Machine Learning Computer Lab 1 (Group A7)

Qinyuan Qi(qinqi464) Satya Sai Naga Jaya Koushik Pilla (satpi345)
Daniele Bozzoli(danbo826)

2023-12-04

Assignment 1: Handwritten digit recognition with K- nearest neighbors (Solved by Qinyuan Qi - qinqi464)

Answer:

(1)

The following code will import the data and divide the data into training, validation and test sets.

```
#######################  Assignment 1.1 ##########################################
# read data
data <- read.csv("optdigits.csv")
row_num <- nrow(data)
cols_num <- ncol(data)

# set the last column name as "label_value" since we need to create a formula later
names(data)[cols_num] <- "label_value"

# set data split ratio to 0.5, 0.25 and 0.25
ratio <- c(train = .5, validate = .25, test = .25)

# data pre-processing
# columns 1-64 are number based and do not need to be normalized, last column
# is integer represent number from 0-9 which don't need to process again

# set random seed
set.seed(12345)

# split data to training, test and validation set
train_id <- sample(1:row_num, floor(row_num * ratio[1]))
train_set <- data[train_id, ]

set.seed(12345)
test_val_id <- setdiff(1:row_num, train_id)
valid_id <- sample(test_val_id, floor(row_num * ratio[2]))
valid_set <- data[valid_id, ]

test_id <- setdiff(test_val_id, valid_id)
test_set <- data[test_id, ]
```

(2)

The code contain a function call to kknn with k=30, kernel = "rectangular" with distance default to 2. We calculate the confusion matrices and the misclassification errors of training and test data set as follows, the code attached in the appendix. As the data showed below, the overall prediction quality is not good enough when k=30. And the misclassification rate of number 0,6 is the lowest. number 8 and 9 have the highest misclassification rate.

Table 1: Confusion matrix for training data set

	0	1	2	3	4	5	6	7	8	9
0	173	0	0	0	0	0	0	0	0	0
1	2	148	0	0	0	0	0	0	0	0
2	2	26	151	0	1	0	0	0	1	0
3	1	6	13	133	8	0	0	1	3	2
4	0	4	5	48	141	4	3	1	8	0
5	0	1	4	17	15	156	2	4	12	9
6	0	2	0	1	8	24	195	12	28	10
7	0	1	0	1	6	13	0	176	44	41
8	0	0	0	0	0	0	0	1	109	76
9	0	0	0	0	0	0	0	0	0	58

Table 2: Confusion matrix for validation data set

	0	1	2	3	4	5	6	7	8	9
0	98	0	0	0	0	0	0	0	0	0
1	0	70	0	0	0	0	0	0	0	0
2	1	26	87	0	1	0	2	0	1	0
3	0	4	11	52	3	0	0	0	1	2
4	0	1	8	25	74	3	0	1	2	0
5	0	1	4	12	18	66	1	0	4	2
6	0	0	1	2	6	19	80	5	10	6
7	0	2	1	0	5	4	0	90	18	13
8	0	0	0	0	1	0	0	1	48	42
9	0	0	0	0	0	0	0	0	0	20

