Πανεπιστήμιο Ιωαννίνων,
Τμήμα Μηχανικών Πληροφορικής

Εξόρυξη Δεδομένων – ΜΥΕ012/ΠΛΕ 059
Χειμερινό Εξάμηνο, 2015

Τρίτη Σειρά Ασκήσεων

Η προθεσμία για την παράδοση της Τρίτης Σειράς Ασκήσεων είναι στις 10 Ιανουαρίου, μέχρι το τέλος της ημέρας. Παραδώστε ηλεκτρονικά τον κώδικα και την αναφορά σας, είτε μέσω turnin είτε μέσω email. Για καθυστερημένες υποβολές ισχύει η πολιτική στην σελίδα του μαθήματος. Λεπτομέρειες για το turn-in, και για το πώς να γράφετε αναφορές είναι στη σελίδα Ασκήσεις του μαθήματος.

Ερώτηση 1

Μία power-law distribution is defined as $P(X = x) = (a - 1)x^{-a}$, όπου a είναι ο εκθέτης της κατανομής. Σας δίνεται ένα σύνολο από παρατηρήσεις $X = \{x_1, \dots, x_n\}$ που έχουν παραχθεί από μία power-law κατανομή. Χρησιμοποιήστε την Maximum Likelihood Estimation τεχνική που περιγράψαμε στην τάξη για να βρείτε τον εκθέτη της power-law κατανομής που ταιριάζει (fits) τα δεδομένα των παρατηρήσεων.

Ερώτηση 2

Ο στόχος αυτής της άσκησης είναι να πειραματιστείτε με αλγορίθμους ομαδοποίησης.

Σας δίνεται ένα αρχείο “bid_cat.csv” με ένα σύνολο από business ids από εστιατόρια στο Λας Βέγκας, καθώς και την κατηγορία τους. Μαζέψτε όλες τις κριτικές για αυτά τα εστιατόρια και δημιουργήστε ένα μεγάλο κείμενο για κάθε εστιατόριο. Χρησιμοποιήστε τον tf-idf vectorizer για να εξάγεται χαρακτηριστικά από αυτό το κείμενο για κάθε εστιατόριο.

Θα πειραματιστείτε πρώτε με τον k-means αλγόριθμο. Κάνετε πάντα την αρχικοποίηση χρησιμοποιώντας τον k-means++. Δημιουργείστε ένα γράφημα του λάθους ως προς το k, για k μέχρι 20. Χρησιμοποιείστε αυτό το γράφημα για να αποφασίσετε τον αριθμό των ομάδων και εξηγήστε την επιλογή σας. Χρησιμοποιώντας τις κατηγορίες από το αρχείο ως το ground truth δημιουργείστε τον πίνακα σύγχυσης και υπολογίστε το precision και recall για κάθε ομάδα όπως το περιγράψαμε στην τάξη (δεν δίνεται αυτόματα από τις συναρτήσεις της Python). Σχολιάστε τα αποτελέσματα ως προς τι περιέχει η κάθε ομάδα, πόσο καλή είναι η ομαδοποίηση. Προσπαθήστε να εξηγήστε την καλή ή κακή απόδοση του αλγορίθμου.

Αν διαφορετικό από αυτό που τρέξατε στο προηγούμενο βήμα, τρέξετε και για k=12, τον αριθμό των κατηγοριών, και κάνετε τις ίδιες μετρήσεις και τον ίδιο σχολιασμό.

Δοκιμάστε επίσης άλλο ένα από τους αλγορίθμους ομαδοποίησης που προσφέρεται από την Python. Κάνετε τις ίδιες μετρήσεις όπως πριν, συγκρίνετε με τον k-means και σχολιάστε τα αποτελέσματα.

Bonus: Στο json αρχείο, κάθε εστιατόριο σχετίζεται με πολλαπλές κατηγορίες. Προτείνετε και υλοποιήσετε ένα τρόπο να χρησιμοποιήσουμε όλες τις κατηγορίες για να αξιολογήσουμε την ομαδοποίηση.

Ερώτηση 3

Ο στόχος αυτής της ερώτησης είναι να πειραματιστείτε με αλγόριθμους κατηγοριοποίησης και πρόβλεψης.

Το συγκεκριμένο πρόβλημα με το οποίο θα ασχοληθείτε είναι να προβλέψετε αν μια επιχείρηση στο Yelp θα γίνει δημοφιλής ή όχι. Μια επιχείρηση θεωρούμε ότι είναι δημοφιλής αν ο αριθμός των reviews που θα πάρει είναι πάνω από τον μέσο όρο των reviews που έχουν επιχειρήσεις που ξεκίνησαν να αποκτούν reviews την ίδια χρονιά.

