

# Statistical Learning for Data Science

## Lecture 15

唐晓颖

电子与电气工程系  
南方科技大学

April 22, 2023

# Linear Discriminant Analysis (LDA)

- LDA for  $p=1$

Parameter estimations in LDA

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

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

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

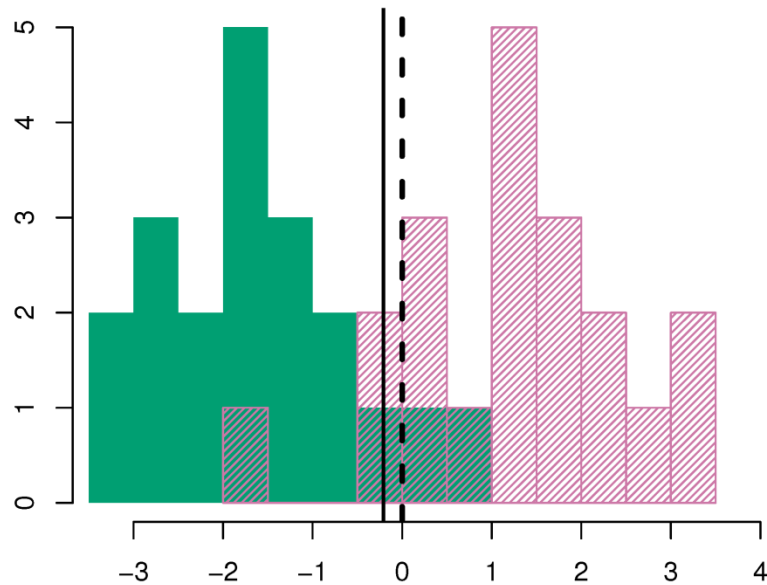
$$= \sum_{k=1}^K \frac{n_k - 1}{n - K} \cdot \hat{\sigma}_k^2$$

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

- $n$  -- the total number of training observations.
- $n_k$  -- the number of training observations in the  $k$ -th class.
- $\hat{\sigma}_k^2$  -- the estimated variance in the  $k$ -th class.
- $\hat{\sigma}^2$  -- a weighted average of the sample variance for each of the  $K$  classes.

# Linear Discriminant Analysis (LDA)

## ■ LDA for $p=1$



A sample of 20 observations for each of the two classes (green versus pink)

Dash – Bayes decision boundary;  
Solid – LDA decision boundary;

- Since  $n_1 = n_2 = 20$ , we have  $\hat{\pi}_1 = \hat{\pi}_2 = 0.5$
- The decision boundary corresponds to the midpoint between the sample means for the two classes,  $(\hat{\mu}_1 + \hat{\mu}_2) / 2$
- The LDA decision boundary is very close to the Bayes decision boundary, indicating the LDA is performing pretty well on this dataset.

# Linear Discriminant Analysis (LDA)

- LDA for  $p > 1$

We assume that  $X = (X_1, X_2, \dots, X_p)$  is drawn from a *multivariate Gaussian* (multivariate normal) distribution, with a class-specific mean vector and a common covariance matrix.

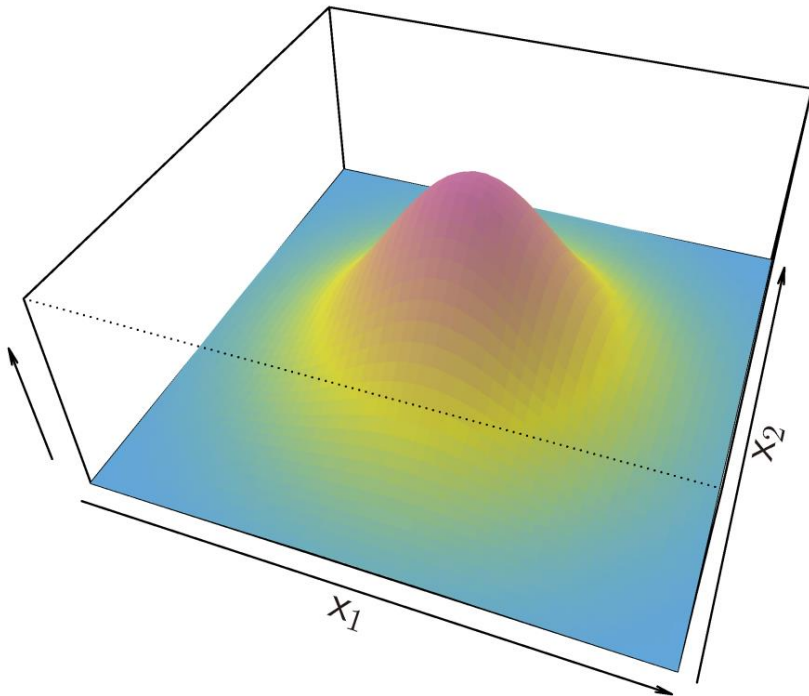
*The multivariate Gaussian assumes that each individual predictor follows a **one-dimension normal distribution**, with **some correlation** between each pair of predictors.*

# Linear Discriminant Analysis (LDA)

- LDA for  $p > 1$

*multivariate Gaussian*

Two predictors are uncorrelated



The surface has a characteristic *bell shape*

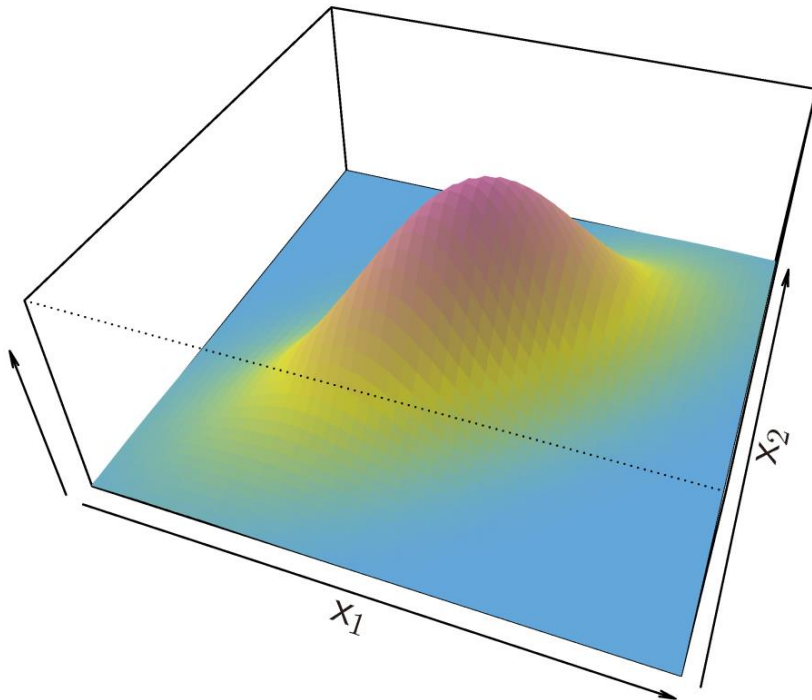
$$Var(X_1) = Var(X_2), Cor(X_1, X_2) = 0$$

