Ομαδοποίηση δεδομένων

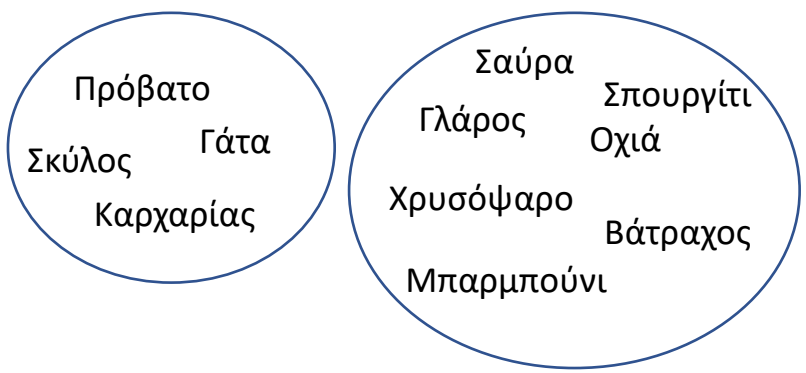
Αθανάσιος Ροντογιάννης

Αν. Καθηγητής, ΣΗΜΜΥ-ΕΜΠ

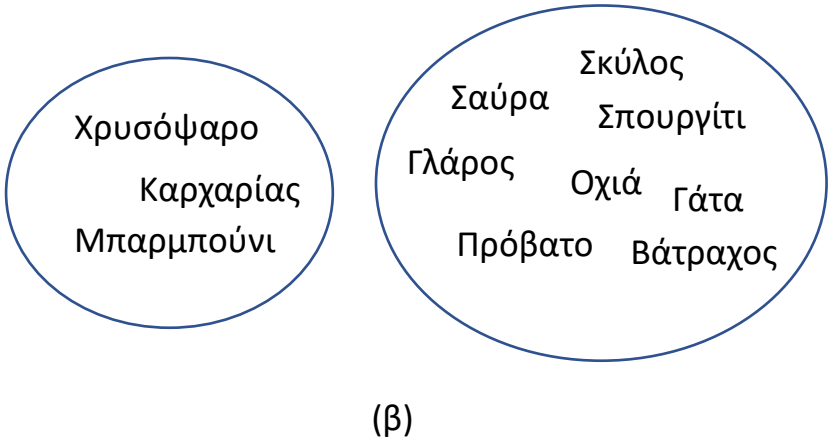
Περίγραμμα

- Ορισμός και βασικές έννοιες ομαδοποίησης
- Μέτρα εγγύτητας
- Αλγόριθμοι ομαδοποίησης
  - Βασικό ακολουθιακό αλγοριθμικό σχήμα
  - Ο αλγόριθμος k-μέσων
  - Ιεραρχικοί αλγόριθμοι ομαδοποίησης
  - Ο αλγόριθμος DBSCAN
- Υπερφασματικές εικόνες
  - Ομαδοποίηση
  - Φασματικός διαχωρισμός

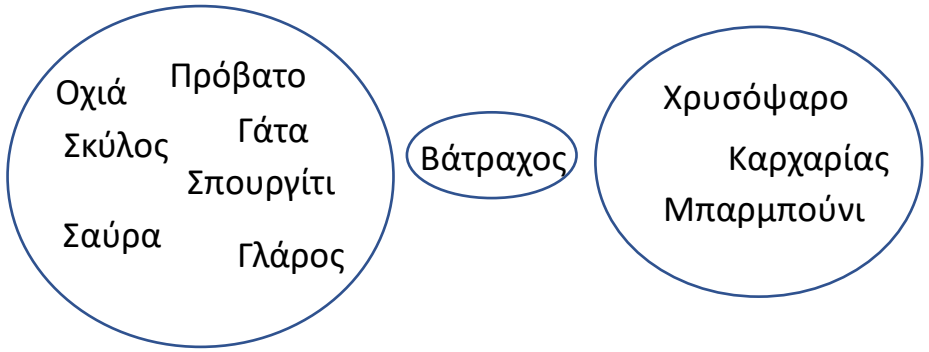
Παράδειγμα ομαδοποίησης



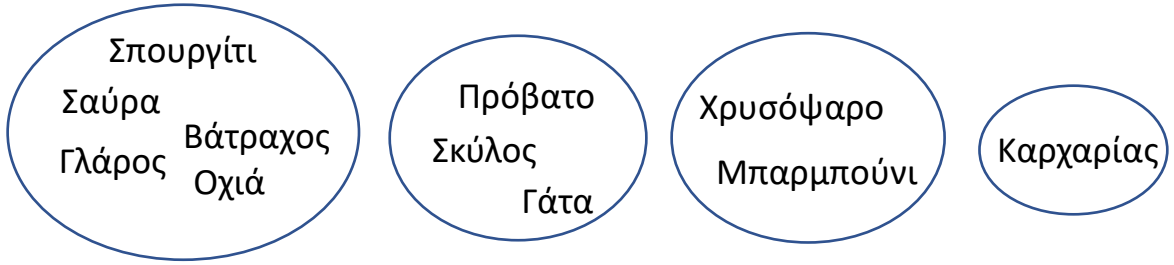
(α)



(β)



(γ)



(δ)

Ομαδοποιήσεις ζώων που προκύπτουν όταν το κριτήριο ομαδοποίησης είναι (α) ο τρόπος με τον οποίο γεννούν τους απογόνους τους, (β) η ύπαρξη πνευμόνων, (γ) το περιβάλλον μέσα στο οποίο ζουν και (δ) ο τρόπος με τον οποίο γεννούν τους απογόνους τους σε συνδυασμό με την ύπαρξη πνευμόνων

Ομαδοποίηση-Βασικές έννοιες

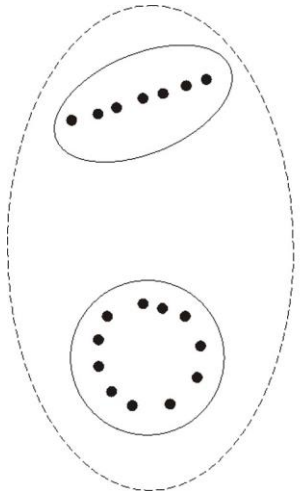
- ✓ Ομαδοποίηση ή συσταδοποίηση (clustering) είναι η διαδικασία καταχώρισης “όμοιων” οντοτήτων (προτύπων) στην ίδια ομάδα (cluster) και “ανόμοιων” οντοτήτων σε διαφορετικές ομάδες
- ✓ Είναι μια μη-επιβλεπόμενη διαδικασία μάθησης: δεν χρησιμοποιούνται δεδομένα εκπαίδευσης
- ✓ Όπως και στην περίπτωση της εκμάθησης με επίβλεψη, τα πρότυπα αναπαρίστανται με τη χρήση χαρακτηριστικών (features), τα οποία σχηματίζουν l-διάστατα διανύσματα χαρακτηριστικών (feature vectors).

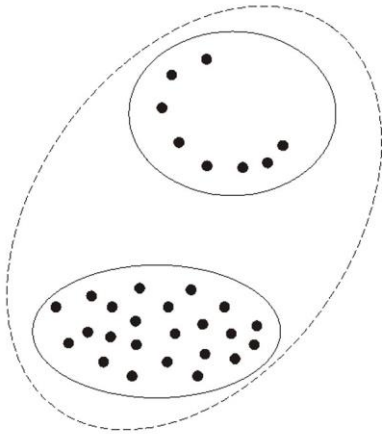
Τα βασικά στάδια της ομαδοποίησης είναι:

- Επιλογή χαρακτηριστικών
- Επιλογή μέτρου εγγύτητας (proximity measure)
- Επιλογή κριτηρίου ομαδοποίησης (clustering criterion)
- Επιλογή αλγόριθμου ομαδοποίησης (clustering algorithm)
- Επικύρωση των αποτελεσμάτων
- Ερμηνεία των αποτελεσμάτων

Ομαδοποίηση-Βασικές έννοιες

- Διαφορετικές επιλογές χαρακτηριστικών, μέτρου εγγύτητας, κριτηρίου ομαδοποίησης και αλγόριθμου μπορούν να οδηγήσουν σε εντελώς διαφορετικές ομαδοποιήσεις δεδομένων.
- Η υποκειμενικότητα είναι μια πραγματικότητα με την οποία θα πρέπει να ζήσουμε από εδώ και πέρα.



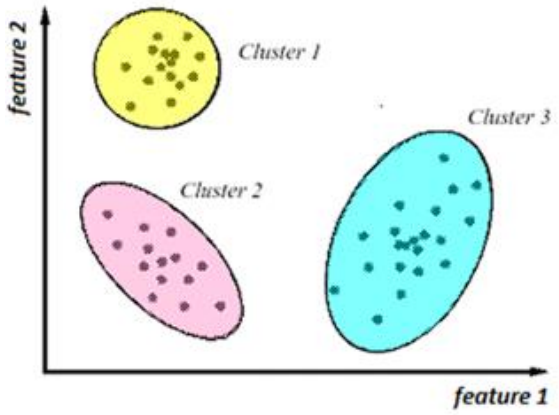


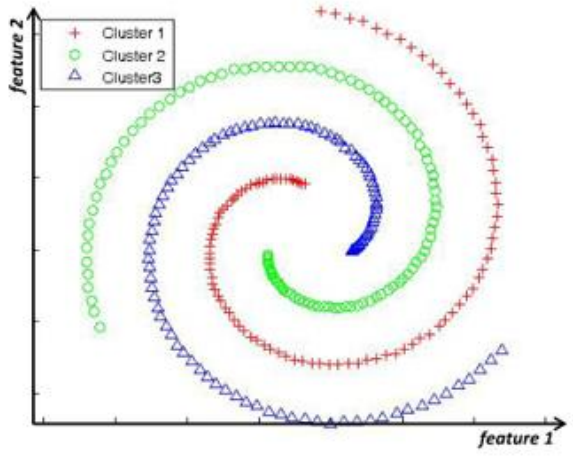
Πόσες ομάδες έχουμε εδώ: 2 ή 4;

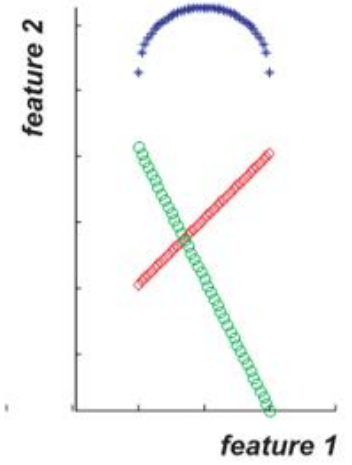
Ομαδοποίηση-Βασικές έννοιες

Παρατηρήσεις σχετικά με τον ορισμό της “ομάδας (cluster)”

- Δεν υπάρχει αυστηρός ορισμός σχετικά με την έννοια της ομάδας (μη καλά ορισμένο πρόβλημα)
- Αυτό που έχουμε συνήθως στο μυαλό μας είναι ότι η ομάδα είναι μια συγκέντρωση σημείων γύρω από:
  - Ένα συγκεκριμένο σημείο στο χώρο των χαρακτηριστικών (και συνήθως μοντελοποιείται με την Gaussian κατανομή)
  - Ένα manifold (π.χ. υπερεπίπεδο, υπερσφαίρα) στο χώρο των χαρακτηριστικών







Ορισμός ομαδοποίησης

Έστω X το υπό εξέταση σύνολο δεδομένων, δηλ.,

$$X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$$

Αυστηρή (hard) ομαδοποίηση

Ορίζουμε ως m-ομαδοποίηση $\mathfrak{R}$ του X, τον διαμερισμό του X σε m σύνολα (ομάδες) $C_1, C_2, \ldots, C_m$, έτσι ώστε να ικανοποιούνται οι ακόλουθες τρεις συνθήκες:

- $C_i \neq \emptyset, i = 1,2, \ldots, m$
- $\bigcup_{i=1}^{m} C_i = X$
- $C_i \cap C_j = \emptyset, i \neq j, i, j = 1,2, \ldots, m$

Κάθε πρότυπο ανήκει σε μία και μόνο ομάδα. Επιπλέον, τα διανύσματα που περιέχονται στην ομάδα C_i είναι πιο «όμοια» μεταξύ τους και λιγότερο «όμοια» με διανύσματα άλλων ομάδων

Ορισμός ομαδοποίησης

Ασαφής (fuzzy) ομαδοποίηση

Η ασαφής ομαδοποίηση του X σε m ομάδες χαρακτηρίζεται από m συναρτήσεις u_j με

$$u_j : X \to [0,1], \qquad j = 1,2, \dots, m$$

και

$$\sum_{j=1}^{m} u_j(\boldsymbol{x}_i) = 1, i = 1,2, \dots, N, \qquad 0 < \sum_{i=1}^{N} u_j(\boldsymbol{x}_i) < N, j = 1,2, \dots, m$$

- Οι συναρτήσεις αυτές καλούνται συναρτήσεις συμμετοχής (membership functions).
- Κάθε διάνυσμα μπορεί να ανήκει ταυτόχρονα σε περισσότερες από μία ομάδες «σε κάποιο βαθμό», ο οποίος ποσοτικοποιείται από την αντίστοιχη τιμή της u_j στο διάστημα [0,1].

Μέτρα εγγύτητας

Ένα μέτρο ανομοιότητας (dissimilarity measure), d, πάνω στο σύνολο X είναι μια συνάρτηση

$$d: X \times X \to \mathbb{R}$$

για την οποία ισχύουν τα παρακάτω:

$$\exists d_0 \in \mathbb{R}: -\infty < d_0 \leq d(\boldsymbol{x}, \boldsymbol{y}) < \infty, \qquad \forall \boldsymbol{x}, \boldsymbol{y} \in X$$

$$d(\boldsymbol{x}, \boldsymbol{x}) = d_0, \quad \forall \boldsymbol{x} \in X$$

και

$$d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{y}, \boldsymbol{x}), \quad \forall \boldsymbol{x}, \boldsymbol{y} \in X$$

Αν επιπλέον

$$d(\boldsymbol{x}, \boldsymbol{y}) = d_0 \text{ αν και μόνο αν } \boldsymbol{x} = \boldsymbol{y} \qquad \text{και}$$

$$d(\boldsymbol{x}, \boldsymbol{z}) \leq d(\boldsymbol{x}, \boldsymbol{y}) + d(\boldsymbol{y}, \boldsymbol{z}), \quad \forall \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in X$$

τότε το d ονομάζεται μετρική ανομοιότητας.

Μέτρα εγγύτητας

Ένα μέτρο ομοιότητας (similarity measure), s, πάνω στο σύνολο X είναι μια συνάρτηση

$$s: X \times X \to \mathbb{R}$$

για την οποία ισχύουν τα παρακάτω:

$$\exists s_0 \in \mathbb{R}: -\infty < s(\boldsymbol{x}, \boldsymbol{y}) \leq s_0 < \infty, \qquad \forall \boldsymbol{x}, \boldsymbol{y} \in X$$

$$s(\boldsymbol{x}, \boldsymbol{x}) = s_0, \quad \forall \boldsymbol{x} \in X$$

και

$$s(\boldsymbol{x}, \boldsymbol{y}) = s(\boldsymbol{y}, \boldsymbol{x}), \quad \forall \boldsymbol{x}, \boldsymbol{y} \in X$$

Αν επιπλέον

$$s(\boldsymbol{x}, \boldsymbol{y}) = s_0 \text{ αν και μόνο αν } \boldsymbol{x} = \boldsymbol{y} \qquad \text{και}$$

$$s(\boldsymbol{x}, \boldsymbol{y}) s(\boldsymbol{y}, \boldsymbol{z}) \leq [s(\boldsymbol{x}, \boldsymbol{y}) + s(\boldsymbol{y}, \boldsymbol{z})] s(\boldsymbol{x}, \boldsymbol{z}), \quad \forall \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in X$$

τότε το s ονομάζεται μετρική ομοιότητας.

Μέτρα εγγύτητας

Στους ιεραρχικούς αλγόριθμους ομαδοποίησης απαιτείται ο υπολογισμός αποστάσεων μεταξύ δύο συνόλων, που περιέχουν διανύσματα του X.

Έστω U ένα σύνολο που περιέχει υποσύνολα του X, δηλαδή $D_i \subset X, i = 1,2, \dots, k$ και $U = \{D_1, \dots, D_k\}$. Ένα μέτρο εγγύτητας Q πάνω στο U είναι μια συνάρτηση

$$Q: U \times U \to \mathbb{R}$$

Οι εξισώσεις για τα μέτρα ανομοιότητας και τα μέτρα ομοιότητας μεταξύ διανυσμάτων, μπορούν να επαναληφθούν εδώ, αντικαθιστώντας τα $\boldsymbol{x}, \boldsymbol{y}$ με τα D_i, D_j, αντίστοιχα, και το X με το U.

Συνήθως τα μέτρα εγγύτητας μεταξύ δύο συνόλων D_i και D_j ορίζονται βάσει μέτρων εγγύτητας μεταξύ των στοιχείων των δύο συνόλων.

