Lab 05 - LDA

[Sicily Xie 54385315]

2022-10-26

In the lectures last week, you learned about regression trees and bagging. Then we switched our focus to classification, for which the logistic regression and linear discriminant analysis (LDA) were introduced.

This week, we will perform LDA on the well-known *Iris* flower data set. The modified data set has two predictors, sepal and petal lengths, that are used to classify observations into one of the two flower species, *versicolor* and *virginica* (coded as 0 and 1, respectively).

```
library(MASS)
library(tidyverse)
theme_set(theme_bw())
data("iris")
iris <- iris %>%
    filter(Species != "setosa") %>%
    select(contains("Length"), Species) %>%
    mutate(Species = fct_drop(Species)) # drop the unused setosa level
```

Questions

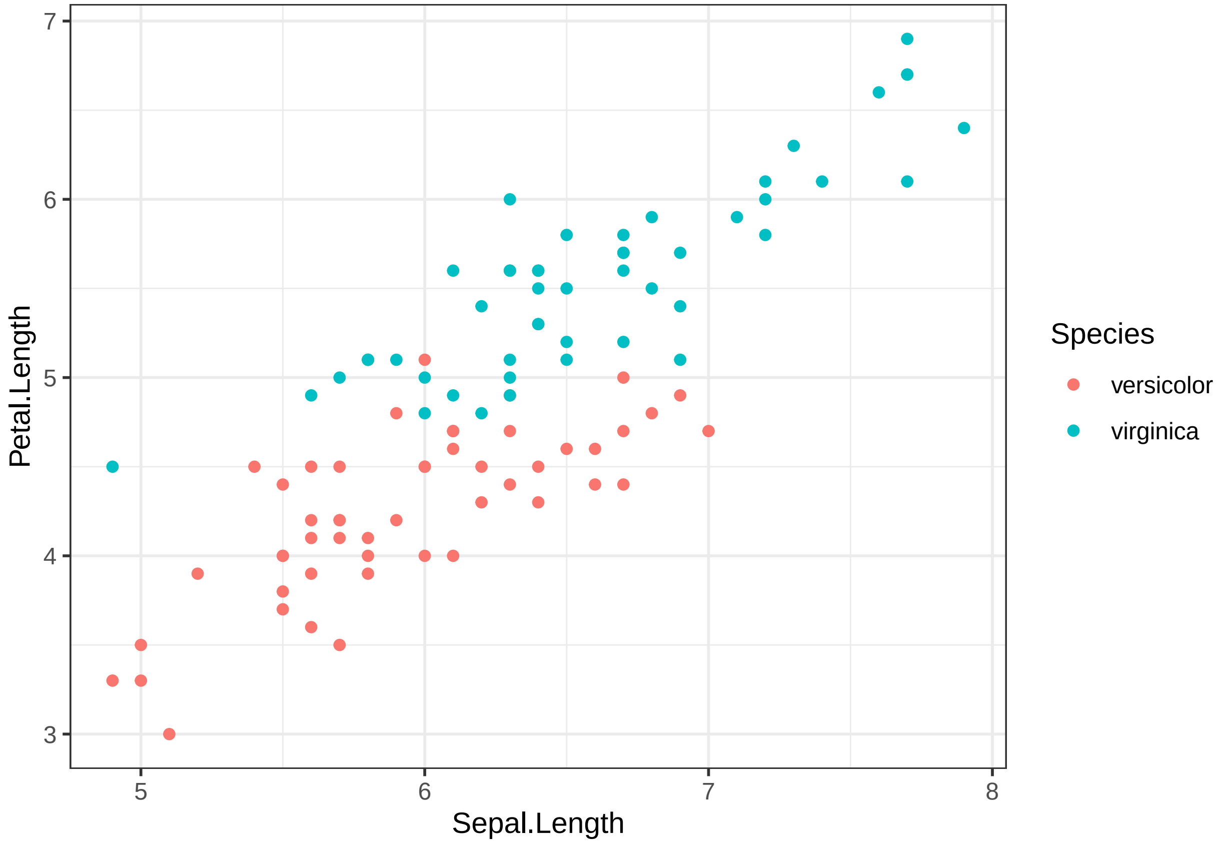
A classification problem has a discrete response. The objective is usually to obtain a rule for the classification of new observations into one of the categories. For predictors $\boldsymbol{X}$ and response (group) G which has possible values of $1, 2, \ldots, K$, we make use of the Bayes' rule to obtain the probability of group membership given $\boldsymbol{X}$:

$$\mathbb{P}(G = i|\boldsymbol{X}) = f_{\boldsymbol{X}|G}(\boldsymbol{x}|G = i)\mathbb{P}(G = i)/f_{\boldsymbol{X}}(\boldsymbol{x}) \propto f_{\boldsymbol{X}|G}(\boldsymbol{x}|G = i)\mathbb{P}(G = i),$$

where $\mathbb{P}(G = i)$ is the prior probability of group membership. Assume the predictors are normally distributed given $G = i$, with constant covariance matrix for each i, i.e., $\boldsymbol{X}|G = i \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$. Then the boundaries between any two classes j and k, i.e., the collection of $\boldsymbol{X}$ such that $\mathbb{P}(G = j|\boldsymbol{X}) = \mathbb{P}(G = k|\boldsymbol{X})$, is linear (hence the name LDA). A new observation is classified into group j if the estimated value of $\mathbb{P}(G = j|\boldsymbol{X})$ is the highest among the K classes.

