

# 4주차 과제 예시답안

진도 : decision tree 개념이해 2 ~ hierarchical clustering

```
# Week_4 Quiz
# 문제 바로 아래에 답안을 작성하여, <<r script 형태>>로 저장 후 제출해주시면 됩니다. (.R 파일업로드 / Rproj파일은 업로드하지 않으셔도 됩니다.)
# 답안인 코드만 작성하시면 됩니다. 코드 실행 결과 포함 X
# [R Script Encoding Error] : File > Reopen with Encoding > 'UTF-8' > OK

# 아래 네 줄의 코드를 실행 후 문제를 풀어주세요.
library(caret)
library(tree)
set.seed(2020)
bcw <- read.csv("경로/BCW.csv") # 경로는 본인의 컴퓨터 환경에 맞게 작성하세요.
bcw <- bcw[, -1] # id 변수 제거

# 1번 문제는 decision tree 모형적합에 관련된 문제로, 코드 작성에 어려움이 있으면 해당 실습파일을 참고하셔서 문제를 풀어보세요
# (5점) 1- (1) bcw의 diagnosis 변수를 factor형 변수로 바꿔주세요.
# hint. as.factor함수 이용
bcw$diagnosis <- as.factor(bcw$diagnosis)

# (5점) 1-(2) train 데이터와 test 데이터를 7:3의 비율로 나누어서 지정해주세요.
flag <- sort(sample(nrow(bcw), nrow(bcw)*.7))
train <- bcw[flag,]
test <- bcw[-flag,]

# (5점) 1-(3) target 변수는 diagnosis 입니다. tree함수를 이용하여 train데이터를 결정나무에 적합시키세요.
treeRaw <- tree(diagnosis ~., data = train)

적합된 나무 형태를 보려면 아래 코드를 실행하시면 됩니다.
plot(treeRaw)
text(treeRaw)

# cv.tree 함수를 이용해 그래프로 나타내보니, size = 7에서 가지치기를 하는 것이 좋은 것으로 나타났습니다.
# (5점) 1-(4) size = 7로 가지치기 한 결정나무를 적합하세요.
prune_tree <- prune.misclass(treeRaw, best=7)

# 2번 문제는 svm에 관련된 문제로, svm 실습 파일을 참고하셔서 문제를 풀어보세요.
ctrl <- trainControl(method="repeatedcv", repeats = 5)

# (5점) 2-(1). trControl로 위에서 작성한 ctrl을 이용하는 선형 svm을 적합하세요. 이 때 target변수는 diagnosis입니다.
# hint. train 함수를 이용. method로 적합방법 지정
svm_linear_fit <- train(diagnosis ~ .,
                        data = train,
                        method = "svmLinear",
                        trControl = ctrl,
                        preProcess = c("center", "scale"),
                        metric="Accuracy")

svm_linear_fit

# (10점) 2-(2). 2- (1)에서 적합한 모형으로 test 데이터의 diagnosis값을 예측한 후, 혼동행렬을 만드는 코드를 작성하세요.
pred_test <- predict(svm_linear_fit, newdata=test)
confusionMatrix(pred_test, test$diagnosis)

# (10점) 3-(1) bcw의 데이터 형태를 보고 "적절한 변수"에만 차원축소를 진행하세요.
# 이 때, 변수의 표준화를 반드시 실행하세요.
str(bcw)
등을 통해 변수 형태를 확인하고, 범주형 변수는 제외한 나머지 변수에만 차원 축소를 진행해야 합니다.
bcw.pca <- prcomp(bcw[2:ncol(bcw)], center = T, scale. = T)

# (5점) 3-(2) 분산의 90% 이상을 설명하기 위해 몇 개의 주성분 변수를 사용해야 하나요? 코드를 작성하고 정답을 작성하세요.
summary(bcw.pca)

# 정답 : 분산의 90% 이상을 설명하기 위해서는 7개의 주성분을 사용해야 한다.

# (10점) 4-(1). 연속형 변수에만 k-means clustering을 실시하려고 합니다.
# 이 때 Elbow plot을 그리고 k의 적절한 개수와 그 이유를 작성해주세요.
# 코드와 정답을 모두 작성하세요.
```

```

library(dplyr)
library(factoextra)
fviz_nbclust(bcw[,2:ncol(bcw)], kmeans, method = "wss", k.max = 15) + # ncol=31
  theme_minimal() +
  ggtitle("Elbow plot")

# 정답 : 개인의 해석을 적어주시면 됩니다.
# 예시 답안 : Elbow plot을 보시면 급격하게 꺾이는 부분은 없지만 k=5에서 기울기가 완만해진다고 보고 k=5로 선택하였다.

# bcw 데이터의 2번째 열부터 31번째 열을 이용하여 계층적 군집분석을 실시하려고 합니다.
# (5점) 5-(1) 유클리드 거리를 기반으로 하는 유사도행렬을 생성하세요.

bcw.dist <- dist(bcw[,2:31], method = "euclidean")
bcw.dist

# (5점) 5-(1) bcw.dist를 이용해 ward's method로 계층적 군집분석을 실시하세요.

bcw.hclust.ward <- hclust(bcw.dist, method = "ward.D2" )
plot(bcw.hclust.ward, cex = 0.6, hang = -1)
rect.hclust(bcw.hclust.ward, k = 5, border = 2:6)

# 5-(1)에서 만든 군집 분석 결과물을 확인하였더니 5개의 군집으로 구성하는 것이 적당해보입니다.
# (5점) 5-(2) 원래의 데이터 bcw에 군집분석한 결과물을 이용해 cluster라는 새로운 변수를 생성하세요.

bcw.clusters <- cutree(bcw.hclust.ward, k = 5)
table(bcw.clusters)

bcw$cluster <- bcw.clusters

```