STAT 32950: Homework 5

Robert Winter

Table of Contents

1 Exercise 1: Two-Class Mini kNN Classification

The training data (x_i, y_i) for $i = 1, 2, 3$ are $(0.3, 1), (0.5, 1), (0.7, 0)$. In the below, for special or ambivalent assignments, state your reasoning.

1.1 Part (a)

For all $x \in [0, 1]$, determine the output of the binary class label y given by a k-Nearest Neighbor (kNN) classifier, using...

1.1.1 Part (i): 1NN

First, note that $\forall x \in [0, 0.3]$, x's nearest neighbor in the training data is clearly $x_1 = 0.3$. Similarly, $\forall x \in [0.7, 1]$, x's nearest neighbor in the training data is clearly $x_3 = 0.7$.

The midpoint between $x_1 = 0.3$ and $x_2 = 0.5$ is 0.4, so for $x \in (0.3, 0.4)$, x's nearest neighbor in the training data is $x_1 = 0.3$, while for $x \in (0.4, 0.5)$, x's nearest neighbor in the training data is $x_2 = 0.5$. Similarly, the midpoint between $x_2 = 0.5$ and $x_3 = 0.7$ is 0.6, so for $x \in (0.5, 0.6)$, x's nearest neighbor in the training data is $x_2 = 0.5$, while for $x \in (0.6, 0.7)$, x's nearest neighbor in the training data is $x_3 = 0.7$.

Note that $x = 0.4$ is equally close to $x_1 = 0.3$ and $x_2 = 0.5$, and $x = 0.6$ is equally close to $x_2 = 0.5$ and $x_3 = 0.7$. We arbitrarily break both of these "ties" by assigning $x = 0.4, 0.6$ to have nearest neighbor $x_2 = 0.5$.

So, the set of points with nearest neighbor x_1 is $S_1 = [0, 0.4)$, the set of points with nearest neighbor x_2 is $S_2 = [0.4, 0.6]$, and the set of points with nearest neighbor x_3 is $S_3 = (0.6, 1]$.

Thus, the binary classifier based on the training data is

$$y(x) = \begin{cases} 0 & \text{if } x \in S_3 \\ 1 & \text{if } x \in S_1 \cup S_2 \end{cases} \Leftrightarrow y(x) = \begin{cases} 0 & \text{if } x \in (0.6, 1] \\ 1 & \text{if } x \in [0, 0.6] \end{cases}.$$

1.1.2 Part (ii): 3NN

Since there are only three points in the training data, if we use 3NN, then *every* $x \in [0, 1]$ has nearest neighbors $\{x_1, x_2, x_3\}$. Since $y_1 = y_2 = 1$ and $y_3 = 0$, the plurality vote is in favor of Class 1, and so *every* $x \in [0, 1]$ will be classified to $y = 1$. That is,

$$y(x) = 1 \ \forall x \in [0, 1].$$

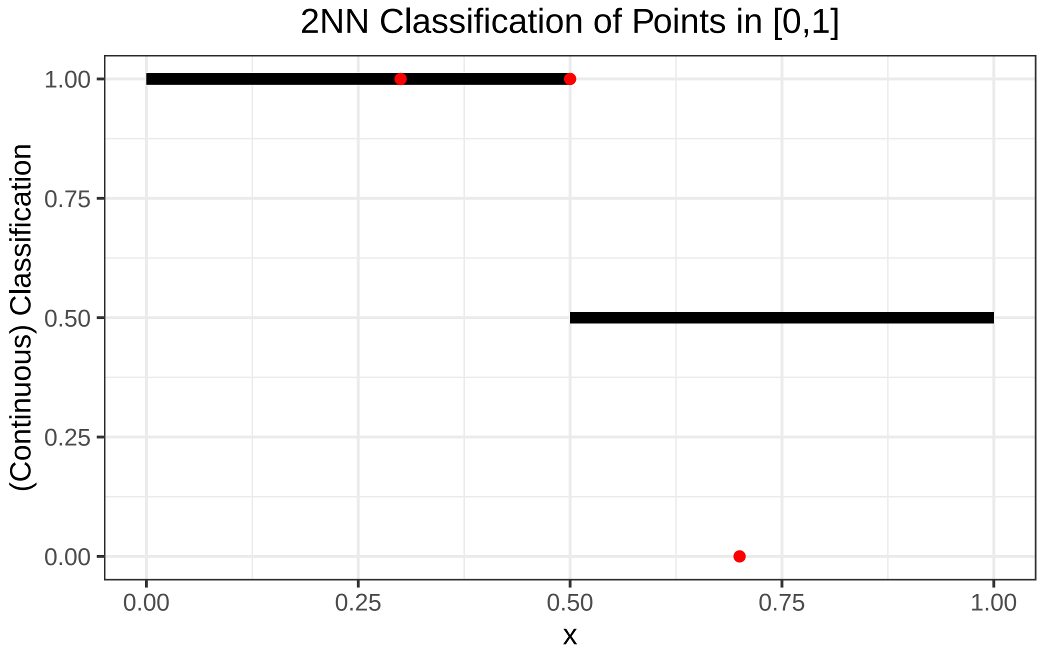
1.2 Part (b)

Using the mean of the k-nearest neighbors of a test point, plot the output y (no longer binary) for all $x \in [0,1]$, by using 2NN.

Notice that $x_1 < x_2 < x_3$. So, if $x \leq x_1$, then x's two nearest neighbors are x_1 and x_2. Similarly, if $x \geq x_3$, then x's two nearest neighbors are x_2 and x_3. Also notice that $x_1 = 0.3$ and $x_3 = 0.7$ are symmetrically arranged around $x_2 = 0.5$. So, if $x < x_2$, then x is closer to x_1 than it is to x_3, and if $x > x_2$, then x is closer to x_3 than it is to x_1.

Thus, $S_{12} = [0, 0.5]$ is the set of points with nearest neighbors $\{x_1, x_2\}$, and $S_{23} = (0.5, 1]$ is the set of points with nearest neighbors $\{x_2, x_3\}$. Note that we have arbitrarily decided that the point $x = 0.5$ has nearest neighbors $\{x_1, x_2\}$ rather than $\{x_2, x_3\}$.

Since $y_1 = y_2 = 1$, all points $x \in S_{12}$ are classified to $y(x) = \frac{1+1}{2} = 1$. Since $y_2 = 1$ and $y_3 = 0$, all points $x \in S_{23}$ are classified to $y(x) = \frac{1+0}{2} = 0.5$. We plot this labeling scheme below, with the training data in red for reference.



2 Exercise 3: Fisher's Linear Discriminants for Three Classes

A business school admissions committee used GPA and GMAT scores to make admission decisions.

The dataset is `GpaGmat.DAT`. The variable `admit = 1,2,3` corresponds to admission decisions "Yes", "No", "Borderline".

```
gpagmat =
 ↪  read.table("C:/Users/rewin/OneDrive/Documents/STAT_32950/Homework/HW5/GpaGmat.DAT")
colnames(gpagmat) = c("GPA", "GMAT", "admit")
```

2.1 Part (a)

Calculate $\bar{x}_i, \bar{x}, S_{pool}$.

```
# Grand mean \bar{x}
q3xbar = colMeans(gpagmat[,1:2]) %>% as.vector()

# Group means \bar{x}_i
q3xbar_groups = gpagmat %>%
  group_by(admit) %>%
  summarize(mean_GPA = mean(GPA),
            mean_GMAT = mean(GMAT),
            count = n())
q3xbar1 = q3xbar_groups[1,2:3] %>% t() %>% as.vector()
q3xbar2 = q3xbar_groups[2,2:3] %>% t() %>% as.vector()
q3xbar3 = q3xbar_groups[3,2:3] %>% t() %>% as.vector()

# Pooled variance
q3n1 = q3xbar_groups[1,4] %>% as.integer()
q3n2 = q3xbar_groups[2,4] %>% as.integer()
q3n3 = q3xbar_groups[3,4] %>% as.integer()

q3S1 = filter(gpagmat, admit == 1) %>% dplyr::select(-admit) %>% cov()
q3S2 = filter(gpagmat, admit == 2) %>% dplyr::select(-admit) %>% cov()
q3S3 = filter(gpagmat, admit == 3) %>% dplyr::select(-admit) %>% cov()

q3Spool = ((q3n1-1)*q3S1 + (q3n2-1)*q3S2 +
 ↪  (q3n3-1)*q3S3)/(q3n1+q3n2+q3n3-3)
```

Thus, we have $\bar{x} \approx (2.975, 488.447)$, $\bar{x}_1 \approx (3.404, 561.226)$, $\bar{x}_2 \approx (2.483, 447.071)$, $\bar{x}_3 \approx (2.993, 446.231)$, and $S_{pool} \approx \begin{bmatrix} 0.036 & -2.019 \\ -2.019 & 3655.901 \end{bmatrix}$.

