Classification

Haixia Liu

September 15, 2020

The materials about this course can be downloaded in
`https://github.com/liuhaixias1/data_mining`

What is classification?

- Data:
  - Quantitative response y
  - p different prdictors $x_1, x_2, \cdots, x_p$
- There is some relationship between y and $\mathbf{x} = (x_1, \cdots, x_p)$

$$y = f(\mathbf{x}) + \epsilon. \tag{1}$$

What is classification?

- Data:
  - Quantitative response y
  - p different prdictors $x_1, x_2, \cdots, x_p$
- There is some relationship between y and $\mathbf{x} = (x_1, \cdots, x_p)$

$$y = f(\mathbf{x}) + \epsilon. \tag{1}$$

- A special form of supervised learning.
- Categorical or qualititative responses: Yes/No; High/Median/Low; ...
- Main task: predict the class of the subject based on the inputs.

Examples

- Email spam detector
- Diagnose a person with a set of syndrome as virus carrier or non-carrier.
- Identify which gene, out of a million genes, is disease-causing or not.
- Judge if a trading activity is a fraud or not.

Why not regression?

Examples: The default data

- Simulated data: 10000 individuals.
- Two inputs: income and balance (monthly)
- One output: When you pay, whether you set credit card default or not (Yes or No)?

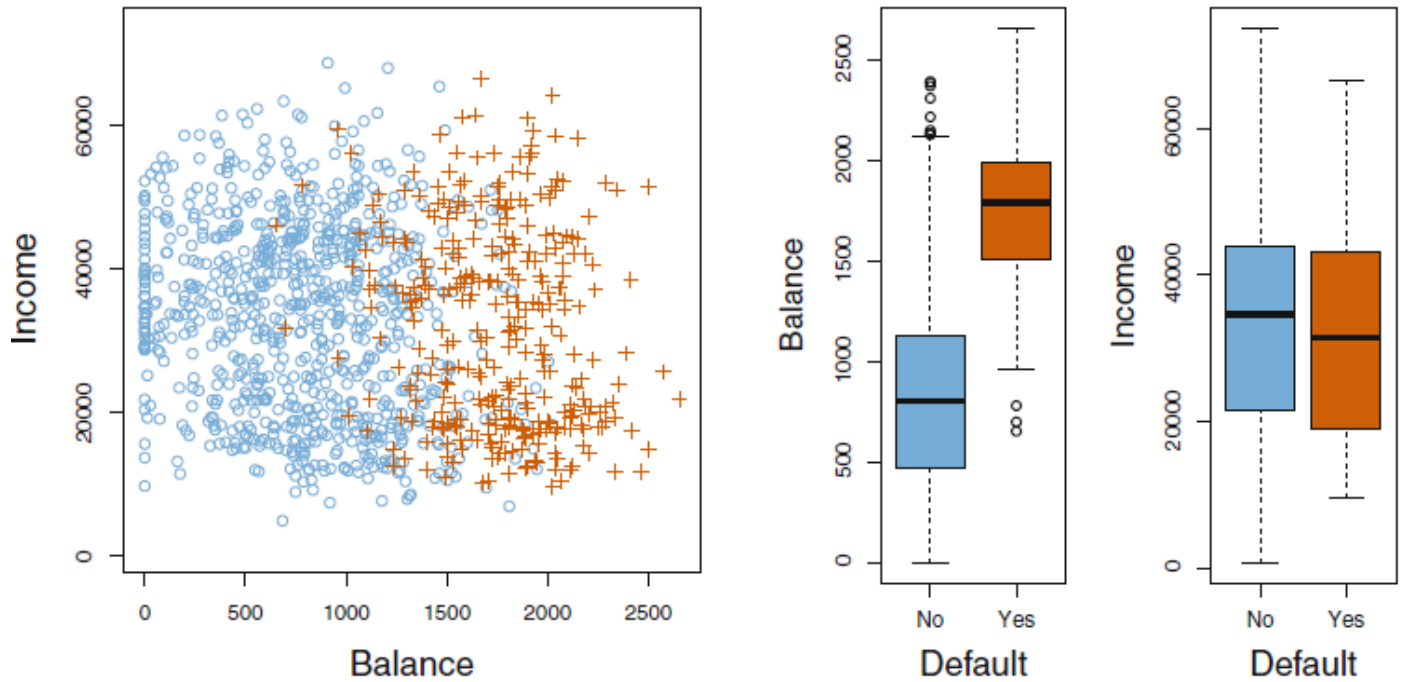


Figure: The Default data set. Left: The annual incomes and monthly credit card balances of a number of individuals. The individuals who defaulted on their credit card payments are shown in orange, and those who did not are shown in blue. Center: Boxplots of balance as a function of default status. Right: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

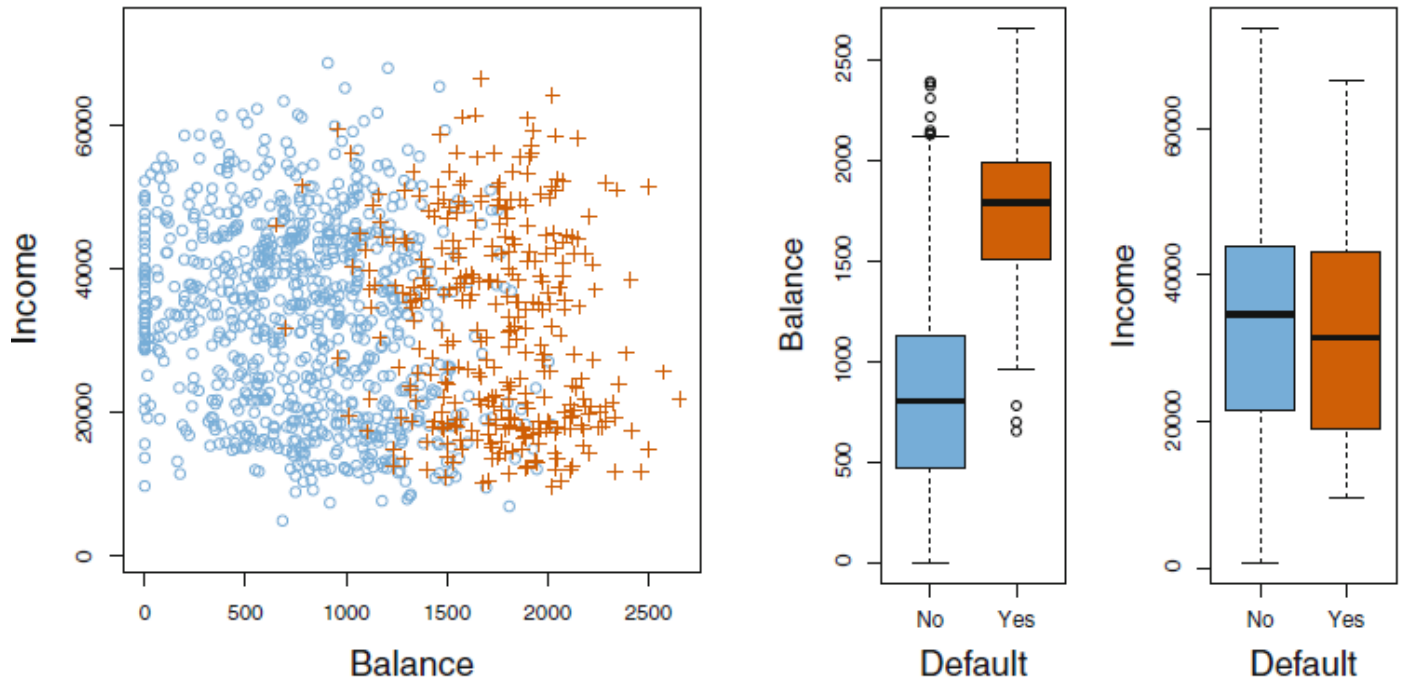


Figure: The Default data set. Left: The annual incomes and monthly credit card balances of a number of individuals. The individuals who defaulted on their credit card payments are shown in orange, and those who did not are shown in blue. Center: Boxplots of balance as a function of default status. Right: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

- Strong relation between balance and default.
- Weaker relation between income and default.