# Linear Discriminant Analysis (LDA)

- LDA for  $p > 1$

*multivariate Gaussian*

Two predictors have a correlation of 0.7



*The bell shape will be distorted if the predictors are correlated or have unequal variances*

# Linear Discriminant Analysis (LDA)

- LDA for  $p > 1$

*multivariate Gaussian*

$$X \sim N(\mu, \Sigma)$$

$E(X) = \mu$  -- the mean of  $X$  (a vector with  $p$  components)

$Cov(X) = \Sigma$  -- the covariance of  $X$  (a matrix of size  $p \times p$  )

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

# Linear Discriminant Analysis (LDA)

- LDA for  $p > 1$

*multivariate Gaussian*

$$X \sim N(\mu, \Sigma)$$

In the case of  $p > 1$  predictors, the LDA classifier assumes that the observations in the  $k$ -th class are drawn from a multi-variate Gaussian distribution  $N(\mu_k, \Sigma)$ , where  $\mu_k$  is a class-specific mean vector, and  $\Sigma$  is a covariance matrix that is common to all  $K$  classes.



# Linear Discriminant Analysis (LDA)

- LDA for  $p > 1$

*multivariate Gaussian*

$$X \sim N(\mu, \Sigma)$$

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

Plugging into  $p_k(x) = \Pr(Y = k \mid X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^K \pi_l f_l(x)}$ , we have that the

Bayes classifier assigns an observation  $X = x$  to the class for which

$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log \pi_k \quad \text{-- discriminant function}$$

$$\delta_k(x) = c_{k0} + c_{k1}x_1 + \dots + c_{kp}x_p \quad \text{-- a linear function (this is the reason for the word *linear* in LDA)}$$

# Linear Discriminant Analysis (LDA)

- LDA for  $p > 1$

Bayes classifier assigns an observation

$X = x$  to the class for which

$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log \pi_k \quad \text{-- discriminant function}$$

$$\delta_k(x) = c_{k0} + c_{k1}x_1 + \dots + c_{kp}x_p \quad \text{-- a linear function (this is the reason for the word *linear* in LDA)}$$

The idea of the discriminant function is to compute one of these discriminant function for each of the classes, and then you classify it to the class for which it's largest. You pick the discriminant function that's largest.

# Linear Discriminant Analysis (LDA)

- LDA for  $p > 1$

Bayes classifier assigns an observation

$X = x$  to the class for which

$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log \pi_k$$

$$\delta_k(x) = c_{k0} + c_{k1}x_1 + \dots + c_{kp}x_p$$



$$c_{k0} = ?$$

$$c_{k1}x_1 + \dots + c_{kp}x_p = ?$$

$$\begin{aligned} c_{k0} &= -\frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log \pi_k \\ c_{k1}x_1 + \dots + c_{kp}x_p &= x^T \Sigma^{-1} \mu_k \end{aligned}$$

# Linear Discriminant Analysis (LDA)

- LDA for  $p > 1$

Note that the classification for an observation  $\vec{x}$  will be the one that maximizes:

$$\delta_k(\vec{x}) = \vec{x}^T \Sigma^{-1} \vec{\mu}_k - \frac{1}{2} \vec{\mu}_k^T \Sigma^{-1} \vec{\mu}_k + \log \pi_k$$

This is just the vector-matrix version of the formula

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

# Linear Discriminant Analysis (LDA)


- LDA for  $p > 1$

Once we have estimates  $\delta_k(x)$ , we can turn these into estimates for class probabilities

$$p_k(x) = \Pr(Y = k \mid X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^K \pi_l f_l(x)}$$

$$f_k(x) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu_k)^T \Sigma^{-1}(x - \mu_k)\right)$$

$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log \pi_k$$


$$p_k(x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^K \pi_l f_l(x)} = \frac{\pi_k \exp\left(x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k\right)}{\sum_{l=1}^K \pi_l \exp\left(x^T \Sigma^{-1} \mu_l - \frac{1}{2} \mu_l^T \Sigma^{-1} \mu_l\right)} = \frac{e^{\hat{\delta}_k(x)}}{\sum_{l=1}^K e^{\hat{\delta}_l(x)}}$$