2.2 Part (b)

Calculate the sample within-group sum of squares and cross products matrix W, its inverse W^{-1}, and the sample between-group sum of squares and cross products matrix B.

Find the eigenvalues λ_1, λ_2 (or rather the estimates $\hat{\lambda}_1, \hat{\lambda}_2$) and eigenvectors a_1, a_2 of $W^{-1}B$.

```
# mean of means for B matrix
q3xbarbar = (q3xbar1 + q3xbar2 + q3xbar3) / 3

# B matrix
q3B = (q3xbar1-q3xbarbar)%*%t(q3xbar1-q3xbarbar) +
      (q3xbar2-q3xbarbar)%*%t(q3xbar2-q3xbarbar) +
      (q3xbar3-q3xbarbar)%*%t(q3xbar3-q3xbarbar)

# W matrix
q3W = (nrow(gpagmat) - 3) * q3Spool

# W^-1
q3Winv = solve(q3W)

# W^-1 B
q3WinvB = q3Winv %*% q3B

# eigenstuffs
q3lambda1 = eigen(q3WinvB)$values[1]
q3lambda2 = eigen(q3WinvB)$values[2]
q3a = eigen(q3WinvB)$vectors
q3a1 = eigen(q3WinvB)$vectors[,1]
q3a2 = eigen(q3WinvB)$vectors[,2]
```

Thus, we have $W \approx \begin{bmatrix} 2.958 & -165.538 \\ -165.538 & 299,783.892 \end{bmatrix}$, $W^{-1} \approx \begin{bmatrix} 0.349 & 1.93 \times 10^{-4} \\ 1.93 \times 10^{-4} & 3.44 \times 10^{-6} \end{bmatrix}$, $B \approx \begin{bmatrix} 0.426 & 50.678 \\ 50.678 & 8751.929 \end{bmatrix}$. So, $W^{-1}B \approx \begin{bmatrix} 0.158 & 19.368 \\ 2.57 \times 10^{-4} & 0.040 \end{bmatrix}$, which has eigenvalues $\hat{\lambda}_1 \approx 0.191$ and $\hat{\lambda}_2 \approx 0.007$ and corresponding eigenvectors $a_1 \approx (1.000, 0.002)$ and $a_2 \approx (-1.000, 0.008)$.

2.3 Part (c)

Use the linear discriminants derived from these eigenvectors to classify two new observations $x = (3.21, 497)$ and $x = (3.22, 497)$. Note: You may use the common rule that x is assigned to class π_k if $a^T x \in \mathbb{R}^2$ is closest to $a^T \bar{x}_k$ for $k = 1, \dots, g$ (here $g = 3$), where $a = \begin{bmatrix} a_1 & a_2 \end{bmatrix}$ is the 2×2 eigenvector matrix.

Let $a = \begin{bmatrix} a_1 & a_2 \end{bmatrix} \approx \begin{bmatrix} 1.000 & -1.000 \\ 0.002 & 0.008 \end{bmatrix}$.

First, let $x_1 = (3.21, 497)$. Then $||a^T(x_1 - \bar{x}_1)||_2 \approx 0.432$, $||a^T(x_1 - \bar{x}_2)||_2 \approx 0.879$, and $||a^T(x_1 - \bar{x}_3)||_2 \approx 0.352$. Since $0.352 < 0.432 < 0.879$, we assign the first candidate to admission group 3, "borderline."

```r
# x = (3.21, 497)
t(q3a)%*%(c(3.21,497)-q3xbar1) %>% norm(type = "2")
```

```
[1] 0.4319065
```

```r
t(q3a)%*%(c(3.21,497)-q3xbar2) %>% norm(type = "2")
```

```
[1] 0.8793717
```

```r
t(q3a)%*%(c(3.21,497)-q3xbar3) %>% norm(type = "2")
```

```
[1] 0.3524615
```

Now, let $x_2 = (3.22, 497)$. Then $||a^T(x_2 - \bar{x}_1)||_2 \approx 0.432$, $||a^T(x_2 - \bar{x}_2)||_2 \approx 0.892$, and $||a^T(x_2 - \bar{x}_3)||_2 \approx 0.356$. Since $0.356 < 0.432 < 0.892$, we assign the second candidate to admission group 3, "borderline," as well.

```r
# x = (3.22, 497)
t(q3a)%*%(c(3.22,497)-q3xbar1) %>% norm(type = "2")
```

```
[1] 0.4322614
```

```r
t(q3a)%*%(c(3.22,497)-q3xbar2) %>% norm(type = "2")
```

```
[1] 0.8924586
```

```r
t(q3a)%*%(c(3.22,497)-q3xbar3) %>% norm(type = "2")
```

```
[1] 0.35624
```

2.4 Part (d)

Plot the original dataset on the plane of the first two discriminants, labeled by admission decisions. Comment on the results in Part (c). Is the admission policy a good one?

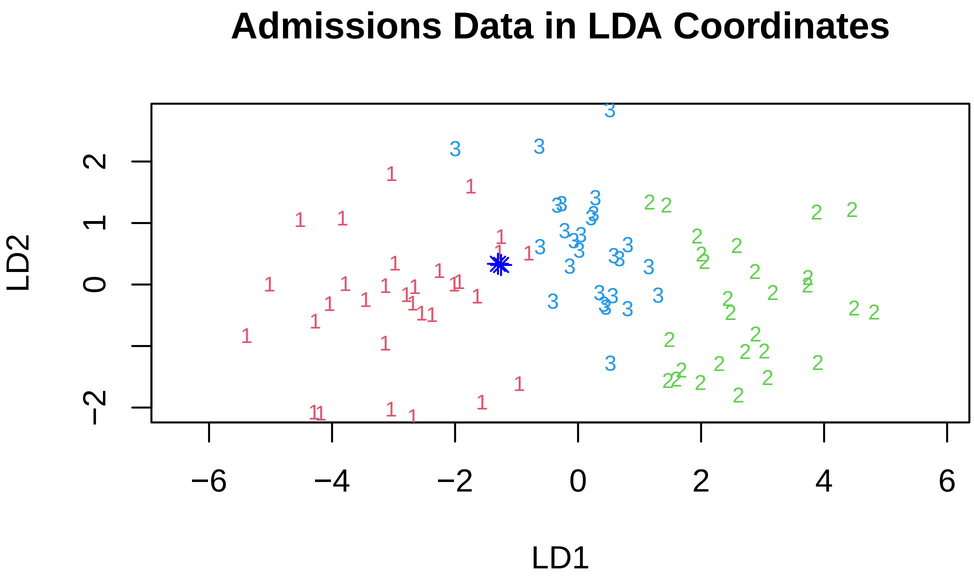
We plot the original dataset in the (LD1, LD2) plane below, with points labeled “1” (red) corresponding to admissions, “2” (green) corresponding to rejections, and “3” (blue) corresponding to “borderline” decisions.

