Εξόρυξη Δεδομένων και Αλγόριθμοι Μάθησης.

Τμήμα Μηχανικών Η/Υ και Πληροφορικής
Πανεπιστήμιο Πατρών

29 Μαρτίου 2023

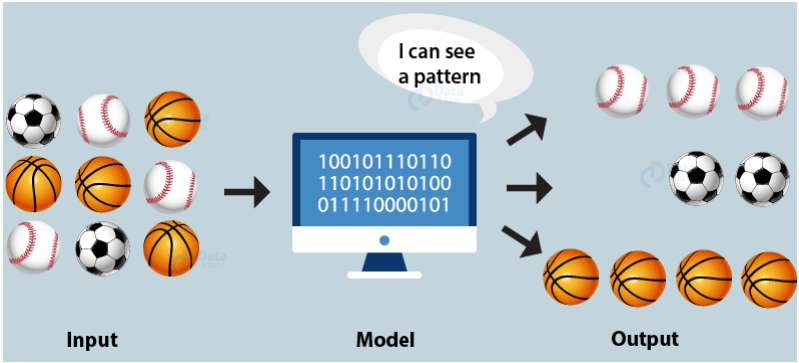
Εισαγωγή (1)

Clustering

Συσταδοποίηση είναι η διαδικασία κατά την οποία ένα σύνολο «αντικείμενων», διαχωρίζονται σε ένα σύνολο από λογικές ομάδες. Η καταχώρηση αντικειμένων σε ίδια ομάδα μεταφράζεται ως ομοιότητα των αντικειμένων αυτών και αντίστροφα.

1. μέθοδος με την οποία μεγάλα σύνολα δεδομένων χωρίζονται σε παρόμοιες ομάδες μικρότερων συνόλων δεδομένων.
2. μέθοδος περιγραφής δεδομένων αλλά και συμπίεσης δεδομένων.

Εισαγωγή (2)


I can see
a pattern
100101110110
110101010100
011110000101
Input
Model
Output

Αλγόριθμος Συσταδοποίησης

- Προσπαθεί να εντοπίσει τις φυσικές ομάδες των στοιχείων με βάση κάποια ομοιότητα.
- Βρίσκει το κέντρο βάρους της ομάδας του συνόλου των δεδομένων.
- Για να καθορίσει τα μέλη της ομάδας, αξιολογεί την απόσταση μεταξύ ενός σημείου και του κεντροειδούς της ομάδας - συστάδας.
- Ως έξοδο, παράγει μια στατιστική περιγραφή των κεντροειδών των συστάδων καθώς και το πλήθος στοιχείων που περιλαμβάνει κάθε ομάδα.

Ποιότητα Συστάδων

Ένας αλγόριθμος συσταδοποίησης είναι καλός αν παράγει συστάδες με

- Μεγάλη ομοιότητα εντός της συστάδας και
- Μικρή ομοιότητα ανάμεσα στις συστάδες

Ποιότητα ομάδας εξαρτάται:

1. Διάμετρο της ομάδας σε σύγκριση με την απόσταση της από τις υπόλοιπες.
2. Απόσταση μεταξύ των σημείων μιας συστάδας από το κέντρο
3. Διάμετρος της μικρότερης συστάδας.

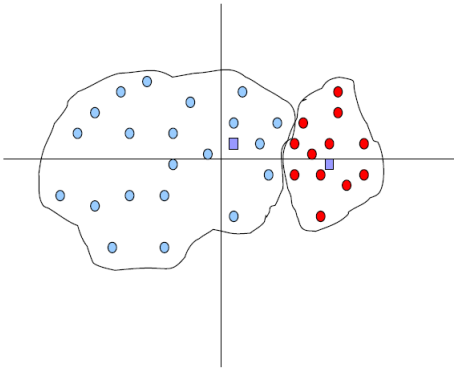
K - Means (1)

- Είσοδος: Σύνολο αντικειμένων, $d-$διαστάσεις και έναν ακέραιο K.
- Στόχος: Εντοπισμός συνόλου σημείων που ελαχιστοποιούν την μέση απόσταση ελαχίστων τετραγώνων κάθε σημείου από το κοντινότερο κέντρο, στον $d-$διάστατο χώρο.
- Μετρική Απόστασης η Ευκλείδεια απόσταση: $d_E(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$

K - Means (2)

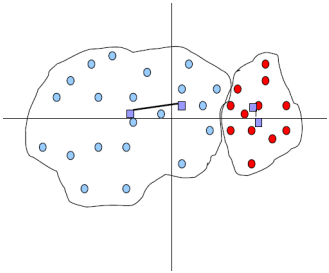
1. Αρχικοποίηση: Τυχαία επιλογή των κ αρχικών σημείων.
2. Για κάθε σημείο εντόπισε το πλησιέστερο κέντρο και ανάθεσε το σημείο στην αντίστοιχη συστάδα.
3. Υπολόγισε τα νέα κέντρα με βάση τα στοιχεία κάθε ομάδας.
4. Επανέλαβε τα βήματα 2 και 3 μέχρι να μην υπάρχει αλλαγή ή η αλλαγή που θα υπάρξει να είναι μικρότερη ενός ορισμένου από τον χρήστη ορίου ή να μέχρι κάποιο κριτήριο σύγκλισης να ικανοποιηθεί.

