70-374 HW2 Solution

Sunan Qian

3/18/2021

In this homework you will do exercises in Shannon entropy, information gain, data transformation, and decision trees. Please follow the instruction given in the homework and answer the question.

Part 1: Partition the Dataset and Run a Predictive Model (20 PTS)

Download: ToyotaCorolla.csv

The dataset ToyotaCorolla.csv contains data on used cars on sale during the late summer of 2004 in the Netherlands. It has 1436 records containing details on 38 attributes, including Price, Age, Kilometers, HP, and other specifications. We plan to analyze the data using linear regression models.

Write the R code for the following:

1. Remove ID, Model, Fuel_Type, and Color variables from the dataset.
2. Change the name of the variable "KM" TO "Kilometers"
3. Randomly partition the dataset(use seed = 1): Partitioning percentages for training (70%), validation (30%).
4. Run a linear regression model using training data to predict the price of the car, using rest of the variables. If needed, appropriately change variables class to factor before running regression.
5. Use the above model to predict the prices of cars in the test data and report accuracy.

```
set.seed(1)

cars<-read.csv("ToyotaCorolla.csv")
cars<-cars[,c(-1, -2)]

names(cars)[which(names(cars)=="KM")]="Kilometers"

# Convert into factor
cars$Automatic<-as.factor(cars$Automatic)

train_rows<-sample(row.names(cars),length(cars$Price)*0.7)
test_rows<-setdiff(row.names(cars),train_rows)

train_data<-cars[train_rows,]
test_data<-cars[test_rows,]

reg<-lm(Price~.,data = train_data)
library(forecast)
```

```
## Warning: package 'forecast' was built under R version 3.6.3
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
pred<-predict(reg,test_data)
```

```
## Warning in predict.lm(reg, test_data): prediction from a rank-deficient fit
## may be misleading
```

```
accuracy(pred,test_data$Price)
```

```
##                 ME     RMSE      MAE       MPE     MAPE
## Test set 141.7395 1115.385 860.9302 0.6927578 8.590586
```

Part 2: Shannon Entropy (log base 2) and Information Gain (10 PTS)

Please use calculation (without using R) to answer the following two questions in Part 2.

2.1 What is the Shannon entropy for the following situations? (Show the calculations) (4 PTS)

Today is a six-letter long day (i.e. Monday length is 6).

Solution:

$$\text{Entropy} = \sum_{i=1}^{n} -P_i \log_2(P_i) = -\frac{3}{7}\log_2\left(\frac{3}{7}\right) - \frac{4}{7}\log_2\left(\frac{4}{7}\right) = 0.5238{+}0.4613 = 0.9851$$

2.2 Consider the following toy dataset with three attributes and an outcome variable Conversion (yes or no).

	Conversion	Previously Shopped	Subscribed to Email marketing	Wish list not empty
User 1	Yes	Yes	No	Yes
User 2	No	No	Yes	Yes
User 3	Yes	Yes	No	Yes
User 4	No	No	Yes	No
User 5	No	No	No	No

What are information gains from splitting on each of those three attributes with Conversion as the y-variable or label? Which one would you choose as the first splitting variable? Show all your work. (6 PTS)

Solution:

Information gain = entropy before split - average entropy of children after split

$$\text{Entropy before split} = \text{Entropy (Yes prob, No prob)} = \text{Entropy}\left(\frac{2}{5}, \frac{3}{5}\right) = -\frac{2}{5}\log\left(\frac{2}{5}\right) - \frac{3}{5}\log\left(\frac{3}{5}\right) = 0.971$$

Attribute 1 – 'Previously shopped'

- Value is "Yes"

  splitted [2,0] = Entropy(1,0) $= -(1)\log_2(1) - (0)\log_2(0) = 0$
- Value is "No"

  splitted [0,3] = Entropy(1,0) $= -(1)\log_2(1) - (0)\log_2(0) = 0$

Average entropy of children after splitting with 'Previously shopped' = 0

Information Gain from 'Previously shopped' = 0.971 - 0 = 0.971

Attribute 2 – 'Subscribed to Email marketing'