Table 3: Misclassification rates

	rate
Training	0.2464678
Validation	0.2827225

Table 4: Misclassification rates of digits using learning data set

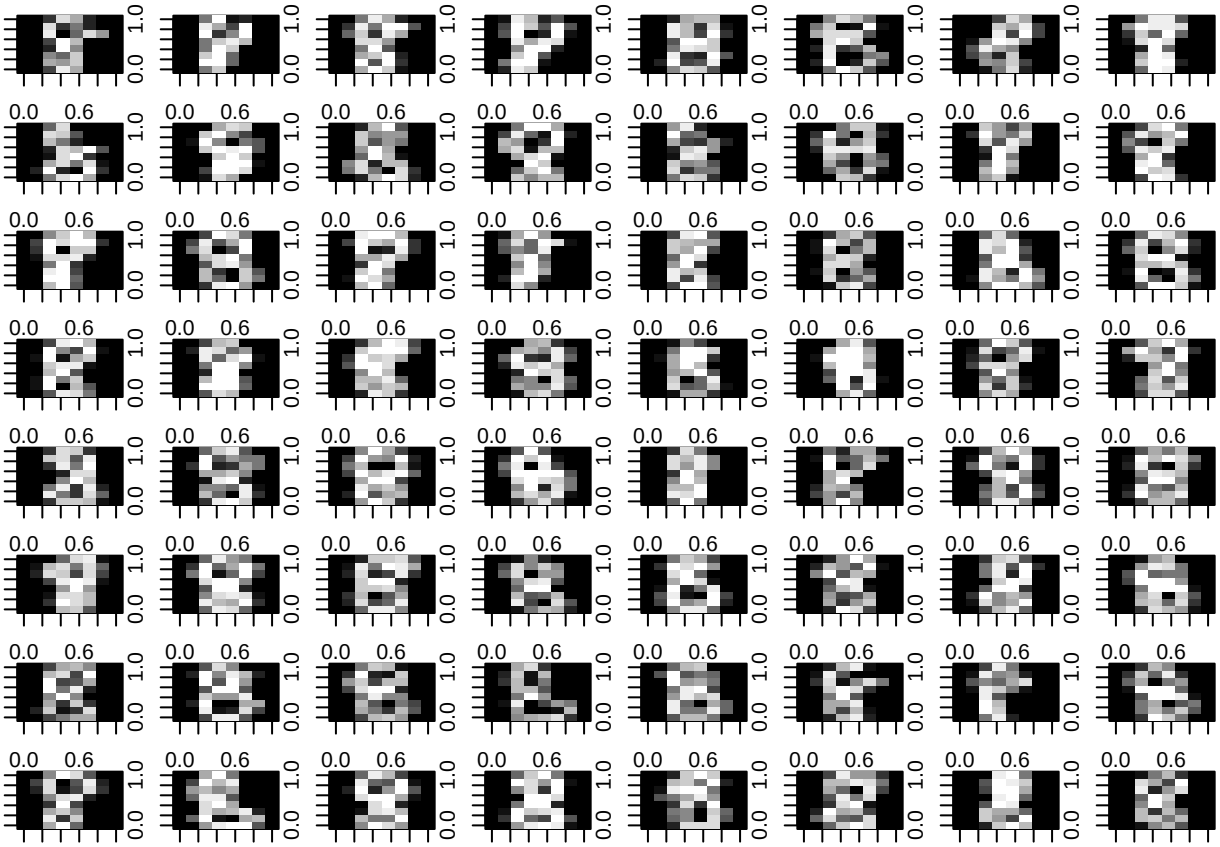
	rate
0	0.0026164
1	0.0209314
2	0.0115123
3	0.0350602

	rate
4	0.0198849
5	0.0214547
6	0.0026164
7	0.0099424
8	0.0502355
9	0.0722135

(3)

By using image function, we generate images on first 64 [8 image] data from learning data set, and we get the following plot.(We use image function instead of heatmap so we can draw more images in one plot)

According to the plot, we can find that image at [3,2] and [5,2] are easy to identify. however, images such as [3,4] [4,2] [4,3] are very hard to figure out the number.



(4)

Using similar code as in 1.2, we fit the knn with k = 1 to 30.

When K increase, since all the points are related to more neighbors, we can say that the model become more complex. misclassification rates also increase. And validation data's misclassification rates always greater than train data's misclassification rates.

According to the plot, we know that optimized K = 1 which will generate the smallest misclassification rate.

When K = 1, we training 3 models, we got the following result in table 5. We can get the conclusion that

the test error rate is higher than the validation error rate. Both validation error rate and test error rate are greater than the value we get from the training data set.

