13 - Linear Classification

Junvie Pailden

SIUE, F2017, Stat 589

October 05, 2017

Introduction

Discrimination and classification are multivariate techniques concerned with separating distinct sets of objects (or observations) and with allocating new objects (observations) to previously defined groups.

Goal

- To describe, either graphically (in three or fewer dimensions) or algebraically, the differential features of objects (observations) from several known collections (populations). We try to find "discriminants" whose numerical values are such that the collections are separated as much as possible.
- To sort objects (observations) into two or more labeled classes. The emphasis is on deriving a rule that can be used to optimally assign new objects to the labeled classes.

Iris Data Species

Observations made on four attributes (sepal length/width, petal length/width) of each of three types of irises.

* Top right: Setosa
* Bottom left: Versicolor
* Bottom right: Virginica

Types of Iris Flower

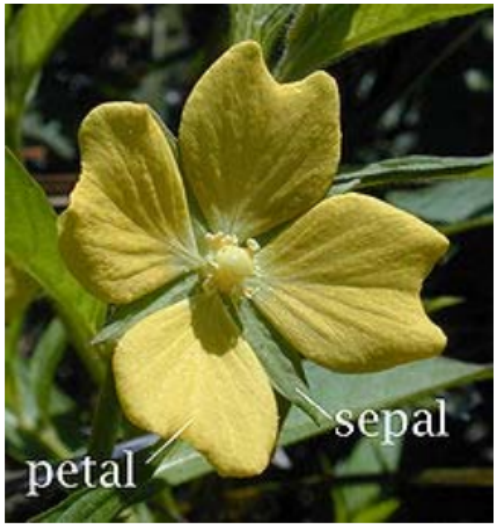


Figure 1

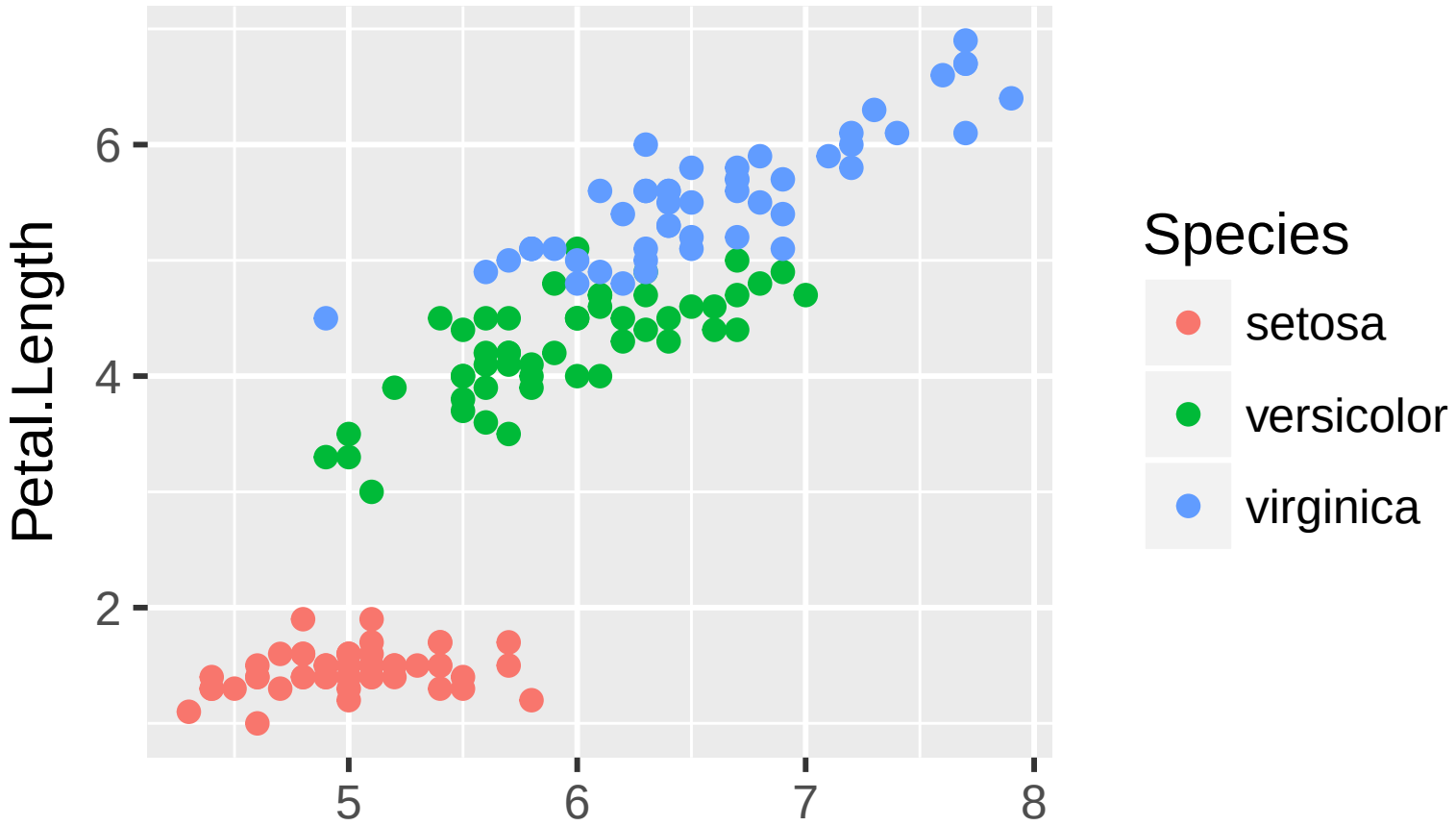
Iris Data

```
# iris data is included in base R
str(iris)
```

```
# 'data.frame': 150 obs. of  5 variables:
#  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9
#  $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3
#  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4
#  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2
#  $ Species     : Factor w/ 3 levels "setosa","versicolor"
```

