ASSIGNMENT 11.2

Kyle Ramirez

3/5/2022

```
{r setup, include=FALSE} knitr::opts_chunk$set(echo = TRUE)
```

Load the ggplot2 package

library(ggplot2) theme_set(theme_minimal()) library(caret) library(pROC) library(mlbench)

K nearest neighbors

Set the working directory to the root of your DSC 520 directory

setwd("/Users/Kyle/Documents/GitHub/KR/Ramirez_Kyle_DSC510/dsc520")

Load the `data/r4ds/Project/binary-classifier-data` to

binary_df <- read.csv("data/Project/binary-classifier-data.csv")

Load the `data/r4ds/Project/trinary-classifier-data` to

trinary_df <- read.csv("data/Project/trinary-classifier-data.csv")

`x` vs. `y` binary

ggplot(binary_df, aes(x=x, y=y)) + geom_point() + geom_smooth()

`x` vs. `y` trinary

ggplot(trinary_df, aes(x=x, y=y)) + geom_point() + geom_smooth()

check data

head(binary_df) head(trinary_df)

Setup

str(binary_df) binary_df$label[binary_df$label == 0] <- 'No' binary_df$label[binary_df$label == 1] <- 'Yes' binary_df$label <- -factor(binary_df$label)

str(binary_df) trinary_df$label[trinary_df$label == 0] <- 'No' trinary_df$label[trinary_df$label == 1] <- 'Yes' trinary_df$label[trinary_df$label == 2] <- 'Unknown' trinary_df$label <- -factor(trinary_df$label)

Data Partition

set.seed(125) ind_bi <- sample(2, nrow(binary_df), replace = T, prob = c(0.7, 0.3)) training_bi <- binary_df[ind ==1,] test_bi <- binary_df[ind ==2,]

ind_tri <- sample(2, nrow(trinary_df), replace = T, prob = c(0.7, 0.3)) training_tri <- trinary_df[ind ==1,] test_tri <- trinary_df[ind ==2,]

KNN Model

trControl <- trainControl(method = "repeatedcv", number = 10, repeats = 3) set.seed(222) fit <- train(label ~ ., data = training, method = 'knn', tuneLength = 20, trControl = trControl, preProc = c("center","scale"))

Model Performance

fit plot(fit) varImp(fit) pred <- predict(fit, newdata = test_bi) confusionMatrix(pred, test_bi$label)

pred <- predict(fit, newdata = test_tri) confusionMatrix(pred, test_tri$label)

Clustering

library(stats) library(dplyr) library(ggplot2) library(ggfortify)

Set the working directory to the root of your DSC 520 directory

setwd("/Users/Kyle/Documents/GitHub/KR/Ramirez_Kyle_DSC510/dsc520")

Load the `data/r4ds/Project/binary-classifier-data` to

cluster_df <- read.csv("data/Project/clustering-data.csv")

`x` vs. `y` cluster

ggplot(cluster_df, aes(x=x, y=y)) + geom_point() + geom_smooth()

unsupervised learning

cluster_data = select(cluster_df,c(1,2))

WSS Plot to choose maximum number of clusters

wssplot <- function(data, nc=15, seed=1234) {
wss <- (nrow(data)-1)*sum(apply(data,2,var)) for (i in 2:nc){ set.seed(seed) wss[i] <- sum(kmeans(data, centers=i)$withinss)} plot(1:nc, wss, type="b", xlab="Number of Clusters", ylab="Within groups sum of squares") }
wssplot(cluster_data)

Spotting the kink in the curve in order to choose the optimum

K-means cluster

KM = kmeans(cluster_data, 2)

Evaluating Cluster Analysis

Cluster Plot

autoplot(KM,cluster_data, frame=TRUE)

Cluster Centers

KM$centers