Exercise 4 - Introduction to Data Science

Exercise 4.1

(Decision Trees) In this assignment you will train a classification tree, using the survey data *survey.csv***.**

a. Import the survey data *survey.csv* into R, storing rows 1 through 600 as training data and rows 601 through 750 as testing data.

b. Build a classification tree from the training data using the "rpart" package, according to the formula "MYDEPV ~ Price + Income + Age". Use the information gain splitting index. Which features were actually used to construct the tree? (see the "printcp" function) Plot the tree using the "rpart.plot" package.

c. Score the model with the training data and create the model's confusion matrix. Which class of MYDEPV was the model better able to classify?

d. Define the resubstitution error rate, and then calculate it using the confusion matrix from the previous step. Is it a good indicator of predictive performance? Why or why not?

e. Using the "ROCR" package, plot the receiver operating characteristic (ROC) curve. Calculate the area under the ROC curve (AUC). Describe the usefulness of this statistic.

f. Score the model with the testing data. How accurate are the tree's predictions?

g. Repeat part (b), but set the splitting index to the *Gini coefficient splitting index*. How does the new tree compare to the previous one?

h. Pruning is a technique that reduces the size/depth of a decision tree by removing sections with low classification power, which helps reduce overfitting and simplifies the model, reducing the computational cost. One way to prune a tree is according to the complexity parameter associated with the smallest cross-validation error. Prune the new tree in this way using the "prune" function. Which features were actually used in the pruned tree? Why were certain variables not used?

i. Create the confusion matrix for the new model, and compare the performance of the model before and after pruning.

Exercise 4.2

(Naïve Bayes) In this assignment you will train a Naïve Bayes classifier on categorical data and predict individuals' incomes.

Part 1

a. Import the *nbtrain.csv* file. Use the first 9010 records as training data and the remaining 1000 records as testing data.

b. Construct the Naïve Bayes classifier from the training data, according to the formula "income ~ age + sex + educ". To do this, use the "naiveBayes" function from the "e1071" package. Provide the model's a priori and conditional probabilities.

c. Score the model with the testing data and create the model's confusion matrix. Also, calculate the overall, 10-50K, 50-80K, and GT 80K misclassification rates. Explain the variation in the model's predictive power across income classes.

Part 2

a. Construct the classifier according to the formula "sex ~ age + educ + income", and calculate the overall, female, and male misclassification rates. Explain the misclassification rates?

b. Divide the training data into two partitions, according to sex, and randomly select 3500 records from each partition. Reconstruct the model from part (a) from these 7000 records. Provide the model's a priori and conditional probabilities.

c. How well does the model classify the testing data? Explain why.

d. Repeat step (b) 4 several times. What effect does the random selection of records have on the model's performance?

e. What conclusions can one draw from this exercise?