Linear Discriminant Analysis

Nate Wells

Math 243: Stat Learning

October 7th, 2020

Outline

In today's class, we will. . .

- Discuss LDA theory and motivation
- Implement LDA in R

Section 1

LDA

Logistic Regression, KNN, and Bayes' Classifier

Recall that for a classification problem, the average test error rate is minimized using the Bayes' classifier:

$$f(x_0) = \text{argmax}_j P(Y = A_j \mid X = x_0)$$

Logistic Regression, KNN, and Bayes' Classifier

Recall that for a classification problem, the average test error rate is minimized using the Bayes' classifier:

$$f(x_0) = \mathrm{argmax}_j P(Y = A_j \mid X = x_0)$$

Both KNN and Logistic regression attempt to estimate the conditional probability $p(X)$:

Logistic Regression, KNN, and Bayes' Classifier

Recall that for a classification problem, the average test error rate is minimized using the Bayes' classifier:

$$f(x_0) = \mathrm{argmax}_j P(Y = A_j \mid X = x_0)$$

Both KNN and Logistic regression attempt to estimate the conditional probability $p(X)$:

- Logistic regression:

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$$

Logistic Regression, KNN, and Bayes' Classifier

Recall that for a classification problem, the average test error rate is minimized using the Bayes' classifier:

$$f(x_0) = \mathrm{argmax}_j P(Y = A_j \mid X = x_0)$$

Both KNN and Logistic regression attempt to estimate the conditional probability $p(X)$:

- Logistic regression:

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$$

- KNN:

$$p(X) = \frac{1}{K}\sum_{i\in N_0} I(y_i = A_j)$$

Bayes' Rule

For any events A and B,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Example

Suppose a test for a certain disease has specificity .9 and sensitivity .8, and that the disease has prior prevalence of 0.01. Find the probability that an individual who tests positive for the disease actually has the disease.

The Bayesian Flip

We want $P(Y = A_j \mid X = x_0)$. Using Bayes' Rule:

$$P(Y = A_j \mid X = x_0) = \frac{P(X = x_0 \mid Y = A_j)P(Y = A_j)}{P(X = X_0)}$$
$$= \frac{P(X = x_0 \mid Y = A_j)P(Y = A_j)}{\sum_i P(X = X_0 | Y = A_i)P(Y = A_i)}$$

The Bayesian Flip

We want $P(Y = A_j \mid X = x_0)$. Using Bayes' Rule:

$$P(Y = A_j \mid X = x_0) = \frac{P(X = x_0 \mid Y = A_j)P(Y = A_j)}{P(X = X_0)}$$
$$= \frac{P(X = x_0 \mid Y = A_j)P(Y = A_j)}{\sum_i P(X = X_0 | Y = A_i)P(Y = A_i)}$$

We estimate the conditional probabilitity of the response using...

- The conditional distribution $P(X = x_0 \mid Y = A_j)$ of each predictor
- The prior distribution $\pi_i = P(Y = A_i)$ of the response

The Bayesian Flip

We want $P(Y = A_j \mid X = x_0)$. Using Bayes' Rule:

$$P(Y = A_j \mid X = x_0) = \frac{P(X = x_0 \mid Y = A_j)P(Y = A_j)}{P(X = X_0)}$$
$$= \frac{P(X = x_0 \mid Y = A_j)P(Y = A_j)}{\sum_i P(X = X_0 | Y = A_i)P(Y = A_i)}$$

We estimate the conditional probabilitity of the response using. . .

- The conditional distribution $P(X = x_0 \mid Y = A_j)$ of each predictor
- The prior distribution $\pi_i = P(Y = A_i)$ of the response

In practice, we don't have access to the conditional distributions of the predictors, so need to estimate them based on data.

LDA

Suppose we have just one predictor X and a multi-level categorical response Y.

LDA

Suppose we have just one predictor X and a multi-level categorical response Y.