- Value is "Yes"

  splitted [0,2] = Entropy(1,0) = $-(1)\log_2(1) - (0)\log_2(0) = 0$

- Value is "No"

  splitted [2,1] = Entropy$\left(\frac{2}{3}, \frac{1}{3}\right) = -\frac{2}{3}\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right) = 0.918$

Average entropy of children after splitting with 'Subscribed to Email marketing' = 0.918

Information Gain = $0.971 - \frac{2}{5} \cdot 0 - \frac{3}{5} \cdot 0.918 = 0.420$

Attribute 3 – 'Wish list not empty'

- Value is "Yes"

  splitted [2,1] = Entropy$\left(\frac{2}{3}, \frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right) = 0.918$

- Value is "No"

  splitted [0,2] = Entropy(1,0) = $-(1)\log_2(1) - (0)\log_2(0) = 0$

Average entropy of children after splitting with 'Wish list not empty' = 0.918

Information Gain = $0.971 - \frac{2}{5} \cdot 0 - \frac{3}{5} \cdot 0.918 = 0.420$

From the information gain of all three attributes, we can state that the 'Previously Shopped' attribute will be the first splitting variable as it has the highest information gain.

Part 3: Dataset Exploration (20 PTS)

Download: bank-data.csv

The key variable in this data set is PEP (Personal Equity Plan, a savings product our bank offers).

Our goal is to predict whether or not a customer will purchase a PEP. We have data from 600 customers as to their purchasing patterns. Data description is in the file called DataDescription4bankdata.docx.

We will use entropy library for computing the entropy values of the attributes.

```
# Setting the working directory where the input file is present
bank  <- read.csv("bank-data.csv")

bank = na.omit(bank)
```

```
# Exploring the data
str(bank)
```

```
## 'data.frame':    600 obs. of  12 variables:
##  $ id          : Factor w/ 600 levels "ID12101","ID12102",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ age         : int  45 61 30 58 50 30 29 35 62 36 ...
##  $ sex         : Factor w/ 2 levels "FEMALE","MALE": 1 2 1 1 1 2 1 2 2 2 ...
##  $ region      : Factor w/ 4 levels "INNER_CITY","RURAL",..: 4 1 1 4 4 1 1 1 1 1 ...
##  $ income      : num  26948 49457 14724 34525 22052 ...
##  $ married     : Factor w/ 2 levels "NO","YES": 1 2 2 2 1 1 2 2 2 2 ...
##  $ children    : int  0 1 0 2 3 3 1 1 0 1 ...
##  $ car         : Factor w/ 2 levels "NO","YES": 1 2 2 2 1 1 1 1 1 2 ...
##  $ save_act    : Factor w/ 2 levels "NO","YES": 2 2 1 2 2 1 2 2 1 2 ...
##  $ current_act: Factor w/ 2 levels "NO","YES": 2 2 2 1 2 2 2 2 2 2 ...
##  $ mortgage    : Factor w/ 2 levels "NO","YES": 2 2 1 1 1 1 1 1 2 1 ...
```

```
##  $ pep        : Factor w/ 2 levels "NO","YES": 1 2 1 2 2 1 1 2 2 2 ...
```

3.1 Explore the data using any plot commands like Scatter Plot, Histogram, Boxplot, Table, etc. Name the top attribute that you think will be informative for predicting the class attribute "pep". Justify your answer. You must provide at least one plot to justify your answer. (Note: this is an open ended question, any reasonable answer is acceptable.) (10 PTS)

Solution to this question is subjective.

