Πανεπιστήμιο Ιωαννίνων,
Τμήμα Μηχανικών Πληροφορικής

Εξόρυξη Δεδομένων – ΜΥΕ012/ΠΛΕ 059
Χειμερινό Εξάμηνο, 2015

Σειρά Ασκήσεων Εξεταστικής Σεπτεμβρίου

Η προθεσμία για την παράδοση αυτής της σειράς Ασκήσεων είναι στις **27 Σεπτεμβρίου**, μέχρι το τέλος της ημέρας. Παραδώστε ηλεκτρονικά τον κώδικα και την αναφορά σας, είτε μέσω turnin είτε μέσω email. Λεπτομέρειες για το turn-in, και για το πώς να γράφετε αναφορές είναι στη σελίδα Ασκήσεις του μαθήματος. Για καθυστερημένες υποβολές ισχύει η πολιτική στην σελίδα του μαθήματος. Δεν υπάρχουν free passes για αυτή την σειρά ασκήσεων. Η προφορική εξέταση θα κανονιστεί μετά την παράδοση των ασκήσεων.

Ερώτηση 1

Υποθέστε ότι σας δίνεται σαν είσοδος ένας πίνακας με *n* γραμμές και *m* στήλες, με 0/1 τιμές. Θέλετε να βρείτε όλα τα *(r,c)-πλακίδια (tiles)* από 1, δηλαδή συνδυασμούς από *r* γραμμές και *c* στήλες ώστε ο υπο-πίνακας με αυτές τις γραμμές και αυτές τις στήλες να έχει μόνο άσσους. Τα πλακίδια μπορεί να είναι επικαλυπτόμενα. Δώσετε ένα αποτελεσματικό αλγόριθμο για το πρόβλημα χρησιμοποιώντας την ιδέα του APriori.

Ερώτηση 2

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 3

Στο Yelp κάποιες επιχειρήσεις έχουν υπερβολικά μεγάλο αριθμό από κριτικές και είναι δύσκολο για ένα χρήστη να τις διαβάσει όλες για να βρει αυτό που ψάχνει. Θέλουμε λοιπόν για κάθε επιχείρηση να διαλέξουμε K κριτικές που να περιγράφουν τα διάφορα **χαρακτηριστικά** της επιχείρησης όσο γίνεται καλύτερα. Υποθέστε ότι υπάρχουν m χαρακτηριστικά $A_1, \dots, A_m$ συνολικά για όλες τις επιχειρήσεις τα οποία είναι γνωστά εκ των προτέρων (π.χ., ποιότητα, τοποθεσία, τιμή, σέρβις, κλπ). Επίσης, μέσω κάποιας προεπεξεργασίας ξέρουμε τα χαρακτηριστικά που αναφέρονται σε κάθε κριτική. Δοθέντων N κριτικών για μια επιχείρηση θέλουμε να διαλέξουμε K ώστε στην τελική συλλογή να αναφέρονται όσο γίνεται περισσότερα από τα m χαρακτηριστικά του ξενοδοχείου σε τουλάχιστον μία κριτική από τη συλλογή.

- Δείξετε ότι υπάρχει ένας greedy αλγόριθμος για το πρόβλημα που έχει σταθερό λόγο προσέγγισης ως προς τον βέλτιστο αλγόριθμο που μεγιστοποιεί τον αριθμό των χαρακτηριστικών που εμφανίζονται στην συλλογή.
- Τι γίνεται αν θέλουμε να εμφανίζονται και τα m χαρακτηριστικά της επιχείρησης?

Ερώτηση 4

Ένα σημαντικό πρόβλημα στα online κοινωνικά δίκτυα είναι η πρόβλεψη μελλοντικών συνδέσεων μεταξύ των χρηστών των δικτύων. Οι αλγόριθμοι για πρόβλεψη χρησιμοποιούνται για συστάσεις φιλίας ώστε να μεγαλώνει το δίκτυο. Στην ερώτηση αυτή θα πειραματιστείτε με διαφορετικούς αλγορίθμους για πρόβλεψη νέων συνδέσεων.

Υπάρχουν πολλές διαφορετικές προσεγγίσεις γι αυτό το πρόβλημα. Εσείς θα κοιτάξετε δύο διαφορετικές προσεγγίσεις σε αυτή την άσκηση.

Στην πρώτη προσέγγιση για κάθε κόμβο v υπολογίζουμε ένα σκορ για την σύνδεση $(v.u)$ μεταξύ του v και κάθε κόμβου u με τον οποίο ο v δεν είναι ήδη συνδεδεμένος. Στη συνέχεια ταξινομούμε τις πιθανές νέες συνδέσεις για τον v με βάση αυτό το σκορ κρατάμε τις K πρώτες, για κάποια σταθερά K. Θα εξετάσετε τρεις διαφορετικούς μεθόδους να υπολογίσετε το σκορ για μία σύνδεση (v, u).

1. Η πρώτη μέθοδος υπολογίζει τον αριθμό των κοινών φίλων που έχουν ο v και ο u.
2. Η δεύτερη μέθοδος κοιτάει πάλι τον αριθμό των κοινών φίλων, αλλά για κάθε κοινό φίλο z προσθέτει ένα βάρος ίσο με $1/\log d_z$ όπου d_z είναι ο βαθμός του z.
3. Η τρίτη μέθοδος είναι η πιο σύνθετη. Για τον κόμβο v κάνουμε ένα τυχαίο περίπατο με επανεκκίνηση όπου η επανεκκίνηση γίνεται πάντα στον κόμβο v. Το σκορ για την σύνδεση (v, u) είναι η πιθανότητα του τυχαίου περίπατου να είναι στον κόμβο u στην κατανομή σύγκλισης.