K - Means: Παράδειγμα 1 (1)



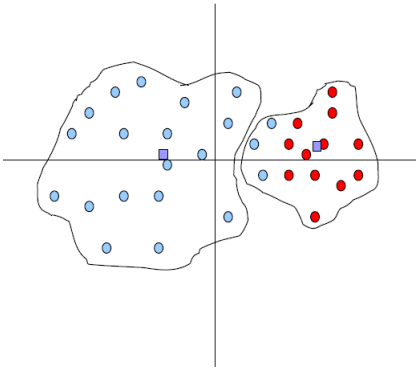
1. Επιλέγονται $k = 2$ κεντροειδή.
2. Ομαδοποίηση με βάση τα επιλεγμένα σημεία.

K - Means: Παράδειγμα 1 (2)



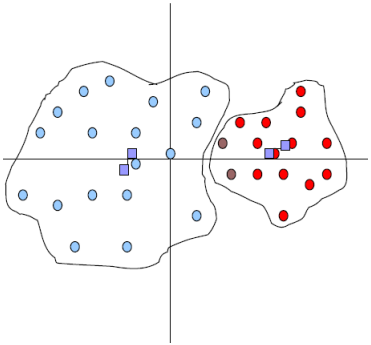
3. Επαναϋπολογισμός κεντροειδών.

K - Means: Παράδειγμα 1 (3)



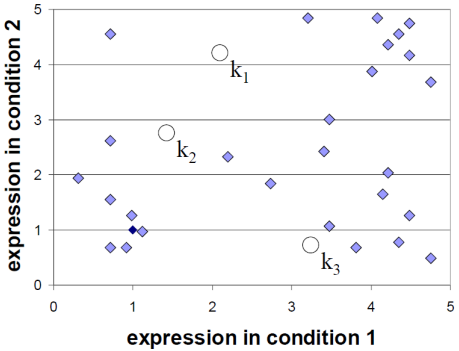
4. Ομαδοποίηση με βάση τα **νέα** επιλεγμένα σημεία.

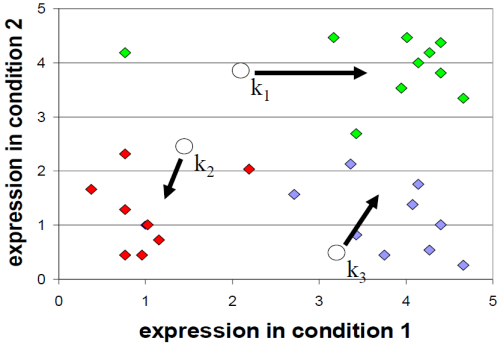
K - Means: Παράδειγμα 1 (4)



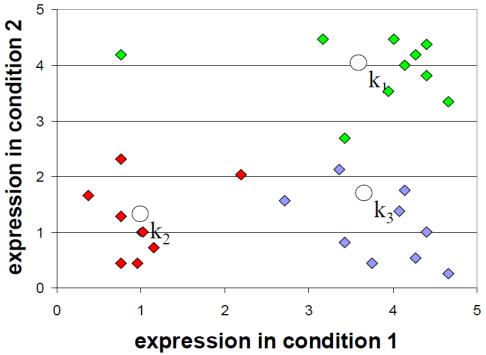
5. Επαναλαμβάνουμε μέχρις ότου να ικανοποιηθεί κάποια από τις συνθήκες σύγκλισης που έχουν οριστεί (π.χ καμία μετακίνηση στοιχείου).

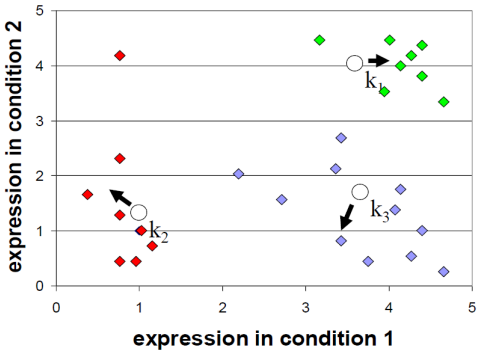
K - Means: Παράδειγμα 2 (1)





K - Means: Παράδειγμα 2 (2)





K - Means: Παρατηρήσεις

- Πολυπλοκότητα: Γραμμική ως προς τον αριθμό σημείων. $O(K \cdot N)$, με k το πλήθος των συστάδων και N το πλήθος των στοιχείων.

