# Stat 432 HW 03

Name: Ahmadreza Eslaminia, netID: ae15

Summer 2024

Include the R code for this HW.

```
knitr::opts_chunk$set(echo = TRUE)
library(ISLR2)
```

```
## Warning: package 'ISLR2' was built under R version 4.3.3
```

```
library(GGally)
```

```
## Warning: package 'GGally' was built under R version 4.3.3
```

```
## Warning: package 'ggplot2' was built under R version 4.3.3
```

```
library(tibble)
```

```
## Warning: package 'tibble' was built under R version 4.3.3
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.3.3
```

```
library(knitr)
```

```
## Warning: package 'knitr' was built under R version 4.3.3
```

```
library(kableExtra)
```

```
## Warning: package 'kableExtra' was built under R version 4.3.3
```

## Question 1 (Classification General)

Consider a categorical response $Y$ which takes possible values 0 and 1 as well as a single numerical predictor $X$. Recall that $p(x) = P(Y = 1|X = x)$.

Consider the model

$$log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x$$

and estimated coefficients

- $\hat{\beta}_0 = 2$
- $\hat{\beta}_1 = 1$

(a) Provide a classification when $x = -1$ that attempts to minimize the classification error.

First, we calculte the logodd for $x = -1$:

$$\log\left(\frac{p(-1)}{1-p(-1)}\right) = 2 + 1(-1) = 1$$

Next, we convert the log-odds to probabiity:

$$\frac{p(-1)}{1-p(-1)} = e^1 = e$$

$$p(-1) = \frac{e}{1+e} \approx 0.73$$

Since $p(-1) > 0.5$, we clasify $x = -1$ as 1.

(b) Calculate an estimate of $P(Y = 1|X = -1)$. Now, let's calculate the estimate of $P(Y = 1|X = -1)$.

Using the probability we calculatd in part (a):

$$P(Y = 1|X = -1) \approx 0.73$$

(c) Find a value $c$ that splits $x$ into regions classified as 1 and 0. let's find a value $c$ that splits $x$ into regions classified as 1 and 0.

The decision boundary occurs where $p(x) = 0.5$:

$$\log\left(\frac{0.5}{1-0.5}\right) = \beta_0 + \beta_1 c$$

$$\log(1) = 2 + 1 \cdot c$$

$$0 = 2 + c$$

$$c = -2$$

So, $x$ values less than -2 are classifid as 0, and $x$ values greater than or equal to -2 are classified as 1.

## Question 2 (Logistic Regression)

We will use data found in (wisc-trn.csv) and (wisc-tst.csv) [check the box folder for files] which contain train and test data respectively.

You should consider coercing the response (`class` variable) to be a factor variable.

