CMSC 478 — Spring 2017 — C. S. Marron
Lab 4: Logistic Regression, LDA, & QDA

Data Description

In this lab, you will work with the `Weekly` dataset of weekly percentage returns for the S&P 500 stock index for the years 1990 to 2010. The dataset is part of the `ISLR` package; if you have not already installed the package, you may download the CSV file of the `Weekly` dataset. The variables in the dataset are described in the appendix.

Exercises

> Whenever you are asked to compute confusion tables or classification rates for models, you need to record that information. I will not accept answers like "We tried a few things and such-and-such ended up the best." You need to be able to show me your results.

Exercise 1: Produce some numerical and graphical summaries of the dtaset. Do there appear to be any patterns?

Exercise 2: Use logistic regression to identify which, if any, of variables `Lag1`, `Lag2`, `Lag3`, `Lag4`, `Lag5`, and `Volume` may be useful for predicting `Direction`. Compute the following for the logistic regression model:

- Confusion matrix
- Overall percentage of misclassified observations
- Percentage of "Up" observations that are misclassified
- Percentage of "Down" observations that are misclassified

What can you conclude about the type of mistakes made by logistic regression?

> For the remainder of the exercises, use the data for 1990 – 2008 as a training dataset and the data for 2009 and 2010 as a test dataset. All confusion matrices and classification rates are to be computed using the test dataset.

Exercise 3: Fit three different models — logistic regression, LDA, and QDA — on the training data with `Lag2` as the only predictor. For each model, compute the confusion matrix and overall correct classification rate when applied to the test data. Which of the methods provides the best results on this data?

Exercise 4: Experiment with different combinations of predictors, including possible transformation and interactions, for each of the methods. Can you find a model that gives better results than any from the previous exercise?

Appendix: Weekly Data Set

The `Weekly` dataset consists of 1089 observations on the following nine variables.

Year

Year that the observation was recorded.

Lag1

Percentage return for previous week.

Lag2

Percentage return for two weeks previous.

Lag3

Percentage return for three weeks previous.

Lag4

Percentage return for four weeks previous.

Lag5

Percentage return for five 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 negative or positive return for the week.

[Top]