Εισαγωγή Ι

Ανάλυση Συστάδων - Εισαγωγικές Έννοιες

- Στην Διαχωριστική Ανάλυση (γνωστή επίσης και ως supervised pattern recognition), είναι εκ των προτέρων γνωστό ότι έχουμε k ομάδες, στις οποίες θέλουμε να κατατάξουμε κάθε νέα παρατήρηση.
- Στην Ανάλυση Συστάδων (Cluster Analysis), το σύνολο των παρατηρήσεων αποτελεί μια (μεγάλη) ομάδα και το στόχος είναι να σχηματίσουμε τις κατάλληλες ομάδες (unsupervised pattern recognition)
- Οι ομάδες αυτές θα πρέπει να περιέχουν 'όμοια' αντικείμενα (ή να έχουν παρόμοιες ιδιότητες).

Εισαγωγή II

Ανάλυση Συστάδων - Εισαγωγικές Έννοιες

- Όταν αναφερόμαστε στη δημιουργία συστάδων, εννοούμε τη δημιουργία k ομάδων με βάση $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ τυχαία διανύσματα (το διαθέσιμο dataset), με την i-οστη ομάδα να περιλαμβάνει n_i από τα n διανύσματα, έτσι ώστε

$$n_1 + n_2 + \ldots + n_k = n.$$

με τα στοιχεία της ομάδας να είναι 'όμοια' (within homogeneity) αλλά οι ομάδες μεταξύ τους 'ανόμοιες'.

- Επιθυμητή είναι η μικρή μεταβλητότητα εντός των ομάδων (within groups variability) και η μεγάλη μεταβλητότητα μεταξύ των ομάδων (between groups variability).

Εισαγωγή III

Ανάλυση Συστάδων - Εισαγωγικές Έννοιες

- **Προβλήματα**: (α) Πώς θα μετρήσουμε την 'ομοιότητα'; (β) πώς θα μετρήσουμε την 'ανομοιότητα' και την 'απόσταση' μεταξύ των ομάδων; (γ) ποια είναι η ελάχιστη απόσταση για το διαχωρισμό δύο ομάδων; (δ) υπάρχει κάποιο κριτήριο 'συσταδοποίησης' (clustering), αφού (όπως θα δούμε και στη συνέχεια) διαφορετικά κριτήρια δίνουν διαφορετικές συστάδες με τα ίδια χαρακτηριστικά;
- Ένα ακόμη πρόβλημα είναι ο αριθμός των συστάδων που πρέπει να προσδιοριστούν. Μπορεί είτε να είναι προκαθορισμένος είτε να καθοριστεί μέσω κάποιου κριτηρίου (π.χ. minimum between-clusters distance)
- Τέλος, αν τα δεδομένα (αντικείμενα) θεωρηθούν ως παρατηρήσεις από μια πεπερασμένη μίξη κατανομών, π.χ. $f(\mathbf{x}) = \sum_{h=1}^{\kappa} q_h f_h(\mathbf{x})$, τότε εκτός από τους συντελεστές $q_1, q_2, \ldots, q_\kappa$ πρέπει να προσδιοριστεί και το κ.

Παραδείγματα Ι

- Έστω η συνήθης τράπουλα (με τα 52 φύλλα). Τότε μπορούμε να ορίσουμε:
  - 13 συστάδες με βάση τα διαφορετικά φύλλα A, 1, 2, ..., J, Q, K
  - 4 συστάδες με βάση το σχέδιο (κούπα,σπαθί,μπαστούνι,καρό)
  - 2 συστάδες με βάση το χρώμα (μαύρο και κόκκινο) κλπ
- Έστω ότι έχουμε 11 γλώσσες στις οποίες χρησιμοποιείται το Λατινικό αλφάβητο. Η ομαδοποίηση των γλωσσών μπορεί να γίνει με βάση το αν οι λέξεις που αντιστοιχούν στους αριθμούς 1, 2, ..., 10 ξεκινούν με το ίδιο γράμμα (Παράδειγμα 12.3, Johnson and Wicher 2007).
- Σε αυτή την περίπτωση, η ομαδοποίηση δίνει **Ομάδα 1**: Αγγλικά, Νορβηγικά, Δανέζικα, Ολλανδικά, Γερμανικά. **Ομάδα 2**: Γαλλικά, Ιταλικά, Ισπανικά, Πολωνικά. **Ομάδα 3**: Ουγγρικά. **Ομάδα 4**: Φινλανδέζικα (Εφαρμογή στο Εργαστήριο).

Παραδείγματα II

- Ως μέτρο ομοιότητας έχει χρησιμοποιηθεί η 'συμφωνία' (concordance) με βάση το παραπάνω κριτήριο. Ανάλογα όμως κριτήρια μπορούν να χρησιμοποιηθούν και για άλλες λέξεις στα διαφορετικά αλφάβητα, οδηγώντας σε διαφορετική συσταδοποίηση.
- Τα παραπάνω παραδείγματα δείχνουν το πόσο σημαντική είναι η επιλογή του κριτηρίου συσταδοποίησης.

Εφαρμογές Ανάλυσης Συστάδων Ι

- **Συμπίεση Δεδομένων (Data Compression)**: Ομαδοποίηση των δεδομένων σε k συστάδες. Από κάθε συστάδα επιλέγω έναν αντιπρόσωπο.
- **Δημιουργία Υποθέσεων (Hypothesis Generation)**: Χρήση συστάδων για δημιουργία ερευνητικών υποθέσεων, τις οποίες πρόκειται να ελέγξουμε και να επιβεβαιώσουμε χρησιμοποιώντας διάφορα *training sets*.
- **Έλεγχος Υποθέσεων**
- **Συσταδοποίηση - Διαχωρισιμότητα (Classification - Discrimination)**
- **Πρόβλεψη**: Μπορεί να βασιστεί σε συστάδες, λαμβάνοντας υπόψη τα κύρια χαρακτηριστικά των μελών της. Για δεδομένες τιμές ενός ατόμου (π.χ. διάνυσμα τιμών $\mathbf{x}$), μπορούμε να το ταξινομήσουμε σε μια από τις διαθέσιμες συστάδες.

Εφαρμογές Ανάλυσης Συστάδων II

- **Παράδειγμα**: Ασθενείς που έχουν προσβληθεί από την ίδια ασθένεια, μπορούν να δώσουν συστάδες με βάση την αντίδρασή τους σε συγκεκριμένη θεραπεία. Στη συνέχεια, για κάθε νέο ασθενή, αναγνωρίζοντας σε ποια συστάδα ανήκει (διάγνωση) μπορεί να προταθεί η καταλληλότερη θεραπεία.
- Ανάλογες εφαρμογές μπορεί κανείς να βρει στις περιοχές της Εκπαίδευσης, του Management, της Ψυχολογίας, των Τηλεπικοινωνιών και της Αρχαιολογίας.

