Συσταδοποίηση (Clustering)

Μηχανική Μάθηση

ΔΠΜΣ Επιστήμη Δεδομένων και Μηχανική Μάθηση

Γιώργος Αλεξανδρίδης – gealexandri@islab.ntua.gr

Εισαγωγή

Χαρακτηριστικά Δεδομένων

- **Μη-επιγεγραμμένα**
- **Διαστατικότητα**
  - Όσο μεγαλώνουν οι διαστάσεις, τόσο μικραίνει η πυκνότητα των δεδομένων ⇒ η εγγύτητα των προτύπων τείνει να γίνεται πιο ομοιόμορφη
- **Πλήθος προτύπων**
  - Οι περισσότεροι αλγόριθμοι συσταδοποίησης λειτουργούν καλά για μικρά ή μεσαία datasets
- **Αραιότητα**
- **Θόρυβος** και **έκτοπες τιμές** (*outliers*)
  - Επηρεάζουν την ποιότητα της συσταδοποίησης
- **Μαθηματικές Ιδιότητες** του χώρου των δεδομένων
  - Πως ορίζεται η εγγύτητα των προτύπων;

Ορισμός

- Χωρισμός δεδομένων σε ομάδες που έχουν **νόημα** ή/και είναι **χρήσιμες**
- Νόημα ⇒ Συσταδοποίηση με στόχο την *κατανόηση*
  - Βιολογία
  - Ανάκτηση Πληροφορίας
  - Περιβάλλον
  - Ιατρική
  - Επιχειρήσεις
  - …
- Χρησιμότητα ⇒ Συσταδοποίηση με στόχο την *ωφέλεια* (utility)
  - Σύνοψη
  - Συμπίεση
  - Εύρεση πλησιέστερων γειτόνων

Συστάδα

- Η έννοια της συστάδας δεν είναι *σαφώς ορισμένη*
- Βασικό επιθυμητό χαρακτηριστικό συστάδας
  - "*Ίδια*" (σχετιζόμενα) πρότυπα *εντός* της *ίδιας* συστάδας
  - "*Διαφορετικά*" (μη-σχετιζόμενα) από τα πρότυπα που εντάσσονται σε *άλλες* συστάδες
- Συσταδοποίηση
  - Ειδική μορφή ταξινόμησης, όπου ο αλγόριθμος δημιουργεί τις ετικέτες των προτύπων
  - **Μη-επιβλεπόμενη μάθηση** (*unsupervised learning*)
- Ταξινόμηση
  - **Επιβλεπόμενη μάθηση** (*supervised learning*)

Τύποι Συστάδων

- **Σαφώς χωρισμένες**
  - Όλα τα σημεία που ανατίθενται σε μια συστάδα είναι εγγύτερα μεταξύ τους (απ' ότι σε σημεία που ανήκουν σε άλλες συστάδες)
- **Σημειακές**
  - Κάθε σημείο είναι εγγύτερα στο κέντρο της δικής του συστάδας απ' ότι στα κέντρο άλλων συστάδων
- **Γραφοθεωρητικές**
  - Σε περίπτωση που τα πρότυπα μπορούν να αναπαρασταθούν υπό τη μορφή γράφου, τότε συστάδες αποτελούν οι *συνεκτικές συνιστώσες* του
- Βασισμένες στην **πυκνότητα**
  - Οι συστάδες οριοθετούνται από τις εναλλαγές πυκνών και αραιών περιοχών στον χώρο των προτύπων
- **Εννοιολογικές**
  - Τα πρότυπα μοιράζονται κάποια ιδιότητα

Χαρακτηριστικά συστάδων

- **Κατανομή** δεδομένων
  - Συγκεκριμένη κατανομή
  - Μείξη κατανομών
- **Σχήμα**
  - Συγκεκριμένης μορφής
  - Οποιασδήποτε μορφής
- **Διαχωρισιμότητα**
  - Μπορούν οι αλγόριθμοι να διαχωρίσουν συστάδες που συμπίπτουν;
- **Σχέσεις** μεταξύ των συστάδων
  - Έχει σημασία σε ορισμένους αλγορίθμους (π.χ. SOM)

Ταξινόμηση συσταδοποίησης

- **Σχέση** μεταξύ των συστάδων
  1. **Διαμεριστική** (*Partitional*): Ανεξάρτητες, μη-επικαλυπτόμενες συστάδες
  2. **Ιεραρχική** (*Hierarchical*): Επικαλυπτόμενες συστάδες που σχηματίζουν μια ιεραρχία (π.χ. δένδρο)

- **Συμμετοχή** των προτύπων
  1. **Αποκλειστική** (*Exclusive* ή *Hard*): Κάθε πρότυπο σε μία μόνο συστάδα
  2. **Μη-αποκλειστική** (*Non-exclusive* ή *Soft*)
     - **Επικαλυπτόμενη** (*Overlapping*): Πρότυπο μπορεί να συμμετέχει σε 2 ή περισσότερες συστάδες
     - **Ασαφής** (*Fuzzy*): Βαθμός συμμετοχής (*membership*) προτύπου σε κάθε συστάδα

- **Τύπος** συσταδοποίησης
  1. **Πλήρης**: Κάθε πρότυπο ανατίθεται σε μια τουλάχιστον συστάδα
  2. **Μερική**: Ορισμένα πρότυπα ανατίθενται σε συστάδες
     - Τα υπόλοιπα μπορεί να είναι έκτοπες τιμές, θόρυβος κλπ

Αλγόριθμοι συσταδοποίησης

- **Διαμεριστικοί**
  - *k*-μέσων (*k*- means), *k*- μεσαίων (*k*- medoids), CLARANS
- **Ιεραρχικοί**
  - DiAna, AgNes, BIRCH, CAMELEON
- Βασισμένοι στην **πυκνότητα**
  - DBSCAN, OPTICS, DenClue
- Βασισμένοι στην κατασκευή **πλέγματος** (*grid-based*)
  - STING, WaveCluster, CLIQUE

Χαρακτηριστικά αλγορίθμων συσταδοποίησης

- **Μη-ντετερμινιστικοί**
  - Εξάρτηση από τη *σειρά* προσπέλασης των *προτύπων*
  - Εξάρτηση από την *αρχικοποίηση* (π.χ. αλγόριθμος k-μέσων)
- **Παραμετρικοί**
  - Μικρές αλλαγές στις τιμές των παραμέτρων ⇒ μεγάλη επίδραση στην ποιότητα της συσταδοποίησης
  - Εύρεση παραμέτρων με εξαντλητική αναζήτηση
- Δυνατότητα **Κλιμάκωσης**
  - Ψευδο-πολυωνυμικοί αλγόριθμοι