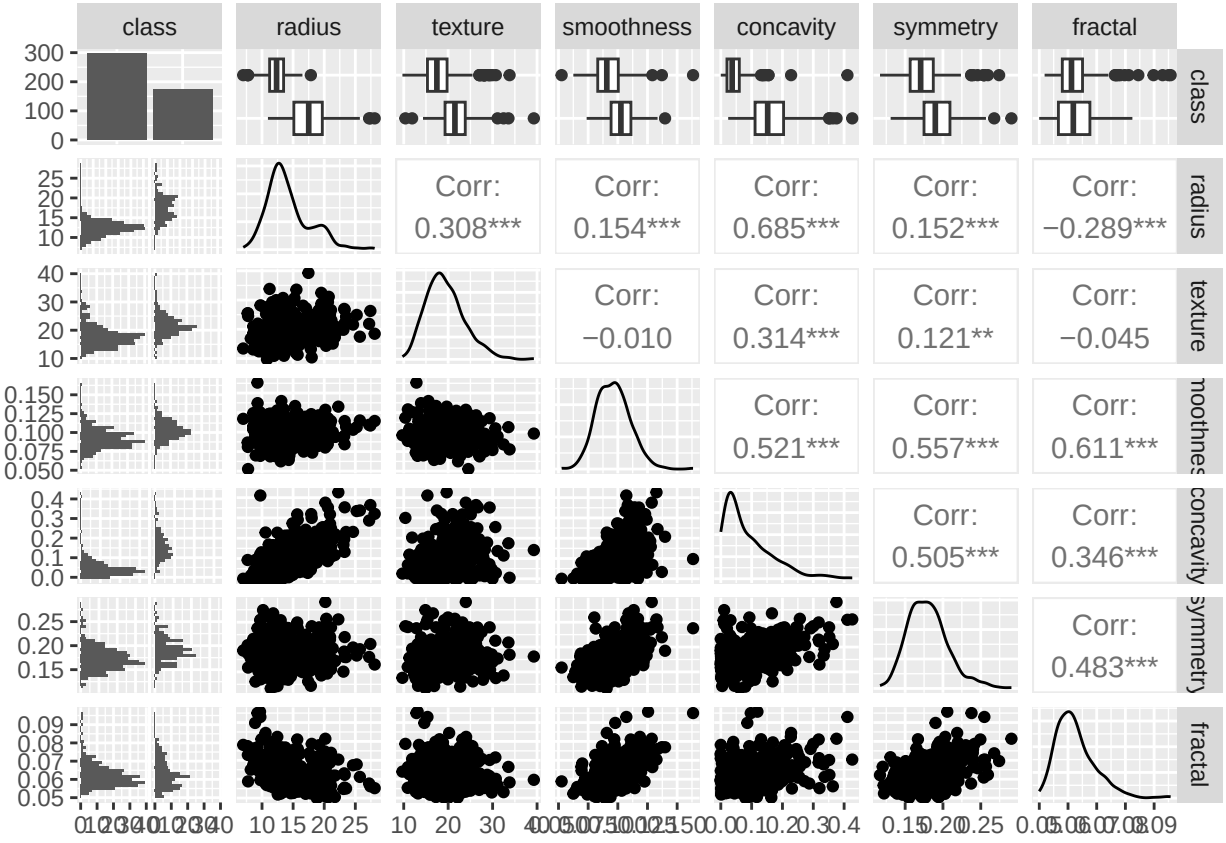
(a) Create plots for EDA analysis with the variables: "class", radius", "texture", "smoothness", "concavity", "symmetry", and "fractal".

```
train_data <- read.csv("wisc-trn.csv")
test_data <- read.csv("wisc-tst.csv")

train_data$class <- as.factor(train_data$class)
test_data$class <- as.factor(test_data$class)

ggpairs(train_data[, c("class" ,"radius", "texture", "smoothness", "concavity", "symmetry", "fractal")]
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




(b) Our task is to conduct classification on `class`. Consider an additive logistic regression that uses "radius", "texture", "smoothness", "concavity", "symmetry", and "fractal" as predictor variables. Use this model to estimate

$$p(x) = P(Y = \texttt{M} \mid X = x).$$

Report test sensitivity, test specificity, and test accuracy for the Bayes Classifier. We will consider `M` (malignant) to be the "positive" class when calculating sensitivity and specificity. Summarize these results.

```
# reg model
logistic_model <- glm(class ~ radius + texture + smoothness + concavity + symmetry + fractal,
                      data = train_data, family = binomial)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(logistic_model)
```

```
##
## Call:
## glm(formula = class ~ radius + texture + smoothness + concavity +
##     symmetry + fractal, family = binomial, data = train_data)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -34.39998    7.39981  -4.649 3.34e-06 ***
## radius         1.16469    0.22663   5.139 2.76e-07 ***
## texture        0.41232    0.07507   5.493 3.96e-08 ***
## smoothness   132.50454   31.52855   4.203 2.64e-05 ***
## concavity     35.56111    8.11476   4.382 1.17e-05 ***
## symmetry      19.92233   12.70048   1.569   0.1167
## fractal     -167.94613   76.94789  -2.183   0.0291 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 616.45  on 468  degrees of freedom
## Residual deviance: 112.54  on 462  degrees of freedom
## AIC: 126.54
##
## Number of Fisher Scoring iterations: 8
```