Μέτρα εγγύτητας μεταξύ δύο σημείων

Μέτρα ανομοιότητας

- Οι σταθμισμένες (weighted) l_p μετρικές ανομοιότητας,

$$d_p(\boldsymbol{x}, \boldsymbol{y}) = \left(\sum_{i=1}^{l} w_i \, |x_i - y_i|^p \right)^{1/p}$$

- Η γενικευμένη σταθμισμένη l_2 μετρική ανομοιότητας,

$$d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{(\boldsymbol{x} - \boldsymbol{y})^T B (\boldsymbol{x} - \boldsymbol{y})}$$

- Η σταθμισμένη l_1ή νόρμα Manhattan

$$d_1(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{l} w_i \, |x_i - y_i|$$

- Η σταθμισμένη l_∞ νόρμα

$$d_\infty(\boldsymbol{x}, \boldsymbol{y}) = \max_{1 \le i \le l} w_i \, |x_i - y_i|$$

Μέτρα εγγύτητας μεταξύ δύο σημείων

Μέτρα ομοιότητας

- Το εσωτερικό γινόμενο (inner product), με τα $\boldsymbol{x}, \boldsymbol{y}$ κανονικοποιημένα,

$$s_{\text{inner}}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^T \boldsymbol{y} = \sum_{i=1}^{l} x_i \, y_i$$

- Το μέτρο ομοιότητας συνημιτόνου (cosine similarity measure)

$$s_{\text{cosine}}(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x}^T \boldsymbol{y}}{\|\boldsymbol{x}\| \|\boldsymbol{y}\|}, \|\boldsymbol{x}\| = \sqrt{\sum_{i=1}^{l} x_i^2}$$

- Το μέτρο κατά Tanimoto ή απόσταση Tanimoto

$$s_{\text{T}}(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x}^T \boldsymbol{y}}{\|\boldsymbol{x}\|^2 + \|\boldsymbol{y}\|^2 - \boldsymbol{x}^T \boldsymbol{y}}$$

Συναρτήσεις εγγύτητας $Q(\boldsymbol{x}, C)$ μεταξύ σημείου και συνόλου σημείων

Α. Όλα τα σημεία του C συνεισφέρουν στον υπολογισμό του $Q(\boldsymbol{x}, C)$

- Η συνάρτηση μέγιστης εγγύτητας

$$Q^{\text{ps}}_{\max}(\boldsymbol{x}, C) = \max_{\boldsymbol{y} \in C} Q(\boldsymbol{x}, \boldsymbol{y})$$

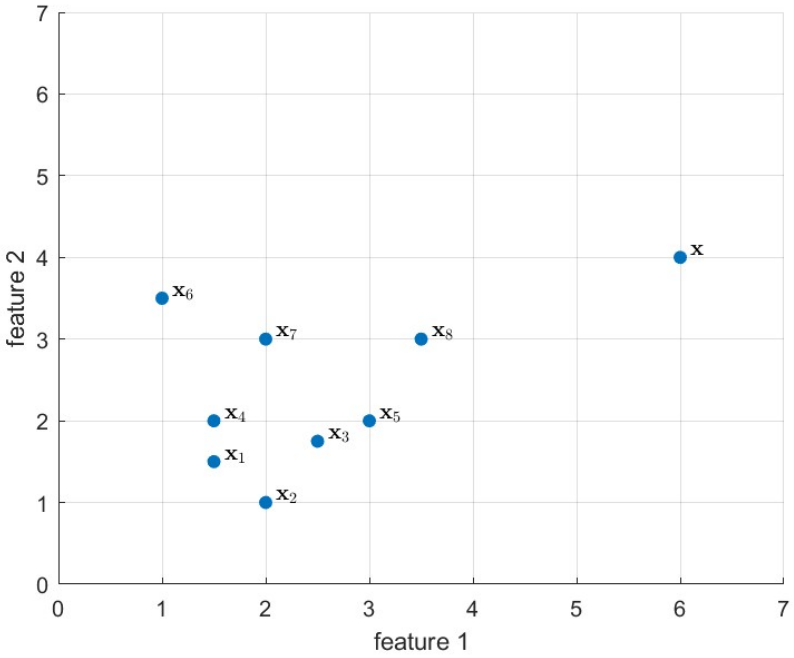
- Η συνάρτηση ελάχιστης εγγύτητας

$$Q^{\text{ps}}_{\min}(\boldsymbol{x}, C) = \min_{\boldsymbol{y} \in C} Q(\boldsymbol{x}, \boldsymbol{y})$$

- Η συνάρτηση μέσης εγγύτητας

$$Q^{\text{ps}}_{\text{avg}}(\boldsymbol{x}, C) = \frac{1}{n_C} \sum_{\boldsymbol{y} \in C} Q(\boldsymbol{x}, \boldsymbol{y})$$

όπου n_C είναι ο πληθάριθμος του C.



$$Q^{\text{ps}}_{\max}(\boldsymbol{x}, C) = d(\boldsymbol{x}, \boldsymbol{x}_1) = 5.15$$
$$Q^{\text{ps}}_{\min}(\boldsymbol{x}, C) = d(\boldsymbol{x}, \boldsymbol{x}_8) = 2.69$$
$$Q^{\text{ps}}_{\text{avg}}(\boldsymbol{x}, C) = 4.33$$

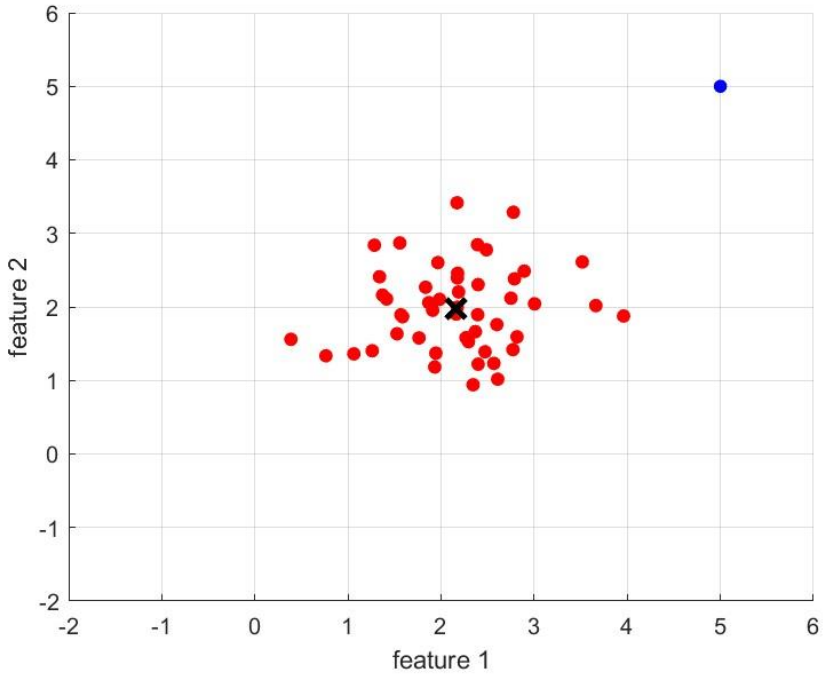
Συναρτήσεις εγγύτητας $Q(\boldsymbol{x}, C)$ μεταξύ σημείου και συνόλου σημείων – Αντιπρόσωπος ομάδας

Β. Το C εφοδιάζεται με έναν αντιπρόσωπο και η εγγύτητα μεταξύ των $\boldsymbol{x}$ και C ορίζεται ως η εγγύτητα μεταξύ του $\boldsymbol{x}$ και του αντιπροσώπου του C.

Για συμπαγείς ομάδες, ο αντιπρόσωπος είναι συνήθως το μέσο διάνυσμα (mean vector) ή μέσο σημείο (mean point)

$$m_p = \frac{1}{n_C} \sum_{\boldsymbol{y} \in C} \boldsymbol{y}$$

Αυτή είναι η πιο συχνά χρησιμοποιούμενη επιλογή, όταν χρησιμοποιούνται σημειακοί αντιπρόσωποι και μελετώνται δεδομένα σε συνεχή χώρο.



Συναρτήσεις εγγύτητας μεταξύ δύο συνόλων

- Η συνάρτηση μέγιστης εγγύτητας

$$\mathcal{Q}^{\mathrm{ss}}_{\max}(D_i, D_j) = \max_{\boldsymbol{x}\in D_i, \boldsymbol{y}\in D_j} \mathcal{Q}(\boldsymbol{x}, \boldsymbol{y})$$

- Η συνάρτηση ελάχιστης εγγύτητας

$$\mathcal{Q}^{\mathrm{ss}}_{\min}(D_i, D_j) = \min_{\boldsymbol{x}\in D_i, \boldsymbol{y}\in D_j} \mathcal{Q}(\boldsymbol{x}, \boldsymbol{y})$$

- Η συνάρτηση μέσης εγγύτητας

$$\mathcal{Q}^{\mathrm{ss}}_{\mathrm{avg}}(D_i, D_j) = \frac{1}{n_{D_i} n_{D_j}} \sum_{\boldsymbol{x}\in D_i} \sum_{\boldsymbol{y}\in D_j} \mathcal{Q}(\boldsymbol{x}, \boldsymbol{y})$$

- Η συνάρτηση εγγύτητας μεταξύ αντιπροσώπων

$$\mathcal{Q}^{\mathrm{ss}}_{\mathrm{mean}}(D_i, D_j) = d(\boldsymbol{m}_{D_i}, \boldsymbol{m}_{D_j})$$

όπου $\boldsymbol{m}_{D_i}, \boldsymbol{m}_{D_j}$ είναι οι σημειακοί αντιπρόσωποι των D_i, D_j.

Αλγόριθμοι ομαδοποίησης

- Ο καλύτερος τρόπος καταχώρισης των διανυσμάτων χαρακτηριστικών $\boldsymbol{x}_i, i = 1,2, \dots, N$ του X σε ομάδες, θα ήταν ο προσδιορισμός όλων των δυνατών διαμερίσεων (ομαδοποιήσεων) του X και η επιλογή της πιο λογικής από αυτές, σύμφωνα με ένα προεπιλεγμένο κριτήριο.
- Αν $S(N, m)$ είναι ο αριθμός όλων των δυνατών ομαδοποιήσεων N διανυσμάτων σε m ομάδες, προκύπτει εύκολα η αναδρομική σχέση:

$$S(N, m) = mS(N-1, m) + S(N-1, m-1)$$

- Οι λύσεις της αναδρομικής σχέσης είναι οι λεγόμενοι αριθμοί Stirling δεύτερης κατηγορίας (Stirling numbers of the second kind)

$$S(N, m) = \frac{1}{m!} \sum_{i=1}^{m} (-1)^{m-i} \binom{m}{i} i^N$$

- Για παράδειγμα, $S(N, 2) = 2^{N-1} - 1, S(15,3) = 2375101, S(20,4) = 45232115901$, …
- Άρα η μέθοδος αυτή δεν μπορεί να εφαρμοστεί στην πράξη, ακόμα και για μέτριες τιμές του N.

Κατηγορίες αλγόριθμων ομαδοποίησης

- Ακολουθιακοί αλγόριθμοι (sequential algorithms)
  - Βασικό ακολουθιακό αλγοριθμικό σχήμα
  - Παραλλαγές του βασικού σχήματος
- Αλγόριθμοι που βασίζονται στη βελτιστοποίηση συνάρτησης κόστους
  - Αυστηροί αλγόριθμοι ομαδοποίησης (isodata ή k-means)
  - Πιθανοτικοί αλγόριθμοι ομαδοποίησης (probabilistic clustering algorithms)
  - Αλγόριθμοι ομαδοποίησης ασαφούς λογικής (fuzzy clustering algorithms)
  - Αλγόριθμοι ομαδοποίησης στη βάση των ενδεχομένων (possibilistic clustering algorithms)
  - Αλγόριθμοι ανίχνευσης ορίων (boundary detection algorithms)
- Ιεραρχικοί αλγόριθμοι ομαδοποίησης (hierarchical clustering algorithms)
  - Συσσωρευτικοί αλγόριθμοι (agglomerative algorithms)
  - Διαιρετικοί αλγόριθμοι (divisive algorithms)
- Αλγόριθμοι με βάση την πυκνότητα (density-based algorithms)
  - Ο αλγόριθμος DBSCAN
  - Παραλλαγές του DBSCAN

Βασικό ακολουθιακό αλγοριθμικό σχήμα

- Τα διανύσματα παρουσιάζονται στον αλγόριθμο μόνο μια φορά.
- Ο αριθμός των ομάδων δεν είναι γνωστός εκ των προτέρων.
- Έστω $d(\boldsymbol{x}, C)$ η απόσταση (ή ανομοιότητα) μεταξύ ενός διανύσματος $\boldsymbol{x}$ και μιας ομάδας C. Αυτό μπορεί να οριστεί λαμβάνοντας υπόψη, είτε όλα τα στοιχεία του C, είτε ενός αντιπροσώπου του.
- Οι παράμετροι που ορίζονται από το χρήστη είναι, το κατώφλι ανομοιότητας, Θ, και ο μέγιστος επιτρεπτός αριθμός ομάδων, q.
- Βασική ιδέα του αλγόριθμου: κάθε νέο διάνυσμα που εξετάζεται από τον αλγόριθμο καταχωρείται, είτε σε μία από τις υπάρχουσες ομάδες, είτε σε μια νέα ομάδα που δημιουργείται, ανάλογα με την απόστασή του από τις ήδη υπάρχουσες ομάδες.

Βασικό ακολουθιακό αλγοριθμικό σχήμα

(Basic Sequential Algorithmic Scheme – BSAS)

- $m = 1$
- $C_m = \{\boldsymbol{x}_1\}$
- Για $i = 2$ έως N
  - Βρες το C_k: $d(\boldsymbol{x}_i, C_k) = \min_{1 \le j \le m} d(\boldsymbol{x}_i, C_j)$
  - Αν $(d(\boldsymbol{x}_i, C_k) > \Theta)$ ΚΑΙ $(m < q)$ τότε
    - $m = m + 1$
    - $C_m = \{\boldsymbol{x}_i\}$
  - Διαφορετικά
    - $C_k = C_k \cup \{\boldsymbol{x}_i\}$
    - Όπου είναι απαραίτητο ενημέρωσε τους αντιπροσώπους
  - Τέλος {Αν}
- Τέλος {Για}

Βασικό ακολουθιακό αλγοριθμικό σχήμα

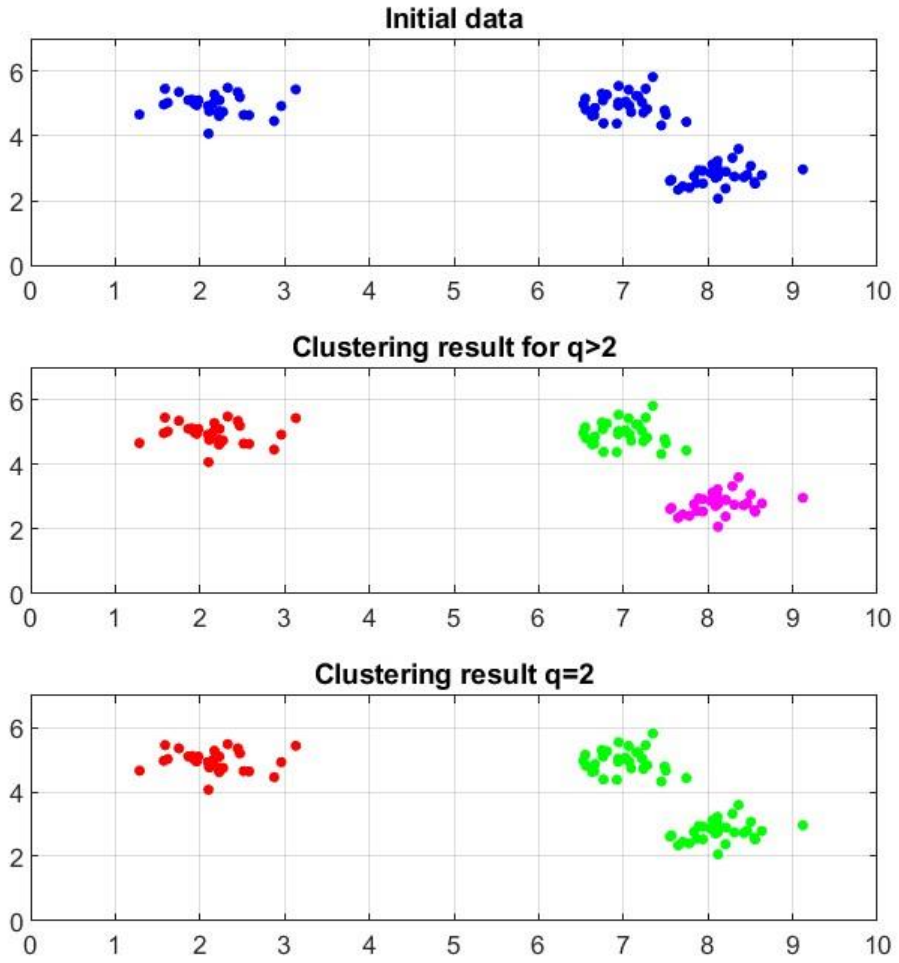
- Αν η ομάδα C αντιπροσωπεύεται από ένα διάνυσμα $\boldsymbol{m}_C$, τότε $d(\boldsymbol{x}, C) = d(\boldsymbol{x}, \boldsymbol{m}_C)$. Αν ο αντιπρόσωπος είναι το μέσο διάνυσμα, τότε αυτό μπορεί να ενημερωθεί ως εξής

$$\boldsymbol{m}_{C_k}^{new} = \frac{\left(n_{C_k}^{new} - 1\right)\boldsymbol{m}_{C_k}^{old} + \boldsymbol{x}}{n_{C_k}^{new}}$$

- Η διάταξη με την οποία τα διανύσματα παρουσιάζονται στον BSAS επηρεάζει σημαντικά τα αποτελέσματα της ομαδοποίησης
- Τα αποτελέσματα του αλγόριθμου επηρεάζονται σημαντικά από την επιλογή της τιμής του κατωφλίου Θ
  - Αν η τιμή του Θ είναι μικρή, θα δημιουργηθούν ομάδες που δεν δικαιολογούνται από τη δομή των δεδομένων
  - Αν η τιμή του Θ είναι μεγάλη, θα δημιουργηθεί μικρότερος αριθμός ομάδων, από αυτόν που απαιτείται για να περιγραφεί η δομή ομαδοποίησης των δεδομένων

Βασικό ακολουθιακό αλγοριθμικό σχήμα

- Αν ο μέγιστος επιτρεπτός αριθμός ομάδων, q, δεν ορίζεται, αφήνουμε τον αλγόριθμο να «αποφασίσει» για τον κατάλληλο αριθμό ομάδων
- Ο BSAS αποτυγχάνει για μη καλά διαχωρισμένες ομάδες
- Όταν κάθε ομάδα αντιπροσωπεύεται από ένα διάνυσμα, ο BSAS ευνοεί τη δημιουργία συμπαγών ομάδων
- Η χρονική πολυπλοκότητα του BSAS είναι $O(N)$



Βασικό ακολουθιακό αλγοριθμικό σχήμα

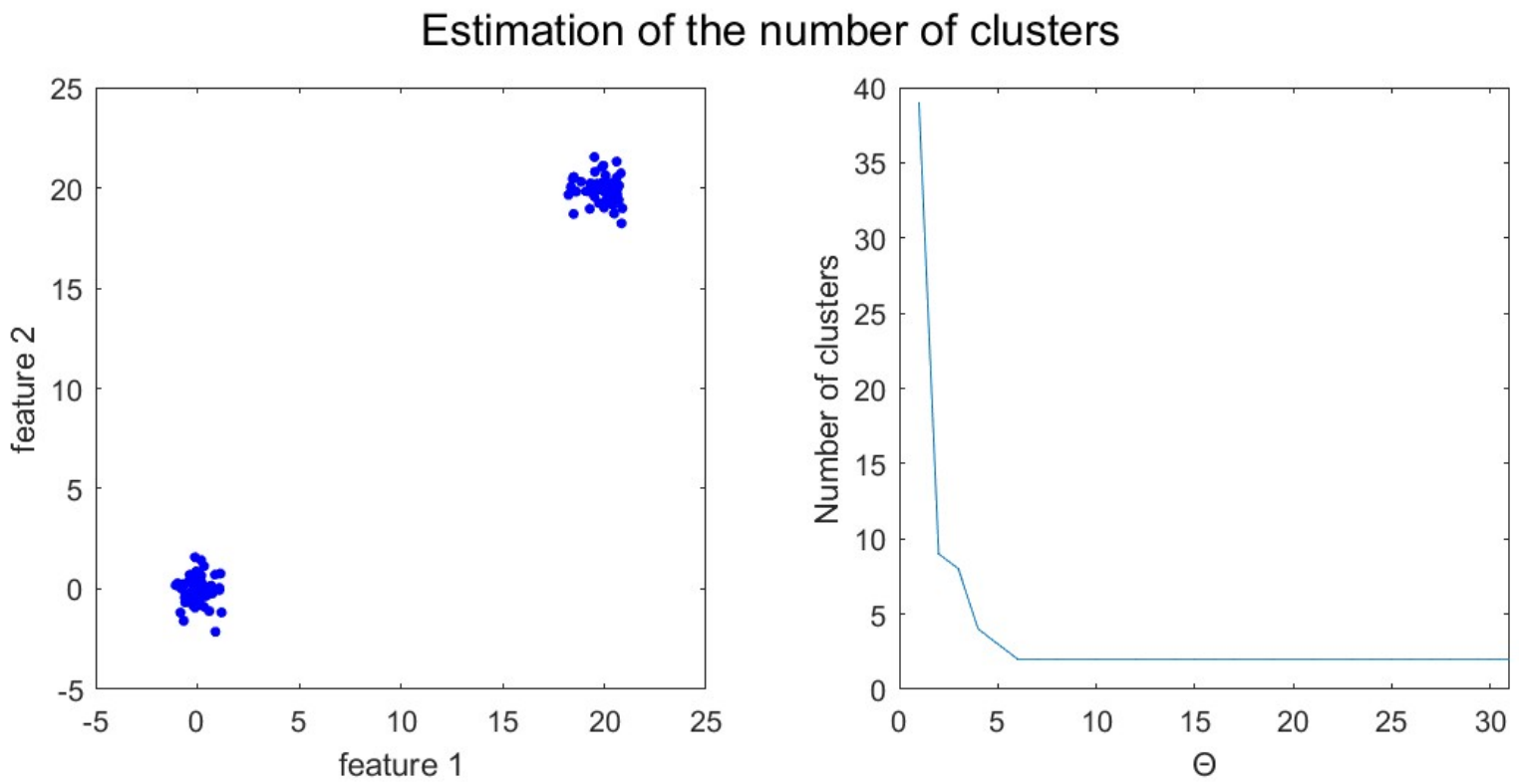
Εκτίμηση του αριθμού των ομάδων

- Για $\Theta = \alpha$ έως b με βήμα c
  - Εκτέλεσε s φορές τον αλγόριθμο BSAS(Θ), παρουσιάζοντας τα δεδομένα με διαφορετική διάταξη κάθε φορά.
  - Χρησιμοποίησε ως εκτίμηση του αριθμού των ομάδων, m_Θ, τον συχνότερα απαντώμενο αριθμό ανάμεσα σε αυτούς που προέκυψαν από τις s εκτελέσεις του BSAS(Θ).
- Τέλος {Για}

➢ Τα α και b είναι αντίστοιχα η ελάχιστη και μέγιστη τιμή ανομοιότητας ανάμεσα σε όλα τα ζεύγη διανυσμάτων του X, δηλ., $\alpha = \min_{i,j=1,2,\ldots,N} d(\boldsymbol{x}_i, \boldsymbol{x}_j)$ και $b = \max_{i,j=1,2,\ldots,N} d(\boldsymbol{x}_i, \boldsymbol{x}_j)$.

