Extentions of Discriminant Analysis

Nate Wells

Math 243: Stat Learning

October 9th, 2020

Outline

In today's class, we will. . .

- Create a handmade LDA model
- Discuss LDA with two or more predictors
- Implement LDA in R
- Define QDA and compare to LDA

Section 1

Handmade LDA model

LDA

Suppose Y is a categorical variable with ℓ levels, and for each level A_j, that

$$X|Y = A_j \sim N(\mu_j, \sigma).$$

LDA

Suppose Y is a categorical variable with ℓ levels, and for each level A_j, that

$$X|Y = A_j \sim N(\mu_j, \sigma).$$

The discriminant function

$$\delta_j(x) = x \cdot \frac{\mu_j}{\sigma^2} - \frac{\mu_j^2}{2\sigma^2} + \ln \pi_j$$

can be used to classify an observation by choosing the level A_j whose discriminant is largest at x.

LDA

Suppose Y is a categorical variable with ℓ levels, and for each level A_j, that

$$X|Y = A_j \sim N(\mu_j, \sigma).$$

The discriminant function

$$\delta_j(x) = x \cdot \frac{\mu_j}{\sigma^2} - \frac{\mu_j^2}{2\sigma^2} + \ln \pi_j$$

can be used to classify an observation by choosing the level A_j whose discriminant is largest at x.

We estimate the values of μ_j and σ from the sample data:

$$\hat{\mu}_j = \frac{1}{n_j} \sum_{i:y_i = A_k} x_i$$

LDA

Suppose Y is a categorical variable with ℓ levels, and for each level A_j, that

$$X|Y = A_j \sim N(\mu_j, \sigma).$$

The discriminant function

$$\delta_j(x) = x \cdot \frac{\mu_j}{\sigma^2} - \frac{\mu_j^2}{2\sigma^2} + \ln \pi_j$$

can be used to classify an observation by choosing the level A_j whose discriminant is largest at x.

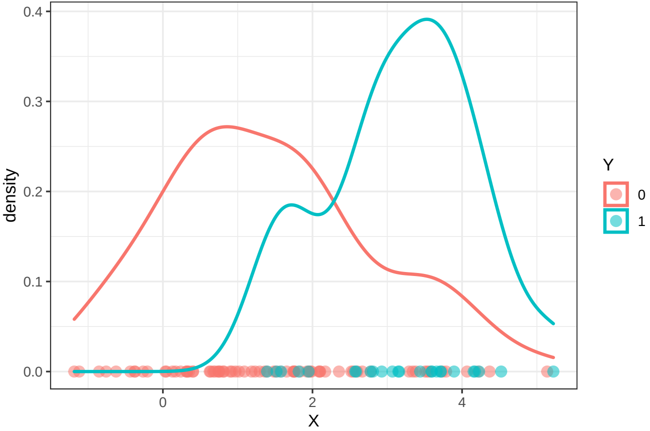
We estimate the values of μ_j and σ from the sample data:

$$\hat{\mu}_j = \frac{1}{n_j} \sum_{i:y_i = A_k} x_i$$

$$\hat{\sigma}^2 = \frac{1}{n - \ell} \sum_{j=1}^{\ell} \sum_{i:y_i = A_k} (x_i - \hat{\mu}_j)^2$$

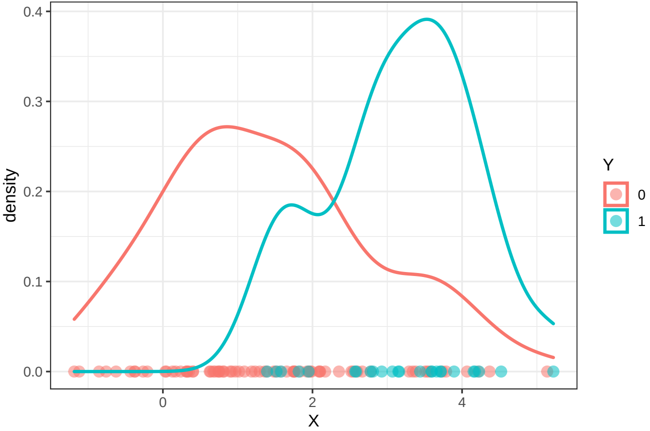
Simulated Data

Suppose $X|Y = 0 \sim N(1, 1)$ and $X|Y = 1 \sim N(3, 1)$, and that $\pi_0 = .75$ and $\pi_1 = .25$.



