Machine Learning with Python

MTH786U/P 2023/24

Week 8: Classification tasks

Nicola Perra, Queen Mary University of London (QMUL)

n.perra@qmul.ac.uk

Classification

Similar to regression, *classification* relates input variables $\{\mathbf{x}_i\}_{i=1}^{s}$ to output variables $\{\mathbf{y}_i\}_{i=1}^{s}$, i.e.

$$\mathbf{y}_i \approx f(\mathbf{x}_i) \qquad \forall i \in \{1,\ldots,s\}$$

Classification

Similar to regression, *classification* relates input variables $\{\mathbf{x}_i\}_{i=1}^{s}$ to output variables $\{\mathbf{y}_i\}_{i=1}^{s}$, i.e.

$$\mathbf{y}_i \approx f(\mathbf{x}_i) \qquad \forall i \in \{1,\ldots,s\}$$

Key-difference in classification:

Classification

Similar to regression, *classification* relates input variables $\{\mathbf{x}_i\}_{i=1}^{s}$ to output variables $\{\mathbf{y}_i\}_{i=1}^{s}$, i.e.

$$\mathbf{y}_i \approx f(\mathbf{x}_i) \qquad \forall i \in \{1,\ldots,s\}$$

Key-difference in classification:

$\mathbf{y}$ can only take on discrete values!

Classification

Similar to regression, *classification* relates input variables $\{\mathbf{x}_i\}_{i=1}^{s}$ to output variables $\{\mathbf{y}_i\}_{i=1}^{s}$, i.e.

$$\mathbf{y}_i \approx f(\mathbf{x}_i) \qquad \forall i \in \{1, \ldots, s\}$$

Key-difference in classification:

$\mathbf{y}$ can only take on discrete values!

Example: $\mathbf{y} \in \{-1, 0, 1\}^s$ with an example vector

$$\mathbf{y} = (0, -1, -1, 1, 0)^\top \quad \text{for} \quad s = 5$$

Binary classification

When y can only take on two values, the classification is called *binary classification*

$y \in \{C_1, C_2\}^s$ C_1, C_2 are called *class labels*

Binary classification

When y can only take on two values, the classification is called *binary classification*

$y \in \{C_1, C_2\}^s$ C_1, C_2 are called *class labels*

Often the class labels are associated with numerical values, e.g.

$y \in \{-1, 1\}^s$ or $y \in \{0, 1\}^s$

Binary classification

When y can only take on two values, the classification is called *binary classification*

$y \in \{C_1, C_2\}^s$ C_1, C_2 are called *class labels*

Often the class labels are associated with numerical values, e.g.

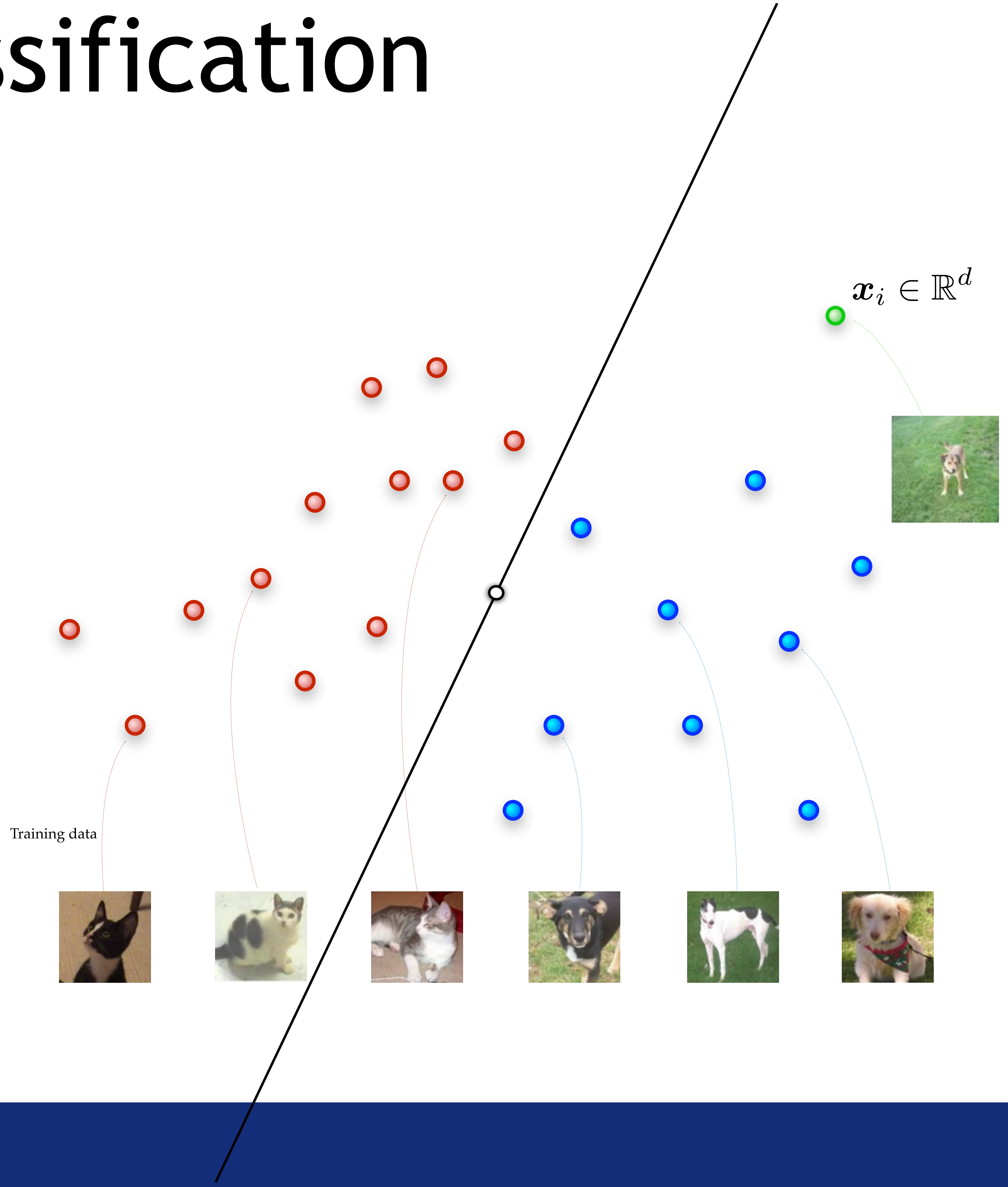
$y \in \{-1, 1\}^s$ or $y \in \{0, 1\}^s$

Note that even if the class labels take on numerical values, there is typically **no ordering** implied between the two classes

Binary classification

Examples:

image source



Binary classification

Examples: Surviving the titanic disaster

©Wikimedia commons

Either you have **survived or not survived** the sinking of the Titanic (assuming you were a passenger on the Titanic)

Binary classification

Examples: Train delays

©Evening Standard

Often formulated and treated as a regression problem

Binary classification

Examples: Train delays

©Evening Standard

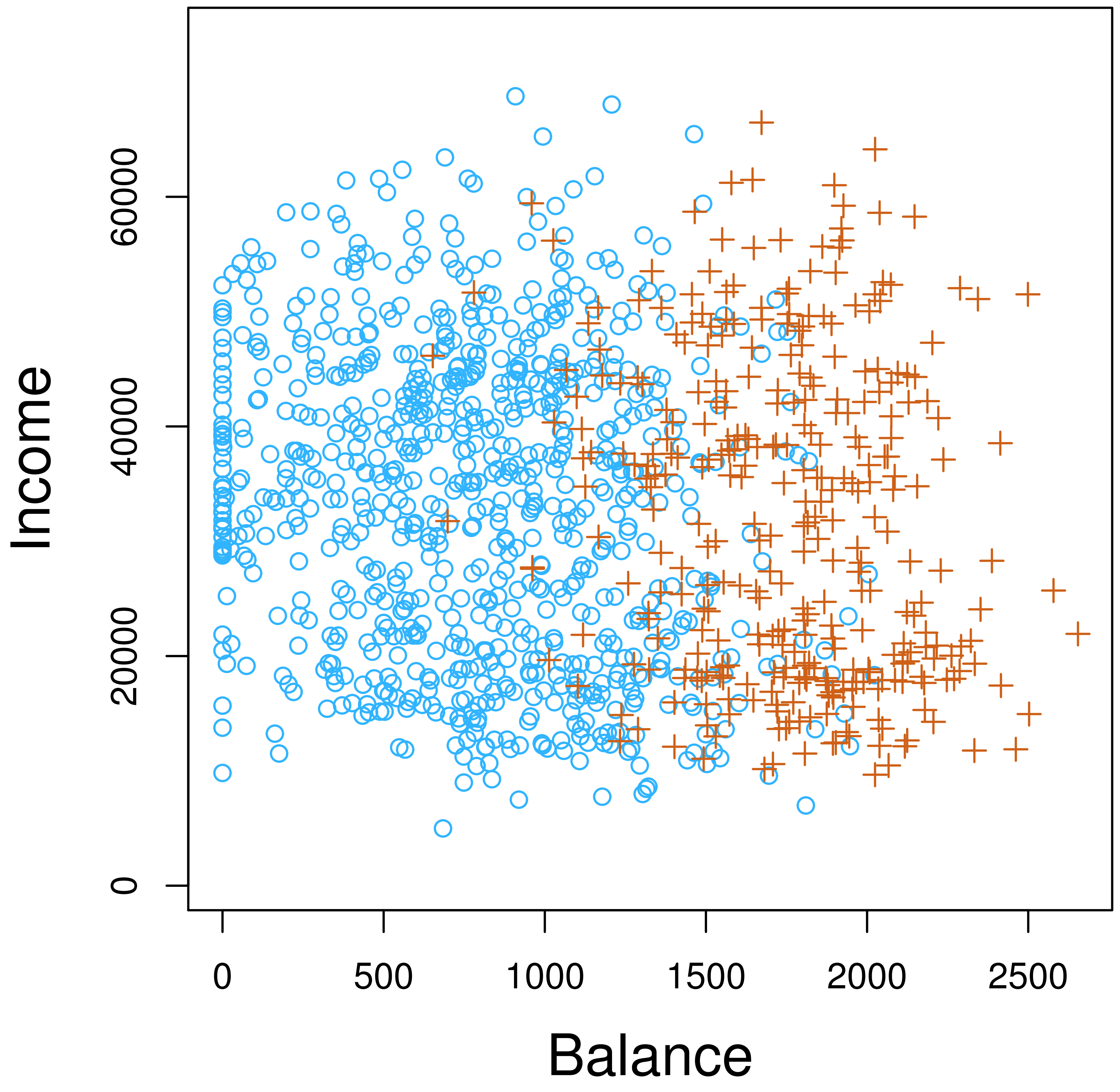
Often formulated and treated as a regression problem

Can also be considered a binary classification problem:

$$y_i = \begin{cases} C_1 & \text{train delay} < 30 \text{ minutes (no refund)} \\ C_2 & \text{train delay} \geq 30 \text{ minutes (refund)} \end{cases}$$

Binary classification

Examples: credit default



+ = individual who defaulted on their credit card payments

o = individual who did not default on their credit card payments

from *Elements of Statistical Learning* by Hastie, Tibshirani and Friedman

Binary classification

Examples: Targeted marketing

Classify customers as **likely buyers vs unlikely buyers** of product X

Image source

Binary classification

Examples: Targeted marketing

Classify customers as **likely buyers vs unlikely buyers** of product X

Image source

Be aware of ethical consequences, e.g. voter targeting

Multi-class classification

In multi-class classification, y can take on more than two values, i.e.

$$y \in \{C_0, C_1, \ldots, C_{K-1}\}^s$$

for the K class labels $C_0, \ldots, C_{K-1}$

Multi-class classification

In multi-class classification, y can take on more than two values, i.e.

$$y \in \{C_0, C_1, \ldots, C_{K-1}\}^s$$

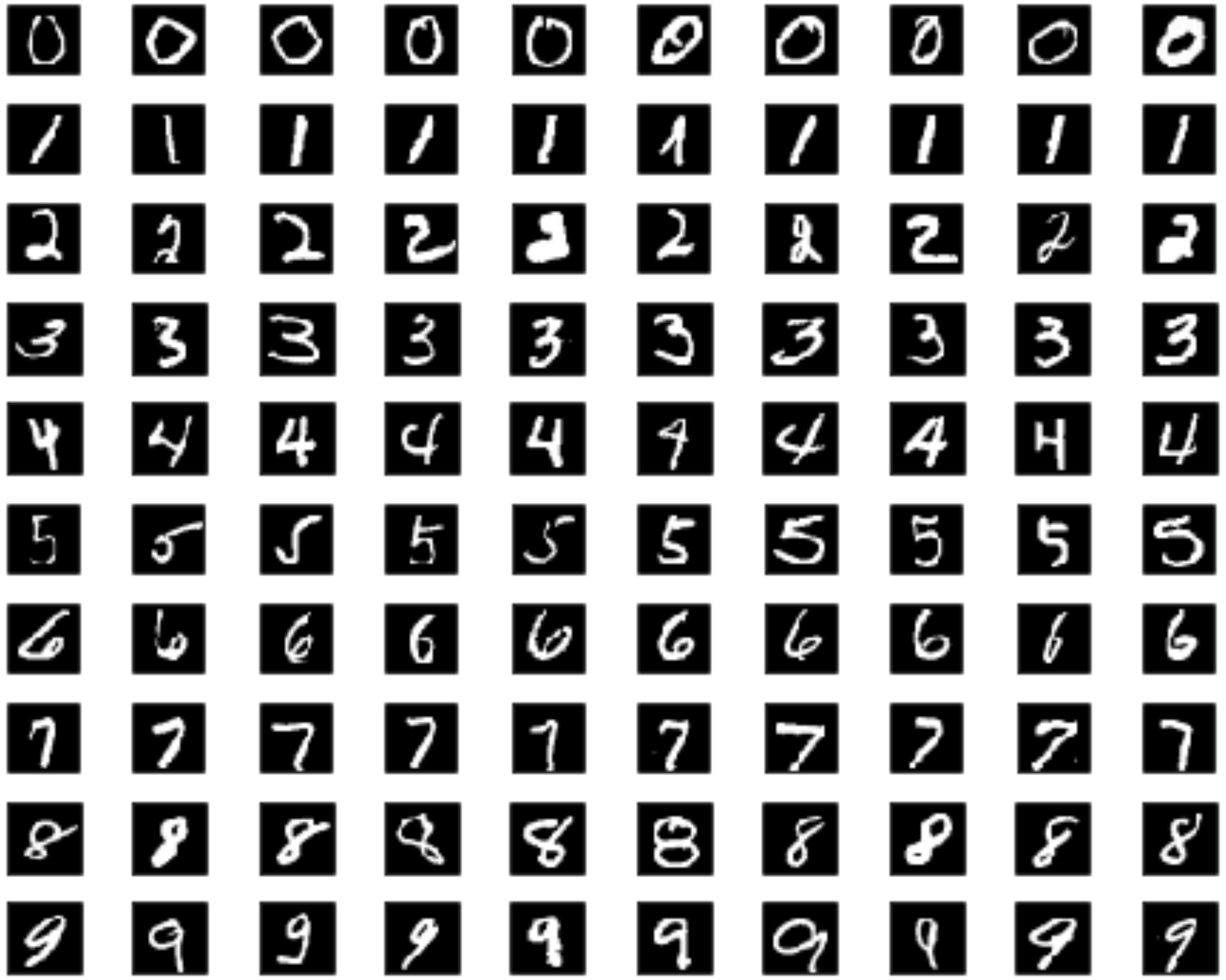
for the K class labels $C_0, \ldots, C_{K-1}$

Again, there is in general no ordering amongst the classes but often we will use numerical values as class labels, e.g.

$$y \in \{0,1,\ldots,K-1\}^s$$

Multi-class classification

Example: classification of hand-written digits



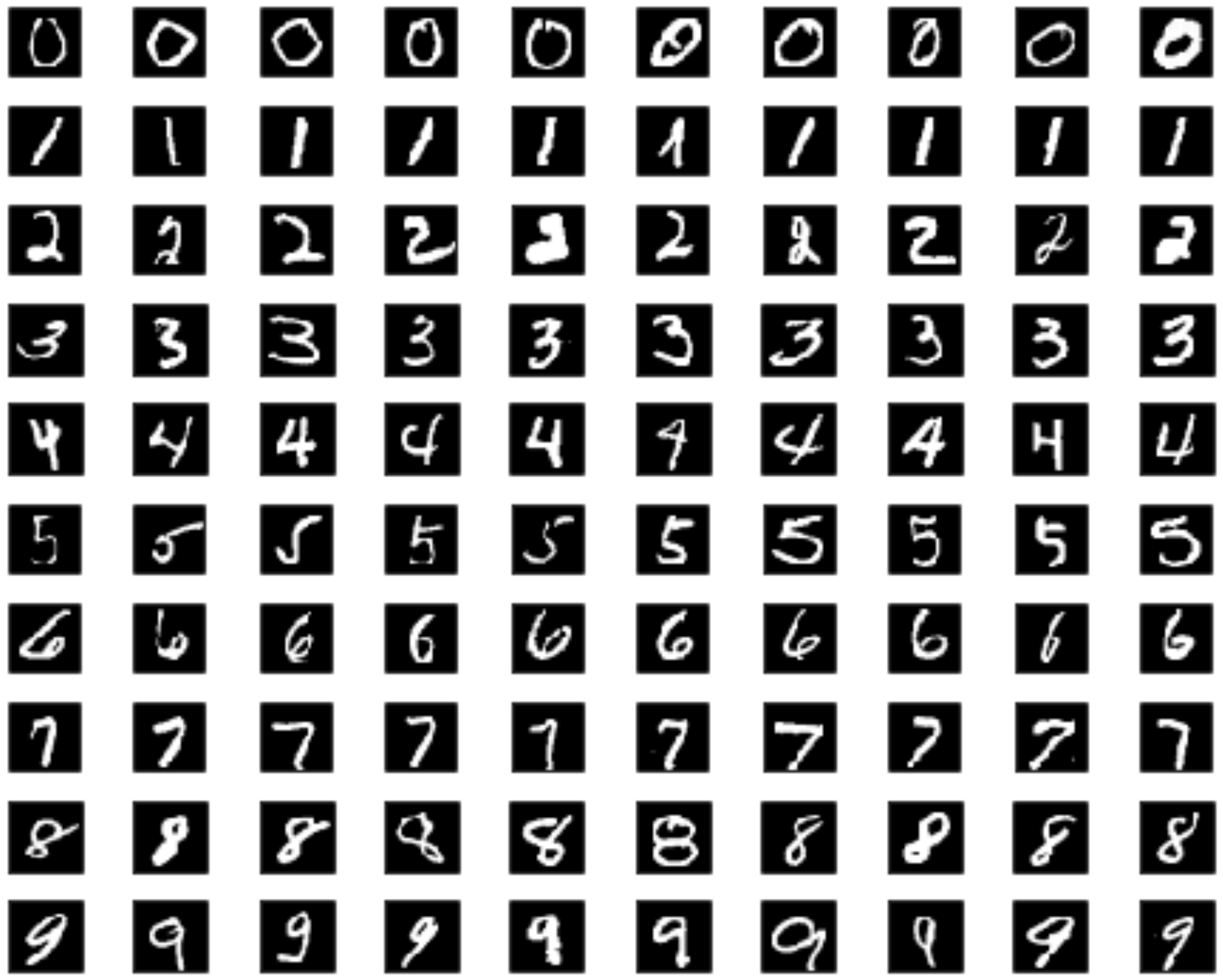
MNIST database

Multi-class classification

Example: classification of hand-written digits

Decide whether an image of a hand-written digit belongs to class

0,1,2,3,4,5,6,7,8 or 9



MNIST database

What is a classifier?

