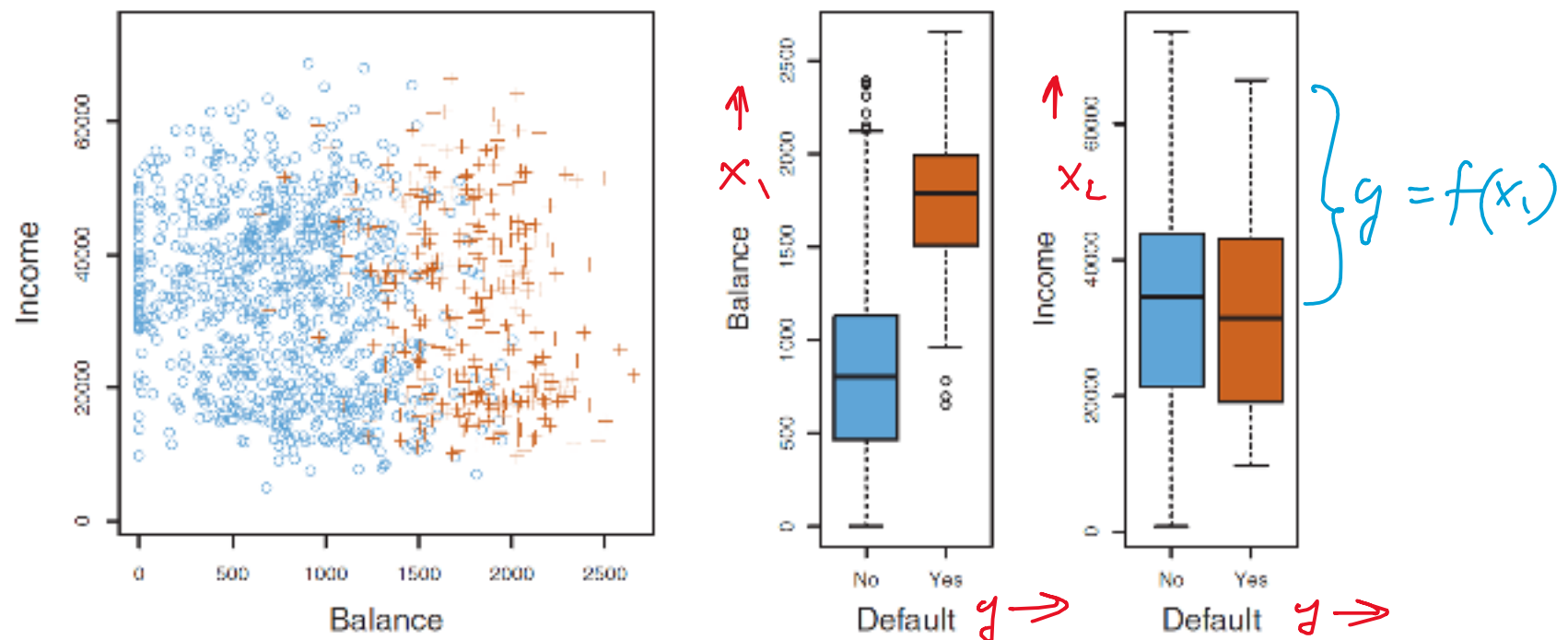


Classification

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.

The Logistic Model

- A linear regression model to represent these probabilities:

