University of Ioannina,
Department of Computer Science & Engineering

Data Mining – CSE 012/CS 059
Fall 2017

Assignment 3

The deadline for the second assignment is Saturday, February 10, at the end of the day. Turn in the code, with instructions on how to run it. The report should include detailed observations on the results. For late submissions the late policy on the page of the course will be applied. Details for the turn-in, and how to write reports are on the Assignments web page of the course. The will be an oral examination of the Assignment.

Question 1

In this question you will practice with algorithms for classification. The goal of this question is to experiment is to discriminate between businesses that belong to the categories 'Nightlife' or 'Bars', and the businesses that belong to the categories 'Cafes', 'Coffee & Tea', or 'Breakfast & Brunch'.

You will work with the businesses in the area of Toronto. Generate the data with the related businesses, and extract features from the text of the reviews for the businesses, as well as any other features you think are useful. You will experiment with four classifiers: Logistic Regression, SVM, Decision Trees, and Naïve Bayes. For the evaluation use 5-fold cross validation. Report the average confusion table, and the values for accuracy, precision, recall, and F1-measure.

In your report, describe the features that you used, and the results you obtained. Examine also the importance that the classifiers give to different features (e.g., the weights produced by the logistic regression classifier, or the first levels of the decision tree), and try to understand which words are the ones that discriminate between the different categories.

Question 2

In this question you will again experiment with classification algorithms. The goal is to build a classifier that can predict if a review for a restaurant (categories Food, Restaurants) is negative (has star rating 1 or 2).

For this question a Kaggle competition was created for the class (here is the link to the competition). Create an account with the email of the university. You will be given access to the competition through the link, and you can then submit a solution for the competition. In the competition you will be provided with the training and test data.

There is a ranking where you can see your position with respect to other solutions. A good position will boost your grade. The important is to obtain a good F1-score compared to the other solutions. You can use any classification algorithm you wish.

In the report, describe the features that you created, and the model that you used, and report your Kaggle account name, and your score. You can also discuss about experiments that you did and did not work out, or how you improved upon a solution that did not work well (add also the numerical results).

Question 3

Prove that for an undirected graph the stationary distribution of a random walk is proportional to the degree of the nodes. If P is the transition matrix of the random walk, and π is the stationary distribution for which $\pi = \pi \cdot P$, show that for node i the probability π_i is proportional to d_i where d_i is the number of edges incident on node i.

Question 4

In this question you will use the data that you created for Assignment 2 on Recommendation Systems. The goal is to use the social network between the Yelp users in order to predict their ratings for new businesses.

Start with the dataset from Assignment 2, consisting of businesses in Toronto that have at least 10 reviews from users with at least 10 reviews. Using these users as nodes of the graph, construct the graph consisting of friendship edges between them, which you will obtain from the file user.json. From this graph keep the largest connected component. This will define the graph G that you will work with, and the nodes in the component the set of users that we are interested in (the set of businesses remains the same).

Remove randomly 10% of the ratings of the users, and try to predict the rating for the user-business pair (u, b) using a random walk with absorbing nodes as follows. Given a pair (u, b) , in the graph G, make every node v that has rated the business b to be absorbing, and assign to that node the value of the rating $R(v, b)$. Using the value propagation method we described in class compute a rating $P(v', b)$ for every non-absorbing node v' in the graph. The predicted rating for node u will be the value $P(u, b)$.

Compute the Root Mean Sum of Square Errors (RMSE) that you obtain with this technique. Then run the algorithms that you implemented in Assignment 2 for this dataset, and compare the RMSE you obtain. Present your results and your observations.

Bonus: Propose, implement, and test a different method for predicting the ratings using random walks on the social graph.

Question 5 (Coverage)

A practical problem when handling graphs is to compute the distance $d(x, y)$ (the length of the shortest path) between two nodes (x, y). The computation is costly, so there is a need for indexes for answering such queries. These indexes keep a small set L of landmark nodes for which we keep the distances to all other nodes, and we use them to answer shortest path queries.

1. Show that for every landmark $\ell \in L$, it holds that $d(x, y) \leq d(x, \ell) + d(\ell, y)$. When does the equality hold?
2. Show that the problem of finding the smallest set L of landmark nodes, so that we can answer exactly shortest path queries for any pair of nodes (x, y) can be expressed as minimum set cover problem.

Question 6

In Kaggle there is an active competition for predicting “toxic” online content in online discussions. (here is the link to the competition). Using the Kaggle account you created for the previous question, submit a solution to the competition. The goal is not to win the competition (although your position will boost your grade), but to work on a real problem that has no known solution.

Create a report that contains the following:

a. A description of the solution you implemented. What features you used, what techniques you applied, and a short explanation of the rationale of your choices.
b. Your results on the Kaggle test dataset.
c. A commentary on the above: What seems to work and why? What insight did you gain into the data and the problem?

Hand in your code, and the report. Make sure to note your user-name in Kaggle, and your standing at the time that you submitted the report.