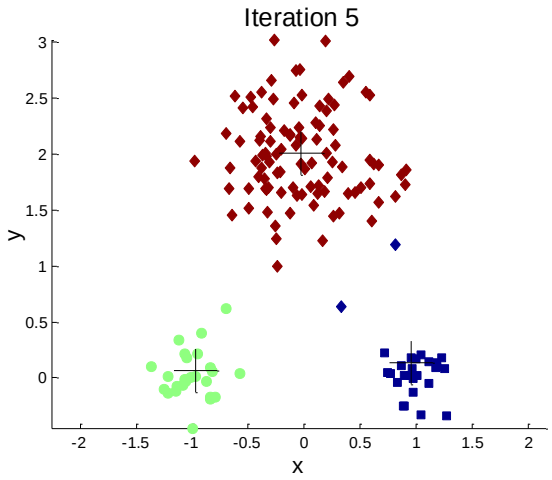
Παράδειγμα 2

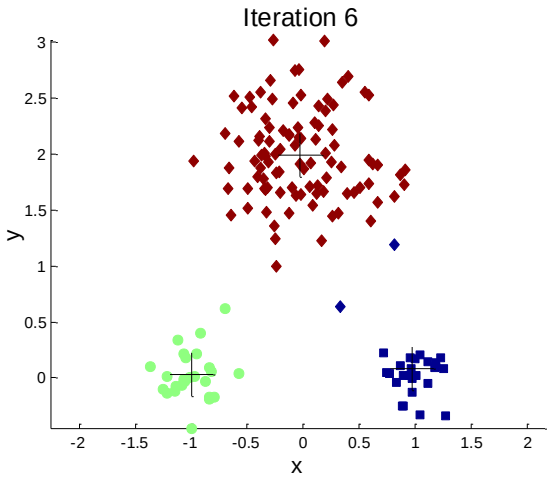




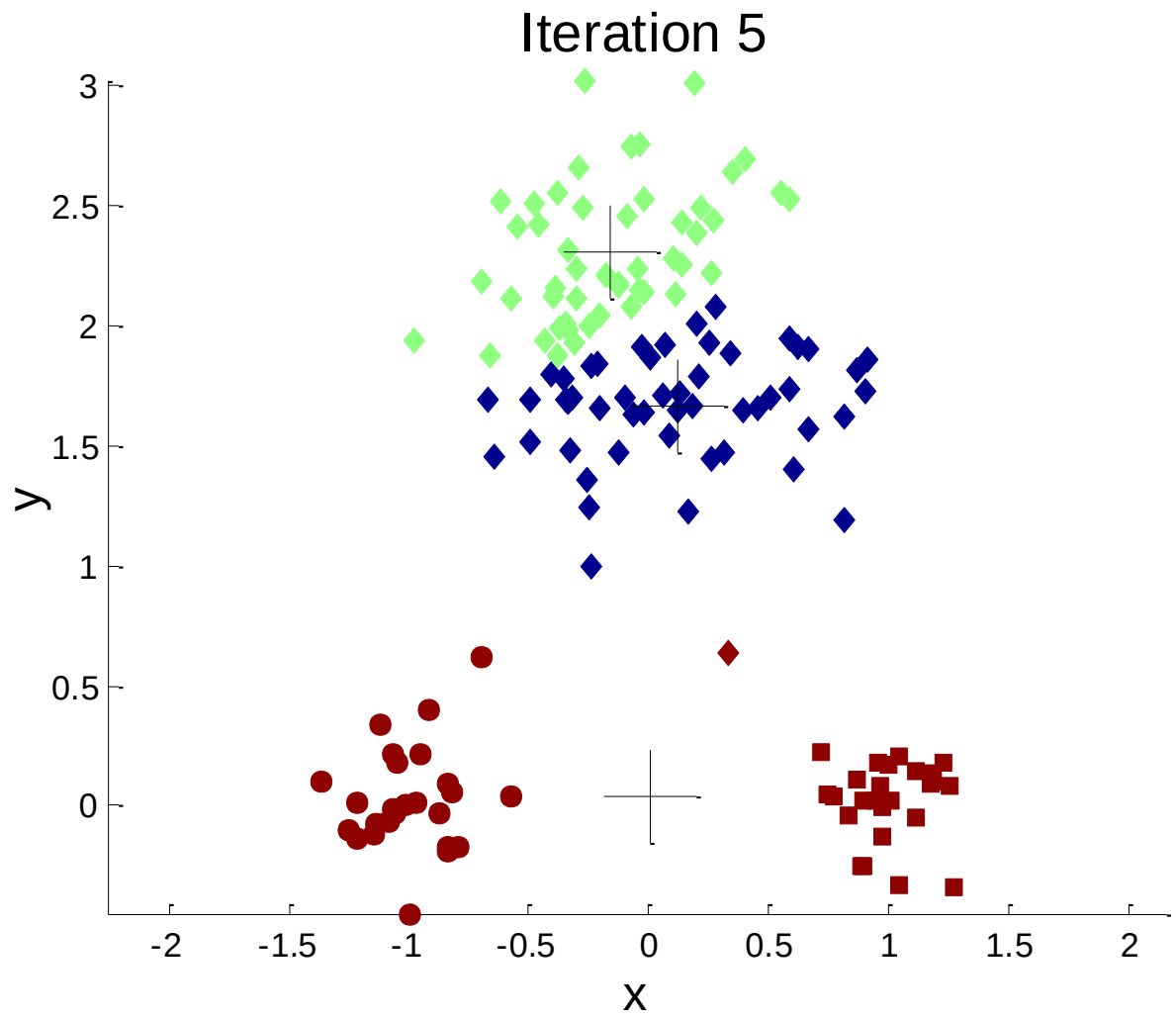




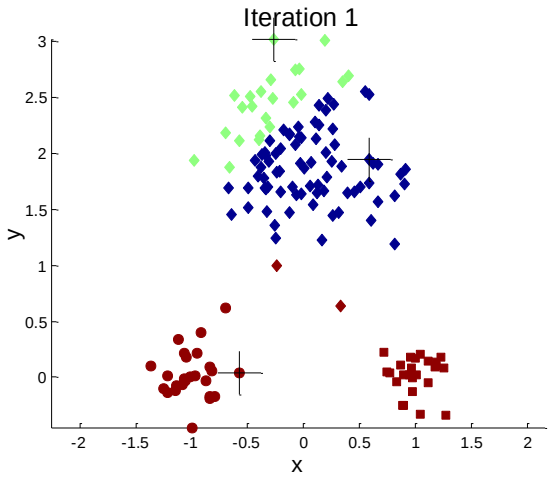


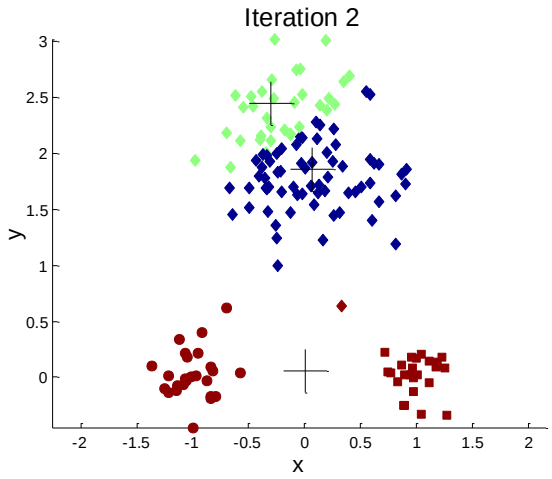


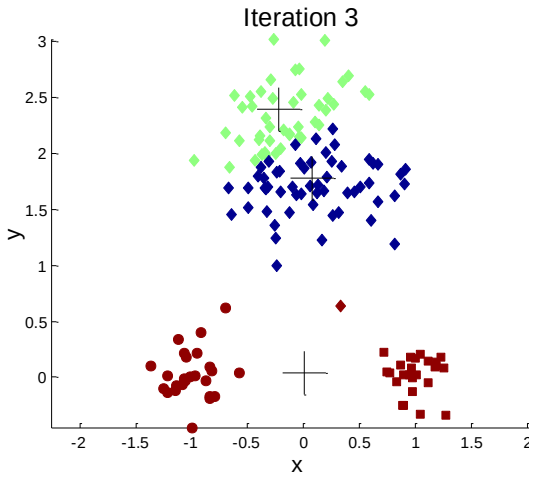
Παράδειγμα 3

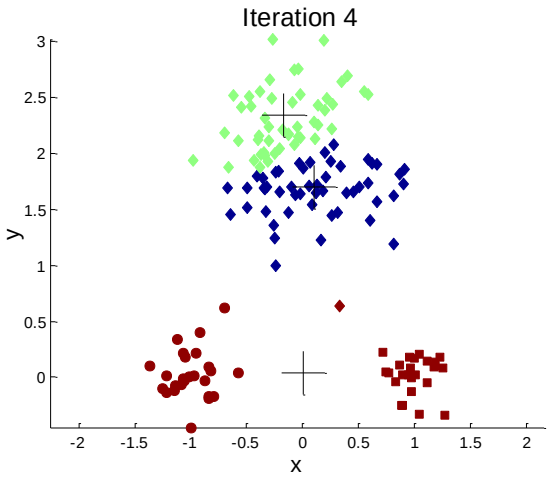


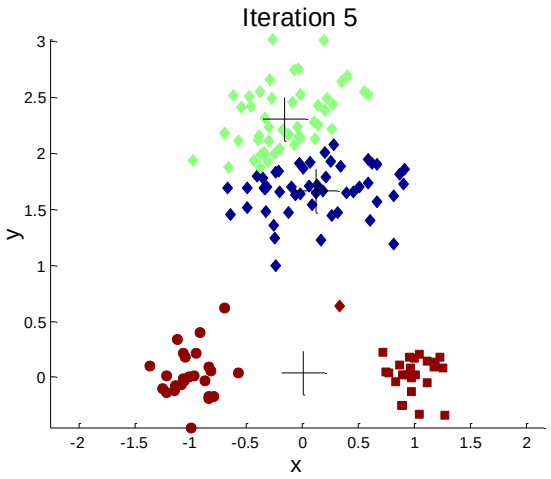
Παράδειγμα 3



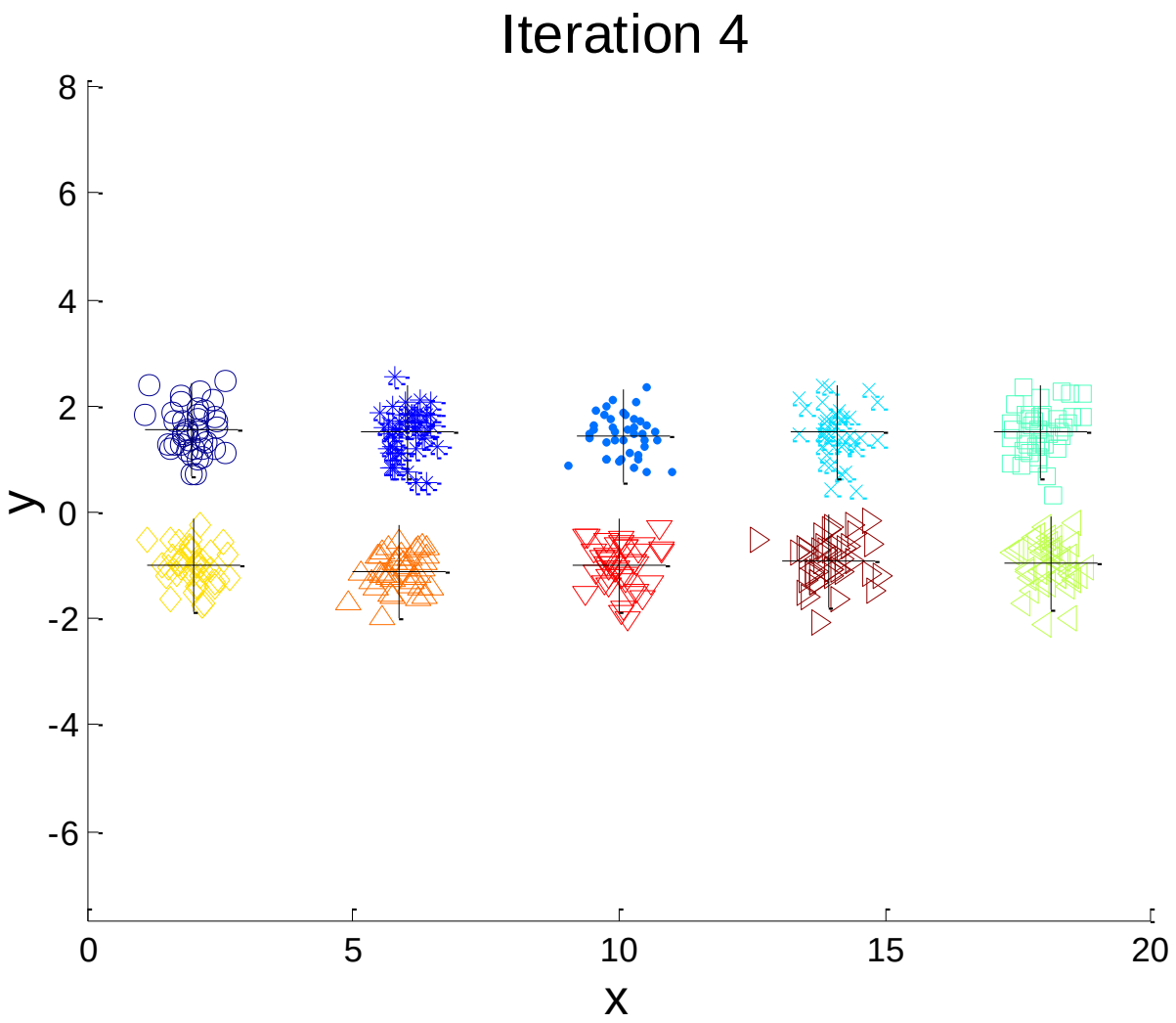






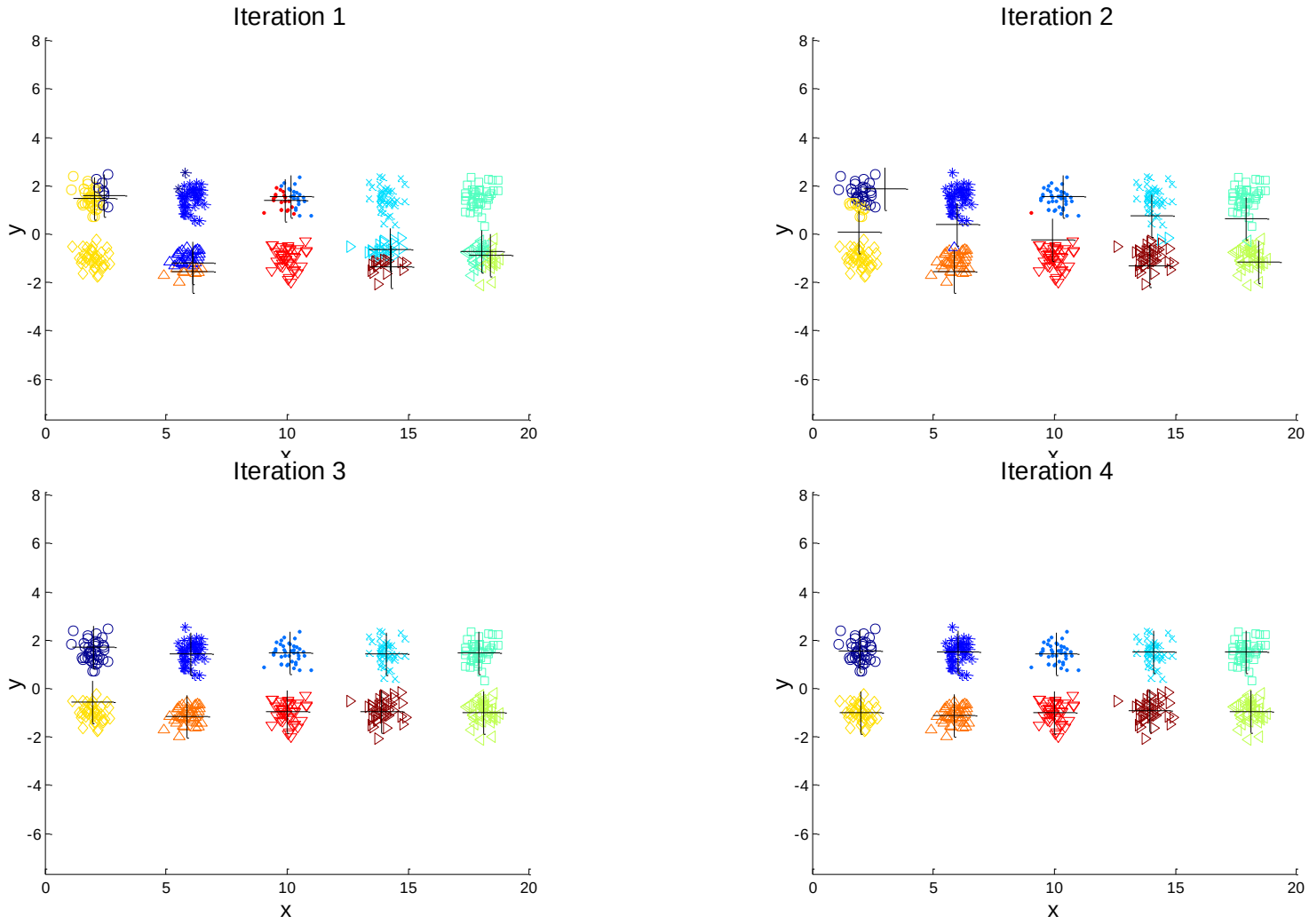


Παράδειγμα 1 10 συστάδων



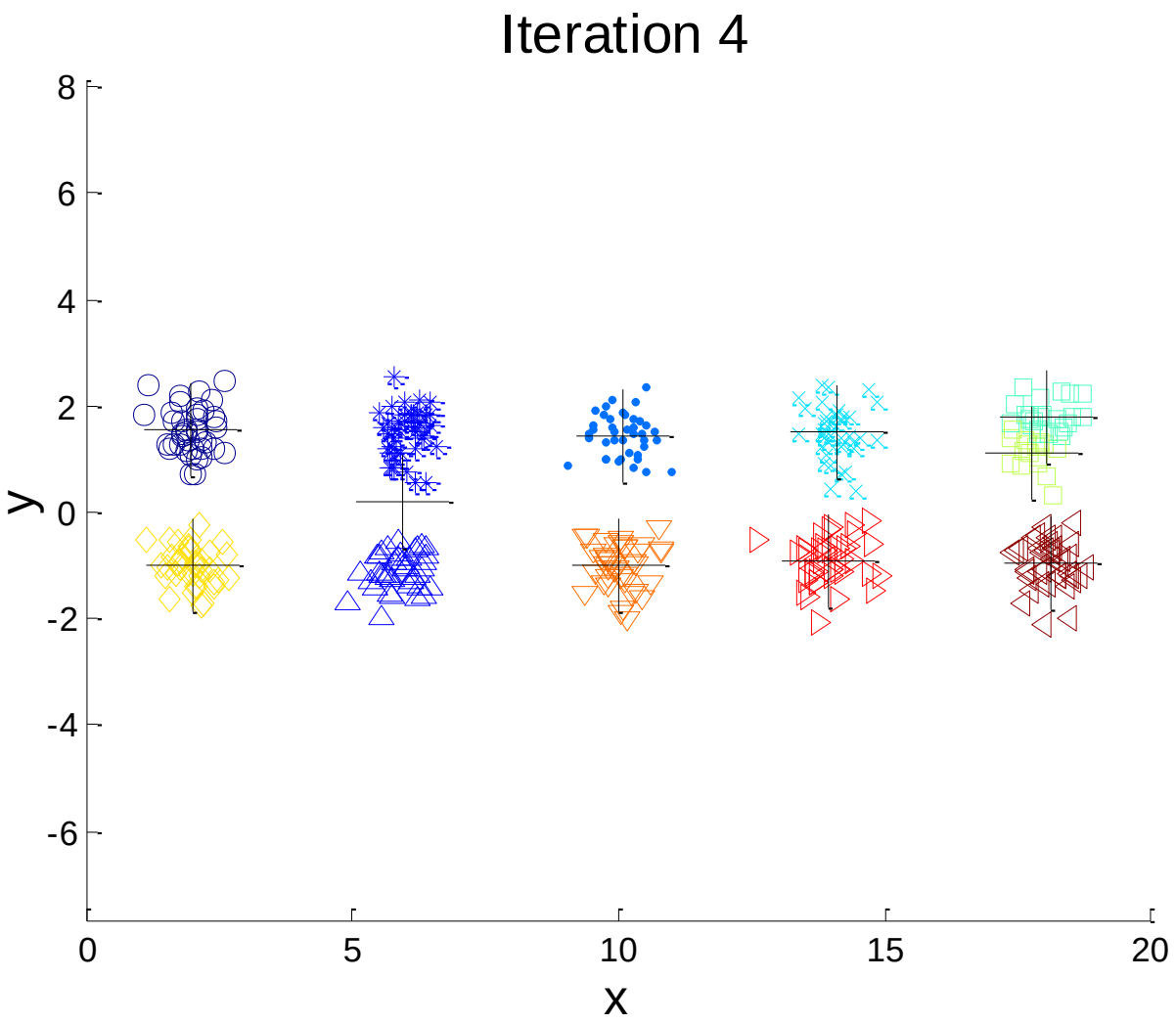
Ξεκινώντας με δύο αρχικά σημεία σε κάθε συστάδα κάθε ζεύγους συστάδων

