#AIR Random-Forest 과제 "auc 0.91 이상으로 올리기"

201811526 이은주

[코드]

rm(list=ls())

install.packages("dplyr")

install.packages("ISLR")

install.packages("glmnet")

install.packages("randomForest")

install.packages("gbm")

install.packages("rpart")

install.packages("boot")

install.packages("ggplot2")

install.packages("ROCR")

library(dplyr)

library(ISLR)

library(glmnet)

library(randomForest)

library(gbm)

library(rpart)

library(boot)

library(ggplot2)

library(ROCR)

adult <- read.csv("C:\\Users\\이은주\\Desktop\\AIR 동아리\\0430과제\\adult.data",

header=FALSE, strip.white=T)

names(adult) <- c('age','workclass','fnlwgt','education','education\_num',

'marital\_status','occupation','relationship','race','sex','capital\_gain',

'capital\_loss','hours\_per\_week','native\_country','wage')

glimpse(adult) #자료 확인:연속형변수 6개, 범주형변수 8개 , wage는 반응변수

adult <- adult[,-c(9,14)] #랜덤포레스트하여 accuracy와 gini를 보고 둘 다 낮은 변수 제거

apply(adult, 2, function(x){sum(x=='?')})

#adult 자료에서 열별로 함수(?인 x의 갯수 합)를 적용

#workclass, occupation, native\_country변수의 ?를 NA로 취급

#반응변수인 wage의 2개범주

levels(adult$wage)

##train, validation, test set설정

set.seed(201311535)

n <- nrow(adult)

idx <- 1:n

training\_idx <- sample(idx, n\*.6)

idx <- setdiff(idx, training\_idx) #idx에는 있고 training\_idx에는 없는 것

validate\_idx <- sample(idx, n\*.2)

test\_idx <- setdiff(idx, validate\_idx)

length(training\_idx)

length(validate\_idx)

length(test\_idx)

training <- adult[training\_idx,]

validation <- adult[validate\_idx,]

test <- adult[test\_idx,]

###########EDA(탐색적 자료 분석)###########

#EDA related with wage

x11() #x11()은 새로운 창에 그림을 그려주는 명령

training %>%

ggplot(aes(x=age, fill=wage)) + geom\_density(alpha=.5) ##중산층 이상의 수입(파랑색)은 나이와 선형적이지 않다.

training %>%

filter(race %in% c('Black','White')) %>%

ggplot(aes(age, fill=wage)) + geom\_density(alpha=.5) + ylim(0,0.1) + facet\_grid(race ~ sex, scales='free\_y')

##흑인이던 백인이던 나이가 40대쯤 넘어갈 때 부터 중산층이 많아짐을 확인

##남자인 경우, 백인과 흑인간의 density의 차이가 없다고 할 수 있다.

##여자인 경우, 흑인은 약 30대 부터 중산층이고 50대 이후로 갈수록 거의 없고, 백인인 경우 25~60까지 중산층이 있다.

x11()

training %>%

group\_by(education\_num, wage) %>%

tally() %>%

group\_by(education\_num) %>%

mutate(rate\_middle = n / sum(n)) %>%

filter(wage==">50K") %>%

ggplot(aes(x=education\_num, y=rate\_middle, fill=education\_num)) + geom\_bar(stat='identity')

##학력이 높아질수록, 중산층의 비율이 높아짐을 확인

#########Random-Forest##########

ad\_rf <- randomForest(wage~., training, importance=T) #500개의 트리, 3개의 변수

ad\_rf

plot(ad\_rf)

tmp <- importance(ad\_rf) #변수의 상대적 중요도

tmp[order(-tmp[,1]), 1, drop=F] #표현1-내림차순으로 나열

varImpPlot(ad\_rf) #표현2-그림

predict(ad\_rf, newdata=adult[1:5,])

predict(ad\_rf, newdata=adult[1:5,], type='prob')

y\_hat\_rf <- predict(ad\_rf, newdata=validation, type='prob')[,'>50K']

y\_obs <- ifelse(validation$wage == '>50K', 1, 0)

pred\_rf <- prediction(y\_hat\_rf, y\_obs)

perf\_rf <- performance(pred\_rf, measure='tpr', x.measure='fpr')

plot(perf\_rf, col='black',main='ROC Curve')

abline(0,1)

legend('bottomright',

legend='RF', col='black', lty=1, lwd=2)

performance(pred\_rf, "auc")@y.values[[1]]

#####test set에 Random-Forest적합########

y\_hat\_rf\_test <- predict(ad\_rf, newdata=test, type='prob')[,">50K"]

y\_obs\_test <- ifelse(test$wage==">50K",1,0)

pred\_rf\_test <- prediction(y\_hat\_rf\_test, y\_obs\_test)

perf\_rf\_test <- performance(pred\_rf\_test, measure='tpr',x.measure='fpr')

plot(perf\_rf\_test, col='black', main='ROC Curve')

abline(0,1)

legend('bottomright',

legend='RF', col='black', lty=1, lwd=2)

performance(pred\_rf\_test, 'auc')@y.values[[1]]

[결과]



-> auc 91.18%로 증가하였다.