Ομαδοποίηση και Κ-Μέσοι

Ιωάννης Μαδεμλής

Ομαδοποίηση

Ορισμός της ομαδοποίησης

- **Ομαδοποίηση** (clustering) είναι η *ανεπίβλεπτη* εργασία διαμέρισης ενός δοθέντος συνόλου δεδομένων σε ομάδες.

- Κάθε στοιχείο της ομάδας μοιράζεται ένα κοινό χαρακτηριστικό με τα υπόλοιπα της ομάδας του, αλλά όχι με στοιχεία άλλων ομάδων.

- Η διαμέριση αυτή μπορεί να έχει:
  - Είτε φυσική ερμηνεία, οπότε η ομαδοποίηση αντανακλά τη φυσική δομή των δεδομένων,
  - Είτε απλώς μεγαλύτερη χρηστική αξία, καθώς τα στοιχεία συνοψίζονται από τις ομάδες στις οποίες κατηγοριοποιούνται.

Ορισμός της ομαδοποίησης

- Η κατάταξη των προτύπων κάποιου συνόλου δεδομένων σε ομάδες απεικονίζει κάθε πρότυπο σε μία κατηγορία: την ομάδα όπου ανήκει.
  - Στο τέλος της ομαδοποίησης, κάθε πρότυπο από το δοθέν σύνολο δεδομένων συνιστά πλέον **στοιχείο** μίας ομάδας.

Ομαδοποίηση

- Διαφέρει από την ταξινόμηση διότι εξάγει συμπεράσματα αποκλειστικά από τα ίδια τα δεδομένα.
  - Δεν αξιοποιεί επιπρόσθετη πληροφορία (π.χ., γνωστές ετικέτες κλάσεων σε κάποιο στάδιο εκπαίδευσης).
  - Ανεπίβλεπτη μάθηση.

Ομαδοποίηση

Ορισμός της ομαδοποίησης

- Ο ορισμός της ομάδας δεν είναι ακριβής.
  - Η καταλληλότερη ομαδοποίηση εξαρτάται από το εκάστοτε πρόβλημα και σύνολο δεδομένων.

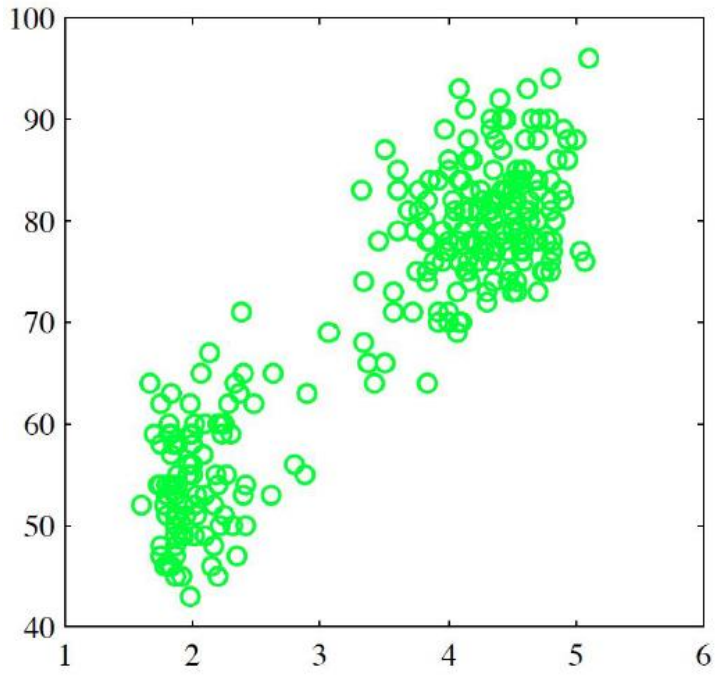
- Ένα κοινό χαρακτηριστικό είναι ότι:
  - όσο μεγαλύτερη είναι η ομοιότητα των προτύπων στο εσωτερικό μίας ομάδας, και
  - όσο μεγαλύτερη η διαφορά τους από στοιχεία άλλων ομάδων,

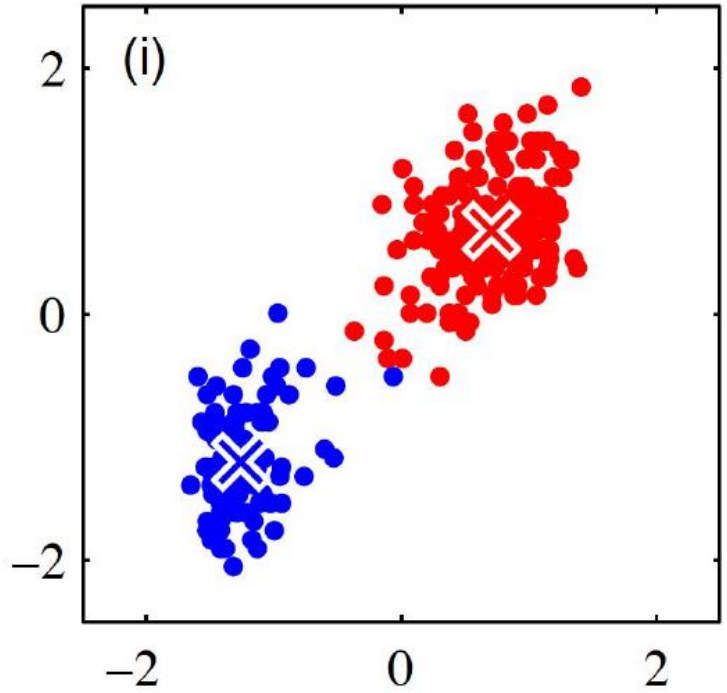
- …τόσο πιο **συμπαγείς** (compact) θεωρούνται οι παραγόμενες ομάδες.
  - Μεγαλύτερη *συμπάγεια* ⟶ επιτυχέστερη ομαδοποίηση.

Ορισμός της ομαδοποίησης

- Ανιχνεύουμε μία δομή στα υπάρχοντα δεδομένα.

Ομαδοποίηση





Πηγή: Διαφάνειες ΜΜ 2020

Ομαδοποίηση

Ορισμός της ομαδοποίησης

- Ωστόσο, είναι θεμιτό να υφίσταται κάποια εκ των προτέρων φυσική τάση διαμέρισης του συνόλου δεδομένων σε ομάδες, διαθέσιμη προς ανακάλυψη από τον αλγόριθμο ομαδοποίησης.
  - Διαφορετικά, τα αποτελέσματά του ενδέχεται να μοιάζουν αυθαίρετα και «αφύσικα».

- Οι περισσότεροι αλγόριθμοι ομαδοποίησης εμπίπτουν σε μία από δύο μεγάλες οικογένειες:
  - **Ιεραρχικοί**, ή
  - **Διαμεριστικοί**.

