CSCI 452 (Data Mining)
Dr. Schwartz
Support Vector Machines
100 pts

This assignment gives you practice with Support Vector Machines. This is an individual assignment. Note that throughout this lab, both the results and the analysis are graded. Be sure to provide your analysis, hypotheses and conclusions in your writeup. You will create an R Markdown file called svms.Rmd and submit both the .Rmd and resulting .html (after knitting) files to autolab. Please be sure to follow the order of the questions and identify your answers/analysis clearly so that I can easily find the individual pieces.

When working with your data files, be sure that you store them in the same directory as your .Rmd file (your working directory). As long as your independently selected data set isn't huge (<1Mb), please include it in the .zip file that you submit. If the data file is on the larger side, be sure to include a link in the comments on where I can find the data.

1) (20 pts) Background (applied example)

Read this paper: Grazyna Suchacka, Magdalena Skolimowska-Kulig, Aneta Potempa: *Classification Of E-Customer Sessions Based On Support Vector Machine*. ECMS(European Conference on Modelling and Simulation) 2015: 594-600 *http://www.scs-europe.net/dlib/2015/ecms2015acceptedpapers/0594-dis_ECMS2015_0120.pdf*

In your submission, answer the following questions about the paper:

a) Briefly describe the problem the authors are trying to solve.
b) How many predictor variables were used?
c) What was the ratio of training data to test data?
d) What kernels did the authors evaluate and how did they tune them?
e) Although overall accuracy was similar in several of the kernels, the linear kernel was strongly preferred. Why?

2) (50 pts) SVM in R Use the Vehicle dataset in the *mlbench* library (https://www.rdocumentation.org/packages/mlbench/versions/2.1-1/topics/Vehicle) and the svm methods in the e1071 library. Document the process and results in your submission file.

a) Examine the data (start from the original data) and do any preprocessing you deem necessary. Document your decisions.
b) Create five (or more!) different training/test set divisions (you can do this more or less programmatically or manually, as you choose). You can use sampling without replacement – the class balance is relatively even in this dataset. You'll want to save and be able to reproduce/reuse these divisions for later steps. Use the same ratio of training to test in each of your sets.

c) Create SVMs with all of your training data for various kernels (linear, polynomial, radial/RBF). Which kernel will you choose to use based on this evaluation?
d) Using your selected kernel, tune the parameters (linear – just cost, radial – gamma and cost, polynomial – gamma, cost and degrees). This piece may take a bit of research on the tune (or svm.tune) function. Show the tuning process and the results (the best parameter(s)).
e) Using the best parameters from the previous step, evaluate all of your training and test data. Analyze your results.

3) **(30 pts) R and Your Choice of Data** Obtain a dataset from the UCI Machine Learning Repository. http://archive.ics.uci.edu/ml/datasets.html (Do not use any of the datasets that we explored in class). Choose a classification problem (no ordinal target values). Perform analysis similar to what you did in part 2 – that is, decide on and do any preprocessing steps, divide the data into multiple training and test sets, evaluate the various kernel possibilities, tune your selected kernel and evaluate your results. Remember that you don't need to get great results, you just need to understand the results and the process and to present your analysis.

Submission instructions: Present your answers to the questions above, analysis, commentary, R code and visualizations in an R Markdown file. Name your file svms.Rmd and, when you knit, create an .html file. Zip these files together (along with your independent data set if it's not too large) and upload your submission to Autolab as the SVM lab. Be sure to just select and zip the files, don't zip a directory.