Iris Data

```
library(tidyverse)
iris %>% ggplot(aes(x = Sepal.Length, y = Petal.Length,
                    colour = Species)) +
  geom_point()
```



Classification for Two Populations

1. Separating two classes of objects.
2. Assigning a new object to one of two classes (or both).

- Label the classes (populations) as π_1 and π_2.
- Objects are classified on basis of of measurements on p associated random variables $\mathbf{X}' = [X_1, X_2, \ldots, X_p]$.
- The observed values of $\mathbf{X}$ differ to some extent from one class (population) to the other.
- Values from first class as being the population of $\mathbf{x}$ values for π_1 and those from the second class as population of $\mathbf{x}$ for π_2.

Classification for Two Populations (cont)

- Let $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ be the pdf's associated with $\mathbf{X}$ for pop'ns π_1 and π_2.
- Let $\Omega = \{\mathbf{x}\}$ be the sample space - collection of all possible observations $\mathbf{x}$.
- Let R_i be that set of $\mathbf{x}$ values for w/c we classify objects as $\pi_i,\ i = 1, 2$.
- Note, $R_2 = \Omega - R_1$ and $R_1 \bigcap R_2 = \emptyset$.

Conditional Probability of Misclassification

- Conditional probability, $P(2|1)$, of classifying an object as π_2 when, in fact, it is from π_1 is

$$P(2|1) = P(\mathbf{X} \in R_2|\pi_1) = \int_{R_2} f_1(\mathbf{x})d\mathbf{x}.$$

- Similarly, conditional probability, $P(1|2)$, of classifying an object as π_1 when, in fact, it is from π_2 is

$$P(1|2) = P(\mathbf{X} \in R_1|\pi_2) = \int_{R_1} f_2(\mathbf{x})d\mathbf{x}.$$

Let p_i be the prior probability of π_i, $i = 1, 2$, where $p_1 + p_2 = 1$.

