MONASH University

ETC3250

Business Analytics

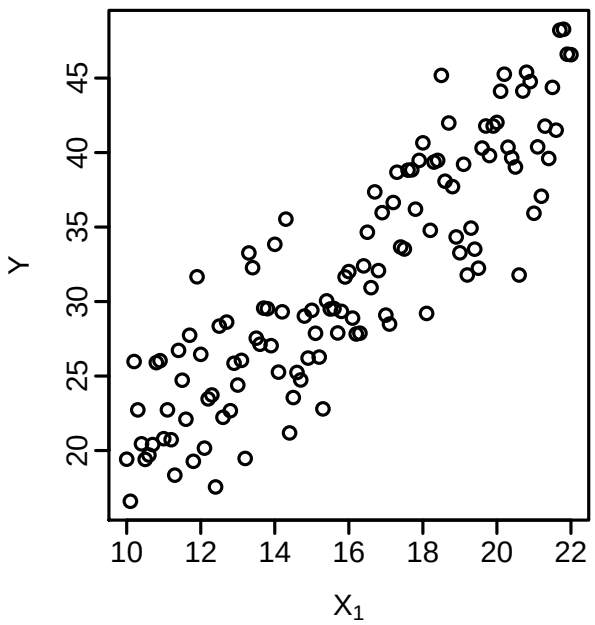
Week 5
Classification

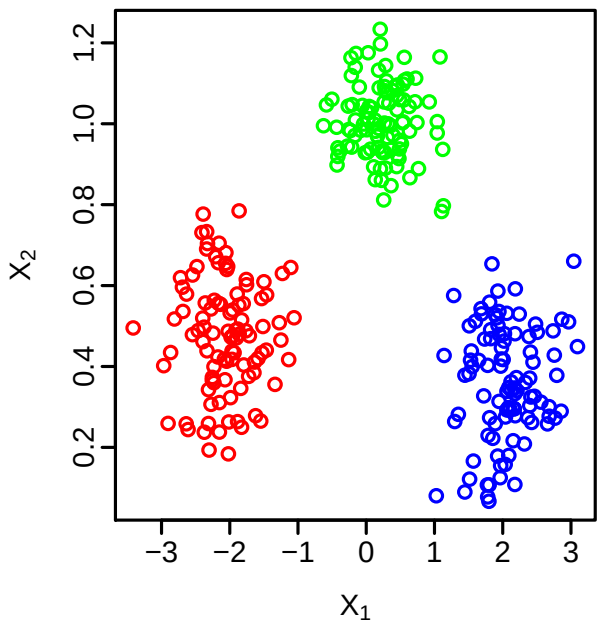
3 October 2016

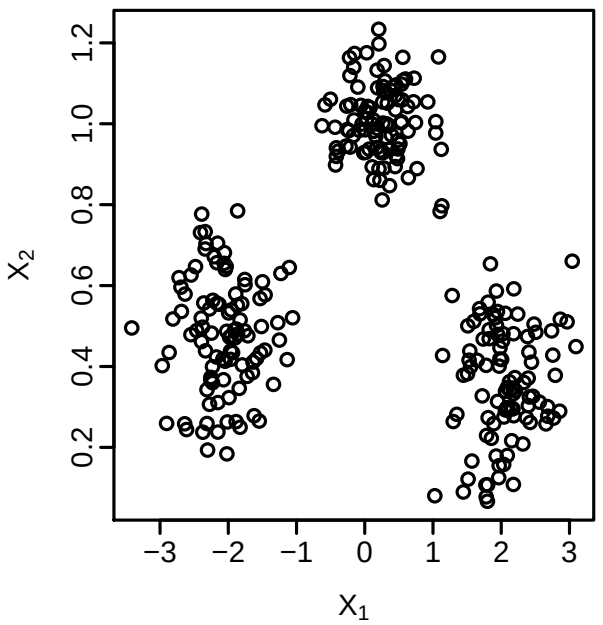
Outline

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

Classification







Optimal classifier

The Bayes classifier is the **optimal classifier** under the error rate:

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

The Bayes classifier at x is given by

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

where

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

Classification methods

- k-Nearest Neighbours
- Linear Discriminant Analysis
- Support Vector Machines
- Trees (+ Random Forests)
- Logistic regression

Random Forests

1. For $b = 1$ to B:

   (a) Draw a bootstrap sample $\mathbf{Z}^*$ of size N from the training data.