What is the most "natural" assumption for the conditional distribution of X?

LDA

Suppose we have just one predictor X and a multi-level categorical response Y.

What is the most "natural" assumption for the conditional distribution of X?

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

LDA

Suppose we have just one predictor X and a multi-level categorical response Y.

What is the most "natural" assumption for the conditional distribution of X?

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

If X is normal, its conditional density is given by

$$P(X = x \mid Y = A_j) = f_j(x) = \frac{1}{\sqrt{2\pi\sigma_j^2}} e^{-(x-\mu_j)^2/2\sigma_j^2}$$

If we assume all conditional distributions have the **same** variance $\sigma_j^2 = \sigma^2$, we can simplify our model.

Log-Likelihood Ratio

To determine to which class an observation belongs, based on the conditional distribution of predictors, we consider likelihood ratio:

$$\begin{aligned}\frac{P(Y=A_j \mid X=x_0)}{P(Y=A_k \mid X=x_0)} &= \frac{P(X=x_0 \mid Y=A_j)P(Y=A_j)/P(X=x_0)}{P(X=x_0 \mid Y=A_k)P(Y=A_k)/P(X=x_0)} \\ &= \frac{P(X=x_0 \mid Y=A_j)P(Y=A_j)}{P(X=x_0 \mid Y=A_k)P(Y=A_k)} \\ &= \frac{e^{-(x_0-\mu_j)^2/2\sigma^2}\pi_j}{e^{-(x_0-\mu_k)^2/2\sigma^2}\pi_k}\end{aligned}$$

Log-Likelihood Ratio

To determine to which class an observation belongs, based on the conditional distribution of predictors, we consider likelihood ratio:

$$\begin{aligned}\frac{P(Y=A_j \mid X=x_0)}{P(Y=A_k \mid X=x_0)} &= \frac{P(X=x_0 \mid Y=A_j)P(Y=A_j)/P(X=x_0)}{P(X=x_0 \mid Y=A_k)P(Y=A_k)/P(X=x_0)} \\ &= \frac{P(X=x_0 \mid Y=A_j)P(Y=A_j)}{P(X=x_0 \mid Y=A_k)P(Y=A_k)} \\ &= \frac{e^{-(x_0-\mu_j)^2/2\sigma^2}\pi_j}{e^{-(x_0-\mu_k)^2/2\sigma^2}\pi_k}\end{aligned}$$

The log-liklihood ratio is obtained by taking natural log above:

$$\ln \frac{P(Y=A_j \mid X=x_0)}{P(Y=A_k \mid X=x_0)} = (x_0-\mu_k)^2/2\sigma^2 - (x_0-\mu_j)^2/2\sigma^2 + \ln \pi_j - \ln \pi_k$$

Log-Likelihood Ratio

To determine to which class an observation belongs, based on the conditional distribution of predictors, we consider likelihood ratio:

$$\begin{aligned}\frac{P(Y=A_j \mid X=x_0)}{P(Y=A_k \mid X=x_0)} &= \frac{P(X=x_0 \mid Y=A_j)P(Y=A_j)/P(X=x_0)}{P(X=x_0 \mid Y=A_k)P(Y=A_k)/P(X=x_0)} \\ &= \frac{P(X=x_0 \mid Y=A_j)P(Y=A_j)}{P(X=x_0 \mid Y=A_k)P(Y=A_k)} \\ &= \frac{e^{-(x_0-\mu_j)^2/2\sigma^2}\pi_j}{e^{-(x_0-\mu_k)^2/2\sigma^2}\pi_k}\end{aligned}$$

The log-liklihood ratio is obtained by taking natural log above:

$$\ln \frac{P(Y=A_j \mid X=x_0)}{P(Y=A_k \mid X=x_0)} = (x_0-\mu_k)^2/2\sigma^2 - (x_0-\mu_j)^2/2\sigma^2 + \ln \pi_j - \ln \pi_k$$