$$
\begin{aligned}
P(\text{obs is correctly classified as } \pi_1) &= P(\text{obs comes from } \pi_1 \text{ and} \\
&\quad \text{is correctly classified as } \pi_1) \\
&= P(\mathbf{X} \in R_1 | \pi_1) P(\pi_1) = P(1|1) p_1 \\
P(\text{obs is misclassified as } \pi_1) &= P(\text{obs comes from } \pi_2 \text{ and} \\
&\quad \text{is misclassified as } \pi_1) \\
&= P(\mathbf{X} \in R_1 | \pi_2) P(\pi_2) = P(1|2) p_2
\end{aligned}
$$

$$
\begin{aligned}
P(\text{obs is correctly classified as } \pi_2) &= P(\text{obs comes from } \pi_2 \text{ and} \\
&\quad \text{is correctly classified as } \pi_2) \\
&= P(\mathbf{X} \in R_2 | \pi_2) P(\pi_2) = P(2|2) p_2 \\
P(\text{obs is misclassified as } \pi_2) &= P(\text{obs comes from } \pi_1 \text{ and} \\
&\quad \text{is misclassified as } \pi_2) \\
&= P(\mathbf{X} \in R_2 | \pi_1) P(\pi_1) = P(2|1) p_1
\end{aligned}
$$

Costs of Misclassification: Cost Matrix

	Classify as	π_1	π_2
True	π_1	0	c(2|1)
Population	π_2	c(1|2)	0

The average or expected cost of misclassification (ECM) is

$$\begin{aligned}
\text{ECM} &= c(2|1)P(2|1)p_1 + c(1|2)P(1|2)p_2 \\
&= c(2|1)p_1 \int_{R_2} f_1(\mathbf{x})d\mathbf{x} + c(1|2)p_2 \int_{R_1} f_2(\mathbf{x})d\mathbf{x} \\
&= \int_{R_1} \left[c(1|2)p_2 f_2(\mathbf{x}) - c(2|1)p_1 f_1(\mathbf{x})\right] d\mathbf{x} + c(2|1)p_1
\end{aligned}$$

A reasonable classification rule should have an ECM as small, or nearly as small, as possible.

Result 11.1 Minimum Expected Cost Regions

The regions R_1 and R_2 that minimize the ECM are defined by the values $\mathbf{x}$ for which the following inequalities hold:

$$R_1 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)$$

$$\text{(density ratio)} \geq \text{(cost ratio)(prior prob ratio)}$$

$$R_2 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)$$

$$\text{(density ratio)} < \text{(cost ratio)(prior prob ratio)}$$

Special Cases of Minimum Expected Cost Regions

- $p_2/p_1 = 1$ (equal prior probabilities)

$$R_1 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{c(1|2)}{c(2|1)} \ , \quad R_2 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \frac{c(1|2)}{c(2|1)}$$

- $c(1|2)/c(2|1) = 1$ (equal misclassification costs)

$$R_1 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{p_2}{p_1} \ , \quad R_2 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \frac{p_2}{p_1}$$

- $p_2/p_1 = c(1|2)/c(2|1) = 1$ (equal prior prob's and equal misclassification costs)

$$R_1 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq 1 \ , \quad R_2 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < 1$$

- When prior probabilities are unknown, they are often taken to be equal.
- If the misclassification cost ratio is indeterminate, it is usually taken to be unity.

Classification with Two Multivariate Normal Populations

- Classification procedures based on normal populations predominate in statistical practice because of their simplicity and reasonably high efficiency across a wide variety of population models.
- We now assume that $f_1(\mathbf{x}) \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $f_1(\mathbf{x}) \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ are multivariate normal densities.
- When $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$
- Suppose that the joint densities of $\mathbf{X}' = [X_1, X_2, \ldots, X_p]$ for popn's π_1 and π_2 are given by

$$f_i(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)\right], \quad i = 1, 2$$

Suppose that $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$, and $\boldsymbol{\Sigma}$ are known. Then, after cancelling the terms $(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}$, the minimum ECM regions become

$$R_1 : \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)+\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)\right] \geq \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)$$

$$R_2 : \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)+\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)\right] < \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)$$

- Allocate new observation $\mathbf{x}_0$ to π_1 if

$$(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}\mathbf{x}_0-\frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2) \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right]$$

- Allocate $\mathbf{x}_0$ to π_2 otherwise.
- When $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$, and $\boldsymbol{\Sigma}$ are unknown, as in most practical cases, we replace the parameters by their sample counterparts.

