Εξόρυξη Δεδομένων

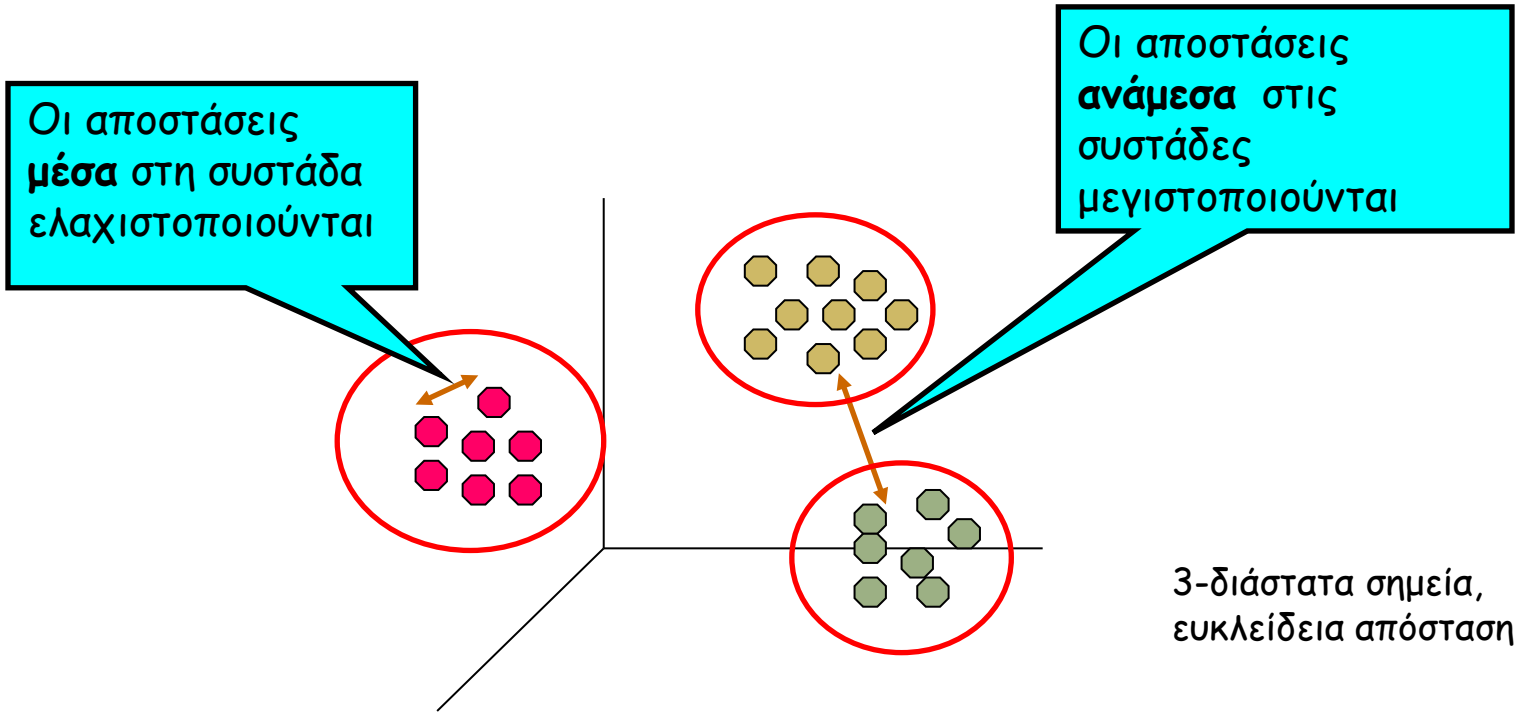
7: Συσταδοποίηση

Περιεχόμενα

- Ορισμός, προβλήματα
- (Μέτρα απόστασης)
- K-Means
  - Bisecting K-Means
  - K-Medoid
- Ιεραρχικοί αλγόριθμοι
  - Συσσωρευτικοί
  - Διαιρετικοί
  - HAC

Ορισμός

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



Εφαρμογές

- Κατανόηση του διαχωρισμού των δεδομένων
- Οπτικοποίηση, συμπεράσματα για την κατανομή

- Προεπεξεργασία
  - Περίληψη: Ελάττωση του μεγέθους μεγάλων συνόλων χρήση αντιπροσωπευτικών σημείων από κάθε συστάδα – πρωτότυπα (prototypes),
  - Συμπίεση ή
  - Αποδοτική κατασκευή ευρετηρίων – εύρεση κοντινότερου γείτονα κλπ

Ποιότητα συσταδοποίησης

- 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
- Προβλήματα:
  - Πόσες συστάδες/ομάδες;
  - Θόρυβος
  - Outliers

Ασάφεια

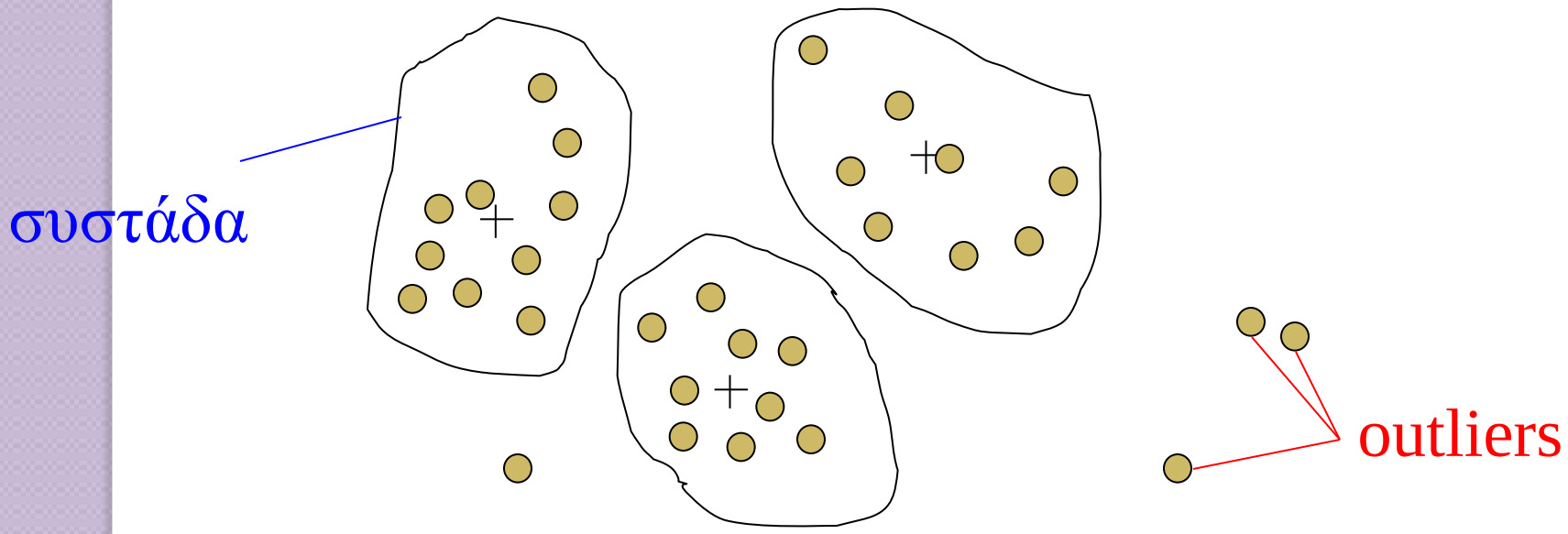
Πόσες Ομάδες?

6 ομάδες

2 ομάδες

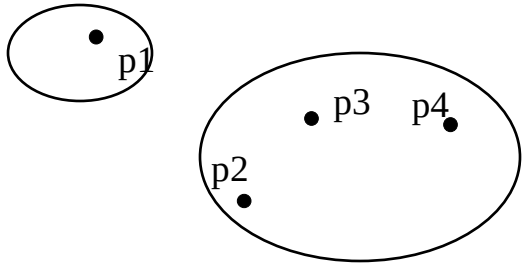
4 ομάδες

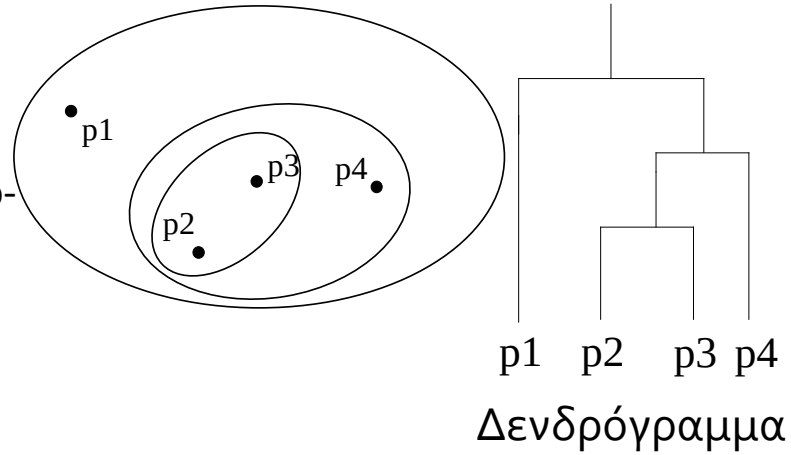
Θόρυβος & Εξαιρέσεις



Τακτικές συσταδοποίησης

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





Άλλοι τρόποι διάκρισης

- Επικαλυπτόμενη συσταδοποίηση
  - Ένα σημείο ανήκει σε περισσότερες από μια συστάδες (πχ οριακά σημεία)
- Ασαφής συσταδοποίηση
  - Στην ασαφή συσταδοποίηση ένα σημείο ανήκει σε κάθε συστάδα με κάποιο βάρος μεταξύ του 0 και του 1
  - Συχνά τα βάρη για κάθε σημείο έχουν άθροισμα 1
  - Η πιθανοτική συσταδοποίηση έχει παρόμοια χαρακτηριστικά
- Μερική - Πλήρης
  - Σε ορισμένες περιπτώσεις θέλουμε να ομαδοποιήσουμε μόνο κάποια από τα δεδομένα (άλλα θόρυβος, ή μη ενδιαφέρουσα πληροφορία)
- Ετερογενή - Ομογενή
  - Συστάδες με πολύ διαφορετικά μεγέθη, σχήματα και πυκνότητες (densities)

Απόσταση και Ομοιότητα

Κριτήρια Ομοιότητας -Απόστασης

- Ομοιότητα
  - 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 (όταν τα ίδια), αλλά το πάνω όρο διαφέρει
- Η συνάρτηση απόστασης ανάμεσα στα αντικείμενα εξαρτάται από το είδος των δεδομένων, δηλαδή από το είδος των γνωρισμάτων τους

Μέτρα απόστασης

- Ευκλείδεια απόσταση

$$d(i,j)=\sqrt{(|x_{i1}-x_{j1}|^2+|x_{i2}-x_{j2}|^2+...+|x_{in}-x_{jn}|^2)}$$

- Manhattan ή city-block

$$L_1(i,j)=|x_{i1}-x_{j1}|+|x_{i2}-x_{j2}|+...+|x_{in}-x_{jn}|$$

- Minkowski (p-norm)

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

- Hamming distance = πλήθος ίδιων bits για δυαδικά διανύσματα

Ομοιότητα σε δυαδικά διανύσματα

- 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
- Invariant ομοιότητα: Συμμετρικές (τιμές 0 και 1 έχουν την ίδια σημασία). Simple matching coefficient.

  SMC = αριθμός ταιριασμάτων / αριθμός γνωρισμάτων

  = $(M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00})$
- Non-invariant (Jaccard): Μη συμμετρικές (η συμφωνία στο 1 πιο σημαντική – πχ όταν το 1 σηματοδοτεί την ύπαρξη κάποιας ασθένειας)

  J = αριθμός 11 ταιριασμάτων / αριθμό μη μηδενικών γνωρισμάτων

  = (M11) / (M01 + M10 + M11)

Παράδειγμα

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

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

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

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

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

Κατηγορικές Μεταβλητές χωρίς Διάταξη (nominal)

- Γενίκευση των δυαδικών μεταβλητών (γνωρισμάτων) όπου μπορούν να πάρουν παραπάνω από 2 τιμές, πχ 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
- Ομοιότητα συνημίτονου (cosine similarity)
  - 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

Cosine similarity: αγνοεί τα 0 (όπως η Jaccard) αλλά να δουλεύει και για μη δυαδικά δεδομένα, επίσης, αγνοεί το μήκος των διανυσμάτων

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

Αλγόριθμοι συσταδοποίησης

K-means και παραλλαγές

Ιεραρχική Συσταδοποίηση

Συσταδοποίηση με βάση την Πυκνότητα (DBSCAN)

BIRCH (δεδομένα στο δίσκο!)

Γενικές απαιτήσεις

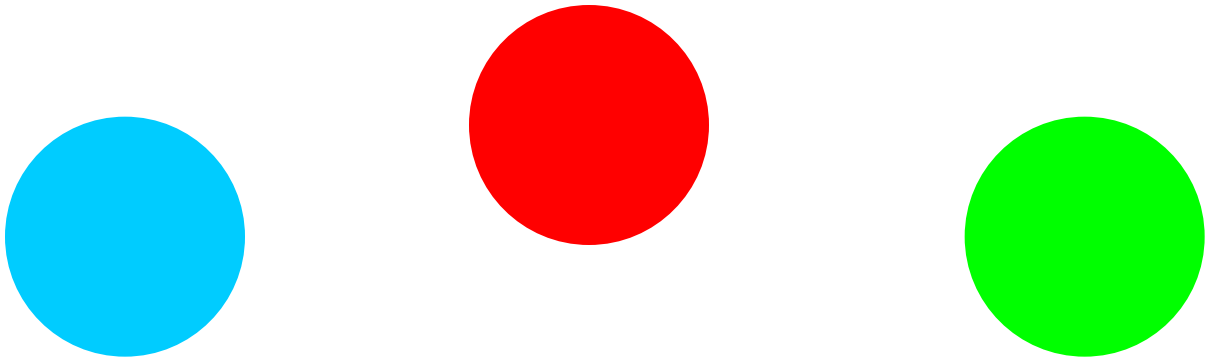
- Κλιμάκωση: στον αριθμό σημείων και διαστάσεων
  - Disk-resident vs Main memory
- Να υποστηρίζει διαφορετικούς τύπους δεδομένων
- Να υποστηρίζει συστάδες με διαφορετικά σχήματα (συνήθως, «σφαίρες»)
- Να είναι εύκολο να δώσουμε τιμές στις παραμέτρους εισόδου (αριθμό συστάδων, μέγεθος κλπ)
- Να μην εξαρτάται από τη σειρά επεξεργασίας των σημείων εισόδου
- Δυναμικά μεταβαλλόμενα δεδομένα
  - Αλλαγή συστάδων με το πέρασμα του χρόνου

Απαιτήσεις

- Καλά διαχωρισμένες συστάδες
- Συνεχείς (contiguous) συστάδες
  - Συστάδες βασισμένες σε κέντρο
  - Συστάδες βασισμένες σε πυκνότητα
  - Βασισμένα σε ιδιότητες ή έννοιες
  - Συστάδες που περιγράφονται από μια αντικειμενική συνάρτηση (Objective Function)

Καλά Διαχωρισμένες Συστάδες

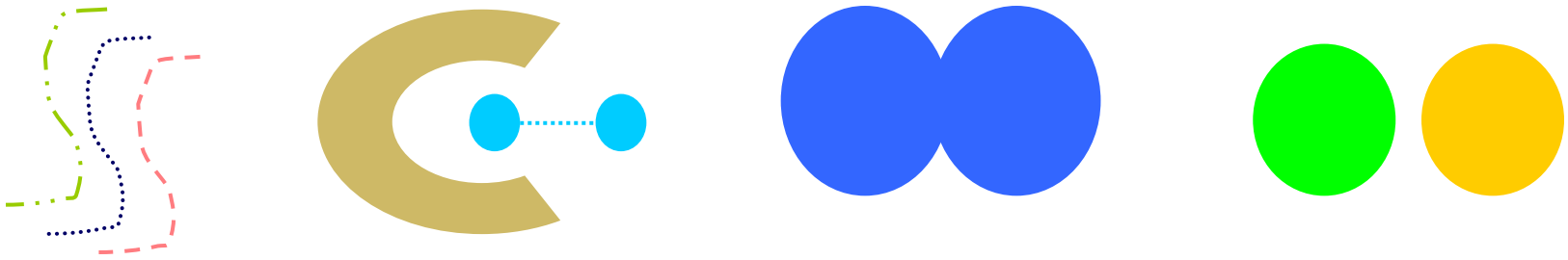
- Μια συστάδα είναι ένα σύνολο από σημεία τέτοια ώστε κάθε σημείο μιας ομάδας είναι κοντινότερο σε (ή πιο όμοιο με) όλα τα άλλα σημεία της ομάδας από ότι σε οποιοδήποτε άλλο σημείο που δεν ανήκει στη συστάδα.
- Συχνά υπάρχει η έννοια του κατωφλιού (threshold)
- Όχι απαραίτητα κυκλικοί (οποιοδήποτε σχήμα)



3 καλά-διαχωρισμένες συστάδες

Συνεχείς συστάδες

- Συνεχείς Συστάδες (Contiguous Clusters) (Κοντινότερος γείτονας ή μεταβατικά)
  - 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 σημείο εκτός συστάδας
- Εμφανίζονται σε περιπτώσεις συστάδων με μη κανονικό σχήμα ή με αλληλοπλεκόμενα σχήματα – ή όταν έχουμε γραφήματα και θέλουμε να βρούμε συνεκτικά υπογραφήματα
- Πρόβλημα με θόρυβο



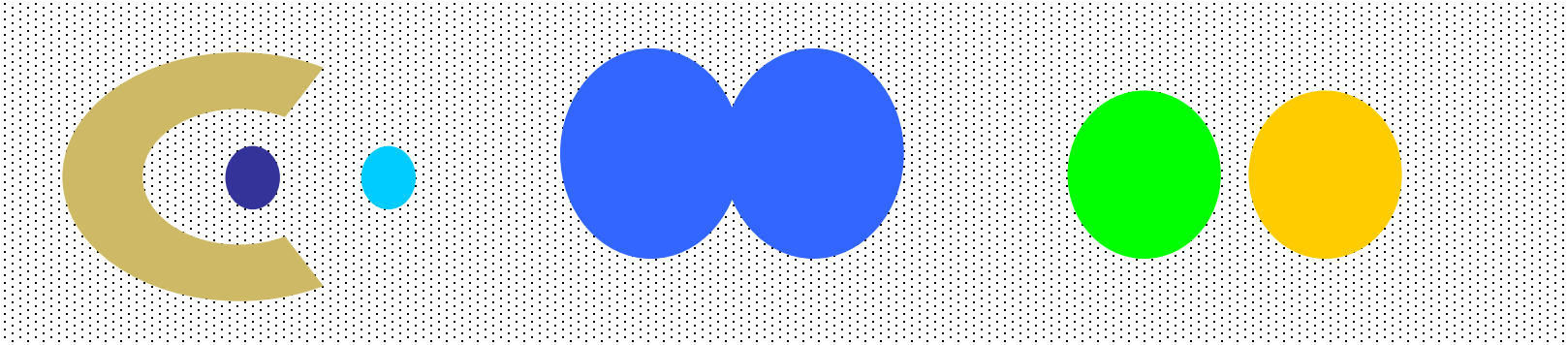
8 συνεχείς συστάδες

Συστάδες βασισμένες σε κέντρο ή πρότυπο

- Μια συστάδα είναι ένα σύνολο από αντικείμενα τέτοιο ώστε ένα αντικείμενο στην ομάδα είναι κοντινότερο στο (ή πιο όμοιο με το) «κέντρο» ή πρότυπο της συστάδας από ότι από το κέντρο οποιασδήποτε άλλης συστάδας .
- Το κέντρο της συστάδας είναι συχνά
  - centroid, ο μέσος όρος των σημείων της συστάδας, ή
  - ένα medoid, το πιο «αντιπροσωπευτικό» σημείο της συστάδας (πχ όταν έχω κατηγορικά γνωρίσματα)
- Προϋποθέτει ότι κάθε φορά που προσθέτουμε ένα νέο αντικείμενο στη συστάδα, ενημερώνουμε το κέντρο της
- Τείνει να δώσει κυκλικές συστάδες

Συστάδες βασισμένες στην πυκνότητα

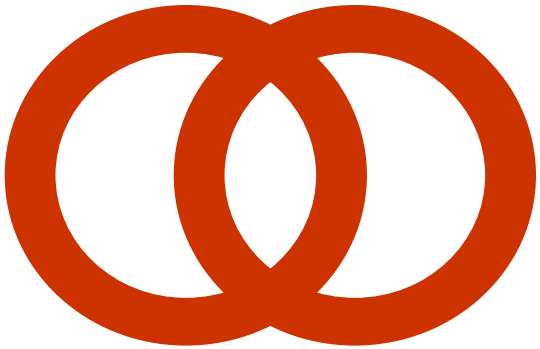
- Μια συστάδα είναι μια πυκνή περιοχή από σημεία την οποία χωρίζουν από άλλες περιοχές μεγάλης πυκνότητας περιοχές χαμηλής πυκνότητας
- Συχνά σε περιπτώσεις συστάδων με μη κανονικό σχήμα ή με αλληλοπλεκόμενα σχήματα ή όταν θόρυβος ή outliers



6 συστάδες βασισμένες στην πυκνότητα

Άλλοι τρόποι σύνδεσης

- Συστάδες με κοινή ιδιότητα ή εννοιολογικές συστάδες.



2 αλληλοκαλυπτόμενοι κύκλοι

- Συστάδες που ελαχιστοποιούν ή μεγιστοποιούν μια αντικειμενική συνάρτηση
- Απαρίθμηση όλων των δυνατών τρόπων χωρισμού των σημείων σε συστάδες και υπολογισμού του «πόσο καλό» (“goodness”) είναι κάθε πιθανό σύνολο από συστάδες χρησιμοποιώντας τη δοθείσα αντικειμενική συνάρτηση (NP-hard)
- Ολικοί ή τοπικοί στόχοι
  - Οι ιεραρχικοί συνήθως τοπικού
  - Οι διαχωριστικοί ολικές

K-means

Γενικά

- 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, Κ, είναι είσοδος στον αλγόριθμο

1: Επιλογή *K* σημείων ως τα αρχικά κεντρικά σημεία
2: **Repeat**
3: Ανάθεση όλων των αρχικών σημείων στο *κοντινότερο* τους από τα *K* κεντρικά σημεία
4: Επανα-υπολογισμός του *κεντρικού σημείου* κάθε συστάδας
5: **Until** τα κεντρικά σημεία να μην αλλάζουν

Ιδιότητες

- Τα αρχικά κέντρα συνήθως επιλέγονται τυχαία
  - Οι συστάδες που παράγονται διαφέρουν από το ένα τρέξιμο του αλγορίθμου στο άλλο
- 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
- Το κεντρικό σημείο είναι (συνήθως) το μέσο (mean) των σημείων της συστάδας (το οποίο μπορεί να μην είναι ένα από τα δεδομένα εισόδου)

Παράδειγμα

Δεδομένα: 2 4 10 12 3 20 30 11 15
Έστω k = 2, και αρχικά επιλέγουμε το 3 και το 4

Cluster A	Cluster B
3: 2	4: 10,12,20,30,11,15

Ξαναυπολογίζω το centroids (π.χ. average)
Και κατηγοριοποιώ

Cluster A	Cluster B
2.5: 2,4,3	14.57: 10,12,20,30,11,15

Ξαναυπολογίζω το centroids (π.χ. average)
Και κατηγοριοποιώ

Cluster A	Cluster B
3: 2,4,3	16.33: 10,12,20,30,11,15

Δεν έχω αλλαγές στις συστάδες

Αντιπροσωπευτικό σημείο συστάδας

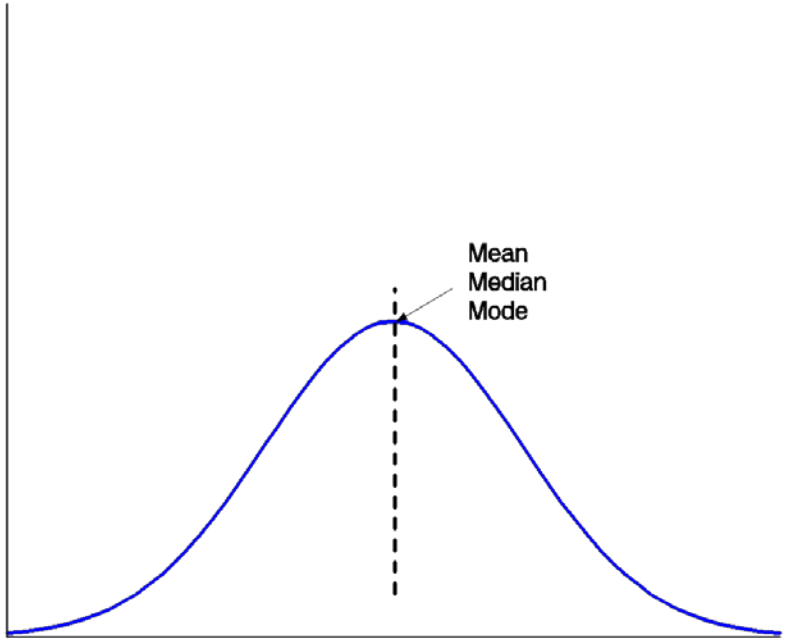
- Αριθμητικό Μέσο/Μέση Τιμή- Mean (αλγεβρική μέτρηση) (sample vs. population):
  - Αριθμητικό μέσο με βάρος (Weighted arithmetic mean)
  - Trimmed mean: κόβουμε τις ακραίες τιμές (πχ τα μεγαλύτερα και μικρότερα)
- Μέσο – μεσαία τιμή (median):
  - Μεσαία τιμή αν μονός αριθμός, ο μέσος όρος των δυο μεσαίων τιμών, αλλιώς
  - Το μέσο συμπεριφέρεται καλύτερα όταν έχω δεδομένα με μη ομοιόμορφη κατανομή (skewed)
- Mode
  - Η πιο συχνά εμφανιζόμενη τιμή
  - Unimodal, bimodal, trimodal
- Midrange (μέσο διαστήματος)
  - (min()+max())/2

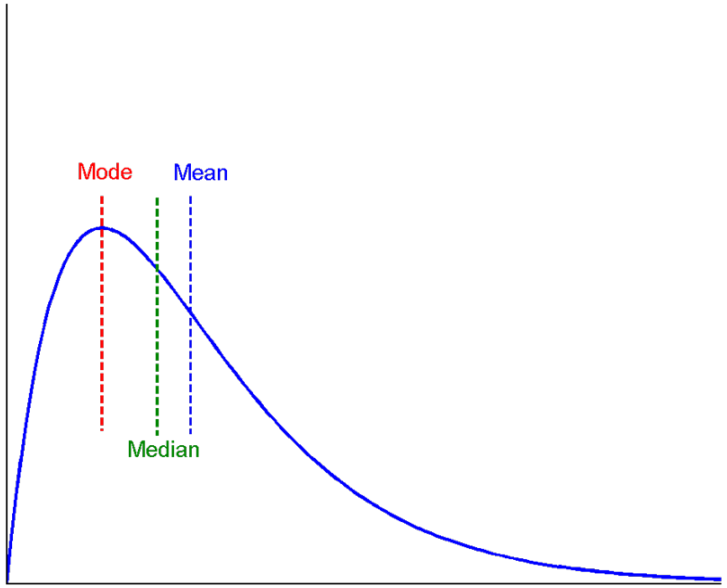
$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

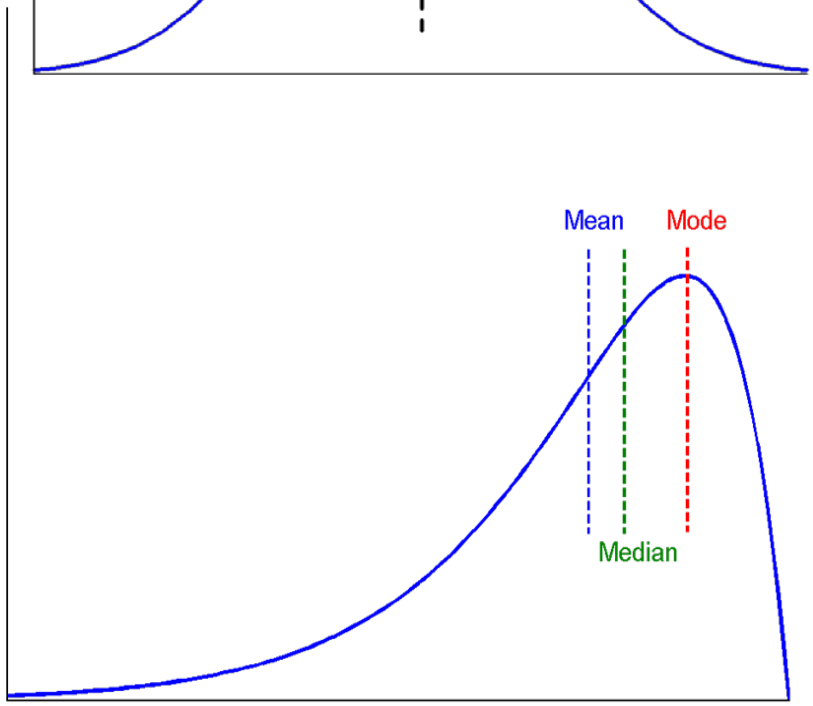
$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