Simulated Data

Suppose $X|Y = 0 \sim N(1, 1)$ and $X|Y = 1 \sim N(3, 1)$, and that $\pi_0 = .75$ and $\pi_1 = .25$.



What feature of the graph shows that $\pi_0 = .75$ and $\pi_1 = .25$?

Find Estimates

Estimates for μ_j and π_j

```
pi0 <- 3/4
pi1 <- 1/4
mu0<-d %>% filter(Y == 0) %>% summarise(mu = mean(X) ) %>% pull()
mu1<-d %>% filter(Y == 1) %>% summarise(mu = mean(X) ) %>% pull()
data.frame(mu0, mu1)
```

```
##       mu0      mu1
## 1 1.42849 3.168335
```

Find Estimates

Estimates for μ_j and π_j

```
pi0 <- 3/4
pi1 <- 1/4
mu0<-d %>% filter(Y == 0) %>% summarise(mu = mean(X) ) %>% pull()
mu1<-d %>% filter(Y == 1) %>% summarise(mu = mean(X) ) %>% pull()
data.frame(mu0, mu1)
```

```
##         mu0      mu1
## 1 1.42849 3.168335
```

Estimates for σ.

```
ssx <- d %>% group_by(Y) %>% summarize(ssx = var(X) * (n() - 1), n()) %>% pull(2,)
ssx
```

```
## [1] 148.19201  23.70648
```

```
sigma2 <- sum(ssx)/(n - 2)
sigma2
```

```
## [1] 1.754066
```

The discriminant function

Solve for intersection of discriminant functions:

The discriminant function

Solve for intersection of discriminant functions:

$$c\frac{\mu_1}{\sigma^2} - \frac{\mu_1^2}{2\sigma^2} + \ln \pi_1 = c\frac{\mu_0}{\sigma^2} - \frac{\mu_0^2}{2\sigma^2} + \ln \pi_0$$

The discriminant function

Solve for intersection of discriminant functions:

$$c\frac{\mu_1}{\sigma^2} - \frac{\mu_1^2}{2\sigma^2} + \ln \pi_1 = c\frac{\mu_0}{\sigma^2} - \frac{\mu_0^2}{2\sigma^2} + \ln \pi_0$$

$$c = \frac{2\sigma^2 \ln \frac{\pi_0}{\pi_1} + \mu_1^2 - \mu_0^2}{2(\mu_1 - \mu_0)}$$

The discriminant function

Solve for intersection of discriminant functions:

$$c\frac{\mu_1}{\sigma^2} - \frac{\mu_1^2}{2\sigma^2} + \ln \pi_1 = c\frac{\mu_0}{\sigma^2} - \frac{\mu_0^2}{2\sigma^2} + \ln \pi_0$$

$$c = \frac{2\sigma^2 \ln \frac{\pi_0}{\pi_1} + \mu_1^2 - \mu_0^2}{2(\mu_1 - \mu_0)}$$

```
c<- (2*sigma2*log(.75/.25) + mu1^2 - mu0^2)/(2*(mu1 - mu0))
c
```

```
## [1] 3.406004
```

The discriminant function

Solve for intersection of discriminant functions:

$$c\frac{\mu_1}{\sigma^2} - \frac{\mu_1^2}{2\sigma^2} + \ln \pi_1 = c\frac{\mu_0}{\sigma^2} - \frac{\mu_0^2}{2\sigma^2} + \ln \pi_0$$

$$c = \frac{2\sigma^2 \ln \frac{\pi_0}{\pi_1} + \mu_1^2 - \mu_0^2}{2(\mu_1 - \mu_0)}$$

```
c<- (2*sigma2*log(.75/.25) + mu1^2 - mu0^2)/(2*(mu1 - mu0))
c
```

```
## [1] 3.406004
```

