University of Ioannina,
Department of Computer Science & Engineering

Data Mining – CSE 012/CS 059
Fall 2015

Assignment 3

The deadline for Assignment 3 is January 10, by the end of the day. Submit everything electronically, either through turn-in or email. 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.

Question 1

A power-law distribution is defined as $P(X = x) = (a - 1)x^{-a}$, where a is the exponent of the distribution. You are given a set of observations $X = \{x_1, \dots, x_n\}$ that are generated from a power-law distribution. Use the Maximum Likelihood Estimation method we described in class to find the exponent of the power-law distribution that best fits the data observations.

Question 2

The goal of this question is to experiment with clustering algorithms.

You are given a file “businesses.tsv” with a set of business ids of restaurants in Las Vegas, as well as their category. Collect all the reviews for these restaurants to create a large text document for each restaurant. Use the tf-idf vectorizer to extract features from this document for each restaurant.

You will first experiment with the k-means algorithm. You should always initialize with k-means++. Create a plot of the error vs k, for k up to 20. Use this plot to decide on the number of clusters, and explain your choice. Using the categories in the file as the ground truth, produce a confusion table, and compute precision and recall per cluster as described in class (not automatically provided by python). Comment on the results, as to what each cluster seems to capture, how well the clustering performs. Try to explain the good or bad performance of the algorithm.

If different from what you have chosen before, do a run for k=12, the number of the categories, and perform the same measurements and commentary.

Try also one more clustering algorithm from those provided by python. Do the same measurements, compare with k-means and comment on the results.

Bonus: In the original json file each restaurant is associated with a list of categories. Propose and implement a way to use all the categories to evaluate the clustering.

Question 3

The goal of this question is to experiment with classification.

The problem you will work on is to predict if a business in Yelp will become popular or not. We consider a business to be popular if the number of reviews it will obtain is above the mean number of reviews of businesses that got their first review in the same year.

You are given the file "bid_class.csv" which contains a set of business ids and their class (True/False), depending on whether the business became popular or not. The goal is to train a classifier that will learn to distinguish between popular and unpopular businesses. There are two tasks:

1. In the first case, you can use for training your model only information from the file `yelp_academic_dataset_business.json` (obviously, you cannot use the field "number of reviews").
2. In the second case you can use also information about the 10 first reviews of each business. You are given the file "bid_rid.csv" which contains the review ids for the first 10 reviews of each business id. You can use any information you want from these reviews.

Design and implement the features that you will use for classification. This is something on which you will need to spend some time. Trivial solutions (e.g., a single feature) will receive zero marks. Experiment with two different classification algorithms of your choice. Use 10-fold cross validation for your evaluation. Report results on the accuracy of the classifier, as well as the precision and recall for the positive class. Keep in mind that the negative class is about 80% of the data, so for a classifier to perform well it should have accuracy more than 80%.

In addition there is a competition in Kaggle for the class, in which you are asked to predict the classes for a new dataset (here is the link to the competition). Create an account with the university email and you should have access to the competition. There is a leaderboard on which you can see how your solution ranks compared to the other solutions. Your position is not important in your grade, but you should submit at least one (reasonable) solution. In the model that you will train for the Kaggle competition you can use all the data for training, for all the businesses in the file. You can use any algorithm you choose.

Hand in your code, and a report which should contain the following:

a. A description of what you did. This should contain a description of the features you created for each case, and a short explanation of the rationale of your choice.
b. Your results on the 10-fold cross validation for the different classification models, and the results on the Kaggle dataset.
c. A commentary on the above: How well can we predict? Which features are important? Is there some feature that makes a big difference in the classification? Does some classification algorithm stand out?