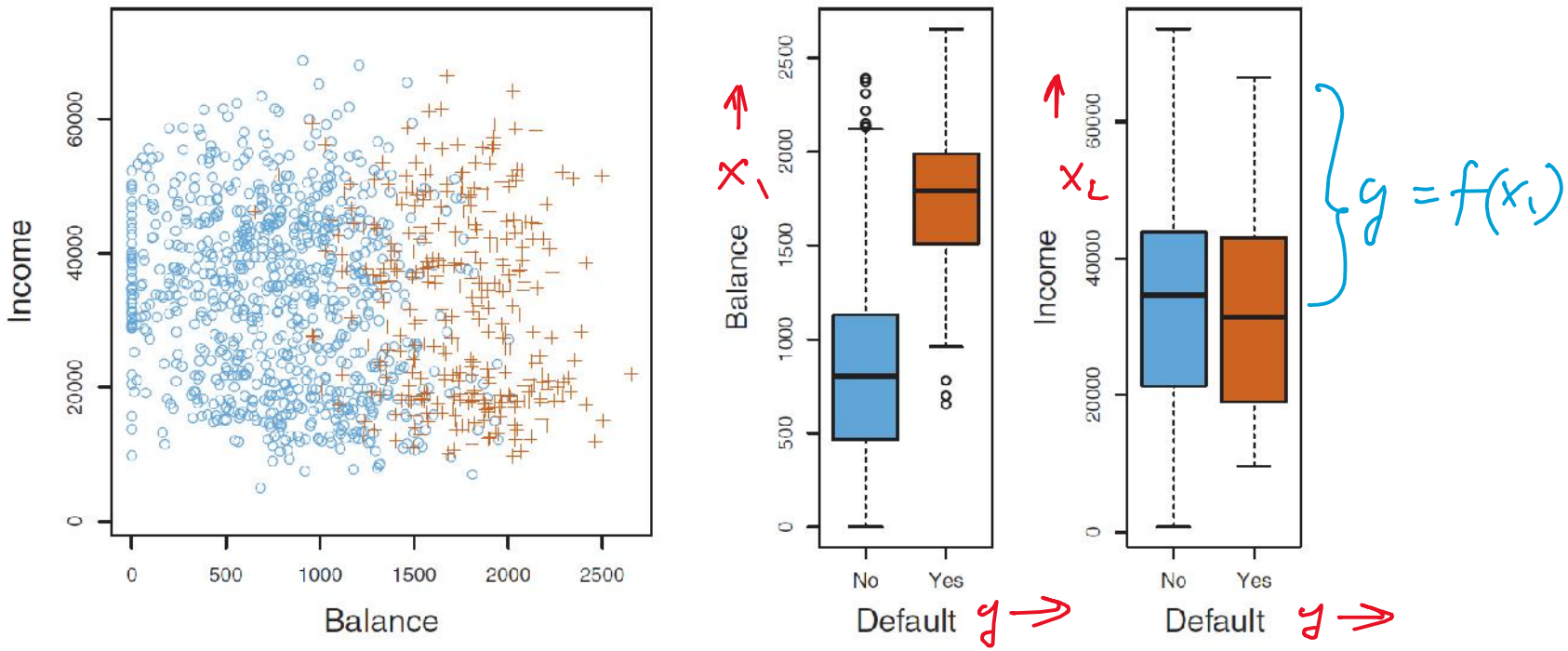
Classification

Prof. Asim Tewari
IIT Bombay

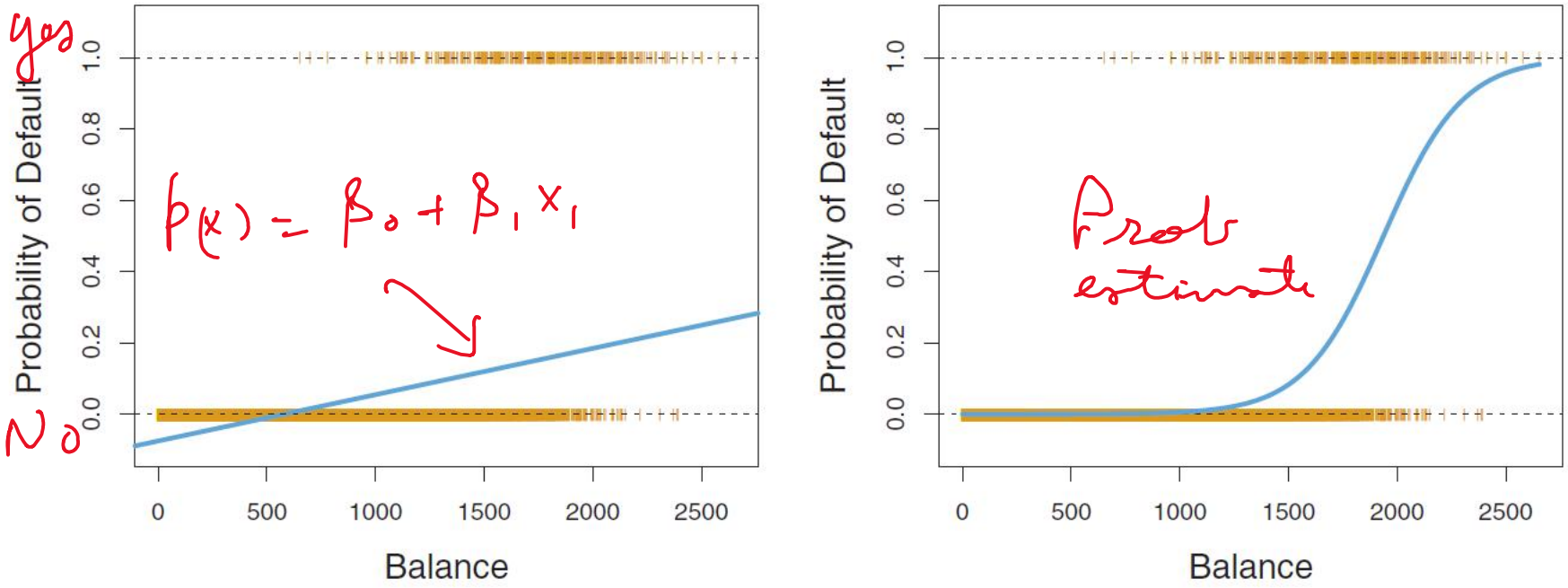
What is *classification?*

- The linear regression models assumes that the response variable *Y* is quantitative. But in many situations, the response variable is instead *qualitative.*
- For example, eye color is qualitative, taking qualitative on values blue, brown, or green. Often qualitative variables are referred to as *categorical.*
- Approaches for predicting qualitative responses, a process that is known as *classification*



The **Default** *data set.* Left: *The annual incomes and monthly credit card balances of a number of individuals. The individuals who defaulted on their credit card payments are shown in orange, and those who did not are shown in blue.* Center: *Boxplots of* balance *as a function of* default *status.* Right: *Boxplots of* income *as a function of* default *status.*

For binary response variable Y



- *Classification using the* Default *data.* Left: *Estimated probability of* default *using linear regression. Some estimated probabilities are negative! The orange ticks indicate the 0/1 values coded for* default*(*No *or* Yes*).* Right: *Predicted probabilities of* default *using logistic regression. All probabilities lie between* 0 *and* 1*.*

The Logistic Model

- A linear regression model to represent these probabilities:

$$p(X) = \beta_0 + \beta_1 X.$$

- In logistic regression, we use the *logistic function*:

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

The Logistic Model

- The quantity $p(X)/[1-p(X)]$ is called the *odds*, and can take on any value odds between 0 and ∞.

$$\frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X}$$

- Values of the odds close to 0 and ∞ indicate very low and very high probabilities of default, respectively. For example, on average 1 in 5 people with an odds of 1/4 will default, since $p(X) = 0.2$ implies an odds of $0.2/(1-0.2) = 1/4$. Likewise on average nine out of every ten people with an odds of 9 will default, since $p(X) = 0.9$ implies an odds of $0.9/(1-0.9) = 9$.

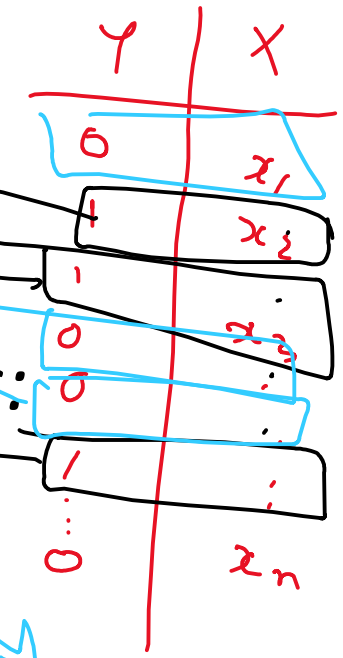
The Logistic Model

- We can define *log-odds* or *logit* as

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

The Logistic Model

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



- *Estimating the Regression Coefficients:*
- *Likelihood function*

$$\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'}))$$

- The basic intuition behind using maximum likelihood to fit a logistic regression model is as follows: we seek estimates for *β*0 and *β*1 such that the predicted probability ˆ*p*(*xi*) of default for each individual, corresponds as closely as possible to the individual's observed default status.

