CSL 603 (Machine Learning)

Assignment 1

Sentiment Analysis of movie reviews

Shreyanshu Shekhar

2016csb1060@iitrpr.acin

Aug 28, 2018

1 Introduction

This program predicts the sentiment of a movie review. Data used was the Large Movie Review Dataset from Stanford for running our experiments. The data set is the string of movie review. I trained my model by supervised learning, I used decision tree and decision forest. It takes 5000 features and1000 instances. Various experiments have been done and their results are shown below.

2 Experiment

2.1 Simple Decision Tree

Taken 5000 features (2500 positive reviews and 2500 negative reviews)

Taken 1000 instances (500 words with positive and 500 with negative emotional sentiments)

Number of nodes : 573

Accuracy on training data set : 83.72781065088756

Accuracy on test data set : 64.22018348623853

2.2 Early Stopping

To implement early stopping, while creating the tree, I checked the number of positive and number of negative reviews, and find their ratio. If the number of the positive reviews is greater than the negative review (or vice-versa) by a certain threshold then it stops creating nodes further. I checked it for different values.

For ratio threshold : 0.03

Number of nodes : 429

Accuracy on training data set : 85.94674556213018

Accuracy on test data set : 65.13761467889908

For ratio threshold : 0.035

Number of nodes : 519

Accuracy on training data set : 85.94674556213018

Accuracy on test data set : 64.98470948012233

For ratio threshold : 0.04

Number of nodes : 420

Accuracy on training data set : 85.94674556213018

Accuracy on test data set : 64.22018348623853

2.3 Noise Addition

To implement noise addition, I picked a review randomly from the list of both from positive and negative review list and interchanged them. I did this process of a certain percentage of data.

For 0.5% of the of the dataset

Accuracy on training data set : 83.72781065088756

Accuracy on test data set : 65.13761467889908

For 1% of the of the dataset

Accuracy on training data set : 83.72781065088756

Accuracy on test data set : 64.52599388379205

For 5% of the of the dataset

Accuracy on training data set : 83.72781065088756

Accuracy on test data set : 65.29051987767585

2.4 Pruning

The decision tree is pruned and checked on the test dataset if it increases the accuracy or not. If it increases the accuracy it is kept. It results in the decrement of height and number of nodes of the tree.

Accuracy on training data set : 83.72781065088756

Accuracy on test data set : 64.37308868501529

2.5 Random Forest (Feature Bagging)

Randomly I removed features from the 5000 feature set and created a new feature set. And for every set, I created a tree and checked their average accuracy.

For 5 trees

Accuracy on training data set : 70.32544378698225

Accuracy on test data set : 67.1559633027523

For 10 trees

Accuracy on training data set : 68.40236686390533

Accuracy on test data set : 67.75229357798165