```
test_probs <- predict(logistic_model, newdata = test_data, type = "response")

#predict
test_pred_class <- ifelse(test_probs > 0.5, "M", "B")


conf_matrix <- table(Predicted = test_pred_class, Actual = test_data$class)

# Calculate metric
sensitivity <- conf_matrix["M", "M"] / sum(conf_matrix[, "M"])
specificity <- conf_matrix["B", "B"] / sum(conf_matrix[, "B"])
accuracy <- sum(diag(conf_matrix)) / sum(conf_matrix)


print(paste("Test Sensitivity:", sensitivity))
```

```
## [1] "Test Sensitivity: 0.825"
```

```
print(paste("Test Specificity:", specificity))
```

```
## [1] "Test Specificity: 0.966666666666667"
```

```
print(paste("Test Accuracy:", accuracy))
```

```
## [1] "Test Accuracy: 0.91"
```

So, based on the logistic regression model results, the sensitivity is like 0.825, which means the model correctly figured out 82.5% of the malignant cases (M) in the test set. The specificity is 0.9667, showing that the model correctly picked out 96.67% of the benin cases (B) in the test set. The accuracy of the model is 0.91, meaning it correctly classified 91% of the cases in the test set. Overall, the model does pretty well in terms of sensitivity, specificity, and accuracy, making it a pretty reliable classifier for telling apart malignant and benign cases in the given dataset.

(c) Repeat part (b) with two different classifiers, each using a different cutoff for predicted probability:

$$\hat{C}(x) = \begin{cases} M & \hat{p}(x) > c \\ B & \hat{p}(x) \leq c \end{cases}$$

- $c = 0.3$
- $c = 0.8$

```
# Function to cal performnce
calculate_metrics <- function(cutoff) {
  test_pred_class <- ifelse(test_probs > cutoff, "M", "B")
  conf_matrix <- table(Predicted = test_pred_class, Actual = test_data$class)
  sensitivity <- conf_matrix["M", "M"] / sum(conf_matrix[, "M"])
  specificity <- conf_matrix["B", "B"] / sum(conf_matrix[, "B"])
  accuracy <- sum(diag(conf_matrix)) / sum(conf_matrix)

  return(c(sensitivity, specificity, accuracy))
}

#  0.3
metrics_0.3 <- calculate_metrics(0.3)
print(paste("Cutoff 0.3 - Sensitivity:", metrics_0.3[1],
            "Specificity:", metrics_0.3[2], "Accuracy:", metrics_0.3[3]))
```

```
## [1] "Cutoff 0.3 - Sensitivity: 0.85 Specificity: 0.916666666666667 Accuracy: 0.89"
```

```
#  0.8
metrics_0.8 <- calculate_metrics(0.8)
print(paste("Cutoff 0.8 - Sensitivity:", metrics_0.8[1],
            "Specificity:", metrics_0.8[2], "Accuracy:", metrics_0.8[3]))
```