Also included in the plot are the (LD1, LD2) coordinates of the two candidates considered in Part (c), $x_1 = (3.21, 497)$ and $x_2 = (3.22, 497)$—which were obtained by multiplying the transpose of the LDA’s `scaling` attribute by the mean-centered GPA and GMAT scores of the candidates. These points are marked with $\star$ symbols and are colored blue, corresponding to our conclusion in Part (c) that they should be allocated to the borderline group. But, looking at the plot, it isn’t obvious that “borderline” was the right decision for these two applicants. While these applicants are both close to the edge between the admitted cloud and the “borderline” cloud, they appear decidedly inside the admitted cloud, and even have LD coordinates very close to those of other points who were admitted. As such, using our LDA analysis to determine admission decisions isn’t infallible, and may make questionable choices for candidates whose LDA coordinates are around the edge between two of the group’s clouds.

```r
q3lda = lda(gpagmat[,1:2], gpagmat[,3])
plot(q3lda, col = rep(2:4, c(q3n1, q3n2, q3n3)),
     main = "Admissions Data in LDA Coordinates")

# LDA coordinates of two candidates
q3x1star = t(q3lda$scaling) %*%
 ↪  c(3.21-mean(gpagmat$GPA),497-mean(gpagmat$GMAT))
q3x2star = t(q3lda$scaling) %*%
 ↪  c(3.22-mean(gpagmat$GPA),497-mean(gpagmat$GMAT))

# Plot two new points
points(q3x1star[1], q3x1star[2], col = "blue", pch = 8)
points(q3x2star[1], q3x2star[2], col = "blue", pch = 8)
```



Admissions Data in LDA Coordinates

3 Exercise 4: Hands-On Classifications for Two Normal Populations

Two data sets X_1, X_2 are sampled from two bivariate normal populations π_1, π_2, with $X_1 = \{(3,7),(2,4),(4,7)\}$, $X_2 = \{(6,9),(5,7),(4,8)\}$.

```
q4X1 = cbind(c(3,2,4),c(7,4,7))
q4X2 = cbind(c(6,5,4),c(9,7,8))

q4mu1 = colMeans(q4X1)
q4mu2 = colMeans(q4X2)

q4S1 = cov(q4X1)
q4S2 = cov(q4X2)
```

3.1 Part (a)

Assuming equal covariance matrices $\Sigma_1 = \Sigma_2$ in the two populations, calculate the estimated linear discriminant function $y = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T S_{pool}^{-1}\mathbf{x}$.

We have $\bar{\mathbf{x}}_1 = (\frac{3+2+4}{3}, \frac{7+4+7}{3}) = (3,6)$ and $\bar{\mathbf{x}}_2 = (\frac{6+5+4}{3}, \frac{9+7+8}{3}) = (5,8)$, so $\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 = (3,6) - (5,8) = (-2,-2)$.

Moreover, we have $S_1 = \begin{bmatrix} 1 & 1.5 \\ 1.5 & 3 \end{bmatrix}$ and $S_2 = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$, so $S_{pool} = \frac{(n_1-1)S_1+(n_2-1)S_2}{n_1+n_2-2} = \frac{(3-1)S_1+(3-1)S_2}{3+3-2} = \frac{1}{2}\left(\begin{bmatrix} 1+1 & 1.5+0.5 \\ 1.5+0.5 & 3+1 \end{bmatrix}\right) = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$. Thus, $S_{pool}^{-1} = \frac{1}{(1)(2)-(1)(1)}\begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix} = \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix}$.

Therefore,

$$\begin{aligned}
y &= (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T S_{pool}^{-1}\mathbf{x} \\
&= \begin{bmatrix} -2 & -2 \end{bmatrix} \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix} \mathbf{x} \\
&= \begin{bmatrix} -2(2) - 2(-1) & -2(-1) - 2(1) \end{bmatrix} \mathbf{x} \\
&= \begin{bmatrix} -2 & 0 \end{bmatrix} \mathbf{x} \\
&= -2x_1,
\end{aligned}$$

where x_1 is the first coordinate of $\mathbf{x} = (x_1, x_2)$.

```
q4Spool = ((3-1)*cov(q4X1) + (3-1)*cov(q4X2))/(3+3-2)
q4y = t(q4mu1 - q4mu2) %*% solve(q4Spool)
```

3.2 Part (b)

Use the rule of minimum expected costs of misclassification (ECM) to classify the new observations $\mathbf{x} = (4.1, 5)$ and $\mathbf{x} = (3.9, 9)$, under the assumption of equal costs and equal priors. (Under these assumptions the classifier is equivalent to Fisher's linear discriminant function.)

We allocate $\mathbf{x}^*$ to π_1 if $y(\mathbf{x}^*) - \frac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T S_{pool}^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2) \geq \log\left(\frac{c(1|2)}{c(2|1)}\frac{p_2}{p_1}\right) = \log\left(\frac{1*1}{1*1}\right) = 0 \Leftrightarrow y(\mathbf{x}^*) \geq \frac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T S_{pool}^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)$. We allocate $\mathbf{x}^*$ to π_2 otherwise. Note that

$$\begin{aligned}
\frac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T S_{pool}^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2) &= \frac{1}{2}\begin{bmatrix} -2 & -2 \end{bmatrix} \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 3+5 \\ 6+8 \end{bmatrix} \\
&= \begin{bmatrix} -1 & -1 \end{bmatrix} \begin{bmatrix} 2(8) - 1(14) \\ -1(8) + 1(14) \end{bmatrix} \\
&= -1(2) - 1(6) \\
&= -8.
\end{aligned}$$

So, we allocate $\mathbf{x}^*$ to π_1 if $y(\mathbf{x}^*) \geq -8$ and to π_2 if $y(\mathbf{x}^*) < -8$.

```
0.5 * t(q4mu1 - q4mu2) %*% solve(q4Spool) %*% (q4mu1 + q4mu2)
```

```
     [,1]
[1,]   -8
```

First, let $\mathbf{x}_1^* = (4.1, 5)$. Then $y(\mathbf{x}_1^*) = \begin{bmatrix}-2 & 0\end{bmatrix}\begin{bmatrix}4.1\\5\end{bmatrix} = -2(4.1) + 0(5) = -8.2 < -8$. Thus, we allocate $\mathbf{x}_1^* = (4.1, 5)$ to the second population, π_2.

```r
q4y %*% c(4.1, 5)
```

```
     [,1]
[1,] -8.2
```

Now, let $\mathbf{x}_2^* = (3.9, 9)$. Then $y(\mathbf{x}_2^*) = \begin{bmatrix}-2 & 0\end{bmatrix}\begin{bmatrix}3.9\\9\end{bmatrix} = -2(3.9) + 0(9) = -7.8 \geq -8$. Thus, we allocate $\mathbf{x}_2^* = (3.9, 9)$ to the first population, π_1.

```r
q4y %*% c(3.9, 9)
```

```
     [,1]
[1,] -7.8
```

3.3 Part (c)

Repeat Part (b) with cost $c(2|1) = \$3$, $c(1|2) = \$20$, and assuming that about 10% of all possible observations belong to population π_1.

Now, we allocate $\mathbf{x}^*$ to π_1 if $y(\mathbf{x}^*) - \frac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T S_{pool}^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2) \geq \log\left(\frac{c(1|2)}{c(2|1)}\frac{p_2}{p_1}\right) = \log\left(\frac{20}{3}\frac{1-0.1}{0.1}\right) = \log(60) \Leftrightarrow y(\mathbf{x}^*) \geq \log(60) - 8 \approx -3.906$, and to π_2 otherwise.

We showed in Part (b) that for $\mathbf{x}_1^* = (4.1, 5)$, we have $y(\mathbf{x}_1^*) = -8.2$. Since $-8.2 < -3.906$, we allocate $\mathbf{x}_1^* = (4.1, 5)$ to the second population, π_2 (as we did before).