Τύποι ομαδοποίησης

- Οι ιεραρχικοί αλγόριθμοι παράγουν ένα σύνολο εμφωλευμένων, δενδρικά ιεραρχημένων ομάδων: *δενδρόγραμμα*.

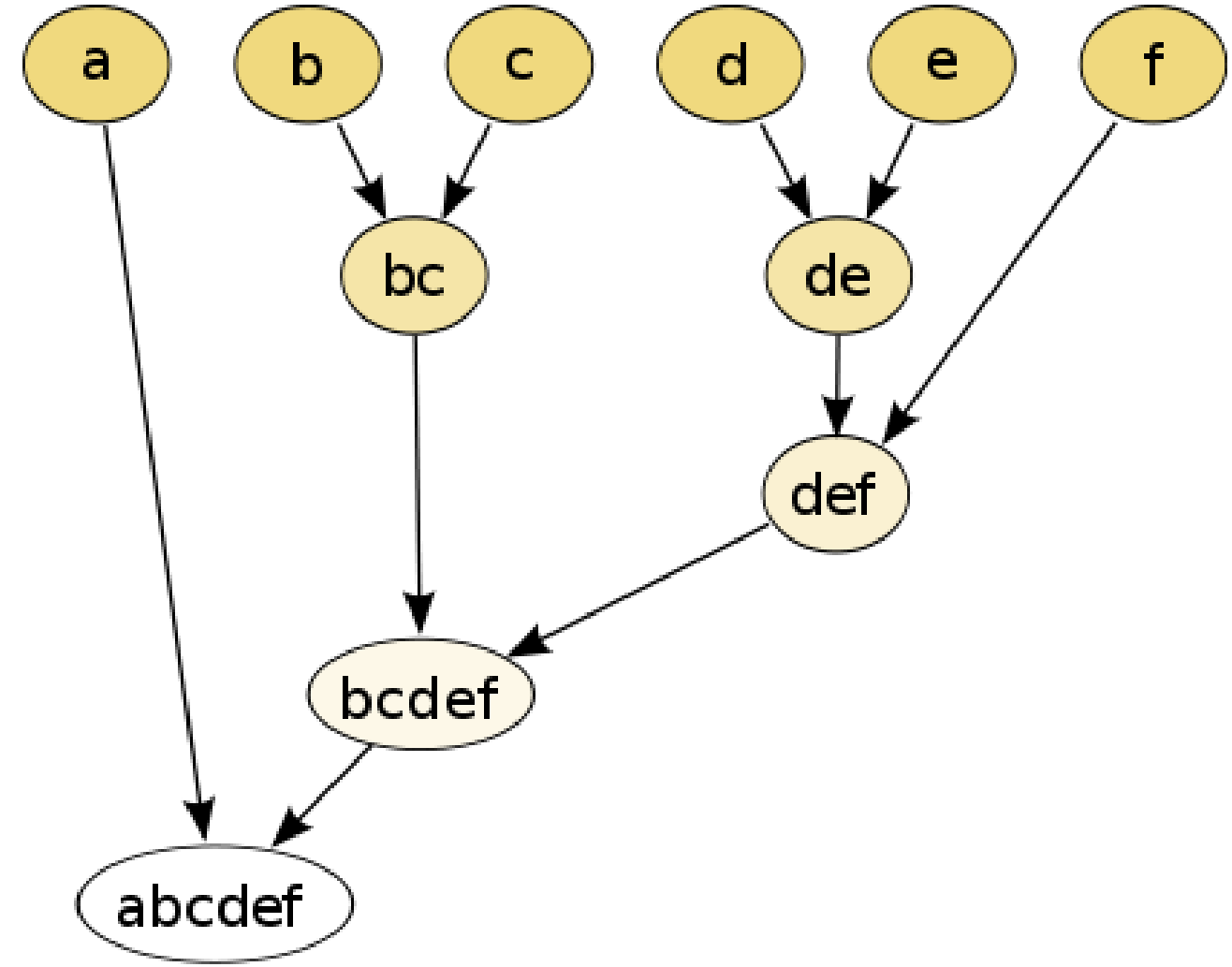
- Κάθε κόμβος-ομάδα του δένδρου ισοδυναμεί με την ένωση των θυγατρικών του κόμβων (υπο-ομάδες).

- Ο κόμβος-ρίζα περιέχει όλα τα πρότυπα του συνόλου δεδομένων.

- Αντιθέτως, τα φύλλα του δένδρου ενδέχεται να περιέχουν ακόμη και μόνο ένα πρότυπο το καθένα.

Τύποι ομαδοποίησης

Ομαδοποίηση



Πηγή: Wikipedia.

Τύποι ομαδοποίησης

- Κατ' αντιδιαστολή, η διαμεριστική ομαδοποίηση παράγει μία διαμέριση του συνόλου δεδομένων σε ξένα υποσύνολα (ομάδες).
  - Κάθε στοιχείο ανήκει σε ένα μόνο τέτοιο υποσύνολο.
- Όμως, η ιεραρχική ομαδοποίηση ισοδυναμεί με μία ακολουθία διαμεριστικών ομαδοποιήσεων.
- Παράλληλα, μία διαμεριστική ομαδοποίηση είναι δυνατόν να εξαχθεί από μία ιεραρχική ως εξής:
  - Αποκόπτουμε το ιεραρχικό δένδρο σε κάποιο ζητούμενο επίπεδο, και
  - Κρατούμε τους κόμβους αυτού του επιπέδου ως τελικούς.
    - Πρόκειται για ξένα μεταξύ τους υποσύνολα.

Τύποι ομαδοποίησης

- Μία εναλλακτική διάκριση είναι μεταξύ:
  - **σκληρής ομαδοποίησης**, όπου κάθε πρότυπο ανήκει αυστηρά σε μία ομάδα,
  - **επικαλυπτόμενης ομαδοποίησης**, όπου κάθε πρότυπο μπορεί να ανήκει ταυτόχρονα σε πολλαπλές ομάδες, και
  - **ασαφούς ομαδοποίησης**, όπου σε κάθε πρότυπο ανατίθεται μία πιθανότητα να ανήκει ξεχωριστά σε κάθε ομάδα.
- Τέλος, διακρίνουμε επίσης μεταξύ:
  - **πλήρους ομαδοποίησης**, όπου όλα τα στοιχεία του συνόλου δεδομένων κατηγοριοποιούνται σε κάποια ομάδα, και
  - **μερικής ομαδοποίησης**, όπου επιτρέπεται σε κάποια στοιχεία να μείνουν εκτός των παραγόμενων ομάδων.