   (b) Grow a random-forest tree T_b to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size n_{min} is reached.

      i. Select m variables at random from the p variables.

      ii. Pick the best variable/split-point among the m.

      iii. Split the node into two daughter nodes.

2. Output the ensemble of trees $\{T_b\}_1^B$.

To make a prediction at a new point x:

Regression: $\hat{f}_{\text{rf}}^B(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$.

Classification: Let $\hat{C}_b(x)$ be the class prediction of the bth random-forest tree. Then $\hat{C}_{\text{rf}}^B(x) = \textit{majority vote}\ \{\hat{C}_b(x)\}_1^B$.

Random Forests: regression

$$\hat{f}_{\text{rf}}^{B}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$$

$$\text{MSE} = E[(y - \hat{f}_{\text{rf}}^{B}(x))^2]$$

$$\text{MSE} = \text{BIAS}^2 + \text{VARIANCE}$$

- BIAS: If trees are sufficiently deep, they have very small bias. Why?
- VARIANCE = $\text{Var}(\frac{1}{B}\sum_{b=1}^{B} T_b(x)) = ?$

Random Forests: regression

- $\text{Var}(T_b(x)) = \sigma^2$
- $\text{Cor}(T_i(x), T_j(x)) = \rho > 0 \ (i \neq j)$

$$\text{Var}\left(\frac{1}{B}\sum_{b=1}^{B} T_b(x)\right) = \rho\sigma^2 + \sigma^2\frac{1-\rho}{B}$$

How do we reduce it?

