HW 2

Fan Bi (fb2234)

We will be predicting whether the housing price is expensive or not using the `sahp` dataset in the **r02pro** package.

You can run the following code to prepare the analysis.

```
library(r02pro)     #INSTALL IF NECESSARY
library(tidyverse)  #INSTALL IF NECESSARY
library(MASS)
my_sahp <- sahp %>%
  na.omit() %>%
  mutate(expensive = sale_price > median(sale_price)) %>%
  dplyr::select(gar_car, liv_area, oa_qual, expensive)
my_sahp_train <- my_sahp[1:100, ]
my_sahp_test <- my_sahp[-(1:100), ]
```

Please answer the following questions.

1. a. Using the training data `my_sahp_train` to fit a logistic regression model of `expensive` on each variable (`gar_car`, `liv_area`, `oa_qual`) separately. For each logistic regression, compute the training and test error. Which variable leads to the smallest training error? Which variable leads to the smallest test error?

Answer:

```
model <- paste0("md_",c(substring(names(my_sahp),1,3)[1:2],substring(names(my_sahp),1,2)[3]))
fm <- paste0("expensive ~ ",names(my_sahp)[1:3])
train_err <- rep(0,3)
names(train_err) <- model
test_err <- rep(0,3)
names(test_err) <- model

for(i in 1:3){
  assign(model[i],glm(as.formula(fm[i]),family = "binomial",data = my_sahp_train))
  pred_train <- predict(get(model[i]),my_sahp_train,type = "response") >= 0.5
  pred_test <- predict(get(model[i]),my_sahp_test,type = "response") >= 0.5
  train_err[i] <- mean(pred_train != my_sahp_train$expensive)
  test_err[i] <- mean(pred_test != my_sahp_test$expensive)
}

train_err
```

```
## md_gar md_liv  md_oa
##   0.29   0.35   0.23
```

```
test_err
```

```
##    md_gar    md_liv     md_oa
## 0.1774194 0.2580645 0.3064516
```

`oa_qual` leads to the smallest training error. By calculating the odds ratio we can learn this model classifies buildings with `oa_qual` not smaller than 7 as expensive.

`gar_car` leads to the smallest test error. By calculating the odds ratio we can learn this model classifies buildings with `gar_car` not smaller than 2 as expensive.

b. Using the training data `my_sahp_train` to fit a logistic regression model of `expensive` on all three variables (`gar_car`, `liv_area`, `oa_qual`). Compute the training and test error. How do the result compare with part a.

Answer:

```
md_all <- glm(expensive~.,family = "binomial",data = my_sahp_train)
model <- c(model,"md_all")
train_err['md_all'] <- mean(I(predict(md_all,my_sahp_train,type = "response") >=0.5) !=
                            my_sahp_train$expensive)
test_err['md_all'] <- mean(I(predict(md_all,my_sahp_test,type = "response") >=0.5) !=
                             my_sahp_test$expensive)
train_err
```

```
## md_gar md_liv  md_oa md_all
##   0.29   0.35   0.23   0.21
```

```
test_err
```

```
##    md_gar    md_liv     md_oa    md_all
## 0.1774194 0.2580645 0.3064516 0.1612903
```

`md_all` that uses all variables as predictors has the smaller training error and test error than all models fitting in part a.

2. Using the training data `my_sahp_train` to fit LDA and QDA models of `expensive` on all three variables (`gar_car`, `liv_area`, `oa_qual`). Compute the training and test error. How do the results compare with Q1?

Answer:

```
md_lda <- lda(expensive~.,data = my_sahp_train)
md_qda <- qda(expensive~.,data = my_sahp_train)
model <- c(model,"md_lda","md_qda")
train_err['md_lda'] <- mean(predict(md_lda,my_sahp_train)$class !=
                              my_sahp_train$expensive)
test_err['md_lda'] <- mean(predict(md_lda,my_sahp_test)$class !=
                           my_sahp_test$expensive)
train_err['md_qda'] <- mean(predict(md_qda,my_sahp_train)$class !=
```

```
                                my_sahp_train$expensive)
test_err['md_qda'] <- mean(predict(md_qda,my_sahp_test)$class !=
                               my_sahp_test$expensive)
data.frame(model = model,
           training_error = unname(train_err),
           test_error = unname(test_err))
```

```
##    model training_error test_error
## 1 md_gar           0.29  0.1774194
## 2 md_liv           0.35  0.2580645
## 3  md_oa           0.23  0.3064516
## 4 md_all           0.21  0.1612903
## 5 md_lda           0.17  0.2419355
## 6 md_qda           0.16  0.1774194
```

LDA and QDA model both give smaller training error than models in Q1. Checking the column of test error, we can find LDA model does not do well in predicting test data when QDA model performs much better with the same accuracy as `md_gar` in Q1.