Ομαδοποίηση

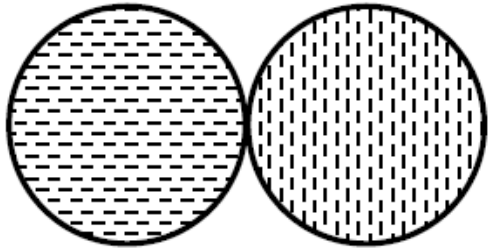
Τύποι ομάδων

- Υπάρχουν ακόμα διαφορετικοί ορισμοί της έννοιας της ομάδας, για χρήση σε διαφορετικά προβλήματα:
  - **καλώς διαχωρισμένων ομάδων**, όπου κάθε στοιχείο τους είναι περισσότερο όμοιο με *όλα τα λοιπά στοιχεία τους*, παρά με οποιοδήποτε στοιχείο άλλων ομάδων,
  - **ομάδων πρωτοτύπων**, όπου κάθε στοιχείο τους είναι περισσότερο όμοιο με *ένα στοιχείο-αντιπρόσωπο της ομάδας* (το **πρωτότυπο**), παρά με το πρωτότυπο οποιασδήποτε άλλης ομάδας,
  - **ομάδων γειτνίασης**, όπου κάθε στοιχείο τους είναι περισσότερο όμοιο με *τουλάχιστον ένα άλλο στοιχείο τους*, παρά με οποιοδήποτε στοιχείο άλλων ομάδων, και
  - **ομάδων πυκνότητας**, όπου κάθε ομάδα είναι μία πυκνή περιοχή στοιχείων περιβαλλόμενη από μία περιοχή αραιής πυκνότητας προτύπων, στον διανυσματικό χώρο των δεδομένων.

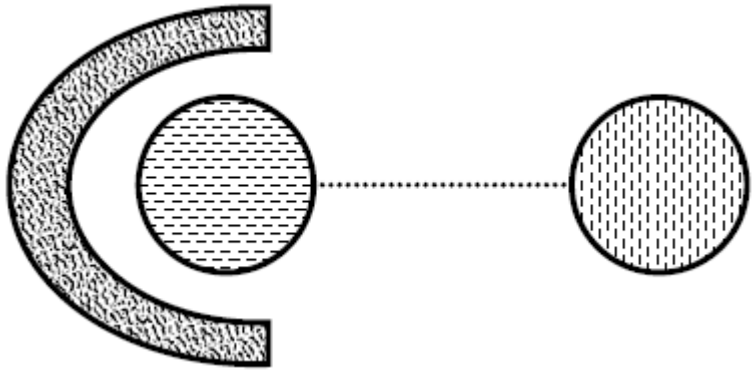
Τύποι ομάδων

Ομαδοποίηση

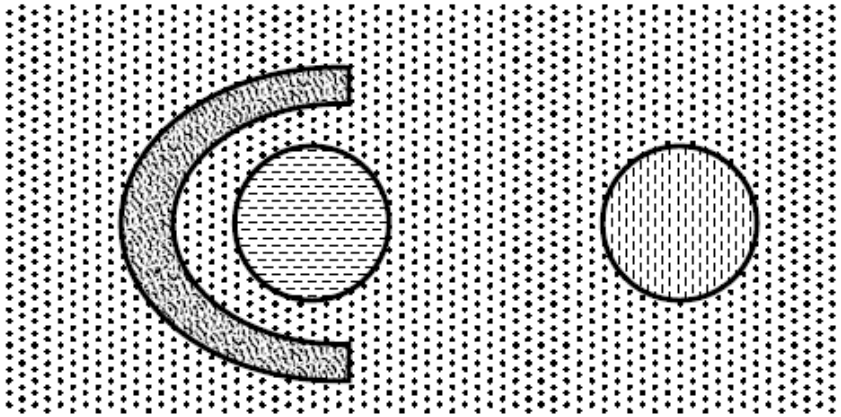
Δύο καλώς διαχωρισμένες ομάδες διδιάστατων δεδομένων.



Δύο ομάδες πρωτοτύπων διδιάστατων δεδομένων.



Δύο ομάδες γειτνίασης διδιάστατων δεδομένων.



Τρεις ομάδες πυκνότητας διδιάστατων δεδομένων.

Πηγή: Tan et al., *Introduction to Data Mining*, 2006.

Ομαδοποίηση

Τύποι ομάδων

- Οι ανωτέρω ορισμοί απαιτούν τη χρήση κάποιου **μέτρου εγγύτητας** (*ομοιότητας* ή *απόστασης*) στον διανυσματικό χώρο των προτύπων.
- Τα συνηθέστερα μέτρα εγγύτητας μεταξύ δύο n-διάστατων προτύπων $\mathbf{x}_i$ και $\mathbf{x}_j$ είναι:
  - Η *απόσταση Μανχάταν* είναι η νόρμα $\mathcal{L}_1$ της διαφοράς των $\mathbf{x}_i$ και $\mathbf{x}_j$:
    $\mathcal{L}_1(\mathbf{x}_i, \mathbf{x}_j) = \left|\left|\mathbf{x}_j - \mathbf{x}_i\right|\right|_1 = \sum_{k=1}^{n} |x_{jk} - x_{ik}|.$
  - Η *ευκλείδεια απόσταση* είναι η νόρμα $\mathcal{L}_2$ της διαφοράς των $\mathbf{x}_i$ και $\mathbf{x}_j$:
    $\mathcal{L}_2(\mathbf{x}_i, \mathbf{x}_j) = \left|\left|\mathbf{x}_j - \mathbf{x}_i\right|\right|_2 = \left(\sum_{k=1}^{n} (x_{jk} - x_{ik})^2\right)^{\frac{1}{2}}.$
  - Ο *συντελεστής ομοιότητας συνημιτόνου* είναι το συνημίτονο της γωνίας μεταξύ $\mathbf{x}_i$ και $\mathbf{x}_j$:
    $cos\theta(\mathbf{x}_i, \mathbf{x}_j) = \frac{\mathbf{x}_i^T \mathbf{x}_j}{||\mathbf{x}_i||_2 ||\mathbf{x}_j||_2}.$