Why not regression?

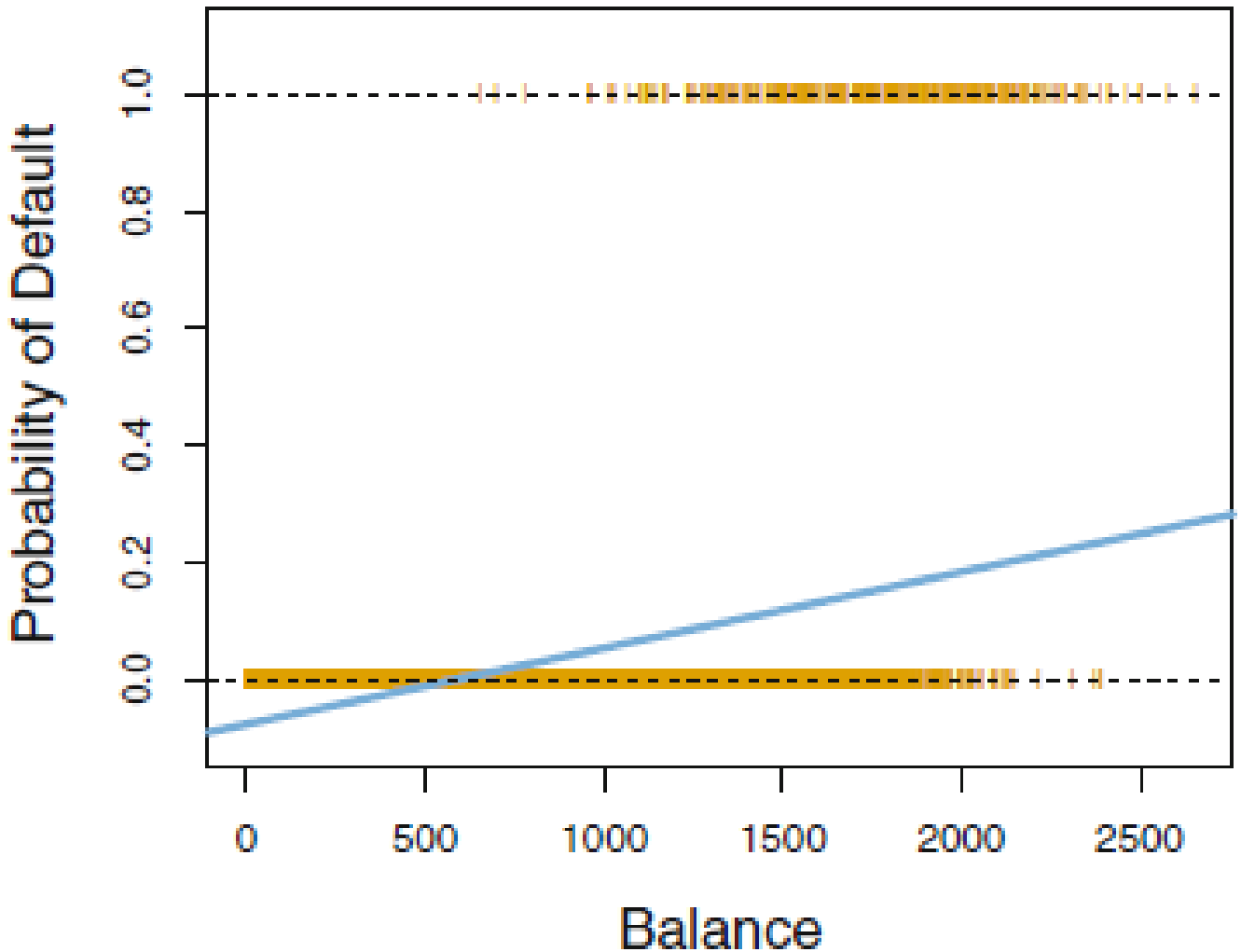


Figure: Linear regression.

Why not regression?

- Coding Qualitative response as numericals, such as 0, 1, 2,..., are generally inappropriate.
- classes could be Asian/European/African, Handwritten numbers (MNIST data); Identify genda (Male/female) from face image.
- In some cases, response classes are ordered, such as severe/moderate/mild; tall/medium/short.
- Coding these into 0, 1, 2 as numericals and apply regression can still be inappropriate, because it implies the differences between the adjacent classes are equal.

Binary response.

- Consider the output is binary: two class,
- Code the response into 0 and 1.
- The training data: $(\mathbf{x}_i, y_i), i = 1, \cdots, n$.
  - $y_i = 1$ for class 1 and $y_i = 0$ for class 0:
  - $x_i = (1; x_{i1}, \cdots, x_{ip})$ are $p + 1$ entries with actually p inputs.
- Key idea: should focus on predicting the probability of the classes.

The logistic regression model

- Assume
$$P(y = 1|\mathbf{x}) = \frac{1}{1 + \exp(-\beta^T \mathbf{x})}.$$
- As a result,
$$P(y = 0|\mathbf{x}) = 1 - P(y = 1|\mathbf{x}) = \frac{1}{1 + \exp(\beta^T \mathbf{x})}.$$
$$\log\left(\frac{P(y = 1|\mathbf{x})}{P(y = 0|\mathbf{x})}\right) = \beta^T \mathbf{x}.$$
- This is called log-odds or logit. And
$$\frac{P(y = 1|\mathbf{x})}{P(y = 0|\mathbf{x})}$$
is called odds.
- Interpretation: one unit increase in variable x_j, increases the log-odds of class 1 by β_j.

The maximum likelihood estimation

- Here, we have independent observations $(\mathbf{x}_i, y_i)$, $i = 1, ..., n$, each follow the (conditional) distribution

$$P(y_i = 1|\mathbf{x}_i) = \frac{1}{1 + \exp(-\beta^T \mathbf{x}_i)} = 1 - P(y_i = 0|\mathbf{x}_i).$$

The maximum likelihood estimation

- Here, we have independent observations $(\mathbf{x}_i, y_i)$, $i = 1, ..., n$, each follow the (conditional) distribution

$$P(y_i = 1|\mathbf{x}_i) = \frac{1}{1 + \exp(-\beta^T \mathbf{x}_i)} = 1 - P(y_i = 0|\mathbf{x}_i).$$

- Recall that, the likelihood is the joint probability function of joint density function of the data.

The maximum likelihood estimation

- Here, we have independent observations $(\mathbf{x}_i, y_i)$, $i = 1, ..., n$, each follow the (conditional) distribution

$$P(y_i = 1|\mathbf{x}_i) = \frac{1}{1 + \exp(-\beta^T \mathbf{x}_i)} = 1 - P(y_i = 0|\mathbf{x}_i).$$

- Recall that, the likelihood is the joint probability function of joint density function of the data.
- So, the joint probability function is

$$\prod_{i=1,...,n;y_i=1} p(y_i = 1|\mathbf{x}_i) \prod_{i'=1,...,n;y_{i'}=0} p(y_{i'} = 0|\mathbf{x}_{i'})$$

which can be conveniently written as

$$\prod_{i=1}^{n} \frac{\exp(y_i \beta^T \mathbf{x}_i)}{1 + \exp(\beta^T \mathbf{x}_i)}.$$

The likelihood and log-likelihood

- The log-likelihood function is

$$\sum_{i=1}^{n}[y_i\beta^T\mathbf{x}_i - \log(1+\exp(\beta^T\mathbf{x}_i))]$$

- The maximizer is denoted as $\hat{\beta}$, which is the MLE of β based on logistic model.

Make predictions

Example: The default data with single input balance.

Make predictions

Example: The default data with single input balance.

- (from ISLR) For the Default data, estimated coefficients of the logistic regression model that predicts the probability of default using balance. $\beta_0 = 0.0055$ and $\beta_1 = -10.6513$.

Make predictions

Example: The default data with single input balance.

