STAT 6340 (Statistical and Machine Learning, Fall 2024)

Mini Project 4

Instructions:

- Due date: Oct 30, 2024 (Use Submission Link on eLearning to upload by 11:59 pm)
- Total points = 35
- Submit a typed report.
- It is OK to consult with others regarding this project or use any publicly available resources, but you must write code and answers on your own. Violation of this policy will be considered academic dishonesty, and you will referred to appropriate university authorities. If you don't do the work by yourself, how will you learn?
- 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: 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 software outputs should be provided in this section that are referred to in the report.
- You can code in `R` or `Python`.
- You are encouraged to use `R Markdown` or `Jupyter Notebook` to write your report.
- Bonus (4 points): Use *both* `R` and `Python` to code, compare the results, and provide an explanation if the results are not identical. This work should be described on a separate page titled Bonus Work with two sections — one to describe the comparison and the other for the code. The first section should be limited to one page.

---

1. (10 points) Consider the plasma volume data stored in `plasma_volume.csv` available on eLearning. The data consist of measurements of plasma volume using two methods: Hurley method (method 1) and Nadler method (method 2). Our interest is in performing inference on the correlation θ between the two methods.

   (a) (1 point) Perform an exploratory analysis of the data. Explain what you observe.

   (b) (1 point) Provide a point estimate $\hat{\theta}$ of θ. Interpret the result. Do you know of any method to compute bias and standard error of $\hat{\theta}$ and 95% confidence interval for θ? If so, use it to provide these estimates. MP3 3c, d, e

   (c) (4 points) Write your own code to perform (nonparametric) bootstrap with $B = 1000$ replications. Display the estimated sampling distribution of $\hat{\theta}$ and compute estimates of bias, standard error of $\hat{\theta}$, and also a 95% confidence interval for θ using the percentile method. Interpret the results.

(d) (2 points) Repeat the computation in (c) using a software package and make sure the results are correct. (Note that even if both results are correct, they may not match exactly due to Monte Carlo variability.)

(e) (2 points) What do you conclude about the correlation between the two methods?

2. (10 points) Consider the Pima Indians Diabetes dataset from Mini Project #3. As before, take all data as training data and standardize all predictors before any analysis. For all the models below, use LOOCV to compute the estimated test error rates.

(a) (2 points) Fit a logistic regression model using all predictors and compute its test error rate.

(b) (3 points) Use ridge regression with penalty parameter chosen optimally via LOOCV to fit a logistic regression model. Report the penalty parameter and compute the test error rate of the model.

(c) (3 points) Use lasso with penalty parameter chosen optimally via LOOCV to fit a logistic regression model. Report the penalty parameter and compute the test error rate of the model.

(d) (2 points) 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?

3. (15 points) Consider the forest fires dataset from Mini Project #2. Take $\log(\texttt{area} + 1)$ as the quantitative response variable. Among the predictors, treat `month` as a categorical variable and all other predictors as quantitative variables. Take all the data as training data. For all the models below, use leave-one-out cross-validation (LOOCV) to compute the estimated test MSE.

MP4 1 ALL

(a) (2 point) Fit a linear regression model using all predictors and compute its test MSE.

(b) (2 points) 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) (2 points) 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) (2 points) 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) (2 points) Use ridge regression with penalty parameter chosen optimally via LOOCV to fit a linear regression model. Report the penalty parameter and compute the test MSE of the model.

(f) (2 points) Use lasso with penalty parameter chosen optimally via LOOCV to fit a linear regression model. Report the penalty parameter and compute the test MSE of the model.

(g) (3 points) Make a tabular summary of the parameter estimates and test MSEs from (a) - (f). Compare the results. Which model(s) would you recommend?