Search Engine Technology Fall 2012 Homework 2

Prof. Dragomir Radev, TA: Siyang Dai, Hongzhi Li

Due Date: Nov. 4th 11:59:59pm

1 Overview

In this assignment, you will build a classifier for sentiment analysis in movie reviews, which contains the following steps:

1. Preprocessing
2. Feature Extraction
3. Feature Selection
4. Classifier Implementation

You can choose whatever language you like. And you can resort to any third party libraries for the preprocessing part. As to the rest of the assignment, we expect that you will implement your own classifier from scratch.

1.1 Dataset

The movie review dataset contains thousands of reviews in csv format. Each row is one example. Each example consists of two columns. The first column is the class label and the second column contains a short sentence. Positive reviews are labeled as 1 while negative reviews are labeled as 0.

1.2 Deliverables

1. Submit your predictions on Kaggle. (one class label per line in the same order of the test set).
   `http://inclass.kaggle.com/c/cs6998` (Please register with your columbia email.)

2. Submit your code and report on courseworks. (Instructions are provided below.)
   `https://courseworks.columbia.edu/`

Your code should be a self-contained package, including any third party library and a README describing how to run your program. Make sure that your code can be run on clic machines. Your code should have two high level programs such as:

./train training_set.csv model_file where training_set.csv is the input, and model_file is the output which contains the saved model. If possible, please save the model in human-readable format. For example, the weights/probabilities on terms ("politics":0.5) for linear classifier/Naive Bayes, the conditions of decision tree ($age < 18$) etc.

./test model_file test_set.csv prediction_file where the first two parameters are the input to the program and the last one is the output file. The model_file is the trained model, test_set.csv is the input test set, and prediction_file is a line separated file with one class label per line in the same order as the test set.

You report should include a detailed description of your approach and your design decisions (as well as your Kaggle id). Your report should contain the following sections: Preprocessing, Feature Extraction, Feature Selection, Classifier Implementation and Reference. We expect a 2-4 pages write-up.

The TAs will randomly choose a fraction of the submissions and check their code. We will also select 4-5 submissions to be presented in class.

2 Preprocessing

You can choose any way to preprocess the data, including stemming and/or text normalization. Note that preprocessing is not guaranteed to improve classification performance. Please refer to chapter 2.2. of the IR book.

3 Feature Selection

You should choose ONE method for feature selection from these two:

1. χ^2 test
2. Mutual Information

Please refer to chapter 13.5 of the IR book.

4 Classifier Implementation

We recommend that you implement one of the following algorithms.

1. Naive Bayes Classifier
2. k-Nearest Neighbor Classifier
3. Decision Tree Classifier
4. Perceptron
5. Support Vector Machines
6. Semi-Supervised Classifier

Please refer to chapters 13 to 15 of the IR book.

For semi-supervised learning, you can obtain more data from other sources like IMDB. Potentially, it might give better performance, but it is more difficult. For fairness, if you obtain other labeled movie review dataset for sentiment analysis, please ignore the labels to fulfill the spirit of semi-supervised learning.

For more information, please refer to Xiaojin Zhu's tutorial (`http://pages.cs.wisc.edu/~jerryzhu/pub/sslicml07.pdf`). A more detailed survey is also available (`http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.99.9681&rep=rep1&type=pdf`).

Bing Liu's tutorial (`http://www.cs.uic.edu/~liub/FBS/Sentiment-Analysis-tutorial-AAAI-2011.pdf`) provides a comprehensive introduction to sentiment analysis. Check the work by Janyce Wiebe as well.

5 Grading Criteria

In general, your grade will be evaluated based on:

1. Methodology (60%).

   In your report, you should describe your approach in detail. Give enough data examples and experimental results as necessary so that we can understand your approach. Discuss the tradeoffs of your choices.

2. Creativity (10%). We encourage you to come up with something new.
3. Competition Rank (30%). We will use a normal bell curve to assign grades for this part based on the highest accuracy and lowest accuracy.

One thing to note is that, it is not very difficult to obtain the original dataset. But if you abuse the dataset, i.e. submit the labels directly, use it as a validation set etc, you will get 0.

5.1 Extra Credit

Some of the top students who do well in the competition will be asked to give a presentation in class and may get extra credit for that.

5.2 Late Submission Policy

Do not submit your report/code directly to the TAs. The courseworks and Kaggle will still accept submission for a few days after the deadline. But each day (or fraction of a day) that your homework is late, your grade will decrease by 10 points. If the courseworks and Kaggle submission pages are closed and you haven't submitted anything, you will get 0.