Big Data Science Course HW#2

Project Description

1 Overview

In this project, you will apply the machine learning techniques learned in this course to a real world dataset. Academia and training courses focus on applying a specific algorithm to carefully preprocessed toy datasets. This project will require that you make decisions such as which classifier to use, which features to use, how to train the classifier, etc. You will find that real world datasets are more difficult to work with, and that careful tuning is required to get optimal performance.

2 Dataset

In this project you will work with a labeled dataset of news articles.

- Each instance in the dataset consists of the text of a New York Times article published between 2000 and 2003. Each article comes from one of four newspaper sections: News, Classifieds, Opinion, and Features. This dataset has been culled from the Linguistic Data Consortium's New York Times Annotated Corpus.
- For your convenience the text has been slightly preprocessed: punctuation and formatting (such as paragraphs) have been removed, and all characters have been converted to lower case.
- The articles are stored in data train.txt and data valid.txt. Each line in the file is a single article. Each article consists of a space-separated list of words.
- The labels (the sections the articles came from) are stored in the labels train.txt and labels valid.txt.
- Each line in the file is a single label. The nth line in the labels train.txt file contains the label for the article stored on the nth line of the data train.txt file.
- The labels are numerical to make it easier to use them, but we have also provided the original text labels in labels train original.txt and labels test original.txt. The mapping is:
    - News: 0
    - Opinion: 1
    - Classifieds: 2
    - Features: 3

YOUR TASK is to write a classifier which can correctly predict the label of a given article. It will not be sufficient to naively implement one of the classification algorithms from class; you will need to implement some more sophisticated methods of feature generation, classification, or both. For instance, you might experiment with various techniques to combat over fitting, or you might use a dimensionality reduction technique to generate better features.

3 Deliverables

This is a 3- to 4-page document that describes how you solved the problem. It should contain detailed discussion about which methods you tried, results on their comparative performance, and a discussion about which methods worked best and why. It should follow the following simple outline: Introduction, Methods, Results, Conclusion. We expect that your methods should produce much better than baseline accuracy (25% correct), 50% is easily achievable, and the best achieved is 69.5%.

Big Data Science Course HW#2

4 Suggested Techniques

A few classification techniques you may want to review/investigate:

1. Logistic regression
2. Neural networks
3. k-nearest neighbors
4. Support vector machines

A few techniques for generating features that you may want to investigate:

1. Dimensionality reduction
2. The “bag of words" model
3. n-gram models (including smoothing techniques)
4. Other kinds of vector space models (see http://www.jair.org/media/2934/live-2934-4846-jair.pdf for several types of such models)
5. TF-IDF