Problem Set 2
CSCI 688

Problem 1 Let $X \in \mathbb{R}^{N \times P}$ be our design matrix and $Y = X\beta + \varepsilon$ where $\beta \in \mathbb{R}^P$ and ε has a multivariate normal distribution so that $\varepsilon \sim N(0, \sigma^2 I)$ where $\sigma^2 > 0$ and I is the $N \times N$ identity matrix. Equivalently $Y \sim N(X\beta, \sigma^2 I)$. If our estimate of β is $\hat{\beta} = (X^T X)^{-1} X^T Y$ show that

$$\hat{\beta} \sim N(\beta, \sigma^2 (X^T X)^{-1}).$$

Hint: If $Z \sim N(\mu, \Sigma)$ where $\mu \in \mathbb{R}^N$ and $\Sigma \in \mathbb{R}^{N \times N}$ then if $B \in \mathbb{R}^{N \times N}$ we have $BZ \sim N(B\mu, B\Sigma B^T)$.

Problem 2 This question should be answered using the `Carseats` data set in the `ISLR` package. You may use `help(Carseats)` to learn more about the data set.

(a) Fit a multiple regression model to predict Sales using Price, Urban, and US.

(b) Provide an interpretation of each coefficient in the model. Be careful, some of the variables in the model are qualitative!

(c) Fit a KNN regression. What is a reasonable value for K?

(d) Compare the performance between the KNN regression and the linear regression. Which would you suggest?

Problem 3 For each of parts (a) through (d), indicate whether we would generally expect the performance of a flexible statistical learning method to be better or worse than an inflexible method. Justify your answer.

(a) The sample size n is extremely large, and the number of predictors p is small.

(b) The number of predictors p is extremely large, and the number of observations n is small.

(c) The relationship between the predictors and response is highly non-linear.

(d) The variance of the error terms, i.e. $\sigma^2 = \text{Var}(\varepsilon)$, is extremely high.

Problem 4 I collect a set of data (n = 100 observations) containing a single predictor and a quantitative response. I then fit a linear regression model to the data, as well as a separate cubic regression, i.e. $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \epsilon$.

(a) Suppose that the true relationship between x and y is linear, i.e. $y = \beta_0 + \beta_1 x + e$. Consider the training residual sum squares (RSS) for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

(b) Answer (a) using test rather than training RSS.