MATH 189 Homework 6

Due Feb 24th, 2023

Q1. 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:

$$\hat{\beta}_0 = -6,\ \hat{\beta}_1 = 0.05,\ \hat{\beta}_2 = 1.$$

(a) Estimate the probability that a student who studies for 40 h and has an undergrad GPA of 3.5 gets an A in the class.

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

Q2. Consider the Weekly data set, which is part of the ISLR package. This data set consists of 1089 weekly percentage returns for the S&P 500 stock index over 21 years, from the beginning of 1990 to the end of 2010. It contains the following 9 variables.

Year: The year that the observation was recorded.

Lag1: Percentage return for previous week.

Lag2: Percentage return for 2 weeks previous.

Lag3: Percentage return for 3 weeks previous.

Lag4: Percentage return for 4 weeks previous.

Lag5: Percentage return for 5 weeks previous.

Volume: Volume of shares traded (average number of daily shares traded in billions).

Today: Percentage return for this week.

Direction: A factor with levels Down and Up indicating whether the market had a positive or negative return on a given week.

(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 covariates/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).