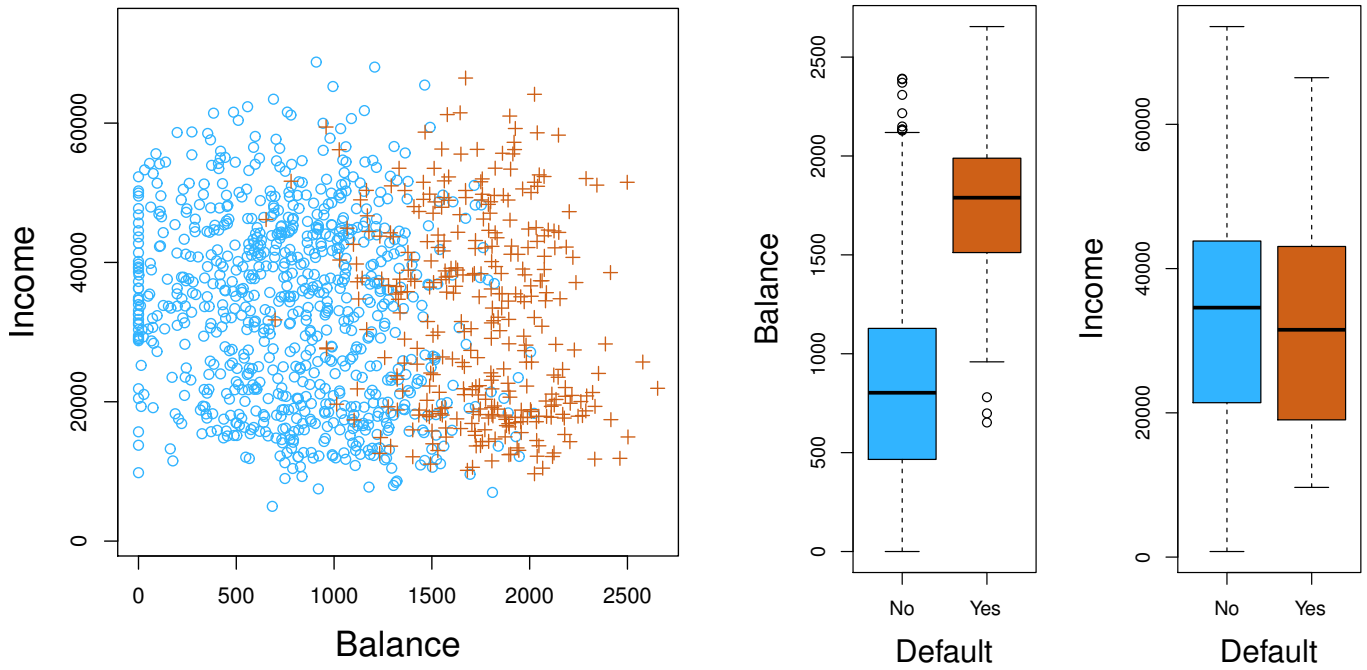
Random Forests: regression

$$\text{Var}\left(\frac{1}{B}\sum_{b=1}^{B} T_b(x)\right) = \rho\sigma^2 + \sigma^2\frac{1-\rho}{B}$$

- $\rho\sigma^2$: decreases if ρ decreases (i.e. if m decreases)
- $\sigma^2\frac{1-\rho}{B}$: decreases if B decreases (irrespective of ρ)

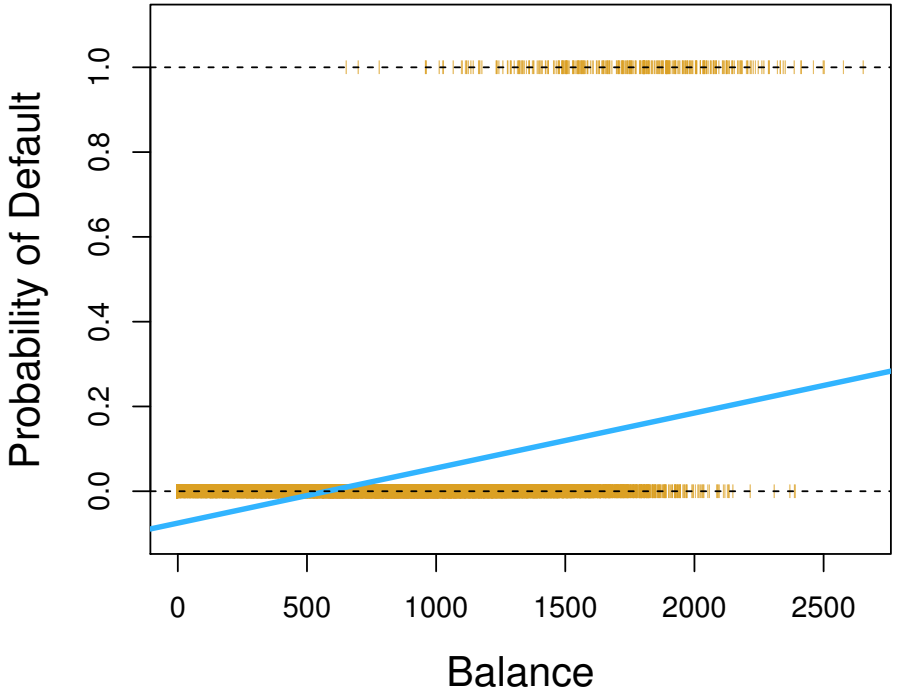
Classification with linear regression

We are interested in predicting whether an individual will default on his or her credit card payment, on the basis of annual income and monthly credit card balance.



How would you use linear regression for classification?