Απεικόνιση

Median, mean and mode of symmetric, positively and negatively skewed data





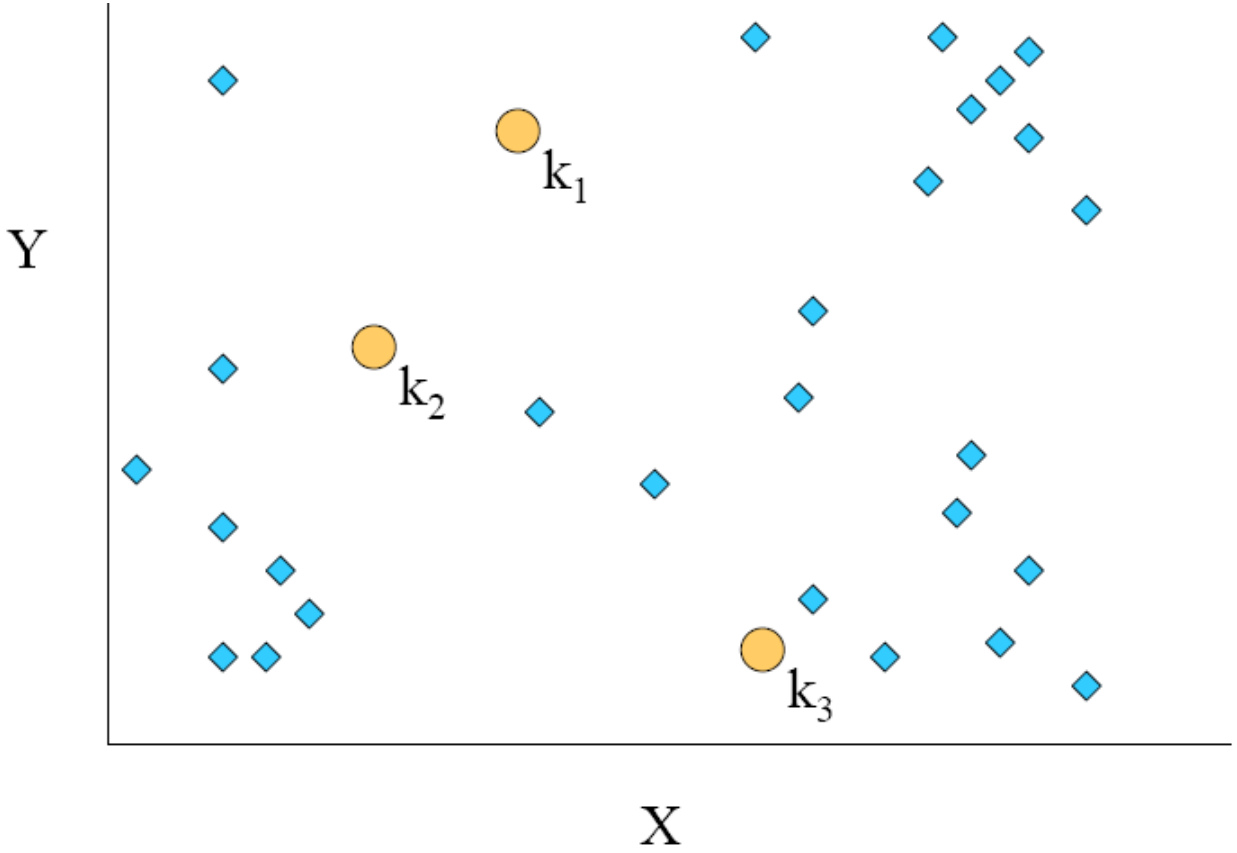


K-means: Βασικός Αλγόριθμος

Αρχική κατάσταση,

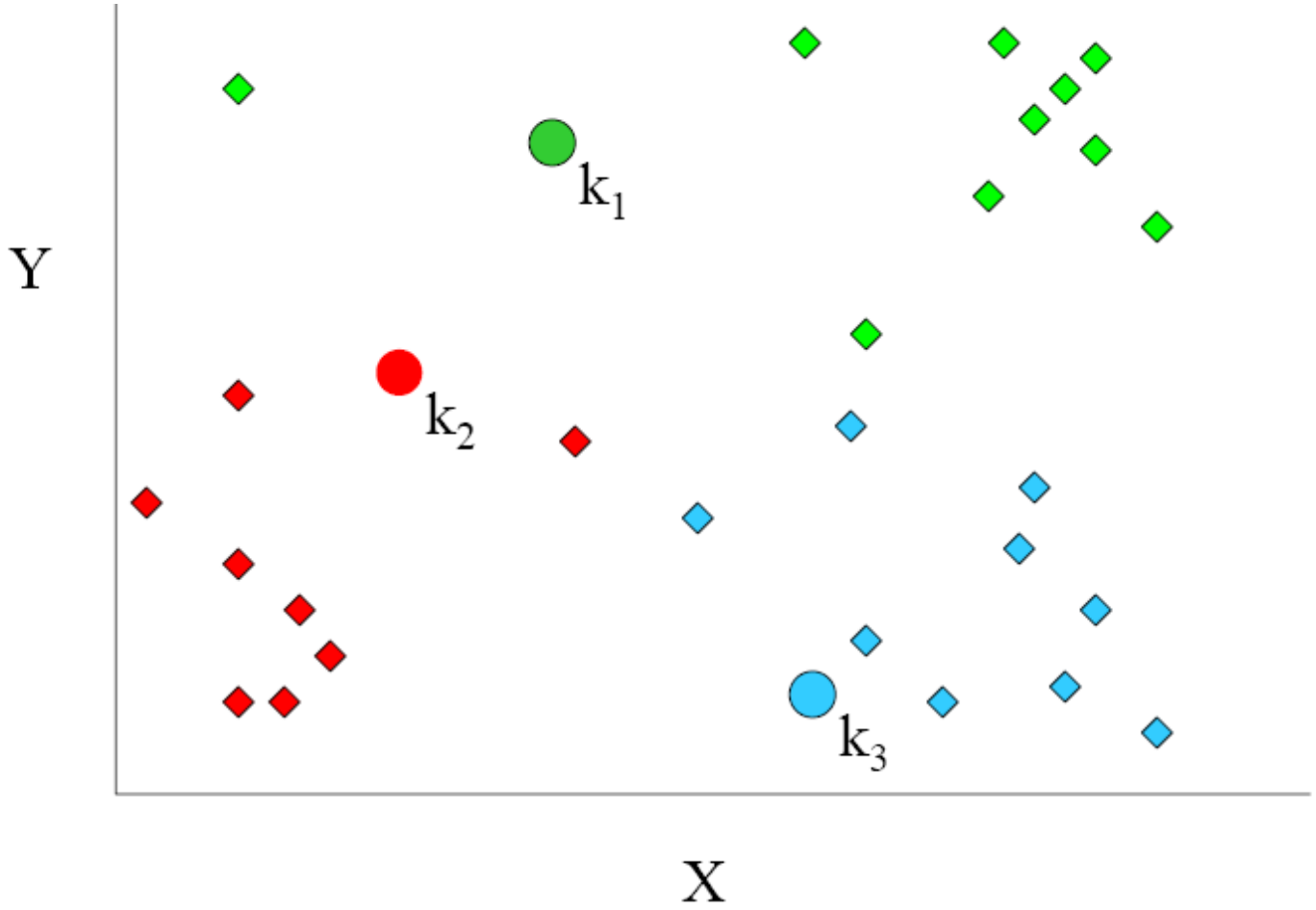
K = 3 συστάδες

Αρχικά σημεία k_1, k_2, k_3



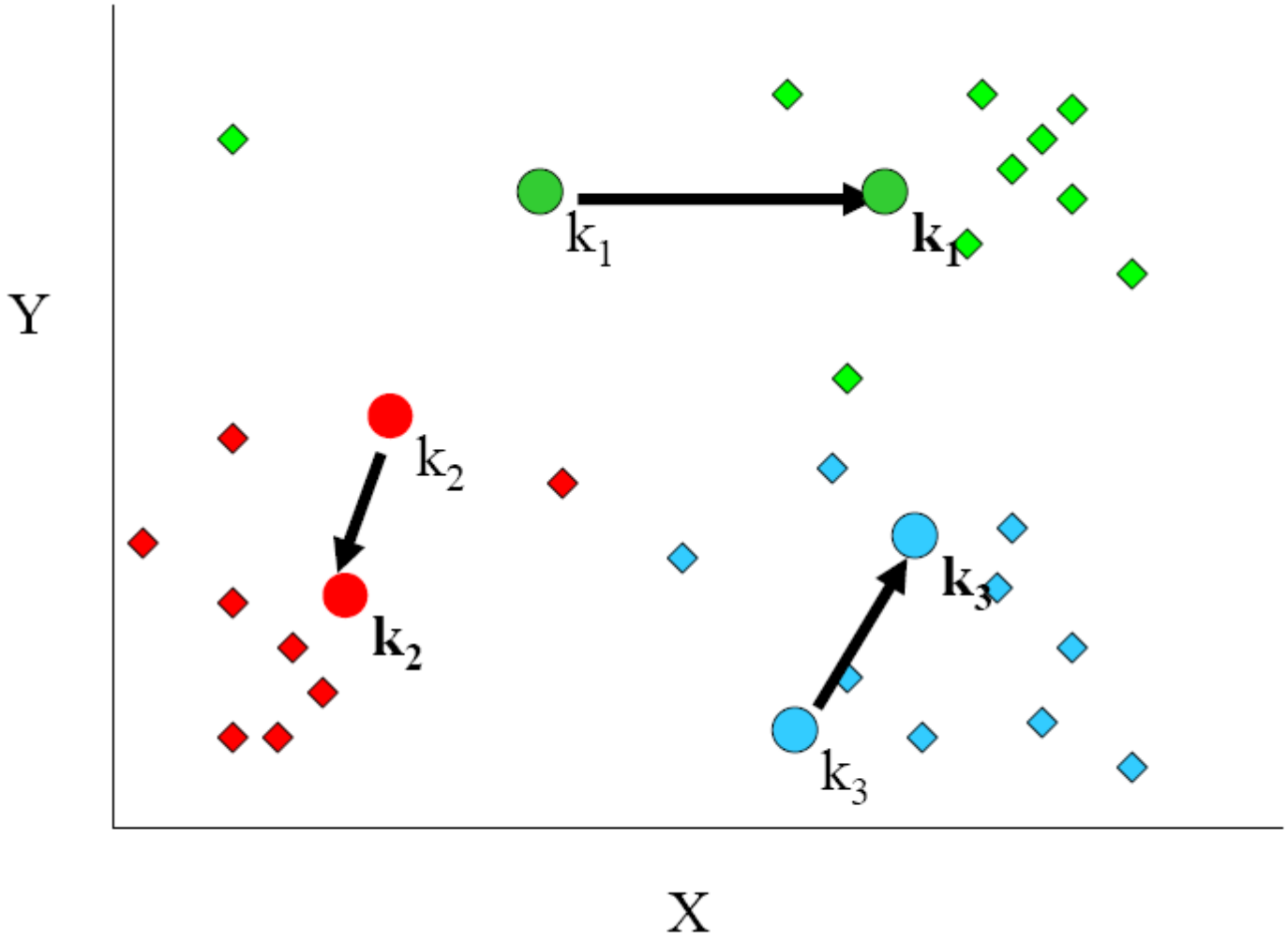
K-means: Βασικός Αλγόριθμος

Τα σημεία ανατίθενται στο πιο γειτονικό από τα 3 αρχικά σημεία



K-means: Βασικός Αλγόριθμος

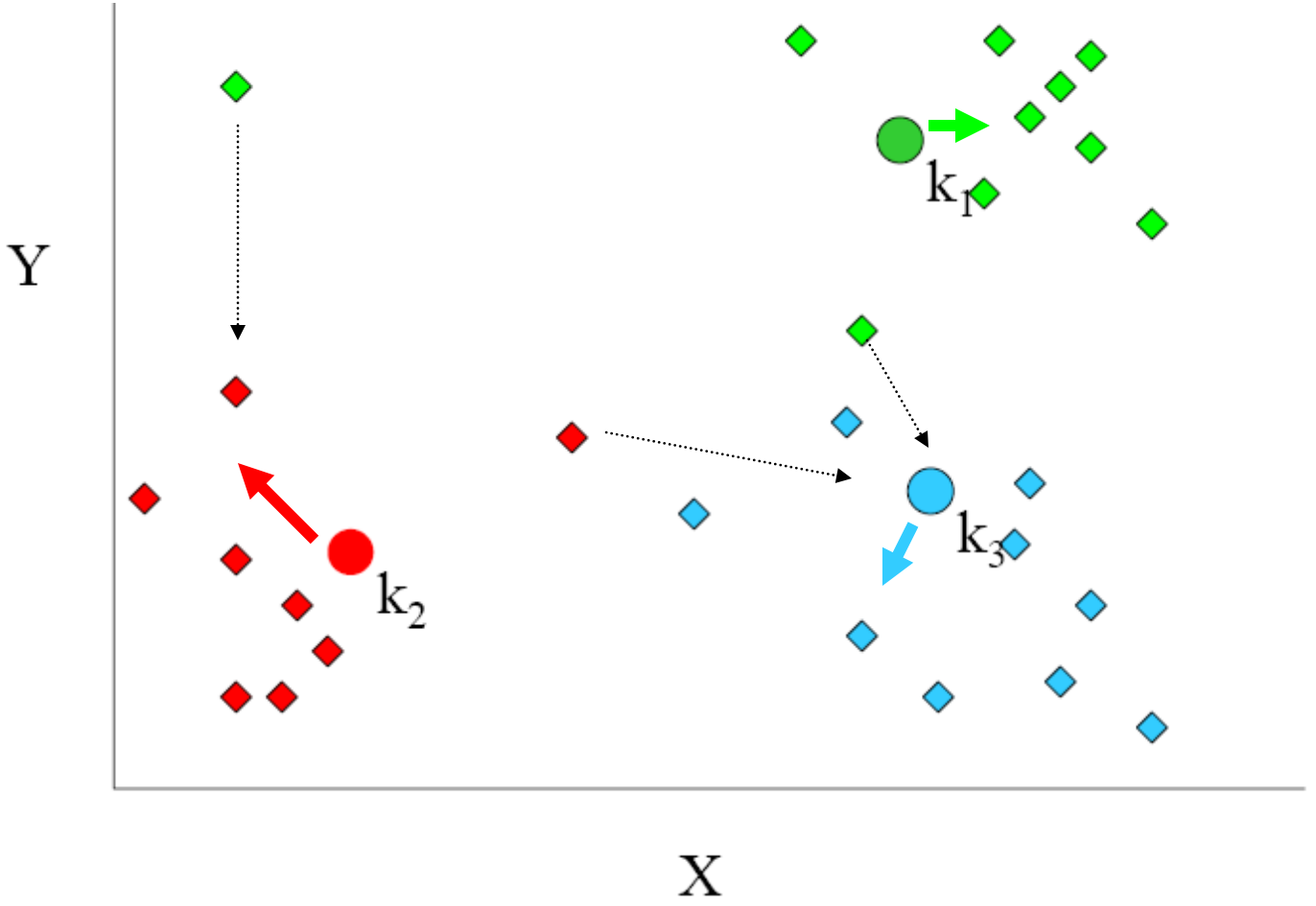
Επανα-υπολογισμός του κέντρου
(κέντρου βάρους) κάθε συστάδας



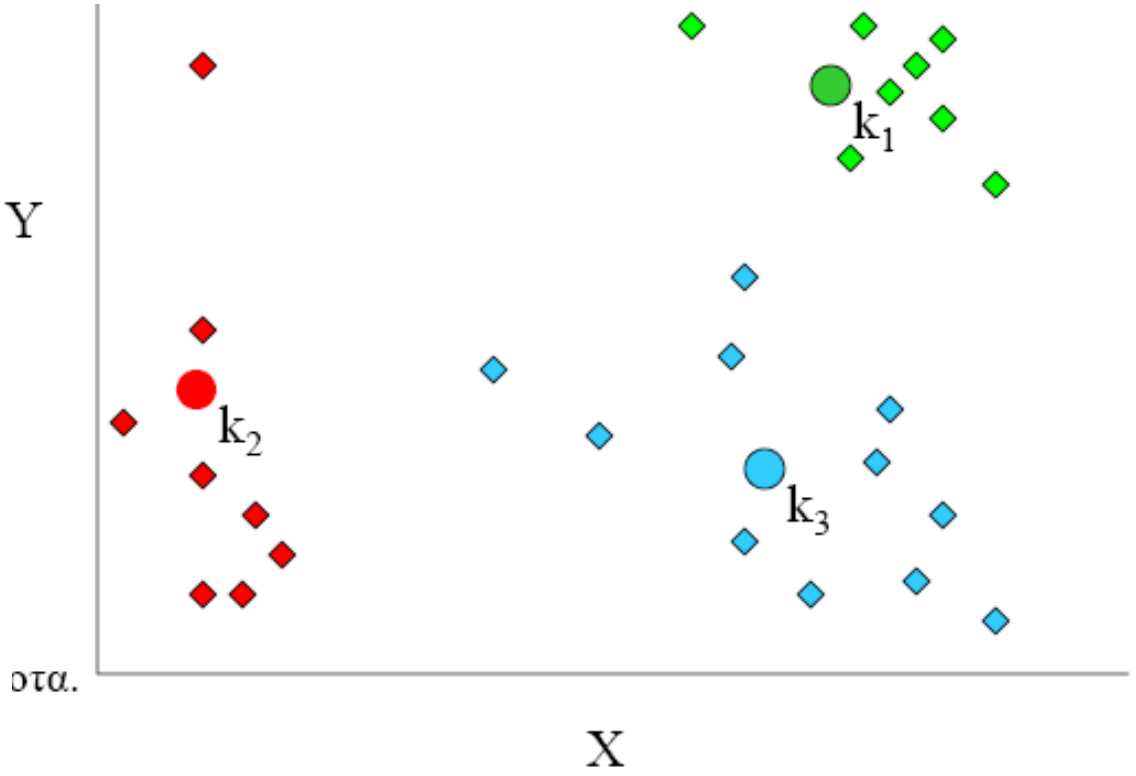
K-means: Βασικός Αλγόριθμος

Νέα ανάθεση των σημείων

Νέα κέντρα βάρους



K-means: Βασικός Αλγόριθμος



Δεν αλλάζει τίποτα -> ΤΕΛΟΣ

Πολυπλοκότητα

- Χώρος: αποθηκεύουμε μόνα τα κέντρα
- Χρόνος: είναι O(I * n * K * d)
  - n = αριθμός σημείων,
  - K = αριθμός συστάδων,
  - I = αριθμός επαναλήψεων,
  - d = αριθμός γνωρισμάτων (διάσταση)
- Για συνηθισμένα μέτρα ομοιότητας, ο αλγόριθμος συγκλίνει
  - Η σύγκλιση συμβαίνει συνήθως τις αρχικές πρώτες επαναλήψεις
- Συχνά η τελική συνθήκη αλλάζει σε
  - Until σχετικά λίγα σημεία να αλλάζουν συστάδα – ή
  - η απόσταση μεταξύ των νέων κεντρικών σημείων από τα παλιά να είναι μικρή
- Το αποτέλεσμα εξαρτάται από την επιλογή των αρχικών σημείων

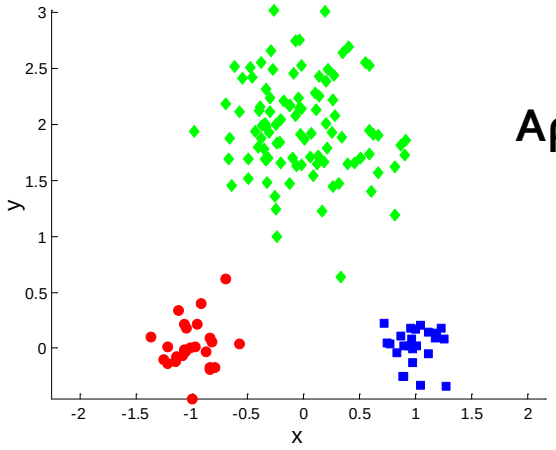
Ποιότητα

- Η πιο συνηθισμένη μέτρηση είναι το άθροισμα των τετράγωνων του λάθους (Sum of Squared Error (SSE))
- Για κάθε σημείο, το λάθος είναι η απόστασή του από την κοντινότερη συστάδα (από το «κέντρο» της)
- Για να πάρουμε το SSE, παίρνουμε το τετράγωνο αυτών των λαθών και τα προσθέτουμε

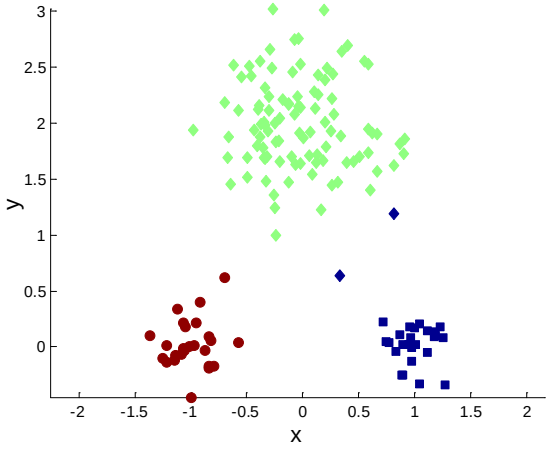
$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$

- Όπου dist η Ευκλείδεια απόσταση, x είναι ένα σημείο στη συστάδα Ci και mi είναι ο αντιπρόσωπος (κεντρικό σημείο) της συστάδας Ci
- Μπορούμε να δείξουμε ότι το σημείο που ελαχιστοποιεί το SSE για τη συστάδα είναι ο μέσος όρος $c_i = 1/m_i \sum_{x \in Ci} x$
- Δοθέντων δύο συστάδων, μπορούμε να επιλέξουμε αυτήν με το μικρότερο λάθος