3. Q3 in Chapter 4 of ISLRv2.

Answer:

Recall the LDA assumption : $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2 = \sigma^2$,so

$$\begin{aligned} p_k(x) &\propto log(p_k(x)) \\ &\propto log(\pi_k \frac{1}{\sqrt{2\pi}\sigma_k} exp\{-\frac{1}{2\sigma_k^2}(x-\mu_k)^2\}) \\ &\propto log(\pi_k \frac{1}{\sqrt{2\pi}\sigma} exp\{-\frac{1}{2\sigma^2}(x-\mu_k)^2\}) \\ &\propto log(\pi_k) + log(\frac{1}{\sqrt{2\pi}\sigma}) - \frac{1}{2\sigma^2}(x-\mu_k)^2 \end{aligned}$$

Then we discard terms not corresponding to k :

$$\begin{aligned} p_k(x) &\propto log(\pi_k) - \frac{1}{2\sigma^2}(-2\mu_k x + \mu_k^2) \\ &\propto \frac{\mu_k}{\sigma^2}x - \frac{\mu_k^2}{2\sigma^2} + log(\pi_k) = \delta_k(x) \end{aligned}$$

So we can say Bayes classifier of LDA is linear :

$$\hat{\delta}_k(x) = \hat{b}_0 + \hat{b}_1 x$$

where

$$\hat{b}_0 = -\frac{\hat{\mu}_k^2}{2\hat{\sigma}^2} + log(\hat{\pi}_k), \hat{b}_1 = \frac{\hat{\mu}_k}{\hat{\sigma}^2}$$

Then we will prove the Bayes classifier of one-dimensional QDA is quadratic :

$$\begin{aligned} p_k(x) &\propto log(p_k(x)) \\ &\propto log(\pi_k \frac{1}{\sqrt{2\pi}\sigma_k} exp\{-\frac{1}{2\sigma_k^2}(x-\mu_k)^2\}) \\ &\propto log(\pi_k) + log(\frac{1}{\sqrt{2\pi}\sigma_k}) - \frac{1}{2\sigma_k^2}(x-\mu_k)^2 \end{aligned}$$

After discarding terms not corresponding to k :

$$\begin{aligned} p_k(x) &\propto log(\frac{\pi_k}{\sigma_k}) - \frac{1}{2\sigma_k^2}(x^2 - 2\mu_k x + \mu_k^2) \\ &\propto -\frac{1}{2\sigma_k^2}x^2 + \frac{\mu_k}{\sigma_k^2}x + log(\frac{\pi_k}{\sigma_k}) - \frac{\mu_k^2}{2\sigma_k^2} \end{aligned}$$

Therefore here the Bayes classifier is quadratic :

$$\hat{\delta}_k(x) = \hat{b}_0 + \hat{b}_1 x + \hat{b}_2 x^2$$

where

$$\begin{aligned} \hat{b}_0 &= log(\frac{\hat{\pi}_k}{\hat{\sigma}_k}) - \frac{\hat{\mu}_k^2}{2\hat{\sigma}_k^2} \\ \hat{b}_1 &= \frac{\hat{\mu}_k}{\hat{\sigma}_k^2} \\ \hat{b}_2 &= -\frac{1}{2\hat{\sigma}_k^2} \end{aligned}$$

4. Q6 in Chapter 4 of ISLRv2.

Answer:

a) Denote this student as $X_1 = (X_{11}, X_{12}) = (40, 3.5)$, and recode Y as below :

$$Y = \begin{cases} 1 & \text{revieve an A} \\ 0 & \text{otherwise} \end{cases}$$

For logistic regression we estimate :

$$\begin{aligned} log\left(\frac{P(Y_1 = 1|X_1)}{P(Y_1 = 0|X_1)}\right) &= \hat{\beta}_0 + \hat{\beta}_1 X_{11} + \hat{\beta}_2 X_{12} = -0.5 \\ P(Y_1 = 1|X_1) &= \frac{e^{-0.5}}{1 + e^{-0.5}} = 0.3775 \end{aligned}$$

So the probability that this student receives an A is 0.3775.

b) Denote the hour this student need to study as X_{11}^*. If $P(Y_1^* = 1|X_1^*) = 0.5$,

$$\begin{aligned} log\left(\frac{P(Y_1^* = 1|X_1^*)}{P(Y_1^* = 0|X_1^*)}\right) = log\left(\frac{0.5}{1 - 0.5}\right) = 0 &= \hat{\beta}_0 + \hat{\beta}_1 X_{11}^* + \hat{\beta}_2 X_{12} \\ \implies X_{11}^* &= 50 \end{aligned}$$

So this student should study 50 hours to have a 50% chance of getting an A in class.