ML 310 Applied Machine Learning

Homework #3 Assigned: 2/11/17 Due: 2/23/17 [100 points]

GOAL: Please become conversant with concepts and implementation using R for the below two topics we discussed in class. Then complete the below exercises.

Section I. Text mining case study

1. Study this Text mining case study [30 points]

https://rstudio-pubs-static.s3.amazonaws.com/31867_8236987cf0a8444e962ccd2aec46d9c3.html

Then, conduct your own Text Analysis on the **textmining.zip** data set provided in Canvas and provide explanation of results. Pl. make sure to include these sections:

- Explore your data
  - Word Frequency
  - Plot Word Frequencies
- Relationships between Terms
  - Term Correlations
  - Word Clouds!
- Clustering by Term Similarity
  - Hierarchal Clustering
  - K-means clustering

Section II. Twitter analytics

2. Study these four R case studies on Twitter analytics as discussed in class: [15 points]

Twitter analytics in R Tutorials:

http://www.r-bloggers.com/getting-started-with-twitter-analysis-in-r/

http://www.r-bloggers.com/in-depth-analysis-of-twitter-activity-and-sentiment-with-r/

http://www.rdatamining.com/examples/text-mining

Then, conduct Twitter Analysis on a data set you can extract/prepare from the link below and provide explanation of results.

https://www.cs.york.ac.uk/semeval-2013/task2/index.php%3Fid=data.html

Useful Ref: Text data sets: http://disi.unitn.it/moschitti/corpora.htm

Section III. Support Vector Machines

1. Study the SVM R coding/tutorials at the below links: [25 points]

   http://www.svm-tutorial.com/2014/10/support-vector-regression-r/

   http://www.r-bloggers.com/learning-kernels-svm/

Then, use the below data set and apply SVM using R libraries to classify Mushrooms using the **Mushroom data set.**

https://archive.ics.uci.edu/ml/datasets/Mushroom

Section IV. Evaluating Credibility [10 points]

Define Sensitivity and Specificity as they relate to a Classifier. How does a ROC curve depict the performance of a Classifier? Please explain using a diagram, and the below data plotted as ROC.

Microscope setting	% of virus strains detected	% of virus strains correctly identified
Off	0	100
Setting 1	35	93
Setting 2	60	85
Setting 3	85	70
Setting 4	92	30
Full	100	0

Section V. Ensemble Methods: Boosting [20 points]

Conduct a Boosting based Ensemble enhancement for the **Mushroom data set** above. Please explain how Boosting enhanced yor model performance compared to the abse algorithm without Boosting.