PREDICTIVE ANALYTICS

Problem Set 4

1 Problem to demonstrate the fitting of a binary response variable using logistic regression

a. Consider the Weekly data available in the ISLR package of R. Use the data from 1990 to 2008 as train data and the data on 2009-2010 as test data.

b. Using the training data, fit a logistic regression with Direction as the response and the five lag variables plus Volume as predictors. Are there any significant predictors? If not, which of the tests is reporting the least p value. Interpret the coefficient corresponding to that test.

c. Estimate the odds of market going "up" at median values of the predictors.

d. Based on the training sample draw the ROC curve and comment.

e. Using the fitted model and the predictor values from the test data, find an optimal cut point that minimizes $TPR(1 - FPR)$. Report the confusion matrix corresponding to the optimal cut-point value. From the confusion matrix compute the test error rate.

f. Is there any change in the test error rate if we have worked with only the two predictors lag 1 and lag 2?

2 Problem to demonstrate the fitting of multi-category logistic regression model

Consider the data file named "literacy" in excel format. A word file containing the data description is also provided. Clean the data by removing all rows with at least one NA entry. Setting seed as 123, select 80% of data as training sample and keep the remaining as test samples.

a. Using the column m12 as your categorical response and m2, m3, m5, m6, c1-c5, c7, c8, c9 as the predictors, fit a multi- category logit model to your data. Interpret the output table.

b. Find the odds of being literate rather than illiterate for an OBC, Hindu female agricultural labourer of age 40 with no formal education, who resides in a joint family. The individual does not possess own land but has electricity connection at home and a black and white TV with cable connection.

3 Problem to demonstrate the use of KNN classification

a. Refer to Problem 1. Fit a KNN classification model for $K = 2, 5, 9$. For each case, train the model using the training sample with all the predictors and test it on the basis of test data.

b. Compute the confusion matrix for each case and find the test error rate.

c. How does KNN compare to the logistic regression.

d. Compute the test error rates for each K, if instead of choosing all predictors we choose only lag 1 and lag 2.

4 Problem to demonstrate the use of Naive Bayes

a. Refer to the Weekly data as in Problem 1. Fit a Naive Bayes estimator to the training sample assuming that all the predictors follow Gaussian distribution.

b. Using the test data and an optimal cutpoint, find the test error rate.

c. Assuming a non parametric modelling of the predictors, fit the Naive Bayes classifier. Find the test error rate in this case.

d. Compare the test error rates of the following cases : (use the all predictor case)

1. Logistic regression
2. KNN with $K = 2$
3. KNN with $K = 5$
4. KNN with $K = 9$
5. Naive Bayes estimator with gaussian assumption
6. Naive Bayes estimator under non-parametric assumption.