Statistical and Machine Learning (Spring 2019)

Mini Project 4

Instructions:

- Due date: March 27, 2019.
- Total points = 20.
- 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.

---

1. Consider the `gpa` data stored in the `gpa.txt` file available on eLearning. The data consist of GPA at the end of freshman year (`gpa`) and ACT test score (`act`) for randomly selected 120 students from a new freshman class.

   (a) Make a scatterplot of `gpa` against `act` and comment on the strength of linear relationship between the two variables.

   (b) Let ρ denote the population correlation between `gpa` and `act`. Provide a point estimate of ρ, bootstrap estimates of bias and standard error of the point estimate, and 95% confidence interval computed using percentile bootstrap. Interpret the results.

   (c) Fit a simple linear regression model for predicting `gpa` on the basis of `act`. Provide the least square estimates of the regression coefficients, standard errors of the estimates, and 95% confidence intervals of the coefficients. Perform model diagnostics to verify the model assumptions and comment on the results.

   (d) Use nonparametric bootstrap to compute the standard errors and 95% confidence intervals (using percentile bootstrap) mentioned in part (c) and compare the two sets of results.

2. Consider the `OJ` dataset which is a part of the `ISLR` package. It consists of `Purchase` as a binary response variable and a number of other variables as predictors. For the response, let "1" indicate `MM` and "0" indicate `CH`. You may try using `crossval` package for computing cross-validation errors, see `https://cran.r-project.org/web/packages/crossval/crossval.pdf`.

   (a) Examine the three store-related variables in the data — `StoreID`, `STORE`, and `Store7`. Should all the three variables be in the model? Regardless of your conclusion, use only `StoreID` among the three, and treat it as a categorical predictor for the remainder of this problem.

(b) Fit a logistic regression model. In addition, compute the confusion matrix, sensitivity, specificity, and overall misclassification rate based on the data, plot the ROC curve, and provide an estimate of the test error rate using 10-fold cross-validation.

(c) Repeat (b) using LDA.

(d) Repeat (b) using QDA.

(e) Repeat (b) using KNN with K chosen optimally using 10-fold cross-validation.

(f) Compare the results in (b)–(e). Which classifier would you recommend? Justify your conclusions.

3. Consider the `Auto` dataset from the `ISLR` package described in the book. Take `mpg` as response and the remaining variables (except `name`) as predictors. Take all data as training data. Use R^2 to compare models of the same size and adjusted R^2 to compare models of different sizes.

(a) Perform an exploratory analysis of the data.

(b) Fit a multiple linear regression model using the least squares method.

(c) Use best-subset selection to find the best model. Display the results graphically and interpret them.

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

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

(f) Compare the results from (b), (c), (d), and (e). Which model(s) would you recommend? Justify your conclusion.