STA 141C: Homework 2

Due 4/25/2025, 8PM

This homework assignment has to be submitted electronically on Gradescope by the due date. Please submit one PDF file containing all your answers and the code used for your data analysis. It is strongly recommended to type your answers rather than submitting handwritten work. If handwritten, please ensure that it is legible and neat.

Problem 1. (2/20 points) 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 $\bar{X} = 10$, while the mean for those that didn't was $\bar{X} = 0$. In addition, the variance of X for these two sets of companies was $\hat{\sigma}^2 = 36$. Finally, 80% of companies issued dividends. Assuming that X follows a normal distribution, predict the probability that a company will issue a dividend this year given that its percentage profit was $X = 4$ last year.

Problem 2. (5/20 points) Suppose that you wish to classify an observation $X \in \mathbb{R}$ into apples and oranges. You fit a logistic regression model and find that

$$\hat{\mathbb{P}}(Y = \text{orange}|X = x) = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x)}.$$

Your friend fits a logistic regression model to the same data using the softmax formulation in (4.13) in the textbook, and finds that

$$\hat{\mathbb{P}}(Y = \text{orange}|X = x) = \frac{\exp(\hat{\alpha}_{\text{orange}0} + \hat{\alpha}_{\text{orange}1} x)}{\exp(\hat{\alpha}_{\text{orange}0} + \hat{\alpha}_{\text{orange}1} x) + \exp(\hat{\alpha}_{\text{apple}0} + \hat{\alpha}_{\text{apple}1} x)}.$$

(a) What is the log odds of orange versus apple in your model?

(b) What is the log odds of orange versus apple in your friend's model?

(c) Suppose that in your model, $\hat{\beta}_0 = 2$ and $\hat{\beta}_1 = -1$. What are the coefficient estimates in your friend's model? Be as specific as possible.

(d) Now suppose that you and your friend fit the same two models on a different data set. This time, your friend gets the coefficient estimates $\hat{\alpha}_{\text{orange}0} = 1.2, \hat{\alpha}_{\text{orange}1} = -2, \hat{\alpha}_{\text{apple}0} = 3, \hat{\alpha}_{\text{apple}1} = 0.6$. What are the coefficient estimates in your model?

(e) Finally, suppose you apply both models from (d) to a data set with $2,000$ test observations. What fraction of the time do you expect the predicted class labels from your model to agree with those from your friend's model? Explain your answer.

Problem 3. (2/20 points)

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 = receive an A. We fit a logistic regression and produce estimated coefficients, $\hat{\beta}_0 = -6$, $\hat{\beta}_1 = 0.05$, and $\hat{\beta}_2 = 1$.

(a) Estimate the probability that a student who studies for 40 hours and has an undergrad GPA of 3.7 gets an A in the class.

(b) How many hours would the student in part (a) need to study to have a 80% chance of getting an A in the class?

Problem 4. (7/20 points)

Choose one version to complete: either the R version or the Python version.

R version. This question should be answered using the `Weekly` data set, which is part of the `ISLR2` package. This data is similar in nature to the `Smarket` data from class, except that it contains 1,089 weekly returns for 21 years, from the beginning of 1990 to the end of 2010.

(a) Produce some numerical and graphical summaries of the `Weekly` data. Do there appear to be any patterns?

(b) Use the full data set to perform a logistic regression with `Direction` as the response and the five lag variables plus `Volume` as predictors. Use the summary function to print the results. Do any of the predictors appear to be statistically significant? If so, which ones?

(c) Compute the confusion matrix and overall fraction of correct predictions. Explain what the confusion matrix is telling you about the types of mistakes made by logistic regression.

(d) Now fit the logistic regression model using a training data period from 1990 to 2008, with `Lag2` as the only predictor. Compute the confusion matrix and the overall fraction of correct predictions for the held out data (that is, the data from 2009 and 2010).

(e) Repeat (d) using LDA.

(f) Repeat (d) using QDA.

(g) Which of these methods appears to provide the best results on this data?

Python version. This question should be answered using the `Weekly` data set, which is part of the `ISLP` package. This data is similar in nature to the `Smarket` data from class, except that it contains 1,089 weekly returns for 21 years, from the beginning of 1990 to the end of 2010.

(a) Produce some numerical and graphical summaries of the `Weekly` data. Do there appear to be any patterns?

(b) Use the full data set to perform a logistic regression with `Direction` as the response and the five lag variables plus `Volume` as predictors. Use the summary function to print the results. Do any of the predictors appear to be statistically significant? If so, which ones?

(c) Compute the confusion matrix and overall fraction of correct predictions. Explain what the confusion matrix is telling you about the types of mistakes made by logistic regression.

(d) Now fit the logistic regression model using a training data period from 1990 to 2008, with `Lag2` as the only predictor. Compute the confusion matrix and the overall fraction of correct predictions for the held out data (that is, the data from 2009 and 2010).

(e) Repeat (d) using LDA.

(f) Repeat (d) using QDA.

(g) Which of these methods appears to provide the best results on this data?

Problem 5. (2/20 points)
Explain how k-fold cross-validation is implemented in practice, and what are the advantages and disadvantages of k-fold cross-validation relative to:

1. The validation set approach?

2. LOOCV?

Problem 6. (2/20 points) Suppose that we use some statistical learning method to make a prediction for the response Y for a particular value of the predictor X. Carefully describe how we might estimate the standard deviation of our prediction. Be as specific as you can.