Παράδειγμα 2 10 συστάδων



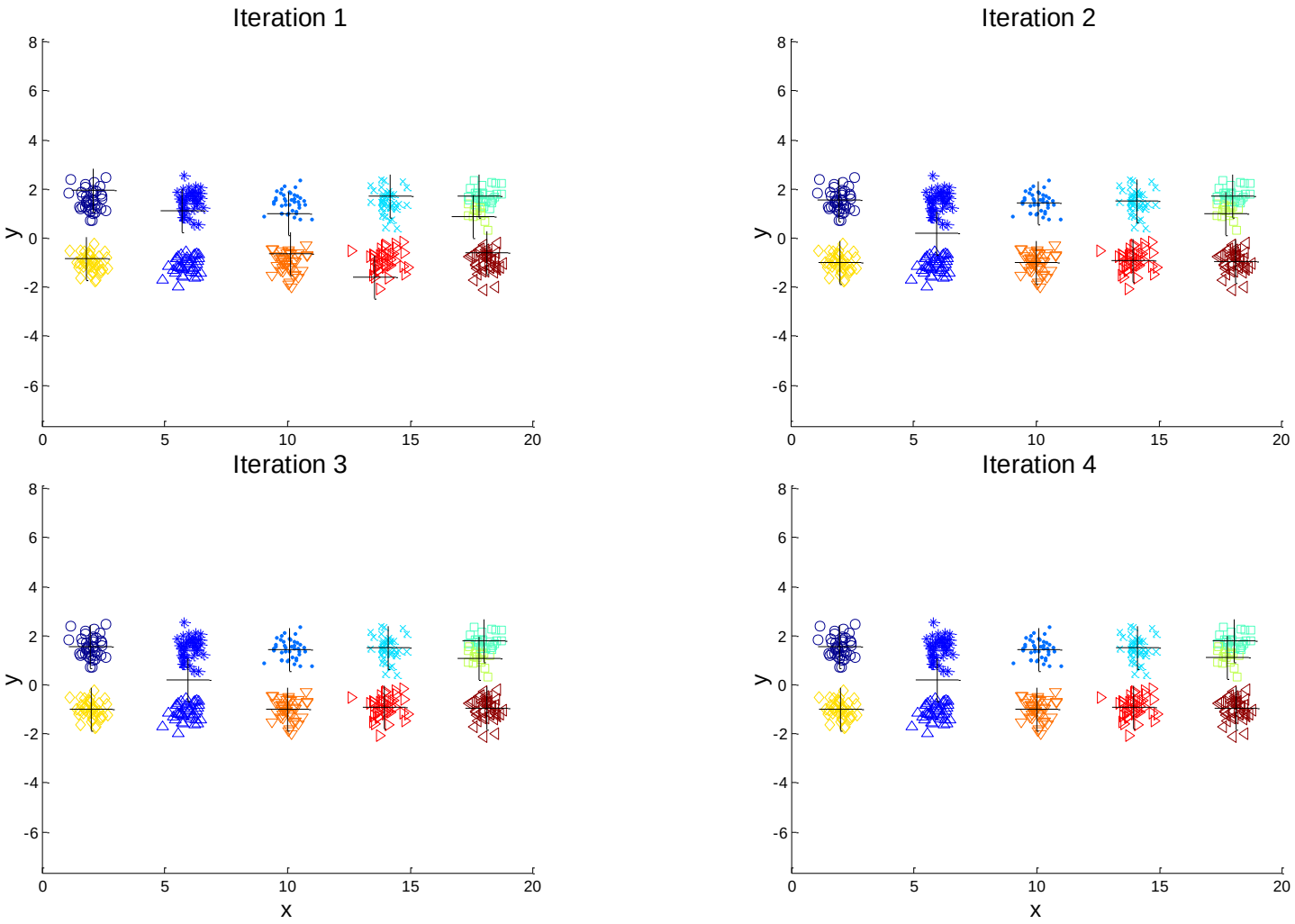
Ξεκινώντας με δύο αρχικά σημεία σε κάθε συστάδα κάθε ζεύγους συστάδων

Παράδειγμα 3 10 συστάδων



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

Παράδειγμα 4 10 συστάδων



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

Επιλογή αρχικών σημείων

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

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

Για παράδειγμα, αν *k* = 10, η πιθανότητα είναι
P = 10!/10^{10} = 0.00036

Λύσεις για την επιλογή αρχικών σημείων

- Πολλαπλά τρεξίματα
  - Βοηθά, αλλά πολλές περιπτώσεις
- Δειγματοληψία και χρήση κάποιας ιεραρχικής τεχνικής
- Επιλογή παραπάνω από *k* αρχικών σημείων
- Σταδιακή επιλογή
  - Επιλογή του πρώτου σημείου τυχαία ή ως το μέσο όλων των σημείων
  - Για καθένα από τα υπόλοιπα αρχικά σημεία επιλογή αυτού που είναι πιο μακριά από τα μέχρι τώρα επιλεγμένα αρχικά σημεία
    - Μπορεί να οδηγήσει στην επιλογή outliers
    - Ο υπολογισμός του πιο απομακρυσμένου σημείου είναι δαπανηρός
    - Συχνά εφαρμόζεται σε δείγματα

Άδειες συστάδες

- Ο βασικός αλγόριθμος μπορεί να οδηγήσει σε *άδειες συστάδες*
- Πολλές στρατηγικές
  - Επιλογή του σημείου που είναι πιο μακριά από όλα τα τωρινά κέντρα = επιλογή του σημείου που συμβάλει περισσότερο στο SSE
  - Ένα σημείο από τη συστάδα με το υψηλότερο SSE θα οδηγήσει σε «σπάσιμο» της, άρα σε μείωση του λάθους
  - Αν πολλές *άδειες συστάδες*, τα παραπάνω βήματα μπορεί να επαναληφθούν πολλές φορές

Σταδιακή ενημέρωση κεντρικών σημείων

- Στο βασικό k-means, τα κέντρα ενημερώνονται αφού όλα τα σημεία έχουν ανατεθεί στο κέντρο
- Μια παραλλαγή είναι να ενημερώνονται τα κέντρα μετά από κάθε ανάθεση (incremental approach)
  - Κάθε ανάθεση ενημερώνει 0 ή 2 κέντρα
  - Πιο δαπανηρό
  - Έχει σημασία η σειρά εισαγωγής/εξέτασης των σημείων
  - Δεν υπάρχουν άδειες συστάδες
  - Μπορεί να χρησιμοποιηθούν βάρη αν υπάρχει κάποια τυχαία αντικειμενική συνάρτηση
    - Έλεγχος τι συμφέρει κάθε φορά

Προ και Μετα Επεξεργασία

- Ολικό SSE και SSE Συστάδας
- Προ-επεξεργασία (Post-processing)
  - Κανονικοποίηση των δεδομένων
  - Απομάκρυνση outliers
- Μετά-επεξεργασία (Post-processing)
  - Split-Merge (διατηρώντας το ίδιο K)
    - Διαχωρισμός (split) συστάδων με το σχετικά μεγαλύτερο SSE
    - Δημιουργία μια νέας συστάδας (π.χ. επιλέγοντας το σημείο που είναι πιο μακριά από όλα τα κέντρα ή τυχαία επιλογή σημείου ή επιλογή του σημείου με το μεγαλύτερο SSE)
    - Συνένωση (merge) συστάδων που είναι σχετικά κοντινές (τα κέντρα τους έχουν την μικρότερη απόσταση) ή τις δυο συστάδες που οδηγούν στην μικρότερη αύξηση του SSE
    - Διαγραφή συστάδας και ανακατανομή των σημείων της σε άλλες συστάδες (αυτό που οδηγεί στην μικρότερη αύξηση του SSE)

k-means με Διχοτόμηση (Bisecting k-means)

- Τροποποίηση του αλγορίθμου k-means
- Μπορεί να παράγει τμηματική/ιεραρχική ομαδοποίηση

k-means με διχοτόμηση (bisecting k-means)

Παραλλαγή που μπορεί να παράγει μια διαχωριστική ή ιεραρχική συσταδοποίηση

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

2: **Repeat**

3: Επιλογή μιας συστάδας από τη λίστα των συστάδων

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

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

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

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

---

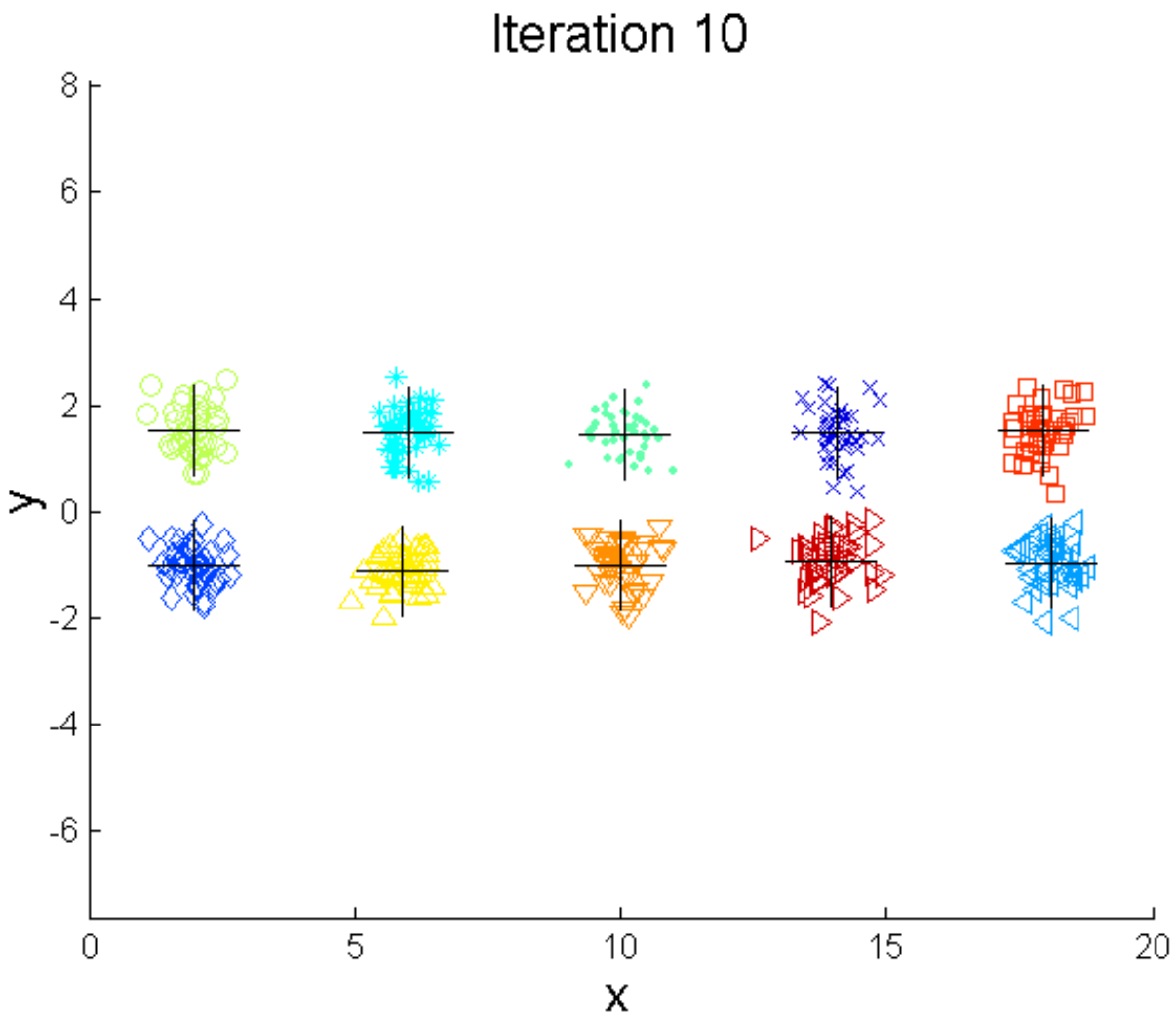
k-means με διχοτόμηση

Ποια συστάδα να διασπάσουμε;

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

Μπορεί να χρησιμοποιηθεί και ως ιεραρχικός

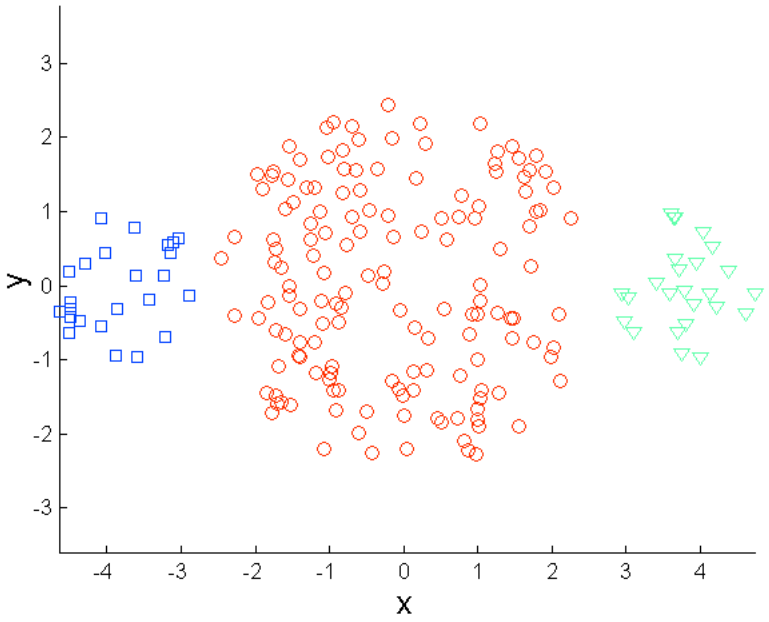
k-means με διχοτόμηση



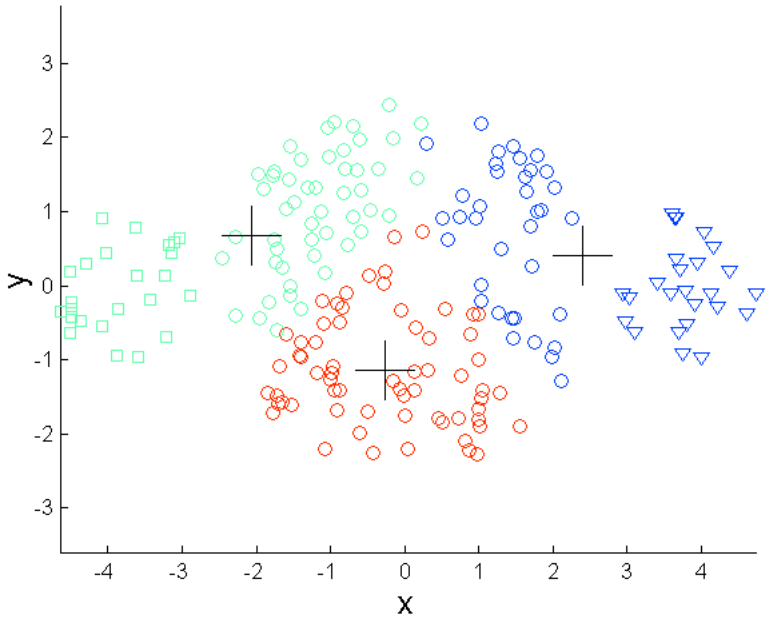
Περιορισμοί k-means

- Όταν οι συστάδες έχουν
  - διαφορετικά μεγέθη
  - διαφορετικές πυκνότητες
  - μη σφαιρικά (non-globular) σχήματα
- Όταν τα δεδομένα έχουν outliers

Περιορισμοί: διαφορετικά μεγέθη



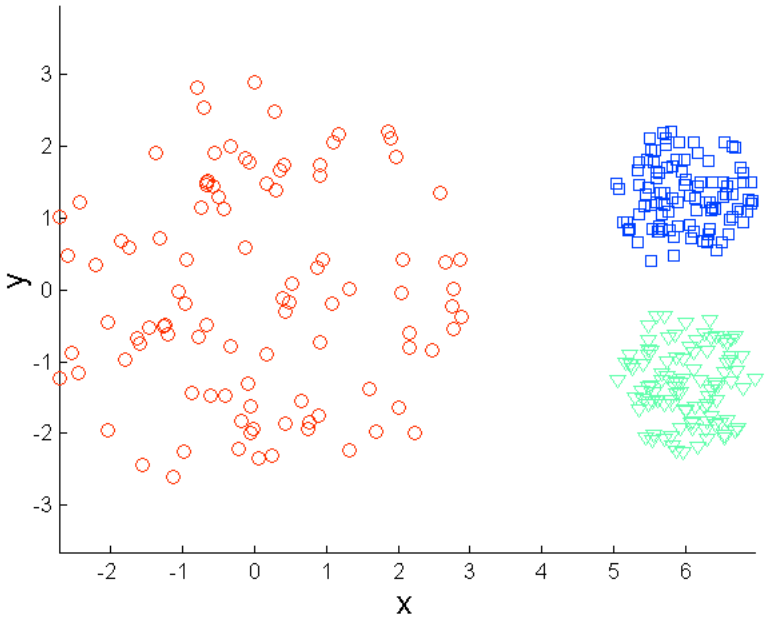
Αρχικά σημεία



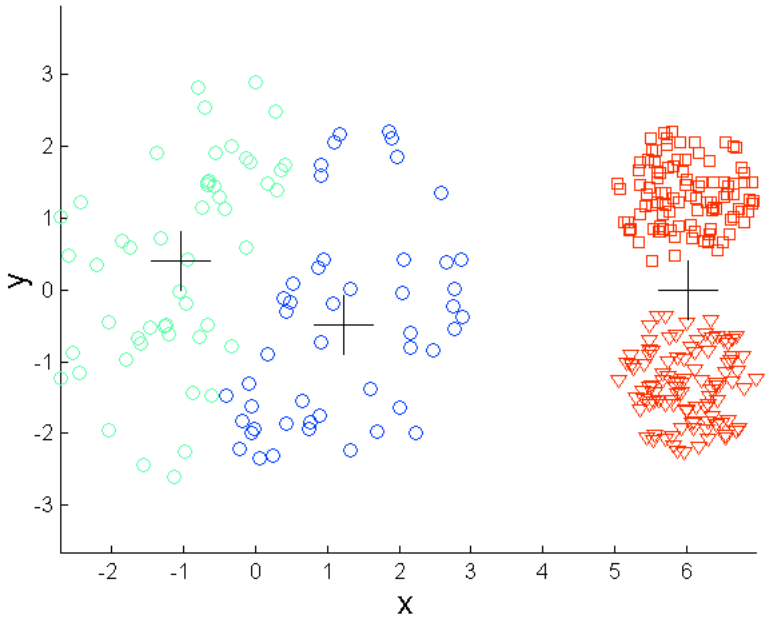
K-means (3 συστάδες)

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

Περιορισμοί: διαφορετικές πυκνότητες



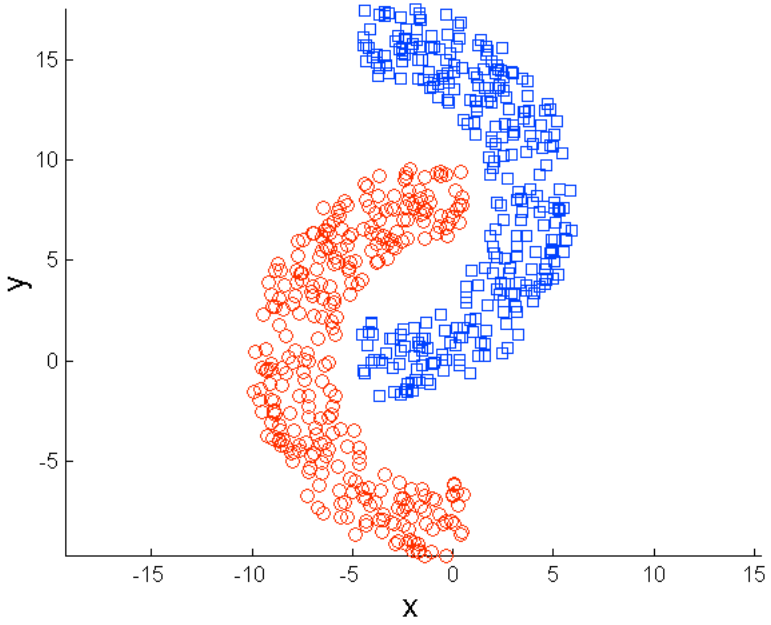
Αρχικά σημεία



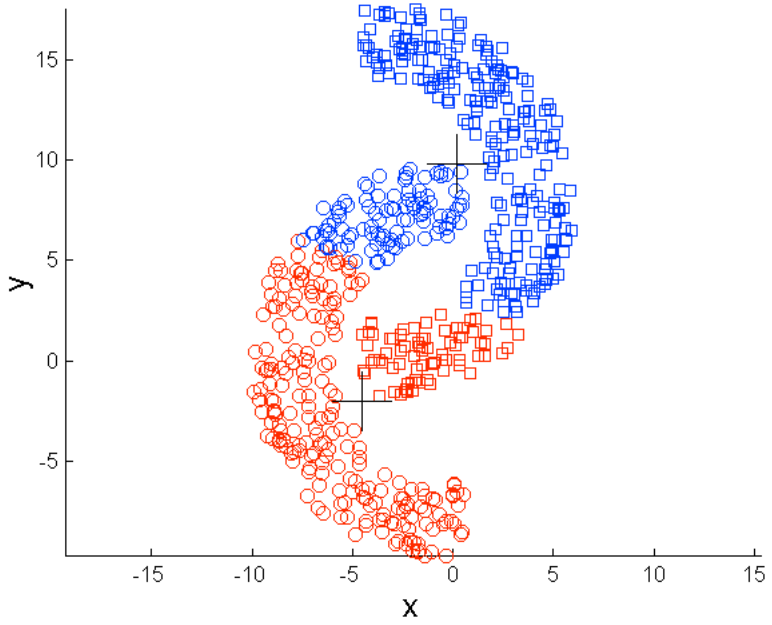
K-means (3 συστάδες)

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

Περιορισμοί: μη κυκλικά σχήματα



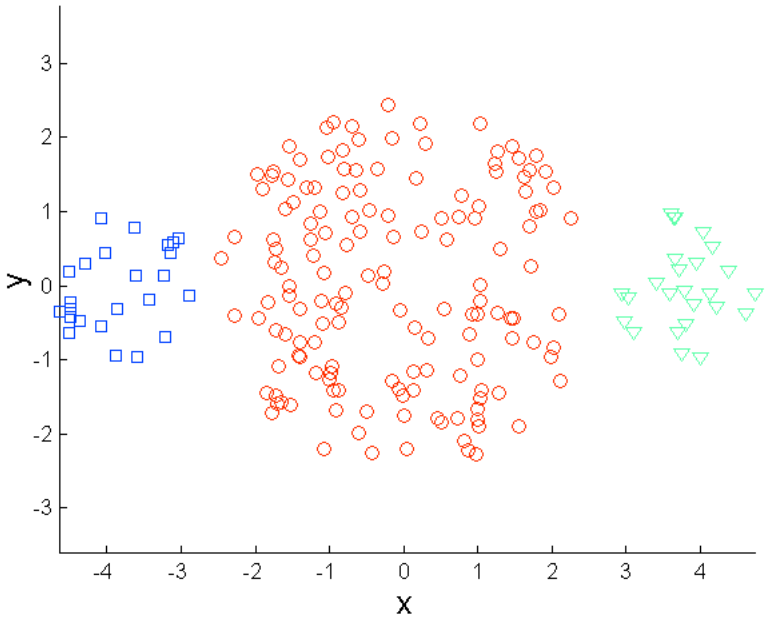
Αρχικά σημεία



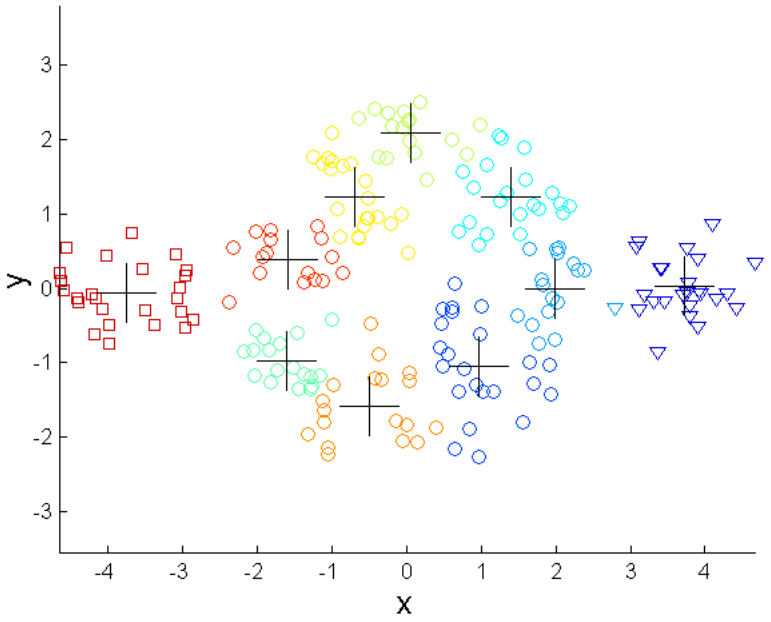
K-means (2 συστάδες)