- (from ISLR) For the Default data, estimated coefficients of the logistic regression model that predicts the probability of default using balance. $\beta_0 = 0.0055$ and $\beta_1 = -10.6513$.
- For the model with balance as input, predict the default probability for an individual with a balance of 1000$ as

$$\begin{aligned} & \hat{P}(\text{default} = \text{Yes}|\text{balance} = 1000) \\ = & \frac{1}{1 + \exp(-(-10.6513 + 0.0055 \times 1000))} = 0.00576 \end{aligned}$$

Remarks

- The function $1/(1+\exp(-x))$ is called logistic *sigmoid* function.

Remarks

- The function $1/(1+\exp(-x))$ is called logistic *sigmoid* function.
- This is the cdf of the logistic distribution.

Remarks

- The function $1/(1+\exp(-x))$ is called logistic *sigmoid* function.
- This is the cdf of the logistic distribution.
- One can use other probabilistic functions to replace the sigmoid function.

Remarks

- The function $1/(1+\exp(-x))$ is called logistic *sigmoid* function.
- This is the cdf of the logistic distribution.
- One can use other probabilistic functions to replace the sigmoid function.
- The probit model uses $\Phi(x)$, where Φ is the cdf of standard normal distribution.

Remarks

- The function $1/(1+\exp(-x))$ is called logistic *sigmoid* function.
- This is the cdf of the logistic distribution.
- One can use other probabilistic functions to replace the sigmoid function.
- The probit model uses $\Phi(x)$, where Φ is the cdf of standard normal distribution.
- For multiple classes (more than 2 classes), the logistic regression model can be adapted by using the *softmax* function.

$$P(y=k|\mathbf{x}) = \frac{\exp(\theta_k^T \mathbf{x})}{\sum_{j=1}^{K} \exp(\theta_j^T \mathbf{x})},$$

where K is the number of classes.

The Bayes Theorem

- General Bayes theorem :

$$P(A_k|B) = \frac{P(B|A_k)P(A_k)}{\sum_{j=1}^{K} P(B|A_j)P(A_j)}$$

for disajoint sets $A_1, ..., A_K$ whose union has probability 1.

The Bayes Theorem

- General Bayes theorem :

$$P(A_k|B) = \frac{P(B|A_k)P(A_k)}{\sum_{j=1}^{K} P(B|A_j)P(A_j)}$$

  for disajoint sets $A_1, ..., A_K$ whose union has probability 1.
- the output $Y = 1, \cdots, K$ and X is the input of p-dimension.
- Both Y and X are random variables.
- Then, Bayes theorem implies

$$\begin{aligned} p_k(x) =& P(Y = k|X = x) = \frac{P(Y = k)f(x|Y = k)}{\sum_{j=1}^{K} P(Y = j)f(x|Y = j)} \\ =& \frac{\pi_k f_k(x)}{\sum_{j=1}^{K} \pi_j f_j(x)}, \qquad \pi_k = P(Y = k) \text{ and } f_k(x) = f(x|Y = k). \end{aligned}$$

The Bayes Theorem

- General Bayes theorem :

$$P(A_k|B) = \frac{P(B|A_k)P(A_k)}{\sum_{j=1}^{K} P(B|A_j)P(A_j)}$$

  for disajoint sets $A_1, ..., A_K$ whose union has probability 1.
- the output $Y = 1, \cdots, K$ and X is the input of p-dimension.
- Both Y and X are random variables.
- Then, Bayes theorem implies

$$\begin{aligned} p_k(x) =& P(Y = k|X = x) = \frac{P(Y = k)f(x|Y = k)}{\sum_{j=1}^{K} P(Y = j)f(x|Y = j)} \\ =& \frac{\pi_k f_k(x)}{\sum_{j=1}^{K} \pi_j f_j(x)}, \qquad \pi_k = P(Y = k) \text{ and } f_k(x) = f(x|Y = k). \end{aligned}$$

- We classify a subject with input x into class k, if its $p_k(x)$ is the largest, for $k = 1, ..., K$.

Bayes classifier

- *Bayes classifier* is a simple classifier that assigns each observation to the most likely class, given its predictor values.
- assign a test observation with predictor vector $\mathbf{x}_0$ to the class j for which

$$P(Y = j|X = \mathbf{x}_0)$$

is largest.

Model assumptions of LDA

X is p-dimensional. $Y = 1, ..., K$, totally K classes.

Model assumptions of LDA

X is p-dimensional. $Y = 1, ..., K$, totally K classes.

- Assume, for $k = 1, ..., K$,

$$X|Y = k \sim \mathcal{N}(\mu_k, \Sigma),$$

where $\mu = (\mu_1, \cdots, \mu_K)$ is p-vector and Σ is p-by-p variance matrix.Note that we assumed the same Σ for all classes $k = 1, ..., K$.

Normal distribution

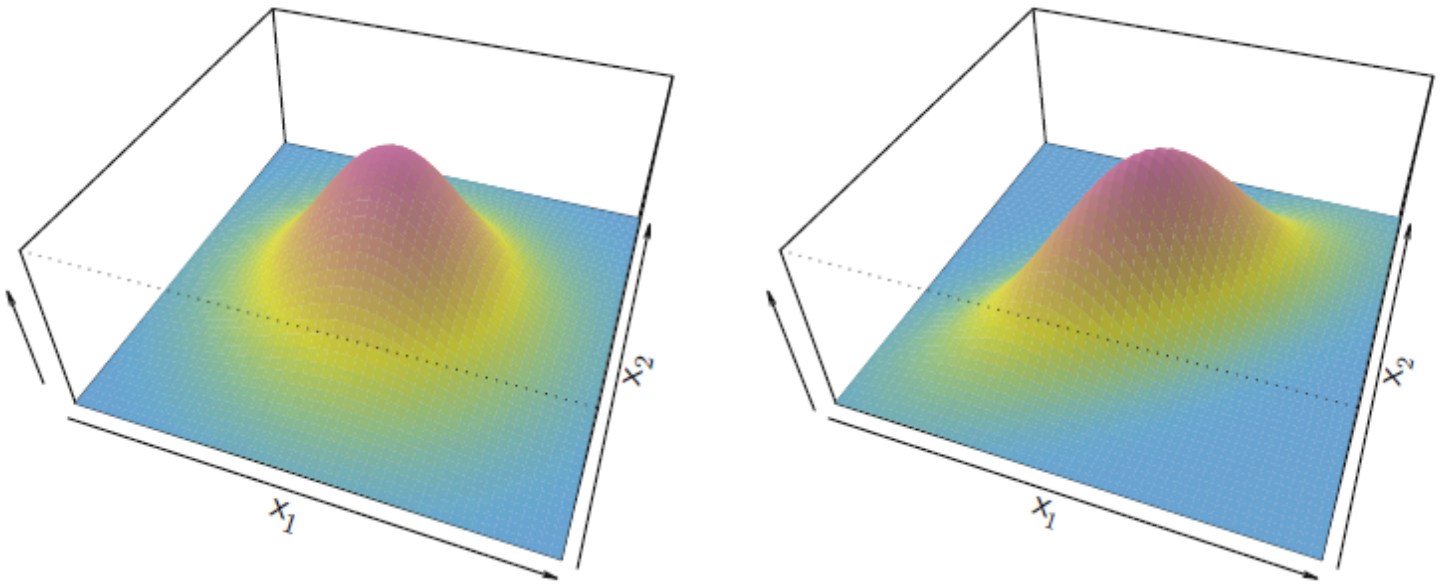


Figure: Left: The two predictors are uncorrelated. Right: The two variables have a correlation of 0.7.