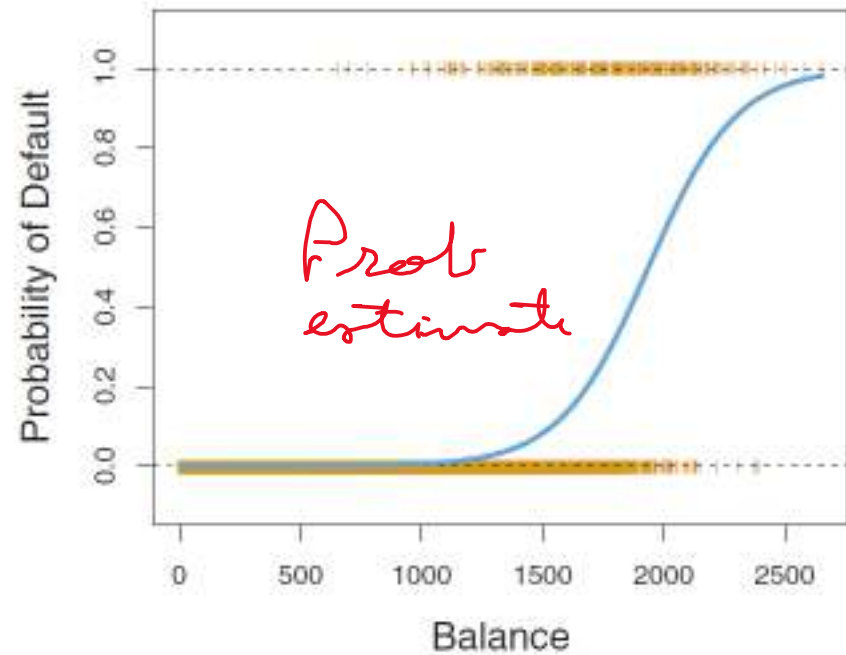
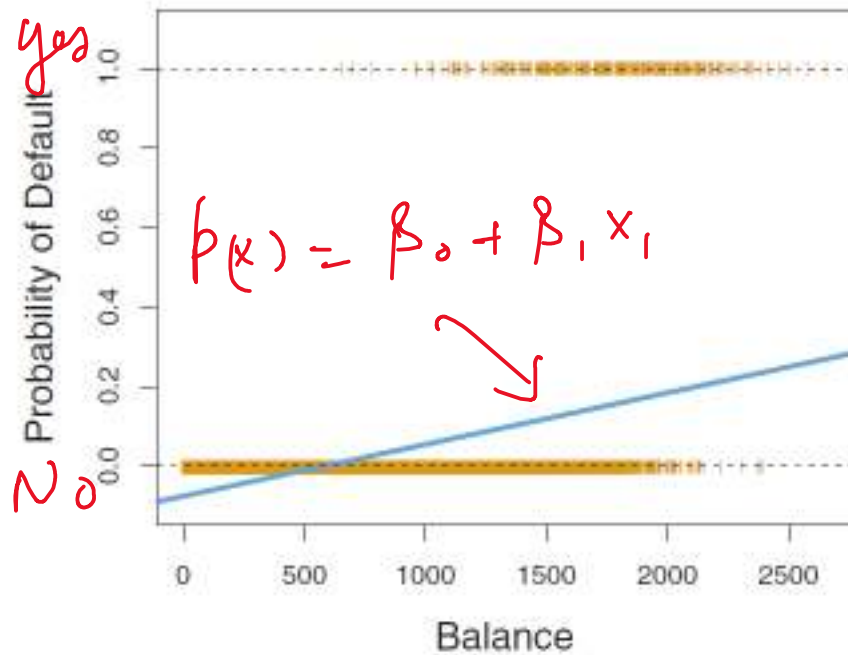
$$p(X) = \beta_0 + \beta_1 X.$$

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

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

$$y = f(x_1)$$

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

- 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

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

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

$p(\beta_0, \beta_1)$

The Logistic Model

Y	X
0	x_1
1	x_2
0	x_3
0	\vdots
1	\vdots
\vdots	x_n

- Estimating the Regression Coefficients:
- Likelihood function

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

$f(\beta_0, \beta_1)$

- 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 $\hat{p}(x_i)$ 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.*

$$y = f(x)$$

$\begin{cases} \rightarrow 0 \\ \rightarrow 1 \end{cases} \quad p(x) \quad p(x)$

The Logistic Model

- Estimating the Regression Coefficients
- Likelihood function

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

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

- The basic intuition behind using maximum likelihood to fit a logistic regression model is as follows: we seek estimates for β_0 and β_1 such that the predicted probability $\hat{p}(x_i)$ 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

$$X = x_1, x_2 \dots x_n$$

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

$$Y = \begin{cases} "1" & \text{if stroke; } \leftarrow p_1(x) \\ "2" & \text{if drug overdose; } \leftarrow p_2(x) \\ "3" & \text{if epileptic seizure. } \leftarrow p_3(x) \end{cases}$$

- 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*

$$\begin{aligned}\log \frac{\Pr(G=1|X=x)}{\Pr(G=K|X=x)} &= \beta_{01} + \beta_{11}^T x \\ \log \frac{\Pr(G=2|X=x)}{\Pr(G=K|X=x)} &= \beta_{02} + \beta_{12}^T x \\ &\vdots \\ \log \frac{\Pr(G=K-1|X=x)}{\Pr(G=K|X=x)} &= \beta_{0(K-1)} + \beta_{1(K-1)}^T x\end{aligned}$$