Sample solutions for one attribute is given below:

```
library("gmodels") # for CrossTable
```

```
## Warning: package 'gmodels' was built under R version 3.6.3
```

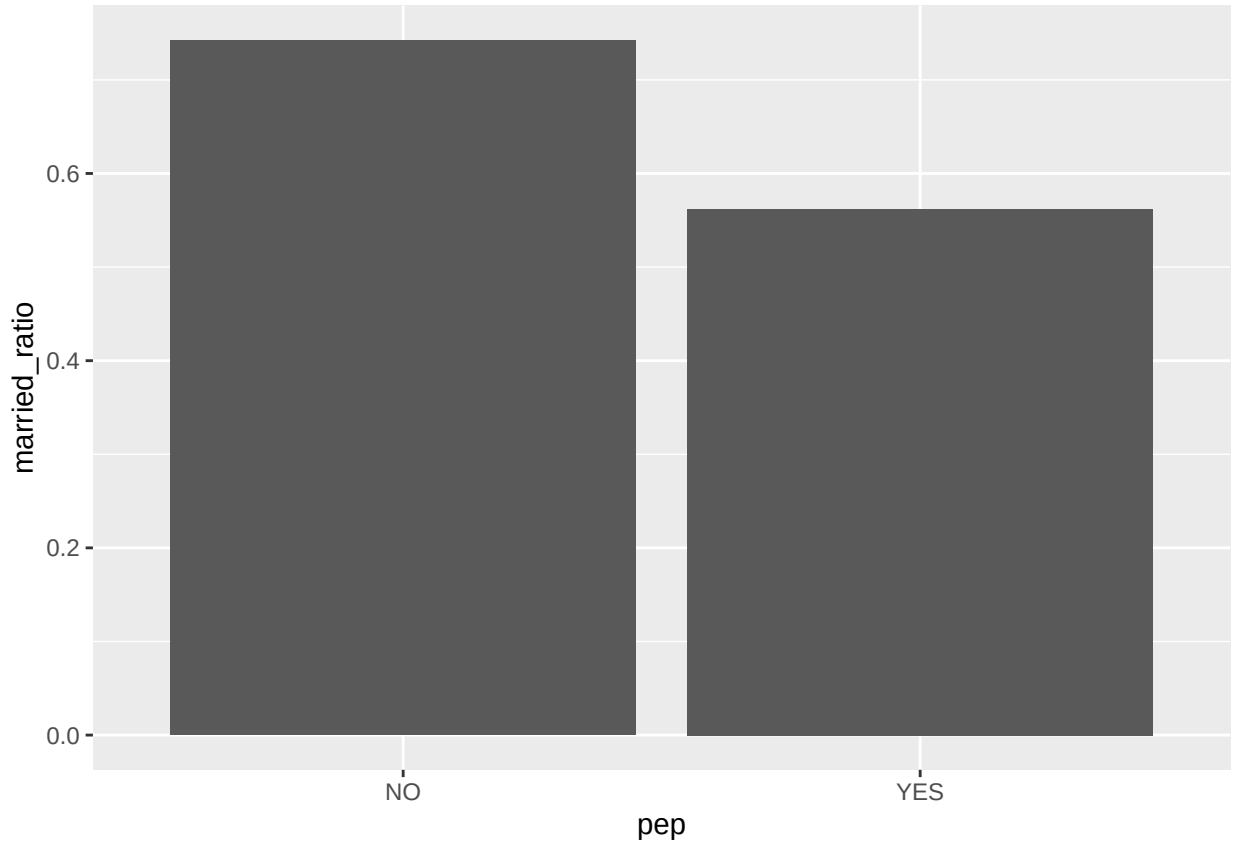
```
CrossTable(bank$married, bank$pep)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  600
##
##
##              | bank$pep
## bank$married |        NO |       YES | Row Total |
## -------------|-----------|-----------|-----------|
##           NO |        84 |       120 |       204 |
##              |     6.499 |     7.733 |           |
##              |     0.412 |     0.588 |     0.340 |
##              |     0.258 |     0.438 |           |
##              |     0.140 |     0.200 |           |
## -------------|-----------|-----------|-----------|
##          YES |       242 |       154 |       396 |
##              |     3.348 |     3.984 |           |
##              |     0.611 |     0.389 |     0.660 |
##              |     0.742 |     0.562 |           |
##              |     0.403 |     0.257 |           |
## -------------|-----------|-----------|-----------|
## Column Total |       326 |       274 |       600 |
##              |     0.543 |     0.457 |           |
## -------------|-----------|-----------|-----------|
##
##
```

```
# Sample plot
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.6.3
```

```
bank$married_yes = ifelse(bank$married=='YES',1,0)
data_sum<-aggregate(bank$married_yes,by=list(bank$pep),FUN=mean)
names(data_sum) <- c("pep","married_ratio")
ggplot(data_sum,aes(pep,married_ratio))+
  geom_bar(stat = "identity")
```




```
bank$married_yes <- NULL
```

As you can see, among pep=Yes, there is a lower married-ratio. It seems that, the attribute "married" can be a good predictor of the class variable "pep", since married people tend to get pep less.

