MONASH University

ETC3250

Business Analytics

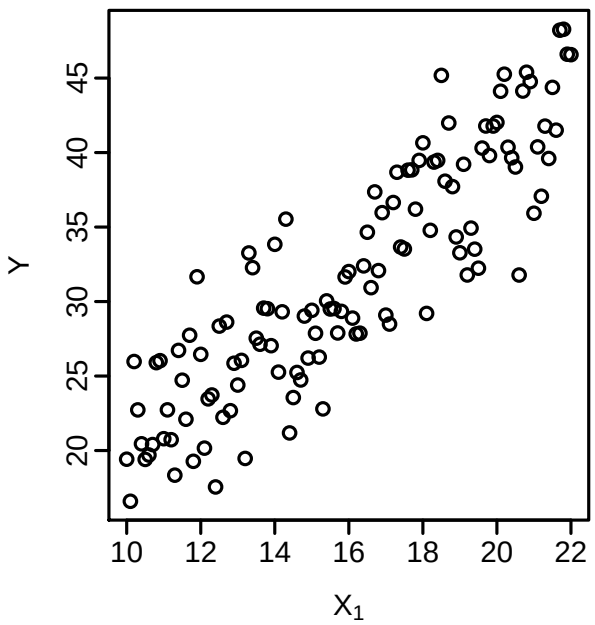
Week 4.
Logistic regression

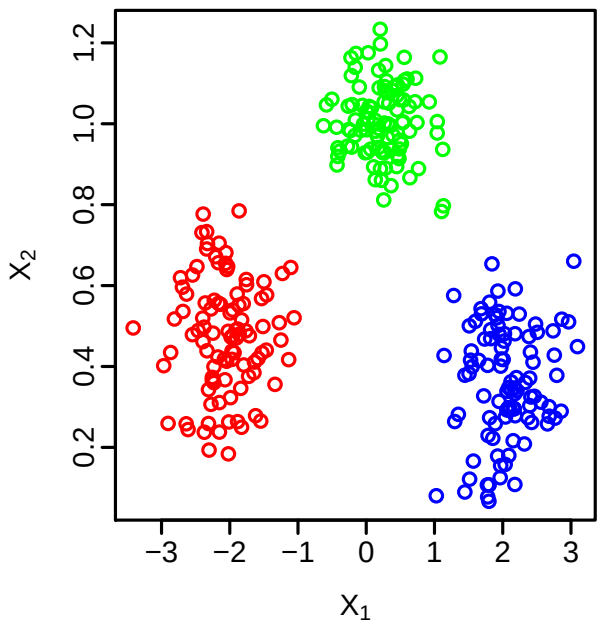
14 August 2017

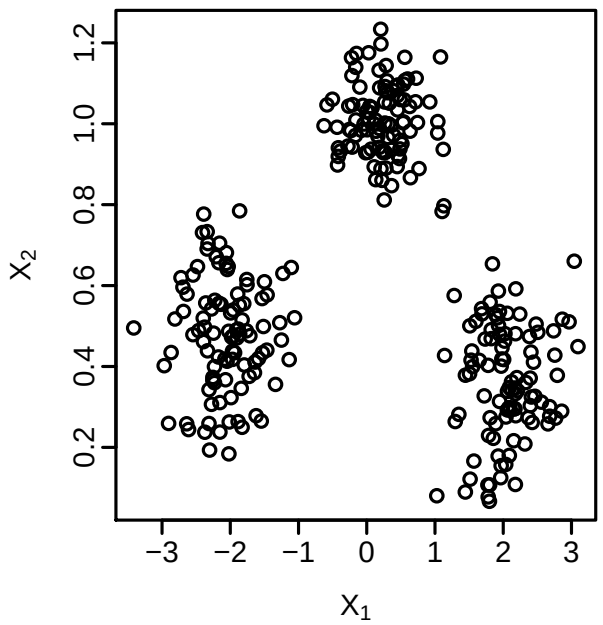
Outline

Week	Topic	Chapter	Lecturer
1	Introduction to business analytics & R	1	Souhaib
2	Statistical learning	2	Souhaib
3	Regression for prediction	3,7	Tas & David
4	Classification	4	Souhaib
	Logistic regression		Souhaib
	Linear discriminant analysis		Souhaib
5	Classification	4, 9	Souhaib
6	Resampling methods	5	Souhaib
7	Dimension reduction	6,10	Souhaib
8	Advanced regression	6	Souhaib
9	Advanced learning methods	8	Souhaib
	Semester break		
10	Clustering	10	Souhaib
11	Visualization		Souhaib
12	Data wrangling		Souhaib

Classification







Classification

$$\mathcal{D} = \{(y_i, \boldsymbol{x}_i)\}_{i=1}^{N},$$

where

- $y_i \in \mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_K\}$ and $\boldsymbol{x}_i \in \mathbb{R}^d$
- $(y_i, \boldsymbol{x}_i) \sim P(Y, \boldsymbol{X}) = P(\boldsymbol{X})\underbrace{P(Y|\boldsymbol{X})}$.