- 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.

$$\begin{aligned}\Pr(G=i|X=x) &= \frac{\exp(\beta_{0i} + \beta_{1i}^T x)}{1 + \sum_{a=1}^{K-1} \exp(\beta_{0a} + \beta_{1a}^T x)}, \quad i = 1, \dots, K-1, \\ \Pr(G=K|X=x) &= \frac{1}{1 + \sum_{a=1}^{K-1} \exp(\beta_{0a} + \beta_{1a}^T x)}\end{aligned}$$

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 $P(\text{Rain} | \text{Cloud})$

So let's put that in the formula:

$$P(\text{Rain} | \text{Cloud}) = P(\text{Rain}) P(\text{Cloud} | \text{Rain}) / P(\text{Cloud})$$

- $P(\text{Rain})$ is Probability of Rain = 10%
- $P(\text{Cloud} | \text{Rain})$ is Probability of Cloud, given that Rain happens = 50%
- $P(\text{Cloud})$ is Probability of Cloud = 40%
- $P(\text{Rain} | \text{Cloud}) = 0.1 \times 0.50 / 0.4 = .125$

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

Naive Bayes Classifier

- The naive Bayes classifier is a probabilistic classifier based on Bayes's theorem. Let A and B denote two random events, then

$$p(A \mid B) \cdot p(B) = p(B \mid A) \cdot p(A)$$

- If the event A can be decomposed into the disjoint events A_1, \dots, A_c , $p(A_i) > 0$, $i=1, 2, \dots, c$, then

$$p(A_i \mid B) = \frac{p(A_i) \cdot p(B \mid A_i)}{\sum_{j=1}^c p(A_j) \cdot p(B \mid A_j)}$$

Naive Bayes Classifier

- The relevant events considered in classification are “object belongs to class i ”, or more briefly, just “ i ”, and “object has the feature vector x ”, or “ x ”. Replacing A_i , A_j and B by these events yields

$$p(i | x) = \frac{p(i) \cdot p(x | i)}{\sum_{j=1}^c p(j) \cdot p(x | j)}$$

Naive Bayes Classifier

If the p features in x are stochastically independent, then

$$p(x \mid i) = \prod_{k=1}^p p(x^{(k)} \mid i)$$

Inserting this into yields the classification probabilities of the naive Bayes classifier.

$$p(i \mid x) = \frac{p(i) \cdot \prod_{k=1}^p p(x^{(k)} \mid i)}{\sum_{j=1}^c p(j) \cdot \prod_{k=1}^p p(x^{(k)} \mid j)}$$

Naive Bayes Classifier

- Naive Bayes classifier: data of the student example

	number of students who	
	passed	failed
went to class	21	4
did not go to class	1	3
studied material	16	2
did not study material	6	3

$$p(\text{went to class} \mid \text{passed}) = \frac{21}{21 + 1} = \frac{21}{22}$$

$$p(\text{studied material} \mid \text{passed}) = \frac{16}{16 + 6} = \frac{16}{22}$$

$$\Rightarrow p(x \mid \text{passed}) = \frac{21 \cdot 16}{22 \cdot 22} = \frac{84}{121}$$

$$p(\text{went to class} \mid \text{failed}) = \frac{4}{4 + 3} = \frac{4}{7}$$

$$p(\text{studied material} \mid \text{failed}) = \frac{2}{2 + 3} = \frac{2}{5}$$

$$\Rightarrow p(x \mid \text{failed}) = \frac{4 \cdot 2}{7 \cdot 7} = \frac{8}{49}$$

Naive Bayes Classifier

$$p(\text{passed}) = \frac{22}{22 + 7} = \frac{22}{29}$$

$$p(\text{failed}) = \frac{7}{22 + 7} = \frac{7}{29}$$

$$p(\text{passed}) \cdot p(x \mid \text{passed}) = \frac{22}{29} \cdot \frac{84}{121} = \frac{168}{319}$$

$$p(\text{failed}) \cdot p(x \mid \text{failed}) = \frac{7}{29} \cdot \frac{8}{49} = \frac{8}{203}$$

$$\Rightarrow p(\text{passed} \mid x) = \frac{\frac{168}{319}}{\frac{168}{319} + \frac{8}{203}}$$

$$= \frac{168 \cdot 203}{168 \cdot 203 + 8 \cdot 319} = \frac{147}{158} \approx 93\%$$

