STAT 847: Assignment 2 - copyright L. Xing, University of Saskatchewan

Date:

(Due date: 11:59pm (Time zone: GMT-6) on Oct 11, 2023. Late assignments will be accepted only for seven (7) days beyond the due date. The penalty for the delay is 10 percent per day of lateness from the value of the assignment, including weekend days.)

TOTAL MARKS

NSID ....

Email ....

Question 1: (4 marks) Suppose that you wish to classify an observation $X \in \mathbb{R}$ into `apples` and `orange`. You fit a logistic regression model and find that

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

Your friend fits a logistic regression model to the same data using the `softmax` formulation below

$$\Pr(Y = k|X = x) = \frac{e^{\beta_{k0}+\beta_{k1}x_1+\ldots+\beta_{kp}x_p}}{\sum_{l=1}^{K} e^{\beta_{l0}+\beta_{l1}x_1+\ldots+\beta_{lp}x_p}},$$

where $k = 1, ..., K$, and finds that

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

(a) (0.5 mark) What is the log odds of `orange` versus `apples` in your model?

(b) (0.5 mark) What is the log odds of `orange` versus `apples` in your friend's model?

(c) (1 mark) 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) (1 mark) 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 $\hat{\alpha}_{\text{orange}0} = 1.2$, $\hat{\alpha}_{\text{orange}1} = -2$, $\hat{\alpha}_{\text{apple}0} = 3$, $\hat{\alpha}_{\text{apple}0} = 0.6$. What are the coefficient estimates in your model?

(e) (1 mark) 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.

Question 2: (10 marks) This question should be answered using the `Weekly` data set, which is part of the I$SLR$2 package. This data is similar in nature to the `Smarket` data from this chapter's lab, except that it contains 1, 089 weekly returns for 21 years, from the beginning of 1990 to the end of 2010.

(a) (2 marks) Produce some numerical and graphical summaries of the `Weekly` data. Do there appear to be any patterns?

(b) (2 marks) 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) (2 mark) 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) (2 marks) 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, the overall fraction of correct predictions for the held-out data (data from 2009 and 2010), and also calculate the precision, recall, F1 score, AUC, and draw the ROC curve.

(e) (2 mark) Experiment with different combinations of predictors, including possible transformations and interactions, for each of the methods. Report the variables, method, and associated confusion matrix that appears to provide the best results on the held out data. Make Comparison between these models base on ROC curve.

Question 3: (7 marks) In this problem, you will develop a model to predict whether a given car gets high or low gas mileage based on the `Auto` data set.

(a) (1 mark) Create a binary variable, `mpg01`, that contains a 1 if `mpg` contains a value above its median, and a 0 if `mpg` contains a value below its median. You can compute the median using the `median()` function. Note you may find it helpful to use the `data.frame()` function to create a single data set containing both `mpg01` and the other `Auto` variables.

(b) (3 mark) Explore the data graphically in order to investigate the association between `mpg01` and the other features. Which of the other features seem most likely to be useful in predicting `mpg01`? Scatterplots and boxplots may be useful tools to answer this question. Describe your findings.

(c) (1 mark) Split the data into a training set and a test set.

(d) (2 mark) Perform logistic regression on the training data in order to predict `mpg01` using the variables that seemed most associated with `mpg01` in **(b)**. What is the test error of the model obtained? Please use different evaluation metrics.

Question 4: (4 marks) Using the `Boston` data set, fit classification models in order to predict whether a given census tract has a crime rate above or below the median. Explore logistic regression model using various subsets of the predictors. Describe your findings.

Hint: You will have to create the response variable yourself, using the variables that are contained in the `Boston` data set.