Linear Discriminant Analysis

Section 4.4

Cathy Poliak, Ph.D.
cpoliak@central.uh.edu

Department of Mathematics
University of Houston

Classification

- The response variable, Y, is **qualitative** or **categorical**.
- Predicting a qualitative response for an observations can be referred to as **classifying** that observation.
- These methods predict the probability of each of the categories of a qualitative variables, as the basis for making the classification.

Logistic Regression

- Logistic regression can be used to model and solve problems when the Y (response) variable is a categorical variable with 2 classes.
- Also called binary classification problems.
- This models the **probability** that Y belongs to one of the two categories.

Multiple Logistic Regression

We now look at predicting a binary response using multiple predictors.

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Where $X = (X_1, \ldots, X_p)$ are p predictors. This can be rewritten as

$$P(Y=1 \mid X) = p(X) = \frac{\exp\left(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p\right)}{1 + \exp\left(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p\right)}$$

We will use the maximum likelihood method to estimate $\beta_0, \beta_1, \ldots, \beta_p$.

Another Model for Classification

We model the distributions of the predictors X separately in each of the response classes (i.e. given Y) and then use Bayes' theorem to flip these around into estimates for $P(Y = k|X = x)$

- When these distributions are assumed to be normal, it turns out that the model is very similar in form to the logistic regression.
- Why use another model?
  - When the classes are well-separated, the parameter estimates for the logistic regression model are surprisingly unstable.
  - If n is small and the distribution of the predictors X is approximately normal in each of the classes, the linear discriminant model is again more stable than the logistic regression model.
  - The linear discriminant model can be used when we have more than two response classes.

$$P(Y=k|X=x) = \frac{P(Y=k \wedge X=x)}{P(X=x)} = \frac{P(Y=k) * P(X=x|Y=k)}{\sum_{k=1}^{K} P(Y=k) * P(X=x|Y=k)}$$

Using Bayes' Theorem for Classification

- Let K be number of classes of a response variable, $K \geq 2$.
- Let π_k represent the overall or *prior* probability that a randomly chosen observation comes from the kth class of Y.
- Let $f_k(x) = P(X = x|Y = k)$ denote the density function of X for an observation that comes from the kth class of Y.
- Then the **Bayes' Theorem** states that

$$P(Y = k|X = x) = p_k(x) = \frac{\pi_k f_k(x)}{\sum_{i=1}^{K} \pi_i f_i(x)}$$

- To estimate π_k we can compute the fraction of the training observations that belong to the kth class.
- The problem is how we estimate $f_k(x)$?

Linear Discriminant Analysis for $p = 1$

- Assume we only have one predictor, $p = 1$.
- Also assume that $X|Y = k \sim N(\mu_k, \sigma_k^2)$. That is X has a normal distribution given the kth class with mean μ_k and variance σ_k^2 for that kth class.

Linear Discriminant Analysis for $p = 1$

- Assume we only have one predictor, $p = 1$.
- Also assume that $X|Y = k \sim N(\mu_k, \sigma_k^2)$. That is X has a normal distribution given the kth class with mean μ_k and variance σ_k^2 for that kth class.
- Then the density function $f_k(x)$ is

$$P(X \mid Y=k) = f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k}\exp\left(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2\right)$$

Linear Discriminant Analysis for $p = 1$

- Assume we only have one predictor, $p = 1$.
- Also assume that $X|Y = k \sim N(\mu_k, \sigma_k^2)$. That is X has a normal distribution given the kth class with mean μ_k and variance σ_k^2 for that kth class.
- Then the density function $f_k(x)$ is

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{1}{2\sigma_k^2}(x - \mu_k)^2\right)$$

- Further assume that $\sigma_1^2 = \cdots = \sigma_k^2 = \sigma^2$.

Linear Discriminant Analysis for $p = 1$

- Assume we only have one predictor, $p = 1$.
- Also assume that $X|Y = k \sim N(\mu_k, \sigma_k^2)$. That is X has a normal distribution given the kth class with mean μ_k and variance σ_k^2 for that kth class.
- Then the density function $f_k(x)$ is

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{1}{2\sigma_k^2}(x - \mu_k)^2\right)$$

- Further assume that $\sigma_1^2 = \cdots = \sigma_k^2 = \sigma^2$.
- Thus with these assumptions we get:

proportion of times in k category

$$p_k(x) = \frac{\pi_k f_k(x)}{\sum_{i=1}^K \pi_i f_i(x)} = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x - \mu_k)^2\right)}{\sum_{i=1}^K \pi_i \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x - \mu_i)^2\right)}$$

The Bayes' Classifier

- The Bayes' classifier involves assigning an observation $X = x$ to the class for which $p_k(x)$ is the largest.