In order to classify, we need a *classifier*

What is a classifier?

In order to classify, we need a *classifier*

A classifier divides the input space into a collection of regions belonging to each class

What is a classifier?

In order to classify, we need a *classifier*

A classifier divides the input space into a collection of regions belonging to each class

The boundaries of these regions are called *decision boundaries*

What is a classifier?

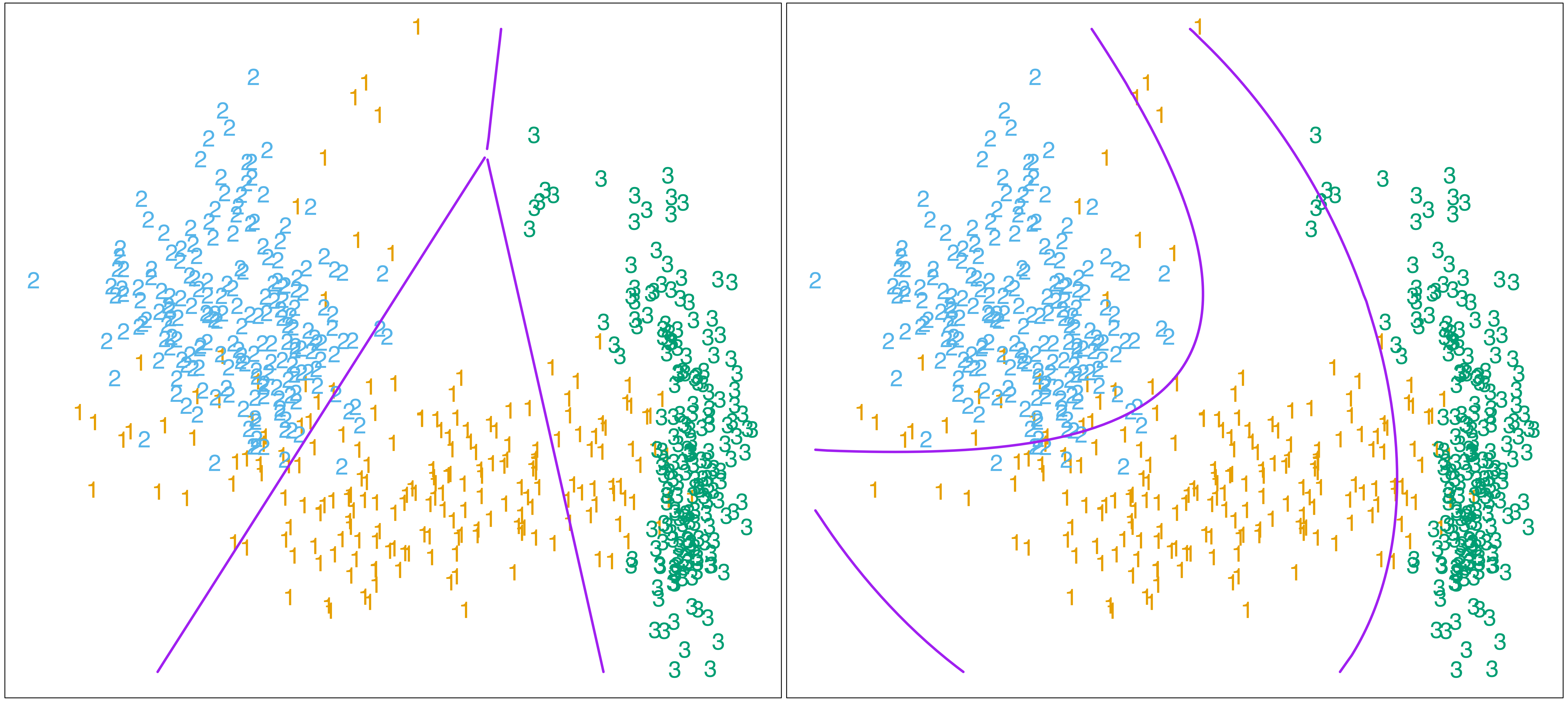
In order to classify, we need a *classifier*

A classifier divides the input space into a collection of regions belonging to each class

The boundaries of these regions are called *decision boundaries*

We distinguish between linear and nonlinear classifiers

What is a classifier?



from *Elements of Statistical Learning*
by Hastie, Tibshirani and Friedman

What is the aim of classification?

Classification itself: we are constructing a predictor based on a training set and are interested in applying the predictor to new data

What is the aim of classification?

Classification itself: we are constructing a predictor based on a training set and are interested in applying the predictor to new data

Example: Credit card company wants to predict if a customer is likely to default or not

What is the aim of classification?

Classification itself:	we are constructing a predictor based on a training set and are interested in applying the predictor to new data
Example:	Credit card company wants to predict if a customer is likely to default or not
Understanding the 'cause' of something:	we are interested in the interpretation of the prediction

What is the aim of classification?

Classification itself:	we are constructing a predictor based on a training set and are interested in applying the predictor to new data
Example:	Credit card company wants to predict if a customer is likely to default or not
Understanding the 'cause' of something:	we are interested in the interpretation of the prediction
Example:	Disease prediction. We do not only want to know if someone is at risk, but also why someone is at risk

What is the aim of classification?

Classification itself: we are constructing a predictor based on a training set and are interested in applying the predictor to new data

Example: Credit card company wants to predict if a customer is likely to default or not

Understanding the ‘cause’ of something: we are interested in the interpretation of the prediction

Example: Disease prediction. We do not only want to know if someone is at risk, but also why someone is at risk

For the second task it is often important to have ‘simple’ models

What is the aim of classification?

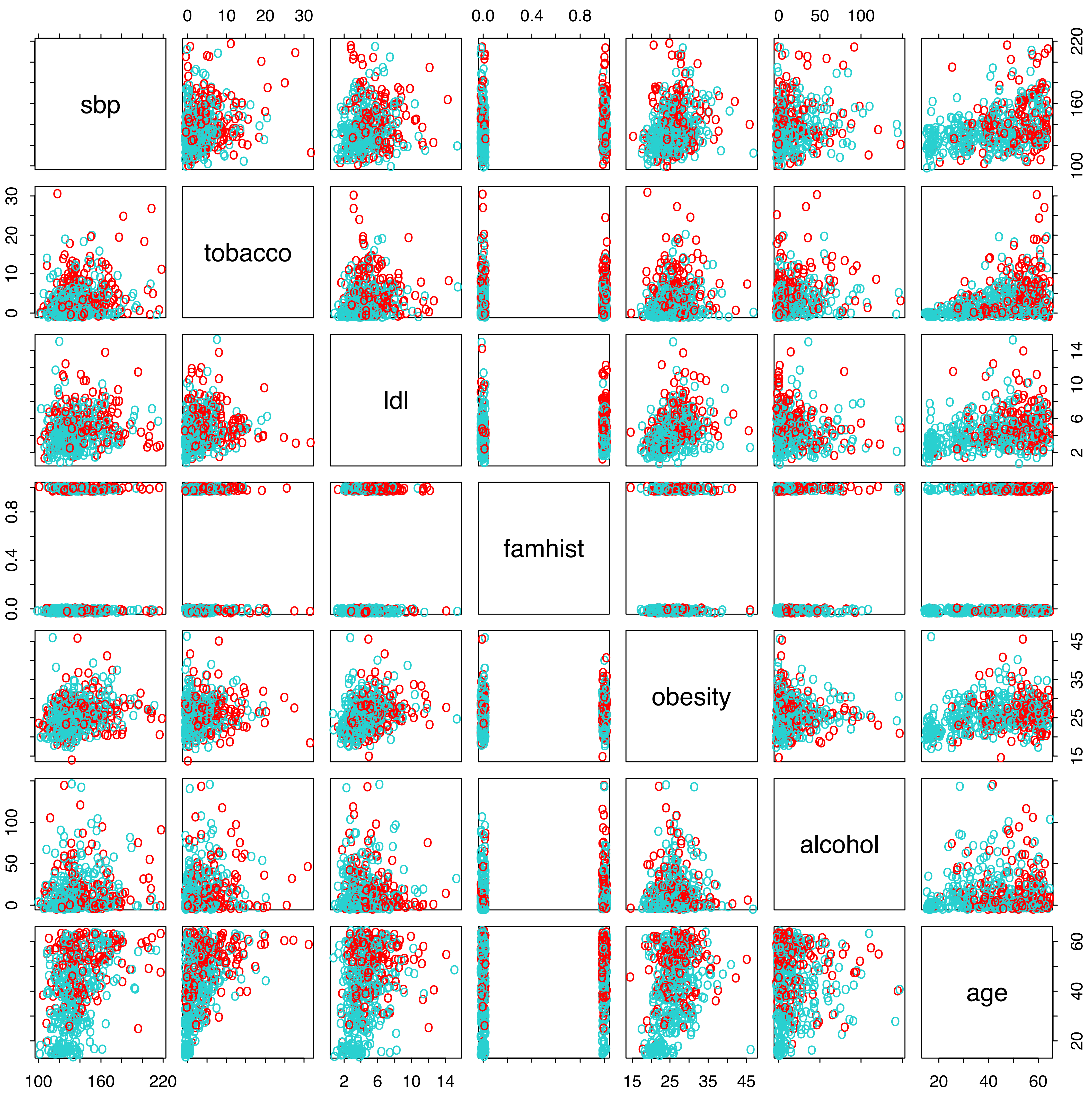
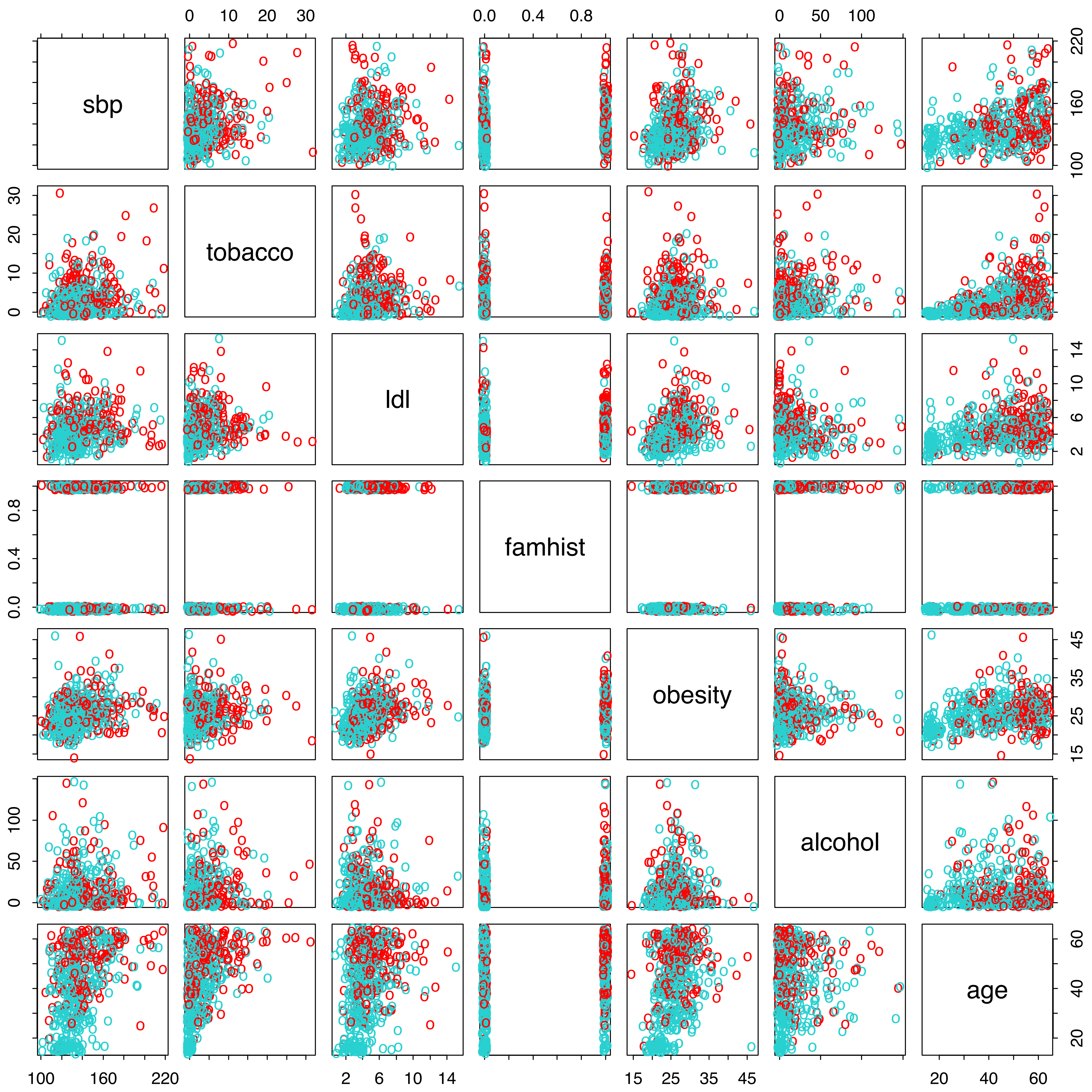


FIGURE 4.12. *A scatterplot matrix of the South African heart disease data. Each plot shows a pair of risk factors, and the cases and controls are color coded (red is a case). The variable* family history of heart disease (`famhist`) *is binary (yes or no).*

from *Elements of Statistical Learning*
by Hastie, Tibshirani and Friedman

What is the aim of classification?



Scatterplots like this one can help to decide which risk factors should be included in a model

from *Elements of Statistical Learning* by Hastie, Tibshirani and Friedman

Curse of dimensionality

Claim 1) "Generalising correctly becomes exponentially harder as the dimensionality grows because fixed-size training sets cover a dwindling fraction of the input space."

Curse of dimensionality

Claim 1) "Generalising correctly becomes exponentially harder as the dimensionality grows because fixed-size training sets cover a dwindling fraction of the input space."