Write a function to create discriminant functions:

```
my_lda <- function(x, pi, mu, sigma2) {
  x * (mu/sigma2) - (mu^2)/(2 * sigma2) + log(pi)
}
```

The discriminant function

Solve for intersection of discriminant functions:

$$c\frac{\mu_1}{\sigma^2} - \frac{\mu_1^2}{2\sigma^2} + \ln \pi_1 = c\frac{\mu_0}{\sigma^2} - \frac{\mu_0^2}{2\sigma^2} + \ln \pi_0$$

$$c = \frac{2\sigma^2 \ln \frac{\pi_0}{\pi_1} + \mu_1^2 - \mu_0^2}{2(\mu_1 - \mu_0)}$$

```
c<- (2*sigma2*log(.75/.25) + mu1^2 - mu0^2)/(2*(mu1 - mu0))
c
```

```
## [1] 3.406004
```

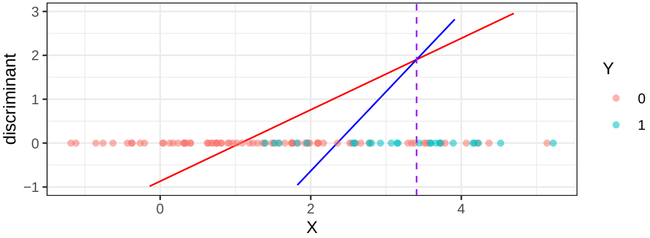
Write a function to create discriminant functions:

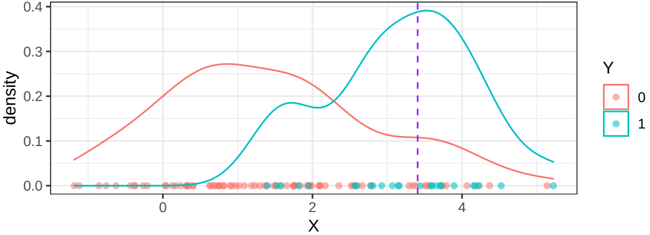
```
my_lda <- function(x, pi, mu, sigma2) {
  x * (mu/sigma2) - (mu^2)/(2 * sigma2) + log(pi)
}
```

Create discriminant function for each class:

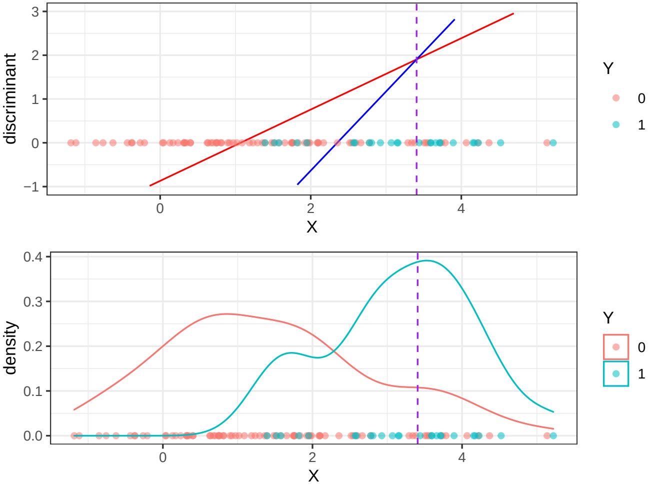
```
d0 <- my_lda(d$X, pi0, mu0, sigma2)
d1 <- my_lda(d$X, pi1, mu1, sigma2)
```

Plots





Plots



Why don't the discriminant functions intersect at the same point as the density curves?

Section 2

LDA with multiple predictors

Multivariate Gaussian Distributions

A vector $X = (X_1, X_2, \ldots, X_p)$ is said to have multivariate gaussian distribution if all linear combinations of coordinates $a1X_1 + a_2X_2 + \cdots + a_pX_p$ have a Normal distribution.