Πλεονεκτήματα:

1. Απλότητα.
2. Τα αντικείμενα τοποθετούνται αυτόματα σε κάποια ομάδα.
3. Ταχύτητα σύγκλισης.

Μειονεκτήματα:

1. Τα αποτελέσματα μπορεί να διαφέρουν ανάλογα με την αρχικοποίηση.
2. Κίνδυνος εγκλωβισμού σε τοπικά ελάχιστα.
3. Προκαθορισμένος αριθμός από συστάδες.
4. Όλα τα αντικείμενα πρέπει να ανήκουν σε κάποια ομάδα υποχρεωτικά.
5. Αναγκαιότητα η ύπαρξη αριθμητικών και μονό δεδομένων

K - Windows

- Επέκταση του K - Means αλγορίθμου.
- Στόχος η διαμέριση στοιχείων σε k συστάδες χρησιμοποιώντας την τεχνική των παραθύρων.
- Επιτρέπει την εξέταση μόνο ενός ορισμένου αριθμού στοιχείων σε κάθε επανάληψη.
- Επιτυγχάνει καλύτερη χρονική πολυπλοκότητα και ακρίβεια στον διαχωρισμό.

K - Windows: Παράθυρο

Τι είναι το παράθυρο - Window

Μια ορθογώνια περιοχή στον d - διάστατο Ευκλείδειο χώρο, όπου d ο αριθμός των διαφορετικών αριθμητικών χαρακτηριστικών. Αποτελεί, δηλαδή μία υποπεριοχή μίας σταθερής περιοχής α και έχει συγκεκριμένο μέγεθος.

Η κεντρική ιδέα του αλγορίθμου είναι η προσπάθεια τοποθέτησης ενός παραθύρου d διαστάσεων στον χώρο, ώστε να περιλαμβάνει μόνο τα στοιχεία που ανήκουν σε μία συστάδα.

- Το παράθυρο επιτρέπει τις πράξεις της μετακίνησης και της μεγέθυνσης.

K - Windows: Μετακίνηση - Μεγέθυνση

Μετακίνηση:

1. Κάθε παράθυρο μετακινείται προς το κέντρο της συστάδας.
2. Η μετακίνηση του στον Ε. χώρο γίνεται ανάλογα με τον μέσο όρο των στοιχείων που περιλαμβάνει.

Μεγέθυνση:

1. Χρήση: Βελτίωση της ποιότητας της συστάδας που περιλαμβάνει κάθε παράθυρο.
2. Προσπάθεια αύξησης του παραθύρου για την συμπερίληψη του μέγιστου αριθμού στοιχείων που ανήκουν στη συστάδα.

K - Windows: Αλγόριθμός - Α Φάση

Είσοδος: ο αριθμός των επιλεγμένων κεντροειδών K, η περιοχή α και v το κάτω όριο μέσης τιμής σημείων που αποτελούν μία συστάδα.

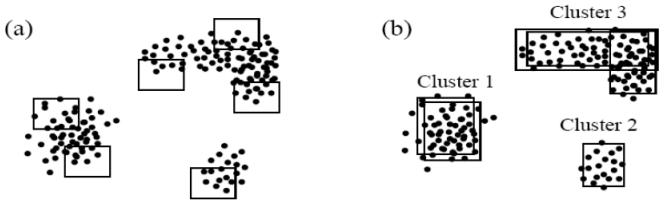
1. k κέντρα επιλέγονται. Και κάθε κέντρο αποτελεί και κέντρο του παραθύρου, που ανήκει στην περιοχή α.
2. Εντοπίζονται τα σημεία που ανήκουν σε κάθε παράθυρο $\rightarrow$ Range Tree.
3. Τα νέα κεντροειδή υπολογίζονται από τα σημεία που ανήκουν στο παράθυρο.
4. Επανάληψη των 2,3 μέχρι να μην υπάρχει αλλαγή.

K - Windows: Αλγόριθμός - Β Φάση

Είσοδος: ο αριθμός των επιλεγμένων κεντροειδών K, η περιοχή α και v το κάτω όριο σημείων που αποτελούν μία συστάδα.

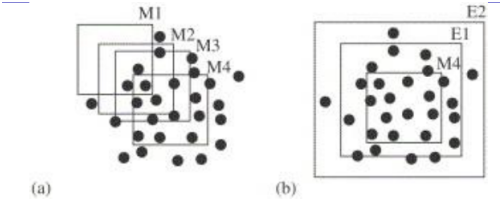
1. Αυξάνονται τα παράθυρα - πράξη μεγέθυνσης - ώστε να περιλαμβάνουν όσο δυνατό περισσότερα στοιχεία της συστάδας διατηρώντας τον μέσο όρο σταθερό*.
2. Αν ο αριθμός στοιχείων που ανήκουν στο παράθυρο μικρότερος του κατωφλίου επαναϋπολογίζεται το παράθυρο ή αυξάνεται η περιοχή α και επαναρχικοποιούνται τα κέντρα για την επανεκκίνηση του αλγορίθμου.