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

Normal distribution

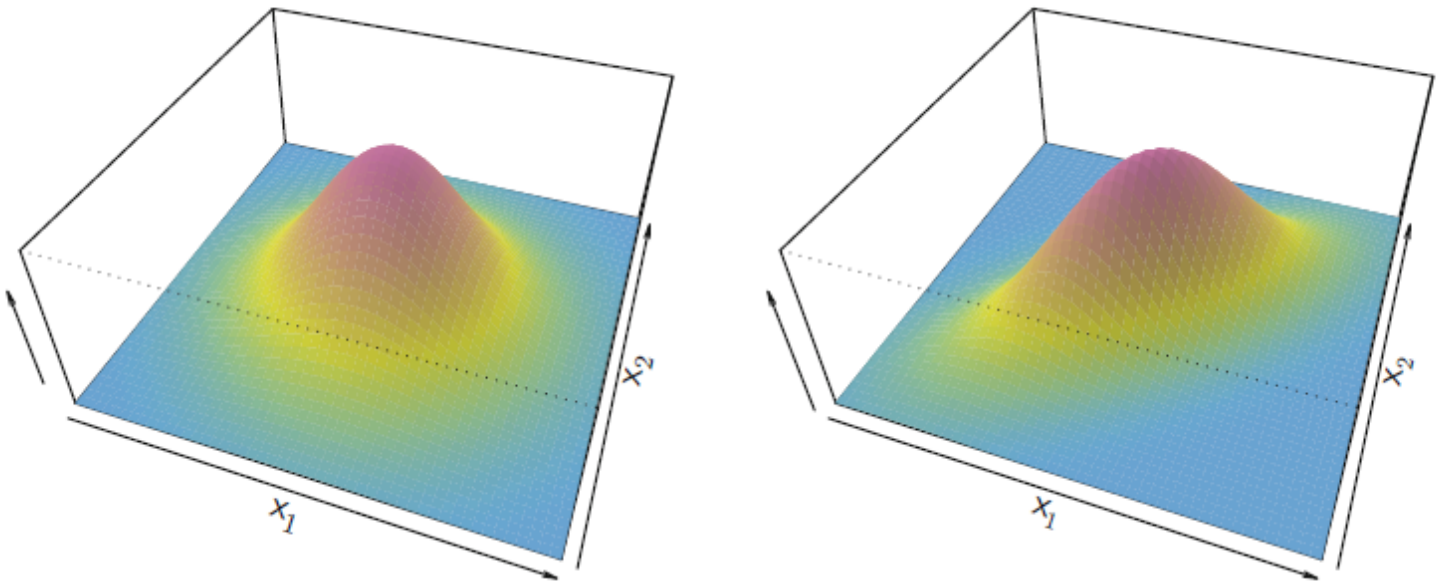


Figure: Left: The two predictors are uncorrelated. Right: The two variables have a correlation of 0.7.

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

$$p_k(x) = P(Y=k|X=\mathbf{x}) = \frac{\pi_k \exp[(-1/2)(\mathbf{x}-\mu_k)^T\Sigma^{-1}(\mathbf{x}-\mu_k)]}{\sum_{l=1}^{K}\pi_l \exp[(-1/2)(\mathbf{x}-\mu_l)^T\Sigma^{-1}(\mathbf{x}-\mu_l)]}.$$

Computing $p_k(\mathbf{x})$ for LDA

$$p_k(\mathbf{x}) = \frac{\pi_k \exp[(-1/2)(\mathbf{x}-\mu_k)^T \Sigma^{-1}(\mathbf{x}-\mu_k)]}{\sum_{l=1}^{K} \pi_l \exp[(-1/2)(\mathbf{x}-\mu_l)^T \Sigma^{-1}(\mathbf{x}-\mu_l)]}.$$

- Comparing $p_k(\mathbf{x})$ is the same as comparing the *numerator*.

Computing $p_k(\mathbf{x})$ for LDA

$$p_k(\mathbf{x}) = \frac{\pi_k \exp[(-1/2)(\mathbf{x}-\mu_k)^T\Sigma^{-1}(\mathbf{x}-\mu_k)]}{\sum_{l=1}^{K}\pi_l \exp[(-1/2)(\mathbf{x}-\mu_l)^T\Sigma^{-1}(\mathbf{x}-\mu_l)]}.$$

- Comparing $p_k(\mathbf{x})$ is the same as comparing the *numerator*.

$$\log(\pi_k \exp[(-1/2)(\mathbf{x}-\mu_k)^T\Sigma^{-1}(\mathbf{x}-\mu_k)])$$
$$= \mu_k^{-1}\Sigma^{-1}\mathbf{x} - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \log\pi_k - \frac{1}{2}\mathbf{x}^T\Sigma^{-1}\mathbf{x}.$$

Computing $p_k(\mathbf{x})$ for LDA

$$p_k(\mathbf{x}) = \frac{\pi_k \exp[(-1/2)(\mathbf{x}-\mu_k)^T\Sigma^{-1}(\mathbf{x}-\mu_k)]}{\sum_{l=1}^{K}\pi_l \exp[(-1/2)(\mathbf{x}-\mu_l)^T\Sigma^{-1}(\mathbf{x}-\mu_l)]}.$$

- Comparing $p_k(\mathbf{x})$ is the same as comparing the *numerator*.

$$\begin{aligned}&\log(\pi_k \exp[(-1/2)(\mathbf{x}-\mu_k)^T\Sigma^{-1}(\mathbf{x}-\mu_k)]) \\ &\qquad = \mu_k^{-1}\Sigma^{-1}\mathbf{x} - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \log\pi_k - \frac{1}{2}\mathbf{x}^T\Sigma^{-1}\mathbf{x}.\end{aligned}$$

- Note that the term $\mathbf{x}^T\Sigma^{-1}\mathbf{x}$ is common for all numerators of $p_k(\mathbf{x})$.

Computing $p_k(\mathbf{x})$ for LDA

$$p_k(\mathbf{x}) = \frac{\pi_k \exp[(-1/2)(\mathbf{x}-\mu_k)^T\Sigma^{-1}(\mathbf{x}-\mu_k)]}{\sum_{l=1}^{K}\pi_l \exp[(-1/2)(\mathbf{x}-\mu_l)^T\Sigma^{-1}(\mathbf{x}-\mu_l)]}.$$

- Comparing $p_k(\mathbf{x})$ is the same as comparing the *numerator*.

$$\log(\pi_k \exp[(-1/2)(\mathbf{x}-\mu_k)^T\Sigma^{-1}(\mathbf{x}-\mu_k)])$$
$$= \mu_k^{-1}\Sigma^{-1}\mathbf{x} - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \log\pi_k - \frac{1}{2}\mathbf{x}^T\Sigma^{-1}\mathbf{x}.$$

- Note that the term $\mathbf{x}^T\Sigma^{-1}\mathbf{x}$ is common for all numerators of $p_k(\mathbf{x})$.
- Set

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

- A subject with input $\mathbf{x}$ is classified into class k, if δ_k is the largest.

Solution: use the sample analogue

The classifer based on δ_k is the Bayesian classifier,

- assuming known μ_k, Σ and π_k, $k = 1, \cdots, K$.

Solution: use the sample analogue

The classifer based on δ_k is the Bayesian classifier,

- assuming known μ_k, Σ and π_k, $k = 1, \cdots, K$.
- A practical problem: μ_j, Σ and π_j may be unknown.

Solution: use the sample analogue

The classifer based on δ_k is the Bayesian classifier,

- assuming known μ_k, Σ and π_k, $k = 1, \cdots, K$.
- A practical problem: μ_j, Σ and π_j may be unknown.
- use *sample mean, pooled sample variance, and sample proportion of class k in the data* instead.

Solution: use the sample analogue

The classifer based on δ_k is the Bayesian classifier,

- assuming known μ_k, Σ and π_k, $k = 1, \cdots, K$.
- A practical problem: μ_j, Σ and π_j may be unknown.
- use *sample mean, pooled sample variance, and sample proportion of class k in the data* instead.
- Let data $(\mathbf{x}_i, y_i), i = 1, ..., n$ and n_k be the number of subjects in class $k \in \{1, \cdots, K\}$ in the data,

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{\{i:y_i=k\}} \mathbf{x}_i; \hat{\pi}_k = \frac{n_k}{n}; \hat{\Sigma} = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{\{i:y_i=k\}} (\mathbf{x}_i - \hat{\mu}_k)(\mathbf{x}_i - \hat{\mu}_k)^T.$$

- The sample analogue of δ_k is

$$\hat{\delta}_k(\mathbf{x}) = \hat{\mu}_k^T \hat{\Sigma}^{-1} \mathbf{x} - \frac{1}{2} \hat{\mu}_k^T \hat{\Sigma}^{-1} \mu_k + \log \hat{\pi}_k.$$

In summary,

- we classify a subject with input $\mathbf{x}$ into class k, if $\hat{\delta}_k$ is the largest.
- $\hat{\delta}_k(\mathbf{x})$ is called discrimination function.
- $\hat{\delta}_k(\mathbf{x})$ is here a linear function of $\mathbf{x}$. This is why we call it LDA.
- The region of values of x being classified into a class has linear boundary, since $\delta_k(\mathbf{x}) = \delta_l(\mathbf{x})$ defines a linear hyperplane for $\mathbf{x}$.