So, a student who went to class regularly and studied the material will pass with a probability of 93 %, and a deterministic naive Bayes classifier will yield “passed” for this student, and all other students with the same history

Naive Bayes Classifier

Naive Bayes classifier: classifier function for the student example

		probability for		classification
		passed	failed	
went to class	studied material	93 %	7 %	passed
went to class	did not study material	67 %	33 %	passed
did not go to class	studied material	46 %	54 %	failed
did not go to class	did not study material	11 %	89 %	failed

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 k th class; this is the probability that a given observation is associated with the k th category of the response variable Y .
- Let $f_k(X) \equiv \Pr(X = x / Y = k)$ denote the *density function* of X for an observation that comes from the k th 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 .

$K=4$

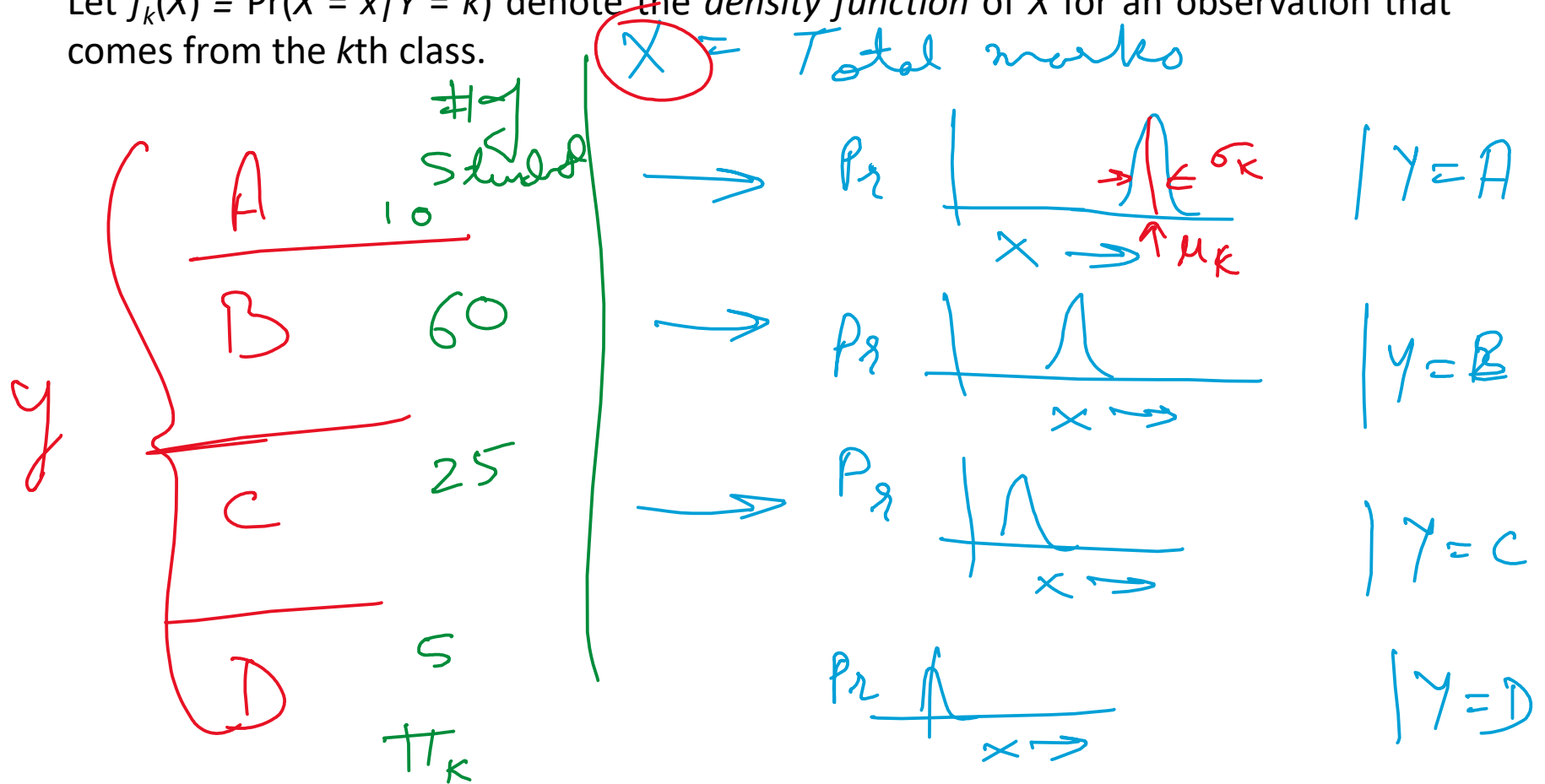
Y	A	10	X = Total marks	$\pi_1 = 0.1$
	B	60		$\pi_2 = 0.6$
	C	25		$\pi_3 = 0.25$
	D	5		$\pi_4 = 0.05$

of Students

freq \rightarrow

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 k th class.