- Suppose that we have n_i observations (IID copies) of $\mathbf{X}' = [X_1, X_2, \ldots, X_p]$ from pop'n π_i, $i = 1, 2$, with $n_1 + n_2 - 2 \geq p$. The data matrices are

$$\underset{(n_1 \times p)}{\mathbf{X}_1} = \begin{bmatrix} \mathbf{x}'_{11} \\ \mathbf{x}'_{12} \\ \vdots \\ \mathbf{x}'_{1n_1} \end{bmatrix}, \quad \underset{(n_2 \times p)}{\mathbf{X}_2} = \begin{bmatrix} \mathbf{x}'_{21} \\ \mathbf{x}'_{22} \\ \vdots \\ \mathbf{x}'_{2n_1} \end{bmatrix}$$

$$\underset{(p \times 1)}{\bar{\mathbf{x}}_i} = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{x}_{ij}, \ \ i = 1, 2$$

$$\underset{(p \times p)}{\mathbf{S}_i} = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)', \ \ i = 1, 2$$

$$\mathbf{S}_{\text{pooled}} = \left[\frac{n_1 - 1}{(n_1 - 1) + (n_2 - 1)} \right] \mathbf{S}_1 + \left[\frac{n_2 - 1}{(n_1 - 1) + (n_2 - 1)} \right] \mathbf{S}_2$$

- $\mathbf{S}_{\text{pooled}}$ is unbiased estimate of $\boldsymbol{\Sigma}$ if the data matrices $\mathbf{X}_1$ and $\mathbf{X}_2$ are random samples from the pop'ns π_1 and π_2.

Sample Classification Rule

The Estimated Minimum ECM Rule for Two Populations

1. Allocate new observation $\mathbf{x}_0$ to π_1 if

$$(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}\mathbf{x}_0 - \frac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2) \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right]$$

2. Allocate $\mathbf{x}_0$ to π_2 otherwise.

- Once parameter estimates are inserted for the corresponding unknown pop'n quantities, there is no assurance that the resulting rule will minimize the ECM.
- However, it seems reasonable to expect that it should perform well if the sample sizes are large.

- Given a new observation $\mathbf{x}_0$, let

$$\begin{aligned}
\hat{y} &= (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}\mathbf{x}_0 = \hat{\mathbf{a}}'\mathbf{x}_0 \\
\bar{y}_1 &= (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}\bar{\mathbf{x}}_1 = \hat{\mathbf{a}}'\bar{\mathbf{x}}_1 \\
\bar{y}_2 &= (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}\bar{\mathbf{x}}_2 = \hat{\mathbf{a}}'\bar{\mathbf{x}}_2 \\
\hat{m} &= \frac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2) = \frac{1}{2}(\bar{y}_1 + \bar{y}_2)
\end{aligned}$$

- If $\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right) = 1$, then $\ln(1) = 0$ (equal prior and equal costs), then estimated minimum ECM rule for two normal populations implies allocating new observation $\mathbf{x}_0$ to π_1 if $\hat{y} \geq \hat{m}$; allocate to π_2 otherwise.

Exercise 11.1

```
# Read the data sets of exercise 11.1 into R:
X1 <- matrix(c(3,7,2,4,4,7), nrow = 3, ncol = 2,
             byrow = T)
X2 <- matrix(c(6,9,5,7,4,8), nrow = 3, ncol = 2,
             byrow = T)
```

Sample ECM Two Population Function

```
ECM.two.popn <- function(X1, X2){
  n1 <- nrow(X1)
  n2 <- nrow(X2)
  xbar1 <- colMeans(X1)
  xbar2 <- colMeans(X2)
  S1 <- cov(X1)
  S2 <-cov(X2)
  Sp <- ((n1-1)/(n1+n2-2))*S1+((n2-1)/(n1+n2-2))*S2
  ta <- t(xbar1-xbar2)%*%solve(Sp)
  list(a = t(ta), m = 0.5*(ta%*%xbar1+ta%*%xbar2))
}
```