Table 5: Misclassification rates of when K=1

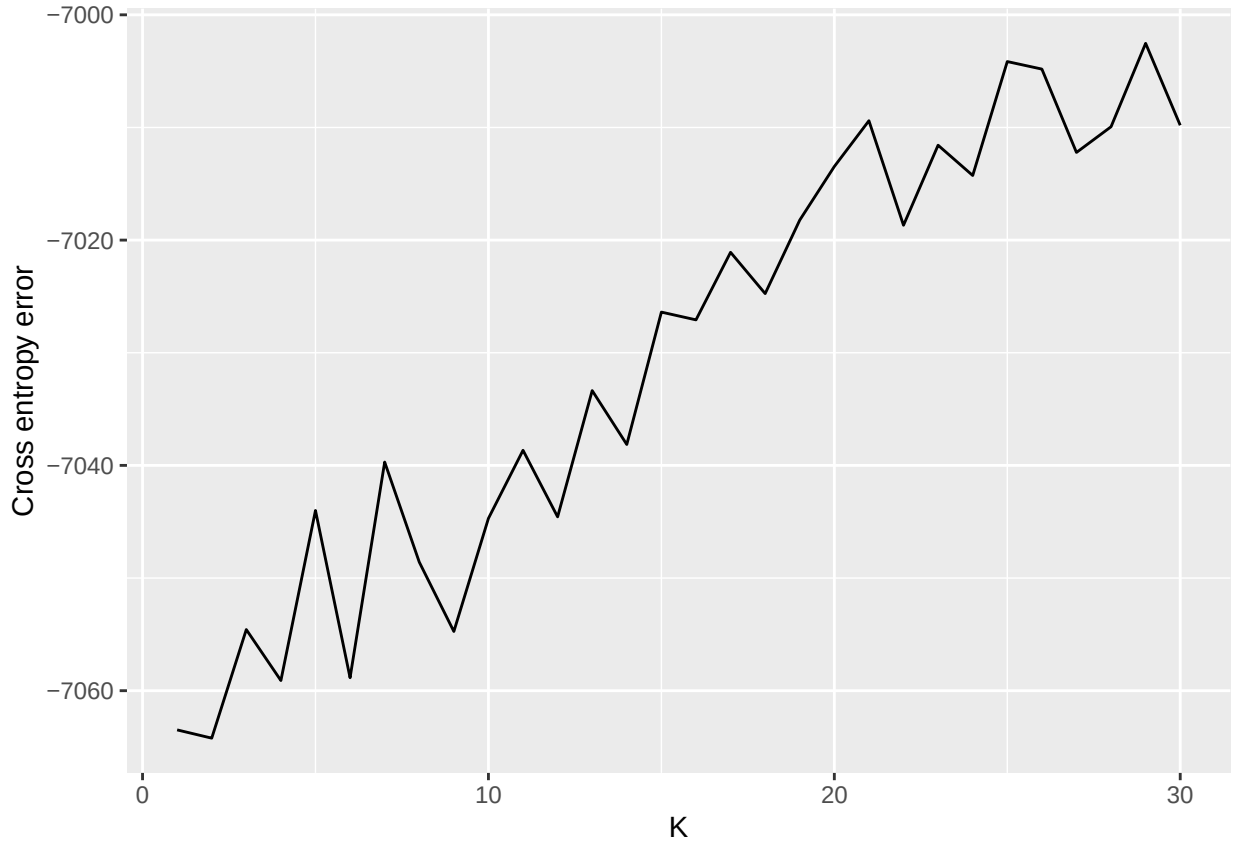
	Test Error
Training	0.0000000
Validation	0.0293194
Test	0.0334728

(5)

We plot the validation error using the code attached in the appendix, the plot of validation error and K as follows.

Since when K=2 get the minimal value, so the optimal K = 2.

The reason why we use cross-entropy is more suitable is: 1) It is sensitive to the predicted probabilities. 2) In multinomial distribution, the predicted value will more likely to be explained as a probability, and cross-entropy loss function is designed to compare the probability based prediction.



