STAT 6340 (Statistical and Machine Learning, Spring 2021)

Mini Project 4

Instructions:

- Due date: March 31, 2021.
- Total points = 30
- Submit a typed report.
- It is OK to discuss the project with other students in the class, but each student must write their own code and answers. If your submitted report (including code and answer) is similar (either partially or fully) to someone else's, this will be considered evidence of academic dishonesty, and you will referred to appropriate university authorities.
- Do a good job.
- You must use the following template for your report:

  Mini Project #
  Name
  Section 1. Answers to the specific questions asked
  Section 2: R code. Your code must be annotated. No points may be given if a brief look at the code does not tell us what it is doing.

- Section 1 of the report must be limited to **three** pages. Also, only those output should be provided in this section that are referred to in the report.

---

1. Consider the wine dataset from Mini Project 2. As there, we will take `Quality` as the quantitative response, the remaining 6 variables as predictors, and all the data as training data. For all the models below, use leave-one-out cross-validation (LOOCV) to compute the estimated test MSE.

   (a) Fit a linear regression model using all predictors and compute its test MSE.

   (b) Use best-subset selection based on adjusted R^2 to find the best linear regression model. Compute the test MSE of the best model.

   (c) Use forward stepwise selection based on adjusted R^2 to find the best linear regression model. Compute the test MSE of the best model.

   (d) Use backward stepwise selection based on adjusted R^2 to find the best linear regression model. Compute the test MSE of the best model.

   (e) Use ridge regression with penalty parameter chosen optimally via LOOCV to fit a linear regression model. Compute the test MSE of the model.

   (f) Use lasso with penalty parameter chosen optimally via LOOCV to fit a linear regression model. Compute the test MSE of the model.

   (g) Make a tabular summary of the parameter estimates and test MSEs from (a) - (f). Compare the results. Which model(s) would you recommend?

2. Consider the diabetes dataset from Mini Project 3. As there, we will take `Outcome` as the binary response, the remaining 8 variables as predictors, and all the data as training data. For all the models below, use 10-fold cross-validation to compute the estimated test error rates. You may use the

bestglm package for parts (b)-(d) of this problem. See, e.g., http://www2.uaem.mx/r-mirror/web/packages/bestglm/vignettes/bestglm.pdf and https://rstudio-pubs-static.s3.amazonaws.com/2897_9220b21cfc0c43a396ff9abf122bb351.html. Note that regsubsets function from leaps package for variable selection will not work for these data as it only works with linear models.

(a) Fit a logistic regression model using all predictors and compute its test error rate.

(b) Use best-subset selection based on AIC to find the best logistic regression model. Compute the test error rate of the best model.

(c) Use forward stepwise selection based on AIC to find the best logistic regression model. Compute the test error rate of the best model.

(d) Use backward stepwise selection based on AIC to find the best logistic regression model. Compute the test error rate of the best model.

(e) Use ridge regression with penalty parameter chosen optimally via 10-fold cross-validation to fit a logistic regression model. Compute the test error rate of the model.

(f) Use lasso with penalty parameter chosen optimally via 10-fold cross-validation to fit a logistic regression model. Compute the test error rate of the model.

(g) Make a tabular summary of the parameter estimates and test error rates from (a) - (f). Compare the results. Which model(s) would you recommend? How does this recommendation compare with what you recommended in Mini Project 3?