Biostatistics 546, Spring 2017
Machine Learning for Biomedical Big Data
Homework 4

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 first 3 problems are from *Introduction to Statistical Learning, by James et al (2012).*

1. Chapter 8, Problem 2.
2. Chapter 8, Problem 4.
3. Chapter 8, Problem 5.
4. This problem uses the `diabetes` data in `lars` package. Specifically, use the **x2** design matrix, and the **y** outcome. First split the data into training and test sets of size 300 and 142 each, using 1234 as the random seed. Make sure to fit the models on a training set and to evaluate their performance on a test set.
   (a) Apply boosting, bagging, and random forests to predict the outcome. Compare the performance of these methods to linear regression with and without penalties.
   (b) Repeat the previous task 10 times, using random seeds 1 to 10. Summarize the ranking of the methods in the 10 different runs.