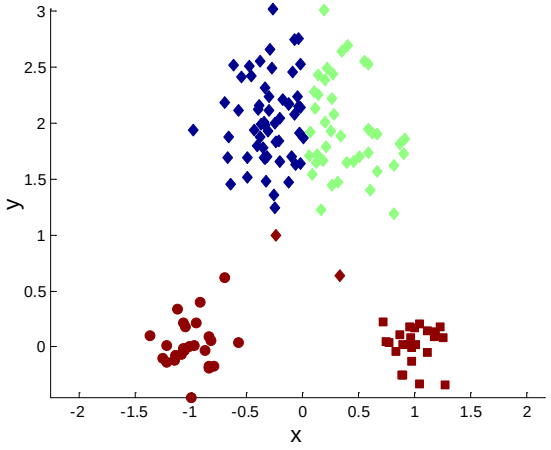
K-means: Παράδειγμα



Αρχικά σημεία

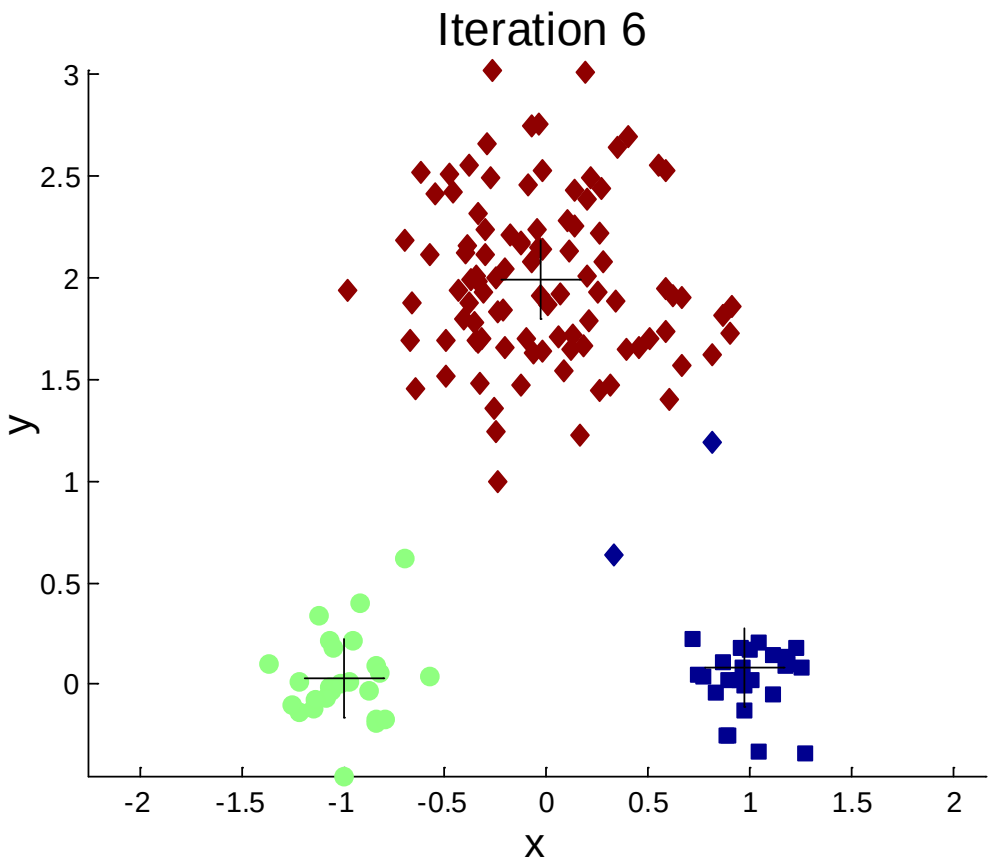


Βέλτιστη συσταδοποίηση

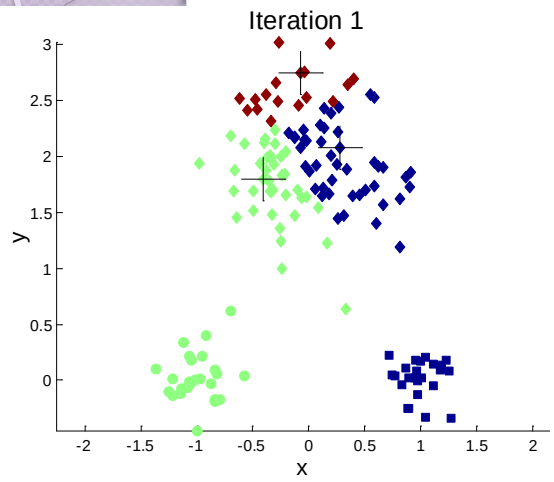


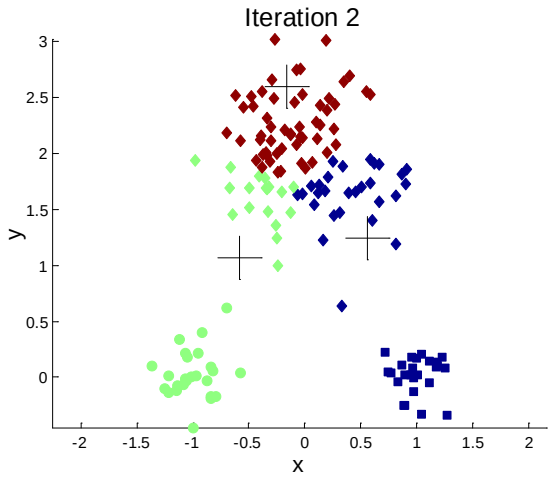
Υπό-βέλτιστη συσταδοποίση

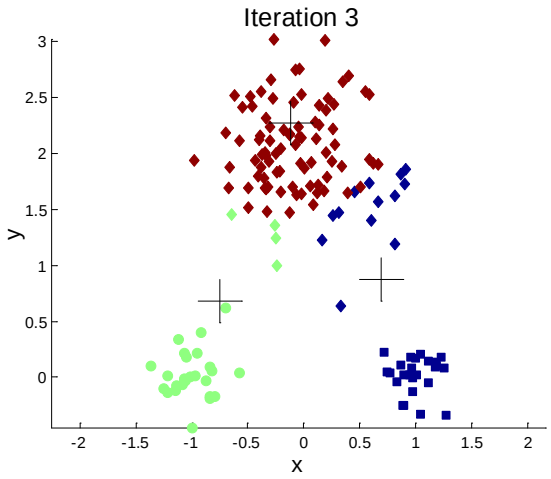
K-means: Επιλογή αρχικών σημείων

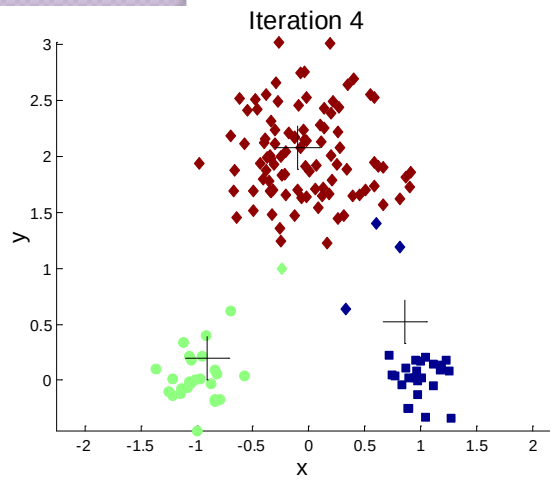


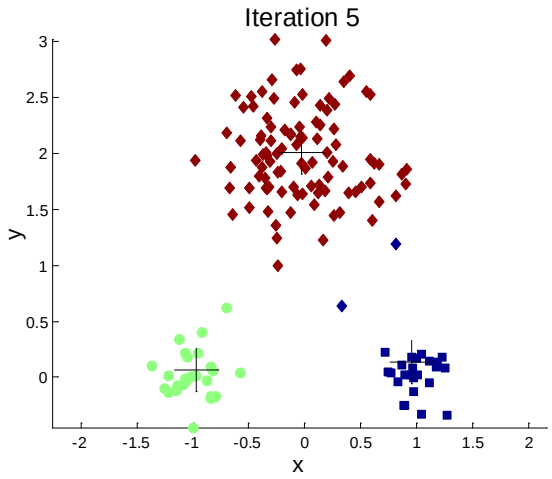
K-means: Επιλογή αρχικών σημείων

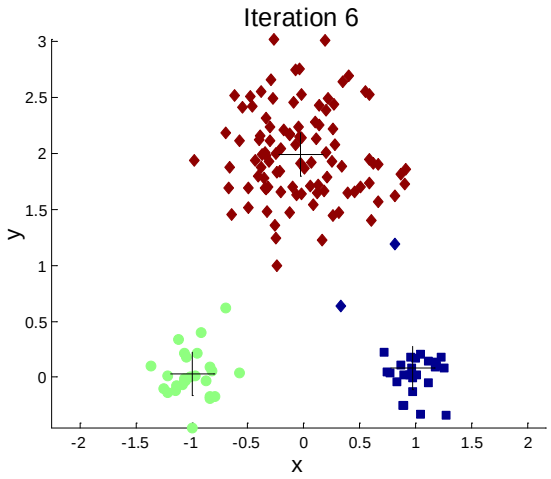




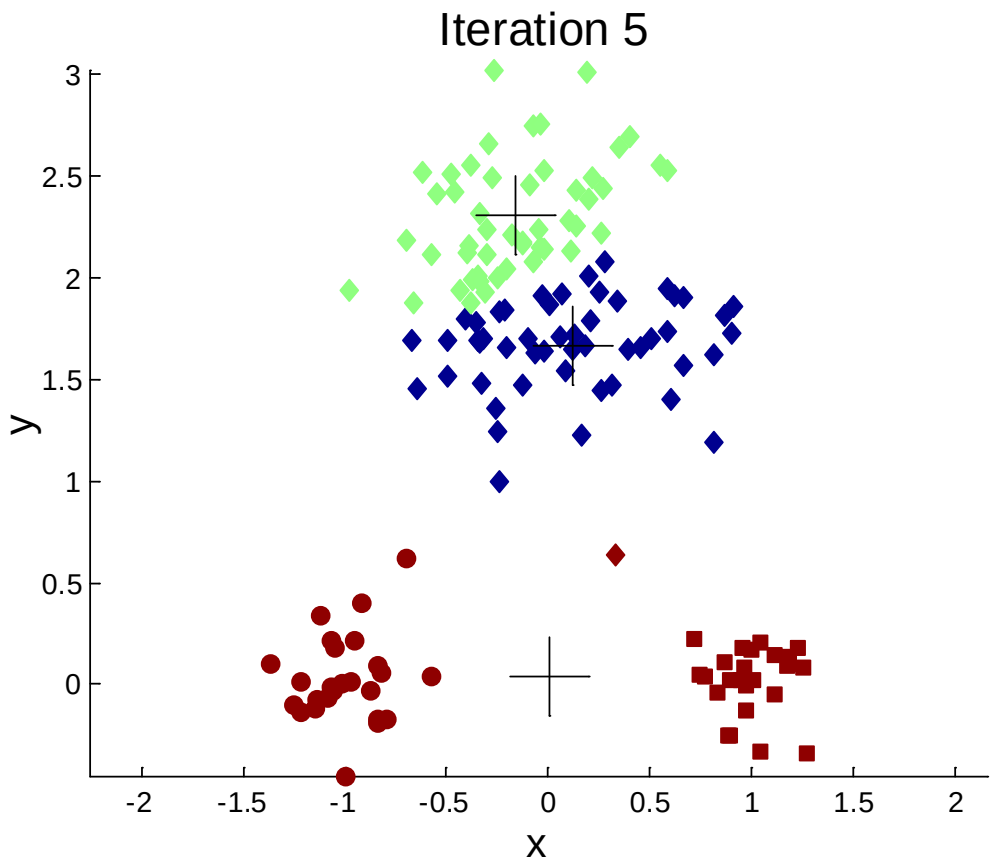




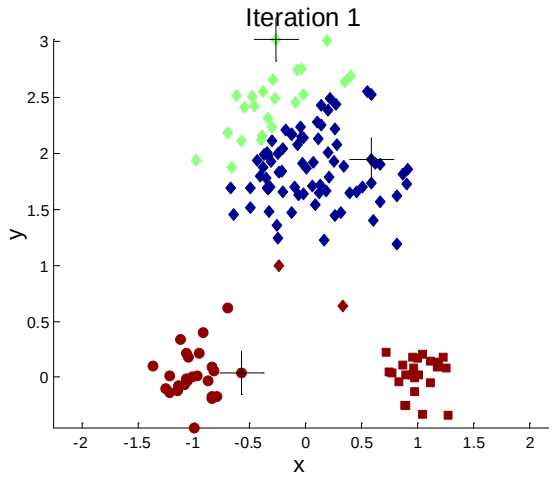


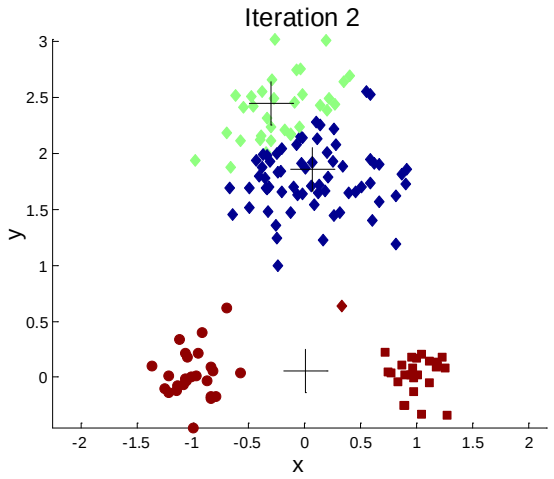


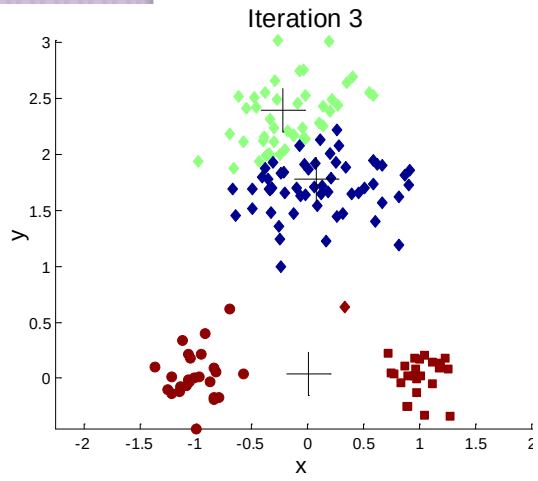
K-means: Επιλογή αρχικών σημείων

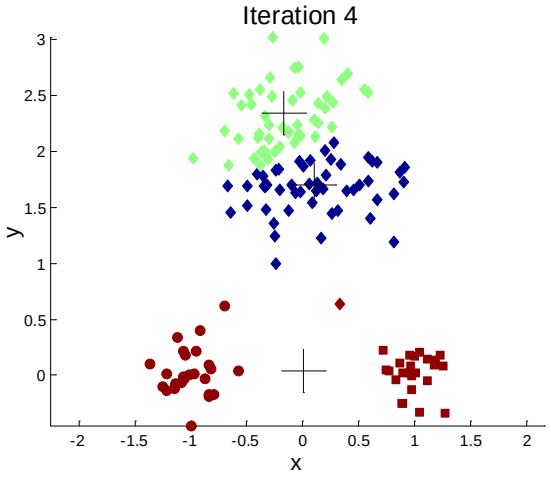


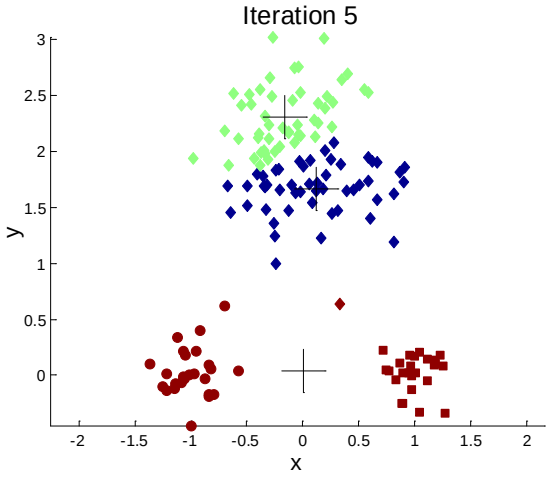
K-means: Επιλογή αρχικών σημείων



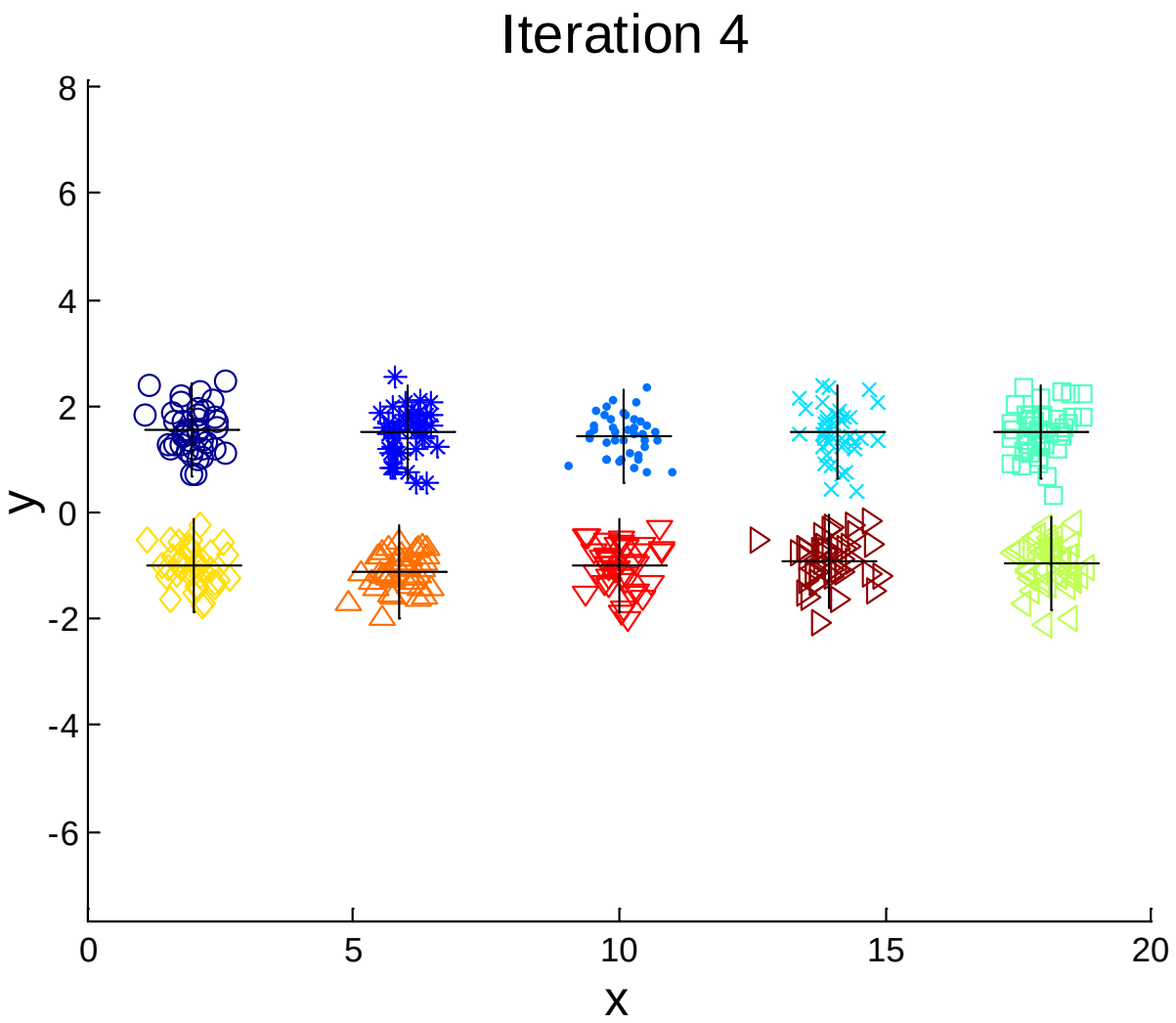






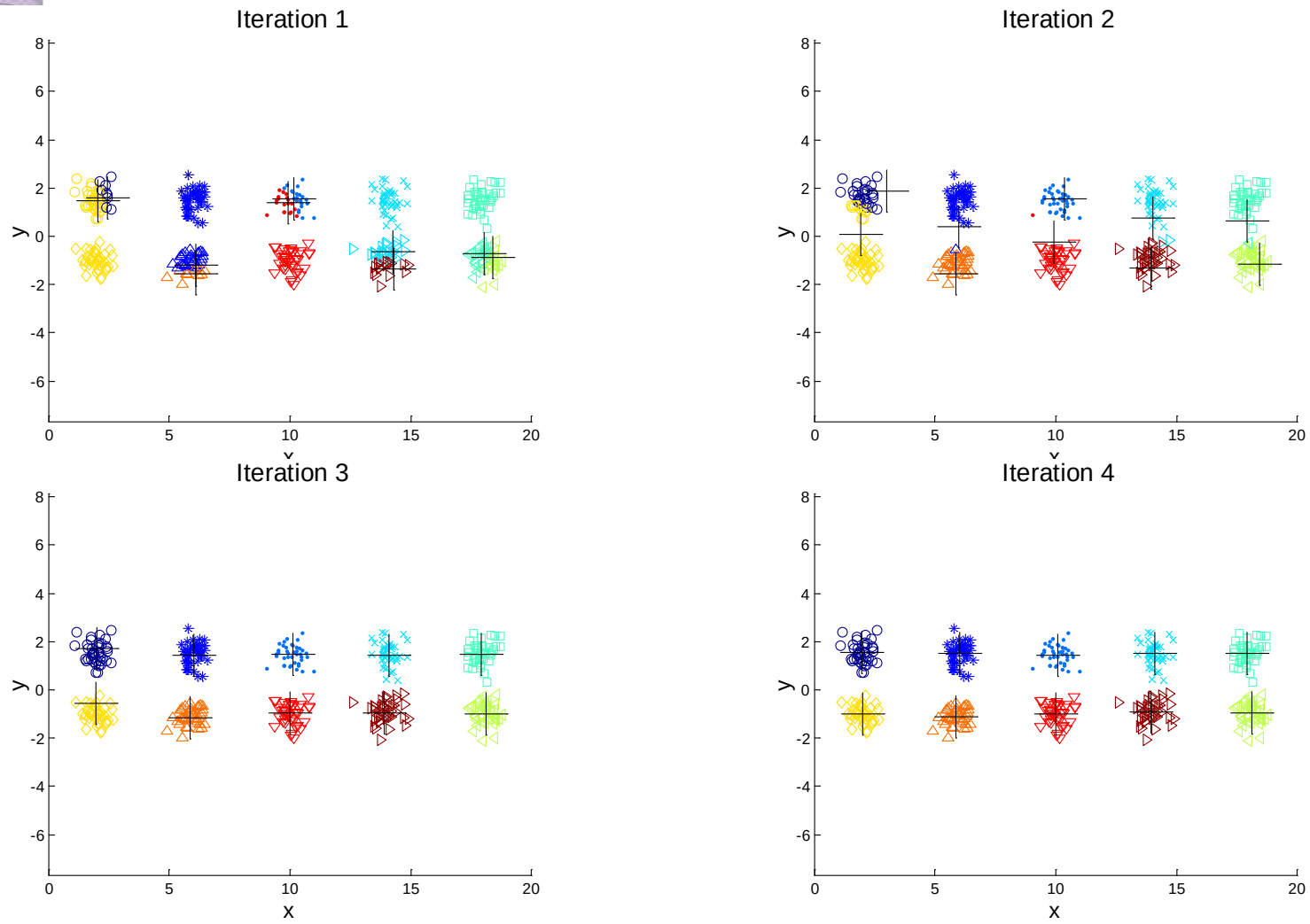


Παράδειγμα 10 συστάδων



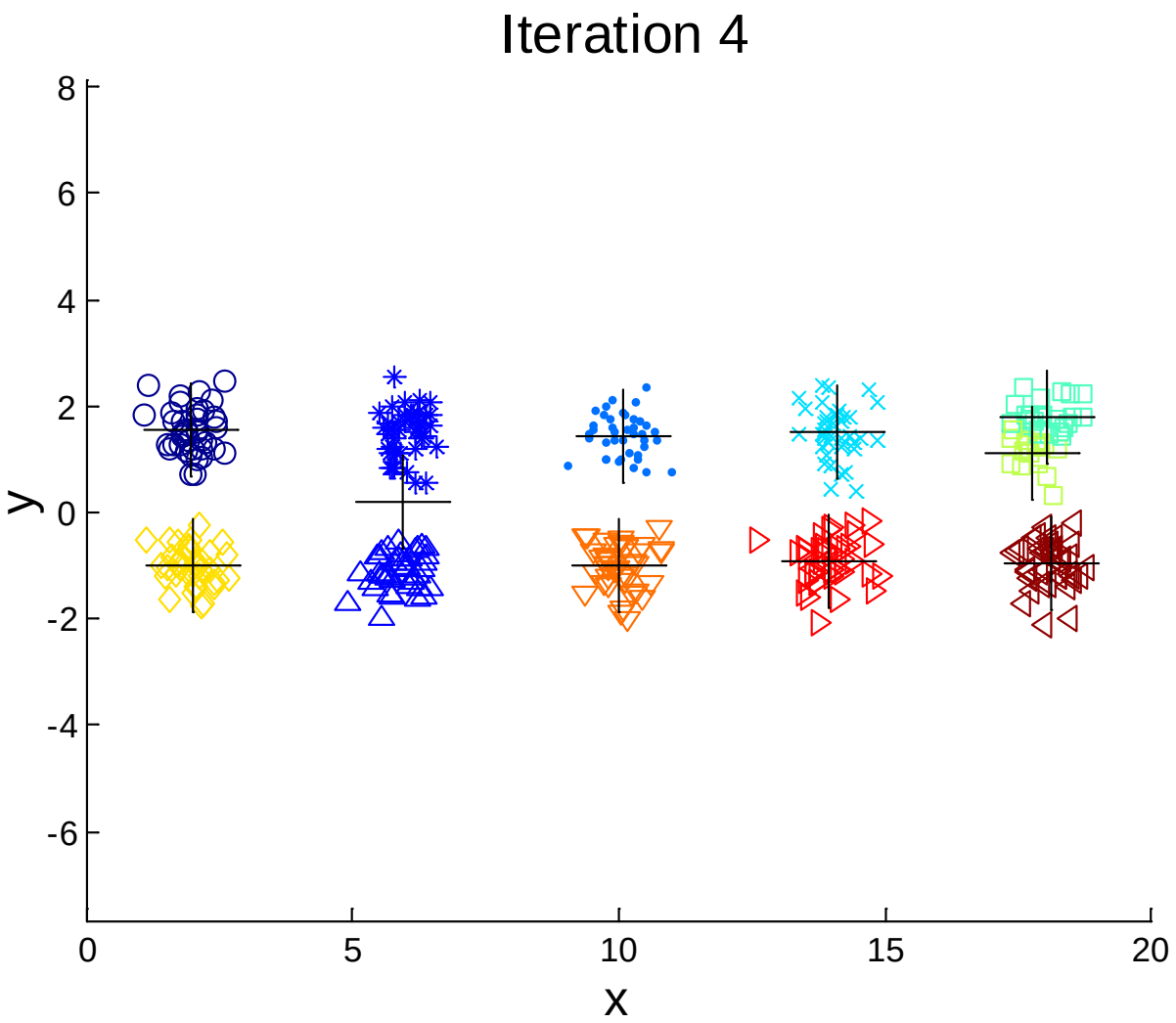
Ξεκινώντας με δύο αρχικά σημεία σε κάθε συστάδα κάθε ζεύγους συστάδων

Παράδειγμα 10 συστάδων



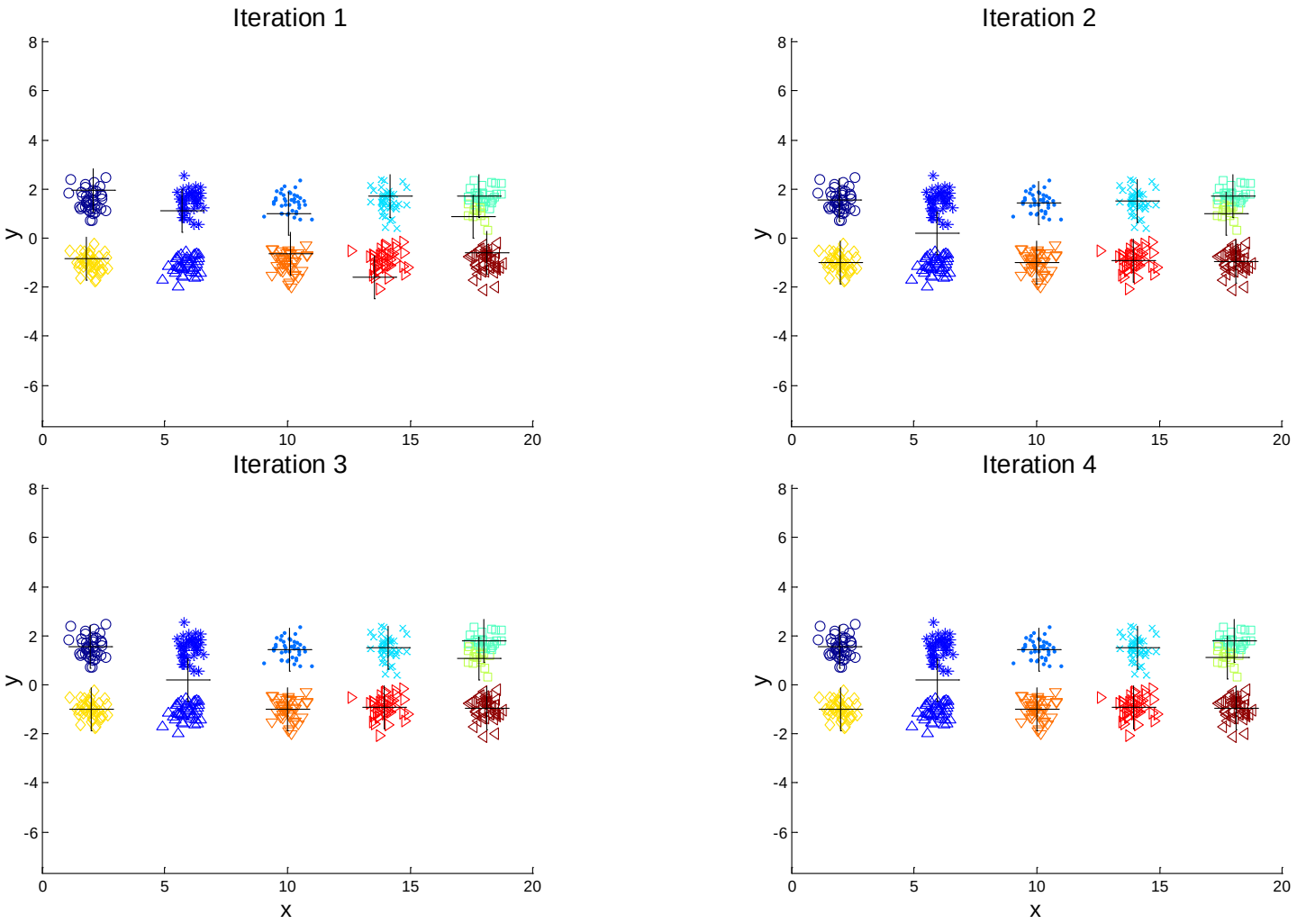
Ξεκινώντας με δύο αρχικά σημεία σε κάθε συστάδα κάθε ζεύγους συστάδων

Παράδειγμα 10 συστάδων



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

Παράδειγμα 10 συστάδων



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

K-means: Λύσεις για την επιλογή αρχικών σημείων

- Πολλαπλά τρεξίματα: Βοηθά, αλλά πολλές περιπτώσεις
- Δειγματοληψία και χρήση κάποιας ιεραρχικής τεχνικής
- Επιλογή m αρχικών σημείων (m>k) και μετά επιλογή k από αυτά τα αρχικά κεντρικά σημεία (πχ τα πιο απομακρυσμένα μεταξύ τους)
- Σταδιακή επιλογή
  - Επιλογή του πρώτου σημείου τυχαία ή ως το μέσο όλων των σημείων
  - Για καθένα από τα υπόλοιπα αρχικά σημεία επέλεξε αυτό που είναι πιο μακριά από τα μέχρι τώρα επιλεγμένα αρχικά σημεία
  - Μπορεί να οδηγήσει στην επιλογή outliers
  - Ο υπολογισμός του πιο απομακρυσμένου σημείου είναι δαπανηρός
  - Συχνά εφαρμόζεται σε δείγματα

Προβλήματα - Παραλλαγές

- Ο βασικός αλγόριθμος μπορεί να οδηγήσει σε άδειες αρχικές συστάδες
  - 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
- Μια παραλλαγή είναι να ενημερώνονται τα κέντρα μετά από κάθε ανάθεση (incremental approach)
  - Πιο δαπανηρό. Έχει σημασία η σειρά εισαγωγής/εξέτασης των σημείων. Δεν υπάρχουν άδειες συστάδες
- Split-Merge (διατηρώντας το ίδιο K)
  - Διαχωρισμός (split) συστάδων με το σχετικά μεγαλύτερο SSE
  - Δημιουργία μια νέας συστάδας: πχ επιλέγοντας το σημείο που είναι πιο μακριά από όλα τα κέντρα ή τυχαία επιλογή σημείου
  - Συνένωση (merge) συστάδων που είναι σχετικά κοντινές (τα κέντρα τους έχουν την μικρότερη απόσταση) ή τις δυο συστάδες που οδηγούν στην μικρότερη αύξηση του SSE
  - Διαγραφή συστάδας και ανακατανομή των σημείων της σε άλλες συστάδες (αυτό που οδηγεί στην μικρότερη αύξηση του SSE)

K-means με διχοτόμηση (bisecting k-means)

Παραλλαγή που μπορεί να παράγει μια διαχωριστική ή ιεραρχική συσταδοποίηση

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

2: **Repeat**

3: Επιλογή[+] μιας συστάδας από τη λίστα των συστάδων

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

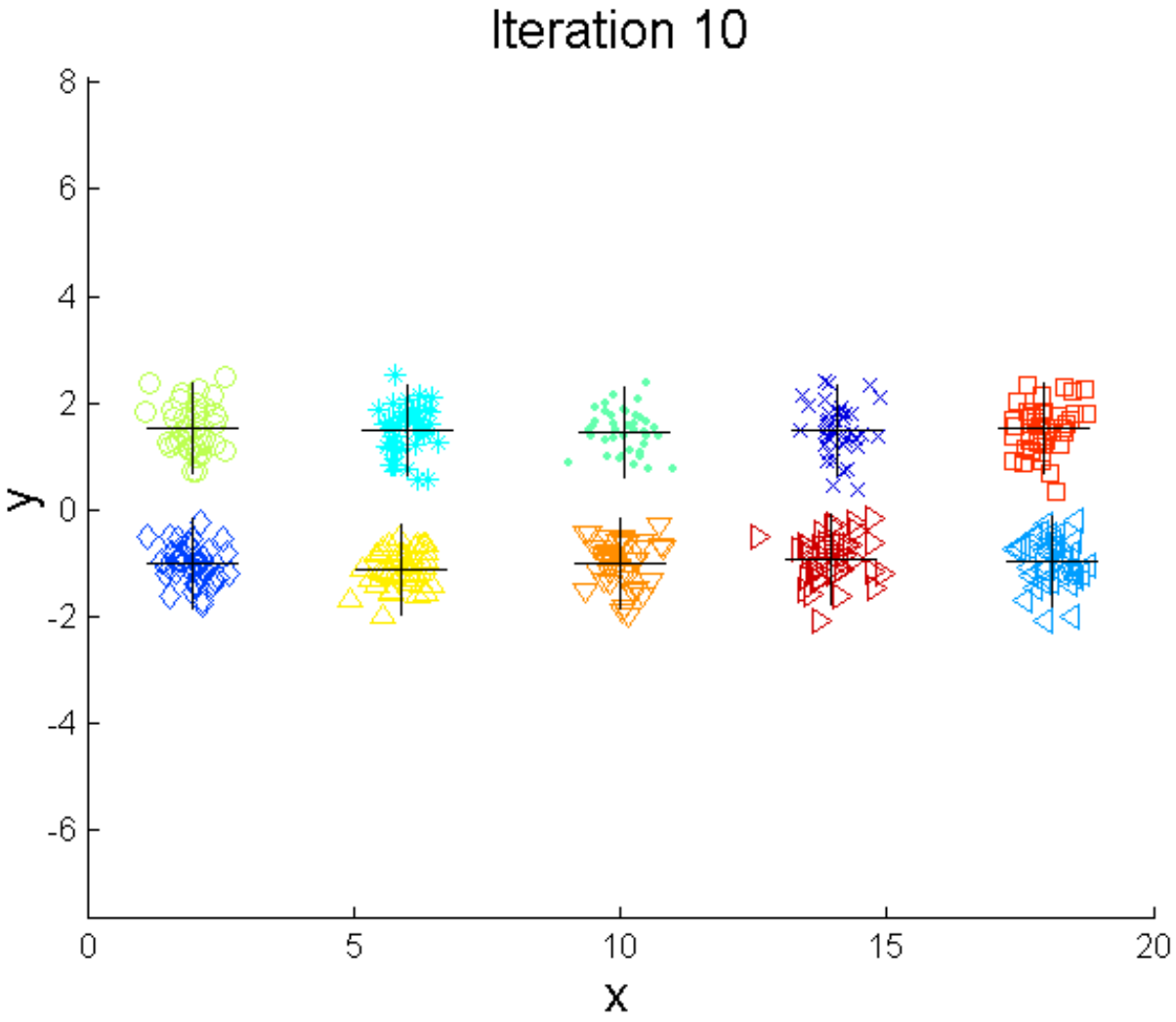
5: διχοτόμησε την επιλεγμένη συστάδα χρησιμοποιώντας το βασικό k-means

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

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

[+] Επιλογή της μεγαλύτερης – Αυτής με το μεγαλύτερο SSE

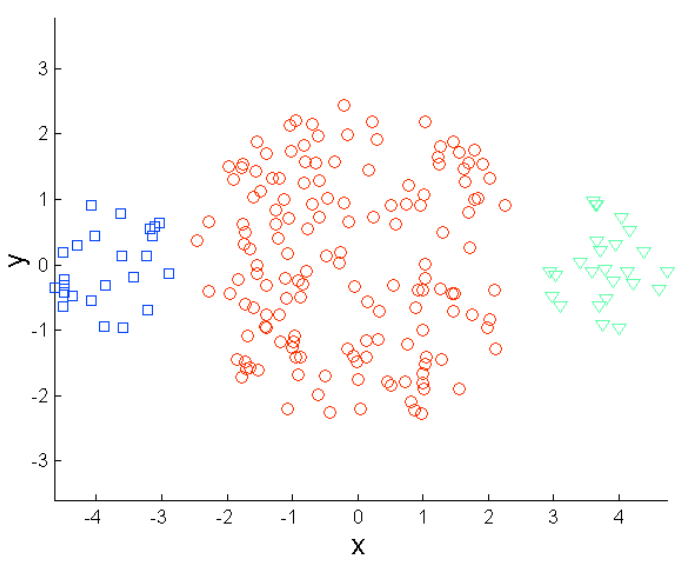
Παράδειγμα



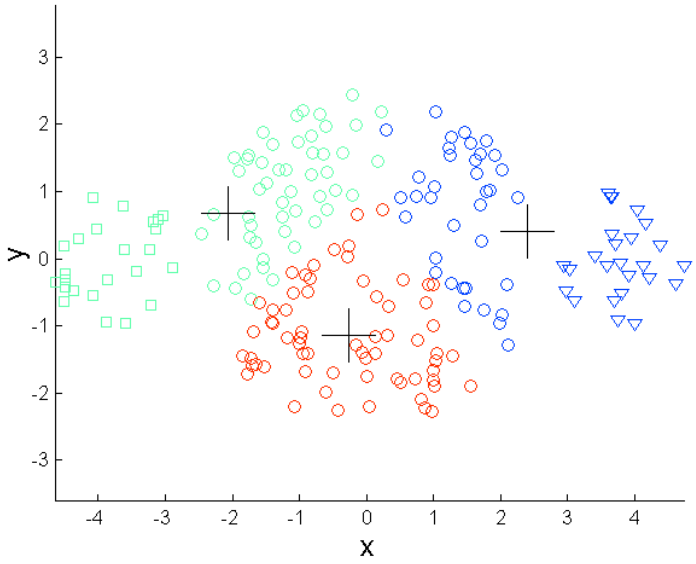
K-means: Περιορισμοί

- Ο K-means έχει προβλήματα όταν οι συστάδες έχουν
  - Διαφορετικά Μεγέθη
  - Διαφορετικές Πυκνότητες
  - Μη κυκλικά σχήματα
- Έχει προβλήματα όταν τα δεδομένα έχουν outliers

K-means: Περιορισμοί – διαφορετικά μεγέθη



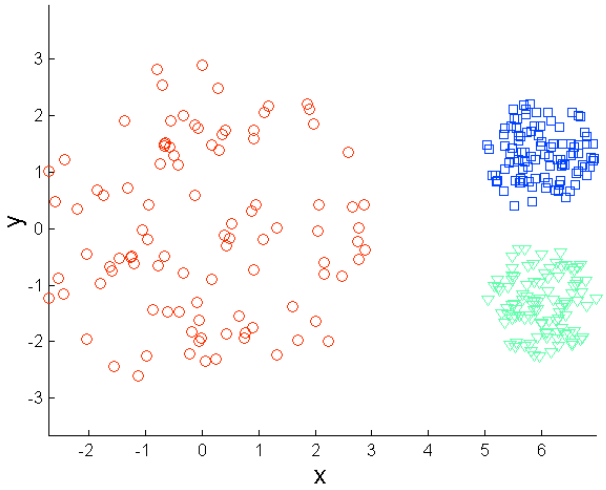
Αρχικά σημεία



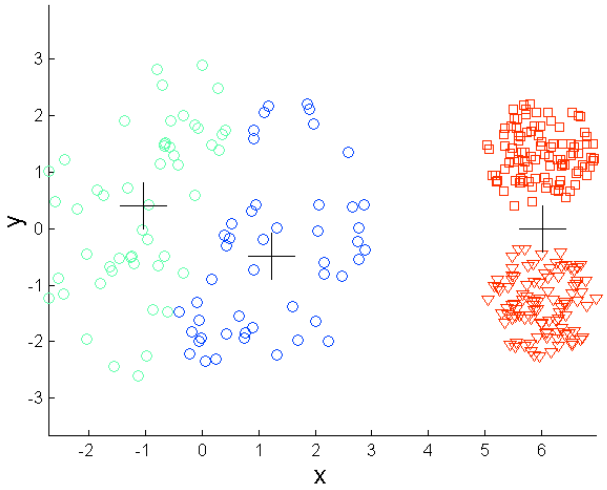
K-means (3 συστάδες)

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

K-means: Περιορισμοί – διαφορετικές πυκνότητες



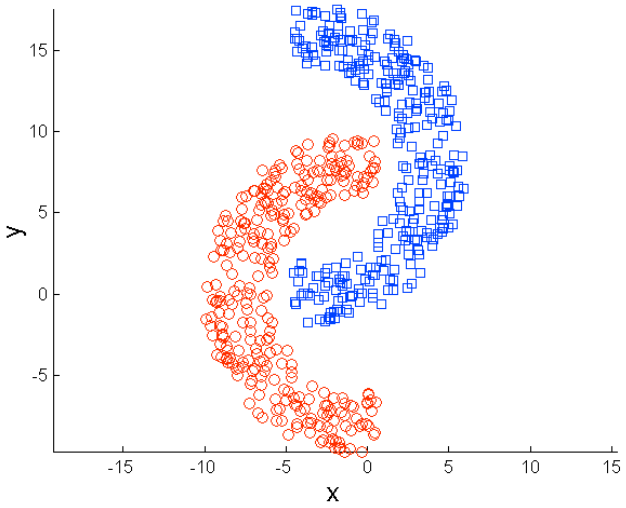
Αρχικά σημεία

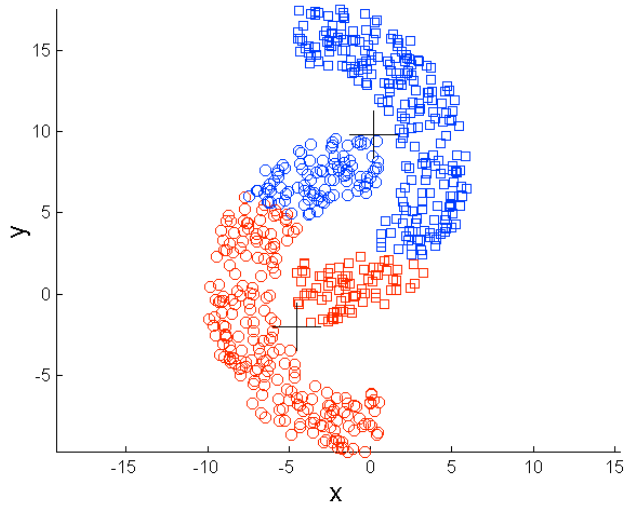


K-means (3 συστάδες)

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

K-means: Περιορισμοί – μη κυκλικά σχήματα





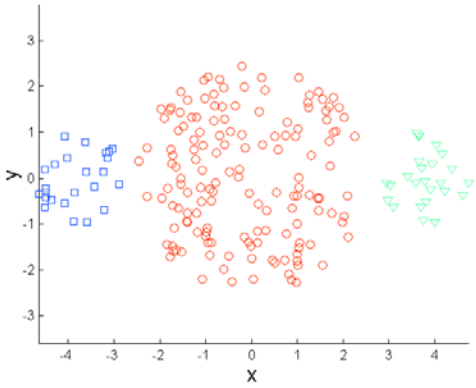
Αρχικά σημεία

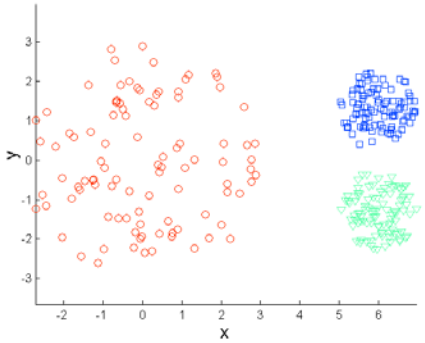
K-means (2 συστάδες)

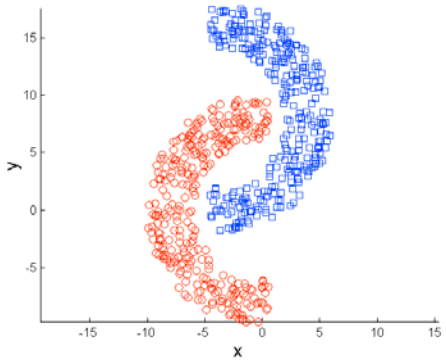
Δεν μπορεί να βρει τις δύο συστάδες γιατί έχουν μη κυκλικά σχήματα

Λύσεις

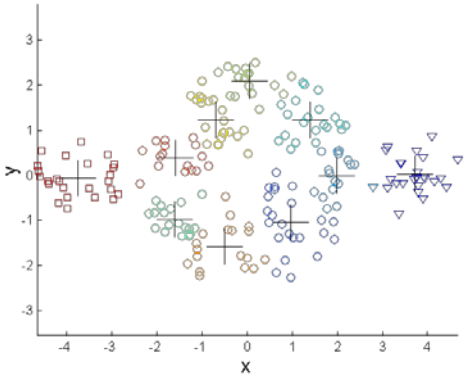
Αρχικά Σημεία

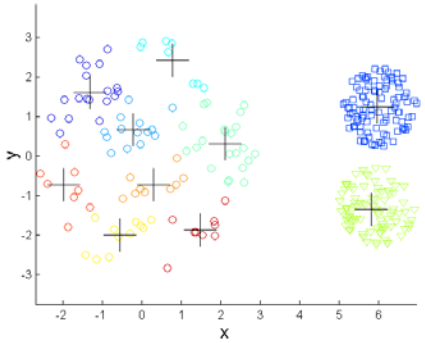


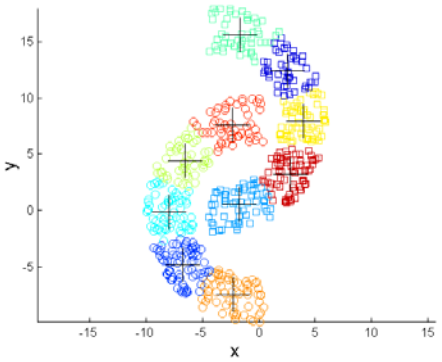




K-means Συστάδες







Μια λύση είναι να χρησιμοποιηθούν πολλές συστάδες

Βρίσκει τμήματα των συστάδων, αλλά πρέπει να τα συγκεντρώσουμε

K-means: Επιλογή αρχικών σημείων

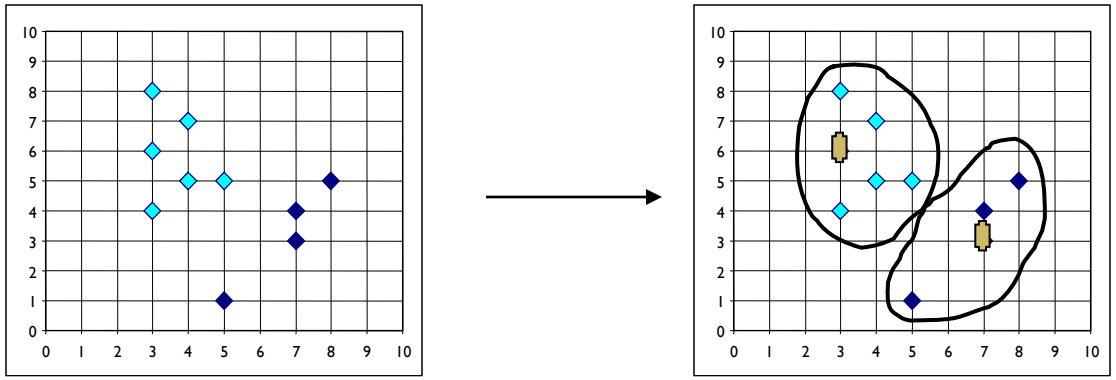
Αν υπάρχουν K «πραγματικές συστάδες» η πιθανότητα να επιλέξουμε ένα κέντρο από κάθε συστάδα είναι μικρή, συγκεκριμένα αν όλες οι συστάδες έχουν το ίδιο μέγεθος n, τότε:

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

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

K-medoid

- Συνήθως συνεχή d-διάστατο χώρο
- Διαλέγει ένα αντιπροσωπευτικό σημείο από τα δεδομένα και ελαχιστοποιεί την απόσταση από αυτό – Medoid: το πιο κεντρικό σημείο της συστάδας (αντί να χρησιμοποιεί το mean)
- Μειώνει την ευαισθησία σε outliers
- Μπορεί να εφαρμοστεί σε δεδομένα οποιουδήποτε τύπου (πχ και για κατηγορικά δεδομένα)



Ιεραρχική Συσταδοποίηση

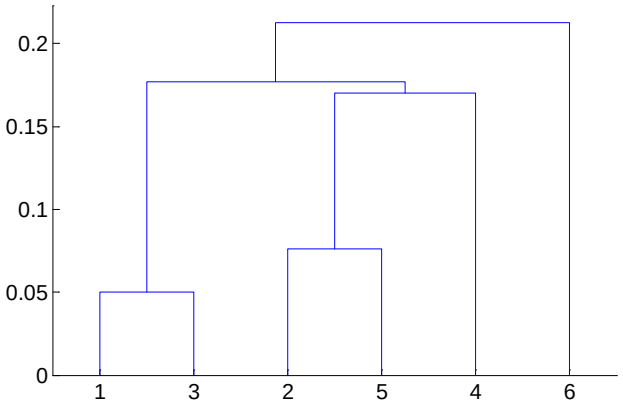
Hierarchical Agglomerative Clustering

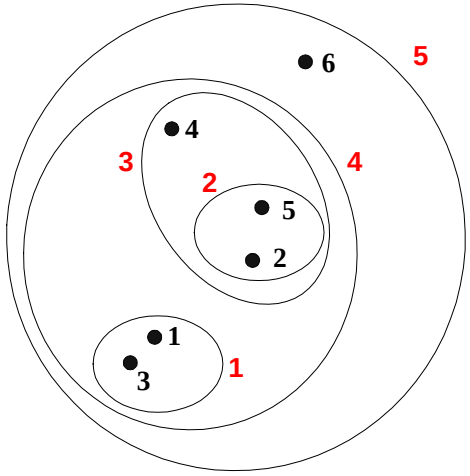
Ιεραρχική Συσταδοποίηση: Βασικά

Παράγει ένα σύνολο από εμφωλευμένες συστάδες οργανωμένες σε ένα ιεραρχικό δέντρο

Μπορεί να παρασταθεί με ένα δένδρο-γραμμα

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





Ιεραρχική Συσταδοποίηση: Πλεονεκτήματα

- Δε χρειάζεται να υποθέσουμε ένα συγκεκριμένο αριθμό από συστάδες
  - Οποιοσδήποτε επιθυμητός αριθμός από συστάδες μπορεί να επιτευχθεί κόβοντας το δενδρόγραμμα στο κατάλληλο επίπεδο
- Χρησιμοποιούμε ένα πίνακα ομοιότητα ή απόστασης
- Διαχωρισμός ή συγχώνευση μιας ομάδας τη φορά
- Μπορεί να αντιστοιχούν σε λογικές ταξινομήσεις
  - Για παράδειγμα στις βιολογικές επιστήμες (ζωικό βασίλειο, phylogeny reconstruction, …)

Ιεραρχική Συσταδοποίηση

Δυο βασικοί τύποι ιεραρχικής συσταδοποίησης

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

Συσσωρευτική Ιεραρχική Συσταδοποίηση (ΣΙΣ-HAC)

Η πιο δημοφιλής τεχνική συσταδοποίησης

Βασικός Αλγόριθμος

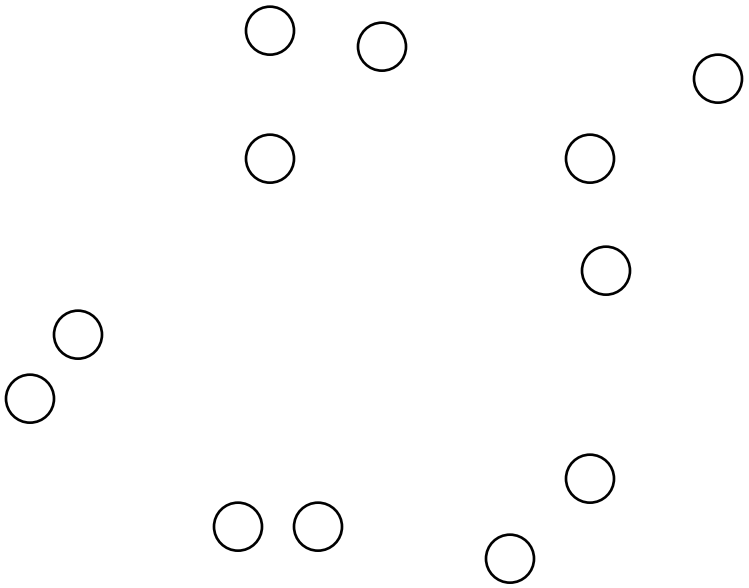
1: Υπολογισμός του Πίνακα Γειτνίασης (Adjacency Matrix)
2: Έστω κάθε σημείο αποτελεί και μια συστάδα
3: **Repeat**
4: Συγχώνευση των δύο κοντινότερων συστάδων
5: Ενημέρωση του Πίνακα Γειτνίασης
6: **Until** να μείνει μία μόνο συστάδα

Βασική λειτουργία είναι ο υπολογισμός της γειτνίασης δυο συστάδων

Διαφορετικοί αλγόριθμοι με βάση το πως ορίζεται η απόσταση ανάμεσα σε δύο συστάδες

Συσσωρευτική Ιεραρχική Συσταδοποίηση (HAC)

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

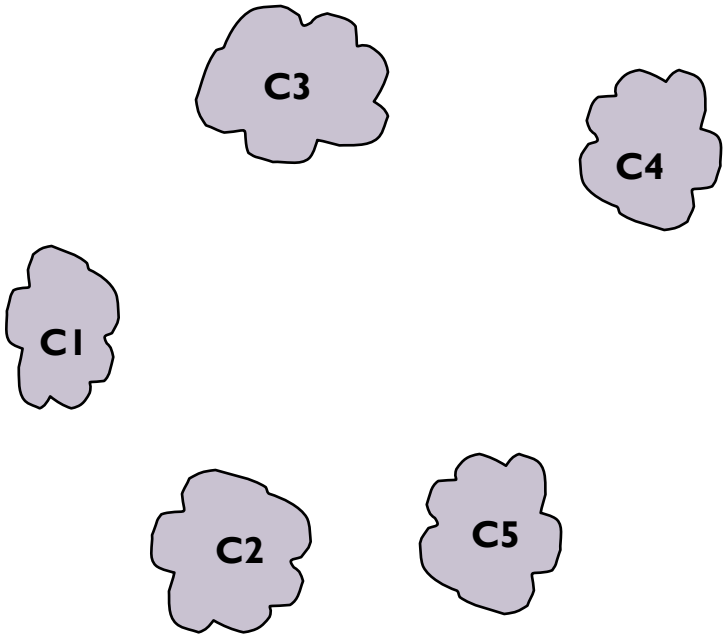


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Πίνακας Γειτνίασης

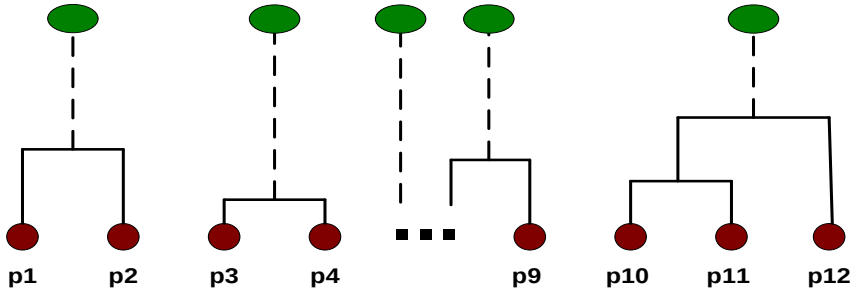
Συσσωρευτική Ιεραρχική Συσταδοποίηση (HAC)

