

Τρίτη Σειρά Ασκήσεων

Η προθεσμία για την τρίτη σειρά ασκήσεων είναι την Δευτέρα 12 Ιανουαρίου, 11:55 μ.μ. Παραδώστε Notebooks με τον κώδικα και τις αναφορές σε .ipynb και .html μορφή. Για την Ερώτηση 1 μπορείτε να παραδώσετε και pdf με την απόδειξη, ή φωτογραφίες από χειρόγραφα. Για καθυστερημένες υποβολές ισχύει η πολιτική στην σελίδα του μαθήματος. Η παράδοση θα γίνει μέσω του course. Λεπτομέρειες στη σελίδα Ασκήσεις του μαθήματος. Η πρώτη ερώτηση είναι ατομική, οι επόμενες μπορούν να γίνουν σε ομάδες των 2. Όπως και στις προηγούμενες ασκήσεις παραδώστε και ένα αρχείο που να καταγράφει την βοήθεια που είχατε online.

Ερώτηση 1

Για την άσκηση αυτή θα δείξετε την σχέση που υπάρχει μεταξύ του Pagerank διανύσματος με ομοιόμορφο jump vector, και των personalized Pagerank διανυσμάτων. Υπενθυμίζω ότι ο Pagerank αλγόριθμος έχει σαν παράμετρο ένα διάνυσμα $\mathbf{v}$ (το jump vector) το οποίο ορίζει μια κατανομή πιθανότητας πάνω στους κόμβους του γραφήματος και η τιμή $\mathbf{v}(i)$ καθορίζει την πιθανότητα να επιλέξουμε τον κόμβο i για επανεκκίνηση. Έστω $\mathbf{p}_u$ το Pagerank διάνυσμα με ομοιόμορφο jump vector (δηλαδή $\mathbf{v}^T = (\frac{1}{n}, \frac{1}{n}, \dots, \frac{1}{n})$), και $\mathbf{p}_i$ το personalized Pagerank διάνυσμα όπου το jump vector δίνει όλη την πιθανότητα στον κόμβο i (δηλαδή, $\mathbf{v}^T = (0, 0, \dots, 0, 1, 0, \dots, 0)$ με το 1 στην i θέση).

Αποδείξτε ότι το διάνυσμα $\mathbf{p}_u$ είναι ο μέσος όρος των διανυσμάτων $\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_n$, δηλαδή, $\mathbf{p}_u = \frac{1}{n} \sum_{i=1}^n \mathbf{p}_i$. Για την απόδειξη θα χρησιμοποιήσετε το γεγονός ότι το Pagerank vector $\mathbf{p}_v$ (το Pagerank διάνυσμα με jump vector $\mathbf{v}$) μπορεί να γραφτεί σαν γραμμική συνάρτηση του jump vector, δηλαδή $\mathbf{p}_v^T = \mathbf{v}^T \mathbf{Q}$, για κάποιο πίνακα $\mathbf{Q}$.

(Υπενθύμιση: Όταν αναφερόμαστε σε διανύσματα υποθέτουμε ότι είναι στήλες. Δηλαδή ένα n -διάστατο διάνυσμα $\mathbf{v}$ είναι ένας $n \times 1$ πίνακας. Αν θέλουμε να χρησιμοποιήσουμε το διάνυσμα σαν γραμμή, δηλαδή σαν ένα $1 \times n$ πίνακα θα το συμβολίζουμε ως $\mathbf{v}^T$)

Απαντήστε στα εξής ερωτήματα:

1. Χρησιμοποιώντας την σχέση $\mathbf{p}_v^T = (1 - a)\mathbf{p}_v^T \mathbf{P} + a\mathbf{v}^T$, δώστε την φόρμουλα για τον πίνακα $\mathbf{Q}$.
2. Δοθείσας της σχέσης $\mathbf{p}_v^T = \mathbf{v}^T \mathbf{Q}$, τι ισχύει για τις γραμμές του πίνακα $\mathbf{Q}$ (ως προς τα personalized Pagerank vectors)?
3. Αποδείξτε ότι $\mathbf{p}_u = \frac{1}{n} \sum_{i=1}^n \mathbf{p}_i$
4. Στην γενική περίπτωση ενός οποιουδήποτε jump vector $\mathbf{v}$ (όχι απαραίτητα το ομοιόμορφο διάνυσμα), πως μπορούμε να εκφράσουμε το $\mathbf{p}_v$ σαν συνάρτηση των $\mathbf{p}_i$?