Διαμεριστικοί αλγόριθμοι

Χαρακτηριστικά

- Χωρίζουν τον χώρο εισόδου σε k συστάδες
- **Σημειακή** ή ***κεντρική*** συσταδοποίηση
  - **Σκληρή** (Hard): Κάθε πρότυπο ανήκει σε μία και μόνο ομάδα
    - Αλγόριθμος k-μέσων (k- means)
    - Αλγόριθμος k-μεσαίων (k-medoids)
      - Διαφορά από k- means: Το κέντρο της συστάδας είναι το πλησιέστερο πρότυπο στο κέντρο βάρους της
  - **Μαλακή** (*Soft*): Μπορεί να ανήκει σε περισσότερες από μια ομάδες
    - **Βαθμός Συμμετοχής** (*Degree of membership*)
    - **Ασαφής Συσταδοποίηση** (*Fuzzy clustering*)
      - Αλγόριθμος fuzzy c-means
- **Χωρική** συσταδοποίηση
  - Διάταξη κέντρων στον χώρο υπό τη μορφή *πλέγματος*
    - Αλγόριθμος Αυτό-οργανούμενων χαρτών (*Self-Organizing Maps* – SOM)
- **Στατιστική** συσταδοποίηση

Στατιστική συσταδοποίηση

- **Υπόθεση**
  - Τα δεδομένα προέρχονται από την **ανάμειξη** (*mixture*) *διαφορετικών* κατανομών
  - Κάθε συστάδα εκφράζει μια κατανομή
- **Ζητούμενο**
  - Να εντοπίσουμε τις εν λόγω κατανομές μαζί με τις παραμέτρους τους
- **Μεθοδολογία**
  - Εντοπισμός παραμέτρων κάθε κατανομής ⇒ **Εκτίμηση Μέγιστης Πιθανοφάνειας** (*Maximum Likelihood Estimation* – MLE)
  - Εντοπισμός παραμέτρων **μοντέλου ανάμειξης** (*mixture model*) ⇒ Αλγόριθμος **Μεγιστοποίησης Αναμονής** (*Expectation-Maximization* – EM)

Μοντέλα Μίξης

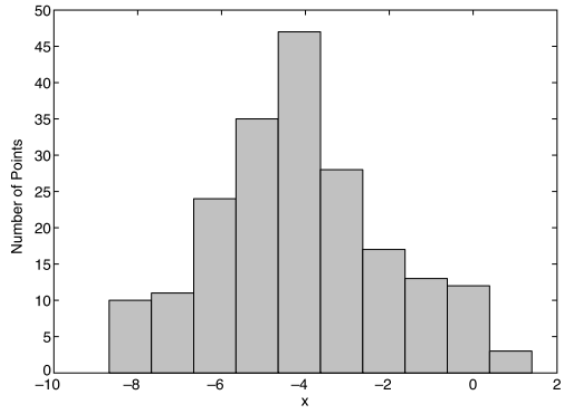
- Mixture Models
- M πρότυπα εισόδου $\mathcal{X} = \{x_1, x_2, \dots, x_M\}$ που δημιουργούνται από K **ανεξάρτητες** κατανομές που περιγράφονται από **σύνολο παραμέτρων** $\Theta = \{\theta_1, \theta_2, \dots, \theta_K\}$
- Πιθανότητα το i-οστό πρότυπο εισόδου να προέρχεται από την j-οστή κατανομή $p(x_i|\theta_j)$
  - Αρκετά συχνά επιλέγεται η *πολυμετάβλητη κανονική κατανομή*, γιατί δημιουργεί συστάδες ελλειπτικού σχήματος γύρω από τη μέση τιμή της
- Πιθανότητα εμφάνισης του i-οστού προτύπου $p(x_i|\Theta) = \sum_{j=1}^{K} w_j p(x_i|\theta_j)$
  - w_j: **συνεισφορά** (*βάρος*) j-οστής κατανομής ($\sum_{j=1}^{K} w_j = 1$)
- Τα πρότυπα δημιουργούνται ανεξάρτητα το ένα από το άλλο
  - $p(\mathcal{X}|\Theta) = \prod_{i=1}^{M} p(x_i|\Theta) = \prod_{i=1}^{M} \sum_{j=1}^{K} w_j p(x_i|\theta_j)$

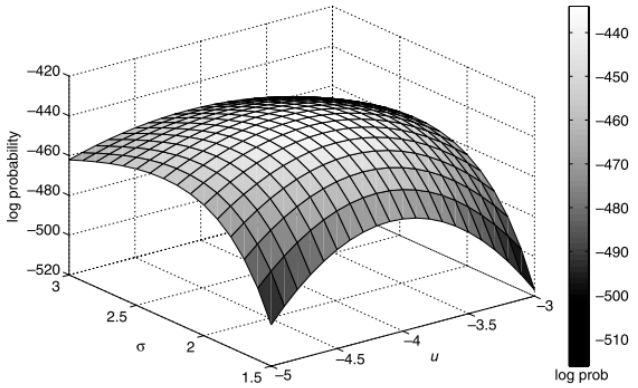
Εκτίμηση μέγιστης πιθανοφάνειας

- $p(\mathcal{X}|\Theta) = \prod_{i=1}^{M} p(x_i|\Theta)$
  - Τα δεδομένα $\mathcal{X}$ είναι *γνωστά* και *αμετάβλητα*.
  - Το **ζητούμενο** είναι οι τιμές των **παραμέτρων** Θ
- Συνάρτηση **πιθανοφάνειας** (*likelihood*) $\mathcal{L}(\Theta|\mathcal{X})$
  - Οι παράμετροι Θ ως συνάρτηση των δεδομένων $\mathcal{X}$
- Εκτίμηση **μέγιστης** πιθανοφάνειας
  - Βρες τις παραμέτρους εκείνες που μεγιστοποιούν την πιθανότητα εμφάνισης των συγκεκριμένων δεδομένων $\mathcal{X}$: $\hat{\theta} \in \left\{\max_{\theta\in\Theta} \mathcal{L}(\theta|\mathcal{X})\right\}$
  - Στην πράξη χρησιμοποιείται συχνά ο *λογάριθμος* της πιθανοφάνειας: $l(\Theta|\mathcal{X}) = \ln \mathcal{L}(\Theta|\mathcal{X})$

Εκτίμηση Μέγιστης Πιθανοφάνειας: Παράδειγμα

- $M = 200$ πρότυπα εισόδου μιας διάστασης που προέρχονται από μία ($K = 1$) κανονική κατανομή, άγνωστης μέσης τιμής μ και τυπικής απόκλισης σ
  - $\mathrm{p}(\mathcal{X}|\Theta) = \mathcal{N}(\mu, \sigma) = \prod_{i=1}^{200} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$
- Εκφράζουμε την κατανομή υπό μορφή πιθανοφάνειας και παίρνουμε τον λογάριθμό της
  - $\ell(\Theta|\mathcal{X}) = -\sum_{i=1}^{200} \frac{(x_i-\mu)^2}{2\sigma^2} - 100 \ln 2\pi - 200 \ln \sigma$
- Λύνουμε τις εξισώσεις $\frac{\partial \ell}{\partial \mu} = 0$ και $\frac{\partial \ell}{\partial \sigma} = 0$
  - Η πιθανοφάνεια μεγιστοποιείται για $\hat{\mu} = -4.1$ και $\hat{\sigma} = 2.1$