➢ Σχηματίζουμε το γράφημα του αριθμού των ομάδων ως προς Θ και εκτιμούμε τον αριθμό των ομάδων ως τον αριθμό που αντιστοιχεί στην επίπεδη περιοχή με το μεγαλύτερο εύρος.

Βασικό ακολουθιακό αλγοριθμικό σχήμα



Εκλέπτυνση της ομαδοποίησης

Διαδικασία συγχώνευσης

Αν στην ομαδοποίηση που προκύπτει μετά το τέλος του αλγόριθμου υπάρχουν ομάδες που είναι πολύ κοντά μεταξύ τους, αυτές είναι λογικό να αποτελούν μια ενιαία ομάδα.

- (Α) Βρες τις ομάδες $C_i, C_j (i < j)$ έτσι ώστε $d(C_i, C_j) = \min_{k,r=1,2,\dots,N} d(C_k, C_r)$
- Αν $d(C_i, C_j) \leq M_1$ τότε
  - Συγχώνευσε τα C_i, C_j στο C_i και εξάλειψε το C_j
  - Ενημέρωσε τον αντιπρόσωπο της ομάδας C_i (αν χρησιμοποιούνται αντιπρόσωποι ομάδων για τον καθορισμό της απόστασης μεταξύ δύο ομάδων).
  - Μετονόμασε τις ομάδες $C_{j+1}, \dots, C_m$ σε $C_j, \dots, C_{m-1}$
  - $m = m - 1$
  - Πήγαινε στο βήμα (Α)
- Διαφορετικά
  - Σταμάτα
- Τέλος {Αν}

Εκλέπτυνση της ομαδοποίησης

Διαδικασία επανεκχώρησης

Το μειονέκτημα των ακολουθιακών αλγόριθμων είναι η ευαισθησία τους στη διάταξη με την οποία παρουσιάζονται τα διανύσματα στον αλγόριθμο. Έτσι, μετά το τέλος του αλγόριθμου μπορεί να πραγματοποιηθεί η παρακάτω διαδικασία επανεκχώρησης διανυσμάτων σε ομάδες.

- Για $i = 1$ έως N
  - Βρες την C_j έτσι ώστε $d(\boldsymbol{x}_i, C_j) = \min_{k=1,\dots,m} d(\boldsymbol{x}_i, C_k)$
  - Θέσε $b(i) = j$
- Τέλος {Για}
- Για $j = 1$ έως m
  - Θέσε $C_j = \{\boldsymbol{x}_i \in X : b(i) = j\}$
  - Ενημέρωσε τους αντιπροσώπους των ομάδων
- Τέλος {Για}

Αλγόριθμοι ομαδοποίησης που βασίζονται στη βελτιστοποίηση συνάρτησης

- Οι τεχνικές αυτές βασίζονται στη βελτιστοποίηση μιας συνάρτησης κόστους J χρησιμοποιώντας τεχνικές διαφορικού λογισμού.
- Το κόστος, J, είναι συνάρτηση των διανυσμάτων του συνόλου δεδομένων, X, και είναι παραμετροποιημένη ως προς ένα άγνωστο διάνυσμα παραμέτρων, $\boldsymbol{\theta}$.
- Στα αλγοριθμικά αυτά σχήματα, ο αριθμός των ομάδων, m, θεωρείται γνωστός.
- Ο στόχος εδώ είναι η εκτίμηση του $\boldsymbol{\theta}$, έτσι ώστε να χαρακτηρίζει κατά τον καλύτερο τρόπο τις ομάδες που σχηματίζουν τα διανύσματα του X.
- Για συμπαγείς ομάδες, είναι λογικό να υιοθετούμε ως παραμέτρους ένα σύνολο m σημείων, $\boldsymbol{\theta}_i$, του l-διάστατου χώρου, καθένα από τα οποία αντιστοιχεί σε μία ομάδα και αποτελεί τον αντιπρόσωπό της. Έτσι θα είναι:

$$\boldsymbol{\theta} = [\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T, \ldots, \boldsymbol{\theta}_m^T]^T$$

Ο αλγόριθμος ομαδοποίησης k-μέσων

- Ο αλγόριθμος k-μέσων (k-means ή isodata ή c-means) είναι ένας αυστηρός (hard) αλγόριθμος ομαδοποίησης που βασίζεται στην βελτιστοποίηση συνάρτησης κόστους.
- Υποθέτουμε ότι ο αριθμός των ομάδων (m) είναι γνωστός και θεωρούμε ότι κάθε διάνυσμα $\boldsymbol{x}_i$ ανήκει σε μια και μόνο ομάδα j.
- Αυτό σημαίνει ότι για τις συναρτήσεις συμμετοχής u_{ij} θα ισχύει:

$$u_{ij} \in \{0,1\}\ \forall i = 1,2,\dots,N, j = 1,2,\dots,m$$

και

$$\sum_{j=1}^{m} u_{ij} = 1, \forall i = 1,2,\dots,N$$

- Μπορούμε να συγκεντρώσουμε τα u_{ij} σε έναν $N \times m$ πίνακα, τον οποίο συμβολίζουμε με U. Τα στοιχεία του πίνακα αυτού είναι 0 ή 1.

Ο αλγόριθμος ομαδοποίησης k-μέσων

- Ο αλγόριθμος k-means προκύπτει από την ελαχιστοποίηση της ακόλουθης συνάρτησης κόστους

$$J(\boldsymbol{\theta}, U) = \sum_{i=1}^{N} \sum_{j=1}^{m} u_{ij} \|\boldsymbol{x}_i - \boldsymbol{\theta}_j\|^2$$

  με αγνώστους τα $\boldsymbol{\theta}, U$ (και γνωστά τα $\boldsymbol{x}_i, i = 1,2, \dots, N$). Το πλήθος των αγνώστων είναι $lm + Nm$.

- Η συνάρτηση αυτή δεν μπορεί να ελαχιστοποιηθεί ως προς τα $\boldsymbol{\theta}, U$ από κοινού, με την κλασσική προσέγγιση. Για την ελαχιστοποίηση της $J(\boldsymbol{\theta}, U)$ χρησιμοποιείται η μέθοδος της εναλλασσόμενης ελαχιστοποίησης (alternating minimization method).
- Είναι μια επαναληπτική μέθοδος σε κάθε βήμα της οποίας η $J(\boldsymbol{\theta}, U)$ ελαχιστοποιείται αρχικά ως προς το U θεωρώντας το $\boldsymbol{\theta}$ σταθερό και στη συνέχεια ελαχιστοποιείται ως προς το $\boldsymbol{\theta}$ θεωρώντας το U σταθερό.

Ο αλγόριθμος ομαδοποίησης k-μέσων

Ελαχιστοποίηση της $J(\boldsymbol{\theta}, U)$ ως προς $\boldsymbol{U}$ (Προσδιορισμός διαμέρισης)

Ας κρατήσουμε σταθερά τα $\boldsymbol{\theta}_j, j = 1,2, \dots, m$. Αφού για κάθε διάνυσμα $\boldsymbol{x}_i$, μόνο ένα u_{ij} ισούται με 1 και όλα τα υπόλοιπα ισούνται με 0, είναι εύκολο να δει κανείς ότι η $J(\boldsymbol{\theta}, U)$ ελαχιστοποιείται αν καταχωρήσουμε κάθε $\boldsymbol{x}_i$ στην εγγύτερή του ομάδα, δηλαδή

$$u_{ij} = \begin{cases} 1, & \alpha\nu\ d(\boldsymbol{x}_i, \boldsymbol{\theta}_j) = \min_{k=1,\dots,m} d(\boldsymbol{x}_i, \boldsymbol{\theta}_j) \\ 0, & \delta\iota\alpha\varphi o\rho\varepsilon\tau\iota\kappa\alpha \end{cases}$$

Ελαχιστοποίηση της $J(\boldsymbol{\theta}, U)$ ως προς $\boldsymbol{\theta}$ (Προσδιορισμός παραμέτρων)

Αν κρατήσουμε σταθερό το U, μπορεί να δειχτεί ότι η ελαχιστοποίηση της $J(\boldsymbol{\theta}, U)$ ως προς τα $\boldsymbol{\theta}_j$ οδηγεί στην ακόλουθη σχέση

$$\boldsymbol{\theta}_j = \frac{1}{n_j} \sum_{i\ \tau.\omega.\ u_{ij}=1} \boldsymbol{x}_i, j = 1,2, \dots, m$$

όπου n_j είναι ο αριθμός των διανυσμάτων $\boldsymbol{x}_i$ για τα οποία ισχύει $u_{ij} = 1$.

Ο αλγόριθμος ομαδοποίησης k-μέσων

k-means algorithm

- Επέλεξε αυθαίρετες αρχικές εκτιμήσεις $\boldsymbol{\theta}_j(0)$ για τα $\boldsymbol{\theta}_j, j = 1,2, \dots, m$.
- Επανέλαβε
  - Για $i = 1$ έως N
    - Προσδιορισμός διαμέρισης: Προσδιόρισε τον κοντινότερο αντιπρόσωπο, έστω $\boldsymbol{\theta}_j$, για το $\boldsymbol{x}_i$
    - Θέσε $b(i) = j$
  - Τέλος {Για}
  - Για $j = 1$ έως m
    - Ενημέρωση παραμέτρων: Προσδιόρισε το $\boldsymbol{\theta}_j$ ως το μέσο διάνυσμα των διανυσμάτων $\boldsymbol{x}_i \in X$ με $b(i) = j$.
  - Τέλος {Για}
- Έως ότου τα $\boldsymbol{\theta}_j$ παραμείνουν αμετάβλητα για δύο συνεχόμενες επαναλήψεις

Ο αλγόριθμος ομαδοποίησης k-μέσων

Αρχικοποίηση του αλγόριθμου

Η αρχικοποίηση των $\boldsymbol{\theta}_j$ μπορεί να πραγματοποιηθεί με διάφορους τρόπους, π.χ.,

- Μπορούμε να επιλέξουμε τυχαία m από τα N διανύσματα $\boldsymbol{x}_i$
- Μπορούμε να διαμερίσουμε με τυχαίο τρόπο τα N διανύσματα σε m ομάδες και να υπολογίσουμε τα κέντρα τους, με τα οποία θα αρχικοποιήσουμε τον αλγόριθμο ή
- Μπορούμε να χρησιμοποιήσουμε το αποτέλεσμα του BSAS για αρχικοποίηση

Κριτήριο τερματισμού του αλγόριθμου

Σε κάθε επανάληψη t του αλγόριθμου υπολογίζεται η ποσότητα $\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}(t-1)\|$, και ο αλγόριθμος τερματίζει όταν

$$\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}(t-1)\| < \varepsilon$$

όπου ε είναι μια πολύ μικρή σταθερά.

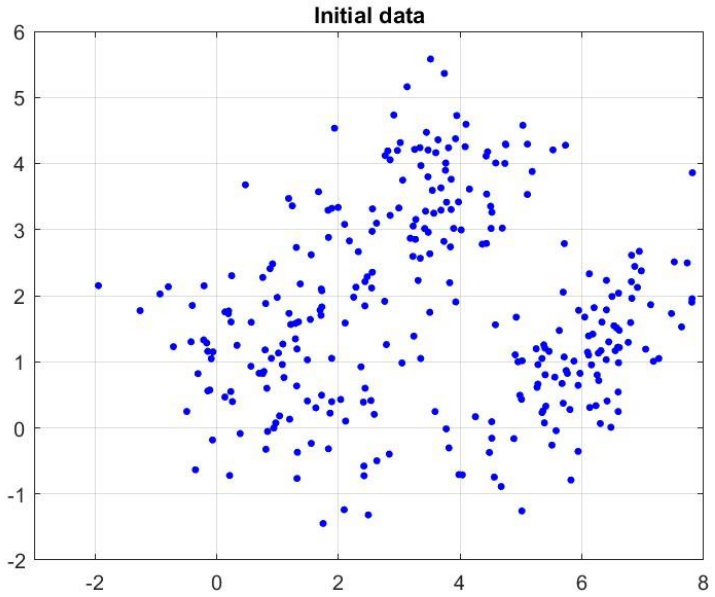
Ο αλγόριθμος ομαδοποίησης k-μέσων

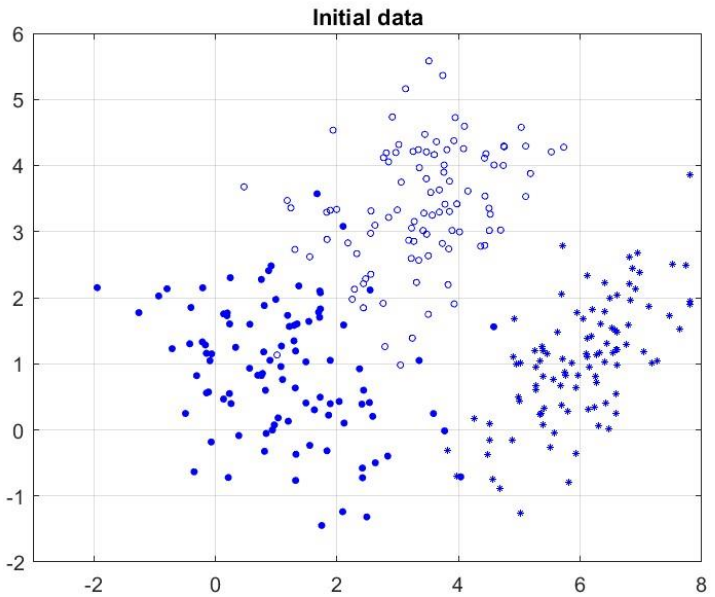
Παρατηρήσεις σχετικά με τον k-means

- Όπως συμβαίνει με όλους τους αλγόριθμους που χρησιμοποιούν σημειακούς αντιπροσώπους, ο k-means είναι κατάλληλος για την ανάδειξη συμπαγών ομάδων.
- Αποδεικνύεται ότι ο αλγόριθμος συγκλίνει σε ένα τοπικό ελάχιστο της συνάρτησης κόστους. Διαφορετικές αρχικοποιήσεις του αλγόριθμου οδηγούν σε διαφορετικά τοπικά ελάχιστα, αλλά κανείς δεν μπορεί να εγγυηθεί σύγκλιση του αλγόριθμου στο ολικό ελάχιστο της $J(\boldsymbol{\theta}, U)$.
- Ένα πλεονέκτημα του αλγόριθμου είναι η χαμηλή υπολογιστική του πολυπλοκότητα που είναι $O(mN)$ για κάθε επανάληψη του αλγόριθμου. Αυτό καθιστά τον αλγόριθμο επιλέξιμο για την επεξεργασία συνόλων δεδομένων μεγάλου μεγέθους.
- Μειονέκτημα του αλγόριθμου είναι ότι «απαιτεί» τη γνώση του πραγματικού αριθμού των ομάδων m. Κακή εκτίμηση του m θα εμποδίσει τον αλγόριθμο να αναδείξει την πραγματική δομή των ομάδων που σχηματίζουν τα σημεία του X.
- Αλγόριθμος k-means είναι ευαίσθητος σε ακραία σημεία (outliers) και θόρυβο (noise). Τα ακραία σημεία, από τη στιγμή που ανήκουν στο X, καταχωρούνται υποχρεωτικά σε μία από τις ομάδες. Έτσι επηρεάζουν τα αντίστοιχα τα μέσα διανύσματα και, συνεπώς, την τελική ομαδοποίηση.

