Homework 5
Statistical Learning, Spring Term 2019
STAT760

The website for the Statistical Learning book has a repository of data:

https://web.stanford.edu/~hastie/ElemStatLearn/

Under the "Data" button we can find several datasets. The prostate cancer data set has eight predictors (columns 1--8)
lcavol
lweight
age
lbph
svi
lcp
gleason
pgg45
and one outcome (column 9), the lpsa measurement.

Problem 1

a) Train a multivariate regression predictor for lpsa. Use the best subset selection method and plot the error rate for the best classifier (for 1, 2,…,8 predictors).
b) Train the same regression using ridge regression and plot the values of the coefficients for different values of lambda (as in Fig. 6.4, left. Notice the log scale for lambda)
c) Train the same regression using the lasso (perform gradient descent and project the vector of coefficients back into the feasible region at every step, if necessary). Use expression 6.8 and plot the value of the coefficients for different s.
(15 points)

Program the classifier using Matlab, R, Python, or any other language of your choice. Do not use a library function for the ridge or lasso regression.

Problem 2 (3 points)

Solve problem 2(a),(b),(c) of the ISLR book. For n very large, what percentage of the data points does not belong to the bootstrap sample?

Problem 3 (2 points)
Solve Problem 4 of the ISLR book.

http://www-bcf.usc.edu/~gareth/ISL/ISLR%20Seventh%20Printing.pdf