Exercise 11.1 I

```
# a)  Compute the vector a for the linear discriminant
# function given by (11-19) in J&W
# b)  Compute the scalar m given by (11-20):
(out1 <- ECM.two.popn(X1, X2))
```

```
# $a
#      [,1]
# [1,]   -2
# [2,]    0
#
# $m
#      [,1]
# [1,]   -8
```

Exercise 11.1 II

```
# Classify the new obs x0 = c(2,7) to pi1
# if t(a)%*%c(2,7) > m
t(out1$a)%*%c(2,7)
```

```
#      [,1]
# [1,]   -4
```

```
# since -4 > -8, then c(2,7) is
# classified as belonging to pi1
# Check original observations
# Classified to pi1?
X1
```

Exercise 11.1 III

```
#      [,1] [,2]
# [1,]    3    7
# [2,]    2    4
# [3,]    4    7
```

```
t(out1$a)%*%t(X1) >= -8
```

```
#      [,1] [,2] [,3]
# [1,] TRUE TRUE TRUE
```

```
X2
```

```
#      [,1] [,2]
# [1,]    6    9
# [2,]    5    7
# [3,]    4    8
```

Exercise 11.1 IV

```
t(out1$a)%*%t(X2) >= -8
```

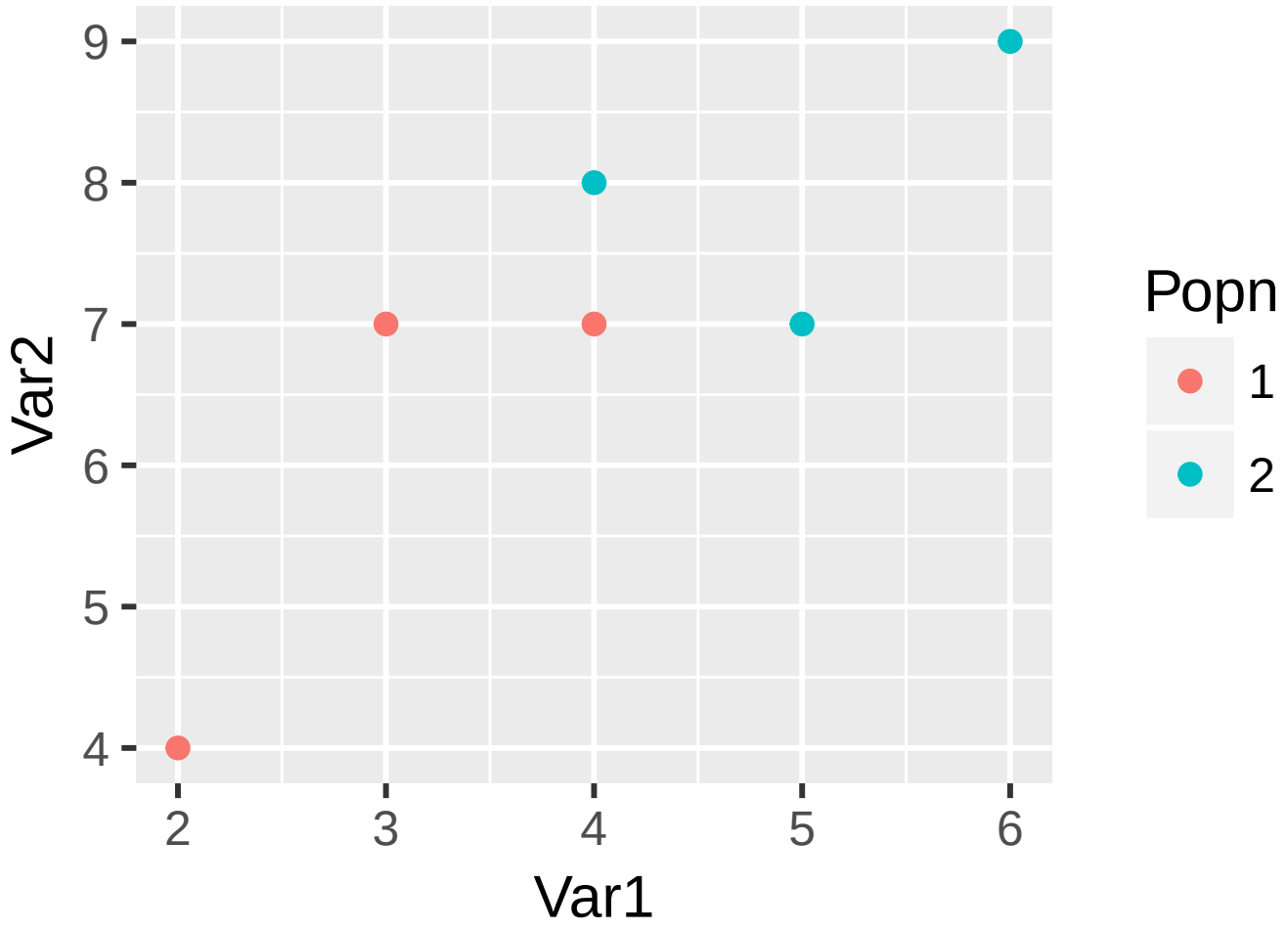
```
#        [,1]  [,2] [,3]
# [1,] FALSE FALSE TRUE
```