We also showed that for $\mathbf{x}_2^* = (3.9, 9)$, we have $y(\mathbf{x}_2^*) = -7.8$. Since $-7.8 < -3.906$, we also allocate $\mathbf{x}_2^* = (3.9, 9)$ to the second population, π_2 (unlike in Part (b)).

3.4 Part (d)

Assume that $\Sigma_1 \neq \Sigma_2$ in the two bivariate normal distributions. Use the general quadratic rule

$$R_1 = \left\{\mathbf{x} : d(\mathbf{x}) \geq \log\left(\frac{c(1|2)}{c(2|1)}\frac{p_2}{p_1}\right)\right\},\; R_2 = \left\{\mathbf{x} : d(\mathbf{x}) < \log\left(\frac{c(1|2)}{c(2|1)}\frac{p_2}{p_1}\right)\right\}$$

to classify the observations $\mathbf{x} = (4.1, 5)$ and $\mathbf{x} = (3.9, 9)$, where

$$d(\mathbf{x}) = -\frac{1}{2}\mathbf{x}^T(S_1^{-1} - S_2^{-1})\mathbf{x} + (\bar{\mathbf{x}}_1^T S_1^{-1} - \bar{\mathbf{x}}_2^T S_2^{-1})\mathbf{x} - \hat{k},\ \hat{k} = \frac{1}{2}(\bar{\mathbf{x}}_1^T S_1^{-1}\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2^T S_2^{-1}\bar{\mathbf{x}}_2) + \frac{1}{2}\log\left(\frac{|S_1|}{|S_2|}\right).$$

Let's make it simple by using equal costs and equal priors as in Part (b).

Since $c(1|2) = c(2|1)$ and $p_1 = p_2$, we have $R_1 = \{\mathbf{x} : d(\mathbf{x}) \geq 0\}$ and $R_2 = \{\mathbf{x} : d(\mathbf{x}) < 0\}$.

Note that $|S_1| = (1)(3) - (1.5)(1.5) = 3 - 2.25 = 0.75$ and $|S_2| = (1)(1) - (0.5)(0.5) = 1 - 0.25 = 0.75$, so

$$\begin{aligned}
\frac{1}{2}\log\left(\frac{|S_1|}{|S_2|}\right) &= \frac{1}{2}\log\left(\frac{0.75}{0.75}\right) \\
&= \frac{1}{2}\log(1) \\
&= 0.
\end{aligned}$$

Moreover,

$$\begin{aligned}
\bar{\mathbf{x}}_1^T S_1^{-1}\bar{\mathbf{x}}_1 &= \begin{bmatrix} 3 & 6 \end{bmatrix} \frac{1}{0.75} \begin{bmatrix} 3 & -1.5 \\ -1.5 & 1 \end{bmatrix} \begin{bmatrix} 3 \\ 6 \end{bmatrix} \\
&= \frac{4}{3}\begin{bmatrix} 3 & 6 \end{bmatrix} \begin{bmatrix} 3(3) - 1.5(6) \\ -1.5(3) + 1(6) \end{bmatrix} \\
&= \begin{bmatrix} 4 & 8 \end{bmatrix} \begin{bmatrix} 0 \\ 1.5 \end{bmatrix} \\
&= 4(0) + 8(1.5) \\
&= 12.
\end{aligned}$$

Similarly,

$$\begin{aligned}
\bar{\mathbf{x}}_2^T S_2^{-1}\bar{\mathbf{x}}_2 &= \begin{bmatrix} 5 & 8 \end{bmatrix} \frac{1}{0.75} \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix} \begin{bmatrix} 5 \\ 8 \end{bmatrix} \\
&= \frac{4}{3}\begin{bmatrix} 5 & 8 \end{bmatrix} \begin{bmatrix} 1(5) - 0.5(8) \\ -0.5(5) + 1(8) \end{bmatrix} \\
&= \begin{bmatrix} \frac{20}{3} & \frac{32}{3} \end{bmatrix} \begin{bmatrix} 1 \\ 5.5 \end{bmatrix} \\
&= \frac{20}{3}(1) + \frac{32}{3}(5.5) \\
&= \frac{196}{3}.
\end{aligned}$$

Thus,

$$
\begin{aligned}
\hat{k} &= \frac{1}{2}(12 - \frac{196}{3}) + 0 \\
&= 6 - \frac{98}{3} \\
&= -\frac{80}{3} \\
&\approx -26.667
\end{aligned}
$$

```
q4khat = 0.5*(t(q4mu1)%*%solve(q4S1)%*%q4mu1 -
 ↪  t(q4mu2)%*%solve(q4S2)%*%q4mu2) +
          0.5*log(det(q4S1)/det(q4S2))
```

Note also that

$$
\begin{aligned}
S_1^{-1} - S_2^{-1} &= \frac{1}{0.75}\begin{bmatrix} 3 & -1.5 \\ -1.5 & 1 \end{bmatrix} - \frac{1}{0.75}\begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix} \\
&= \frac{4}{3}\begin{bmatrix} 2 & -1 \\ -1 & 0 \end{bmatrix}.
\end{aligned}
$$

```
solve(q4S1) - solve(q4S2)
```

```
          [,1]      [,2]
[1,]  2.666667 -1.333333
[2,] -1.333333  0.000000
```

Also,

$$
\begin{aligned}
\bar{\mathbf{x}}_1^T S_1^{-1} - \bar{\mathbf{x}}_2^T S_2^{-1} &= \begin{bmatrix} 3 & 6 \end{bmatrix} \frac{1}{0.75}\begin{bmatrix} 3 & -1.5 \\ -1.5 & 1 \end{bmatrix} - \begin{bmatrix} 5 & 8 \end{bmatrix} \frac{1}{0.75}\begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix} \\
&= \frac{4}{3}\begin{bmatrix} 3(3) + 6(-1.5) & 3(-1.5) + 6(1) \end{bmatrix} - \frac{4}{3}\begin{bmatrix} 5(1) + 8(-0.5) & 5(-0.5) + 8(1) \end{bmatrix} \\
&= \frac{4}{3}\begin{bmatrix} 0 & 1.5 \end{bmatrix} - \frac{4}{3}\begin{bmatrix} 1 & 5.5 \end{bmatrix} \\
&= \frac{4}{3}\begin{bmatrix} -1 & -4 \end{bmatrix}.
\end{aligned}
$$

```
t(q4mu1)%*%solve(q4S1) - t(q4mu2)%*%solve(q4S2)
```

```
           [,1]      [,2]
[1,] -1.333333 -5.333333
```

Thus,

$$\begin{aligned} d(\mathbf{x}) &= -\frac{1}{2}\mathbf{x}^T \frac{4}{3}\begin{bmatrix} 2 & -1 \\ -1 & 0 \end{bmatrix}\mathbf{x} + \frac{4}{3}\begin{bmatrix} -1 & -4 \end{bmatrix}\mathbf{x} + \frac{80}{3} \\ &= -\frac{2}{3}\mathbf{x}^T \begin{bmatrix} 2 & -1 \\ -1 & 0 \end{bmatrix}\mathbf{x} + \frac{4}{3}\begin{bmatrix} -1 & -4 \end{bmatrix}\mathbf{x} + \frac{80}{3}. \end{aligned}$$

```
q4d = function(x){
  d = -0.5*t(x)%*%(solve(q4S1)-solve(q4S2))%*%x +
      (t(q4mu1)%*%solve(q4S1)-t(q4mu2)%*%solve(q4S2))%*%x -
      (0.5*(t(q4mu1)%*%solve(q4S1)%*%q4mu1 -
 ↪   t(q4mu2)%*%solve(q4S2)%*%q4mu2) +
       0.5*log(abs(det(q4S1))/abs(det(q4S2))))
  return(d)
}
```