We want to build a classifier $C(\boldsymbol{x})$ that assigns a class label from $\mathcal{C}$ to a new unlabeled observation $\boldsymbol{x}$.

Classification

The optimal classifier at $\boldsymbol{x}$, i.e the classifier which minimizes $\mathbb{E}[I(Y \neq C(\boldsymbol{x}))]$ is

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

where

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

- We do not know $p_k(\boldsymbol{x})$.
- We only observe $\mathcal{D} = \{(y_i, \boldsymbol{x}_i)\}_{i=1}^{N}$.

$\rightarrow$ Estimate $p_k(\boldsymbol{x})$ using $\mathcal{D}$

$\rightarrow$ Use the class which maximizes $\hat{p}_k(\boldsymbol{x})$

Classification problem

Instead of estimating and then maximizing $p_k(\boldsymbol{x})$, why not try to directly find the classifier that minimize the (sample) error rate $\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq C(\boldsymbol{x}_i))$?

If the true class is $y \in \{-1, 1\}$, and the prediction is $\hat{y}$, the zero-one loss is

$$I(y, \hat{y}) = I(y \neq \hat{y}) = \begin{cases} 1 & \text{if } y \neq \hat{y} \\ 0 & \textit{otherwise} \end{cases} = \begin{cases} 1 & \text{if } y\hat{y} < 0 \ (\neq \text{ signs}) \\ 0 & \textit{otherwise} \end{cases}$$

$\rightarrow$ the loss function is *non-convex* and *discontinuous* ☹

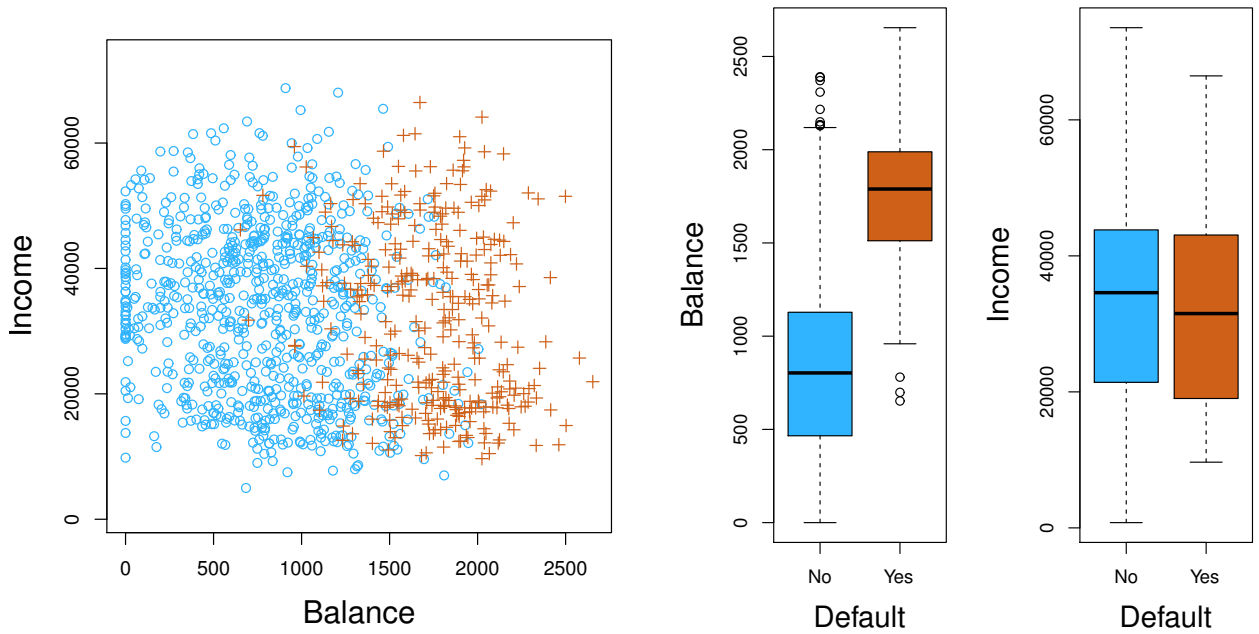
In binary classification, i.e. $Y \in \{\mathcal{C}_1, \mathcal{C}_2\}$, we have

$$E[Y|\boldsymbol{X} = \boldsymbol{x}] = \mathcal{C}_2 + P(Y = \mathcal{C}_1|\boldsymbol{X} = \boldsymbol{x})(\mathcal{C}_1 - \mathcal{C}_2)$$