Σας δίνεται ένα αρχείο "bid_class.csv" ο οποίο περιέχει ένα σύνολο από business ids και την κλάση τους (True/False), ανάλογα με το αν η επιχείρηση έγινε δημοφιλής ή όχι. Ο στόχος είναι να εκπαιδεύσετε ένα κατηγοριοποιητή που θα μάθει να ξεχωρίζει τις δημοφιλείς από τις μη δημοφιλείς επιχειρήσεις. Έχουμε δύο περιπτώσεις

1. Στην πρώτη περίπτωση μπορείτε να χρησιμοποιήσετε για την δημιουργία του μοντέλου μόνο τα δεδομένα από το αρχείο `yelp_academic_dataset_business.json` (φυσικά όχι τον αριθμό των reviews)
2. Στην δεύτερη περίπτωση μπορείτε να χρησιμοποιήσετε πληροφορία για τα 10 πρώτα reviews της κάθε επιχείρησης. Το αρχείο "bid_rid.csv" το οποίο επίσης σας δίνεται περιέχει τα review ids για τα πρώτα 10 reviews για κάθε επιχείρηση. Μπορείτε να χρησιμοποιήσετε οποιαδήποτε πληροφορία για αυτά τα reviews.

Σχεδιάστε και εξάγετε από τα δεδομένα τα χαρακτηριστικά (features) που θα χρησιμοποιήσετε για την κατηγοριοποίηση. Αυτό είναι κάτι στο οποίο θα πρέπει να ξοδέψετε κάποιο χρόνο, τετριμμένες λύσεις (π.χ., με ένα μόνο χαρακτηριστικό) δεν θα βαθμολογηθούν. Πειραματιστείτε με δύο διαφορετικούς αλγόριθμους κατηγοριοποίησης που θα επιλέξετε εσείς. Κάνετε 10-fold cross validation για την αξιολόγηση των αποτελεσμάτων ενός αλγορίθμου. Αναφέρετε τα αποτελέσματα ως προς την ακρίβεια της πρόβλεψης καθώς και για το precision και recall ως προς την θετική κλάση. Λάβετε υπόψιν σας ότι η αρνητική κλάση είναι περίπου 80% των δεδομένων, άρα για να είναι καλός ο classifier σας θα πρέπει να έχει ακρίβεια πάνω από 80%.

Επιπλέον, υπάρχει ένας διαγωνισμός στο Kaggle για το μάθημα στο οποίο πρέπει να προβλέψετε τις κλάσεις για ένα νέο σύνολο από δεδομένα (εδώ είναι ο σύνδεσμος για τον διαγωνισμό). Δημιουργήστε ένα account με το email του πανεπιστημίου. Θα σας δοθεί πρόσβαση στον διαγωνισμό του μαθήματος και θα μπορέστε να καταθέσετε μια λύση για τον διαγωνισμό. Υπάρχει μία κατάταξη στο οποίο μπορείτε να δείτε την θέση σας σε σχέση με άλλες λύσεις. Η θέση σας δεν έχει σημασία στον βαθμό σας, αλλά θα πρέπει να καταθέσετε τουλάχιστον μία λύση. Στο μοντέλο που θα εκπαιδεύσετε για το Kaggle μπορείτε να χρησιμοποιήσετε για εκπαίδευση όλα τα δεδομένα για όλες τις επιχειρήσεις στο αρχείο. Μπορείτε να χρησιμοποιήσετε όποιο αλγόριθμο θέλετε.

Παραδώσετε τον κώδικα σας καθώς και μια αναφορά η οποία θα πρέπει να περιλαμβάνει:

a. Μια περιγραφή του τι κάνατε. Αυτό θα πρέπει να περιλαμβάνει και μια περιγραφή των χαρακτηριστικών που δημιουργήσατε σε κάθε περίπτωση και μια σύντομη εξήγηση γιατί τα διαλέξατε.

b. Τα αποτελέσματα σας και στο 10-fold cross validation, μια σύγκριση των αλγορίθμων, και τα αποτελέσματα στο σύνολο των δεδομένων στο Kaggle.
c. Ένα σχολιασμό πάνω στα αποτελέσματα: Πόσο καλά μπορούμε να προβλέψουμε? Ποια χαρακτηριστικά είναι σημαντικά? Υπάρχει κάποιο χαρακτηριστικό το οποίο να κάνει μεγάλη διαφορά στην κατηγοριοποίηση? Ξεχωρίζει κάποιος από τους αλγορίθμους?