Ο αλγόριθμος ομαδοποίησης k-μέσων

Πείραμα με 3 κλάσεις – 100 σημεία από κάθε κλάση (Gaussian κατανομές)



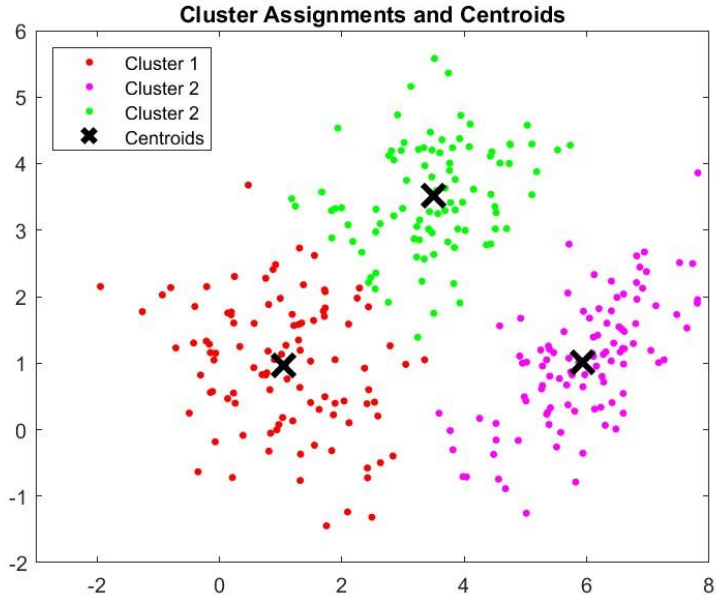


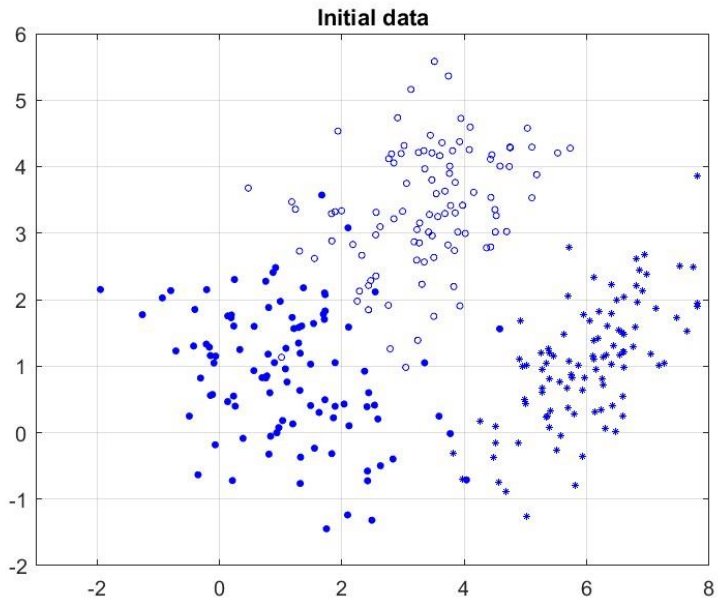
$$\boldsymbol{\mu}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \boldsymbol{\mu}_2 = \begin{bmatrix} 3.5 \\ 3.5 \end{bmatrix}, \boldsymbol{\mu}_3 = \begin{bmatrix} 6 \\ 1 \end{bmatrix}$$

$$\Sigma_1 = \begin{bmatrix} 1 & -0.3 \\ -0.3 & 1 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} 1 & 0.3 \\ 0.3 & 1 \end{bmatrix}, \Sigma_3 = \begin{bmatrix} 1 & 0.7 \\ 0.7 & 1 \end{bmatrix}$$

Ο αλγόριθμος ομαδοποίησης k-μέσων

Αποτέλεσμα ομαδοποίησης





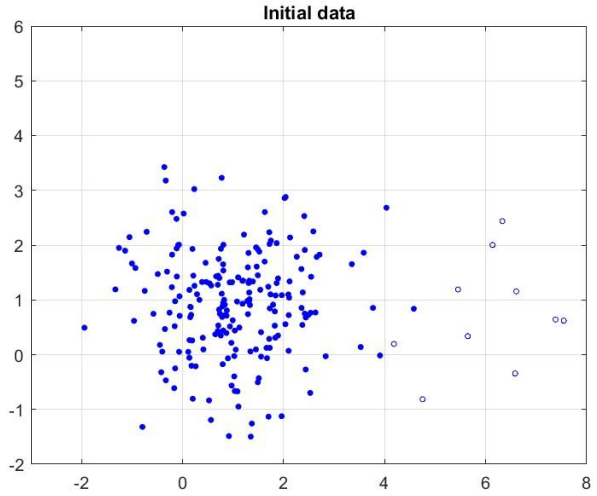
$$\boldsymbol{\theta}_1 = \begin{bmatrix} 1.06 \\ 0.96 \end{bmatrix}, \boldsymbol{\theta}_2 = \begin{bmatrix} 5.94 \\ 1.01 \end{bmatrix}, \boldsymbol{\theta}_3 = \begin{bmatrix} 3.49 \\ 3.52 \end{bmatrix}$$

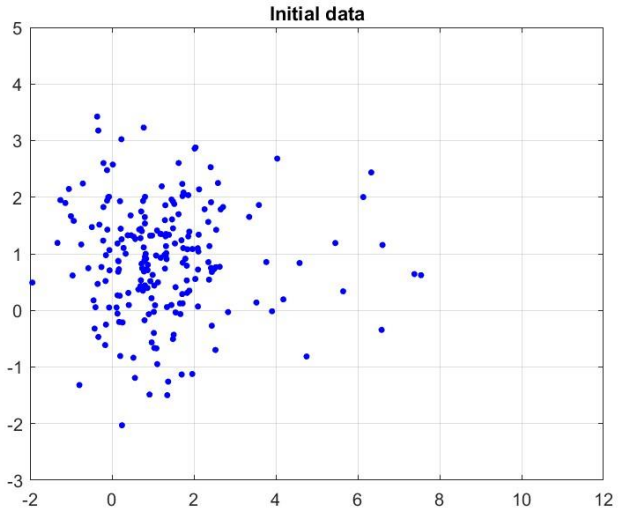
$$\mu_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \mu_2 = \begin{bmatrix} 3.5 \\ 3.5 \end{bmatrix}, \mu_3 = \begin{bmatrix} 6 \\ 1 \end{bmatrix}$$

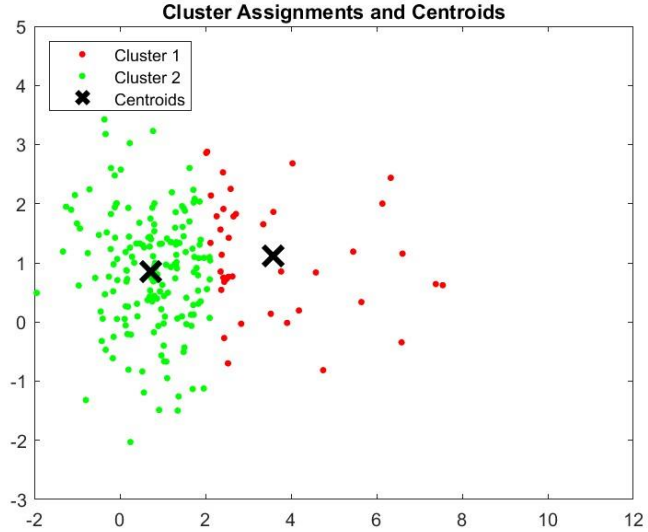
Confusion Matrix: $A = \begin{bmatrix} 93 & 4 & 3 \\ 0 & 100 & 0 \\ 9 & 0 & 91 \end{bmatrix}$, Success Rate: $\mathrm{SR} = \frac{284}{300} \cong 0.95$

Ο αλγόριθμος ομαδοποίησης k-μέσων

Πείραμα με 2 κλάσεις – 200 σημεία από τη μια και 10 σημεία από την άλλη (Gaussian κατανομές)





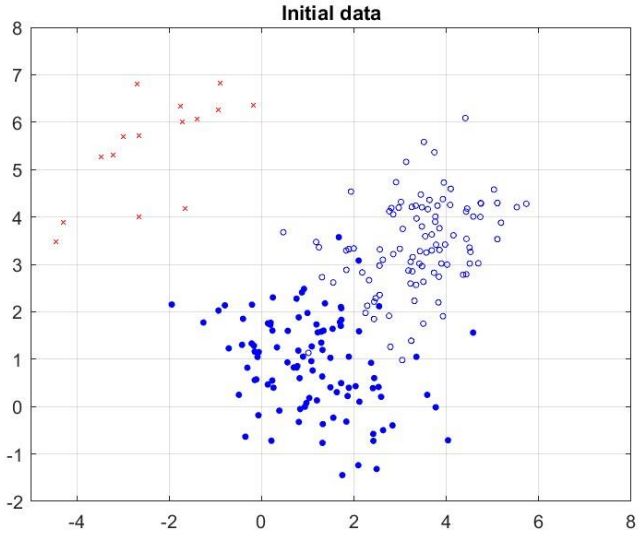


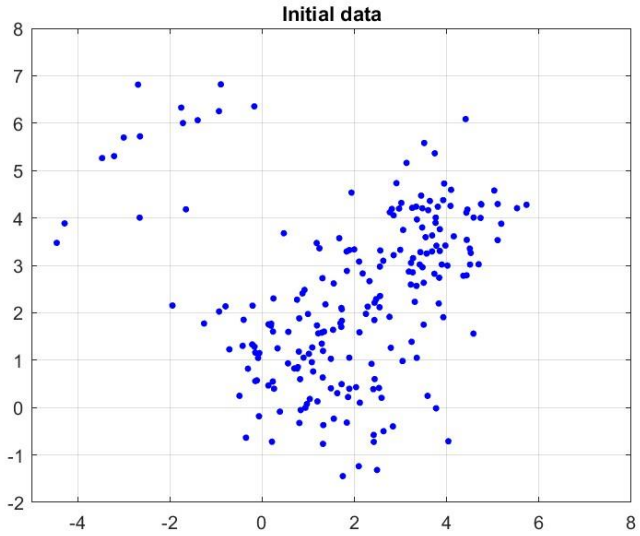
$$\boldsymbol{\mu}_1 = \begin{bmatrix}1\\1\end{bmatrix}, \boldsymbol{\mu}_2 = \begin{bmatrix}6\\1\end{bmatrix} \qquad \Sigma_1 = \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}, \Sigma_2 = \begin{bmatrix}1.5 & 0\\0 & 1.5\end{bmatrix} \qquad \boldsymbol{\theta}_1 = \begin{bmatrix}3.58\\1.12\end{bmatrix}, \boldsymbol{\theta}_2 = \begin{bmatrix}0.72\\0.85\end{bmatrix}$$

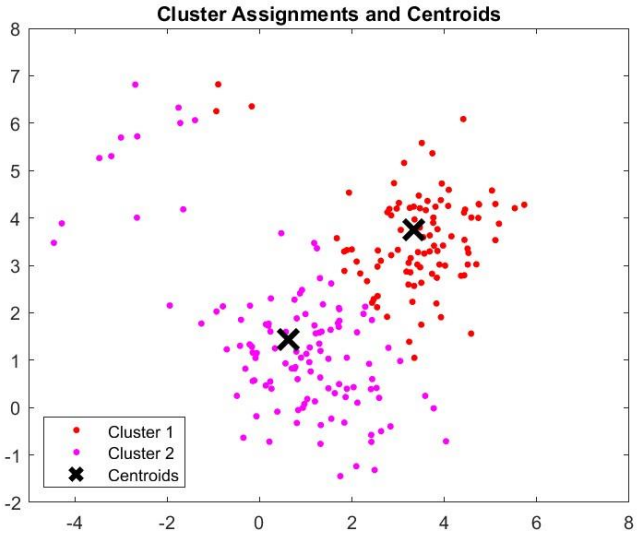
Ο αλγόριθμος k-μέσων μπορεί να αποτύχει αν κάποια κλάση υπο-αντιπροσωπεύεται

Ο αλγόριθμος ομαδοποίησης k-μέσων

Πείραμα με 2 κλάσεις και θόρυβο (outliers)







$$\boldsymbol{\mu}_1 = \begin{bmatrix}1\\1\end{bmatrix}, \boldsymbol{\mu}_2 = \begin{bmatrix}3.5\\3.5\end{bmatrix} \qquad \Sigma_1 = \begin{bmatrix}1 & 0.3\\0.3 & 1\end{bmatrix}, \Sigma_2 = \begin{bmatrix}1 & 0.7\\0.7 & 1\end{bmatrix} \qquad \boldsymbol{\theta}_1 = \begin{bmatrix}3.34\\3.75\end{bmatrix}, \boldsymbol{\theta}_2 = \begin{bmatrix}0.62\\1.44\end{bmatrix}$$

Η απόδοση του k-means επηρεάζεται από την παρουσία ακραίων σημείων και θορύβου

Ένα σχόλιο για τον αλγόριθμο Fuzzy c-means

- Οι αλγόριθμοι ασαφούς ομαδοποίησης προκύπτουν από την ελαχιστοποίηση μιας συνάρτησης κόστους της μορφής

$$J_q(\boldsymbol{\theta}, \mathrm{U}) = \sum_{i=1}^{N}\sum_{j=1}^{m} u_{ij}^q d(\boldsymbol{x}_i, \boldsymbol{\theta}_j)\,, q > 1$$

  λαμβάνοντας υπ' όψιν τους περιορισμούς

$$\sum_{j=1}^{m} u_{ij} = 1, i = 1,2, \ldots, N$$

- Αν $d(\boldsymbol{x}_i, \boldsymbol{\theta}_j)$ είναι το τετράγωνο της Ευκλείδειας απόστασης, το πρόβλημα αυτό μπορεί να λυθεί σε κλειστή μορφή με τη μέθοδο της εναλλασσόμενης ελαχιστοποίησης, χρησιμοποιώντας πολλαπλασιαστές Lagrange, και δίνει

$$u_{ij} = \frac{1}{\sum_{k=1}^{m}\left(\frac{d(\boldsymbol{x}_i, \boldsymbol{\theta}_j)}{d(\boldsymbol{x}_i, \boldsymbol{\theta}_k)}\right)^{\frac{1}{q-1}}}, \qquad \boldsymbol{\theta}_j = \frac{\sum_{i=1}^{N} u_{ij}^q \boldsymbol{x}_i}{\sum_{i=1}^{N} u_{ij}^q}$$

Ιεραρχικοί αλγόριθμοι ομαδοποίησης

- Οι ιεραρχικοί αλγόριθμοι ομαδοποίησης παράγουν μια ιεραρχία από ομαδοποιήσεις, αντί να παράγουν μια και μοναδική ομαδοποίηση των δεδομένων.
- Ορίσαμε μια m-ομαδοποίηση πάνω στο σύνολο X των διανυσμάτων, ως εξής:

$$\Re = \{C_j, j = 1,2, \dots, m\}$$

με $C_j \subseteq X$.

- Θα λέμε ότι μια ομαδοποίηση $\Re_1$ που αποτελείται από k ομάδες, είναι εμφωλιασμένη (nested) στην ομαδοποίηση $\Re_2$ που αποτελείται από $(r < k)$ ομάδες, αν κάθε ομάδα της $\Re_1$ είναι υποσύνολο μιας ομάδας της $\Re_2$ και γράφουμε $\Re_1 \sqsubset \Re_2$.
- Οι ιεραρχικοί αλγόριθμοι παράγουν μια ιεραρχία εμφωλιασμένων ομαδοποιήσεων. Πιο συγκεκριμένα, οι αλγόριθμοι αυτοί περιλαμβάνουν N βήματα, όσα δηλαδή είναι τα διανύσματα του X. Σε κάθε βήμα t, παράγεται μια νέα ομαδοποίηση, η οποία βασίζεται στην ομαδοποίηση που προέκυψε από το προηγούμενο βήμα $t - 1$.
- Υπάρχουν δύο κύριες κατηγορίες αυτών των αλγορίθμων, οι συσσωρευτικοί (agglomerative) και οι διαιρετικοί (divisive) αλγόριθμοι ομαδοποίησης.

Ιεραρχικοί αλγόριθμοι ομαδοποίησης

- Στους συσσωρευτικούς αλγόριθμους, η αρχική ομαδοποίηση, $\mathfrak{R}_0$, αποτελείται από N ομάδες, η κάθε μία από τις οποίες περιέχει ένα μόνο στοιχείο του X. Κατά το πρώτο βήμα παράγεται η ομαδοποίηση $\mathfrak{R}_1$, η οποία περιέχει $N-1$ σύνολα, έτσι ώστε $\mathfrak{R}_0 \sqsubset \mathfrak{R}_1$. Η διαδικασία αυτή συνεχίζεται μέχρις ότου παραχθεί και η τελευταία ομαδοποίηση, $\mathfrak{R}_{N-1}$, η οποία περιέχει ένα μόνο σύνολο, δηλαδή, το σύνολο X των δεδομένων. Έτσι θα έχουμε,

$$\mathfrak{R}_0 \sqsubset \mathfrak{R}_1 \sqsubset \cdots \sqsubset \mathfrak{R}_{N-1}$$

- Οι διαιρετικοί αλγόριθμοι ακολουθούν την αντίθετη πορεία. Σε αυτή την περίπτωση η αρχική ομαδοποίηση, $\mathfrak{R}_0$, αποτελείται από ένα μόνο σύνολο, το X. Κατά το πρώτο βήμα παράγεται η ομαδοποίηση $\mathfrak{R}_1$, η οποία αποτελείται από δύο σύνολα, έτσι ώστε $\mathfrak{R}_1 \sqsubset \mathfrak{R}_0$. Η διαδικασία αυτή συνεχίζεται έως ότου παραχθεί και η τελευταία ομαδοποίηση, $\mathfrak{R}_{N-1}$, η οποία περιέχει N σύνολα, καθένα από τα οποία αποτελείται από ένα μόνο στοιχείο του X. Έτσι, τώρα, θα έχουμε,

$$\mathfrak{R}_{N-1} \sqsubset \mathfrak{R}_{N-2} \sqsubset \cdots \sqsubset \mathfrak{R}_0$$

Συσσωρευτικοί αλγόριθμοι ομαδοποίησης

[Έστω $g(C_i, C_j)$ μια συνάρτηση που μετράει το βαθμό εγγύτητας μεταξύ των C_i και C_j.]

Γενικευμένο συσσωρευτικό σχήμα (generalized agglomerative scheme – GAS)

1. Αρχικοποίηση:
   1.1. Επέλεξε την $\Re_0 = \{C_i = \{\boldsymbol{x}_i\},\ i = 1,2, \dots, N\}$ ως αρχική ομαδοποίηση
   1.2. $t = 0$.
2. Επανέλαβε
   2.1. $t = t + 1$
   2.2. Ανάμεσα σε όλα τα δυνατά ζεύγη ομάδων (C_r, C_s) της $\Re_{t-1}$, βρες εκείνο, έστω το (C_i, C_j) για το οποίο ισχύει

$$g(C_i, C_j) = \begin{cases} \min\limits_{r,s} g(C_r, C_s), & \text{αν η } g \text{ ειναι συναρτηση ανομοιοτητας} \\ \max\limits_{r,s} g(C_r, C_s), & \text{αν η } g \text{ ειναι συναρτηση ομοιοτητας} \end{cases}$$

   2.3. Όρισε την ομάδα $C_q = C_i \cup C_j$ και δημιούργησε τη νέα ομαδοποίηση $\Re_t = (\Re_{t-1} - \{C_i, C_j\}) \cup \{C_q\}$

Έως ότου όλα τα διανύσματα συγκεντρωθούν σε μία ομάδα.