# Linear Discriminant Analysis (LDA)

- LDA for  $p > 1$

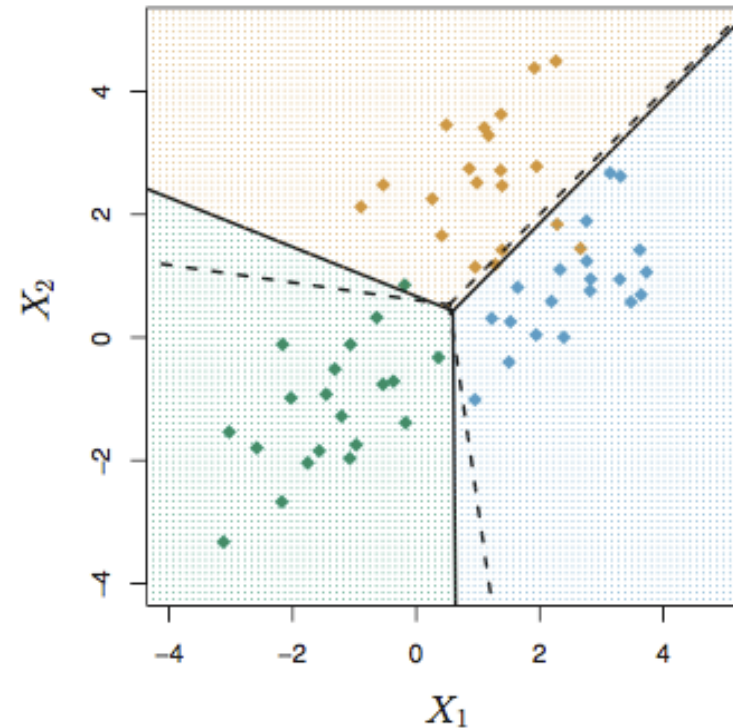
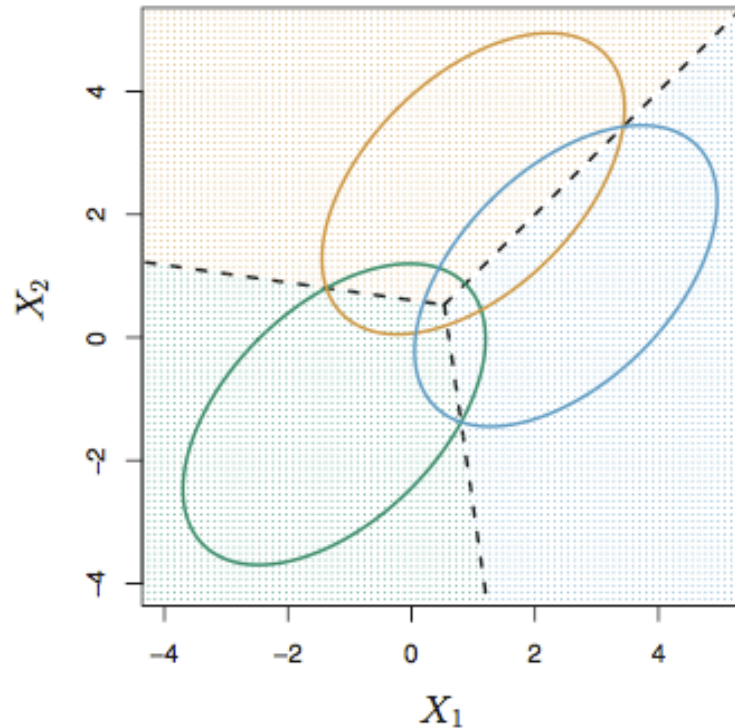
So classifying to the largest  $\hat{\delta}_k(x)$  amounts to classifying to the class for which  $\hat{\Pr}(Y = k \mid X = x)$  is largest.

When  $K = 2$ , we classify to class 2 if  $\hat{\Pr}(Y = k \mid X = x) \geq 0.5$  else to class 1.

# Linear Discriminant Analysis (LDA)

- LDA for  $p > 1$  and  $k > 2$

The linear discriminant nature of LDA still holds not only when  $p > 1$ , but also when  $K > 2$  for that matter as well. A picture can be very illustrative:

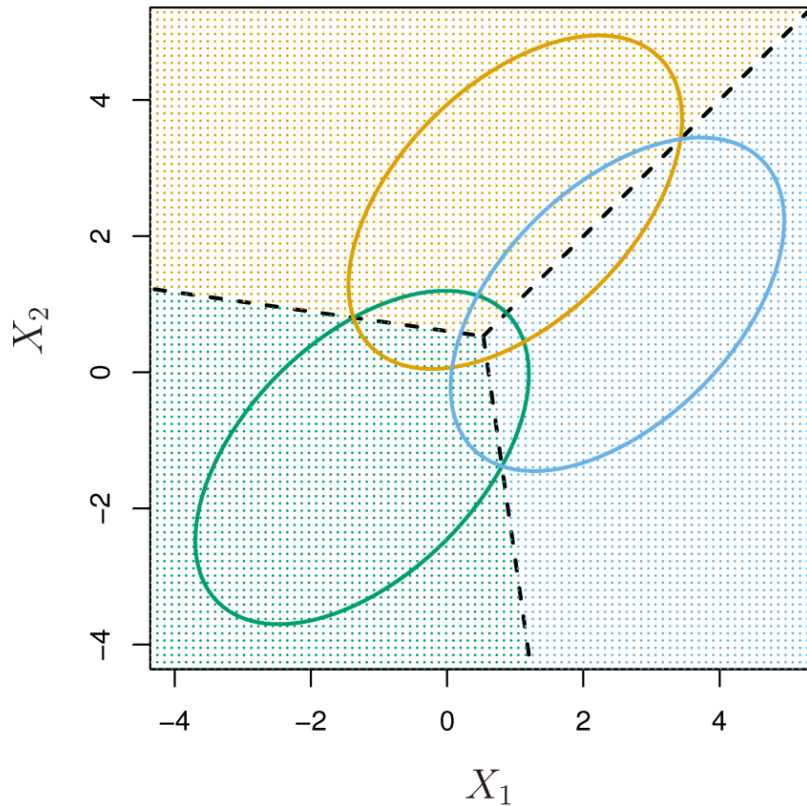


# Linear Discriminant Analysis (LDA)

- LDA for  $p > 1$

*multivariate Gaussian*

$$X \sim N(\mu, \Sigma)$$



- Three equally-sized Gaussian classes
- Class-specific mean vectors
- A common covariance matrix

Question:

$$\pi_1 = ? \quad \pi_1 = 1/3$$

$$\pi_2 = ? \quad \pi_2 = 1/3$$

$$\pi_3 = ? \quad \pi_3 = 1/3$$



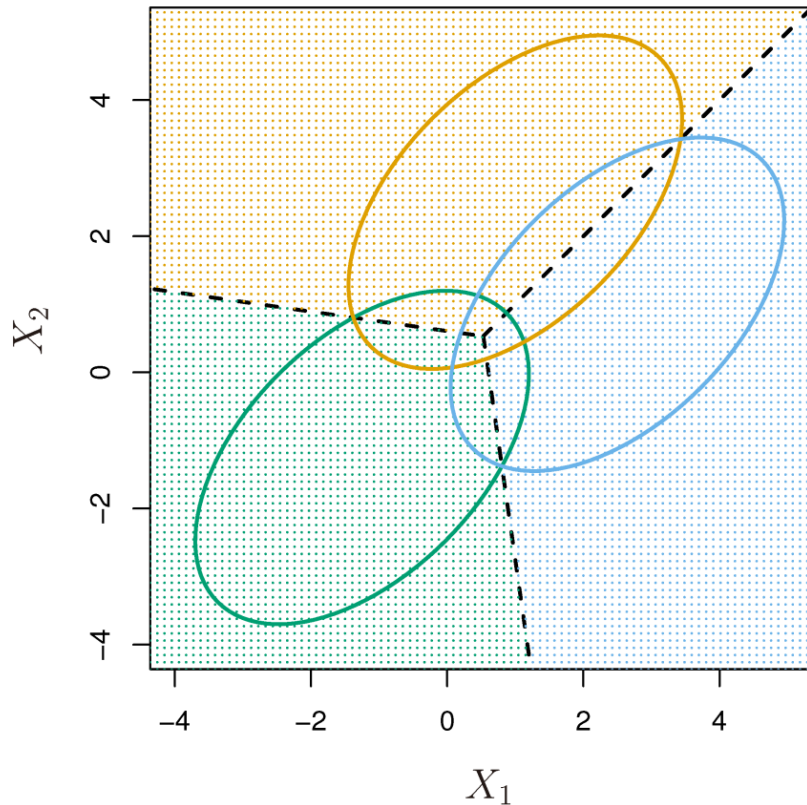


# Linear Discriminant Analysis (LDA)

- LDA for  $p > 1$

*multivariate Gaussian*

$$X \sim N(\mu, \Sigma)$$



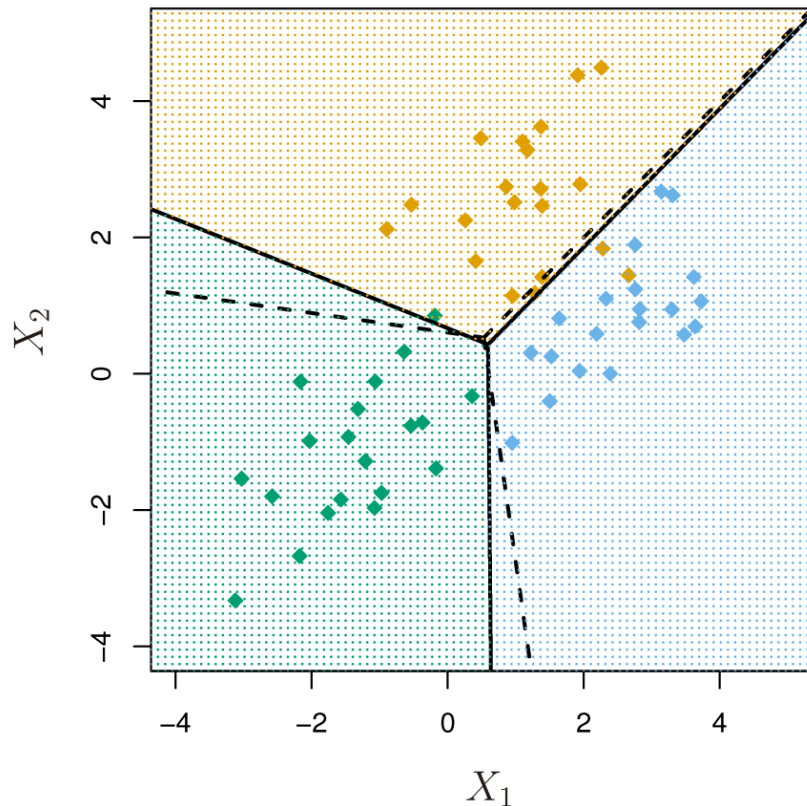
- Three equally-sized Gaussian classes
  - Class-specific mean vectors
  - A common covariance matrix
- 
- *Three ellipses* – regions that contain 95% of the probability for each of the three classes.
  - *Dash lines* – the Bayesian decision boundaries. Were they known, they would yield the fewest misclassification errors, among all possible classifiers.

# Linear Discriminant Analysis (LDA)

- LDA for  $p > 1$

*multivariate Gaussian*

$$X \sim N(\mu, \Sigma)$$



20 observations were generated from each class.

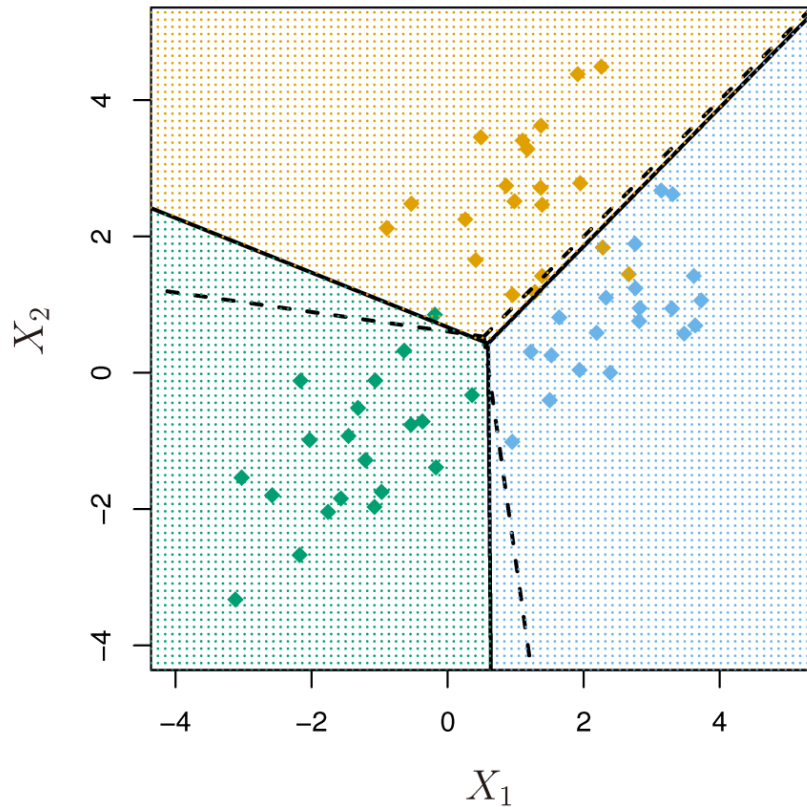
- *Dash lines* – the Bayesian decision boundaries. Were they known, they would yield the fewest misclassification errors, among all possible classifiers.
- *Solid lines* – the corresponding LDA decision boundaries.