Plots - Exercise 11.1

```
X <- data.frame(c(rep(1,3), rep(2,3)), rbind(X1, X2))
colnames(X) <- c("Popn", "Var1", "Var2")
X$Popn <- as.factor(X$Popn)
X
```

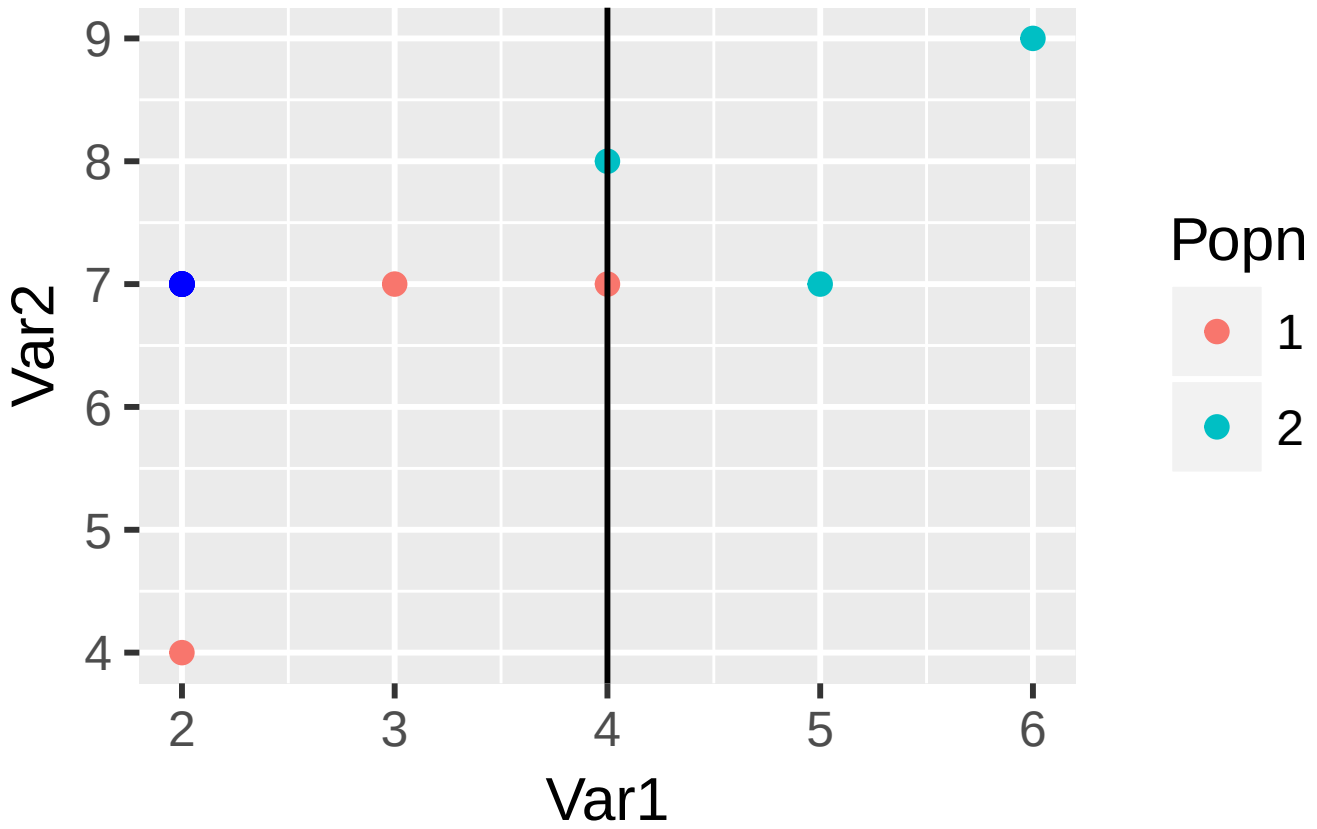
```
#   Popn Var1 Var2
# 1    1    3    7
# 2    1    2    4
# 3    1    4    7
# 4    2    6    9
# 5    2    5    7
# 6    2    4    8
```

```
p1 <-  X %>% ggplot(aes(x = Var1, y = Var2,
                        colour = Popn)) +
  geom_point()
p1
```