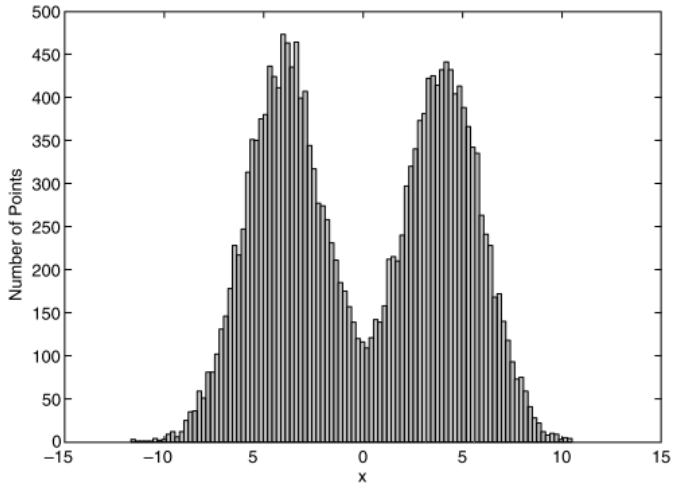


Αλγόριθμος Μεγιστοποίησης Αναμονής

- Στη γενική περίπτωση των μοντέλων μίξης, δεν γνωρίζουμε από ποια από τις K ανεξάρτητες κατανομές προέρχεται το πρότυπο x_i
- **Βήματα** αλγορίθμου
  1. **Αρχικοποίηση** των παραμέτρων του μοντέλου
  2. Επανάληψη
     1. Βήμα **Αναμονής** (*Expectation Step*): Υπολογισμός της πιθανότητας το πρότυπο x_i να προέρχεται από την j-οστή κατανομή $p(j|x_i, \Theta)$
     2. Βήμα **Μεγιστοποίησης** (*Maximization Step*): Χρησιμοποιώντας τα $p(j|x_i, \Theta)$, εκτίμησε τις παραμέτρους Θ που μεγιστοποιούν την πιθανοφάνεια
  3. Μέχρι τη σύγκληση (ή για ένα καθορισμένο αριθμό βημάτων)
- Ομοιότητα με αλγόριθμο k-μέσων
  - Αποτελεί ειδική περίπτωση του ΕΜ για *κανονικές* κατανομές *σφαιρικού* σχήματος με *ίδιους πίνακες συνδιασποράς* αλλά *διαφορετικές* μέσες τιμές
  - Βήμα Αναμονής ⇒ Ανάθεση κάθε προτύπου σε ένα κέντρο
  - Βήμα Μεγιστοποίησης ⇒ Υπολογισμός των νέων κέντρων

Αλγόριθμος Μεγιστοποίησης Αναμονής: Παράδειγμα

- $M = 20.000$ πρότυπα, τα οποία υποθέτουμε ότι προέρχονται από 2 κανονικές κατανομές $\mathcal{N}_1(-2,2), \mathcal{N}_2(3,2)$ με ίσα βάρη ($w_1 = w_2 = \frac{1}{2}$)
- <u>Βήμα Αναμονής</u>:
  - Χρήση κανόνα Bayes $p(j|x_i, \Theta) = \frac{w_j p(x_i|\theta_j)}{\sum_{j=1}^{2} w_j p(x_i|\theta_j)}$
- <u>Βήμα Μεγιστοποίησης</u>:
  - $\mu_j = \sum_{i=1}^{20.000} x_i \frac{p(j|x_i,\Theta)}{\sum_{i=1}^{20.000} p(j|x_i,\Theta)}$
  - Στην περίπτωση της κανονικής κατανομής, η εκτίμηση για τη μέγιστη τιμή του μ είναι ένας ζυγισμένος μέσος όρος των δειγμάτων της



Iteration	μ_1	μ_2
0	−2.00	3.00
1	−3.74	4.10
2	−3.94	4.07
3	−3.97	4.04
4	−3.98	4.03
5	−3.98	4.03

Αλγόριθμος Μεγιστοποίησης Αναμονής: Πλεονεκτήματα και Μειονεκτήματα

- **Πλεονεκτήματα**
  - Πιο *γενικό* μοντέλο από k-μέσους ή fuzzy c-means
    - Μπορεί να χρησιμοποιηθεί πληθώρα κατανομών
  - Εύρεση συστάδων *διαφορετικού* μεγέθους και σχήματος
  - Ευκολότερος *χαρακτηρισμός* των συστάδων από τις παραμέτρους του μοντέλου
    - Αντί ενός κέντρου
- **Μειονεκτήματα**
  - Μπορεί να είναι *αργός*
  - *Μη-πρακτικός* για μοντέλα με πολλές παραμέτρους
  - Δεν λειτουργεί σωστά σε συστάδες με *λίγα* πρότυπα
  - Πρέπει να γίνεται *ορθή εκτίμηση* του πλήθους των συστάδων
    - Ύπαρξη σχετικών τεχνικών προσδιορισμού
  - Επηρεάζεται από *θόρυβο* και *έκτοπες τιμές*
    - Ύπαρξη σχετικών τεχνικών αντιμετώπισης

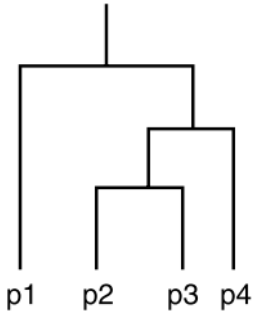
Αλγόριθμοι ιεραρχικής συσταδοποίησης

Προσεγγίσεις

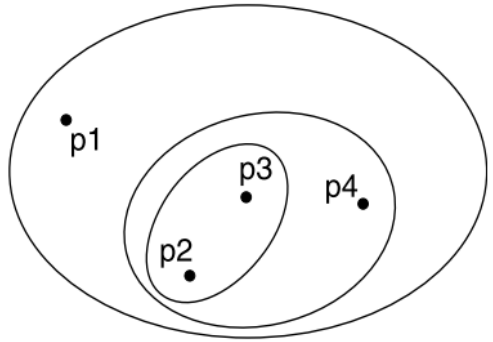
- **Συσσωρευτική** (*Agglomerative*)
  - Αρχικά το κάθε πρότυπο αποτελεί και *συστάδα*
  - Σε κάθε βήμα ενώνεται το *πλησιέστερο* ζεύγος συστάδων, μέχρι να μείνει μόνο μια
    - *Πως* αποφασίζεται ποιο ζεύγος είναι *πλησιέστερο*;
- **Διαχωριστική** (*Divisive*)
  - Αρχικά όλος ο χώρος των προτύπων αποτελεί μια *ενιαία συστάδα*
  - Σε κάθε βήμα, μια συστάδα *διαχωρίζεται*, μέχρις ότου να μείνουν *μεμονωμένες* συστάδες ή πρότυπα
    - *Πως* αποφασίζεται ποια συστάδα θα διαχωριστεί και *πως* πραγματοποιείται ο διαχωρισμός;
  - Αλγόριθμος DiAna (Divisive Analysis)

Απεικόνιση

- Μπορεί να απεικονιστεί υπό τη μορφή *δενδρογράμματος*
- Φαίνονται οι **σχέσεις** μεταξύ των συστάδων και η **σειρά** με την οποία αυτές *σχηματίστηκαν*
  - Ενώθηκαν ⇒ **Σωρευτική** ιεραρχική συσταδοποίηση
  - Χωρίστηκαν ⇒ **Διαχωριστική** ιεραρχική συσταδοποίηση
- Στην περίπτωση δισδιάστατων σημείων μπορεί να χρησιμοποιηθεί και **εμφωλευμένο** (*nested*) διάγραμμα



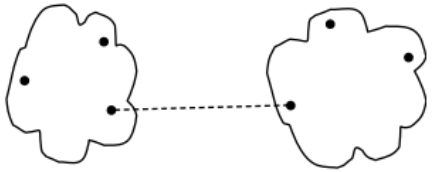
(a) Dendrogram.