Αλγόριθμοι Συσταδοποίησης Ι

- **Στόχος**: Η διαμέριση του διαθέσιμου dataset $\mathbf{X}$, το οποίο αποτελείται από n διανύσματα $\mathbf{x}_1, \ldots, \mathbf{x}_n$ διαστάσεων $p \times 1$ το καθένα, σε k μη-κενές συστάδες μεγέθους $n_1 > 0, \ldots, n_k > 0$.
- Οι δυνατοί τρόποι για να γίνει αυτό αντιστοιχούν στο πλήθος των ακέραιων λύσεων της εξίσωσης $n_1 + \ldots + n_k = n$ και ισούνται με

$$S(n,k) = \frac{1}{k!} \sum_{i=0}^{n} (-1)^{k-i} \binom{n}{k} i^n,$$

  ο οποίος είναι γνωστός και ως ο *Αριθμός Stirling* 2ου είδους.
- Ενδεικτικά, $S(16,2) = 32767$ (οι τρόποι για να διαμερίσουμε τα 16 χαρτιά σε 2 ομάδες διαφορετικών μεγεθών) και $S(16,3) = 7141686$ (οι τρόποι για να διαμερίσουμε τα 16 χαρτιά σε 3 ομάδες διαφορετικών μεγεθών).

Αλγόριθμοι Συσταδοποίησης ΙΙ

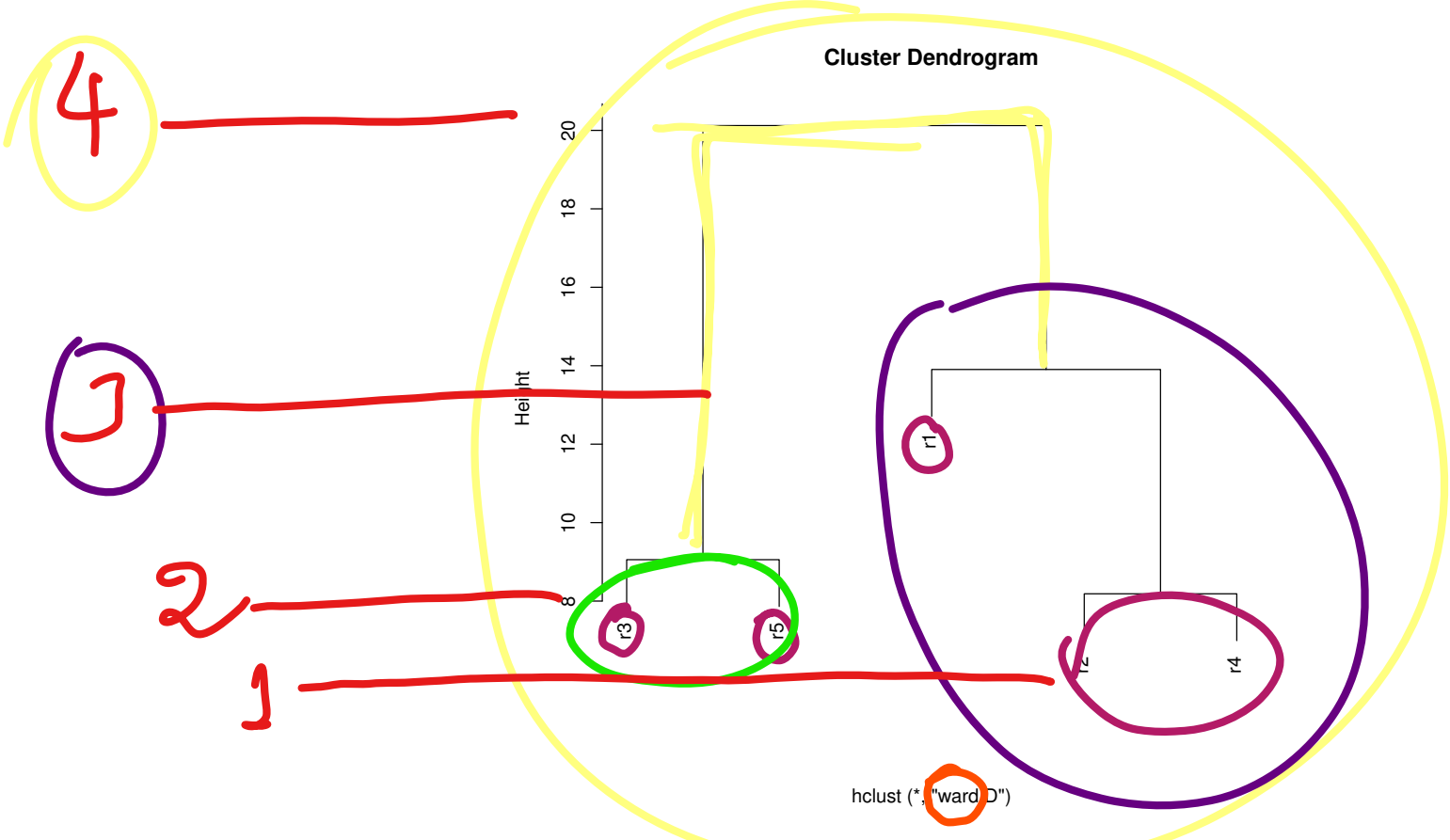
- Ακόμη και οι σύγχρονοι Η/Υ συχνά συναντούν προβλήματα στον προσδιορισμό των δυνατών συστάδων (καθώς το n και το k αυξάνουν).
- Γιάυτό το λόγο, οι αλγόριθμοί που έχουν αναπτυχθεί συνήθως δίνουν 'καλές' αλλά όχι τις 'βέλτιστες' συστάδες.
- Υπάρχουν 2 βασικές κατηγορίες τέτοιων αλγορίθμων, οι Ιεραρχικοί και οι Μη-Ιεραρχικοί. Στη συνέχεια θα δούμε βασικά χαρακτηριστικά αυτών.

Ιεραρχικοί Αλγόριθμοι Συσταδοποίησης Ι

- Αποτελούν την κυριότερη κατηγορία και λειτουργούν είτε (α) με διαδοχικές συνενώσεις ομάδων (Agglomerative algorithms) ή (β) με διαδοχικές διαμερίσεις (Divisive algorithms).
- Οι Agglomerative αλγόριθμοι ξεκινούν με τις n διαφορετικές παρατηρήσεις ως μια ομάδα η καθεμία. Οι περισσότερο όμοιες παρατηρήσεις τοποθετούνται σε μια ομάδα και άρα οι ομάδες δημιουργούνται με βάση την ομοιότητα των παρατηρήσεων. Τελικά, όλες οι παρατηρήσεις θα αποτελέσουν (στο τελευταίο βήμα) μια μεγάλη ομάδα.
- Οι Divisive αλγόριθμοι ξεκινούν αντίστροφα, με τις n διαφορετικές παρατηρήσεις να αποτελούν μια μεγάλη ομάδα. Στο επόμενο βήμα οι n παρατηρήσεις χωρίζονται σε δύο ομάδες, και οι παρατηρήσεις που είναι 'πιο μακριά' από τις υπόλοιπες, θα αποτελούν μια ξεχωριστή ομάδα. Ο αλγόριθμος συνεχίζει μέχρι τελικά, όλες οι παρατηρήσεις να αποτελέσουν (στο τελευταίο βήμα) n ξεχωριστές ομάδες.

Ιεραρχικοί Αλγόριθμοι Συσταδοποίησης II

- Οι διαδοχικές διαιρέσεις/συνενώσεις σε κάθε βήμα των παραπάνω αλγορίθμων, αναπαρίστανται με ένα **dendrogram** (δενδρόγραμμα).