Assignment 2: Linear regression and ridge regressions (Solved by Satya Sai Naga Jaya Koushik Pilla)

Answer:

(1)

We read the data and divide data into training and test data (60/40) and normalize them.

```
############################  Assignment 2.1 ####################################
# Load the data
data <- read.csv("parkinsons.csv")
row_num <- nrow(data)
cols_num <- ncol(data)

# Divide the data into training and test data (60/40)
# set data split ratio
ratio <- c(train = .6, test = .4)

# set random seed
set.seed(12345)

# split data
train_id <- sample(1:row_num, floor(row_num * ratio[1]))
train_set <- data[train_id, ]
test_id <- setdiff(1:row_num, train_id)
test_set <- data[test_id, ]
```

```
# normalize data.
train_set <- as.data.frame(lapply(train_set, normalize))
test_set <- as.data.frame(lapply(test_set, normalize))
```

(2)

We create a linear model using following code, we get the parameters as show below.

```
############################  Assignment 2.2 ####################################
# Linear regression model

# apply linear regression
model <- lm(motor_UPDRS ~ ., data = train_set)

# predict the test value using the linear regression just created
test_pred <- predict(model, test_set)

# calculate test MSE
test_mse <- mean((test_pred - test_set$motor_UPDRS)^2)
model
```

And we get test MSE = 0.005370424 we know that Shimmer.APQ3 and Shimmer.DDA contribute significantly to the model

(3)

Functions implemented as below.

```
############################  Assignment 2.3 ####################################
# Loglikelihood function
loglikelihood <- function(theta, sigma) {
  n <- length(Y)
  log_likelihood_values <- -0.5 * (log(2 * pi * sigma^2) + ((Y - theta %*% X) / sigma)^2)
  return(total_log_likelihood <- sum(log_likelihood_values))
}

# ridge
ridge <- function(param) {
  param_length <- length(param)
  theta <- param[1:param_length-2]
  sigma <- param[param_length-1]
  lambda <- param[param_length]
  loglikelihood_result <- loglikelihood(theta, sigma)
  ridge_penalty <- lambda * sum(theta^2)
  return(loglikelihood_result - ridge_penalty)
}

# ridgeopt
ridgeopt <- function(theta, sigma, lambda) {
  size_theta <- length(theta)
  param <- rep(0,size_theta+2)
  for(i in 1:size_theta){
    param[i] <- theta[i]
  }
  param[size_theta + 1] <- sigma
```

```
  param[size_theta + 2] <- lambda
  ridge_result <- optim(par = param, fn = ridge, method="BFGS")

}

# df
df <- function(X,lambda) {
  n <- nrow(X)
  # calculate hat_matrix
  hat_matrix <- X %*% solve(t(X) %*% X + lambda * diag(ncol(X)))
  # get degree of the hat_matrix
  df_ridge <- sum(diag(hat_matrix))
  return(df_ridge)
}
```

(4)

Functions implemented as below.

```
############################  Assignment 2.4 ####################################

#sigma <- 0.01
#theta <- rep(1,21)
#X <- train_set[]
#lambda <- 1
#ridgeopt(theta,sigma,lambda)

#lambda <- 100
#ridgeopt(theta,sigma,lambda)

#lambda <- 1000
#ridgeopt(theta,sigma,lambda)
```

Assignment 3. Logistic regression and basis function expansion (Solved by Daniele Bozzoli)

Answer:

(1)

```
data <- read.csv("pima-indians-diabetes.csv")

age <- data[, 8]
plasma <- data[, 2]

names(data)[9] <- "diabetes"
names(data)[2] <- "plasma"
names(data)[8] <- "age"

x <- age
y <- plasma
df <- data.frame(x, y)

ggplot2::ggplot(df, ggplot2::aes(age, plasma)) +
```

```
  ggplot2::geom_point(ggplot2::aes(colour = data[, 9]))+
  ggplot2::labs(x = "Age", y="Plasma glucose")
```

We observe how a lot of people without diabete is highly concentrated in the lower part of the age axis, explaining how younger individuals tend to be non-diabetic, on the other hand, we notice how especially people with higher plasma glucose concentration tend to be diabetic, less concentrated in a specific age group, more spread across the x-age axis.