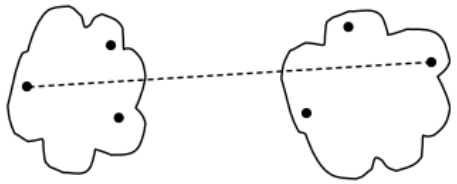
(b) Nested cluster diagram.

Σωρευτική Ιεραρχική Συσταδοποίηση: Εγγύτητα Συστάδων

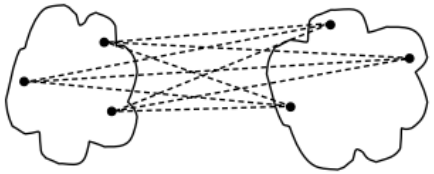
- **Γραφοθεωρητική** προσέγγιση
  - MIN (*single linkage*) ⇒ *εγγύτητα* των δύο πλησιέστερων προτύπων που ανήκουν σε διαφορετικές συστάδες
    - Αλγόριθμος AgNes (*Agglomerative Nesting*)
  - MAX (*complete linkage*) ⇒ *εγγύτητα* των δύο πιο απομακρυσμένων προτύπων που ανήκουν σε διαφορετικές συστάδες
  - Group Average ⇒ *μέσος όρος* της εγγύτητας μεταξύ όλων των προτύπων των δύο συστάδων
- **Σημειακή** προσέγγιση
  - Μέθοδος του Ward ⇒ *ελαχιστοποίηση* της διακύμανσης των προτύπων της συστάδας (*in-cluster variance*)



(a) MIN (single link.)



(b) MAX (complete link.)



(c) Group average.

Σωρευτική Ιεραρχική Συσταδοποίηση: Πολυπλοκότητα

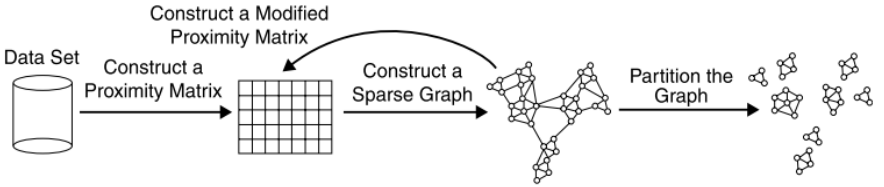
- M πρότυπα εισόδου
- **Χωρική**
  - Πίνακας εγγύτητας (*proximity matrix*): $\mathcal{O}(M^2)$
- **Χρονική**
  - Υπολογισμός ομοιότητας μεταξύ όλων των προτύπων: $\mathcal{O}(M^2)$
  - Γραμμική αναζήτηση στον πίνακα εγγύτητας: $\mathcal{O}(M^3)$
  - Χρήση ταξινομημένης λίστας: $\mathcal{O}(M^2 \log M)$
- *Ψευδο*-πολυωνυμική πολυπλοκότητα

Χαρακτηριστικά

- Μη-ύπαρξη μιας *ολικής* αντικειμενικής συνάρτησης
  - Αποφασίζεται **τοπικά** ποιες συστάδες θα ενωθούν/χωριστούν
  - Δεν "παγιδεύονται" σε *ολικά* ελάχιστα, δεν εξαρτώνται από την *αρχικοποίηση*
  - Δεν επιλύουν ένα *δύσκολο* πρόβλημα **γενικής** βελτιστοποίησης
    - Ωστόσο είναι ψευδο-πολυωνυμικοί ως προς χώρο/χρόνο
- Μπορούν να επεξεργάζονται συστάδες **διαφορετικού** μεγέθους
  - **Ζυγισμένη** (*weighted*) και **μη ζυγισμένη** προσέγγιση (*unweighted*)
- Ο σχηματισμός των συστάδων είναι **τελικός**
  - Δεν μπορεί να *αναστραφεί* σε μεταγενέστερο βήμα
  - Το *τοπικό* κριτήριο βελτιστοποίησης δεν μπορεί να γίνει *ολικό*

Γραφοθεωρητική Συσταδοποίηση

- **Βασική αρχή**
  - Κατασκευή *γράφου εγγύτητας* από τα πρότυπα εισόδου
  - Θεωρητικά M-κλίκα ⇒ στην πράξη ορισμένα μόνο πρότυπα μοιάζουν μεταξύ τους
- Τεχνικές **αραίωσης**
  - Αφαιρούμε ακμές με εγγύτητα κάτω από *κατώφλι*
  - Κρατάμε μόνο τους k πλησιέστερους γείτονες

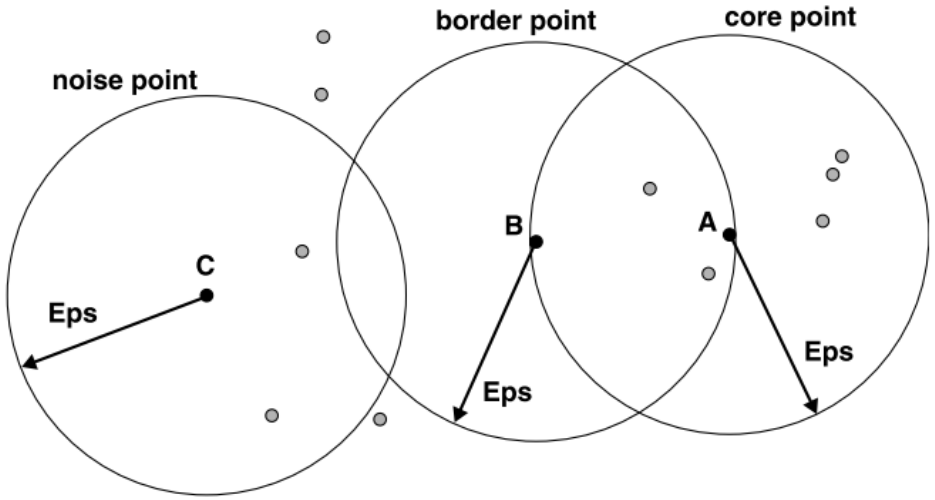


- Εφαρμογή **τεχνικών διαμέρισης** *αραιών* γράφων
  - **Ελάχιστο συνδετικό δέντρο** (*Minimum spanning tree* – MST)
  - Αλγόριθμος METIS ⇒ Τεχνικη OPOSSUM (Optimal Partitioning of Sparse Smiliarities Using METIS)

Αλγόριθμοι συσταδοποίησης βασισμένοι στην πυκνότητα

Πυκνότητα δεδομένων: "κεντρική" προσέγγιση

- Το *πλήθος* των προτύπων εντός καθορισμένης ακτίνας Eps από ένα καθορισμένο πρότυπο
- Τα υπόλοιπα πρότυπα ταξινομούνται ως
  - **Σημεία πυρήνα** (*core points*)
    - Εντός ακτίνας Eps και εντός γειτονίας μεγέθους $MinPts$
  - **Οριακά σημεία** (*border points*)
    - Πάνω στην ακτίνα
  - **Σημεία θορύβου** (*noise points*)
    - Εκτός ακτίνας