Συσσωρευτικοί αλγόριθμοι ομαδοποίησης

- Είναι σαφές ότι το σχήμα αυτό δημιουργεί μια ιεραρχία από N ομαδοποιήσεις, έτσι ώστε η κάθε μία από αυτές να είναι εμφωλιασμένη σε όλες τις επόμενες ομαδοποιήσεις, δηλ. $\Re_{t_1} \sqsubset \Re_{t_2}, t_1 < t_2 = 1,2, \dots, N-1.$
- Εναλλακτικά, μπορούμε να πούμε ότι αν δύο διανύσματα βρεθούν στην ίδια ομάδα στο επίπεδο t της ιεραρχίας, θα παραμείνουν στην ίδια ομάδα για όλες τις επόμενες ομαδοποιήσεις.
- Ένα μειονέκτημα της ιδιότητας του εμφωλιασμού είναι ότι δεν υπάρχει τρόπος ανάνηψης από μια «κακή» ομαδοποίηση που εμφανίστηκε σε προηγούμενο επίπεδο της ιεραρχίας.
- Ο συνολικός αριθμός ζευγών ομάδων, που πρέπει να εξεταστούν κατά τη διάρκεια της διαδικασίας ομαδοποίησης είναι

$$\sum_{t=0}^{N-1} \binom{N-t}{2} = \sum_{k=1}^{N} \binom{k}{2} = \frac{N(N-1)(N+1)}{6}$$

- Δηλαδή, ο συνολικός αριθμός πράξεων που απαιτούνται σε ένα συσσωρευτικό αλγόριθμο είναι ανάλογος του N^3.

Συσσωρευτικοί αλγόριθμοι που βασίζονται στη θεωρία πινάκων

- Υπάρχουν δύο κύριες κατηγορίες συσσωρευτικών αλγορίθμων. Οι αλγόριθμοι της πρώτης κατηγορίας που βασίζονται σε ιδέες από τη θεωρία πινάκων, ενώ οι αλγόριθμοι της δεύτερης κατηγορίας βασίζονται σε ιδέες από τη θεωρία γράφων.

Ορισμοί

- Ο πίνακας προτύπων (pattern matrix) $D(X)$ είναι ο $N \times l$ η i-οστή γραμμή του οποίου είναι το (αντεστραμμένο) i-οστό διάνυσμα του X.
- Ο πίνακας ομοιότητας (ανομοιότητας), $P(X)$, είναι ένας $N \times N$, του οποίου το στοιχείο (i, j) ισούται με το βαθμό ομοιότητας $s(\boldsymbol{x}_i, \boldsymbol{x}_j)$ (ανομοιότητας $d(\boldsymbol{x}_i, \boldsymbol{x}_j)$) των διανυσμάτων $\boldsymbol{x}_i$ και $\boldsymbol{x}_j$. Ο πίνακας αυτός είναι γνωστός και ως πίνακας εγγύτητας.
- Ο πίνακας P είναι ένας συμμετρικός πίνακας. Επιπλέον, αν ο P είναι πίνακας ομοιότητας τα διαγώνια στοιχεία του είναι ίσα με τη μέγιστη τιμή του s (s_0). Αν ο P είναι πίνακας ανομοιότητας τα διαγώνια στοιχεία του είναι ίσα με την ελάχιστη τιμή του d (d_0).
- Σε ένα πίνακα προτύπων αντιστοιχούν περισσότεροι του ενός πίνακες εγγύτητας, ανάλογα με την επιλογή του μέτρου εγγύτητας $Q(\boldsymbol{x}_i, \boldsymbol{x}_j)$.

Συσσωρευτικοί αλγόριθμοι που βασίζονται στη θεωρία πινάκων

Παράδειγμα

Έστω $X = \{\boldsymbol{x}_i, i = 1,2,\dots,5\}$ με $\boldsymbol{x}_1 = [1,1]^T, \boldsymbol{x}_2 = [2,1]^T, \boldsymbol{x}_3 = [5,4]^T, \boldsymbol{x}_4 = [6,5]^T, \boldsymbol{x}_5 = [6.5,6]^T$. Ο πίνακας προτύπων του X, ο αντίστοιχος πίνακας ανομοιότητας όταν χρησιμοποιείται η Ευκλείδεια απόσταση και ο πίνακας ομοιότητας όταν χρησιμοποιείται η απόσταση κατά Tanimoto είναι αντίστοιχα

$$D(X) = \begin{bmatrix} 1 & 1 \\ 2 & 1 \\ 5 & 4 \\ 6 & 5 \\ 6.5 & 6 \end{bmatrix}, \; P_1(X) = \begin{bmatrix} 0 & 1 & 5 & 6.4 & 7.4 \\ 1 & 0 & 4.2 & 5.7 & 6.7 \\ 5 & 4.2 & 0 & 1.4 & 2.5 \\ 6.4 & 5.7 & 1.4 & 0 & 1.1 \\ 7.4 & 6.7 & 2.5 & 1 & 0 \end{bmatrix}, P_2(X) = \begin{bmatrix} 1 & 0.75 & 0.26 & 0.21 & 0.18 \\ 0.75 & 1 & 0.44 & 0.35 & 0.20 \\ 0.26 & 0.44 & 1 & 0.96 & 0.90 \\ 0.21 & 0.35 & 0.96 & 1 & 0.98 \\ 0.18 & 0.20 & 0.90 & 0.98 & 1 \end{bmatrix}$$

- Σημειώνουμε ότι όλα τα διαγώνια στοιχεία του $P_1(X)$ είναι 0, ενώ όλα τα διαγώνια στοιχεία του $P_2(X)$ είναι 1.

Συσσωρευτικοί αλγόριθμοι που βασίζονται στη θεωρία πινάκων

- Ένα δεντρόγραμμα κατωφλίου (threshold dendrogram), ή απλά δεντρόγραμμα, είναι ένα αποτελεσματικό μέσο για την αναπαράσταση της ακολουθίας των ομαδοποιήσεων που παράγει ένας συσσωρευτικός αλγόριθμος.
- Για το προηγούμενο παράδειγμα, αν $g(C_i, C_j) = d_{\min}^{ss}(C_i, C_j)$ οι ομαδοποιήσεις που πραγματοποιούνται διαδοχικά και το αντίστοιχο δεντρόγραμμα θα είναι

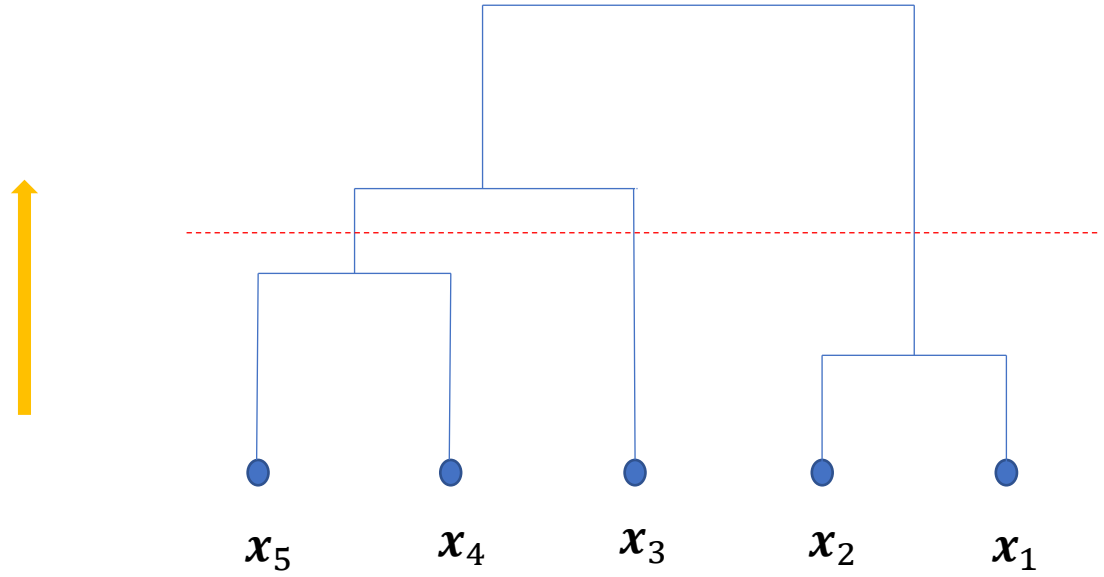
$$\{\{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4, \boldsymbol{x}_5\}\}$$

$$\{\{\boldsymbol{x}_1, \boldsymbol{x}_2\}, \{\boldsymbol{x}_3, \boldsymbol{x}_4, \boldsymbol{x}_5\}\}$$

$$\{\{\boldsymbol{x}_1, \boldsymbol{x}_2\}, \{\boldsymbol{x}_3\}, \{\boldsymbol{x}_4, \boldsymbol{x}_5\}\}$$

$$\{\{\boldsymbol{x}_1, \boldsymbol{x}_2\}, \{\boldsymbol{x}_3\}, \{\boldsymbol{x}_4\}, \{\boldsymbol{x}_5\}\}$$

$$\{\{\boldsymbol{x}_1\}, \{\boldsymbol{x}_2\}, \{\boldsymbol{x}_3\}, \{\boldsymbol{x}_4\}, \{\boldsymbol{x}_5\}\}$$

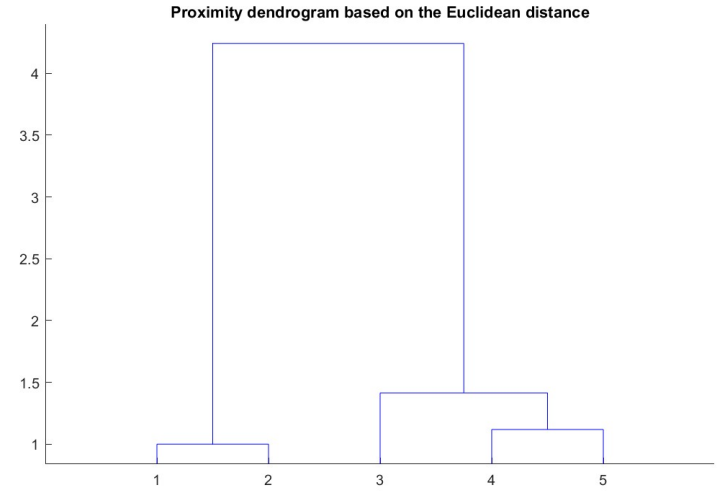


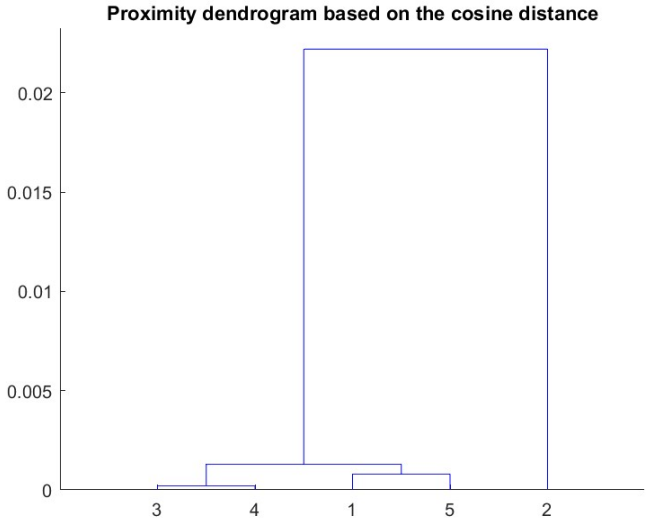
- Τεμαχίζοντας το δεντρόγραμμα σε ένα συγκεκριμένο επίπεδο κάθε φορά, παίρνουμε και μια διαφορετική ομαδοποίηση.

Συσσωρευτικοί αλγόριθμοι που βασίζονται στη θεωρία πινάκων

- Ένα δεντρόγραμμα εγγύτητας (proximity dendrogram) είναι ένα δεντρόγραμμα που λαμβάνει υπ’ όψιν το επίπεδο εγγύτητας στο οποίο δύο ομάδες συγχωνεύονται πρώτη φορά.
- Το εργαλείο αυτό μπορεί να δώσει μια ένδειξη για το κατά πόσο ο σχηματισμός ομάδων σε κάποιο επίπεδο είναι φυσικός ή αναγκαστικός, δηλαδή, μια ένδειξη σχετικά με εκείνη την ομαδοποίηση που ταιριάζει καλύτερα στα δεδομένα.

$$D(X) = \begin{bmatrix} 1 & 1 \\ 2 & 1 \\ 5 & 4 \\ 6 & 5 \\ 6.5 & 6 \end{bmatrix}, P(X) = \begin{bmatrix} 0 & 1 & 5 & 6.4 & 7.4 \\ 1 & 0 & 4.2 & 5.7 & 6.7 \\ 5 & 4.2 & 0 & 1.4 & 2.5 \\ 6.4 & 5.7 & 1.4 & 0 & 1.1 \\ 7.4 & 6.7 & 2.5 & 1 & 0 \end{bmatrix}$$





Συσσωρευτικοί αλγόριθμοι που βασίζονται στη θεωρία πινάκων

- Οι αλγόριθμοι ενημέρωσης πινάκων (matrix updating algorithms) δέχονται σαν είσοδο τον $N \times N$ πίνακα ανομοιότητας $P_0 = P(X)$, ο οποίος προέρχεται από το X. Σε κάθε επίπεδο t το μέγεθος του πίνακα ανομοιότητας P_0 γίνεται $(N-t) \times (N-t)$. Ο πίνακας P_t προκύπτει από τον P_{t-1} (α) απαλείφοντας τις δύο γραμμές και τις δύο στήλες που αντιστοιχούν στις ομάδες που συγχωνεύτηκαν και (β) προσθέτοντας μια νέα γραμμή και μια νέα στήλη, που περιέχει τις αποστάσεις μεταξύ της νέας ομάδας που σχηματίστηκε και των υπολοίπων ομάδων που δεν επηρεάστηκαν σε αυτό το βήμα.
- Η απόσταση της νέας ομάδας που σχηματίστηκε, C_q (λόγω της συγχώνευσης των C_i και C_j) και μιας εκ των υπολοίπων ομάδων, C_s, είναι μια συνάρτηση της μορφής

$$d(C_q, C_s) = f\left(d(C_i, C_s), d(C_j, C_s), d(C_i, C_j)\right)$$

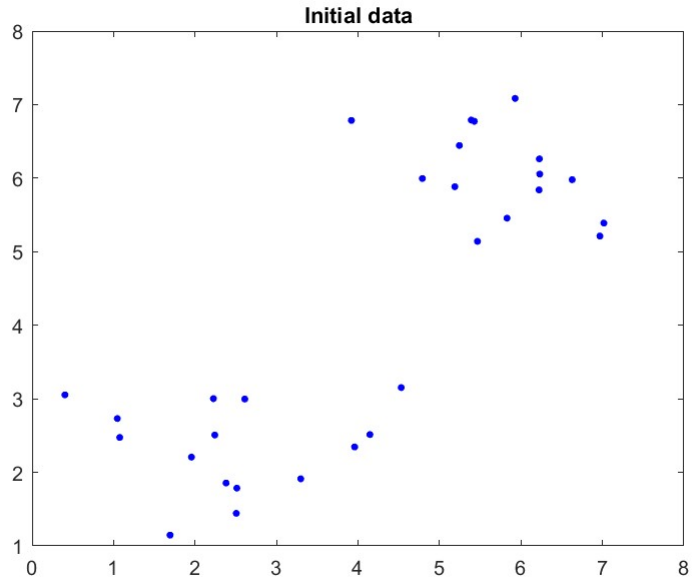
- Οι δύο πιο γνωστοί αλγόριθμοι που προκύπτουν με βάση την επιλογή του $d(C_q, C_s)$ είναι
  - Ο αλγόριθμος απλού δεσμού με $d(C_q, C_s) = \min\{d_{\min}^{ss}(C_i, C_s), d_{\min}^{ss}(C_j, C_s)\}$
  - Ο αλγόριθμος πλήρους δεσμού με $d(C_q, C_s) = \max\{d_{\min}^{ss}(C_i, C_s), d_{\min}^{ss}(C_j, C_s)\}$

Συσσωρευτικοί αλγόριθμοι που βασίζονται στη θεωρία πινάκων

Αλγοριθμικό σχήμα ενημέρωσης πινάκων (matrix updating algorithmic scheme – MUAS)

- Αρχικοποίηση
  - $\Re_0 = \{\{\boldsymbol{x}_i\}, i = 1, \dots, N\}$
  - $P_0 = P(X)$
  - $t = 0$
- Επανέλαβε
  - $t = t + 1$
  - Βρες τις C_i, C_j για τις οποίες $d(C_i, C_j) = \min_{r,s=1,2,\dots,N} d(C_r, C_s)$
  - Συγχώνευσε τις C_i, C_j σε μια ομάδα C_q και σχημάτισε την ομαδοποίηση $\Re_t = \left(\Re_{t-1} - \{C_i, C_j\}\right) \cup \{C_q\}$
  - Όρισε τον πίνακα εγγύτητας P_t από τον P_{t-1}, όπως εξηγήθηκε πριν
- Έως ότου σχηματιστεί η ομαδοποίηση $\Re_{N-1}$, δηλαδή, έως ότου όλα τα διανύσματα να βρεθούν στην ίδια ομάδα.