The Bayes' Classifier

- The Bayes' classifier involves assigning an observation $X = x$ to the class for which $p_k(x)$ is the largest.
- Taking the log of $p_k(x)$ and discarding terms that do not depend on k, we get the **discriminat score**

$$\delta_k(x) = x \times \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + log(\pi_k)$$

The Bayes' Classifier

- The Bayes' classifier involves assigning an observation $X = x$ to the class for which $p_k(x)$ is the largest.
- Taking the log of $p_k(x)$ and discarding terms that do not depend on k, we get the **discriminat score**

$$\delta_k(x) = x \times \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + log(\pi_k)$$

- Thus we get a linear function of x.
- Classify $X = x$, where $p_k(x)$ is the largest is equivalent to were the discriminant score is the largest.

Example $K = 2$: Bayes Decision Boundary

- If $K = 2$ and we assign x to class 1, then $\delta_1(x) > \delta_2(x)$. Assume $\pi_1 = \pi_2$.

Example $K = 2$: Bayes Decision Boundary

- If $K = 2$ and we assign x to class 1, then $\delta_1(x) > \delta_2(x)$. Assume $\pi_1 = \pi_2$.

$$x\frac{\mu_1}{\sigma^2} - \frac{\mu_1^2}{2\sigma^2} + log(\pi_1) > x\frac{\mu_2}{\sigma^2} - \frac{\mu_2^2}{2\sigma^2} + log(\pi_2)$$

$$2x\mu_1 - \mu_1^2 > 2x\mu_2 - \mu_2^2$$

$$2x(\mu_1 - \mu_2) > \mu_1^2 - \mu_2^2$$

Example $K = 2$: Bayes Decision Boundary

- If $K = 2$ and we assign x to class 1, then $\delta_1(x) > \delta_2(x)$. Assume $\pi_1 = \pi_2$.

$$x\frac{\mu_1}{\sigma^2} - \frac{\mu_1^2}{2\sigma^2} + log(\pi_1) > x\frac{\mu_2}{\sigma^2} - \frac{\mu_2^2}{2\sigma^2} + log(\pi_2)$$
$$2x\mu_1 - \mu_1^2 > 2x\mu_2 - \mu_2^2$$
$$2x(\mu_1 - \mu_2) > \mu_1^2 - \mu_2^2$$

- The Bayes decision boundary is the point where $\delta_1(x) = \delta_2(x)$. Which means

$$\delta_1(x) = \delta_2(x)$$
$$2x(\mu_1 - \mu_2) = \mu_1^2 - \mu_2^2$$
$$x = \frac{\mu_1^2 - \mu_2^2}{2(\mu_1 - \mu_2)}$$
$$x = \frac{\mu_1 + \mu_2}{2}$$

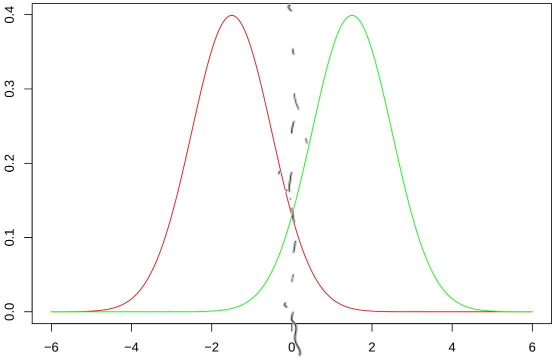
Example

Suppose we have two normal density functions with $\mu_1 = -1.5$, $\mu_2 = 1.5$, $\sigma_1^2 = \sigma_2^2 = 1$

Boundary $X = \frac{-1.5 + 1.5}{2} = 0$

If $X < 0$ class 1

If $x > 0$ class



Linear Discriminant Analysis

- In practice we will not know the true value of $\mu_1, \dots, \mu_K$, $\pi_1, \dots, \pi_k$ and σ^2.
- The **linear discriminant analysis** (LDA) is the method that approximates the Bayes classifier by plugging estimates for π_k, μ_k, and σ^2.
- The following are used:

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i$$

Poole variance

$$\frac{(n-1)S_1^2 + (n-1)S_2^2}{n-K}$$

$$\hat{\sigma}^2 = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2$$

$$\hat{\pi}_k = \frac{n_k}{n}$$

The LDA Classifier

The LDA classifier plugs the estimates given for $\hat{\mu}_k$, $\hat{\sigma}^2$, and $\hat{\pi}_k$ in the Bayes classifier and assigns and observation $X = x$ to the class for which

$$\hat{\delta}_k(x) = x\frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\hat{\mu}_k^2}{2\hat{\sigma}^2} + \log(\hat{\pi}_k)$$

is the largest.