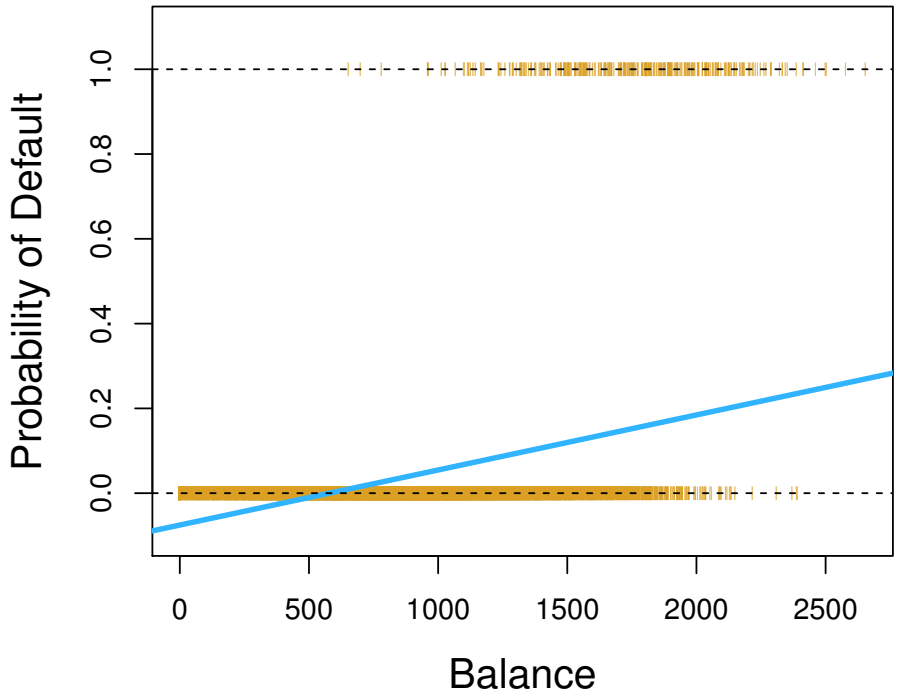
$\rightarrow$ we can use regression ☺

Credit card example

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.



Credit card example



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

Classification with linear regression

- Binary classification
  - $Y = 0$ (stroke) and $Y = 1$ (drug overdose)
  - 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

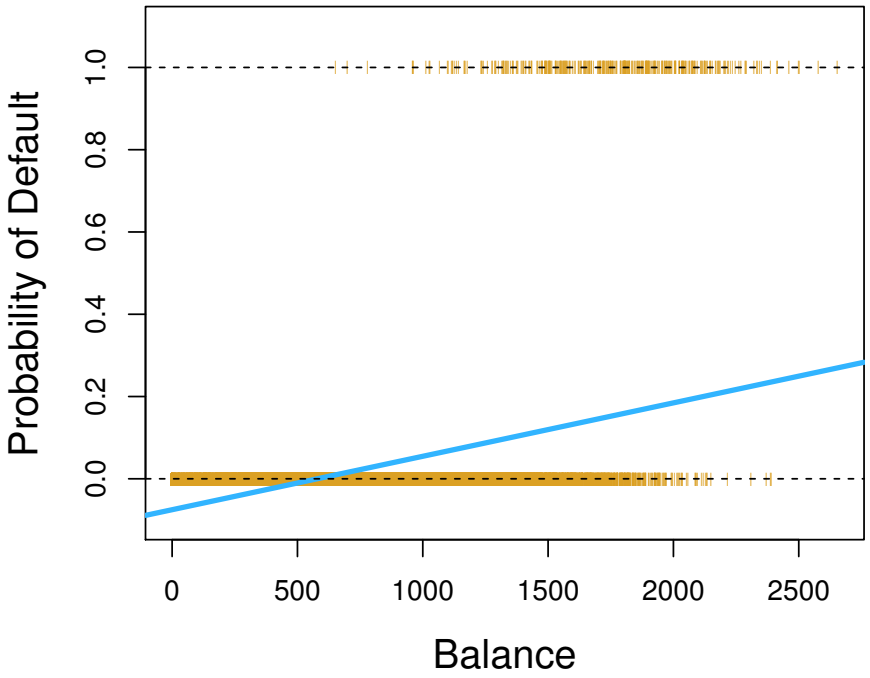
We model the probability of Y to be equal $y \in \{-1, 1\}$, given a data point $\boldsymbol{x}$, as:

