Lab 11: Comparing Classifiers

Donggyun Kim
27008257

4/11/2018

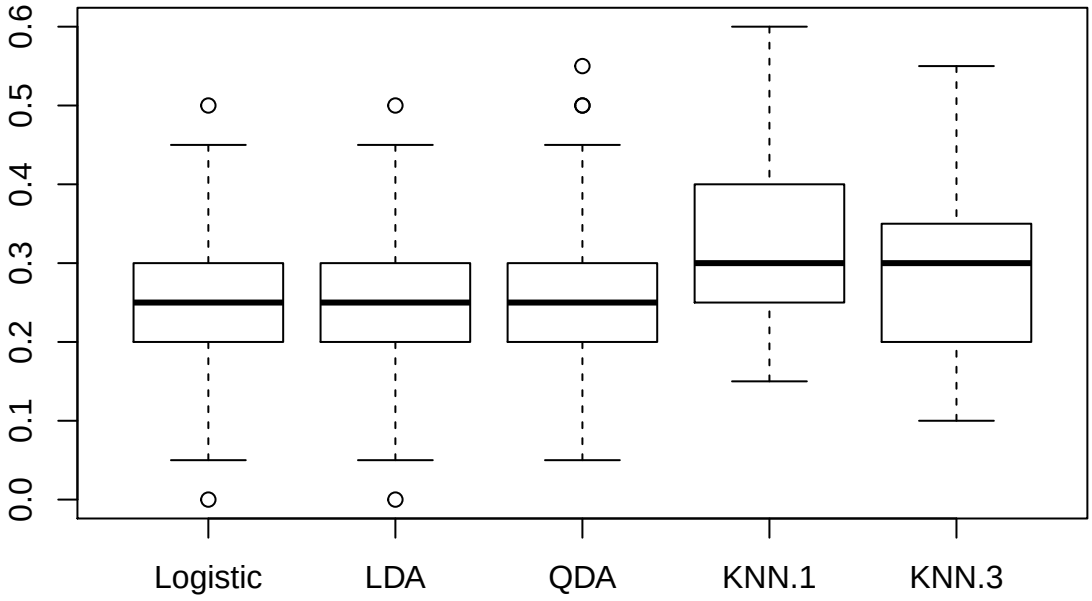
```
library(MASS)
library(mvtnorm)
library(ggplot2)
library(caret)
library(e1071)
library(class)


expit <- function(x) {
  exp(x) / (1 + exp(x))
}

gen_datasets <- function() {
  id <- diag(c(1,1))
  df1 <- data.frame(y = factor(rep(c(0, 1), each = 50)),
                    rbind(rmvnorm(50, mean = c(0, 0), sigma = id),
                          rmvnorm(50, mean = c(1, 1), sigma = id)))

  covmat <- matrix(c(1, -1/2, -1/2, 1), nrow = 2)
  df2 <- data.frame(y = factor(rep(c(0, 1), each = 50)),
                    rbind(rmvnorm(50, mean = c(0, 0), sigma = covmat),
                          rmvnorm(50, mean = c(1, 1), sigma = covmat)))

  mu <- c(0, 0)
  nu <- 4
  sigma <- matrix(c(1, 1/2, 1/2, 1), nrow = 2)
  n <- 50
  x_first <- t(t(mvrnorm(n, rep(0, length(mu)), sigma) *
                   sqrt(nu / rchisq(n, nu))) + mu)
  mu <- c(1, 1)
  x_second <- t(t(mvrnorm(n, rep(0, length(mu)), sigma) *
                    sqrt(nu / rchisq(n, nu))) + mu)
  df3 <- data.frame(y = factor(rep(c(0, 1), each = 50)),
                    rbind(x_first, x_second))

  covmat2 <- matrix(c(1, 1/2, 1/2, 1), nrow = 2)
  df4 <- data.frame(y = factor(rep(c(0, 1), each = 50)),
                    rbind(rmvnorm(50, mean = c(0, 0), sigma = covmat2),
                          rmvnorm(50, mean = c(1, 1), sigma = covmat)))

  x <- matrix(rnorm(200), ncol = 2)
  df5_temp <- data.frame(x^2, x[, 1] *x[, 2])
  beta <- c(0, 2, -1, -2)
  y <- apply(df5_temp, 1, function(row) {
    p <- expit(sum(c(1, row) * beta))
    sample(x = c(0, 1), size = 1, prob = c(1 - p, p))
```

```
  })
  df5 <- data.frame(y = factor(y), x)

  x <- matrix(rnorm(200), ncol = 2)
  y <- 1 * (x[, 1]^2 + x[, 2]^2 > qchisq(p=0.5, df=2))
  df6 <- data.frame(y = factor(y), x)
  list(df1, df2, df3, df4, df5, df6)
}
I <- 5  # number of models
J <- 6  # number of scenarios