3.2 Calculate the Shannon entropy for categorical attributes "sex", "region", "married", "current_act", "save_act" in the dataset and identify which one among the listed five attributes is with the highest entropies. (10 PTS)

```
# loading entropy library
library(entropy)

# A function that calculates the entopy
shannonEntropy<-function(x){
  Col_ent<-entropy.empirical(table(x)/length(x), unit="log2")
  return(Col_ent)
}

# apply function lets us perform column-wise or row-wise operation
```

```
# we pass the data set "bank" and value "2" for column-wise operation
# followed by the function "shannonEntropy"
all_entropy<-apply(bank[,c("sex","region",'married','current_act',"save_act")],2,shannonEntropy)
all_entropy
```

```
##         sex      region     married current_act    save_act
##   1.0000000   1.7976010   0.9248187   0.7978009   0.8931735
```

From the result, attributes with the highest entropy are region.

Part 4: Decision Trees (30 PTS)

Continue with: bank-data.csv

Using bank-data, we are going to build a decision tree with "pep" being the target variable. Note:

- Please remove column "id" before training the model.
- For this example we will use all the data to train the decision tree. So you don't need to partition the data and reserve any data for testing.
- Use: set.seed(5)
- We will use "caret" library, generic training function called "train", and method called "rpart". Use the help function or look at the caret library pdf "https://cran.rproject.org/web/packages/caret/caret.pdf".
- Use 10-fold cross-validation and train the decision tree model.
- Set: tuneLength = 30

Once you have finished train the decision tree, answer the questions below:

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 3.6.3
```

```
## Loading required package: lattice
```

```
set.seed(5)
```

4.1 Why are we using cross-validation? (5 PTS)

Solution:

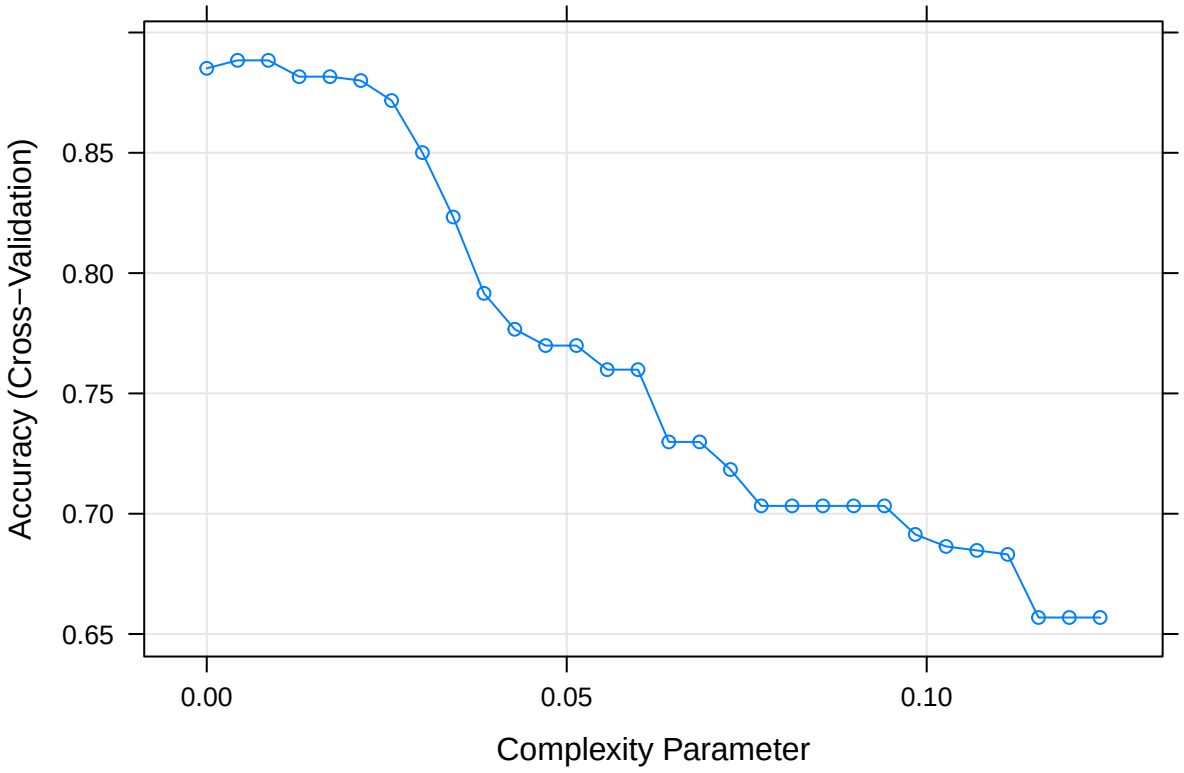
To prevent over-fitting

4.2 Provide the decision tree accuracy plot (across different complexity parameters) and the decision tree plot for the final model. (5 PTS)

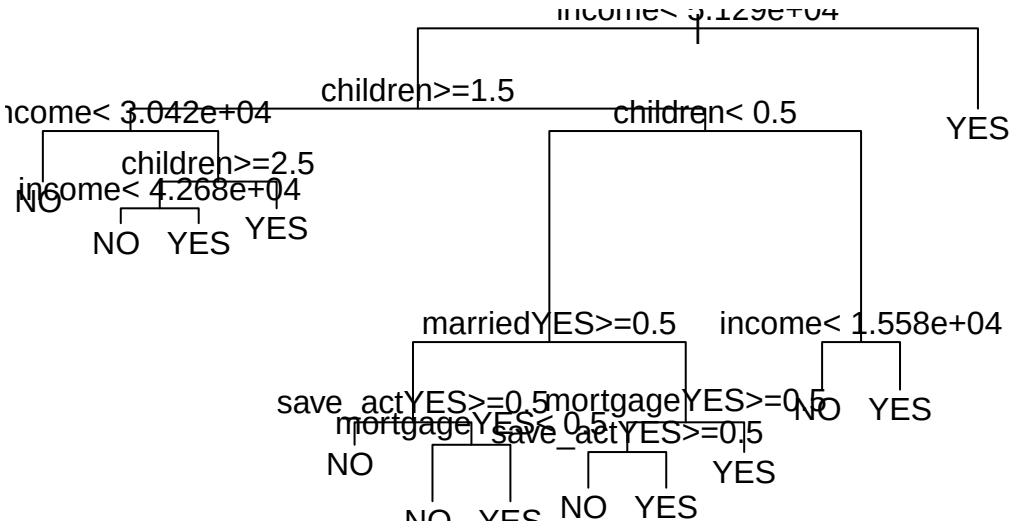
```
bank = bank[,-1]

cvtree <- train(pep~., data=bank,
                method = 'rpart',
                trControl=trainControl(method = 'cv', number=10),
                tuneLength = 30,
                #tuneGrid = data.frame(cp = c(0.04,0.05))
                )

