STA 545 Statistical Data Mining I, Fall 2020

Homework 5, due: Wednesday 10/7/2020 (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. (25 points) Please use the datasets shown in Question 1 of your homework 4 to fit a PLS 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. In addition, please fit a lasso regression model on the set A, with the tuning parameter λ chosen by the set B. Report the test error obtained, along with the the estimated regression coefficients of the original input variables.

2. (25 points) The coordinate descent algorithm can be used to solve the following lasso optimization problem

$$\min_{\beta} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j|,$$

where $\beta = (\beta_1, \ldots, \beta_p)^T$, $\{(y_i, x_i) : i = 1, 2, \ldots, n\}$ are n training data points, and $\lambda \geq 0$ is a given parameter. As shown in the glmnet R package, for some data analysis problems, in order to deliver a good linear model, we need to use a mixture of the lasso penalty and the ridge penalty, and therefore need to solve the following optimization problem to estimate the regression coefficients

$$\min_{\beta} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda[\alpha \sum_{j=1}^{p} \beta_j^2 + (1-\alpha) \sum_{j=1}^{p} |\beta_j|],$$

where α is a given tuning parameter in the interval $(0, 1)$ and λ is a given positive parameter. Please show that the above optimization problem is equivalent to a lasso optimization problem, and therefore we can use the coordinate descent algorithm for the lasso method to solve it.

3. (50 points) Please write your own R function for the coordinate descent algorithm to fit lasso regression models. Please use the prostate cancer data to compare the results from your own R function with the results from the glmnet R function in the glmnet R package.