Assignment #3: Logistic Regression and KNN (4 pts)

Group Submission

Due: October 17 2:00 pm.

In this problem, we predict *Direction* using the data Weekly.csv.

a. i. Split the original data into one training set and one testing set. The training set contains observations in the period from 1990 to 2008 (Hint: we can use a Boolean vector train=(Year < 2009)). The testing set contains observations in 2009 and 2010 (Hint: since train is a Boolean vector here, should use ! symbol to reverse the elements of a Boolean vector to obtain the testing set, e.g. [!train]). Using the training data, develop a logistic regression model for Direction with the five lag variables as predictors. Which predictors appear to be statistically significant? What are their own odds ratio respectively?

ii. Use the model got in (i) to make prediction using the testing data. Use 0.5 as the threshold for prediction. Show the confusion matrix and compute the prediction accuracy.

iii. Perform 10-NN using the five lag variables as predictors. Show the confusion matrix. Does the KNN model provide a better result?

b. Perform 5-fold cross-validation on the original data using logistic regression and 10-NN respectively with the five lag variables as predictors. Which model is better?

Deliverables

1. Group submission. Each group submits one set of report and code. Please include a cover page on the report listing all team members' names.
2. Two files: R code and the report are submitted as two separate files to Blackboard. Screenshot of R code is not accepted.
3. The report should contain the answer to each question. No R raw outputs or software screenshot should be included in the report except plots.