To begin, load the data and plot the observations using the following command:

```
ggplot(iris, aes(Sepal.Length, Petal.Length)) +
  geom_point(aes(color = Species))
```



Q1 Do you think the assumption of equal covariance matrices in LDA is reasonable here? Explain.

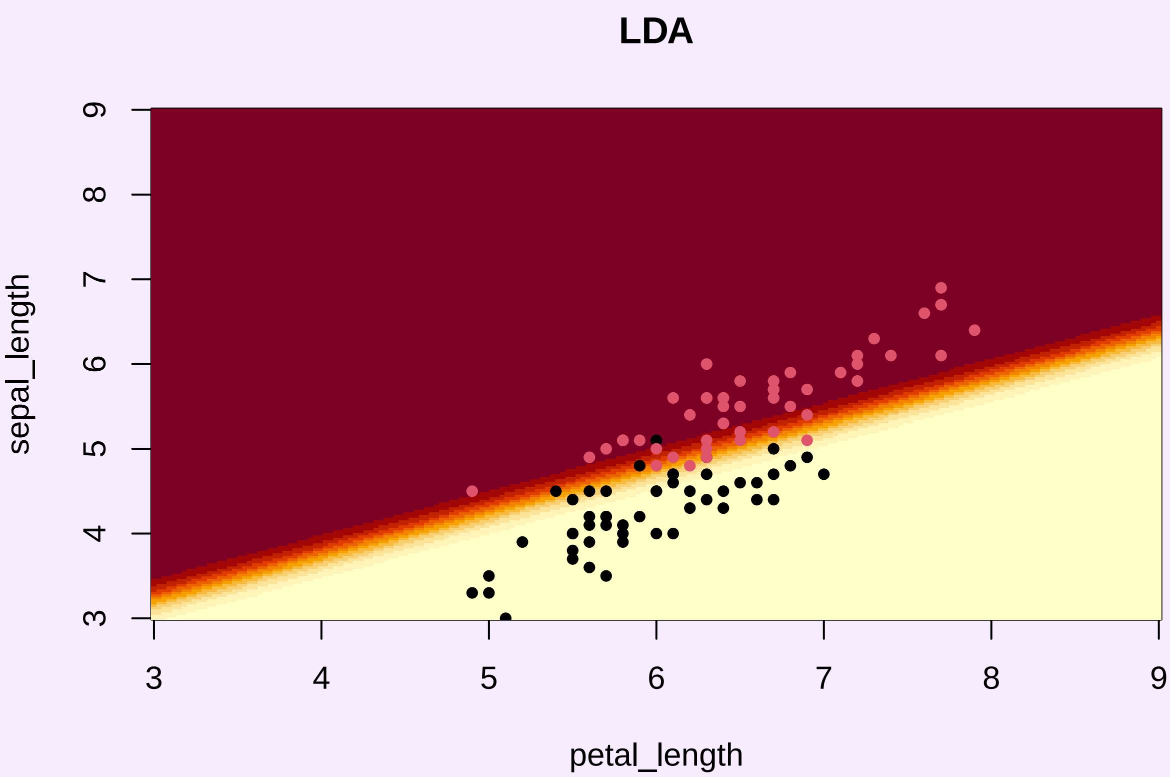
Hint: The covariance matrix is related to the shape or distribution of the points on a scatterplot.

Since the points show constant variance along the diagonal, the assumption of equal covariance matrices in LDA is reasonable.

Q2 Conduct LDA with the function `lda` in package `MASS`. Obtain the class boundary and draw it in the plot below.

Hint: To plot the boundary, first make predictions on a fine grid of sepal and petal lengths, and then use the `contour()` function in base `R` with appropriate arguments (or `geom_tile()` in `{ggplot}`)

```
model_lda <- lda(Species ~ Sepal.Length + Petal.Length, data = iris)
xSepal <- seq(3,9, length = 200); xPetal <- seq(3,9, length = 200)
dd <- expand.grid(xSepal, xPetal)
names(dd) <- c("Sepal.Length", "Petal.Length")
pr.lda <- predict(model_lda, newdata = dd)$posterior
image(xSepal, xPetal, matrix(pr.lda[,2],200,200),
ylab = "sepal_length", xlab = "petal_length", main = "LDA")
points(Petal.Length ~ Sepal.Length,
data = iris, pch = 20, cex = 1,
col = Species)
```



Q3 Suppose we have two new observations $\boldsymbol{x}_1 = (4.9, 5.3)$ and $\boldsymbol{x}_2 = (5.9, 4.8)$. What are the estimated posterior probabilities of these flowers being *versicolor* (reddish dots)? Which species do you predict them to be?

