STAT 6340 (Statistical and Machine Learning, Spring 2020)

Mini Project 4

Instructions:

- Due date: Mar 25, 2020.
- Total points = 30.
- Submit a typed report.
- It is OK to discuss the project with others and to search on the internet, 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 **five** pages. Also, only those output should be provided in this section that are referred to in the report.

---

1. Consider the oxygen saturation data stored in `oxygen_saturation.txt` file available on eLearning. The data consist of measurements of percent saturation of hemoglobin with oxygen in 72 adults, obtained using an oxygen saturation monitor (`OSM`, method 1) and a pulse oximetry screener (`POS`, method 2). You can read about oxygen saturation on Wikipedia, `https://en.wikipedia.org/wiki/Oxygen_saturation_(medicine)`. We are primarily interested in evaluating *agreement* between the two methods for measuring oxygen saturation.

   (a) Make a scatterplot of the data and superimpose the 45^o line. Comment on the extent of agreement between the two methods for measuring oxygen saturation. Note that the two methods would have *perfect agreement* if all the points in the scatterplot fell on the 45^o line.

   (b) Let μ_j and σ_j^2 respectively represent the population mean and variance of the jth method, $j = 1, 2$; and σ_{12} and ρ respectively represent the covariance and correlation between the two methods. Argue that perfect agreement corresponds to $\{\mu_1 = \mu_2, \sigma_1 = \sigma_2, \rho = 1\}$.

   (c) Let θ be the *concordance correlation coefficient* (CCC) between the two methods. It is a measure of agreement between the methods and is defined as

   $$\theta = \frac{2\sigma_{12}}{(\mu_1 - \mu_2)^2 + \sigma_1^2 + \sigma_2^2} = \rho \frac{2\sigma_1\sigma_2}{(\mu_1 - \mu_2)^2 + \sigma_1^2 + \sigma_2^2}.$$

   Argue that:

   i. $0 \leq |\theta| \leq |\rho| \leq 1$. What is the practical implication of this result?

   ii. $\theta = 1 \iff \{\mu_1 = \mu_2, \sigma_1 = \sigma_2, \rho = 1\}$, implying perfect agreement.

(d) Provide a point estimate $\hat{\theta}$ of θ.

(e) Write your own code to compute (nonparametric) bootstrap estimates of bias and standard error of $\hat{\theta}$, and a 95% *lower confidence bound* for θ computed using the percentile method. Interpret the results.

(f) Repeat the computation in (e) using `boot` package and compare your results.

(g) State your conclusion about the extent of agreement between the two methods. Would you say that the methods agree well enough to be used interchangeably in practice?

2. Consider the wine quality dataset from Mini Project 3. As in that project, we will take `good` as the binary response and all the 11 physiochemical characteristics of as predictors. Take all data as training data. Use AIC to compare models and 10-fold cross-validation to compute test error rates. Further, for any fitting method (see below) that has a complexity parameter, choose the best value for it using 10-fold cross-validation test error. 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. This is a repeat of what you did in the previous project.

   (b) Use best-subset selection to find the best logistic regression model. Compute its test error rate.

   (c) Repeat (b) using forward stepwise selection.

   (d) Repeat (b) using backward stepwise selection.

   (e) Repeat (b) using ridge regression.

   (f) Repeat (b) using lasso.

   (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 the previous project?