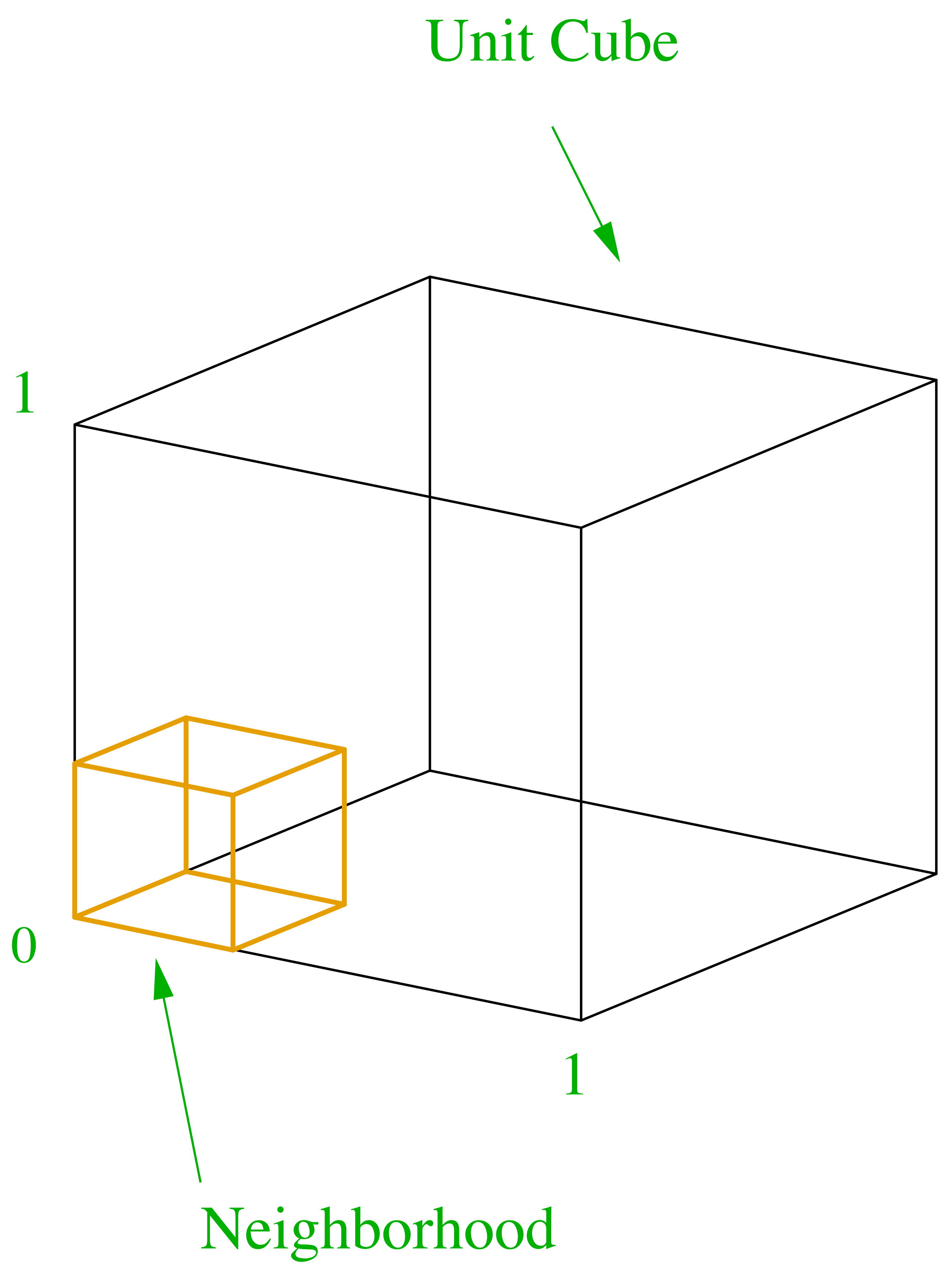
Imagine all points lie in d-dimensional unit cube $[0,1]^d$

Curse of dimensionality

Imagine all points lie in d-dimensional unit cube $[0,1]^d$

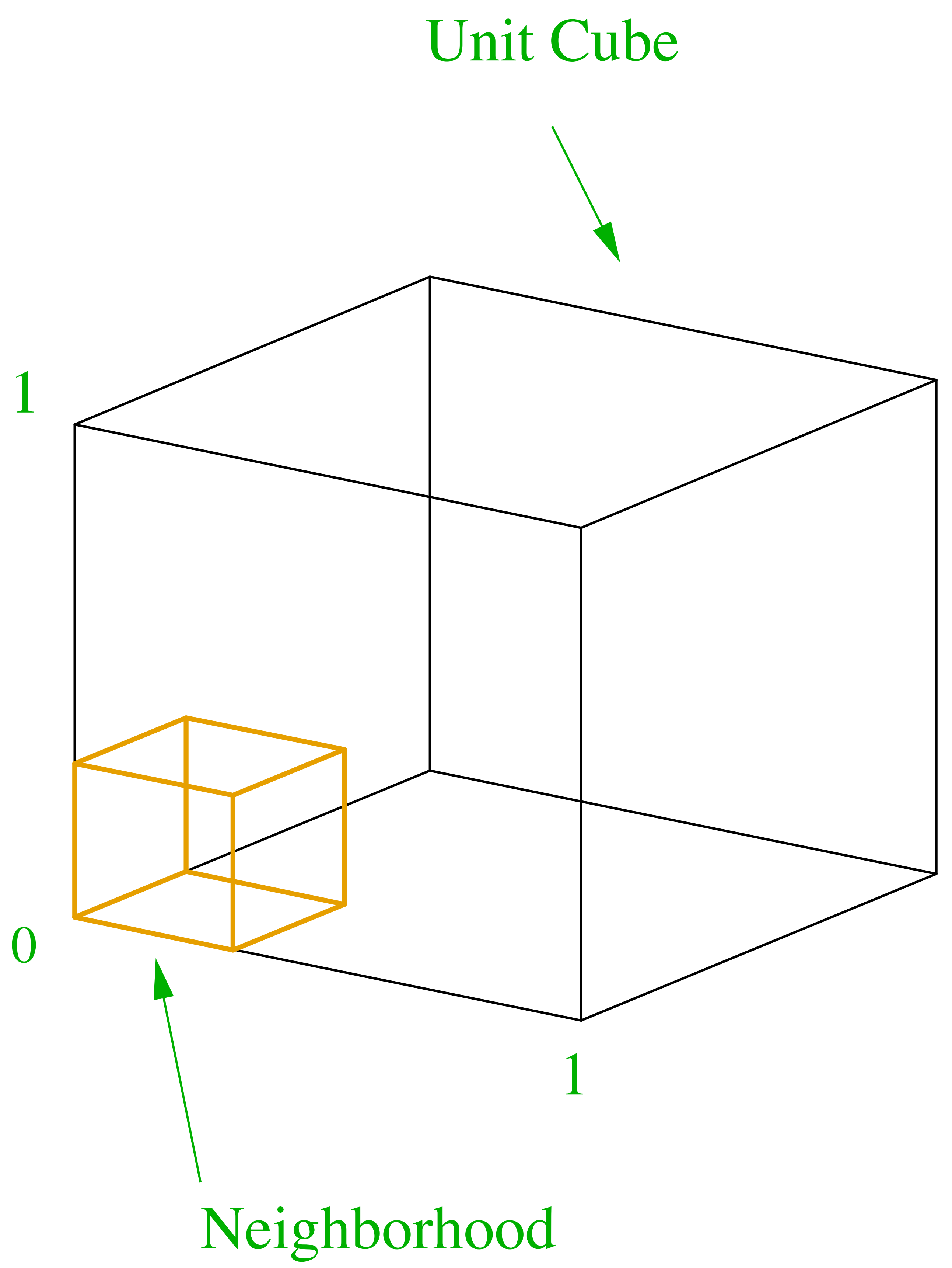
Curse of dimensionality

Imagine all points lie in d-dimensional unit cube $[0,1]^d$



Curse of dimensionality

Imagine all points lie in d-dimensional unit cube $[0,1]^d$

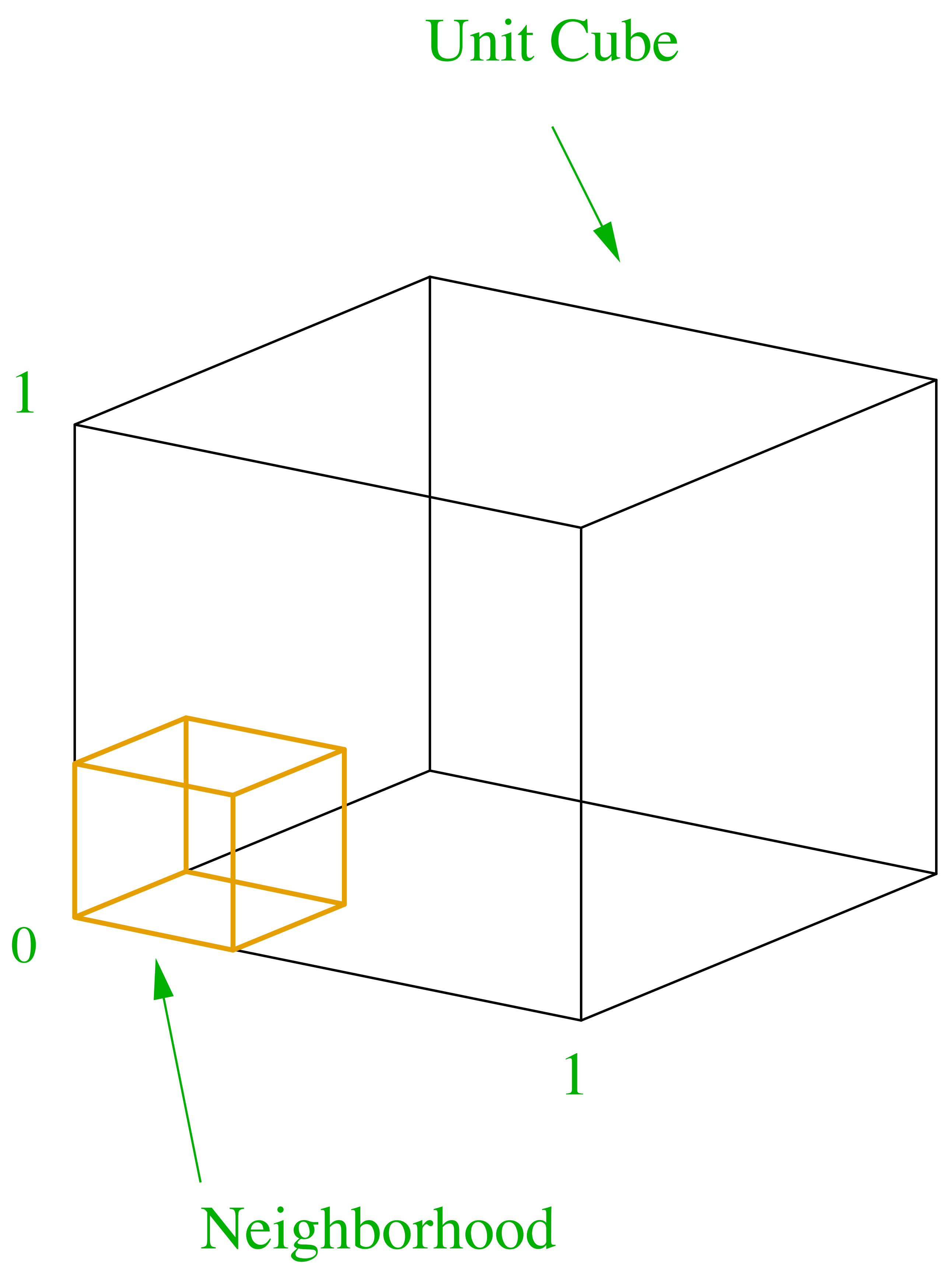


Consider sub-cube $[a, a+r]^d \subset [0,1]^d$ with

$0 \leq a$ and $a + r \leq 1$

Curse of dimensionality

Imagine all points lie in d-dimensional unit cube $[0,1]^d$



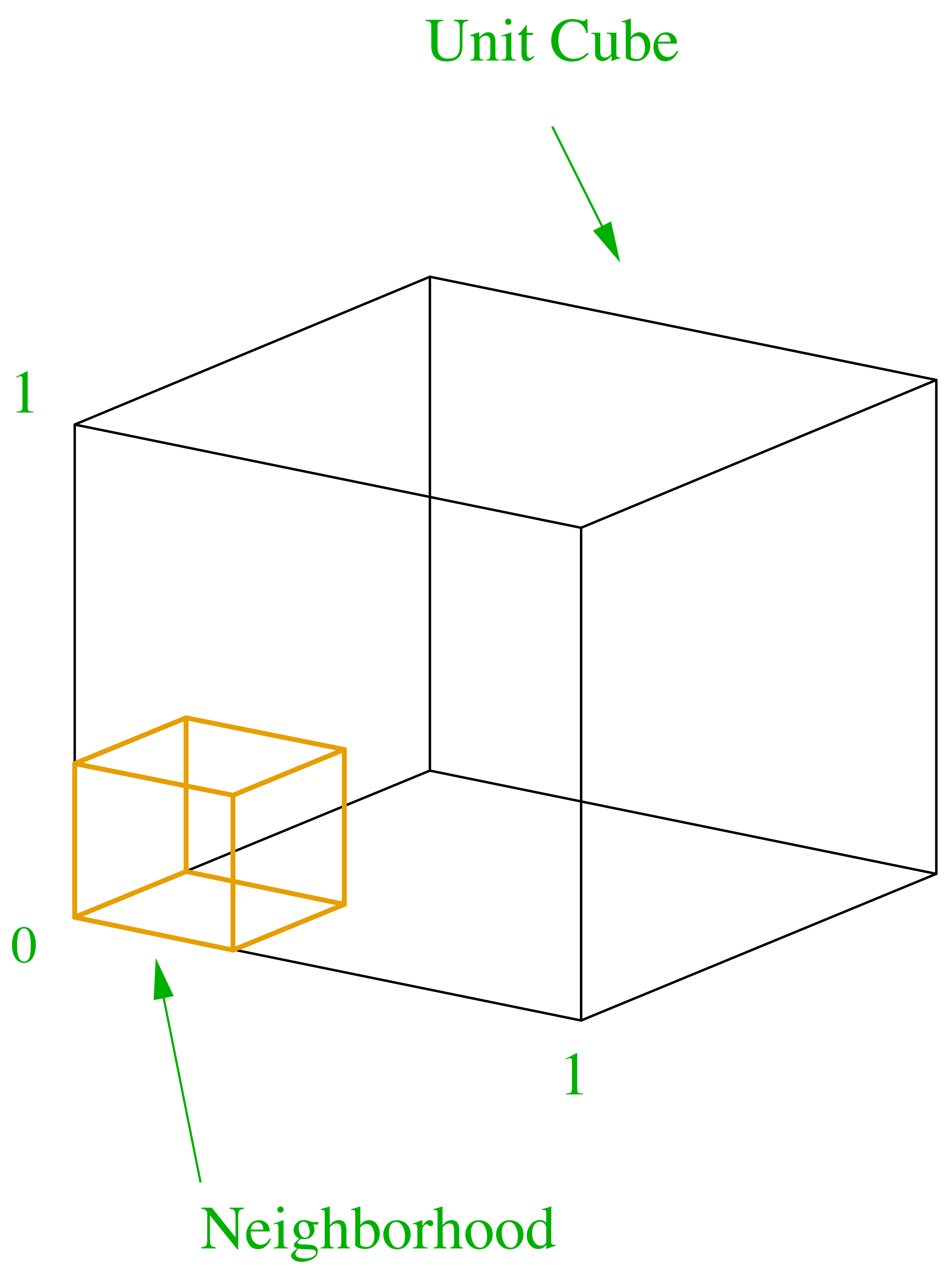
Consider sub-cube $[a, a+r]^d \subset [0,1]^d$ with

$0 \leq a$ and $a + r \leq 1$

What fraction of the total volume does this cube cover?

Curse of dimensionality

Imagine all points lie in d-dimensional unit cube $[0,1]^d$



Consider sub-cube $[a, a+r]^d \subset [0,1]^d$ with

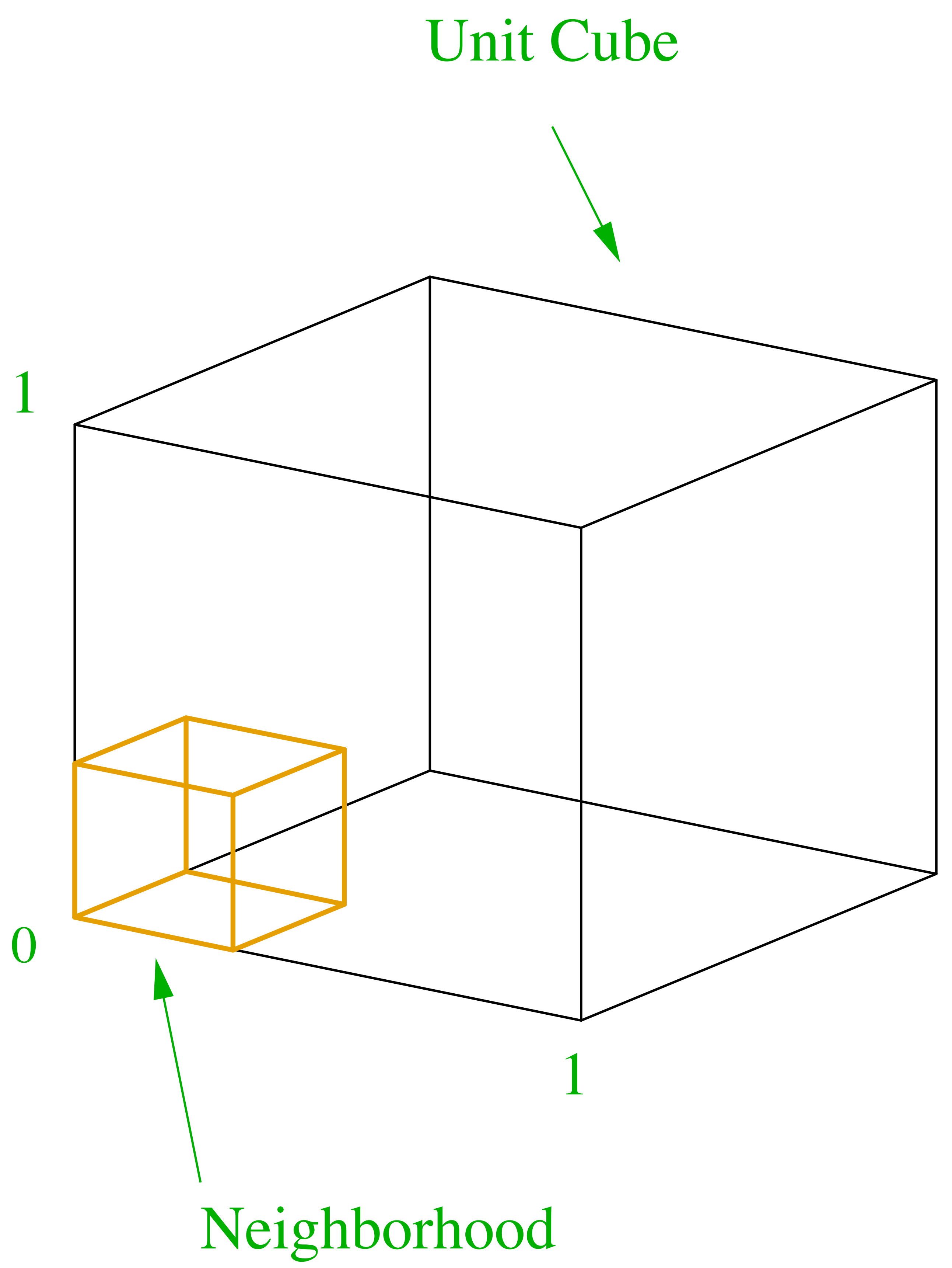
$0 \leq a$ and $a + r \leq 1$

What fraction of the total volume does this cube cover?

