INFO 7390–Machine Learning and Data Sciences

Northeastern University, Fall 2017

Problem Set 2, Due: November 28, 2017

Problem Set Rules:

1. Each student should hand in an individual problem set at the beginning of class.
2. Discussing problem sets with other students is permitted. Copying from another person or solution set is *not* permitted.
3. Late assignments will *not* be accepted. No exceptions.
4. How to get data sets: See Package ISLR

1. Remember that

$$p(y \mid \beta; x) = \frac{1}{1 + e^{-y\beta^T X}}$$

if we use 1 for a positive example and -1 for a negative example. Assuming that the n training examples were generated independently, we can then write down the likelihood of the parameters as

$$L(\beta) = \prod_{i=1}^{n} \frac{1}{1 + e^{-y_i\beta^T X_i}}$$

Use Gradient Descent to derive an update rule to determine the parameters $\beta_0, ...\beta_p$ that maximize the log likelihood.

2. This question should be answered using the Weekly S&P Stock Market Data.

   (a) Produce some numerical and graphical summaries of 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) Use *Forward Stepwise Selection* approach for feature selection. In each step, choose the best model using the highest AUC. Report the features, confusion matrix, and AUC that appears to provide the best results on the held out data.

3. Using the Boston data set, fit classification models in order to predict whether a given suburb has a crime rate above or below the median. Explore logistic regression using various subsets of the predictors. Describe your findings. In particular, report the features, confusion matrix, AUC that appears to provide the best results on a held out data.

4. We want to use logistic regression to predict the probability of default using income and balance on the Default data set. Our goal is to estimate the test error of this logistic regression model using the validation set approach. Do not forget to set a random seed before beginning your analysis.

   (a) Fit a logistic regression model that uses income and balance to predict default.

   (b) Using the validation set approach, estimate the test error of this model. In order to do this, you must perform the following steps:

      i. Split the sample set into a training set and a validation set.

      ii. Fit a multiple logistic regression model using only the training observations.

      iii. Obtain a prediction of default status for each individual in the validation set by computing the posterior probability of default for that individual, and classifying the individual to the default category if the posterior probability is greater than 0.5.

      iv. Compute the validation set error, which is the fraction of the observations in the validation set that are misclassified.

   (c) Repeat the process in (b) three times, using three different splits of the observations into a training set and a validation set. Describe your findings and comment on the results obtained.

   (d) Now consider a logistic regression model that predicts the probability of default using income, balance, and a dummy variable for student. Estimate the test error for this model using the validation set approach. Comment on whether or not including a dummy variable for student leads to a reduction in the test error rate.