(2)(3)

Code as below. Looking at the graph, we notice how the classification is not very accurate. Though, it catches the fact that people with higher plasma glucose concentration are more commonly diabetic. We obtain a misclassification error = 0.266.

```
model <- glm(diabetes ~ plasma + age, fam = binomial(link="logit"), data = data)
summary(model)

# r = 0.5

c1 <- fitted(model) >= .5

missC_error <- sum(data$diabetes != c1) / nrow(data)
cat("Misclassification error:", missC_error, "\n")

ggplot2::ggplot(df, ggplot2::aes(age, plasma)) +
  ggplot2::geom_point(ggplot2::aes(colour = c1))+
  ggplot2::labs(x = "Age", y="Plasma glucose")     # T = Diabetic, F = Non-diabetic
```

(4)

It looks like the line that splits the predicted observations into diabetic and non diabetic into the two cases (r= 0.2 and r=0.8) has the same slope, but different intercept. Both cases are pretty bad predictors as 0.2 and 0.8 are respectively too low and too high to get a good classification for our case.

```
# r = 0.2

c2 <- fitted(model) >= .2

ggplot2::ggplot(df, ggplot2::aes(age, plasma)) +
  ggplot2::geom_point(ggplot2::aes(colour = c2))+
  ggplot2::labs(x = "Age", y="Plasma glucose")

# r = 0.8

c3 <- fitted(model) >= .8

ggplot2::ggplot(df, ggplot2::aes(age, plasma)) +
  ggplot2::geom_point(ggplot2::aes(colour = c3))+
  ggplot2::labs(x = "Age", y="Plasma glucose")
```

(5)

The outcome of this analysis is better than the first case, as we obtain a lower misclassification error than before, and also graphically we notice how this version captures what we were saying in the first part of the

analysis, how people with higher plasma glucose concentration are more exposed to the risk of having diabete. We also notice how we strangely get a misclassificated predicted value in the bottom right corner of the graph.

```r
# Adding new variables, logit function with r= 0.5

z1 <- data$plasma^4
z2 <- data$plasma^3 * data$age
z3 <- data$plasma^2 * data$age^2
z4 <- data$plasma^1 * data$age^3
z5 <- data$age^4

newdata <- cbind(data, z1, z2, z3, z4, z5)

newmodel <- glm(diabetes ~ plasma + age + z1 + z2 + z3 + z4 + z5 , family = binomial(link = "logit") , c

summary(newmodel)

c4 <- fitted(newmodel) >= .5

new_missC_error <- sum(data$diabetes != c4) / nrow(data)
cat("Misclassification error:", new_missC_error, "\n")

ggplot2::ggplot(df, ggplot2::aes(age, plasma)) +
  ggplot2::geom_point(ggplot2::aes(colour = c4))+
  ggplot2::labs(x = "Age", y="Plasma glucose")
```