So,

$$\begin{aligned} d(\mathbf{x}_1^*) &= d(4.1, 5) \\ &= -\frac{2}{3}\begin{bmatrix} 4.1 & 5 \end{bmatrix}\begin{bmatrix} 2 & -1 \\ -1 & 0 \end{bmatrix}\begin{bmatrix} 4.1 \\ 5 \end{bmatrix} + \frac{4}{3}\begin{bmatrix} -1 & -4 \end{bmatrix}\begin{bmatrix} 4.1 \\ 5 \end{bmatrix} + \frac{80}{3} \\ &= -\frac{2}{3}\begin{bmatrix} 4.1 & 5 \end{bmatrix}\begin{bmatrix} 2(4.1) - 1(5) \\ -1(4.1) + 0(5) \end{bmatrix} + \frac{4}{3}[-1(4.1) - 4(5)] + \frac{80}{3} \\ &= -\frac{2}{3}[(4.1)(3.2) + (5)(-4.1)] - \frac{4}{3}(24.1) + \frac{80}{3} \\ &= -\frac{2}{3}(-7.38) - \frac{4}{3}(24.1) + \frac{80}{3} \\ &= -\frac{41}{75} \\ &\approx -0.547 \\ &< 0. \end{aligned}$$

```
q4d(c(4.1,5))
```

```
           [,1]
[1,] -0.5466667
```

Since $d(\mathbf{x}_1^*) < 0$, we allocate $\mathbf{x}_1^* = (4.1, 5)$ to the second population, π_2.

Analogously,

$$\begin{aligned}
d(\mathbf{x}_2^*) &= d(3.9, 9) \\
&= -\frac{2}{3}\begin{bmatrix} 3.9 & 9 \end{bmatrix}\begin{bmatrix} 2 & -1 \\ -1 & 0 \end{bmatrix}\begin{bmatrix} 3.9 \\ 9 \end{bmatrix} + \frac{4}{3}\begin{bmatrix} -1 & -4 \end{bmatrix}\begin{bmatrix} 3.9 \\ 9 \end{bmatrix} + \frac{80}{3} \\
&= -\frac{2}{3}\begin{bmatrix} 3.9 & 9 \end{bmatrix}\begin{bmatrix} 2(3.9) - 1(9) \\ -1(3.9) + 0(9) \end{bmatrix} + \frac{4}{3}[-1(3.9) - 4(9)] + \frac{80}{3} \\
&= -\frac{2}{3}[3.9(-1.2) + 9(-3.9)] - \frac{4}{3}(39.9) + \frac{80}{3} \\
&= -\frac{2}{3}(-39.78) - \frac{4}{3}(39.9) + \frac{80}{3} \\
&= -\frac{1}{75} \\
&\approx -0.013 \\
&< 0
\end{aligned}$$

```r
q4d(c(3.9,9))
```

```
              [,1]
[1,] -0.01333333
```

Since $d(\mathbf{x}_2^*) < 0$, we allocate $\mathbf{x}_2^* = (3.9, 9)$ to the second population, π_2, as well.

3.5 Part (e)

Compare and comment on the classification results in Parts (b) and (d).

In both Parts (b) and (d), we assume equal costs and equal priors. However, whereas Part (b) assumes $\Sigma_1 = \Sigma_2$ and therefore uses a linear classification rule, Part (d) assumes that $\Sigma_1 \neq \Sigma_2$ and therefore employs a quadratic classification rule.

Our analyses in both Parts (b) and (d) agree that the observation $\mathbf{x}_1^* = (4.1, 5)$ should be classified to the second population, π_2. However, the two analyses disagree on the classification of $\mathbf{x}_2^* = (4.9, 9)$. In Part (b), we recovered $y(\mathbf{x}_2^*) = -7.8 > -8$, and therefore classified $\mathbf{x}_2^* \in \pi_1$. In Part (d), on the other hand, we recovered $d(\mathbf{x}_2^*) = -\frac{1}{75} < 0$, and therefore classified $\mathbf{x}_2^* \in \pi_2$. It is interesting that in both analyses, the classification of $\mathbf{x}_2^*$ was really on the margin: $y(\mathbf{x}_2^*) = -7.8$ was just barely above the -8 cutoff in Part (b), and $d(\mathbf{x}_2^*) = -\frac{1}{75}$ was just barely below the 0 cutoff in Part (d). Since the approach in Part (d) makes a less restrictive assumption about the populations' covariance matrices than the approach in Part (b), it is probably safest to allocate $\mathbf{x}_2^*$ to π_2 — but the narrow margins in each analysis suggest that this observation could plausibly be allocated to either population.

3.6 Part (f)

Test for the difference in population mean vectors using Hotelling's two-sample T^2 test statistic.

Under the null hypothesis that $\mu_1 = \mu_2$, Hotelling's T^2 test statistic is

$$\begin{aligned}
T^2 &= [(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)]^T \left[\left(\frac{1}{n_1} + \frac{1}{n_2} \right) S_{pool} \right]^{-1} [(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)] \\
&= \left[\left(\begin{bmatrix} 3 \\ 6 \end{bmatrix} - \begin{bmatrix} 5 \\ 8 \end{bmatrix} \right) - \begin{bmatrix} 0 \\ 0 \end{bmatrix} \right]^T \left[\left(\frac{1}{3} + \frac{1}{3} \right) \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix} \right]^{-1} \left[\left(\begin{bmatrix} 3 \\ 6 \end{bmatrix} - \begin{bmatrix} 5 \\ 8 \end{bmatrix} \right) - \begin{bmatrix} 0 \\ 0 \end{bmatrix} \right] \\
&= \frac{3}{2} \begin{bmatrix} -2 \\ -2 \end{bmatrix}^T \frac{1}{(1)(2) - (1)(1)} \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} -2 \\ -2 \end{bmatrix} \\
&= \frac{3}{2} \begin{bmatrix} -2 & -2 \end{bmatrix} \begin{bmatrix} 2(-2) - 1(-2) \\ -1(-2) + 1(-2) \end{bmatrix} \\
&= \begin{bmatrix} -3 & -3 \end{bmatrix} \begin{bmatrix} -2 \\ 0 \end{bmatrix} \\
&= -3(-2) - 3(0) \\
&= 6.
\end{aligned}$$

Moreover, under the null hypothesis,

$$\begin{aligned}
T^2 &\sim \frac{(n_1 + n_2 - 2)p}{n_1 + n_2 - p - 1} F_{p, n_1 + n_2 - p - 1} \\
&= \frac{(3 + 3 - 2)(2)}{3 + 3 - 2 - 1} F_{2, 3+3-2-1} \\
&= \frac{8}{3} F_{2,3}.
\end{aligned}$$

$$\mathbb{P}\left(\frac{8}{3} F_{2,3} > 6 \right) = \mathbb{P}\left(F_{2,3} > \frac{9}{4} \right) \approx 0.253 > 0.05.$$

Thus, there is *not* statistical evidence that the two populations have significantly different mean vectors.

```r
t(q4mu1 - q4mu2) %*% solve(((1/3)+(1/3))*q4Spool) %*% (q4mu1 - q4mu2) #
↪  T^2
```

```
     [,1]
[1,]    6
```

```r
pf(9/4, 2, 3, lower.tail = F) # p-value
```

```
[1] 0.2529822
```

