Decision Tree Example: Statlog Heart

We will use statlog heart dataset again for decision tree model practice. Statlog (Heart) Data Set is downloaded from UCI machine learning repository. It has 13 different attributes and 1 class variable. The detailed feature attributes are presented in the below. You could also find it here: https://archive.ics.uci.edu/ml/datasets/Statlog+%28Heart%29

The attributes include: 1. Age; 2. Sex (female =0, male =1); 3. chestpain = chest pain type (4 values); 4. restbloodpr = resting blood pressure; 5. serumchol = serum cholestoral in mg/dl; 6. blsugar.120 = fasting blood sugar > 120 mg/dl (have =1, don’t have =0); 7. restelectro = resting electrocardiographic results (values 0,1,2); 8. maxheartrate = maximum heart rate achieved; 9. angina = exercise induced angina (don’t have =0, have =1); 10. oldpeak = ST depression induced by exercise relative to rest; 11. st.segment = the slope of the peak exercise ST segment; 12. majorvessels = number of major vessels (0-3) colored by flourosopy; 13. thal: 3 = normal; 6 = fixed defect; 7 = reversable defect;

The class variable is statlogheart: 1=Absence of heart disease; 2= presence of heart disease;

Among all attributes, we have real attributes: 1,4,5,8,10,12; ordered attributes: 11; binary attributes: 2,6,9 and nominal attributes: 3, 7, 13;

1. Import dataset to R and check the data structure of the dataset.

**data <- read.csv(file.choose(), stringsAsFactors = F)**

**str(data)**

1. Change attributes 2,3,6,7,9,13, and 14 to factor variables. Hint:

dt[c(2,3,6,7,9,13,14)]<-data.frame(lapply(dt[c(2,3,6,7,9,13,14)],factor))

**data[c(2,3,6,7,9,13,14)]<-data.frame(lapply(data[c(2,3,6,7,9,13,14)],factor))**

1. Add labels (No disease, Heart disease) to statlogheart

**data$statlogheart <- factor(data$statlogheart);levels(data$statlogheart)[2]<-"Heart Disease"**

**data$statlogheart <- factor(data$statlogheart);levels(data$statlogheart)[1]<-"No Disease"**

1. Create a training dataset with 99% of records, and use it for the rest of the analysis. (Discussion: Typically, you will not want to keep such a high percentage of records in training set. I made this exception because results were not stable at lower percentage split.)

**library(C50)**

**set.seed(1234)**

**totalNoOfRows <- nrow(data)**

**train.size <- floor(totalNoOfRows \* 0.99)**

**train.indices <- sample(1:totalNoOfRows, train.size, replace = F)**

**str(train.indices)**

**data.train <- data[train.indices,]**

**str(data.train)**

1. Build a decision tree model with statlogheart as outcome variable and the rest variables as predictors. Then display simple facts about the tree. What tree size does this decision tree model have?

**m <- C5.0(data.train[-14], data.train$statlogheart, trials = 1, costs = NULL)**

**m**

**Tree size: 13**

1. Summarize this decision tree model and display the detailed information of the tree model. How many patients have been misclassified? How many attributes have been utilized to generate the tree.

**summary(m)**

**25 errors or misclassifications**

**9 attributes utilized**

1. If a patient is female, has one major vessel colored by flourosopy (majorvessels) and has thal = 6, does this patient have heart disease base on the decision tree?

***Yes,* she does have heart disease.**

1. If a patient has no major vessel colored, angina is one, oldpeak is less than 1.5, has thal = 7, serumchol is less than 255, does this patient have heart disease base on the decision tree?

***No,* this patient doesn’t have heart disease.**

1. Boost the accuracy of decision trees by boosted decision tree with 10 trials. What is the average tree size for 10 iterations? How many patients in the boosted model are misclassified?

**Average tree size is 10.3**

**In the boosted model 7 patients are misclassified**

1. Create a cost sensitive decision tree on training data. The sequence of the costs should be c (0, 1, 10, 0) in the R syntax.

**error\_cost <- matrix(c(0,10,10,0), nrow = 2)**

**m\_cost <- C5.0(data.train[-14], data.train$statlogheart, costs = error\_cost)**

1. Use the above cost sensitive decision tree to create a factor vector of predictions for in-sample data. How many cases of false negatives are there? Is there any difference in CrossTable results and contingency matrix in summary command?

**m\_cost\_pred <- predict(m\_cost, data.train)**

**summary(m\_cost\_pred)**

**library(gmodels)**

**CrossTable(data.train$statlogheart, m\_cost\_pred, prop.chisq = FALSE, prop.c = FALSE, prop.r = FALSE, dnn = c('acutal statlogheart', 'predicted statlogheart'))**

**141 false negatives on the cross table**

**The contingency matrix in summary command shows**

**153 No Disease, 114 Heart Disease**

**and the column total in the CrossTable shows**

**153 No Disease, 114 Heart Disease**

**Therefore, there is no difference between the results.**