# Linear Discriminant Analysis (LDA)

- LDA for  $p > 1$

*multivariate Gaussian*

$$X \sim N(\mu, \Sigma)$$



20 observations were generated from each class.

Overall, the LDA decision boundaries are pretty close to the Bayes decision boundaries. The test error rates for the Bayes and LDA classifiers are 0.0746 and 0.0770.



LDA is performing well on this data.

# Linear Discriminant Analysis (LDA)

- LDA for  $p > 1$

## Discriminant plot

When there are  $K$  classes, linear discriminant analysis can be viewed exactly in a  $K - 1$  dimensional plot. Because it essentially classifies to the closest centroid, and they span a  $K - 1$  dimensional plane. Even when  $K > 3$ , we can find the “best” 2-dimensional plane for visualizing the discriminant rule.

Purpose of discriminant plot: displays the observations in the feature space along the two or three most important linear discriminants. Each point in the plot represents an observation, and its color or shape indicates its class membership.

# Linear Discriminant Analysis (LDA)

- LDA for  $p > 1$

## Discriminant plot

Example:

For example, let Group A =  $\{(1,2), (2,1)\}$  and Group B =  $\{(9,10), (13,13)\}$ .

Now, the centroids are calculated as the centroids of the groups of data points so

Centroid of group A =  $((1+2)/2, (2+1)/2) = (1.5, 1.5)$

Centroid of group B =  $((9+13)/2, (10+13)/2) = (11, 11.5)$

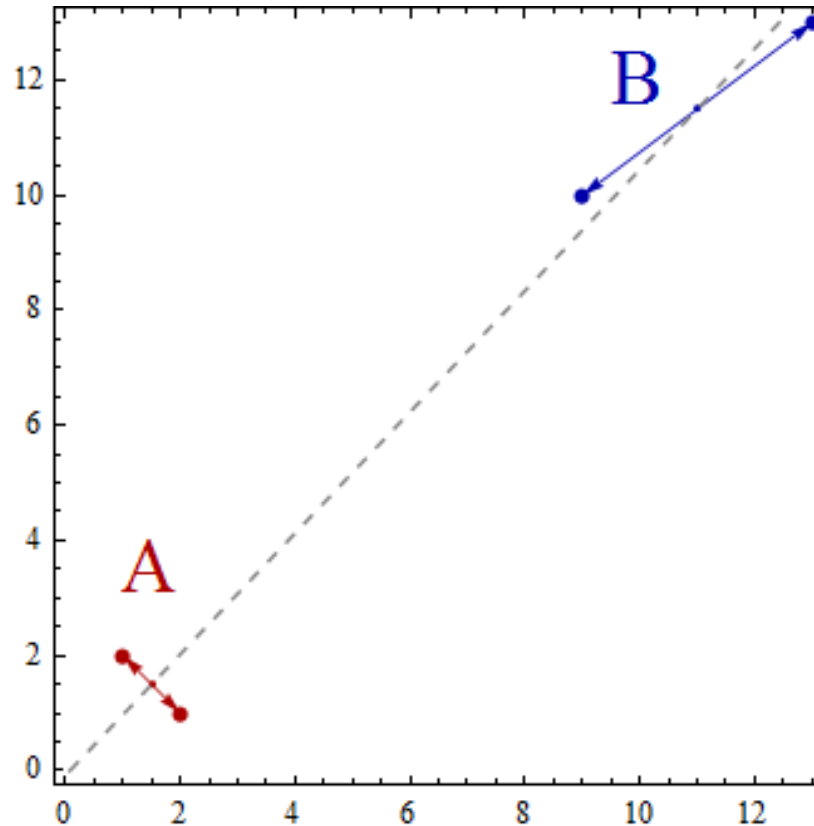
The Centroids are simply 2 points and they span a 1-dimensional line which joins them together.

# Linear Discriminant Analysis (LDA)

- LDA for  $p > 1$

Discriminant plot

Example:



# Linear Discriminant Analysis (LDA)

- Computations for LDA

$$\pi_k f_k(x) = \frac{\pi_k}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu_k)^T \Sigma^{-1}(x - \mu_k)\right)$$



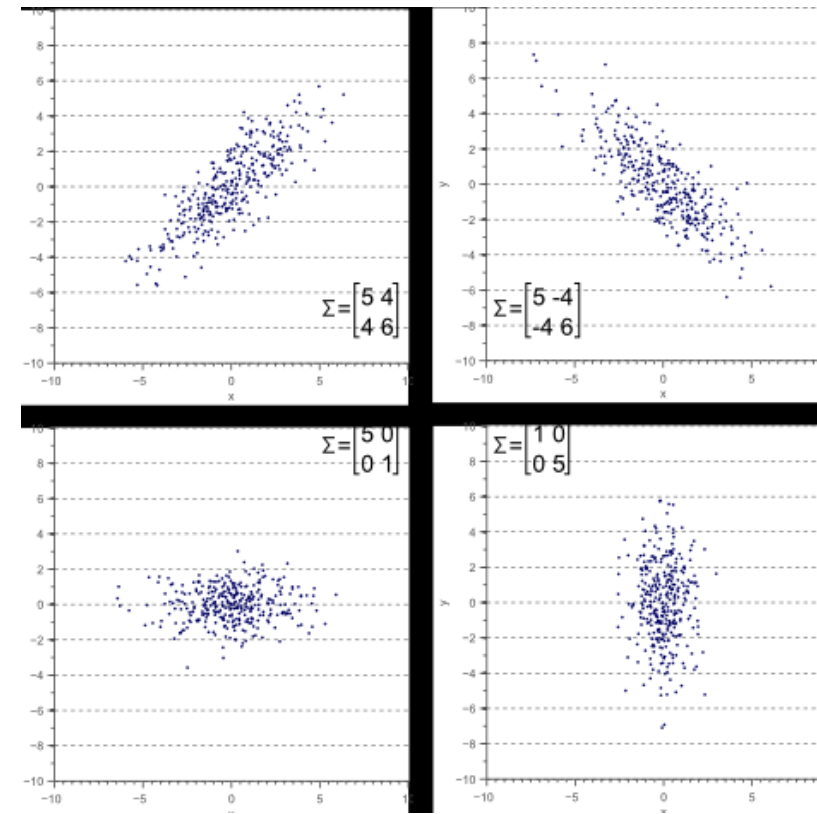
$$\delta_k(x) = -\frac{1}{2} \log |\Sigma| - \frac{1}{2}(x - \mu_k)^T \Sigma^{-1}(x - \mu_k) + \log \pi_k$$

Let's perform an eigen-decomposition of the covariance matrix

$$\Sigma = UDU^T$$

$D$  is diagonal matrix with elements  $d_l$ ,  $l = 1, 2, \dots, p$ .  $U$  is a  $p \times p$  orthonormal matrix.

