Predicting Iris Species

Jordan Schupbach

September 15, 2017

1 Introduction

The famous (Fisher 1936) (or (Anderson 1936)) iris data set gives the measurements (in centimeters) of the sepal length and width, and petal length and width for 50 flowers, each from one of three species of iris. The species are *I. setosa*, *I versicolor*, and *I. virginica*. We provide images for each of the three species in Figure 1. Researchers are interested in whether they can identify the species from the set of the four measurements taken on the plants. For this study, the focus is only on predicting *I. Setosa* versus *I. Versicolor.*

1.1 Description of Data

This dataset has 1 categorical variable (species) and 4 quantitative variables (sepal length and width, and petal length and width; all measured in centimeters). Again, the researchers are interested in predicting species from the four measurements. We provide the first six lines of the dataset in Table 1 and a five-number summary of the data in Table 2. We also may want to subset our data by species and look at summary statistics. Means for the two species of interest are given in Table 3.

These data were gathered by Dr. Edgar Anderson. The *Iris setosa* and *Iris versicolor* were found growing together in the same colony. The third species, *Iris virginica*, was from a different colony, as it would be a "circumstance which might considerably disturb both the mean values and their variabilities."

2 Statistical Methods Used

To classify the two species of interest, *I. setosa* and *I. versicolor*, we use the method of linear discriminant analysis following (Fisher 1936). One assumption of the method is that the covariates be independently normally distributed. We check the assumption of normality by plotting histograms of four variables grouped by species type in Figure 1. Further, we provide beanplots from the `beanplot` R package (Kampstra 2008), which assume a normal kernel density. The two figures would suggest the assumption of normality is reasonably met.

Linear discriminant analysis allows us to write an objective function of the form

$$X = \lambda_1 x_1 + \lambda_2 x_2 + \lambda_3 x_3 + \lambda_4 x_4 \tag{1}$$

for which we classify an observation as either *I. Setosa* or *I. Versicolor* depending on its value.

3 Summary of Statistical Findings

We provide means for the two groups in Table 3. The LDA was conducted in R (R Core Team 2017) using the `lda()` function from the `MASS` package (Venables and Ripley 2002). Coefficients of discrimination are given in Table 4. We see that we can write equation 1 as

$$X = -1.77x_1 - 0.3x_2 + 3.04x_3 + 2.14x_4 \tag{2}$$

Following (Fisher 1936), we can take the coefficient for sepal width to be unity (see Table 5) and rewrite equation 2 as

$$X = x_1 + 0.17x_2 - 1.71x_3 - 1.21x_4 \tag{3}$$

These equations can be used to score each observation on this new scale, and classify the observation as either *I. setosa* or *I. virginica*. Figure 5 shows the values of the scores for the two species according to equation 2. We can see from the plot that the scores do an impressive job in separating *I. setosa* and *I. virginica*. A confusion matrix is given in Table 6. We see that the equation can perfectly separate *I. setosa* plants from *I. virginica*.

4 Scope of Inference

Since we did not randomly assign any treatment to treatment groups, no causal inference can be inferred in this study. Further, since observations did not come from a random sample, we cannot infer to a larger population than that of the sample.

5 Appendix

5.1 Tables

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.10	3.50	1.40	0.20	setosa
2	4.90	3.00	1.40	0.20	setosa
3	4.70	3.20	1.30	0.20	setosa
4	4.60	3.10	1.50	0.20	setosa
5	5.00	3.60	1.40	0.20	setosa
6	5.40	3.90	1.70	0.40	setosa

Table 1: First 6 lines of the dataset

Table 2: Summary table of the four quantitative variables

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Sepal.Length	150	5.843	0.828	4.300	5.100	6.400	7.900
Sepal.Width	150	3.057	0.436	2.000	2.800	3.300	4.400
Petal.Length	150	3.758	1.765	1.000	1.600	5.100	6.900
Petal.Width	150	1.199	0.762	0.100	0.300	1.800	2.500

