

Δεύτερη Σειρά Ασκήσεων

Η προθεσμία για την παράδοση αυτού του κομματιού είναι στις 16 Δεκεμβρίου, στο ξεκίνημα του μαθήματος. Για καθυστερημένες υποβολές ισχύει η πολιτική στην σελίδα του μαθήματος. Οι λεπτομέρειες για το turn-in είναι στη σελίδα του μαθήματος. Παραδώσετε το απαιτούμενο υλικό, καθώς και όποιο κώδικα έχετε γράψει, και οδηγίες για το πώς τρέχει ο κώδικας σας.

Ερώτηση 1

Στο βιβλίο του μαθήματος (Εισαγωγή στην Εξόρυξη Δεδομένων, των Tan, Steinbach, και Kumar) εκτός από τους αλγόριθμους clustering που περιγράψαμε στην τάξη, περιγράφονται και οι αλγόριθμοι CLARANS, BIRCH, ROCK, CHAMELEON, DENCLUE και CURE (περιγράφεται επίσης στο δωρεάν online βιβλίο Mining Massive Datasets των Rajaraman και Ullman). Διαλέξτε ένα από τους παραπάνω αλγορίθμους και περιγράψτε την κεντρική ιδέα του με δικά σας λόγια σε 2-3 παραγράφους. Μπορείτε να διαβάσετε και την σχετική δημοσίευση αν χρειάζεστε περισσότερες λεπτομέρειες.

Ερώτηση 2

Σας δίνεται ένα σύνολο από N αντικείμενα τα οποία θέλουμε να ομαδοποιήσουμε. Αντί για πίνακα αποστάσεων, σας δίνεται και ένας $N \times N$ πίνακας, όπου για κάθε ζεύγος (x, y) έχουμε μία τιμή $+1/-1$ (ή true/false) για το αν πρέπει να μπουν στην ίδια ομάδα ή σε διαφορετική. Θα πρέπει να προτείνετε μια μεθοδολογία για την ομαδοποίηση των αντικειμένων. Η μεθοδολογία σας θα πρέπει να έχει τα ακόλουθα δύο βήματα:

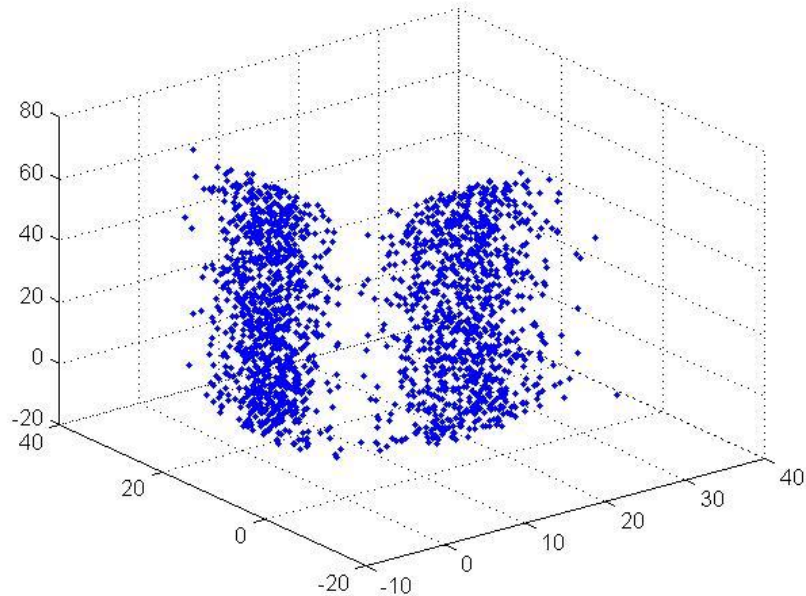
1. Ορίσετε μια συνάρτηση Q , η οποία δεδομένης μιας διαμέρισης C των N χρηστών σε k ομάδες $C = \{c_1, c_2, \dots, c_k\}$ θα υπολογίζει την «ποιότητα» $Q(C)$ της διαμέρισης.
2. Προτείνετε ένα αλγόριθμο ο οποίος προσπαθεί να βρει μια διαμέριση C η οποία βελτιστοποιεί την ποιότητα $Q(C)$. Αν το χρειάζεστε μπορείτε να υποθέσετε ότι ο αριθμός των ομάδων k δίνεται σαν είσοδος στον αλγόριθμο.

Ο σκοπός αυτής της άσκησης είναι να σκεφτείτε πώς να προσεγγίσετε ένα «καινούριο» πρόβλημα. Τετριμμένες απαντήσεις που δείχνουν ελλιπή προσπάθεια θα πάρουν χαμηλή βαθμολογία.

Ερώτηση 3

Για την ερώτηση αυτή συστήνεται η χρήση του MATLAB.

Σας δίνεται το αρχείο “3d-data.txt” (και “3d-data.mat” για απευθείας φόρτωση στο MATLAB) το οποίο περιέχει 2000 τρισδιάστατα σημεία. Η τρισδιάστατη απεικόνιση των σημείων φαίνεται στο παρακάτω σχήμα.



