Τέταρτη Σειρά Ασκήσεων

Η προθεσμία για την παράδοση της Τέταρτης Σειράς Ασκήσεων είναι στις 12 Φεβρουαρίου, μέχρι το τέλος της ημέρας. Παραδώστε ηλεκτρονικά τον κώδικα και την αναφορά σας, είτε μέσω turnin είτε μέσω email. Για καθυστερημένες υποβολές ισχύει η πολιτική στην σελίδα του μαθήματος. Λεπτομέρειες για το turn-in, και για το πώς να γράφετε αναφορές είναι στη σελίδα Ασκήσεις του μαθήματος.

Ερώτηση 1

Έστω $U = \{x_1, \dots, x_N\}$ ένα σύμπαν από στοιχεία. Ορίζουμε ένα σταθμισμένο υποσύνολο S του U ως ένα N-διάστατο διάνυσμα W_S, όπου $W_S[x]$ είναι το βάρος του στοιχείου x στο S, αν $x \in S$, και μηδέν αν $x \notin S$. Έστω $C = \{S_1, \dots, S_k\}$ μια συλλογή από σταθμισμένα υποσύνολα του U. Ορίζουμε το διάνυσμα βάρους της συλλογής C ως ένα N-διάστατο διάνυσμα W_C, όπου $W_C[x] = \max_{S \in C} W_S[x]$.

Δεδομένης μια συλλογής C ορίζουμε την συνάρτηση $f(C) = \sum_{x \in U} W_C[x]$ να είναι το συνολικό βάρος του C για όλα τα στοιχεία. (Προσέξτε ότι η συνάρτηση ορίζεται και για ένα μόνο σύνολο, το οποίο είναι μια ειδική περίπτωση της συλλογής). Αποδείξτε ότι η συνάρτηση f είναι submodular, δηλαδή, για οποιεσδήποτε δύο συλλογές A, B, όπου $A \subseteq B$, και για οποιοδήποτε σύνολο $S \notin B$, $f(A \cup \{S\}) - f(A) \geq f(B \cup \{S\}) - f(B)$.

Ερώτηση 2

Έστω $G = (V, E)$ ένα μη κατευθυνόμενο γράφημα το οποίο αναπαριστά ένα κοινωνικό δίκτυο. Μία τριάδα από κόμβους $\{u, v, w\}$, ορίζει ένα **ανοιχτό τρίγωνο**, αν $(u, v) \in E$ και $(v, w) \in E$ αλλά $(u, w) \notin E$. Θεωρείστε ότι οι ακμές του γραφήματος έχουν ετικέτες, είτε Strong (Ισχυρή) είτε Weak (Αδύναμη). Λέμε ότι μία ανάθεση ετικετών στις ακμές του γραφήματος ικανοποιεί την **Strong Triadic Closure ιδιότητα** αν για κάθε ανοιχτό τρίγωνο $\{u, v, w\}$, τουλάχιστον μία από τις ακμές (u, v), (v, w) έχει την ετικέτα Weak. Διαισθητικά ο ορισμός αυτός λέει ότι δεν μπορεί ο v να έχει ισχυρές σχέσεις και με τον u και με τον w, αλλά ο u και ο w να μην γνωρίζονται.

Μας δίνεται σαν είσοδος το γράφημα $G = (V, E)$ χωρίς ετικέτες στις ακμές. Θέλουμε να βρούμε μια ανάθεση ετικετών σε όλες τις ακμές ώστε να ικανοποιείται η Strong Triadic Closure property, **και** ο αριθμός των Weak ετικετών να ελαχιστοποιείται. Με άλλα λόγια, θέλουμε να βρούμε το μικρότερο υποσύνολο από ακμές $S \subseteq E$, έτσι ώστε αν τους αναθέσουμε την ετικέτα Weak, όλα τα ανοιχτά τρίγωνα θα ικανοποιούν την απαίτηση της Strong Triadic Closure ιδιότητας.

Δείξτε πως αυτό το πρόβλημα μπορεί να αναχθεί σε ένα πρόβλημα κάλυψης. Με βάση την αναγωγή προτείνετε ένα προσεγγιστικό αλγόριθμο και προσδιορίστε τον λόγο προσέγγισης.

Ερώτηση 3

Σε αυτή την ερώτηση θα χρησιμοποιήσετε τα δεδομένα που δημιουργήσατε για την Δεύτερη Σειρά Ασκήσεων, για τα συστήματα συστάσεων. Ο στόχος είναι να χρησιμοποιήσουμε το κοινωνικό δίκτυο μεταξύ των χρηστών του Yelp για να προβλέψουμε τα ratings τους για νέες επιχειρήσεις.

