Πανεπιστήμιο Ιωαννίνων,
Τμήμα Μηχανικών Η/Υ & Πληροφορικής

Εξόρυξη Δεδομένων – ΜΥΕ012/ΠΛΕ 059
Χειμερινό Εξάμηνο, 2017

Τρίτη Σειρά Ασκήσεων

Η προθεσμία για την δεύτερη σειρά ασκήσεων είναι το Κυριακή 11 Φεβρουαρίου 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

Στην άσκηση αυτή θα εξασκηθείτε στην εφαρμογή αλγορίθμων κατηγοριοποίησης. Ο στόχος σε αυτή την άσκηση είναι να διαχωρίσετε μεταξύ επιχειρήσεων που ανήκουν στις κατηγορίες ‘Nightlife’, ή ‘Bars’, και τις επιχειρήσεις που ανήκουν στις κατηγορίες ‘Cafes’, ‘Coffee & Tea’, ή ‘Breakfast & Brunch’.

Θα δουλέψετε με τις επιχειρήσεις από την περιοχή του Τορόντο. Δημιουργήστε τα δεδομένα με τις επιχειρήσεις, και εξάγετε χαρακτηριστικά από το κείμενο των reviews των επιχειρήσεων, καθώς και όποια άλλα χαρακτηριστικά πιστεύετε μπορούν να βοηθήσουν. Θα πειραματιστείτε με τέσσερεις classifiers: Logistic Regression, SVM, Decision Trees, και Naïve Bayes. Για την αξιολόγηση θα χρησιμοποιήσετε 5-fold cross validation. Αναφέρετε τον μέσο confusion matrix και τις μετρικές accuracy, precision, recall και F1-measure.

Στην αναφορά σας περιγράψετε τα χαρακτηριστικά που χρησιμοποιήσατε, και τα αποτελέσματα της κατηγοριοποίησης. Εξετάσετε επίσης τη σημασία που δίνουν στα διάφορα χαρακτηριστικά οι κατηγοριοποιητές (π.χ., τα βάρη που έχουν στον logistic regression classifier, ή τα πρώτα επίπεδα του decision tree), και προσπαθήστε να καταλάβετε ποιες λέξεις είναι που ξεχωρίζουν τις δύο κατηγορίες.

Ερώτηση 2

Και σε αυτή στην άσκηση θα πειραματιστείτε με αλγορίθμους κατηγοριοποίησης. Ο στόχος είναι να φτιάξετε ένα classifier ο οποίος θα προβλέπει αν ένα review για ένα εστιατόριο (κατηγορίες Food, Restaurants) είναι αρνητικό (έχει 1 ή 2 stars).

Για την άσκηση αυτή δημιουργήθηκε ένας διαγωνισμός στο Kaggle για το μάθημα (εδώ είναι ο σύνδεσμος για τον διαγωνισμό). Δημιουργήστε ένα account με το email του πανεπιστημίου. Θα σας δοθεί πρόσβαση στον διαγωνισμό μέσω του link και θα μπορέστε να καταθέσετε μια λύση για τον διαγωνισμό. Εκεί σας δίνονται τα training data και τα test data.

Υπάρχει μία κατάταξη στην οποία μπορείτε να δείτε την θέση σας σε σχέση με άλλες λύσεις. Μια καλή θέση θα ενισχύσει τον βαθμό σας. Το σημαντικό είναι να πετύχετε ένα καλό σκορ στο F1-measure σε σχέση με τις υπόλοιπες λύσεις. Μπορείτε να χρησιμοποιήσετε όποιο αλγόριθμο θέλετε.

Στην αναφορά περιγράψετε τα χαρακτηριστικά που δημιουργήσατε και το μοντέλο που χρησιμοποιήσατε, και αναφέρετε το όνομα σας στο Kaggle, και το σκορ σας στον διαγωνισμό. Επίσης, αναφέρετε και πειράματα που κάνατε και δεν δούλεψαν, ή πως βελτιώσατε μια λύση που δεν απέδιδε καλά (περιλάβετε και αποτελέσματα από τα πειράματα).