The decision boundary between A_j and A_k is the point c where

$$(c-\mu_k)^2/2\sigma^2 + \ln \pi_j = (c-\mu_j)^2/2\sigma^2 + \ln \pi_k$$

Binary Classfication

Suppose Y is binary, and that each of $X|Y = 0$ and $X|Y = 1$ are Normal with common variance σ and means μ_1 and μ_2. Moreover, assume a uniform prior $\pi_1 = \pi_0 = \frac{1}{2}$

Binary Classfication

Suppose Y is binary, and that each of $X|Y = 0$ and $X|Y = 1$ are Normal with common variance σ and means μ_1 and μ_2. Moreover, assume a uniform prior $\pi_1 = \pi_0 = \frac{1}{2}$

Solve for c in

$$(c - \mu_k)^2/2\sigma^2 + \ln \pi_j = (c - \mu_j)^2/2\sigma^2 + \ln \pi_k$$

Binary Classfication

Suppose Y is binary, and that each of $X|Y=0$ and $X|Y=1$ are Normal with common variance σ and means μ_1 and μ_2. Moreover, assume a uniform prior $\pi_1 = \pi_0 = \frac{1}{2}$

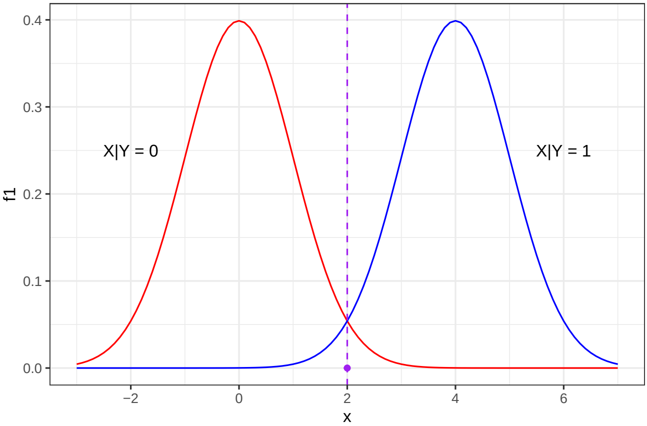
Solve for c in

$$(c-\mu_k)^2/2\sigma^2 + \ln \pi_j = (c-\mu_j)^2/2\sigma^2 + \ln \pi_k$$

We get $c = \frac{\mu_1+\mu_2}{2}$

Plots

Suppose $X|Y = 0 \sim N(0,1)$ and $X|Y = 1 \sim N(4,1)$



What is LDA?

If we **knew** the conditional distribution of the predictors, we could easily create decision boundaries.

What is LDA?

If we **knew** the conditional distribution of the predictors, we could easily create decision boundaries.

- But we only have data, so we need to estimate those distributions.

What is LDA?

If we **knew** the conditional distribution of the predictors, we could easily create decision boundaries.

- But we only have data, so we need to estimate those distributions.

A normal distribution requires only 2 parameters: μ and σ.

What is LDA?

If we **knew** the conditional distribution of the predictors, we could easily create decision boundaries.

- But we only have data, so we need to estimate those distributions.

A normal distribution requires only 2 parameters: μ and σ.

- We need one estimate of μ for each level of Y.
- Since we assumed each conditional distribution had the same variance, we need only 1 estimate for σ

What is LDA?

If we **knew** the conditional distribution of the predictors, we could easily create decision boundaries.

- But we only have data, so we need to estimate those distributions.

A normal distribution requires only 2 parameters: μ and σ.

- We need one estimate of μ for each level of Y.
- Since we assumed each conditional distribution had the same variance, we need only 1 estimate for σ

LDA is an algorithm for obtaining these estimates and then classifying based on log-likelihood ratio:

What is LDA?

If we **knew** the conditional distribution of the predictors, we could easily create decision boundaries.