The Logistic Model

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

	Coefficient	Std. error	Z-statistic	P-value
Intercept	−10.6513	0.3612	−29.5	<0.0001
balance	0.0055	0.0002	24.9	<0.0001

- *For the* Default *data, estimated coefficients of the logistic regression model that predicts the probability of* default *using* balance*. A one-unit increase in* balance *is associated with an increase in the log odds of* default *by* 0.0055 *units.*

The Logistic Model

- *Estimating the Regression Coefficients:*
- *Likelihood function*

$$\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'}))$$

$$p(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

- The basic intuition behind using maximum likelihood to fit a logistic regression model is as follows: we seek estimates for $\beta 0$ and $\beta 1$ such that the predicted probability $\hat{p}(xi)$ of default for each individual, corresponds as closely as possible to the individual's observed default status.

The Logistic Model

Making Predictions

- *We* predict that the default probability for an individual with a balance of $1, 000 is

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 1,000}}{1+e^{-10.6513+0.0055\times 1,000}} = 0.00576$$

- Which is below 1%. In contrast, the predicted probability of default for an individual with a balance of $2, 000 is much higher, and equals 0.586 or 58.6%.

The Logistic Model

- Predictors with more than two levels: One can use qualitative predictors (with levels) with the logistic regression model using the dummy variable approach.
- *Multiple Logistic Regression*

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

Classification of response variable with more than two classes

- Why Not Linear Regression?
- We could consider encoding these values as a quantitative response variable, Y , as follows:

$$Y = \begin{cases} 1 & \text{if stroke;} \\ 2 & \text{if drug overdose;} \\ 3 & \text{if epileptic seizure.} \end{cases}$$

- Unfortunately, this coding implies an ordering on the outcomes, putting drug overdose in between stroke and epileptic seizure, and insisting that the difference between stroke and drug overdose is the same as the difference between drug overdose and epileptic seizure.

Classification of response variable with more than two classes

- *The Logistic Model*

$$\log \frac{\Pr(G=1|X=x)}{\Pr(G=K|X=x)} = \beta_{10} + \beta_1^T x$$
$$\log \frac{\Pr(G=2|X=x)}{\Pr(G=K|X=x)} = \beta_{20} + \beta_2^T x$$
$$\vdots$$
$$\log \frac{\Pr(G=K-1|X=x)}{\Pr(G=K|X=x)} = \beta_{(K-1)0} + \beta_{K-1}^T x$$

  - The model is specified in terms of K − 1 log-odds or logit transformations (reflecting the constraint that the probabilities sum to one).
  - Although the model uses the last class as the denominator in the odds-ratios, the choice of denominator is arbitrary in that the estimates are equivariant under this choice.

$$\Pr(G=k|X=x) = \frac{\exp(\beta_{k0} + \beta_k^T x)}{1 + \sum_{\ell=1}^{K-1} \exp(\beta_{\ell 0} + \beta_\ell^T x)}, \quad k = 1, \ldots, K-1,$$
$$\Pr(G=K|X=x) = \frac{1}{1 + \sum_{\ell=1}^{K-1} \exp(\beta_{\ell 0} + \beta_\ell^T x)},$$

Bayes' Theorem

You are planning a picnic today, but the morning is cloudy

- Oh no! 50% of all rainy days start off cloudy!
- But cloudy mornings are common (about 40% of days start cloudy)
- And this is usually a dry month (only 3 of 30 days tend to be rainy, or 10%)
- **What is the chance of rain during the day?**

Bayes' Theorem

Bayes' Theorem is a way of finding a probability when we know certain other probabilities.
The formula is:

$$P(A|B) = P(A)\, P(B|A)/P(B)$$

Which tells us:
how often A happens given that B happens, written P(A|B),
When we know:
how often B happens given that A happens, written P(B|A)
and how likely A is on its own, written P(A) and
how likely B is on its own, written P(B)

Bayes' Theorem

We will use Rain to mean rain during the day, and Cloud to mean cloudy morning.

