IE 425

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

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 1000, partition the dataset into training and test sets where 70% of goes into the training set and 30% goes into the test set. What is the percentage of spam messages in the overall, training, and test 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 error rate, false positive rate and false negative rate.

d) What is the size of the tree which makes the cross-validation (CV) error (measured by the deviance in the tree package) 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? (Note that in the tree package there is no standard deviation of the CV error) How many leaf nodes does it have? Call this last tree "opttree".optOur o

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

f) Repeat (b)-(e) using the tree package.