# plot the accuracies
plot(cvtree)
```




```
# draw the tree plot
plot(cvtree$finalModel)
text(cvtree$finalModel)
```

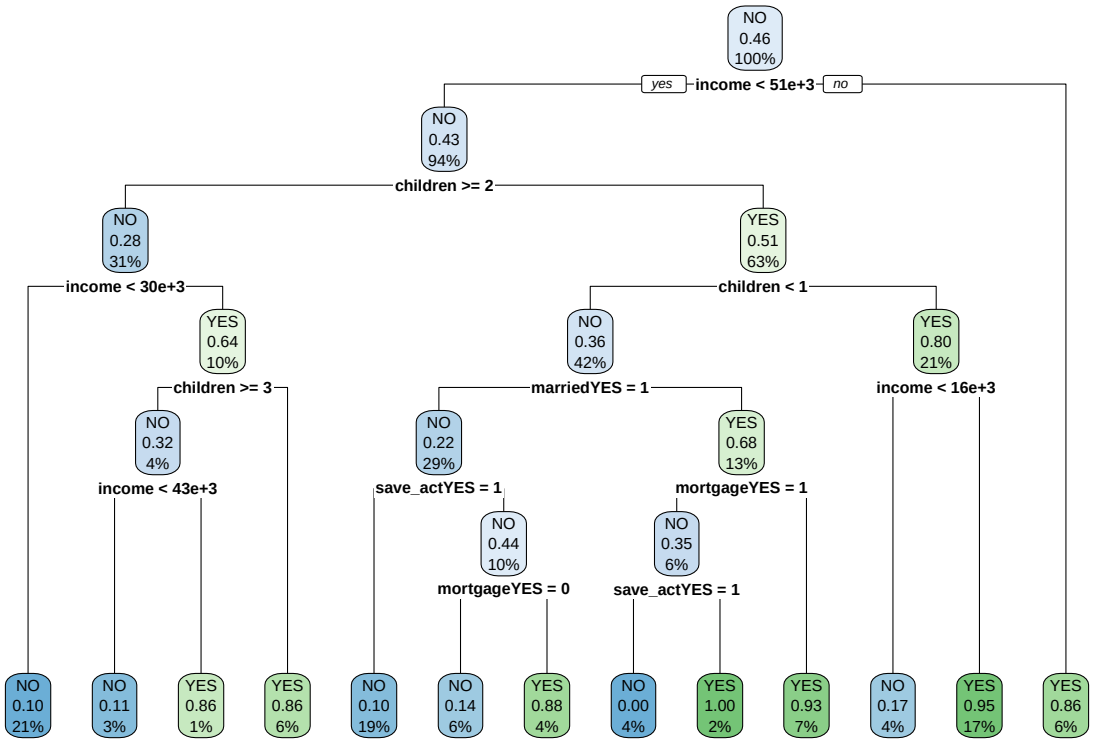



```
# more beautiful plots
library(rpart.plot)
```

```
## Warning: package 'rpart.plot' was built under R version 3.6.3
```

```
## Loading required package: rpart
```

```
rpart.plot(cvtree$finalModel)
```




```
#rpart.plot(tree.fit1$finalModel,fallen.leaves = FALSE)

varImp(cvtree)
```

```
## rpart variable importance
##
##                  Overall
## income         100.00000
## children        67.58018
## save_actYES     58.67768
## marriedYES      56.84271
## age             43.10144
## mortgageYES     41.73816
## regionTOWN       6.98008
## sexMALE          4.12164
## regionRURAL      1.76623
## regionSUBURBAN   1.13846
## carYES           0.04233
## current_actYES   0.00000
```

4.3 Based on the tree plot, answer the following:

A) What percentage of people are with income > 51e+3 and what percentage of people with income > 51e+3 buy PEP? (5 PTS)

There are about 6% people with income greater than 51e+3. About 86% of people with income greater than 51284.3 buy PEP.

B) What are the characteristics of people with the lowest percentage of buying PEP? (5 PTS)

There is a leave with 0 % of people buying PEP. Describe those people.

People who are

- not married and
- less than 51284.3 income and
- has mortgage and
- has saving account and
- no child

C) Does the decision tree make sense (subjective answer)? Justify your answer. (5 PTS)

All relevant answers are given full marks.

Yes it says high income individuals with children will buy the savings product while single non-married people with low income who has mortgage and savings account already, will not.

4.4 Given the characteristics of the potential 3 customers below, predict using your decision tree plot, the probability that these customers will buy PEP if you send them targeted marketing letters. Then rank customers from most likely to buy to least likely to buy by their ID. (5 PTS)

16783: 50% chance to buy

19832: 36.7% chance to buy

32334: 91.3% chance to buy

Order is 32334, 16783, then 19832.

Part 5 Companion Dataset Analyses (20 PTS)

All relevant answers are given full marks.