Example: Lab Questions

Consider data from two populations assuming a Normal distribution and $\sigma_1^2 = \sigma_2^2$:

Population 1			Population 2		
3	-3	0	6	-5	1
2	-3	1	5	-5	0
4	-3	1	4	-5	1
1	-3	4	5	-5	0
5	-3	4	5	-5	0
		10			2

1. Determine $\hat{\mu}_1$.

   a) 3
   b) 5
   c) 4
   d) 2

2. Determine $\hat{\mu}_2$.

   a) 3
   b) 5
   c) 4
   d) 2

3. Determine $\sum_{i:y_i=1} (x_i - \hat{\mu}_1)^2$

a) 10
b) 5
c) 2
d) 3

4. Determine $\sum_{i:y_i=2} (x_i - \hat{\mu}_2)^2$

a) 10
b) 5
c) 2
d) 3

5. Determine $\hat{\sigma}^2$. $= \frac{10+2}{10-2} = \frac{12}{8} = \frac{\Sigma(x_i - \hat{\mu}_1)^2 + \Sigma(x_i - \hat{\mu}_2)^2}{N-K}$

a) 0.5
b) 1.5
c) 2.5
d) 2.0

Classify $x = 2$ into population 1 or population 2.

$\hat{\pi}_1 = 0.5 \quad \hat{\pi}_2 = 0.5 \quad P_1(x) = \frac{\hat{\pi}_1 P(X = x \mid k=1)}{\hat{\pi}_1 P(X = x \mid k=1) + \hat{\pi}_2 P(X=x \mid k=2)}$

Result

X	Population	Posterior for Population 1	Posterior Population 2	Predicted Population
3	1	$P(Y=1 \mid X=3)$ 0.791	0.209 $=P(Y=2 \mid X=3)$	1
2	1	0.935	0.065	1
4	1	0.500	0.500	2
1	1	0.982	0.018	1
5	1	0.209	0.791	2
6	2	0.065	0.935	2
5	2	0.209	0.791	2
4	2	0.500	0.500	1
5	2	0.209	0.791	2
5	2	0.209	0.791	2

LDA in R

- Uses the MASS package.
- Function: lda(class∼ x1,prior = proportions).

```
> lda.r = lda(class.x~x[,1])
> pred.lda = predict(lda.r)
> pred.lda
$class
[1] 1 1 1 1 2 2 2 1 2 2
Levels: 1 2

$posterior
            1          2
1  0.79139147 0.20860853
2  0.93503083 0.06496917
3  0.50000000 0.50000000
4  0.98201379 0.01798621
5  0.20860853 0.79139147
6  0.06496917 0.93503083
7  0.20860853 0.79139147
8  0.50000000 0.50000000
9  0.20860853 0.79139147
10 0.20860853 0.79139147
```

```
> table(class.x,pred.lda$class)
```

Pred.

	class.x	1	2
actual	1	3	2
	2	1	4

$\frac{3}{10} = 0.3$ error rate

Example with Breast Cancer

```
> bc.lda = lda(Class ~ Cell.size,data = train)
> bc.lda

Prior probabilities of groups:
        0         1
0.6542969 0.3457031

Group means:
  Cell.size
0  1.340299
1  6.581921

Coefficients of linear discriminants:
                LD1
Cell.size 0.5788146
> plot(bc.lda)
> lda.pred = predict(bc.lda,test)
> table(test$Class,lda.pred$class)

    0   1
0 109   0
1  22  40
```

LDA for $p > 1$

p = # of predictors

- Assume that $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ is drawn from a multivariate normal distribution, with a class specific mean vector $\boldsymbol{\mu}_k$ and common covariance matrix $\underset{p \times p}{\boldsymbol{\Sigma}}$.

