IE 425

Homework 1 (due April 5, Wednesday @23:59)

You have to submit a report which includes the answers to the questions below and the R code generated.

1. Consider the dataset called spam in the kernlab package. You can load this data set by writing data("spam", package = "kernlab")

A data set collected at Hewlett-Packard Labs, that classifies 4601 e-mails as spam or non-spam. There are 57 variables indicating the frequency of certain words and characters in the e-mail. The first 48 variables contain the frequency of the variable name (e.g., business) in the e-mail. If the variable name starts with "num" (e.g., num650), then it indicates the frequency of the corresponding number (e.g., 650). The variables 49-54 indicate the frequency of the characters ';', '(', '[', '!', '\$', and '\#'. The variables 55-57 contain the average, longest and total run-length of capital letters. Variable 58 indicates the type of the mail and is either "nonspam" or "spam".

Our goal is to predict the type of messages as either spam or nonspam.

a) Using a seed value of 425, partition the dataset into training and test sets where 80% of goes into the training set and 20% goes into the test set. Make sure that the proportion of classes remains the same in both sets.

b) Using the rpart package and training set, determine the largest possible tree. How many leaf nodes do exist in the tree?

c) Make predictions in the test set and report the accuracy, error rate, false positive rate, false negative rate, and precision.

d) What is the size of the tree in terms of the number of leaf nodes which makes the cross-validation (CV) error the smallest? Note that rpart function provides this automatically. What is the smallest the tree which has a CV error smaller than the smallest CV error plus one standard deviation of the error? Call this last tree "opttree".

e) Make predictions on the test set with opttree and report the accuracy, error rate, false positive rate, false negative rate, and precision. Compare the result with part c)

2. Consider the "ToyotaCorolla" dataset. Our goal is to predict the Price attribute.

a. Partition the data set into training and test sets with 75% going into the training set by using a seed value of 582.

b. Using the rpart package and training set, determine the tree which gives the smallest cross-validation error? How many leaf nodes do exist in this tree? Which attributes are the most important?

c. Make predictions in the test set and report the RMSE, MAE, and MAPE.

d. Using the randomForest package and training set, generate models by playing with “mtry”, nodesize”, and “ntree” parameters. What parameter combination gives the smallest RMSE in the test set?

e. Comment on which input attributes are most important in making predictions.

f. Compare RMSE, MAE, and MAPE in the test set obtained by rpart and randomForest models.