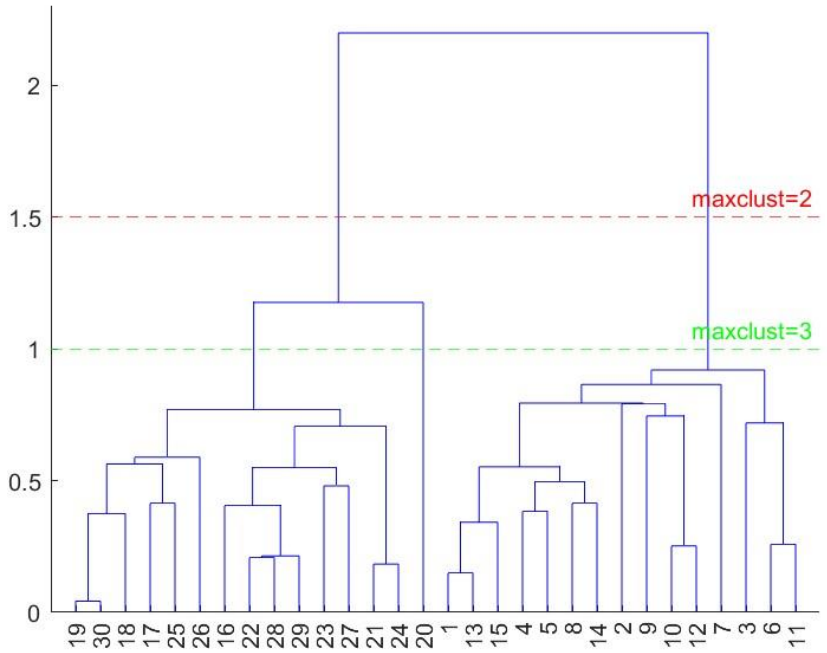
Συσσωρευτικοί αλγόριθμοι που βασίζονται στη θεωρία πινάκων

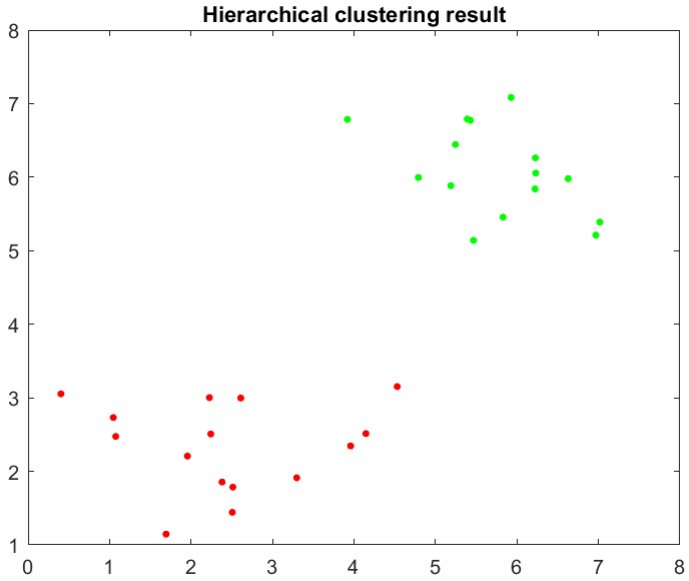


Πείραμα με 2 κλάσεις – 15 σημεία από κάθε κλάση (Gaussian κατανομές)

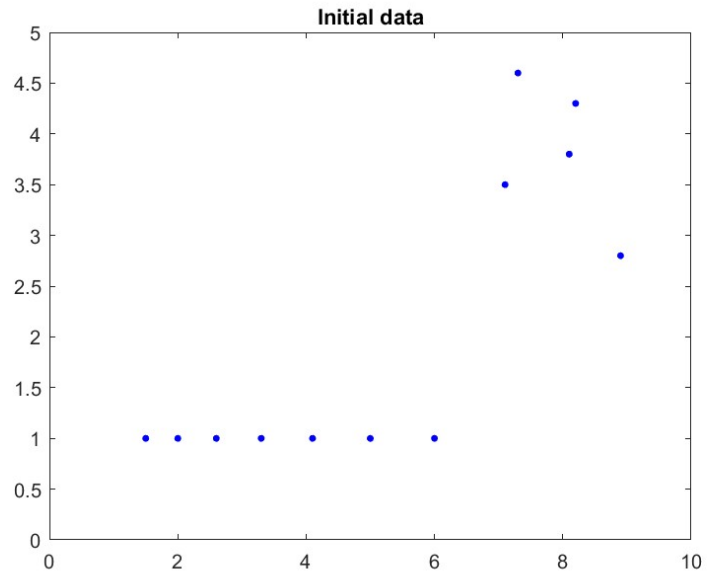
$$\boldsymbol{\mu}_1 = \begin{bmatrix} 2 \\ 2 \end{bmatrix}, \boldsymbol{\mu}_2 = \begin{bmatrix} 6 \\ 6 \end{bmatrix}$$

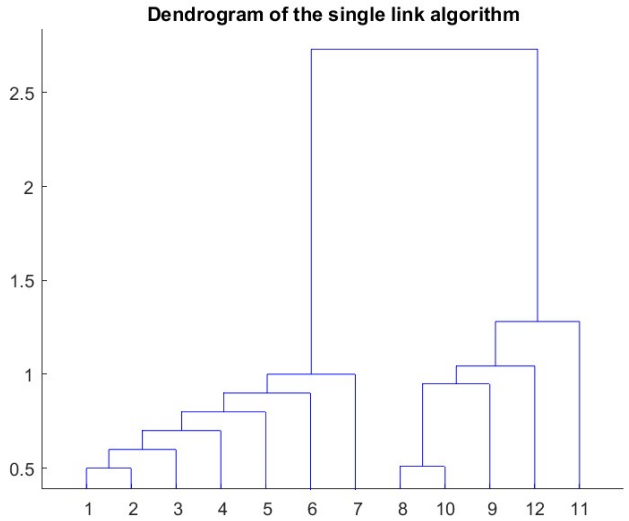
$$\Sigma_1 = \Sigma_2 = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix}$$

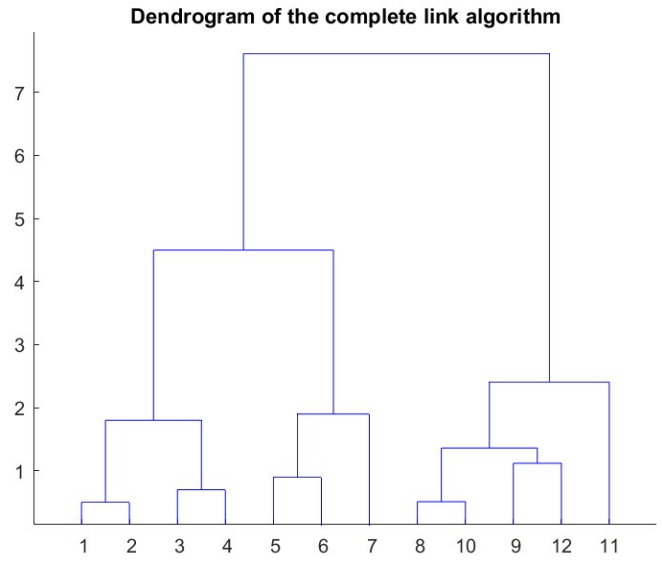




Συσσωρευτικοί αλγόριθμοι που βασίζονται στη θεωρία πινάκων







- Οι ομάδες που παράγονται από αλγόριθμο απλού δεσμού σχηματίζονται σε χαμηλά επίπεδα ανομοιότητας στο δεντρόγραμμα, ενώ οι ομάδες που παράγονται από αλγόριθμο πλήρους δεσμού σχηματίζονται σε υψηλά επίπεδα ανομοιότητας.
- Ο αλγόριθμος απλού δεσμού έχει την τάση να ευνοεί το σχηματισμό επιμηκών ομάδων, ενώ ο αλγόριθμος πλήρους δεσμού αναδεικνύει πρώτα μικρές συμπαγείς ομάδες.

Αλγόριθμοι ομαδοποίησης που βασίζονται στην πυκνότητα

- Στους αλγόριθμους αυτούς, οι ομάδες θεωρούνται ως περιοχές του l-διάστατου χώρου που είναι «πυκνές» σε σημεία του συνόλου δεδομένων, X.
- Οι αλγόριθμοι δεν θέτουν κανένα περιορισμό στο σχήμα των ομάδων που θα προκύψουν και έτσι έχουν τη δυνατότητα να αναδεικνύουν ομάδες οποιουδήποτε σχήματος.
- Οι αλγόριθμοι που βασίζονται στην πυκνότητα είναι, επίσης, ικανοί να χειρίζονται αποτελεσματικά τα ακραία σημεία (outliers).
- Η υπολογιστική πολυπλοκότητα των αλγορίθμων αυτών είναι μικρότερη από $O(N^2)$, πράγμα που τους καθιστά επιλέξιμους για την επεξεργασία μεγάλων συνόλων δεδομένων.
- Οι τεχνικές αυτές δεν είναι κατάλληλες για την επεξεργασία δεδομένων σε χώρους υψηλής διάστασης, όπου η πυκνότητα των δεδομένων είναι μικρή και όταν δεν έχουμε καλά διαχωρισμένες κλάσεις.

Ο αλγόριθμος DBSCAN

- Στον αλγόριθμο DBSCAN (Density-Based Spatial Clustering of Applications with Noise) η πυκνότητα γύρω από ένα σημείο $\boldsymbol{x}$, εκτιμάται ως ο αριθμός των σημείων του X που βρίσκονται μέσα σε μία ορισμένη περιοχή του l-διάστατου χώρου που περικλείει το $\boldsymbol{x}$.

- Θεωρούμε ότι η περιοχή αυτή είναι υπερσφαίρα $V_\varepsilon(\boldsymbol{x})$ με κέντρο το $\boldsymbol{x}$ και ακτίνα ε, η οποία είναι παράμετρος οριζόμενη από το χρήστη. Έστω $N_\varepsilon(\boldsymbol{x})$ ο αριθμός των σημείων του X που βρίσκονται μέσα στην $V_\varepsilon(\boldsymbol{x})$.

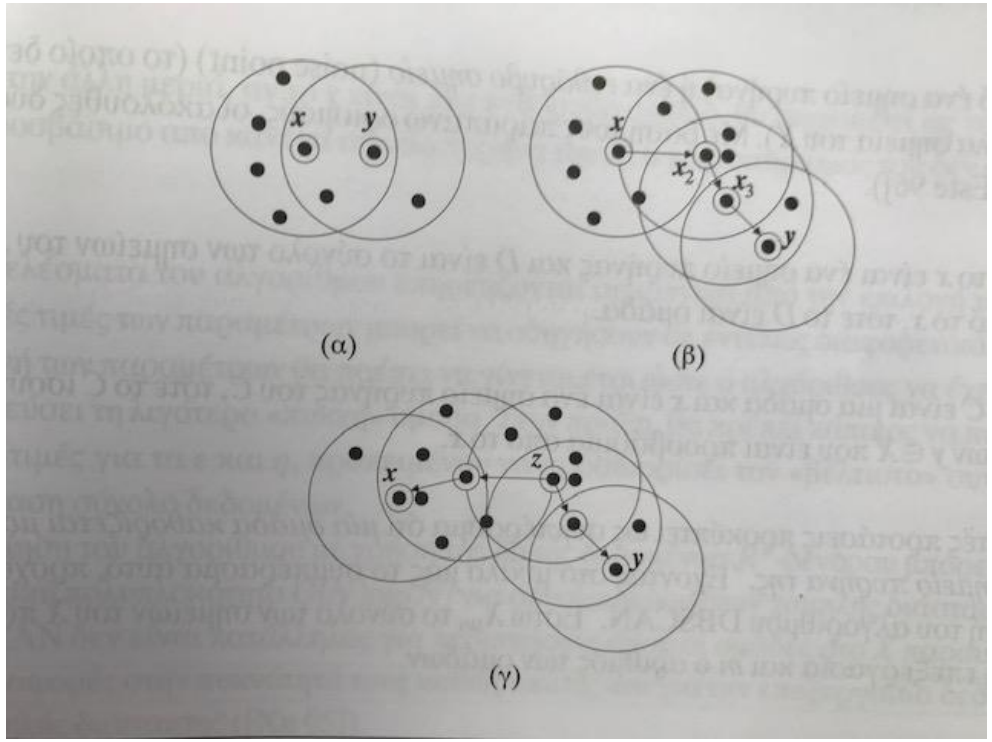
- Μια δεύτερη παράμετρος που ορίζεται από το χρήστη είναι ο ελάχιστος αριθμός σημείων, q, που πρέπει να περιέχονται στην $V_\varepsilon(\boldsymbol{x})$ προκειμένου το $\boldsymbol{x}$ να θεωρηθεί «εσωτερικό» σημείο μιας ομάδας.

Ο αλγόριθμος DBSCAN

- **Ορισμός 1**. Ένα σημείο $\boldsymbol{y}$ είναι άμεσα προσβάσιμο ως προς την πυκνότητα (directly density reachable) από ένα σημείο $\boldsymbol{x}$ αν (i) $\boldsymbol{y} \in V_\varepsilon(\boldsymbol{x})$ και (ii) $N_\varepsilon(\boldsymbol{x}) \geq q$

- **Ορισμός 2**. Ένα σημείο $\boldsymbol{y}$ είναι προσβάσιμο ως προς την πυκνότητα (density reachable) από ένα σημείο $\boldsymbol{x}$ του X αν υπάρχει μια ακολουθία σημείων $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p \in X$ με $\boldsymbol{x}_1 = \boldsymbol{x}$ και $\boldsymbol{x}_p = \boldsymbol{y}$, έτσι ώστε το $\boldsymbol{x}_{i+1}$ να είναι άμεσα προσβάσιμο από το $\boldsymbol{x}_i$.

- **Ορισμός 3**. Ένα σημείο $\boldsymbol{x}$ είναι συνδεδεμένο με βάση την πυκνότητα με ένα σημείο $\boldsymbol{y} \in \boldsymbol{X}$ αν υπάρχει $\mathbf{z} \in \boldsymbol{X}$ έτσι ώστε αμφότερα τα $\boldsymbol{x}$ και $\boldsymbol{y}$ να είναι προσβάσιμα ως προς την πυκνότητα από το $\mathbf{z}$.



q=5

Ο αλγόριθμος DBSCAN

- Μια ομάδα C, στα πλαίσια του DBSCAN, ορίζεται ως ένα μη κενό υποσύνολο του X που ικανοποιεί τις ακόλουθες συνθήκες:

  (i) Αν το $\boldsymbol{x}$ ανήκει στο C και το $\boldsymbol{y} \in X$ είναι προσβάσιμο από το $\boldsymbol{x}$, τότε $\boldsymbol{y} \in C$.

  (ii) Για κάθε ζεύγος $(\boldsymbol{x}, \boldsymbol{y}) \in C$, τα $\boldsymbol{x}$ και $\boldsymbol{y}$ είναι συνδεδεμένα.

- Αν $C_1, C_2, \ldots, C_m$ είναι οι ομάδες του X, τότε, το σύνολο των σημείων που δεν περιέχονται σε καμία από αυτές ονομάζεται θόρυβος (noise).

- Ένα σημείο $\boldsymbol{x}$ ορίζεται ως σημείο πυρήνας (core point), αν έχει τουλάχιστον q σημεία στη γειτονιά του. Διαφορετικά, το $\boldsymbol{x}$ καλείται μη βασικό σημείο (noncore point). Ένα μη βασικό σημείο μπορεί να είναι είτε ένα συνοριακό σημείο (border point) μιας ομάδας (προσβάσιμο από ένα σημείο πυρήνα) ή ένα ενθόρυβο σημείο (το οποίο δεν είναι προσβάσιμο από άλλα σημεία του X.

Πρόταση 1. Αν το $\boldsymbol{x}$ είναι ένα σημείο πυρήνας και D είναι το σύνολο των σημείων του X που είναι προσβάσιμα από το $\boldsymbol{x}$, τότε το D είναι ομάδα.

Πρόταση 2. Αν C είναι μια ομάδα και $\boldsymbol{x}$ είναι ένα σημείο πυρήνας του C, τότε το C ισούται με το σύνολο σημείων $\boldsymbol{y} \in X$ που είναι προσβάσιμα από το $\boldsymbol{x}$.

Ο αλγόριθμος DBSCAN

Από τις παραπάνω προτάσεις προκύπτει ότι μια ομάδα καθορίζεται μοναδικά από οποιοδήποτε σημείο πυρήνα της. Έστω X_{un} το σύνολο των σημείων του X που δεν έχουν υποστεί ακόμη επεξεργασία και m ο αριθμός των ομάδων.

Αλγόριθμος DBSCAN

- Θέσε $X_{un} = X$ και $m = 0$.
- Ενόσω $X_{un} \neq \emptyset$
  - Επέλεξε αυθαίρετα ένα $\boldsymbol{x} \in X_{un}$
  - Αν το $\boldsymbol{x}$ είναι ένα μη βασικό σημείο
    - Σημείωσε το $\boldsymbol{x}$ ως ενθόρυβο σημείο
    - $X_{un} = X_{un} - \{\boldsymbol{x}\}$
  - Διαφορετικά αν το $\boldsymbol{x}$ είναι ένα βασικό σημείο τότε
    - $m = m + 1$
    - Προσδιόρισε όλα τα σημεία του X που είναι προσβάσιμα από το $\boldsymbol{x}$
    - Καταχώρησε το $\boldsymbol{x}$ και τα προηγούμενα σημεία στην ομάδα C_m. Τα συνοριακά σημεία που έχουν πιθανόν σημειωθεί ως ενθόρυβα καταχωρούνται επίσης στο C_m.
    - $X_{un} = X_{un} - \{C_m\}$
  - Τέλος {Αν)
- Τέλος {Ενόσω}

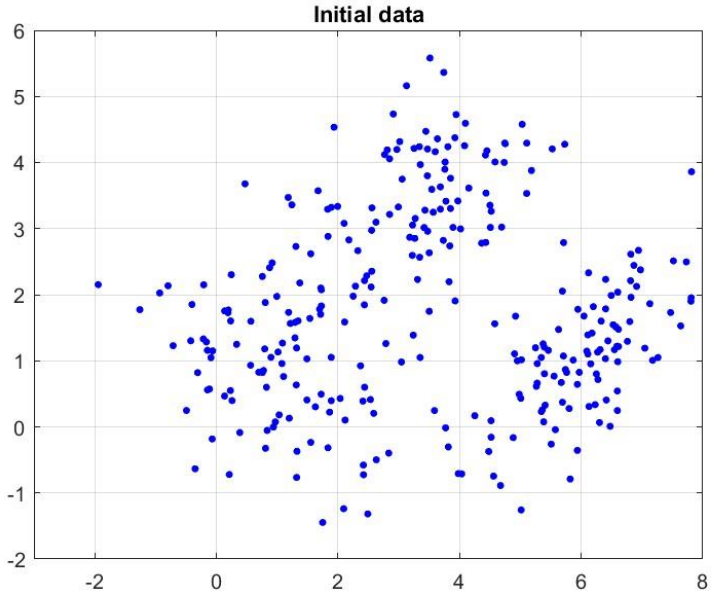
Ο αλγόριθμος DBSCAN

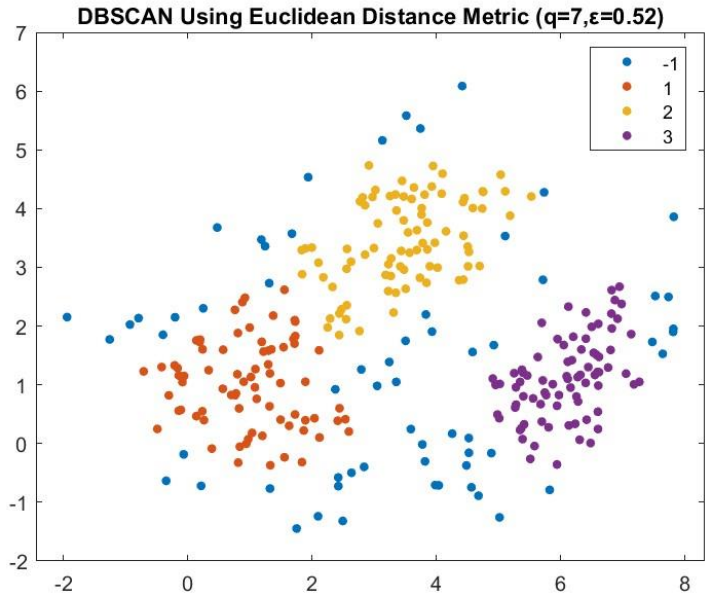
Παρατηρήσεις σχετικά με τον αλγόριθμο DBSCAN

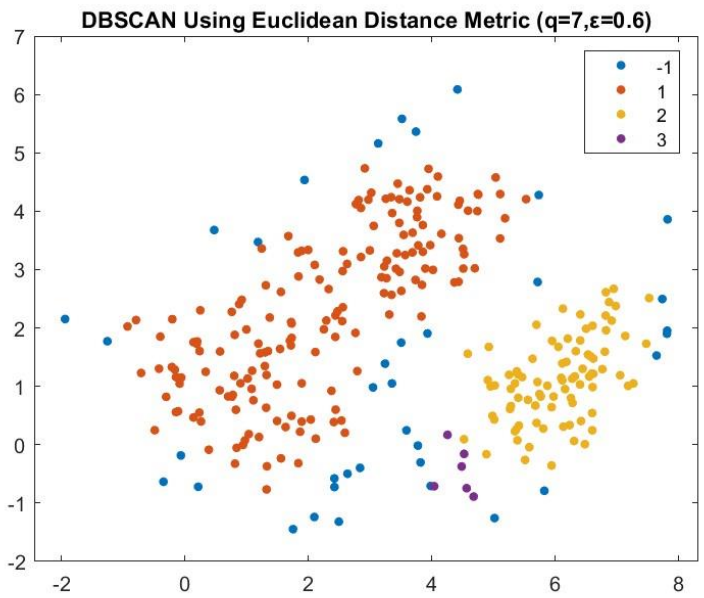
- Ακόμα και αν ένα σημείο $\boldsymbol{y}$, που αποτελεί ακραίο σημείο μιας ομάδας C, χαρακτηριστεί αρχικά από τον αλγόριθμο ως ενθόρυβο σημείο, θα καταχωρηθεί αργότερα στην ομάδα C, όταν εξεταστεί ένα σημείο πυρήνας $\boldsymbol{x}$ του C από το οποίο το $\boldsymbol{y}$ είναι προσβάσιμο.
- Τα αποτελέσματα του αλγόριθμου επηρεάζονται σημαντικά από την επιλογή των ε και q. Η επιλογή των παραμέτρων θα πρέπει να γίνεται έτσι ώστε ο αλγόριθμος να έχει την ικανότητα να ανιχνεύσει τη λιγότερο πυκνή ομάδα. Στην πράξη θα πρέπει κάποιος να πειραματιστεί με αρκετές τιμές για τα ε και q, προκειμένου να προσδιορίσει το βέλτιστο συνδυασμό.
- Για σύνολα δεδομένων χαμηλής διάστασης, μπορεί να αναπτυχθεί υλοποίηση του αλγόριθμου με αριθμητική πολυπλοκότητα $O(N \log_2 N)$.
- Ο DBSCAN δεν είναι κατάλληλος για περιπτώσεις όπου οι ομάδες στο X παρουσιάζουν σημαντικές διαφορές στην πυκνότητά τους καθώς επίσης και για την επεξεργασία δεδομένων σε χώρους υψηλής διάστασης.