K <- 100  # number of iterations
arr <- array(0, c(I, J, K))
rownames(arr) <- c("Logistic", "LDA", "QDA", "KNN-1", "KNN-3")
colnames(arr) <- paste("SCENARIO", 1:6, sep = " ")

for (k in 1:K) {
  SRs <- gen_datasets()
  index <- sample(nrow(SRs[[1]]), size = 0.8 * nrow(SRs[[1]]))  # training index
  for (j in 1:J) {
    dat <- SRs[[j]]
    obs <- dat$y[-index]  # test observation y
    for (i in 1:I) {
      if (i == 1) {
        glm_obj <- glm(y ~ ., family = binomial, data = dat, subset = index)
        glm_prob <- predict(glm_obj, newdata = dat[-index, ], type = "response")
        glm_pred <- numeric(length(glm_prob))
        glm_pred[glm_prob >= 0.5] <- 1
        tbl <- table(obs, glm_pred)
        arr[i, j, k] <- 1 - sum(diag(tbl)) / sum(tbl)  # test error rate
      } else if (i == 2) {
        lda_obj <- lda(y ~ ., data = dat, subset = index)
        lda_prob <- predict(lda_obj, newdata = dat[-index, ])$posterior
        lda_pred <- numeric(nrow(lda_prob))
        lda_pred[lda_prob[, 1] <= lda_prob[, 2]] <- 1
        tbl <- table(obs, lda_pred)
        arr[i, j, k] <- 1 - sum(diag(tbl)) / sum(tbl)
      } else if (i == 3) {
        qda_obj <- qda(y ~ ., data = dat, subset = index)
        qda_prob <- predict(qda_obj, newdata = dat[-index, ])$posterior
        qda_pred <- numeric(nrow(qda_prob))
        qda_pred[qda_prob[, 1] <= qda_prob[, 2]] <- 1
        tbl <- table(obs, qda_pred)
        arr[i, j, k] <- 1 - sum(diag(tbl)) / sum(tbl)
      } else if (i == 4) {
        knn_obj1 <- knn(train = dat[index, -1], test = dat[-index, -1],
                        cl = dat[index, 1])
        tbl <- table(obs, knn_obj1)
        arr[i, j, k] <- 1 - sum(diag(tbl)) / sum(tbl)
      } else {
        knn_obj2 <- knn(train = dat[index, -1], test = dat[-index, -1],
                        cl = dat[index, 1], k = 3)
        tbl <- table(obs, knn_obj2)
        arr[i, j, k] <- 1 - sum(diag(tbl)) / sum(tbl)
```