Ομαδοποίηση

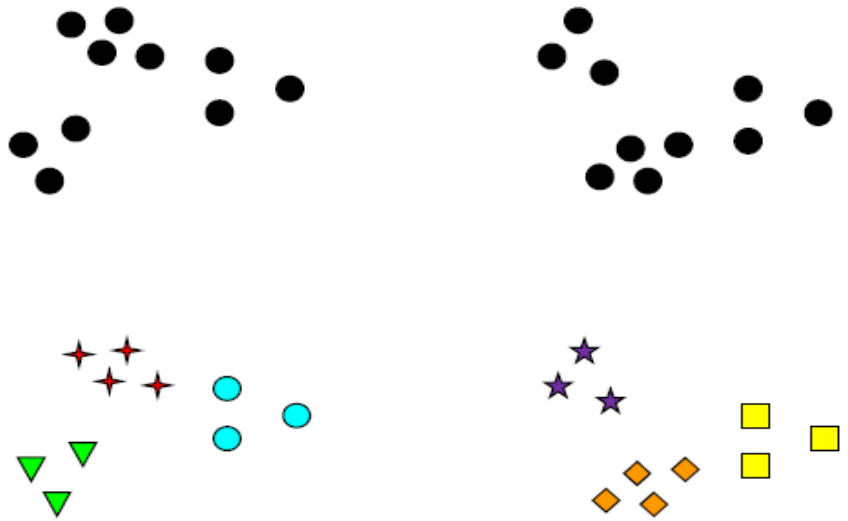
Τύποι ομάδων

- Τα πρωτότυπα σε χώρους συνεχών γνωρισμάτων τοποθετούνται στο μέσον των αντίστοιχων ομάδων και ονομάζονται **κεντροειδή** (centroid).
- Τα πρωτότυπα σε χώρους διακριτών γνωρισμάτων είναι τα πιο αντιπροσωπευτικά στοιχεία των αντίστοιχων ομάδων και ονομάζονται **μεσοειδή** (medoid).
- Τα κεντροειδή και τα μεσοειδή ονομάζονται από κοινού και **κέντρα** των αντίστοιχων ομάδων.
- Ένα μεσοειδές εξ ορισμού αποτελεί στοιχείο της ομάδας του, ενώ ένα κεντροειδές όχι.
- Ένα κεντροειδές πιθανόν να μην ταυτίζεται με *κανένα* πρότυπο.

Ομαδοποίηση

Αμφισημία και εφαρμογές ομαδοποίησης

- Εξαιτίας της έλλειψης ενός σταδίου επιβλεπόμενης εκπαίδευσης με τη βοήθεια γνωστών ετικετών, το πρόβλημα είναι συνήθως πολύ δύσκολο σε πραγματικές συνθήκες.
- Οι παραγόμενες ομάδες είναι αμφίσημες και εξαρτώνται πολύ από τις επιλεγμένες τιμές υπερπαραμέτρων.
  - Παράδειγμα: η φυσική διαμέριση των δεδομένων αυτών είναι σε μία, σε δύο, ή σε τρεις ομάδες;
  - Μία ομαδοποίηση τριών ομάδων δεν είναι *εμφανώς* προτιμότερη από μία των δύο ομάδων.



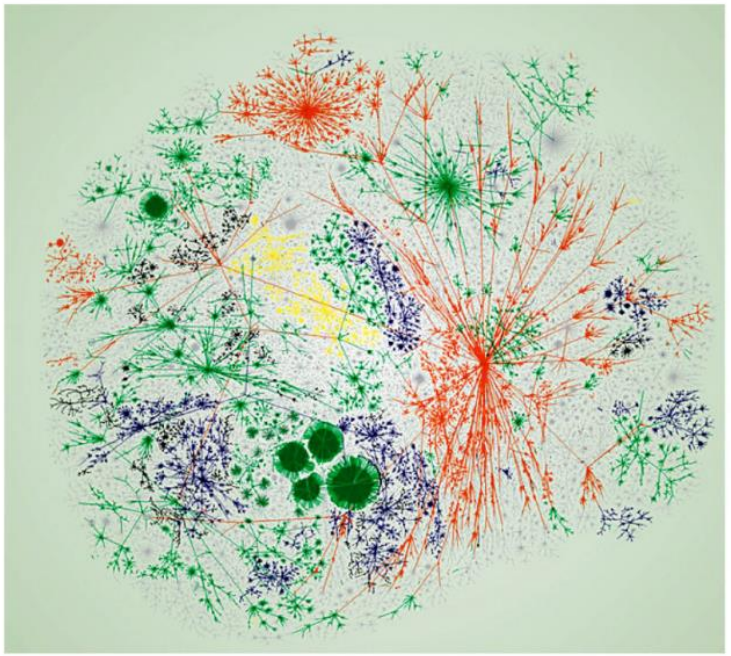
Πηγή: Διαφάνειες ΜΜ 2021.

Αμφισημία και εφαρμογές ομαδοποίησης

- Ωστόσο, υπάρχει τεράστιο πεδίο εφαρμογών της ομαδοποίησης και, επομένως, μεγάλο ενδιαφέρον.

Ομαδοποίηση

Κατάτμηση εικόνας (Image segmentation)



Ομαδοποίηση κόμβων σε γράφους (π.χ. κοινωνικά δίκτυα)

Πηγή: Διαφάνειες ΜΜ 2020.

Αλγόριθμος των k-μέσων

- Ο πιο διαδεδομένος διαμεριστικός αλγόριθμος ομαδοποίησης πρωτοτύπων είναι ο αλγόριθμος των **k-μέσων**.
- Απαιτεί προκαθορισμό του πλήθους των ζητούμενων ομάδων (υπερπαράμετρος k).
- Εφαρμόζεται κυρίως σε χώρους συνεχών γνωρισμάτων.
- Τα πρωτότυπα που ορίζει είναι κεντροειδή.
- Γνωστός και ως **αλγόριθμος του Lloyd** (1957).
- Υπάρχουν δεκάδες παραλλαγές/βελτιώσεις του.

Αλγόριθμος των k-μέσων

- Στις περισσότερες περιπτώσεις ζητούμε απλώς τη διαμέριση των δοθέντων προτύπων σε k ομάδες.

- Όμως ορισμένες φορές είναι θεμιτό να έχουμε τη δυνατότητα αντιστοίχισης *καινούργιων* μελλοντικών προτύπων, όταν αυτά εμφανίζονται, σε μία από τις k ομάδες όπου κατέληξε η ομαδοποίηση των αρχικών δεδομένων.

- Σε αυτές τις περιπτώσεις, τα k πρωτότυπα τα οποία δίνουν ως έξοδο οι k-μέσοι μπορούν να χρησιμοποιηθούν ως βάση δεδομένων ενός αλγορίθμου ταξινόμησης kNN με $k = 1$.
  - Έτσι αναθέτουμε κάθε *νέο* πρότυπο στην εγγύτερή του ομάδα, με τα πρωτότυπα ως αντιπροσώπους των ομάδων.

k-Μέσοι

Αλγόριθμος των k-μέσων

- Είσοδος: N n-διάστατα πρότυπα.

- Έξοδος: 1 N-διάστατο διάνυσμα ανάθεσης ομάδων $\mathbf{c}$.
  - Η i-οστή τιμή του είναι φυσικός αριθμός στο [1, k]: ο αύξων αριθμός της ομάδας όπου ανήκει το i-οστό πρότυπο.

  k n-διάστατα κεντροειδή $\boldsymbol{\mu}_i$, $1 \leq i \leq k$.