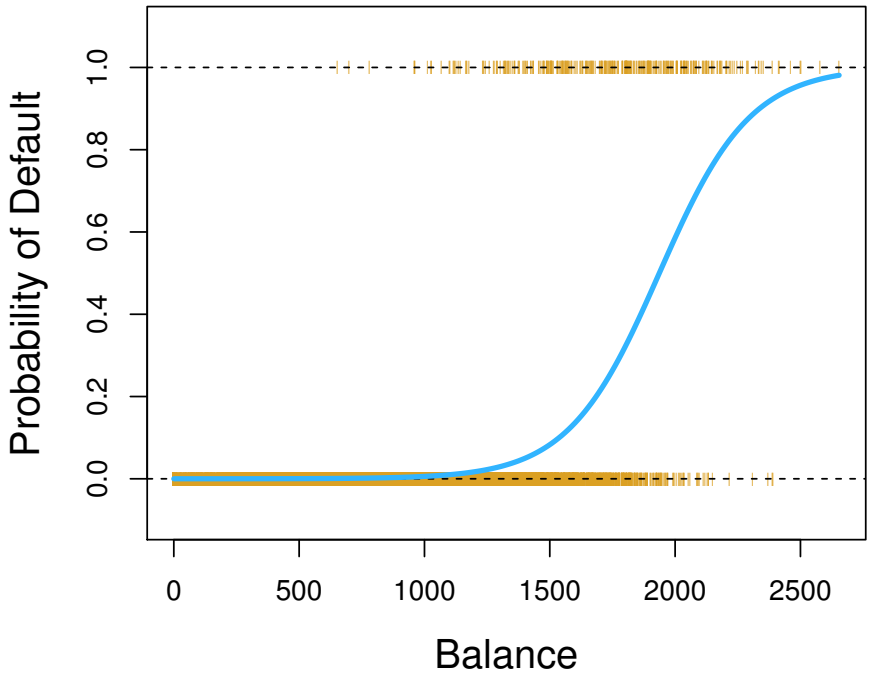
$$P(Y = y|\boldsymbol{x}) = \frac{1}{1 + e^{-y\beta'\boldsymbol{x}}} = \text{Logistic}(y\beta'\boldsymbol{x}).$$

This amounts to modeling the *log-odds ratio* as a linear function of $\boldsymbol{x}$:

$$log\left(\frac{P(Y = 1|\boldsymbol{x})}{P(Y = -1|\boldsymbol{x})}\right) = \beta'\boldsymbol{x}.$$

Credit card example





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

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

Interpretation: univariate case

- $E[Y|X] = \beta_0 + \beta_1 X$
  - β_1 gives the average change in Y associated with a one-unit increase in X
- $log(\frac{p(X)}{1-p(X)}) = \beta_0 + \beta_1 X \implies p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$
  - Increasing X by one unit changes the log odds by β_1, or equivalently it multiplies the odds by e^{β_1}
  - However, because $p(X)$ is not linear in X, β_1 does not correspond to the change in $p(X)$ associated with a one-unit increase in 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

	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

Interpretation: multivariate case

	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

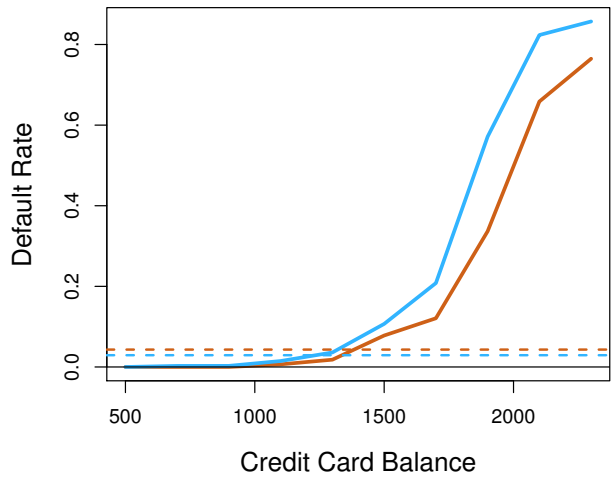
$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$$