Σχήμα: Παράδειγμα Δενδρογράμματος

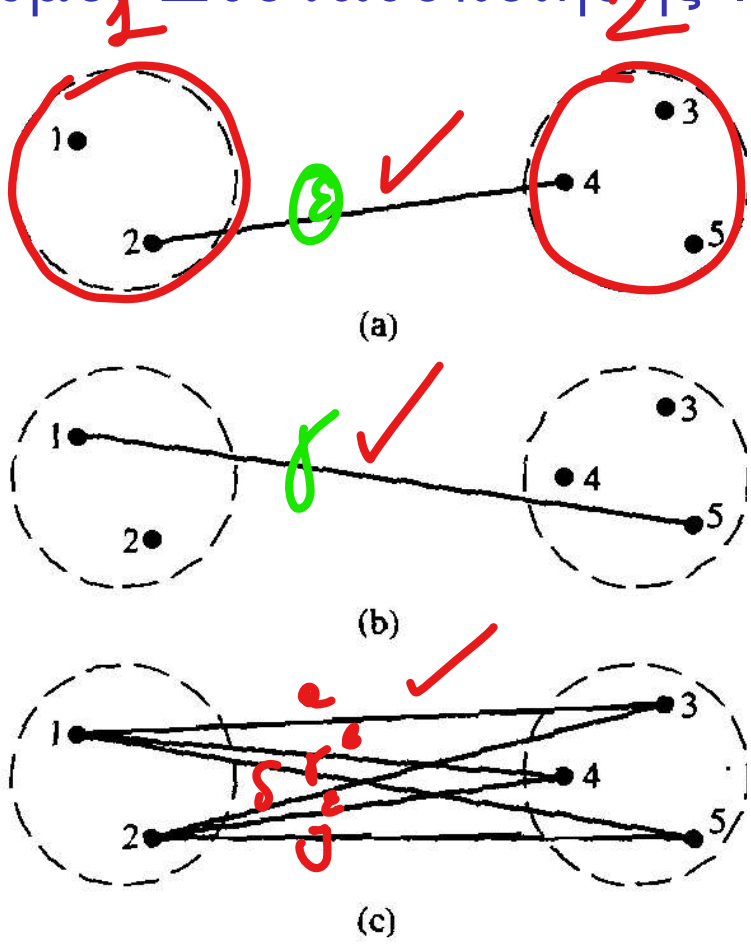
Ιεραρχικοί Αλγόριθμοι Συσταδοποίησης III

- Οι μέθοδοι σύνδεσης (Linkage Methods) είναι κατάλληλοι για τη δημιουργία συστάδων, ως μέρος των Ιεραρχικών Αλγόριθμων Συσταδοποίησης.
- Η *απλή σύνδεση* (**single linkage**) αναφέρεται στη δημιουργία ομάδων σύμφωνα με μικρότερη απόσταση μεταξύ των μελών.
- Η *πλήρης σύνδεση* (**complete linkage**) αναφέρεται στη δημιουργία ομάδων σύμφωνα με τη μεγαλύτερη απόσταση μεταξύ των μελών.
- Τέλος, η *μέση σύνδεση* (**average linkage**) αναφέρεται στη δημιουργία ομάδων σύμφωνα με τη μέση απόσταση μεταξύ των μελών. Άλλες μέθοδοι είναι οι Centroid, Ward's method, Median (Παράδειγμα στο εργαστήριο).
- Χαρακτηριστικό είναι το παρακάτω σχήμα.

Ιεραρχικοί Αλγόριθμοι Συσταδοποίησης IV

Single Linkage

Complete Linkage



min απόσταση

Max απόσταση

average απόσταση
μέση τιμή των d_ij

D = πίνακας αποστάσεων

	1	2
1	0	ε
2	ε	0

Σχήμα: Δημιουργία Συστάδων με Ιεραρχικές Μεθόδους, (a) Single Linkage (b) Complete Linkage, (c) Average Linkage (Πηγή: Johnson and Wichern 2007)

Ιεραρχικοί Αλγόριθμοι Συσταδοποίησης V

- Για την εφαρμογή των Ιεραρχικών Αλγορίθμων Συσταδοποίησης, επιλέγουμε κάποιο μέτρο απόστασης ή ομοιότητας μεταξύ των αντικειμένων (παρατηρήσεων).
- Προφανώς, η ομοιότητα είναι μια φθίνουσα συνάρτηση της απόστασης $d(\mathbf{x}, \mathbf{y})$ μεταξύ των $\mathbf{x}$ και $\mathbf{y}$ ενώ αντίθετα, η ανομοιότητα είναι μια αύξουσα συνάρτηση.
- Γενικά, στην ανάλυση συστάδων, χρησιμοποιείται κάποιο μέτρο (ή δείκτης) ομοιότητας/ανομοιότητας/απόστασης, με την Ευκλείδεια Απόσταση να είναι μια συνήθης επιλογή.
- Θα πρέπει να σημειωθεί ότι η επιλογή της απόστασης $d(\mathbf{x}, \mathbf{y})$ μεταξύ των $\mathbf{x}$ και $\mathbf{y}$ εξαρτάται από τη φύση των διαθέσιμων παρατηρήσεων και από τους σκοπούς για τους οποίους εφαρμόζεται η ανάλυση σε συστάδες, η οποία (κάποιες φορές) μπορεί να είναι και αυθαίρετη.

Ιεραρχικοί Αλγόριθμοι Συσταδοποίησης

Η Περίπτωση των Κατηγορικών Μεταβλητών

- Έστω $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})$. $i = 1, 2, \ldots, n$, όπου τα X_{ij}, $j = 1, 2, \ldots, p$ είναι δίτιμες 0-1 τ.μ. Δηλαδή, $X_{ij} = 1$ σημαίνει παρουσία του j-οστού χαρακτηριστικού (συμπτώματος) στο i-οστό άτομο, διαφορετικά έχουμε απουσία του χαρακτηριστικού (συμπτώματος).
- **Παράδειγμα**: $X_{i1} = 1$ το i-οστο άτομο έχει ξανθό μαλλί, $X_{i2} = 1$ το i-οστό άτομο έχει γαλάζια μάτια κ.ο.κ.
- Τότε, μπορούμε να ορίσουμε δείκτες ομοιότητας με βάση τις συμφωνίες/ασυμφωνίες, όπου $a =$ πλήθος των $(X_{i1} = 1, X_{i2} = 1)$, $b =$ πλήθος των $(X_{i1} = 1, X_{i2} = 0)$, $c =$ πλήθος των $(X_{i1} = 0, X_{i2} = 1)$, $d =$ πλήθος των $(X_{i1} = 0, X_{i2} = 0)$.
- Ένας δείκτης σε αυτή την περίπτωση είναι το

$$(a + d) / (a + b + c + d)$$

.

Αποστάσεις Ι

Η απόσταση d μεταξύ $\mathbf{x} = (x_1, \ldots, x_p)$ και $\mathbf{y} = (y_1, \ldots, y_p)$ ικανοποιεί τις παρακάτω ιδιότητες:

- Μη-αρνητική: $d(\mathbf{x}, \mathbf{y}) \geq 0$ με $d(\mathbf{x}, \mathbf{y}) = 0$ αν-ν $\mathbf{x} = \mathbf{y}$.
- Συμμετρική: $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$
- Τριγωνική Ανισότητα: $d(\mathbf{x}, \mathbf{y}) \leq d(\mathbf{y}, \mathbf{z}) + d(\mathbf{x}, \mathbf{z})$

Κάποιες Αποστάσεις:

- Ευκλείδια: $d_E = |\mathbf{x} - \mathbf{y}| = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2} = \sqrt{(\mathbf{x} - \mathbf{y})'(\mathbf{x} - \mathbf{y})}$
- City-Block ή Manhattan: $d_C = \sum_{i=1}^{p} |x_i - y_i|$
- Minkowski τάξης s: $d_M = \left(\sum_{i=1}^{p} |x_i - y_i|^s\right)^{1/s}$
- Chebyshev: $d_T = \max\{|x_i - y_i|, i = 1, 2, \ldots, p\}$
- Quadratic: Αν $\mathbf{Q}$ είναι ένας θετικά ορισμένος πίνακας, ορίζεται η απόσταση $d_Q = (\mathbf{x} - \mathbf{y})'\mathbf{Q}(\mathbf{x} - \mathbf{y})$.