Multivariate Gaussian Distributions

A vector $X = (X_1, X_2, \ldots, X_p)$ is said to have multivariate gaussian distribution if all linear combinations of coordinates $a1X_1 + a_2X_2 + \cdots + a_pX_p$ have a Normal distribution.

A multivariate gaussian distribution is specified by mean vector $\mu = (\mu_1, \mu_2, \ldots, \mu_p)$ and covariance matrix

$$\Sigma = \begin{pmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) & \cdots & \mathrm{Cov}(X_1, X_p) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}(X_2) & \cdots & \mathrm{Cov}(X_2, X_p) \\ \vdots & & \ddots & \vdots \\ \mathrm{Cov}(X_p, X_1) & \mathrm{Cov}(X_p, X_2) & & \mathrm{Var}(X_p) \end{pmatrix}$$

Multivariate Gaussian Distributions

A vector $X = (X_1, X_2, \ldots, X_p)$ is said to have multivariate gaussian distribution if all linear combinations of coordinates $a1X_1 + a_2X_2 + \cdots + a_pX_p$ have a Normal distribution.

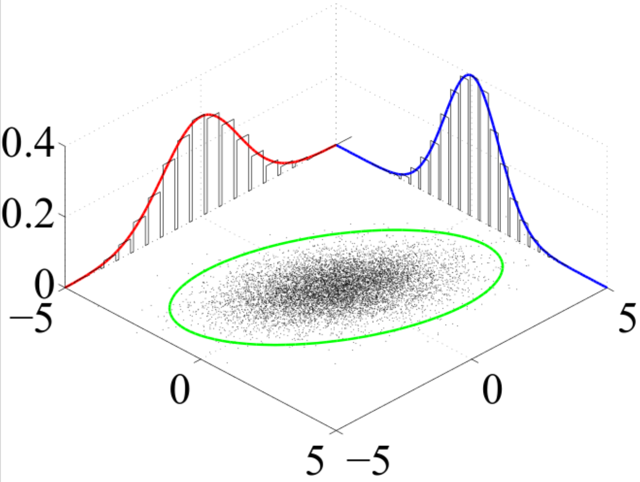
A multivariate gaussian distribution is specified by mean vector $\mu = (\mu_1, \mu_2, \ldots, \mu_p)$ and covariance matrix

$$\Sigma = \begin{pmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) & \cdots & \mathrm{Cov}(X_1, X_p) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}(X_2) & \cdots & \mathrm{Cov}(X_2, X_p) \\ \vdots & & \ddots & \vdots \\ \mathrm{Cov}(X_p, X_1) & \mathrm{Cov}(X_p, X_2) & & \mathrm{Var}(X_p) \end{pmatrix}$$

The multivariate Gaussian density f on $x \in \mathbb{R}^p$ is

$$f(x) = \frac{1}{(2\pi)^{p/2}(|\det\Sigma|)^{1/2}} \exp\Big(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\Big)$$

Multivariate Scatterplot



LDA with multiple predictors

Suppose that Y is categorical with ℓ levels and that $X = (X_1, \ldots, X_p)$ are a vector of predictors. Assume that $X|Y = A_j \sim N(\mu_j, \Sigma)$ with conditional density f_j, where Σ is common to all conditional densities.

LDA with multiple predictors

Suppose that Y is categorical with ℓ levels and that $X = (X_1, \ldots, X_p)$ are a vector of predictors. Assume that $X|Y = A_j \sim N(\mu_j, \Sigma)$ with conditional density f_j, where Σ is common to all conditional densities.

As before, we consider the log-likelihood ratio:

$$\ln \frac{P(X = x \mid Y = A_j)}{P(X = x \mid Y = A_k)} = \ln \frac{f_j(x)\pi_j}{f_k(x)\pi_k}$$

LDA with multiple predictors

Suppose that Y is categorical with ℓ levels and that $X = (X_1, \ldots, X_p)$ are a vector of predictors. Assume that $X|Y = A_j \sim N(\mu_j, \Sigma)$ with conditional density f_j, where Σ is common to all conditional densities.