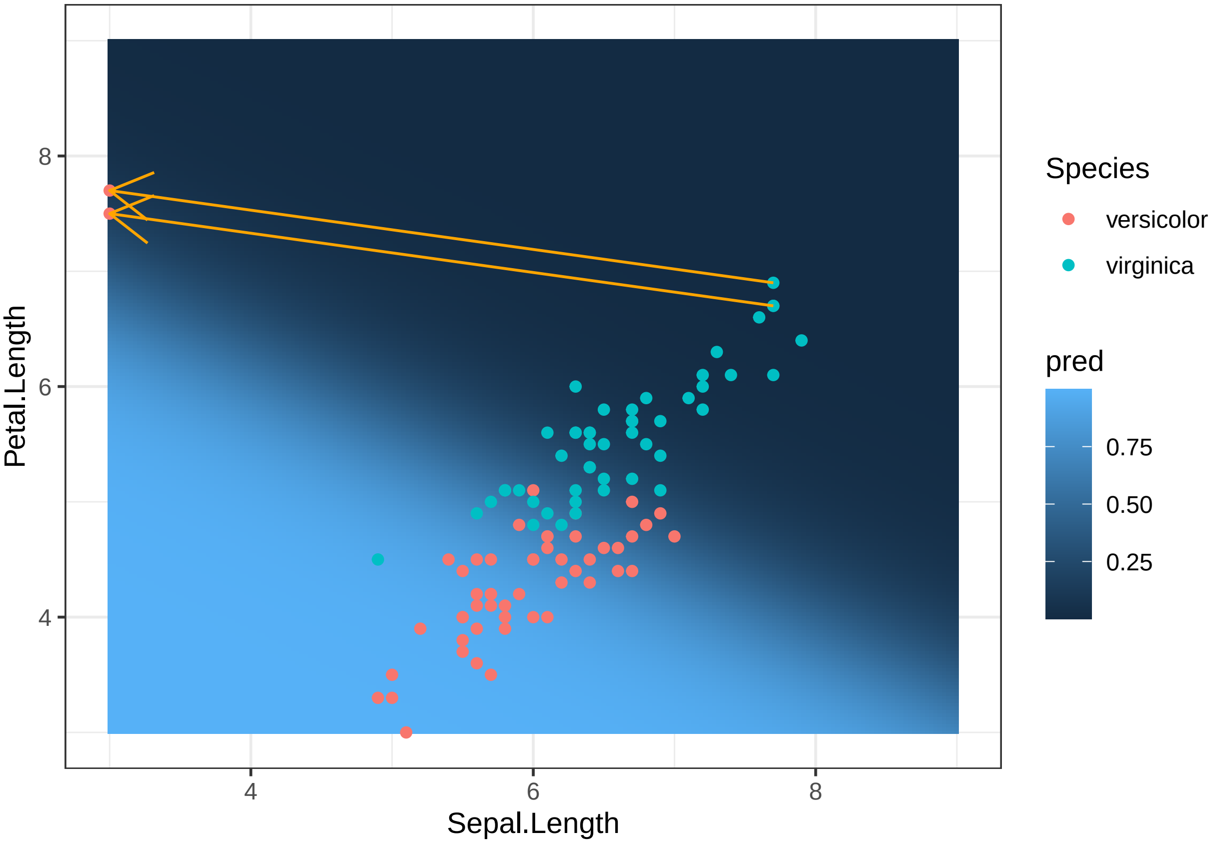
Note: Use the "highest posterior probability" to make your class prediction.

```
x1 <- c(4.9,5.3)
x2 <- c(5.9,4.8)
data <- data.frame(rbind(x1,x2))
names(data)<-c("Sepal.Length","Petal.Length")
new.predict <-predict(model_lda,newdata = data)
new.predict$posterior
```

```
##      versicolor virginica
## x1 2.936104e-05 0.9999706
## x2 3.148328e-01 0.6851672
```

The estimated posterior probabilities of these flowers being versicolor are 2.936104e-05 and 3.148328e-01, respectively. I predict them to be virginica based on the "highest posterior probability" criteria.

Q4 Suppose we change the sepal and petal lengths of two *virginica* observations, so that they are in the top-left corner of the plot. Running LDA on this modified data set yields the boundary shown below. Why is it so different from the one in **Q2**, even though we are only moving points far away from the boundary? (Answer in the space below.)



Since we move two data points far away from our original one, the equal covariance matrix assumption may no longer holds. Therefore, the boundary is different.