STAT Project 1

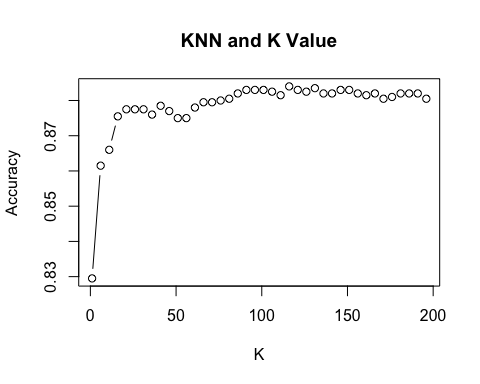
Yebom Kim

2024-09-09

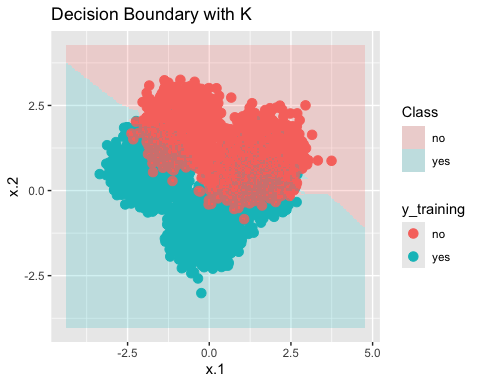
setwd("/Users/springkim/Downloads/")  
test\_data <- read.csv("1-test\_data.csv")  
training\_data <- read.csv("1-tranining-data.csv")  
library(ggplot2)  
library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ readr 2.1.5  
## ✔ forcats 1.0.0 ✔ stringr 1.5.1  
## ✔ lubridate 1.9.3 ✔ tibble 3.2.1  
## ✔ purrr 1.0.2 ✔ tidyr 1.3.1  
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(class)  
#a)  
x\_training <- training\_data[, -ncol(training\_data)]  
y\_training <- training\_data[, ncol(training\_data)]  
x\_test <- test\_data[, -ncol(test\_data)]  
y\_test <- test\_data[, ncol(test\_data)]  
  
k\_vals <- seq(1, 200, by = 5)  
output\_1a <- data.frame(K = k\_vals, Accuracy = numeric(length(k\_vals)))  
  
for (i in 1:length(k\_vals)) {  
 K <- k\_vals[i]  
 y\_pred <- knn(train=x\_training, test=x\_test, cl=y\_training, k=K)  
 accuracy <- sum(y\_pred == y\_test) / length(y\_test)  
 output\_1a$Accuracy[i] <- accuracy  
}  
plot(output\_1a$K, output\_1a$Accuracy, type = "b",   
 xlab = "K", ylab = "Accuracy", main = "KNN and K Value")



#b)  
output\_1b <- data.frame(K = k\_vals, Train\_Error = numeric(length(k\_vals)),   
 Test\_Error = numeric(length(k\_vals)))  
  
#c)  
opt\_index <- which.min(output\_1b$Test\_Error)  
opt\_K <- output\_1b$K[opt\_index]  
opt\_train\_error <- output\_1b$Train\_Error[opt\_index]  
opt\_test\_error <- output\_1b$Test\_Error[opt\_index]  
  
#d)  
x\_min <- min(x\_training[,1]) - 1  
x\_max <- max(x\_training[,1]) + 1  
y\_min <- min(x\_training[,2]) - 1  
y\_max <- max(x\_training[,2]) + 1  
grid <- expand.grid(seq(x\_min, x\_max, length.out = 200), seq(y\_min, y\_max, length.out = 200))  
colnames(grid) <- colnames(x\_training)[1:2]  
  
grid\_pred <- knn(train = x\_training, test = grid, cl = y\_training, k = opt\_K)  
grid\_pred\_df <- data.frame(grid, Class = grid\_pred)  
  
ggplot() +  
 geom\_point(data = as.data.frame(x\_training),   
 aes(x = x\_training[, 1], y = x\_training[, 2], color = y\_training), size = 3) +  
 geom\_tile(data = grid\_pred\_df,   
 aes(x = x.1, y = x.2, fill = Class), alpha = 0.2) +  
 labs(title = paste("Decision Boundary with K"),  
 x = "x.1", y = "x.2")

 ```