Decsion trees

Subramani.M

18 May 2018

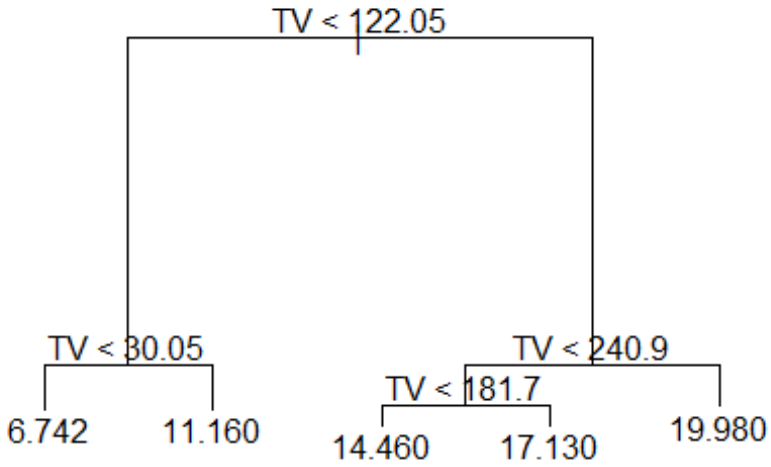
* Decision trees IN REGRESSION ANALYSIS

in lm method we use mathematical stuffs to calculate the co efficients but in decision trees

we use rules to take decision

* Top down approach, greedy algorithm

```
library(tree)

## Warning: package 'tree' was built under R version 3.4.4

library(dplyr)

## Warning: package 'dplyr' was built under R version 3.4.3

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

setwd("C:/Users/Administrator/Desktop/Machine Learning/DATA SETS")
advertising <- read.csv("C:/Users/Administrator/Desktop/Machine Learning/DATA
SETS/Advertising.csv")

model_tree = tree(sales ~ TV,data = advertising)
{{plot(model_tree)
  text(model_tree)}}
```




```
advertising %>% filter(TV < 134.75) %>% summarise(count =n() )

## Warning: package 'bindrcpp' was built under R version 3.4.3

##   count
## 1    89

n <- advertising %>% filter(TV < 2.4) %>% summarise(predict=mean(sales))

library(dplyr)
TV_uniq = sort(unique(advertising$TV))
cuts = c()
mses=c()

for( i in seq(1,length(TV_uniq)-1)){
  curr_cut = (TV_uniq[i] + TV_uniq[i+1]) / 2
  #print(curr_cut)
  cuts = c(cuts,curr_cut)
  samples_left = advertising %>% filter(TV < curr_cut)
  samples_right = advertising %>%  filter(TV > curr_cut)
  avg_left = mean(samples_left$sales)
  avg_right = mean(samples_right$sales)
  advertising$predicted_sales = if_else(advertising$TV <
curr_cut,avg_left,avg_right)
  curr_mse = sum((advertising$sales - advertising$predicted_sales) ^ 2) /
nrow(advertising)
  mses = c(mses,curr_mse)
```

```
}
models_perf <- data.frame("TV_cut"=cuts,"MSE"=mses)
models_perf %>% arrange(MSE) %>% head(1)

##   TV_cut      MSE
## 1 122.05 14.29269

TV_uniq = sort(unique(advertising$TV))
advertising_temp = advertising %>% filter(TV < 122.05)
cuts = c()
mses=c()

for( i in seq(1,length(TV_uniq)-1)){
  curr_cut = (TV_uniq[i] + TV_uniq[i+1]) / 2
  #print(curr_cut)
  cuts = c(cuts,curr_cut)
  samples_left = advertising_temp %>% filter(TV < curr_cut)
  samples_right = advertising_temp%>%  filter(TV > curr_cut)
  avg_left = mean(samples_left$sales)
  avg_right = mean(samples_right$sales)
  advertising_temp$predicted_sales = if_else(advertising_temp$TV <
curr_cut,avg_left,avg_right)
  curr_mse = sum((advertising_temp$sales - advertising_temp$predicted_sales)
^ 2) / nrow(advertising)
  mses = c(mses,curr_mse)
}
models_perf <- data.frame("TV_cut"=cuts,"MSE"=mses)
models_perf %>% arrange(MSE) %>% head(1)

##   TV_cut      MSE
## 1  30.05 1.569203
```