Θα πρέπει να απαντήσετε τις εξής ερωτήσεις:

1. Τι παρατηρείτε στο παραπάνω σχήμα? Υπάρχει κάποια εμφανής ομαδοποίηση για τα δεδομένα?
2. Εφαρμόστε τον k-means αλγόριθμο στα δεδομένα. Δημιουργήστε ένα τρισδιάστατο γράφημα των δύο ομάδων με διαφορετικά χρώματα. Τι παρατηρείτε? Εξηγήστε γιατί.
3. Πάρτε τα δύο πρώτα αριστερά ιδιάζοντα διανύσματα (singular vectors) του πίνακα των δεδομένων, και εφαρμόστε τον k-means στη μειωμένη διάσταση. Δημιουργήστε ένα τρισδιάστατο γράφημα των αρχικών σημείων με τις δύο ομάδες να έχουν διαφορετικά χρώματα. Τι παρατηρείτε? Εξηγήστε γιατί.
4. Υλοποιήστε τον EM αλγόριθμο που περιγράψαμε στην τάξη (για τρισδιάστατα δεδομένα) και εφαρμόστε τον στα τρισδιάστατα δεδομένα και εφαρμόστε τον στα δεδομένα σας. Και πάλι, δημιουργήστε ένα τρισδιάστατο γράφημα των δύο ομάδων με διαφορετικά χρώματα. Τι παρατηρείτε? Εξηγήστε γιατί.

Τεχνικές λεπτομέρειες:

- Για την υλοποίηση του EM σε πολυδιάστατα δεδομένα, πρέπει να υπολογίσετε τις ποσότητες που περιγράψαμε στην τάξη για όλες τις διαστάσεις.
- Οι παρακάτω εντολές στο MATLAB μπορεί να σας φανούν χρήσιμες:
 - `help <command>`: δίνει πληροφορίες για την εντολή `<command>`
 - `D = load 'data.txt'`: φορτώνει το ascii αρχείο με στήλες σε ένα πίνακα D

- `load 'data.mat'`: φορτώνει το `mat` αρχείο και δημιουργεί τους πίνακες που είχαν σωθεί στο `mat` αρχείο
- `kmeans`: υπολογίζει την ομαδοποίηση ενός πίνακα δεδομένων χρησιμοποιώντας τον `k-means` αλγόριθμο.
- `find(X==1)`: επιστρέφει ένα διάνυσμα με τις θέσεις του πίνακα `X` που έχουν τιμή 1.
- `scatter3(D(:,1),D(:,2),D(:,3), 'b')`: Δημιουργεί το παραπάνω `plot`. Το τελευταίο όρισμα ορίζει ότι θέλουμε σημεία ('.') σε μπλε χρώμα ('b').
- `D(x,:)`: Επιστρέφει ένα πίνακα με τις γραμμές από το `D` που προσδιορίζονται στο διάνυσμα `x`
- `svds`: Υπολογίζει το Singular Value Decomposition ενός πίνακα
- `hold`: «Κρατάει» ένα σχήμα ώστε η επόμενη εντολή να τυπώσει πάνω στο ίδιο σχήμα (το πιο πρόσφατο).

Ερώτηση 4

Χρησιμοποιήστε τα δεδομένα από την Ερώτηση 2 της Άσκησης 1β, και κάνετε ομαδοποίηση των μαθημάτων. Για την ομαδοποίηση μπορείτε να χρησιμοποιήσετε κάποια υπάρχουσα υλοποίηση κάποιου από τους αλγόριθμους που είδαμε στην τάξη (υπάρχουν δείκτες στη σελίδα του μαθήματος), ή μια δικιά σας υλοποίηση. Για να βρείτε τον «σωστό» αριθμό από ομάδες χρησιμοποιήστε κάποια από τις τεχνικές που περιγράψαμε στην τάξη. Εξετάστε τις ομάδες που δημιουργούνται και προσπαθήστε να ερμηνεύσετε τα αποτελέσματα. Γράψετε μια αναφορά στην οποία θα περιγράψετε την μεθοδολογία σας και τα συμπεράσματά σας.

Τεχνικές Λεπτομέρειες

1. Για το pre-processing, οι παρακάτω `unix` εντολές μπορεί να σας φανούν χρήσιμες:
 - a. `cut`: επιτρέπει να πάρουμε συγκεκριμένες κολώνες από ένα αρχείο με διαχωριζόμενες τιμές
 - b. `sort`: ταξινομεί τις γραμμές ενός αρχείου σε αλφαβητική σειρά. `-n` for αριθμητική σειρά
 - c. `uniq`: αφαιρεί συνεχόμενες γραμμές που είναι ίδιες.
2. Για την υλοποίηση του LSH δεν χρειάζεται να υλοποιήσετε δικό σας `hash table` ή `list`. Χρησιμοποιείτε υπάρχουσες υλοποιήσεις που έρχονται με τη γλώσσα που χρησιμοποιείτε (π.χ., στην C++ υπάρχει η STL υλοποίηση). Επίσης ίσως είναι βολικό όταν κάνετε `hash` τα `mini-signatures` να τα μετατρέψετε σε `strings` με '+' και '-' χαρακτήρες.
3. Αν και δεν είναι αποτελεσματικό, προς χάριν απλότητας μπορείτε να δημιουργήσετε τον πλήρη πίνακα φοιτητών και βαθμών. Θα πάρετε `bonus` βαθμούς για μια πιο αποτελεσματική υλοποίηση.