Classification with linear regression



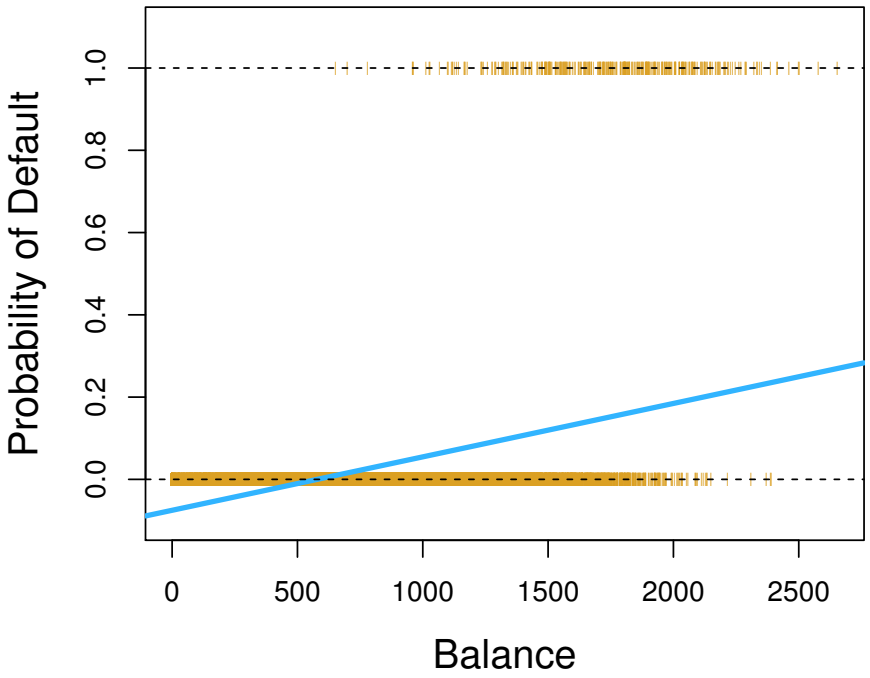
- $p(X) = P(Y = 1|X) = \beta_0 + \beta_1 X$
  $p(\text{balance}) = p(\text{default} = \text{Yes}|\text{balance})$
- default = Yes if $\hat{p}(\text{balance}) > 0.5$
- Problem I: $\hat{p}(X) > 0$ or $\hat{p}(X) < 0$

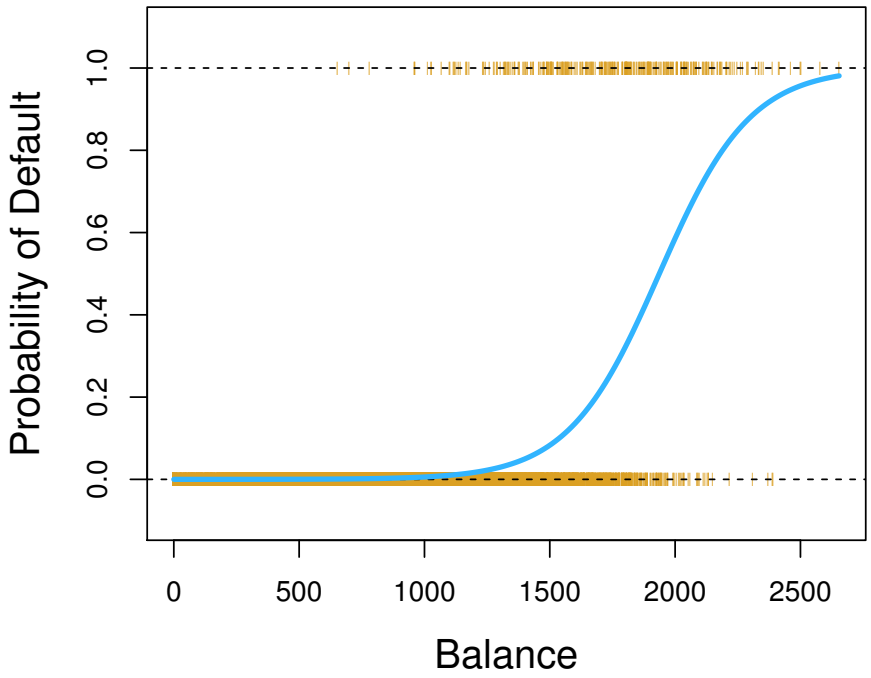
Classification with linear regression

- Binary classification
  - $Y = 0$ (stroke) and $Y = 1$ (drug overdose)
  - $E[Y|X] = P(Y = 1|X)$
  - Problem I: 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 II: Each coding will produce fundamentally different linear models

