MONASH University

ETC3250

Business Analytics

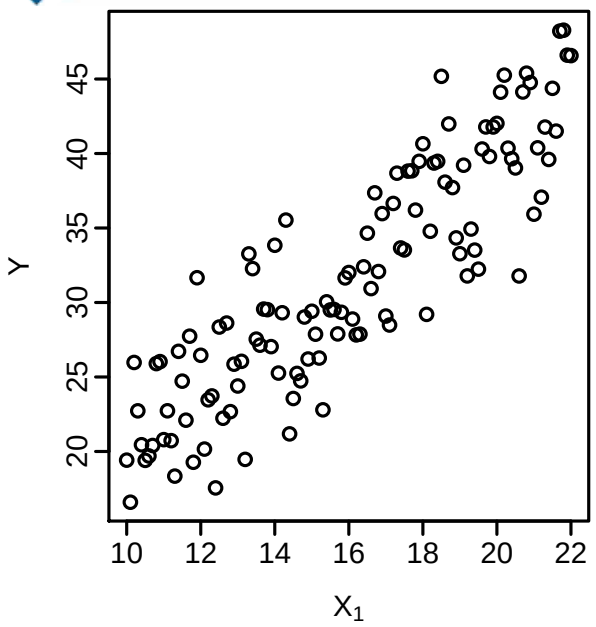
Week 5
Classification

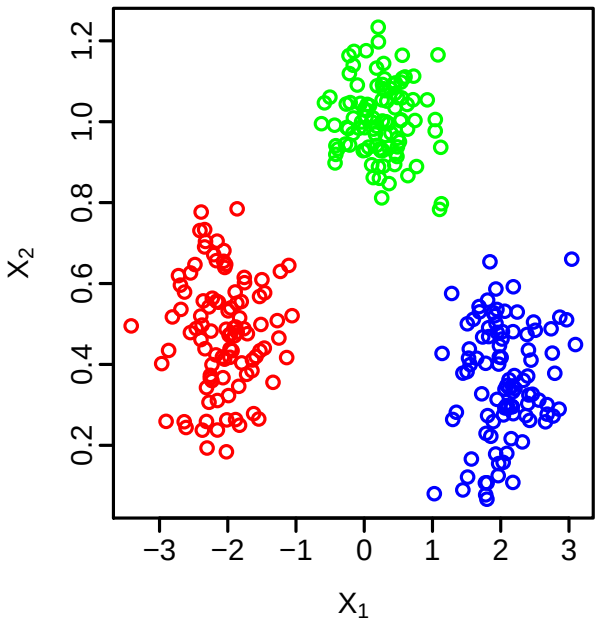
14 September 2015

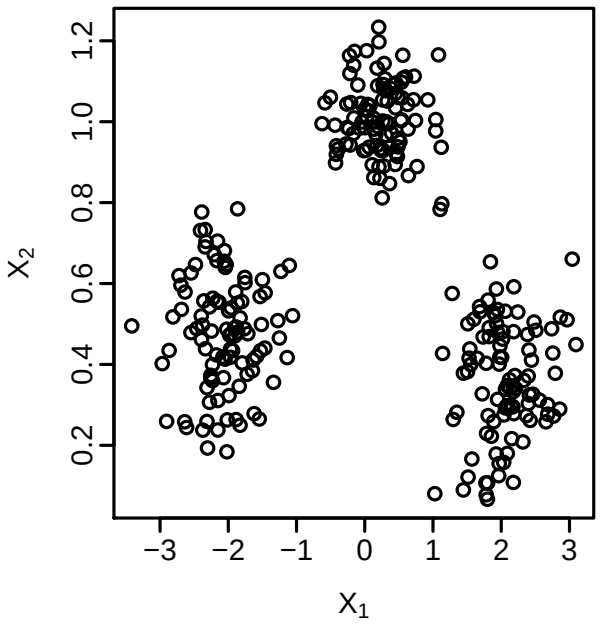
Outline

Week	Topic	Chapter	Lecturers
1	Introduction to business analytics & R	1	Rob,Souhaib
2	Statistical learning	2	Rob,Souhaib
3	Regression for prediction	3	Rob
4	Resampling	5	Rob, Souhaib
5	Dimension reduction	6,10	Rob, Souhaib
6	Visualization		Di
7	Visualization		Di
8	Classification	4	Souhaib, Di
9	Classification	4,9	Di, Souhaib
-	Semester Break		
10	Advanced classification	8	Di
11	Advanced regression	6	Di
12	Clustering	10	Di

What is classification?