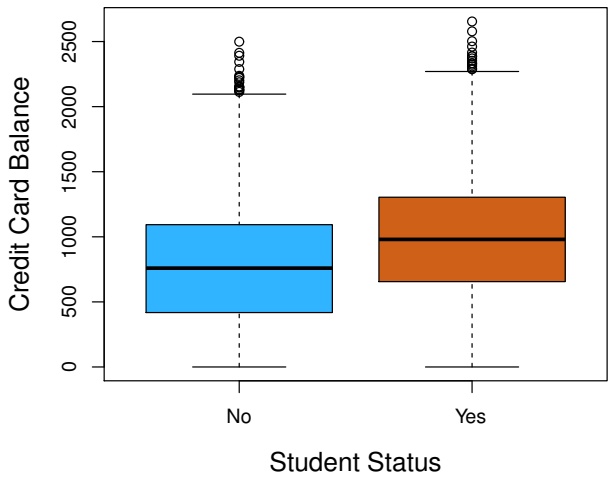
	Coefficient	Std. error	Z-statistic	P-value
Intercept	−10.8690	0.4923	−22.08	<0.0001
balance	0.0057	0.0002	24.74	<0.0001
income	0.0030	0.0082	0.37	0.7115
student[Yes]	−0.6468	0.2362	−2.74	0.0062

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.

Maximum Likelihood Estimation

$$P(Y = y|\boldsymbol{x}) = \frac{1}{1 + e^{-y\beta'\boldsymbol{x}}} = \text{Logistic}(y\beta'\boldsymbol{x}); \quad y \in \{-1, 1\}.$$

We choose parameters β to maximize the likelihood of the data given the model. The likelihood function is

$$\mathcal{L}_n(\beta) = \prod_{i=1}^{n} \frac{1}{1 + e^{-y_i\beta'\boldsymbol{x}_i}}.$$

It is more convenient to maximize the *log-likelihood function* $l_n(\beta) = log(\mathcal{L}_n(\beta))$:

$$\max_\beta l_n(\beta) := -\sum_{i=1}^{n} log(1 + e^{-y_i\beta'\boldsymbol{x}_i})$$

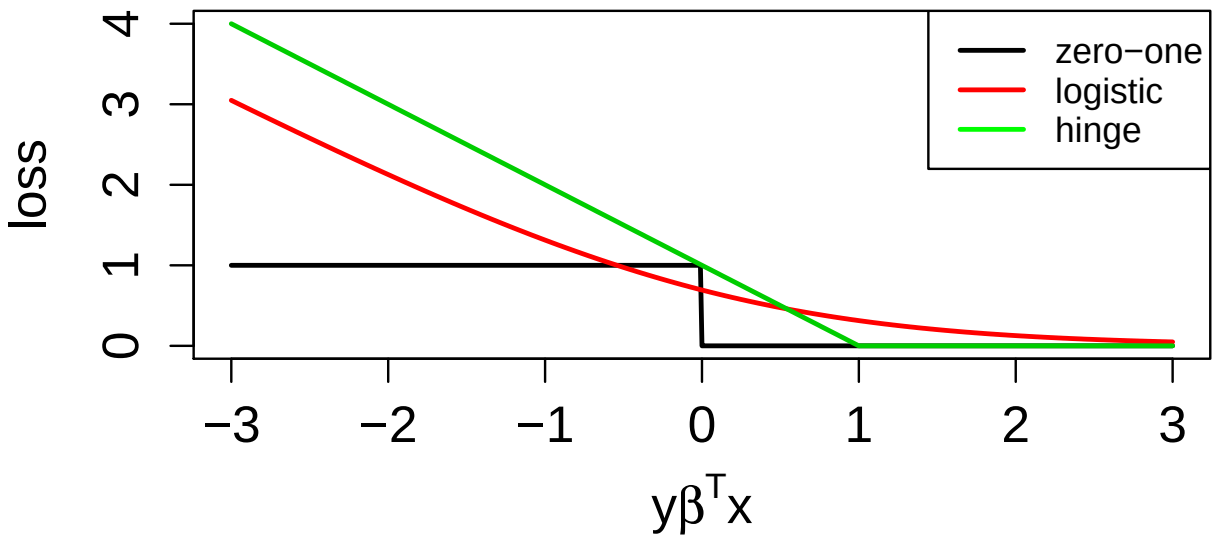
Using the Newton–Raphson algorithm, the parameters can be computed using *iteratively reweighted least squares*.