Αποστάσεις II

- Mahalanobis: Αν $\mathbf{x}, \mathbf{y}$ είναι τυχαία διανύσματα με τον ίδιο πίνακα διακυμάνσεων-συνδιακυμάνσεων $\mathbf{\Sigma}$, τότε η απόσταση Mahalanobis ορίζεται ως $\Delta^2 = (\mathbf{x} - \mathbf{y})' \mathbf{\Sigma}^{-1} (\mathbf{x} - \mathbf{y})$. Σημειώνεται ότι η απόσταση αυτή είναι γνωστή και ως 'στατιστική απόσταση' μεταξύ των $\mathbf{x}, \mathbf{y}$.

Ιεραρχικές Μέθοδοι Ομαδοποίησης Ι

Συμπληρωματικό Υλικό

- Στην ομαδοποίηση αντικειμένων (ατόμων, σημείων), οι λεγόμενες ιεραρχικές μέθοδοι προχωρούν είτε με μια ακολουθία διαδοχικών συγχωνεύσεων είτε με ακολουθία διαδοχικών διαιρέσεων (διαμερίσεων).
- Οι συγχωνευτικές (ή ανιούσες, agglomerative) ιεραρχικές μέθοδοι ξεκινούν με τα n άτομα ως 'ομάδες' (clusters).
- Πρώτα τοποθετούνται σε ομάδες τα πιο όμοια μεταξύ τους άτομα και οι αρχικές αυτές ομάδες συγχνεύονται σύμφωνα με κάποιο βαθμό ομοιότητας (similarity).
- Τελικά με την ομοιότητα να μικραίνει, όλες οι υποομάδες (sub-clusters ή sub-groups) ενώνονται σε μια ομάδα, με στοιχεία όλα τα n άτομα.
- Οι διαιρετικές (ή διαμεριστικές, divisive) ιεραρχικές μέθοδοι, προχωρούν ανάποδα.

Ιεραρχικές Μέθοδοι Ομαδοποίησης ΙΙ

Συμπληρωματικό Υλικό

- Η αρχική (ενιαία) cluster των n σημείων διαιρείται σε 2 υπο-ομάδες ώστε τα άτομα της μια υποομάδας να 'απέχουν' των ατόμων της άλλης.
- Οι υποομάδες αυτές υποδιαιρούνται περαιτέρω σε ανόμοιες υποομάδες έως ότου φτάσουμε σε n υποομάδες, δηλάδή κάθε άτομο να αποτελεί και ένα cluster.
- Όπως είδαμε και προηγουμένως, η διαδικασία αυτή μπορεί να παρασταθεί γραφικά με ένα δενδρόγραμμα, το οποίο δείχνει τις διαδοχιές συγχωνεύσεις ή διαιρέσεις.

Συγχωνευτικές Ιεραρχικές Μέθοδοι Ομαδοποίησης Ι

Linkage Methods

- Οι μέθοδοι Linkage χρησιμεύουν για την ομαδοποίηση είτε ατόμων (είτε μεταβλητών, αν και συστήνεται η αποφυγή της), κάτι το οποίο δεν ισχύει για τις συγχωνευτικές μεθόδους ιεραρχικής ταξινόμησης.
- Έχουμε ήδη ορίσει (α) την απλή σύνδεση ή σύνδεση του εγγύτερου γείτονα (Simple Linkage, Nearest Neighbor), (β) την πλήρη σύνδεση ή σύνδεση απώτερου γείτονα (Complete Linkage, Furthest Neighbor), (γ) τη σύνδεση της μέσης απόστασης (Average Linkage).
- Τα βήματα του αλγορίθμου περιγράφονται παρακάτω:
  - **Β1**: Ξεκίνα με n ομάδες, όσες και τα αντικείκενα $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ και έστω $\mathbf{D} = (d_{ij})$ ο πίνακας των αποστάσεων (ή ομοιοτήτων) των n αντικειμένων.
  - **Σχόλιο**:Το d_{ij} εκφράζει την απόσταση του i από το j, $i \neq j$. Προφανώς, ο $\mathbf{D}$ είναι συμμετρικός με $d_{ij} = d_{ji}$ και $d_{ii} = 0$.
  - **Β2**:Βρες στον πίνακα $\mathbf{D}$ τις όμοιες ομάδες, έστω αυτές U, V (εκεί δηλαδή που είναι η μικρότερη τιμή d_{ij}).

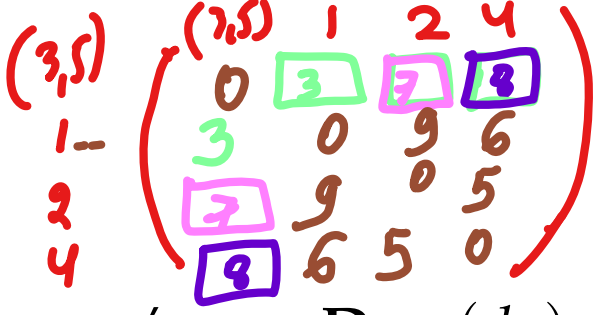
Συγχωνευτικές Ιεραρχικές Μέθοδοι Ομαδοποίησης II

Linkage Methods

- **B3:**Συγχώνευσε τις U και V και έστω (UV) η νέα ομάδα. 'Επικαιροποίησε' (update) τον $\mathbf{D}$ ως εξής: (1) απαλείφοντας τις γραμμές και στήλες που αντιστοιχούν στις ομάδες U και V και (2) προσθέτοντας μια γραμμή και μια στήλη, στις οποίες θα βρίσκεται πλέον η απόσταση της νέας ομάδας (UV) από τις υπόλοιπες ομάδες.
- Επανάλαβε τα B2 και B3 για $n-1$ φορές ώστε τα n άτομα να βρεθούν σε μια ομάδα. Σημειώνουμε ποιές ομάδες συγχωνεύονται και σε ποιες αποστάσεις (ή ομοιότητες) έγιναν οι συγχωνεύσεις.

Παράδειγμα Ι

Single Linkage - Nearest Neighbor



- Έστω $n = 5$ αντικείμενα με πίνακα αποστάσεων $\mathbf{D} = (d_{ij})$

$$\mathbf{D} = (d_{ij}) = \begin{pmatrix} 0 & 9 & 3 & 6 & 11 \\ 9 & 0 & 7 & 5 & 10 \\ 3 & 7 & 0 & 9 & 2 \\ 6 & 5 & 9 & 0 & 8 \\ 11 & 10 & 2 & 8 & 0 \end{pmatrix}$$

- Θεωρούμε τα άτομα 1, 2, 3, 4, 5 ως 5 διαφορετικές ομάδες (συστάδες) και η ομαδοποίηση αρχίζει με τη τη συγχώνευση των 2 πλησιέστερων ατόμων. Προφανώς

$$\min\{d_{ij}, 1 \le i, j \le 5\} = d_{53} = 2$$

και έτσι οι (5) και (3) ενώνονται και σχηματίζουν τη συστάδα $(3,5)$.