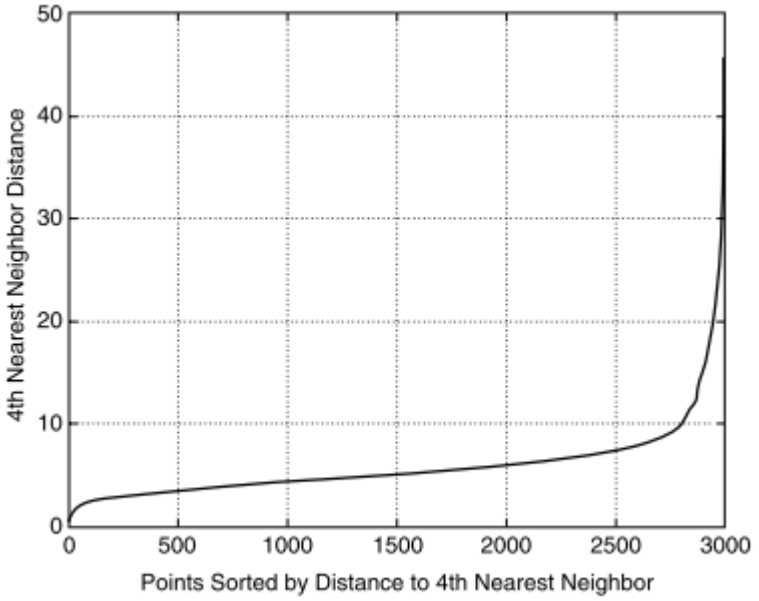
Αλγόριθμος DBSCAN

- Γενικά, οι αλγόριθμοι συσταδοποίησης βασισμένοι στην πυκνότητα εντοπίζουν περιοχές δεδομένων υψηλής πυκνότητας διαχωριζόμενες μεταξύ τους από περιοχές χαμηλής πυκνότητας

- Αλγόριθμος **DBSCAN**
  - **D**ensity-**B**ased **S**patial **C**lustering of **A**pplications with **N**oise
  - *Απλός* και *αποτελεσματικός* αλγόριθμος συσταδοποίησης βασισμένος στην *πυκνότητα*, αποτελεί τη βάση για άλλους, περισσότερο σύνθετους αλγορίθμους
  - Λειτουργία
    1. *Χαρακτηρισμός* του κάθε σημείου σε μια από τις 3 προαναφερόμενες κατηγορίες
    2. Αφαίρεση των σημείων *θορύβου*
    3. Σύνδεση με ακμή των *κέντρων* που απέχουν μεταξύ τους απόσταση *μικρότερη* από *Eps*
    4. Δημιουργία συστάδων από τις *συνεκτικές συνιστώσες*
    5. Ανάθεση των *οριακών* σημείων σε κάποια από τις *σχηματισμένες* συστάδες

DBSCAN: Προσδιορισμός παραμέτρων

- **Παράμετροι**
  - Μέτρο ακτίνας: Eps
  - Ελάχιστο Μέγεθος Γειτονιάς: $MinPts$
- Προσδιορισμός Eps και $MinPts$
  - $k - dist$ η απόσταση του k-οστού πλησιέστερου γείτονα
    - Τα σημεία πυρήνα θα έχουν μικρό $k - dist$, σε αντίθεση με τα σημεία θορύβου
  - Για δεδομένο k
    - Σχεδίασε την *απόσταση* όλων των σημείων από τον k-οστό πλησιέστερο γείτονα σε *αύξουσα* σειρά
    - Θέσε το Eps στο σημείο που παρατηρείται *απότομη μεταβολή* στη γραφική παράσταση και $MinPts = k$
  - Στον αρχικό αλγόριθμο DBSCAN, $k = 4$

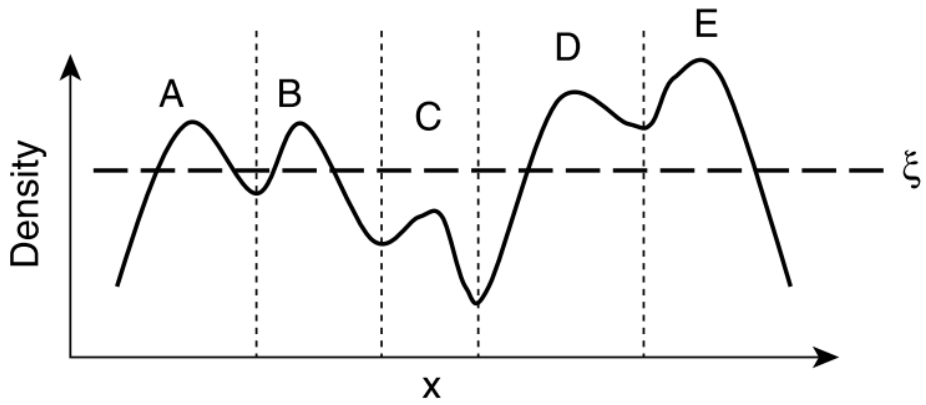


DBSCAN: Χαρακτηριστικά

- **Πολυπλοκότητα** (M πλήθος προτύπων)
  - Χωρική: $\mathcal{O}(M)$
  - Χρονική: $\mathcal{O}(M^2)$ η οποία μπορεί να μειωθεί μέχρι $\mathcal{O}(M \log M)$ σε χώρους χαμηλών διαστάσεων
  - Ψευδο-πολυωνυμικός
- Μήκος ακτίνας Eps
  - Εξαρτάται από το k, αλλά δεν αλλάζει πολύ όσο το k μεταβάλλεται
    - k μικρό ⇒ περιοχές με "θόρυβο" ενδέχεται να χαρακτηριστούν ως *συστάδες*
    - k μεγάλο ⇒ μικρές συστάδες ενδέχεται να χαρακτηριστούν ως "*θόρυβος*"
- Πλεονεκτήματα
  - Ανθεκτικός στο θόρυβο, μπορεί να εντοπίσει συστάδες οποιουδήποτε μεγέθους
- Μειονεκτήματα
  - Δυσκολεύεται όταν οι συστάδες έχουν μεταβλητή πυκνότητα
  - Ψευδο-πολυωνυμικότητα, ειδικά σε χώρους μεγάλων διαστάσεων

DENCLUE

- **DEN**sity **CLU**st**E**ring
- Η **συνολική πυκνότητα** των δεδομένων *μοντελοποιείται* ως **συνάρτηση**
  - Ως το *άθροισμα* συναρτήσεων επιρροής που σχετίζονται με κάθε πρότυπο
    - Συναρτήσεις πυρήνα (*kernel functions*)
  - Αναλογία με αλγορίθμους ταξινόμησης RBF
- Ύπαρξη τοπικών μεγίστων ⇒ **τοπικά σημεία έλξης** (*local density attractors*)
- Τα πρότυπα που έχουν ανατεθεί σε κορυφές κάτω από **κατώφλι** ξ
  - Αντιμετωπίζονται ως **θόρυβος**
  - Αν βρίσκονται σε άλλη "κοντινή" κορυφή, ανατίθενται σε εκείνη τη συστάδα