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

Περιορισμοί k-means



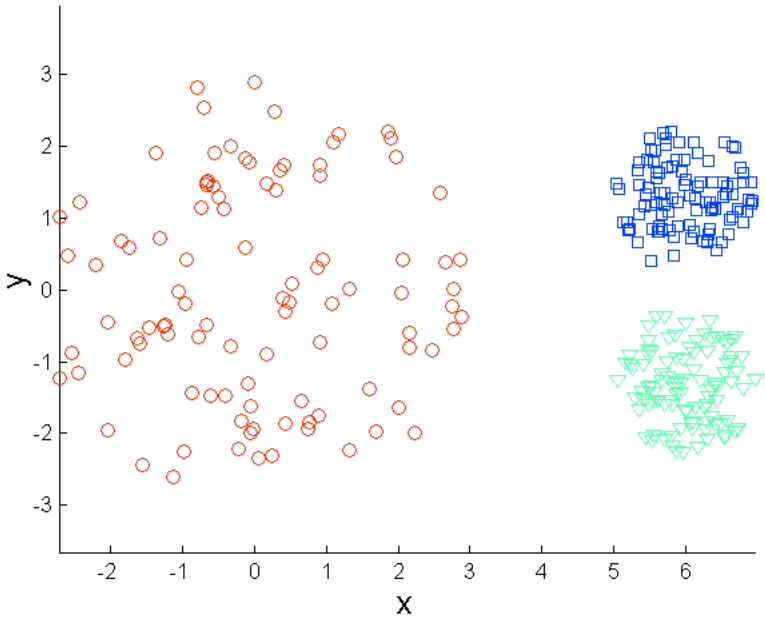
Αρχικά σημεία



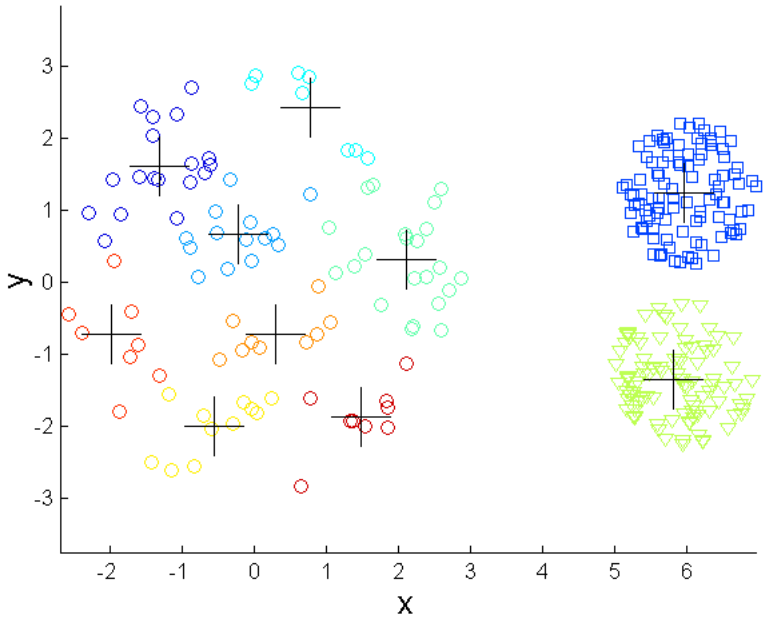
K-means συστάδες

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

Περιορισμοί k-means

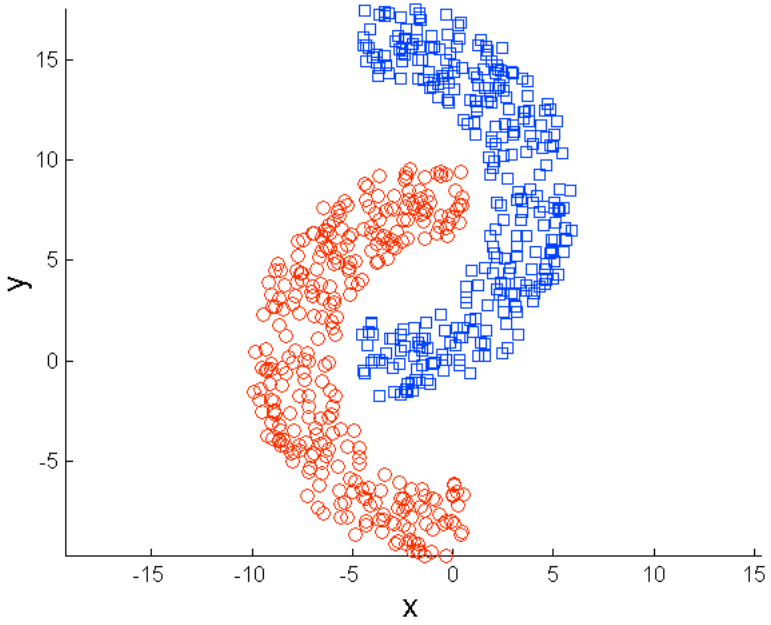


Αρχικά σημεία

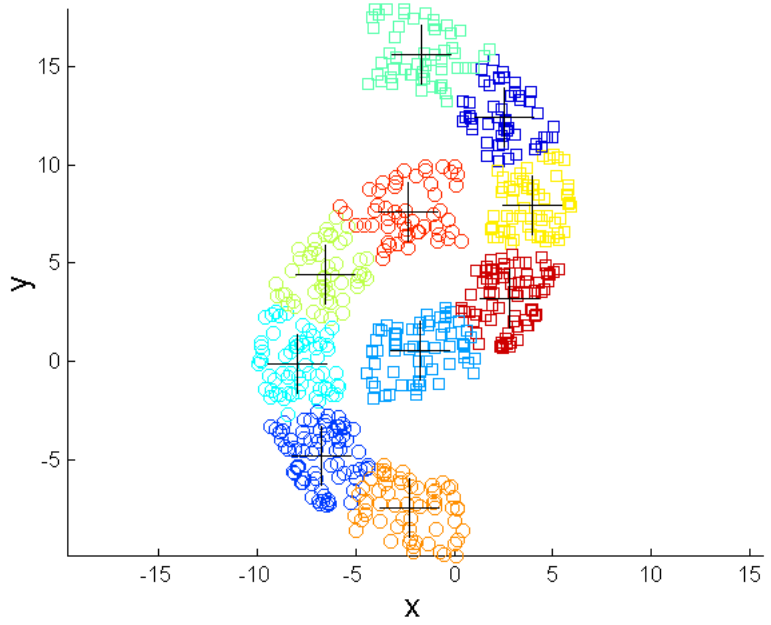


K-means συστάδες

Περιορισμοί: διαφορετικά μεγέθη



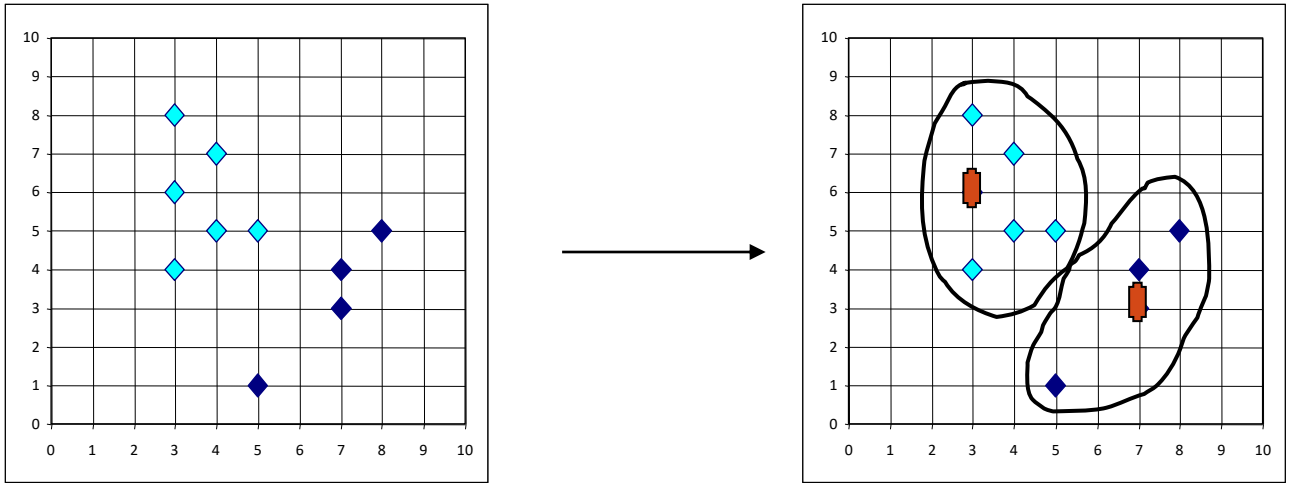
Αρχικά σημεία



K-means συστάδες

K-medoid

- Διαλέγει ένα αντιπροσωπευτικό σημείο από τα δεδομένα (σε αντίθεση με το centroid) και ελαχιστοποιεί την απόσταση από αυτό
- Μειώνει την ευαισθησία σε outliers
- Μπορεί να εφαρμοστεί σε δεδομένα οποιουδήποτε τύπου (πχ και για κατηγορικά δεδομένα)



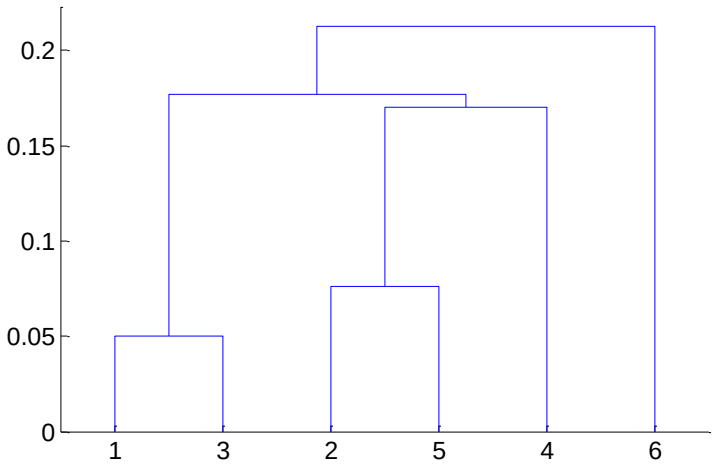
Γενικές Απαιτήσεις

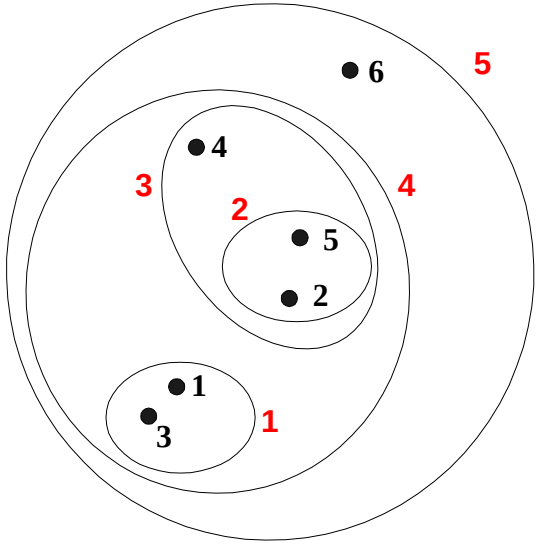
- Κλιμακωσιμότητα στον αριθμό σημείων και διαστάσεων
- Να υποστηρίζει διαφορετικούς τύπους δεδομένων
- Να υποστηρίζει συστάδες με διαφορετικά σχήματα
  - Συνήθως σφαιρικά
- Να είναι εύκολο να δώσουμε τιμές στις παραμέτρους εισόδου (π.χ. αριθμό συστάδων, μέγεθος)
- Να μην εξαρτάται από τη *σειρά* επεξεργασίας των σημείων εισόδου
- Δυναμικά μεταβαλλόμενα δεδομένα
  - Αλλαγή συστάδων με το πέρασμα του χρόνου
- Απόδοση
  - Disk-resident vs Main memory

Ιεραρχική Συσταδοποίηση

Ιεραρχική Συσταδοποίηση: Βασικά

- Παράγει ένα σύνολο από εμφωλευμένες συστάδες οργανωμένες σε ένα ιεραρχικό δέντρο
- Μπορεί να παρασταθεί με ένα δενδρόγραμμα
  - Ένα διάγραμμα που μοιάζει με δένδρο και καταγράφει τις ακολουθίες από συγχωνεύσεις (merges) και διαχωρισμούς (splits)





Πλεονεκτήματα

- Δε χρειάζεται να υποθέσουμε ένα συγκεκριμένο αριθμό από συστάδες
- Οποιοσδήποτε επιθυμητός αριθμός από συστάδες μπορεί να επιτευχθεί κόβοντας το δενδρόγραμμα στο κατάλληλο επίπεδο
- Μπορεί να αντιστοιχούν σε λογικές ταξινομήσεις
  - Π.χ. βιολογικές επιστήμες

Τύποι Ιεραρχικής Συσταδοποίησης

Δυο βασικοί τύποι ιεραρχικής συσταδοποίησης

- Συσσωρευτικός (Agglomerative)
  - Αρχίζει με τα σημεία ως ξεχωριστές συστάδες
  - Σε κάθε βήμα, συγχωνεύει το πιο κοντινό ζευγάρι συστάδων μέχρι να μείνει μόνο μία (ή k) συστάδες
- Διαιρετικός (Divisive)
  - Αρχίζει με μία συστάδα που περιέχει όλα τα σημεία
  - Σε κάθε βήμα, διαχωρίζει μία συστάδα, έως κάθε συστάδα να περιέχει μόνο ένα σημείο (ή να δημιουργηθούν k συστάδες)
- Οι παραδοσιακοί αλγόριθμοι χρησιμοποιούν έναν πίνακα ομοιότητα ή απόστασης
  - διαχωρισμός ή συγχώνευση μιας ομάδας τη φορά

Αλγόριθμος Συσσωρευτικής Ιεραρχικής Συσταδοποίησης (ΣΙΣ)

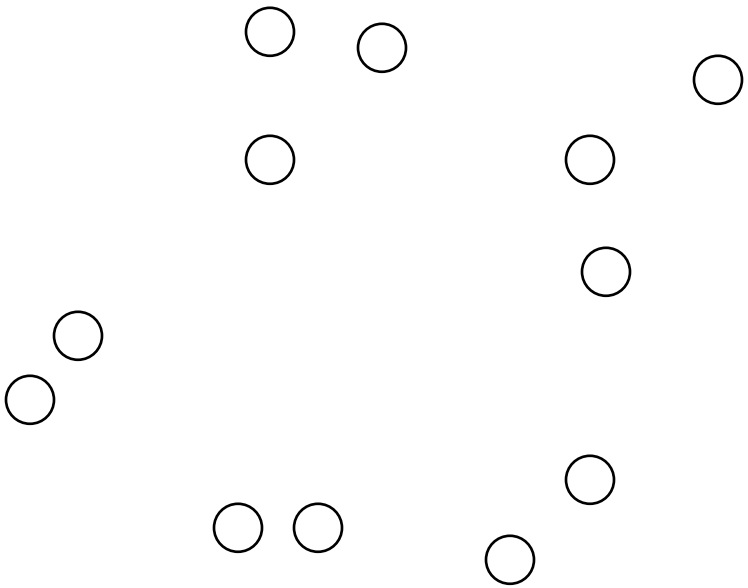
Η πιο δημοφιλής τεχνική συσταδοποίησης

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

- Βασική λειτουργία είναι ο υπολογισμός της γειτνίασης δυο συστάδων
- Διαφορετικοί αλγόριθμοι με βάση το πώς ορίζεται η απόσταση ανάμεσα σε δύο συστάδες

Βήματα ΣΙΣ

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

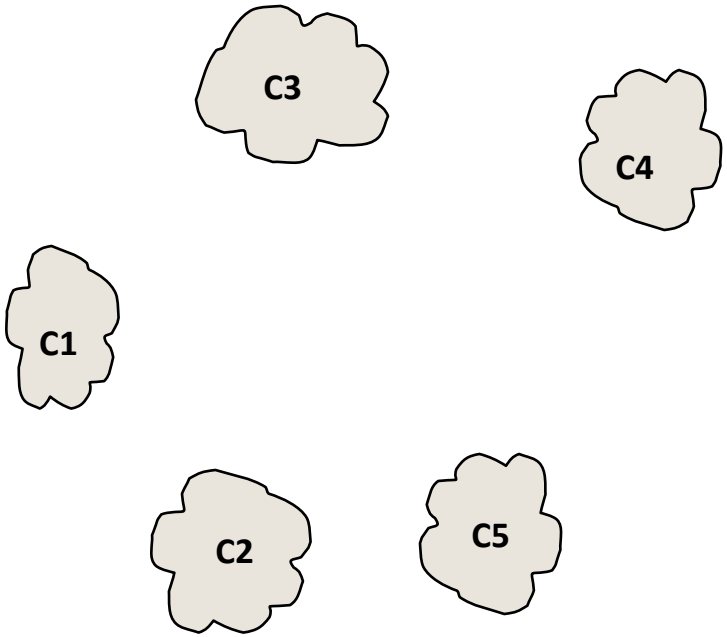


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Πίνακας Γειτνίασης

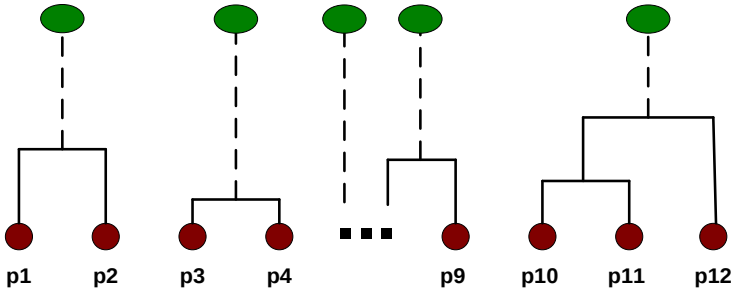
Βήματα ΣΙΣ

Μετά από κάποιες συγχωνεύσεις, έχουμε κάποιες συστάδες



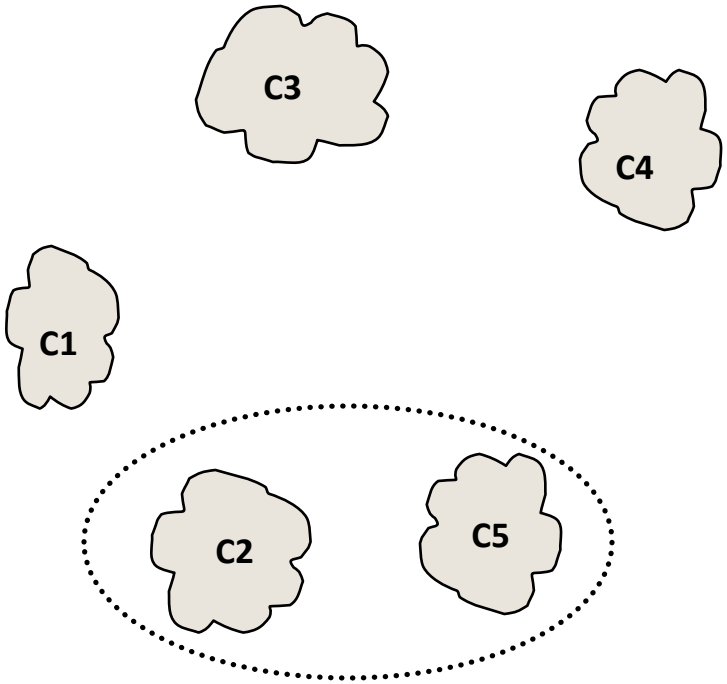
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Πίνακας Γειτνίασης



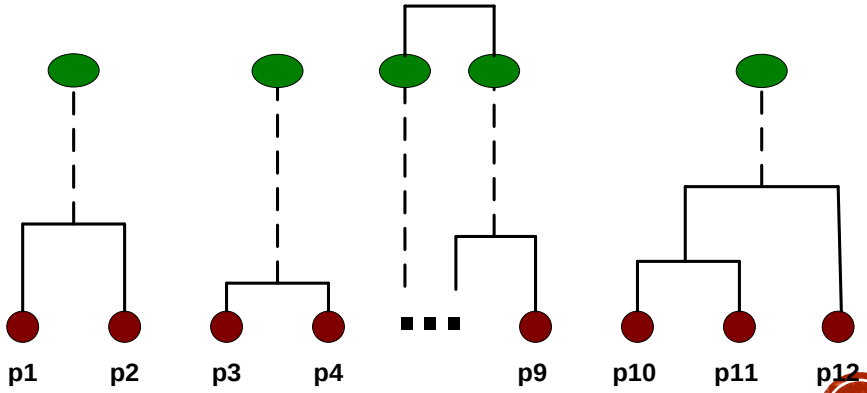
Βήματα ΣΙΣ

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



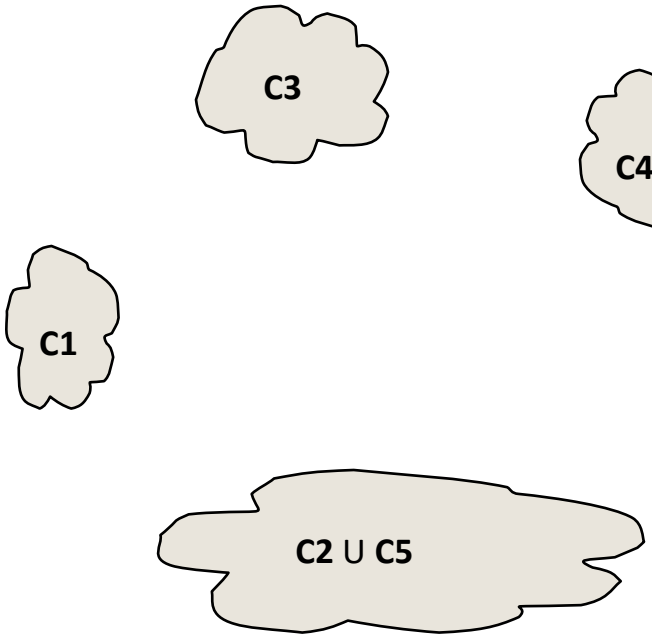
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Πίνακας Γειτνίασης



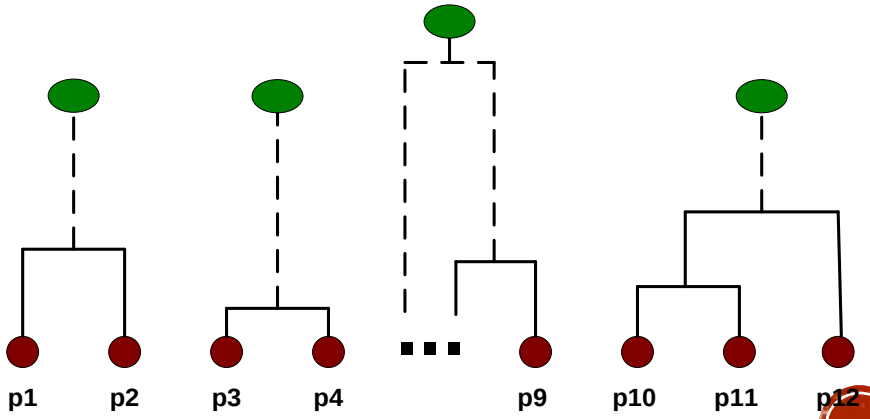
Βήματα ΣΙΣ

Μετά τη συγχώνευση η ερώτηση είναι: Πώς ενημερώνουμε τον πίνακα γειτνίασης;

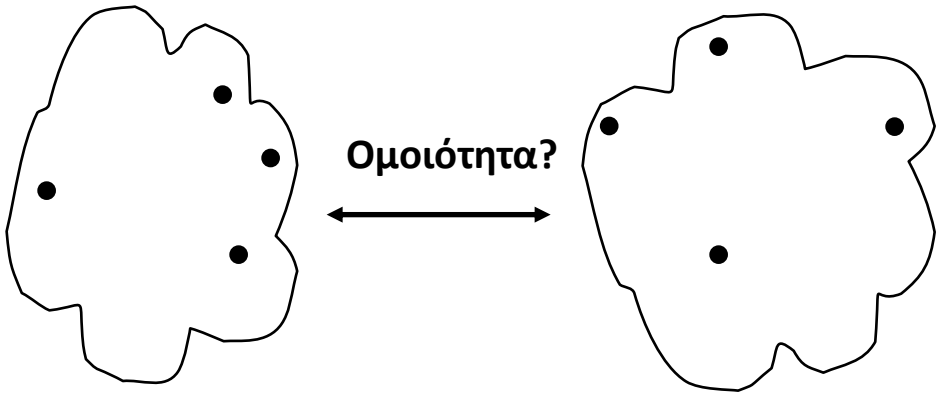


	C1	C2 ∪ C5	C3	C4
C1		?		
C2 ∪ C5	?	?	?	?
C3		?		
C4		?		

