Biostatistics 546, Spring 2017
Machine Learning for Biomedical Big Data
Homework 1

Instructions:
Answer the following questions, in full sentences. Solutions should be word-processed. For problems that include coding, pasting output from your `R` *session is not acceptable, unless otherwise indicated; your goal should be to perform statistical learning and the* `R` *output is relevant only if it is in incorporated as part of the analysis. Append your* `R` *code separately, with comments for your future reference. Please upload your solutions as a* `.pdf` *file.*

NOTE: While you can (and are encouraged to) work together, your solution to the homework, including the code and the writeup, should be *your own work*.

The following problems are all from the *Introduction to Statistical Learning*, by James et al (2012).

1. Problem 4, Chapter 3
2. Problem 9, Chapter 3
3. Problem 3, Chapter 5
4. Problem 8, Chapter 5
5. (*Optional*) Problem 7, Chapter 3
6. Download `dat4hw1.RData` from Canvas. This data set contains two objects, a 300×30 covariate matrix x and a response vector y.

   Set the random seed to 1 and randomly split the data into 200 training observations and 100 test observations. Build the best predictive model that you can (with the best possible transformations of the best subset of variables) and report the training and test error for your model. Describe your steps for choosing the model and report the final model.