```
## [1] "Cutoff 0.8 - Sensitivity: 0.8 Specificity: 0.983333333333333 Accuracy: 0.91"
```

(d) Of the three classifiers proposed in (b) and (c), which one do you prefer? For the cancer study, we want to chose the model labels cancer as cancer, and we value this over labeling non-cancer as non-cancer. Which metric should you use?

Based on the results from part (b) and (c), the classifier with Cutoff 0.3 is prefered for the cancer study becuse it has the highest sensitivity of 0.850. This means it correctly identifies 85% of malignat cases, which is crusial for ensuring that canceer cases are not missed. Sensitivity is the most important metric to use here, as we prioritize identifying cancer cases over correctly labling non-cancer cass.

## Question 3 (LDA)

Consider the following estimates and information from data for a three-class classifications boundary.

| Class A | Class B | Class C |
|---|---|---|
| $\hat{\mu}_A = 5$ | $\hat{\mu}_B = 7$ | $\hat{\mu}_C = 10$ |
| $n_A = 100$ | $n_B = 40$ | $n_C = 60$ |

$$\hat{\sigma} = 1$$

(a) Use LDA to estimate the probability $P(Y = B|X = 7.5)$. Show your work. Calculating the prior probablities:

$$\pi_A = \frac{100}{200} = 0.5, \quad \pi_B = \frac{40}{200} = 0.2, \quad \pi_C = \frac{60}{200} = 0.3$$

Calculate the density functins:

$$\phi(7.5 \mid \mu_A) = \frac{1}{\sqrt{2\pi}} e^{-3.125}, \quad \phi(7.5 \mid \mu_B) = \frac{1}{\sqrt{2\pi}} e^{-0.125}, \quad \phi(7.5 \mid \mu_C) = \frac{1}{\sqrt{2\pi}} e^{-3.125}$$

Now, posteror probabilities:

$$P(Y = B \mid X = 7.5) = \frac{0.2e^{-0.125}}{0.5e^{-3.125} + 0.2e^{-0.125} + 0.3e^{-3.125}} \approx \frac{0.2}{0.23984} \approx 0.834$$

(b) Provide a classification when $x = 5.5$. (Use the Bayes classifier.) Show your work.

$$\phi(5.5 \mid \mu_A) = \frac{1}{\sqrt{2\pi}} e^{-0.125}, \quad \phi(5.5 \mid \mu_B) = \frac{1}{\sqrt{2\pi}} e^{-1.125}, \quad \phi(5.5 \mid \mu_C) = \frac{1}{\sqrt{2\pi}} e^{-10.125}$$

$$P(Y = A \mid X = 5.5) = \frac{0.5e^{-0.125}}{0.5e^{-0.125} + 0.2e^{-1.125} + 0.3e^{-10.125}} \approx 0.713$$

$$P(Y = B \mid X = 5.5) = \frac{0.2e^{-1.125}}{0.5e^{-0.125} + 0.2e^{-1.125} + 0.3e^{-10.125}} \approx 0.285$$

$$P(Y = C \mid X = 5.5) \approx 0$$

Classify as A.

(c) Calculate the LDA decision boundary. (Use the Bayes classifier.) Show your work.

For each pair of classes, we solve the following equation for $x$:

$$\log\left(\frac{\pi_i \phi(x \mid \mu_i)}{\pi_j \phi(x \mid \mu_j)}\right) = \log\left(\frac{\pi_i}{\pi_j}\right) + \frac{(x - \mu_i)^2 - (x - \mu_j)^2}{2\sigma^2} = 0$$

Decision boundary betwen Class A and Class B:

$$\log\left(\frac{0.5}{0.2}\right) + \frac{(x - 5)^2 - (x - 7)^2}{2} = 0$$

$$\log(2.5) + \frac{4x - 24}{2} = 0$$
$$\log(2.5) + 2x - 12 = 0$$
$$2x = 12 - \log(2.5)$$
$$x = 6 - \frac{\log(2.5)}{2}$$

Decision boundary between Class A and C:

$$\log\left(\frac{0.5}{0.3}\right) + \frac{(x-5)^2 - (x-10)^2}{2} = 0$$
$$\log\left(\frac{5}{3}\right) + \frac{10x - 75}{2} = 0$$
$$\log\left(\frac{5}{3}\right) + 5x - 37.5 = 0$$
$$5x = 37.5 - \log\left(\frac{5}{3}\right)$$
$$x = 7.5 - \frac{\log\left(\frac{5}{3}\right)}{5}$$

Decision boundry between Class B and C:

$$\log\left(\frac{0.2}{0.3}\right) + \frac{(x-7)^2 - (x-10)^2}{2} = 0$$
$$\log\left(\frac{2}{3}\right) + \frac{6x - 51}{2} = 0$$
$$\log\left(\frac{2}{3}\right) + 3x - 25.5 = 0$$
$$3x = 25.5 - \log\left(\frac{2}{3}\right)$$
$$x = 8.5 - \frac{\log\left(\frac{2}{3}\right)}{3}$$

## Question 4 (QDA)

Consider the following estimates and information from data for a three-class classifications boundary.

| Class A | Class B | Class C |
|---|---|---|
| $\hat{\mu}_A = 5$ | $\hat{\mu}_B = 7$ | $\hat{\mu}_C = 10$ |
| $n_A = 100$ | $n_B = 40$ | $n_C = 60$ |
| $\hat{\sigma}_A = 1$ | $\hat{\sigma}_B = 2$ | $\hat{\sigma}_C = 3$ |

(a) Use QDA to estimate the probability $P(Y = B|X = 7.5)$. Show your work. the posterior probability is given by:

$$P(Y = k \mid X = x) = \frac{\pi_k \cdot \phi(x \mid \mu_k, \sigma_k^2)}{\sum_{l=1}^{K} \pi_l \cdot \phi(x \mid \mu_l, \sigma_l^2)}$$

the prior probabilties:

$$\pi_A = \frac{n_A}{n_A + n_B + n_C} = \frac{100}{100 + 40 + 60} = \frac{100}{200} = 0.5$$

$$\pi_B = \frac{40}{200} = 0.2, \quad \pi_C = \frac{60}{200} = 0.3$$

Calculat the density functions:

$$\phi(7.5 \mid \mu_A, \sigma_A) = \frac{1}{\sqrt{2\pi \cdot 1}} e^{-\frac{(7.5-5)^2}{2 \cdot 1^2}} = \frac{1}{\sqrt{2\pi}} e^{-3.125}$$

$$\phi(7.5 \mid \mu_B, \sigma_B) = \frac{1}{\sqrt{2\pi \cdot 2^2}} e^{-\frac{(7.5-7)^2}{2 \cdot 2^2}} = \frac{1}{2\sqrt{2\pi}} e^{-0.03125}$$

$$\phi(7.5 \mid \mu_C, \sigma_C) = \frac{1}{\sqrt{2\pi \cdot 3^2}} e^{-\frac{(7.5-10)^2}{2 \cdot 3^2}} = \frac{1}{3\sqrt{2\pi}} e^{-0.17361}$$

posterior probabilites:

$$P(Y = B \mid X = 7.5) = \frac{0.2 \cdot \frac{1}{2\sqrt{2\pi}} e^{-0.03125}}{0.5 \cdot \frac{1}{\sqrt{2\pi}} e^{-3.125} + 0.2 \cdot \frac{1}{2\sqrt{2\pi}} e^{-0.03125} + 0.3 \cdot \frac{1}{3\sqrt{2\pi}} e^{-0.17361}}$$

$$P(Y = B \mid X = 7.5) = \frac{0.2 \cdot \frac{1}{2} e^{-0.03125}}{0.5 \cdot e^{-3.125} + 0.2 \cdot \frac{1}{2} e^{-0.03125} + 0.3 \cdot \frac{1}{3} e^{-0.17361}}$$

Approximating

$$P(Y = B \mid X = 7.5) \approx \frac{0.09691}{0.02185 + 0.09691 + 0.08406}$$

$$P(Y = B \mid X = 7.5) \approx \frac{0.09691}{0.20282} \approx 0.478$$