$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N \text{ or } \{x_i\}_{i=1}^N, \quad x_i = (x_{i1}, \ldots, x_{ip})^T$$

- Supervised learning: classification and regression
- Unsupervised learning: clustering

What is classification?

The response variable Y is qualitative.

- e.g., email is one of $\mathcal{C} = (\text{spam}, \text{ham})$
- e.g., voters are one of $\mathcal{C} = (\text{Liberal}, \text{Labor}, \text{Green}, \text{National}, \text{Other})$

Our goals are:

1. Build a classifier $C(x)$ that assigns a class label from $\mathcal{C}$ to a future unlabeled observation x.
2. Assess the uncertainty in each classification (i.e., the probability of misclassification).
3. Understand the roles of the different predictors among $X = (X_1, X_2, \ldots, X_p)$.

Optimal classifier

In place of MSE, we now use the error rate:

$$E[I(Y \neq \hat{f}(X))]$$

Suppose the K classes in $\mathcal{C}$ are numbered $1, 2, \ldots, K$. Let

$$p_k(x) = \Pr(Y = k \mid X = x), \qquad k = 1, 2, \ldots, K.$$

These are the conditional class probabilities at x. Then the Bayes classifier at x is

$$C(x) = j \quad \text{if } p_j(x) = \max\{p_1(x), p_2(x), \ldots, p_K(x)\}$$

Optimal classifier

In place of MSE, we now use the error rate:

$$E[I(Y \neq \hat{f}(X))]$$

Suppose the K classes in $\mathcal{C}$ are numbered $1, 2, \ldots, K$. Let

$$p_k(x) = \Pr(Y = k \mid X = x), \qquad k = 1, 2, \ldots, K.$$

These are the conditional class probabilities at x.

Then the Bayes classifier at x is

$$C(x) = j \quad \text{if } p_j(x) = \max\{p_1(x), p_2(x), \ldots, p_K(x)\}$$

Optimal classifier

In place of MSE, we now use the error rate:

$$E[I(Y \neq \hat{f}(X))]$$

Suppose the K classes in $\mathcal{C}$ are numbered $1, 2, \ldots, K$. Let

$$p_k(x) = \Pr(Y = k \mid X = x), \qquad k = 1, 2, \ldots, K.$$

These are the conditional class probabilities at x. Then the Bayes classifier at x is

$$C(x) = j \quad \text{if } p_j(x) = \max\{p_1(x), p_2(x), \ldots, p_K(x)\}$$

Optimal classifier

- The Bayes classifier gives the minimum average test error rate, called the “Bayes error rate”: $1 - \mathrm{E}\left(\max_j \Pr(Y = j|X)\right)$
- The “Bayes error rate” is the lowest possible error rate that could be achieved if we knew exactly the “true” probability distribution of the data.
- It is analagous to the “irreducible error” in regression.
- In reality, the Bayes classifier is not known.

Classification of classification methods

- Generative and discriminative models
  - $P(X, Y)$ vs $P(Y|X)$
- Logistic regression
  - Today
- Linear Discriminant Analysis
  - Thursday
- Support Vector Machines
  - Monday, next week
- k-Nearest Neighbours
  - Thursday, next week
- Advanced methods: Boosting, Random Forests, etc
  - Week 10

Classification with linear regression

- How would you use linear regression for classification?
- Binary classification
  - $Y = 0$ (stroke) and $Y = 1$ (drug overdose)
  - $E[Y|X] = P(Y = 1|X)$
  - Problem: estimates outside $[0, 1]$
- Multi-class classification
  - $Y = 1$ (stroke), $Y = 2$ (drug overdose) and $Y = 3$ (epileptic seizure)
  - $Y = 1$ (epileptic seizure), $Y = 2$ (stroke) and $Y = 3$ (drug overdose)
  - Problem: Each coding will produce fundamentally different linear models