Appendix: All code for this report

```
##########################  Init code For Assignment 1 ########################
rm(list = ls())
knitr::opts_chunk$set(echo = TRUE)
library(kknn)
library(ggplot2)
library(Metrics)

######################  Assignment 1.1 ############################################
# read data
data <- read.csv("optdigits.csv")
row_num <- nrow(data)
cols_num <- ncol(data)

# set the last column name as "label_value" since we need to create a formula later
names(data)[cols_num] <- "label_value"

# set data split ratio to 0.5, 0.25 and 0.25
ratio <- c(train = .5, validate = .25, test = .25)

# data pre-processing
# columns 1-64 are number based and do not need to be normalized, last column
# is integer represent number from 0-9 which don't need to process again

# set random seed
set.seed(12345)

# split data to training, test and validation set
train_id <- sample(1:row_num, floor(row_num * ratio[1]))
train_set <- data[train_id, ]

set.seed(12345)
test_val_id <- setdiff(1:row_num, train_id)
valid_id <- sample(test_val_id, floor(row_num * ratio[2]))
valid_set <- data[valid_id, ]

test_id <- setdiff(test_val_id, valid_id)
test_set <- data[test_id, ]
######################  Assignment 1.2 ############################################

k_value <- 30
kknn_train <- kknn(label_value ~ .,
                        train_set,
                        train_set,
                        k = k_value,
                        kernel = "rectangular",scale = TRUE)

# generate confusion matrix for training data
confusion_matrices_train <- table(round(kknn_train$fit), train_set$label_value)

# print confusion matrix
#print(confusion_matrices_train)
```

```r
# calculate error rate
error_rate_train <- mean(round(kknn_train$fit) != train_set$label_value)
#cat("Misclassification error for training data is: ", error_rate_train, "\n")

kknn_valid <- kknn(label_value ~ .,
                   train_set,
                   valid_set,
                   k = k_value,
                   kernel = "rectangular",scale = TRUE)

# generate confusion matrix for validate data
confusion_matrices_valid <- table(round(kknn_valid$fit), valid_set$label_value)


# print confusion matrix
#print(confusion_matrices_valid)

# calculate error rate
error_rate_valid <- mean(round(kknn_valid$fit) != valid_set$label_value)
#cat("Misclassification error for valid data is: ", error_rate_valid, "\n")

# misclassification rates
misclassification_rates_data_1_2 <- data.frame(name=c("Training","Validation"),
                                               rate=c(error_rate_train,
                                                      error_rate_valid))
names(misclassification_rates_data_1_2)[1] <- ""

# misclassification errors for each digits on training data
misclassification_rates_for_each_digits_data_1_2 <- data.frame(c(0,9),rate=rep(0,10))
names(misclassification_rates_for_each_digits_data_1_2)[1] <- ""

for(i in 1:10){
  misclassification_rates_for_each_digits_data_1_2[i,] <- c(i-1,mean(
    round(kknn_train$fit) != train_set$label_value &
    train_set$label_value==(i-1)))
}


# render result to tables

knitr::kable(confusion_matrices_train,
             caption = "Confusion matrix for training data set")
knitr::kable(confusion_matrices_valid,
             caption = "Confusion matrix for validation data set")
knitr::kable(misclassification_rates_data_1_2,
             caption = "Misclassification rates")
knitr::kable(misclassification_rates_for_each_digits_data_1_2,
             caption = "Misclassification rates of digits using learning data set")
######################  Assignment 1.3 ##########################################

# Find all the 8 images
train_eight <- train_set[train_set[,65]==8, 1:64]
```

```r
# draw 64 [8 images]
par(mfrow = c(8, 8), mar=c(1,1,1,1))

for(i in 1:64){
  row.id <- i
  m <- matrix(as.numeric(train_eight[row.id, 1:64]), nrow=8, ncol=8)
  image(m[,8:1], col=grey(seq(0, 1, length=16)))
}

######################  Assignment 1.4 ##########################################
