Πανεπιστήμιο Ιωαννίνων, Εξόρυξη Δεδομένων – ΜΥΕ 012/ΠΛΕ 059
Τμήμα Μηχανικών Πληροφορικής Χειμερινό Εξάμηνο, 2014

Τρίτη Σειρά Ασκήσεων

Η προθεσμία για την παράδοση αυτού του κομματιού είναι στις 13 Ιανουαρίου, στο ξεκίνημα του μαθήματος. Για καθυστερημένες υποβολές ισχύει η πολιτική στην σελίδα του μαθήματος. Οι λεπτομέρειες για το turn-in είναι στη σελίδα του μαθήματος. Παραδώσετε το απαιτούμενο υλικό, καθώς και όποιο κώδικα έχετε γράψει, και οδηγίες για το πώς τρέχει ο κώδικας σας.

Ερώτηση 1

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 2

Στην άσκηση αυτή θα δημιουργήσετε ένα classifier για spam emails. Θα χρησιμοποιήσετε το SpanAssassin dataset το οποίο διατίθεται online (το link είναι στη σελίδα του μαθήματος). Για να εκπαιδεύσετε τον classifier, χρειάζεστε δεδομένα εκμάθησης τα οποία να ανήκουν στην θετική κλάση (spam) και στην αρνητική κλάση (non-spam), τα οποία θα πάρετε από το site του SpamAssasin, και θα τα συνδυάσετε σε ένα dataset (τουλάχιστον 2000 παραδείγματα). Το dataset αυτό θα το χωρίσετε σε training (80%) και testing (20%) υποσύνολα.

Κάθε παράδειγμα είναι το κείμενο από ένα email σε ξεχωριστό αρχείο. Από το κείμενο θα δημιουργήσετε features για τον classifier σας. Κάποια από τα features θα πρέπει να είναι λέξεις από τα emails για τις οποίες θα δώσετε τιμή το tf-idf της λέξης. Από τις λέξεις μπορείτε να κάνετε μία επιλογή με αυτές που θεωρείτε πιο σημαντικές, π.χ., αυτές με το μεγαλύτερο βάρος, ή τις λέξεις που εμφανίζονται στα subjects, κλπ. Κατά τα άλλα, features μπορεί να είναι οτιδήποτε πιστεύετε ότι βοηθάει να διαχωριστούν τα spam emails από τα non-spam emails. Π.χ., το μέγεθος του μηνύματος, ή το domain του αποστολέα, κλπ, καθώς και συνδυασμός των παραπάνω. Ο classifier σας θα πρέπει να έχει τουλάχιστον 20 features (αν είναι όλες οι λέξεις θα είναι πολύ παραπάνω). Το κάθε email γίνεται πλέον ένα διάνυσμα στο χώρο των features.

Αφού αποφασίσετε τα features θα τα εξάγετε από τα κείμενα και θα δημιουργήσετε έτσι την είσοδο για training και testing. Για τον αλγόριθμο κατηγοριοποίησης χρησιμοποιήσετε το LibLinear πακέτο (link στη σελίδα του μαθήματος) το οποίο υλοποιεί τον Logistic Regression classifier (θα πρέπει να

μετατρέψετε την είσοδο σε αυτό που ζητάει το πακέτο). Τρέξτε μετά τον classifier που δημιουργήσατε πάνω στο test set.

Παραδώσετε μια αναφορά όπου θα περιγράφετε όλη την παραπάνω διαδικασία: Το πώς διαλέξατε τα δεδομένα, πως τα σπάσατε σε training και test σύνολα, τα features που δημιουργήσατε, τον αλγόριθμο που χρησιμοποιήσατε και τα αποτελέσματα του classifier στο test set. Για τα αποτελέσματα δώστε την ακρίβεια του classifier, καθώς και το confusion matrix. Βάλετε ένα κατώφλι θ στην πιθανότητα να είναι ένα email spam, και δημιουργείστε την precision-recall καμπύλη για την θετική κλάση, για $\theta \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$. Παραδώσετε επίσης τον κώδικα σας καθώς και τα τροποποιημένα αρχεία εισόδου και εξόδου (εφόσον τα έχετε τροποποιήσει).