Πίνακας Γειτνίασης



ΣΙΣ: Ορισμός απόστασης μεταξύ συστάδων

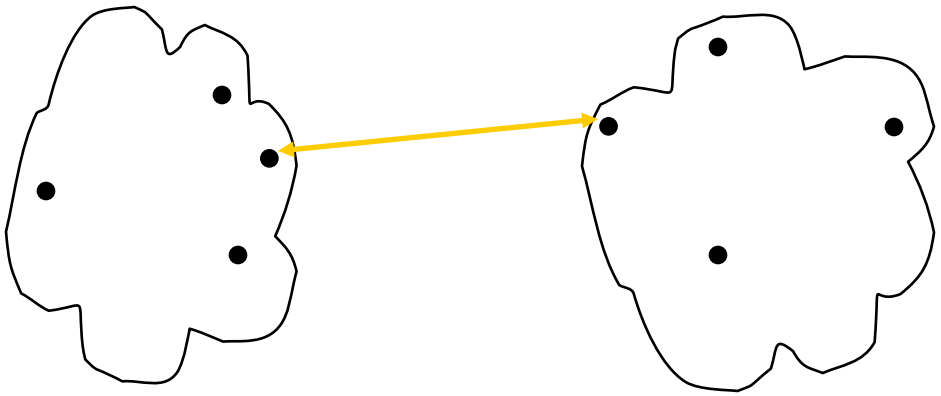


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Πίνακας Γειτνίασης

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

ΣΙΣ MIN



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

. .

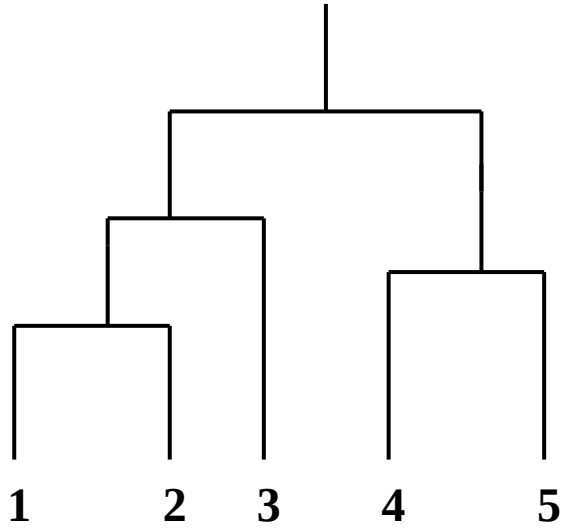
Πίνακας Γειτνίασης

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

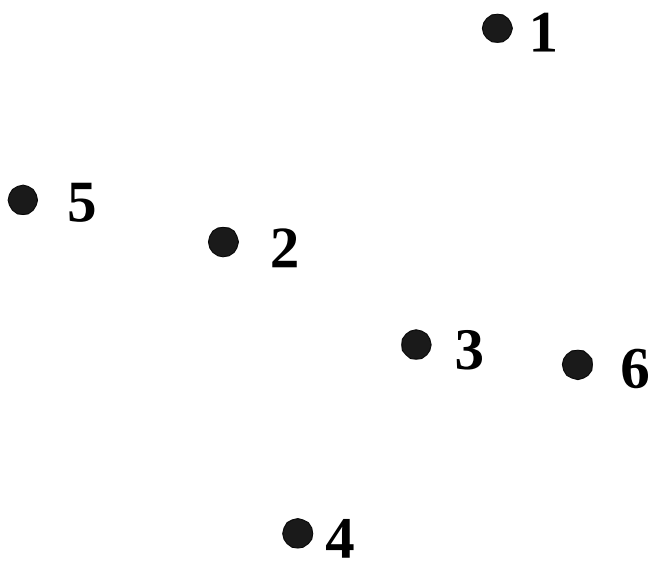
ΣΙΣ MIN

- 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

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



ΣΙΣ MIN Παράδειγμα



1 (0.4, 0.53)

2 (0.22, 0.38)

3 (0.35, 0.32)

4 (0.26, 0.19)

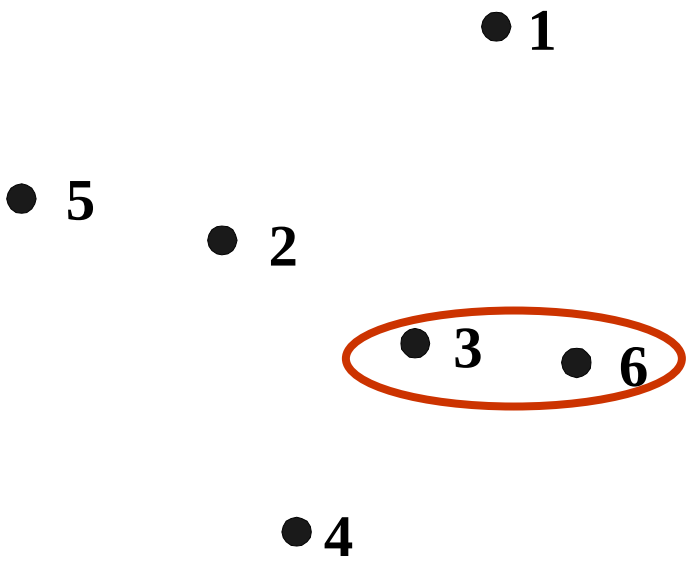
5 (0.08, 0.41)

6 (0.45, 0.30)

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

Πίνακας απόστασης (Ευκλείδεια)

ΣΙΣ MIN Παράδειγμα



1 (0.4, 0.53)

2 (0.22, 0.38)

3 (0.35, 0.32)

4 (0.26, 0.19)

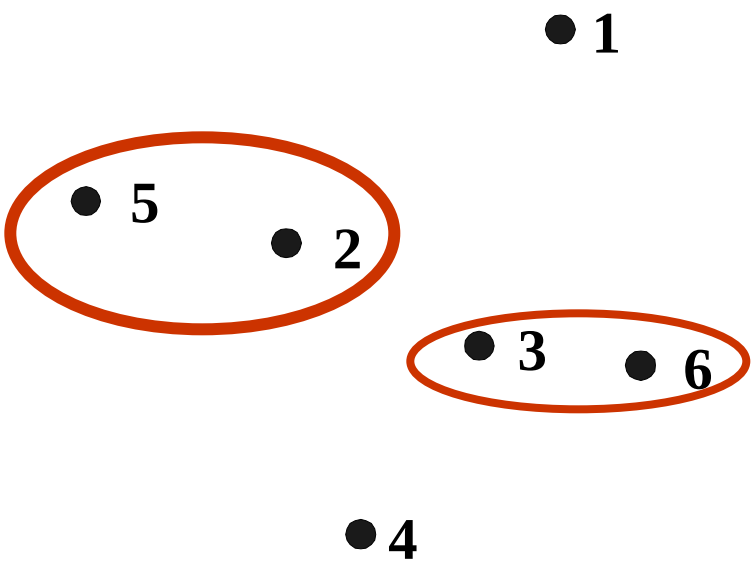
5 (0.08, 0.41)

6 (0.45, 0.30)

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

Καθορίζεται μόνο από μια ακμή – την μικρότερη

ΣΙΣ MIN Παράδειγμα



1 (0.4, 0.53)

2 (0.22, 0.38)

3 (0.35, 0.32)

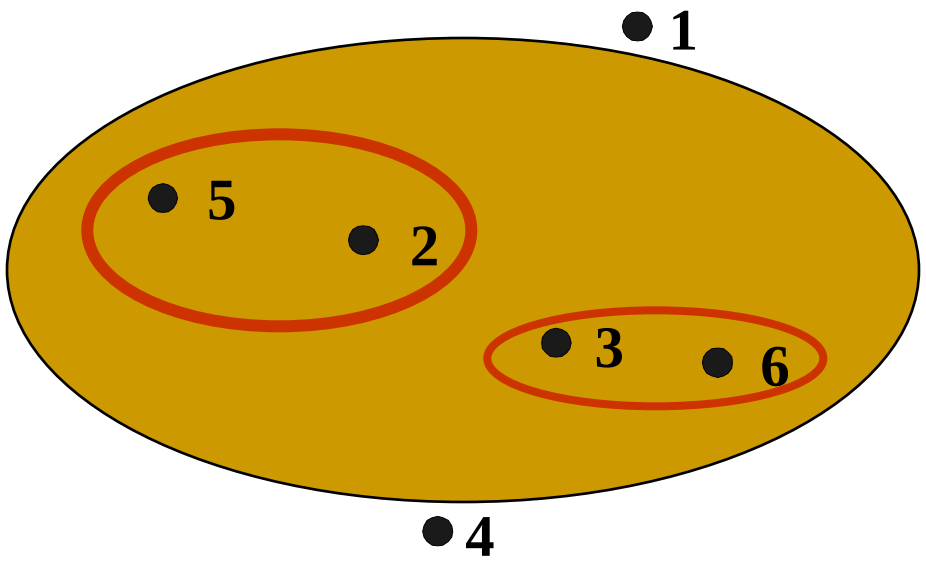
4 (0.26, 0.19)

5 (0.08, 0.41)

6 (0.45, 0.30)

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

ΣΙΣ MIN Παράδειγμα



1 (0.4, 0.53)

2 (0.22, 0.38)

3 (0.35, 0.32)

4 (0.26, 0.19)

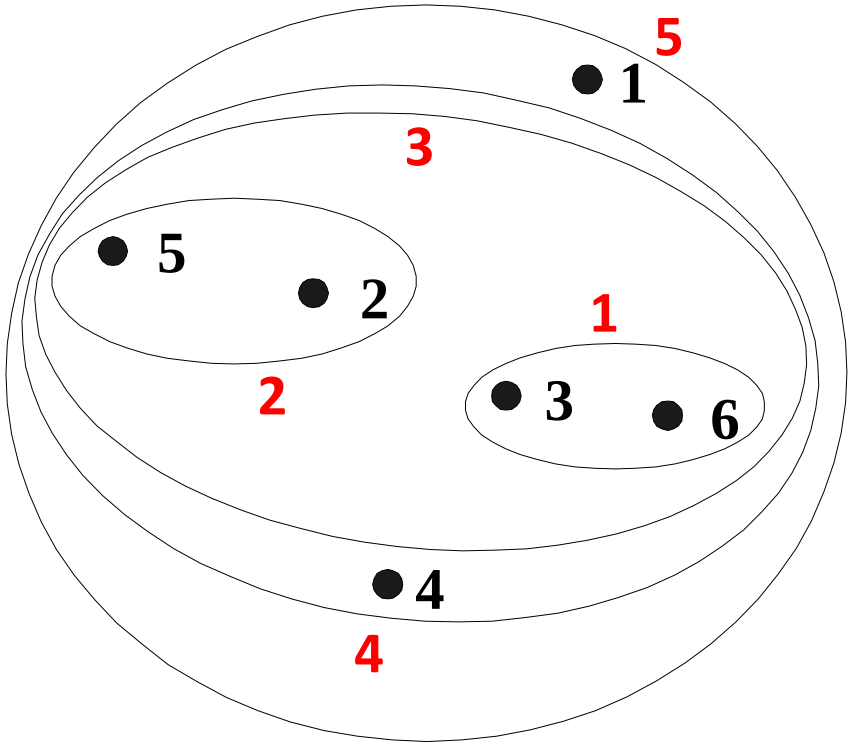
5 (0.08, 0.41)

6 (0.45, 0.30)

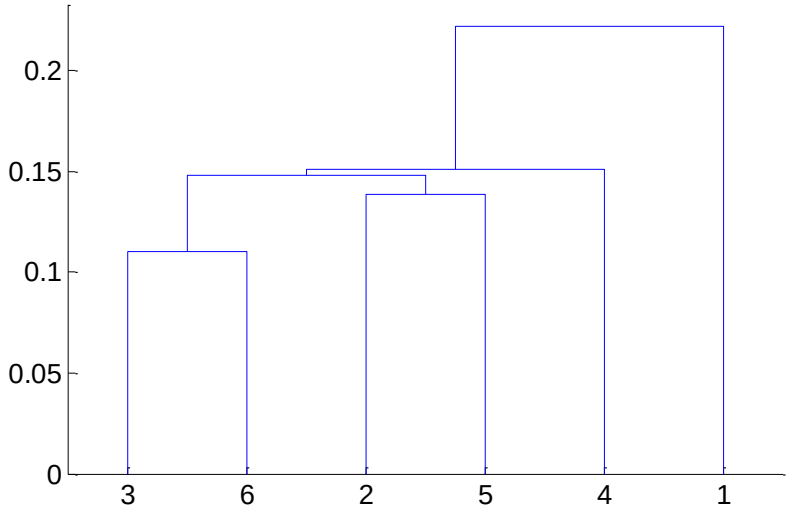
Αρκεί να «δω» μια ακμή

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

ΣΙΣ MIN Παράδειγμα



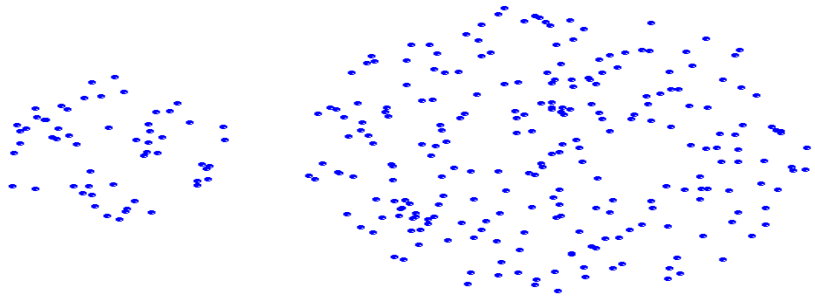
Εμφωλευμένες Συστάδες



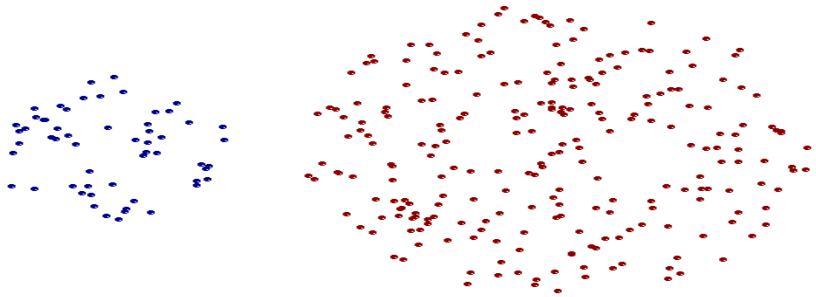
Δεντρόγραμμα

Το δεντρόγραμμα (y-άξονας) δίνει και τις αποστάσεις

ΣΙΣ ΜΙΝ Πλεονεκτήματα



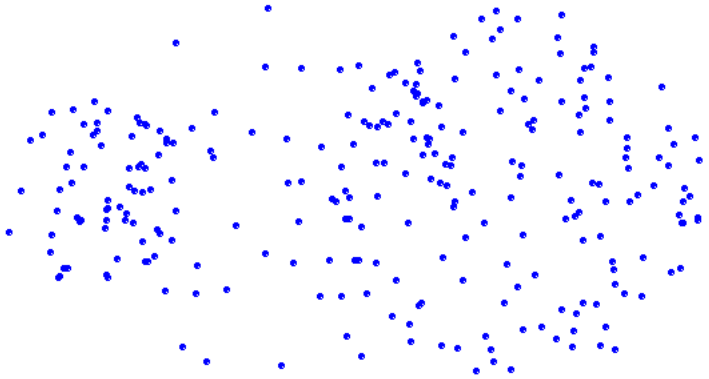
Αρχικά σημεία



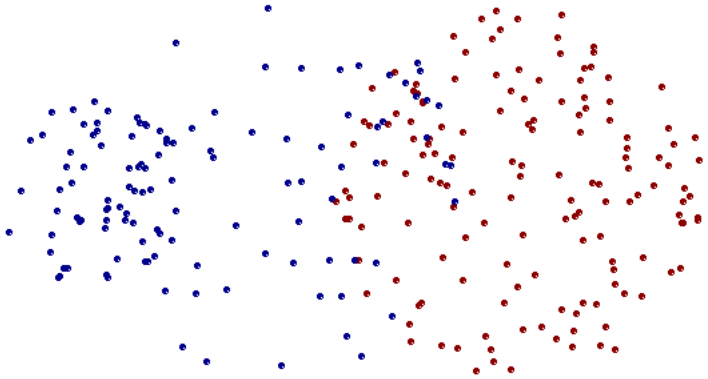
Δύο συστάδες

- Contiguity-based (συνεχόμενες συστάδες)
- Μπορεί να χειριστεί μη ελλειπτικά (non-elliptical) σχήματα

ΣΙΣ MIN Μειονεκτήματα



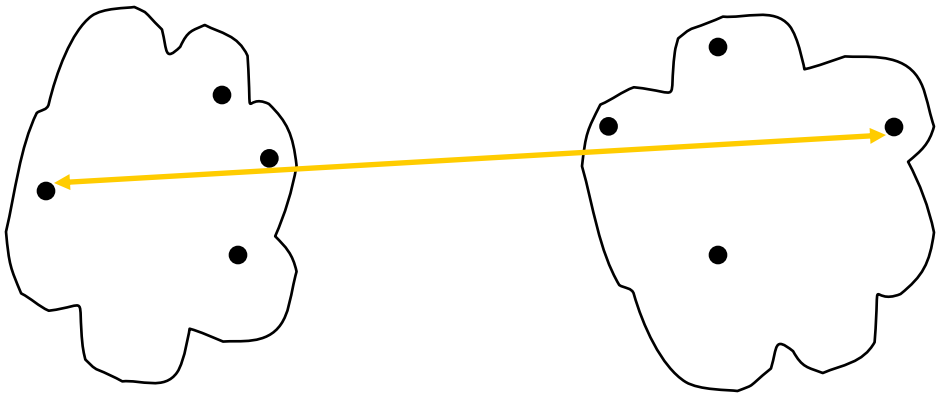
Αρχικά σημεία



Δύο συστάδες

- Ευαίσθητο σε θόρυβο και outliers!