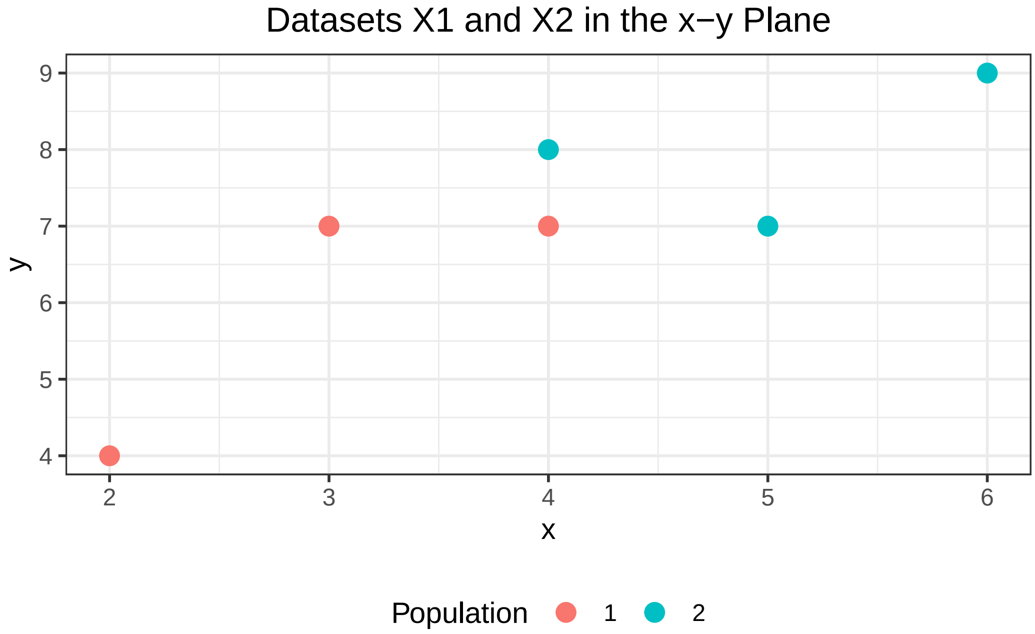
4 Exercise 5: Graphical Exercise on Linear Discriminants and Conceptual SVM

This geometric-graphical exercise is based on the data in Question 4 in this assignment.

The two datasets sampled from two classes are $X_1 = \{(3,7),(2,4),(4,7)\}$ and $X_2 = \{(6,9),(5,7),(4,8)\}$.

Plot the data points (as $(x_{ci}, y_{ci}), c = 1,2; i = 1,2,3$) on the x-y plane, with clear labels on the plot indicating the class of each point. This will be the base plot for the following parts.

We plot X_1 and X_2 below, with points color-coded by class.



4.1 Part (a)

On your base plot, based on your results in Parts (a) and (b) in Question 4, plot the linear classification border $\hat{y} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T S_{pool}^{-1}\mathbf{x} - \hat{m} = 0$ obtained by

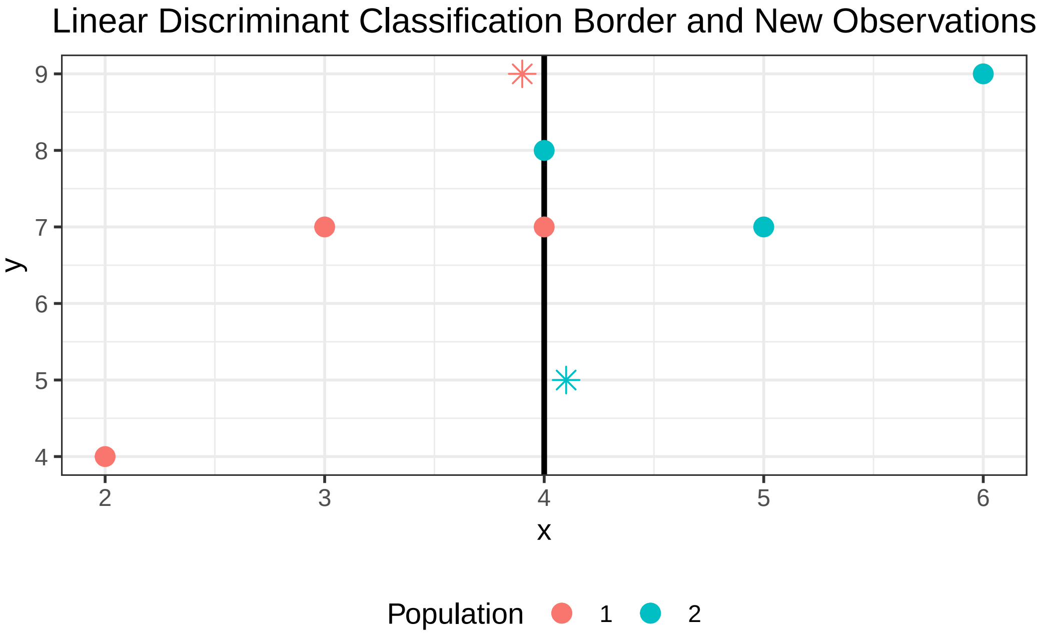
the linear discriminant function, and plot the two new observations $(4.1, 5)$ **and** $(3.9, 9)$**. (Assuming equal covariance** $\Sigma_1 = \Sigma_2$**, equal costs, and equal priors.)**

We showed in Exercise 4(a) that $(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T S_{pool}^{-1}\mathbf{x} = \begin{bmatrix}-2 & 0\end{bmatrix}\mathbf{x}$. We also showed in Exercise 4(b) that $\hat{m} = \frac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T S_{pool}^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2) = -8$. Thus, our classification border is

$$\begin{aligned}
&\hat{y} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T S_{pool}^{-1}\mathbf{x} - \hat{m} = 0 \\
&\Leftrightarrow \begin{bmatrix}-2 & 0\end{bmatrix}\begin{bmatrix}x \\ y\end{bmatrix} - (-8) = 0 \\
&\Leftrightarrow -2x + 0y = 8 \\
&\Leftrightarrow x = 4,
\end{aligned}$$

where we denote the coordinates of a vector $\mathbf{x}$ by $x = (x, y)$.

We also showed in Exercise 4(b) that, using this classification border, we should allocate $\mathbf{x}_1^* = (4.1, 5)$ to π_2 and $\mathbf{x}_2^* = (3.9, 9)$ to π_1. In the plot below, we include the classification border at $x = 4$, as well as the new datapoints $\mathbf{x}_1^*$ and $\mathbf{x}_2^*$, each marked with a $\star$ shape. As we can see in the plot, $\mathbf{x}_1^* = (4.1, 5)$ clearly belongs on the right side of $x = 4$, with the other points in π_2 (blue), while $\mathbf{x}_2^* = (3.9, 9)$ clearly belongs on the left side of $x = 4$, with the other points in π_1 (red). This is a nice visual corroboration of our finding in Exercise 4(b).



4.2 Part (b)

(Conceptual plots only. No calculations or usage of software which would use different, more sophisticated criteria.)

Start with another base plot containing points from samples X_1, X_2.

i. **Add the linear classifier obtained by the method of SVM.**

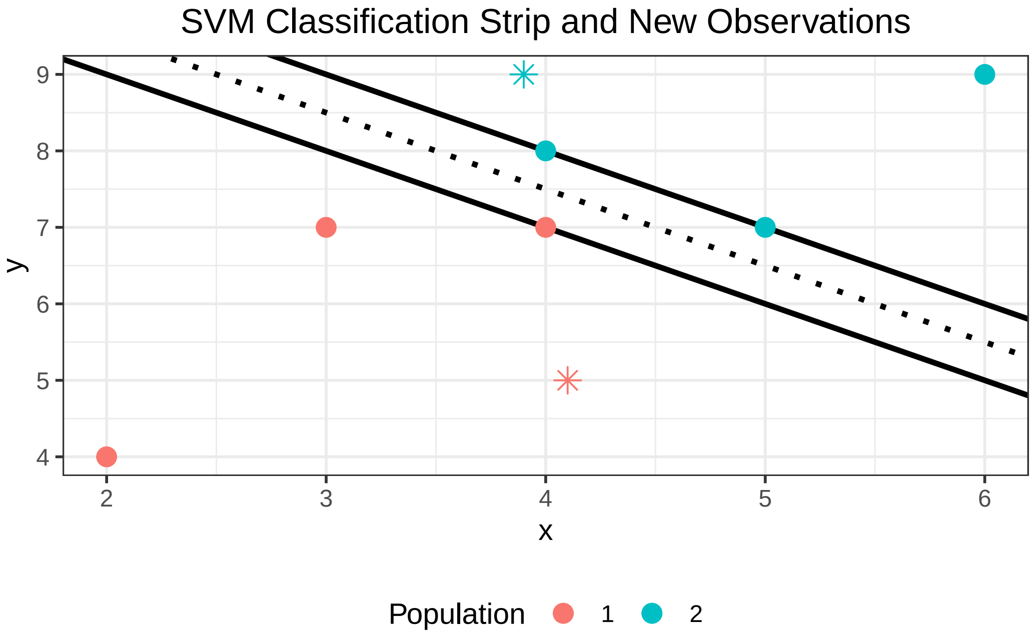
ii. **Identify the supporting vectors.**

iii. **Plot the two observations $(4.1, 5)$ and $(3.9, 9)$. To which classes are they assigned by linear SVM?**

