Data Science 1 IE 594 HMW1

Due Tuesday, Nov 19, 5:00 pm

1. Load the Boston data set from sklearn library.

```
from sklearn.datasets import load_boston
boston = load_boston()
```

We will now try to predict per capita crime rate using the other variables in this data set. In other words, per capita crime rate is the response, and the other variables are the predictors.

a) For each predictor, fit a simple linear regression model to predict the response. Describe your results. In which of the models is there a statistically significant association between the predictor and the response? Create some plots to back up your assertions.

b) Fit a multiple regression model to predict the response using all of the predictors. Describe your results. For which predictors can we reject the null hypothesis H0 : βj = 0?

c) How do your results from (a) compare to your results from (b)?
Create a plot displaying the univariate regression coefficients from (a) on the x -axis, and the multiple regression coefficients from (b) on the y -axis. That is, each predictor is displayed as a single point in the plot. Its coefficient in a simple linear regression model is shown on the x -axis, and its coefficient estimate in the multiple linear regression model is shown on the y -axis.

d) Is there evidence of non-linear association between any of the predictors and the response? To answer this question, for each predictor X , fit a model of the form

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon$$

e) Fit Ridge regression and Lasso. Compare the results. Find the optimal regularization parameter for each using a cross-validation.

2. Using the Boston data set, fit classification models in order to predict whether a given suburb has a crime rate above or below the median. Explore logistic regression, Naïve Bayes Classifier, and KNN models using various subsets of the predictors. Describe your findings using a cross-validation.