- Μέτρο εγγύτητας: Ευκλείδεια απόσταση.

- Συνάρτηση κόστους: **Αθροιστική διασπορά ομαδοποίησης**: ολικό Sum of Squared Errors (SSE).
  - **Διασπορά** είναι το άθροισμα των τετραγώνων των ευκλείδειων αποστάσεων κάθε στοιχείου μίας ομάδας από το κέντρο της.
  - **Αθροιστική διασπορά** είναι το άθροισμα των διασπορών όλων των k παραχθέντων ομάδων.
  - Μικρή τελική αθροιστική διασπορά σημαίνει ομαδοποίηση υψηλής συμπάγειας.

k-Μέσοι

Αλγόριθμος των k-μέσων

- Με βάση τα $\boldsymbol{\mu}_i$ και το $\mathbf{c}$, κατασκευάζουμε ένα σύνολο $C = \{C_1 \cup C_2 \cup \ldots \cup C_k\}$ όπου το υποσύνολο C_i περιέχει τα N_i στοιχεία της i-οστής ομάδας.

- Άρα προσπαθούμε να λύσουμε το εξής πρόβλημα ελαχιστοποίησης:

$$argmin_C \, \text{SSE} = argmin_C \sum_{i=1}^{k} \frac{1}{N_i} \sum_{\mathbf{x},\mathbf{y} \in C_i} ||\mathbf{x} - \mathbf{y}||_2^2 .$$

- Δυσεπίλυτο πρόβλημα:
  - Το $\mathbf{c}$, με βάση το οποίο κατασκευάζεται το ζητούμενο C, περιέχει διακριτές τιμές: ⟶ άρα δεν μπορούμε να χρησιμοποιήσουμε απευθείας αλγορίθμους συνεχούς βελτιστοποίησης, αφού η συνάρτηση SSE δεν είναι παραγωγίσιμη ως προς $\mathbf{c}$.
  - Μη κυρτή συνάρτηση κόστους: ⟶ πολλαπλά τοπικά ελάχιστα.
  - Έχει αποδειχθεί πως εμπίπτει στην κλάση υπολογιστικής πολυπλοκότητας των *NP-δύσκολων* προβλημάτων.

Αλγόριθμος των k-μέσων

- Ο αλγόριθμος του Lloyd είναι μία **προσεγγιστική**, **επαναληπτική**, **άπληστη** μέθοδος επίλυσης του προβλήματος.

- Έχει αποδειχθεί πως σταδιακά συγκλίνει σε τοπικό ελάχιστο του κόστους.

- Έχει υπολογιστική πολυπλοκότητα $O(Nnk)$ ανά επανάληψη.

- Απαιτεί χειροκίνητη αρχικοποίηση των k κεντροειδών (π.χ., τυχαία) στον χώρο των προτύπων.
  - Το k υπερπαράμετρος.

k-Μέσοι

Αλγόριθμος των k-μέσων

k-Μέσοι

Βήματα

- 1. Αρχικοποίηση των k κεντροειδών.
- 2. *Επανάλαβε*:
- 3. { Ανάθεσε καθένα από τα N πρότυπα στην ομάδα την οποία αντιπροσωπεύει το εγγύτερό του κεντροειδές (με βάση την τετραγωνική ευκλείδεια απόσταση).
- 4. Επανυπολόγισε καθένα από τα k κεντροειδή, ώστε να συνιστά τον αριθμητικό μέσο των προτύπων/στοιχείων που έχουν ανατεθεί στην ομάδα του.
- 5. }
- 6. *Ώσπου*: τα κεντροειδή παύουν να μετακινούνται σημαντικά μεταξύ διαδοχικών επαναλήψεων.

Προβλήματα των k-μέσων

- Είναι βέβαιο ότι αυτός ο επαναληπτικός βρόχος θα τερματίσει σε τοπικό βέλτιστο του κόστους.

- Όμως η ποιότητα της εν λόγω λύσης εξαρτάται *πολύ* από την **αρχικοποίηση** των κεντροειδών.
  - Κακή ποιότητα λύσης ισοδυναμεί με παγίδευση σε ανεπαρκές τοπικό ελάχιστο του κόστους αθροιστικής διασποράς.

- Ενδέχεται να προκύψουν στο τέλος **κενές ομάδες**.

- Η ομαδοποίηση είναι ευάλωτη σε παραμορφώσεις εξαιτίας **ανώμαλων προτύπων** τα οποία δύσκολα ανατίθενται σε κάποια ομάδα.
  - Οι ανωμαλίες πρέπει να ανιχνεύονται και να απομακρύνονται σε ένα στάδιο προεπεξεργασίας, *πριν* την ομαδοποίηση.

Παραλλαγές των k-μέσων

- Τα ανωτέρω ζητήματα αντιμετωπίζονται σε κάποιον βαθμό με παραλλαγές του βασικού αλγορίθμου.
  - Κατάλληλες *διασπάσεις* ή *συγχωνεύσεις ομάδων*.
- Ένα παράδειγμα αποτελεί ο αλγόριθμος ISODATA.
  - Δύο ομάδες συγχωνεύονται σε μία αν:
    - η απόσταση των κέντρων τους είναι μικρότερη από ένα κατώφλι, ή
    - το πλήθος των στοιχείων τους είναι μικρότερο από ένα κατώφλι.
  - Μία ομάδα διασπάται σε δύο αν:
    - η διασπορά της είναι μεγαλύτερη από ένα κατώφλι, ή
    - το πλήθος των στοιχείων της είναι μεγαλύτερο από ένα κατώφλι.
- Στο τέλος ενδέχεται να έχουμε πλήθος ομάδων διάφορο του k.
  - Έτσι εισάγουμε όμως επιπρόσθετες υπερπαραμέτρους προς βελτιστοποίηση: τα κατώφλια.