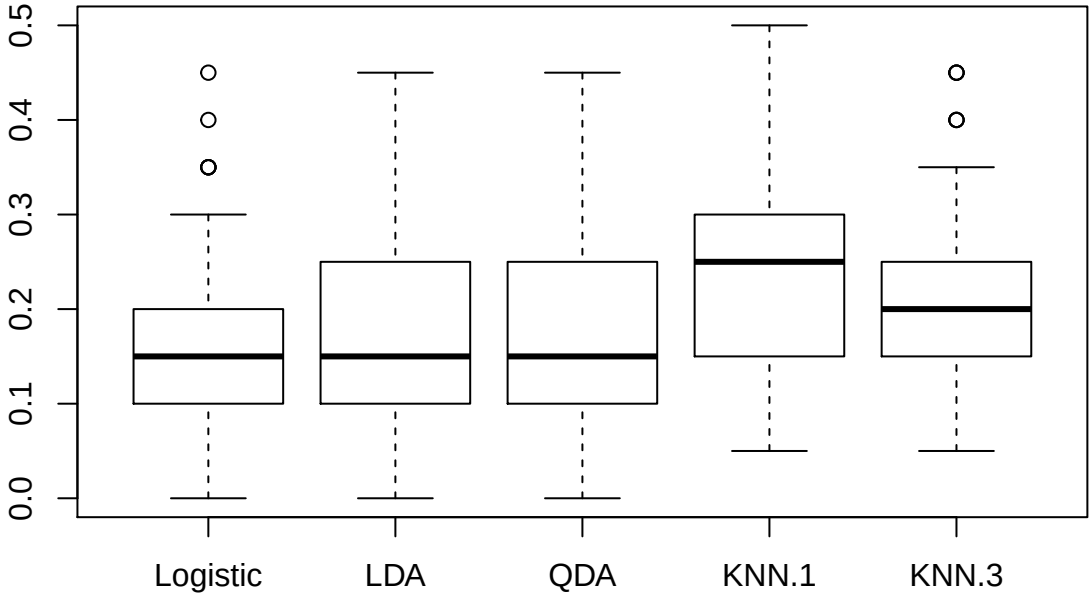
```
      }
    }
  }
}

arr[, , 1:3]
```

```
## , , 1
##
##          SCENARIO 1 SCENARIO 2 SCENARIO 3 SCENARIO 4 SCENARIO 5 SCENARIO 6
## Logistic       0.25       0.15       0.35       0.05       0.25       0.60
## LDA            0.20       0.15       0.35       0.10       0.25       0.55
## QDA            0.20       0.15       0.30       0.20       0.20       0.05
## KNN-1          0.20       0.15       0.30       0.20       0.30       0.25
## KNN-3          0.15       0.20       0.40       0.10       0.20       0.15
##
## , , 2
##
##          SCENARIO 1 SCENARIO 2 SCENARIO 3 SCENARIO 4 SCENARIO 5 SCENARIO 6
## Logistic       0.25       0.15       0.25       0.25        0.3       0.15
## LDA            0.20       0.20       0.25       0.15        0.3       0.15
## QDA            0.25       0.15       0.60       0.15        0.2       0.10
## KNN-1          0.45       0.20       0.35       0.20        0.2       0.00
## KNN-3          0.20       0.15       0.15       0.20        0.2       0.05
##
## , , 3
##
##          SCENARIO 1 SCENARIO 2 SCENARIO 3 SCENARIO 4 SCENARIO 5 SCENARIO 6
## Logistic       0.25       0.15       0.40       0.35       0.60       0.20
## LDA            0.20       0.15       0.40       0.35       0.60       0.20
## QDA            0.25       0.15       0.35       0.40       0.20       0.10
## KNN-1          0.15       0.25       0.40       0.40       0.25       0.10
## KNN-3          0.20       0.15       0.30       0.40       0.10       0.05
```