Μετά από κάποιες συγχωνεύσεις, έχουμε κάποιες συστάδες



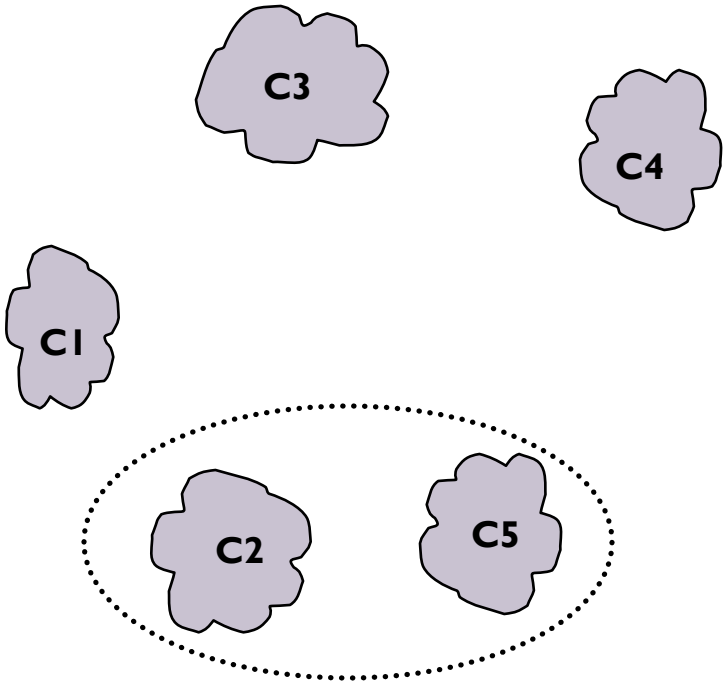
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Πίνακας Γειτνίασης



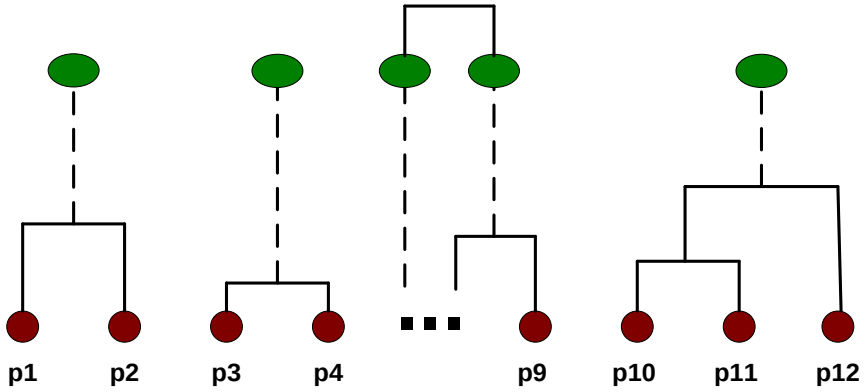
Συσσωρευτική Ιεραρχική Συσταδοποίηση (HAC)

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



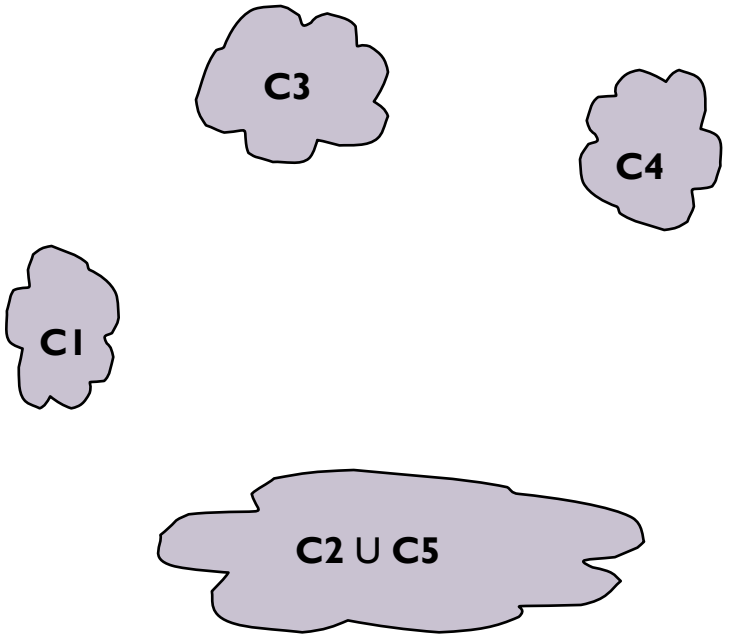
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Πίνακας Γειτνίασης



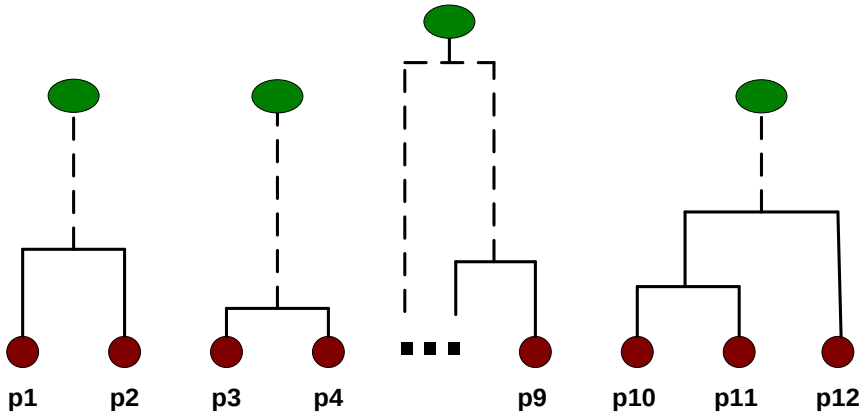
Συσσωρευτική Ιεραρχική Συσταδοποίηση (HAC)

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

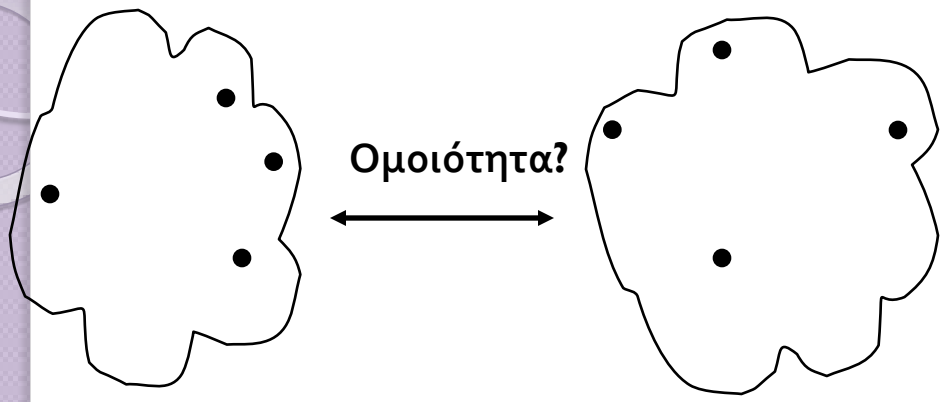


	C1	C2 U C5	C3	C4
C1		?		
C2 U C5	?	?	?	?
C3		?		
C4		?		

Πίνακας Γειτνίασης



HAC: Ορισμός απόστασης μεταξύ συστάδων



- MIN
- MAX
- Μέσος όρος της συστάδας
- Η απόσταση μεταξύ των κεντρικών σημείων
- Άλλες μέθοδοι βασισμένες σε μια αντικειμενική συνάρτηση

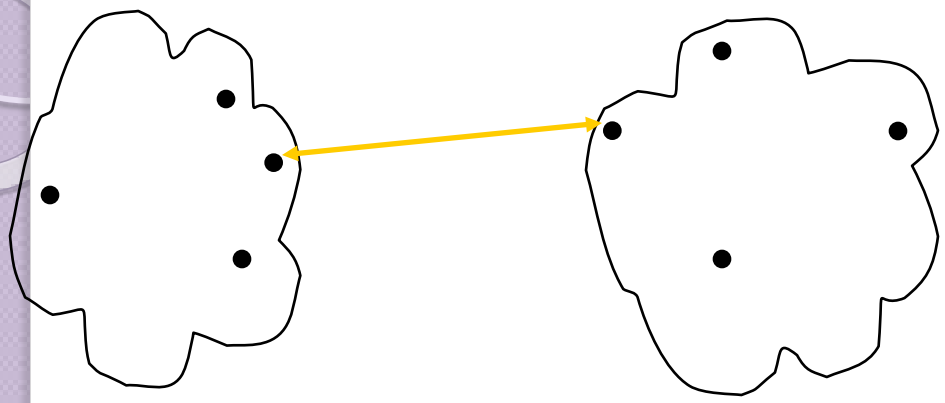
	p1	p2	p3	p4	p5	...
p1						
p2						
p3						
p4						
p5						
.						

.

.

Πίνακας Γειτνίασης

HAC: Ορισμός απόστασης μεταξύ συστάδων



- **MIN**
- MAX
- Μέσος όρος της ομάδας
- Η απόσταση μεταξύ των κεντρικών σημείων
- Άλλες μέθοδοι βασισμένες σε μια αντικειμενική συνάρτηση

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

. .

Πίνακας Γειτνίασης

HAC: Ορισμός απόστασης μεταξύ συστάδων

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

HAC: Ορισμός απόστασης μεταξύ συστάδων

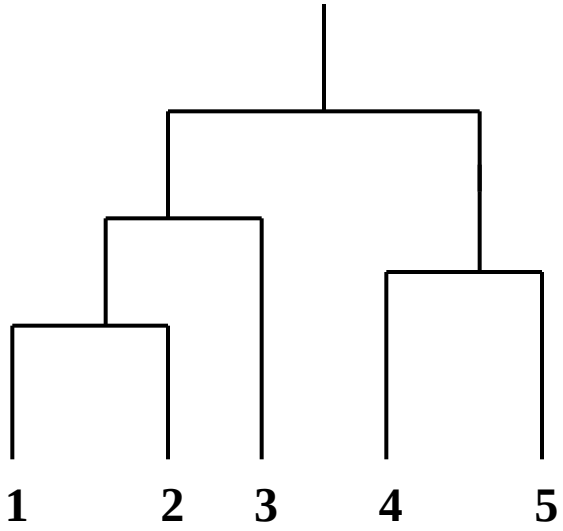
MIN ή μοναδικής ακμής ή απλού συνδέσμου (single link)

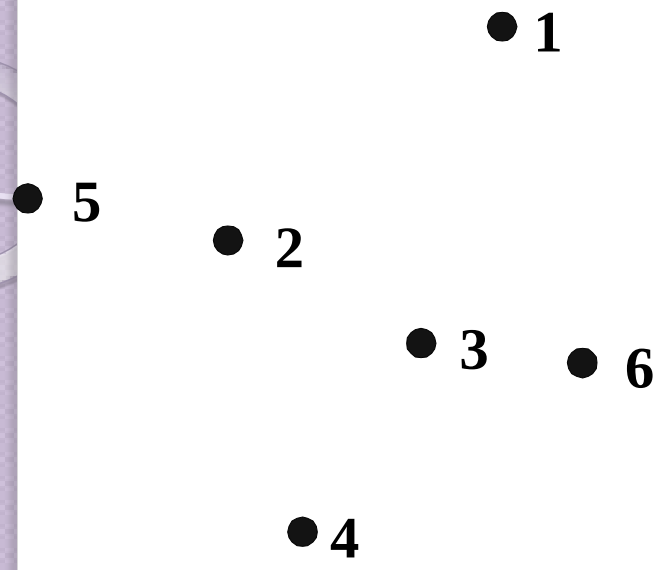
Η ομοιότητα μεταξύ δυο συστάδων βασίζεται στα δυο πιο όμοια (πιο γειτονικά) σημεία στις διαφορετικές συστάδες (με όρους γραφημάτων – shortest edge)

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

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

Προσοχή: ομοιότητα!!





1 (0.4, 0.53)

2 (0.22, 0.38)

3 (0.35, 0.32)

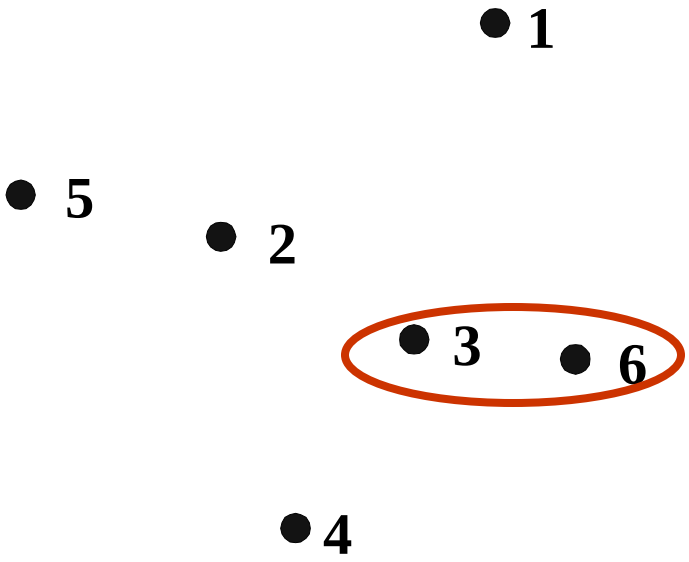
4 (0.26, 0.19)

5 (0.08, 0.41)

6 (0.45, 0.30)

Πίνακας απόστασης (Ευκλείδεια)

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



1 (0.4, 0.53)

2 (0.22, 0.38)

3 (0.35, 0.32)

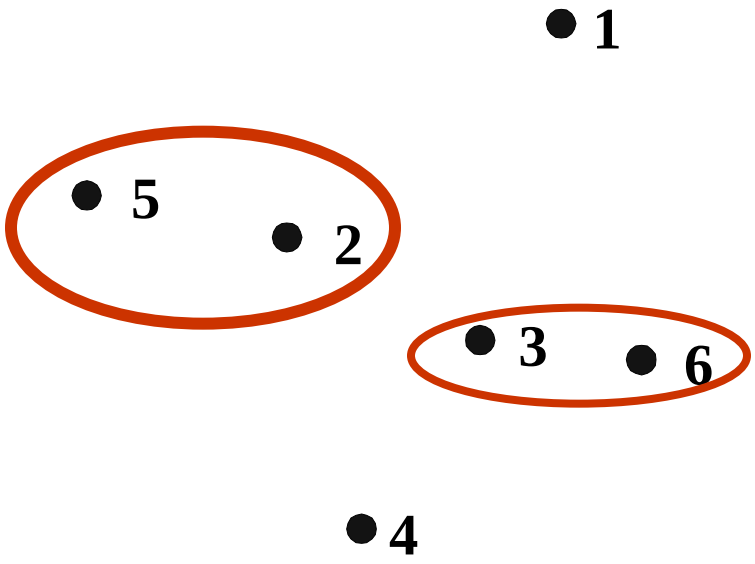
4 (0.26, 0.19)

5 (0.08, 0.41)

6 (0.45, 0.30)

Καθορίζεται μόνο από μια ακμή – την μικρότερη

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



1 (0.4, 0.53)

2 (0.22, 0.38)

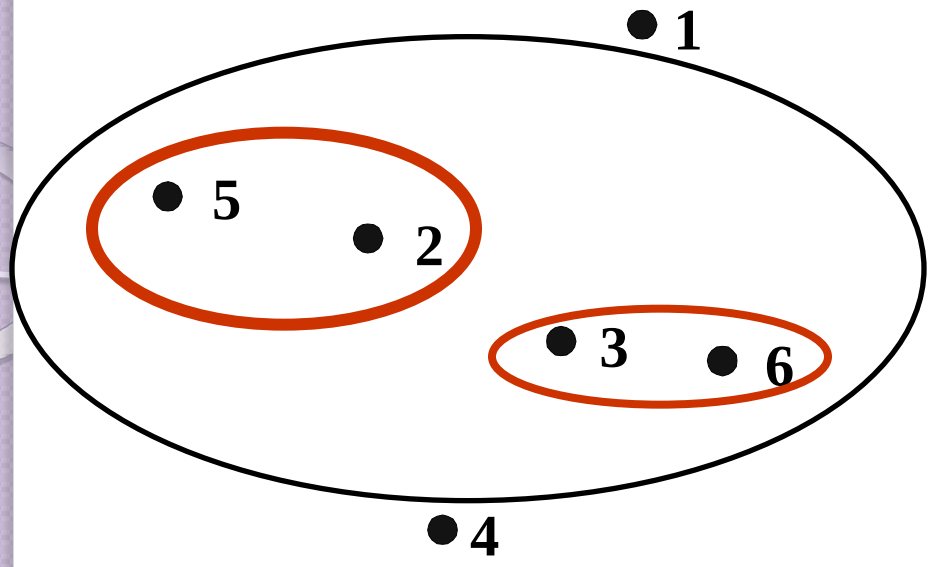
3 (0.35, 0.32)

4 (0.26, 0.19)

5 (0.08, 0.41)

6 (0.45, 0.30)

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



1 (0.4, 0.53)

2 (0.22, 0.38)

3 (0.35, 0.32)

4 (0.26, 0.19)

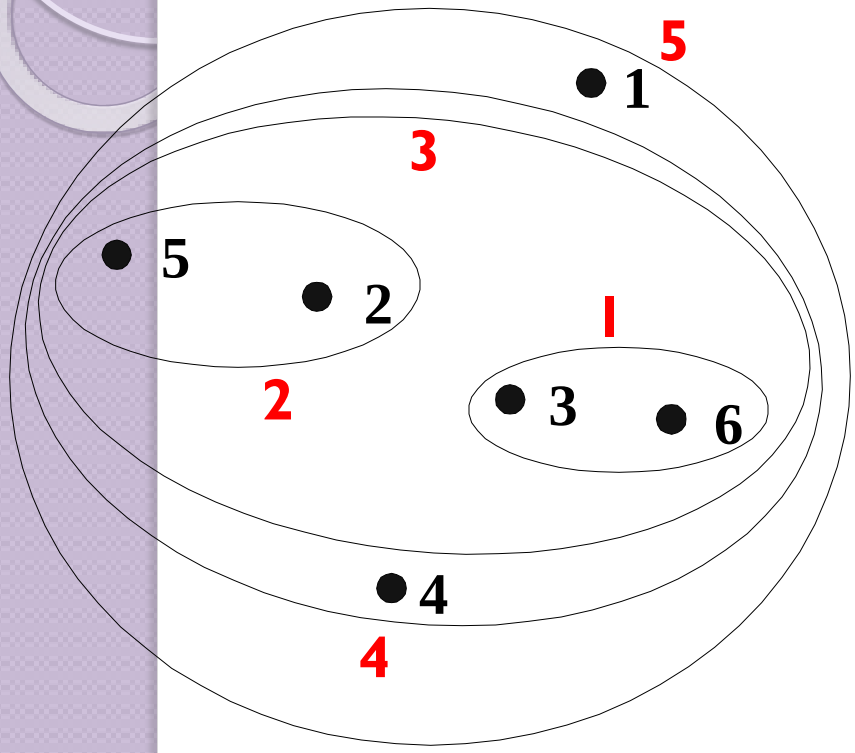
5 (0.08, 0.41)

6 (0.45, 0.30)

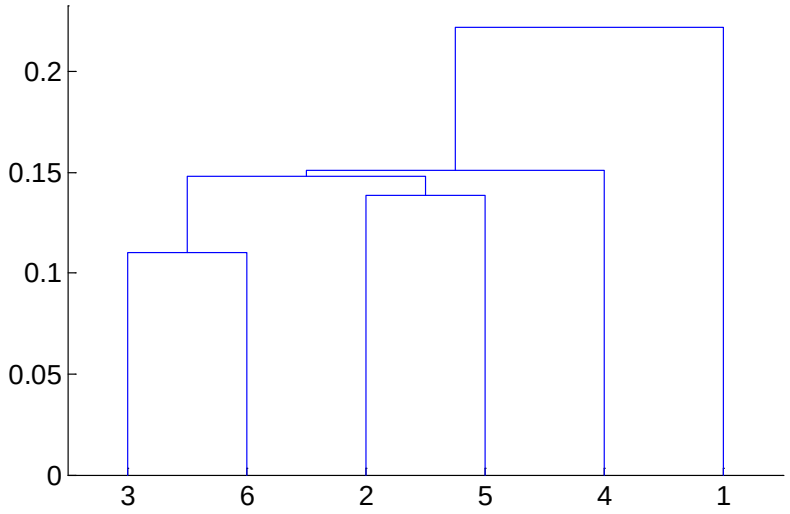
Αρκεί να «δω» μια ακμή

	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**

HAC: Ορισμός απόστασης μεταξύ συστάδων: MIN



Φωλιασμένες Συστάδες

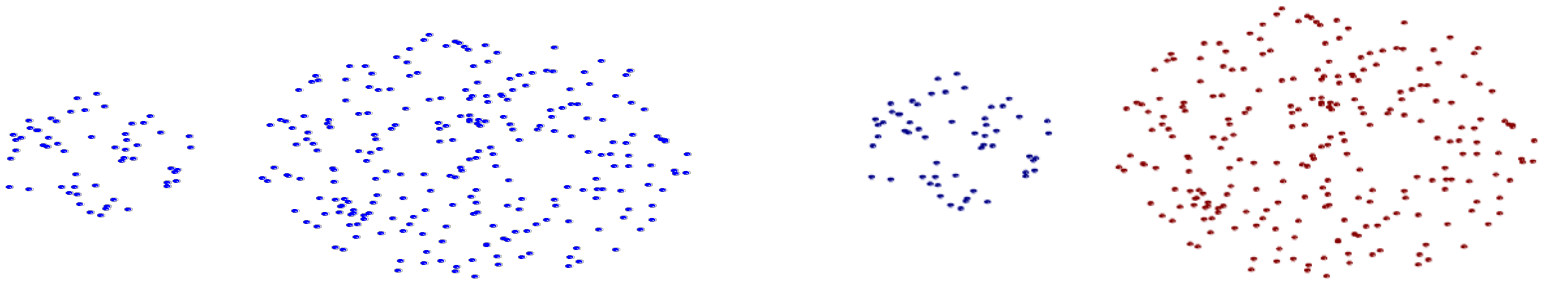


Δεντρόγραμμα

Το δεντρόγραμμα (y-άξονας) δίνει και τις αποστάσεις

HAC: Ορισμός απόστασης μεταξύ συστάδων: MIN

Προτερήματα



Αρχικά σημεία

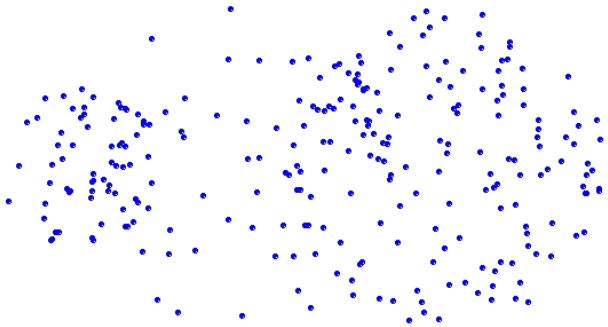
Δύο συστάδες

Contiguity-based (συνεχόμενες συστάδες)

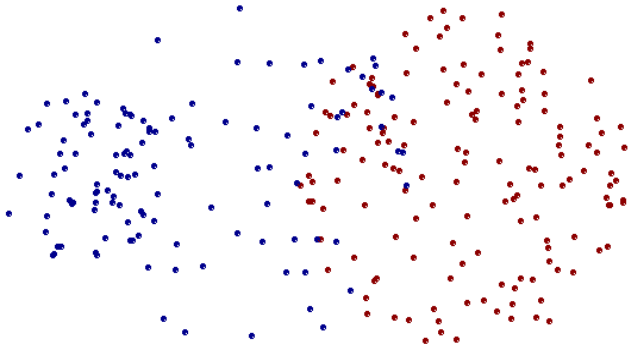
Μπορεί να χειριστεί μη ελλειπτικά (non-elliptical) σχήματα

HAC: Ορισμός απόστασης μεταξύ συστάδων: MIN

Μειονεκτήματα



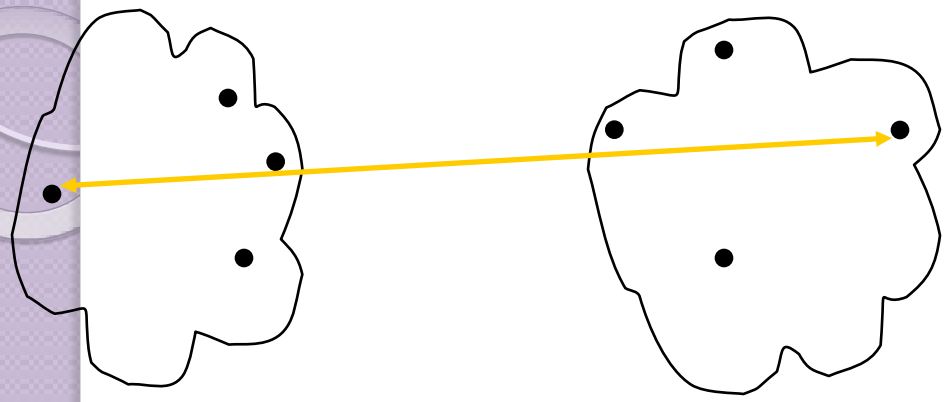
Αρχικά σημεία



Δύο συστάδες

- **Ευαίσθητο σε θόρυβο και outliers**

HAC: Ορισμός απόστασης μεταξύ συστάδων



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

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

· **Πίνακας Γειτνίασης**

·

HAC: Ορισμός απόστασης μεταξύ συστάδων: MAX

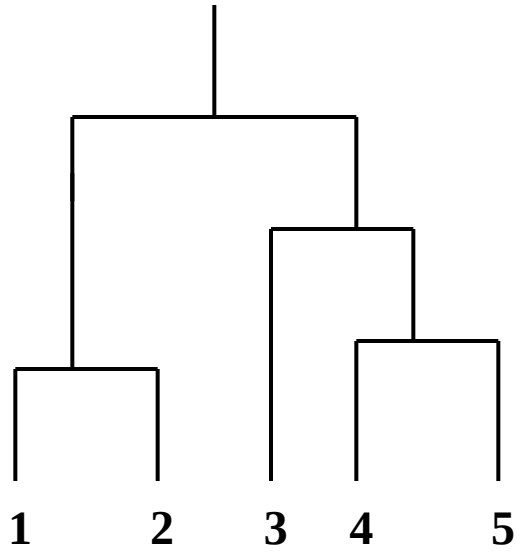
MAX ή πλήρους συνδεσιμότητας (complete linkage)
- Αναζητά κλίκες

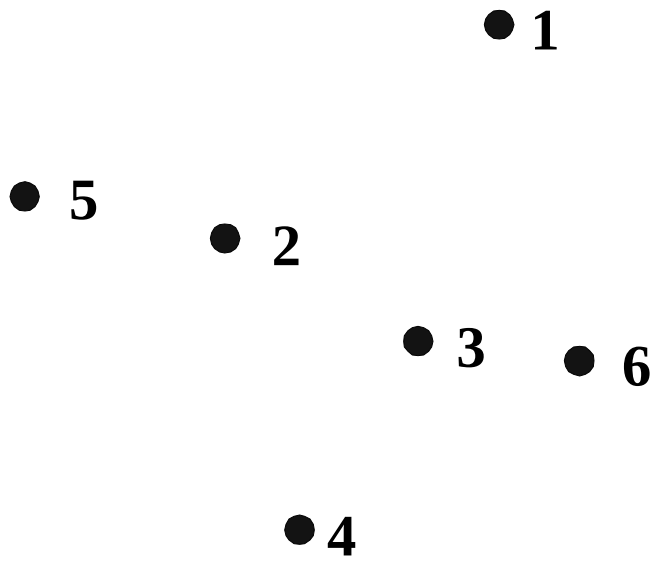
Η ομοιότητα μεταξύ δυο συστάδων βασίζεται στα δυο λιγότερο όμοια (πιο μακρινά) σημεία στις διαφορετικές συστάδες (longest edge)

Καθορίζεται από όλα τα ζεύγη τιμών στις δύο συστάδες.

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

ομοιότητα





1 (0.4, 0.53)

2 (0.22, 0.38)

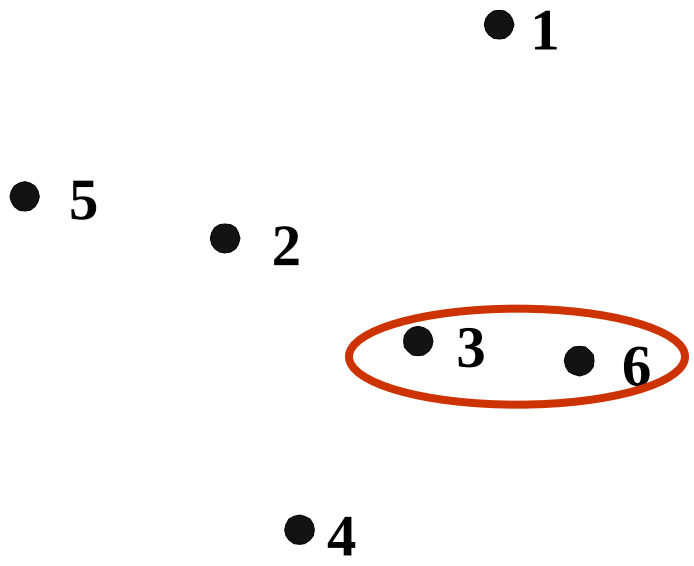
3 (0.35, 0.32)

4 (0.26, 0.19)

5 (0.08, 0.41)

6 (0.45, 0.30)

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



1 (0.4, 0.53)

2 (0.22, 0.38)

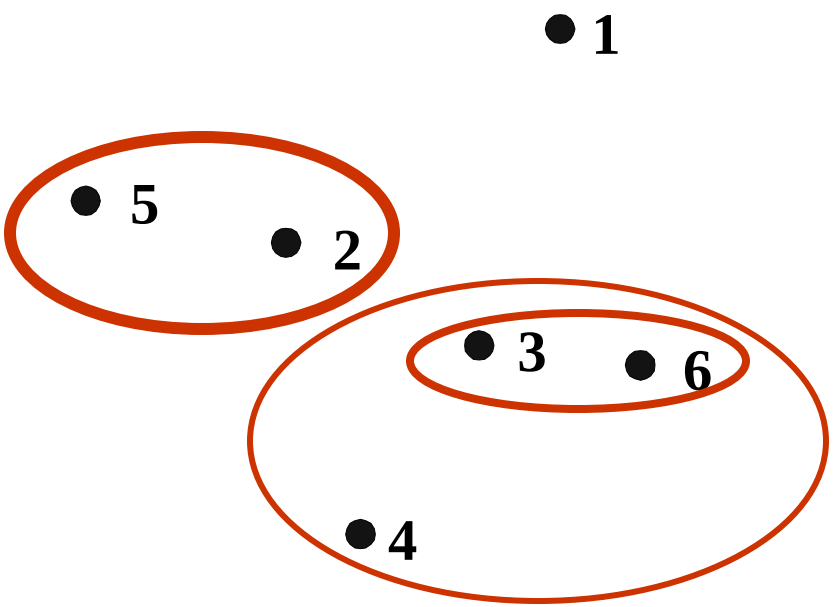
3 (0.35, 0.32)

4 (0.26, 0.19)

5 (0.08, 0.41)

6 (0.45, 0.30)

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



1 (0.4, 0.53)

2 (0.22, 0.38)

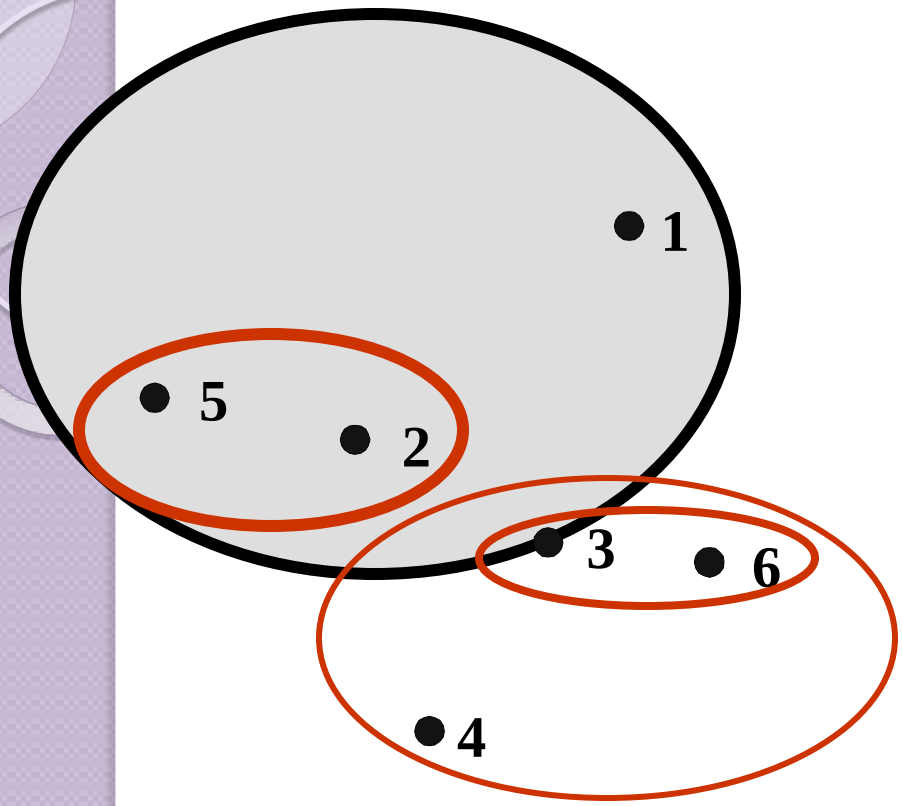
3 (0.35, 0.32)

4 (0.26, 0.19)

5 (0.08, 0.41)

6 (0.45, 0.30)

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



1 (0.4, 0.53)

2 (0.22, 0.38)

3 (0.35, 0.32)

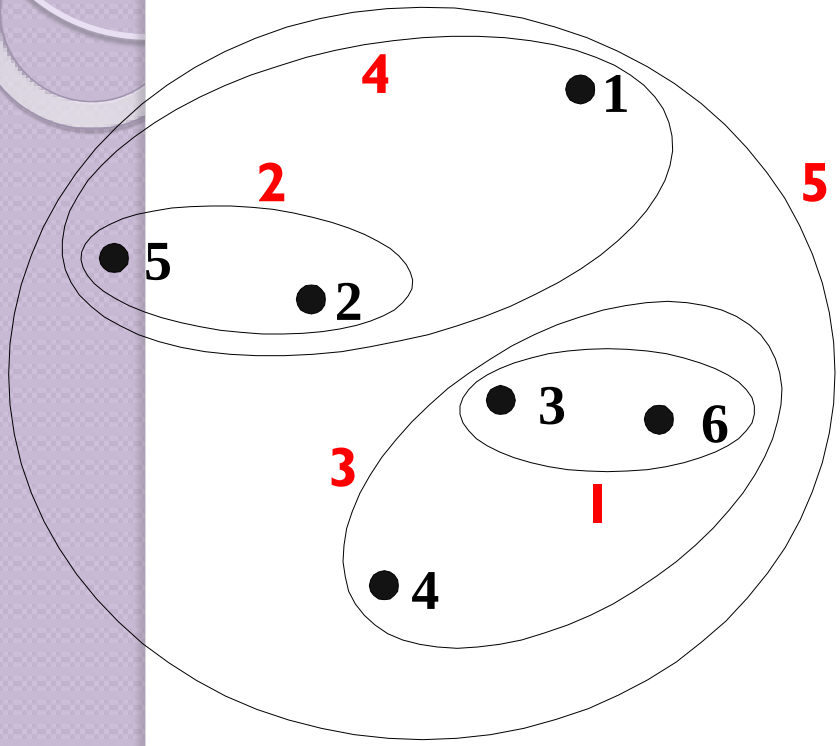
4 (0.26, 0.19)

5 (0.08, 0.41)

