

Υλοποίηση Αλγορίθμου k-means

Στόχος της άσκησης αυτής είναι να υλοποιήσετε τον αλγόριθμο K-means σε οποιαδήποτε γλώσσα προγραμματισμού. Να φτιάξετε μια συνάρτηση:

`[ClusterCenters, IDC] = mykmeans(Data, K)`

Είσοδος

- `TrainData`: ένας πίνακας από N γραμμές (όσα και τα δεδομένα) και M στήλες από πραγματικούς αριθμούς. Κάθε γραμμή αναπαριστά ένα δεδομένο και κάθε στήλη την τιμή ενός χαρακτηριστικού.
- `K`: το πλήθος των ομάδων.
- Η συνάρτηση απόστασης η οποία θα χρησιμοποιήσετε είναι η Ευκλείδεια απόσταση.

Έξοδος

- `ClusterCenters`: ένας πίνακας από K -γραμμές και M -στήλες. Κάθε μια γραμμή αποτελεί το κέντρο της κάθε ομάδας
- `IDC`: ένας πίνακας από N γραμμές ο οποίος αντιστοιχεί στο label της ομάδας για κάθε ένα από τα δεδομένα.

Να δοκιμάσετε το πρόγραμμά σας στα ακόλουθα 2 διαστάσεων δεδομένα: $X = \{X_1, X_2, X_3\}$ όπου $X \sim N(\mu, \Sigma)$:

- $X_1 = (\mu, \Sigma) = \left((4, 0), \begin{pmatrix} 0.29 & 0.4 \\ 0.4 & 4 \end{pmatrix} \right)$
- $X_2 = (\mu, \Sigma) = \left((5, 7), \begin{pmatrix} 0.29 & 0.4 \\ 0.4 & 0.9 \end{pmatrix} \right)$
- $X_3 = (\mu, \Sigma) = \left((7, 4), \begin{pmatrix} 0.64 & 0 \\ 0 & 0.64 \end{pmatrix} \right)$

Να παράγετε συνολικά 150 σημεία δηλαδή 50 σημεία από κάθε κατανομή για τα οποία θα τρέξετε τον αλγόριθμο k-means. Να δημιουργήσετε ένα γράφημα όπου θα φαίνονται όλα τα σημεία στο επίπεδο και να χρωματίσετε τα σημεία από κάθε ομάδα ξεχωριστά (προτείνετε Κόκκινο, Πράσινο, Μπλε) για κάθε βήμα update του αλγορίθμου k-means. Για τα κέντρα της κάθε ομάδας να χρησιμοποιήσετε διαφορετικό σύμβολο π.χ. '+'.

Τέλος να υπολογίσετε το Sum of Squared Error και να δημιουργήσετε ένα γράφημα για κάθε βήμα update του αλγορίθμου k-means :

$$SSE = \sum_{i=1}^K \sum_{x \in C_i} \|x - m_{C_i}\|^2$$

Όπου $C_i, i = 1, \dots, K$ οι K -ομάδες και m_{C_i} το κέντρο της i -ομάδας.

Υποδείξεις:

1. Για αρχικά κέντρα θεωρείστε K σημεία από τα σημεία `TrainData` (η επιλογή να γίνεται τυχαία χωρίς επανατοποθέτηση κάθε φορά που καλείται η συνάρτηση).

2. Θεωρείστε ότι ο αλγόριθμος τερματίζεται όταν τα κέντρα δεν μετακινούνται ή καλύτερα $\|m_C^{new} - m_C^{old}\| \leq \varepsilon$

π.χ. $\varepsilon = 10^{-3}$.

Προσοχή: Να παραδώσετε τον κώδικα καθώς επίσης και μια περιγραφή με τα αποτελέσματά σας.