K - Windows: Παράδειγμα (1)



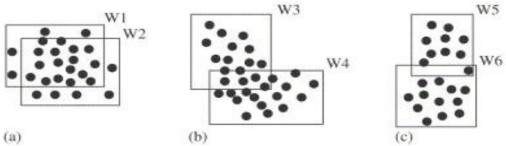
1. 3 ξεχωριστές ομάδες στοιχείων και 6 παράθυρα.
2. Ο αλγόριθμος έχει προσδιορίσει σωστά της συστάδες.

K - Windows: Παράδειγμα (2)



- Πρώτη φάση του αλγορίθμου (σχήμα *a*): Επιλογή κέντρων και μετακινήσεις παραθύρου.
- Δεύτερη φάση του αλγορίθμου (σχήμα *b*): Αύξηση μεγέθους κατά ποσοστό.

Unsupervised K - Windows



- Αρχικοποίηση με μεγάλο αριθμό παραθύρων.
- Συγχώνευση για να καθοριστεί αυτόματα ο αριθμός των συστάδων.
- Για κάθε επικαλυπτώμενο παράθυρο υπολογίζεται το ποσοστό επικάλυψης και αν ξεπερνά ένα όριο τότε το ένα διαγράφεται.

K - Windows: Παρατηρήσεις

Πλεονεκτήματα:

1. Μειώνει τον αριθμό των στοιχείων προς εξέταση για πιθανή ομοιότητα.
2. Μικρή χρονική πολυπλοκότητα.
3. Αποτελέσματα υψηλής ποιότητας.

Μειονεκτήματα:

1. Εξαρτάται και περιορίζεται από τις υπερ-γραμμικές απαιτήσεις του Range Tree και δε μπορεί να εφαρμοστεί σε μεγάλες περιοχές.

K - Means VS K - Windows

K - Means:

1. Πιο διαδεδομένος.
2. Τοπικά ελάχιστα
3. Ανάλογα την αρχικοποίηση μπορεί να γίνει πολύ δαπανηρός.

K - Windows:

1. Καλύτερος χρόνος εκτέλεσης.
2. Καλύτερη ακρίβεια διαίρεσης αντικειμένων.
3. Εντοπίζει τον αριθμό των συστάδων.
4. Παραλληλοποιείται εύκολα.

Ιεραρχική Συσταδοποίηση

Αλγόριθμοι Ιεραρχικής Συσταδοποίησης

Προσπαθούν να δημιουργήσουν μια ιεραρχία συστάδων με μορφή όμοια ενός δέντρου.

Διαχωρίζονται σε:

- Συσσωρευτικούς - Agglomerative
  1. Από κάτω προς τα πάνω προσέγγιση.
  2. Αρχικά, κάθε σημείο αποτελεί μία ξεχωριστή συστάδα.
  3. Σε κάθε βήμα συγχωνεύει το πιο κοντινό ζευγάρι, μέχρις ότου να υπάρχει μία και μόνο μεγάλη συστάδα.
- Διαιρετικούς - Divisive
  1. Από πάνω προς τα κάτω προσέγγιση.
  2. Ξεκινά από μία μεγάλη συστάδα και σε κάθε βήμα την διαιρεί σε μικρότερες.

Απόσταση Μεταξύ των Συστάδων

- Απλή ή μονή σύνδεση (single link)
- Πλήρης σύνδεση (complete link)
- Μέση σύνδεση (average linkage)
- Απόσταση μεταξύ κεντροειδών.

Άσκηση

Πραγματοποιήστε hierachical clustering πλήρους σύνδεσης με τον ακόλουθο πίνακα ευκλείδειων αποστάσεων:

	a	b	c	d	e
a	0	4	7	9	1
b	4	0	3	5	3
c	7	3	0	2	6
d	9	5	2	0	8
e	1	3	6	8	0

Λύση(1)

Πραγματοποιήστε hierachical clustering πλήρους σύνδεσης με τον ακόλουθο πίνακα ευκλείδειων αποστάσεων:
$a \rightarrow e$: Πλησιέστερο σημείο.
Πλήρους σύνδεσης: $d(b, \{a, e\}) = MAX(d(b, a), d(b, e)) = MAX(4, 3) = 4 \cdots$

	ae	b	c	d
ae	0	4	7	9
b	4	0	3	5
c	7	3	0	2
d	9	5	2	0

Λύση(2)

	ae	b	cd
ae	0	4	9
b	4	0	5
cd	9	5	0

Λύση(3)

	abe	*cd*
abe	0	9
cd	9	0