The chance of Rain given Cloud is written P(Rain|Cloud)

So let's put that in the formula:

P(Rain|Cloud) = P(Rain) P(Cloud|Rain)/P(Cloud)

- P(Rain) is Probability of Rain = 10%
- P(Cloud|Rain) is Probability of Cloud, given that Rain happens = 50%
- P(Cloud) is Probability of Cloud = 40%
- P(Rain|Cloud) = 0.1 x 0.50.4 = .125

Or a 12.5% chance of rain. Not too bad, let's have a picnic!

Fitting Logistic Regression Models

- For Binary response

Classification of response variable with more than two classes

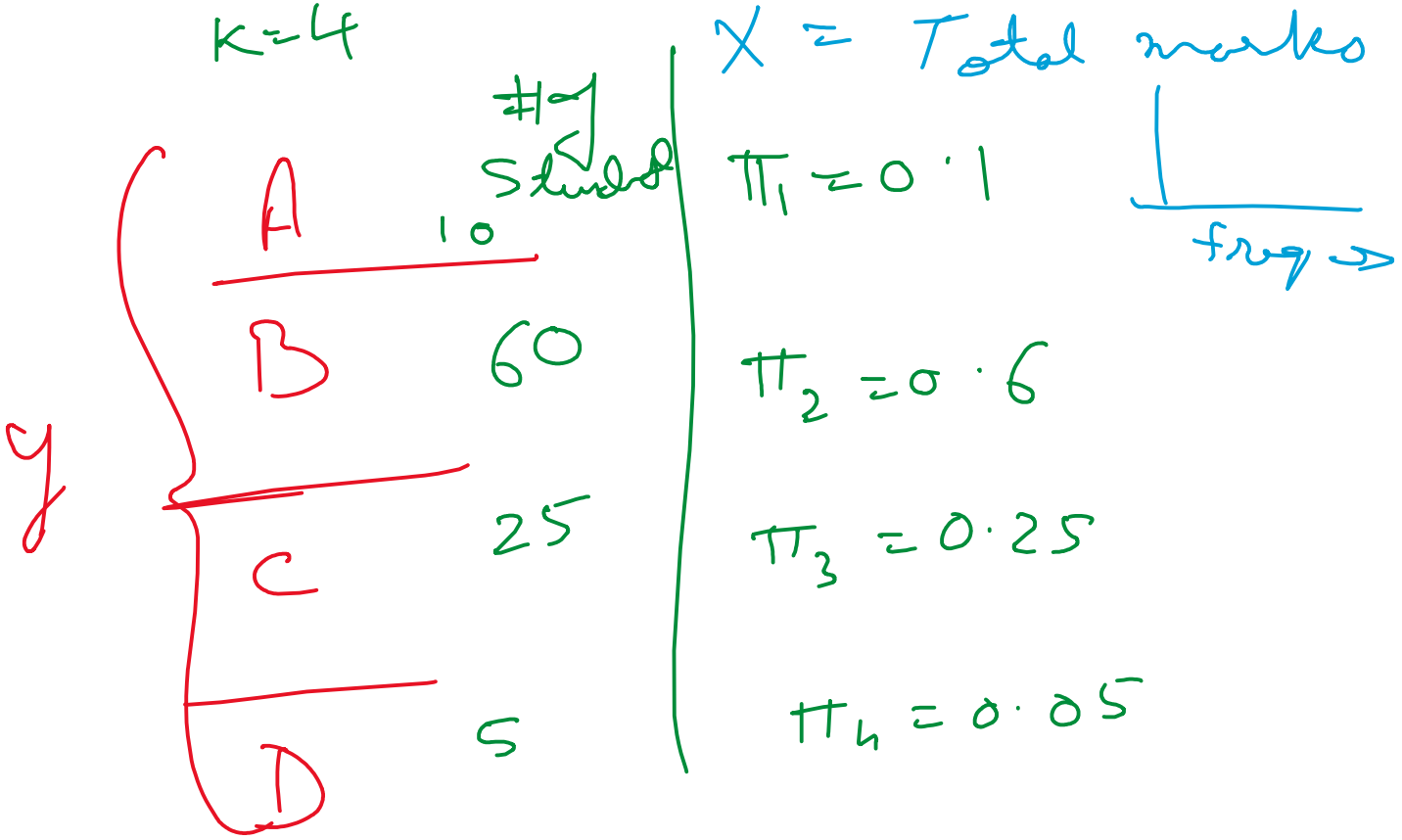
Classification of response variable with more than two classes