Ερώτηση 2

Στην άσκηση αυτή θα εξασκηθείτε στην εφαρμογή αλγορίθμων ομαδοποίησης. Θα δουλέψετε και πάλι με τα δεδομένα για τις ταινίες. Κατεβάστε από το Kaggle repository το αρχείο ratings_small.csv, links_small.csv και movies_metadata.csv. Το δεύτερο αρχείο χρησιμεύει για να συνδέσει τα ids του πρώτου με το τρίτο. Ο στόχος είναι να κάνουμε ομαδοποίηση των ταινιών βάσει την κοινή συμπεριφορά στα ratings. Κρατήστε μόνο τις ταινίες που έχουν τουλάχιστον 10 ratings, και δημιουργήστε ένα αραιό πίνακα ταινίες-χρήστες με τα ratings.

Εφαρμόστε δύο αλγορίθμους ομαδοποίησης: (1) Τον k-means, (2) Ένα agglomerative clustering αλγόριθμο. Για να βρείτε τον αριθμό από clusters κάνετε ένα γράφημα με δύο Y άξονες, με το sum of squares error και το silhouette coefficient του k-means αλγόριθμου, όπως είδατε στο μάθημα και το φροντιστήριο. Εξετάστε αριθμό clusters από 2 έως και 20, και επιλέξτε το καλύτερο k. Αιτιολογήστε την απόφασή σας.

Για να αξιολογήσετε τα clusters θα χρησιμοποιήσετε τα genres των ταινιών που είναι σε κάθε cluster. Για κάθε cluster θα υπολογίσετε την συχνότητα των genres. Τι γίνεται αν χρησιμοποιήσουμε τα πιο συχνά genres για κάθε cluster? Μια εναλλακτική της συχνότητας είναι το lift, το οποίο είναι η συχνότητα του genre μέσα στο cluster προς την συχνότητα συνολικά. Χρησιμοποιήστε το lift για να βρείτε τα 3 genres με το μεγαλύτερο lift. Χρησιμοποιώντας αυτά προσπαθήστε να καταλάβετε τι αντιπροσωπεύει το κάθε cluster και αξιολογήστε την ομαδοποίηση.

Εφαρμόστε την ίδια διαδικασία αλλά αυτή τη φορά εφαρμόστε PCA στον πίνακα με τα ratings και κρατήστε 50 ποιο ισχυρά χαρακτηριστικά (latent factors). Αναφέρετε αν βλέπετε διαφορά στα αποτελέσματα.

Προτείνετε κάποιο άλλο τρόπο να αξιολογήσετε τα clusters, χρησιμοποιώντας τα χαρακτηριστικά των ταινιών.

Παραδώστε ένα Notebook με τον κώδικά σας, τα αποτελέσματά σας και τον σχολιασμό.

Ερώτηση 3

Στην Ερώτηση αυτή θα εξασκηθείτε στην εφαρμογή αλγορίθμων κατηγοριοποίησης. Χρησιμοποιώντας τα metadata για τις ταινίες, θα προσπαθήσετε να προβλέψετε το genre της ταινίας χρησιμοποιώντας την περίληψη (overview).

Θα περιορίσουμε το dataset σε τρεις διακριτές κατηγορίες: War (Πολεμικές), Music (Μουσικές) και Animation (Κινουμένων Σχεδίων). Πάρτε από τα δεδομένα τις ταινίες σε αυτές τις τρεις κατηγορίες, και αφαιρέστε ταινίες που ανήκουν σε παραπάνω από μία κατηγορία από αυτές τις τρεις (π.χ. ταινίες που είναι και Animation και Music).

Θα εξετάσετε 2 τρόπους για εξαγωγή χαρακτηριστικών (feature extraction).

- **TF-IDF:** Χρησιμοποιήστε τον TfidfVectorizer (προτείνεται max_features=5000 και αφαίρεση stop_words).
- **Word2Vec:** Χρησιμοποιήστε ένα προ-εκπαιδευμένο μοντέλο από τη βιβλιοθήκη gensim (π.χ. glove-wiki-gigaword-50 ή word2vec-google-news-300). Για κάθε ταινία, υπολογίστε το μέσο διάνυσμα (mean vector) όλων των λέξεων της περιλήψης.