ΣΙΣ MAX

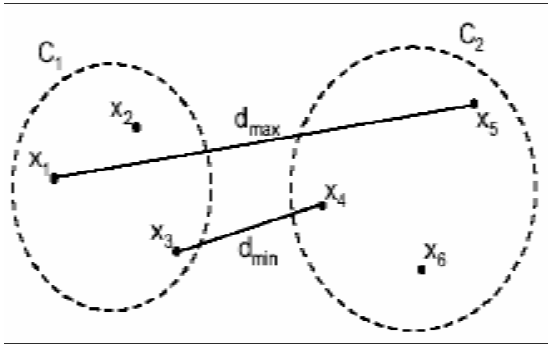


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						

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

· Πίνακας Γειτνίασης

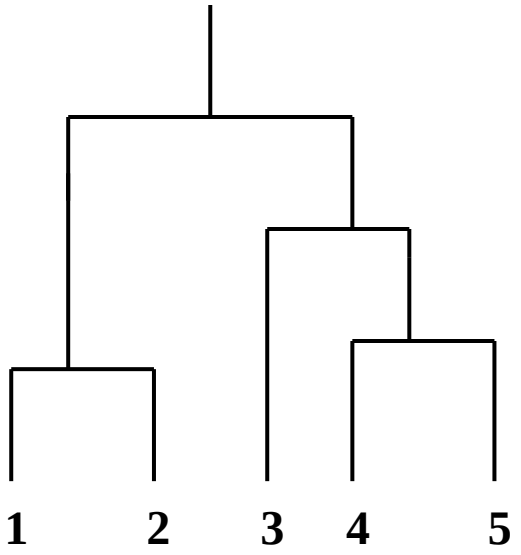
·



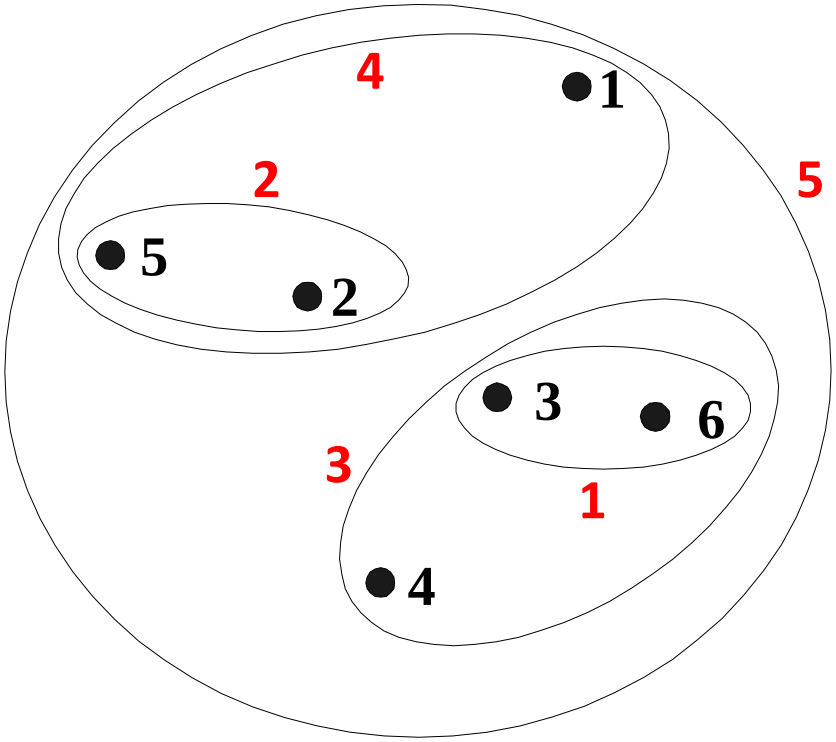
ΣΙΣ MAX

- MAX ή πλήρους συνδεσιμότητας (complete linkage)
  - Αναζητά κλίκες
- Η ομοιότητα μεταξύ δυο συστάδων βασίζεται στα δυο λιγότερο όμοια (πιο μακρινά) σημεία στις διαφορετικές συστάδες (longest edge), δηλαδή οι συστάδες με την μικρότερη τέτοια απόσταση
- Καθορίζεται από όλα τα ζεύγη τιμών στις δύο συστάδες

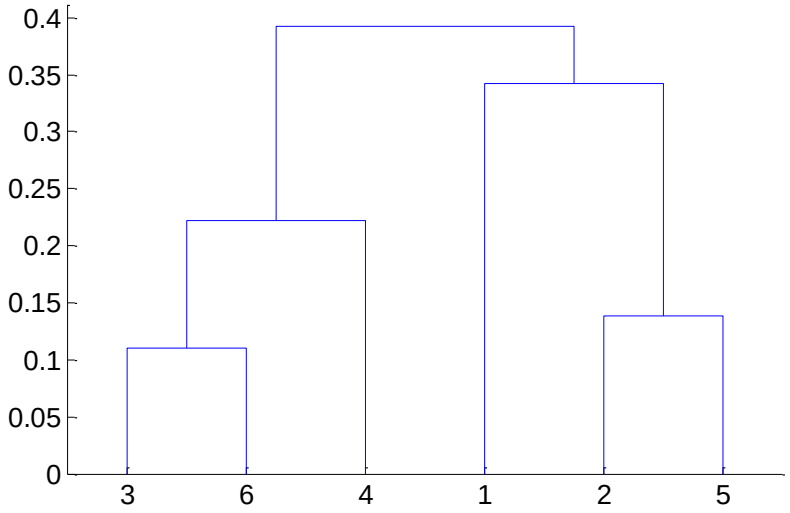
	I1	I2	I3	I4	I5
I1	1.00	0.90	0.10	0.65	0.20
I2	0.90	1.00	0.70	0.60	0.50
I3	0.10	0.70	1.00	0.40	0.30
I4	0.65	0.60	0.40	1.00	0.80
I5	0.20	0.50	0.30	0.80	1.00



ΣΙΣ MAX Παράδειγμα

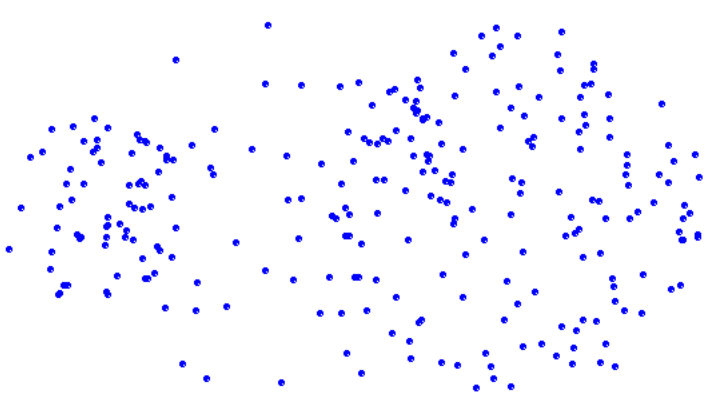


Εμφωλευμένες Συστάδες

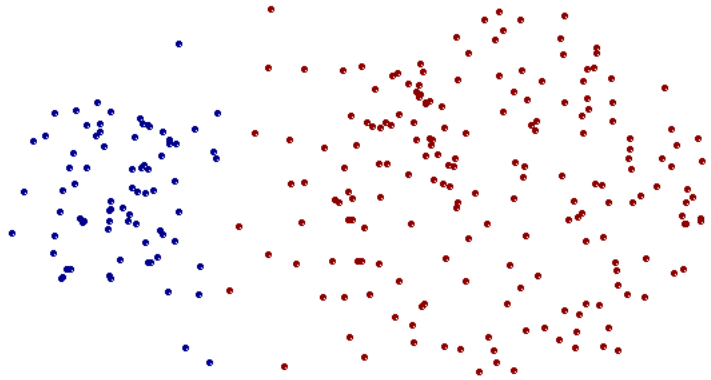


Δεντρόγραμμα

ΣΙΣ MAX Πλεονεκτήματα



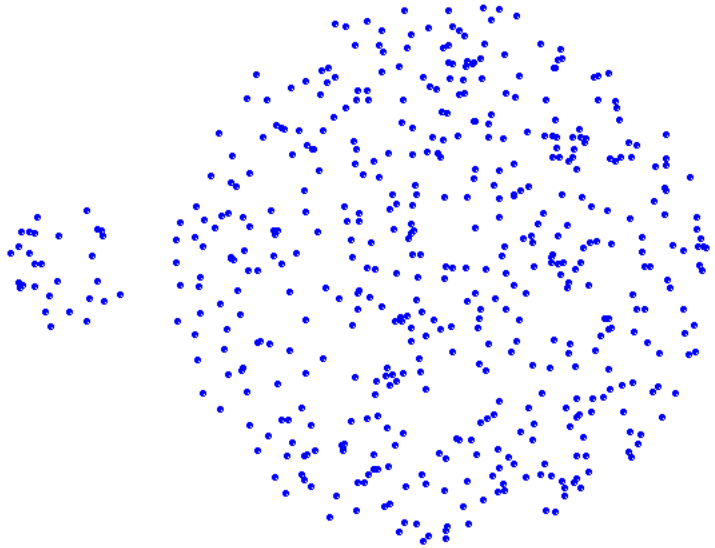
Αρχικά Σημεία



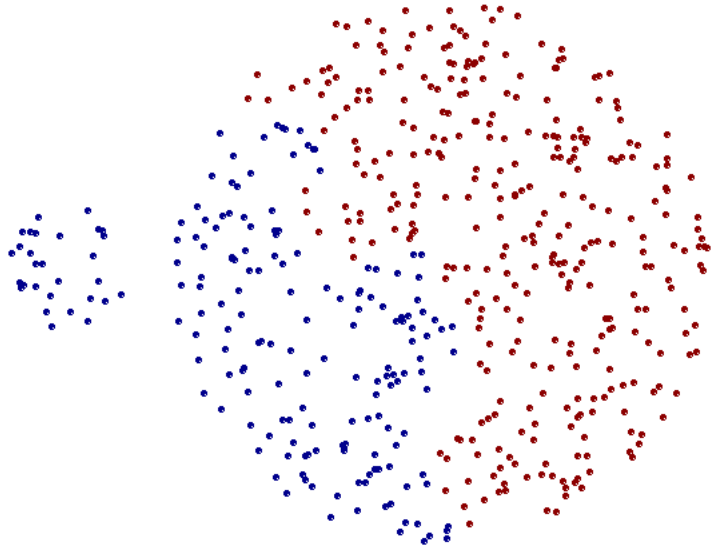
Δύο Συστάδες

- Λιγότερη εξάρτηση σε θόρυβο και outliers

ΣΙΣ ΜΑΧ Πλεονεκτήματα



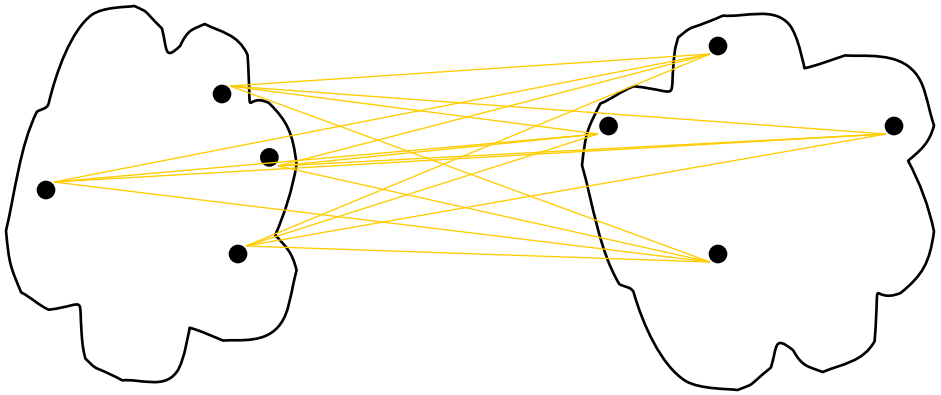
Αρχικά σημεία



Δύο συστάδες

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

ΣΙΣ Μέσος Όρος Ομάδας



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

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

- Πίνακας Γειτνίασης
-

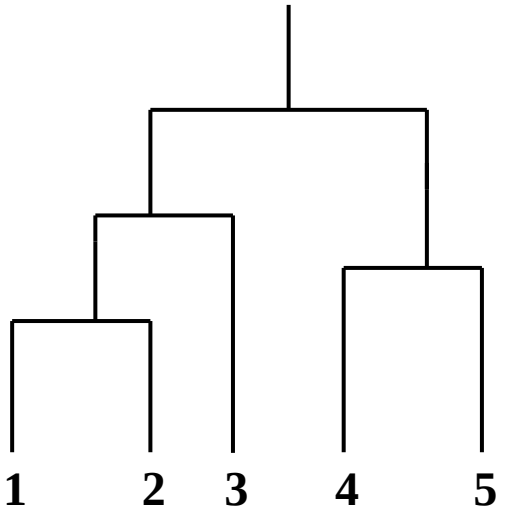
ΣΙΣ Μέσος Όρος Ομάδας

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

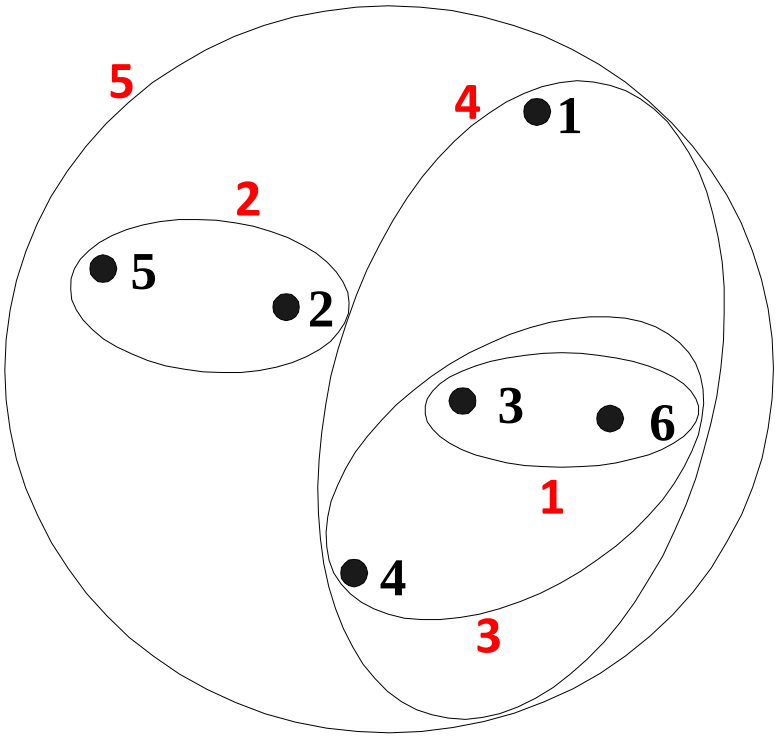
$$\text{proximity}(\text{Cluster}_i, \text{Cluster}_j) = \frac{\sum_{\substack{p_i \in \text{Cluster}_i \\ p_j \in \text{Cluster}_j}} \text{proximity}(p_i, p_j)}{|\text{Cluster}_i| * |\text{Cluster}_j|}$$

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

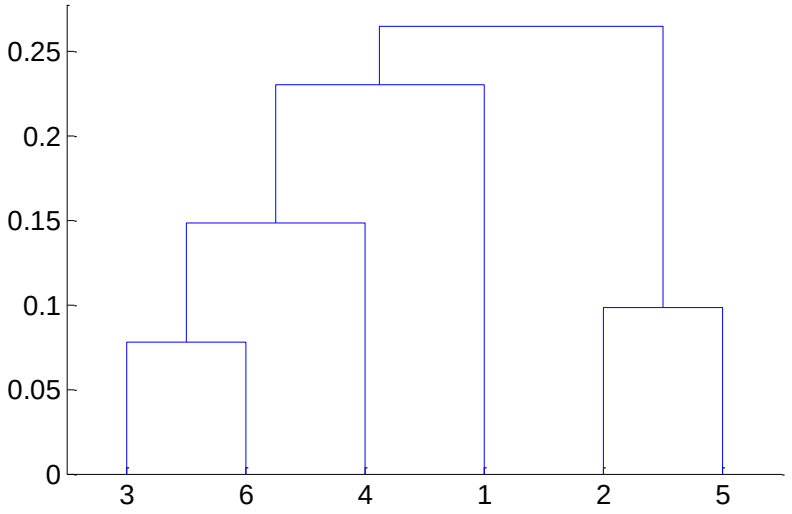
	I1	I2	I3	I4	I5
I1	1.00	0.90	0.10	0.65	0.20
I2	0.90	1.00	0.70	0.60	0.50
I3	0.10	0.70	1.00	0.40	0.30
I4	0.65	0.60	0.40	1.00	0.80
I5	0.20	0.50	0.30	0.80	1.00



ΣΙΣ Μέσος Όρος Ομάδας



Φωλιασμένες Συστάδες

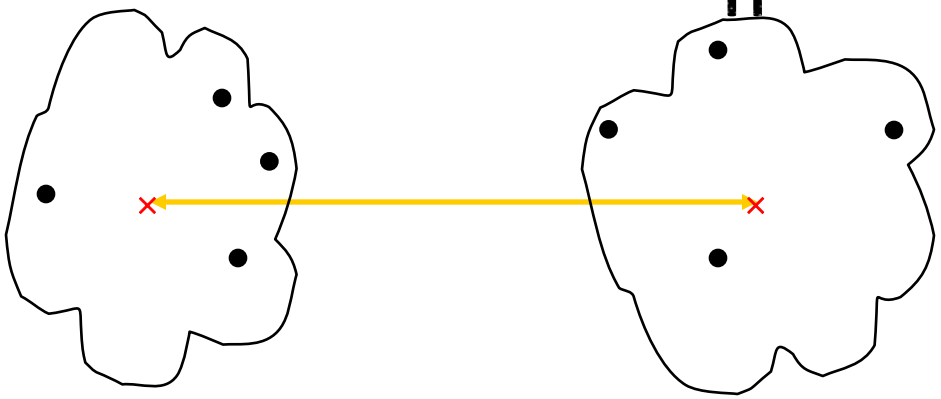


Δενδρόγραμμα

ΣΙΣ Μέσος Όρος Ομάδας

- Ανάμεσα σε MIN-MAX
- Πλεονεκτήματα: μικρότερη ευαισθησία σε θόρυβο και outliers
- Μειονεκτήματα: Ευνοεί κυκλικές συστάδες

ΣΙΣ Απόσταση Μεταξύ Κεντρικών Σημείων



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

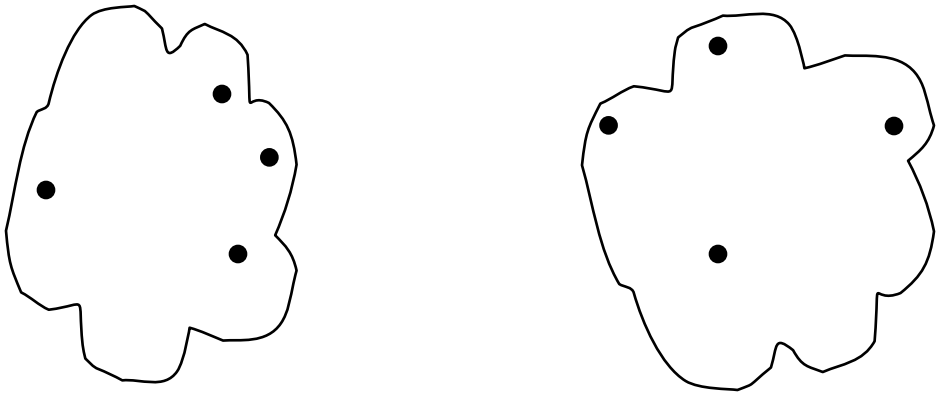
- MIN
- MAX
- Μέσος όρος της ομάδας
- **Η απόσταση μεταξύ των κεντρικών σημείων**
- Άλλες μέθοδοι βασισμένες σε μια αντικειμενική συνάρτηση
  - Η μέθοδος του Ward χρησιμοποιεί τετραγωνικά λάθη

· Πίνακας Γειτνίασης

·

- Πρόβλημα: μη μονότονη αύξηση της απόστασης
  - Δυο συστάδες που συγχωνεύονται μπορεί να έχουν μικρότερη απόσταση από συστάδες που έχουν συγχωνευτεί σε προηγούμενα βήματα

ΣΙΣ Μέθοδος του Ward



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

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

· Πίνακας Γειτνίασης

·

ΣΙΣ Μέθοδος του Ward

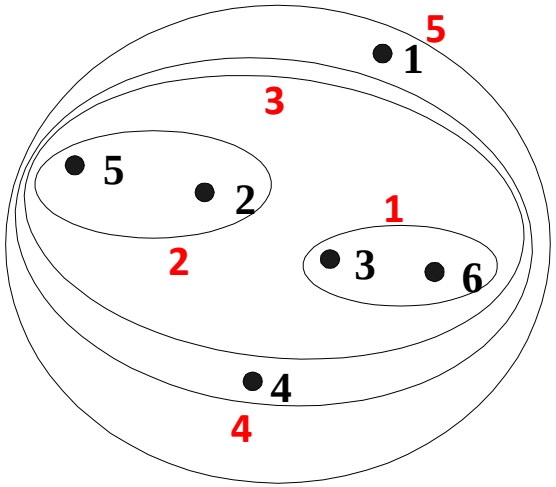
- Βασισμένο στην αύξηση του SSE όταν συγχωνεύονται οι δύο συστάδες
- Ιεραρχικό ανάλογο του k-means
- Μπορεί να χρησιμοποιηθεί για την αρχικοποίηση του k-means

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

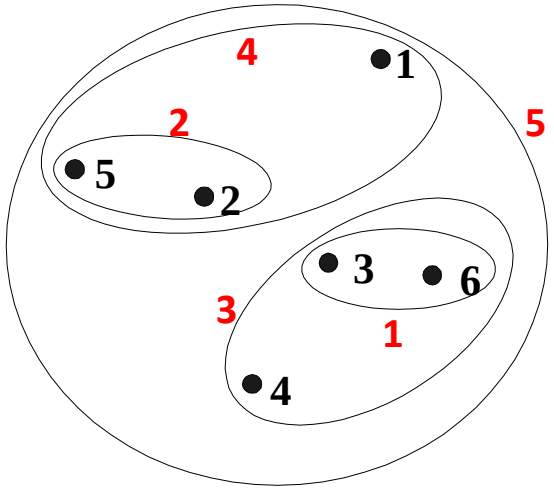
Ward απόσταση μεταξύ συστάδων C_i and C_j είναι η διαφορά μεταξύ του ολικού λάθους των δύο συστάδων και του ολκού λάθους αν τις ενώσουμε σε μία συστάδα C_{ij}

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

ΣΙΣ: Σύγκριση απόστασης μεταξύ συστάδων

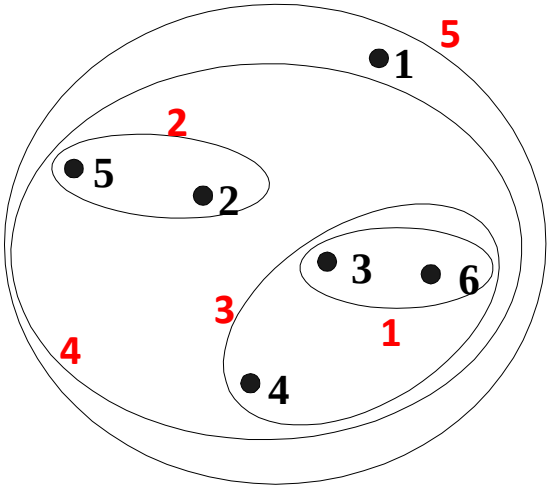


MIN

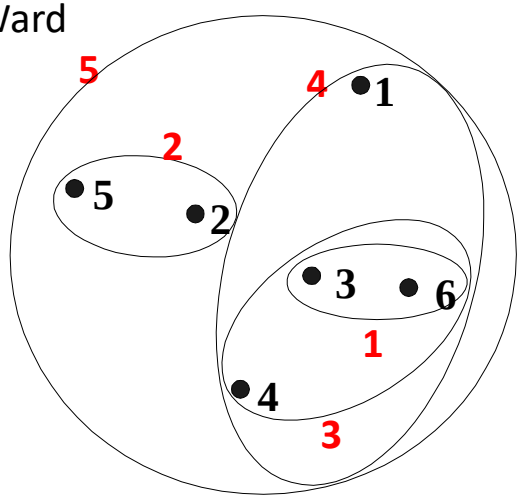


MAX

Μέθοδος του Ward



Μέσο Ομάδας



ΣΙΣ: Πολυπλοκότητα Χρόνου και Χώρου

- $O(m^2)$ χώρος για την αποθήκευση του πίνακα γειτνίασης
  - m αριθμός σημείων
- $O(m^3)$
  - Ξεκινάμε με m συστάδες και μειώνουμε 1 τη φορά
  - Αν γραμμική αναζήτηση του πίνακα $O(m^2)$
  - Καλύτερος χρόνος αν διατηρούμε κάποια ταξινόμηση των αποστάσεων (π.χ. heap)

ΣΙΣ: Περιορισμοί και Προβλήματα

- Οι αποφάσεις είναι τελικές
  - Μόλις δυο συστάδες συγχωνευτούν αυτό δεν μπορεί να αλλάξει
- Δεν ελαχιστοποιούν άμεσα κάποια αντικειμενική συνάρτηση

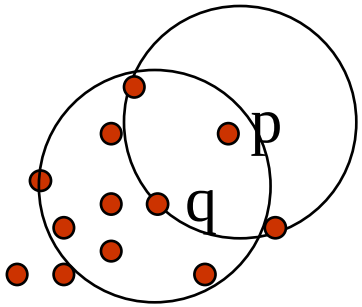
Συσταδοποίηση με βάση την Πυκνότητα (DBSCAN)

Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

DBSCAN Εισαγωγή

- Αλγόριθμος βασισμένος στην πυκνότητα
  - Πυκνότητα για ένα σημείο = αριθμός σημείων (**MinPts**) μέσα σε μια προκαθορισμένη ακτίνα (**Eps**) από αυτό (συμπεριλαμβανομένου του σημείου)
- Δύο παράμετροι:
  - Eps: Μέγιστη ακτίνα της γειτονιάς
  - MinPts: Ελάχιστος αριθμός σημείων στην Eps-γειτονιά ενός σημείου
- Γειτονιά ενός σημείου p = όλα τα σημεία σε απόσταση Eps από το p:
  NEps(p) = {q | dist(p,q) <= Eps }

Για το p έχουμε 4
Για το q έχουμε >5

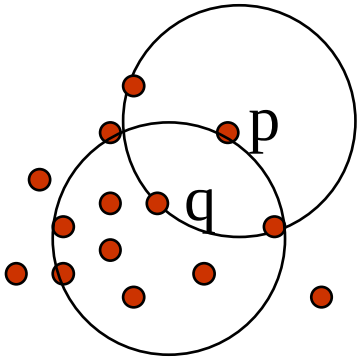


MinPts = 5
e = 1 cm

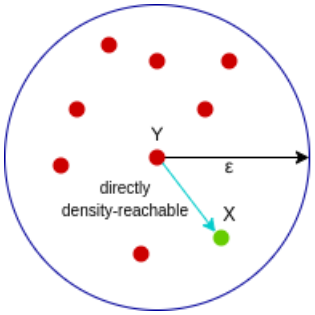
DBSCAN Εισαγωγή

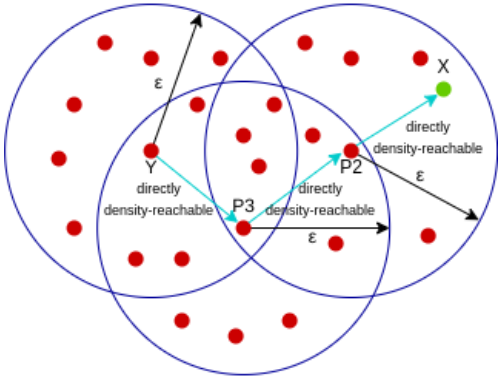
Τα σημεία διαχωρίζονται σε:

- Βασικά (core) - σημεία πυρήνα: ένα σημείο για το οποίο υπάρχουν περισσότερα από ένα προκαθορισμένο αριθμό (***MinPts***) σημεία σε ακτίνα ***Eps***
  - Αυτά είναι τα σημεία που είναι στο εσωτερικό μιας συστάδας (ομάδας πυκνών σημείων)
- Οριακά (border) - σημεία ορίου: ένα σημείο για το οποίο υπάρχουν λιγότερα από ένα προκαθορισμένο αριθμό (MinPts) σημεία σε ακτίνα Eps, αλλά είναι στη γειτονιά (τουλάχιστον) ενός βασικού σημείου
- Θορύβου (noise): ένα σημείο που δεν είναι ούτε σημείο πυρήνα ούτε σημείο ορίου

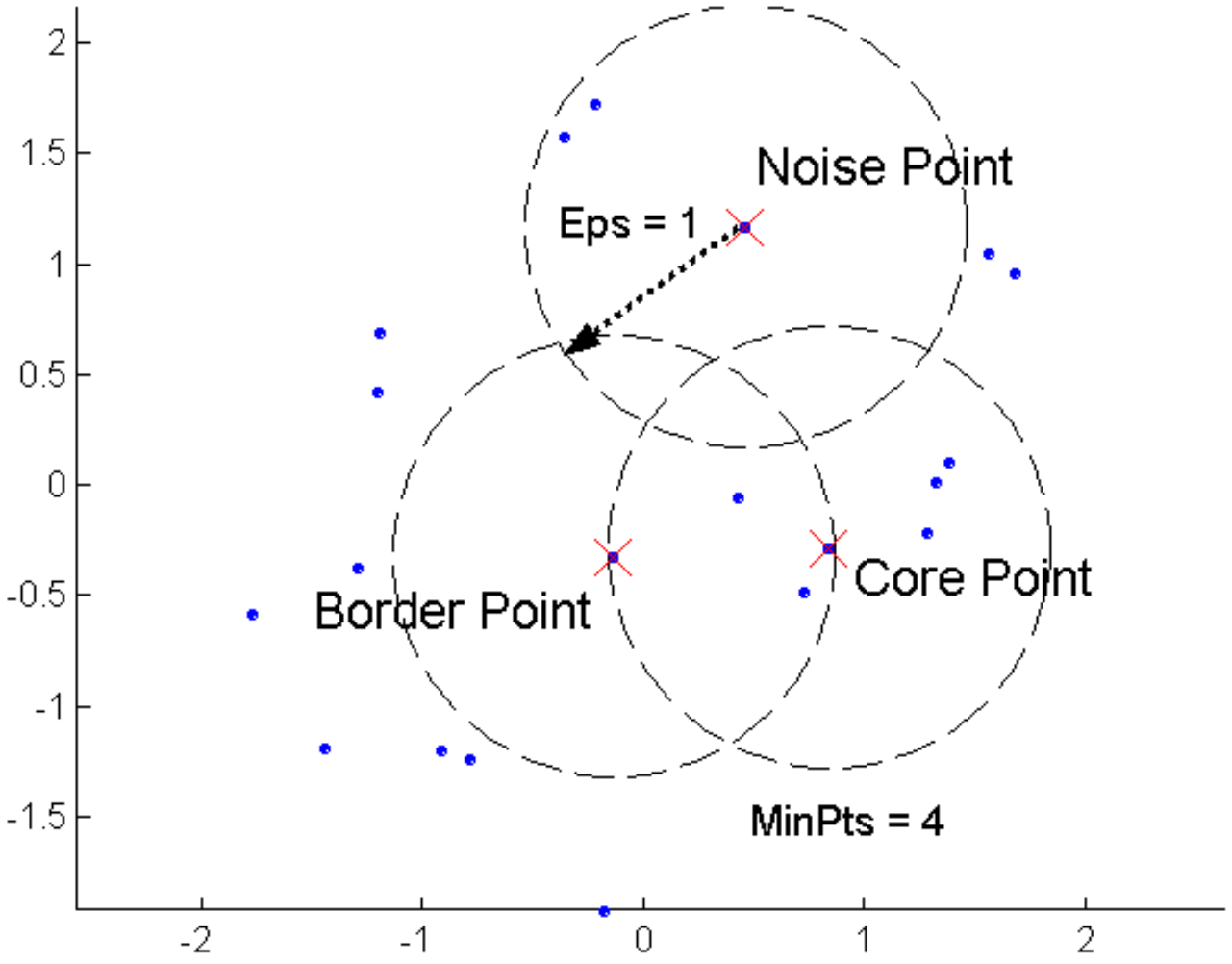


MinPts = 5
e = 1 cm





DBSCAN Εισαγωγή



DBSCAN Αλγόριθμος

1: Χαρακτήρισε κάθε σημείο ως **πυρήνα**, **ορίου** ή **θορύβου**

2: Διέγραψε τα σημεία θορύβου

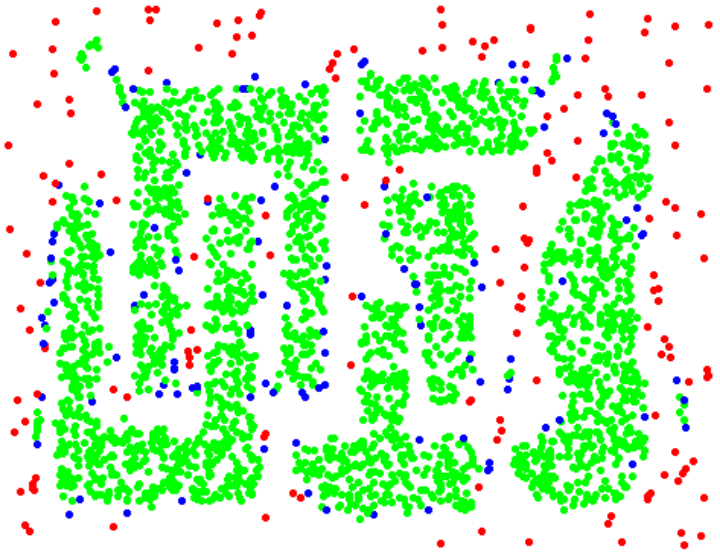
3: Τοποθέτησε μια ακμή μεταξύ όλων των σημείων πυρήνα που είναι σε απόσταση έως *Eps* μεταξύ τους

4: Κάνε κάθε ομάδα συνδεδεμένων σημείων πυρήνα μια διαφορετική συστάδα

5: Ανάθεσε κάθε σημείο ορίου σε μία από τις συστάδες των συσχετιζόμενων του σημείων πυρήνα

DBSCAN Αλγόριθμος

Βήμα 1 και 2



Αρχικά σημεία

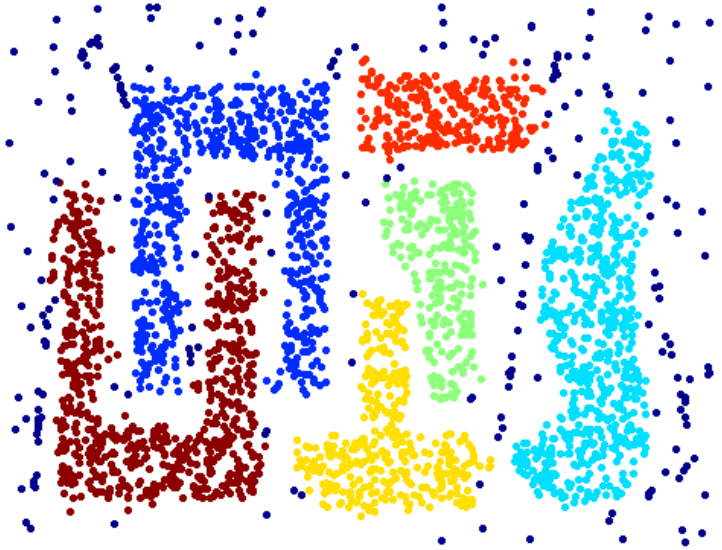
Τύποι σημείων: core, border και noise

Eps = 10, MinPts = 4

DBSCAN Αλγόριθμος και Πλεονεκτήματα

Βήμα 3 και 4

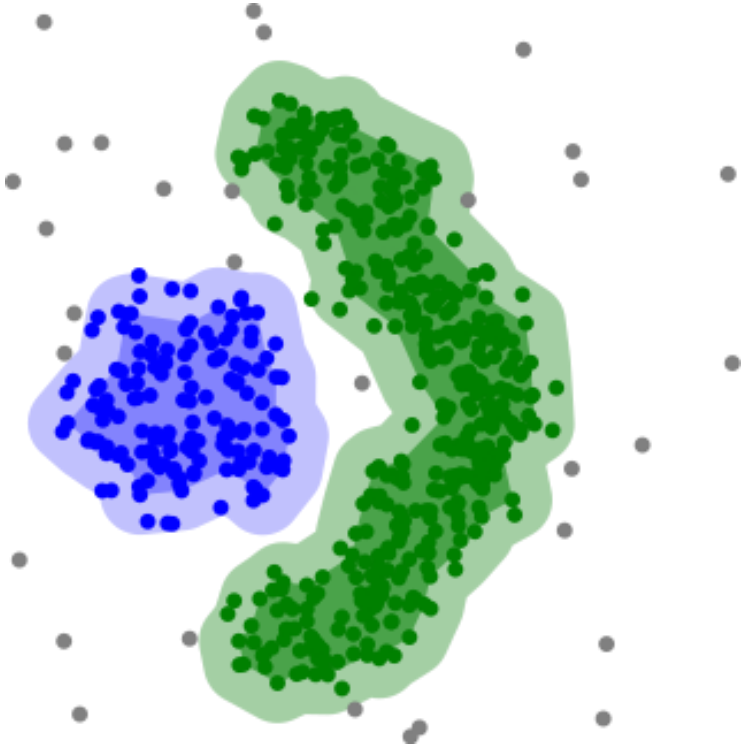
Αρχικά Σημεία



Συστάδες

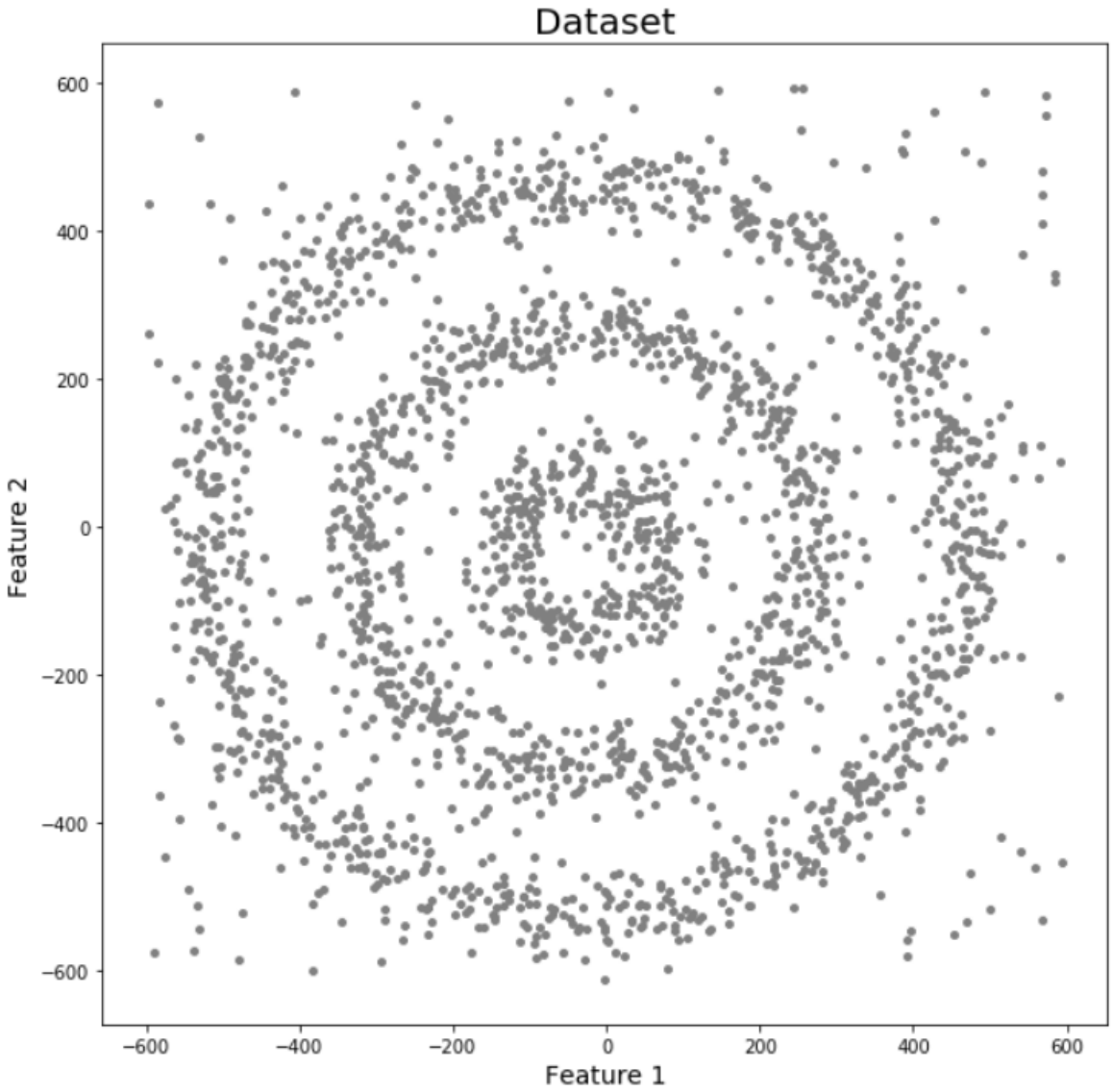
- Δεν επηρεάζεται από το θόρυβο
- Μπορεί να χειριστεί συστάδες με διαφορετικά σχήματα και μεγέθη

DBSCAN Αλγόριθμος και Πλεονεκτήματα

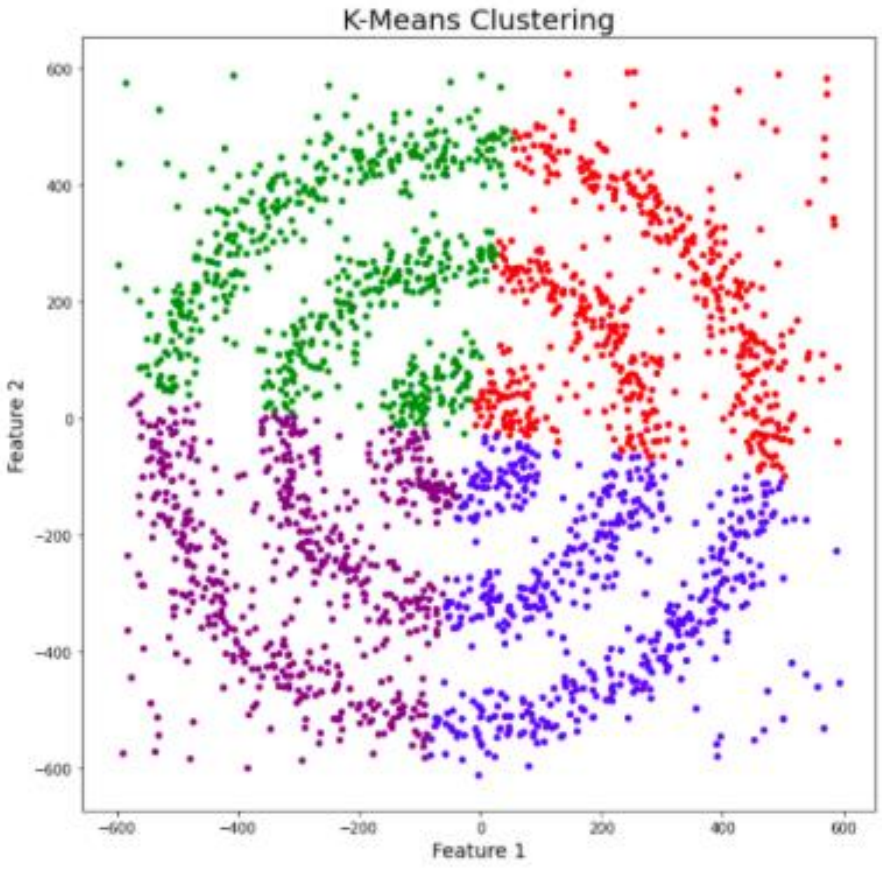


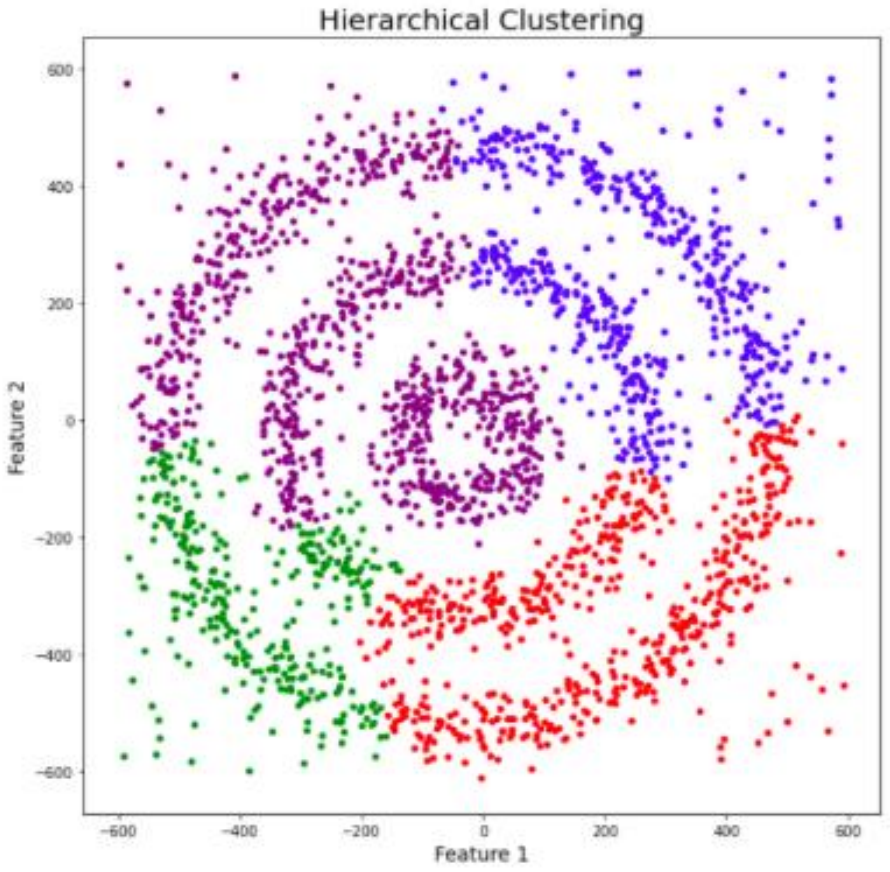
- Το DBSCAN μπορεί να βρει μη γραμμικά διαχωρίσιμες συστάδες
  - Αυτό το σύνολο δεδομένων δεν μπορεί να ομαδοποιηθεί επαρκώς με k-means συσταδοποίηση