(b) Provide a classification when $x = 5.5$. (Use the Bayes classifier.) Show your work. density functions:

$$\phi(5.5 \mid \mu_A, \sigma_A) = \frac{1}{\sqrt{2\pi}} e^{-0.125}, \quad \phi(5.5 \mid \mu_B, \sigma_B) = \frac{1}{2\sqrt{2\pi}} e^{-0.3125}, \quad \phi(5.5 \mid \mu_C, \sigma_C) = \frac{1}{3\sqrt{2\pi}} e^{-2.25}$$

posterior probbilities:

$$P(Y = A \mid X = 5.5) = \frac{0.5 \cdot e^{-0.125}}{0.5 \cdot e^{-0.125} + 0.2 \cdot \frac{1}{2} e^{-0.3125} + 0.3 \cdot \frac{1}{3} e^{-2.25}} \approx \frac{0.44125}{0.44125 + 0.08638 + 0.02193} \approx 0.804$$

$$P(Y = B \mid X = 5.5) \approx \frac{0.08638}{0.54956} \approx 0.157$$

$$P(Y = C \mid X = 5.5) \approx \frac{0.02193}{0.54956} \approx 0.04$$

Classify as A.

(c) Calculate the QDA decision boundary. (Use the Bayes classifier.) Show your work. the decision boundary between Class $i$ and Class $j$ is given by solving:

$$\log\left(\frac{\pi_i \phi(x \mid \mu_i, \sigma_i^2)}{\pi_j \phi(x \mid \mu_j, \sigma_j^2)}\right) = 0$$

Decision boundry between Class A and B:

$$\log\left(\frac{0.5}{0.2}\right) + \log\left(\frac{\frac{1}{\sqrt{2\pi\cdot 1}} e^{-\frac{(x-5)^2}{2\cdot 1^2}}}{\frac{1}{2\sqrt{2\pi}} e^{-\frac{(x-7)^2}{2\cdot 2^2}}}\right) = 0$$

$$\log\left(\frac{0.5}{0.2}\right) + \log\left(\frac{2}{1}\right) + \frac{(x-5)^2}{2} - \frac{(x-7)^2}{8} = 0$$

$$\log(2.5) + \log(2) + \frac{4(x-5)^2 - (x-7)^2}{8} = 0$$

$$\log(5) + \frac{3x^2 - 26x + 51}{8} = 0$$

$$3x^2 - 26x + 51 + 8\log(5) = 0$$

Decision boundary between Class A and Class C:

$$\log\left(\frac{0.5}{0.3}\right) + \log\left(\frac{\frac{1}{\sqrt{2\pi\cdot 1}} e^{-\frac{(x-5)^2}{2\cdot 1^2}}}{\frac{1}{3\sqrt{2\pi}} e^{-\frac{(x-10)^2}{2\cdot 3^2}}}\right) = 0$$

$$\log\left(\frac{5}{3}\right) + \log\left(\frac{3}{1}\right) + \frac{(x-5)^2}{2} - \frac{(x-10)^2}{18} = 0$$

$$\log(5) + \frac{8x^2 - 70x + 125}{18} = 0$$

$$8x^2 - 70x + 125 + 18\log(5) = 0$$

Decision boundary between Class B and Class C:

$$\log\left(\frac{0.2}{0.3}\right) + \log\left(\frac{\frac{1}{2\sqrt{2\pi}} e^{-\frac{(x-7)^2}{2\cdot 2^2}}}{\frac{1}{3\sqrt{2\pi}} e^{-\frac{(x-10)^2}{2\cdot 3^2}}}\right) = 0$$

$$\log\left(\frac{2}{3}\right) + \log\left(\frac{3}{2}\right) + \frac{(x-7)^2}{8} - \frac{(x-10)^2}{18} = 0$$

$$\frac{9(x^2 - 14x + 49) - 4(x^2 - 20x + 100)}{72} = 0$$

$$\frac{5x^2 - 46x - 49}{72} = 0$$

$$5x^2 - 46x + 49 = 0$$

## Question 5 (LDA-2 dimension)

Consider the following estimates and information from data for a three-class classifications boundary with two feature variables.

| Class A | Class B | Class C |
|---|---|---|