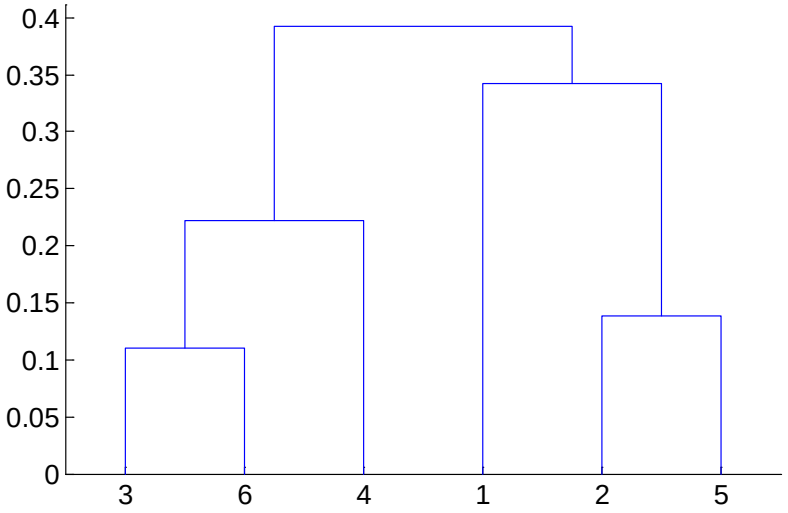
6 (0.45, 0.30)

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

HAC: Ορισμός απόστασης μεταξύ συστάδων: MAX



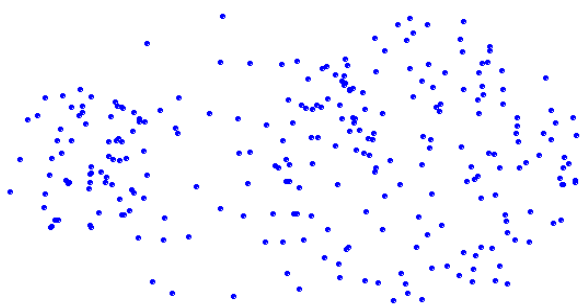
Φωλιασμένες Συστάδες



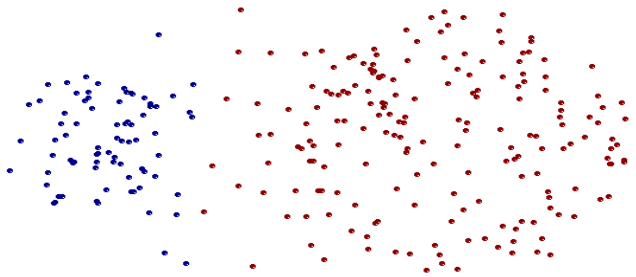
Δεντρόγραμμα

HAC: Ορισμός απόστασης μεταξύ συστάδων: MAX

Πλεονεκτήματα



Αρχικά Σημεία

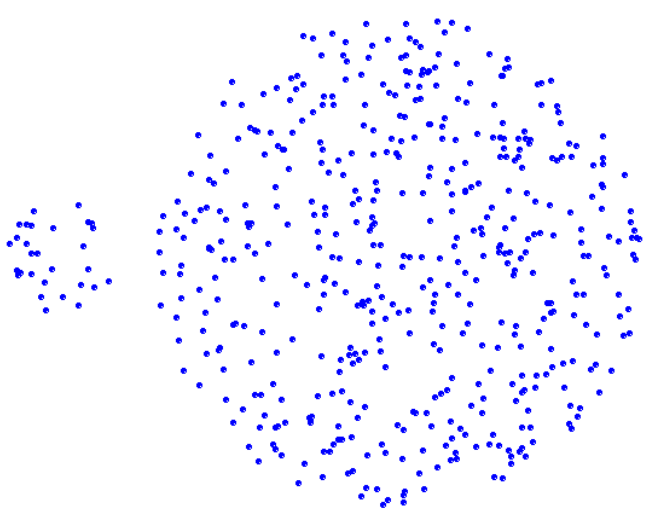


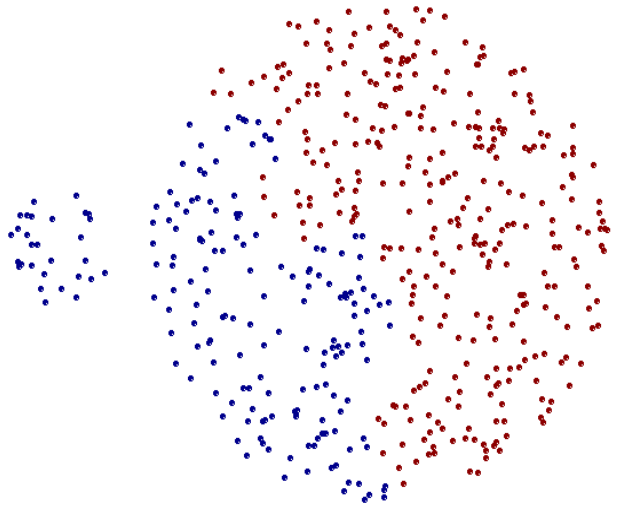
Δύο Συστάδες

- **λιγότερη εξάρτηση σε θόρυβο και outliers**

HAC: Ορισμός απόστασης μεταξύ συστάδων: MAX

Μειονεκτήματα



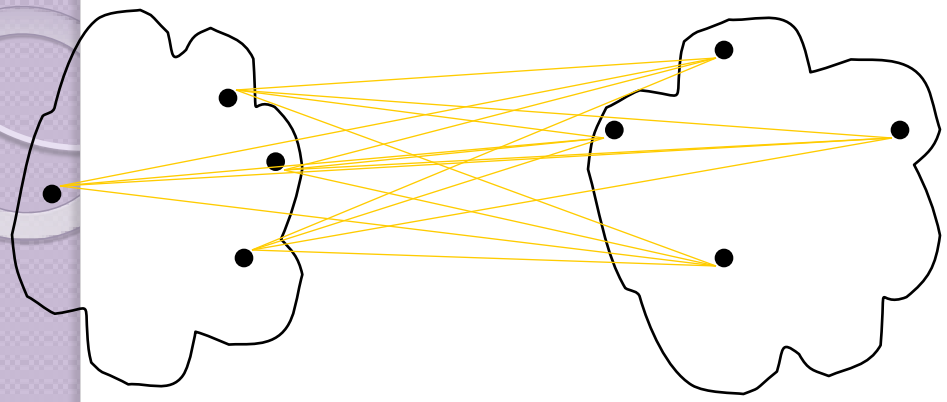


Δύο συστάδες

Αρχικά σημεία

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

HAC: Ορισμός απόστασης μεταξύ συστάδων



	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**

·

HAC: Ορισμός απόστασης μεταξύ συστάδων: Μέσο Ομάδας

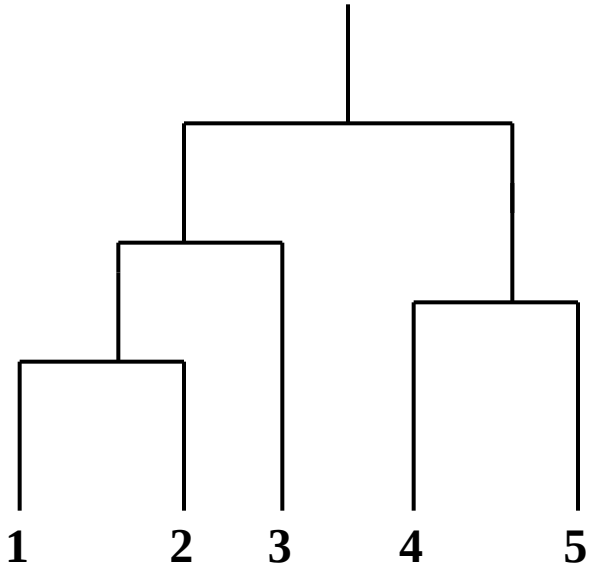
- Κοντινότητα δύο συστάδων είναι η μέση τιμή της ανα-δύο κοντινότητας (average of pairwise proximity) μεταξύ των σημείων των δύο συστάδων.

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

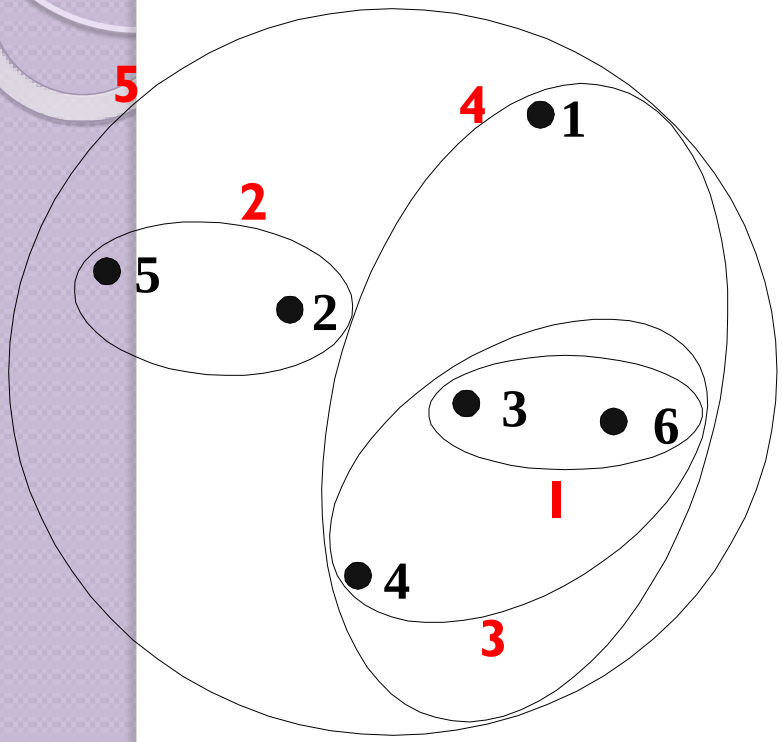
- Χρήση μέσης γιατί η ολική θα έδινε προτίμηση στις μεγάλες συστάδες

ομοιότητα

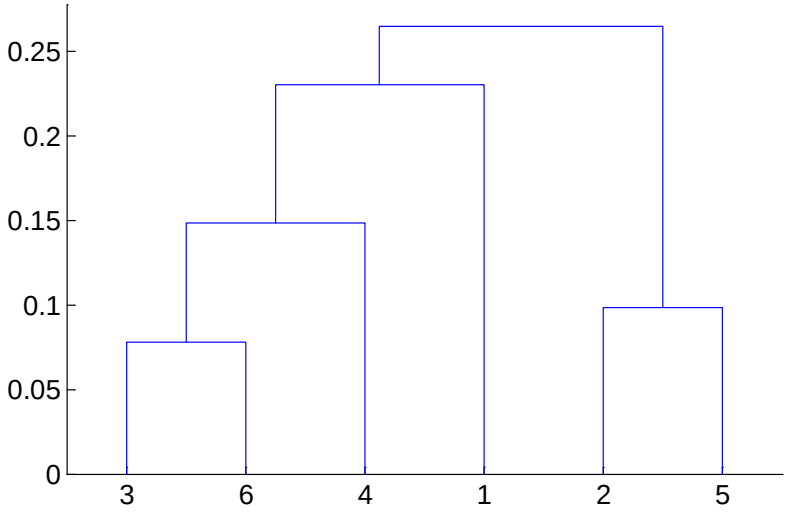
	I1	I2	I3	I4	I5
I1	1.00	0.90	0.10	0.65	0.20
I2	0.90	1.00	0.70	0.60	0.50
I3	0.10	0.70	1.00	0.40	0.30
I4	0.65	0.60	0.40	1.00	0.80
I5	0.20	0.50	0.30	0.80	1.00



HAC: Ορισμός απόστασης μεταξύ συστάδων: Μέσο Ομάδας



Φωλιασμένες Συστάδες

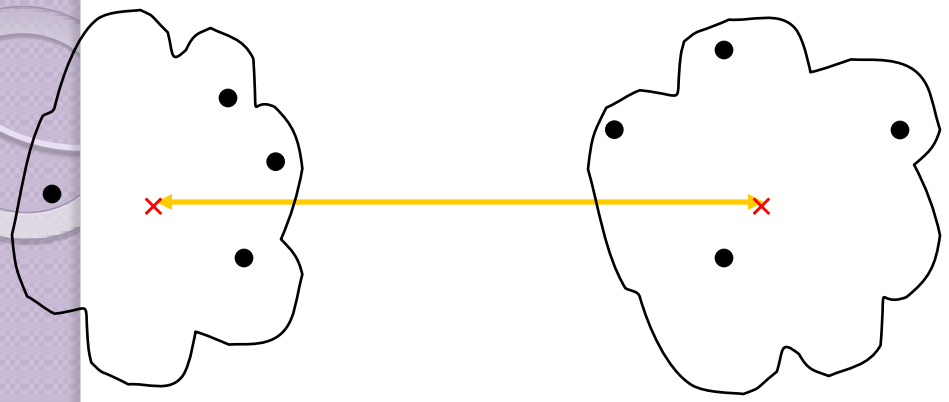


Dendrogram

HAC: Ορισμός απόστασης μεταξύ συστάδων: Μέσο Ομάδας

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

HAC: Ορισμός απόστασης μεταξύ συστάδων



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

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

- **Πίνακας Γειτνίασης**

Πρόβλημα: μη μονότονη αύξηση της απόστασης

Δηλαδή, δυο συστάδες που συγχωνεύονται μπορεί να έχουν μικρότερη απόσταση από συστάδες που έχουν συγχωνευτεί σε προηγούμενα βήματα

Εξόρυξη Δεδομένων

Συσταδοποίηση ΙΙ

Περιεχόμενα

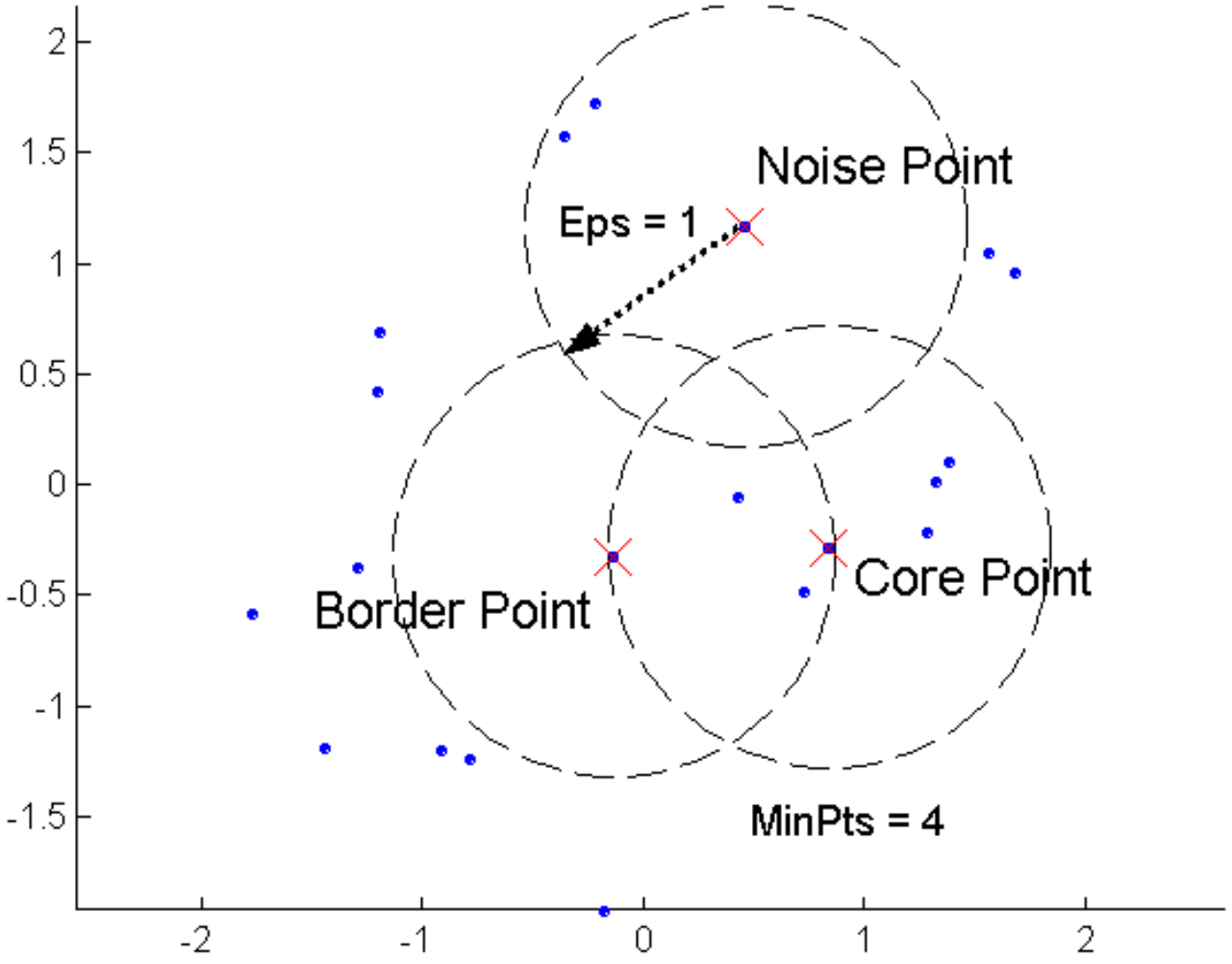
- Αλγόριθμοι βασισμένοι στην πυκνότητα
  - DB-SCAN
- Μέτρα εγκυρότητας συσταδοποίησης - Cluster Validity
  - Cohesion
  - Separation
  - Άλλοι αλγόριθμοι
  - BIRCH

DBSCAN

DBSCAN: Γενικά

- Ο DBSCAN είναι ένας αλγόριθμος βασισμένος στην πυκνότητα
- Πυκνότητα για ένα σημείο = αριθμός σημείων (MinPts) μέσα σε ποια προκαθορισμένη ακτίνα (Eps) από αυτό (συμπεριλαμβανομένου του σημείου)
- Τα σημεία διαχωρίζονται σε:
  - Βασικά (core): ένα σημείο για το οποίο υπάρχουν περισσότερα από ένα προκαθορισμένο αριθμό (***MinPts***) σημεία σε ακτίνα ***Eps***
    - Αυτά είναι τα σημεία που είναι στο εσωτερικό μιας συστάδας (ομάδας πυκνών σημείων)
  - Οριακά (border): 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 βασικό ούτε οριακό

DBSCAN: Γενικά



Αλγόριθμος

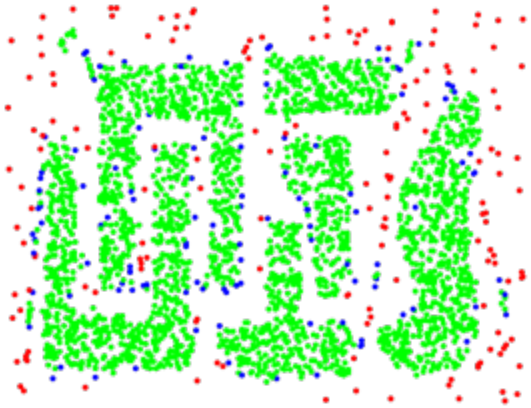
1: Χαρακτήρισε κάθε σημείο ως βασικό, οριακό ή θόρυβο
2: Διέγραψε τα σημεία θορύβου

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

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

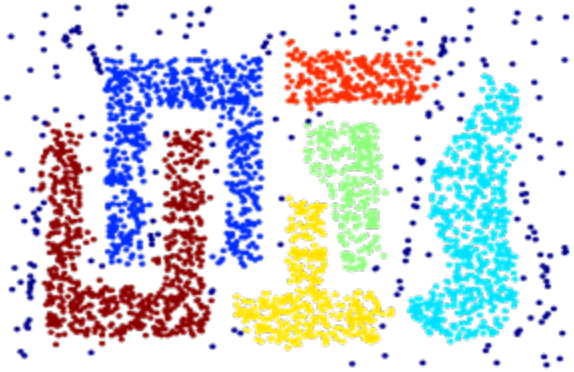
Αρχικά σημεία
Eps = 10, MinPts = 4

Βήμα 1&2



Τύποι σημείων:
core, **border** και
noise

Βήμα 3&4



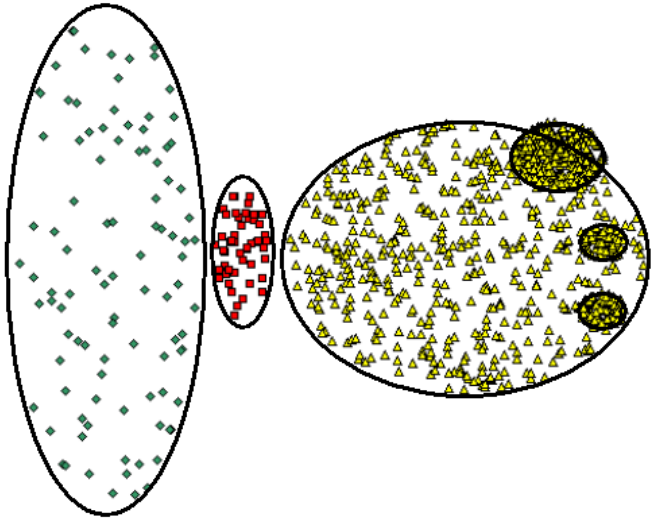
Συστάδες

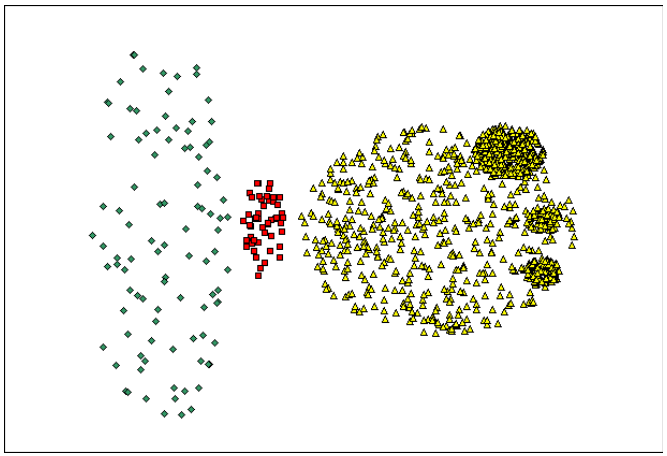
- Δεν επηρεάζεται από το θόρυβο
- Μπορεί να χειριστεί συστάδες με διαφορετικά σχήματα και μεγέθη

Πολυπλοκότητα

- Για m σημεία εισόδου:
- Χρόνου
  - O(m x χρόνος εντοπισμού σημείων σε eps-γειτονιά) = $O(m^2)$
  - Για μικρό αριθμό διαστάσεων, υπάρχουν δομές που υποστηρίζουν την πράξη σε O(m logm)
- Χώρου
  - O(m) χώρος (για κάθε σημείο κρατάμε μόνο ένα label σε μια συστάδα ανήκει και το είδος του (βασικό, οριακό, θόρυβος)

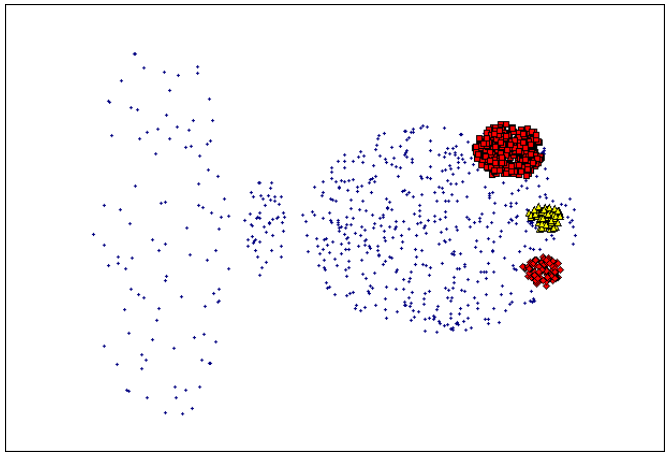
Περιορισμοί





(MinPts=4, Eps=9.75).

- **Η επιλογή των αρχικών σημείων επηρεάζει τις τελικές συστάδες**
- **Μεγάλη ευαισθησία σε σημεία με διαφορετικές πυκνότητες**
- **Σε πολυ-διάστατα δεδομένα είναι δύσκολος ο ορισμός πυκνότητας και δαπανηρός ο υπολογισμός γειτόνων**



(MinPts=4, Eps=9.92)

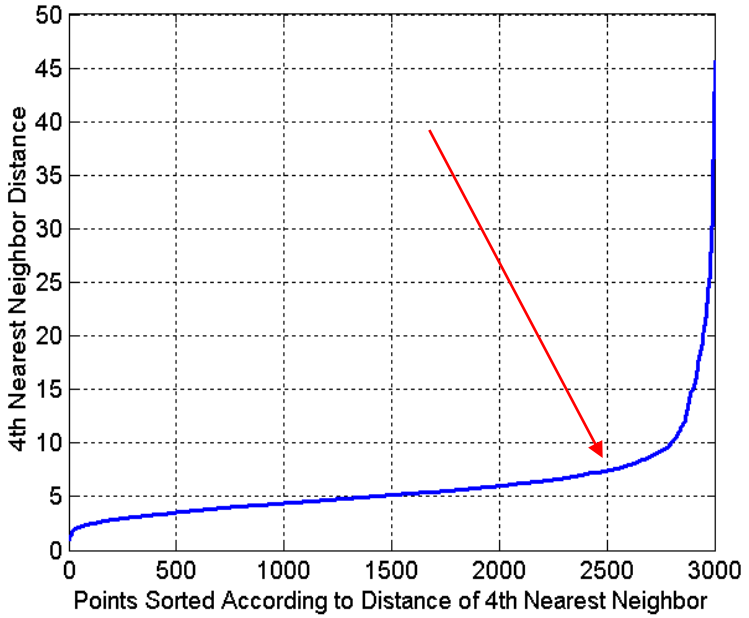
Καθορισμός των MinPts και Eps

- Κοιτάζουμε την απόσταση ***k-dist*** ενός σημείου από τον k-οστό κοντινότερο γείτονα του
- Κατά μέσο όρο, 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



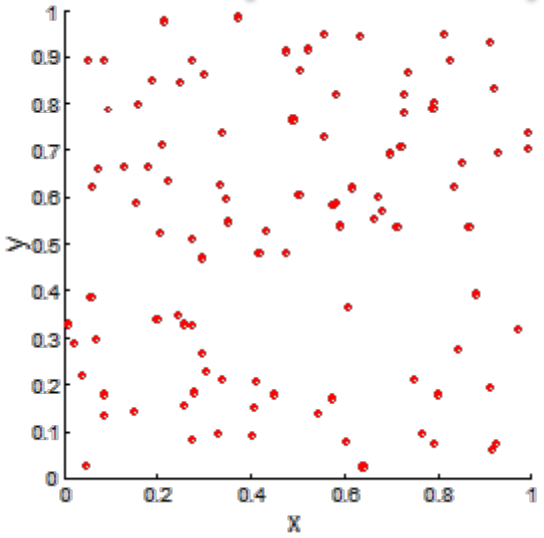
Ποιότητα της συσταδοποίησης Cluster validity

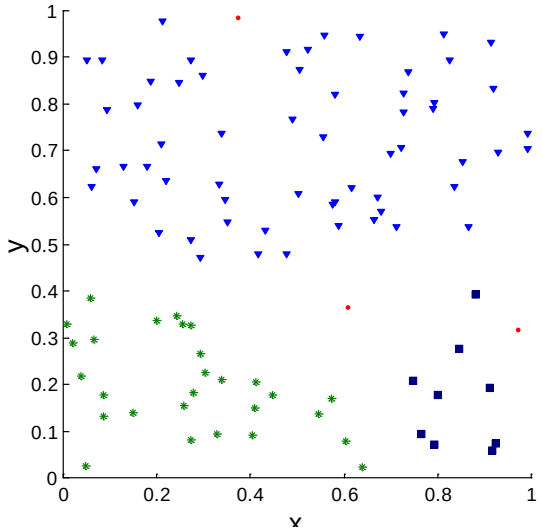
Θέματα ποιότητας

- 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

Παράδειγμα

Τυχαία Σημεία

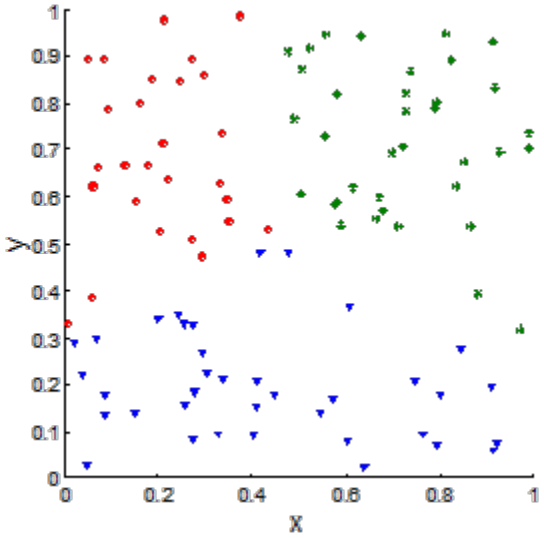


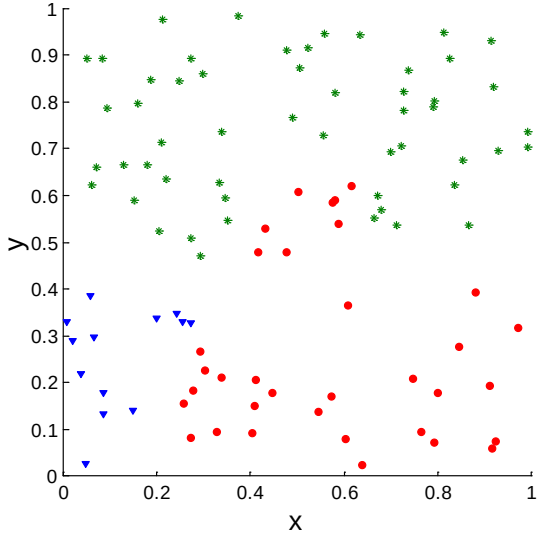


DBSCAN

3 ομάδες κοιτώντας την απόσταση του 4ου γείτονα

K-means





HAC 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, πχ 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 τη κάθε συστάδα χωριστά

Χρήση κριτηρίων ορθότητας

- Καθορίζουν
  - Το πόσο καλή είναι μια συσταδοποίηση
  - Το ποσό καλή είναι μια συστάδα
  - Το ποσό καλό είναι ένα σημείο σε μια συστάδα
- Μπορούν να χρησιμοποιηθούν για τη βελτίωση της συσταδοποίησης
  - Πχ μια συστάδα με κακή συνεκτικότητα μπορεί να χρειαστεί να διασπαστεί
  - Δυο συστάδες όχι καλά διαχωρισμένες μπορεί να συγχωνευτούν

Μετρήσεις Ποιότητας Συσταδοποίησης

Πόσο καλή είναι μια συσταδοποίηση

- Με επίβλεψη (supervised - External Index):
  - Υπάρχει εξωτερική πληροφορία (πληροφορία εκτός των δεδομένων), πχ ετικέτες για τις συστάδες
  - Πόσο οι περιγραφές των συστάδων ταιριάζουν με τις ετικέτες των κλάσεων. – πχ Εντροπία
- Χωρίς επίβλεψη (unsupervised - Internal Index):
  - Συνεκτικότητα (cohesion)
  - Διακριτότητα ή διαχωρισμός (separation)
  - Χρήση Πίνακα Γειτνίασης
- Συγκριτικοί -Σχετικό Ευρετήριο (Relative Index):
  - Χρησιμοποιείται για τη σύγκριση δυο διαφορετικών συσταδοποιήσεων ή συστάδων, πχ δυο k-means συσταδοποιήσεις με διαφορετικό k

Χωρίς επίβλεψη

Συνεκτικότητα και Διαχωρισμός

- Μέτρα χωρίς επίβλεψη
  - Ένα για να χαρακτηρίσουμε κάθε συστάδα ξεχωριστά (cohesion – συνεκτικότητα: πόσο κοντά (όμοια) είναι τα σημεία κάθε συστάδας)
  - Ένα για τις συστάδες μεταξύ τους (separation – διαχωρισμός: πόσο μακριά (ανόμοιες) είναι δύο συστάδες

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

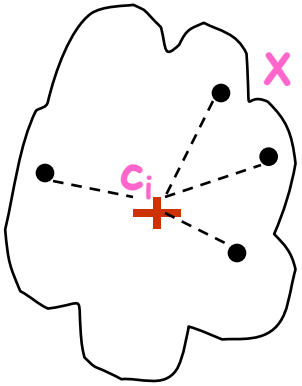
$$BSS = \sum |C_i| (m - m_i)^2$$

- Ορίζονται είτε
  - Prototype-based (centroid based): με βάση το «κεντρικό σημείο» κάθε συστάδας είτε
  - Graph-based: με βάση τις ανά-δύο αποστάσεις των σημείων

Συνεκτικότητα και Διαχωρισμός

Centroid-based clustering (πχ k-means)

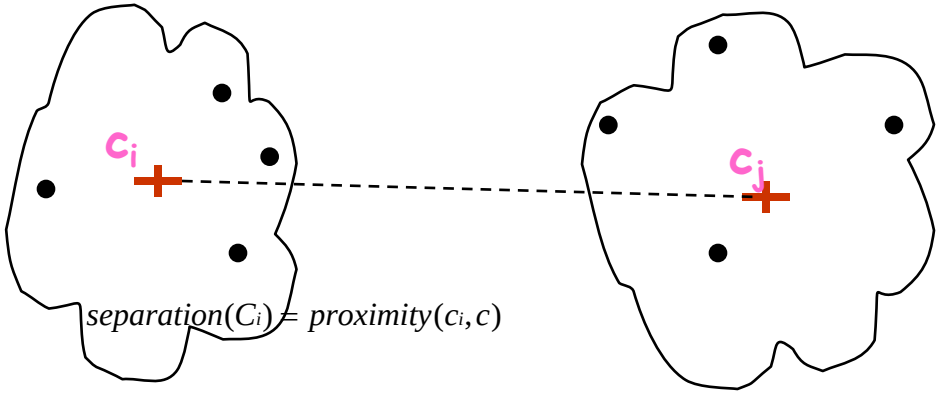
Συνεκτικότητα (cohesion)



$$cohesion(C_i) = \sum_{x \in C_i} proximity(x, c_i)$$

Αν proximity = τετράγωνο της Ευκλείδειας, τότε ESS

Διαχωρισμός (separation)



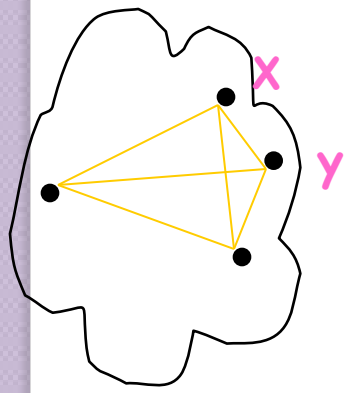
$$separation(C_i, C_j) = proximity(c_i, c_j)$$

Συνεκτικότητα και Διαχωρισμός

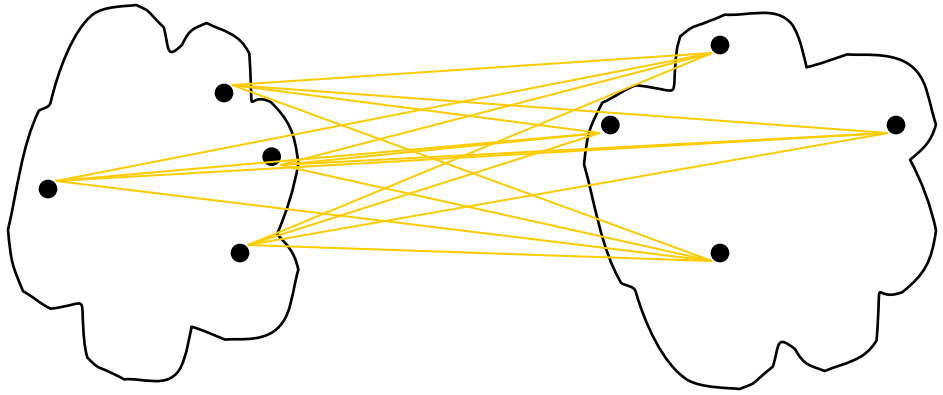
Graph-based συσταδοποίηση

- Η συνεκτικότητα μιας συστάδας (cluster cohesion) είναι το άθροισμα των βαρών (συνήθως απόσταση) μεταξύ όλων των συνδέσεων σε μια συστάδα.
- Ο διαχωρισμός (cluster separation) είναι το άθροισμα των βαρών (συνήθως απόσταση) μεταξύ κόμβων της συστάδας και των κόμβων εκτός συστάδας

Συνεκτικότητα (cohesion)