<http://www.visiondummys.com/2014/04/geometric-interpretation-covariance-matrix/>



# Linear Discriminant Analysis (LDA)

- Computations for LDA

$$\delta_k(x) = -\frac{1}{2} \log |\Sigma| - \frac{1}{2} (x - \mu_k)^T \Sigma^{-1} (x - \mu_k) + \log \pi_k$$

Let's perform an eigen-decomposition of the covariance matrix

$$\Sigma = U D U^T$$

$D$  is diagonal matrix with elements  $d_l$ ,  $l = 1, 2, \dots, p$ .  $U$  is a  $p \times p$  orthonormal matrix.

$$\begin{aligned} (x - \mu_k)^T \Sigma^{-1} (x - \mu_k) &= \left[ U^T (x - \mu_k) \right]^T D^{-1} \left[ U^T (x - \mu_k) \right] \\ &= \left[ D^{-\frac{1}{2}} U^T (x - \mu_k) \right]^T \left[ D^{-\frac{1}{2}} U^T (x - \mu_k) \right] \end{aligned} \quad U^T = U^{-1}$$



# Linear Discriminant Analysis (LDA)

- Computations for LDA

$$\delta_k(x) = -\frac{1}{2} \log |\Sigma| - \frac{1}{2} (x - \mu_k)^T \Sigma^{-1} (x - \mu_k) + \log \pi_k$$

$$(x - \mu_k)^T \Sigma^{-1} (x - \mu_k) = \left[ D^{-\frac{1}{2}} U^T (x - \mu_k) \right]^T \left[ D^{-\frac{1}{2}} U^T (x - \mu_k) \right]$$

## Why do I do this?

What we are trying to establish here is a linear transform on the original data.  $D^{-\frac{1}{2}} U^T$  is a square matrix. You use this square matrix to multiply the original column vector to get the linear transform of the data. This quadratic term is equivalent to **doing a linear transform on the data and then after the transform, multiplied its own transpose**. This gives you the square of the norm of the transformed vector.

### Euclidean norm [ edit ]

*Main article: Euclidean distance*

On an  $n$ -dimensional [Euclidean space](#)  $\mathbb{R}^n$ , the intuitive notion of length of the vector  $\mathbf{x} = (x_1, x_2, \dots, x_n)$  is captured by the formula

$$\|\mathbf{x}\|_2 := \sqrt{x_1^2 + \dots + x_n^2}.$$

This is the Euclidean norm, which gives the ordinary distance from the origin to the point  $\mathbf{x}$ , a consequence of the [Pythagorean theorem](#). This operation may also be referred to as "SRSS" which is an acronym for the square root of the sum of squares.<sup>[2]</sup>

The Euclidean norm is by far the most commonly used norm on  $\mathbb{R}^n$ , but there are other norms on this vector space as will be shown below. However, all these norms are equivalent in the sense that they all define the same topology.

On an  $n$ -dimensional [complex space](#)  $\mathbb{C}^n$  the most common norm is

$$\|\mathbf{z}\| := \sqrt{|z_1|^2 + \dots + |z_n|^2} = \sqrt{z_1 \bar{z}_1 + \dots + z_n \bar{z}_n}.$$

In both cases the norm can be expressed as the [square root](#) of the [inner product](#) of the vector and itself:

$$\|\mathbf{x}\| := \sqrt{\mathbf{x}^* \mathbf{x}},$$

where  $\mathbf{x}$  is represented as a [column vector](#)  $[(x_1; x_2; \dots; x_n)]$ , and  $\mathbf{x}^*$  denotes its [conjugate transpose](#).

This formula is valid for any [inner product space](#), including Euclidean and complex spaces. For Euclidean spaces, the inner product is equivalent to the [dot product](#). Hence, in this specific case the formula can be also written with the following notation:

$$\|\mathbf{x}\| := \sqrt{\mathbf{x} \cdot \mathbf{x}}.$$

The Euclidean norm is also called the [Euclidean length](#), [L<sup>2</sup> distance](#), [l<sup>2</sup> distance](#), [L<sup>2</sup> norm](#), or [l<sup>2</sup> norm](#); see [L<sup>p</sup> space](#).

The set of vectors in  $\mathbb{R}^{n+1}$  whose Euclidean norm is a given positive constant forms an [n-sphere](#).

# Linear Discriminant Analysis (LDA)

- Computations for LDA

$$\delta_k(x) = -\frac{1}{2} \left[ D^{-\frac{1}{2}} U^T (x - \mu_k) \right]^T \left[ D^{-\frac{1}{2}} U^T (x - \mu_k) \right] + \log \pi_k$$

