LDA Extensions

Nate Wells

Math 243: Stat Learning

November 5th, 2021

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.

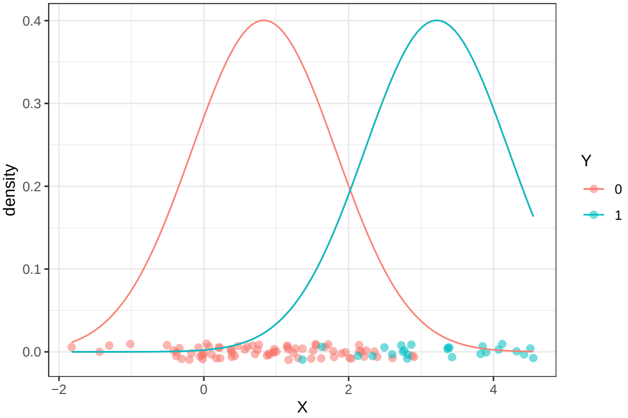
We estimate the values of μ_j and σ from the sample data:

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

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

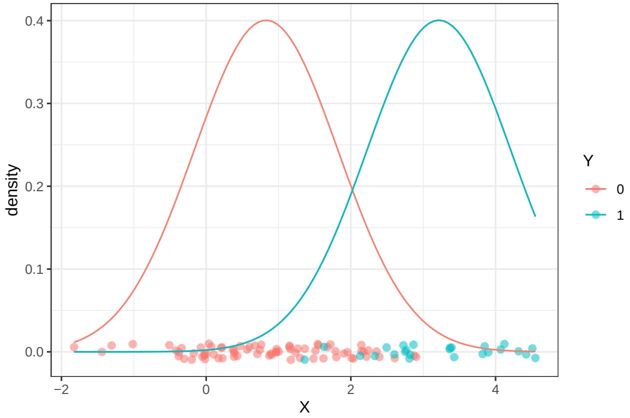
Simulated Data

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

```
d %>% group_by(Y) %>% summarize(pi = n()/n, mu = mean(X))
```

```
## # A tibble: 2 x 3
##   Y        pi    mu
##   <chr> <dbl> <dbl>
## 1 0      0.75 0.828
## 2 1      0.25 3.22
```

Find Estimates

Estimates for μ_j and π_j

```
d %>% group_by(Y) %>% summarize(pi = n()/n, mu = mean(X))
```

```
## # A tibble: 2 x 3
##   Y        pi    mu
##   <chr> <dbl> <dbl>
## 1 0      0.75 0.828
## 2 1      0.25 3.22
```

Estimate for σ^2.

```
d %>% group_by(Y) %>% summarize(ssx = var(X) * (n() - 1)) %>%
  summarize(sigma_sq = sum(ssx)/(n-2))
```

```
## # A tibble: 1 x 1
##   sigma_sq
##      <dbl>
## 1    0.992
```

The discriminant function

Solve for intersection of discriminant functions: $\delta_0(c) = \delta_1(c)$ when

The discriminant function

Solve for intersection of discriminant functions: $\delta_0(c) = \delta_1(c)$ when

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

The discriminant function

Solve for intersection of discriminant functions: $\delta_0(c) = \delta_1(c)$ when

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

```
c<- (mu0 + mu1)/2 + (sigma2*log(pi0) - log(pi1))/(mu1-mu0)
c
```

```
## [1] 2.483001
```

The discriminant function

Solve for intersection of discriminant functions: $\delta_0(c) = \delta_1(c)$ when

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

```
c<- (mu0 + mu1)/2 + (sigma2*log(pi0) - log(pi1))/(mu1-mu0)
c
```

```
## [1] 2.483001
```

Write a function to create discriminant functions:

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

The discriminant function

Solve for intersection of discriminant functions: $\delta_0(c) = \delta_1(c)$ when

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

```
c<- (mu0 + mu1)/2 + (sigma2*log(pi0) - log(pi1))/(mu1-mu0)
c
```

```
## [1] 2.483001
```

