STATS 415 Homework 6

Due at 11:59pm, November 5, 2020

1. (15 points) Consider the following linear model with fixed design:

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta}^* + \epsilon_i, \quad i \in [n],$$

where $\{\mathbf{x}_i\}_{i\in[n]}$ are p-dimensional, $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ and $\{\epsilon_i\}_{i=1}^n$ are independent. Derive the AIC of a candidate model $\mathcal{S} \subset [p]$. We assume that σ^2 is known in advance.

2. (15 points) Suppose we fit a linear regression model with a ridge penalty, minimizing

$$\begin{aligned} & \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij})^2 \\ \text{subject to} \quad & \sum_{j=1}^p |\beta|_j^2 \le s \end{aligned}$$

for a particular value of s. Complete each sentence (a)-(e) below by choosing the best option among (1)-(6). Explain your answers. All options refer to the overall trends, which may be locally affected by noise. (3 points per question)

   (a) As we increase s from 0, the number of variables included in the model ...

   (b) As we increase s from 0, the training error ...

   (c) As we increase s from 0, the test error ...

   (d) As we increase s from 0, the variance of $\hat{\beta}$ ...

   (e) As we increase s from 0, the squared bias of $\hat{\beta}$ ...

   Answer options:

   (1) changes in ways that are impossible to predict.

   (2) increases initially, and then eventually starts decreasing.

(3) decreases initially, and then eventually starts increasing.

(4) steadily increases.

(5) steadily decreases.

(6) remains constant.

3. (70 points) In this exercise, we will predict the acceptance rate of a college (number of applications accepted / number of applications received) using the `College` dataset from the ISLR package. (10 points each question)

   (a) Split the data set into a training set and a test set. Fix the random seed to the value 234, choose 30% (rounded down to the nearest integer) of the data at random for testing, and use the rest for training. Define a new response variable `Accept/Apps` and remove `Accept` and `Apps` from the dataset in the following training and testing. Plot this variable against every variable in the dataset (make sure you use the appropriate type of plot for each predictor). Comment on which variables appear to be most predictive.

   (b) Fit a linear model using least squares on the training set, and report the training and test error obtained, with `Accept/Apps` as the response variable and all the other variables except `Accept` and `Apps` as predictors.

   (c) Use AIC, BIC, and adjusted R^2 to select a potentially smaller model instead, from the set of all possible predictors used in 3b. Report which model each method chose, and the training and test errors for their chosen model(s).

   (d) Use 5-fold cross-validation to estimate the test error from the training data, for the candidate smaller model(s) you found so far, and for the full model from 3b. Compare the training, CV, and test errors and comment on the results.

   (e) Fit a ridge regression model on the training set, with λ chosen by 10-fold cross-validation. Plot the coefficients' solution paths. Report the training, cross-validated, and test errors.

   (f) Fit a lasso model on the training set, with λ chosen by 10-fold cross-validation. Plot the coefficients' solution paths. Report

which variables are included in the model, and the training, cross-validated, and test errors.

(g) Comment on the results obtained, comparing the results of ridge and lasso with your best reduced model from (b) and (c). How accurately can we predict the acceptance rate overall? How much difference is there among the test errors resulting from different approaches? Which approach would you recommend for this dataset and why?

Please limit your answer to Q3 to 8 pages, organized into a coherent typed data analysis report. Answers to Q1 and Q2 may be either typed or handwritten. Please clearly write your name, your UMID, and your GSI/lab number on the homework, and submit it on Canvas.