Nathan Canterbury

Directed Study Results – 11/3

Models

Two models in WEKA, varying accuracy and ROC curve

1.

SMO Model, hyper parameters

General Classifier

Correctly Classified Instances – 74.9%

Incorrectly Classified Instances – 25.08%

ROC Area – 0.732

However, stratified cross-validation is not as good.

Correctly Classified Instances – 66.5%

Incorrectly Classified Instances – 33.44%

ROC Area – 0.648

2. J48 Tree

General thoughts

* Worried about over-fitting? ROC curve is among the best I’ve been able to achieve, but lack of data may contribute to over-fitting? However, this is with a modified attribute list. Only using 5.
* Seem to be much better at predicting pro-Obama than pro-Romney. Look at confusion matrix.

General Classifier

Correctly Classified Instances – 68.47%

Incorrectly Classified Instances – 31.52%

ROC Area – 0.675

Cross-validation

Correctly Classified Instances – 70.40%

Incorrectly Classified Instances – 29.59%

ROC Area – 0.717

Attribute Selection

I looked at four different models of attribute evaluation.

1. CFS Subset Evaluator

- Solely selects Obama and Romney.

2. CorrelationAttributeEval

- Obama and Romney are also the highest ranked attributed within the CorrelationAttributeEval, though the rankings are only 0.28 and 0.27, respectively.

3. InformationGain

- Obama and Romney are the only attributes that produce any attribute gain, and even then they are both low values.

- Obama: 0.0585

- Romney: 0.0547

- Thoughts: It makes some sense to me that there would be very little information gain from our attributes. I wouldn’t expect there to be a significant information gain off splitting the data on single attributes such as pos tags. I think that with the data we are dealing with, it is hard to find any single attribute that effectively splits the data to the extent that it will discriminate classes. Rather, I think a tree model is where we are best able to focus in on a specific class, with the combination of several attributes. A worry here is of course over-fitting the data. This is where having a larger data set would be valuable.

4. WrapperSubsetEval with BestFirst Search

- I primarily ran simulations of the J48 tree as I consistently saw the best results with the model.

- Merit of best subset found: 0.709

- Selects 5 attributes: negative\_words, O, L, R, Romney.

- I think that this is informative as to what is the most effective combination of attributes.

Final thoughts:

* Any thoughts on boosting the accuracy a little bit more?
* Did not see difference in accuracy upon normalizing and standardizing.
* Should we consider using the amazon services to get more labeled data?
* Thinking about which two models I would now like to build myself. Will begin building this week. Tree-based (random forest?) vs. more general? (Naïve Bayes)

Comments – Building model based on context of problem, not necessarily existing ideas

Results section – have to build a graph that shows the Obama popularity and Romney popularity. Show correlation between rating of candidate and a certain event.