Classification with linear regression

- How would you use linear regression for classification?
- Binary classification
  - $Y = 0$ (stroke) and $Y = 1$ (drug overdose)
  - $E[Y|X] = P(Y = 1|X)$
  - Problem: estimates outside $[0, 1]$
- Multi-class classification
  - $Y = 1$ (stroke), $Y = 2$ (drug overdose) and $Y = 3$ (epileptic seizure)
  - $Y = 1$ (epileptic seizure), $Y = 2$ (stroke) and $Y = 3$ (drug overdose)
  - Problem: Each coding will produce fundamentally different linear models

Classification with linear regression

- How would you use linear regression for classification?
- Binary classification
  - $Y = 0$ (stroke) and $Y = 1$ (drug overdose)
  - $E[Y|X] = P(Y = 1|X)$
  - Problem: estimates outside $[0, 1]$
- Multi-class classification
  - $Y = 1$ (stroke), $Y = 2$ (drug overdose) and $Y = 3$ (epileptic seizure)
  - $Y = 1$ (epileptic seizure), $Y = 2$ (stroke) and $Y = 3$ (drug overdose)
  - Problem: Each coding will produce fundamentally different linear models

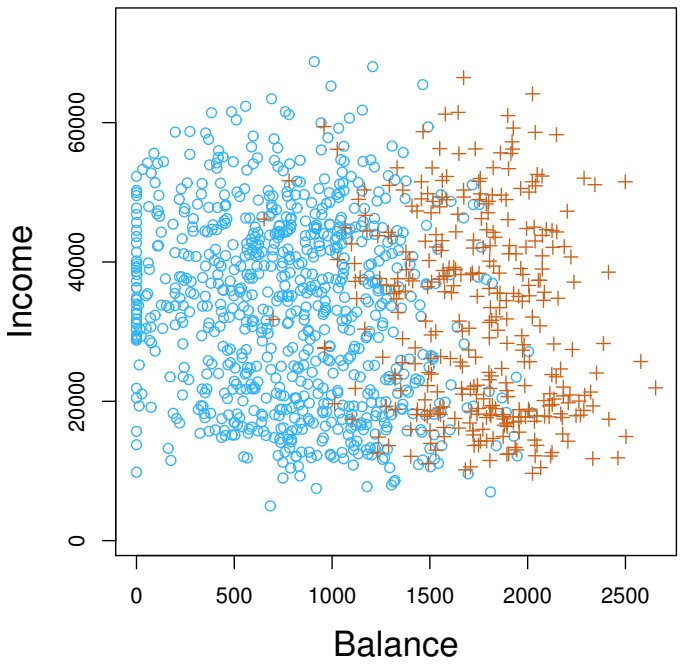
Classification with linear regression

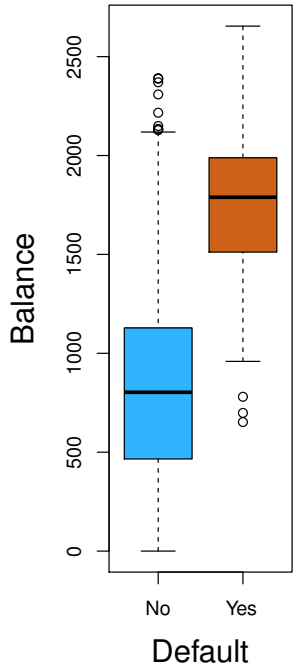
- How would you use linear regression for classification?
- Binary classification
  - $Y = 0$ (stroke) and $Y = 1$ (drug overdose)
  - $E[Y|X] = P(Y = 1|X)$
  - Problem: estimates outside $[0, 1]$
- Multi-class classification
  - $Y = 1$ (stroke), $Y = 2$ (drug overdose) and $Y = 3$ (epileptic seizure)
  - $Y = 1$ (epileptic seizure), $Y = 2$ (stroke) and $Y = 3$ (drug overdose)
  - Problem: Each coding will produce fundamentally different linear models