As before, we consider the log-likelihood ratio:

$$\ln \frac{P(X = x \,|\, Y = A_j)}{P(X = x \,|\, Y = A_k)} = \ln \frac{f_j(x)\pi_j}{f_k(x)\pi_k}$$

The discriminant function $\delta_j(x)$ for $x \in \mathbb{R}^p$ is

$$\delta_j(x) = x^T \Sigma^{-1} \mu_j - \frac{1}{2} \mu_j^T \Sigma^{-1} \mu_j + \ln \pi_j$$

LDA with multiple predictors

Suppose that Y is categorical with ℓ levels and that $X = (X_1, \ldots, X_p)$ are a vector of predictors. Assume that $X|Y = A_j \sim N(\mu_j, \Sigma)$ with conditional density f_j, where Σ is common to all conditional densities.

As before, we consider the log-likelihood ratio:

$$\ln \frac{P(X = x \mid Y = A_j)}{P(X = x \mid Y = A_k)} = \ln \frac{f_j(x)\pi_j}{f_k(x)\pi_k}$$

The discriminant function $\delta_j(x)$ for $x \in \mathbb{R}^p$ is

$$\delta_j(x) = x^T \Sigma^{-1} \mu_j - \frac{1}{2}\mu_j^T \Sigma^{-1} \mu_j + \ln \pi_j$$

We classify a point x by assigning it to the level with largest discriminant function at x.

LDA with multiple predictors

Suppose that Y is categorical with ℓ levels and that $X = (X_1, \dots, X_p)$ are a vector of predictors. Assume that $X|Y = A_j \sim N(\mu_j, \Sigma)$ with conditional density f_j, where Σ is common to all conditional densities.

As before, we consider the log-likelihood ratio:

$$\ln \frac{P(X = x \mid Y = A_j)}{P(X = x \mid Y = A_k)} = \ln \frac{f_j(x)\pi_j}{f_k(x)\pi_k}$$

The discriminant function $\delta_j(x)$ for $x \in \mathbb{R}^p$ is

$$\delta_j(x) = x^T \Sigma^{-1} \mu_j - \frac{1}{2} \mu_j^T \Sigma^{-1} \mu_j + \ln \pi_j$$

We classify a point x by assigning it to the level with largest discriminant function at x.

Decision boundaries are given by solving for intersections of the $\binom{p}{2}$ pairs of discriminant functions:

$$x^T \Sigma^{-1} \mu_j - \frac{1}{2} \mu_j^T \Sigma^{-1} \mu_j + \ln \pi_j = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \ln \pi_k$$

Classification

Let's investigate the classic `iris` dataset:

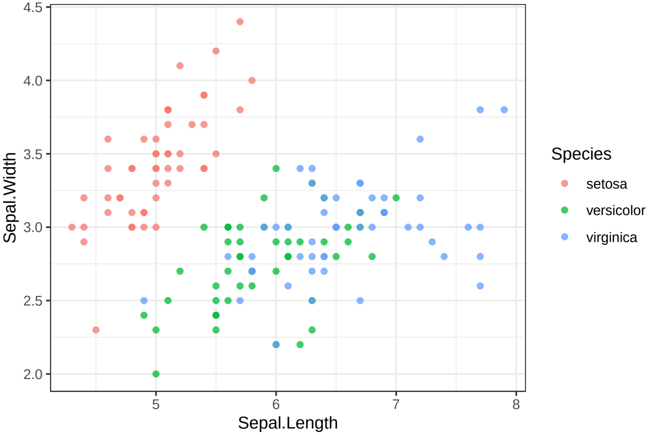
Classification

Let's investigate the classic `iris` dataset:

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
## 1          4.8         3.4          1.6         0.2     setosa
## 2          6.1         2.9          4.7         1.4 versicolor
## 3          5.7         2.8          4.1         1.3 versicolor
## 4          6.8         3.2          5.9         2.3  virginica
## 5          6.7         2.5          5.8         1.8  virginica
```

Can we classify `Species` based on `Sepal.Length` and `Sepal.Width`?

Iris Plot



Where should we place our **linear** decision boundaries?

LDA in R

It would be tedious to compute LDA discriminant functions by hand. So we use the `lda` function in the `mass` package.