DBSCAN vs Αλγορίθμους Συσταδοποίησης

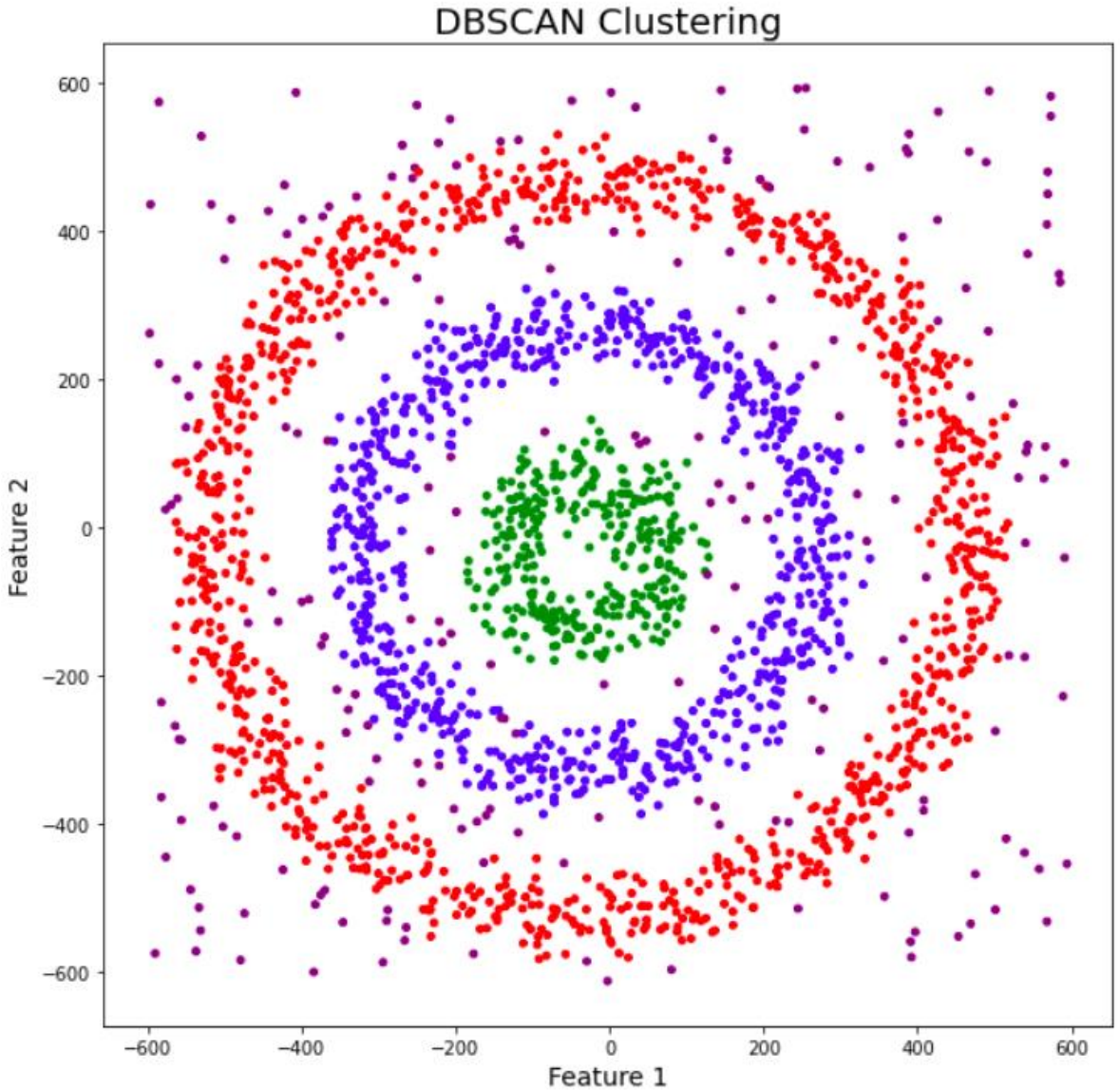


DBSCAN vs Αλγορίθμους Συσταδοποίησης

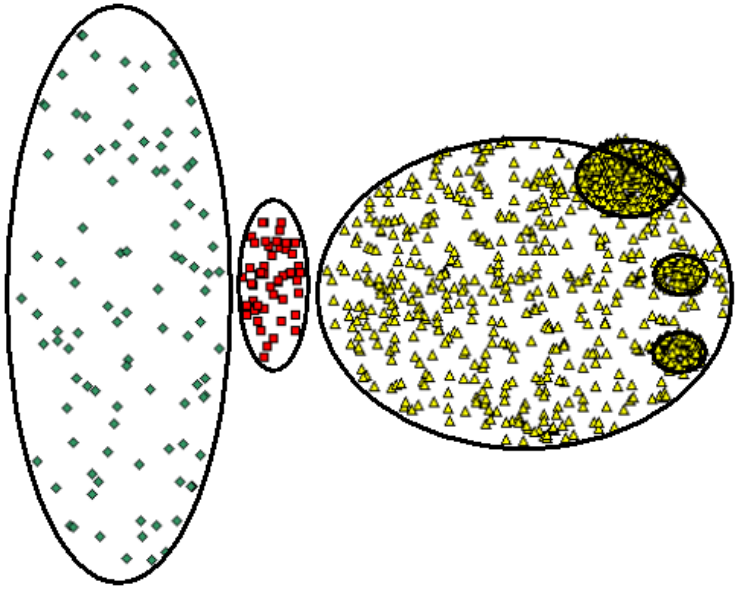




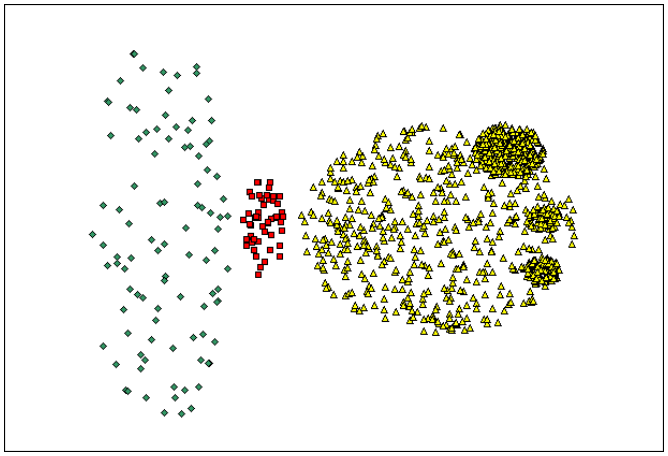
DBSCAN vs Αλγορίθμους Συσταδοποίησης



DBSCAN Περιορισμοί

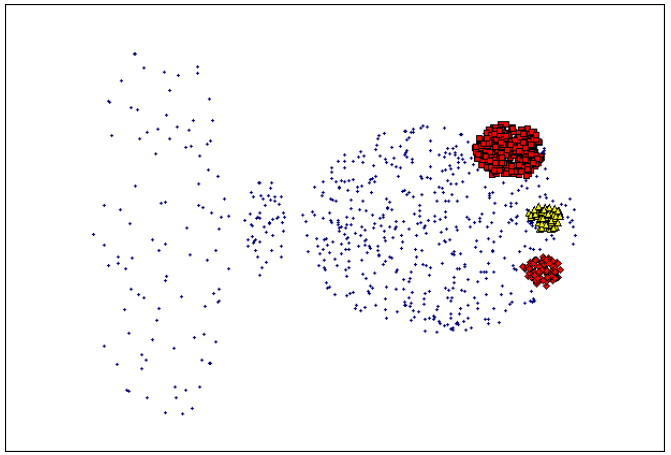


Αρχικά Σημεία



(MinPts=4, Eps=9.75)

- Διαφορετικές πυκνότητες
- Πολυδιάστατα δεδομένα
  - δύσκολος ορισμός πυκνότητας και δαπανηρός υπολογισμός γειτόνων



(MinPts=4, Eps=9.92)

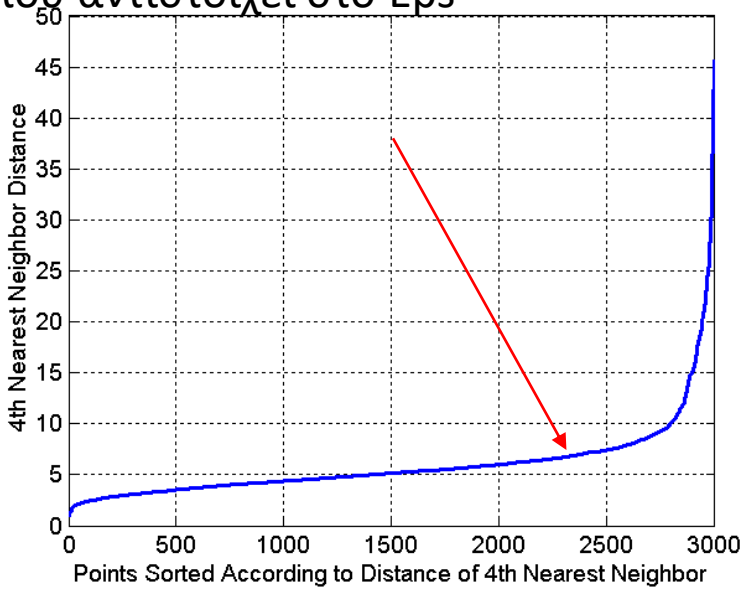
Καθορισμός των MinPts και Eps

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

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

Eps ~ 7

MinPts = 4



DBSCAN Συνοπτικά

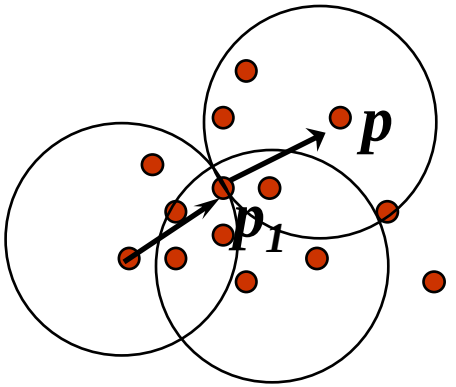
Συσταδοποίηση βασισμένη στην πυκνότητα (τοπικό κριτήριο)

Βασικά χαρακτηριστικά:

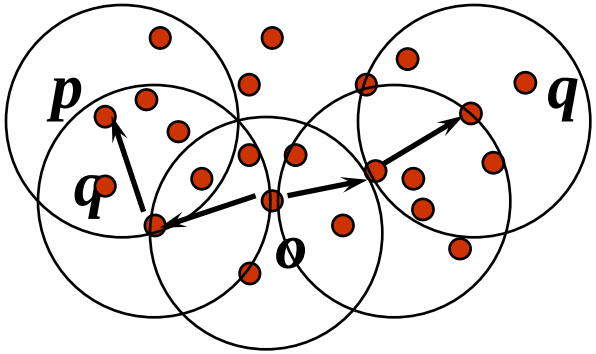
- Ανακαλύπτουν συστάδες οποιουδήποτε σχήματος
- Αντιμετωπίζουν το θόρυβο
- Μία διάσχιση (scan) των δεδομένων
- Χρειάζονται παραμέτρους εισόδου για την πυκνότητα
- Δύσκολο να ανακαλύψουν συστάδες με διαφορετική πυκνότητα
- Στις πολλές διαστάσεις, η έννοια της πυκνότητας είναι ασαφής
- Το κόστος εξαρτάται από το κόστος υπολογισμού του κοντινότερου γείτονα

DBSCAN Τυπικός ορισμός

- Density-reachable (προσπελάσιμο με βάση τη πυκνότητα)
  - Ένα σημείο p είναι density-reachable από ένα σημείο q αν υπάρχει μια αλυσίδα από σημεία
    $p_1, \ldots, p_n, p_1 = q, p_n = p, p_n = p$
    τέτοια ώστε το p_{i+1} να είναι στη γειτονιά του p_i



- Density-connected
  - Ένα σημείο p είναι density-connected σε ένα σημεία q αν υπάρχει ένα σημείο o τέτοιο ώστε και το p and q να είναι density-reachable από το o
  - Συστάδα είναι το μέγιστο (maximal) σύνολο από density-connected σημεία

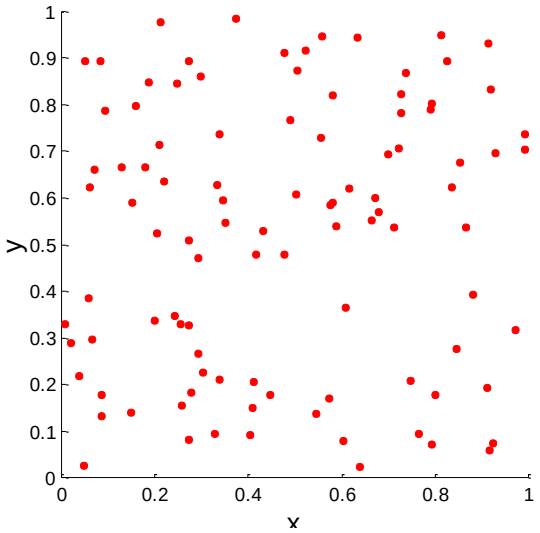


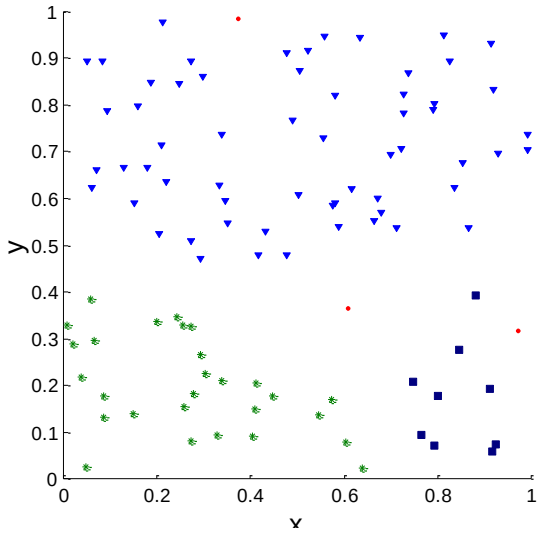
Εγκυρότητα Συσταδοποίησης (Cluster validity)

- Ποιότητα ή εγκυρότητα συσταδοποίησης
  - Πόσο καλή είναι η συσταδοποίηση που επιτύχαμε;
- Οι αλγόριθμοι που είδαμε παράγουν κάποιες συστάδες ακόμα και όταν τα δεδομένα παράγονται τυχαία
- Δύσκολη η αξιολόγηση, ιδιαίτερα σε πολλές διαστάσεις

Συστάδες σε Τυχαία Δεδομένα

Τυχαία Σημεία

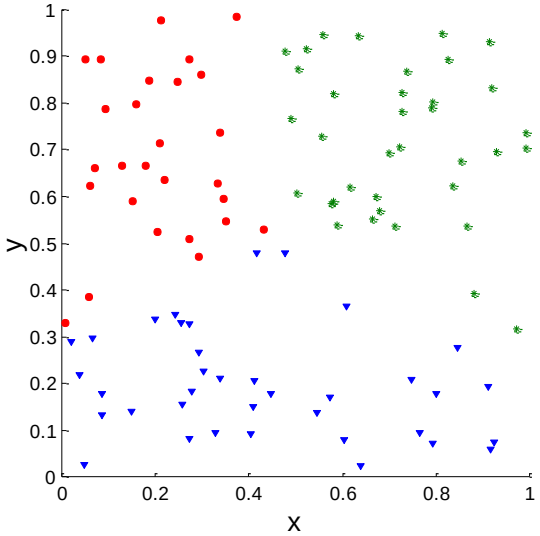


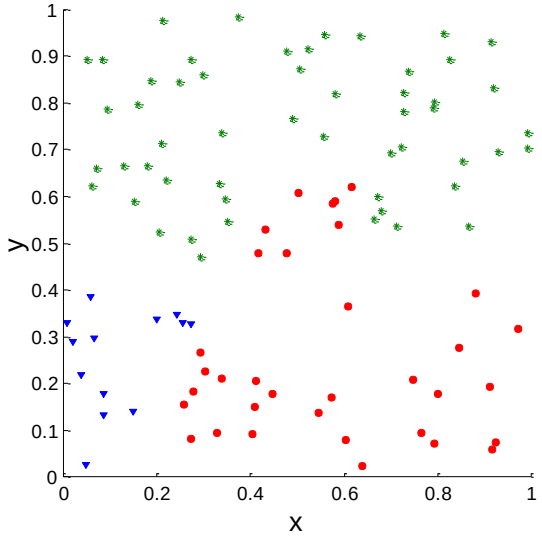


DBSCAN

3 ομάδες κοιτώντας την απόσταση του 4ου γείτονα

K-means





ΣΙΣ με MAX-link

Κριτήρια Ορθότητας Συσταδοποίησης

1. Υπάρχει τάση ομαδοποίησης (clustering tendency), δηλαδή μη τυχαία δομή στο σύνολο των δεδομένων;
2. Σύγκριση των αποτελεσμάτων της ανάλυσης της ομαδοποίησης με κάποια ήδη γνωστά αποτελέσματα, π.χ. κάποια ετικέτα που ήδη έχει δοθεί για μια συστάδα
3. Πόσο καλά ταιριάζουν τα αποτελέσματα της ανάλυσης με τα δεδομένα χωρίς αναφορά σε εξωτερική πληροφορία, χρησιμοποιώντας μόνο τα δεδομένα
4. Σύγκριση των αποτελεσμάτων δυο διαφορετικών συσταδοποιήσεων για να αποφασιστεί ποια είναι καλύτερη
5. Καθορισμός του «σωστού» αριθμού συστάδων

Τα 2, 3, και 4 μπορεί να αφορούν είτε την ολική συσταδοποίηση είτε την κάθε συστάδα χωριστά

Μετρήσεις Ποιότητας Συσταδοποίησης

- Με επίβλεψη (supervised)
  - Εξωτερικό Ευρετήριο (External Index): Υπάρχει εξωτερική πληροφορία (π.χ. ετικέτες για τις συστάδες)
  - Μετράμε πόσο οι περιγραφές των συστάδων ταιριάζουν με τις ετικέτες των κλάσεων (π.χ. Εντροπία)
- Χωρίς επίβλεψη (unsupervised)
  - Εσωτερικό Ευρετήριο (Internal Index): Εκτιμάμε το πόσο καλή είναι μια συσταδοποίηση χωρίς παροχή εξωτερικής πληροφορίας
    - Συνεκτικότητα (cohesion)
    - Διακριτότητα ή διαχωρισμός (separation)
- Συγκριτικοί - Σχετικό Ευρετήριο (Relative Index)
  - Χρησιμοποιείται για τη σύγκριση δυο διαφορετικών συσταδοποιήσεων ή συστάδων
    - Συχνά για αυτό το σκοπό χρησιμοποιείται ένα εσωτερικό (π.χ. δυο k-means συσταδοποιήσεις με βάση το SSE) ή εξωτερικό ευρετήριο

Χαρακτηρισμός Ποιότητας Συσταδοποίησης *Με Επίβλεψη*

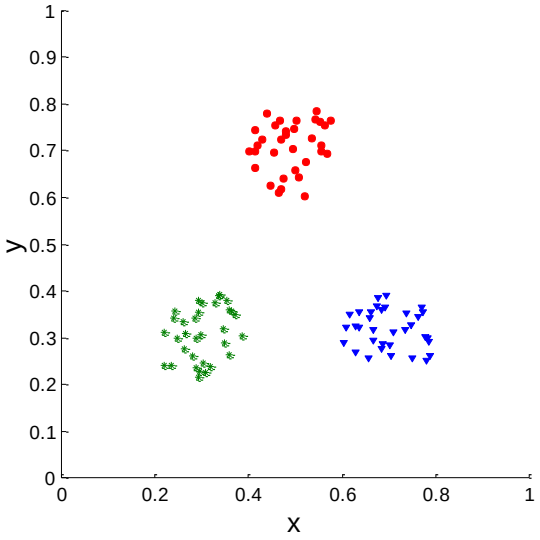
- Χρήση Πίνακα Εγγύτητας

Πίνακας Εγγύτητας (Συσχέτιση)

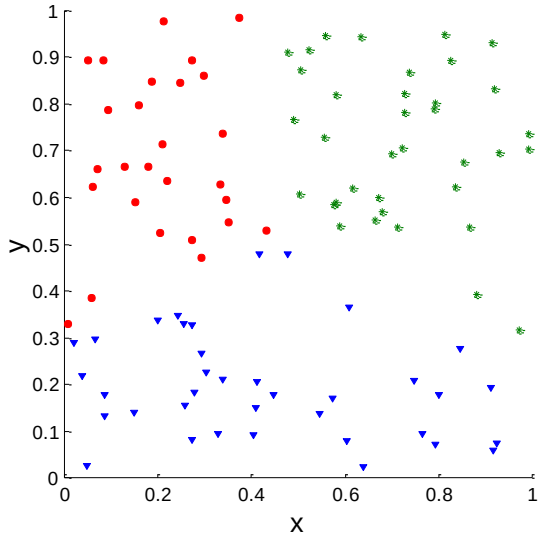
- Δύο Πίνακες
  - **Πίνακας Εγγύτητας** (proximity matrix)
    - Πίνακας με την ομοιότητα των σημείων
  - **Πίνακας Εμφάνισης** (incidence matrix)
    - Μια γραμμή και μια στήλη για κάθε σημείο
    - Μια εγγραφή είναι **1** αν το αντίστοιχο ζευγάρι σημείων ανήκει στην ίδια συστάδα
    - Μια εγγραφή είναι **0** αν το αντίστοιχο ζευγάρι σημείων ανήκει σε διαφορετική συστάδα
- Υπολογισμός της **συσχέτισης (correlation)** των δύο πινάκων

Πίνακας Εγγύτητας (Συσχέτιση)

Υπολογισμός correlation των δύο πινάκων όταν χρησιμοποιείται ο K-means στα παρακάτω σύνολα



Corr = -0.9235



Corr = -0.5810

Πίνακας Εγγύτητας (Συσχέτιση)

Υψηλή συσχέτιση σημαίνει ότι τα σημεία που ανήκουν στην ίδια συστάδα είναι κοντινά μεταξύ τους

- Δεν είναι καλή μέτρηση για κάποιες συστάδες που βασίζονται σε πυκνότητα και σε συνέχεια (contiguity)
- Επειδή, οι δυο πίνακες είναι συμμετρικοί, χρειάζεται ο υπολογισμός
  n(n-1) / 2 εγγραφών

Χαρακτηρισμός Ποιότητας Συσταδοποίησης *Χωρίς Επίβλεψη*

- Χρήση Συνοχής και Διαχωρισμού
- Χρήση Πίνακα Γειτνίασης

Συνοχή και Διαχωρισμός

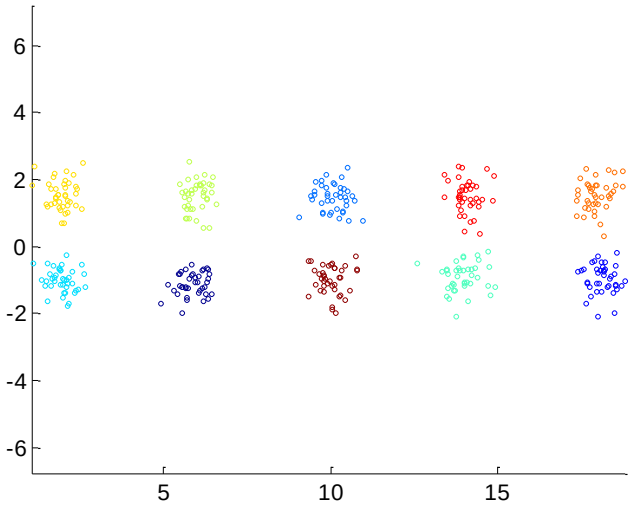
- Έχουμε μια συσταδοποίηση (ένα σύνολο συστάδων);
  - Πόσο «καλή/έγκυρη» είναι
- Δύο μέτρα:
  - Για **κάθε συστάδα ξεχωριστά** (cohesion - συνοχή): πόσο κοντά (όμοια) είναι τα σημεία κάθε συστάδας
  - Για τις **συστάδες μεταξύ τους** (separation - διαχωρισμός): πόσο μακριά (ανόμοιες) είναι δύο συστάδες
- Συνδυασμός της συνοχής και του διαχωρισμού για το χαρακτηρισμό συνολικά της συσταδοποίησης