Διαχωρισμός (separation)



$$separation(C_i, C_j) = \sum_{\substack{x \in C_i \\ y \in C_j}}^{n} proximity(x, y)$$

$$cohesion(C_i) = \sum_{\substack{x \in C_i \\ y \in C_i}}^{n} proximity(x, y)$$

Συνεκτικότητα και Διαχωρισμός

Συνολική Συνεκτικότητα

$$overall - cohesion = \sum_{i=1}^{k} w_i \, \text{cohesion}(C_i)$$

Άθροισμα συνεκτικότητας κάθε συστάδας

Συνολικός Διαχωρισμός

$$overall - separation = \sum_{i=1}^{k} w_i \, \text{separation}(C_i)$$

Άθροισμα διαχωρισμού των συστάδων

Συνολικός Χαρακτηρισμός Ποιότητας για τη συσταδοποίηση

$$overall - validity = \sum_{i=1}^{k} \frac{\text{seperation}(C_i)}{\text{cohesion}(C_i)}$$

$$overall - validity = \sum_{i=1}^{k} w_i \, validity(C_i)$$

Όπου το βάρος (w_i) μπορεί να είναι πχ ανάλογο του μεγέθους της συστάδας ή η τετραγωνική ρίζα της συνεκτικότητας ή 1

Σχέση prototype και graph-based συνεκτικότητας (για Ευκλείδειες αποστάσεις)

Έστω Ευκλείδεια απόσταση, **σχέση SSE με συνεκτικότητα** (πόσο στενά σχετιζόμενα είναι τα αντικείμενα μιας συστάδας);

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

$$cluster - SSE = \sum_{x \in C_i} dist^2(c_i, x)$$

Αποδεικνύεται ότι

$$cluster - SSE = \sum_{x \in C_i} dist^2(x, c_i) = \frac{1}{2m_i} \sum_{x \in C_i} \sum_{y \in C_i} dist(x, y)^2$$

Δηλαδή, είτε πάρουμε την απόσταση από το κέντρο είτε το μέσο όρο των ανά δύο αποστάσεων των σημείων είναι το ίδιο

Σχέση prototype και graph-based διαχωρισμού (για Ευκλείδειες αποστάσεις)

Έστω Ευκλείδεια απόσταση, σχέση SSB (group sum of squares) με διαχωρισμό (πόσο μακριά είναι οι συστάδες);

$$cluster - SSB = dist(c_i, c)^2$$

$$(ολικό-)SSB = \sum_{i=1}^{K} m_i \, dist(c_i, c)^2$$

Το ολικό κέντρο (σημείο c στους τύπους) είναι το σημείο που προκύπτει αν πάρουμε το μέσο (mean) των κέντρων όλων των συστάδων

Για ισομεγέθεις συστάδες: $m_i = m / K$

Αποδεικνύεται ότι

$$ολικό - SSB = \sum_{x \in C_i} m_i \, dist^2(c_i, c) = \frac{1}{2K} \sum_{i=1}^{K} \sum_{j=1}^{K} \frac{m}{K} dist(c_i, c_j)^2$$

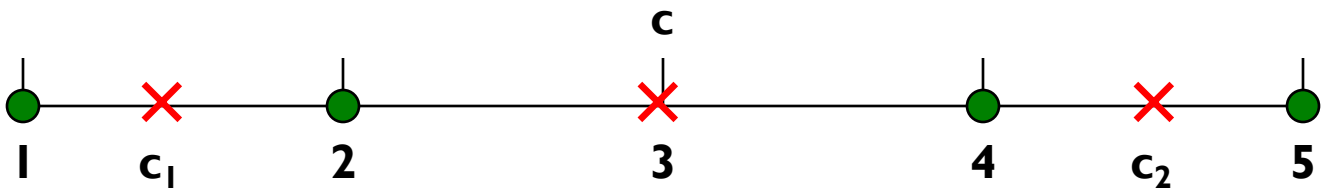
Δηλαδή, είτε πάρουμε την απόσταση των κέντρων κάθε συστάδας από το ολικό κέντρο είτε το μέσο όρο των ανά δύο αποστάσεων των κέντρων κάθε συστάδας είναι το ίδιο

Παράδειγμα

Total-SSE + Total-SSB = σταθερά

Ελαχιστοποίηση της SSE (συνεκτικότητας)

=> Μεγιστοποίηση του SSB (διαχωρισμού)



K = 1 cluster:

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

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

$$Total = 10 + 0 = 10$$

K = 2 clusters:

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

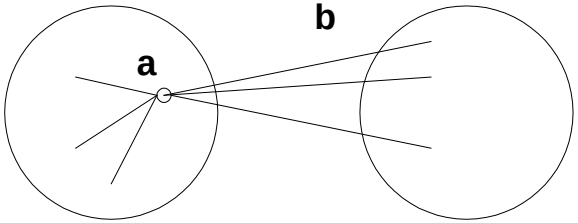
$$total - SSB = 2 \times (3-1.5)^2 + 2 \times (4.5-3)^2 = 9$$

$$Total = 1 + 9 = 10$$

Άλλες τεχνικές

- Silhouette Coefficient (συντελεστής σκιαγράφησης)
  - 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 if a < b, (or s = b/a - 1 if a >= b, not the usual case)
  - Συνήθως μεταξύ του 0 και του 1
  - Όσο πιο κοντά στο 1, τόσο το καλύτερο

- Δείχνει πόσο «κεντρικό» είναι ένα σημείο για μία συστάδα



Άλλες τεχνικές

- Δύο Πίνακες
  - Πίνακας Γειτνίασης (proximity matrix): ο πίνακας με την ομοιότητα των σημείων
  - Πίνακας Εμφάνισης ("incidence" matrix)

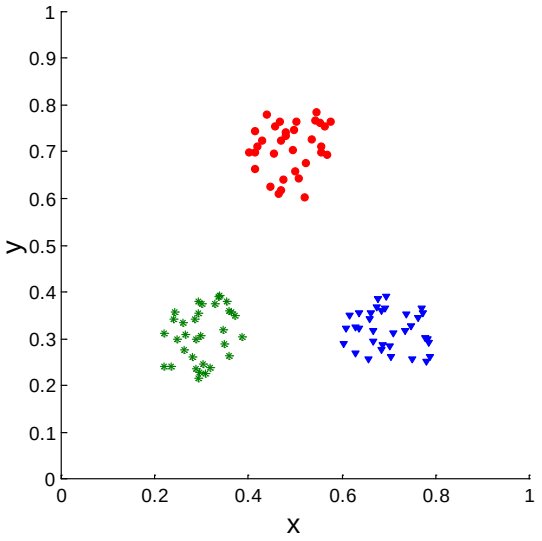
    Μια γραμμή και μια στήλη για κάθε σημείο

    Μια εγγραφή είναι 1 αν το ζευγάρι σημείων ανήκει στην ίδια συστάδα

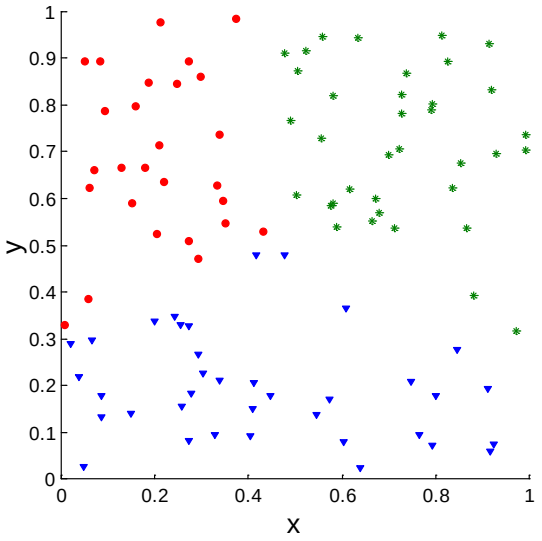
    Μια εγγραφή είναι 0 αν το ζευγάρι σημείων ανήκει σε διαφορετική συστάδα
- Υπολογισμός της συσχέτισης (correlation) των δύο πινάκων
  - Υψηλή συσχέτιση σημαίνει ότι τα σημεία που ανήκουν στην ίδια συστάδα είναι κοντινά μεταξύ τους

Παράδειγμα – K-Means

- 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



Corr = -0.9235



Corr = -0.5810

Με επίβλεψη

Τεχνικές

- Μας δίνονται κάποιες ετικέτες κλάσεων και θέλουμε να δούμε πόσο καλά ταιριάζουν με τα δεδομένα
- Classification-oriented (μετρήσεις για ταξινόμηση): κατά πόσο μια συστάδα περιέχει αντικείμενα μίας μόνο κλάσης
- Similarity-oriented: κατά πόσο δύο αντικείμενα που ανήκουν στην ίδια κλάση, ανήκουν και στην ίδια συστάδα

EM (expectize maximization)

Expectation Maximization -EM

- Επιλέγει αρχικά k κέντρα συστάδων, με τυχαίο τρόπο
- Επαναληπτικά επαναορίζει τις συστάδες με βάση δύο βήματα:
  - Expectation step: ανέθεσε κάθε σημείο Xi στη συστάδα Ci με τη μεγαλύτερη πιθανότητα

$$P(X_i \in C_k) = p(C_k|X_i) = \frac{p(C_k)p(X_i|C_k)}{p(X_i)}$$

  - Maximization step: Εκτίμηση των παραμέτρων του μοντέλου

$$m_k = \frac{1}{N}\sum_{i=1}^{N}\frac{X_i P(X_i \in C_k)}{\sum_j P(X_i \in C_j)}$$

Για δύο συστάδες (A,B)

1. Initialization: $\mu_A^0, \sigma_A^0, , P_A^0, \mu_B^0, \sigma_B^0,$ and P_B^0, ϵ;
2. At iteration j: compute the probabilities

$$Pr[A|x] = \frac{Pr^j[x|A]P_A^j}{Pr^j[x]}, Pr^j[B|x] = \frac{Pr^j[x|B]P_B^j}{Pr^j[x]}$$

3. Update the new mixture parameters:

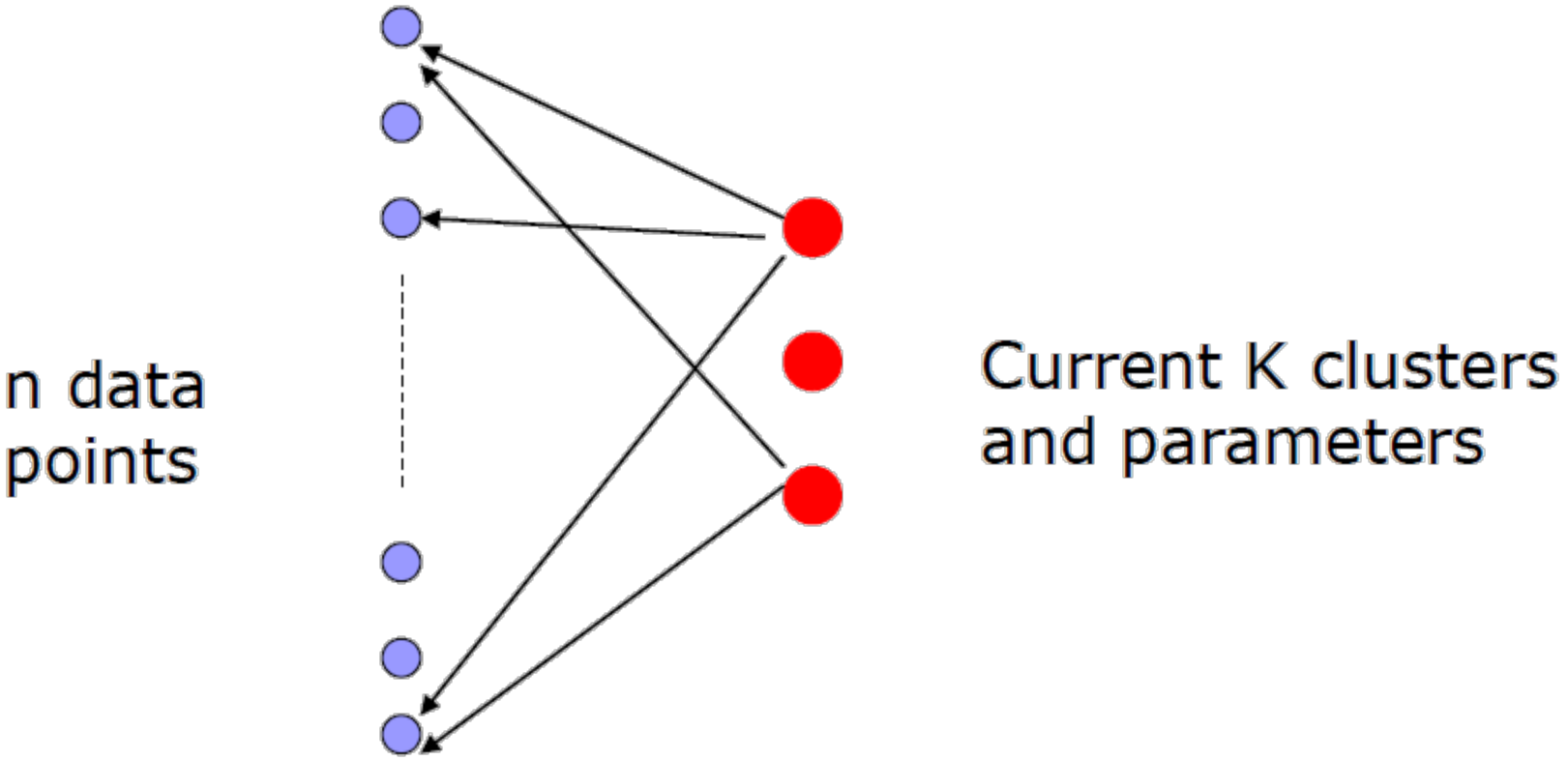
$$P_A^{j+1} = \frac{1}{n}\sum_x Pr[A|x], \qquad P_B^{j+1} = \frac{1}{n}\sum_x Pr[B|x];$$

$$\mu_A^{j+1} = \frac{\sum_x xPr[A|x]}{\sum_x Pr[A|x]}, \qquad \mu_B^{j+1} = \frac{\sum_x xPr[B|x]}{\sum_x Pr[B|x]};$$

$$\sigma_A^{j+1} = \frac{\sum_x P[A|x](x-\mu_A^{j+1})^2}{\sum_x P[A|x]}, \qquad \sigma_B^{j+1} = \frac{\sum_x P[B|x](x-\mu_B^{j+1})^2}{\sum_x P[B|x]};$$

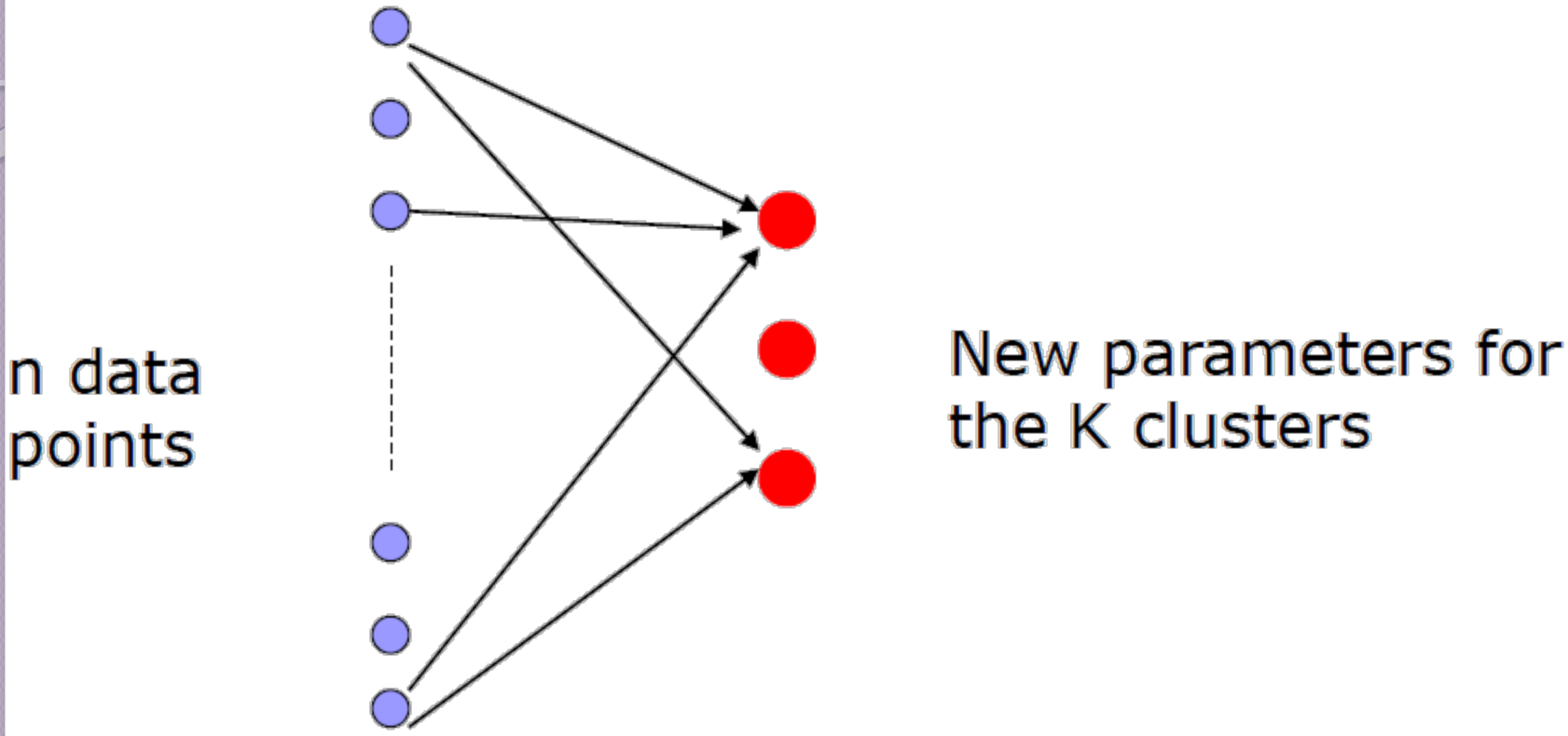
4. Compute the log estimate $E_j = \sum_x \log(Pr^j(x))$. Stop if $|E_j - E_{j+1}| \leq \epsilon$; Otherwise set $j = j+1$ and go to Step 2.

The E (Expectation) Step



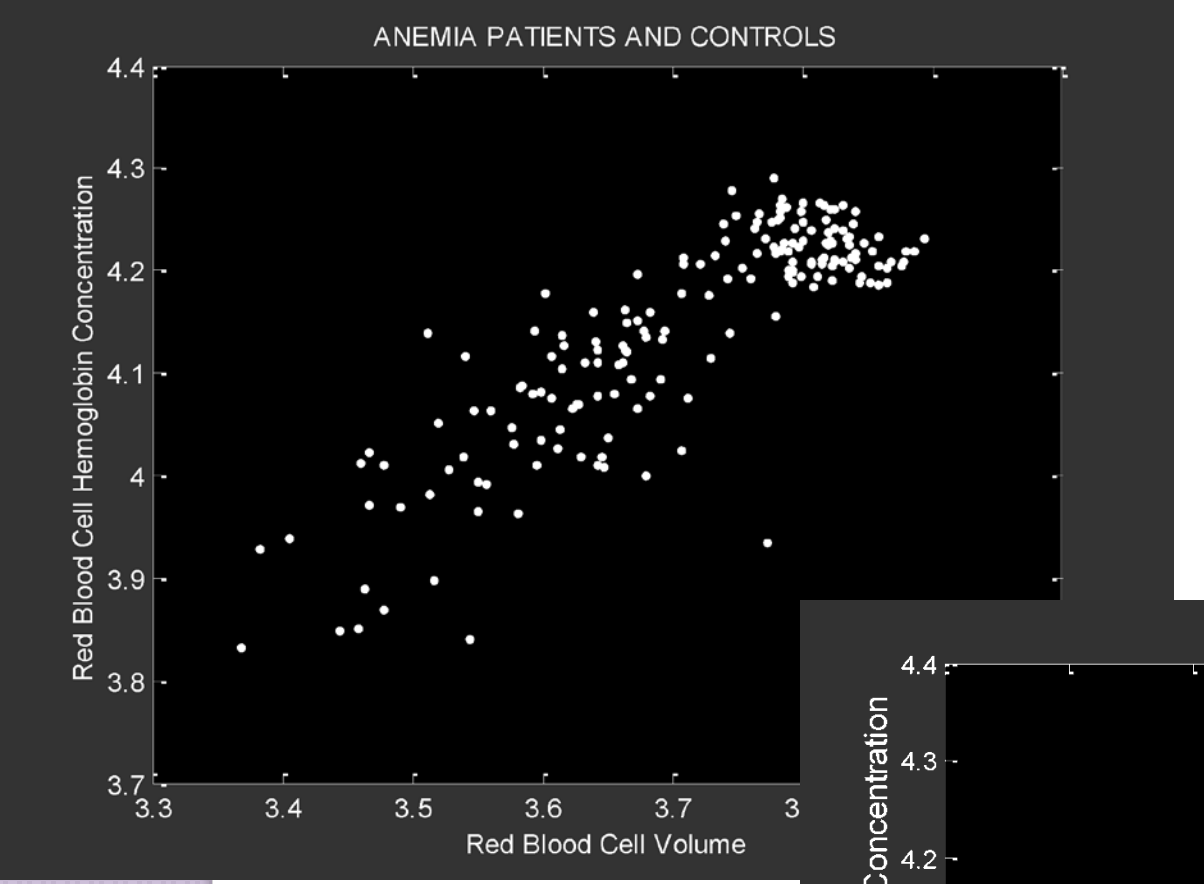
- E step: Υπολόγισε την πιθανότητα p(το σημείο i να ανήκει στη συστάδα k)

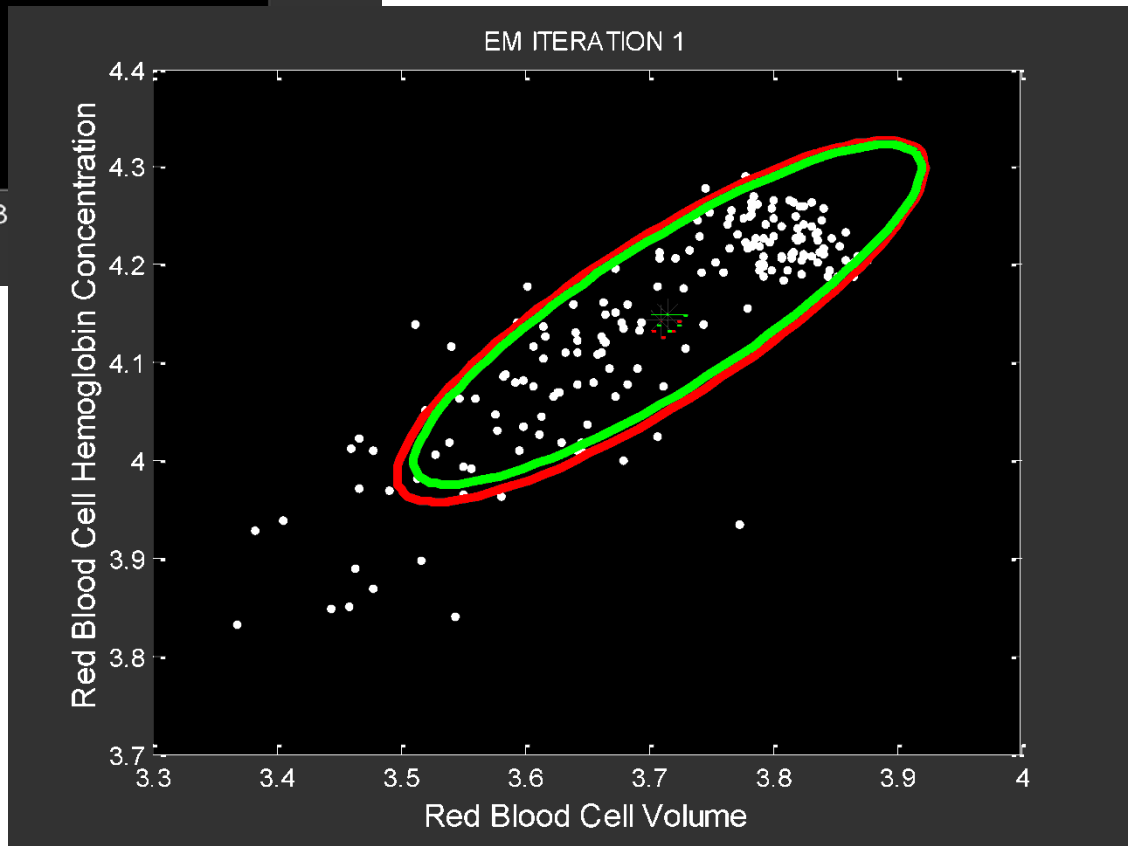
The M (Maximization) Step

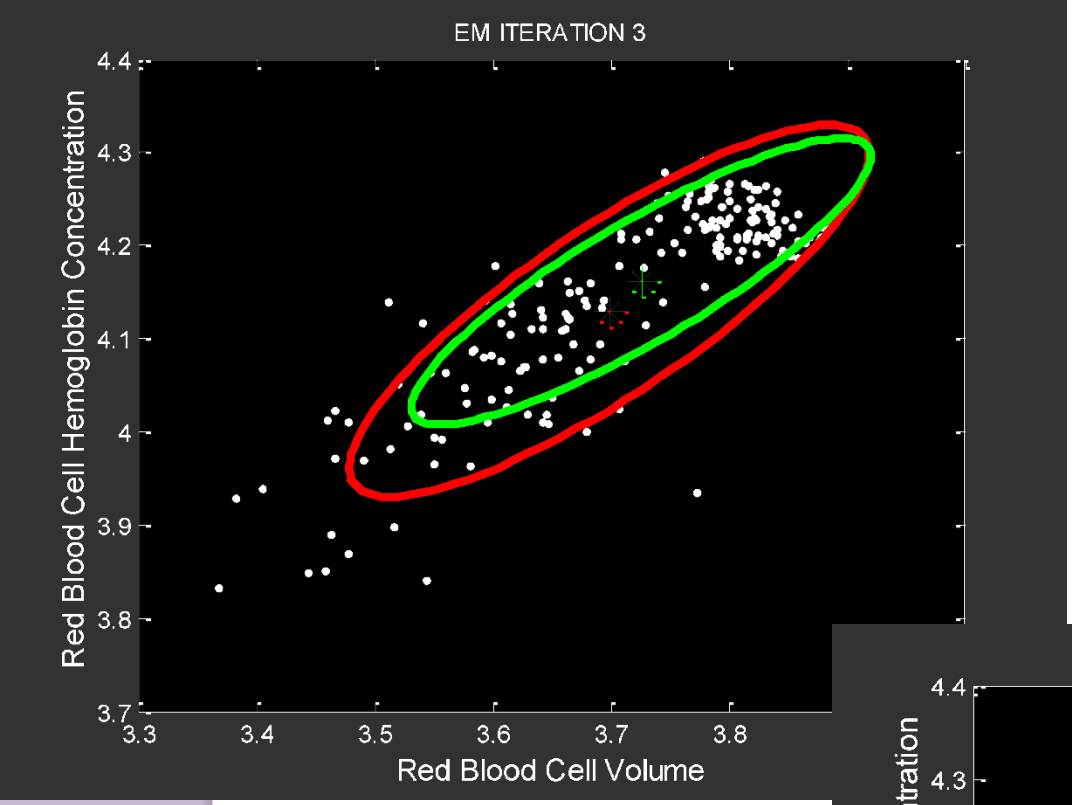


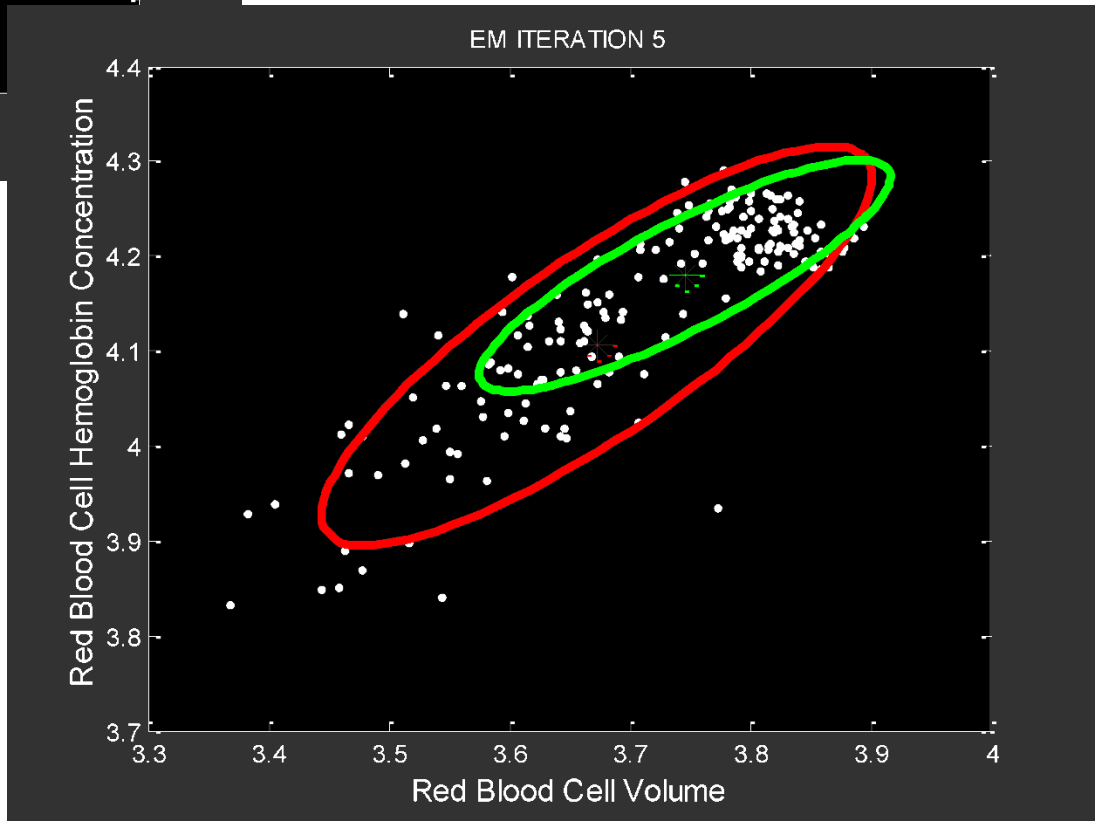
- M step: Υπολόγισε τις νέες τιμές για τις συστάδες με δεδομένο ότι έχεις **n** σημεία που μοιράζονται στις **k** συστάδες
- Δες αν θα επαναλάβεις

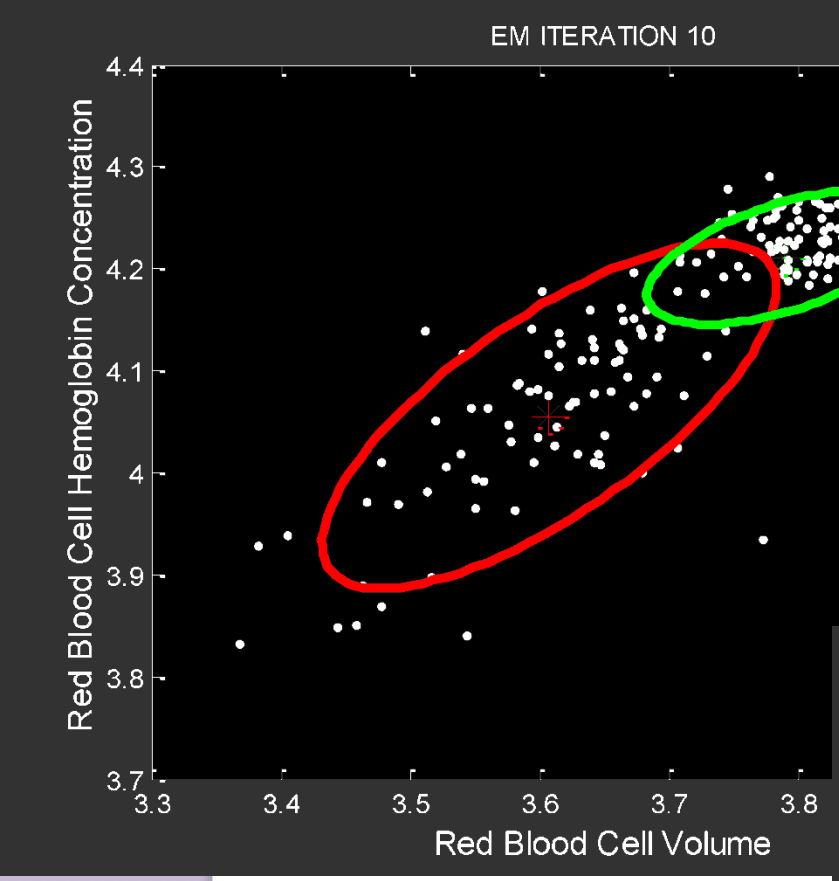
- Πολυπλοκότητα κάθε επανάληψης:
  - O(nK f(p))
- Εξαρτάται από το πιθανοτικό μοντέλο που χρησιμοποιείται
  - Για Gaussians: Estep → O(nK),
    Mstep →O(nKp2)
- Ο K-Means είναι ειδική περίπτωση του EM
  - Επέλεξε k τυχαία κέντρα
  - Βάλε τα σημεία σε k clusters
  - Υπολόγισε τα νέα κέντρα

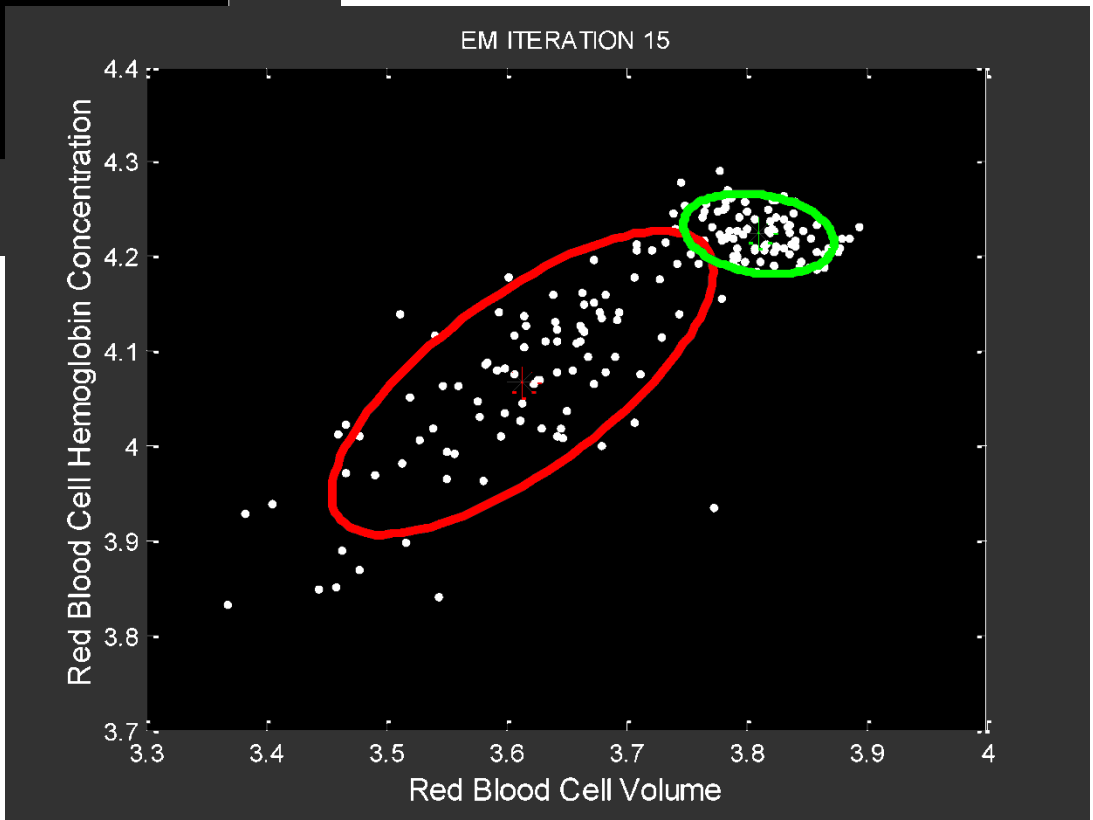

ANEMIA PATIENTS AND CONTROLS
Red Blood Cell Hemoglobin Concentration
Red Blood Cell Volume
4.4
4.3
4.2
4.1
4
3.9
3.8
3.7
3.3
3.4
3.5
3.6
3.7
3