Linear Discriminant Analysis

- Then Bayes' theorem states that $P_r(X=x | Y=k)$

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

Handwritten annotations: A green arrow points from $X=x$ to x . A red checkmark is above the fraction. A red box highlights $\pi_k f_k(x)$. A green circle encloses the entire fraction. A green arrow labeled "Train" points to the circle.

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

Handwritten annotations: A green arrow points to $f_k(x)$. A red checkmark is under σ_k^2 . A red checkmark is under μ_k .

Linear Discriminant Analysis

- Linear Discriminant Analysis for $p = 1$

given x_i

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

$\frac{1}{1} \frac{P_1(x_i)}{P_1(x_i)}$

- let us further assume that $\sigma_1^2 = \dots = \sigma_K^2$
- $\frac{2}{3} \frac{P_2(x_i)}{P_2(x_i)}$

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

Linear Discriminant Analysis

If all variance (σ) are equal, then:

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

Linear Discriminant Analysis

π_k, μ_k

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

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

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

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

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

$$\left. \begin{array}{l} \ln[p_k(x)] \\ \max \\ \ln(\pi_k) + \frac{2\mu_k x}{2\sigma^2} - \frac{\mu_k^2}{2\sigma^2} \end{array} \right\}$$

Linear Discriminant Analysis

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

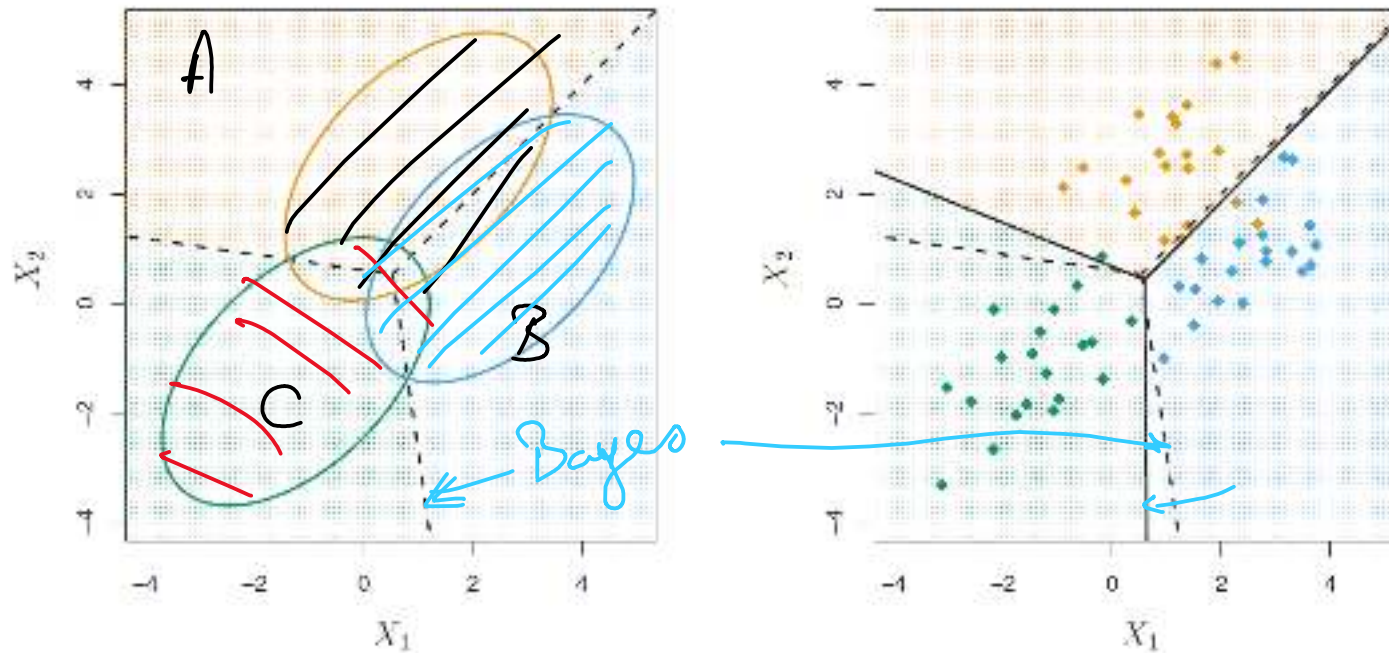
$$\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

