STA 545 Statistical Data Mining I, Fall 2020

Homework 4, due: Wednesday 9/30/2020 (1PM)

Please submit the pdf file generated by R markdown in UBlearns. Please use tables, figures, or a few sentences to answer data analysis questions.

1. (70 points) In this exercise, we will predict the number of applications received using the other variables in the College data set. Please install the ISLR R package to download this data set.

   (a) (5 points) This data set has 777 observations. Please randomly split the data set into a training set (500 observations) and a test set (277 observations). Please use the set.seed() function in this step so that you can reproduce your following analysis results.

   (b) (15 points) Fit a linear model using least squares on the training set. Report the estimated regression coefficients and the test error obtained.

   (c) (15 points) If we fit the ridge regression model on the training set considering all possible values of the tuning parameter, which ridge regression model has the lowest training error? If we fit the PCR model on the training set considering all possible values of the tuning parameter M, which PCR model has the lowest training error? Are these two models always the same as the linear model in part (b)? Why?

   (d) (5 points) Further split the training set into two parts randomly: set A (250 observations) and set B (250 observations). Please use the set.seed() function in this step so that you can reproduce your following analysis results.

   (e) (15 points) Fit a ridge regression model on the set A, with the tuning parameter λ chosen by the set B. Report the estimated regression coefficients and the test error obtained.

(f) (15 points) Fit a PCR model on the set A, with the parameter M chosen by the set B. Report the value of M selected by the set B, the estimated regression coefficients of the original input variables, and the test error obtained.

2. (15 points) Suppose that $n = 2$, $p = 2$, $x_{11} = x_{12}$, $x_{21} = x_{22}$. Furthermore, suppose that $y_1 + y_2 = 0$, $x_{11} + x_{21} = 0$, $x_{12} + x_{22} = 0$. Please show that in this setting, the ridge coefficient estimates satisfy $\hat{\beta}_0 = 0$, $\hat{\beta}_1 = \hat{\beta}_2$.

3. (15 points) Using the singular value decomposition of the $n \times p$ data matrix $X = UDV^T$, please show that

$$\hat{y}^{ridge} = X\hat{\beta}^{ridge} = \sum_{j=1}^{p} u_j \frac{d_j^2}{d_j^2 + \lambda} u_j^T y,$$

where u_j is the j-th column of the matrix U and d_j is the j-th diagonal element of the diagonal matrix D.