STATS 503 - Homework 2

Due Wednesday, February 24, 2021
Online submission on Canvas required by 11:59pm EDT

Note this is a group homework, and the groups are defined by People/Groups/Homework 2 on Canvas.

1. Suppose that we wish to predict whether a given stock will issue a dividend this year ("Yes" or "No") based on X, last year's percent profit. We examine a large number of companies and discover that the mean value of X for companies that issued a dividend was $\overline{X} = 10$ while the mean for those that didn't was $\overline{X} = 0$. In addition, the variance of X for these two sets of companies was $\widehat{\sigma}^2 = 36$. Finally 80% of companies issued dividends. Predict the probability that a company will issue a dividend this year given that its percentage return was $X = 4$ last year. (Specify your assumptions)

2. Suppose we collect data for a group of students in a statistics class with variables X_1 = hours studied, X_2 = undergrad GPA, and $Y = c_1$ if receive an A and c_2 otherwise. We fit a logistic regression and produce estimated coefficients, $\hat{\beta}_0 = -6$, $\hat{\beta}_1 = 0.05$, $\hat{\beta}_2 = 1$.

   (a) Write out the resulting logistic regression equation.

   (b) How many hours would a student who has an undergrad GPA of 3.5 need to study to have a 50% chance of getting an A in the class?

3. In this problem, you will develop models to predict the wine type based on the `Wine` data set.

   (a) Explore the data graphically in order to investigate the association between `Type` and the other features. Which of the other features seem most likely to be useful in predicting `Type`? Scatterplots and boxplots may be useful tools to answer this question. Describe your findings.

(b) Perform LDA, QDA and Naive Bayes on the training data in order to predict `Type`. What are the test errors of the models obtained?

4. Use the k-nearest neighbor classifier on the `Theft` dataset. Use cross-validation to select the best k and use the test data to evaluate the performance of the selected model. Show the training, cross-validation and test errors for each choice of k and report your findings.

5. The textbook (*"An Introduction to Statistical Learning with Applications in R"*) describes that the cv.glm() function can be used in order to compute the LOOCV error estimate. Alternatively, one could compute those quantities using just the glm() and predict.glm() functions, and a for loop. You will now take this approach in order to compute the LOOCV error for a logistic regression model on the `Weekly` data set (in the ISLR package).

   (a) Fit a logistic regression model that predicts `Direction` using `Lag1` and `Lag2`. Report and comment on the result.

   (b) Fit a logistic regression model that predicts `Direction` using `Lag1` and `Lag2` using all but the first observation. Report and comment on the result.

   (c) Use the model from (b) to predict the direction of the first observation. You can do this by predicting that the first observation will go up if $Pr(\text{Direction=``Up''} \mid \text{Lag1}, \text{Lag2}) > 0.5$. Was this observation correctly classified?

   (d) Write a for loop from $i = 1$ to $i = n$, where n is the number of observations in the data set, that performs each of the following steps:

      i. Fit a logistic regression model using all but the ith observation to predict `Direction` using `Lag1` and `Lag2`.

      ii. Compute the posterior probability of the market moving up for the ith observation.

iii. Use the posterior probability for the *i*th observation in order to predict whether or not the market moves up.

iv. Determine whether or not an error was made in predicting the direction for the *i*th observation. If an error was made, then indicate this as a 1, and otherwise indicate it as a 0.

(e) Take the average of the *n* numbers obtained in (d)iv in order to obtain the LOOCV estimate for the test error. Comment on the results.

Limit your solutions to at most 10 pages (including code and figures).