CSC581 Midterm Spring 2018
due Wednesday 3/21 in Sakai

version 1.1

Problems

For the midterm examination you are to build and evaluate support vector machine models for the data set that you selected in your midterm proposal.

Part A Perform an exploratory data analysis using summary statistics and histograms. Briefly explain your findings. You can reuse material from your midterm proposal here.

Part B Build the best model possible for your data set:

1. Document your grid search/model evaluation process carefully, including the type of kernel you are using, the values of its free parameters, and the value of C.
2. Use the cross-validated error/accuracy in order to determine your best model.
3. Select the **two best performing** models.

Part C Construct a confusion matrix for each of the two top models.

1. Are the models balanced in terms of the type of errors they commit?
2. Is one model preferable over the other given the type of data you are analyzing?

Part D Investigate whether the difference in performance of your top two models is statistically significant or not using the bootstrap. You should use 95% confidence intervals for this investigation. Also, use the hold-out method in order to compute the error of the model(s). **Do not use cross-validation - the svm internal cross-validation constructs a stratified sample thereby effectively destroying your bootstrap sample.**

1. What are the 95% error confidence intervals for your two models?
2. Is the performance difference statistically significant? If yes, which model would you pick? if no, which model would you pick and why?

Write a brief report summarizing your findings from Parts A, B, C, and D.