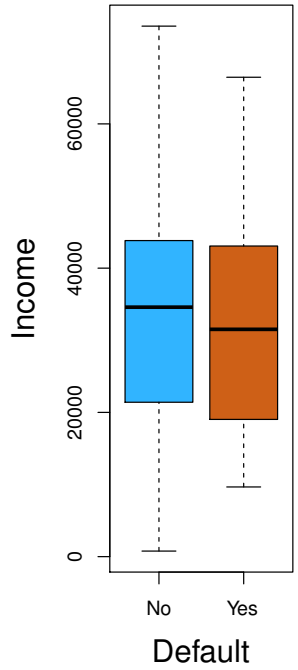
Classification with linear regression

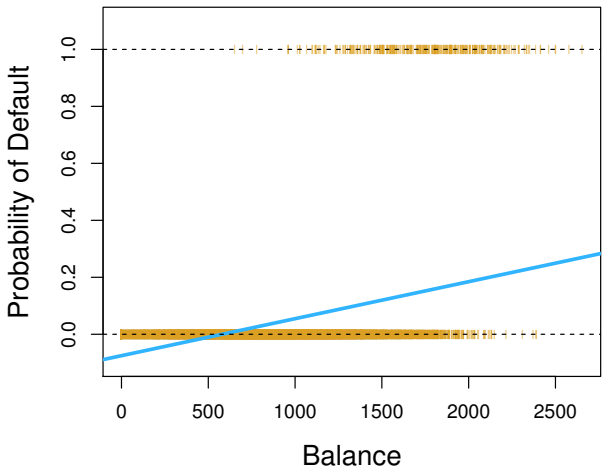
- How would you use linear regression for classification?
- Binary classification
  - $Y = 0$ (stroke) and $Y = 1$ (drug overdose)
  - $E[Y|X] = P(Y = 1|X)$
  - Problem: estimates outside $[0, 1]$
- Multi-class classification
  - $Y = 1$ (stroke), $Y = 2$ (drug overdose) and $Y = 3$ (epileptic seizure)
  - $Y = 1$ (epileptic seizure), $Y = 2$ (stroke) and $Y = 3$ (drug overdose)
  - Problem: Each coding will produce fundamentally different linear models

Why not linear regression?









$$p(X) = P(Y = 1|X) = \beta_0 + \beta_1 X$$

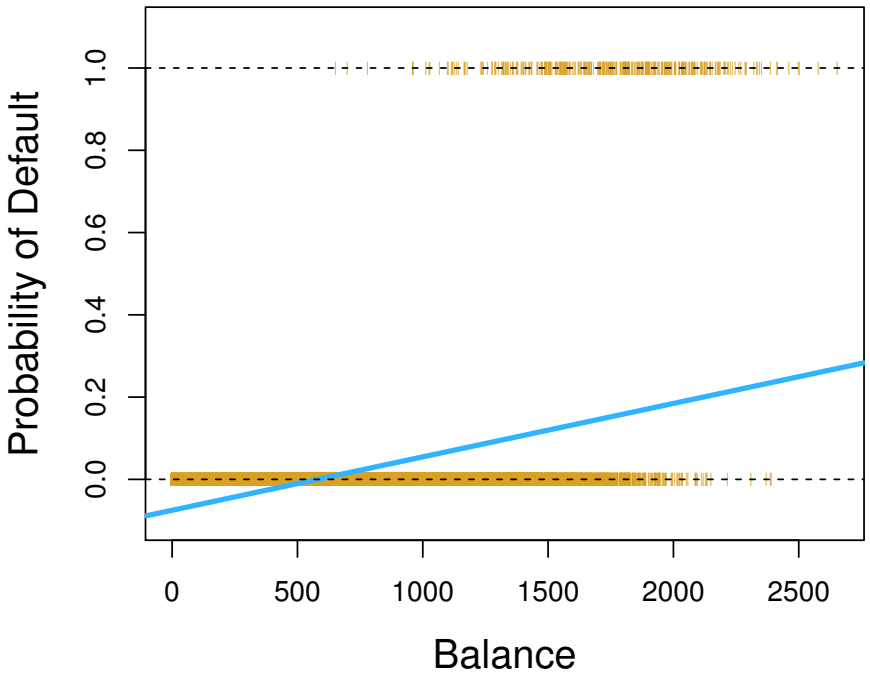
$$p(\text{balance}) = p(\text{default} = \text{Yes}|\text{balance})$$

