STATS 415 Homework 7

Due at 2:30pm in class on Nov 21, 2019

1. Suppose we fit a linear regression model with a ridge penalty, minimizing

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

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. (8 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.

2. This exercise continues Q3 of Homework 6. Use the same training and test datasets. The goal is to predict the acceptance rate from the other variables in the `College` data set. (15 points per question)

(a) 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.

(b) 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.

(c) Fit a PCR model on the training set, with M chosen by cross-validation. Report the test error obtained, along with the value of M selected by cross-validation.

(d) Comment on the results obtained, comparing the results of ridge and lasso to your best reduced model from Homework 6. (That is, the reduced model from HW 6 that had the lowest test error). How accurately can we predict the acceptance rateoverall? 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 Q2 to 6 pages, organized into a coherent typed data analysis report. Answers to Q1 may be either typed or handwritten. Please staple everything together and clearly write your name, your UMID, and your GSI/lab number on the homework.