Lab 11: Comparing Classifiers

Stat 154, Spring 2018

Introduction

In this lab we will compare the classification algorithms we have studied so far (logistic regression, LDA, QDA, k-NN) based on six different synthetic datasets.

We will follow closely Chapter 4.5 in *ISL*. The objective is to compare the predictive power for various classification algorithms we have studied so far under various hypothetical scenarios. See *ISL* for a more detailed discussion.

You may need to use the following packages:

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

Data Simulation

Simulate six datasets (each has $p = 2$ predictors):

Scenario 1: Simulate 100 observations, half are from $N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\right)$, half from $N\left(\begin{bmatrix}1\\1\end{bmatrix}, \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\right)$. Treat the first half as class 1 and the rest as class 2.

Scenario 2: Simulate 100 observations, half are from $N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & -0.5\\-0.5 & 1\end{bmatrix}\right)$, half from $N\left(\begin{bmatrix}1\\1\end{bmatrix}, \begin{bmatrix}1 & -0.5\\-0.5 & 1\end{bmatrix}\right)$. Treat the first half as class 1 and the rest as class 2.

Scenario 3: Simulate 100 observations, half are from $t_4\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & 0.5\\0.5 & 1\end{bmatrix}\right)$, half from $t_4\left(\begin{bmatrix}1\\1\end{bmatrix}, \begin{bmatrix}1 & 0.5\\0.5 & 1\end{bmatrix}\right)$. Treat the first half as class 1 and the rest as class 2.

Scenario 4: Simulate 100 observations, half are from $N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & 0.5\\0.5 & 1\end{bmatrix}\right)$, half from $N\left(\begin{bmatrix}1\\1\end{bmatrix}, \begin{bmatrix}1 & -0.5\\-0.5 & 1\end{bmatrix}\right)$. Treat the first half as class 1 and the rest as class 2.

Scenario 5: Simulate two independent sequences of $N(0,1)$ random variables $(X_{1,1}, X_{2,1}, ...X_{100,1})$ and $(X_{1,2}, X_{2,2}, ...X_{100,2})$. Let $Y_i \sim \text{Ber}(p_i)$ for $i = 1, ..., 100$, where

$$\text{logit}(p_i) = \beta_0 + \beta_1 X_{i,1}^2 + \beta_2 X_{i,2}^2 + \beta_3 X_{i,1} X_{i,2},$$

with $(\beta_0, \beta_1, \beta_2, \beta_3) = (0, 2, -1, 2)$.

Scenario 6: Simulate two independent sequences of $N(0,1)$ random variables $(X_{1,1}, X_{2,1}, ...X_{n,1})$ and $(X_{1,2}, X_{2,2}, ...X_{n,2})$. Let

$$Y_i = \begin{cases} 1 & \text{if } X_{i,1}^2 + X_{i,2}^2 > \chi_2^2(0.5) \approx 1.386 \\ 0 & \text{otherwise.} \end{cases},$$

where $\chi_2^2(0.5)$ is the median of a χ^2 distribution with 2 degrees of freedom.

Remark: You can use the following function `gen_datasets` to simulate the six datasets. The code for drawing from a non-central multivariate t-distribution is from https://stats.stackexchange.com/questions/68476/drawing-from-the-multivariate-students-t-distribution.

```
set.seed(100)

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, -0.5, -0.5, 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); sigma <- matrix(c(1, 1/2, 1/2, 1), 2); nu <- 4
  n <- 50 # Number of draws
```

```r
  x_first <- t(t(mvrnorm(n, rep(0, length(mu)), sigma)
                 * sqrt(nu / rchisq(n, nu))) + mu)
  mu <- c(1, 1); sigma <- matrix(c(1, 1/2, 1/2, 1), 2); nu <- 4
  n <- 50 # Number of draws
  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, 0.5, 0.5, 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)
}
```

Repeat the following 100 times:

- Simulate 6 datasets via `gen_datasets()`.
- For each dataset, use 80% of the data as the training set and the remaining 20% as the test set.
- Fit logistic regression, LDA, QDA, k-NN with 1 neighbor, and k-NN-CV using the training set.
  - logistic regression: `glm()` from `"stats"`
  - LDA: `lda()` from `"MASS"`
  - QDA: `qda()` from `"MASS"`
  - k-NN: `knn()` from `"class"`

- For each model, compute the test error rate and generate predictions on the test set.
- Store the 5×6 matrix of error rates.
- You should now have a $5 \times 6 \times 100$ array of test error rates.
- For each of the six scenario, make a boxplot of test error rates. See Figure 4.10 and Figure 4.11 in ISL (screenshots below).

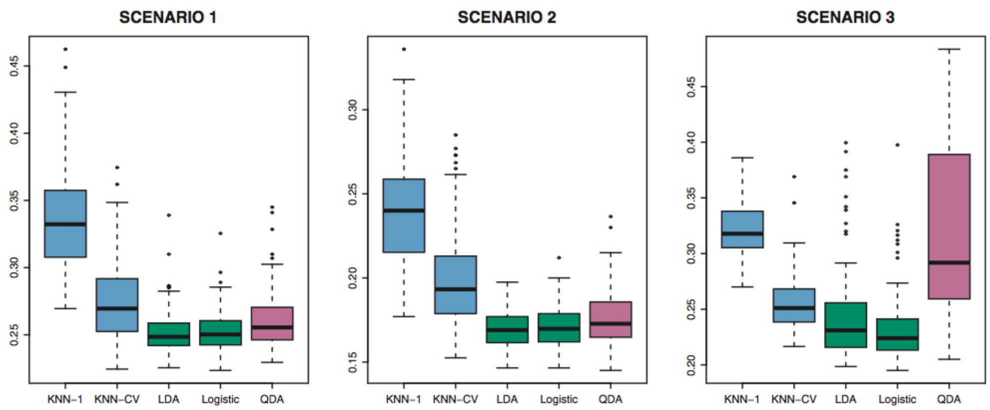

FIGURE 4.10. *Boxplots of the test error rates for each of the linear scenarios described in the main text.*

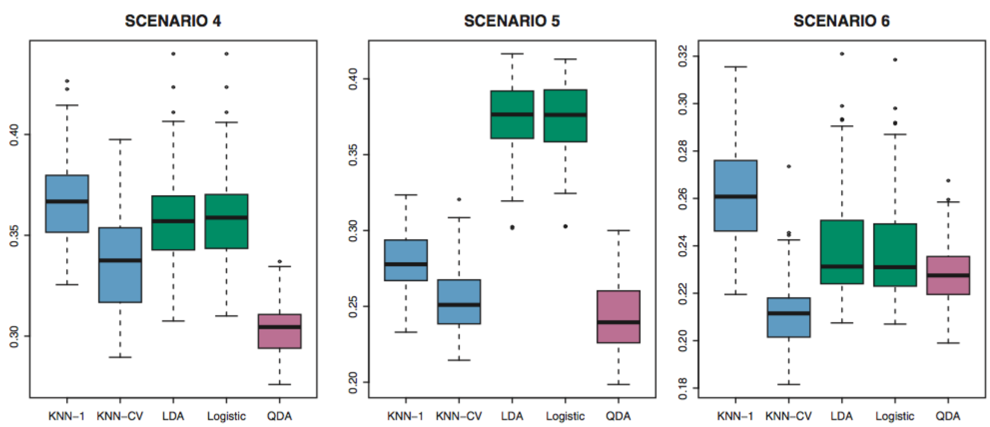

FIGURE 4.11. *Boxplots of the test error rates for each of the non-linear scenarios described in the main text.*