Relationship with error rate minimization?

Some common loss functions used for binary classification:

- zero-one loss: $l(y, \beta'\boldsymbol{x}) = \mathbf{1}\{y\beta'\boldsymbol{x} < 0\}$
- **logistic loss**: $l(y, \beta'\boldsymbol{x}) = log(1 + e^{-y\beta'\boldsymbol{x}})$
- hinge loss (later): $l(y, \beta'\boldsymbol{x}) = \max(0, 1 - y\beta'\boldsymbol{x})$

If $y = 1$ and $\hat{y} = \beta'\boldsymbol{x}$:



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 *overall error rate*? Is it good?

- The overall error rate is low. The error rate among individuals who defaulted ($252/333 = 75.7\%$) is very high.
- For a credit card company that is trying to identify high-risk individuals, an error rate of 75.7% among individuals who default may well be unacceptable.

$\rightarrow$ **Class-specific performance is also important**

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 *overall error rate*? Is it good?

- The overall error rate is low. The error rate among individuals who defaulted ($252/333 = 75.7\%$) is very high.
- For a credit card company that is trying to identify high-risk individuals, an error rate of 75.7% among individuals who default may well be unacceptable.

$\rightarrow$ **Class-specific performance is also important**

Evaluation of classifiers

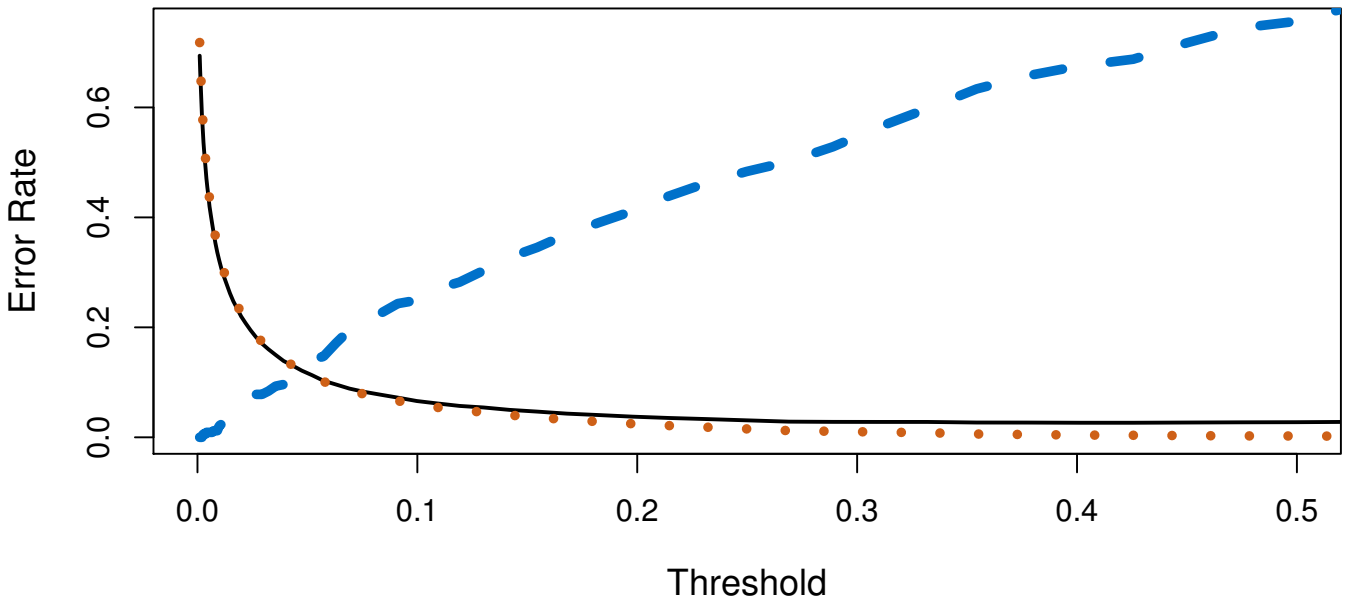
default = Yes if $\hat{p}(\text{balance}) > 0.5$

		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?

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

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 is a plot of the true positive rate (TPR) vs. false positive rate (FPR):

- TPR = TP/P: % “+” points correctly labeled “+” (sensitiviy).
- FPR = FP/N: % “-” points incorrectly labeled “+” (1 - specificity).