Πανεπιστήμιο Ιωαννίνων,
Τμήμα Μηχανικών Πληροφορικής

Εξόρυξη Δεδομένων – ΠΛΕ 059
Χειμερινό Εξάμηνο, 2013

Δεύτερη Σειρά Ασκήσεων

Η σειρά πρέπει να παραδοθεί στην αρχή του μαθήματος της 17ης Δεκεμβρίου. Για καθυστερημένες υποβολές ισχύει η πολιτική στην σελίδα του μαθήματος. Οι λεπτομέρειες για το turn-in είναι στη σελίδα του μαθήματος.

Ερώτηση 1 (Μετρικές απόστασης)

Θεωρείστε ένα σύμπαν U από αντικείμενα και μια συνάρτηση απόστασης $dist: U \times U \to \mathbb{R}$ έτσι ώστε για κάθε ζεύγος από αντικείμενα x, y μπορούμε να υπολογίσουμε την απόσταση τους $dist(x, y)$. Η συνάρτηση της απόστασης $dist$ είναι μια **μετρική**. Θεωρείστε ένα σύνολο $S \subseteq U$ από N αντικείμενα. Έστω ότι $m \in U$ είναι το αντικείμενο **στο U** τέτοιο ώστε $m = \arg\min_{y \in U} \sum_{x \in S} dist(y, x)$. Για παράδειγμα, στην περίπτωση που το U είναι το $\mathbb{R}^d$ και $dist$ είναι η Ευκλείδεια απόσταση, το m η μέση τιμή των σημείων στο S. Επίσης έστω y^* είναι το αντικείμενο **στο S** έτσι ώστε $y^* = \arg\min_{y \in S} \sum_{x \in S} dist(y, x)$. Αποδείξτε ότι

$$\sum_{x \in S} dist(x, y^*) \leq 2 \cdot \sum_{x \in S} dist(x, m)$$

Η εύρεση του m είναι ένα NP-hard πρόβλημα για κάποιες μετρικές, ενώ η εύρεση του y^* είναι εύκολη. Η παραπάνω ανισότητα δείχνει ότι αν χρησιμοποιήσουμε το y^* αντί για το m βρίσκουμε ένα αντικείμενο με άθροισμα αποστάσεων κοντά στο ελάχιστο.

Hint: Θα να είναι χρήσιμο να θεωρήσετε το αντικείμενο $y_m \in S$ που είναι το πιο κοντινό στο αντικείμενο m.

Ερώτηση 2 (Min-Hashing)

Κάνετε την **Άσκηση 3.3.2** από το βιβλίο Mining Massive Datasets των Anand Rajaraman and Jeff Ullman. Παραδώστε τα ενδιάμεσα αποτελέσματα του υπολογισμού της Min-Hashing υπογραφής (όπως κάναμε στην τάξη και όπως κάνουν στο βιβλίο), την τελική υπογραφή και με τις τέσσερεις συναρτήσεις, και την εκτιμώμενη και πραγματική ομοιότητα μεταξύ όλων των ζευγών από στήλες.

Ερώτηση 3 (Clustering)

Σε αυτή την άσκηση θα υλοποιήσετε τον k-means αλγόριθμο που περιγράψαμε στην τάξη. Μπορείτε να χρησιμοποιήσετε οποιαδήποτε προγραμματιστική γλώσσα για την υλοποίηση σας. Εφαρμόστε τον αλγόριθμο στα iris, και spambase datasets που είναι στη σελίδα Υλικό του μαθήματος. Οι κλάσεις δεν θα πρέπει να είναι μέρος της εισόδου όταν τρέχετε τον αλγόριθμο. Ο αριθμός των clusters θα πρέπει να είναι ίσος με τον αριθμό των κλάσεων στα δεδομένα. Δημιουργείστε τον πίνακα σύγχυσης μεταξύ κλάσεων και clusters, και υπολογίστε μία από τις μετρικές αξιολόγησης που χρησιμοποιούν τις κλάσεις. Παραδώσετε τον κώδικα σας, την έξοδο του clustering, τον πίνακα σύγχυσης και τη μετρική αξιολόγησης, και μία σύντομη αναφορά για τα αποτελέσματα σας.

Ερώτηση 4 (Clustering)

Ο στόχος αυτής της άσκησης είναι να πειραματιστείτε με την εφαρμογή του k-means αλγόριθμου σε πραγματικά δεδομένα. Για την ερώτηση αυτή θα χρησιμοποιήσετε το dataset με τα FourSquare tips που χρησιμοποιήσατε στην πρώτη σειρά ασκήσεων. Επιπλέον της προ-εξεργασίας που κάνατε στην Άσκηση 1, θα υπολογίσετε επίσης το tf-idf σκορ για κάθε διακριτή λέξη που εμφανίζεται σε ένα tip (ο ορισμός του tf-idf σκορ δόθηκε στην δεύτερη διάλεξη). Κάθε tip θα αναπαρίσταται ως ένα διάνυσμα με τα tf-idf σκορ των λέξεων που εμφανίζονται στο tip.

Μπορείτε να χρησιμοποιήσετε την δική σας υλοποίηση του k-means από την προηγούμενη ερώτηση, ή κάποια άλλη (υπάρχει υλοποίηση του k-means στο WEKA, και στο MATLAB). Για να βρείτε τον αριθμό των clusters δημιουργείστε την SSE καμπύλη και διαλέξτε την τιμή του k που φαίνεται «σωστή». Για κάθε cluster που δημιουργείτε, πάρετε το κέντρο του cluster και εκτυπώσετε τις top-20 λέξεις (ή παραπάνω αν χρειάζεται) με το μεγαλύτερο σκορ. Εξετάσετε τις λέξεις και προσπαθήστε να επιχειρηματολογήσετε ως προς το τι αντιπροσωπεύει το κάθε cluster.

Παραδώστε τον κώδικα σας και μια αναφορά όπου θα περιγράφετε τα αποτελέσματα σας. Η αναφορά θα πρέπει να περιέχει την SSE καμπύλη, μια συζήτηση για την τιμή του k, τις top λέξεις των κέντρων, και την εκτίμηση σας για το τι αντιπροσωπεύει το κάθε cluster.