Πανεπιστήμιο Ιωαννίνων,
Τμήμα Μηχανικών Πληροφορικής

Εξόρυξη Δεδομένων – ΜΥΕ012/ΠΛΕ 059
Χειμερινό Εξάμηνο, 2015

Δεύτερη Σειρά Ασκήσεων

Η προθεσμία για την παράδοση της Δεύτερης Σειράς Ασκήσεων είναι στις 6 Δεκεμβρίου, 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 (Μετρικές απόστασης)

Δεδομένου ενός σύμπαντος από N αντικείμενα $U = \{x_1, x_2, \dots, x_N\}$, μια ιεράρχηση (ranking) του συνόλου U είναι μια ταξινόμηση των στοιχείων του U από το πρώτο προς το τελευταίο. Μαθηματικά, ένα ranking ορίζεται ως μια 1-προς-1 συνάρτηση $R: U \to \{1, .., N\}$, όπου $R(x_i)$ είναι η σειρά (rank) του στοιχείου x_i, π.χ., αν $R(x_i) = 2$, τότε το στοιχείο x_i είναι το δεύτερο στοιχείο στο ranking. Θα χρησιμοποιούμε $\mathcal{R}$ για να αναπαριστούμε το σύνολο όλων των $N!$ rankings (ο αριθμός όλων των πιθανών αναδιατάξεων).

Σε αυτή την ερώτηση:

- Θα πρέπει να ορίσετε μια συνάρτηση απόστασης $d: \mathcal{R} \times \mathcal{R} \to \mathbb{R}$ μεταξύ δύο rankings. Ο ορισμός σας θα πρέπει να είναι «λογικός» ώστε οι αποστάσεις να αντικατοπτρίζουν το πόσο διαφορετικά είναι δύο rankings; Τετριμμένοι ορισμοί δεν θα βαθμολογηθούν.
- Αποδείξτε ότι η συνάρτηση σας ορίζει ή δεν ορίζει μια μετρική.

Ερώτηση 2 (Συστήματα συστάσεων)

Ο στόχος αυτής της άσκησης είναι να πειραματιστείτε με αλγορίθμους για συστήματα συστάσεων.

Θα χρησιμοποιήσετε το Yelp dataset που χρησιμοποιήσατε και για την πρώτη σειρά ασκήσεων. Σε αυτή την άσκηση θα χρησιμοποιήσετε τα αρχεία `yelp_academic_dataset_business.json` και `yelp_academic_dataset_review.json`. Χρησιμοποιώντας αυτά τα δεδομένα θα δημιουργήσετε ένα user-business πίνακα με τα ratings των χρηστών για όλες τις επιχειρήσεις στην πόλη του "Las Vegas". Κρατήσετε μόνο τους χρήστες που έχουν κάνει τουλάχιστον 10 ratings, και τις επιχειρήσεις που έχουν δεχτεί τουλάχιστον 10 ratings (επαναλάβετε αυτή τη διαδικασία επαναληπτικά μέχρι όλοι οι χρήστες και όλες οι επιχειρήσεις στον πίνακα σας να έχουν τουλάχιστον 10 ratings). Μετά, κανονικοποιήστε τον πίνακα αφαιρώντας από κάθε γραμμή την μέση τιμή των ratings του χρήστη.

Αφαιρέστε τυχαία ένα 10% των ratings. Ο στόχος είναι να υπολογίσετε αυτά τα ratings εφαρμόζοντας τις τεχνικές collaborative filtering που μάθαμε στην τάξη χρησιμοποιώντας το υπόλοιπο 90% των δεδομένων. Θα δοκιμάσετε τους παρακάτω αλγόριθμους για να προβλέψετε το rating του χρήστη u για την επιχείρηση b:

1. Χρησιμοποιήστε την μέση τιμή $\overline{r(u)}$ των ratings του u για την πρόβλεψη (πριν την κανονικοποίηση).

2. Χρησιμοποιήστε την μέση τιμή $\overline{r(b)}$ των ratings της επιχείρησης b για την πρόβλεψη (πριν την κανονικοποίηση).
3. Για τον χρήστη u υπολογίστε το σύνολο $N_k(u)$ με τους k πιο όμοιους χρήστες (σύμφωνα με το cosine similarity) οι οποίοι έχουν βαθμολογήσει την επιχείρηση b. Στη συνέχεια χρησιμοποιήστε την εξής εξίσωση για την πρόβλεψη σας:
$$p(u,b) = \overline{r(u)} + \frac{\sum_{u' \in N_k(u)} s(u,u')(r(u',b) - \overline{r(u')})}{\sum_{u' \in N_k(u)} s(u,u')}$$
4. Για την επιχείρηση b υπολογίστε το σύνολο $N_k(b)$ με τα k πιο όμοιες επιχειρήσεις (σύμφωνα με το cosine similarity) οι οποίες έχουν βαθμολογηθεί από τον χρήστη u. Στη συνέχεια χρησιμοποιήστε την εξής εξίσωση για την πρόβλεψη σας:
$$p(u,b) = \overline{r(u)} + \frac{\sum_{b' \in N_k(m)} s(b,b')(r(u,b') - \overline{r(u)})}{\sum_{b' \in N_k(b)} s(b,b')}$$
5. Εφαρμόστε το Singular Value Decomposition στον κανονικοποιημένο πίνακα R, και κρατήστε τα k μεγαλύτερα singular vectors για να πάρετε ένα rank-k πίνακα R_k. (Χρησιμοποιήστε τις singular τιμές για να αποφασίσετε το μέγεθος του k). Στη συνέχεια χρησιμοποιήστε την εξής εξίσωση για την πρόβλεψη σας: $p(u,b) = \overline{r(u)} + R_k(u,b)$