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

	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

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

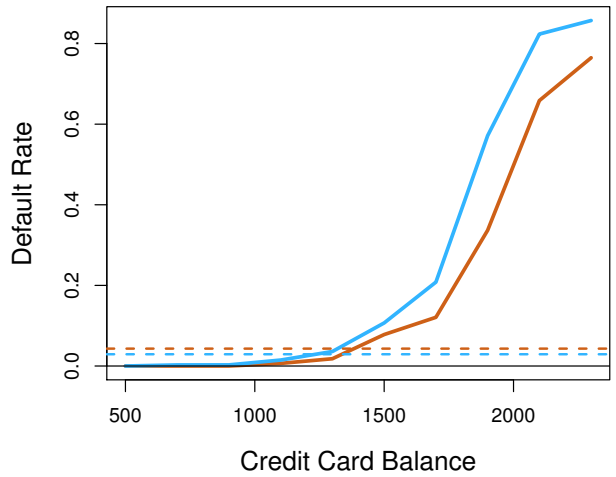
	Coefficient	Std. error	Z-statistic	P-value
Intercept	−10.8690	0.4923	−22.08	<0.0001
balance	0.0057	0.0002	24.74	<0.0001
income	0.0030	0.0082	0.37	0.7115
student[Yes]	−0.6468	0.2362	−2.74	0.0062

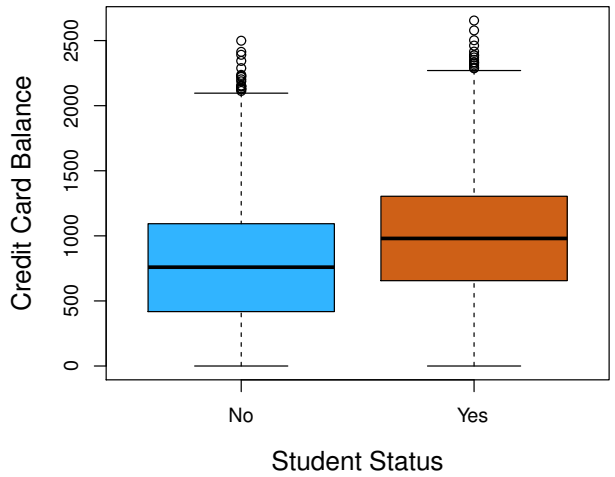
	Coefficient	Std. error	Z-statistic	P-value
Intercept	−3.5041	0.0707	−49.55	<0.0001
student[Yes]	0.4049	0.1150	3.52	0.0004

How is it possible for student status to be associated with an **increase** in probability of default in one case and a **decrease** in probability of default in the other case?!

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 $\frac{p(X)}{1-p(X)}$
- 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

