Overview

Sometimes we make decisions beyond the rating of a restaurant. For example, if a restaurant has a high rating but it often fails to pass hygiene inspections, then this information can dissuade many people to eat there. Using this hygiene information could lead to a more informative system; however, it is often the case where we don’t have such information for all the restaurants, and we are left to make predictions based on the small sample of data points.

In this task, you are going to predict whether a set of restaurants will pass the public health inspection tests given the corresponding Yelp text reviews along with some additional information such as the locations and cuisines offered in these restaurants. Making a prediction about an unobserved attribute using data mining techniques represents a wide range of important applications of data mining. Through working on this task, you will gain direct experience with such an application. Due to the flexibility of using as many indicators for prediction as possible, this would also give you an opportunity to potentially combine many different algorithms you have learned from the courses in the Data Mining Specialization to solve a real world problem and experiment with different methods to understand what’s the most effective way of solving the problem.

About the Dataset

You should first download the dataset. The dataset is composed of a training subset containing 546 restaurants used for training your classifier, in addition to a testing subset of 12753 restaurants used for evaluating the performance of the classifier. In the training subset, you will be provided with a binary label for each restaurant, which indicates whether the restaurant has passed the latest public health inspection test or not, whereas for the testing subset, you will not have access to any labels. The dataset is spread across three files such that the first 546 lines in each file correspond to the training subset, and the rest are part of the testing subset. Below is a description of each file:

- hygiene.dat: Each line contains the concatenated text reviews of one restaurant.
- hygiene.dat.labels: For the first 546 lines, a binary label (0 or 1) is used where a 0 indicates that the restaurant has passed the latest public health inspection test, while a 1 means that the restaurant has failed the test. The rest of the lines have "[None]" in their label field implying that they are part of the testing subset.
- hygiene.dat.additional: It is a CSV (Comma-Separated Values) file where the first value is a list containing the cuisines offered, the second value is the zip code, which gives an idea about the location, the third is the number of reviews, and the fourth is the average rating, which can vary between 0 and 5 (5 being the best).

Note that the training subset is perfectly balanced, i.e., the number of restaurants with label 1 is equal to those with label 0. However, the testing subset is imbalanced where the majority of restaurants have a label of 0 (meaning that they have passed the inspection). Due to this imbalance, the classification accuracy may not be a suitable measure for evaluating the performance of classifiers. Therefore, we will use the F1 measure, which is the harmonic mean of precision and recall, to rank the submissions in the leaderboard. The F1 measure will be based on the macro-averages of precision and recall (macro-averaging is used here to ensure that the two classes are given equal weight as we do not want class 0 to dominate the measure).

Instructions

As you have probably noticed, this task is similar to Task 4 in the programming assignment of the Text Mining and Analytics course; however, there are three major differences:

1. The training data is perfectly balanced, whereas the testing data is skewed, which creates a new challenge since the training and testing data have different distributions.

2. The main performance metric is the F1 score as opposed to the classification accuracy that was used in the Text Mining course. This means that a good classifier is expected to perform well on both classes.
3. Extra non-textual features such as the cuisines, locations, and average rating are given. This might help in further improving the prediction performance and provide an opportunity to experiment with many more strategies for solving the problem.

You are free to use whatever toolkit or programming language you prefer. See the Getting Help section below for some useful toolkits that are good candidates to complete this task. You should train a classifier over the 546 training instances and then submit the binary predictions for the remaining 12753 instances, each on a separate line. The first line should contain the nickname that you want to have on the leaderboard, i.e., the output file should have the following format:

NicknameLabel1Label2...Label12573

Submission

Report. We suggest that it be 1-3 pages long. The report will be peer-graded. Your report will need to include the following elements:

- A brief description and comparison of all the methods you tried. By "methods" we are referring to the text processing techniques, feature representation and selection, and learning algorithms you experimented with. Try to explain why some methods are performing better than others and include a failure analysis (i.e., looking at particular cases where prediction is incorrect to understand where you might be able to further improve a method).
- Details about the method that gave the highest F1 score.