Knowing Normal distribution

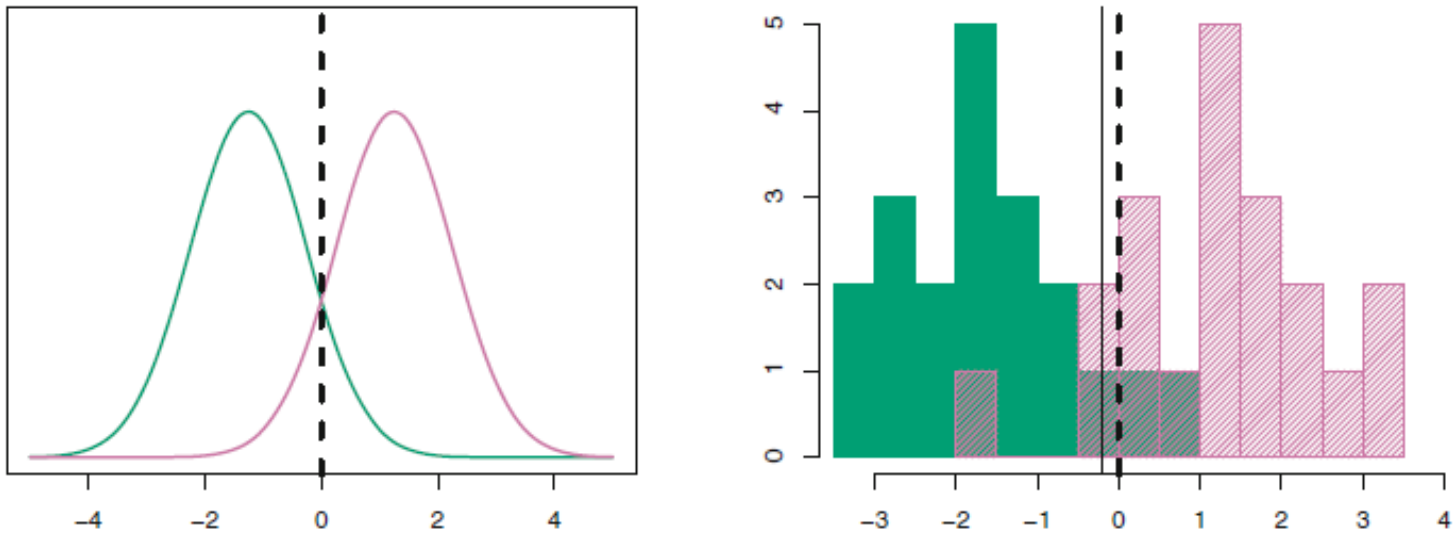


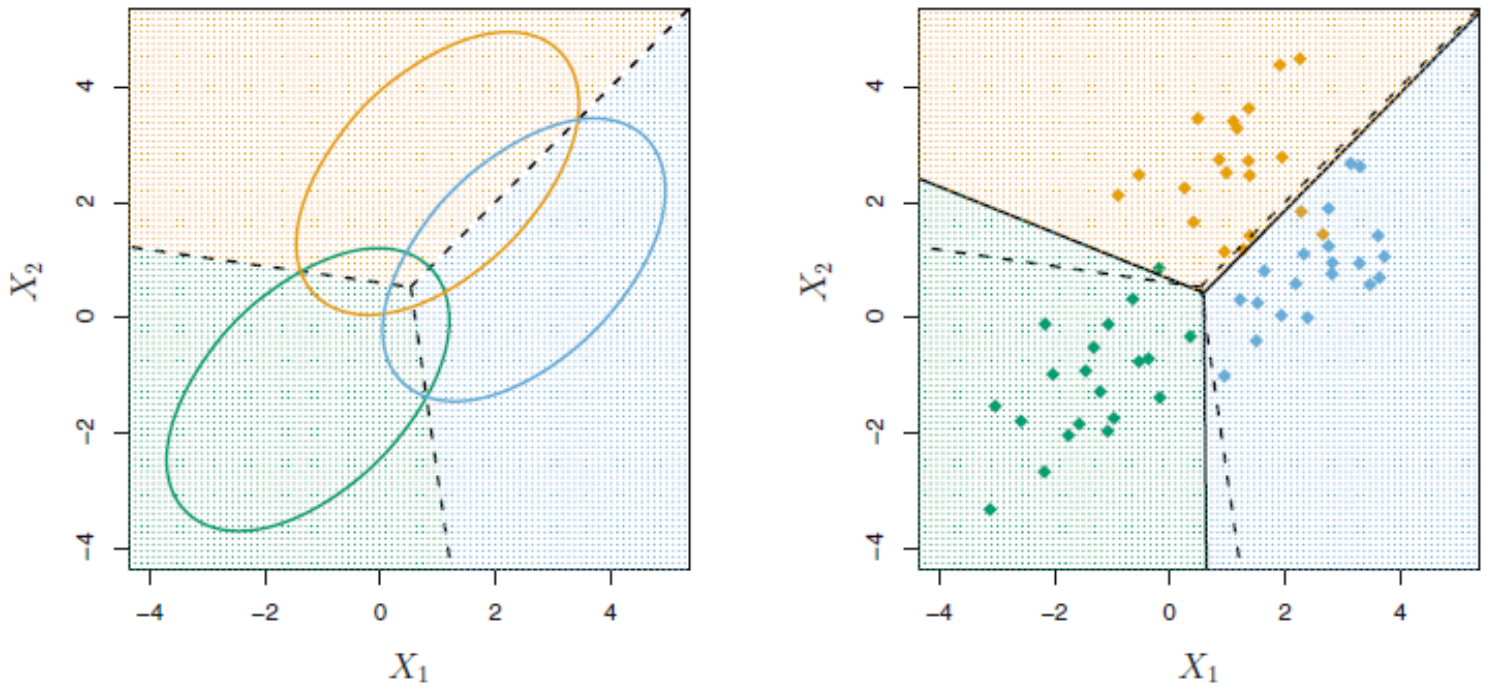
FIGURE 4.4. Left: Two one-dimensional normal density functions are shown. The dashed vertical line represents the Bayes decision boundary. Right: 20 observations were drawn from each of the two classes, and are shown as histograms. The Bayes decision boundary is again shown as a dashed vertical line. The solid vertical line represents the LDA decision boundary estimated from the training data.

The graph of LDA

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.
- The dashed lines are the Bayes decision boundaries.

The graph of LDA



- **Left**: Ellipses that contain 95% of the probability for each of the three classes are shown.
- **Right**: 20 observations were generated from each class, and the corresponding LDA decision boundaries are indicated using solid black lines.

Two Types of Errors

In practice, a binary classifier such as this one can make two types of errors:

- it can incorrectly assign an individual who defaults to the no default category,
- it can incorrectly assign an individual who does not default to the default category.

It is often of interest to determine which of these two types of errors are being made.
The detailed result is shown in the *confusion matrix.*

The Confusion matrix.

- Consider the true default statuses for the 10, 000 (9644 No and 252 YES) training observations in the Default data set.
- In this case with two classes (Yes/No), i.e., $K = 2$. Bayes classifier classifies into *default* class if

$$Pr(default = Yes|X = x) > 0.5.$$

		True default status		
		No	Yes	Total
Predicted default status	No	9644	252	9896
	Yes	23	81	104
	Total	9667	333	10000

Class-specific performance

		True default status		
		No	Yes	Total
Predicted default status	No	9644	252	9896
	Yes	23	81	104
	Total	9667	333	10000

- Overall error rate: $(252 + 23)/10000 = 2.75\%$.