	Sepal.Width	Sepal.Length	Petal.Width	Petal.Length
setosa	3.43	5.01	0.25	1.46
versicolor	2.77	5.94	1.33	4.26

Table 3: Group means

	LD1
Sepal.Width	-1.77
Sepal.Length	-0.30
Petal.Width	3.04
Petal.Length	2.14

Table 4: Coefficients of linear discriminants

	LD1
Sepal.Width	1.00
Sepal.Length	0.17
Petal.Width	-1.71
Petal.Length	-1.21

Table 5: Coefficients of linear discriminants with first scaled to unity

	setosa	versicolor
setosa	50	0
versicolor	0	50

Table 6: Confusion matrix

5.2 Figures

Figure 1: Three species of iris: I. Setosa, I. Virginica, I. Versicolor

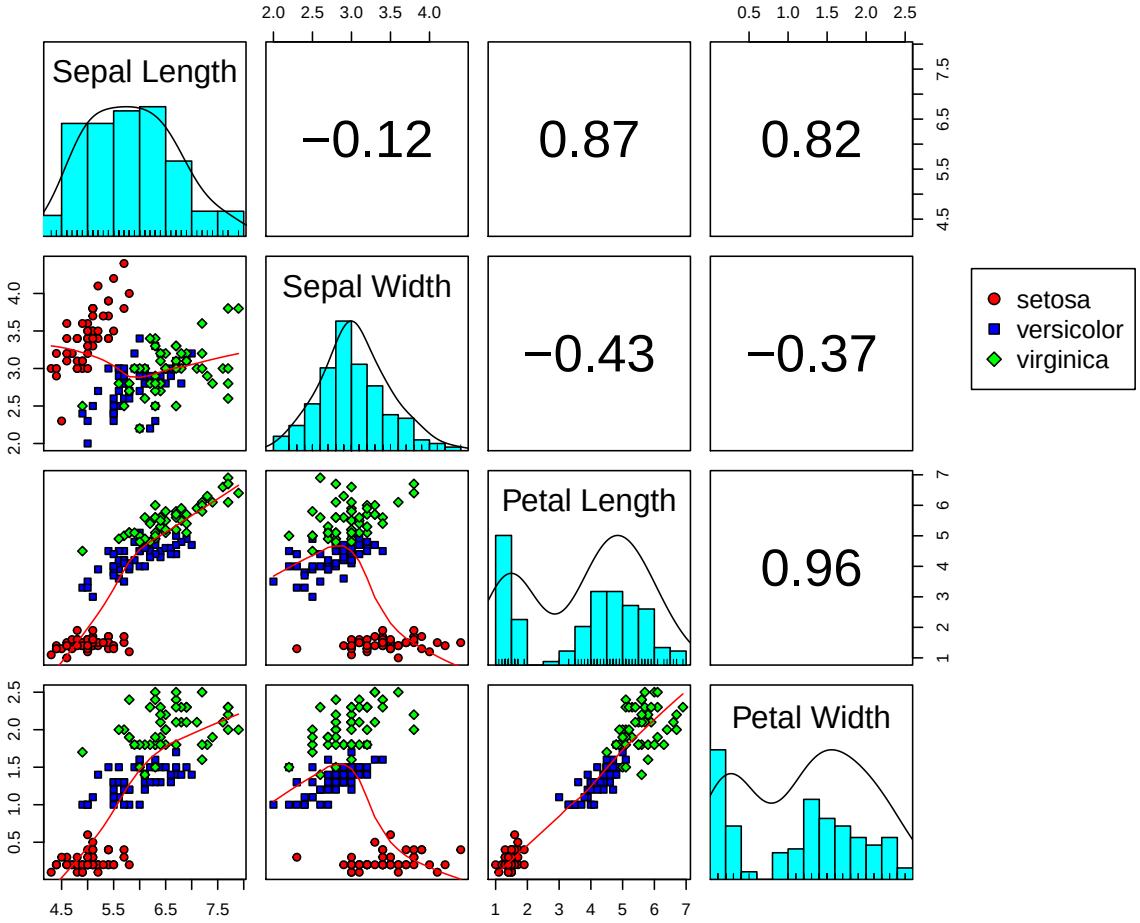

Figure 2: A matrix of scatterplots of the quantitative variables

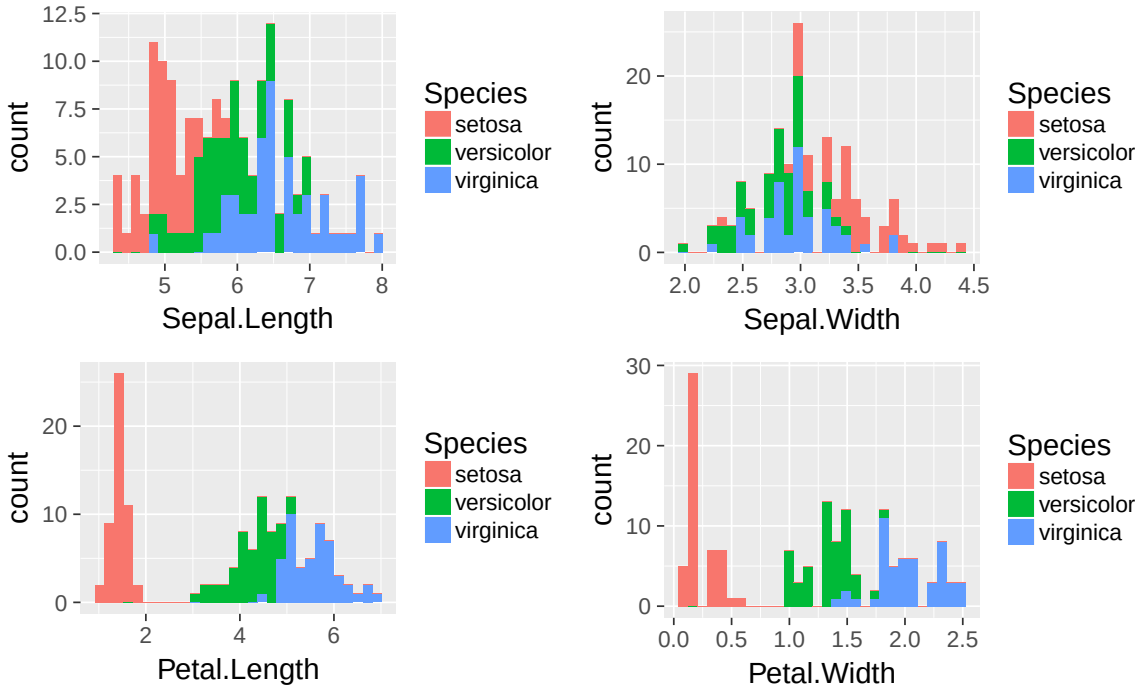

