Τρίτη Σειρά Ασκήσεων

Η προθεσμία για την τρίτη σειρά ασκήσεων είναι στις 21 Μαΐου στο τέλος της μέρας. Κάνετε turn-in τον κώδικα σας, με οδηγίες πώς για το πώς τρέχει. Η αναφορά θα πρέπει να έχει λεπτομερείς παρατηρήσεις για τα αποτελέσματα σας. Για καθυστερημένες υποβολές ισχύει η πολιτική στην σελίδα του μαθήματος. Λεπτομέρειες για το turn-in στη σελίδα Ασκήσεις του μαθήματος. Θα γίνει προφορική εξέταση των ασκήσεων.

Ερώτηση 1

Μία power-law κατανομή ορίζεται ως $P(X = x) = (a-1)x^{-a}$, όπου a είναι ο εκθέτης της κατανομής. Σας δίνεται ένα σύνολο από παρατηρήσεις $X = \{x_1, \dots, x_n\}$ που έχουν παραχθεί από μία power-law κατανομή. Χρησιμοποιήστε την Maximum Likelihood Estimation τεχνική που περιγράψαμε στην τάξη για να βρείτε τον εκθέτη της power-law κατανομής που ταιριάζει (fits) τα δεδομένα των παρατηρήσεων.

Ερώτηση 2

Θεωρείστε το παρακάτω πρόβλημα:

Δεδομένου ενός συνόλου X με n πραγματικούς αριθμούς και ενός ακεραίου k, βρείτε μια διαμέριση του X σε k ομάδες $\{X_1, X_2, \dots, X_k\}$, όπου $\cup_i X_i = X$ και $X_i \cap X_j = \emptyset$, για κάθε $i, j, i \neq j$, ώστε το άθροισμα των διαμέτρων των ομάδων $\sum_{i=1}^{k} diam(X_i)$ ελαχιστοποιείται, όπου η διάμετρος της ομάδας X_i $diam(X_i) = \max_{x,y \in X_i} |x - y|$, είναι η μεγαλύτερη διαφορά μεταξύ δύο αριθμών στην ομάδα X_i.

Σχεδιάστε ένα **βέλτιστο** αλγόριθμο για το πρόβλημα που τρέχει σε γραμμικό χρόνο ως προς τον αριθμό των σημείων. Αποδείξτε ότι ο αλγόριθμος σας είναι βέλτιστος.

Ερώτηση 3 (Clustering)

Για την ερώτηση αυτή θα εξασκηθείτε στην εφαρμογή αλγορίθμων ομαδοποίησης (clustering).

Θα χρησιμοποιήσετε το αρχείο “clinton_trump_tweets.txt” (**Προσοχή**: Είναι καινούριο dataset) που μπορείτε να κατεβάσετε από τη σελίδα του μαθήματος. Κάνετε την ίδια επεξεργασία όπως και στην Άσκηση 2: αφαιρέστε τα retweets, και κάνετε επαναληπτικό κλάδεμα ώστε τελικά να έχουμε χρήστες που έχουν χρησιμοποιήσει τουλάχιστον 20 διαφορετικά hashtags/handles, και hashtags/handles με που έχουν χρησιμοποιηθεί από τουλάχιστον 20 διαφορετικούς χρήστες. Για άσκηση αυτή θα χρησιμοποιήσουμε για κάθε και την συχνότητα με την οποία ένας χρήστης χρησιμοποιεί ένα hashtag/handle.

Θα εξετάσουμε δύο διαφορετικά clustering προβλήματα.

1. Στο πρώτο πρόβλημα θα κοιτάξετε το clustering των hashtags/handles. Αναπαραστήστε κάθε hashtag/handle ως ένα διάνυσμα με τον αριθμό των εμφανίσεων του hashtag/handle για κάθε χρήστη. Μπορείτε να χρησιμοποιήσετε τις βιβλιοθήκες της python για feature extraction για να κατασκευάσετε την αναπαράσταση.

Αρχικά θα εφαρμόσετε τον k-means αλγόριθμο. Δημιουργήστε την γραφική παράσταση του SSE λάθους ως συνάρτηση του αριθμού των clusters για k μέχρι 20 για να αποφασίστε τον αριθμό των clusters. Τρέξετε τον αλγόριθμο για τον αριθμό που θα επιλέξετε και εξετάστε με το μάτι τα clusters που δημιουργούνται. Προσπαθήστε από τα hashtags/handles που μπαίνουν σε κάθε cluster να συμπεράνετε ποιο θέμα αφορά. Περιγράψετε τα συμπεράσματα σας στην αναφορά.

Στη συνέχεια τρέξετε τον agglomerative αλγόριθμο για ιεραρχικό clustering με τον ίδιο αριθμό από clusters, καθώς και τον DBSCAN αλγόριθμο. Δημιουργήστε το confusion matrix για κάθε ζεύγος από αλγορίθμους και σχολιάστε κατά πόσο οι διαφορετικοί αλγόριθμοι βρίσκουν παρόμοια αποτελέσματα.

2. Στο δεύτερο πρόβλημα θα κοιτάξετε το clustering των χρηστών. Αναπαραστήστε κάθε user ως ένα διάνυσμα με τον αριθμό των φορών που ο χρήστης έχει χρησιμοποιήσει το κάθε hashtag/handle. Μπορείτε να χρησιμοποιήσετε τις βιβλιοθήκες της python για feature extraction για να κατασκευάσετε την αναπαράσταση.

Ο στόχος μας στο δεύτερο κομμάτι είναι να εξετάσουμε την λύση που μας δίνει το clustering ως προς κάποιο γνωστό ground truth. Στο αρχείο “clinton_trump_user_classes.txt” που σας δίνεται στη σελίδα του μαθήματος έχουμε την «κλάση» του για κάθε user id που εμφανίζεται στα δεδομένα μας. Κλάση 0 αντιστοιχεί σε followers του Trump, ενώ κλάση 1 αντιστοιχεί σε followers της Clinton.

Τρέξετε τον k-means αλγόριθμο και τις τέσσερεις διαφορετικές παραλλαγές του agglomerative hierarchical clustering αλγορίθμου (single-link, complete-link, average, Ward) για 2 clusters. Υπολογίσετε το confusion matrix και το precision, recall και F-measure για κάθε αλγόριθμο και συγκρίνεται την απόδοση τους.

Για τον k-means κοιτάξετε τα δύο κέντρα και εξετάσετε τα 30 hashtags/handles με τις μεγαλύτερες τιμές. Μπορούμε να βγάλουμε κάποιο συμπέρασμα από τα πιο συχνά hashtags/handles σε κάθε cluster για το τι κάνει τα δύο cluster να ξεχωρίζουν?

Παραδώσετε τον κώδικα σας, καθώς και όλες τις γραφικές παραστάσεις, τα αποτελέσματα και την συζήτηση πάνω σε αυτά.

Bonus: Για τους χρήστες που πρέπει να κάνετε cluster στο δεύτερο κομμάτι της ερώτησης μπορείτε να πάρετε κι άλλη πληροφορία από το αρχείο με τα tweets. Χρησιμοποιήστε αυτή την πληροφορία για να εξάγετε επιπλέον χαρακτηριστικά ώστε να βελτιώσετε το αποτέλεσμα του clustering. Συγκρίνετε με την αρχική σας λύση ως προς το precision, recall και F-measure.