Ερώτηση 3

Αποδείξτε ότι για ένα μη κατευθυνόμενο γράφο η κατανομή σύγκλισης (stationary distribution) ενός τυχαίου περίπατου είναι ανάλογη του βαθμού του κάθε κόμβου. Δηλαδή αν P είναι ο πίνακας μετάβασης του τυχαίου περίπατου, και π η κατανομή σύγκλισης για την οποία ισχύει ότι $\pi = \pi \cdot P$, δείξτε ότι για τον κόμβο i, η πιθανότητα π_i είναι ανάλογη του d_i, όπου d_i είναι ο αριθμός των ακμών με άκρο την κορυφή i.

Ερώτηση 4

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 Toronto 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 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 $P(u, b)$.

Υπολογίσετε το Root Mean Square Error (RMSE) για αυτή τη μέθοδο. Στη συνέχεια, τρέξετε τους αλγορίθμους που υλοποιήσατε στην Δεύτερη Σειρά (Ερώτηση 3) για αυτό το dataset και συγκρίνετε το Root Mean Square Error (RMSE). Παρουσιάστε τα αποτελέσματα σας και γράψετε τις παρατηρήσεις σας.

Bonus: Προτείνετε, υλοποιήστε και τεστάρετε μια διαφορετική μέθοδο που να προβλέπει τα ratings των χρηστών χρησιμοποιώντας το γράφημα των φιλιών μεταξύ των χρηστών. Περιγράψετε την μέθοδο σας και τα αποτελέσματα της.

Ερώτηση 5 (Κάλυψη)

Ένα πρόβλημα το οποίο προκύπτει στην διαχείριση γραφημάτων είναι να υπολογίσουμε την απόσταση $d(x, y)$ (το μήκος του πιο σύντομου μονοπατιού) μεταξύ δύο κόμβων (x, y). Ο υπολογισμός είναι δαπανηρός οπότε υπάρχει ανάγκη για ευρετήρια (indexes) που απαντάνε τέτοια ερωτήματα. Τα ευρετήρια αυτά κρατάνε ένα μικρό σύνολο L από κόμβους-σημεία-αναφοράς (landmarks) για τα οποία κρατάμε τις αποστάσεις τους από

όλους τους κόμβους στο γράφημα, και τις χρησιμοποιούμε για να απαντήσουμε ερωτήματα συντομότερων μονοπατιών.

1. Δείξετε ότι για κάθε landmark $\ell \in L$, ισχύει ότι $d(x, y) \leq d(x, \ell) + d(\ell, y)$. Πότε ισχύει η ισότητα?
2. Δείξετε ότι το πρόβλημα του να βρούμε το μικρότερο σύνολο L από σημεία αναφοράς ώστε να μπορούμε να απαντήσουμε ακριβώς shortest path queries για οποιοδήποτε ζευγάρι από κορυφές (x, y) μπορεί να εκφραστεί ως ένα πρόβλημα ελάχιστης κάλυψης (minimum set cover).

Ερώτηση 6 (bonus)

Στο Kaggle υπάρχει ένας ενεργός διαγωνισμός για την πρόβλεψη «τοξικού» περιεχομένου σε συζητήσεις (εδώ το link του διαγωνισμού). Χρησιμοποιώντας τον λογαριασμό που δημιουργήσατε στην προηγούμενη ερώτηση, υποβάλετε μια λύση στον διαγωνισμό. Ο στόχος δεν είναι να κερδίσετε τον διαγωνισμό (αν και η θέση σας στον διαγωνισμό θα ενισχύσει το βαθμό σας), αλλά να δουλέψετε σε ένα πραγματικό πρόβλημα που δεν έχει γνωστή λύση.

Δημιουργήστε μια αναφορά που θα περιέχει τα παρακάτω:

a. Μια περιγραφή της λύσης που υλοποιήσατε. Τι χαρακτηριστικά χρησιμοποιήσατε, ποιες τεχνικές. Μια σύντομη περιγραφή της λογικής πίσω από τις επιλογές σας.
b. Τα αποτελέσματα σας στο Kaggle test dataset.
c. Ένα σχολιασμό στα παραπάνω: Τι δούλεψε και τι δεν δούλεψε τόσο καλά? 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.