Class-specific performance

		True default status		
		No	Yes	Total
Predicted default status	No	9644	252	9896
	Yes	23	81	104
	Total	9667	333	10000

- Overall error rate: $(252 + 23)/10000 = 2.75\%$.
- Error rate with default people: $252/333 = 75.7\%$

Class-specific performance

		True default status		
		No	Yes	Total
Predicted default status	No	9644	252	9896
	Yes	23	81	104
	Total	9667	333	10000

- Overall error rate: $(252 + 23)/10000 = 2.75\%$.
- Error rate with default people: $252/333 = 75.7\%$
- Sensitivity: $1 - 75.7\% = 24.3\%$

Class-specific performance

		True default status		
		No	Yes	Total
Predicted default status	No	9644	252	9896
	Yes	23	81	104
	Total	9667	333	10000

- Overall error rate: $(252 + 23)/10000 = 2.75\%$.
- Error rate with default people: $252/333 = 75.7\%$
- Sensitivity: $1 - 75.7\% = 24.3\%$
- Error rate within people no-default: $23/9667 = 0.24\%$.

Class-specific performance

		True default status		
		No	Yes	Total
Predicted default status	No	9644	252	9896
	Yes	23	81	104
	Total	9667	333	10000

- Overall error rate: $(252 + 23)/10000 = 2.75\%$.
- Error rate with default people: $252/333 = 75.7\%$
- Sensitivity: $1 - 75.7\% = 24.3\%$
- Error rate within people no-default: $23/9667 = 0.24\%$.
- Specificity: $1 - 0.24\% = 99.8\%$

A modification

- The Bayes classifier classifies a subject into class k, if the posterior probability $p_k(x)$ is the largest.
- A modification is classifies into *default* class if

$$Pr(default = Yes|X = x) > 0.2.$$

		True default status		
		No	Yes	Total
Predicted default status	No	9432	138	9570
	Yes	235	195	430
	Total	9667	333	10000

Class-specific performance

- Overall error rate: $(235 + 138)/10000 = 3.73\%$. (increased)

Class-specific performance

- Overall error rate: $(235 + 138)/10000 = 3.73\%$. (increased)
- Error rate with default people: $138/333 = 41.4\%$ (lower after modification)

Class-specific performance

- Overall error rate: $(235 + 138)/10000 = 3.73\%$. (increased)
- Error rate with default people: $138/333 = 41.4\%$ (lower after modification)
- Sensitivity: $1 - 41.4\% = 58.6\%$

Class-specific performance

- Overall error rate: $(235 + 138)/10000 = 3.73\%$. (increased)
- Error rate with default people: $138/333 = 41.4\%$ (lower after modification)
- Sensitivity: $1 - 41.4\% = 58.6\%$
- Error rate within people no-default: $235/9667 = 2.43\%$. (increased)

Class-specific performance

- Overall error rate: $(235 + 138)/10000 = 3.73\%$. (increased)
- Error rate with default people: $138/333 = 41.4\%$ (lower after modification)
- Sensitivity: $1 - 41.4\% = 58.6\%$
- Error rate within people no-default: $235/9667 = 2.43\%$. (increased)
- Specificity: $1 - 2.43\% = 97.57\%$

Class-specific performance

- Overall error rate: $(235 + 138)/10000 = 3.73\%$. (increased)
- Error rate with default people: $138/333 = 41.4\%$ (lower after modification)
- Sensitivity: $1 - 41.4\% = 58.6\%$
- Error rate within people no-default: $235/9667 = 2.43\%$. (increased)
- Specificity: $1 - 2.43\% = 97.57\%$
- Idenfication of defaulter (sensitivity) is more important to credit card company!

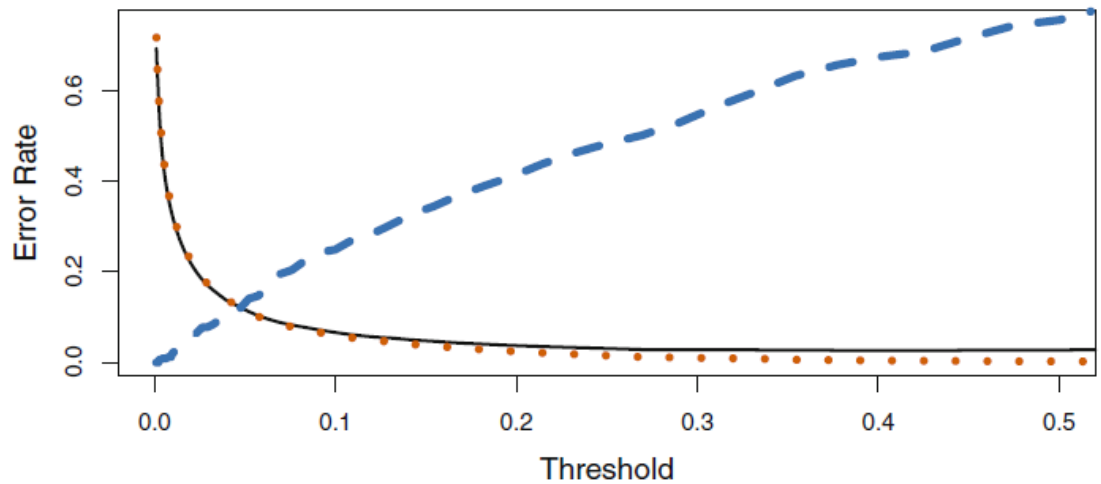
Class-specific performance

- Overall error rate: $(235 + 138)/10000 = 3.73\%$. (increased)
- Error rate with default people: $138/333 = 41.4\%$ (lower after modification)
- Sensitivity: $1 - 41.4\% = 58.6\%$
- Error rate within people no-default: $235/9667 = 2.43\%$. (increased)
- Specificity: $1 - 2.43\% = 97.57\%$
- Idenfication of defaulter (sensitivity) is more important to credit card company!
- This modification may be helpful to the campany. A *tradeoff* of specificity for sensitivity.

The tradeoff

For the Default data set, error rates are shown as a function of the threshold value. $\mathbf{x}$ is classifed to *default* class if

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



- The black solid line displays the overall error rate.
- The blue dashed line – the fraction of defaulting customers that are incorrectly classified.
- The orange dotted line – the fraction of errors among the non-defaulting customers.

Clarifying the terminology

		Predicted class		
		− or Null	+ or Non-null	Total
True class	− or Null	True Neg. (TN)	False Pos. (FP)	N
	+ or Non-null	False Neg. (FN)	True Pos. (TP)	P
	Total	N*	P*	

Name	Definition	Synonyms
False Pos. rate	FP/N	Type I error, 1 - specificity
True Pos. rate	TP/P	1- Type II error, power, sensitivity, recall
Pos. Pred.value	TP/P*	Precision 1- false discovery rate
Neg.Pred.value	TN/N*	

ROC curve (receiver operating characteristics)

How can we decide which threshold value is best?

ROC curve (receiver operating characteristics)

How can we decide which threshold value is best?
The ROC curve is a popular graphic for simultaneously displaying the two types of errors for all possible thresholds.

ROC curve (receiver operating characteristics)

How can we decide which threshold value is best?
The ROC curve is a popular graphic for simultaneously displaying the two types of errors for all possible thresholds.

- The *true positive rate* is the sensitivity: the fraction of defaulters that are correctly identified, using a given threshold value.
- The *false positive rate* is 1 − specificity: the fraction of non-defaulters that we classify incorrectly as defaulters, using that same threshold value.

ROC curve (receiver operating characteristics)

How can we decide which threshold value is best?
The ROC curve is a popular graphic for simultaneously displaying the two types of errors for all possible thresholds.

- The *true positive rate* is the sensitivity: the fraction of defaulters that are correctly identified, using a given threshold value.
- The *false positive rate* is 1 – specificity: the fraction of non-defaulters that we classify incorrectly as defaulters, using that same threshold value.
- ROC curve is a plot of true positive rate versus false positive rate for all possible thresholding.