Linear Discriminant Analysis

- Suppose that we wish to classify an observation into one of K classes, where $K \geq 2$. In other words, the qualitative response variable Y can take on K possible distinct and unordered values.
- Let π_k represent the overall or *prior* probability that a randomly chosen observation comes from the kth class; this is the probability that a given observation is associated with the kth category of the response variable Y .
- Let $f_k(X) \equiv \Pr(X = x \mid Y = k)$ denote the *density function* of X for an observation that comes from the kth class.

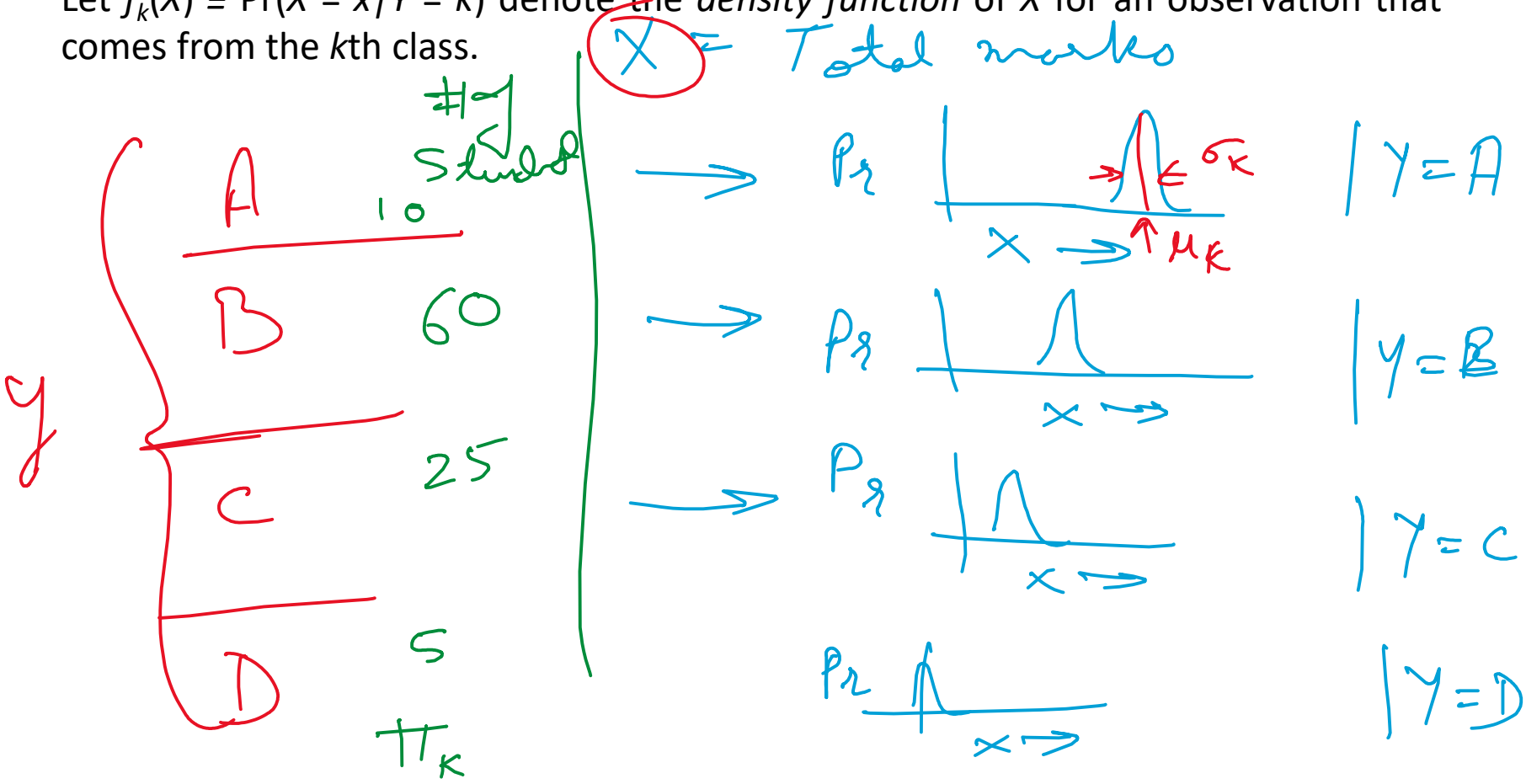
Linear Discriminant Analysis

Let π_k represent the overall or *prior* probability that a randomly chosen observation comes from the *k*th class; this is the probability that a given observation is associated with the *k*th category of the response variable *Y* .