Ερώτηση 3

Για την άσκηση αυτή θα υλοποιήσετε τον τυχαίο περίπατο με απορροφητικούς κόμβους που περιγράψαμε στην τάξη. Δοθέντος ενός γραφήματος και ενός συνόλου απορροφητικών κόμβων για κάθε μη απορροφητικό κόμβο θα υπολογίσετε την πιθανότητα να απορροφηθεί ο τυχαίος περίπατος σε έναν από τους απορροφητικούς κόμβους.

Θα εφαρμόσετε τον τυχαίο περίπατο σε ένα γράφημα που θα κατεβάσετε από εδώ. Οι κόμβοι του γραφήματος είναι blogs και συνδέονται μεταξύ τους αν το ένα αναφέρετε στο άλλο. Επίσης οι κόμβοι ανήκουν σε 39 διαφορετικά groups. Διαλέξετε τυχαία 10 κόμβους από κάθε group και κάνετε τους απορροφητικούς. Για κάθε μη απορροφητικό κόμβο υπολογίστε την πιθανότητα να καταλήξει σε κάποιο από τα 39 groups (η πιθανότητα να καταλήξει σε ένα group είναι το άθροισμα των πιθανοτήτων να καταλήξει σε κάποιο από τους 10 απορροφητικούς κόμβους του group). Τοποθετήστε τον κόμβο στο group με την μεγαλύτερη πιθανότητα. Με βάση την τελική ανάθεση, υπολογίστε τον πίνακα σύγχυσης και το precision και recall για κάθε group.

Παραδώστε τον κώδικα σας και μια αναφορά με τα αποτελέσματα σας με ένα σχολιασμό για το πόσο καλά αποδίδει αυτή η μέθοδος κατηγοριοποίησης των blogs.

Υπόδειξη: Αν και έχετε 390 απορροφητικούς κόμβους, ουσιαστικά χρειάζεστε να κρατήσετε μόνο 39 πιθανότητες.

Ερώτηση 4

Δουλεύετε στο TripAdvisor και είσαστε υπεύθυνοι για τη διαχείριση των online κριτικών (reviews). Επειδή το κάθε ξενοδοχείο έχει πάρα πολλές κριτικές θέλετε να διαλέξετε K από αυτές (όπου το K είναι μικρό, όχι πάνω από 10) που να περιγράφουν τα διάφορα **χαρακτηριστικά** του ξενοδοχείου όσο γίνεται καλύτερα. Υποθέστε ότι υπάρχουν m χαρακτηριστικά $A_1, \dots, A_m$ συνολικά για όλα τα ξενοδοχεία τα οποία είναι γνωστά εκ των προτέρων (π.χ., καθαριότητα, τοποθεσία, τιμή, προσωπικό, κλπ). Επίσης, μέσω κάποιας προεπεξεργασίας ξέρουμε τα χαρακτηριστικά που αναφέρονται σε κάθε κριτική. Δοθέντων N κριτικών για ένα ξενοδοχείο θέλουμε να διαλέξουμε K ώστε στην τελική συλλογή

να αναφέρονται (σε τουλάχιστον μία κριτική από τη συλλογή) όσο γίνεται περισσότερα από τα m χαρακτηριστικά του ξενοδοχείου.

- 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 του ξενοδοχείου?

Τεχνικές Λεπτομέρειες

Για το pre-processing, οι παρακάτω unix εντολές μπορεί να σας φανούν χρήσιμες:

- **cut**: επιτρέπει να πάρουμε συγκεκριμένες κολώνες από ένα αρχείο με διαχωριζόμενες τιμές
- **sort**: ταξινομεί τις γραμμές ενός αρχείου σε αλφαβητική σειρά . –n for αριθμητική σειρά
- **uniq**: αφαιρεί συνεχόμενες γραμμές που είναι ίδιες.
- **grep**: βρίσκει μια έκφραση μέσα σε ένα αρχείο.

Κάνετε “man <εντολή>” σε unix/linux για περισσότερες πληροφορίες για κάθε εντολή.