  $\underset{p \times 1}{\mu_k} =$

- The multivariate normal density function is

$$\underset{p \times p}{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \ddots & & \\ \vdots & & & \\ \sigma_{p1} & & & \sigma_p^2 \end{bmatrix} \quad f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2} \det(\boldsymbol{\Sigma}^{1/2})} e^{-(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})/2}$$

- Plugging the density function for the kth class, $f_k(X = x)$, into $p_k(x)$ reveals the Bayes classifier assigns and observation $X = x$ to the class for which

$$\delta_k(x) = \underbrace{\underset{n \times p}{x^T} \underset{p \times p}{\boldsymbol{\Sigma}^{-1}} \underset{p \times 1}{\boldsymbol{\mu}_k}}_{n \times 1} - \frac{1}{2} \boldsymbol{\mu}_k^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k + log(\pi_k)$$

is largest.

Estimates for $p > 1$

Given a data set, the estimates for $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}$, and π_k are as follows:

$$\hat{\boldsymbol{\mu}}_k = \begin{bmatrix} \mu_{1_k} \\ \mu_{2_k} \\ \vdots \\ \mu_{p_k} \end{bmatrix}$$

$$\hat{\boldsymbol{\Sigma}} = \sum_{i=1}^{K} \frac{n_i - 1}{N - K} \hat{\boldsymbol{\Sigma}}_i$$

$$\hat{\pi}_k = \frac{n_k}{n}$$

Where $\boldsymbol{\Sigma}_k$ is the variance-covariance matrix for the kth class and N is the total number of observations.

Example

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \qquad \sigma_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Consider two data sets:

$$\boldsymbol{X}_1 = \begin{bmatrix} 3 & 7 \\ 2 & 4 \\ 4 & 7 \end{bmatrix} \text{ and } \boldsymbol{X}_2 = \begin{bmatrix} 6 & 9 \\ 5 & 7 \\ 4 & 8 \end{bmatrix}$$

Then the estimates are:

$$\hat{\pi}_1 = \hat{\pi}_2 = 0.5$$

$$\hat{\mu}_1 = \begin{bmatrix} 3 \\ 6 \end{bmatrix} \text{ and } \hat{\mu}_2 = \begin{bmatrix} 5 \\ 8 \end{bmatrix}$$

$$\hat{\Sigma}_1 = \begin{bmatrix} 1 & 1.5 \\ 1.5 & 3 \end{bmatrix} \text{ and } \hat{\Sigma}_2 = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$

$$\hat{\Sigma} = \frac{3-1}{6-2} \begin{bmatrix} 1 & 1.5 \\ 1.5 & 3 \end{bmatrix} + \frac{3-1}{6-2} \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$$

```
> #For more than one predictor
> X = matrix(c(3,2,4,6,5,4,7,4,7,9,7,8),nrow = 6)
> class.x = rep(c(1,2),each = 3)
> X = as.matrix(cbind(X,class.x))
> lda.x2 = lda(class.x ~ X[,1:2])
> pred.lda2 = predict(lda.x2)
> pred.lda2
$class
[1] 1 1 2 2 2 2
Levels: 1 2

$posterior
           1          2
1 0.88079708 0.11920292
2 0.98201379 0.01798621
3 0.50000000 0.50000000
4 0.01798621 0.98201379
5 0.11920292 0.88079708
6 0.50000000 0.50000000

$x
            LD1
1 -1.000000e+00
2 -2.000000e+00
3  0.000000e+00
4  2.000000e+00
5  1.000000e+00
6  5.551115e-17
```

Breast Cancer with 3 variables

```
> bc.lda2 = lda(Class ~ Cell.size + Cl.thickness + Cell.shape,data = train)
> bc.lda2
Prior probabilities of groups:
        0         1
0.6542969 0.3457031

Group means:
  Cell.size Cl.thickness Cell.shape
0  1.340299     3.020896   1.429851
1  6.581921     7.158192   6.519774

Coefficients of linear discriminants:
                   LD1
Cell.size    0.2884712
Cl.thickness 0.2226278
Cell.shape   0.2320740
> lda.pred2 = predict(bc.lda2,test)
> table(test$Class,lda.pred2$class)

    0   1
0 109   0
1  11  51
```

Applications of LDA

- Bankruptcy
- Face recognition
- Biomedical studies: assessment of severity state of a patient and prognosis of disease outcome.
- Linear discriminant analysis also allows us to reduce dimensions.