Πανεπιστήμιο Ιωαννίνων,
Τμήμα Μηχανικών Η/Υ & Πληροφορικής

Εξόρυξη Δεδομένων – ΜΥΕ012/ΠΛΕ 059
Χειμερινό Εξάμηνο, 2017

Δεύτερη Σειρά Ασκήσεων

Η προθεσμία για την δεύτερη σειρά ασκήσεων είναι την Παρασκευή 15 Δεκεμβρίου μέχρι το τέλος της ημέρας. Κάνετε turn-in τον κώδικα σας, με οδηγίες πώς για το πώς τρέχει, και την αναφορά σας. 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. Αν δεν μπορείτε να κάνετε turnin στείλετε την εργασία σας μέσω email. Θα γίνει προφορική εξέταση των δύο πρώτων ασκήσεων την επόμενη εβδομάδα.

Ερώτηση 1

Μία εκθετική κατανομή ορίζεται με συνάρτηση πυκνότητας πιθανότητας $f(x) = \lambda e^{-\lambda x}$, για $x \geq 0$, όπου λ είναι η παράμετρος της κατανομής. Σας δίνεται ένα σύνολο από παρατηρήσεις $X = \{x_1, \dots, x_n\}, x_i \geq 0$, που έχουν παραχθεί από μία εκθετική κατανομή. Χρησιμοποιήστε την Maximum Likelihood Estimation τεχνική που περιγράψαμε στην τάξη για να βρείτε την παράμετρο της κατανομής που ταιριάζει (fits) τα δεδομένα των παρατηρήσεων.

Ερώτηση 2 (Principal Component Analysis)

Σας δίνεται το αρχείο “data2.txt” το οποίο περιέχει ένα sparse πίνακα 500×20. Ο πίνακας είναι αποθηκευμένος σε tab-separated τριάδες (γραμμή, στήλη, τιμή). Ο πίνακας κρατάει πληροφορία για την κατανάλωση 20 χημικών ουσιών από 500 χρήστες. Οι τιμές είναι ο αριθμός των φορών που ο κάθε χρήστης έχει χρησιμοποιήσει μια ουσία. Υπάρχουν 10 νόμιμες ουσίες (είναι οι στήλες 0-9) και 10 παράνομες (οι στήλες 10-19).

Φορτώσετε τα δεδομένα και εφαρμόστε Principal Component Analysis (χρησιμοποιήσετε το πακέτο του sklearn για PCA). Εξετάσετε τα δύο πρώτα components. Κάνετε ένα plot των σημείων στη μία και στις δύο διαστάσεις. Τι παρατηρείτε? Πως ερμηνεύετε τα αποτελέσματα? Παραδώστε ένα notebook με τις γραφικές παραστάσεις και μια αναφορά με την ανάλυση σας.

Ερώτηση 3 (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 `business.json` και `review.json` (το τελευταίο είναι πάνω από 3GB οπότε θα χρειαστείτε χώρο, και πρέπει να το λάβετε υπόψιν σας κατά την επεξεργασία). Χρησιμοποιώντας αυτά τα δεδομένα θα δημιουργήσετε ένα user-business πίνακα με τα ratings των χρηστών για όλες τις επιχειρήσεις στην πόλη του “Toronto”. 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).

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. **User Average (UA):** Χρησιμοποιήστε την μέση τιμή $\overline{r(u)}$ των ratings του u για την πρόβλεψη.
2. **Business Average (BA):** Χρησιμοποιήστε την μέση τιμή $\overline{r(b)}$ των ratings της επιχείρησης b για την πρόβλεψη.
3. **User-based Collaborative Filtering** (**UCF**): Για τον χρήστη u υπολογίστε το σύνολο $N_k(u)$ με τους k πιο όμοιους χρήστες οι οποίοι έχουν βαθμολογήσει την επιχείρηση b. Στη συνέχεια χρησιμοποιήστε την εξής εξίσωση για την πρόβλεψη σας:
$$p(u,b) = \overline{r(u)} + \frac{\sum_{u' \in N_k(u)} s(u,u')(r(u',b) - \overline{r(u')})}{\sum_{u' \in N_k(u)} s(u,u')}$$
Για την ομοιότητα χρησιμοποιήστε το correlation coefficient (το cosine similarity, μετά την αφαίρεση της μέσης τιμής από την κάθε γραμμή). Αν η τιμή γίνει μικρότερη του 1, ή μεγαλύτερη του 5, στρογγυλοποιείστε στο 1 ή το 5.
4. **Item-based Collaborative Filtering** (**ICF**): Για την επιχείρηση b υπολογίστε το σύνολο $N_k(b)$ με τις 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:
$$p(u,b) = \frac{\sum_{b' \in N_k(m)} s(b,b')r(u,b')}{\sum_{b' \in N_k(b)} s(b,b')}$$
5. **Singular Value Decomposition** (**SVD**)**:** Εφαρμόστε το Singular Value Decomposition στον 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 τιμή $p(u,b) = R_k(u,b)$ για την πρόβλεψη σας. (Bonus: Πειραματιστείτε και με Principal Component Analysis – κάνει διαφορά?)

Για την αξιολόγηση και σύγκριση των αλγορίθμων θα χρησιμοποιήσετε την RMSE (Root Mean Square Error) μετρική. 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 RMSE του αλγορίθμου ορίζεται ως

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

Δημιουργήστε γραφικές παραστάσεις με το RMSE για διαφορετικές τιμές του k για όλους τους αλγορίθμους.

