Predicting Congressional Bill Outcomes

Vivian Hu and David Wu

January 23, 2015

1 Problem

Congressional bills can be long, dense, and esoteric. Whether or not they will be voted into law depends on many factors, but the results affect each and every one of us. How well can we predict whether a Congressional bill will be made into a law? We propose to predict legislative bill voting outcomes in Congress as a function of bill text.

This is a classification problem; specifically, a binary text classification problem in which the inputs are legislative bill texts, and the output is a prediction of the bill's success. Further analysis can be done on limiting data sets to certain congresses or legislative categories.

2 Method

We plan to process topics from text using topic models, which we will use as parameters in a binary classification algorithm. We also plan to explore simpler models that only consider the frequency of words like bag-of-words, but hypothesize that they are more suited to attribution than prediction.

In order to predict bill voting outcomes based on text, we must first extract "topics" from the text. For topic extraction, we plan to use a topic model [1] such as latent Dirichlet allocation (LDA). LDA [2] is a natural language processing model that allows us to derive a small number of topics from a text. We may also consider Latent semantic indexing [3], which might be helpful. From these unsupervised learning methods, we can generate abstract "topics" that represent textual themes of a bill.

Unlike Yano, Smith, and Wilkerson [4], we don't plan to use supervised learning techniques to trained binary logistic regression models to predict bill categories. Unless we find good data sets to use as training data, we plan to use unsupervised techniques like those detailed above to extract topics.

Finally, we plan to use the topics as features, perhaps along with relevant keywords, in a binary classifier. The main classifier we plan to consider is the random forests method [5], which essentially trains multiple decision trees and outputs the average of the results of the trees. Other classifiers we may look at include Linear Discriminant Analysis and the meta-algorithm AdaBoost in conjunction with the previous two. There is also much to be learned in comparing the effectiveness of the classification algorithms detailed above with ones learned in class, such as logistic regression and SVMs.

3 Data Sets

One comprehensive data-set we plan to use is found at `www.govtrack.us` [6]. This contains legislative bill text and voting results for all congresses up to the present, with good data from 1973 onwards. Specifically, it contains a bulk data dump that represents bill text and their voting results as JSON or TXT files.

4 Milestone

By the project milestone, we plan to have selected and thoroughly understood a topic model and classification algorithm. We also plan to pre-process our congressional bill dataset, sanitize, and regularize it if needed. Finally, we will have implemented a good deal of the topic model and classifier.

References

[1] http://www.cs.princeton.edu/~blei/papers/Blei2012.pdf

[2] http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

[3] http://www.cse.msu.edu/~cse960/Papers/LSI/LSI.pdf

[4] http://dl.acm.org/citation.cfm?id=2382157

[5] https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm

[6] https://www.govtrack.us/congress/votes