Θα εξετάσετε τους εξής classifiers: Decision Tree, kNN, Logistic Regression, SVM, και MLP. Για την αξιολόγηση θα κάνετε 5-fold cross validation. Για να μην υπάρχει μεταφορά πληροφορίας μεταξύ train και test δεδομένων, το TF-IDF θα πρέπει να το κάνετε για κάθε training fold ξεχωριστά. Για κάθε classifier θα υπολογίσετε το accuracy, precision, recall, και το συνολικό confusion table. Κάνετε ένα πίνακα με όλα τα αποτελέσματα και συγκρίνετε τους classifiers.

Μπορείτε να δοκιμάσετε δικά σας παραδείγματα για να δείτε πως συμπεριφέρονται οι classifiers σας. Π.χ., δοκιμάστε την παρακάτω περίληψη: *"A lonely soldier plays the piano while waiting for the battle to begin"*

Bonus: Εξερευνήστε επιπλέον features από τα δεδομένα της ταινίας για να βελτιώσετε τους classifiers.

Παραδώστε ένα Notebook με τον κώδικά σας, τα αποτελέσματα και τον σχολιασμό.

Ερώτηση 4

Στην Ερώτηση αυτή θα υλοποιήσετε μια παραλλαγή του User Collaborative Filtering (UCF) αλγόριθμου που υλοποιήσατε στην Δεύτερη Άσκηση, χρησιμοποιώντας μια συνάρτηση ομοιότητας μεταξύ των χρηστών που χρησιμοποιεί το Personalized PageRank (PPR). Για την υλοποίηση της ομοιότητας, θα δημιουργήσετε ένα γράφο μεταξύ των χρηστών, δημιουργώντας ακμές μεταξύ κάθε χρήστη και των m πιο όμοιων χρηστών. Πάνω σε αυτό το γράφο για κάθε χρήστη θα τρέξετε ένα Personalized PageRank για να βρείτε τους k πιο όμοιους χρήστες για τον UCF αλγόριθμο.

Θα χρησιμοποιήσετε τα δεδομένα που χρησιμοποιήσατε και στην Δεύτερη Άσκηση. Θα χρησιμοποιήσετε το train set για να φτιάξετε τον γράφο μεταξύ των χρηστών και να υπολογίσετε τα PPR scores. Στη συνέχεια θα κάνετε πρόβλεψη για τα (user, movie) ζευγάρια στο test set. Συγκεκριμένα θα κάνετε τα εξής βήματα:

- Υπολογίστε το cosine similarity μεταξύ όλων των χρηστών στο training set, χρησιμοποιώντας έτοιμη συνάρτηση από το sklearn. Δημιουργήστε ένα γράφημα προσθέτοντας για κάθε χρήστη u , m ακμές με τους m πιο όμοιους χρήστες. Οι ακμές θα έχουν βάρος ίσο με το cosine similarity μεταξύ των χρηστών. Μια ακμή θα προστίθεται μόνο αν δεν υπάρχει ήδη στο γράφημα. Θα πειραματιστείτε με τιμές του $m \in \{3, 5, 10\}$.
- Για κάθε κόμβο u , τρέξετε ένα Personalized PageRank (στον γράφο με τα βάρη) με σημείο εκκίνησης τον κόμβο u . Θα χρησιμοποιήσετε την τιμή του Personalized PageRank για κάθε κόμβο v , $PPR(u, v)$, ως την ομοιότητα του κόμβου u με τον κόμβο v .
- Υλοποιήστε μια παραλλαγή του UCF αλγόριθμου που υλοποιήσατε στην Δεύτερη Άσκηση χρησιμοποιώντας το $PPR(u, v)$ για την ομοιότητα μεταξύ των χρηστών. Θα χρησιμοποιήσετε τις τιμές του k που χρησιμοποιήσατε και στην Δεύτερη Άσκηση.
- Κάνετε plot το RMSE ως συνάρτηση του k , για τις διαφορετικές τιμές του m (στο ίδιο γράφημα). Βρείτε το χαμηλότερο RMSE και συγκρίνετε το με αυτό του UCF που υλοποιήσατε στην Δεύτερη Άσκηση. Σχολιάστε τα αποτελέσματα.

Bonus: Πειραματιστείτε με διαφορετικές τιμές του restart probability και εξετάστε αν βελτιώνονται τα αποτελέσματα.

Παραδώστε ένα notebook με τους υπολογισμούς σας, τα αποτελέσματα, και το κείμενο του σχολιασμού.