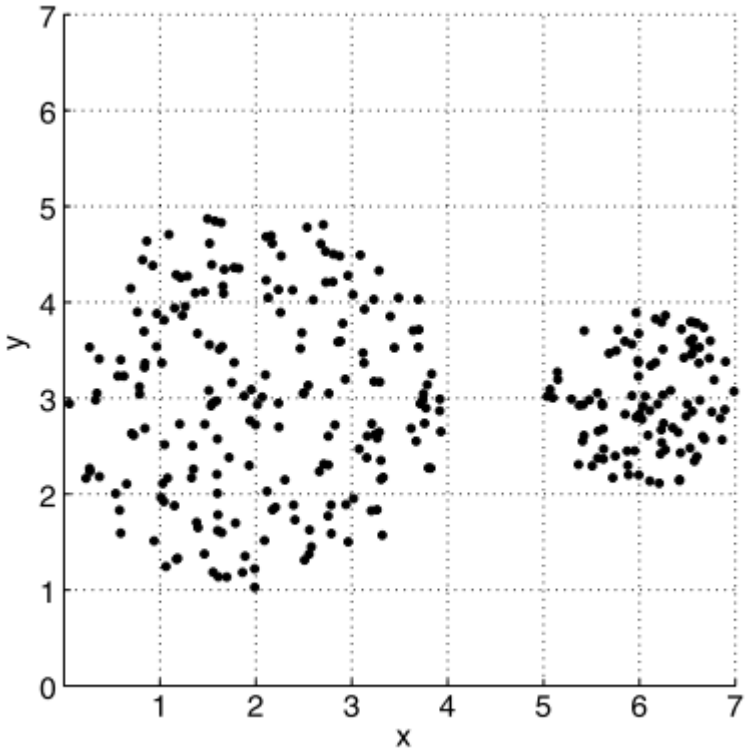


DENCLUE: Χαρακτηριστικά

- **Χρονική Πολυπλοκότητα**
  - Υπολογισμός πυκνότητας σε κάθε σημείο συναρτήσει των υπόλοιπων σημείων $\mathcal{O}(M^2)$
  - Μπορεί να μειωθεί με τη χρήση πλέγματος ⇒ Δεν λαμβάνονται υπόψη όλα τα σημεία αλλά μόνο τα γειτονικά στο πλέγμα
- **Πλεονεκτήματα**
  - Καλή θεωρητική βάση
  - Συνήθως *ακριβέστερος* υπολογισμός της πυκνότητας σε σχέση με DBSCAN
    - DBSCAN *ειδική περίπτωση* του DENCLUE
- **Μειονεκτήματα**
  - Υπολογιστικό κόστος (ακόμα και με χρήση πλέγματος)
  - Τα μειονεκτήματα των τεχνικών χρήσης πλέγματος

Αλγόριθμοι κατασκευής πλέγματος

- Υποκατηγορία αλγορίθμων βασισμένων στην πυκνότητα
  - "Κλασσικός" ορισμός πυκνότητας: *πλήθος* προτύπων στη *μονάδα* του όγκου
- **Πλέγμα** χωρίζει τον **χώρο** σε **κελιά**
  - Συστάδες σχηματίζονται από κελιά συγκεκριμένης πυκνότητας προτύπων
- **Κατασκευή πλέγματος**
  1. Τεμαχισμός του χώρου των δεδομένων σε ίσα μεγέθη
  2. Τεμαχισμός του χώρου των δεδομένων σε μεγέθη που περιέχουν ίσο αριθμό προτύπων ⇒ **διακριτοποίηση συχνότητας** (frequency discretization)
- **Κατασκευή συστάδων** από "γειτονικά" πυκνά κελιά
  - Κατώφλι πυκνότητας τ
  - Προσοχή στα κελιά στις οριακές περιοχές της συστάδας



Αλγόριθμοι κατασκευής πλέγματος: Χαρακτηριστικά

- **Πλεονεκτήματα**
  - Χρονική πολυπλοκότητα κατασκευής πλέγματος: $\mathcal{O}(M)$
    - Σε ειδικές περιπτώσεις φτάνει και σε $\mathcal{O}(M \log M)$
- **Μειονεκτήματα**
  - Εξάρτηση από **κατώφλι** τ
    - Υψηλό $\tau \Rightarrow$ απώλεια "μικρών" συστάδων
    - Χαμηλό $\tau \Rightarrow$ συνένωση διαφορετικών συστάδων
  - Εξάρτηση από τη **μορφή** των συστάδων
    - Στην προηγούμενη διαφάνεια το τετραγωνικό πλέγμα δεν προσεγγίζει καλά τις κυκλικές συστάδες
  - **Εκθετική αύξηση** του **πλήθους** των **κελιών** όσο αυξάνουν οι διαστάσεις

Κλιμάκωση αλγορίθμων συσταδοποίησης

Τεχνικές κλιμάκωσης

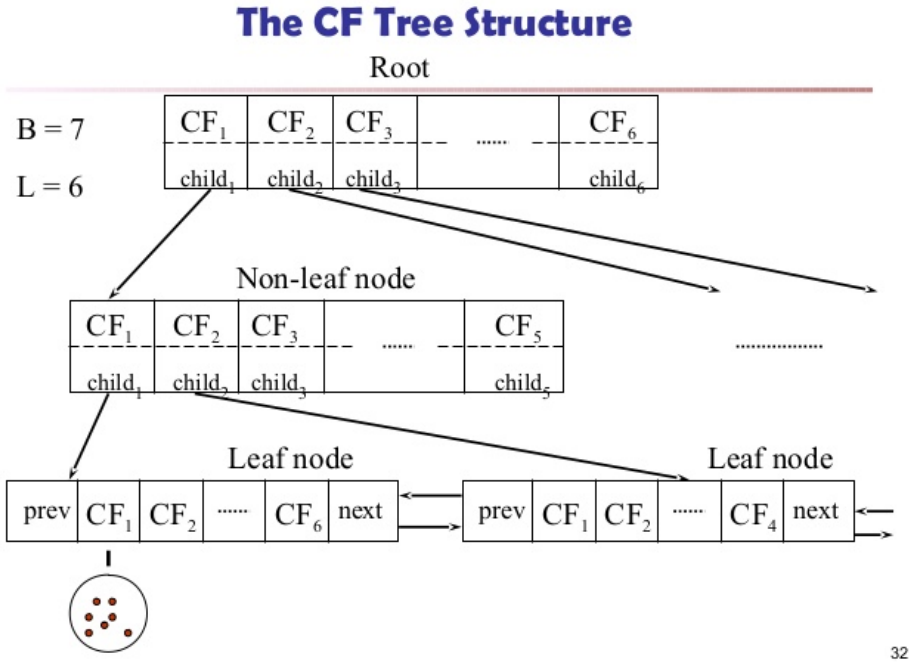
- Ψευδο-πολυωνυμική χρονική και χωρική πολυπλοκότητα
- Τεχνικές κλιμάκωσης
  - **Ιεραρχική διαμέριση** του χώρου
    - Γρηγορότερος υπολογισμός αποστάσεων και εύρεσης γειτόνων
    - Δέντρα k διαστάσεων, $R*$ δέντρα, κλπ
  - Χρήση **οριακών τεχνικών**
    - π.χ χρήση *τριγωνικής ανισότητας* για τον υπολογισμό αποστάσεων
  - **Δειγματοληψία**
    - Συσταδοποίηση δειγματοληπτούμενων προτύπων και κατόπιν ανάθεση των υπολοίπων στις συστάδες που έχουν δημιουργηθεί
    - Μπορεί να εξαφανιστούν "μικρές" συστάδες
  - **Διαμέριση** προτύπων
    - Χωρισμός προτύπων σε μη-επικαλυπτόμενα σύνολα, συσταδοποίηση αυτών και κατόπιν ένωση/συνδυασμός τους
  - **Σύνοψη** δεδομένων
    - Δημιουργία σύνοψης δεδομένων σε ένα πέρασμα και κατόπιν συσταδοποίηση της σύνοψης
  - **Παράλληλος/Κατανεμημένος** Υπολογισμός