Answer: r^d

Curse of dimensionality

Imagine all points lie in d-dimensional unit cube $[0,1]^d$



Consider sub-cube $[a, a+r]^d \subset [0,1]^d$ with

$0 \leq a$ and $a + r \leq 1$

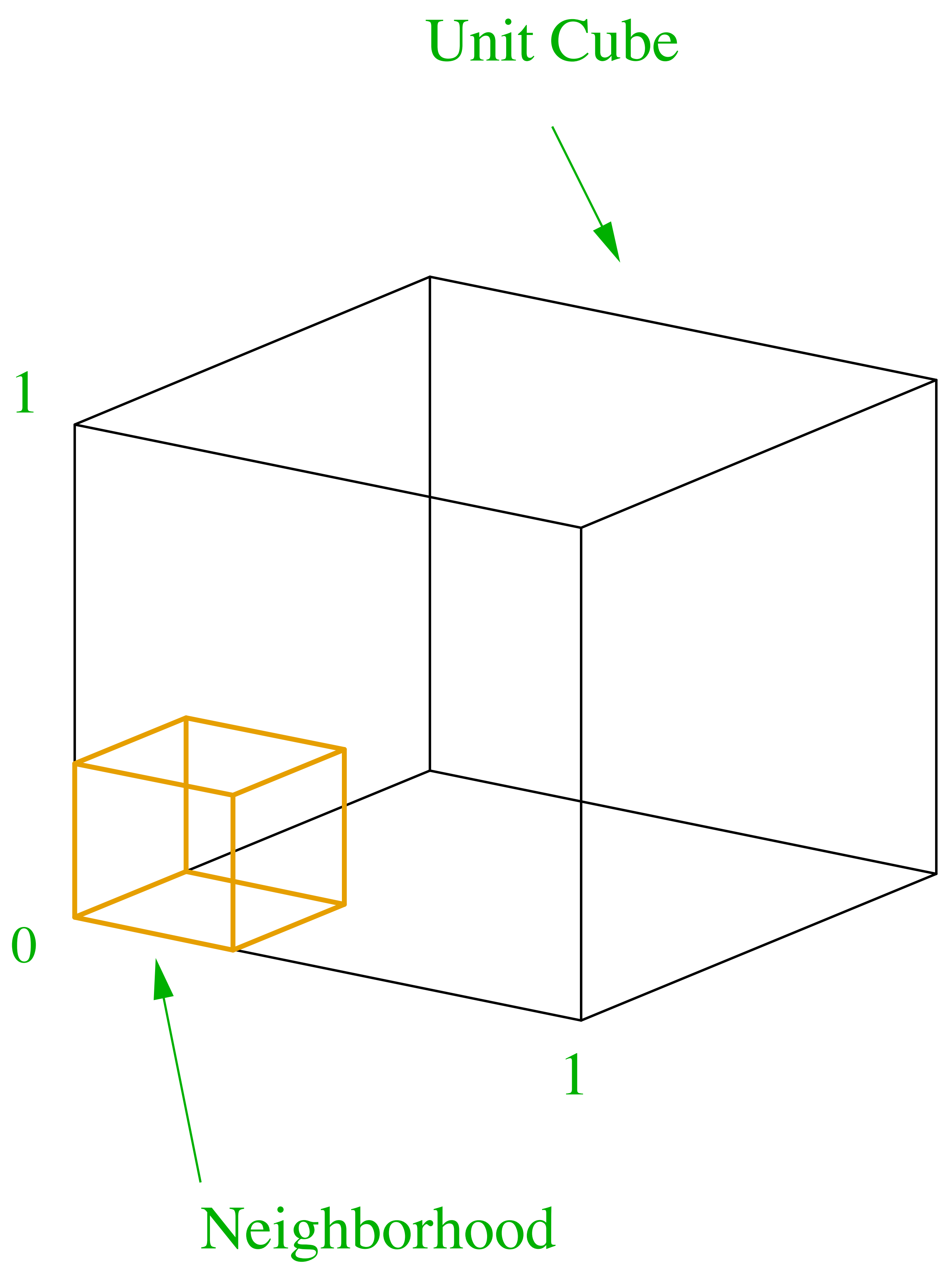
What fraction of the total volume does this cube cover?

Answer: r^d

Hence, in expectation a fraction $\alpha = r^d$ of the total data lies in this cube

Curse of dimensionality

Imagine all points lie in d-dimensional unit cube $[0,1]^d$



Consider sub-cube $[a, a+r]^d \subset [0,1]^d$ with

$0 \leq a$ and $a + r \leq 1$

What fraction of the total volume does this cube cover?

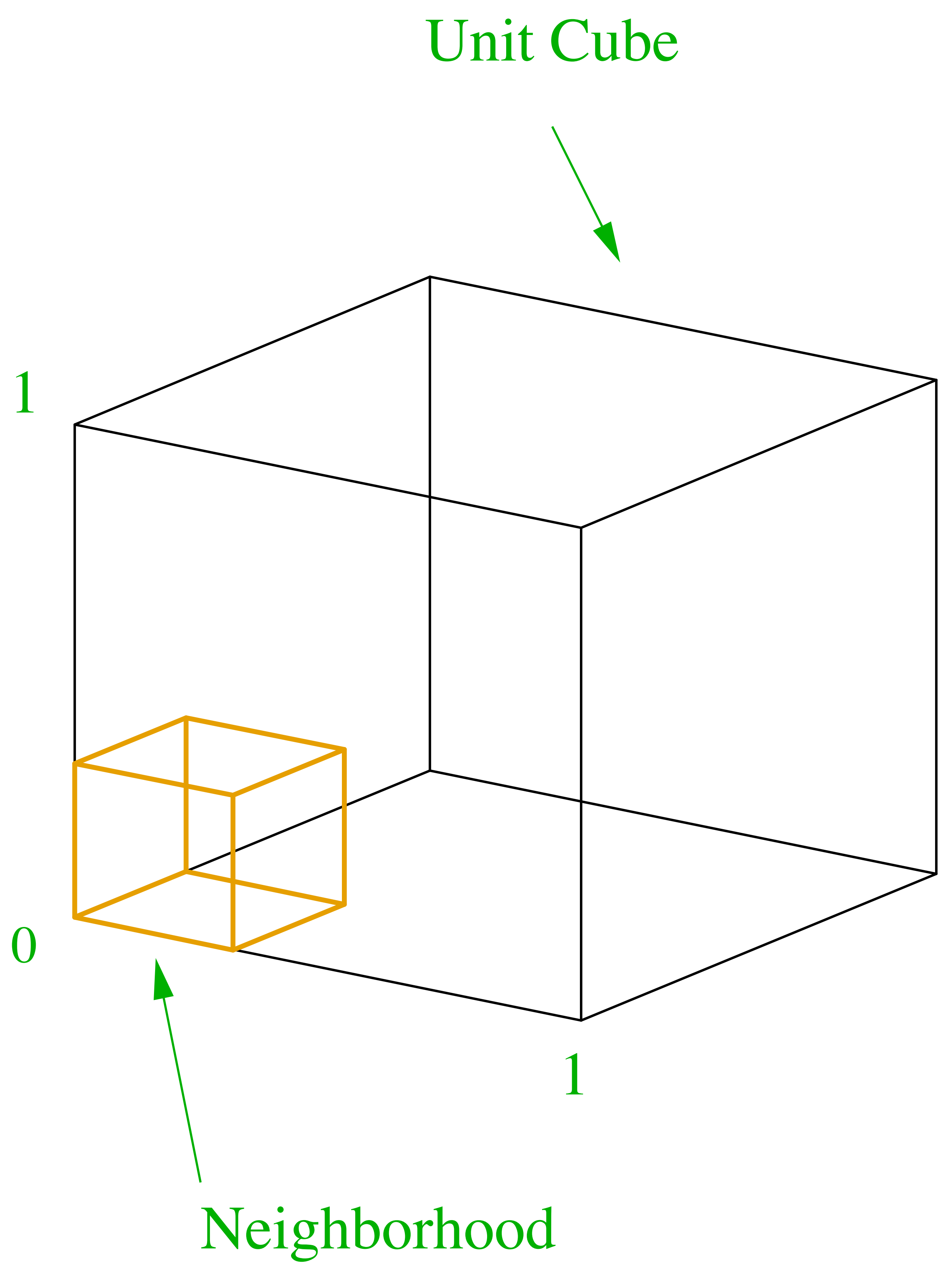
Answer: r^d

Hence, in expectation a fraction $\alpha = r^d$ of the total data lies in this cube

$\alpha = 1\,\%$
$d = 5$ $\Rightarrow$ $r \approx 0.4$

Curse of dimensionality

Imagine all points lie in d-dimensional unit cube $[0,1]^d$



Consider sub-cube $[a, a+r]^d \subset [0,1]^d$ with

$0 \leq a$ and $a + r \leq 1$

What fraction of the total volume does this cube cover?

Answer: r^d

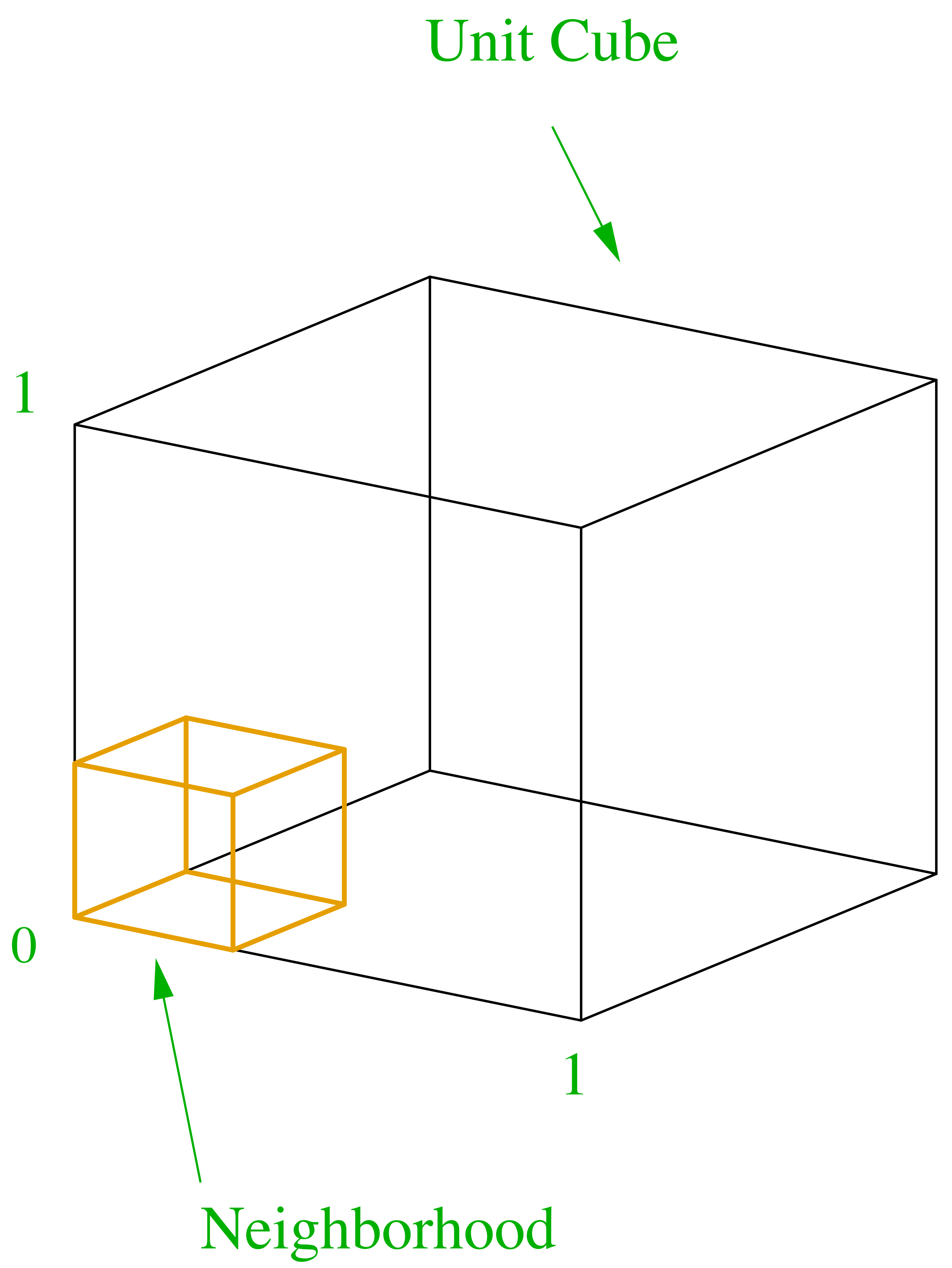
Hence, in expectation a fraction $\alpha = r^d$ of the total data lies in this cube

$\alpha = 1\,\%$, $d = 5$ $\Rightarrow$ $r \approx 0.4$

$\alpha = 1\,\%$, $d = 10$ $\Rightarrow$ $r \approx 0.63$

Curse of dimensionality

Imagine all points lie in d-dimensional unit cube $[0,1]^d$



Consider sub-cube $[a, a+r]^d \subset [0,1]^d$ with

$0 \leq a$ and $a + r \leq 1$

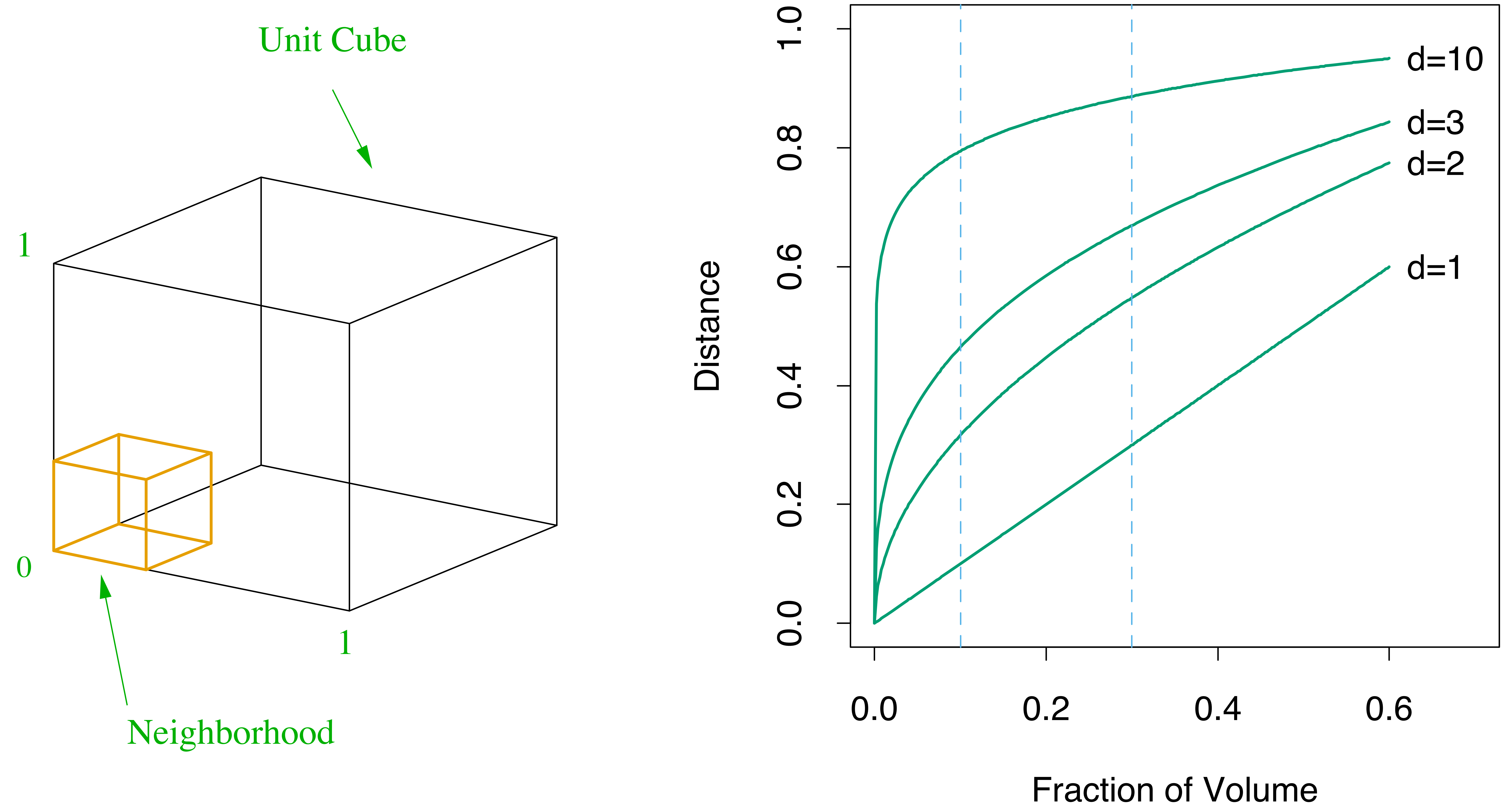
What fraction of the total volume does this cube cover?