is largest. For instance, if $K = 2$ and $\pi_1 = \pi_2$, then the Bayes classifier assigns an observation to class 1 if $2x(\mu_1 - \mu_2) > \mu_1^2 - \mu_2^2$, and to class 2 otherwise. In this case, the Bayes decision boundary corresponds to the point where

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

$$y = f(x_1, x_2)$$

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 $X = (X_1, X_2, \dots, X_p)$ 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 $X \sim N(\mu, \Sigma)$. Here $E(X) = \mu$ is the mean of X (a vector with p components), and $\text{Cov}(X) = \Sigma$ is the $p \times p$ covariance matrix of X . Formally, the multivariate Gaussian density is defined as

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

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

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

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

		<i>True default status</i>		
		No	Yes	Total
<i>Predicted default status</i>	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

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



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

$$P_n(\text{yes} | x) > 0.5 \uparrow$$

$$P_n(\text{yes} | x) > 0.2 \leftarrow$$

		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

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 %.

Classification Confusion Matrix

	Class 1 Predicted	Class 2 Predicted
Class 1 Actual	True Positive (TP)	False Negative (FN)
Class 2 Actual	False Positive (FP)	True Negative (TN)

- Here,
 - Class 1 : Positive
 - Class 2 : Negative
 - **Definition of the Terms:**
 - Positive (P) : Observation is positive (for example: You have Covid).
 - Negative (N) : Observation is not positive (for example: You do not have Covid).
-
- True Positive (TP) : Observation is positive, and is predicted to be positive.
 - False Negative (FN) : Observation is positive, but is predicted negative.
 - True Negative (TN) : Observation is negative, and is predicted to be negative.
 - False Positive (FP) : Observation is negative, but is predicted positive.

Prof. Asim Tewari, IIT Bombay

Classification Rate/Accuracy

	Class 1 Predicted	Class 2 Predicted
Class 1 Actual	True Positive (TP)	False Negative (FN)
Class 2 Actual	False Positive (FP)	True Negative (TN)

- Classification Rate or Accuracy is given by the relation:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

99.9% accuracy can be excellent, good, mediocre, poor or terrible depending upon the problem.

Prof. Asim Tewari, IIT Bombay

	Class 1 Predicted	Class 2 Predicted
Class 1 Actual	True Positive (TP)	False Negative (FN)
Class 2 Actual	False Positive (FP)	True Negative (TN)

$$= \frac{TP}{TP + FN} \quad \text{Recall}$$

- Total number of correctly classified positive examples divide to the total number of positive examples.
- High Recall indicates the class is correctly recognized (small number of FN).

$$\text{Precision} = \frac{TP}{TP + FP}$$

- Total number of correctly classified positive examples by the total number of predicted positive examples.
- High Precision indicates an example labeled as positive is indeed positive (small number of FP).

High recall, low precision: This means that most of the positive examples are correctly recognized (low FN) but there are a lot of false positives.

Low recall, high precision: This shows that we miss a lot of positive examples (high FN) but those we predict as positive are indeed positive (low FP)

The Logistic Model

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

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

- 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 $\hat{p}(x_i)$ of default for each individual, corresponds as closely as possible to the individual's observed default status.

F-measure

	Class 1 Predicted	Class 2 Predicted
Class 1 Actual	True Positive (TP)	False Negative (FN)
Class 2 Actual	False Positive (FP)	True Negative (TN)

- Since we have two measures (Precision and Recall) it helps to have a measurement that represents both of them. We calculate an F-measure which uses Harmonic Mean in place of Arithmetic Mean as it punishes the extreme values more.
The F-Measure will always be nearer to the smaller value of Precision or Recall.

$$F - measure = \frac{2 * Recall * Precision}{Recall + Precision}$$

Positive predictive value (PPV)

$$PPV = \frac{\text{True positive}}{\text{Tested positive}}$$