In the plot below, we include the SVM classification strip, as well as the new datapoints $\mathbf{x}_1^* = (4.1, 5)$ and $\mathbf{x}_2^* = (3.9)$, each marked with a $\star$ shape.

By inspection, the SVM linear classification strip is bounded between the line $y = -x + 11$ and the line $y = -x + 12$. The lower side of the strip (the line $y = -x + 11$) is supported by the observation $(4, 7) \in X_1$. The upper side of the strip (the line $y = -x + 12$) is supported by the observations $(5, 7), (4, 8) \in X_2$. Thus, the full set of support vectors is $\{(4, 7), (5, 7), (4, 8)\}$.

Strikingly, we see that allocations of $\mathbf{x}_1^* = (4.1, 5)$ and $\mathbf{x}_2^* = (3.9, 9)$ are the opposites of what they were in Part (a). Now, $\mathbf{x}_1^* = (4.1, 5)$ is below the classification strip, so is allocated to π_1 along with the remaining red points. Moreover, $\mathbf{x}_2^* = (3.9, 9)$ is now above the classification strip, so is allocated to π_2 along with the remaining blue points.



4.3 Part (c)

Compare Parts (a) and (b). Which classifier do you prefer? Your reasons?

Strikingly, our analyses in Parts (a) and (b) allocated the points $\mathbf{x}_1^* = (4.1, 5)$ and $\mathbf{x}_2^* = (3.9, 9)$ to opposite populations. In particular, whereas our linear discriminant analysis in Part (a) allocated $\mathbf{x}_1^*$ to π_2, our SVM analysis in Part (b) allocated $\mathbf{x}_1^*$ to π_1. Moreover, whereas our linear discriminant analysis in Part (a) allocated $\mathbf{x}_2^*$ to π_1, our SVM analysis in Part (b) allocated $\mathbf{x}_2^*$ to π_2.

I find the SVM classifier explored in Part (b) far more compelling for these data than the linear discriminant classifier explored in Part (a).

Firstly, the linear discriminant classifier simply splits the x-y plane into a “left” half and a “right” half, while the SVM classifier exploits the “diagonal” arrangement of the data to split the x-y plane into a “northeast” half and a “southwest” half. So, the SVM classifier seems to take better advantage of the arrangement of the data than the linear discriminant classifier.

Moreover, the SVM classifier creates a clean split in the training data, with all training points in Populations 1 and 2 separated by a strip of width 1. The linear discriminant classifier, on the other hand, fails to establish such a clean split, and even has two points—one from each population—sitting on the classification boundary. Since the SVM classifier makes the training data from the two populations truly distinguishable, it seems like the better tool here.

4.4 Part (d)

In Part (d) of Question 4, under the assumptions of equal costs, equal priors, $\Sigma_1 \neq Sigma_2$, and bivariate normal distributions for the two classes, two new observations $(4.1, 5)$ and $(3.9, 9)$ were classified by the quadratic function $d(x)$.

Now for this visualization exercise...

4.4.1 Part (i)

On another base plot, plot $d(\mathbf{x}) = d(x, y) = 0$, the class border(s) by the quadratic discriminant function. (Hint: It is numerically easier to use $3d(x, y) = 0$, which is a degenerate conic section equation, now with integer coefficients.)

Recall that in Exercise 4(d), we found that

$$d(\mathbf{x}) = -\frac{2}{3}\mathbf{x}^T \begin{bmatrix} 2 & -1 \\ -1 & 0 \end{bmatrix} \mathbf{x} + \frac{4}{3} \begin{bmatrix} -1 & -4 \end{bmatrix} \mathbf{x} + \frac{80}{3}.$$

Thus,

$$\begin{aligned}
3d(\mathbf{x}) &= -2\mathbf{x}^T \begin{bmatrix} 2 & -1 \\ -1 & 0 \end{bmatrix} \mathbf{x} + 4 \begin{bmatrix} -1 & -4 \end{bmatrix} \mathbf{x} + 80 \\
&= -2 \begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} 2 & -1 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + 4 \begin{bmatrix} -1 & -4 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + 80 \\
&= -2 \begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} 2x - y \\ -x + 0y \end{bmatrix} + 4(-x - 4y) + 80 \\
&= -2[x(2x-y) + y(-x)] - 4x - 16y + 80 \\
&= -2[2x^2 - xy - xy] - 4x - 16y + 80 \\
&= -4x^2 + 4xy - 4x - 16y + 80
\end{aligned}$$

Hence,

$$\begin{aligned}
& d(\mathbf{x}) = 0 \\
\Leftrightarrow\ & 3d(\mathbf{x}) = 0 \\
\Leftrightarrow\ & -4x^2 + 4xy - 4x - 16y + 80 = 0.
\end{aligned}$$

This is the equation of a degenerate hyperbola, plotted below. Note that the classification region for the first population, π_1, is

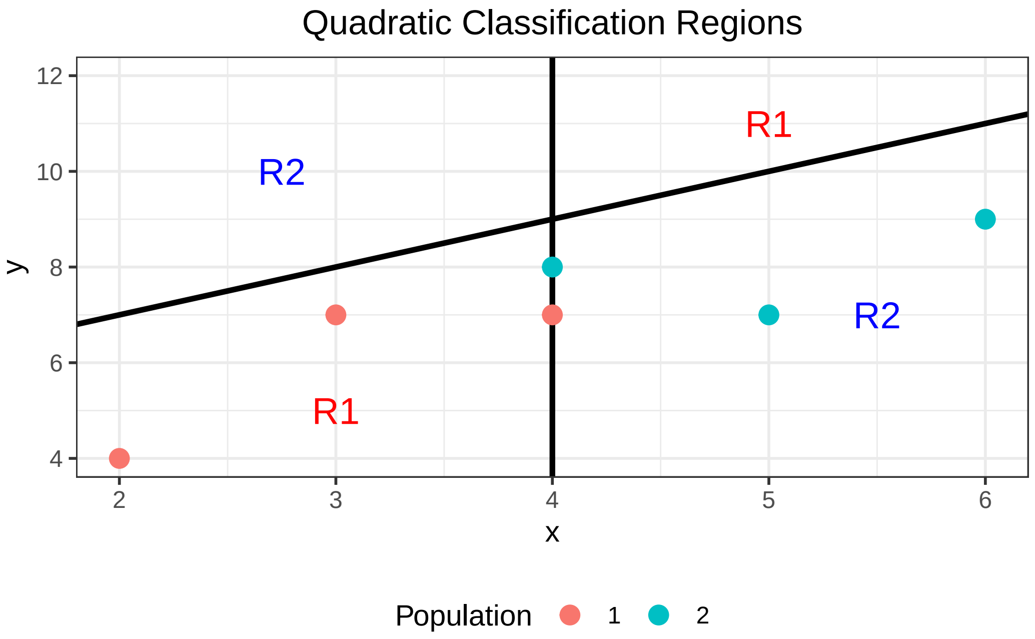
$$\begin{aligned}
R_1 &= \{\mathbf{x} : d(\mathbf{x}) \geq 0\} \\
&= \{(x,y) : -4x^2 + 4xy - 4x - 16y + 80 \geq 0\},
\end{aligned}$$

which consists of a "northeast" piece and a "southwest" piece

Similarly, the classification region for the second population, π_2, is

$$\begin{aligned}
R_1 &= \{\mathbf{x} : d(\mathbf{x}) < 0\} \\
&= \{(x,y) : -4x^2 + 4xy - 4x - 16y + 80 < 0\},
\end{aligned}$$