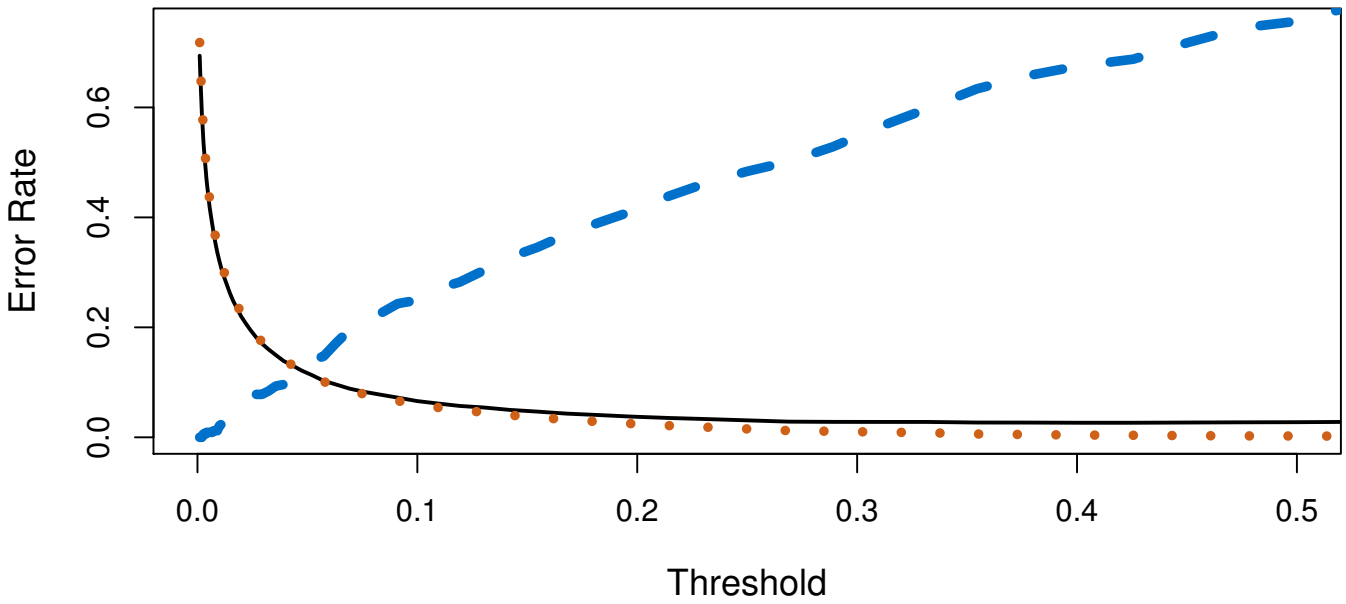
Evaluation of classifiers

		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

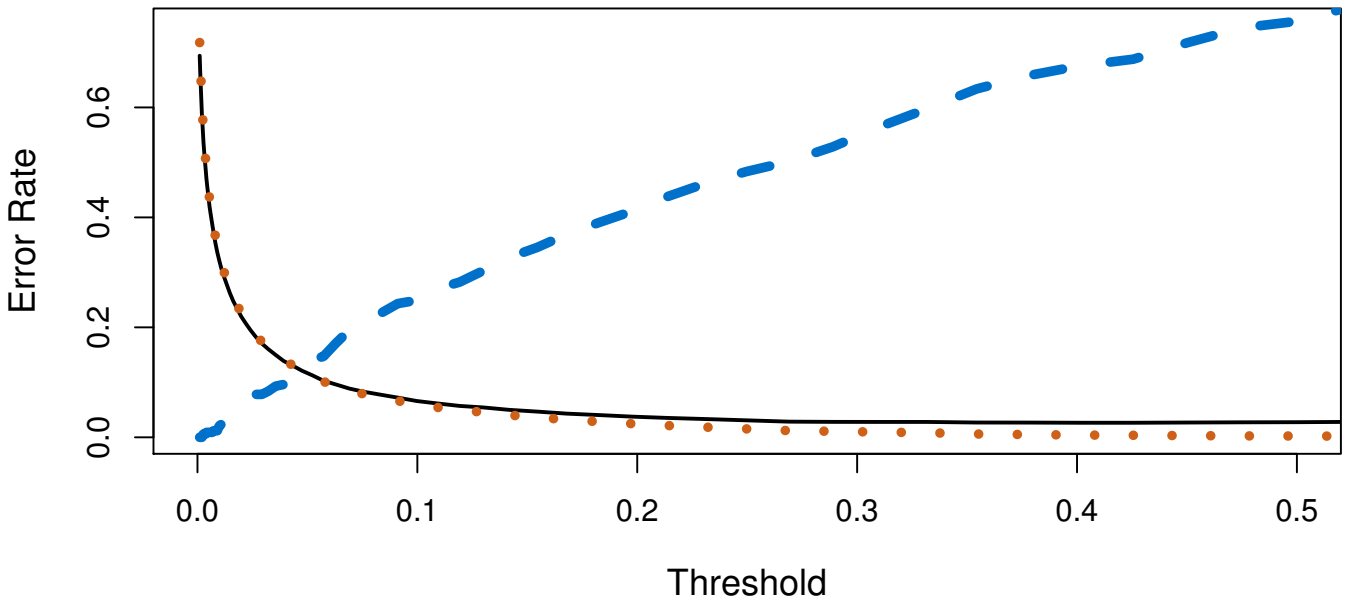


- Overall error rate (black solid)
- False positive (blue dashed)
- False negative (orange dotted)

How to choose the threshold?

Use prior knowledge about the cost associated with default for example.

Evaluation of classifiers



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

How to choose the threshold?

Use prior knowledge about the cost associated with default for example.

Evaluation of classifiers

The receiver operating characteristic (ROC) curve