Παράδειγμα II

Single Linkage - Nearest Neighbor

- Στο επόμενο βήμα, ο νέος πίνακας $\mathbf{D}$ θα περιέχει τις (εγγύτερες) αποστάσεις της $(3,5)$ από τα αντικείμενα 1, 2 και 4:

$$d_{(35),1} = \min\{d_{31}, d_{51}\} = \min\{3, 11\} = 3,$$

$$d_{(35),2} = \min\{d_{32}, d_{52}\} = \min\{7, 10\} = 7,$$

$$d_{(35),4} = \min\{d_{34}, d_{54}\} = \min\{9, 8\} = 8.$$

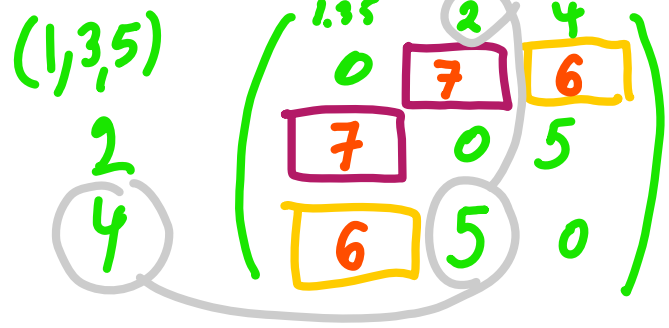
- Διαγράφουμε τις γραμμές και στήλες που αντιστοιχούν στα 3 και 5 και στη συνέχεια προσθέτουμε μια γραμμή και στήλη που αντιστοιχεί στη συστάδα $(3,5)$, οπότε και προκύπτει ο νέος πίνακας αποστάσεων

$$\mathbf{D} = (d_{ij}) = \begin{array}{c} \\ (35) \\ 1 \\ 2 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} (35) & 1 & 2 & 4 \end{array} \\ \begin{pmatrix} 0 & 3 & 7 & 8 \\ 3 & 0 & 9 & 6 \\ 7 & 9 & 0 & 5 \\ 8 & 6 & 5 & 0 \end{pmatrix} \end{array}$$

Συμμετρικός

Παράδειγμα III

Single Linkage - Nearest Neighbor

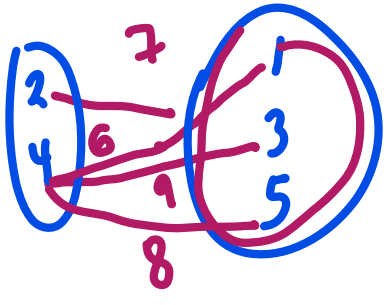


- Πλέον η μικρότερη απόσταση είναι η $d_{(35),1} = 3$ και συνεπώς συγχωνεύουμε τις $(3,5)$ και (1) για το σχηματισμό της $(1,3,5)$.
- Με τρόπο ανάλογο, έχουμε

$$d_{(135),2} = \min\{d_{1,2}, d_{(35),2}\} = \min\{9,7\} = 7,$$

$$d_{(135),4} = \min\{d_{1,4}, d_{(35),4}\} = \min\{6,8\} = 6,$$

- Ο νέος πίνακας αποστάσεων είναι



$$\mathbf{D} = (d_{ij}) = \begin{array}{c} \\ (135) \\ 2 \\ 4 \end{array} \begin{array}{c} \begin{array}{ccc} (135) & 2 & 4 \end{array} \\ \begin{pmatrix} 0 & 7 & 6 \\ 7 & 0 & 5 \\ 6 & 5 & 0 \end{pmatrix} \end{array}$$

$$\begin{array}{c} \\ (135) \\ (24) \end{array} \begin{array}{c} \begin{array}{cc} (135) & (24) \end{array} \\ \begin{pmatrix} 0 & 6 \\ 6 & 0 \end{pmatrix} \end{array}$$

Παράδειγμα IV

Single Linkage - Nearest Neighbor

- Η νέα απόσταση του εγγύτερου γείτονα είναι η $d_{4,2} = 5$ και άρα τα αντικείμενα 4 και 2 ενώνονται σε μια ομάδα $(2,4)$. Πλέον έχουν δημιουργηθεί 2 ομάδες, η (135) και η (24). Η απόσταση εγγύτερου γείτονα είναι η

$$d_{(135),(24)} = \min\{d_{(135),2}, d_{(135),4}\} = \min\{7,6\} = 6.$$

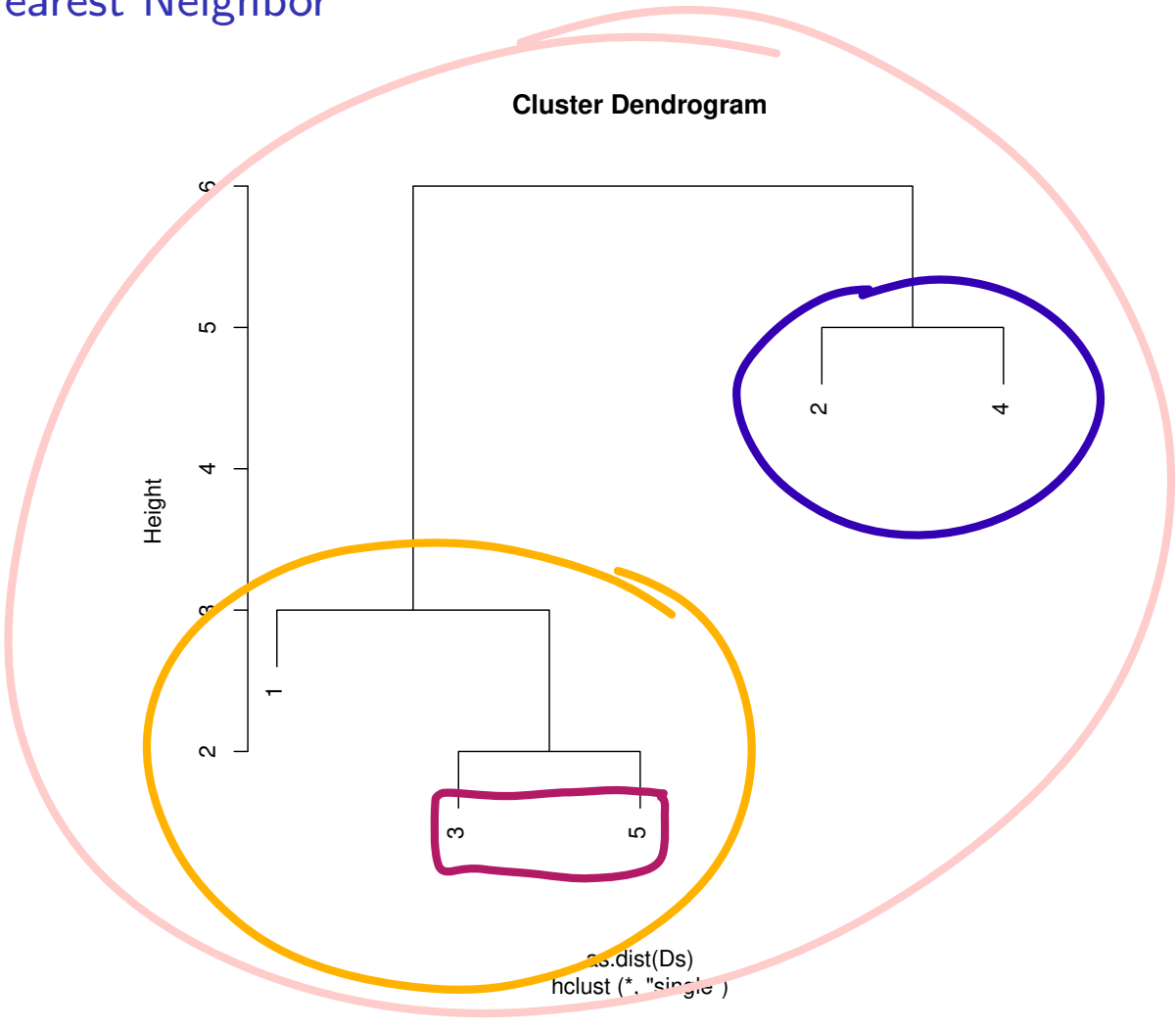
- Ο τελικός 2×2 πίνακας αποστάσεων περιέχει μία μόνο απόσταση, την $d_{(135),(24)} = 6$ και μορφή