Figure 3: Histograms of covariates separated by species

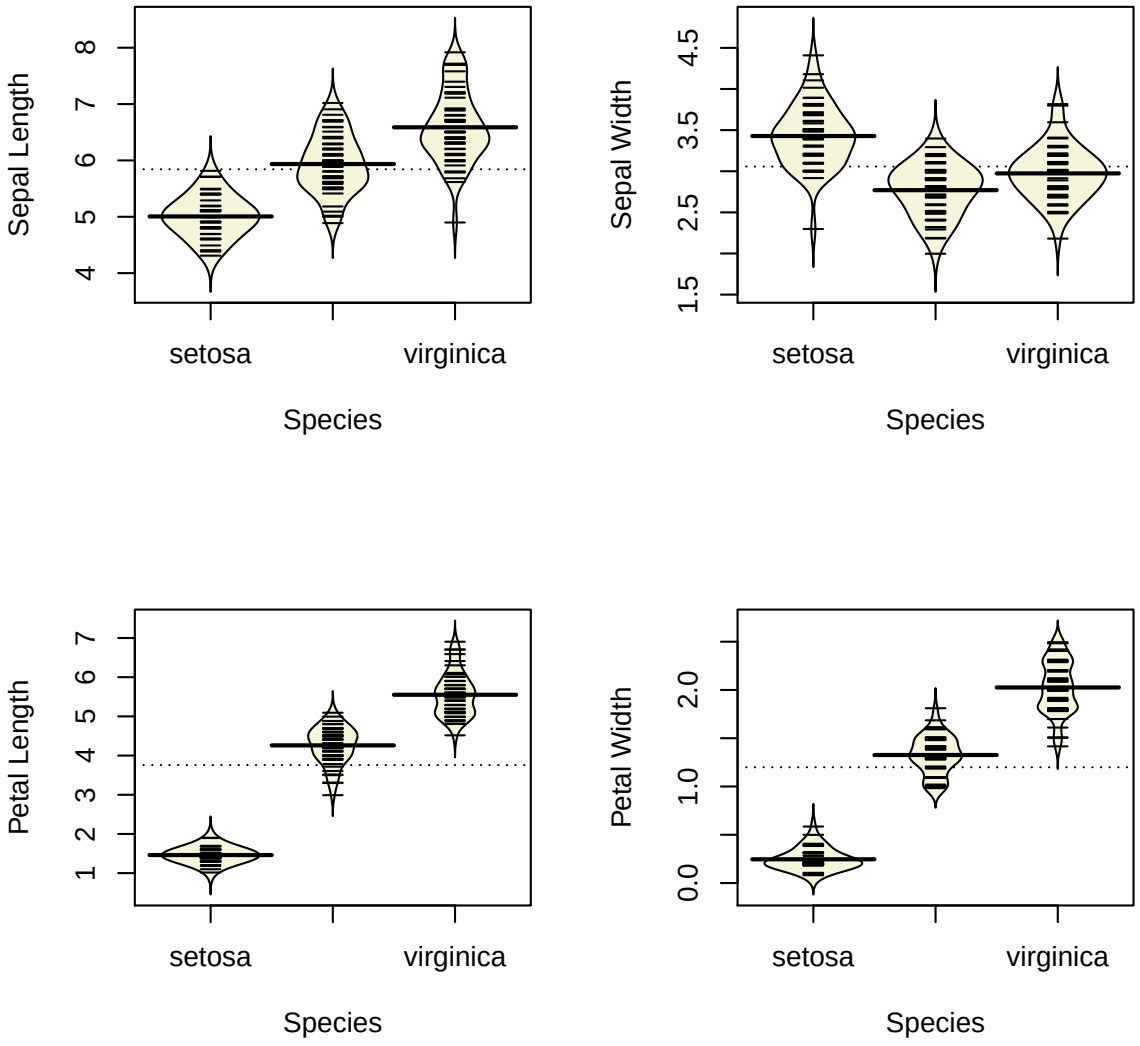

Figure 4: Beanplots of the four quantitative variables by species

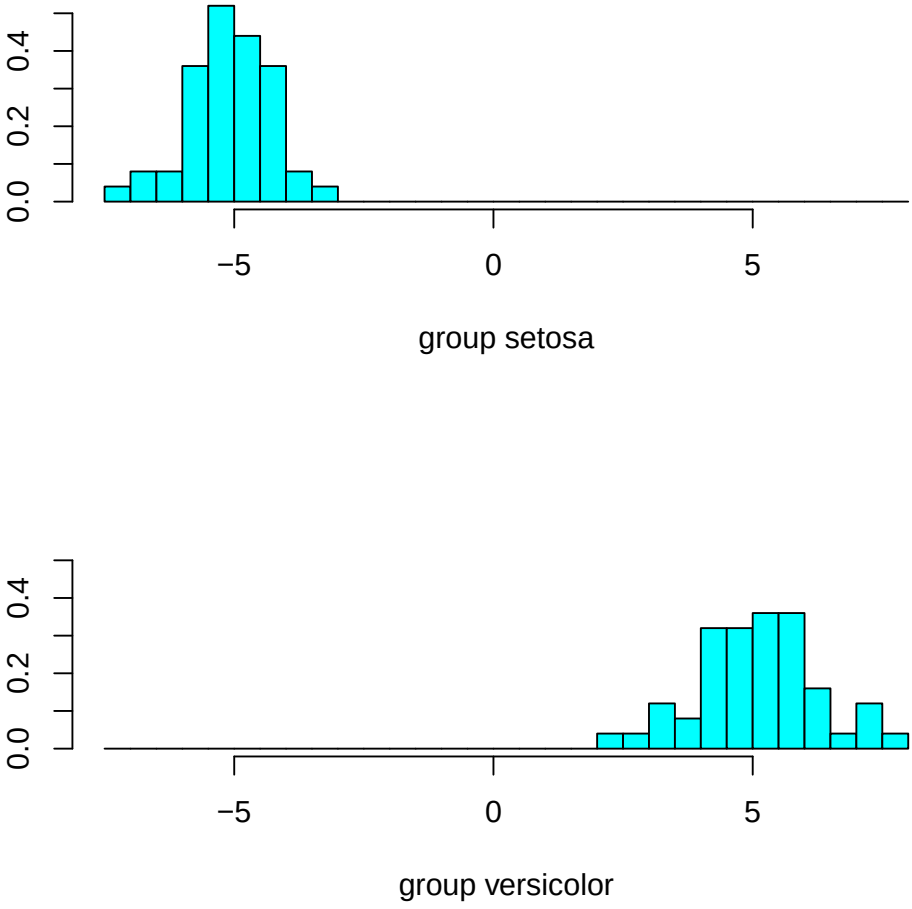

Figure 5: Linear discriminant scores for setosa and versicolor

5.3 R-code

```
# The dataframe iris is available in R's dataset package that is automatically loa
# Subset the data
iris_sub <- subset(iris, iris$Species == "setosa" |
                   iris$Species == "versicolor")
iris_sub$Species <- factor(iris_sub$Species)
head(iris_sub)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

```
# LDA
z <- lda(Species ~ Sepal.Width + Sepal.Length + Petal.Width +
         Petal.Length, iris_sub, prior = c(1,1)/2)
preds <- predict(z)$class