Answer: r^d

Hence, in expectation a fraction $\alpha = r^d$ of the total data lies in this cube

$$\begin{aligned} \alpha &= 10\,\% \\ d &= 10 \end{aligned} \quad \Rightarrow \quad r \approx 0.8$$

Curse of dimensionality



Curse of dimensionality

Claim 2) In high-dimension, data-points are far from each other. Consequently, “as the dimensionality increases, the choice of nearest neighbour becomes effectively random.”

Curse of dimensionality

Claim 2) In high-dimension, data-points are far from each other. Consequently, “as the dimensionality increases, the choice of nearest neighbour becomes effectively random.”

Consider s data points uniformly distributed in the cube $[0,1]^d$

and a nearest neighbour estimate at point $(1/2, 1/2, \ldots, 1/2)$

Curse of dimensionality

Claim 2) In high-dimension, data-points are far from each other. Consequently, "as the dimensionality increases, the choice of nearest neighbour becomes effectively random."

Consider s data points uniformly distributed in the cube $[0,1]^d$ and a nearest neighbour estimate at point $(1/2,1/2,\ldots,1/2)$

For our sub-cube $[(1-r)/2,(1+r)/2]^d \subset [0,1]^d$ with $0 \leq r \leq 1$, what is the chance that a random sample is in $[0,1]^d$ but not in the sub-cube?

Curse of dimensionality

Claim 2) In high-dimension, data-points are far from each other. Consequently, "as the dimensionality increases, the choice of nearest neighbour becomes effectively random."

Consider s data points uniformly distributed in the cube $[0,1]^d$ and a nearest neighbour estimate at point $(1/2, 1/2, \ldots, 1/2)$

For our sub-cube $[(1-r)/2, (1+r)/2]^d \subset [0,1]^d$ with $0 \leq r \leq 1$, what is the chance that a random sample is in $[0,1]^d$ but not in the sub-cube?

$$1 - r^d$$

Curse of dimensionality

Claim 2) In high-dimension, data-points are far from each other. Consequently, “as the dimensionality increases, the choice of nearest neighbour becomes effectively random.”

Consider s data points uniformly distributed in the cube $[0,1]^d$ and a nearest neighbour estimate at point $(1/2,1/2,\ldots,1/2)$

For our sub-cube $[(1-r)/2,(1+r)/2]^d \subset [0,1]^d$ with $0 \leq r \leq 1$, what is the chance that all s i.i.d. random samples are in $[0,1]^d$ but not in the sub-cube?

$$\left(1 - r^d\right)^s$$

Curse of dimensionality

Claim 2) In high-dimension, data-points are far from each other. Consequently, "as the dimensionality increases, the choice of nearest neighbour becomes effectively random."

Consider s data points uniformly distributed in the cube $[0,1]^d$ and a nearest neighbour estimate at point $(1/2,1/2,\ldots,1/2)$

For our sub-cube $[(1-r)/2,(1+r)/2]^d \subset [0,1]^d$ with $0 \le r \le 1$, what is the chance that all s i.i.d. random samples are in $[0,1]^d$ but not in the sub-cube?

$$\left(1 - r^d\right)^s$$

How to choose r such that this probability is 1/2?

Curse of dimensionality

Claim 2) In high-dimension, data-points are far from each other. Consequently, “as the dimensionality increases, the choice of nearest neighbour becomes effectively random.”

Consider s data points uniformly distributed in the cube $[0,1]^d$

and a nearest neighbour estimate at point $(1/2, 1/2, \ldots, 1/2)$

$$r = \sqrt[d]{1 - \sqrt[s]{\frac{1}{2}}}$$

Curse of dimensionality

Claim 2) In high-dimension, data-points are far from each other. Consequently, “as the dimensionality increases, the choice of nearest neighbour becomes effectively random.”

Consider s data points uniformly distributed in the cube $[0,1]^d$

and a nearest neighbour estimate at point $(1/2, 1/2, \ldots, 1/2)$

$$r = \sqrt[d]{1 - \sqrt[s]{\frac{1}{2}}}$$

Example: $d = 10$, $s = 500$ $\Rightarrow$ $r \approx 0.52$

Nearest neighbour classification

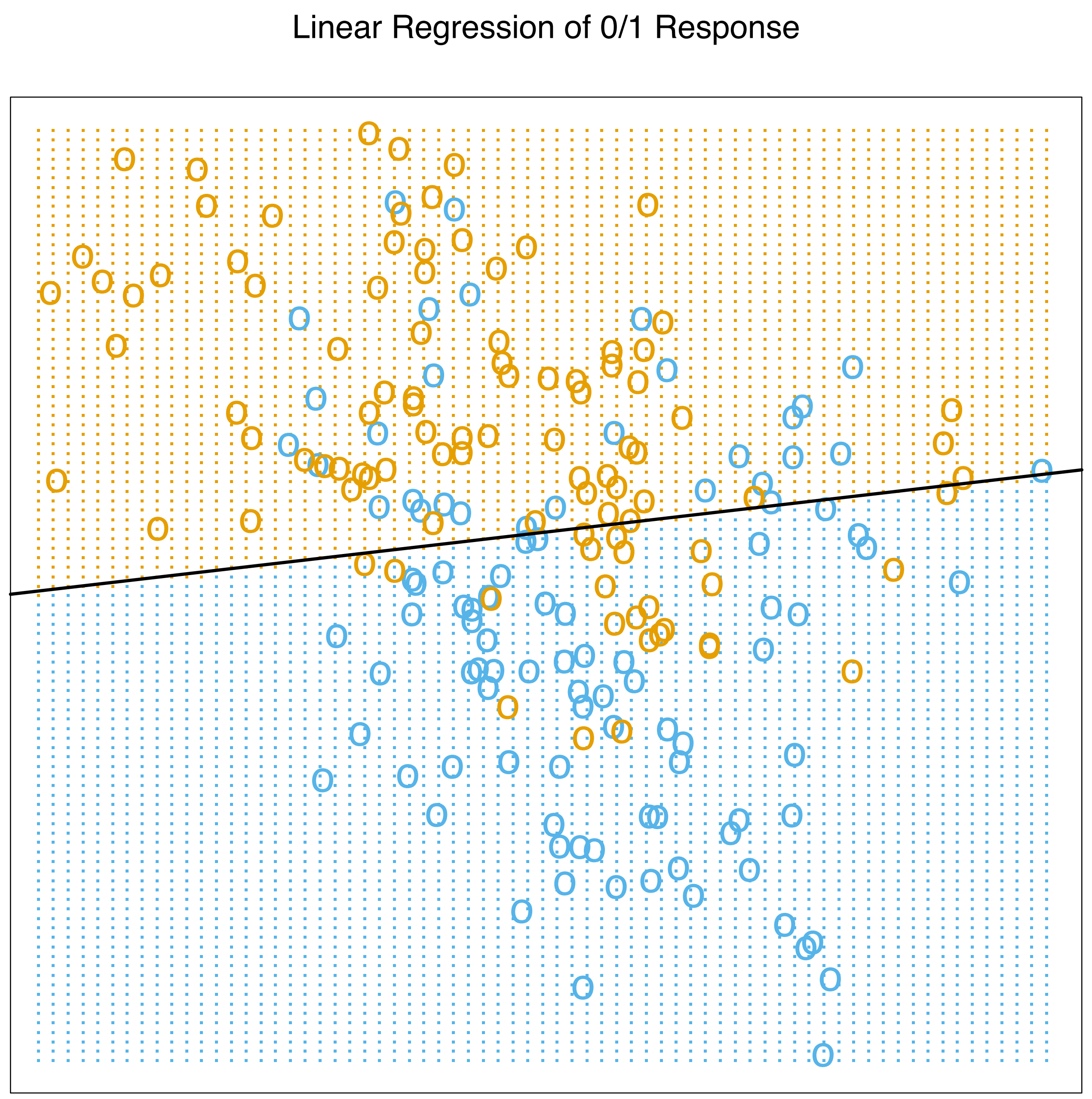


FIGURE 2.1. *A classification example in two dimensions. The classes are coded as a binary variable (*BLUE *$= 0$,* ORANGE *$= 1$), and then fit by linear regression. The line is the decision boundary defined by $x^T\hat{\beta} = 0.5$. The orange shaded region denotes that part of input space classified as* ORANGE*, while the blue region is classified as* BLUE.

Nearest neighbour classification

Suppose $\{(x_i, y_i)\}_{i=1}^{s}$ are points sampled from the distribution $\mathscr{D}$

Nearest neighbour classification

Suppose $\{(x_i, y_i)\}_{i=1}^{s}$ are points sampled from the distribution $\mathscr{D}$

For a new input x we determine the corresponding label y
by looking at the K points in the training set that are "nearest" to x:

Nearest neighbour classification

Suppose $\{(x_i, y_i)\}_{i=1}^{s}$ are points sampled from the distribution $\mathcal{D}$

For a new input x we determine the corresponding label y
by looking at the K points in the training set that are “nearest” to x:

$$p(y = c \,|\, x, \mathcal{D}, K) = \frac{1}{K} \sum_{i \in N_K(x, \mathcal{D})} \iota(y_i = c)$$

with $\iota(z) := \begin{cases} 1 & \text{if } z \text{ is true} \\ 0 & \text{if } z \text{ is false} \end{cases}$

Nearest neighbour classification

Suppose $\{(x_i, y_i)\}_{i=1}^{s}$ are points sampled from the distribution $\mathscr{D}$

For a new input x we determine the corresponding label y
by looking at the K points in the training set that are "nearest" to x:

$$p(y = c \,|\, x, \mathscr{D}, K) = \frac{1}{K} \sum_{i \in N_K(x, \mathscr{D})} \iota(y_i = c)$$

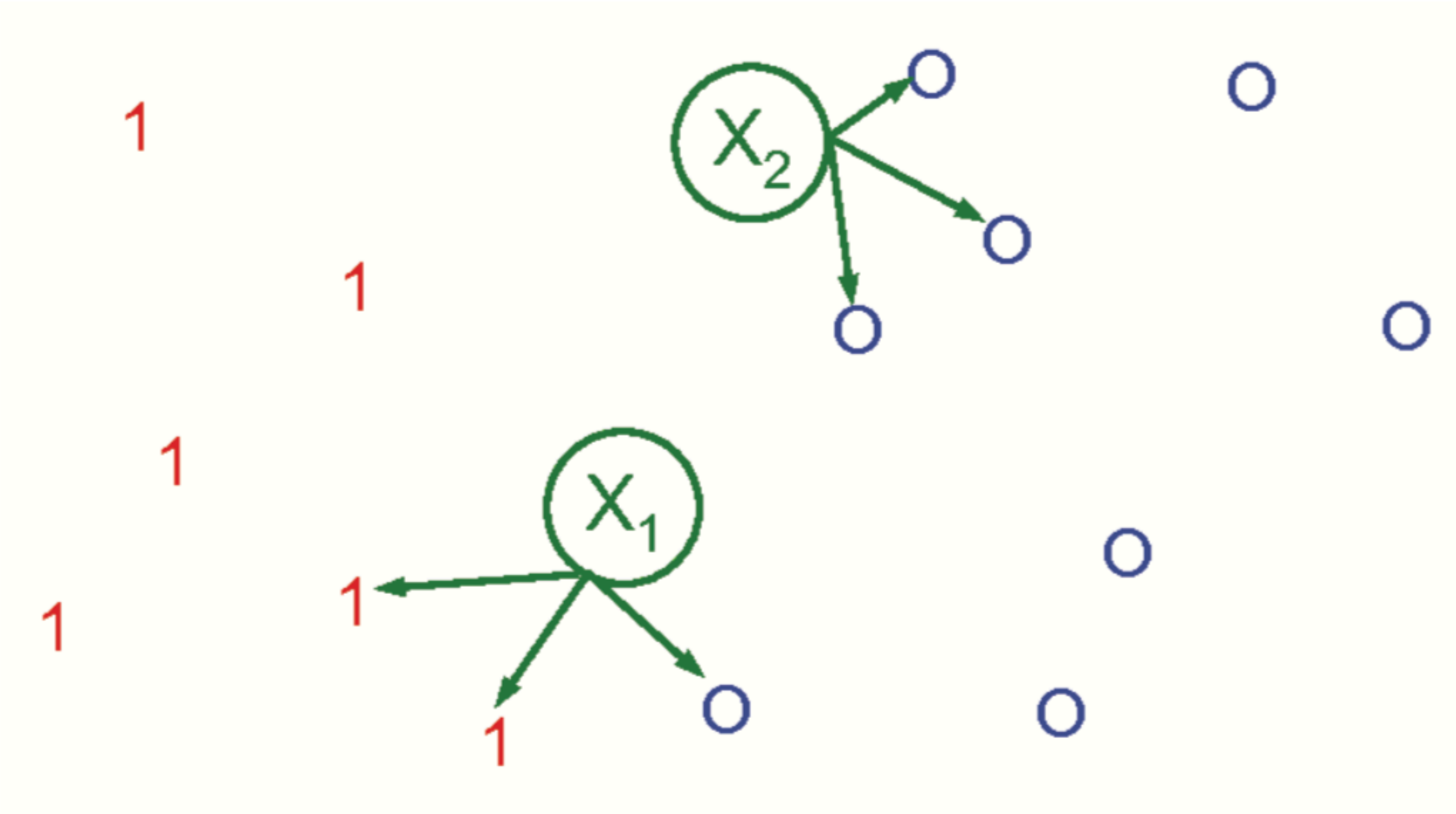
with $\iota(z) := \begin{cases} 1 & \text{if } z \text{ is true} \\ 0 & \text{if } z \text{ is false} \end{cases}$

Subsequently we set $f(x) = \arg\max_{c} p(y = c \,|\, x, \mathscr{D}, K)$

Nearest neighbour classification

Example:

$$N_K(x, \mathcal{D}) = \text{ indices of } K \text{ nearest points to } x \text{ in } \mathcal{D}$$

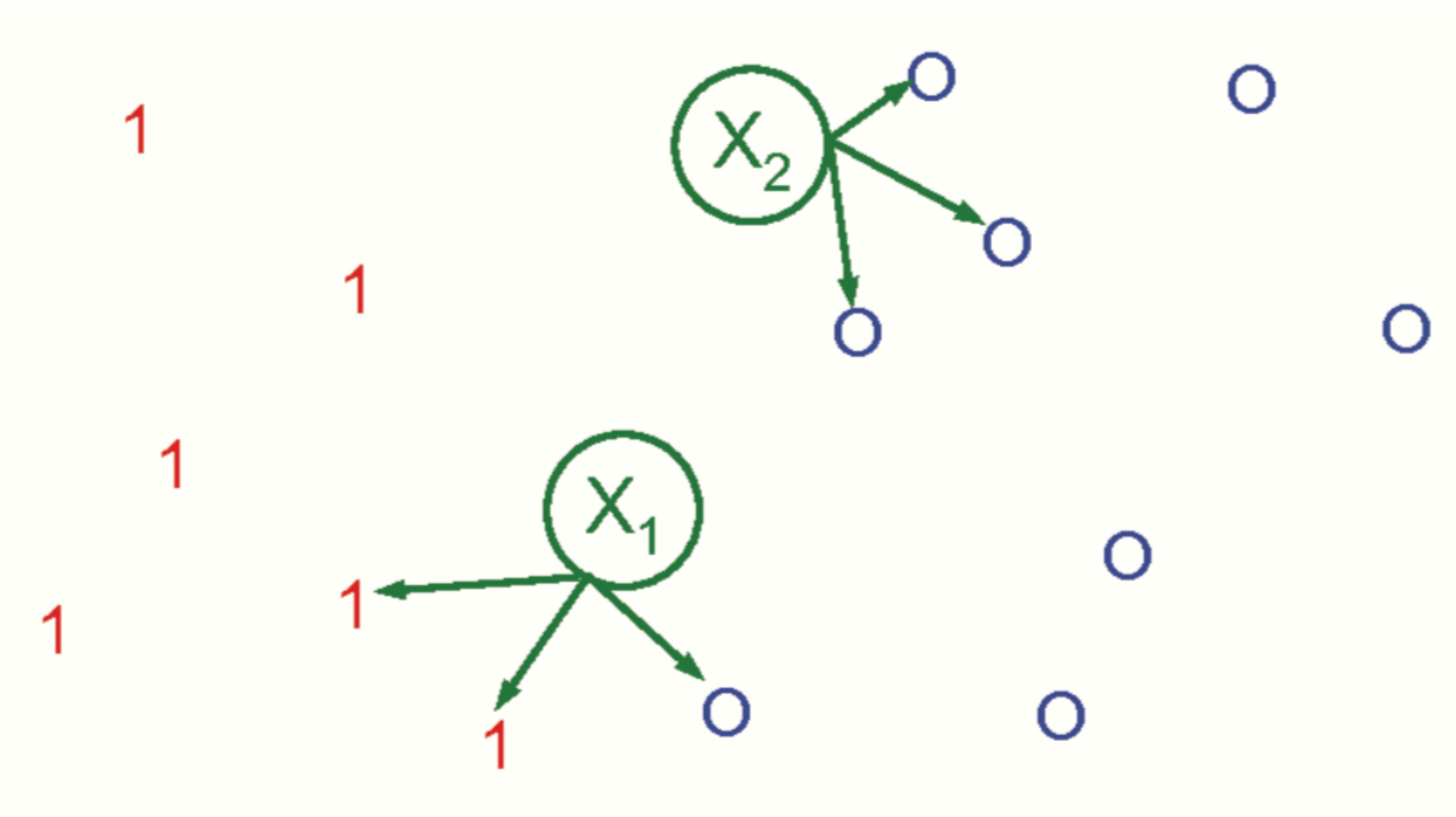


Kevin Murphy. *Machine Learning*

Nearest neighbour classification

Example:

$N_K(x, \mathcal{D}) =$ indices of K nearest points to x in $\mathcal{D}$



$p(y = 1 \,|\, x_1, \mathcal{D}, K = 3) = 2/3$

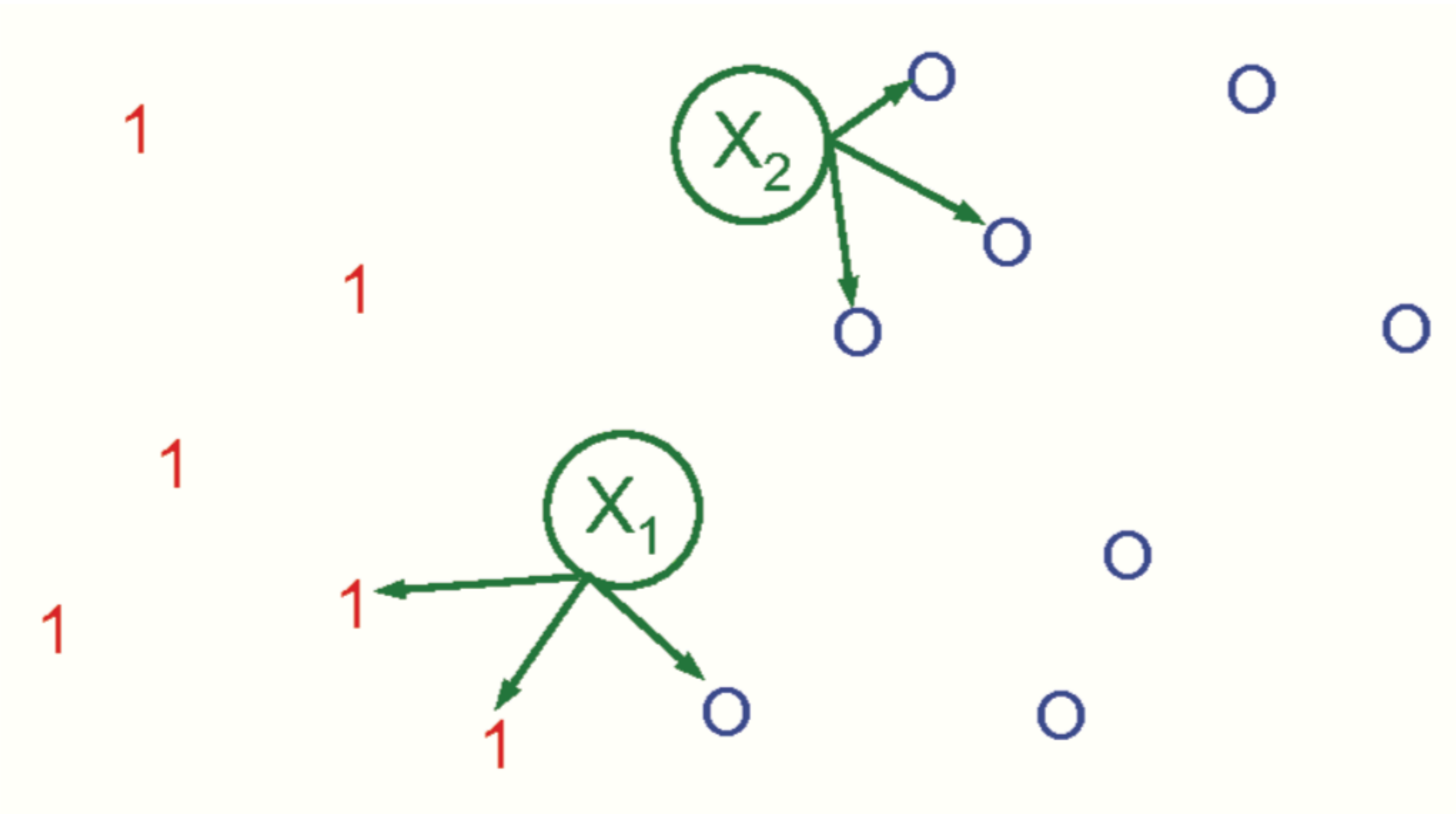
$p(y = 0 \,|\, x_1, \mathcal{D}, K = 3) = 1/3$

Kevin Murphy. *Machine Learning*

Nearest neighbour classification

Example:

$$N_K(x, \mathscr{D}) = \text{ indices of } K \text{ nearest points to } x \text{ in } \mathscr{D}$$



Kevin Murphy. *Machine Learning*

$$p(y = 1 \,|\, x_1, \mathscr{D}, K = 3) = 2/3$$

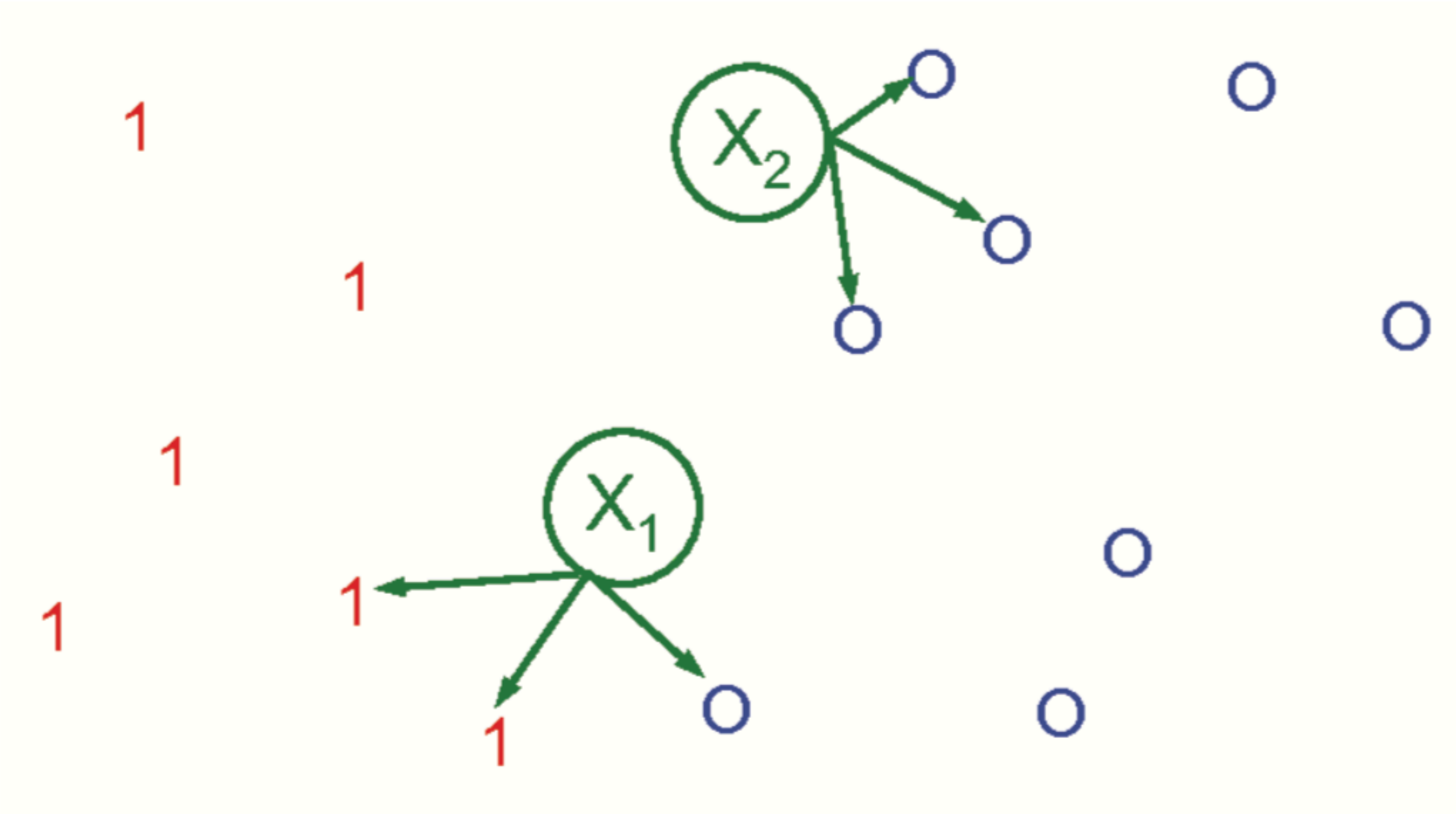
$$p(y = 0 \,|\, x_1, \mathscr{D}, K = 3) = 1/3$$

$$\Rightarrow \qquad f(x_1) = 1$$

Nearest neighbour classification

Example:

$N_K(x, \mathcal{D}) =$ indices of K nearest points to x in $\mathcal{D}$



Kevin Murphy. *Machine Learning*

$$p(y = 1 \mid x_1, \mathcal{D}, K = 3) = 2/3$$

$$p(y = 0 \mid x_1, \mathcal{D}, K = 3) = 1/3$$

$$\Rightarrow \qquad f(x_1) = 1$$

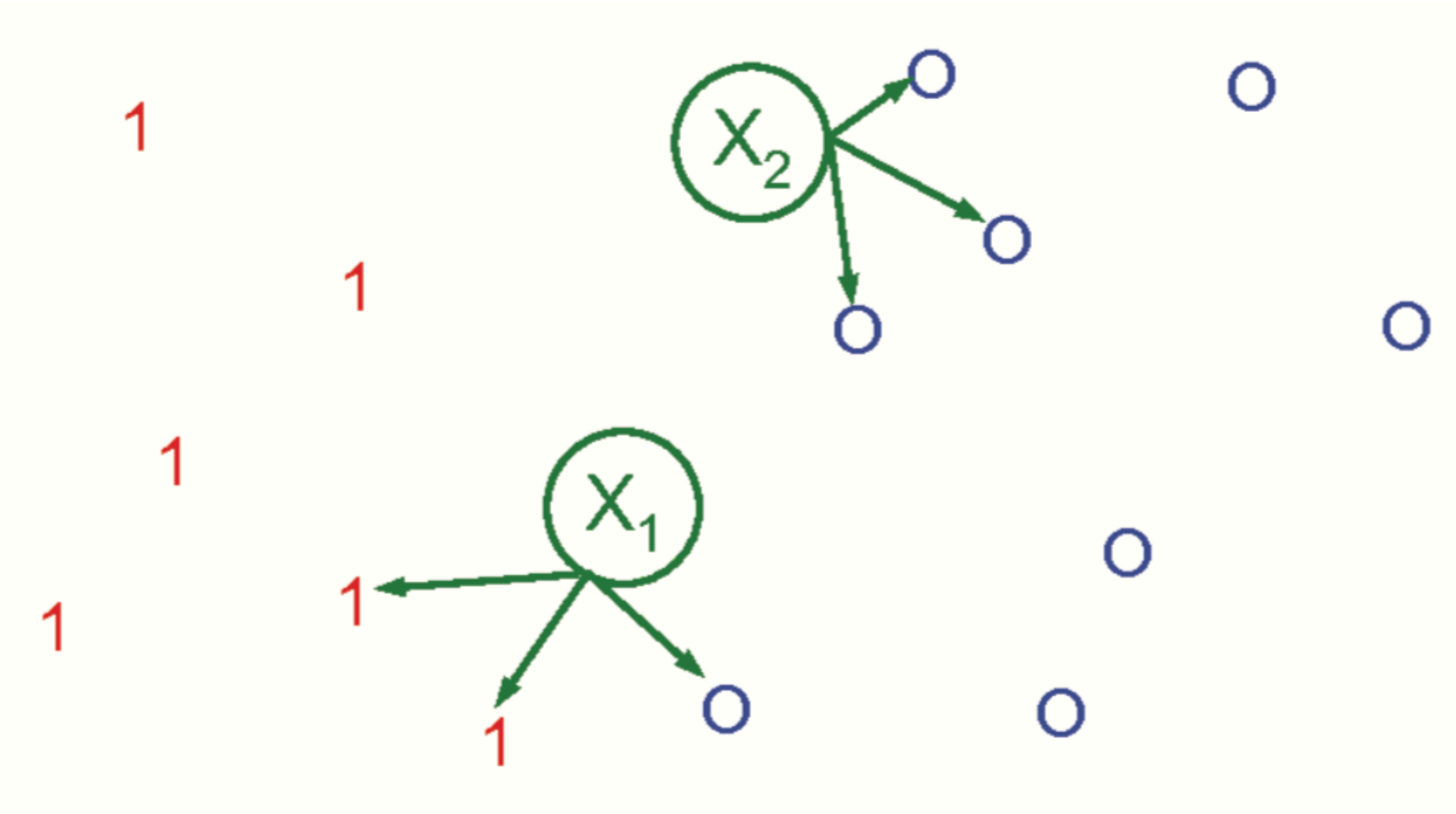
$$p(y = 1 \mid x_2, \mathcal{D}, K = 3) = 0$$

$$p(y = 0 \mid x_2, \mathcal{D}, K = 3) = 1$$

Nearest neighbour classification

Example:

$N_K(x, \mathcal{D}) =$ indices of K nearest points to x in $\mathcal{D}$



Kevin Murphy. *Machine Learning*

$$p(y=1 \,|\, x_1, \mathcal{D}, K=3) = 2/3$$

$$p(y=0 \,|\, x_1, \mathcal{D}, K=3) = 1/3$$

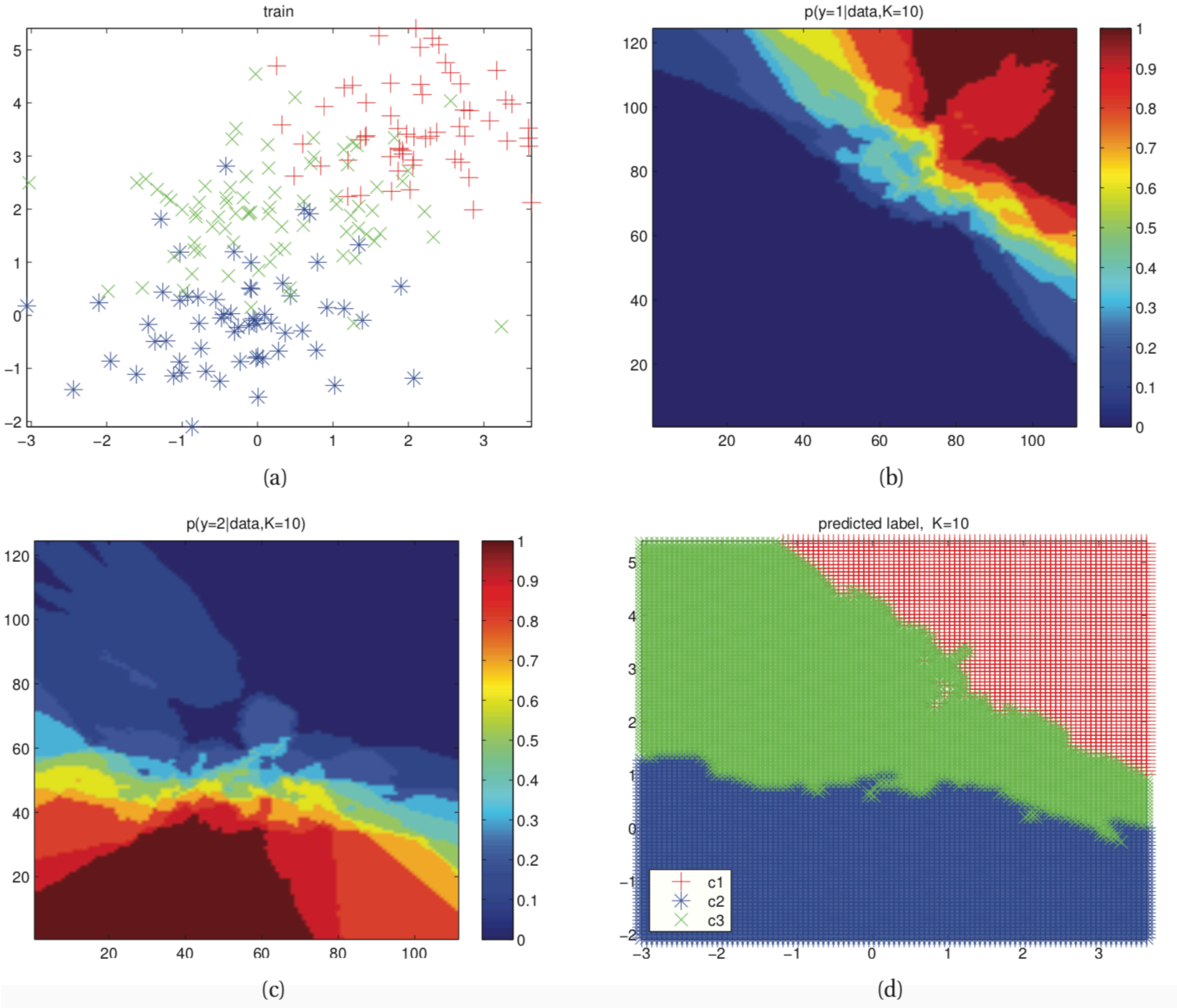
$$\Rightarrow \qquad f(x_1) = 1$$

$$p(y=1 \,|\, x_2, \mathcal{D}, K=3) = 0$$

$$p(y=0 \,|\, x_2, \mathcal{D}, K=3) = 1$$

$$\Rightarrow \qquad f(x_2) = 0$$

Nearest neighbour classification



Kevin Murphy.
Machine Learning

Nearest neighbour classification

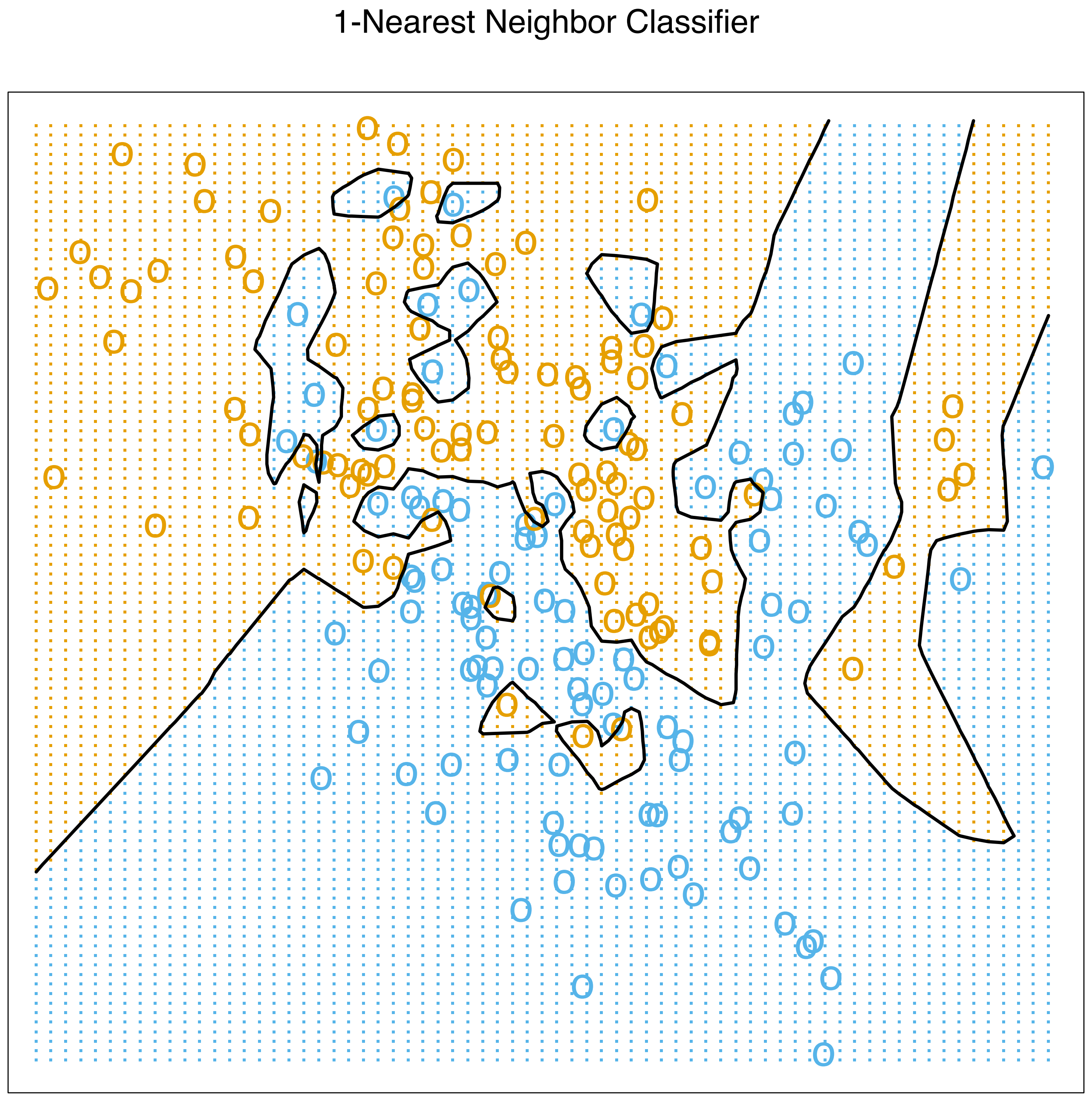


FIGURE 2.3. *The same classification example in two dimensions as in Figure 2.1. The classes are coded as a binary variable* (BLUE $= 0$, ORANGE $= 1$)*, and then predicted by 1-nearest-neighbor classification.*

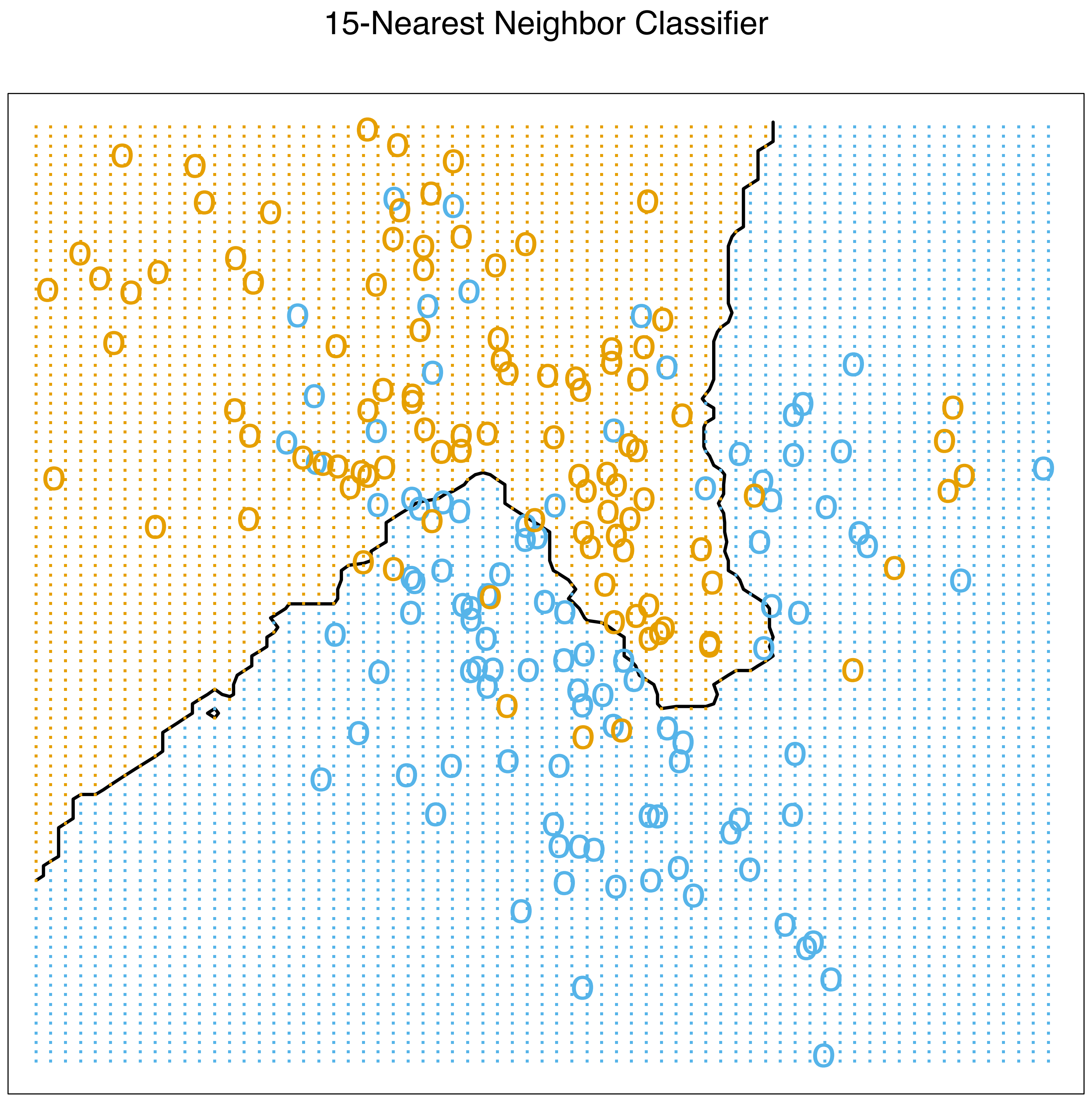


FIGURE 2.2. *The same classification example in two dimensions as in Figure 2.1. The classes are coded as a binary variable* (BLUE $= 0$, ORANGE $= 1$) *and then fit by 15-nearest-neighbor averaging as in (2.8). The predicted class is hence chosen by majority vote amongst the 15-nearest neighbors.*

Classification as a special case of regression

Solve $\hat{w} = \arg\min_{w} \left\{ \frac{1}{2s} \sum_{i=1}^{s} |f(x_i, w) - y_i|^2 + \frac{\alpha}{2} \|w\|^2 \right\}$ for $y_i = \begin{cases} 0 & \text{for class label } C_1 \\ 1 & \text{for class label } C_2 \end{cases}$

Classification as a special case of regression

Solve $\hat{w} = \arg\min_{w} \left\{ \frac{1}{2s} \sum_{i=1}^{s} |f(x_i, w) - y_i|^2 + \frac{\alpha}{2} \|w\|^2 \right\}$ for $y_i = \begin{cases} 0 & \text{for class label } C_1 \\ 1 & \text{for class label } C_2 \end{cases}$

Then it is natural to decide

$$f(x_i, \hat{w}) < \frac{1}{2} \quad \Rightarrow \quad \text{The predicted output is in class with label } C_1$$

$$f(x_i, \hat{w}) > \frac{1}{2} \quad \Rightarrow \quad \text{The predicted output is in class with label } C_2$$

Classification as a special case of regression

Solve $\hat{w} = \arg\min_{w} \left\{ \frac{1}{2s} \sum_{i=1}^{s} |f(x_i, w) - y_i|^2 + \frac{\alpha}{2} \|w\|^2 \right\}$ for $y_i = \begin{cases} 0 & \text{for class label } C_1 \\ 1 & \text{for class label } C_2 \end{cases}$

Then it is natural to decide

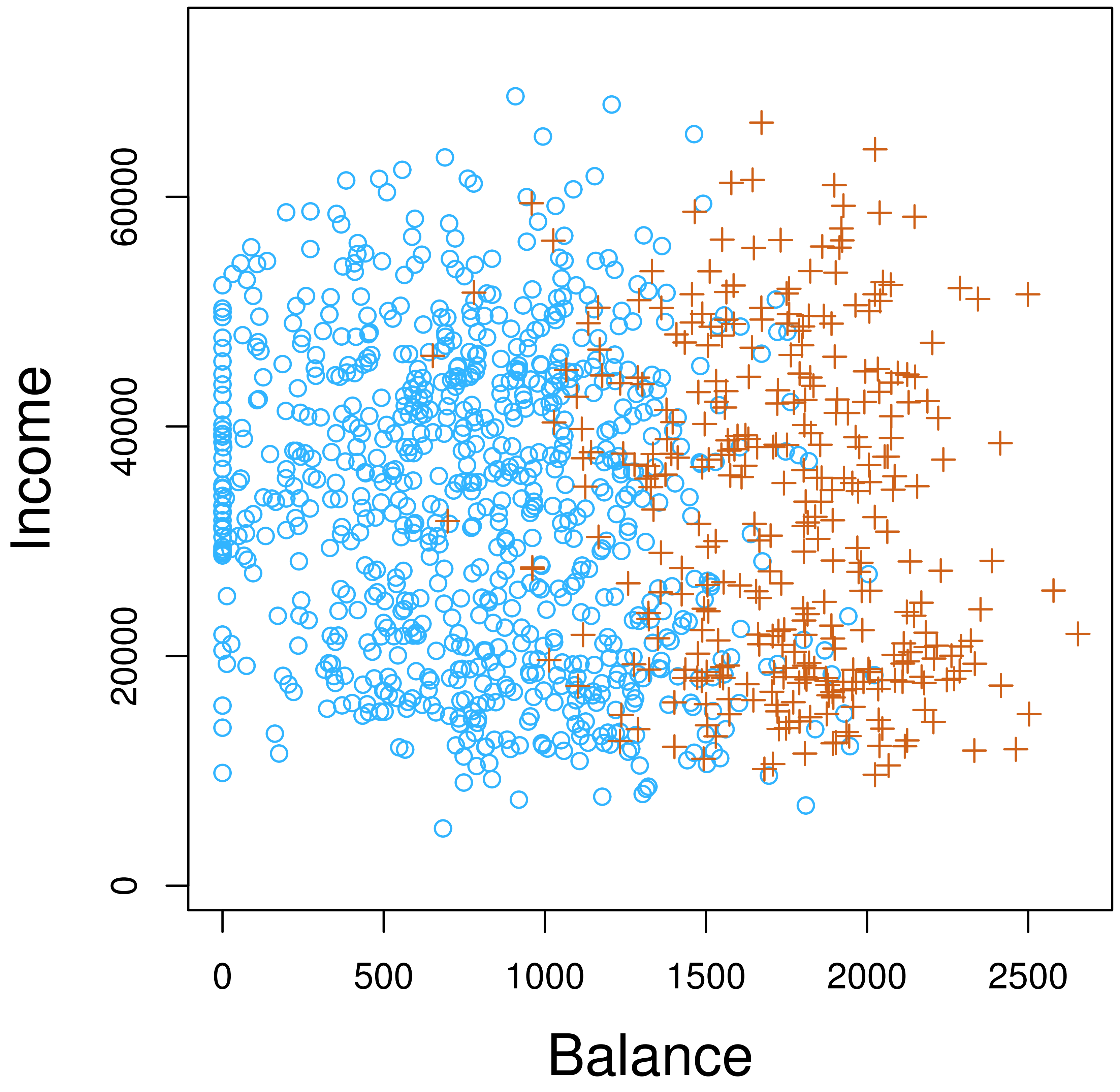
$$f(x_i, \hat{w}) < \frac{1}{2} \quad \Rightarrow \quad \text{The predicted output is in class with label } C_1$$

$$f(x_i, \hat{w}) > \frac{1}{2} \quad \Rightarrow \quad \text{The predicted output is in class with label } C_2$$

Let's look at an example, shall we?

Classification as a special case of regression

Example: credit default



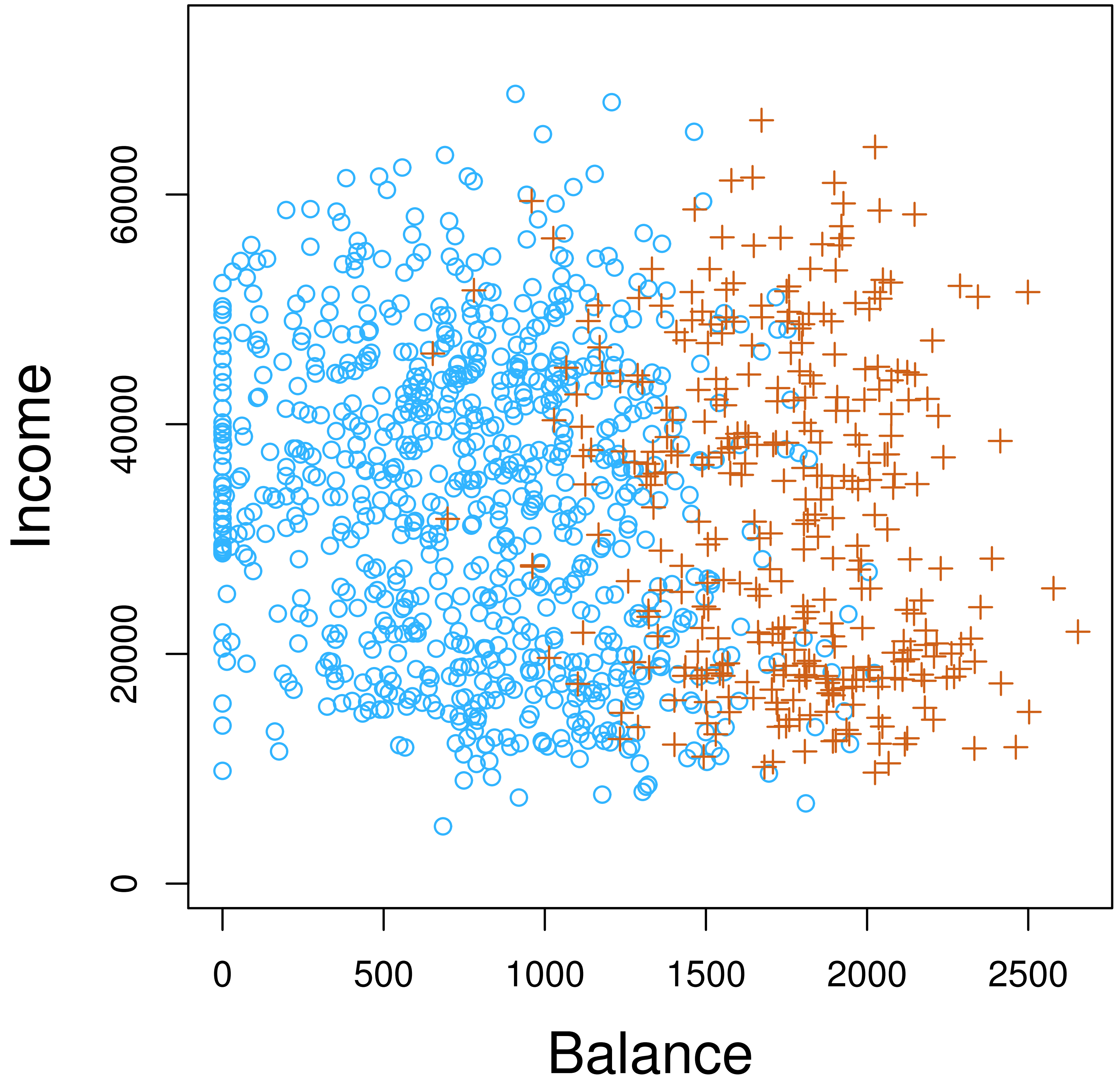
+ = individual who defaulted on their credit card payments