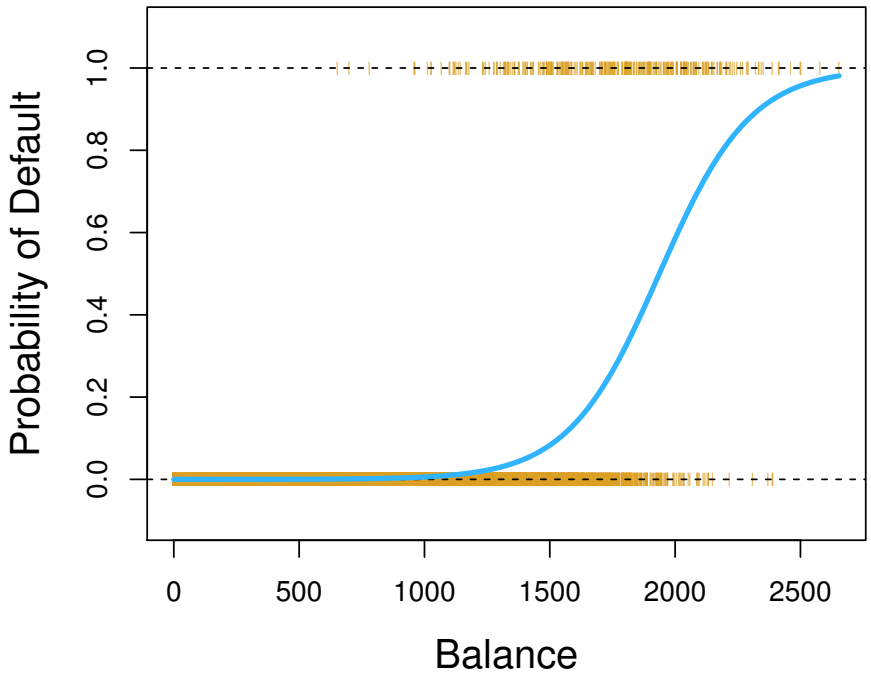
default = Yes if $\hat{p}(\text{balance}) > 0.5$

Problem: $\hat{p}(X) > 0$ or $\hat{p}(X) < 0$

Logistic regression

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





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

Logistic regression

Logistic regression

	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

	Coefficient	Std. error	Z-statistic	P-value
Intercept	−3.5041	0.0707	−49.55	<0.0001
student[Yes]	0.4049	0.1150	3.52	0.0004

Linear regression

	Coefficient	Std. error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

Logistic regression

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$$

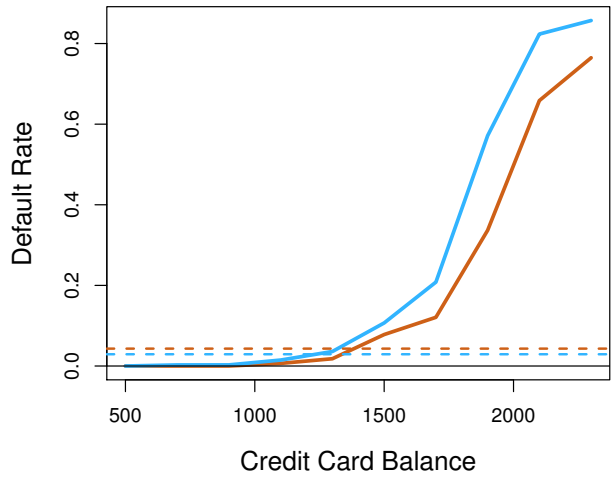
	Coefficient	Std. error	Z-statistic	P-value
`Intercept`	−10.8690	0.4923	−22.08	<0.0001
`balance`	0.0057	0.0002	24.74	<0.0001
`income`	0.0030	0.0082	0.37	0.7115
`student[Yes]`	−0.6468	0.2362	−2.74	0.0062

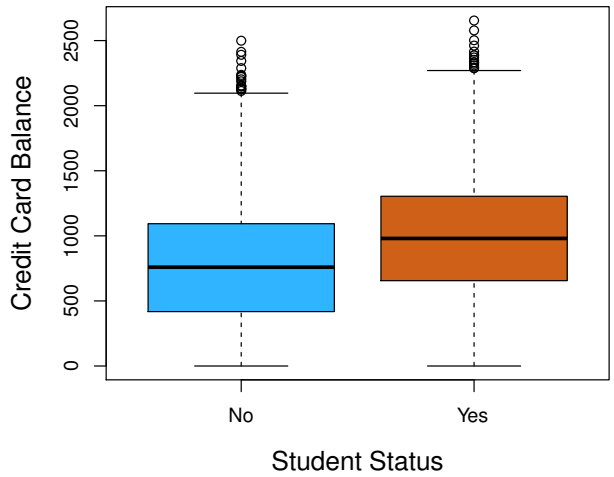
	Coefficient	Std. error	Z-statistic	P-value
`Intercept`	−10.6513	0.3612	−29.5	<0.0001
`balance`	0.0055	0.0002	24.9	<0.0001