Παραλλαγές των k-μέσων

- Εναλλακτική στρατηγική αντιμετώπισης των κενών ομάδων: **σταδιακή ομαδοποίηση** (incremental).
- Τα 2 βασικά, εναλλασσόμενα βήματα του αλγορίθμου στο εσωτερικό μίας επανάληψης πολυπλέκονται:
  - Στον βασικό αλγόριθμο, η ενημέρωση των κεντροειδών συμβαίνει αφού πρώτα ανατεθούν όλα τα πρότυπα σε ομάδες.
  - Στον σταδιακό αλγόριθμο, η ενημέρωση των κεντροειδών γίνεται μετά από *κάθε* ανάθεση μεμονωμένου στοιχείου σε ομάδα.
  - Άρα σε κάθε βήμα απαιτούνται μόνο 0 ή 2 ενημερώσεις κεντροειδών:
    - 0 αν το τελευταίο πρότυπο δεν άλλαξε ομάδα.
    - 2 αν το τελευταίο πρότυπο άλλαξε ομάδα.
  - Κάθε επανάληψη αποτελείται από N τέτοια βήματα.
- Έτσι είναι αδύνατον να παραχθούν κενές ομάδες.
  - Αν μία ομάδα μείνει με μόνο ένα τελευταίο πρότυπο $\mathbf{x}_i$, το κεντροειδές της επανυπολογίζεται κατευθείαν ώστε να συμπίπτει με το $\mathbf{x}_i$.
  - Όμως η ομαδοποίηση εξαρτάται πλέον σημαντικά από τη διάταξη των προτύπων εισόδου.

Παραλλαγές των k-μέσων

- Τι γίνεται με την εξάρτηση από την αρχικοποίηση;
  - <u>Αφελής λύση</u>:
    - Επανειλημμένη εκτέλεση ανεξάρτητων στιγμιοτύπων του αλγορίθμου με διαφορετικά αρχικά κεντροειδή, και
    - Εν τέλει, διατήρηση της ομαδοποίησης με τη μικρότερη αθροιστική διασπορά.
  - <u>Προτιμότερη λύση</u>:
    - Δειγματοληψία των προτύπων,
    - Ιεραρχική ομαδοποίηση του δείγματος,
    - Εξαγωγή k ομάδων από κατάλληλο επίπεδο του δενδρογράμματος,
    - Εν τέλει, χρήση των κεντροειδών αυτών των ομάδων ως αρχικοποίηση για τον κύριο αλγόριθμο k-μέσων στο ολικό σύνολο δεδομένων.
    - Έτσι υπάρχει κάποια εγγύηση ότι τα αρχικά κέντρα δεν διαφέρουν πολύ από τα ζητούμενα πραγματικά.
    - Άρα ο αλγόριθμος είναι περισσότερο πιθανόν να συγκλίνει **ταχύτερα** σε **περισσότερο βέλτιστη** λύση.

Παραλλαγές των k-μέσων

- Εναλλακτικά, μπορούμε να επιλέγουμε διαδοχικά τα αρχικά κεντροειδή, σε k βήματα:
  - Λαμβάνουμε κάθε φορά ως επόμενο κεντροειδές το πρότυπο που βρίσκεται **μακρύτερα** από τα ήδη επιλεγμένα στο χώρο των προτύπων.
    - Αλγόριθμος ***k*-μέσων++**: πιθανοκρατική εκδοχή αυτής της ιδέας, όπου η πιθανότητα επιλογής ενός προτύπου ως επόμενο κεντροειδές είναι ανάλογη του τετραγώνου της απόστασής του από το εγγύτερό του ήδη επιλεγμένο κεντροειδές.
- Η προσέγγιση αυτή είναι ευάλωτη σε ανωμαλίες.
  - Ενδέχεται να επιλέξουμε ως αρχικό κέντρο ομάδας ένα πρότυπο το οποίο είναι πολύ μακριά από όλα τα άλλα.
- Συνηθίζεται να εφαρμόζεται σε μικρό τυχαίο δείγμα του αρχικού συνόλου δεδομένων.
  - Έτσι μειώνεται σημαντικά η πιθανότητα επιλογής ανωμαλίας ως αρχικού κεντροειδούς.
- Μία ακόμα καλύτερη λύση είναι ο **διχοτομικός αλγόριθμος αρχικοποίησης**.

Παραλλαγές των k-μέσων

- Διχοτομικός αλγόριθμος:
  - Εκτελούμε k-1 διαδοχικές φορές τους απλούς k-μέσους:
    - Κάθε φορά διαιρούμε σε 2 ομάδες ένα *διαφορετικό* υποσύνολο των ολικών δεδομένων μας.
  - Στην πρώτη εκτέλεση, ομαδοποιούμε/διχοτομούμε όλο το σύνολο δεδομένων.
  - Στην αρχή της i-οστής εκτέλεσης έχουμε στη διάθεσή μας i ομάδες/κεντροειδή από τις προηγούμενες εκτελέσεις.
  - Επιλέγουμε ως υποσύνολο προς διχοτόμηση (μέσω ομαδοποίησης) την κληρονομημένη ομάδα με τη μεγαλύτερη διασπορά.
  - Στο τέλος της i-οστής εκτέλεσης έχουμε πλέον στη διάθεσή μας $i+1$ ομάδες/κεντροειδή.
  - Τα τελικά αποτελέσματα για $i = k$-1, τα αξιοποιούμε ως αρχικοποίηση για μία κανονική εκτέλεση των k-μέσων.
    - Αναγκαίο, διότι οι k-1 προκαταρκτικές εκτελέσεις ελαχιστοποιούν μόνο τη διασπορά *υποσυνόλων* των δεδομένων.

Παραλλαγές των k-μέσων

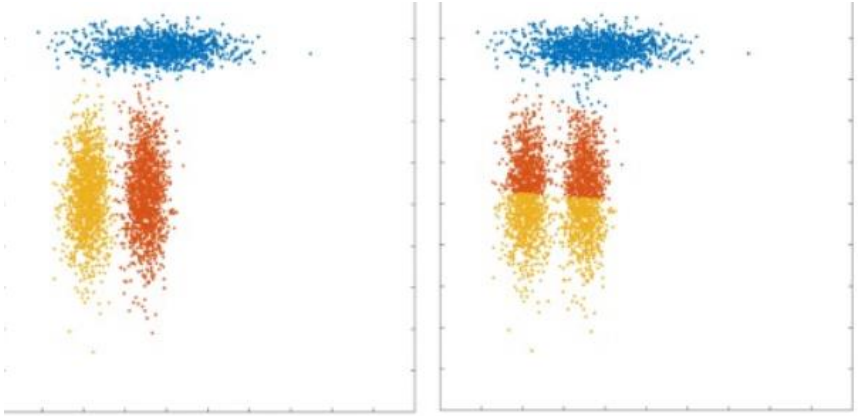
- Ο διχοτομικός αλγόριθμος αρχικοποίησης μας δίνει τη δυνατότητα να απεικονίσουμε τα αποτελέσματα των προκαταρκτικών ομαδοποιήσεων σε ένα δενδρόγραμμα.
  - Έτσι προκύπτει και μία βοηθητική ιεραρχική ομαδοποίηση.
- Τι γίνεται όμως την επιλογή του κατάλληλου πλήθους ομάδων k;
- Στόχος είναι η εύρεση της ομαδοποίησης με το ελάχιστο ολικό SSE για δεδομένο k.
  - Άρα μπορούμε να επαναλάβουμε την εκτέλεση των k-μέσων για διαφορετικές τιμές του k και να επιλέξουμε εν τέλει την ομαδοποίηση με την ελάχιστη αθροιστική διασπορά;