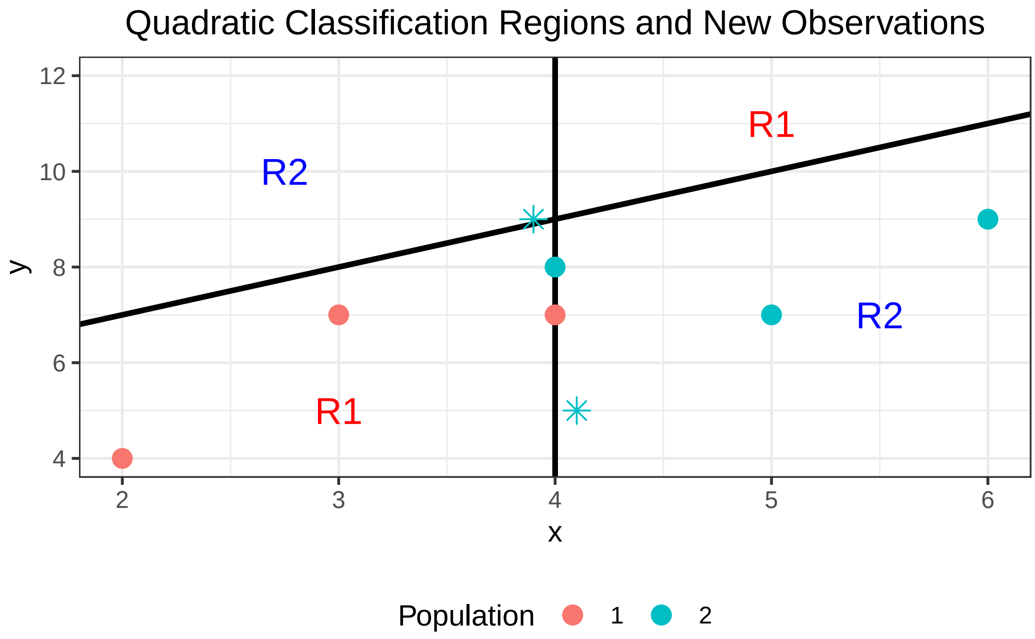
which consists of a "southeast" piece and a "northwest" piece. These regions are included in the plot below.



4.4.2 Part (ii)

Plot the two observations $(4.1, 5)$ **and** $(3.9, 9)$**. Which classes are they assigned into by the quadratic discriminant rule?**

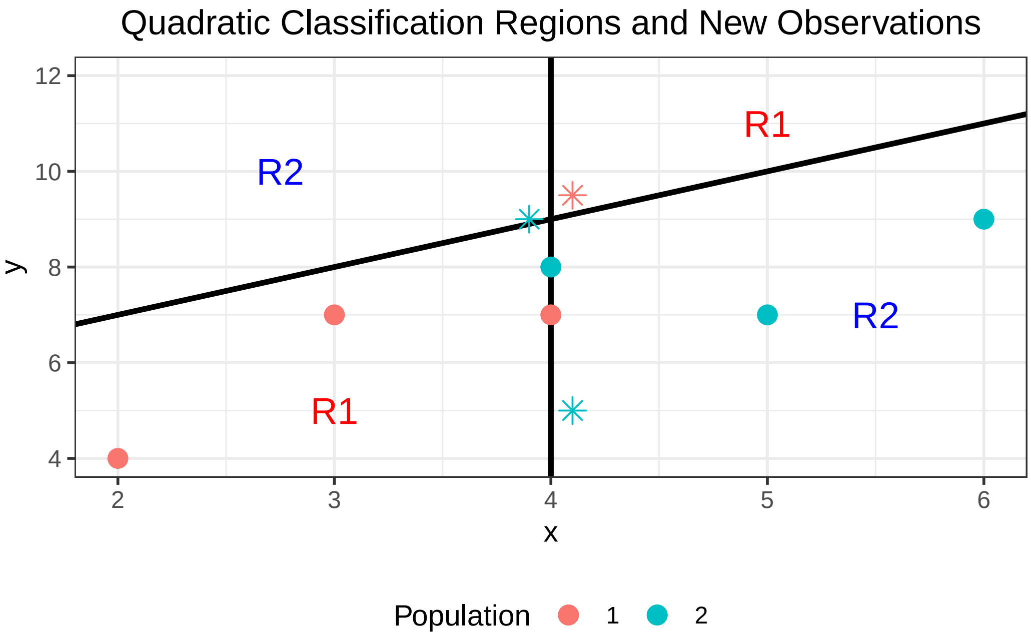
We showed in Exercise 4(d) that, using this classification rule, we should allocate both $\mathbf{x}_1^* = (4.1, 5)$ and $\mathbf{x}_2^* = (3.9, 9)$ to the second population, π_2. We add these points to the plot below, each marked with a $\star$ shape. We see that $\mathbf{x}_1^*$ falls in the "southeast" piece of R_2, while $\mathbf{x}_2^*$ falls in the "northwest" piece of R_2.



4.4.3 Part (iii)

Classify another new observation $(4.1, 9.5)$**. Is the classification reasonable?**

We add the point $\mathbf{x}_3^* = (4.1, 9.5)$ in the plot below. It falls in the "northeast" piece of R_1, hence is classified to the first population, π_1. This is admittedly somewhat strange: none of the observations from the training data X_1 fell in this "northeast" piece, so $\mathbf{x}_3^*$ is actually not very close to any other points it is classified with. In fact, it is actually closest to points that are allocated to π_2, such as $(3.9, 9)$ and $(4, 8)$.



4.4.4 Part (iv)

Usually, we go for higher order (such as from linear to quadratic) and less assumptions for more refined or "better" classification rules. Is the quadratic rule here better than the linear discriminant clasifier in Part (a)? Can you explain the reason by the pattern of the quadratic classification regions?

In this case, the quadratic classification rule actually seems worse than the linear discriminant classifier explored in Part (a). For example, as we just showed in Part (d)(iii), the point $\mathbf{x}_3^* = (4.1, 9.5)$ was allocated to π_1, even though it is actually closer to many points in R_2 than it is to some points in R_1. The linear discriminant examined in Part (a) would have assigned $\mathbf{x}_3^* = (4.1, 9.5)$ to π_2 since $4.1 > 4$, which seems to be a more appropriate allocation.

Because our quadratic classification rule has the form of a degenerate hyperbola, it partitions the x-y plane into four pieces, with classification regions comprising two "kitty-corner" pieces. Since all of the training points in π_1 are relatively close to one another, and are generally "southwest" of the training points in π_2, it is natural that the first piece of R_1 is the "southwest" piece of the plane. This means that the other piece of R_1 must be the "kitty-corner" piece: the "northeast" piece — even though there is no evidence from the training data that points in π_1 fall in this region. The same could be said for R_2: it is natural that the first piece of this region is the "southeast" piece of the plane, but this means that the other portion of R_2 must be the "northeast" piece — even though there are no training points from X_2 in this region.

Perhaps if the training data were larger, or if if the points in the two populations were more intermingled in the x-y plane, the quadratic classification rule would be more helpful than it was here.