Linear Discriminant Analysis

Let $f_k(X) \equiv \Pr(X = x | Y = k)$ denote the *density function* of X for an observation that comes from the kth class.



Linear Discriminant Analysis

- *Then Bayes' theorem* states that

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

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

Linear Discriminant Analysis

- *Linear Discriminant Analysis for p* = 1

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

- let us further assume that $\sigma^2_1 = \ldots = \sigma^2_K$

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_l)^2\right)}$$

Linear Discriminant Analysis

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_l)^2\right)}$$

π_k, μ_k

$x^2 - 2\mu_k x + \mu_k^2$

$\sigma_1 = \sigma_2 = \sigma_3 = \sigma$

$$p_1(x_i) = \frac{g_1(x_i, k=1)}{g_0(x_i)}$$

$$p_2(x_i) = \frac{g_1(x_i, k=2)}{g_0(x_i)}$$

$$p_3(x_i) = \frac{g_1(x_i, k=3)}{g_0(x_i)}$$

$\ln[p_k(x)]$

$$\max \ln(\pi_k) + \frac{2\mu_k x}{2\sigma^2} - \frac{\mu_k^2}{2\sigma^2}$$

Linear Discriminant Analysis

- this is equivalent to assigning the observation to the class for which

$$2x(\mu_1 - \mu_2) > \mu_1^2 - \mu_2^2 \Rightarrow \text{Class } 1$$

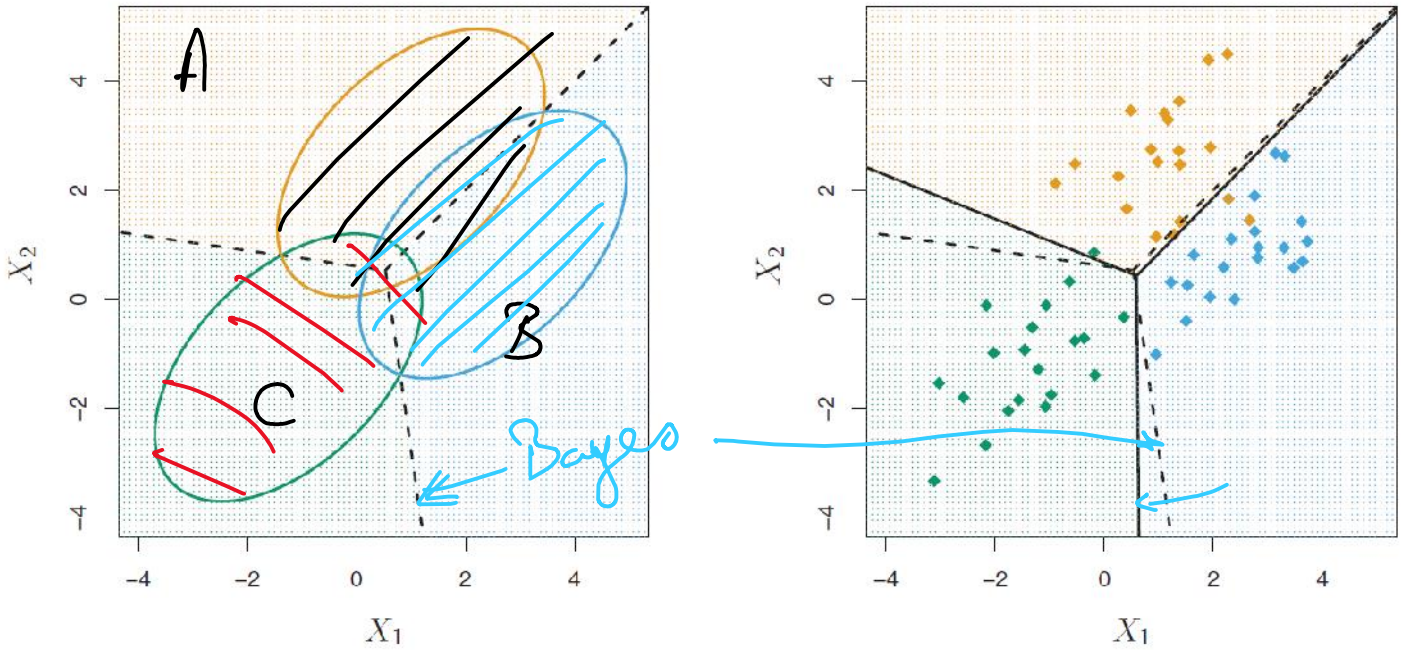
$$\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