## Tables:
# Table 1
head(iris)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

```
#print(xtable(head(iris), caption = "First 6 lines of the dataset",
#              label = "data-head"),
#      comment = FALSE)

# Table 2
summary(iris)
```

```
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
##  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
##  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
##  Median :5.800   Median :3.000   Median :4.350   Median :1.300
##  Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
##  3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
##  Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
##        Species
##  setosa    :50
##  versicolor:50
##  virginica :50
##
##
##
```

```
#stargazer(iris, iqr = TRUE,
#          type = "latex", header = FALSE, label = "summary",
#          title = "Summary table of the four quantitative variables")

# Table 3
z$mean
```

```
##            Sepal.Width Sepal.Length Petal.Width Petal.Length
## setosa           3.428        5.006       0.246        1.462
## versicolor       2.770        5.936       1.326        4.260
```

```
#print(xtable(z£mean, label = "means-tbl", caption = "Group means"),
#      comment = FALSE)

# Table 4
z$scaling
```

```
##                    LD1
## Sepal.Width  -1.773845
## Sepal.Length -0.300458
## Petal.Width   3.035726
## Petal.Length  2.142260
```

```
#print(xtable(z£scaling, label = "scaling-tbl",
#             caption = "Coefficients of linear discriminants"),
#      comment = FALSE)
```

```r
# Table 5
(1/z$scaling[1]) * z$scaling
```

```
##                      LD1
## Sepal.Width   1.0000000
## Sepal.Length  0.1693823
## Petal.Width  -1.7113818
## Petal.Length -1.2076926
```

```r
#print(xtable((1/z£scaling[1]) * z£scaling, label = "scaling-unity-tbl",
#  caption = "Coefficients of linear discriminants with first scaled to unity"),
#      comment = FALSE)

# Table 6: Confusion Matrix
table(preds, iris_sub$Species)
```

```
##
## preds        setosa versicolor
##   setosa         50          0
##   versicolor      0         50
```

```r
#print(xtable(table(preds, iris_sub£Species), label = "confusion",
#              caption = "Confusion matrix"),
#      comment = FALSE)

## Figures
# Figure 1:
fig1 <- rasterGrob(readPNG("images/iris_setosa.png"), interpolate=TRUE)
fig2 <- rasterGrob(readPNG("images/iris_virginica.png"), interpolate=TRUE)
fig3 <- rasterGrob(readPNG("images/iris_versicolor.png"), interpolate=TRUE)
grid.arrange(fig1, fig2, fig3, ncol=3)

# Figure 2:
par(mar = c(5,4,4,3))
pairs.panels(iris[,1:4],
             labels = c("Sepal Length",
                        "Sepal Width",
                        "Petal Length",
                        "Petal Width"),
             bg=c("red",
```

```
                  "blue",
                  "green")[iris$Species],
             pch=20+as.numeric(iris$Species),
             ellipses = FALSE,
             oma=c(4,4,6,12))
par(xpd=TRUE)
legend(.9*par("usr")[2],  ## 1.05 times x1 limit
       .7*par("usr")[4],   ## .8 times y2 limit
       levels(iris$Species),
       pch = c(21,22,23),
       pt.bg = c("red", "blue", "green"))

# Figure 3:
plot1 <- ggplot(iris, aes(x=Sepal.Length))
plot1<- plot1 + geom_histogram(aes(fill = Species))
plot2 <- ggplot(iris, aes(x=Sepal.Width))
plot2<- plot2 + geom_histogram(aes(fill = Species))
plot3 <- ggplot(iris, aes(x=Petal.Length))
plot3<- plot3 + geom_histogram(aes(fill = Species))
plot4 <- ggplot(iris, aes(x=Petal.Width))
plot4<- plot4 + geom_histogram(aes(fill = Species))
theme_set(theme_gray(base_size = 18))
m<-grid.arrange(plot1,plot2, plot3,plot4, ncol=2)
# print(m)


# Figure 4:
par(mfrow = c(2,2))
beanplot(Sepal.Length ~ Species, data = iris, ylab = "Sepal Length",
         xlab = "Species", col = "beige", method = "jitter")
beanplot(Sepal.Width ~ Species, data = iris, ylab = "Sepal Width",
         xlab = "Species", col = "beige", method = "jitter")
beanplot(Petal.Length ~ Species, data = iris, ylab = "Petal Length",
         xlab = "Species", col = "beige", method = "jitter")
beanplot(Petal.Width ~ Species, data = iris, ylab = "Petal Width",
         xlab = "Species", col = "beige", method = "jitter")

# Figure 5:
plot(z)
```

References

Anderson, Edgar. 1936. "The Species Problem in Iris." *Annals of the Missouri Botanical Garden* 23 (3). Missouri Botanical Garden Press: 457–509. http://www.jstor.org/stable/2394164.

Fisher, R. A. 1936. "The Use of Multiple Measurements in Taxonomic Problems." *Annals of Eugenics* 7 (2). Blackwell Publishing Ltd: 179–88. doi:10.1111/j.1469-1809.1936.tb02137.x.

Kampstra, Peter. 2008. "Beanplot: A Boxplot Alternative for Visual Comparison of Distributions." *Journal of Statistical Software, Code Snippets* 28 (1): 1–9. http://www.jstatsoft.org/v28/c01/.

R Core Team. 2017. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with S*. Fourth. New York: Springer. http://www.stats.ox.ac.uk/pub/MASS4.