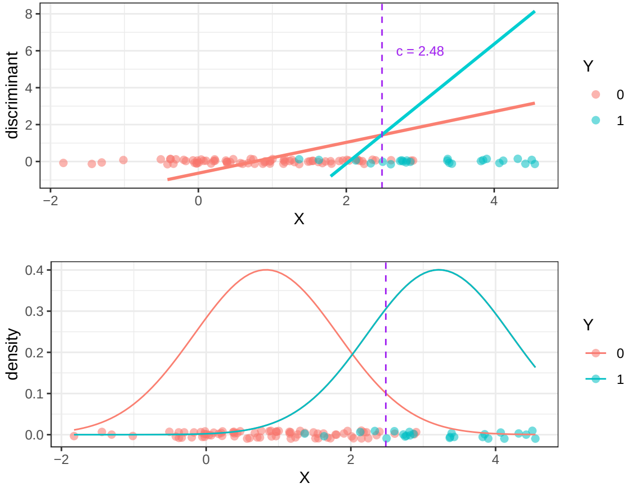
Write a function to create discriminant functions:

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

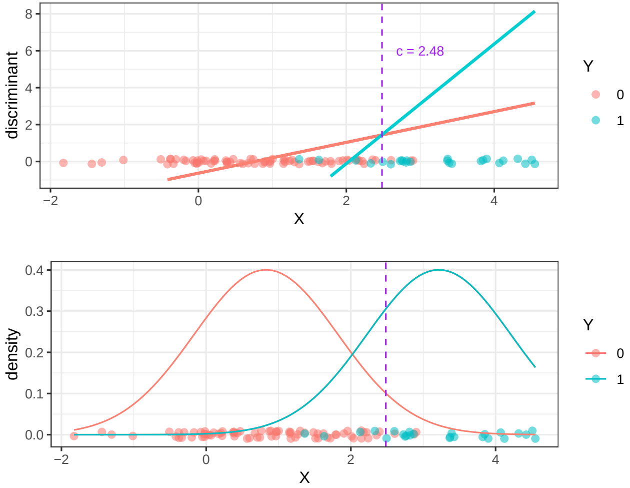
Evaluate discriminant function on data for each class:

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

Plots



Plots



- Why don't discriminant functions intersect at the same point as 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 $X = (X_1, X_2, \dots, X_p)$ 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 $\mu = (\mu_1, \mu_2, \dots, \mu_p)$ 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.

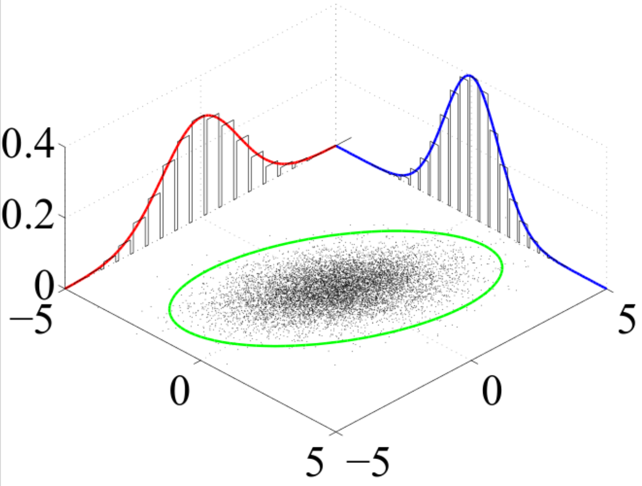
A multivariate gaussian distribution is specified by mean vector $\mu = (\mu_1, \mu_2, \ldots, \mu_p)$ and covariance matrix

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

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:

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

LDA with multiple predictors

Suppose that Y is categorical with ℓ levels and that $X = (X_1, \dots, X_p)$ 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:

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

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:

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

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 $X = (X_1, \ldots, X_p)$ 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:

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

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:

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

Section 3

LDA in R

Classification

- The `penguins` data set from the `palmerpenguins` package collected by Dr. Kristen Gorman on several attributes of antarctic penguins:

Classification

- The `penguins` data set from the `palmerpenguins` package collected by Dr. Kristen Gorman on several attributes of antarctic penguins:

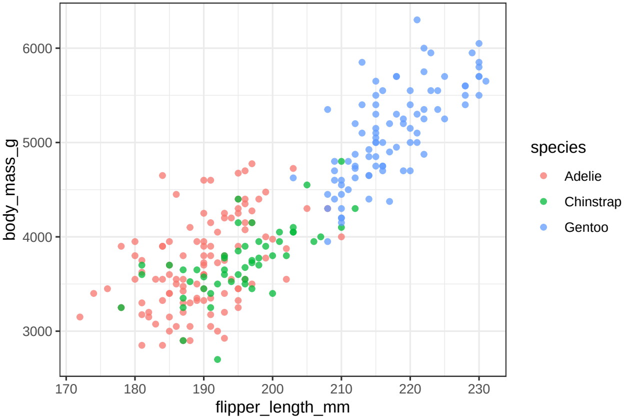
```
library(palmerpenguins)
penguins <- penguins %>% drop_na()
glimpse(penguins)
```

```
## Rows: 333
## Columns: 8
## $ species           <fct> Adelie, Adelie, Adelie, Adelie, Adelie, Adelie, Adel~
## $ island            <fct> Torgersen, Torgersen, Torgersen, Torgersen, Torgerse~
## $ bill_length_mm    <dbl> 39.1, 39.5, 40.3, 36.7, 39.3, 38.9, 39.2, 41.1, 38.6~
## $ bill_depth_mm     <dbl> 18.7, 17.4, 18.0, 19.3, 20.6, 17.8, 19.6, 17.6, 21.2~
## $ flipper_length_mm <int> 181, 186, 195, 193, 190, 181, 195, 182, 191, 198, 18~
## $ body_mass_g       <int> 3750, 3800, 3250, 3450, 3650, 3625, 4675, 3200, 3800~
## $ sex               <fct> male, female, female, female, male, female, male, fe~
## $ year              <int> 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007~
```

```
library(rsample)
set.seed(115)
penguins_split <- initial_split(penguins , strata = species)
penguins_train <- training(penguins_split)
penguins_test <- testing(penguins_split)
```

Penguins Plot

- Can we classify `species` based on `body_mass_g` and `flipper_length_mm`?



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)
penguin_lda <- lda(species ~ flipper_length_mm + body_mass_g,data = penguins_train)
```

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)
penguin_lda <- lda(species ~ flipper_length_mm + body_mass_g,data = penguins_train)
```

- The `lda` function creates an LDA model which can be used to `predict`. It also has the following useful elements.
  - `prior`, the prior probabilities used (defaults to class proportions in training data)
  - `means`, means for predictors within each group

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)
penguin_lda <- lda(species ~ flipper_length_mm + body_mass_g,data = penguins_train)
```

- The `lda` function creates an LDA model which can be used to `predict`. It also has the following useful elements.
  - `prior`, the prior probabilities used (defaults to class proportions in training data)
  - `means`, means for predictors within each group

```
penguin_lda$prior
```

```
##    Adelie Chinstrap    Gentoo
## 0.4377510 0.2048193 0.3574297
```

```
penguin_lda$means
```

```
##           flipper_length_mm body_mass_g
## Adelie             189.8991    3710.550
## Chinstrap          195.4902    3739.706
## Gentoo             217.3820    5101.966
```

Predictions

- The `mass` package has a `predict` function for `lda`, which creates a list with two objects:
  - `class`, the predicted class for each observation
  - `posterior`, the posterior probabilities for each class

Predictions

- The mass package has a predict function for lda, which creates a list with two objects:
  - class, the predicted class for each observation
  - posterior, the posterior probabilities for each class

```
## [1] Adelie Adelie Adelie Adelie Adelie Adelie
## Levels: Adelie Chinstrap Gentoo
```

```
##      Adelie  Chinstrap       Gentoo
## 1 0.8829216 0.11704567 3.278079e-05
## 2 0.5821311 0.41748669 3.822059e-04
## 3 0.7546054 0.24523589 1.586998e-04
## 4 0.7256116 0.24942525 2.496315e-02
## 5 0.9420053 0.05799351 1.179860e-06
## 6 0.8176350 0.18222767 1.373104e-04
```

Error Rate

How well does LDA do?

Error Rate

How well does LDA do?