	Coefficient	Std. error	Z-statistic	P-value
`Intercept`	−3.5041	0.0707	−49.55	<0.0001
`student[Yes]`	0.4049	0.1150	3.52	0.0004

Logistic regression

	Coefficient	Std. error	Z-statistic	P-value
Intercept	−10.8690	0.4923	−22.08	<0.0001
balance	0.0057	0.0002	24.74	<0.0001
income	0.0030	0.0082	0.37	0.7115
student[Yes]	−0.6468	0.2362	−2.74	0.0062





A student is **riskier** than a non-student without information about the student's credit card balance. However, that student is **less risky** than a non-student with the same credit card balance.

Logistic regression

- We define the *odds* as $\frac{p(X)}{1-p(X)}$
- Example: $p(X) = 0.2$ and $p(X) = 0.9$
- Odds close to 0 and $\infty \rightarrow$ very low and very high probabilities of default, respectively.
- Log-odds (or logit) is linear in X: $log(\frac{p(X)}{1-p(X)}) = \beta_0 + \beta_1 X$

Evaluation of classifiers

default = Yes if $\hat{p}(\text{balance}) > 0.5$

		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

What is the problem with the *overall error rate*?

Evaluation of classifiers

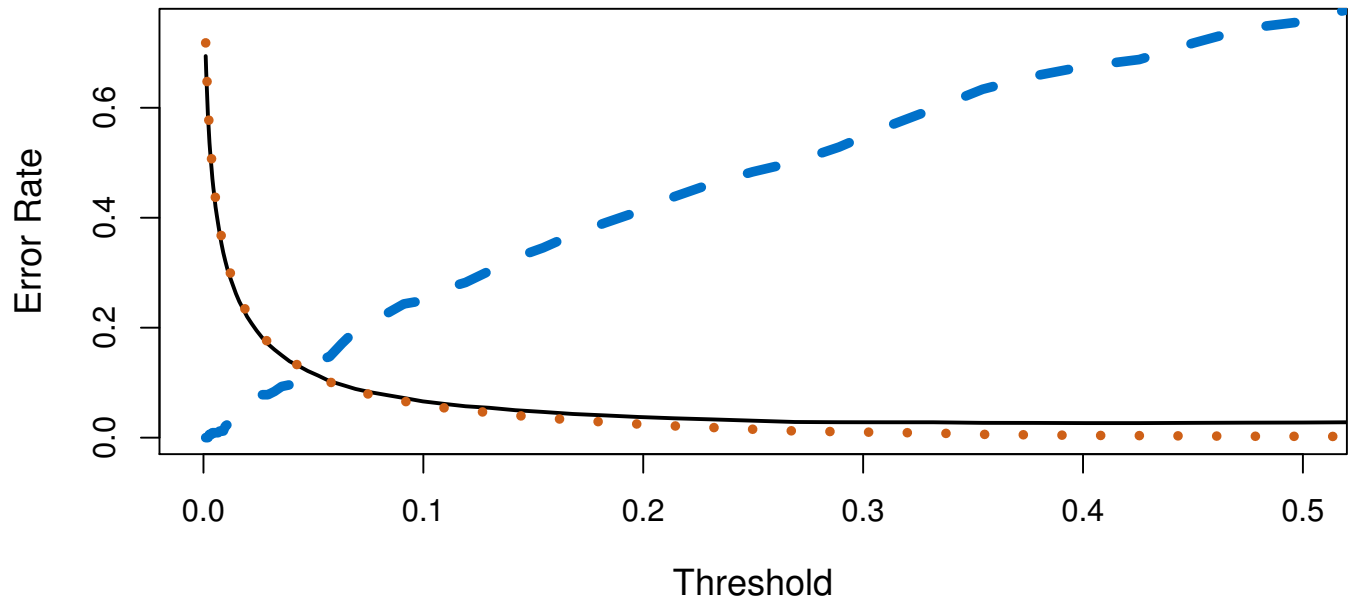
default = Yes if $\hat{p}(\text{balance}) > 0.5$

		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

default = Yes if $\hat{p}(\text{balance}) > 0.2$

		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

Evaluation of classifiers



- Overall error rate (black solid)
- False positive (blue dashed)
- False negative (orange dotted)

How to choose the threshold?

