STAT 515 Applied Statistics and Visualization for Analytics (Fall 2016)

Assignment 4

Due Date: 11/17/2016 Thursday 11:59 PM

1. Web Robot Detection Study (50pts)

Web Robot Detection

Web usage mining is the task of applying statistical techniques to extract useful patterns from web access logs. These patterns can reveal interesting characteristics of site visitors; e.g., people who repeatedly visit a website and view the same product description page are more likely to buy the product if certain incentives such as rebates or free shipping are offered.

In web usage mining, it is important to distinguish accesses made by human users from those due to web robots. A web robot (also known as web crawler) is a software program that automatically locates and retrieves information from the internet by following the hyperlinks embedded in web pages. These programs are deployed by search engine portals to gather the documents necessary for indexing the web. Web robot accesses must be discarded before applying web mining techniques to analyze human browsing behavior. The objective is to detect the web robot access.

The file **Web Robot** is a sample of the data obtained from a web server log. Each line corresponds to a web session. A web session is a sequence of requests made by a client during a single visit to a website. To classify the web sessions, features are constructed to describe the characteristics of each session. The followings are the features used in web robot detection task:

TotalPages	Total number of pages retrievd in a web session
ImagePages	Percentage of imge pages retrievd in a web session (out of total pages)
TotalTime	Total amount of time spent by website visitor
RepeatedAccess	The same page requested more than once in a web session
ErrorRequest	Errors in requestin for webpages
GET	Percentage of requestes made using GET method
POST	Percentage of requestes made using POST method
HEAD	Percentage of requestes made using HEAD method
Breadth	Breadth of web traversal
Depth	Depth of web traversal
MultiIP	Session with multiple IP addresses
MultiAgent	Session with multiple user agents
Web robot	Whether the page was accessed by a web robot (1) or a human (0)

(Depth determines the maximum distance of a requested page, where distance is measured in terms of number of hyperlinks away from the entry point of the website. The breadth attribute measures the width of the corresponding web graph).

a. What is the overall proportion of the Web Robot in the original dataset?

b. Partition the data into training (60%) and validation (40%) (set the random seed to 12345). Using all the predictors, build a logistic regression model with backward selection method. Write down the logistic regression equation (with only the significant predictors). By setting the cutoff level at 0.5, report the classification/confusion table (frequency counts), the overall accuracy rate and misclassification error on the validation dataset.

c. Using the same training and validation sets, build a classification tree with the training data, report the tree graph. Calculate and report the Gini index for the split of the top node. Show your work on how to calculate the Gini index for left and right child (branch) and the overall Gini for the split. Apply the model on the validation set, report the classification table, the overall accuracy and misclassification error rate on the validation data.

d. Suppose we now have new observation with some missing values. The only information available to us is as follows: Depth=1, Breadth=1, ImagePages=0.75. Following the structure of classification tree you obtained in part (c), is it possible to make a predictor whether this web access is from Robot or not? If yes, please classify the new observation as Robot_Yes or Robot_No. Provide a brief explanation.

e. With the same training and validation sets, use the random forest method to build a classification tree with the training data, generate the variable importance plot and report the top three most important variables. Report the classification table (frequency counts), the overall accuracy rate, misclassification error rate on the **validation** dataset.

f. Compare the models you obtained in parts (b), (c) and (e), which model is the best with respect to the misclassification error rate performance on the validation dataset.

Submission:

1. Prepare a pdf (or MS word) file with answers to above questions, brief explanations for the analysis you perform. Please provide your R codes at the end of the submitted file as appendix.
2. Name the file as "LastName, FirstName-HW4.pdf" (for example, "Ji,Ran-HW4.pdf") and submit it on blackboard.