

Biostatistics 546, Spring 2016
Machine Learning for Biomedical Big Data
Homework 3

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 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 four problems below are from Introduction to Statistical Learning, by James et al (2012).

1. (*Optional*) Chapter 4, Problem 3.
2. Chapter 4, Problem 5
3. Chapter 4, Problem 6
4. Chapter 9, Problem 3
5. Chapter 9, Problem 5
6. Using the `Boston` data set from the `MASS` package, fit classification models in order to predict whether a given suburb has a crime rate above or below the median. Explore logistic regression, LDA, QDA and KNN models using various subsets of the predictors. Also try penalized logistic regression (ridge and lasso), as well as SVM using the optimal choices of tuning parameters for each method. Describe your findings.