| $\hat{\mu}_A = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ | $\hat{\mu}_B = \begin{bmatrix} 3 \\ 1 \end{bmatrix}$ | $\hat{\mu}_C = \begin{bmatrix} 2 \\ 4 \end{bmatrix}$ |
| $n_A = 100$ | $n_B = 40$ | $n_C = 60$ |

$$\hat{\Sigma} = \begin{bmatrix} 1 & 0.3 \\ 0.3 & 1 \end{bmatrix}$$

(a) Provide a classification when $x = (2, 3)$. (Use the Bayes classifier.) Show your work. For LDA, the linear discriminant function for class $k$ is:

$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log(\pi_k)$$

Calculate $\Sigma^{-1}$:

$$\Sigma^{-1} = \frac{1}{1 - (0.3)^2} \begin{pmatrix} 1 & -0.3 \\ -0.3 & 1 \end{pmatrix} = \begin{pmatrix} 1.086 & -0.326 \\ -0.326 & 1.086 \end{pmatrix}$$

Calculate the linear disciminant functions:

$$\delta_A(x) = (2, 3) \begin{pmatrix} 1.086 & -0.326 \\ -0.326 & 1.086 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} - \frac{1}{2} \begin{pmatrix} 1 & 1 \end{pmatrix} \begin{pmatrix} 1.086 & -0.326 \\ -0.326 & 1.086 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} + \log(0.5)$$

$$= 2.526 - \frac{1}{2} \times 1.44 + \log(0.5) \approx 2.526 - 0.72 - 0.693 = 1.113$$

$$\delta_B(x) = (2, 3) \begin{pmatrix} 1.086 & -0.326 \\ -0.326 & 1.086 \end{pmatrix} \begin{pmatrix} 3 \\ 1 \end{pmatrix} - \frac{1}{2} \begin{pmatrix} 3 & 1 \end{pmatrix} \begin{pmatrix} 1.086 & -0.326 \\ -0.326 & 1.086 \end{pmatrix} \begin{pmatrix} 3 \\ 1 \end{pmatrix} + \log(0.2)$$

$$= 3.268 - \frac{1}{2} \times 4.716 + \log(0.2) \approx 3.268 - 2.358 - 1.609 = -0.699$$

$$\delta_C(x) = (2, 3) \begin{pmatrix} 1.086 & -0.326 \\ -0.326 & 1.086 \end{pmatrix} \begin{pmatrix} 2 \\ 4 \end{pmatrix} - \frac{1}{2} \begin{pmatrix} 2 & 4 \end{pmatrix} \begin{pmatrix} 1.086 & -0.326 \\ -0.326 & 1.086 \end{pmatrix} \begin{pmatrix} 2 \\ 4 \end{pmatrix} + \log(0.3)$$

$$= 4.072 - \frac{1}{2} \times 18.2 + \log(0.3) \approx 4.072 - 9.1 - 1.204 = -6.232$$

Classify as A since $\delta_A(x) > \delta_B(x)$ and $\delta_A(x) > \delta_C(x)$.

(b) Use LDA to estimate the probability $P(Y = B|X = (2, 3))$. Show your work.

Posterior probability for class $k$ is given by:

$$P(Y = k \mid X = x) = \frac{e^{\delta_k(x)}}{\sum_l e^{\delta_l(x)}}$$

Calculate the poserior probabilities:

$$P(Y = A \mid X = (2,3)) = \frac{e^{1.113}}{e^{1.113} + e^{-0.699} + e^{-6.232}} \approx \frac{3.045}{3.045 + 0.497 + 0.002} \approx 0.85$$

$$P(Y = B \mid X = (2,3)) = \frac{e^{-0.699}}{3.045 + 0.497 + 0.002} \approx \frac{0.497}{3.045 + 0.497 + 0.002} \approx 0.14$$

$$P(Y = C \mid X = (2,3)) = \frac{e^{-6.232}}{3.045 + 0.497 + 0.002} \approx \frac{0.002}{3.045 + 0.497 + 0.002} \approx 0.001$$

The estimated probability $P(Y = B \mid X = (2,3))$ is 0.14