```
library(MASS)
mlda <- lda(Species ~ Sepal.Length + Sepal.Width,data = iris)
mlda_pred <- predict(mlda)
conf_mlda <- table(mlda_pred$class,iris$Species)
conf_mlda
```

```
##
##              setosa versicolor virginica
##   setosa         49          0         0
##   versicolor      1         36        15
##   virginica       0         14        35
```

LDA in R

It would be tedious to compute LDA discriminant functions by hand. So we use the `lda` function in the `mass` package.

```
library(MASS)
mlda <- lda(Species ~ Sepal.Length + Sepal.Width,data = iris)
mlda_pred <- predict(mlda)
conf_mlda <- table(mlda_pred$class,iris$Species)
conf_mlda
```

```
##
##              setosa versicolor virginica
##   setosa         49          0         0
##   versicolor      1         36        15
##   virginica       0         14        35
```

It looks like LDA had a hard time distinguishing between vesicolor and virginica.

LDA in R

It would be tedious to compute LDA discriminant functions by hand. So we use the `lda` function in the `mass` package.

```
library(MASS)
mlda <- lda(Species ~ Sepal.Length + Sepal.Width,data = iris)
mlda_pred <- predict(mlda)
conf_mlda <- table(mlda_pred$class,iris$Species)
conf_mlda
```

```
##
##              setosa versicolor virginica
##   setosa         49          0         0
##   versicolor      1         36        15
##   virginica       0         14        35
```

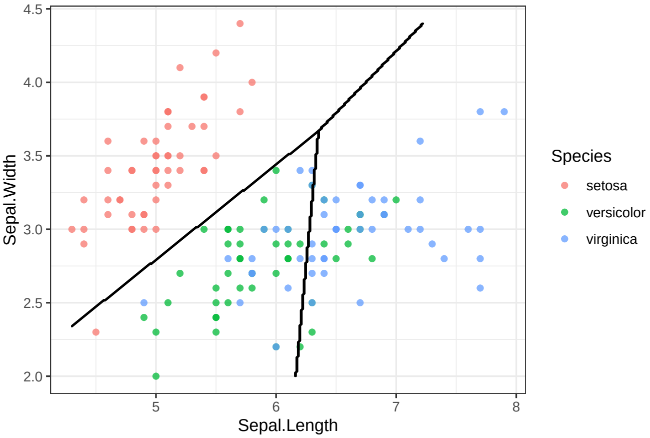
It looks like LDA had a hard time distinguishing between vesicolor and virginica.

Overall error rate