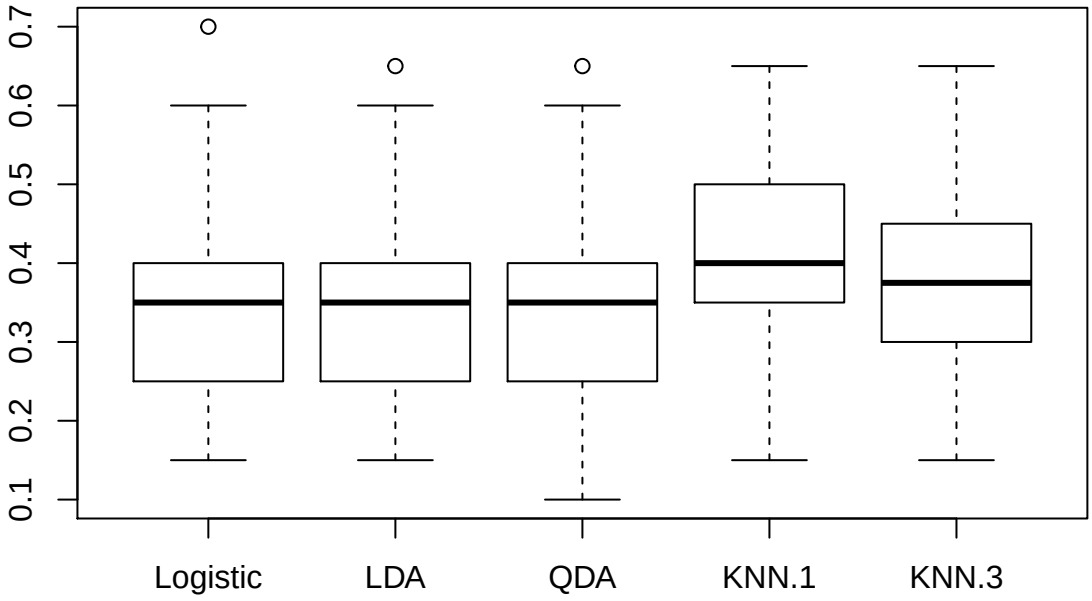
```
vrbl_names <- paste0("SR", 1:J)
for (j in 1:J) {
  assign(vrbl_names[j], data.frame(t(arr[, j, ])))
}

# boxplots of each scenario
boxplot(SR1)
```



