Alex Gibbons

May 2, 2024

CMPT 363 – Data Mining

Dr. Ankur Agrawal

Assignment #10

# Alex Gibbons

# Assignment 10

url <-

"https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data"

adult <- read.table(url,sep=",",header=FALSE,na.strings="?")

names(adult) <- c("age","workClass","fnlwgt","education","educationNum","maritalStatus","occupation","relationship","race","sex","capitalGain","capitalLoss","hoursPerWeek","nativeCountry","earnings")

df <- adult

set.seed(1234)

# generate row numbers for training dataset

train <- sample(nrow(df),0.7\*nrow(df))

# training sample with 70% of tuples

df.train <- df[train,]

# validation sample with 30% of tuples

df.validate <- df[-train,]

# create classical decision tree

library("rpart")

dtree <- rpart(earnings ~ ., data=df.train,method="class",parms=list(split="information"))

print(dtree$cptable)

# min xerror is 0.6443572 with 5 splits w/ std error 0.00999

# smallest tree within that threshold is 5 splits

# I have gone back and tested with less splits...

# 4 gives same results as 5 and 3 is much less accurate

# prune tree

dtree.pruned <- prune(dtree, cp=0.01)

library("rpart.plot")

# plot decision tree

prp(dtree.pruned, type=2,extra=104,main="Decision Tree 5")

dtree.pred <- predict(dtree.pruned, df.validate, type ="class")

dtree.perf <- table(df.validate$earnings,dtree.pred, dnn=c("Actual","Predicted"))

print(dtree.perf)

tn <- dtree.perf[1,1]

fp <- dtree.perf[1,2]

fn <- dtree.perf[2,1]

tp <- dtree.perf[2,2]

accuracy <- (tp+tn)/(tp+tn+fp+fn)

error.rate <- (fp+fn)/(tp+tn+fp+fn)

sensitivity <- tp/(tp+fn)

specificity <- tn/(tn+fp)

precision <- tp/(tp+fp)

recall <- tp/(tp+fn)

f.measure <- (2\*precision\*recall)/(precision+recall)

cat("Accuracy: ", accuracy, "\n")

cat("Error Rate: ", error.rate, "\n")

cat("Sensitivity: ", sensitivity, "\n")

cat("Specificity: ", specificity, "\n")

cat("Precision: ", precision, "\n")

cat("Recall: ", recall, "\n")

cat("F-measure: ", f.measure, "\n")

Results:

A computer screen shot of a computer error

Description automatically generated