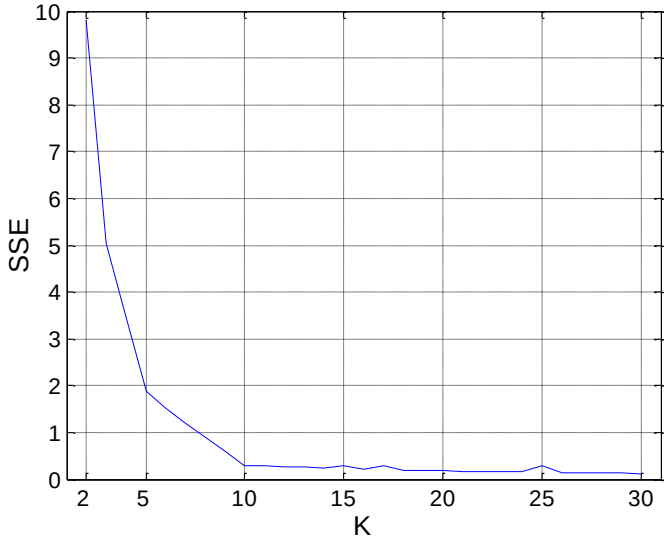
Συνοχή και Διαχωρισμός

- Μπορούν να χρησιμοποιηθούν για τη βελτίωση της συσταδοποίησης
  - Π.χ. μια συστάδα με κακή συνοχή μπορεί να χρειαστεί να διασπαστεί
- Δυο συστάδες όχι καλά διαχωρισμένες μπορεί να συγχωνευτούν
- Αποτελέσματα:
  - Πόσο καλή είναι μια συσταδοποίηση
  - Ποσό καλή είναι μια συστάδα
  - Ποσό καλό είναι ένα σημείο σε μια συστάδα

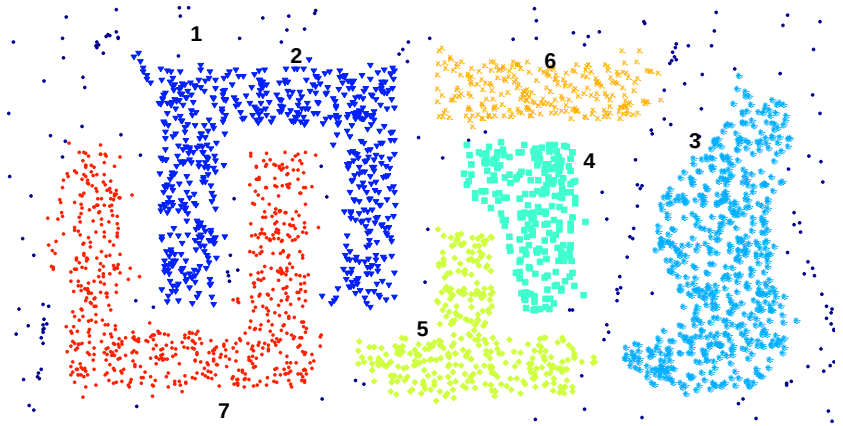
Χρήση για καθορισμό του Πλήθους Συστάδων

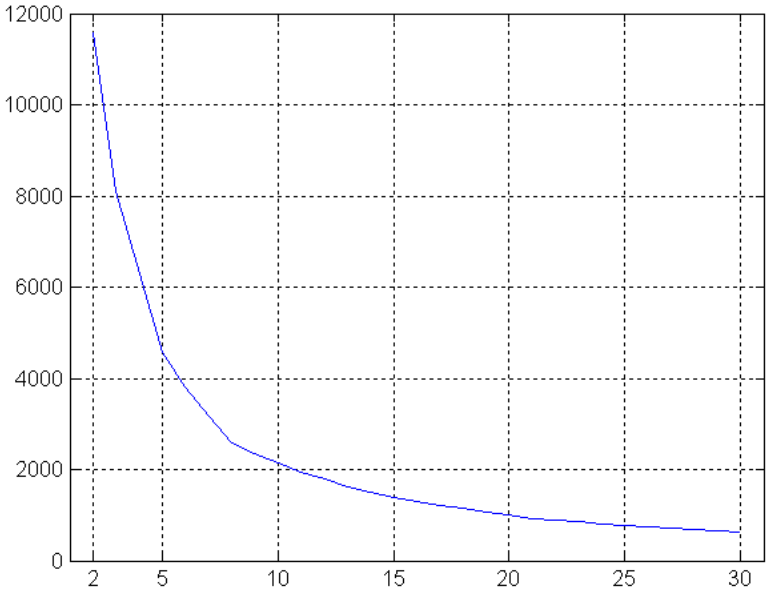
Χρήση SSE (Άθροισμα του Τετραγωνικού Σφάλματος) για υπολογισμό του σωστού αριθμού συστάδων χρησιμοποιώντας τον K-means
(K = 5 και 10 φαίνονται καλές τιμές)





Χρήση για καθορισμό του Πλήθους Συστάδων

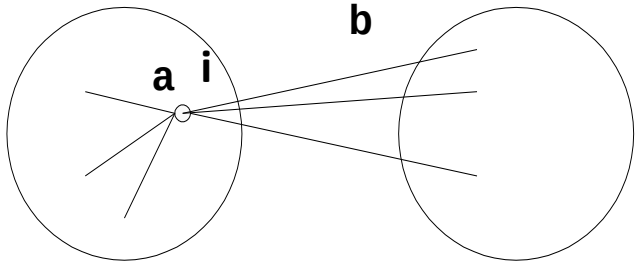




Συντελεστής Σκιαγράφησης (Silhouette Coefficient)

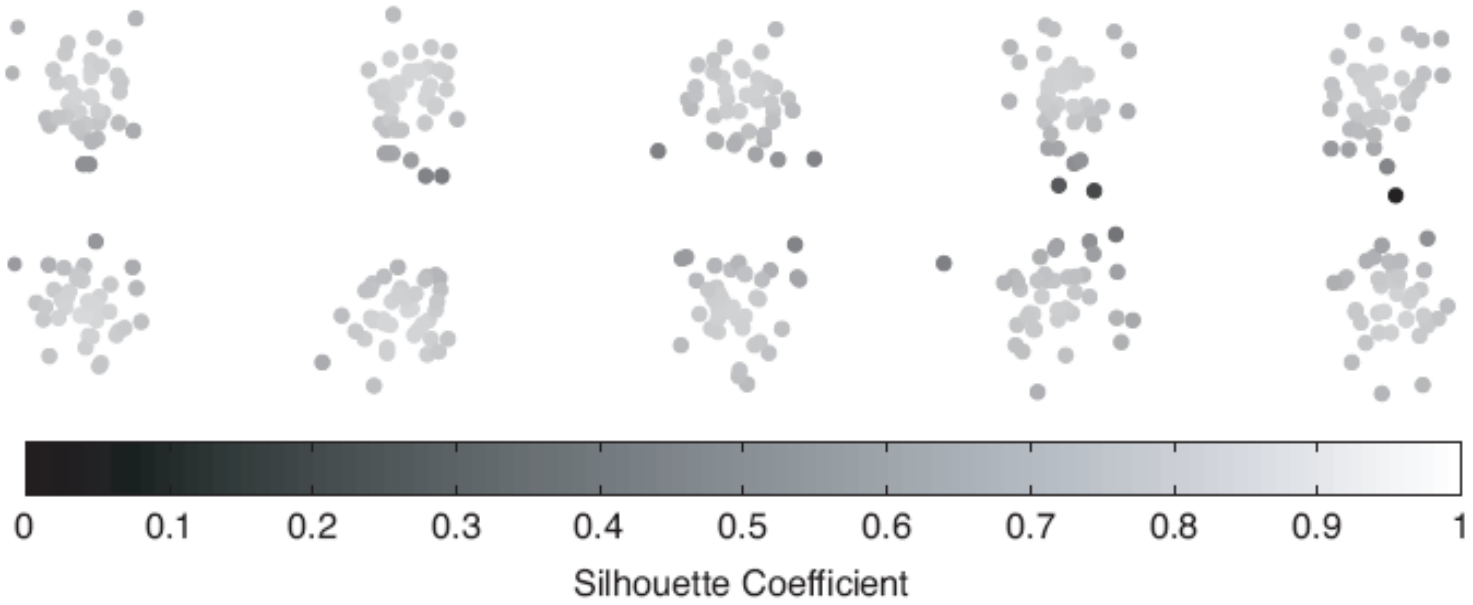
Για κάθε σημείο *i*

- Υπολογισμός ***a*** = μέση απόσταση του *i* από τα σημεία της συστάδας
- Υπολογισμός ***b*** = μέση απόσταση του *i* από όλα τα σημεία κάθε άλλης συστάδας
  - Επιλογή του μικρότερου, δηλαδή μέση απόσταση από την κοντινότερη συστάδα
- s = 1 – a/b, αν a < b, (ή s = b/a - 1 αν a ≥ b, η μη συνηθισμένη περίπτωση)
  - Συνήθως μεταξύ του 0 και του 1
  - Όσο πιο κοντά στο 1, τόσο το καλύτερο



Μπορεί να χρησιμοποιηθεί και για μια συστάδα ή συσταδοποίηση θεωρώντας μέσες τιμές για όλα τα σημεία τους ή συστάδες

Συντελεστής Σκιαγράφησης

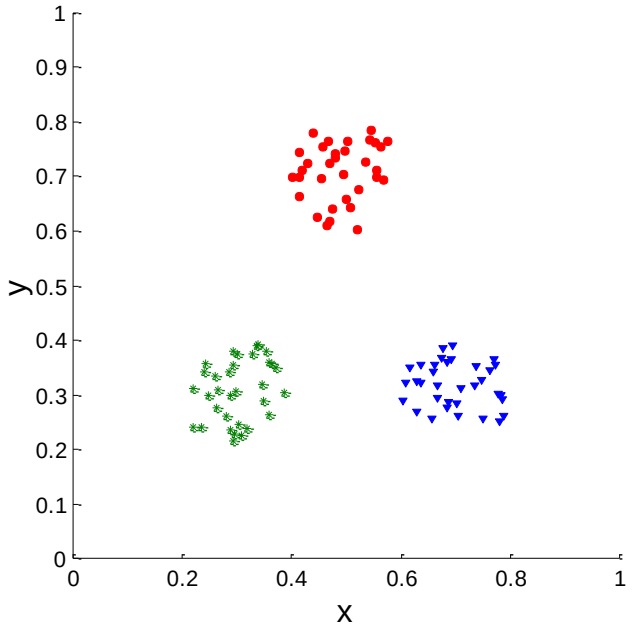


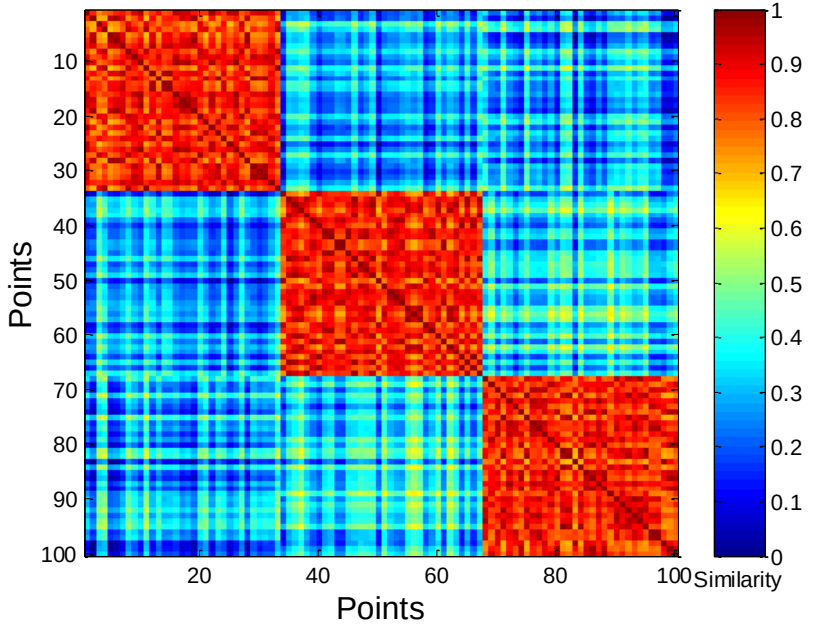
Ο συντελεστής σκιαγράφησης για σημεία στις 10 συστάδες

Πόσο «κεντρικό» είναι ένα σημείο για μία συστάδα (όσο πιο ανοιχτό-χρωμο τόσο το καλύτερο)

Πίνακας Εγγύτητας

- Αναδιατάσσουμε τα σημεία στον πίνακα γειτνίασης ή εγγύτητας έτσι ώστε τα σημεία που ανήκουν στην ίδια συστάδα να είναι γειτονικά
- Συγκεκριμένα, τα διατάσσουμε με βάση τη συστάδα:
  - Σημεία Συστάδας 1, Σημεία Συστάδας 2, Σημεία Συστάδας 3

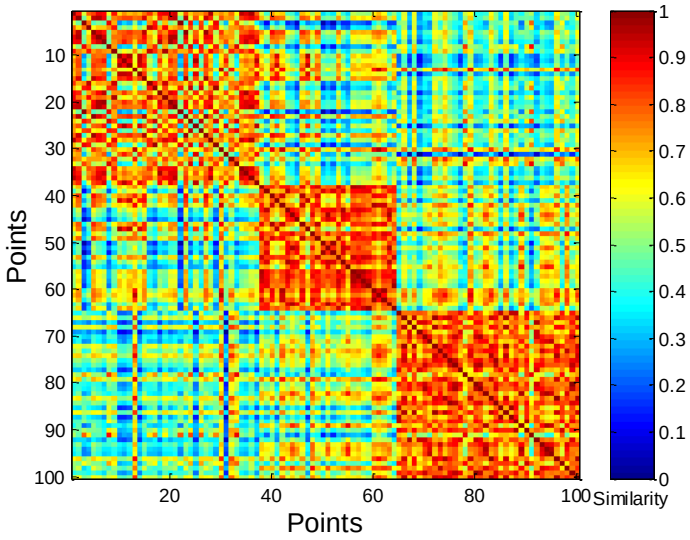


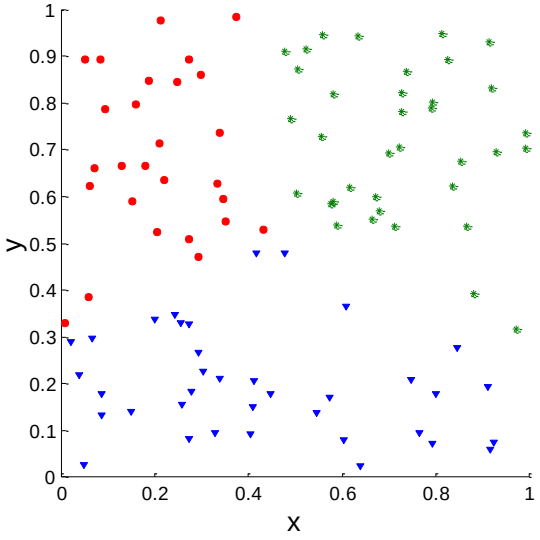


Σημείωση: similarity = 1 – (d –min_d)/(max_d – min_d)

ΔΙΑΓΩΝΙΟΣ ΜΠΛΟΚ ΠΙΝΑΚΑΣ

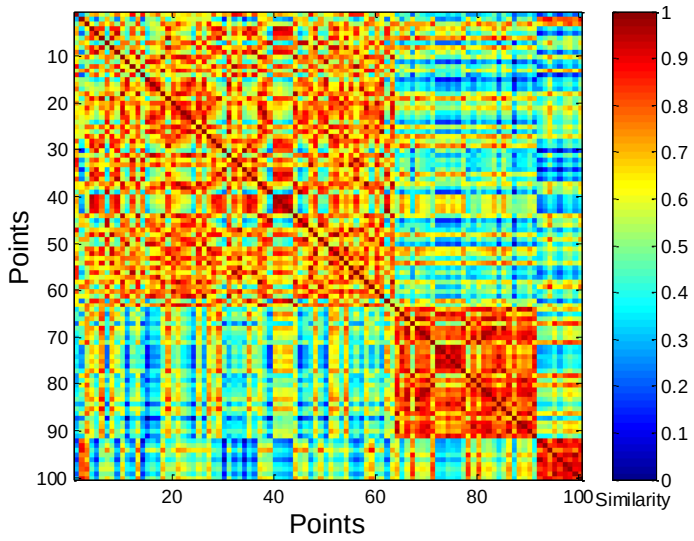
Πίνακας Εγγύτητας

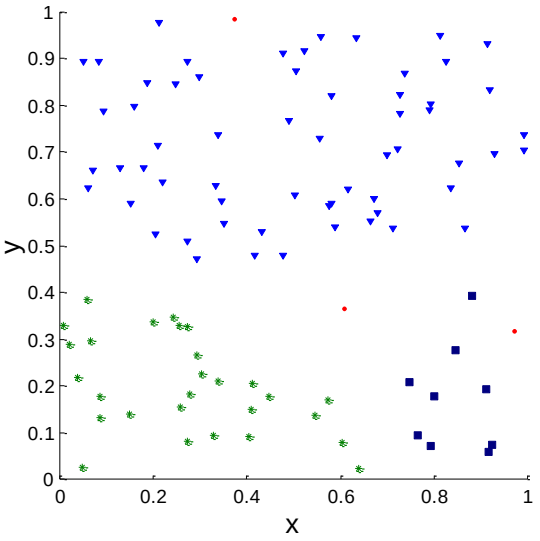




K-means

Πίνακας Εγγύτητας

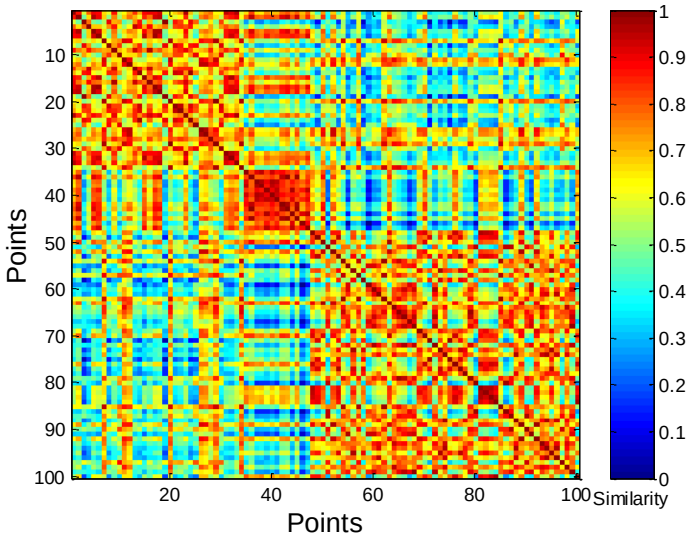


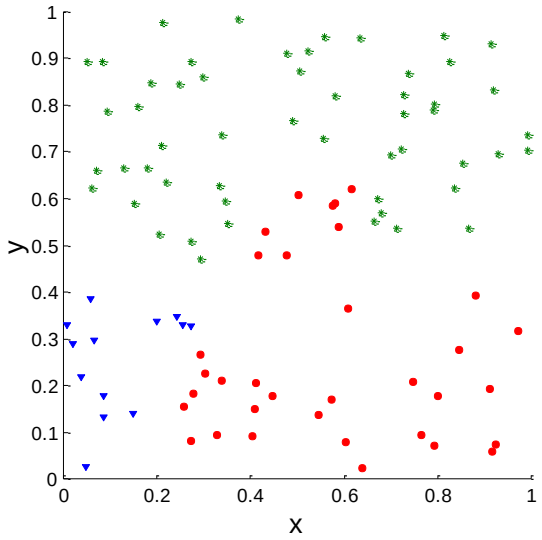


DBSCAN

Κάποιες συστάδες ακόμα και σε τυχαία δεδομένα

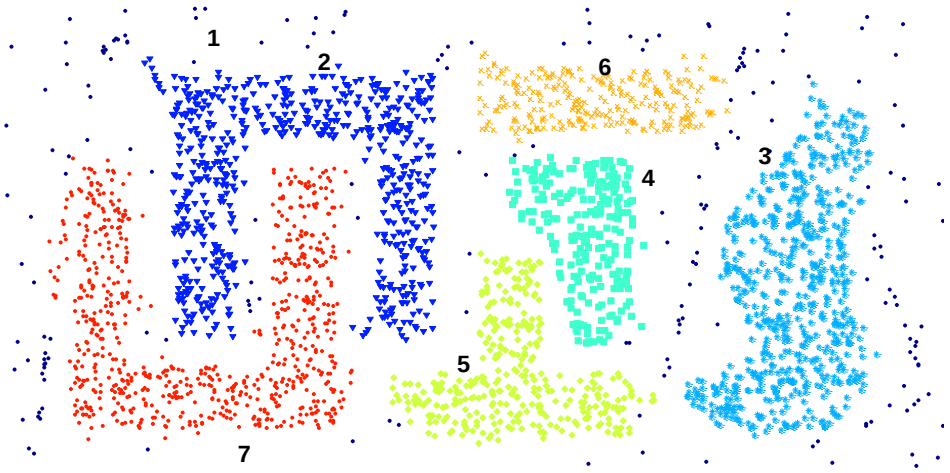
Πίνακας Εγγύτητας

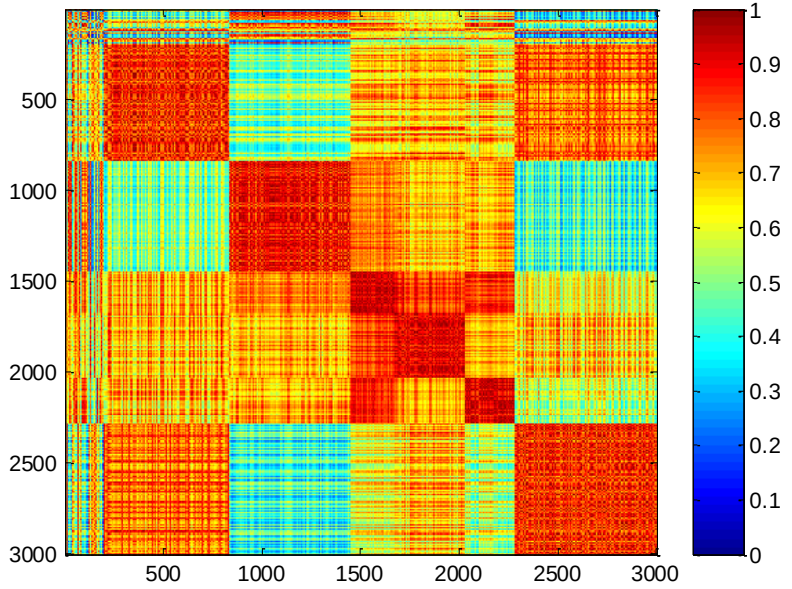




ΣΙΣ-max

Πίνακας Εγγύτητας





DBSCAN

Αναφορές

- Pang-Ning Tan, Michael Steinbach, Vipin Kumar, «Introduction to Data Mining», 2nd edition, Pearson, 2018
- Jure Leskovec, Anand Rajaraman, Jeff Ullman, «Mining of Massive Datasets», Cambridge University Press, 2019, http://www.mmds.org/
- https://www.cs.uoi.gr/~pitoura/courses/dm/
- Ευστάθιος Κύρκος, «Επιχειρηματική ευφυΐα και εξόρυξη δεδομένων», 2015, ISBN: 978-960-603-109-0
- https://www.analyticsvidhya.com/blog/2020/09/how-dbscan-clustering-works/