Classify New Point (2,7)

```
# add new point (2,7)
p1 + geom_point(aes(x = 2, y = 7),
                          colour = "blue" ) +
     geom_vline(xintercept = out1$m/out1$a[1])
```



Iris Data I

```
str(iris)
```

```
# 'data.frame': 150 obs. of  5 variables:
#  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9
#  $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3
#  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4
#  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2
#  $ Species     : Factor w/ 3 levels "setosa","versicolor"
```

Iris Data II

```
# create data subset for Species = versicolor,
# columns Sepal.Length and Petal.Length
X.versicolor <- iris %>%
  filter(Species == "versicolor") %>%
  select(Sepal.Length, Petal.Length)
# same for Specifies = virginica
X.virginica <- iris %>%
  filter(Species == "virginica") %>%
  select(Sepal.Length, Petal.Length)

(out.iris <- ECM.two.popn(X.versicolor, X.virginica))
```

Iris Data III

```
# $a
#                [,1]
# Sepal.Length  4.9
# Petal.Length -9.5
#
# $m
#      [,1]
# [1,]  -16
```

Iris Data IV

```
# Classifier Function
belong.iris <- function(X){
  if(t(out.iris$a)%*%X >= out.iris$m){
    type = "versicolor"
  } else {
    type = "virginica"
  }
  return(type)
}

# Check the classifier on existing observations
res1 <- apply(X.versicolor, 1, belong.iris)
table(res1)
```

Iris Data V

```
# res1
# versicolor  virginica
#         47          3
```

```
# misclassified to pi2
sum(res1 == "virginica")
```

```
# [1] 3
```

```
res2 <- apply(X.virginica, 1, belong.iris)
table(res2)
```

```
# res2
# versicolor  virginica
#          3         47
```

Iris Data VI

```
# misclassified to pi3
sum(res2 == "versicolor")
```

```
# [1] 3
```

Linear discrimator between Species Versicolor and Virginica

```
# select variables, filter species
iris %>% select(Sepal.Length, Petal.Length, Species) %>%
  filter(Species != "setosa") %>%
  ggplot(aes(x = Sepal.Length, y = Petal.Length,
             colour = Species)) +
  geom_point() +
  # draw the linear discrimination border
  geom_abline(intercept = out.iris$m/out.iris$a[2],
              slope = -out.iris$a[1]/out.iris$a[2])
```

Linear discrimator between Species Versicolor and Virginica