Θα πρέπει να παραδώσετε τα ακόλουθα:

- Όλο τον κώδικα που θα γράψετε εσείς.
- Μια αναφορά που θα περιγράφει τι κάνατε, θα συγκρίνει τους διαφορετικούς αλγορίθμους, και θα σχολιάζει τα αποτελέσματα.

Σημειώσεις:

- Το αρχείο με τα reviews είναι πολύ μεγάλο και άρα δεν μπορείτε να το φορτώσετε στην μνήμη. Επίσης, ο αριθμός των user/business ζευγών είναι μεγάλος και για να τον διαχειριστείτε θα πρέπει να δημιουργήσετε κατάλληλες δομές.
- Χρησιμοποιήστε τις συναρτήσεις της python για τον υπολογισμό αποστάσεων ή πράξεων με πίνακες, είναι πολύ πιο γρήγορες.

Ερώτηση 4 (Clustering)

Στην ερώτηση αυτή θα εξασκηθείτε στην εφαρμογή αλγορίθμων ομαδοποίησης (clustering). Θα χρησιμοποιήσουμε ένα σύνολο δεδομένων παρόμοιο με αυτό που χρησιμοποιήσατε στην πρώτη σειρά ασκήσεων. Θα πάρετε πάλι τις επιχειρήσεις από το Τορόντο και θα κρατήσετε αυτές που βρίσκονται σε γειτονιές με τουλάχιστον **300** επιχειρήσεις. Χρησιμοποιώντας αυτά τα δεδομένα θα κοιτάξετε τα παρακάτω προβλήματα:

1. Πάρετε τις γεωγραφικές συντεταγμένες των επιχειρήσεων και εφαρμόστε αλγορίθμους clustering στα δισδιάστατα σημεία. Θα χρησιμοποιήσετε τους αλγορίθμους k-means, agglomerative και DBSCAN. Χρησιμοποιήστε (όποτε απαιτείται) αριθμό clusters ίσο με τον αριθμό των γειτονιών που έχετε στα δεδομένα σας. Για τον DBSCAN πειραματιστείτε με διαφορετικές τιμές για τις παραμέτρους του (Bonus: δημιουργείστε το διάγραμμα με την απόσταση από τον minPts-πιο κοντινό γείτονα για να αποφασίσετε την τιμή του eps). Δημιουργείστε τον πίνακα σύγχυσης και για τους τρεις αλγορίθμους και αναφέρετε το precision και recall. Οπτικοποιήσετε τα δεδομένα και τα αποτελέσματα με plots και τοποθετήσετε τα πάνω στον χάρτη της πόλης. Εξετάσετε τα δεδομένα με το μάτι και σχολιάστε τι βλέπετε.
2. Για κάθε επιχείρηση στην λίστα σας, πάρετε τις κατηγορίες τους. Κρατήσετε τις επιχειρήσεις που έχουν στην λίστα των κατηγοριών τους τις κατηγορίες ‘Food’ και ‘Restaurants’ (και τις δύο). Κρατήστε τις κατηγορίες που εμφανίζονται σε τουλάχιστον 5% των επιχειρήσεων (εξαιρέστε τις κατηγορίες ‘Food’ και ‘Restaurant’ που εμφανίζονται παντού). Κρατήστε τις επιχειρήσεις για τις οποίες έχουμε τουλάχιστον μια κατηγορία από αυτές που κρατήσατε. Για τις επιχειρήσεις που έχουνε παραπάνω από μία κατηγορία, κρατήστε αυτή που είναι πιο συχνή στα δεδομένα σας. Πετάξετε τις κατηγορίες (και τις αντίστοιχες επιχειρήσεις) που εμφανίζονται σε λιγότερες από 15 επιχειρήσεις
   Για κάθε μία από τις επιχειρήσεις στο τελικό σύνολο επιχειρήσεων, από το αρχείο review.json πάρετε όλα τα reviews για την επιχείρηση και φτιάξετε ένα μεγάλο κείμενο. Χρησιμοποιείστε αυτά τα κείμενα για να δημιουργήσετε την tf-idf αναπαράσταση των επιχειρήσεων (χρησιμοποιήσετε την έτοιμη βιβλιοθήκη της python για να πάρετε αυτή την αναπαράσταση - μπορείτε επίσης να κάνετε επιπλέον επιλογές για τις παραμέτρους της βιβλιοθήκης). Κάνετε cluster τις επιχειρήσεις χρησιμοποιώντας k-means και agglomerative clustering με αριθμό clusters ίσο με τον αριθμό των κατηγοριών. Εξετάστε αν τα clusters που βρίσκετε αντιστοιχούν στις κατηγορίες χρησιμοποιώντας τον πίνακα σύγχυσης και τις μετρικές precision και recall. Σχολιάστε τα αποτελέσματα.
3. Είναι πιθανό τα κείμενα να μην γίνονται clustered με τρόπο που συμφωνεί με τις κατηγορίες. Για τα δεδομένα που δημιουργήσατε στο προηγούμενο βήμα, χρησιμοποιήστε τον k-means αλγόριθμο και δημιουργήστε το silhouette plot για να αποφασίσετε για τον αριθμό των clusters. Για τον αριθμό που επιλέξατε, εξετάσετε τα clusters του k-means, και χρησιμοποιώντας τις λέξεις και τις κατηγορίες

(μπορείτε να χρησιμοποιήσετε την πλήρη λίστα των κατηγοριών), προσπαθήσετε να εξηγήσετε σε τι είδους εστιατόρια αντιστοιχεί το κάθε cluster.

Παραδώσετε τον κώδικα σας και μια αναφορά με λεπτομερή σχολιασμό και ανάλυση των αποτελεσμάτων σας.