Ξεκινήστε με τα δεδομένα από την Δεύτερη Σειρά, τα οποία αποτελούνται από τις επιχειρήσεις στο Las Vegas που έχουν τουλάχιστον 10 κριτικές από χρήστες με τουλάχιστον 10 κριτικές. Χρησιμοποιώντας αυτούς τους χρήστες ως τις κορυφές, δημιουργείστε ένα γράφημα με ακμές τις φιλίες μεταξύ των χρηστών, τις οποίες θα πάρετε από το αρχείο yelp_academic_dataset_user.json . Από αυτό το γράφημα κρατήστε τη μεγαλύτερη συνεκτική συνιστώσα. Αυτή θα ορίσει το γράφημα G με το οποίο θα δουλέψετε, και οι κόμβοι της συνιστώσας το σύνολο των χρηστών που μας ενδιαφέρουν (το σύνολο των επιχειρήσεων παραμένει το ίδιο).

Αφαιρέστε τυχαία 10% των ratings χρηστών και προσπαθήστε να προβλέψετε το rating για το ζευγάρι χρήστη-επιχείρηση (u, b) χρησιμοποιώντας ένα τυχαίο περίπατο με απορροφητικούς κόμβους, ως εξής: Δεδομένου του ζεύγους (u, b) και το γράφημα G, κάνετε κάθε κόμβο v ο οποίος έχει δώσει rating για την επιχείρηση b να είναι απορροφητικός, και αναθέστε του τιμή ίση με το rating $R(v, b)$. Χρησιμοποιώντας την τεχνική για την διάχυση (propagation) τιμών που περιγράψαμε στην τάξη, υπολογίστε ένα rating $P(v', b)$ για κάθε μη-απορροφητικό κόμβο v'στο γράφημα. Η πρόβλεψη για τον κόμβο u θα είναι η τιμή $P(u, b)$.

Συγκρίνεται το Μέσο Άθροισμα Τετραγώνων Λαθών (MSSE) για αυτή τη μέθοδο, με το MSSE που παίρνετε με τις τεχνικές collaborative filtering με βάση τον χρήστη και με βάση το αντικείμενο, τις οποίες υλοποιήσατε για την Δεύτερη Σειρά (μέθοδοι 3,4 από την Ερώτηση 2).

Υπενθύμιση: Το MMSE ορίζεται ως:

$$SSE = \frac{1}{n}\sum_{i=1}^{n}(r_i - p_i)^2$$

Όπου r_i είναι το πραγματικό rating και p_i είναι η πρόβλεψη.

Bonus: Προτείνετε, υλοποιήστε και τεστάρετε μια διαφορετική μέθοδο που να προβλέπει τα ratings των χρηστών χρησιμοποιώντας τυχαίους περιπάτους στο κοινωνικό γράφημα.

Ερώτηση 4

Στο Kaggle υπάρχει ένας ενεργός διαγωνισμός από το AirBnB για την πρόβλεψη της πρώτης κράτησης ενός νέου χρήστη (εδώ το link του διαγωνισμού). Χρησιμοποιώντας τον λογαριασμό που δημιουργήσατε στην προηγούμενη άσκηση, υποβάλετε μια λύση στον διαγωνισμό. Ο στόχος δεν είναι να κερδίσετε τον διαγωνισμό (αν και πάλι, η θέση σας στον διαγωνισμό θα προσμετρηθεί για βαθμούς bonus), αλλά να δουλέψετε σε ένα πραγματικό πρόβλημα που δεν έχει γνωστή λύση.

Δημιουργήστε μια αναφορά που θα περιέχει τα παρακάτω:

a. Μια περιγραφή για το πώς προσεγγίσατε το πρόβλημα (το είδατε σαν πρόβλημα κατηγοριοποίησης, ομαδοποίησης, συχνά στοιχειοσύνολα, κλπ).

b. Μια περιγραφή της λύσης που υλοποιήσατε. Τι χαρακτηριστικά χρησιμοποιήσατε, ποιες τεχνικές. Μια σύντομη περιγραφή της λογικής πίσω από τις επιλογές σας.
c. Τα αποτελέσματα σας στο Kaggle test dataset.
d. Ένα σχολιασμό στα παραπάνω: Τι δουλεύει και γιατί? Τι καταλάβατε για τα δεδομένα και το πρόβλημα?

Παραδώσετε το κώδικα σας και την αναφορά. Αναφέρετε το όνομα σας στο Kaggle, και την θέση σας στην κατάταξη όταν κάνατε την υποβολή.