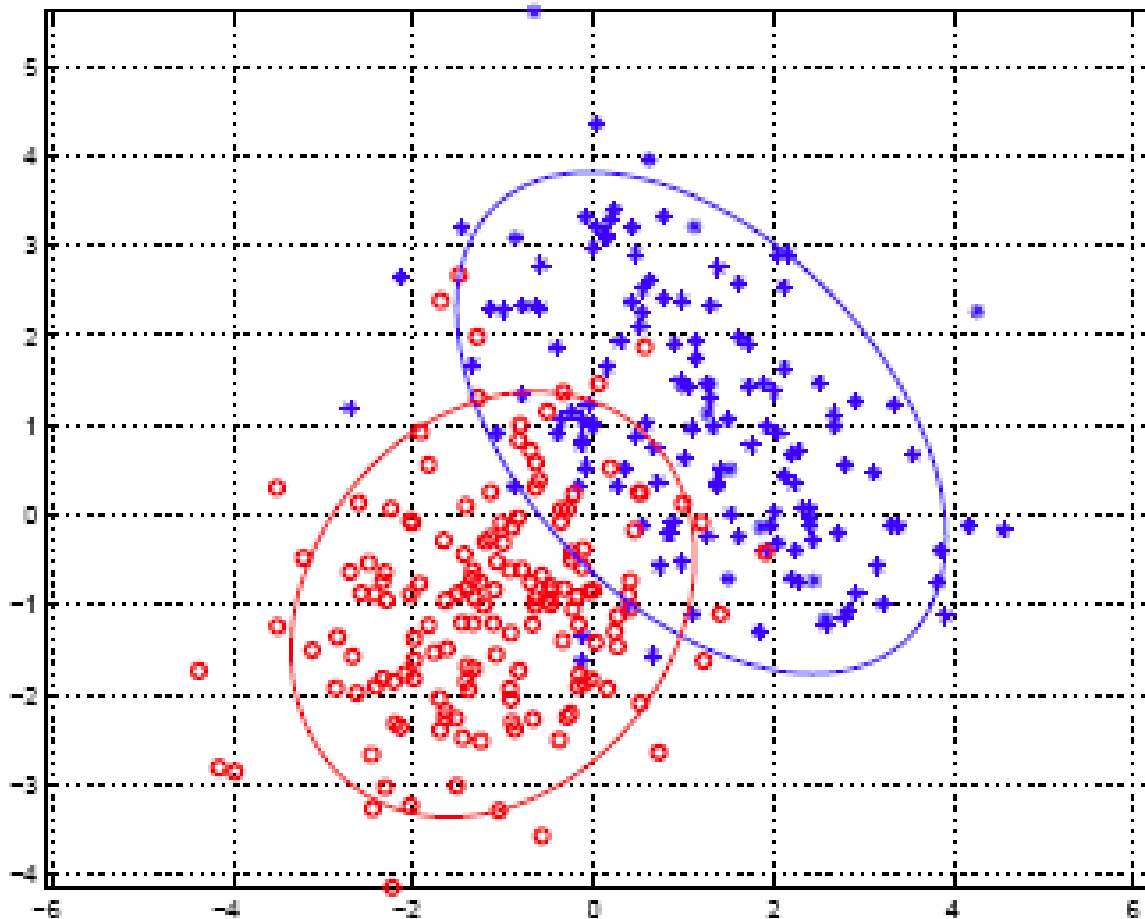
Let's take a look at what is going on in LDA geometrically. For simplicity, let's assume uniform prior probabilities, i.e. all the classes have the same prior probabilities.

- Sphere the data  $D^{-\frac{1}{2}} U^T x \rightarrow x^*$  and  $D^{-\frac{1}{2}} U^T \mu_k \rightarrow \mu_k^*$ .
- For the transformed data and class centroids, classify  $x^*$  to the closest class centroid in the transformed space, modulo the effect of the class prior probabilities  $\pi_k$ . For equal prior, simply find the closest class centroid (in Euclidean distance) and classify to the corresponding class.

