Machine Learning Project Update

Krishna Shukla
Simon Fang

December 2016

1 Dataset

We are going to use the dataset *Political Social Media Post* that is obtained from Kaggle and uploaded by the user CrowdFlower. It contains 5000 messages from US politicians' social media accounts. For each message, the dataset has a human judgment about the purpose, partisanship, and audience of the messages. The politicians range from US senators to local politicians.

2 Problem

In the midst of political uncertainty, we thought it would be interesting to predict the bias of a message on social media. Our efforts will be threefold. (1) Our first goal is to predict what words would indicate whether a message is partisan or neutral. (2) Secondly, we would like to predict whether certain words indicate a support message or an attack message. (3) Finally, we investigate the usage of different social media, i.e., do politicians use Twitter and Facebook for different purposes?

3 Approach

For our first two goals, we will be making use of a Decision Tree Learning algorithm. For our last goal, we will try to make use of support vector machines (SVM).

4 Evaluation

Does this refer to ROC/AP? If this is the case, then I think it would make sense to make use of Average Precison. Especially when we are going to make use of Support Vector Machines, then we can calculate the Accuracy and Recall of our decision boundary.

5 Progress update

We have been trying to pre-process the data, but we are not really familiar with the processing of natural language data. We ar trying to make some progress in the coming days. Moreover, we are trying to understand what features we are going to use for our Decision Tree Learning algorithm. The same question also arises for our Support Vector Machine. We are planning on tackling these problems in the coming days and otherwise we would like to discuss with the teachers on Monday.