Για την αξιολόγηση και σύγκριση των αλγορίθμων θα χρησιμοποιήσετε το Άθροισμα Τετραγώνων των Λαθών. Αν $r_1, r_2, \dots, r_n$ είναι τα ratings που θέλουμε να προβλέψουμε, και $p_1, p_2, \dots, p_n$είναι οι προβλέψεις του αλγορίθμου, το Άθροισμα Τετραγώνων των Λαθών του αλγορίθμου ορίζεται ως

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

Θα πρέπει να παραδώσετε τα ακόλουθα:

- Όλο τον κώδικα που θα γράψετε εσείς.
- Το αρχείο με το 10% που προσπαθείτε να προβλέψετε και το υπόλοιπο 90%
- Μια αναφορά που θα περιγράφει τι κάνατε, θα συγκρίνει τους διαφορετικούς αλγορίθμους, και θα σχολιάζει τα αποτελέσματα.

Σημειώσεις:

- Όταν αφαιρέσετε την μέση τιμή από τα ratings ενός χρήστη υπάρχει η περίπτωση να πάρουμε ένα διάνυσμα με όλο μηδενικά. Επίσης κάποιοι χρήστες ή επιχειρήσεις μπορεί να έχουν μηδενικές τιμές ομοιότητας με όλους. Σε αυτή την περίπτωση η ομοιότητα δεν έχει νόημα και θα χρησιμοποιήσετε τους αλγορίθμους 1,2 για την πρόβλεψη σας στους αλγόριθμους 3,4 αντίστοιχα.
- Χρησιμοποιήστε τις συναρτήσεις της python για τον υπολογισμό αποστάσεων ή πράξεων με πίνακες, είναι πολύ πιο γρήγορες.

Ερώτηση 3 (Locality Sensitive Hashing)

Ο στόχος αυτής της ερώτησης είναι να πειραματιστείτε με το Locality Sensitive Hashing για πραγματικά διανύσματα. Θα χρησιμοποιήσετε τον ίδιο πίνακα δεδομένων όπως και στην ερώτηση 2, όπου οι γραμμές έχουν κανονικοποιήθει αφαιρώντας την μέση τιμή. Αγνοήστε τις γραμμές με μόνο μηδενικές τιμές.

Καταρχάς, υπολογίστε την ακριβή cosine ομοιότητα μεταξύ όλων των ζευγαριών των χρηστών, και βρείτε τα ζεύγη με ομοιότητα τουλάχιστον θ.

Στη συνέχεια, δημιουργήστε τυχαία διανύσματα διάστασης K, με τιμές -1/+1 (επιλεγμένες τυχαία), και υπολογίστε την προβολή των διανυσμάτων των χρηστών (γραμμές του πίνακα) πάνω σε αυτά τα διανύσματα. Μετατρέψτε τις τιμές σε τιμές -1/+1 ανάλογα με το πρόσημο της προβολής. Έτσι ορίζουμε μια «υπογραφή» μεγέθους K για κάθε χρήστη.

Σπάστε την υπογραφή σε b μίνι-υπογραφές μεγέθους r ($K = b/r$). Δημιουργήστε b hash tables όπου οι αντίστοιχες μίνι-υπογραφές των χρηστών θα χρησιμοποιηθούν ως κλειδιά. Βρείτε όλα τα ζεύγη των χρηστών οι οποίοι γίνονται hashed στο ίδιο bucket για τουλάχιστον ένα hash table. Στη συνέχεια υπολογίστε τα εξής: (1) Το ποσοστό των ζευγαριών που χρειάζεται να ελέγξουμε σε σχέση με το σύνολο όλων των δυνατών ζευγαριών; (2) Το ποσοστό των πολύ όμοιων ζευγαριών (πάνω από θ) τα οποία βρίσκουμε (recall). Υπολογίστε την μέση τιμή για αυτά τα νούμερα, κάνοντας 10 πειράματα.

Τρέξετε πειράματα για θ = 0.7, 0.75 και 0.8, K = 100 και 200, και r = 5, 10 και 20. Παραδώσετε τον κώδικα σας, και μια αναφορά με τα νούμερα (μέσες τιμές) και ένα σχολιασμό των αποτελεσμάτων.