```
(sum(conf_mlda) - sum(diag(conf_mlda)))/sum(conf_mlda)
```

```
## [1] 0.2
```

Iris Decision Boundaries



Section 3

QDA

Generalized Model

For a data set with 15 predictors and 1000 observations, would you be more worried about bias (Y) or variance (N) for an LDA model?

Generalized Model

For a data set with 15 predictors and 1000 observations, would you be more worried about bias (Y) or variance (N) for an LDA model?

- With lots of data, variance is likely low. But the modeling restrictions of LDA might make bias problematic.

Generalized Model

For a data set with 15 predictors and 1000 observations, would you be more worried about bias (Y) or variance (N) for an LDA model?

- With lots of data, variance is likely low. But the modeling restrictions of LDA might make bias problematic.
- We might be able to improve MSE by considering a more **complex** model.

Generalized Model

For a data set with 15 predictors and 1000 observations, would you be more worried about bias (Y) or variance (N) for an LDA model?

- With lots of data, variance is likely low. But the modeling restrictions of LDA might make bias problematic.
- We might be able to improve MSE by considering a more **complex** model.

One underlying assumption for LDA was that all conditional distribution of predictors $P(X = x \mid Y = y_j)$ had the same variance (or covariance matrix, for $p \geq 2$).

Generalized Model

For a data set with 15 predictors and 1000 observations, would you be more worried about bias (Y) or variance (N) for an LDA model?

- With lots of data, variance is likely low. But the modeling restrictions of LDA might make bias problematic.
- We might be able to improve MSE by considering a more **complex** model.

One underlying assumption for LDA was that all conditional distribution of predictors $P(X = x \mid Y = y_j)$ had the same variance (or covariance matrix, for $p \geq 2$).

Lifting this restriction leads to **Quadratic Discriminant Analysis** (QDA)

QDA

Suppose that Y is categorical with ℓ levels and that $X = (X_1, \ldots, X_p)$ are a vector of predictors. Assume that $X|Y = A_j \sim N(\mu_j, \Sigma_j)$ with conditional density f_j.

QDA

Suppose that Y is categorical with ℓ levels and that $X = (X_1, \ldots, X_p)$ are a vector of predictors. Assume that $X|Y = A_j \sim N(\mu_j, \Sigma_j)$ with conditional density f_j.

As with LDA, we consider the log likelihood ratios

$$\ln \frac{P(X = x \mid Y = A_j)}{P(X = x \mid Y = A_k)} = \ln \frac{f_j(x)\pi_j}{f_k(x)\pi_k}$$

QDA

Suppose that Y is categorical with ℓ levels and that $X = (X_1, \ldots, X_p)$ are a vector of predictors. Assume that $X|Y = A_j \sim N(\mu_j, \Sigma_j)$ with conditional density f_j.

As with LDA, we consider the log likelihood ratios

$$\ln \frac{P(X = x \mid Y = A_j)}{P(X = x \mid Y = A_k)} = \ln \frac{f_j(x)\pi_j}{f_k(x)\pi_k}$$

But now when we substitute the formula for multivariate densities f_i, the variance (or covariance) terms in numerator and denominator do **not** cancel.

QDA

Suppose that Y is categorical with ℓ levels and that $X = (X_1, \dots, X_p)$ are a vector of predictors. Assume that $X|Y = A_j \sim N(\mu_j, \Sigma_j)$ with conditional density f_j.

As with LDA, we consider the log likelihood ratios

$$\ln \frac{P(X = x \mid Y = A_j)}{P(X = x \mid Y = A_k)} = \ln \frac{f_j(x)\pi_j}{f_k(x)\pi_k}$$

But now when we substitute the formula for multivariate densities f_i, the variance (or covariance) terms in numerator and denominator do **not** cancel.

This leads to the QDA discriminant function $\delta_j(x)$:

$$\delta_j(x) = -\frac{1}{2}x^T\Sigma_j^{-1}x + x^T\Sigma_j^{-1}\mu_j - \frac{1}{2}\mu_j^T\Sigma_j^{-1}\mu_j - \frac{1}{2}\ln\det\Sigma_j + \ln\pi_j$$

QDA

Suppose that Y is categorical with ℓ levels and that $X = (X_1, \ldots, X_p)$ are a vector of predictors. Assume that $X|Y = A_j \sim N(\mu_j, \Sigma_j)$ with conditional density f_j.

As with LDA, we consider the log likelihood ratios

$$\ln \frac{P(X = x \mid Y = A_j)}{P(X = x \mid Y = A_k)} = \ln \frac{f_j(x)\pi_j}{f_k(x)\pi_k}$$

But now when we substitute the formula for multivariate densities f_i, the variance (or covariance) terms in numerator and denominator do **not** cancel.

This leads to the QDA discriminant function $\delta_j(x)$:

$$\delta_j(x) = -\frac{1}{2}x^T\Sigma_j^{-1}x + x^T\Sigma_j^{-1}\mu_j - \frac{1}{2}\mu_j^T\Sigma_j^{-1}\mu_j - \frac{1}{2}\ln\det\Sigma_j + \ln\pi_j$$

Which simplifes to the following when $p = 1$:

$$\delta_j(x) = -x^2\frac{1}{2\sigma_j} + x\frac{\mu_j}{\sigma_j} - \frac{\mu_j^2}{2\sigma_j} - \frac{1}{2}\ln\sigma_j + \ln\pi_j$$

In R

We use the qda function in the mass package.