Προβλήματα των k-μέσων

- Δυστυχώς, η αθροιστική διασπορά δεν μπορεί να χρησιμοποιηθεί ως κριτήριο εκτίμησης της βέλτιστης τιμής του k.
  - Μεγαλύτερο k οδηγεί *πάντα* σε χαμηλότερο ολικό SSE.
- Ο αλγόριθμος είναι **απλός** και **αποδοτικός**, αλλά έχει πολλαπλά προβλήματα.
- Η καλή λειτουργία του προϋποθέτει πως το σύνολο δεδομένων είναι *φυσικά διαμερισμένο* σε k ομάδες οι οποίες:
  - Έχουν χονδρικά σχήμα υπερσφαίρας στον χώρο των προτύπων.
  - Έχουν παρόμοια πυκνότητα προτύπων στο εσωτερικό τους.
  - Έχουν παρόμοιο πλήθος περιεχόμενων στοιχείων.

Προβλήματα των k-μέσων

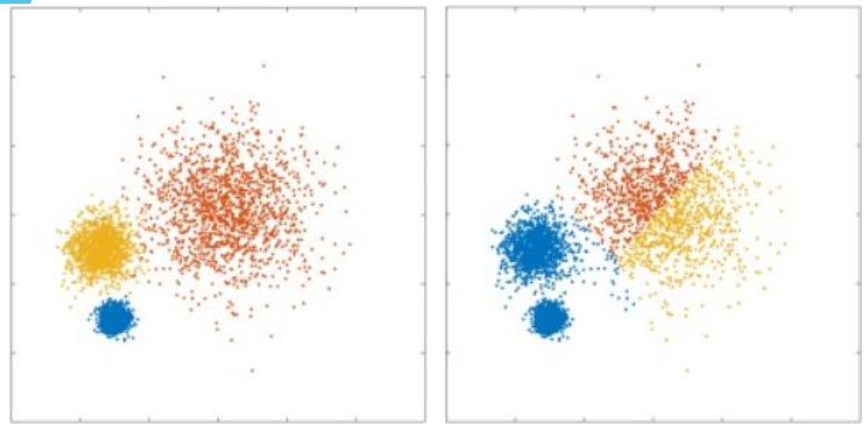
Μη σφαιρικά σχήματα ομάδων.



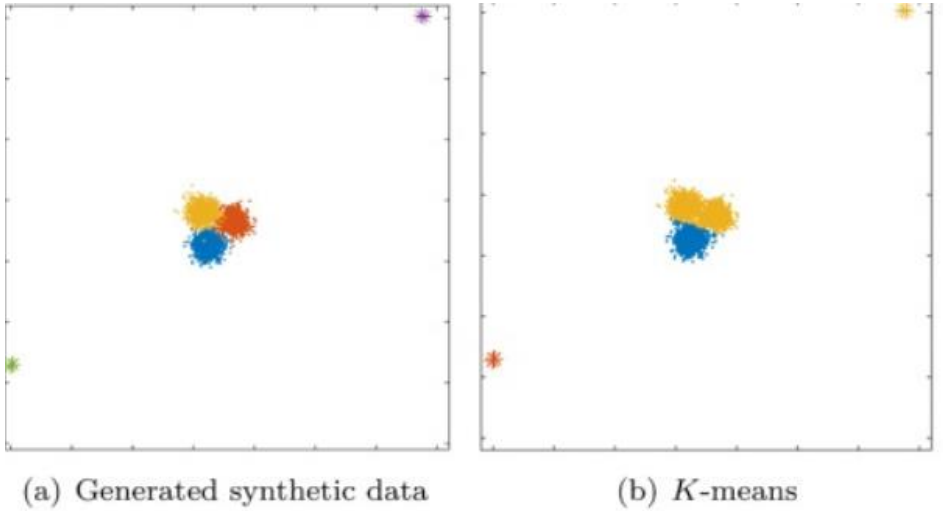
(a) Generated synthetic data (b) K-means

Πηγή: Raykov et al., "*What to Do When K-Means Clustering Fails: A Simple yet Principled Alternative Algorithm*", 2016.

Διαφορετικές πυκνότητες ανά ομάδα.



(a) Generated synthetic data (b) K-means



(a) Generated synthetic data (b) K-means

Ευάλωτος σε ανωμαλίες.

Αξιολόγηση ομαδοποίησης

- Η αξιολόγηση μίας ομαδοποίησης δεν είναι εύκολη εργασία, κυρίως για δύο λόγους:
  - Ο κάθε αλγόριθμος ουσιαστικά ορίζει το δικό του είδος ομάδων.
    - Π.χ., η αθροιστική διασπορά λειτουργεί καλά ως μέτρο αξιολόγησης μόνο σε ευκλείδειους χώρους προτύπων όπου οι ομάδες είναι κατά προσέγγιση σφαιρικές.
  - Συνήθως δεν είναι γνωστό εκ των προτέρων αν υπάρχει φυσική τάση διαμέρισης των προτύπων σε ομάδες και ποια είναι αυτή. ·

Μέτρα αξιολόγησης

- Οι περισσότερες μέθοδοι αξιολόγησης επιχειρούν να δώσουν κάποιες απαντήσεις και να παράσχουν ένα πλαίσιο σύγκρισης μεταξύ διαφορετικών ομαδοποιήσεων στα ίδια δεδομένα.

Αξιολόγηση ομαδοποίησης

Μέτρα αξιολόγησης

- Οι μέθοδοι αξιολόγησης διακρίνονται σε:
  - **Επιβλεπόμενες** ή **εξωτερικές**.
    - Είναι γνωστή μία εξωτερικά παρεχόμενη ομαδοποίηση αναφοράς και οι παραγόμενες ομάδες συγκρίνονται με αυτές τις ομάδες αναφοράς.
  - **Ανεπίβλεπτες** ή **εσωτερικές**.
    - Επιχειρούν να εκτιμήσουν πόσο *συμπαγείς* είναι οι παραχθείσες ομάδες.

- Στην ανεπίβλεπτη αξιολόγηση, η εκτίμηση του πόσο συμπαγείς είναι οι ομάδες γίνεται μετρώντας τη **συνοχή** (cohesion) και τον **διαχωρισμό** τους (separation).

