Exercises - random forests

Jan-Philipp Kolb

23 Mai, 2019

1. Read in the `adult.csv` file with `header=FALSE`. Store this in df. Use `str()` command to see the dataframe. Download the Data from here
2. You are given the meta_data that goes with the CSV. You can download this **here**. Use that to add the column names for your dataframe. Notice the df is ordered going from V1,V2,V3,... and so on.
3. Use the table command and print out the distribution of the class feature.
4. Change the class column to binary.
5. Use the `cor()` command to see the corelation of all the numeric and integer columns including the class column.
6. Split the dataset into Train and Test sample. You may use `sample.split()` and use the ratio as 0.7 and set the seed to be 1000. Make sure to install and load `caTools` package.
7. Check the number of rows of Train and Test
8. We are ready to use decision tree in our dataset. Load the package `rpart` and `rpart.plot`
9. Use `rpart` to build the decision tree on the Train set. Include all

Solution

1

```
df=read.csv("C:/Contract/adult.csv",header=FALSE)
str(df)
```

2

```
colnames(df)=c("age","workclass","fnlwgt","education","education
```

3

```
table(df$class)
```

4

```
df$class=ifelse(df$class==" >50K", 1, 0)
```

5

```
cor(df[,c(1,3,5,11,12,13,15)])
```

Exercise

1. use the `predict()` command to make predictions on the Train data. Set the method to `class`. Class returns classifications instead of probability scores. Store this prediction in pred_dec.
2. Print out the confusion matrix
3. What is the accuracy of the model. Use the confusion matrix.
4. What is the misclassification error rate? Refer to Basic_decision_tree exercise to get the formula.
5. Lets say we want to find the baseline model to compare our prediction improvement. We create a base model using this code

```r
length(Test$class)
base=rep(1,3183)
```

- Use the table() command to create a confusion matrix between the base and Test$class.

6. What is the number difference between the confusion matrix accuracy of dec and base?
7. Remember the predict() command in question 1. We will use the

Solution

```
df=read.csv("C:/Contract/adult.csv",header=FALSE)
library(caTools)
colnames(df)=c("age","workclass","fnlwgt","education","education
df$class=ifelse(df$class==" >50K", 1, 0)
df$class=as.factor(df$class)
set.seed(1000)

split=sample.split(df$class, SplitRatio=0.8)

Train=df[split==TRUE,]

Test=df[split==FALSE,]
library(rpart)
library(rpart.plot)
dec=rpart(class~., data=Train)
par(mar = rep(2, 4))
```

1.