- is largest. For instance, if $K = 2$ and $\pi 1 = \pi 2$, then the Bayes classifier assigns an observation to class 1 if $2x(\mu 1 - \mu 2) > \mu^2 1 - \mu^2 2$, and to class 2 otherwise. In this case, the Bayes decision boundary corresponds to the point where

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

A
B
C
$y = f(x_1, x_2)$

Linear Discriminant Analysis



- *An example with three classes. The observations from each class are drawn from a multivariate Gaussian distribution with p = 2, with a class-specific mean vector and a common covariance matrix.* Left: *Ellipses that contain 95% of the probability for each of the three classes are shown. The dashed lines are the Bayes decision boundaries.* Right: 20 *observations were generated from each class, and the corresponding LDA decision boundaries are indicated using solid black lines. The Bayes decision boundaries are once again shown as dashed lines.*

Linear Discriminant Analysis

Linear Discriminant Analysis for p >1

To do this, we will assume that $X = (X1, X2, \ldots, Xp)$ is drawn from a *multivariate Gaussian* (or multivariate normal) distribution, with a class-specific multivariate mean vector and a common covariance matrix.

Linear Discriminant Analysis

Linear Discriminant Analysis for p >1

To indicate that a p-dimensional random variable X has a multivariate Gaussian distribution, we write $X \sim N(\mu, \mathbf{\Sigma})$. Here $E(X) = \mu$ is the mean of X (a vector with p components), and $\text{Cov}(X) = \mathbf{\Sigma}$ is the $p \times p$ covariance matrix of X. Formally, the multivariate Gaussian density is defined as

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

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

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

is largest.

Linear Discriminant Analysis

		True default status		
		No	Yes	Total
Predicted default status	No	9,644	252	9,896
	Yes	23	81	104
	Total	9,667	333	10,000

- *A confusion matrix compares the LDA predictions to the true default statuses for the* 10, 000 *training observations in the* Default *data set. Elements on the diagonal of the matrix represent individuals whose default statuses were correctly predicted, while off-diagonal elements represent individuals that were misclassified. LDA made incorrect predictions for* 23 *individuals who did not default and for* 252 *individuals who did default.*

Linear Discriminant Analysis

		True default status		
		No	Yes	Total
Predicted default status	No	9,644	252	9,896
	Yes	23	81	104
	Total	9,667	333	10,000

Since only 3.33% 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%.

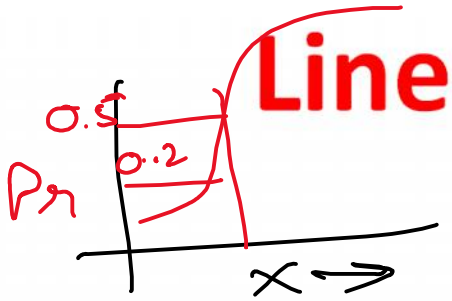
→ 96.67 %

Linear Discriminant Analysis

The Bayes classifier works by assigning an observation to the class for which the posterior probability *pk*(*X*) is greatest. In the two-class case, this amounts to assigning an observation to the *default* class if

$$\Pr(\texttt{default} = \texttt{Yes}|X = x) > 0.5.$$

Linear Disc



		True default status		
		No	Yes	Total
Predicted default status	No	9,644	252	9,896
	Yes	23	81	104
	Total	9,667	333	10,000

$Pr(\text{Yes}|x) > 0.5$

		True default status		
		No	Yes	Total
Predicted default status	No	9,432	138	9,570
	Yes	235	195	430
	Total	9,667	333	10,000

$Pr(\text{Yes}|x) > 0.2$

A confusion matrix compares the LDA predictions to the true default statuses for the 10, 000 *training observations in the* Default *data set, using a modified threshold value that predicts default for any individuals whose posterior default probability exceeds* 20 *%.*

Linear Discriminant Analysis