```
##            Truth
## Prediction  Adelie Chinstrap Gentoo
##   Adelie        31         9      0
##   Chinstrap      6         6      0
##   Gentoo         0         2     30
```

Error Rate

How well does LDA do?

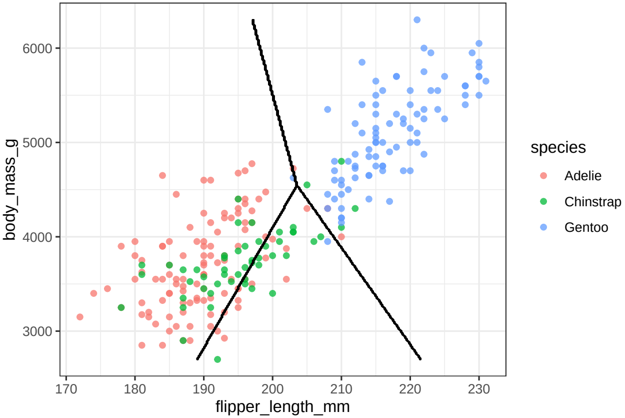
```
##             Truth
## Prediction  Adelie Chinstrap Gentoo
##   Adelie        31         9      0
##   Chinstrap      6         6      0
##   Gentoo         0         2     30
```

- It looks like the model had some trouble distinguishing between Adelie and Chinstrap penguins.

```
accuracy(lda_results, truth = obs, estimate = preds)
```

```
## # A tibble: 1 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy multiclass     0.798
```

Penguin Decision Boundaries



Section 4

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

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

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

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

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 $X = (X_1, \ldots, X_p)$ 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

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

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

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

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)
penguin_qda <- qda(species ~ flipper_length_mm + body_mass_g,data = penguins_train)
penguin_results <- data.frame(obs = penguins_test$species,
                              preds = predict(penguin_qda, penguins_test)$class)
conf_mat(penguin_results, truth = obs, estimate = preds)
```

```
##             Truth
## Prediction  Adelie Chinstrap Gentoo
##   Adelie        32        12      0
##   Chinstrap      5         3      0
##   Gentoo         0         2     30
```

In R

We use the qda function in the mass package.

```
library(MASS)
penguin_qda <- qda(species ~ flipper_length_mm + body_mass_g,data = penguins_train)
penguin_results <- data.frame(obs = penguins_test$species,
                              preds = predict(penguin_qda, penguins_test)$class)
conf_mat(penguin_results, truth = obs, estimate = preds)
```

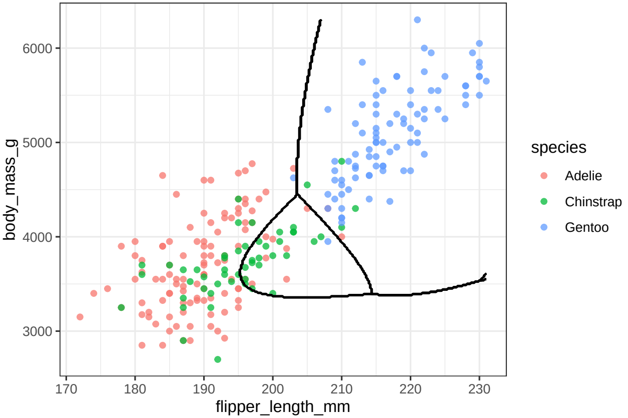
```
##            Truth
## Prediction  Adelie Chinstrap Gentoo
##   Adelie        32        12      0
##   Chinstrap      5         3      0
##   Gentoo         0         2     30
```

How did we do?

```
accuracy(penguin_results, truth = obs, estimate = preds)
```

```
## # A tibble: 1 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy multiclass     0.774
```

QDA Decision Boundaries



LDA - QDA Comparison