Μέτρα αξιολόγησης

Εσωτερική αξιολόγηση

- Η συνοχή εκφράζει το πόσο όμοια είναι τα στοιχεία κάθε ομάδας μεταξύ τους.
- Ο διαχωρισμός εκφράζει πόσο ανόμοια είναι τα πρότυπα κάθε ομάδας με τα πρότυπα των υπολοίπων ομάδων.
- Η συνοχή και ο διαχωρισμός:
  - Υπολογίζονται με βάση κάποια μετρική εγγύτητας.
  - Μπορούν να υπολογιστούν ξεχωριστά ανά ομάδα και κατόπιν να συναθροιστούν.
- Η τιμή συνοχής, η τιμή διαχωρισμού ή κάποια σύνθεσή τους αξιοποιείται για τον υπολογισμό μίας τιμής **εγκυρότητας** ομάδας.

Εσωτερική αξιολόγηση

Μέτρα αξιολόγησης

- Ολικό **μέτρο εγκυρότητας** μιας ομαδοποίησης είναι το σταθμισμένο άθροισμα της επιμέρους εγκυρότητας όλων των παραχθέντων ομάδων.
  - Οι συντελεστές στάθμισης των ομάδων συνήθως υπολογίζονται με βάση το σχετικό μέγεθος της κάθε ομάδας, δηλαδή το ποσοστό των ολικών προτύπων τα οποία αυτή περιέχει.

- Συνήθεις ορισμοί:
  - *Συνοχή*: άθροισμα της εγγύτητας όλων των στοιχείων μίας ομάδας μεταξύ τους ανά δύο.
  - *Διαχωρισμός*: άθροισμα της εγγύτητας όλων των στοιχείων δύο διαφορετικών ομάδων μεταξύ τους ανά δύο.

Εσωτερική αξιολόγηση

Μέτρα αξιολόγησης

- Σε περίπτωση ομάδων πρωτοτύπων οι ορισμοί αυτοί τροποποιούνται:
  - *Συνοχή*: άθροισμα της εγγύτητας όλων των στοιχείων μίας ομάδας με το πρωτότυπο της.
  - *Διαχωρισμός*: εγγύτητα μεταξύ των πρωτοτύπων δύο ομάδων.
- Αν επιπροσθέτως το μέτρο εγγύτητας είναι η ευκλείδεια απόσταση, τότε:
  - Η συνοχή συμπίπτει με τη διασπορά της ομάδας.
  - Ο διαχωρισμός υπολογίζεται ως η **ενδο-ομαδική διασπορά** (SSB).
    - SSB είναι το τετράγωνο της ευκλείδειας απόστασης του κέντρου της ομάδας από το αριθμητικό μέσον **όλων** των στοιχείων του συνόλου δεδομένων.

Μέτρα αξιολόγησης

Εσωτερική αξιολόγηση

- Η **αθροιστική ενδο-ομαδική διασπορά** (ολικό SSB) είναι το άθροισμα των SSB όλων των ομάδων.

- Το άθροισμα του ολικού SSB με την αθροιστική διασπορά της ομαδοποίησης (ολικό SSE) καλείται **ολική διασπορά** (TSS).
  - Εκφράζει το άθροισμα των τετραγώνων των ευκλείδειων αποστάσεων κάθε στοιχείου από το συνολικό μέσον όλων των προτύπων.
    - Η εκάστοτε ομαδοποίηση δεν εμφανίζεται σε αυτόν τον ορισμό!
  - Το TSS παραμένει σταθερό για δοθέν σύνολο δεδομένων, ανεξαρτήτως της ομαδοποίησης.
    - Είναι σταθερή ιδιότητα του συνόλου δεδομένων, όχι της εκάστοτε ομαδοποίησης.

Εσωτερική αξιολόγηση

Μέτρα αξιολόγησης

- Ελαχιστοποίηση της αθροιστικής διασποράς (ολικού SSE) της ομαδοποίησης ισοδυναμεί με:
  - Μεγιστοποίηση της ολικής συνοχής, και
  - Μεγιστοποίηση του ολικού διαχωρισμού της ομαδοποίησης (ολικό SSB).
- Τα μέτρα εγκυρότητας τα οποία βασίζονται στη συνοχή και στον διαχωρισμό μπορούν να χρησιμοποιηθούν για να συγκρίνουμε διαφορετικές ομαδοποιήσεις και να επιλέξουμε την «καλύτερη».
  - Είναι όμως περισσότερο κατάλληλα αν ο φυσικός διαμερισμός των δεδομένων είναι σε ομάδες υπερσφαιρικού σχήματος.

Μέτρα αξιολόγησης

Εσωτερική αξιολόγηση

- Εκτός από την αξιολόγηση μίας ολόκληρης ομαδοποίησης μπορεί να γίνει αξιολόγηση μίας μεμονωμένης ομάδας.
  - Π.χ., ώστε οι μη ικανοποιητικές ομάδες να διασπαστούν ή να συγχωνευτούν.
- Για το σκοπό αυτό χρησιμοποιείται ο **συντελεστής σιλουέτας**.
  - Αρχικά υπολογίζεται ξεχωριστά για το μεμονωμένο *i*-οστό πρότυπο μίας ομάδας ως η ποσότητα $(b_i - a_i) / \max(b_i, a_i)$.
    - a_i είναι η μέση απόσταση του *i*-οστού προτύπου από τα υπόλοιπα στοιχεία της ομάδας του.
    - b_i είναι η ελάχιστη τιμή μεταξύ των μέσων αποστάσεων του *i*-οστού προτύπου από τα στοιχεία καθεμίας από τις *υπόλοιπες* ομάδες.

Μέτρα αξιολόγησης

Εσωτερική αξιολόγηση

- Ιδανικά, ο συντελεστής σιλουέτας ενός προτύπου πρέπει να είναι θετικός και το a_i του να είναι όσο το δυνατόν εγγύτερα στο 0.

- Αρνητικός συντελεστής σιλουέτας σηματοδοτεί προβληματική ομαδοποίηση.
  - Συνεπάγεται πως η μέση απόσταση του προτύπου από τα υπόλοιπα της δικής του ομάδας είναι μεγαλύτερη από τη μέση απόστασή του από τα στοιχεία κάποιας άλλης ομάδας.

- Ο συντελεστής σιλουέτας λαμβάνει τιμή στο διάστημα [-1,1] και μεγαλύτερη τιμή σημαίνει καλύτερη ομαδοποίηση.

Μέτρα αξιολόγησης

Εσωτερική αξιολόγηση

- Ο μέσος όρος των συντελεστών σιλουέτας όλων των προτύπων μίας ομάδας ονομάζεται **συντελεστής της ομάδας**.

- Ο μέσος όρος των συντελεστών σιλουέτας όλων των προτύπων του συνόλου δεδομένων ονομάζεται **συντελεστής ομαδοποίησης**.

- Έτσι μπορούμε να αξιολογήσουμε τα αποτελέσματα σε όποιο επίπεδο θέλουμε.

Thank you for your attention!

Q & A

Contact: imademlis@aueb.gr