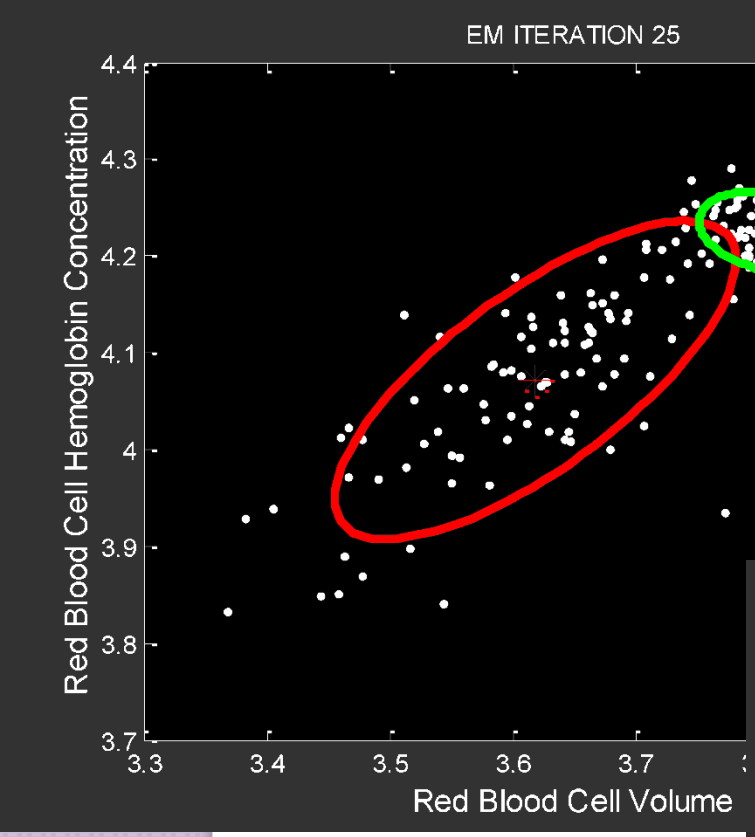

EM ITERATION 1
Red Blood Cell Hemoglobin Concentration
Red Blood Cell Volume
4.4
4.3
4.2
4.1
4
3.9
3.8
3.7
3.3
3.4
3.5
3.6
3.7
3.8
3.9
4

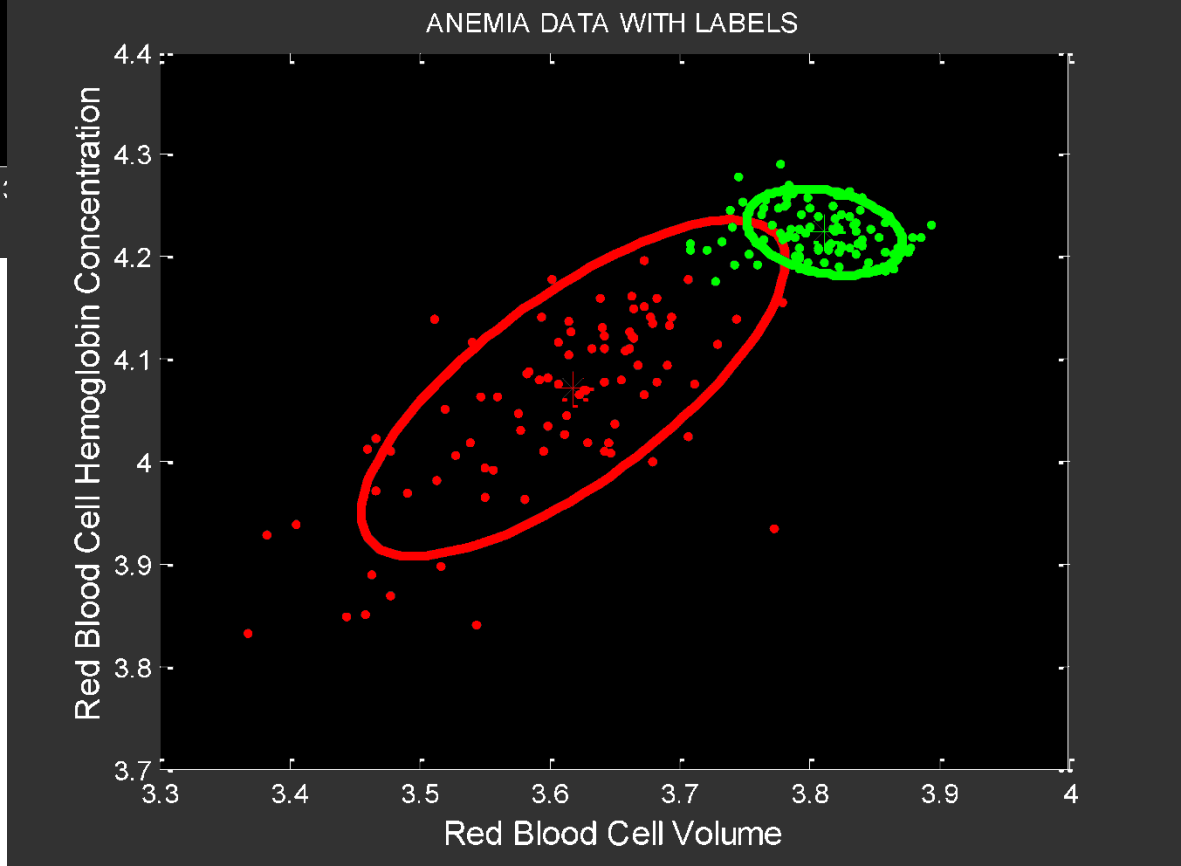

EM ITERATION 3
Red Blood Cell Hemoglobin Concentration
Red Blood Cell Volume
4.4
4.3
4.2
4.1
4
3.9
3.8
3.7
3.3
3.4
3.5
3.6
3.7
3.8

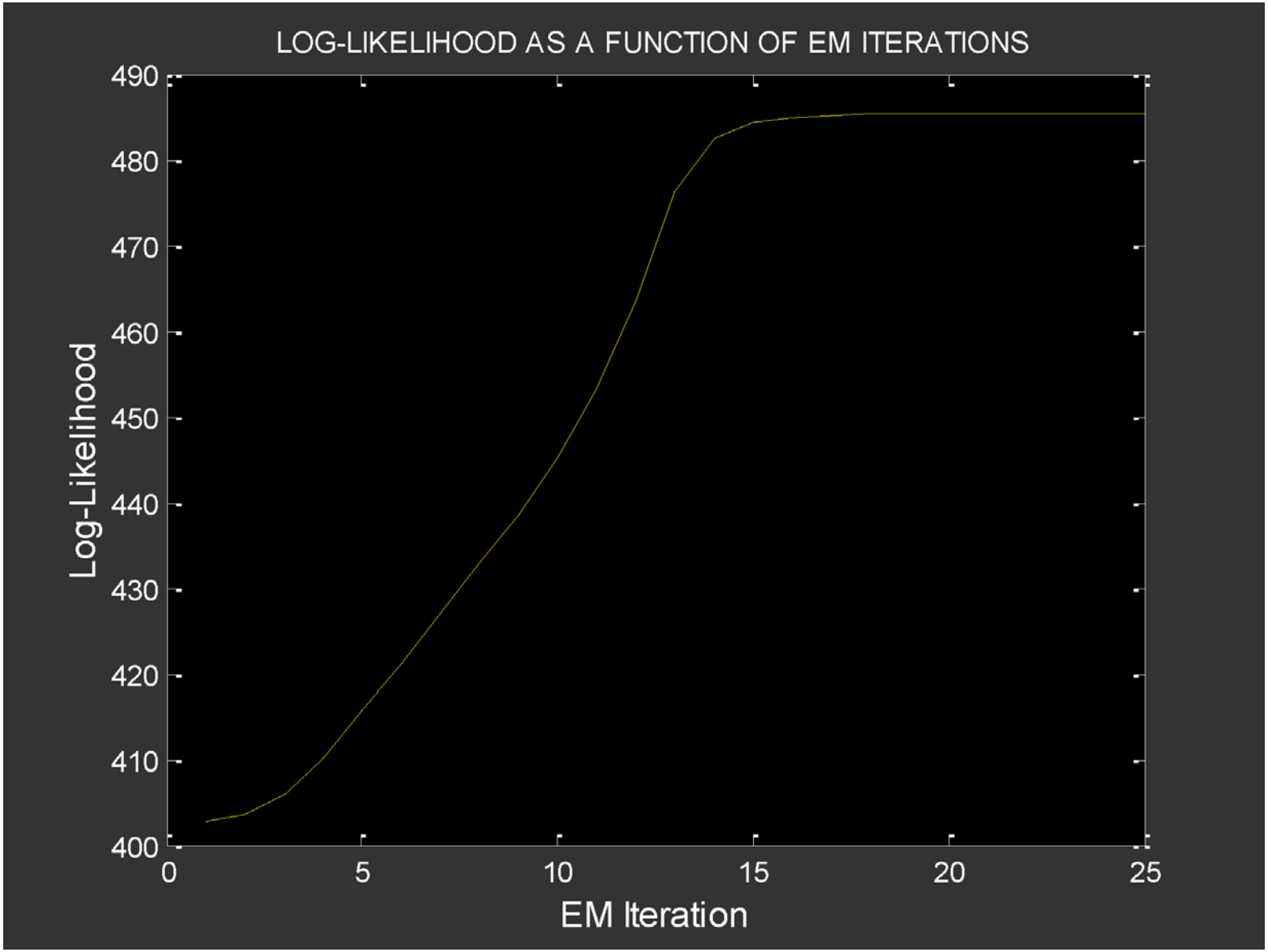

EM ITERATION 5
Red Blood Cell Hemoglobin Concentration
Red Blood Cell Volume
4.4
4.3
4.2
4.1
4
3.9
3.8
3.7
3.3
3.4
3.5
3.6
3.7
3.8
3.9
4


EM ITERATION 10
Red Blood Cell Hemoglobin Concentration
Red Blood Cell Volume
4.4
4.3
4.2
4.1
4
3.9
3.8
3.7
3.3
3.4
3.5
3.6
3.7
3.8


EM ITERATION 15
Red Blood Cell Hemoglobin Concentration
Red Blood Cell Volume
4.4
4.3
4.2
4.1
4
3.9
3.8
3.7
3.3
3.4
3.5
3.6
3.7
3.8
3.9
4


EM ITERATION 25
Red Blood Cell Hemoglobin Concentration
Red Blood Cell Volume
4.4
4.3
4.2
4.1
4
3.9
3.8
3.7
3.3
3.4
3.5
3.6
3.7


ANEMIA DATA WITH LABELS
Red Blood Cell Hemoglobin Concentration
Red Blood Cell Volume
4.4
4.3
4.2
4.1
4
3.9
3.8
3.7
3.3
3.4
3.5
3.6
3.7
3.8
3.9
4


LOG-LIKELIHOOD AS A FUNCTION OF EM ITERATIONS
Log-Likelihood
490
480
470
460
450
440
430
420
410
400
0
5
10
15
20
25
EM Iteration

BIRCH

T. Zhang, R. Ramakrishnan and M. Linvy. BIRCH: An Efficient Data Clustering Method for Very Large Databases, SIGMOD 1996

BIRCH

- BIRCH: Balanced Iterative Reducing and Clustering using Hierarchies
- ΣΤΟΧΟΣ: μείωση του χρόνου εισόδου/εξόδου (I/O)
  - Μεγάλα Σύνολα Δεδομένων
  - Περιορισμένη μνήμη (πολύ μικρότερη από το μέγεθος των δεδομένων)
  - Κόστος I/O γραμμικό στο μέγεθος του συνόλου δεδομένων
- ΛΥΣΗ: Αντί να κρατάμε όλα τα σημεία μιας συστάδας κρατάμε κάποια «στατιστικά» για κάθε συστάδα και για τις σχέσεις μεταξύ των συστάδων
  - Αρκεί ένα πέρασμα (scan) των δεδομένων
  - Ένα ή περισσότερα επιπρόσθετα περάσματα για βελτίωση της ποιότητας της συσταδοποίησης

Χαρακτηριστικά του BIRCH

- Τοπικότητα: κάθε απόφαση σχετικά με συσταδοποίηση παίρνεται χωρίς να χρειάζεται να διαβαστούν όλα τα σημεία ή όλες οι υπάρχουσες συστάδες
- Σημεία σε αραιές περιοχές θεωρούνται οριακά (outliers) και (προαιρετικά) αφαιρούνται
- Λαμβάνει υπ’ όψιν τη διαθέσιμη μνήμη

Ορισμοί (για μια συστάδα)

Έστω μια συστάδα σημείων: $\{\vec{X}_i\}$

Centroid(κεντρικό σημείο): $\vec{X0} = \frac{\sum_{i=1}^{N} \vec{X}_i}{N}$

Radius (ακτίνα) *μέση απόσταση των σημείων της συστάδας από το κεντρικό σημείο*

$$R = \left(\frac{\sum_{i=1}^{N} (\vec{X}_i - \vec{X0})^2}{N}\right)^{\frac{1}{2}}$$

Diameter (διάμετρος): *μέση ανα-δύο απόσταση των σημείων της συστάδας*

$$D = \left(\frac{\sum_{i=1}^{N} \sum_{j=1}^{N} (\vec{X}_i - \vec{X}_j)^2}{N(N-1)}\right)^{\frac{1}{2}}$$

Περιορισμός: Παράγει σφαιρικές συστάδες

Ορισμοί (μεταξύ συστάδων)

Μας ενδιαφέρει και η απόσταση των κεντρικών σημείων δυο συστάδων

Απόσταση δυο συστάδων = απόσταση των κεντρικών σημείων των συστάδων

centroid Euclidean distance

$$D0 = ((\vec{X}0_1 - \vec{X}0_2)^2)^{\frac{1}{2}}$$

centroid Manhattan distance

$$D1 = |\vec{X}0_1 - \vec{X}0_2| = \sum_{i=1}^{d} |\vec{X}0_1^{(i)} - \vec{X}0_1^{(i)}|$$

Ορισμοί (μεταξύ συστάδων)

average inter-cluster (D2) μέση απόσταση των σημείων της μιας συστάδας από τα σημεία της άλλης

$$D2 = \left(\frac{\sum_{i=1}^{N_1}\sum_{j=N_1+1}^{N_1+N_2}(\vec{X}_i - \vec{X}_j)^2}{N_1 N_2}\right)^{\frac{1}{2}}$$

intra-cluster (D3) μέση απόσταση όλων των σημείων D της συστάδας

$$D3 = \left(\frac{\sum_{i=1}^{N_1+N_2}\sum_{j=1}^{N_1+N_2}(\vec{X}_i - \vec{X}_j)^2}{(N_1+N_2)(N_1+N_2-1)}\right)^{\frac{1}{2}}$$

variance increase (D4)

Νέα Απόσταση

Σε συστάδα μετά τη συγχώνευση

$$D4 = \left(\sum_{k=1}^{N_1+N_2}\left(\vec{X}_k - \frac{\sum_{l=1}^{N_1+N_2}\vec{X}_l}{N_1+N_2}\right)^2 - \sum_{i=1}^{N_1}\left(\vec{X}_i - \frac{\sum_{l=1}^{N_1}\vec{X}_l}{N_1}\right)^2 - \sum_{j=N_1+1}^{N_1+N_2}\left(\vec{X}_j - \frac{\sum_{l=N_1+1}^{N_1+N_2}\vec{X}_l}{N_2}\right)^2\right)^{\frac{1}{2}}$$

Απόσταση στο C_i

Απόσταση στο C_j

Περίληψη της συστάδας

- Clustering Feature (CF): μια περίληψη μιας υπο-συστάδας δεδομένων - Μια τριάδα (αριθμός-σημείων, γραμμικό-άθροισμα-σημείων-συστάδας, άθροισμα-τετραγώνου-σημείων-συστάδας)

$$\text{Given a cluster } \{\vec{X}_1, \vec{X}_2, \ldots, \vec{X}_N\}$$

$$\mathbf{CF} = (N, \vec{LS}, SS)$$

N is the number of data points

$$\vec{LS} = \sum_{i=1}^{N} \vec{X}_i$$

$$SS = \sum_{i=1}^{N} \vec{X}_i^{\,2}$$

- Σημαντική (προσθετική) ιδιότητα:

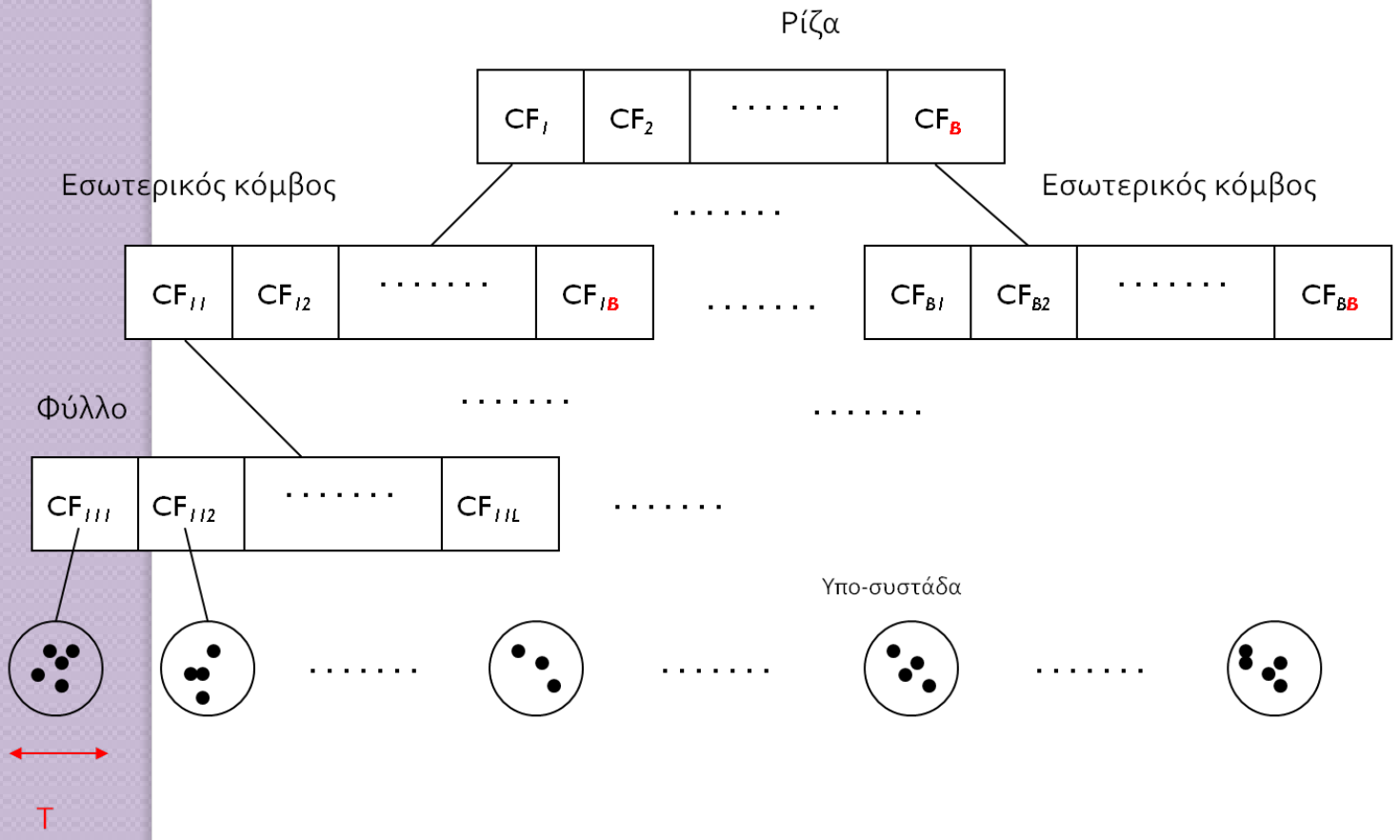
$$CF_1+CF_2=(N_1+N_2,LS_1+LS_2,SS_1+SS_2)$$

Χρήση του Clustering Feature (CF)

- Με το CF οι εγγραφές είναι συνοπτικές – πολύ λιγότερη πληροφορία από ότι όλα τα σημεία της υπο-συστάδας
- Λόγω της προσθετικής ιδιότητας μπορούμε να συγχωνεύσουμε δυο υπο-συστάδες σταδιακά
- Μια εγγραφή CF έχει αρκετή πληροφορία για να υπολογίσουμε τα D0-D4

BIRCH: Ιεραρχικός αλγόριθμος

- Χτίζει σταδιακά καθώς διαβάζει τα δεδομένα ένα δεντρό-γραμμα του οποίου κόμβοι είναι οι τιμές CF που περιγράφουν τα δεδομένα κάθε υπο-συστάδας



Ένα δέντρο ζυγισμένο σε ύψος με 3 παραμέτρους: branching factor B, threshold T και ένα φύλλο που περιέχει το πολύ L στοιχεία

BIRCH: CF-δέντρο εισαγωγή στοιχείου

- Ο αλγόριθμος διαβάζει τα δεδομένα και τα εισάγει στο CF δέντρο ένα-ένα

- Η εισαγωγή ενός στοιχείου στο CF-δέντρο γίνεται με top-down διάσχιση ξεκινώντας από τη ρίζα με βάση μια συνάρτηση απόστασης Distance(σημείο, cluster)
- Χρήση της D0, D1, D2, D3 ή D4

- Κάθε σημείο εισάγεται στην κοντινότερη υπο-συστάδα που υπάρχει σε κάποιο από τα φύλλα

Βήματα

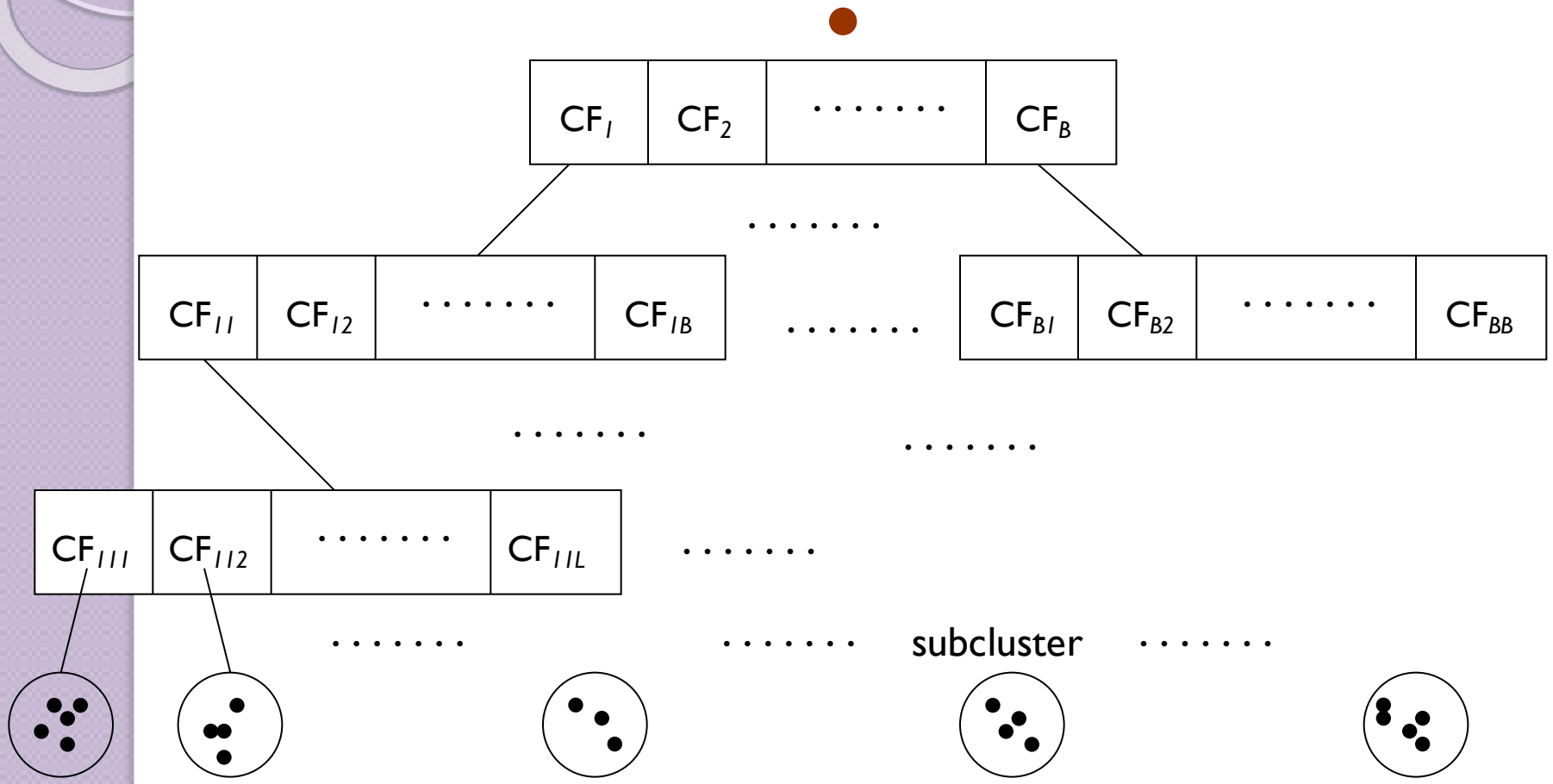
1. Εύρεση κατάλληλου φύλου

   αν το φύλλο μπορεί να το απορροφήσει (διάμετρος παραμένει <= T) ok,

   Αλλιώς 3

2. Ενημέρωση του φύλλου
3. Διάσπαση φύλλου
4. Ενημέρωση τιμής CF

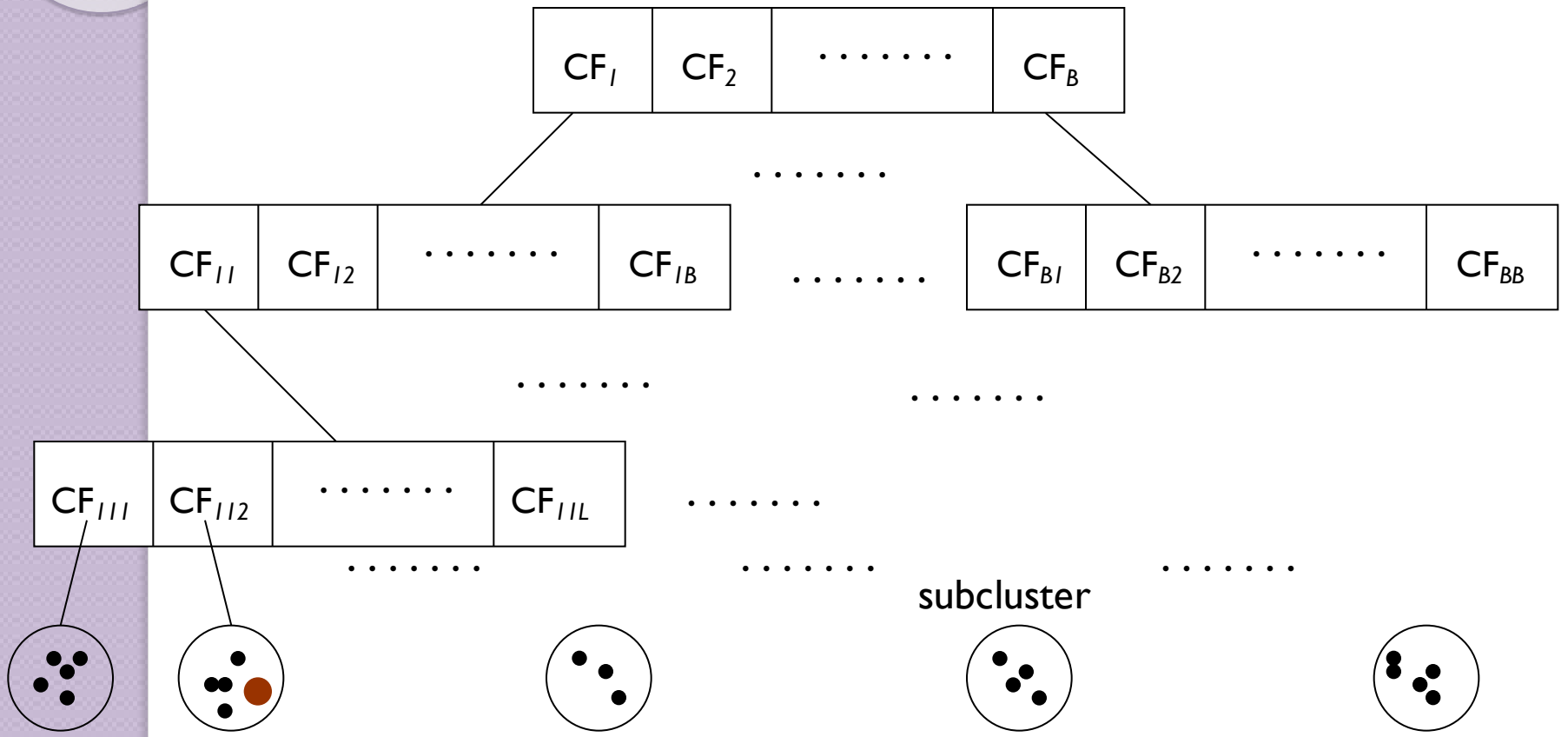
Παράδειγμα: 1. Εύρεση φύλλου



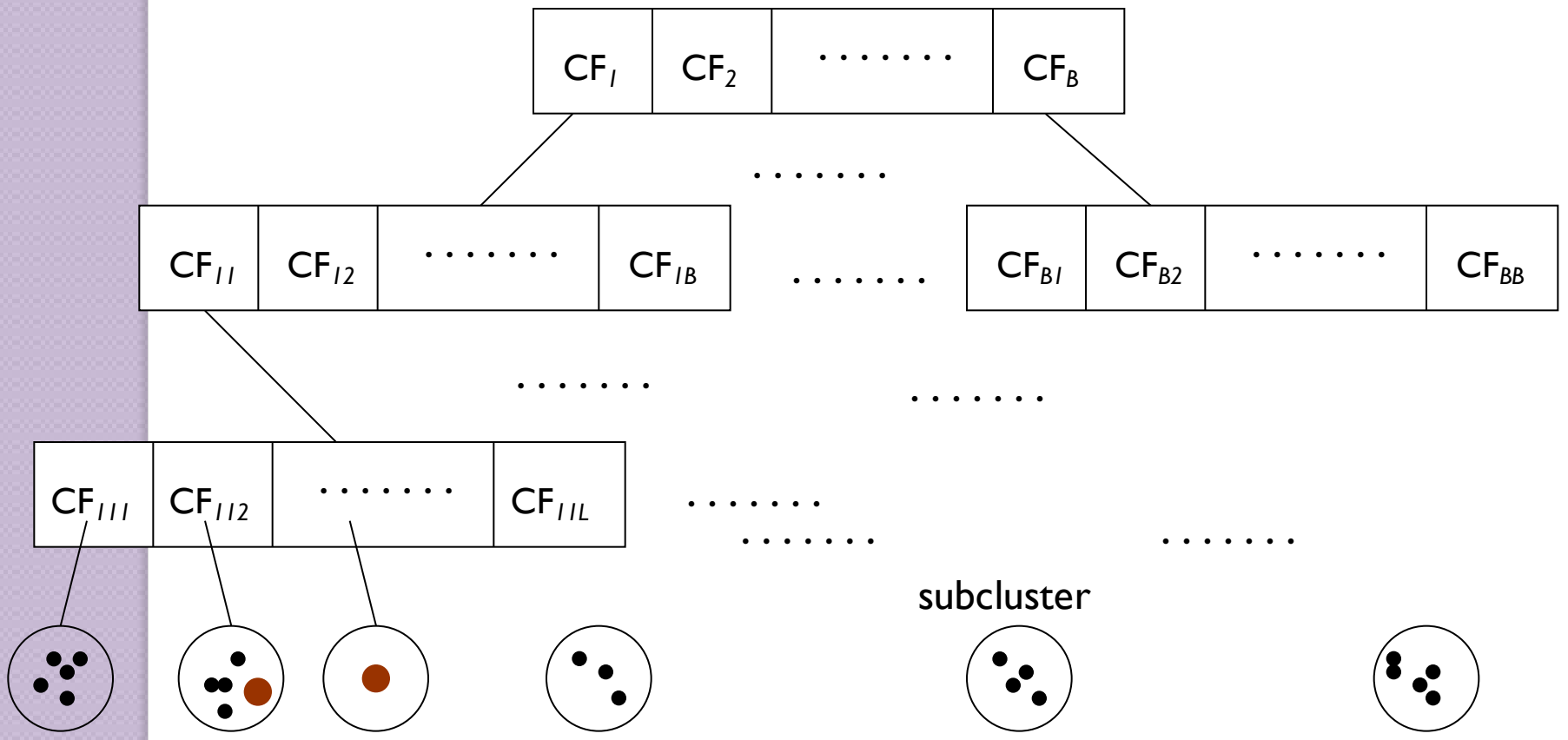
Ενημέρωση CF

- Κάθε σημείο εισάγεται στο κοντινότερη υπο-συστάδα που υπάρχει σε κάποιο από τα φύλλα
  - Αν η εισαγωγή ενός σημείου μεγαλώσει τη διάμετρο της υποσυστάδας πάνω από T, τότε έχουμε δημιουργία νέας υποσυστάδας
    - Αν η νέα συστάδα χωρά στο φύλλο, ok -> ενημέρωση προγόνων
- Αν η νέα συστάδα δε χωρά (σε μια σελίδα) -> υπερχείλιση στο φύλλο