- But we only have data, so we need to estimate those distributions.

A normal distribution requires only 2 parameters: μ and σ.

- We need one estimate of μ for each level of Y.
- Since we assumed each conditional distribution had the same variance, we need only 1 estimate for σ

LDA is an algorithm for obtaining these estimates and then classifying based on log-likelihood ratio:

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

What is LDA?

If we **knew** the conditional distribution of the predictors, we could easily create decision boundaries.

- But we only have data, so we need to estimate those distributions.

A normal distribution requires only 2 parameters: μ and σ.

- We need one estimate of μ for each level of Y.
- Since we assumed each conditional distribution had the same variance, we need only 1 estimate for σ

LDA is an algorithm for obtaining these estimates and then classifying based on log-likelihood ratio:

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

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

The Discriminant

Rather than comparing log likelihoods, we could instead look at the log conditional probability for each level. This function $\delta_j(x)$ is called the *discriminant* for level j:

The Discriminant

Rather than comparing log likelihoods, we could instead look at the log conditional probability for each level. This function $\delta_j(x)$ is called the *discriminant* for level j:

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

The Discriminant

Rather than comparing log likelihoods, we could instead look at the log conditional probability for each level. This function $\delta_j(x)$ is called the *discriminant* for level j:

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

We can then assign an observation to the class whose discriminant is largest.

The Discriminant

Rather than comparing log likelihoods, we could instead look at the log conditional probability for each level. This function $\delta_j(x)$ is called the *discriminant* for level j:

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

We can then assign an observation to the class whose discriminant is largest.

Why is LDA called **Linear** Discriminant Analysis?

The Discriminant

Rather than comparing log likelihoods, we could instead look at the log conditional probability for each level. This function $\delta_j(x)$ is called the *discriminant* for level j:

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

We can then assign an observation to the class whose discriminant is largest.

Why is LDA called **Linear** Discriminant Analysis?

- Because the discriminant function is linear in x.

The Discriminant

Rather than comparing log likelihoods, we could instead look at the log conditional probability for each level. This function $\delta_j(x)$ is called the *discriminant* for level j:

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

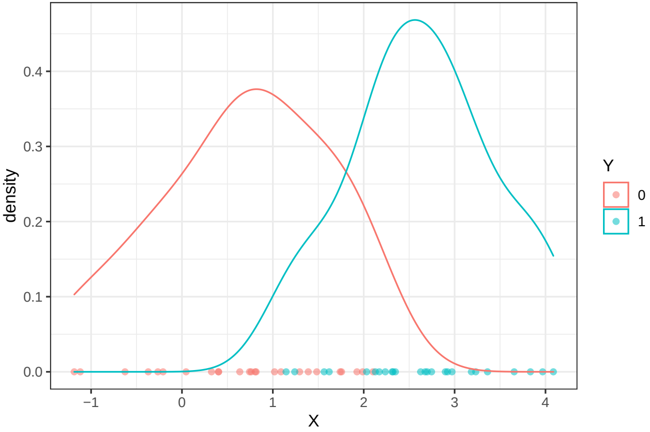
We can then assign an observation to the class whose discriminant is largest.

Why is LDA called **Linear** Discriminant Analysis?

- Because the discriminant function is linear in x.
- Using this classification algorithm will result in linear decision boundaries.

Simulated Data

Suppose $X|Y = 0 \sim N(1,1)$ and $X|Y = 1 \sim N(3,1)$, and that each class is of the same size.



Find Estimates

Estimates for μ_j.

```
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 0.6587046 3.068198
```

Find Estimates

Estimates for μ_j.

```
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 0.6587046 3.068198
```

Estimates for σ.

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

```
## [1] 74.31554 94.41776
```

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

```
## [1] 1.721768
```

The discriminant function

Write a function to create discriminant functions:

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

The discriminant function

Write a function to create discriminant functions:

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

Create discriminant function for each class:

Plot