Evaluation of classifiers

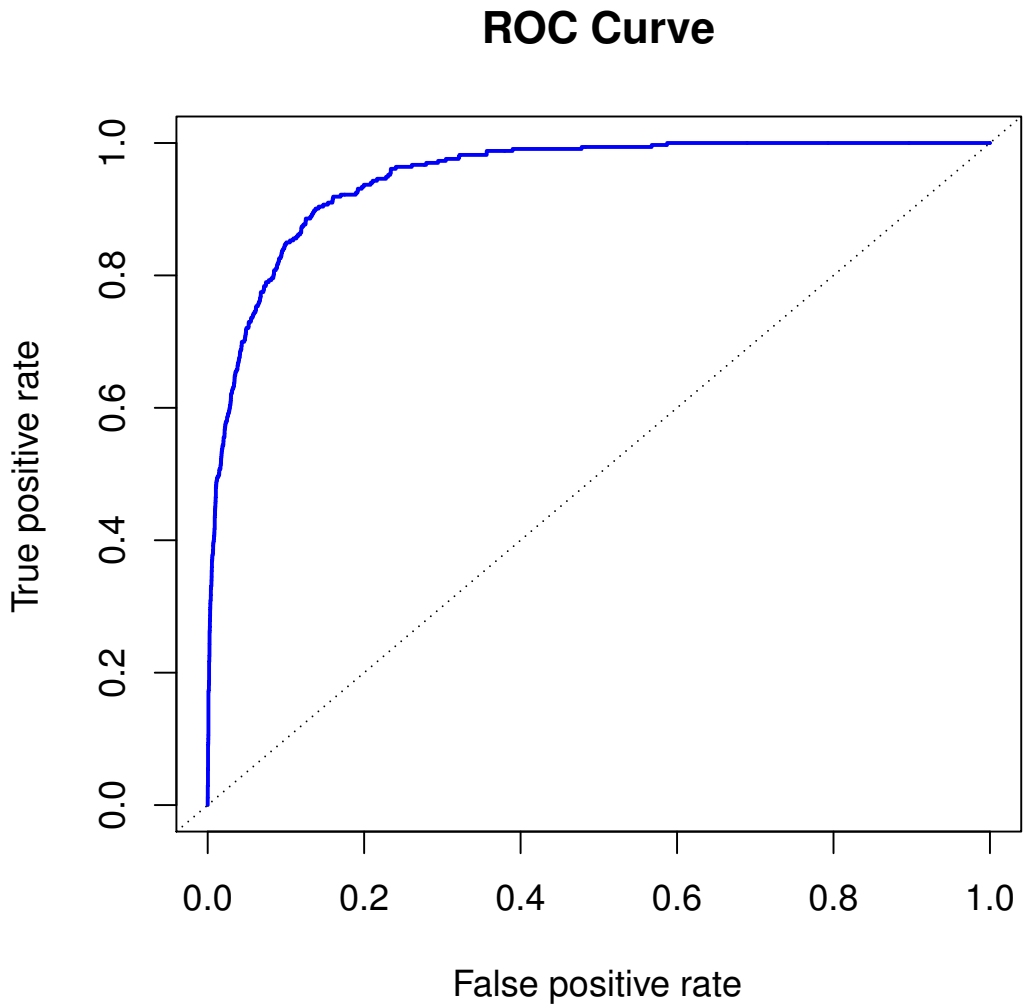
		True default status		
		No	Yes	Total
Predicted default status	No	9,644	252	9,896
	Yes	23	81	104
	Total	9,667	333	10,000

		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 proportion
Neg. Pred. value	TN/N*	

Evaluation of classifiers

The receiver operating characteristic (ROC) curve



Next class

Linear Discriminant Analysis (LDA)

- Logistic regression is a discriminative model while LDA is a generative model
- How would you use binary logistic regression for multi-class classification? LDA is a better approach for multi-class classification