# Linear Discriminant Analysis (LDA)

- Geometric illustration of LDA

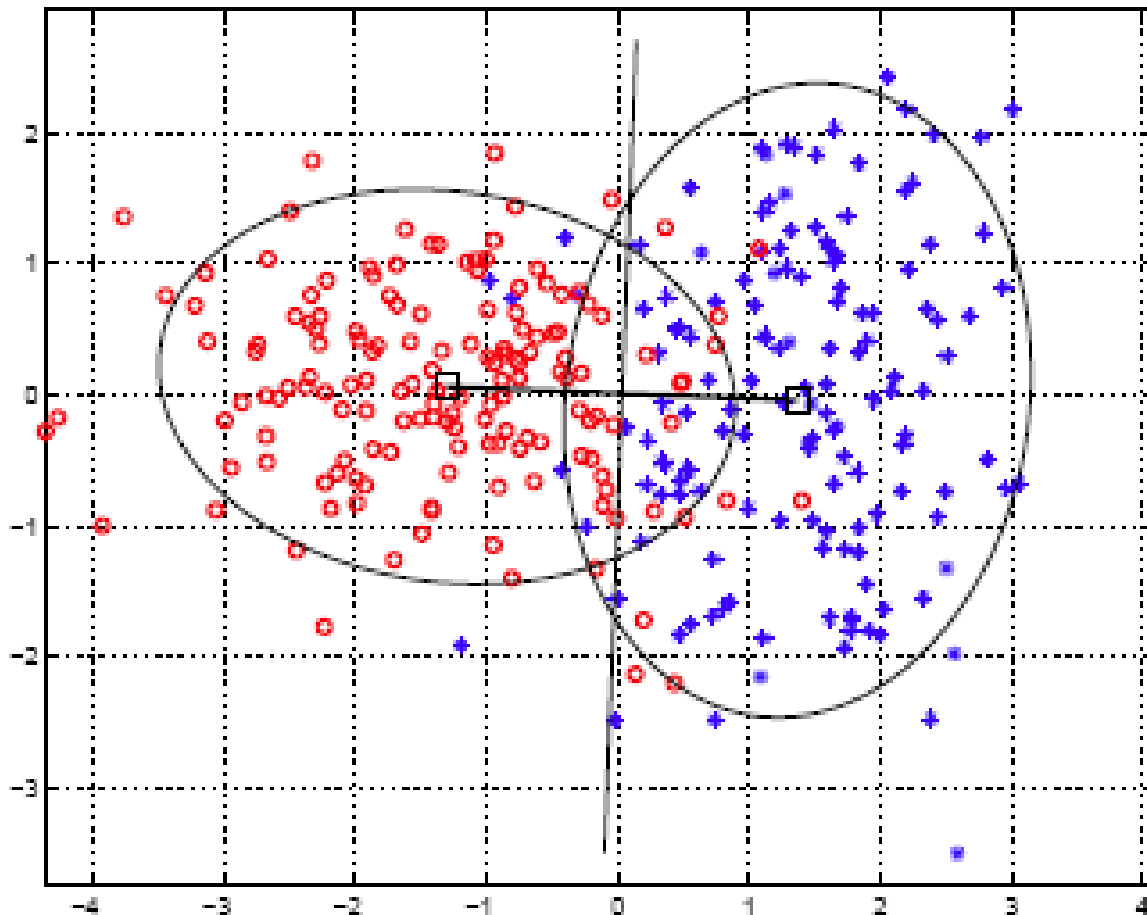
Original data



# Linear Discriminant Analysis (LDA)

- Geometric illustration of LDA

Transformed data

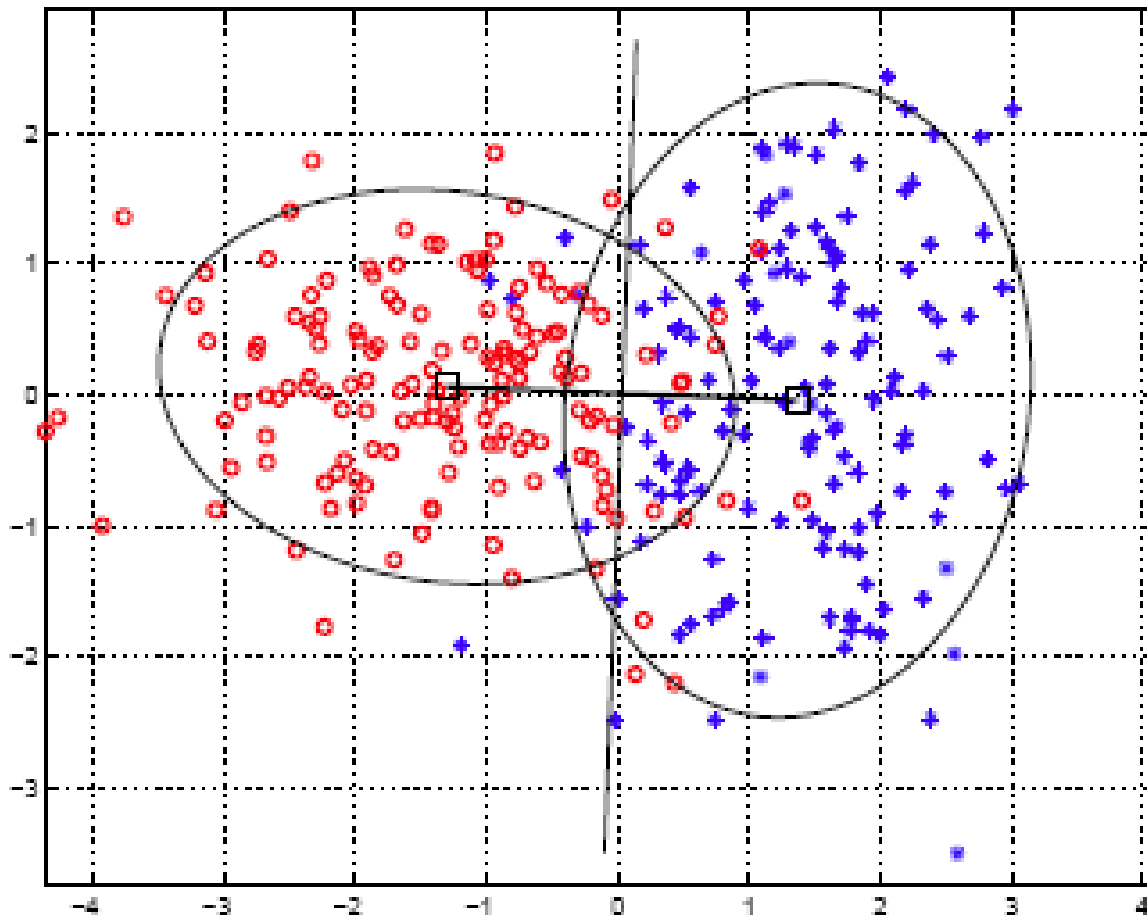


To classify a new test data point, we first apply the same transform to the data, then we will compute the distance to either mean and see which one is closer to the transformed test data point. If we have equal priors in this case, the decision boundary is simply a line perpendicular to the line connecting the two mean vectors (in the transformed space).

# Linear Discriminant Analysis (LDA)

- Geometric illustration of LDA

Transformed data



If you do not have uniform priors, i.e. one prior is higher than the other, then this decision boundary would be shifted, but it will still be perpendicular to the line connecting the two mean factors in the transformed space.

# Linear Discriminant Analysis (LDA)

- LDA on the default data

Goal: predict whether or not an individual will default on the basis of credit card balance and student status based on 10,000 training samples

		<i>True Default Status</i>		
		No	Yes	Total
<i>Predicted Default Status</i>	No	9644	252	9896
	Yes	23	81	104
Total		9667	333	10000

A confusion matrix

$$\text{Error rate} = (23+252)/10000 = 2.75\%$$

# Linear Discriminant Analysis (LDA)

- LDA on the default data

Goal: predict whether or not an individual will default on the basis of credit card balance and student status based on 10,000 training samples

Error rate =  $(23+252)/10000 = 2.75\%$

Some caveats:

- This is *training* error. And we may be overfitting. (The higher the ratio of  $p/n$ , the more we expect this overfitting to play a role. For this data, we have  $p/n = 2/10000$ . So overfitting won't be a big concern for this case.)
- Since only 3.33% ( $333/10000$ ) of the individuals in the training sample defaulted, a simple but useless classifier that always predicts that each individual will not default, regardless of his or her credit card balance and student status, will result in an error rate of 3.33%, only a bit higher than the LDA training set error rate.

# Linear Discriminant Analysis (LDA)

- Class-specific performance

Sensitivity – the percentage of true defaulters that are identified  
(81/333 = 24.3%)

Specificity – the percentage of non-defaulters that are correctly identified (9644/9667 = 99.8%)

**Comment:** LDA is trying to approximate the Bayes classifier, which has the lowest *total error rate* out of all classifiers. That is , the Bayes classifier will yield the smallest possible total number of misclassified observations, irrespective of which class the errors come from.

		<i>True Default Status</i>		
		No	Yes	Total
<i>Predicted Default Status</i>	No	9644	252	9896
	Yes	23	81	104
Total		9667	333	10000



# Linear Discriminant Analysis (LDA)

- Types of errors

False positive rate – the fraction of negative samples that are classified as positive ( $23/9667 = 0.2\%$ )

False negative rate – the fraction of positive samples that are classified as negative ( $252/333 = 75.7\%$ )

		<i>True Default Status</i>		
		No	Yes	Total
<i>Predicted Default Status</i>	No	9644	252	9896
	Yes	23	81	104
	Total	9667	333	10000