Αλγόριθμος BIRCH

- **B**alanced **I**terative **R**educing and **C**lustering using **H**ierarchies
- Εφαρμογή σε *ευκλείδειους χώρους*
- Συσταδοποίηση στο πρώτο πέρασμα των δεδομένων
  - Βελτίωση της συσταδοποίησης στα επόμενα
- **Χαρακτηριστικό Συστάδας** (*Clustering Feature* – CF)
  - Αναπαράσταση συστάδας από τριπλέτα
    - **Πλήθος** των προτύπων της, το **άθροισμα** τους και το **άθροισμα** των **τετραγώνων** τους
    - Χρήση για τον υπολογισμό
      - *κέντρου* της συστάδας
      - *διαμέτρου* συστάδας (μέσω υπολογισμού της διασποράς)
      - *Απόστασης* μεταξύ κέντρων (πχ L_1, L_2)

Δέντρο CF

- Ζυγισμένο ως προς το ύψος, αποθηκεύει τα χαρακτηριστικά της συστάδας
  - Κάθε κόμβος που δεν είναι φύλλο αποθηκεύει τα αθροίσματα των χαρακτηριστικών συστάδας των παιδιών του
- **Δύο** παράμετροι
  - **Παράγοντας διακλάδωσης** (*branching factor*)
    - *Μέγιστος* αριθμός *παιδιών* ανά κόμβο
  - **Κατώφλι**
    - *Μέγιστη διάμετρος* υπό-συστάδων που είναι αποθηκευμένες στα φύλλα του δέντρου



Αλγόριθμος BIRCH (συνέχεια)

- Για κάθε πρότυπο εισόδου
  1. Βρες το *πλησιέστερο* φύλλο στο δέντρο
  2. Προσέθεσε *δείκτη* στο συγκεκριμένο φύλλο και *ενημέρωσε* το δένδρο CF
  3. Αν με την προσθήκη του προτύπου η διάμετρος υπερβεί το κατώφλι, τότε το φύλλο *διασπάται* (και ενδεχομένως και οι γονείς του)
- **Χρονική** πολυπλοκότητα: $\mathcal{O}(M)$
- **Χαρακτηριστικά**
  - *Ευαίσθητος* ως προς τη *σειρά εισαγωγής* των προτύπων
  - Με την *εκ των υστέρων διόρθωση* τους, οι συστάδες μπορεί να γίνουν **μη-"φυσικές"**
  - Δεδομένου ότι η παράμετρος που εξετάζεται είναι η *ακτίνα*, οι συστάδες τείνουν να γίνονται *σφαιρικές*

Αξιολόγηση ποιότητας συσταδοποίησης

Αξιολόγηση και Επαλήθευση

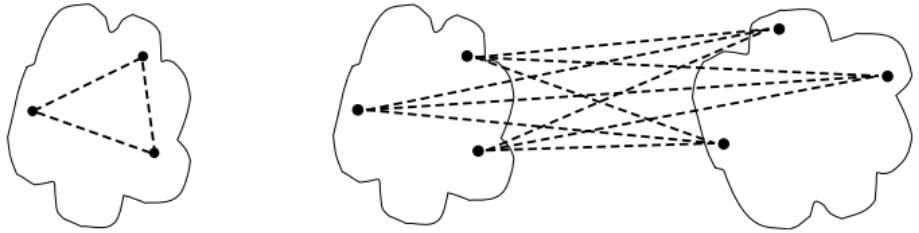
- **Μη επιγεγραμμένα δεδομένα**
  - Δεν μπορούν να χρησιμοποιηθούν οι μετρικές της επιβλεπόμενης μάθησης (Ακρίβεια, Ανάκληση, κλπ)
- **Αξιολόγηση συσταδοποίησης** (*cluster evaluation*) ⇒ **Επαλήθευση συσταδοποίησης** (*cluster validation*)
- Βασικό δομικό στοιχείο της **διερευνητικής ανάλυσης δεδομένων** (*exploratory data analysis*)
- Γιατί είναι **απαραίτητη**;
  - Οι αλγόριθμοι συσταδοποίησης τείνουν να ανακαλύπτουν δομή σε δεδομένα ακόμα και αν αυτή δεν υπάρχει (**τάση συσταδοποίησης** - *clustering tendency*)
  - Αν υπάρχει δομή, αυτή *ανταποκρίνεται* στις συστάδες που βρέθηκαν;

Ταξινόμηση μετρικών επαλήθευσης

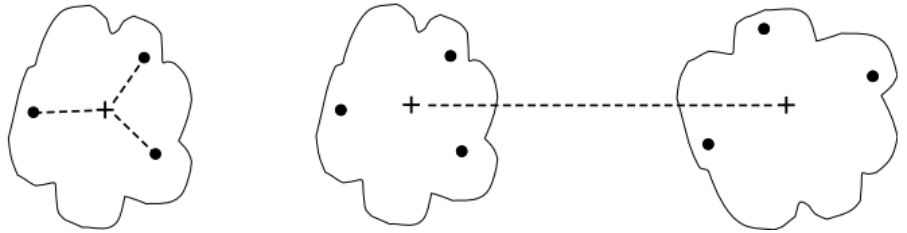
1. **Μη-επιβλεπόμενες** (*un-supervised*)
   - Δεν χρησιμοποιούν **εξωτερική πληροφορία** (*internal indices*)
   - Κατηγορίες
     1. Μέτρηση **συνοχής συστάδας** (*cluster cohesion*)
     2. Μέτρηση **διαχωρισμού συστάδων** (*cluster seperation*)
2. **Επιβλεπόμενες** (*supervised*)
   - Χρησιμοποιούν *εξωτερική πληροφορία*, όταν υπάρχει (*external indices*)
     - π.χ εντροπία
3. **Σχετικές** (*relative*)
   - Συγκρίνουν διαφορετικές συσταδοποιήσεις

Συνοχή και διαχωρισμός

- Συνάρτηση εγγύτητας $prox$
  - *Ομοιότητα, απόσταση* κλπ
- **Γραφοθεωρητική** προσέγγιση
  - Χρήση *γράφου εγγύτητας*
  - Συνοχή $coh(C_i) = \sum_{x \in C_i} \sum_{y \in C_i} prox(x, y)$
  - Διαχωρισμός $\text{sep}(C_i, C_j) = \sum_{x \in C_i} \sum_{y \in C_j} prox(x, y)$