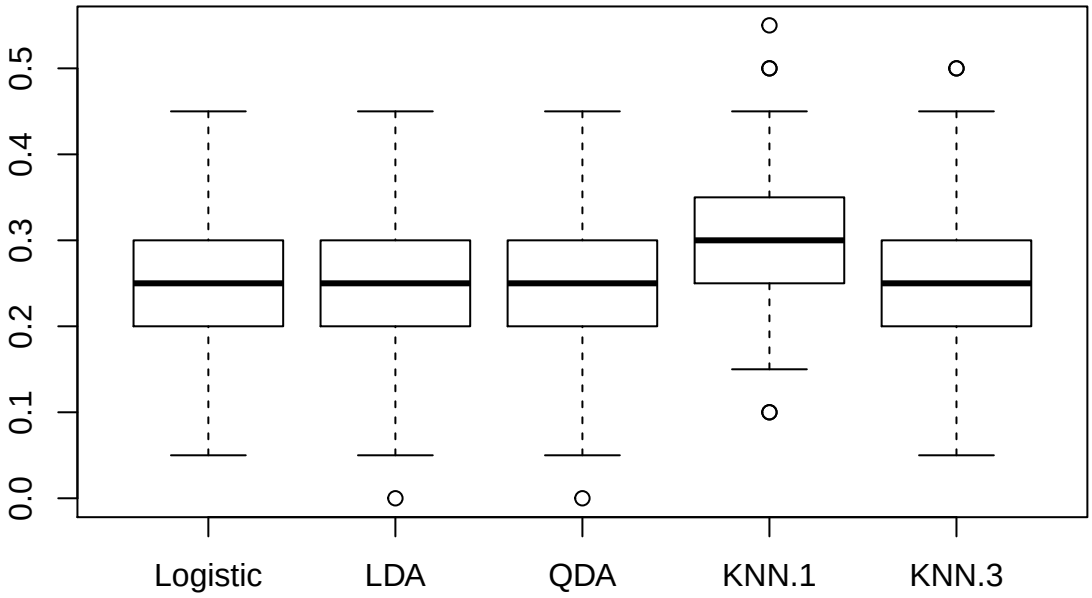

```
boxplot(SR2)
```



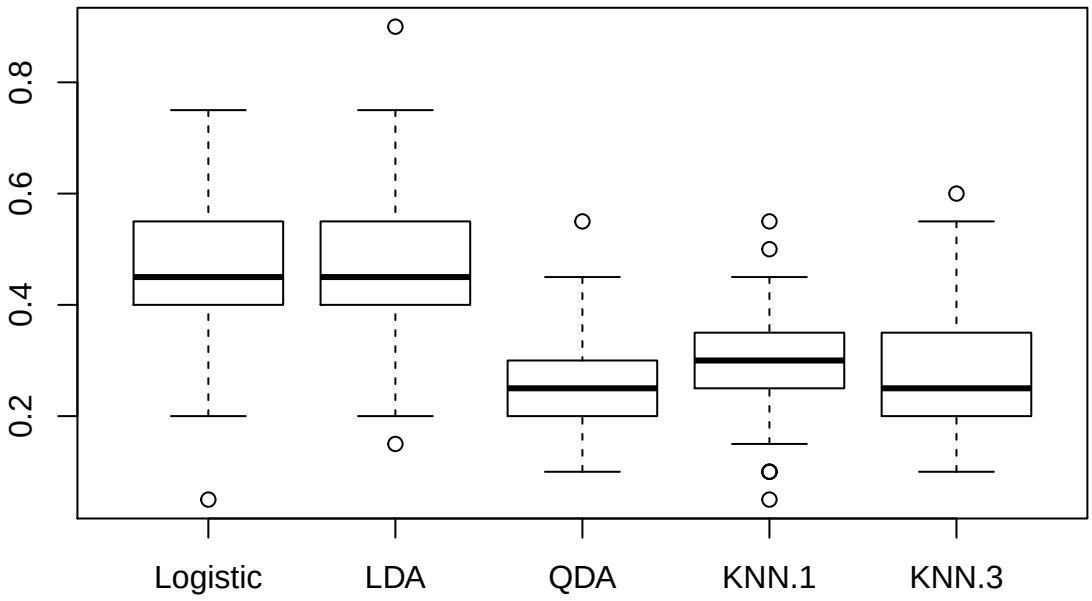

```
boxplot(SR3)
```



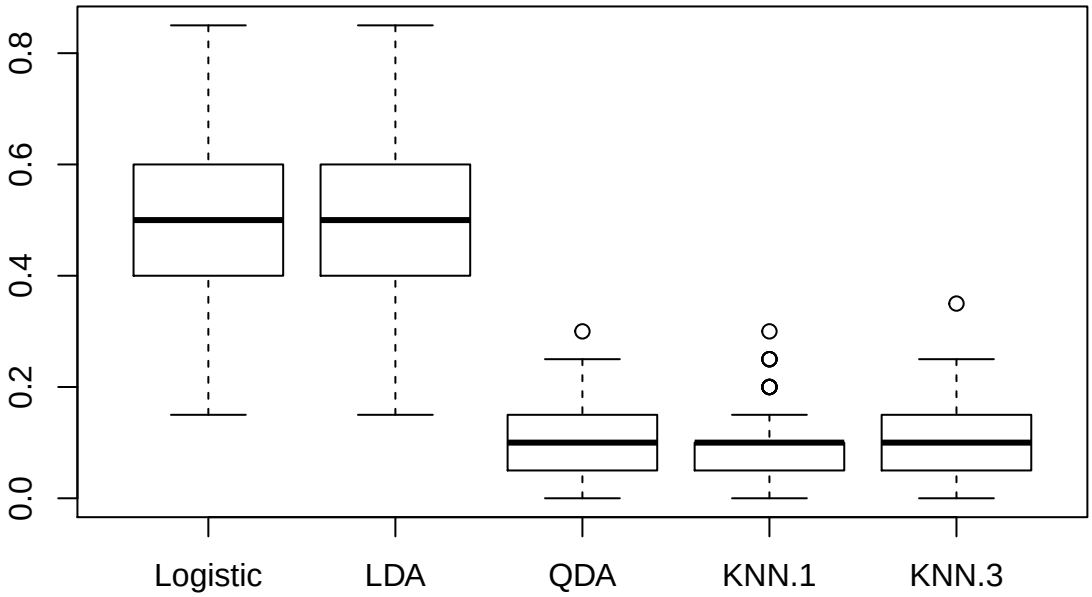

```
boxplot(SR4)
```




```
boxplot(SR5)
```


0.2
0.4
0.6
0.8
Logistic
LDA
QDA
KNN.1
KNN.3

```
boxplot(SR6)
```


0.0
0.2
0.4
0.6
0.8
Logistic
LDA
QDA
KNN.1
KNN.3