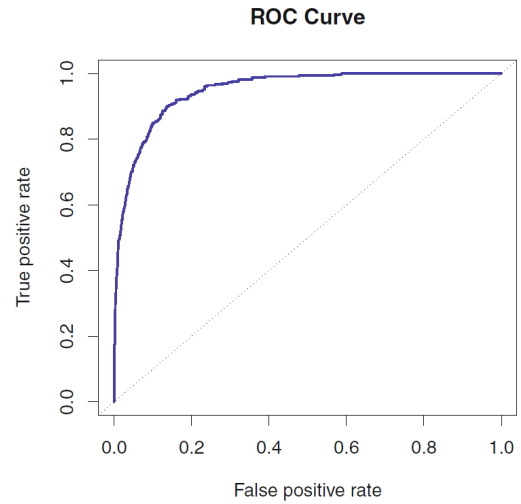


Figure: A ROC curve for the LDA classifier on the Default data.

- The dotted line represents the "no information" classifier.

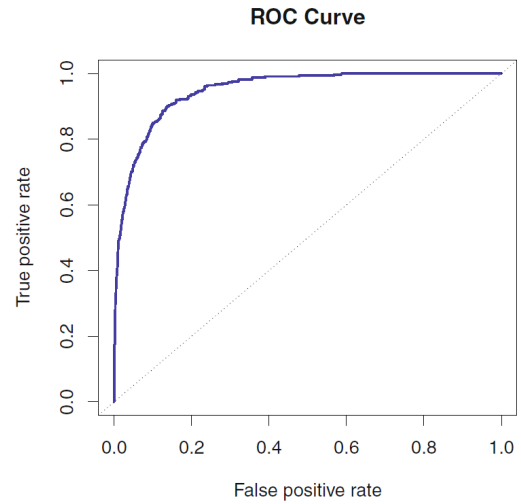


Figure: A ROC curve for the LDA classifier on the Default data.

- The dotted line represents the "no information" classifier.
- The actual thresholds are not shown in curve.

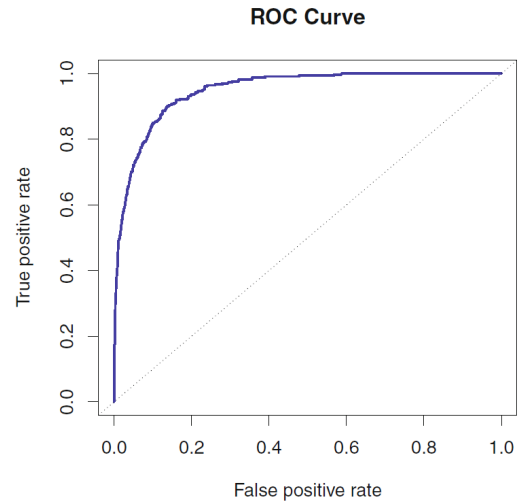


Figure: A ROC curve for the LDA classifier on the Default data.

- The dotted line represents the "no information" classifier.
- The actual thresholds are not shown in curve.
- The overall performance of a classifier is given by *area under the (ROC) curve* (AUC).

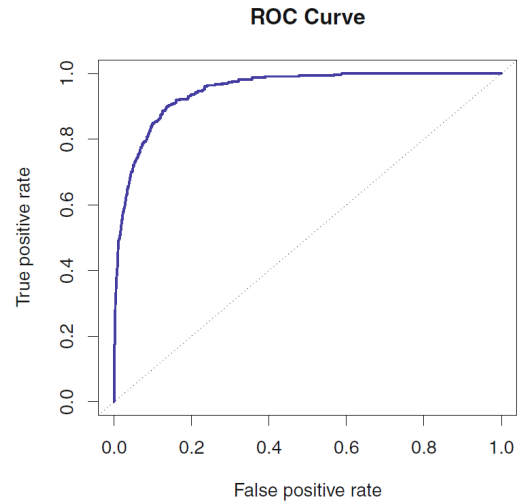


Figure: A ROC curve for the LDA classifier on the Default data.

- The dotted line represents the "no information" classifier.
- The actual thresholds are not shown in curve.
- The overall performance of a classifier is given by *area under the (ROC) curve* (AUC).
- The ideal ROC curve hugs the top left corner, indicating a high true positive rate and a low false positive rate.

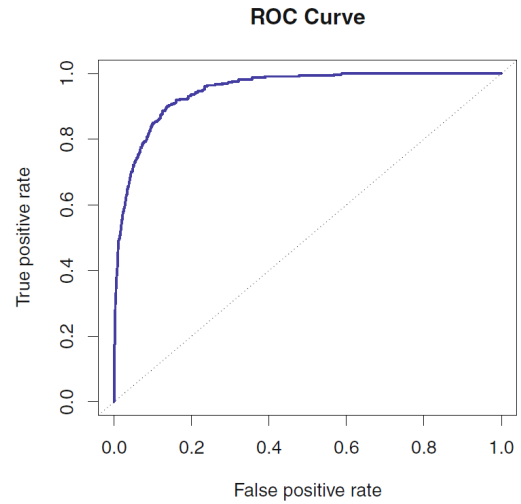


Figure: A ROC curve for the LDA classifier on the Default data.

- The dotted line represents the "no information" classifier.
- The actual thresholds are not shown in curve.
- The overall performance of a classifier is given by *area under the (ROC) curve* (AUC).
- The ideal ROC curve hugs the top left corner, indicating a high true positive rate and a low false positive rate.
- The larger area under the (ROC) curve, the better the classifier.

The model assumption of QDA

- Recall that LDA assume the class-specific normal distribution with class means μ_k and *same* class variance Σ.
- The QDA assume the class-specific normal distribution with class means μ_k and class variance Σ_k.
- The rest of the derivations follow an analogous line.

The discrimination function

$$f_k(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}} \exp\Big(-\frac{1}{2}(\mathbf{x}-\mu_k)^T\Sigma_k^{-1}(\mathbf{x}-\mu_k)\Big).$$

The discrimination function

$$f_k(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}} \exp\Big(-\frac{1}{2}(\mathbf{x}-\mu_k)^T\Sigma_k^{-1}(\mathbf{x}-\mu_k)\Big).$$

$$p_k(\mathbf{x}) = \frac{\frac{\pi_k}{(2\pi)^{p/2}|\Sigma_k|^{1/2}} \exp[(-1/2)(\mathbf{x}-\mu_k)^T\Sigma_k^{-1}(\mathbf{x}-\mu_k)]}{\sum_{l=1}^{K} \frac{\pi_l}{(2\pi)^{p/2}|\Sigma_l|^{1/2}} \exp[(-1/2)(\mathbf{x}-\mu_l)^T\Sigma_k^{-1}(\mathbf{x}-\mu_l)]}.$$

The discrimination function

$$f_k(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}} \exp\Big(-\frac{1}{2}(\mathbf{x}-\mu_k)^T\Sigma_k^{-1}(\mathbf{x}-\mu_k)\Big).$$

$$p_k(\mathbf{x}) = \frac{\frac{\pi_k}{(2\pi)^{p/2}|\Sigma_k|^{1/2}} \exp[(-1/2)(\mathbf{x}-\mu_k)^T\Sigma_k^{-1}(\mathbf{x}-\mu_k)]}{\sum_{l=1}^K \frac{\pi_l}{(2\pi)^{p/2}|\Sigma_l|^{1/2}} \exp[(-1/2)(\mathbf{x}-\mu_l)^T\Sigma_k^{-1}(\mathbf{x}-\mu_l)]}.$$

- The discrimination function:

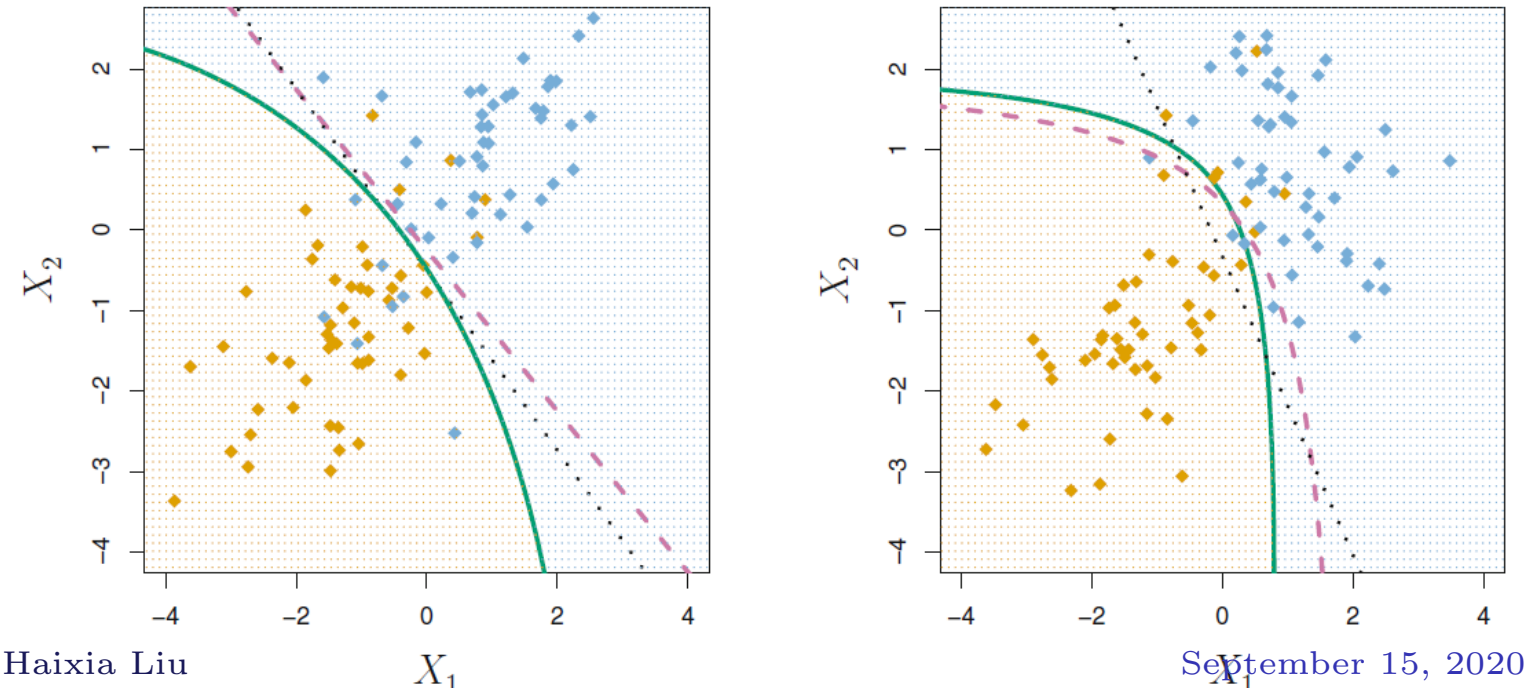
$$\begin{aligned}\delta_k(\mathbf{x}) &= -\frac{1}{2}(\mathbf{x}-\mu_k)^T\Sigma_k^{-1}(\mathbf{x}-\mu_k) - \frac{1}{2}\log(|\Sigma_k|) + \log\pi_k \\ &= -\frac{1}{2}\mathbf{x}^T\Sigma_k^{-1}\mathbf{x} + \mathbf{x}^T\Sigma_k^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma_k^{-1}\mu_k - \frac{1}{2}\log(|\Sigma_k|) + \log\pi_k\end{aligned}$$

This is a *quadratic funciton* of $\mathbf{x}$.

- In actual implementation, need to use class specific sample mean and class specific sample variance to estimate π_k, μ_k and Σ_k.
- More complex than LDA, more parameters to estimate.
- If the class variances are equal or close, LDA is better. Otherwise, QDA is better.

Comparing LDA and QDA

FIGURE 4.9. Left: The Bayes (purple dashed), LDA (black dotted), and QDA (green solid) decision boundaries for a two-class problem with $\Sigma_1 = \Sigma_2$. The shading indicates the QDA decision rule. Since the Bayes decision boundary is linear, it is more accurately approximated by LDA than by QDA. Right: Details are as given in the left-hand panel, except that $\Sigma_1 \neq \Sigma_2$. Since the Bayes decision boundary is non-linear, it is more accurately approximated by QDA than by LDA.



K-nearest neighbors (KNN) classifier

Given

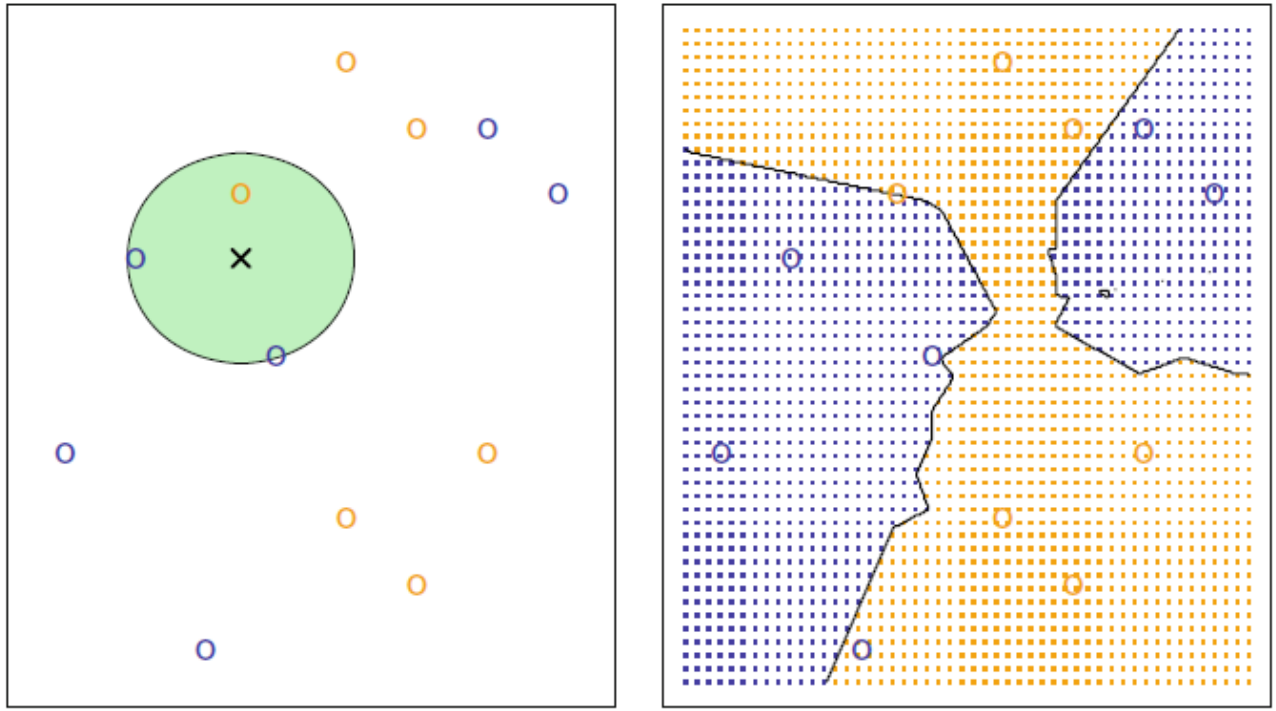
- training data $(\mathbf{x}_i, y_i),\ i = 1, \cdots, n,$
- a positive integer K,
- a test observation $\mathbf{x}_0$.

the KNN classifier

- identifies the neighbors K points in the training data that are closest to x_0, represented by $\mathcal{N}_0$.
- estimates the conditional probability for class j as the fraction of points in N0 whose response values equal j:

$$P(Y = j|X = \mathbf{x}_0) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} I(y_i = j).$$

- KNN applies Bayes rule and classifies the test observation x_0 to the class with the largest probability.



- The KNN approach, using $K = 3$, training: six blue observations and six orange observations.
- Left: a test observation at which a predicted class label is desired is shown as a black cross.
- Right: The KNN decision boundary. The blue grid indicates the region in which a test observation will be assigned to the blue class, and the orange grid indicates the region in which it will be assigned to the orange class.