Η δεύτερη προσέγγιση αντιμετωπίζει το πρόβλημα ως ένα πρόβλημα κατηγοριοποίησης και φτιάχνει ένα classifier, ο οποίος, για ένα ζεύγος (v, u), προσπαθεί να προβλέψει αν η σύνδεση αυτή θα εμφανιστεί ή όχι στο μέλλον. Για τα χαρακτηριστικά του classifier μπορούν να χρησιμοποιηθούν τα σκορ που αναφέραμε πριν καθώς και άλλα χαρακτηριστικά που έχουν να κάνουν με τις ιδιότητες του κόμβου, την συμπεριφορά του στο σύστημα, ή το δίκτυο.

Για την άσκηση αυτή θα πειραματιστείτε και με τις δύο προσεγγίσεις. Θα χρησιμοποιήσετε τα δεδομένα και το κοινωνικό δίκτυο που δημιουργήσατε για την Ερώτηση 3 στην Τέταρτη Σειρά Ασκήσεων (αναφέρετε ξανά τις λεπτομέρειες του δικτύου που δημιουργήσατε στην αναφορά σας). Διαλέξετε τυχαία 100 κόμβους οι οποίοι να έχουν τουλάχιστον 20 γείτονες. Αυτό θα είναι το test set. Αυτοί είναι οι κόμβοι για τους οποίους θέλουμε να κάνουμε πρόβλεψη. Θα κάνετε τρία διαφορετικά πειράματα:

1. Στο πρώτο πείραμα θα ακολουθήσετε την πρώτη προσέγγιση. Για καθένα από τους κόμβους στο test set, αφαιρέστε έναν από τους γείτονες τυχαία. Αυτή την ακμή θα προσπαθήσετε να προβλέψετε. Θα υπολογίσετε τα σκορ με τις τρεις μεθόδους που αναφέρουμε παραπάνω και θα κρατήσετε τους K συνδέσεις με τα μεγαλύτερα σκορ για $K = 1, 2, 5, 10$. Στην συνέχεια θα υπολογίσετε την ακρίβεια της πρόβλεψης ως το ποσοστό των κόμβων για τους οποίους η σύνδεση που αφαιρέσατε είναι μέσα στις K πρώτες προτάσεις. (Αν υπάρχουν ισοπαλίες υποθέστε ότι ο σωστός κόμβος προωθείται στην κορυφή). Για σύγκριση θα υλοποιήσετε και ένα αλγόριθμο Random ο οποίος διαλέγει τυχαία K κόμβους να προτείνει σαν πιθανές συνδέσεις. Φτιάξετε μια γραφική παράσταση με την μέση ακρίβεια (για τους 100 κόμβους) για τις τέσσερεις αυτές μεθόδους.
   Bonus: Προτείνετε, υλοποιήστε και τεστάρετε μια διαφορετική μέθοδο για να υπολογίζετε το σκορ μιας σύνδεσης.

2. Στο δεύτερο πείραμα θα φτιάξετε ένα classifier που χρησιμοποιεί χαρακτηριστικά που θα ορίσετε εσείς και θα προσπαθεί να προβλέψει για ένα ζευγάρι (u, v) αν θα δημιουργηθεί ακμή μεταξύ τους. Ο classifier θα πρέπει να έχει τουλάχιστον 10 features (μπορείτε να χρησιμοποιήσετε και τις επιχειρήσεις για να βγάλετε features). Εκπαιδεύστε τον classifier με τις υπάρχουσες ακμές καθώς και ένα ίσο αριθμό από μη υπάρχουσες ακμές (αυτές αντιστοιχούν στην αρνητική κλάση). Στο σύνολο εκμάθησης δεν πρέπει να συμπεριλαμβάνονται δεδομένα για τους 100 κόμβους που θα χρησιμοποιήσετε για τεστ. Για αυτούς τους 100 κόμβου κρατήστε 10 από τους γείτονες τους, και για όλες τις υπόλοιπες πιθανές συνδέσεις τρέξτε τον classifier για να προβλέψετε αν θα εμφανιστεί αυτή η ακμή στο μέλλον ή όχι. Τεστάρετε τουλάχιστον 2 μεθόδους κατηγοριοποίησης και αναφέρετε το μέσο (για τους 100 κόμβους) precision και recall των classifiers.
3. Τέλος θα εξετάσετε μια μέθοδο που συνδυάζει αυτές τις δύο προσεγγίσεις. Δημιουργήστε ένα Logistic Regression classifier για το πείραμα 2. Για ένα κόμβο v και μία πιθανή νέα ακμή (v, u) υπολογίστε την πιθανότητα που δίνει ο Logistic Regression classifier γι αυτή την ακμή. Χρησιμοποιήστε αυτές τις πιθανότητες για να βρείτε τις K πιο πιθανές ακμές και υπολογίστε τις ίδιες μετρικές όπως στο πείραμα 1. Βάλετε τη νέα καμπύλη στο ίδιο plot μαζί με τις υπόλοιπες.

Για όλα τα παραπάνω πειράματα, πάρετε τη μέση τιμή από 10 διαφορετικές επαναλήψεις με τυχαία επιλεγμένους 100 κόμβους.

Παραδώσετε τον κώδικα σας και μια αναφορά στην οποία θα περιγράφετε τις επιλογές που κάνατε στην υλοποίηση, καθώς και την λογική πίσω από τις επιλογές σας. Κάνετε την δική σας υλοποίηση του τυχαίου περίπατου με επανεκκίνηση. Στην αναφορά σας κάνετε ένα σχολιασμό στα αποτελέσματα. Όπως πάντα η αναφορά είναι πολύ σημαντική στην αξιολόγηση της εργασίας σας.