```
library(MASS)
mqda <- qda(Species ~ Sepal.Length + Sepal.Width,data = iris)
mqda_pred <- predict(mlda)
conf_mqda <- table(mlda_pred$class,iris$Species)
conf_mqda
```

```
##
##              setosa versicolor virginica
##   setosa         49          0         0
##   versicolor      1         36        15
##   virginica       0         14        35
```

In R

We use the qda function in the mass package.

```
library(MASS)
mqda <- qda(Species ~ Sepal.Length + Sepal.Width,data = iris)
mqda_pred <- predict(mlda)
conf_mqda <- table(mlda_pred$class,iris$Species)
conf_mqda
```

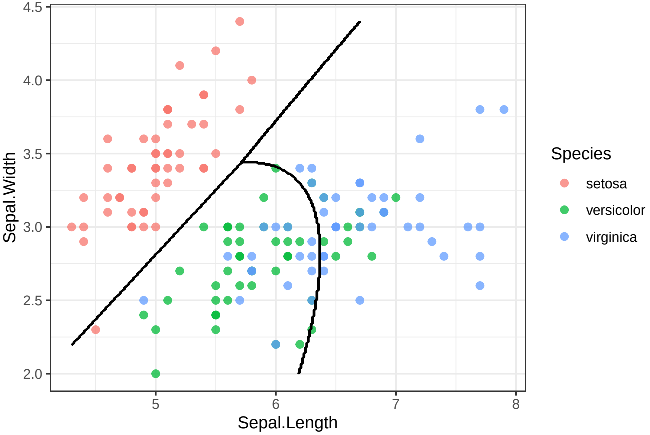
```
##
##              setosa versicolor virginica
##   setosa         49          0         0
##   versicolor      1         36        15
##   virginica       0         14        35
```

How did we do?

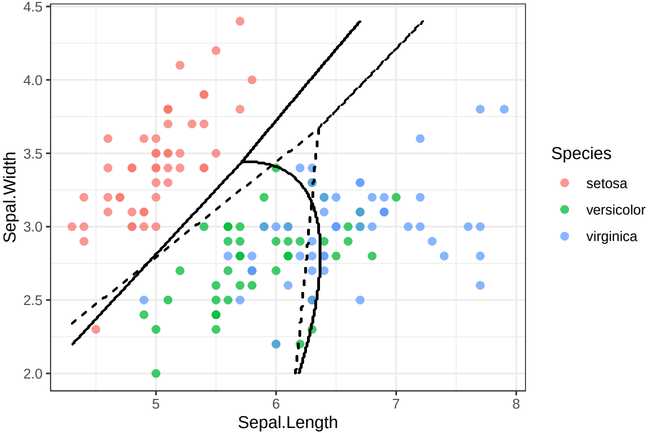
```
(sum(conf_mqda) - sum(diag(conf_mqda)))/sum(conf_mqda)
```

```
## [1] 0.2
```

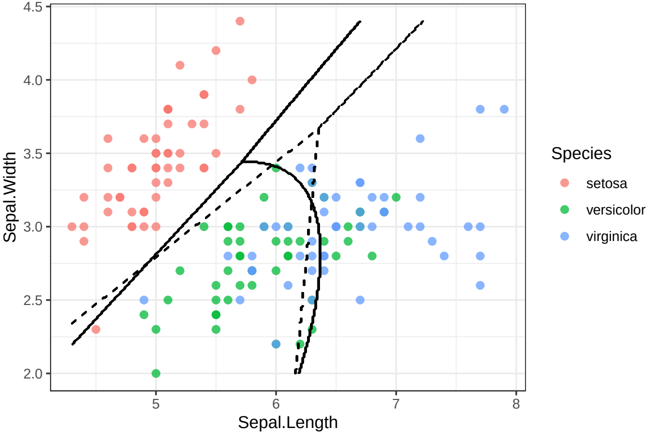
QDA Decision Boundaries



LDA - QDA Comparison



LDA - QDA Comparison



Which model do you think would perform better on test data? LDA(Y) or QDA (N)