# calculate error rate for different k values
error_rates_train <- c()
error_rates_valid <- c()

for(k_value in 1:30){

  kknn_train <- kknn(label_value ~ .,
                           train_set,
                           train_set,
                           k = k_value,
                           kernel = "rectangular",scale = TRUE)

  kknn_valid <- kknn(label_value ~ .,
                         train_set,
                         valid_set,
                         k = k_value,
                         kernel = "rectangular",scale = TRUE)

  # train error rate
  error_rates_train <- append(error_rates_train,
                       mean(round(predict(kknn_train)) != train_set$label_value))


  # valid error rate
  error_rates_valid <- append(error_rates_valid,
                       mean(round(predict(kknn_valid)) != valid_set$label_value))


}

# plot the error rate graph
k <- 1:30

error_rate_data <- data.frame(k, error_rates_train,error_rates_valid)

ggplot(data = error_rate_data) +
  geom_line(mapping = aes(x=k,y=error_rates_train,colour="train")) +
  geom_line(mapping = aes(x=k,y=error_rates_valid,colour="validation")) +
  labs(x = "K",y="Misclassification rates") +
  scale_color_manual(name = "Data set", values = c("train" = "blue", "validation" = "red"))


k_value <- 1
```

```r
kknn_train <- kknn(label_value ~ .,
                           train_set,
                           train_set,
                           k = k_value,
                           kernel = "rectangular",scale = TRUE)

kknn_valid <- kknn(label_value ~ .,
                          train_set,
                          valid_set,
                          k = k_value,
                          kernel = "rectangular",scale = TRUE)

kknn_test <- kknn(label_value ~ .,
                          train_set,
                          test_set,
                          k = k_value,
                          kernel = "rectangular",scale = TRUE)

# train error rate
error_rate_train <- mean(round(predict(kknn_train)) != train_set$label_value)


# valid error rate
error_rate_valid <- mean(round(predict(kknn_valid)) != valid_set$label_value)

# test error rate
error_rate_test <- mean(round(predict(kknn_test)) != test_set$label_value)

data1_4 <- data.frame(c("Training","Validation","Test"),
                      c(error_rate_train,error_rate_valid,error_rate_test))
names(data1_4)[1] <- ""
names(data1_4)[2] <- "Test Error"
knitr::kable(data1_4,
             caption = "Misclassification rates of when K=1")

######################  Assignment 1.5 ##########################################
# init value

k_values <- 1:30
epsilon <- 1e-15
validation_errors <- numeric(length(k_values))

for(k in k_values){
    kknn_train <- kknn(label_value ~ .,
                           train_set,
                           valid_set,
                           k = k,
                           kernel = "rectangular",scale = TRUE)

    pred <- round(predict(kknn_train))
    validation_errors[k] <- -sum(valid_set$label_value * log(pred + epsilon))
}
```

```r
error_cross_entropy_data <- data.frame(k_values,validation_errors)

ggplot(data = error_cross_entropy_data) +
  geom_line(mapping = aes(x=k_values,y=validation_errors))  +
  labs(x = "K",y="Cross entropy error")
############################  Init code For Assignment 2 ########################
rm(list = ls())
knitr::opts_chunk$set(echo = TRUE)

############################  Common Functions ##################################
# normalize data
normalize <- function(x) {
  return((x - min(x)) / (max(x) - min(x)))
}
############################  Assignment 2.1 ####################################
# Load the data
data <- read.csv("parkinsons.csv")
row_num <- nrow(data)
cols_num <- ncol(data)

# Divide the data into training and test data (60/40)
# set data split ratio
ratio <- c(train = .6, test = .4)

# set random seed
set.seed(12345)

# split data
train_id <- sample(1:row_num, floor(row_num * ratio[1]))