Ο αλγόριθμος DBSCAN

Πείραμα με 3 κλάσεις – 100 σημεία από κάθε κλάση (Gaussian κατανομές)





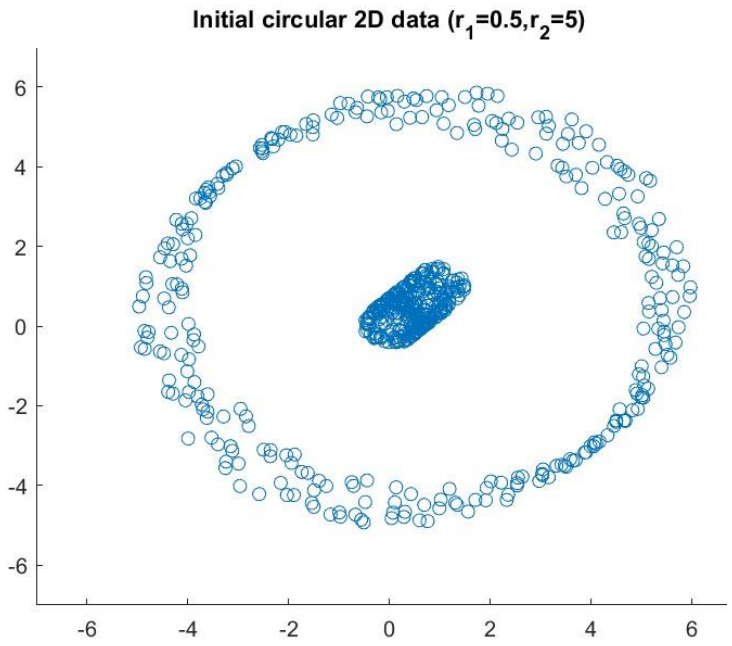


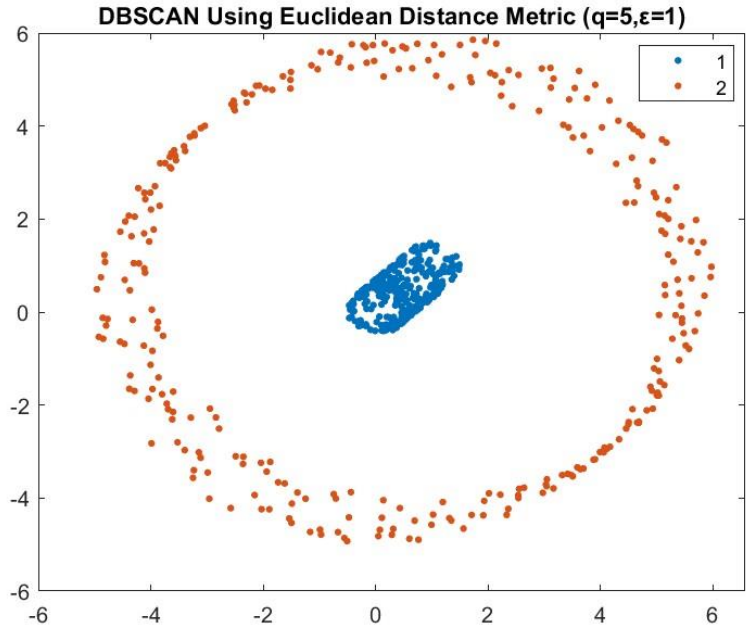
$$\boldsymbol{\mu}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \boldsymbol{\mu}_2 = \begin{bmatrix} 3.5 \\ 3.5 \end{bmatrix}, \boldsymbol{\mu}_3 = \begin{bmatrix} 6 \\ 1 \end{bmatrix}$$

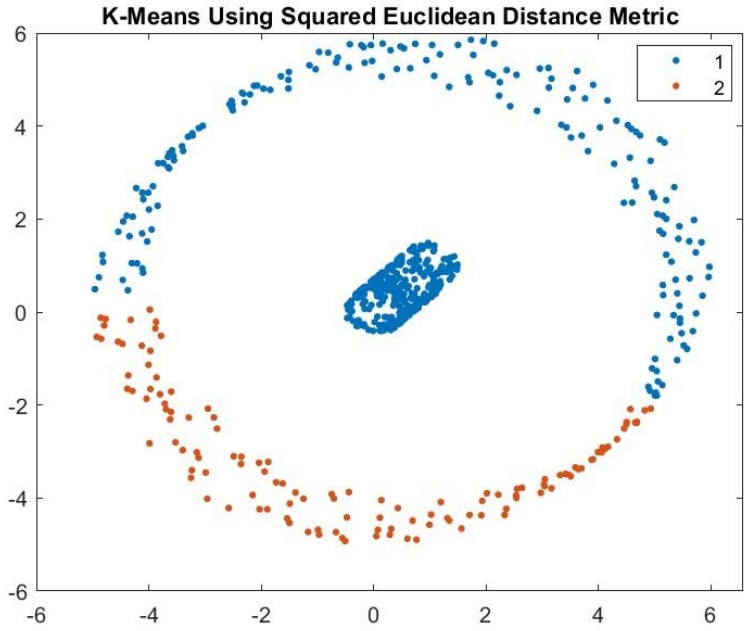
$$\Sigma_1 = \begin{bmatrix} 1 & -0.3 \\ -0.3 & 1 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} 1 & 0.3 \\ 0.3 & 1 \end{bmatrix}, \Sigma_3 = \begin{bmatrix} 1 & 0.7 \\ 0.7 & 1 \end{bmatrix}$$

Ο αλγόριθμος DBSCAN

Πείραμα με 2 δισδιάστατες δακτυλιοειδείς κλάσεις – 300 σημεία από κάθε κλάση



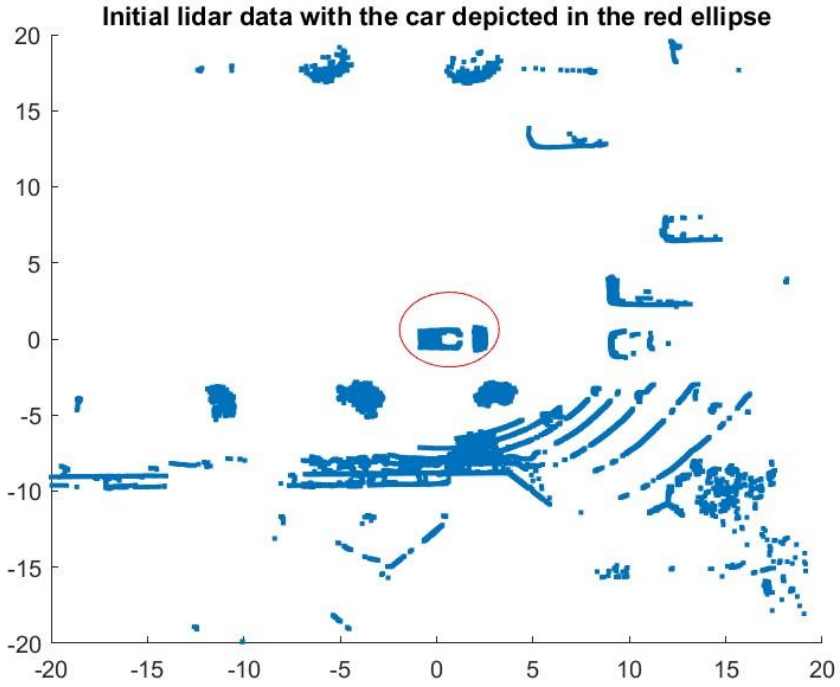


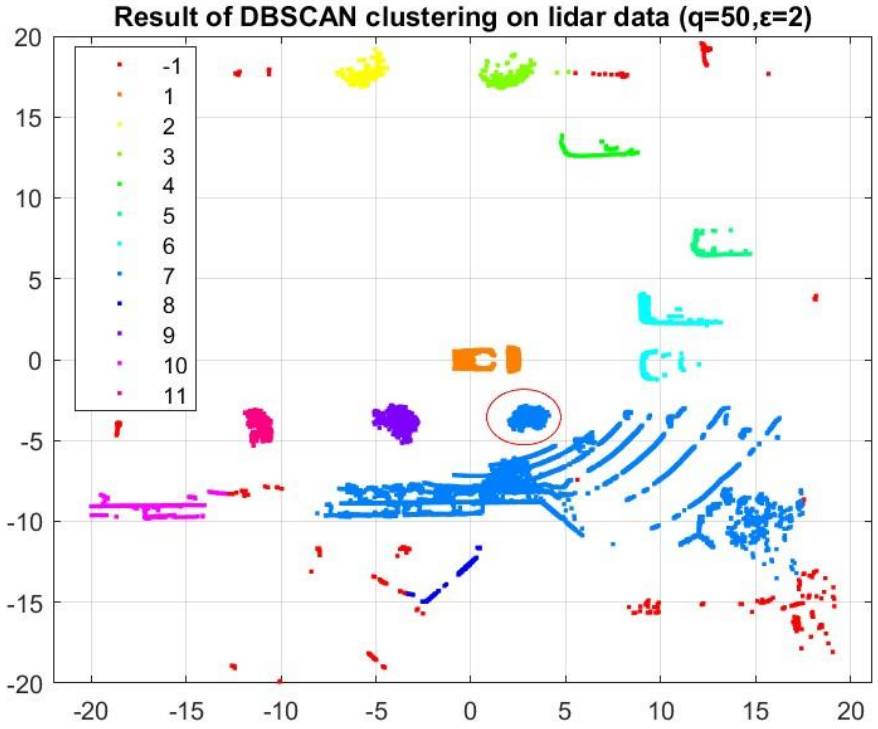


Source: https://www.mathworks.com/help/stats/dbscan.html

Ο αλγόριθμος DBSCAN

Ομαδοποίηση δεδομένων Lidar



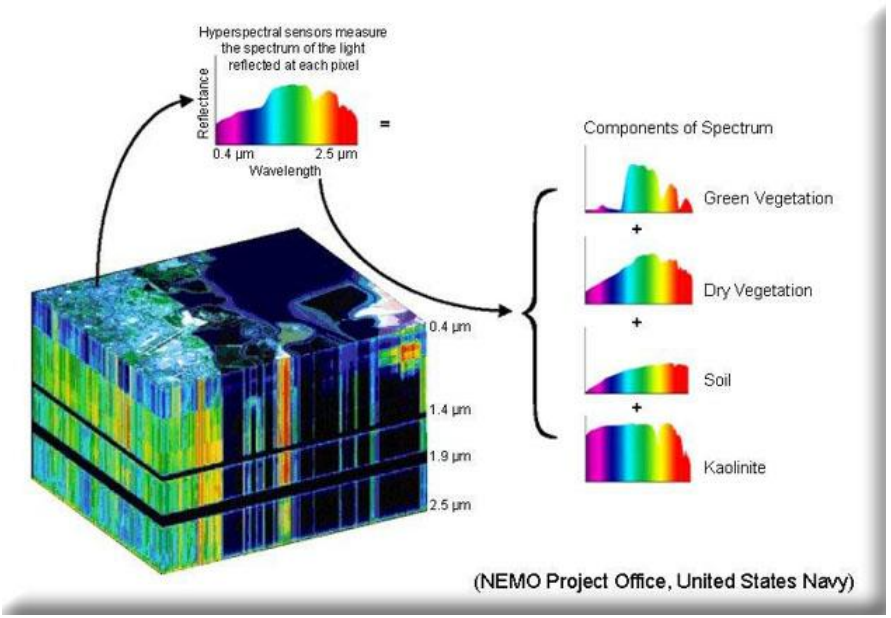


Το πλήθος των αρχικών σημείων είναι 19070. Ο αλγόριθμος αναγνωρίζει 18446 σημεία πυρήνα και 412 outliers (με κόκκινο χρώμα)

Source: https://www.mathworks.com/help/stats/dbscan.html

Υπερφασματικές εικόνες

Οι υπερφασματικές εικόνες - ΥΕ (hyperspectral images – HSI) είναι τρισδιάστατες «οντότητες» (φασματικοί κύβοι) που αποτελούνται από ένα μεγάλο αριθμό μονοχρωματικών εικόνων που έχουν ληφθεί σε πολύ μικρού εύρους κανάλια (μπάντες) του ΗΜ φάσματος, τις περισσότερες φορές στην περιοχή του ορατού και του κοντινού υπέρυθρου.

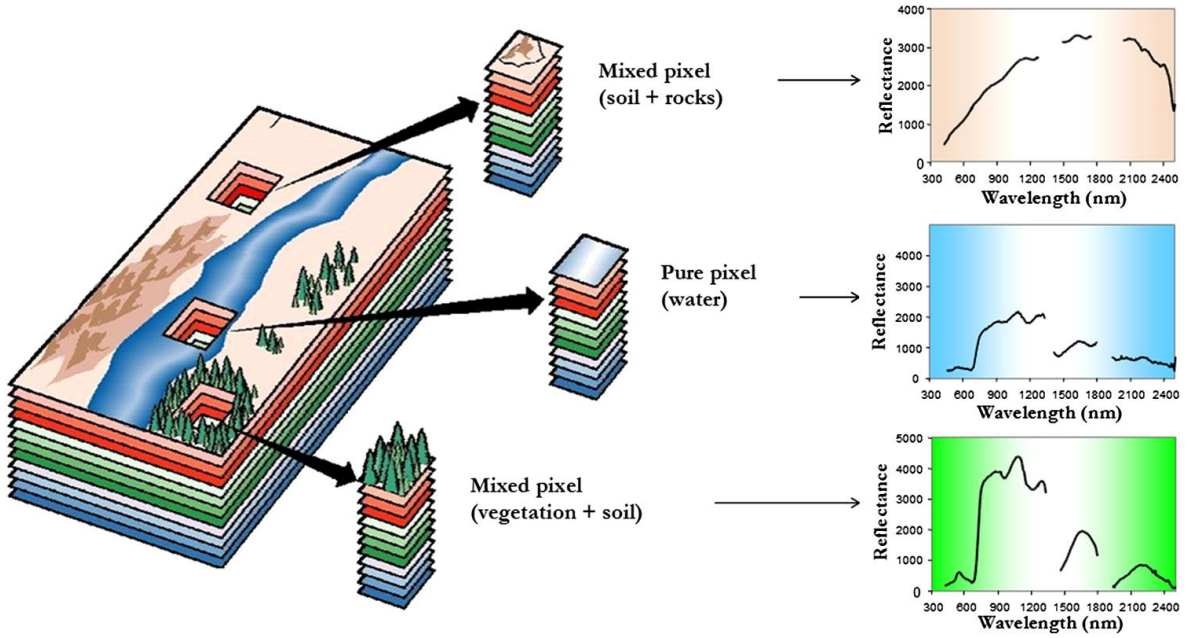


Υπερφασματικές κάμερες χρησιμοποιούνται σε όλα τα σύγχρονα εναέρια μέσα παρατήρησης της Γης, καθώς και στα σύγχρονα δορυφορικά συστήματα. Επίσης, σε πολλές άλλες εφαρμογές.

- AVIRIS. 224 φασματικά κανάλια στην περιοχή 0.4-2.5μm. Χωρική ανάλυση: 20m.
- HYDICE. 210 φασματικά κανάλια στην περιοχή 0.4-2.5μm. Χωρική ανάλυση: 1-4m.
- OMEGA. 186 φασματικά κανάλια στην περιοχή 0.5-5.2μm. Χωρική ανάλυση: 300m.
- CRISM. 544 φασματικά κανάλια στην περιοχή 0.35-3.95μm. Χωρική ανάλυση: 18m.

Υπερφασματικές εικόνες

- Οι υπερφασματικές εικόνες περιέχουν πλούσια πληροφορία, την οποία μπορούμε να αξιοποιήσουμε όχι μόνο για να αναγνωρίσουμε αντικείμενα ή υλικά στην σκηνή της εικόνας, αλλά και για να προσδιορίσουμε την κατανομή τους πάνω στην χωρική περιοχή που απεικονίζεται.
- Σε κάθε pixel μιας ΥΕ αντιστοιχεί ένα διάνυσμα του οποίου τα στοιχεία είναι οι τιμές της ανακλώμενης ενέργειας από την φυσική περιοχή του pixel σε όλες τις φασματικές ζώνες. Το μέγεθος του διανύσματος ισούται με τον αριθμό των φασματικών καναλιών και ονομάζεται φασματική υπογραφή του pixel.
- Τα pixels μιας ΥΕ διακρίνονται σε pure και mixed. Ένα pixel ονομάζεται pure αν υπάρχει ένα μόνο υλικό στην χωρική περιοχή που αντιστοιχεί στο pixel. Αν περισσότερα του ενός υλικού «συνυπάρχουν», το pixel ονομάζεται mixed.
- Η φασματική υπογραφή ενός pure υλικού ονομάζεται endmember.



Source: G. Shaw and D. Manolakis, “Signal processing for hyperspectral image exploitation,” IEEE Signal Process. Mag., vol. 19, pp. 12–16, Jan. 2002.