Παράδειγμα: 4. Ενημέρωση CF τιμών



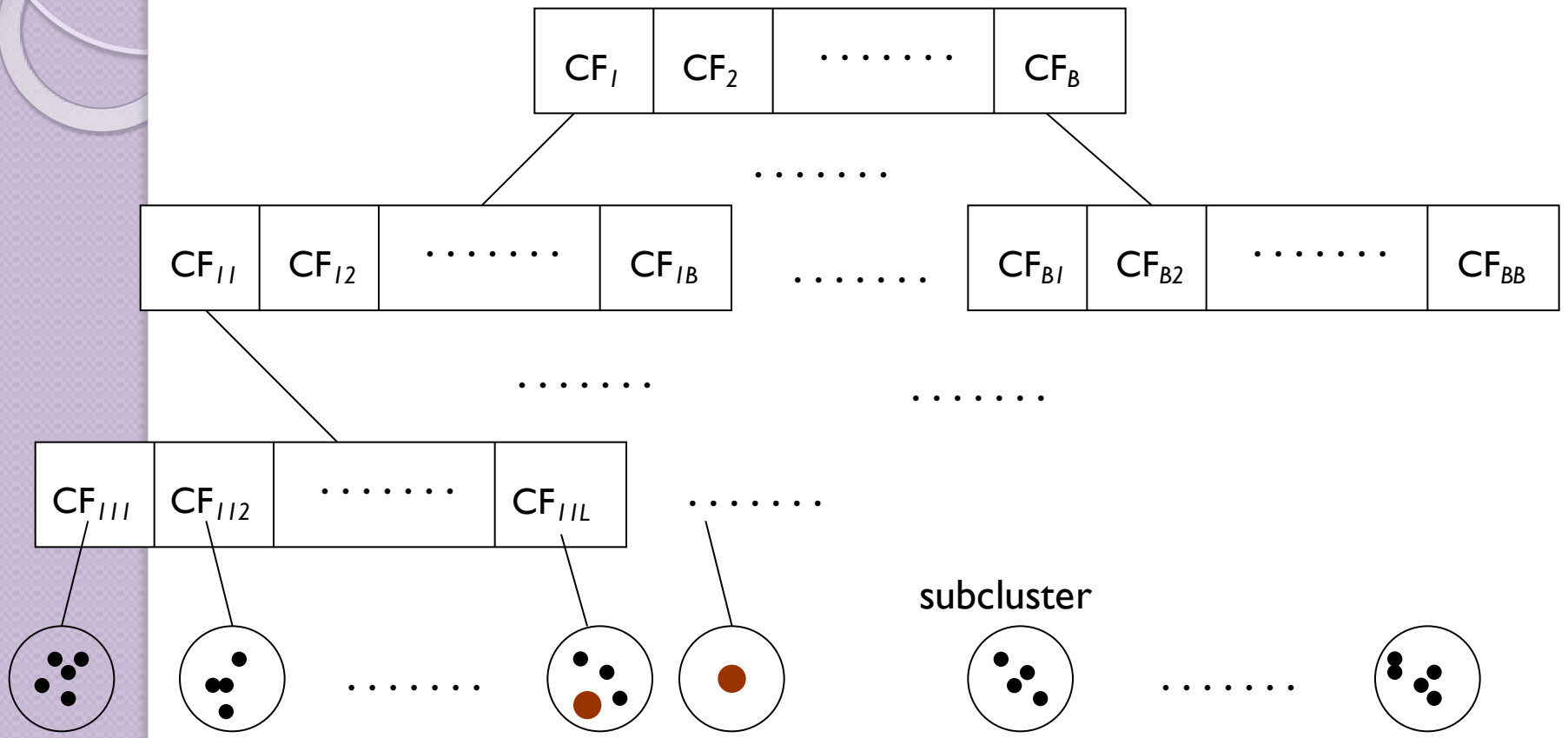
Παράδειγμα: 2. Ενημέρωση φύλλου



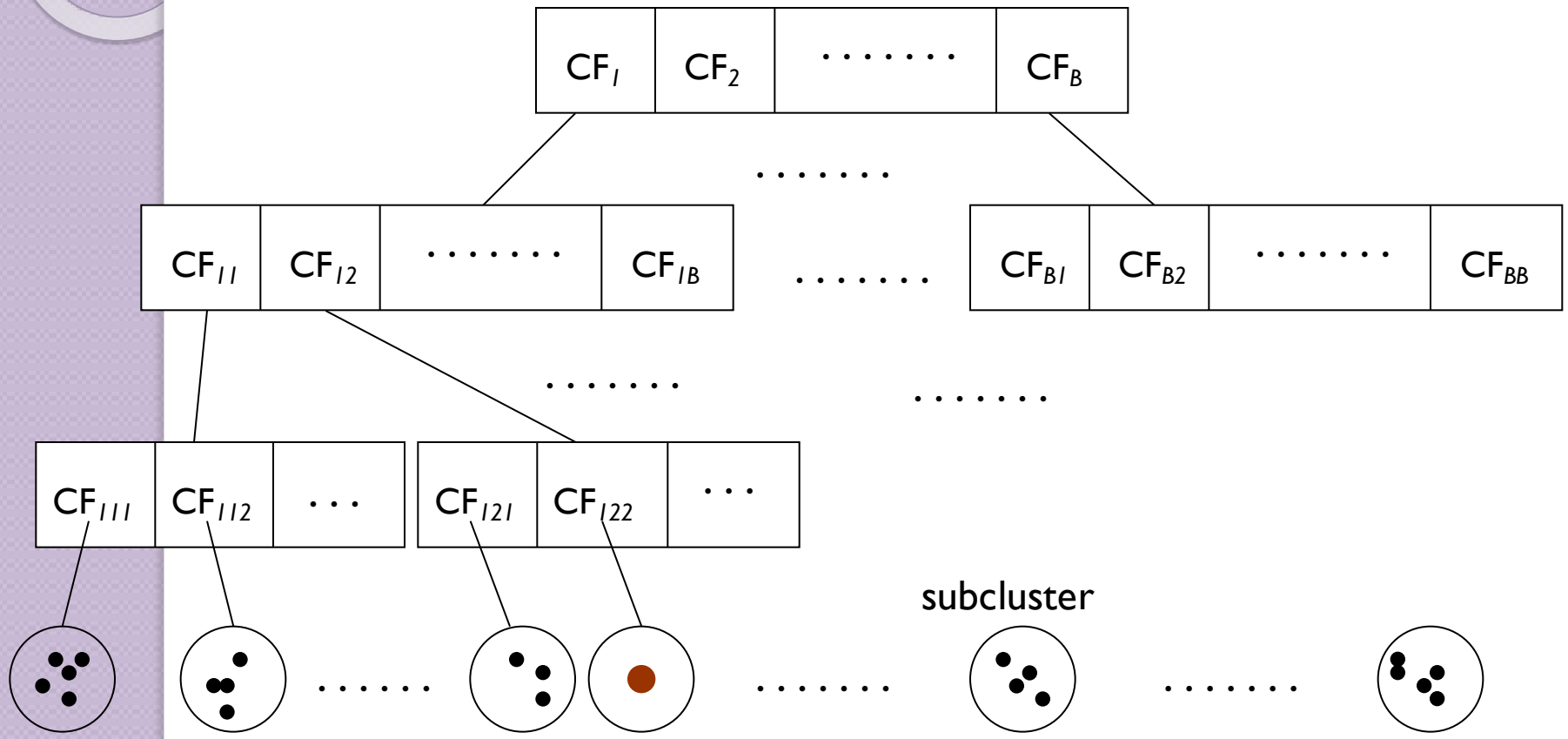
Διάσπαση φύλλου (split)

- Δημιουργία νέου φύλλου και μοίρασμα των συστάδων. Πώς;
- Εύρεση των δύο υπο-συστάδων του φύλλου πού έχουν τη μεγαλύτερη απόσταση μεταξύ τους, έστω Ci και Cj
  - Αυτές οι δύο αποτελούν το κριτήριο διάσπασης των υπο-συστάδων του φύλλου – κάθε μια από αυτές σε ένα από τα δύο νέα φύλλα
  - όλες οι άλλες υπο-συστάδες C ανατίθενται στο φύλλο της Ci ή στο φύλλο της Cj με βάση ποια από τις δύο είναι πιο όμοιά της

Παράδειγμα: 3-4 Διάσπαση και ενημέρωση του μονοπατιού από τη ρίζα



Παράδειγμα: 3-4 Διάσπαση και ενημέρωση του μονοπατιού από τη ρίζα

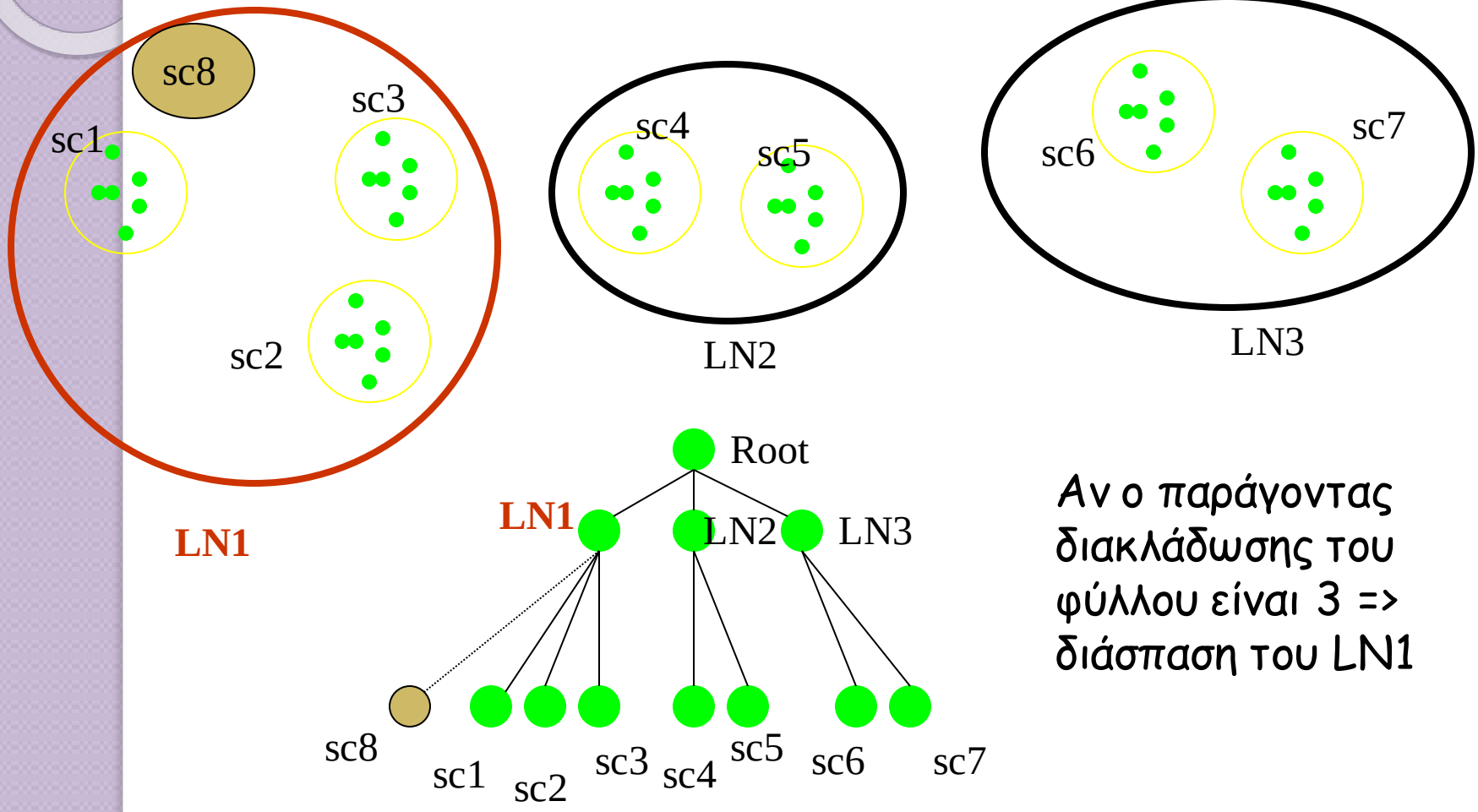


Διασπάσεις

- Διάσπαση φύλλου μπορεί να οδηγήσει σε υπερχείλιση εσωτερικού κόμβου (όταν περιέχει περισσότερα παιδιά από ότι ο παράγοντας διακλάδωσης)
- Διάσπαση εσωτερικού κόμβου
  - Οι εσωτερικοί κόμβοι διασπούνται αναδρομικά με βάση μια μέτρηση της απόσταση των συστάδων τους
- Διάσπαση της ρίζας, οδηγεί σε αύξηση του ύψους του δέντρου κατά 1

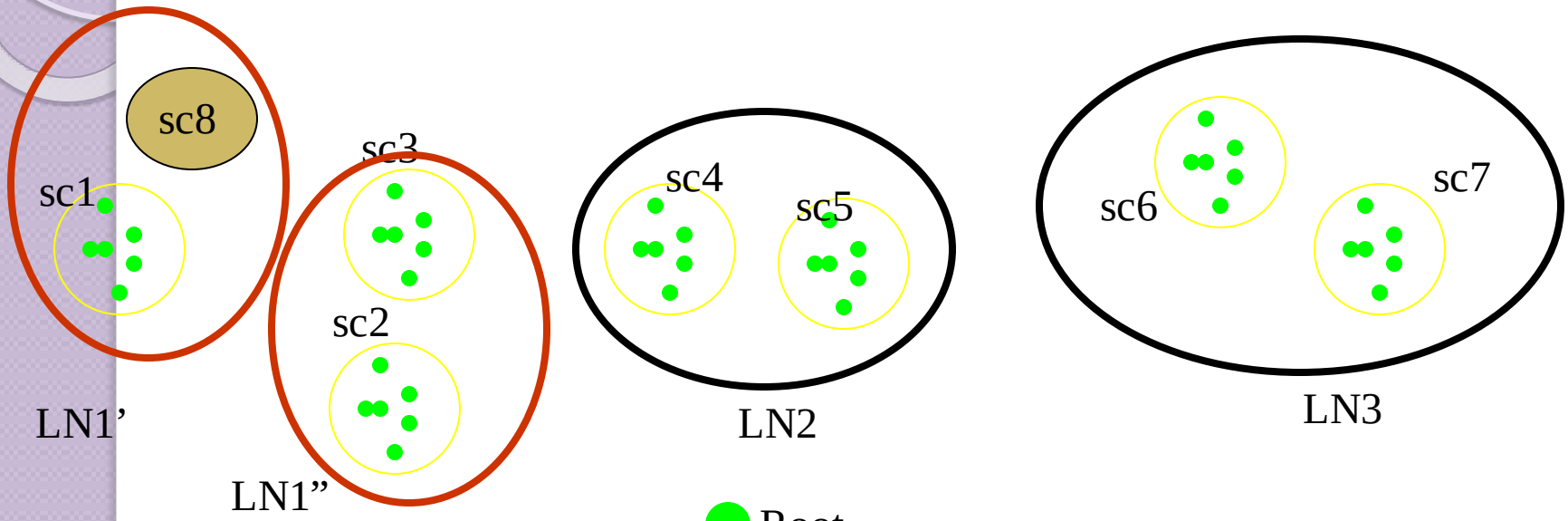
BIRCH: CF-δέντρο

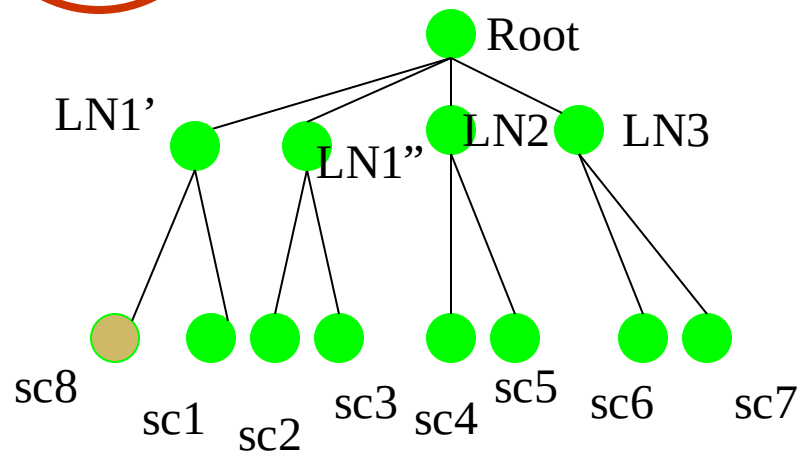
Νέα υπο-συστάδα



Αν ο παράγοντας διακλάδωσης του φύλλου είναι 3 => διάσπαση του LN1

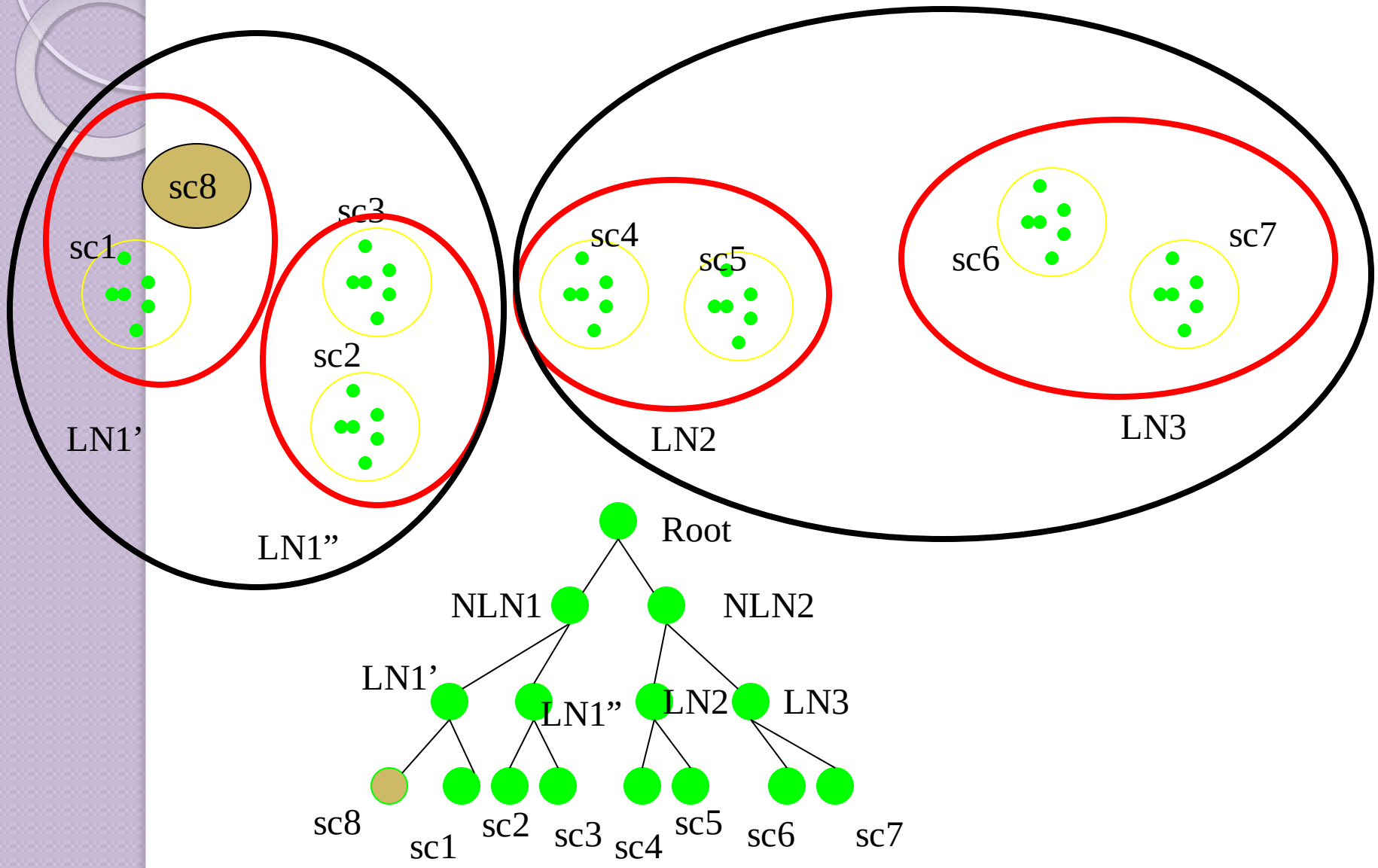
BIRCH: CF-δέντρο





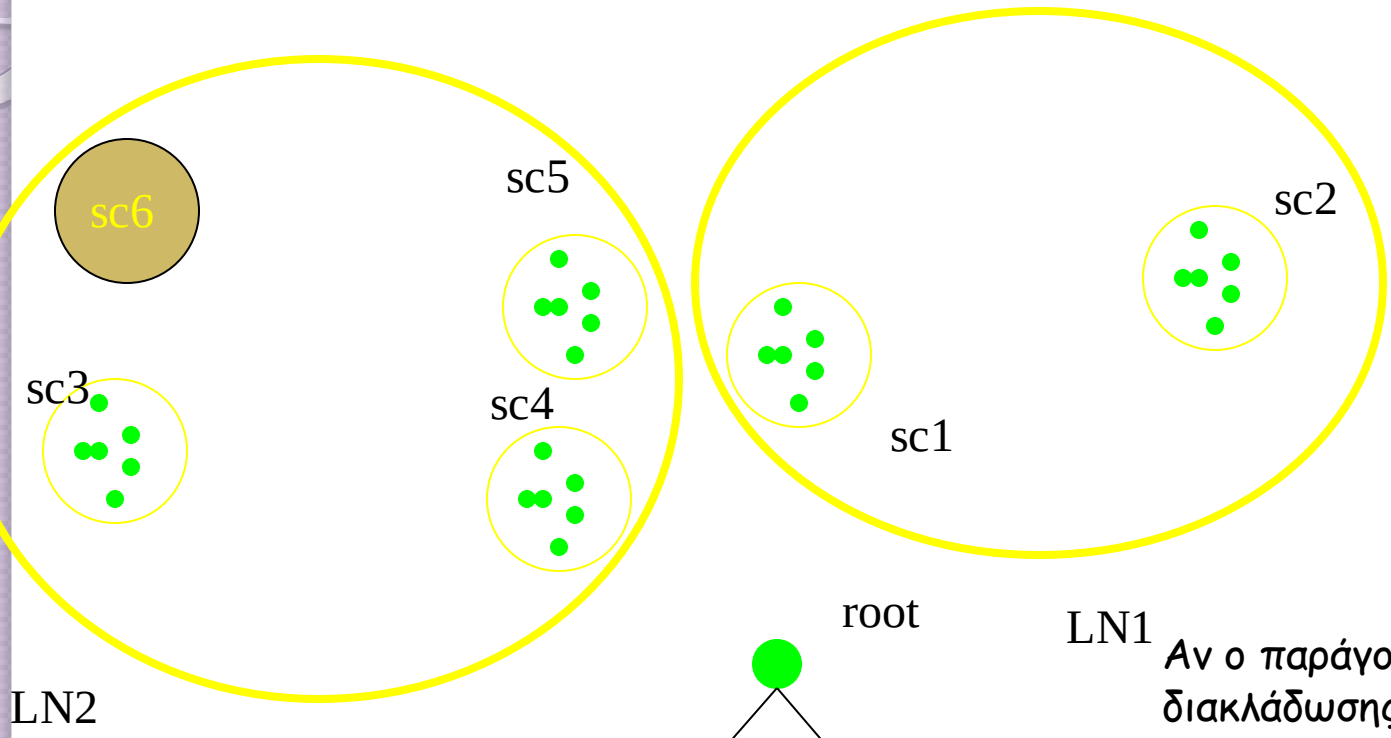
Αν ο παράγοντας διακλάδωσης εσωτερικού κόμβου είναι είναι 3 => διάσπαση της ρίζας και αύξηση του ύψους

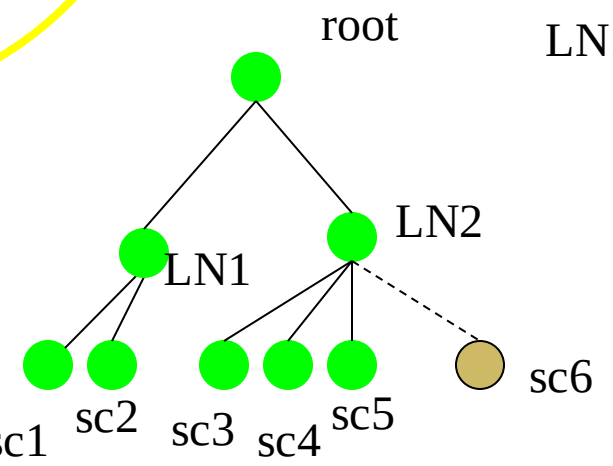
BIRCH: CF-δέντρο



BIRCH: CF-δέντρο

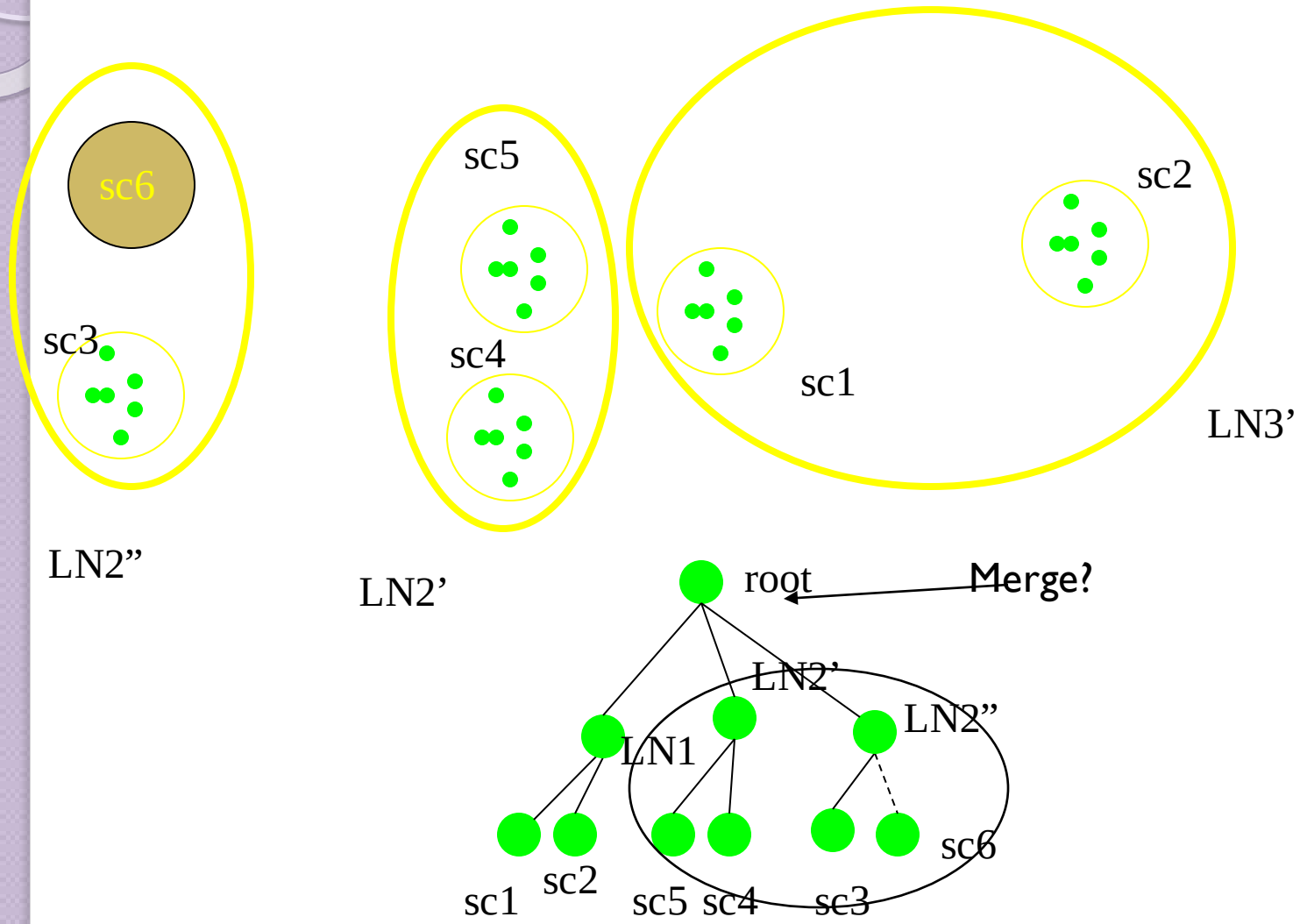
Έστω ότι η αρίθμηση των υποσυστάδων αντιστοιχεί στη σειρά δημιουργίας τους



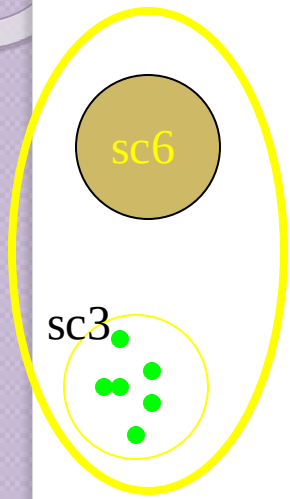


Αν ο παράγοντας διακλάδωσης του φύλλου είναι 3 => διάσπαση του LN2

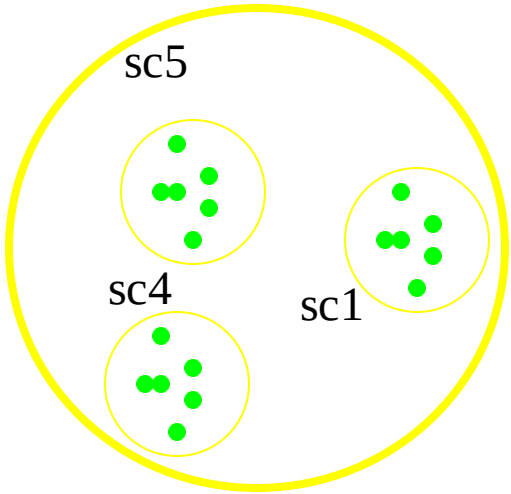
BIRCH: CF-δέντρο



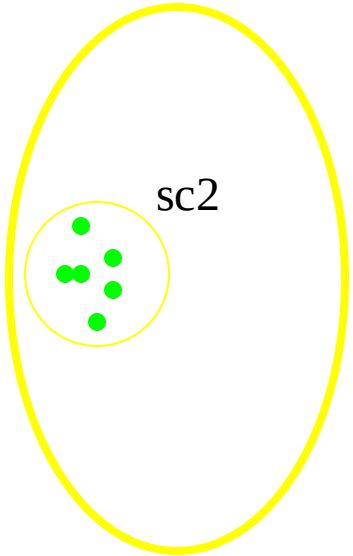
BIRCH: CF-δέντρο



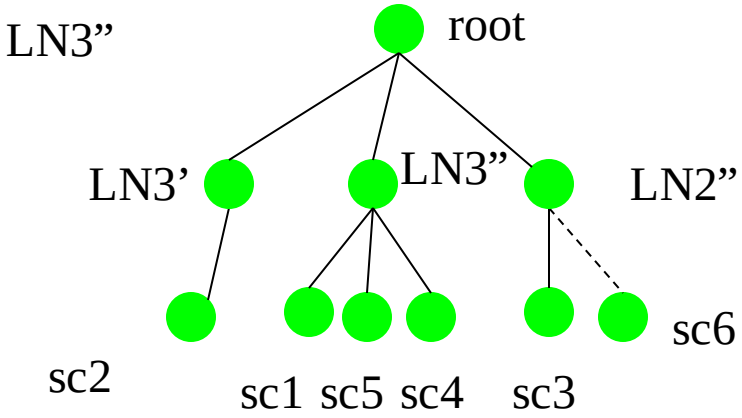
LN2”



LN3”



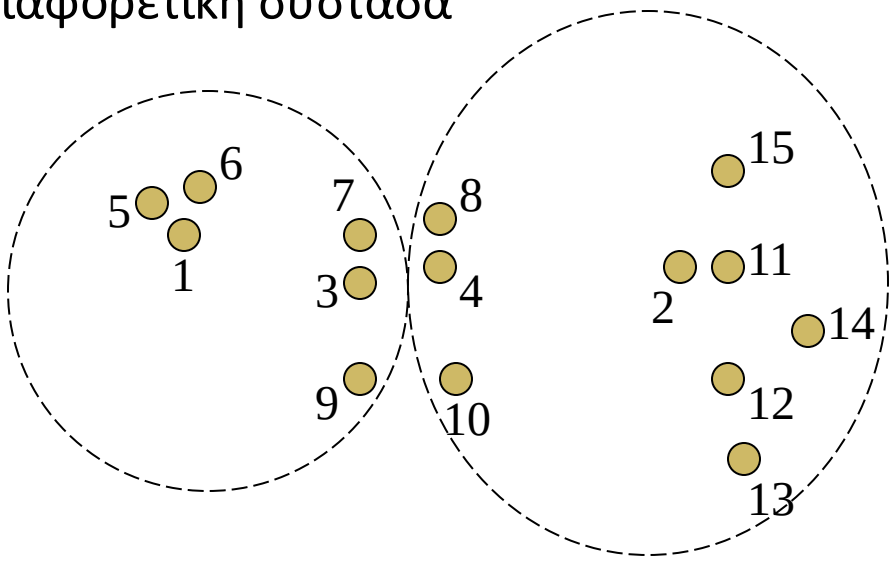
LN3’



Συγχώνευση LN2’ και LN1 και ο καινούργιος κόμβος θα διασπαστεί πάλι

BIRCH: αλγόριθμος

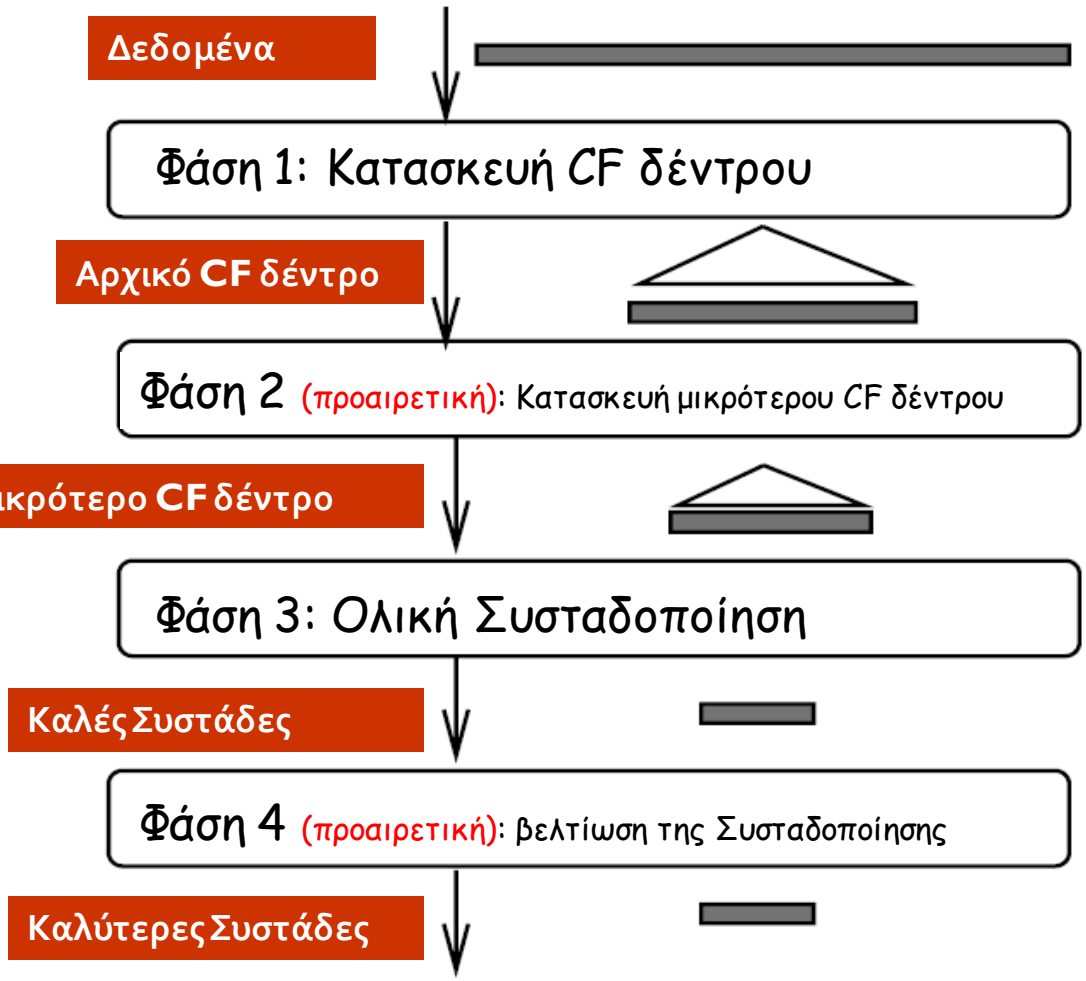
- Επειδή η κατασκευή επηρεάζεται από το μέγεθος της σελίδας:
  - Οι συστάδες που δημιουργούνται μπορεί να μην είναι πραγματικές
  - ανάλογα με το skew (κατανομή) και τη σειρά που έρχονται τα δεδομένα
- Επίσης, αν ξανά-εισάγουμε ένα σημείο μπορεί να εισαχθεί σε διαφορετική συστάδα



Αριθμός αντιστοιχεί στη σειρά εισαγωγής,

Έστω dist(1, 2) > T

BIRCH-αλγόριθμος



Φάση 1: Μια δομή κύριας μνήμης που συνοψίζει τα δεδομένα

Φάση 2: Κοιτά τα φύλλα και προσπαθεί να διώξει τους outliers και να ενοποιήσει «όμοιες» συστάδες που αντιστοιχούν σε περιοχές με πολλά σημεία

Χρειάζεται για να βελτιώσει τη Φάση 3

BIRCH

Φάση 3

Ξανα-συσταδοποιεί τα φύλλα του δέντρου

Γιατί;

Πχ κοντινές συστάδες που (έτυχε να) είναι σε διαφορετικά φύλλα

Πως;

- Για κάθε συστάδα που εμφανίζεται στα φύλλα, υπολογίζουμε το κεντρικό της σημείο (centroid) και τα θεωρούμε ως αρχικά σημεία – αυτά τα αρχικά σημεία μπορούμε να τα συσταδοποιήσουμε χρησιμοποιώντας έναν οποιαδήποτε αλγόριθμο συσταδοποίησης
- Εναλλακτικά, μπορούμε να συσταδοποιήσουμε τις συστάδες ως έχουν – πχ με έναν ιεραρχικό συγκεντρωτικό αλγόριθμο

BIRCH

Φάση 4 (προαιρετική)

Χρησιμοποιεί τα κεντρικά σημεία των συστάδων που παράγει η Φάση 3 ως seeds, και αναδιανέμει όλα τα στοιχεία εισόδου (δεύτερο πέρασμα!)

Μπορεί να έχουμε και παραπάνω από ένα επιπρόσθετα περάσματα (έχει αποδειχτεί σύγκλιση)

- Εξασφαλίζει ότι όλα τα αντίγραφα ενός σημείου πάνε στην ίδια συστάδα
- Μπορούμε επίσης να βάλουμε ως ετικέτα σε κάθε σημείο, τη συστάδα που ανήκει
- Μπορούμε να απαλλαγούμε από outliers (πχ σημεία πολύ μακριά από όλα τα seeds)