$$\mathbf{D} = (d_{ij}) = \begin{pmatrix} 0 & 6 \\ 6 & 0 \end{pmatrix}$$

- Επομένως, οι (135) και (24) συγχωνεύονται σε μια μοναδική ομάδα των 5 ατόμων. Παρακάτω δίνεται και το αντίστοιχο δενδρόγραμμα.

Παράδειγμα V

Single Linkage - Nearest Neighbor



Σχήμα: Δενδρόγραμμα για το παραπάνω παράδειγμα

Παράδειγμα Ι

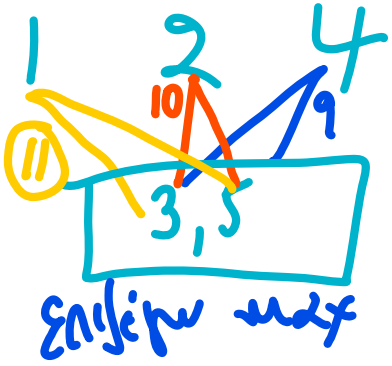
Complete Linkage - Furthest Neighbor

- Ας θεωρήσουμε το ίδιο πλήθος αντικειμένων όπως στο προηγούμενο παράδειγμα και ας εφαρμόσουμε τη μέθοδο της Πλήρους Ένωσης - Απώτατου Γείτονα για τη δημιουργία συστάδων.
- Όπως και στην Απλή Ένωση, βρίσκουμε ότι

$$\min\{d_{ij}, 1 \leq i, j \leq 5\} = d_{53} = 2$$

και έτσι οι (5) και (3) ενώνονται και σχηματίζουν τη συστάδα $(3,5)$.

- Στο επόμενο βήμα, ο νέος πίνακας $\mathbf{D}$ θα περιέχει τις (απώτερες) αποστάσεις της $(3,5)$ από τα αντικείμενα 1, 2 και 4:



$$d_{(35),1} = \max\{d_{31}, d_{51}\} = \max\{3, 11\} = 11,$$

$$d_{(35),2} = \max\{d_{32}, d_{52}\} = \max\{7, 10\} = 10,$$

$$d_{(35),4} = \max\{d_{34}, d_{54}\} = \max\{9, 8\} = 9.$$

Παράδειγμα II

Complete Linkage - Furthest Neighbor

- Διαγράφουμε τις γραμμές και στήλες που αντιστοιχούν στα 3 και 5 και στη συνέχεια προσθέτουμε μια γραμμή και στήλη που αντιστοιχεί στη συστάδα $(3,5)$, οπότε και προκύπτει ο νέος πίνακας αποστάσεων

$$\mathbf{D} = (d_{ij}) = \begin{array}{c} \\ (35) \\ 1 \\ 2 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} (35) & 1 & 2 & 4 \end{array} \\ \begin{pmatrix} 0 & 11 & 10 & 9 \\ 11 & 0 & 9 & 6 \\ 10 & 9 & 0 & 5 \\ 9 & 6 & 5 & 0 \end{pmatrix} \end{array}$$

- Πλέον η μικρότερη απόσταση είναι η $d_{(35),1} = 5$ και συνεπώς συγχωνεύουμε τις (2) και (4) στην ομάδα $(2,4)$.
- Με τρόπο ανάλογο, έχουμε

$$d_{(24),(35)} = \max\{d_{2,(35)}, d_{4,(35)}\} = \max\{10,9\} = 10,$$

$$d_{(24),1} = \max\{d_{2,1}, d_{4,1}\} = \max\{9,6\} = 9.$$

Παράδειγμα III

Complete Linkage - Furthest Neighbor

- Ο νέος πίνακας αποστάσεων είναι

$$\mathbf{D} = (d_{ij}) = \begin{pmatrix} 0 & 10 & 11 \\ 10 & 0 & 9 \\ 11 & 9 & 0 \end{pmatrix}$$

- Η νέα μικρότερη απόσταση είναι η $d_{(2,4),1} = 9$ και άρα τα αντικείμενα 2, 4 και 1 ενώνονται σε μια ομάδα $(1,2,4)$. Πλέον έχουν δημιουργηθεί 2 ομάδες, η (35) και η (124). Η απόσταση απώτερου γείτονα είναι η

$$d_{(35),(124)} = \max\{d_{1,(35)}, d_{(24),(35)}\} = \max\{11, 10\} = 11.$$

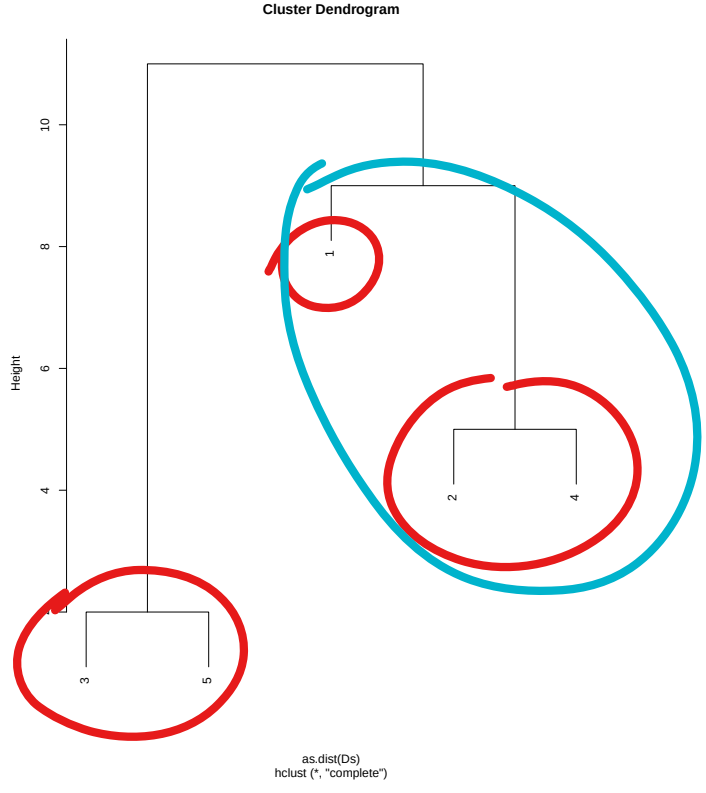
- Ο τελικός 2×2 πίνακας αποστάσεων περιέχει μία μόνο απόσταση, την $d_{(35),(124)} = 11$ και μορφή

$$\mathbf{D} = (d_{ij}) = \begin{pmatrix} 0 & 11 \\ 11 & 0 \end{pmatrix}$$

Παράδειγμα IV

Complete Linkage - Furthest Neighbor

- Επομένως, οι (35) και (124) συγχωνεύονται σε μια μοναδική ομάδα των 5 ατόμων. Παρακάτω δίνεται και το αντίστοιχο δενδρόγραμμα.