Υπερφασματικές εικόνες

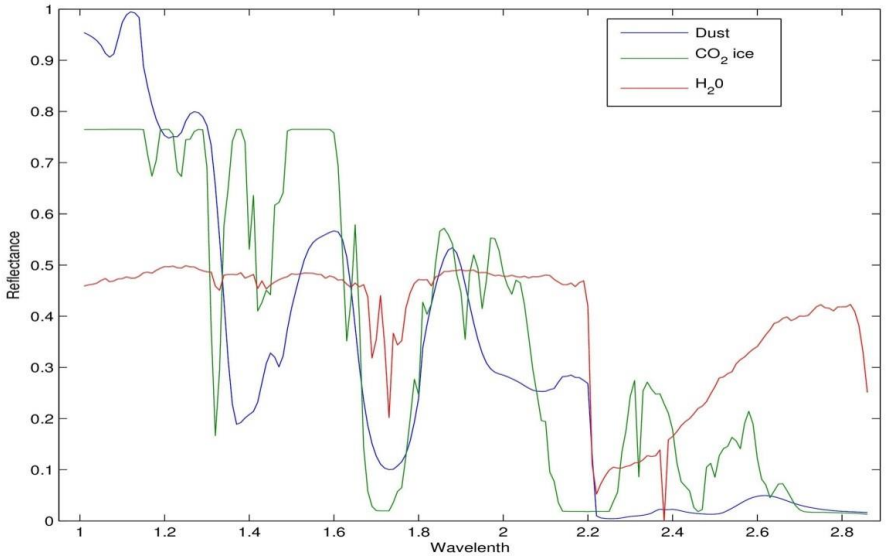
OMEGA Mars South Polar Cap HSI $(871 \times 128 \times 186)$

Band 20

Band 80

Band 100

Band 120

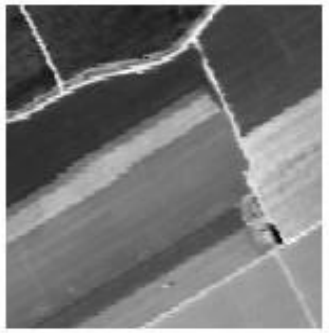


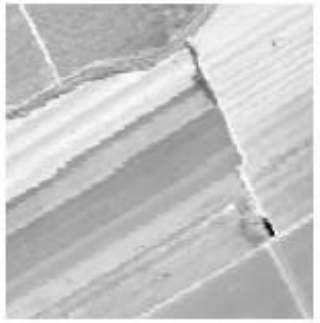
Τα endmembers της εικόνας

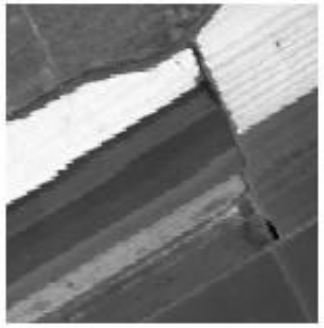
Υπερφασματικές εικόνες

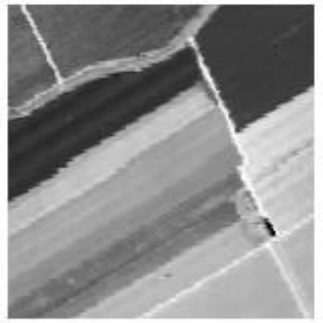
AVIRIS Salinas HSI $(150 \times 150 \times 224)$

RGB representation of the image









Four different band images of Salinas HSI

Ομαδοποίηση των pixels της ΥΕ

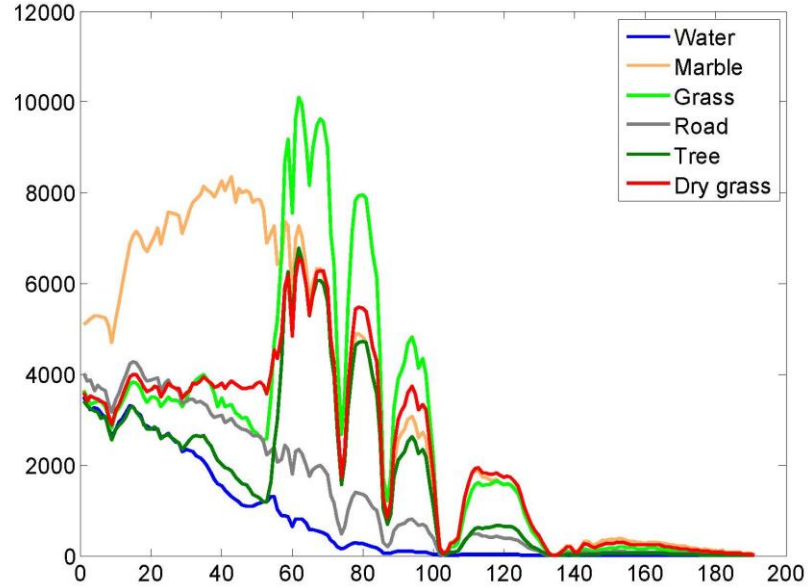
- Έστω μια ΥΕ διάστασης $n \times k \times l$, όπου n, k είναι οι χωρικές διαστάσεις και l η φασματική διάσταση (αριθμός καναλιών). Στόχος είναι η ομαδοποίηση των pixels της εικόνας σε κλάσεις που αντιστοιχούν στα διαφορετικά υλικά ή/και αντικείμενα που υπάρχουν στην εικόνα.
- Οι φασματικές υπογραφές των pixels της εικόνας μπορούν να θεωρηθούν ως σημεία στο $\mathbb{R}_+^l$. Έτσι, είναι αναμενόμενο ότι pixels που αντιστοιχούν στο ίδιο υλικό, θα έχουν φασματικές υπογραφές που θα δημιουργούν μια σχετικά συμπαγή ομάδα στον l-διάστατο χώρο.
- Κατά συνέπεια, μπορεί να χρησιμοποιηθεί ένας αλγόριθμος ομαδοποίησης που έχει την ικανότητα να αναδεικνύει συμπαγείς ομάδες, π.χ., ο k-means. Στην περίπτωση αυτή, τα διανύσματα χαρακτηριστικών θα είναι οι φασματικές υπογραφές των pixels. Κάθε pixel θα καταχωρηθεί σε μία και μόνο ομάδα, παρόλο που μπορεί να είναι mixed pixel.
- Ο αλγόριθμος θα επιστρέψει μια κατάτμηση της φυσικής εικόνας με βάση τα διαφορετικά υλικά που υπάρχουν σε αυτή.
- Επειδή τα δεδομένα αυτά είναι πλεονάζοντα και έχουν μεγάλο βαθμό συσχέτισης, μπορεί να χρησιμοποιηθεί μια μέθοδος μείωσης διάστασης (π.χ. PCA) και στη συνέχεια να εφαρμοστεί ένας αλγόριθμος ομαδοποίησης με διανύσματα χαρακτηριστικών πολύ μικρότερης διάστασης.

Ομαδοποίηση των pixels της ΥΕ

HYDICE Washington DC Mall HSI $n = k = 150$ (22500 pixels), $l = 191$ spectral bands

RGB αναπαράσταση της εικόνας

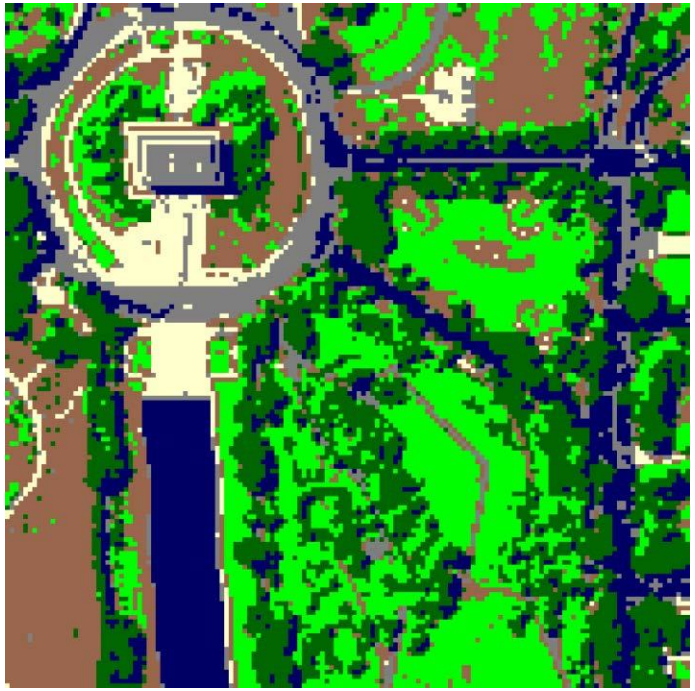
Η 90η μπάντα της εικόνας



Τα endmembers της εικόνας

Ομαδοποίηση των pixels της ΥΕ

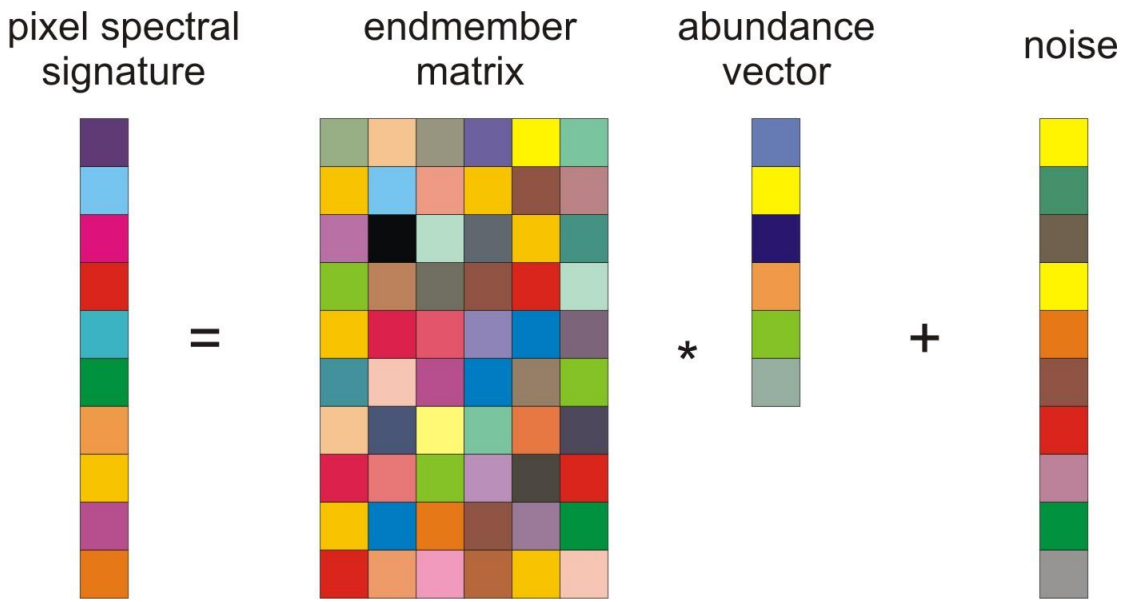
Αποτέλεσμα κατάτμησης της εικόνας μέσω ομαδοποίησης των pixels με χρήση του k-means με $m = 6$



Φασματικός διαχωρισμός ΥΕ

Γραμμικό μοντέλο μίξης (Linear mixing model – LMM)

Στο γραμμικό μοντέλο μίξης υποθέτουμε ότι η φασματική υπογραφή κάθε pixel μπορεί να εκφρασθεί ως γραμμικός συνδυασμός των endmembers συν κάποιο θόρυβο, ο οποίος θεωρείται συνήθως Gaussian. Οι συντελεστές αυτού του γραμμικού συνδυασμού ονομάζονται abundances. Ο πίνακας του οποίου οι στήλες είναι τα endmembers, ονομάζεται endmember matrix.



Φασματικός διαχωρισμός ΥΕ

Μαθηματικά αυτό εκφράζεται ως εξής:

$$\boldsymbol{x}_i = E\boldsymbol{w}_i + \boldsymbol{\eta}_i, \qquad i = 1,2,\dots,N(=n\cdot k)$$

όπου,

- N είναι ο συνολικός αριθμός των pixels της εικόνας
- m είναι ο αριθμός των endmembers
- l είναι ο αριθμός των φασματικών καναλιών
- $\boldsymbol{x}_i$ είναι η $l \times 1$ φασματική υπογραφή του i-οστού pixel
- E είναι ο $l \times m$ endmember matrix, κοινός για όλα τα pixels
- $\boldsymbol{w}_i$ είναι το άγνωστο $m \times 1$ διάνυσμα των abundances του i-οστού pixel. Το j-οστό στοιχείο του είναι ένα μέτρο της «συνεισφοράς» του j-οστού endmember στο pixel.
- $\boldsymbol{\eta}_i$ είναι ένα $l \times 1$ διάνυσμα θορύβου

Φασματικός διαχωρισμός ΥΕ

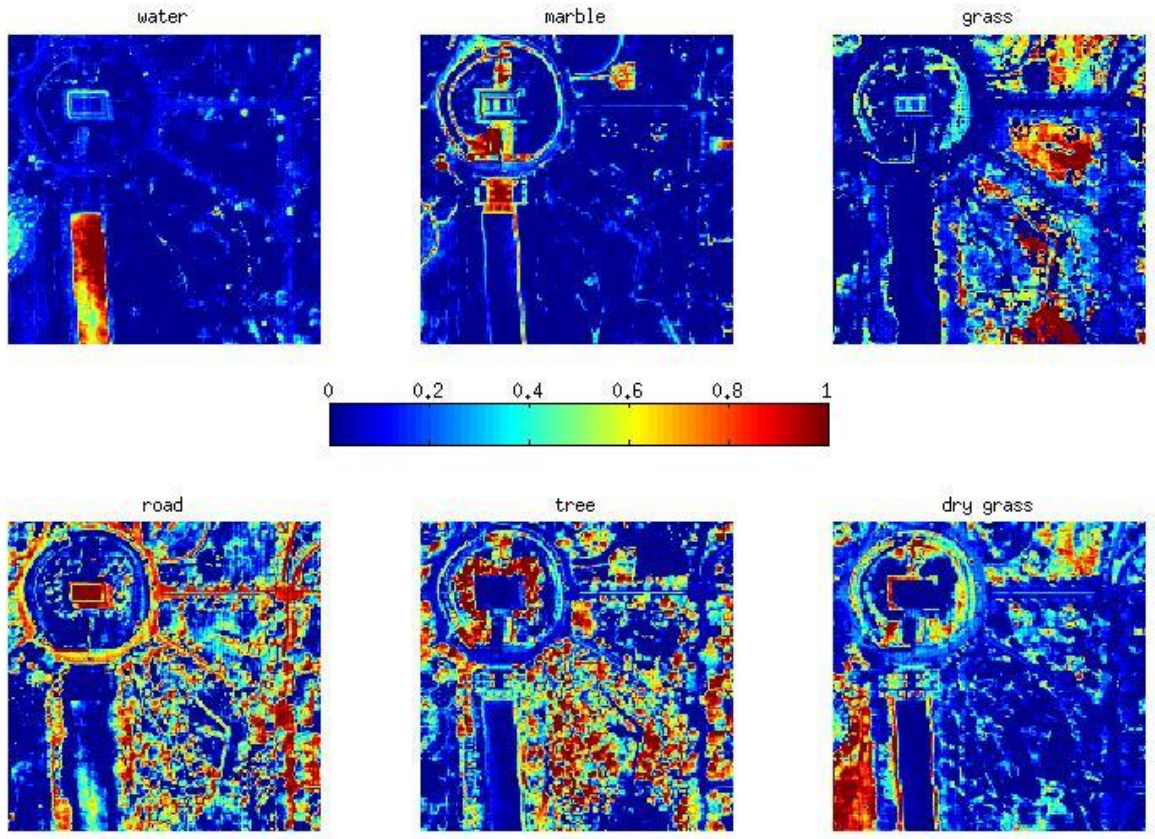
Αν γνωρίζουμε τα $\boldsymbol{x}_i, E$, με βάση το προηγούμενο μοντέλο, μπορούμε να εκτιμήσουμε τα $\boldsymbol{w}_i$ με τη μέθοδο ελαχίστων τετραγώνων, η οποία δίνει

$$\boldsymbol{w}_i = (E^T E)^{-1} E^T \boldsymbol{x}_i, \qquad i = 1,2, \dots, N.$$

Παρατηρήσεις

- Είναι συνήθως $l \gg m$ (ο αριθμός των καναλιών είναι πολύ μεγαλύτερος από τον αριθμό των endmembers). Κατά συνέπεια ο πίνακας $E^T E$ είναι σχεδόν πάντα αντιστρέψιμος.
- Τα $\boldsymbol{x}_i$ είναι γνωστά από την ΥΕ, ενώ ο πίνακας E μπορεί να κατασκευαστεί, α) από την ίδια την εικόνα, επιλέγοντας με προσοχή pure pixels, β) αν διαθέτουμε ground truth, ή γ) από σχετικές φασματικές βιβλιοθήκες υλικών.
- Τα διανύσματα $\boldsymbol{w}_i$ που εκτιμήθηκαν, μπορούν να χρησιμοποιηθούν για την παραγωγή των λεγόμενων χαρτών abundances (abundance maps), έναν για κάθε endmember, που απεικονίζουν τη χωρική κατανομή των διάφορων υλικών/αντικειμένων πάνω στην εικόνα.

Φασματικός διαχωρισμός ΥΕ



Abundance maps obtained using LS spectral unmixing on HYDICE Washington Mall hyperspectral cube.

Φασματικός διαχωρισμός ΥΕ

Επιβολή περιορισμών στο διάνυσμα παραμέτρων $\boldsymbol{w}_i$

- No constraints (conventional LS)

$$\min_{\boldsymbol{w}_i} \|\boldsymbol{x}_i - E\boldsymbol{w}_i\|^2$$

- Sum-to-one constraint

$$\min_{\boldsymbol{w}_i,\lambda} \left[\|\boldsymbol{x}_i - E\boldsymbol{w}_i\|^2 + \lambda \left(\sum_{j=1}^{m} w_{ij} - 1 \right) \right]$$

Λύνεται με χρήση πολλαπλασιαστών Lagrange.

- Nonnegativity constraint

$$\min_{\boldsymbol{w}_i \geq 0} \|\boldsymbol{x}_i - E\boldsymbol{w}_i\|^2$$

Λύνεται με διάφορες επαναληπτικές μεθόδους που σε κάθε επανάληψη επιβάλλουν τον περιορισμό.

- Sparsity constraint

$$\min_{\boldsymbol{w}_i,\lambda} [\|\boldsymbol{x}_i - E\boldsymbol{w}_i\|^2 + \lambda \|\boldsymbol{w}_i\|_0]$$

Λύνεται με convex relaxation (αντικατάσταση της l_0 από την l_1νόρμα).

Βιβλιογραφία

- Σ. Θεοδωρίδης, Κ. Κουτρούμπας, Αναγνώριση Προτύπων, Εκδ. Π.Χ. Πασχαλίδης, Αθήνα, 2011.
- S. Theodoridis, Machine Learning: A Bayesian and Optimization Perspective, 2nd Edition, Academic Press, 2020.