train_set <- data[train_id, ]
test_id <- setdiff(1:row_num, train_id)
test_set <- data[test_id, ]

# normalize data.
train_set <- as.data.frame(lapply(train_set, normalize))
test_set <- as.data.frame(lapply(test_set, normalize))
############################  Assignment 2.2 ####################################
# Linear regression model

# apply linear regression
model <- lm(motor_UPDRS ~ ., data = train_set)

# predict the test value using the linear regression just created
test_pred <- predict(model, test_set)

# calculate test MSE
test_mse <- mean((test_pred - test_set$motor_UPDRS)^2)
model
############################  Assignment 2.3 ####################################
# Loglikelihood function
loglikelihood <- function(theta, sigma) {
  n <- length(Y)
  log_likelihood_values <- -0.5 * (log(2 * pi * sigma^2) + ((Y - theta %*% X) / sigma)^2)
```

```
  return(total_log_likelihood <- sum(log_likelihood_values))
}

# ridge
ridge <- function(param) {
  param_length <- length(param)
  theta <- param[1:param_length-2]
  sigma <- param[param_length-1]
  lambda <- param[param_length]
  loglikelihood_result <- loglikelihood(theta, sigma)
  ridge_penalty <- lambda * sum(theta^2)
  return(loglikelihood_result - ridge_penalty)
}

# ridgeopt
ridgeopt <- function(theta, sigma, lambda) {
  size_theta <- length(theta)
  param <- rep(0,size_theta+2)
  for(i in 1:size_theta){
    param[i] <- theta[i]
  }
  param[size_theta + 1] <- sigma
  param[size_theta + 2] <- lambda
  ridge_result <- optim(par = param, fn = ridge, method="BFGS")

}

# df
df <- function(X,lambda) {
  n <- nrow(X)
  # calculate hat_matrix
  hat_matrix <- X %*% solve(t(X) %*% X + lambda * diag(ncol(X)))
  # get degree of the hat_matrix
  df_ridge <- sum(diag(hat_matrix))
  return(df_ridge)
}
############################  Assignment 2.4 ####################################

#sigma <- 0.01
#theta <- rep(1,21)
#X <- train_set[]
#lambda <- 1
#ridgeopt(theta,sigma,lambda)

#lambda <- 100
#ridgeopt(theta,sigma,lambda)

#lambda <- 1000
#ridgeopt(theta,sigma,lambda)

############################  Init code For Assignment 3 ########################
rm(list = ls())
knitr::opts_chunk$set(echo = TRUE)
```

```r
library(ggplot2)
data <- read.csv("pima-indians-diabetes.csv")

age <- data[, 8]
plasma <- data[, 2]

names(data)[9] <- "diabetes"
names(data)[2] <- "plasma"
names(data)[8] <- "age"

x <- age
y <- plasma
df <- data.frame(x, y)

ggplot2::ggplot(df, ggplot2::aes(age, plasma)) +
  ggplot2::geom_point(ggplot2::aes(colour = data[, 9]))+
  ggplot2::labs(x = "Age", y="Plasma glucose")


model <- glm(diabetes ~ plasma + age, fam = binomial(link="logit"), data = data)
summary(model)

# r = 0.5

c1 <- fitted(model) >= .5

missC_error <- sum(data$diabetes != c1) / nrow(data)
cat("Misclassification error:", missC_error, "\n")

ggplot2::ggplot(df, ggplot2::aes(age, plasma)) +
  ggplot2::geom_point(ggplot2::aes(colour = c1))+
  ggplot2::labs(x = "Age", y="Plasma glucose")     # T = Diabetic, F = Non-diabetic


# r = 0.2

c2 <- fitted(model) >= .2

ggplot2::ggplot(df, ggplot2::aes(age, plasma)) +
  ggplot2::geom_point(ggplot2::aes(colour = c2))+
  ggplot2::labs(x = "Age", y="Plasma glucose")

# r = 0.8

c3 <- fitted(model) >= .8

ggplot2::ggplot(df, ggplot2::aes(age, plasma)) +
  ggplot2::geom_point(ggplot2::aes(colour = c3))+
  ggplot2::labs(x = "Age", y="Plasma glucose")


# Adding new variables, logit function with r= 0.5
```

```r
z1 <- data$plasma^4
z2 <- data$plasma^3 * data$age
z3 <- data$plasma^2 * data$age^2
z4 <- data$plasma^1 * data$age^3
z5 <- data$age^4

newdata <- cbind(data, z1, z2, z3, z4, z5)

newmodel <- glm(diabetes ~ plasma + age + z1 + z2 + z3 + z4 + z5 , family = binomial(link = "logit") , c

summary(newmodel)

c4 <- fitted(newmodel) >= .5

new_missC_error <- sum(data$diabetes != c4) / nrow(data)
cat("Misclassification error:", new_missC_error, "\n")

ggplot2::ggplot(df, ggplot2::aes(age, plasma)) +
  ggplot2::geom_point(ggplot2::aes(colour = c4))+
  ggplot2::labs(x = "Age", y="Plasma glucose")
```