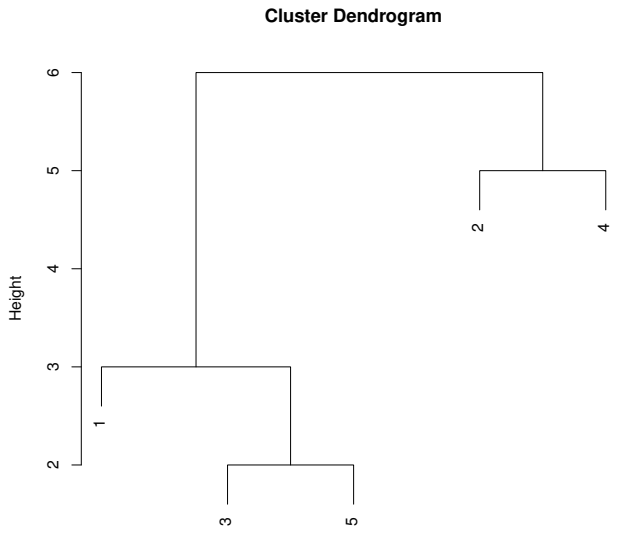


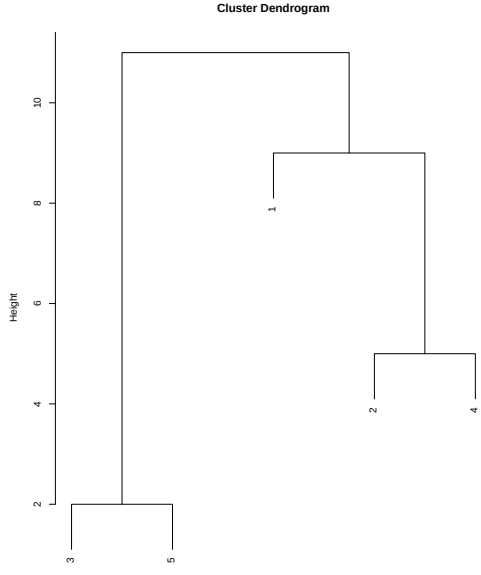
Σχήμα: Δενδρόγραμμα για το παραπάνω παράδειγμα, Furthest Neighbor

Παράδειγμα V

Complete Linkage - Furthest Neighbor

- Συγκρίνοντας τα δύο δενδρογράμματα της απλής και της πλήρους ένωσης, βλέπουμε ότι οι σχηματισμοί των ομάδων (35) και (24) γίνονται σε διαφορετικά στάδια της διαδικασίας συσταδοποίησης.
- Και οι 2 τεχνικές οδηγούν στις ομάδες (24) και (35) με κάποια διαφοροποίηση ως προς την τοποθέτηση του 1, είτε στην (124) (simple linkage) είτε στην (135) (complete linkage).





Μη-Ιεραρχικοί Αλγόριθμοι Συσταδοποίησης Ι

k-Means Method

- Οι πιο συχνά χρησιμοποιούμενες μέθοδοι δημιουργίας συστάδων βασίζονται στο κριτήριο του τετραγωνικού σφάλματος (square-error criterion)
- Έστω $\mathbf{X}$ το σύνολο $n \times p$ των δεδομένων, το οποίο έχει διαμεριστεί (με κάποιον τρόπο) στις συστάδες $C_1, C_2, \ldots, C_k$, καθεμία από τις οποίες περιέχει n_i παρατηρήσεις εκ των $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$.
- Έστω

$$\bar{\mathbf{x}}_i = \frac{1}{n} \sum_{s=1}^{n_i} \mathbf{x}_s^{(i)}, \quad i = 1, 2, \ldots, k$$

το οποίο αποτελεί το κέντρο της i-οστης συστάδας.

Μη-Ιεραρχικοί Αλγόριθμοι Συσταδοποίησης ΙΙ

k-Means Method

- Η μεταβλητότητα εντός της C_i (within-cluster variation) ισούται με το άθροισμα των τετραγώνων της Ευκλείδειας απόστασης μεταξύ $\mathbf{x}_s^{(i)}$ και $\overline{x}_i$, δηλ.

$$e_i^2 = \sum_{s=1}^{n_i} (\mathbf{x}_s^{(i)} - \overline{\mathbf{x}}_i)' (\mathbf{x}_s^{(i)} - \overline{\mathbf{x}}_i).$$

- Το ολικό σφάλμα συσταδοποίησης (overall clustering error) είναι το

$$\text{Sum of squared errors} = E_k^2 = \sum_{i=1}^{k} e_i^2$$

και ο στόχος είναι να βρεθεί μια διαμέριση του συνόλου δεδομένων $\mathbf{X}$ η οποία να ελαχιστοποιεί το E_k^2 για δεδομένο k.

- Ο αλγόριθμος για την εύρεση αυτής της διαμέρισης είναι γνωστός ως k-means Algorithm και τα βήματά του είναι τα εξής:

Μη-Ιεραρχικοί Αλγόριθμοι Συσταδοποίησης III

k-Means Method

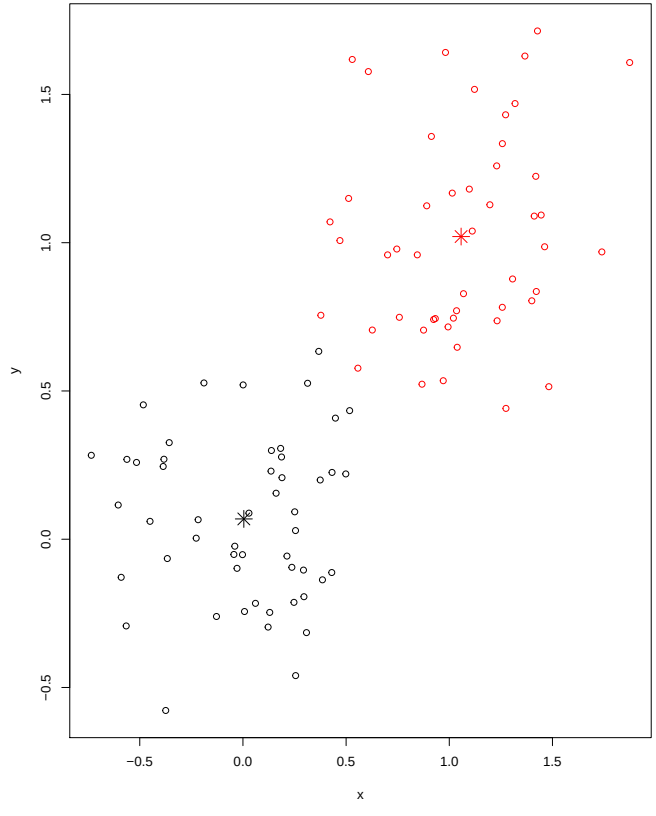
- **B1:** Επιλέγω τα k αρχικά κέντρα (seed points) για τις k (μη προσδιορισμένες) ομάδες.
- **B2:** Τοποθετούμε καθένα από τα $\mathbf{x}_i$, $i = 1, \ldots, n$, αντικείμενα στην πιο κοντινή ομάδα (μικρότερη απόσταση με το κέντρο της ομάδας).
- **B3:** Υπολογίζουμε εκ νέου τα κέντρα των νέων ομάδων (μετά την τοποθέτηση των αντικειμένων στο B2)
- **B4:** Επαναλαμβάνουμε τα B2 και B3 μέχρι να προκύψει μια βέλτιση τιμή για το E_k^2.
- **Σχόλιο:** Αφού σχηματιστούν οι ομάδες, μπορούμε να τροποποιήσουμε τον τελικό αριθμό τους είτε συγχωνεύοντας είτε διαιρώντας κάποιες από τις υπάρχουσες, είτε ακόμη και απομακρύνοντας μικρές ή εξωκείμενες ομάδες.