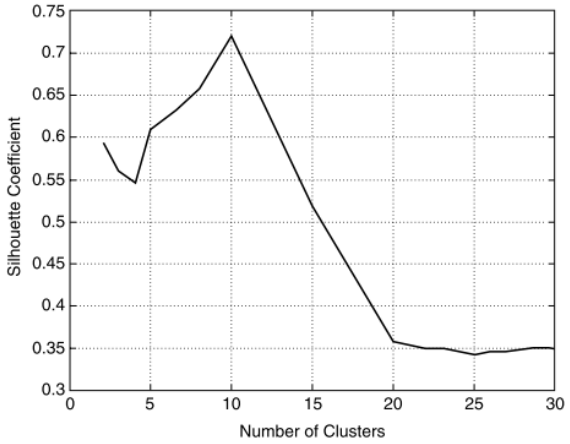


- **"Κεντρική" προσέγγιση**
  - Χρήση πίνακα εγγύτητας
  - Υπολογίζεται ως προς το κέντρο της συστάδας c_i
  - Συνοχή $coh(C_i) = \sum_{x \in C_i} prox(x, c_i)$
  - Διαχωρισμός $sep(C_i, C_j) = prox(c_i, c_j)$



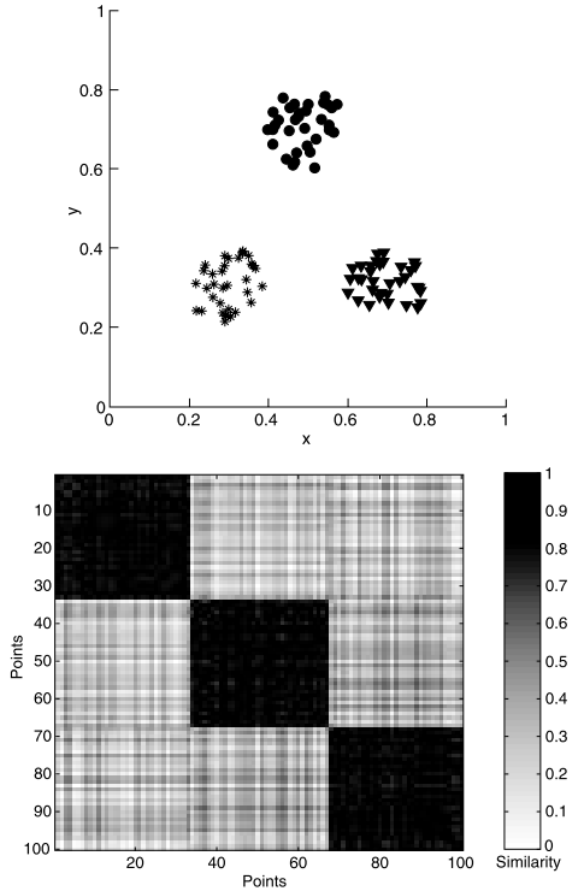
Silhouette Coefficient

- **Συνδυάζει** τις έννοιες της *συνοχής* και του *διαχωρισμού*
- Χρησιμοποιείται και για την **εύρεση** του *βέλτιστου αριθμού συστάδων*
- Για το i-οστό πρότυπο
  - a_i η *μέση απόσταση* του από τα υπόλοιπα πρότυπα της συστάδας του
  - b_i η *απόσταση* του από το *κοντινότερο* πρότυπο εκτός της συστάδας του
  - $s_i = \frac{b_i - a_i}{\max(a_i, b_i)}$
- Λαμβάνει τιμές στο $[-1, +1]$
  - Αρνητικές τιμές μη επιθυμητές
    - Σημαίνει ότι η μέση απόσταση του από τα πρότυπα της συστάδας του είναι μεγαλύτερη από την απόσταση από το κοντινότερο πρότυπο εκτός συστάδας
  - Γενικά θέλουμε τιμές όσο πιο κοντά στο +1



Χρήση του πίνακα εγγύτητας

- Ταξινόμηση του πίνακα σχεδίασης με βάση τις σχηματισμένες συστάδες και κατόπιν σχεδίασή του
- Αν η συσταδοποίηση είναι επιτυχής, θα έχει **διαγώνια δομή**
- Αν όχι, θα εμφανιστούν *μοτίβα* επάνω στον πίνακα που θα υποδεικνύουν και *άλλες σχέσεις* μεταξύ των προτύπων και των συστάδων



Τάση συσταδοποίησης

- Χρήση στατιστικών ελέγχων για **χωρική τυχαιότητα** (*spatial randomness*)
- **Στατιστικό Hopkins**
  - Δημιουργία p *τυχαίων σημείων* στο χώρο δεδομένων και δειγματοληψία *άλλων τόσων προτύπων*
  - Για τα δύο σύνολα σημείων υπολογίζεται η *απόσταση* από το *πλησιέστερο πρότυπο* (u_i και w_i αντίστοιχα)

$$H = \frac{\sum_p w_i}{\sum_p u_i + \sum_p w_i}$$

  - Αν το H είναι **κοντά** στο $\mathbf{0.5}$, τότε οι *αποστάσεις* των σημείων των δύο συνόλων είναι **περίπου ίδιες** ⇒ δεν υπάρχει δομή συστάδων
  - Αν είναι **κοντά** στο $\mathbf{0}$, υπάρχει *έντονη δομή* συστάδων που απέχουν πολύ μεταξύ τους
  - Αν είναι **κοντά** στο $\mathbf{1}$, οι συστάδες είναι πιο *ομοιόμορφα κατανεμημένες* στο χώρο

Επιβλεπόμενες μετρικές επαλήθευσης

- Ετικέτες για τα πρότυπα
  - Εξωτερική πληροφορία εν γένει
- **Ταξινόμησης**
  - *Εντροπία*, Καθαρότητα: πόσο η κάθε συστάδα περιέχει δεδομένα μιας κλάσης
  - *Ακρίβεια*: Ποσοστό προτύπων συστάδας που ανήκουν σε συγκεκριμένη κλάση
  - *Ανάκληση*: Ποσοστό που μια συστάδα περιέχει όλα τα πρότυπα μιας συγκεκριμένης κλάσης
  - *Μετρική* F_1
- **Ομοιότητας**
  - Στατιστικό Rand
  - Συντελεστής ομοιότητας Jaccard

Κριτήρια επιλογής βέλτιστου αλγορίθμου

- Εξαρτάται από τη *φύση* του *προβλήματος*
- **Κριτήρια**
  - **Τύπος** συσταδοποίησης
    - π.χ είναι χρήσιμη η ιεραρχία ή όχι; πρότυπα ανήκουν σε πολλές συστάδες ή όχι;
  - Χαρακτηριστικά **συστάδων**
    - π.χ. μας ενδιαφέρει η χωρική διάταξη των συστάδων ή όχι;
  - Χαρακτηριστικά **δεδομένων**
    - Πως υπολογίζεται η εγγύτητα μεταξύ των προτύπων;
  - **Θόρυβος** και **έκτοπες τιμές**
  - **Πλήθος** προτύπων
  - …

Βιβλιογραφία

1. P. Tan, M. Steinbach, V. Kumar – *Introduction to Data Mining* (New International Edition)
   - Κεφάλαια 8 και 9
2. J. Han, M. Kamber, J. Pei – *Datamining, Concepts and Techniques* (3rd edition)
   - Κεφάλαια 10 και 11