- Ο αλγόριθμος k-means είναι υπολογιστικά αποδοτικός και δίνει πολύ καλά αποτελέσματα (συμπαγείς ομάδες, σφαιρικές και καλώς διαχωρισμένες).

Μη-Ιεραρχικοί Αλγόριθμοι Συσταδοποίησης IV

k-Means Method

- Επίσης, έχει καλή απόδοση στην ανίχνευση/δημιουργία ομάδων ελλειψοειδούς σχήματος, αν χρησιμοποιηθεί η απόσταση Mahalanobis αντί της Ευκλείδειας.



Μη Ιεραρχικές Μέθοδοι Ταξινόμησης Ι

Συμπληρωματικό Υλικό

- Οι τεχνικές αυτές δεν απαιτούν τον προσδιορισμό του πίνακα αποστάσεων $\mathbf{D}$ ενώ τα δεδομένα $\mathbf{x}_i$, $i = 1, \ldots, n$ δεν είναι ανάγκη να αποθηκεύονται κατά την εκτέλεση του αλγορίθμου.
- Κατά συνέπεια, οι μη-ιεραρχικές μέθοδοι ταξινόμησης είναι προτιμότερες για μεγάλα σύνολα δεδομένων.
- **Παράδειγμα**: Έστω $k = 2$ ο προκαθορισμένος αριθμός ομάδων και $n = 4$ άτομα (έστω A, B, C, D), για τα οποία έχουν σημειωθεί οι τιμές των (X_1, X_2)

Άτομο	x_1	x_2
A	5	3
B	-1	1
C	1	-2
D	-3	-2

Μη Ιεραρχικές Μέθοδοι Ταξινόμησης ΙΙ

Συμπληρωματικό Υλικό

- Παίρνουμε αυθαίρετα τη διαμέριση (AB) και (CD) και υπολογίζουμε τα κέντρα $(\overline{x}_1, \overline{x}_2)$ για την κάθε ομάδα.

Ομάδα	$\overline{x}_1$	$\overline{x}_2$
(AB)	$\frac{5+(-1)}{2} = 2$	$\frac{3+1)}{2} = 2$
(CD)	$\frac{1+(-3)}{2} = -1$	$\frac{-2+(-2)}{2} = -2$

- Στο επόμενο βήμα, υπολογίζουμε την Ευκλείδεια απόσταση κάθε ατόμου από τα κέντρα των 2 ομάδων. Τοποθετούμε το κάθε άτομο στην πλησιέστερη ομάδα.
- Όταν ένα άτομο φύγει από την αρχική του ομάδα, τα κέντρα των (νέων) ομάδων 'διορθώνονται' προτού προχωρήσουμε στο επόμενο βήμα.

Μη Ιεραρχικές Μέθοδοι Ταξινόμησης ΙΙΙ

Συμπληρωματικό Υλικό

- Υπολογίζουμε τις Ευκλείδειες αποστάσεις (τα τετράγωνά τους), αρχικά για το Α:

$$d^2(A,(AB)) = (5-2)^2 + (3-2)^2 = 10,$$

$$d^2(A,(CD)) = (5+1)^2 + (3+2)^2 = 61.$$

- Άρα το Α είναι πιο κοντά στην (ΑΒ) απ΄ότι στη (CD), οπότε και παραμένει στην (ΑΒ).
- Ομοίως, για το Β:

$$d^2(B,(AB)) = (-1-2)^2 + (-1-2)^2 = 10,$$

$$d^2(B,(CD)) = (-1+1)^2 + (1+2)^2 = 9.$$

- Επομένως, το Β τοποθετείται στη (CD), δίνοντας την ομάδα (BCD). Τα νέα κέντρα είναι:

Μη Ιεραρχικές Μέθοδοι Ταξινόμησης IV

Συμπληρωματικό Υλικό

Ομάδα	$\overline{x}_1$	$\overline{x}_2$
A	5	3
(BCD)	$\frac{-1+1+(-3)}{3} = -1$	$\frac{1-2+(-2)}{3} = -1$

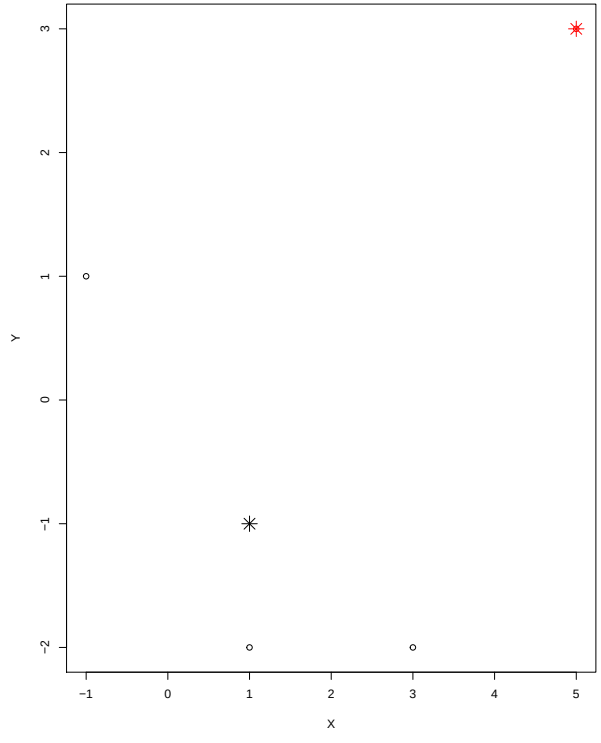
- Πάλι ελέγχουμε κάθε άτομο για ενδεχόμενο μετακίνηση σε άλλη ομάδα. Βρίσκουμε τον πίνακα με τις Ευκλείδειες αποστάσεις από τα κέντρα των (A) και (BCD).

Ομάδα	A	B	C	D
A	0	40	41	89
(BCD)	52	4	5	5

- Βλέπουμε ότι κάθε άτομο παραμένει στην ομάδα του και η διαδικασία σταμτά με $k = 2$ ομάδες, την (A) και την (BCD).

Μη Ιεραρχικές Μέθοδοι Ταξινόμησης V

Συμπληρωματικό Υλικό



Σχήμα: Διαγραμματική Απεικόνιση των 2 Ομάδων, K-Means example