o = individual who did not default on their credit card payments

from *Elements of Statistical Learning* by Hastie, Tibshirani and Friedman

Classification as a special case of regression

Example: credit default

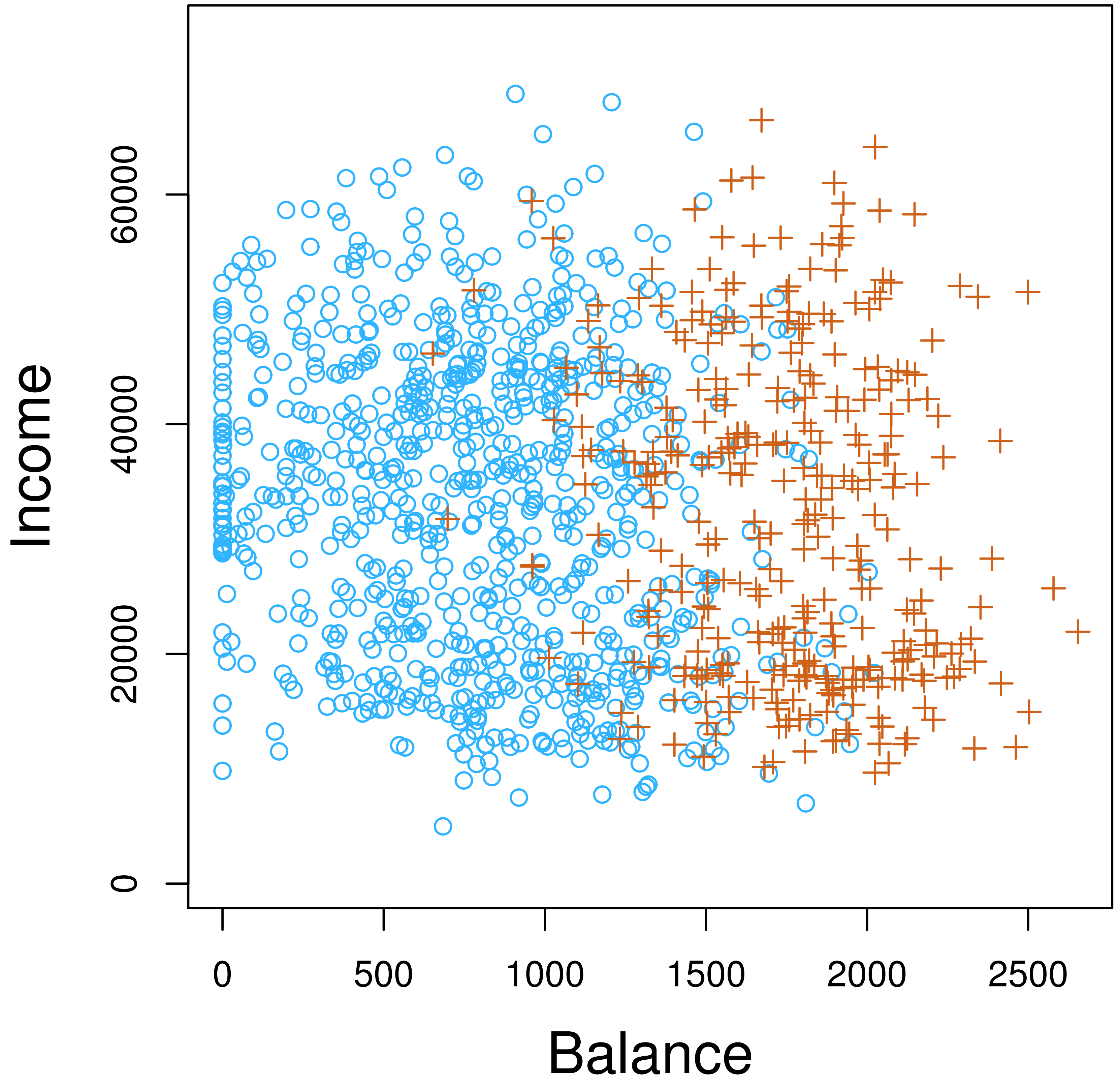


It seems that the balance is the dominant contributing factor towards predicting a default

from *Elements of Statistical Learning* by Hastie, Tibshirani and Friedman

Classification as a special case of regression

Example: credit default



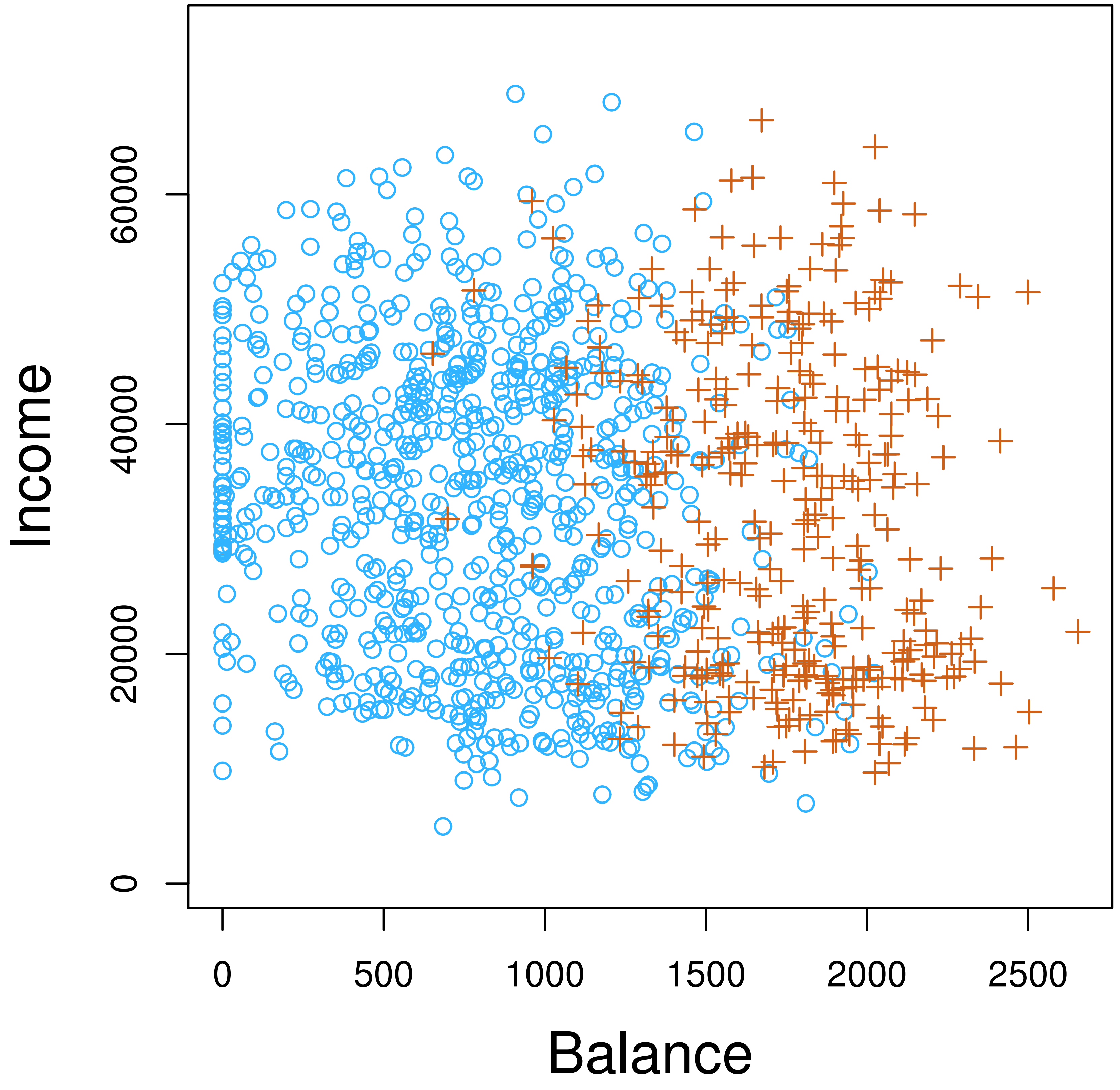
It seems that the balance is the dominant contributing factor towards predicting a default

We therefore ignore the income for now and focus on the balance

from *Elements of Statistical Learning* by Hastie, Tibshirani and Friedman

Classification as a special case of regression

Example: credit default



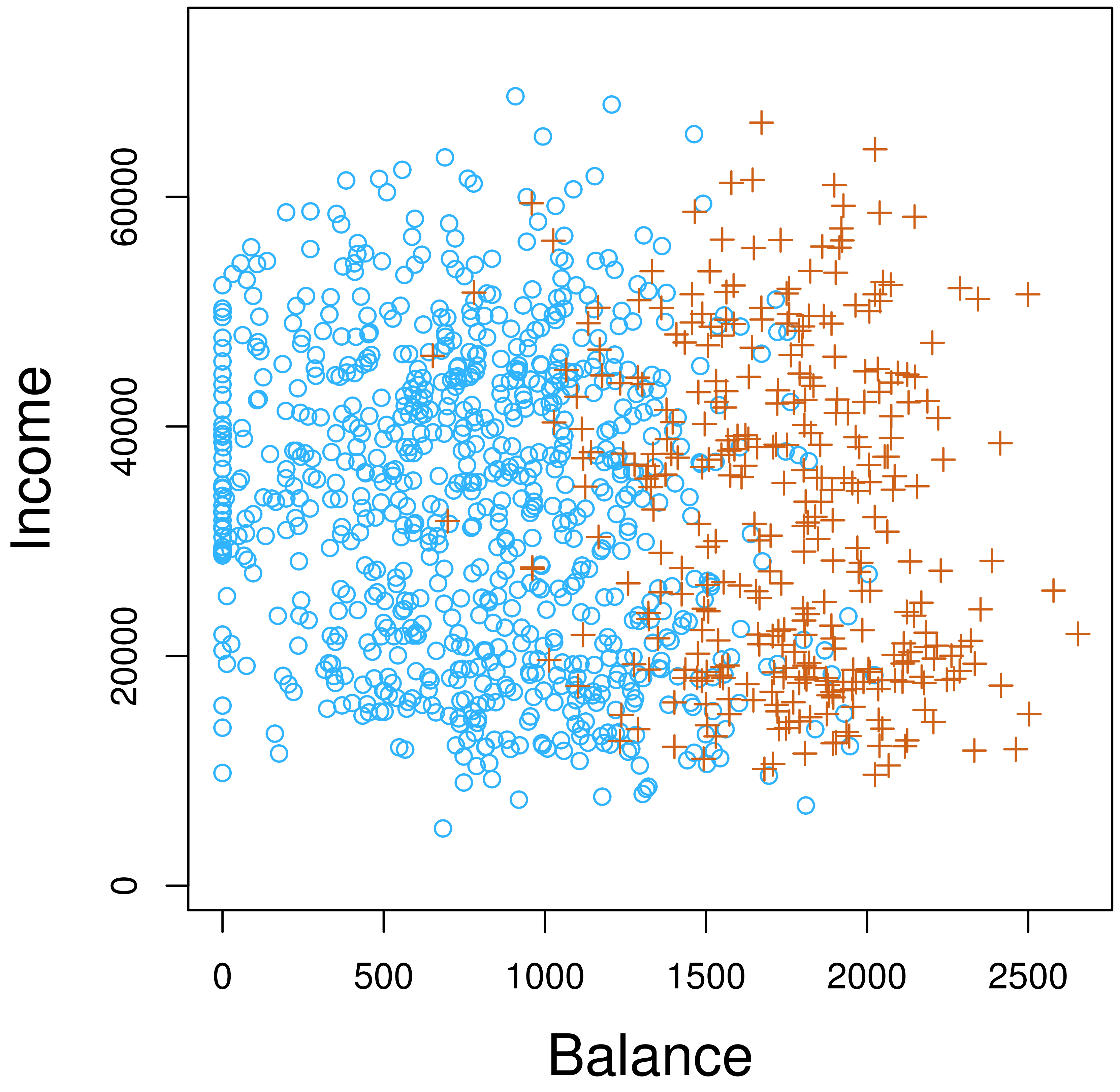
Model assumption

$$f(x, w_0, w_1) = w_0 + w_1 x$$

from *Elements of Statistical Learning*
by Hastie, Tibshirani and Friedman

Classification as a special case of regression

Example: credit default



Model assumption

$$f(x, w_0, w_1) = w_0 + w_1 x$$

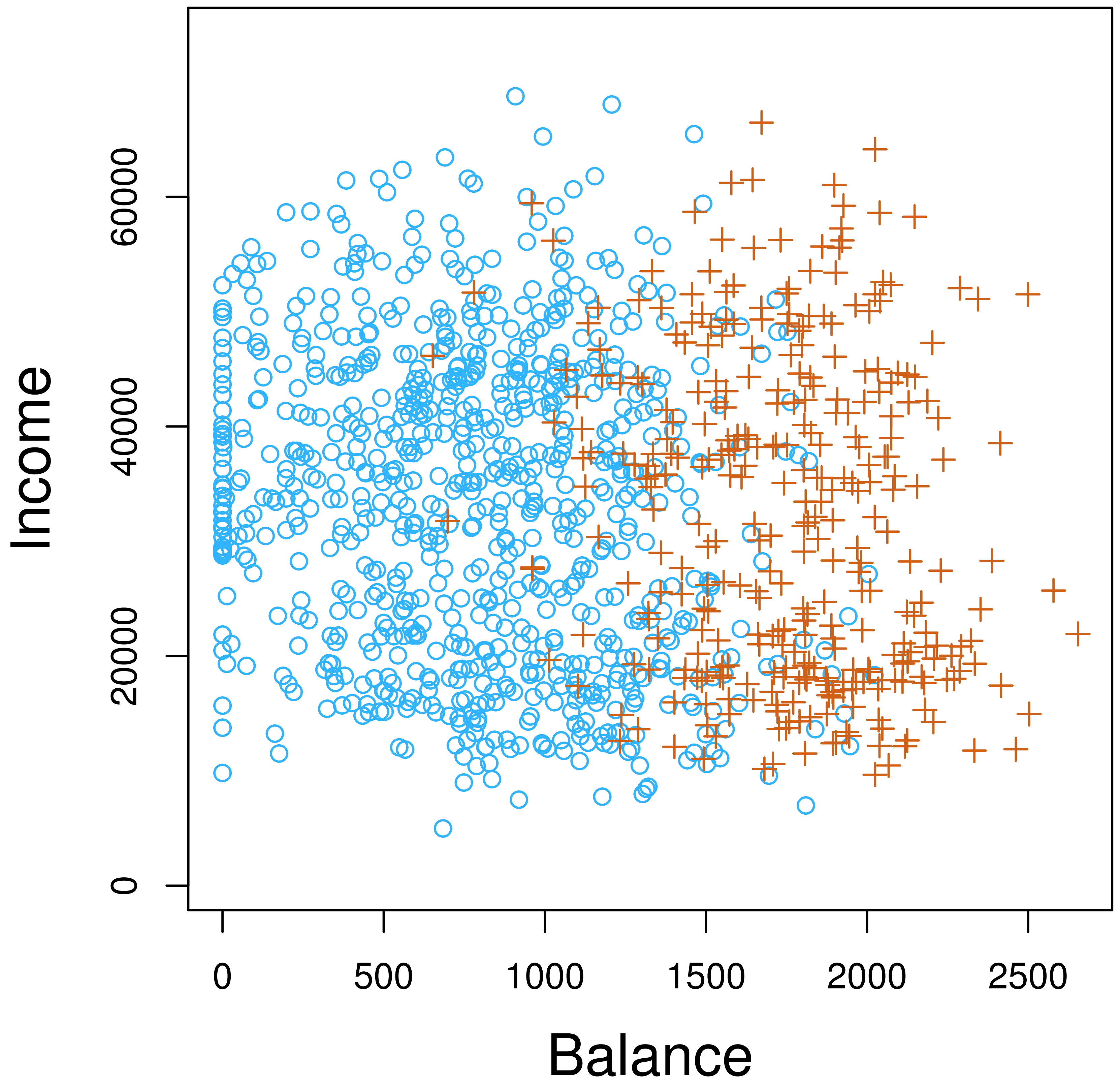
Input $\{x_i\}_{i=1}^{s}$

$$x_i = \text{Balance of customer } i$$

from *Elements of Statistical Learning*
by Hastie, Tibshirani and Friedman

Classification as a special case of regression

Example: credit default



Model assumption

$$f(x, w_0, w_1) = w_0 + w_1 x$$

Input $\{x_i\}_{i=1}^{s}$

$$x_i = \text{Balance of customer } i$$

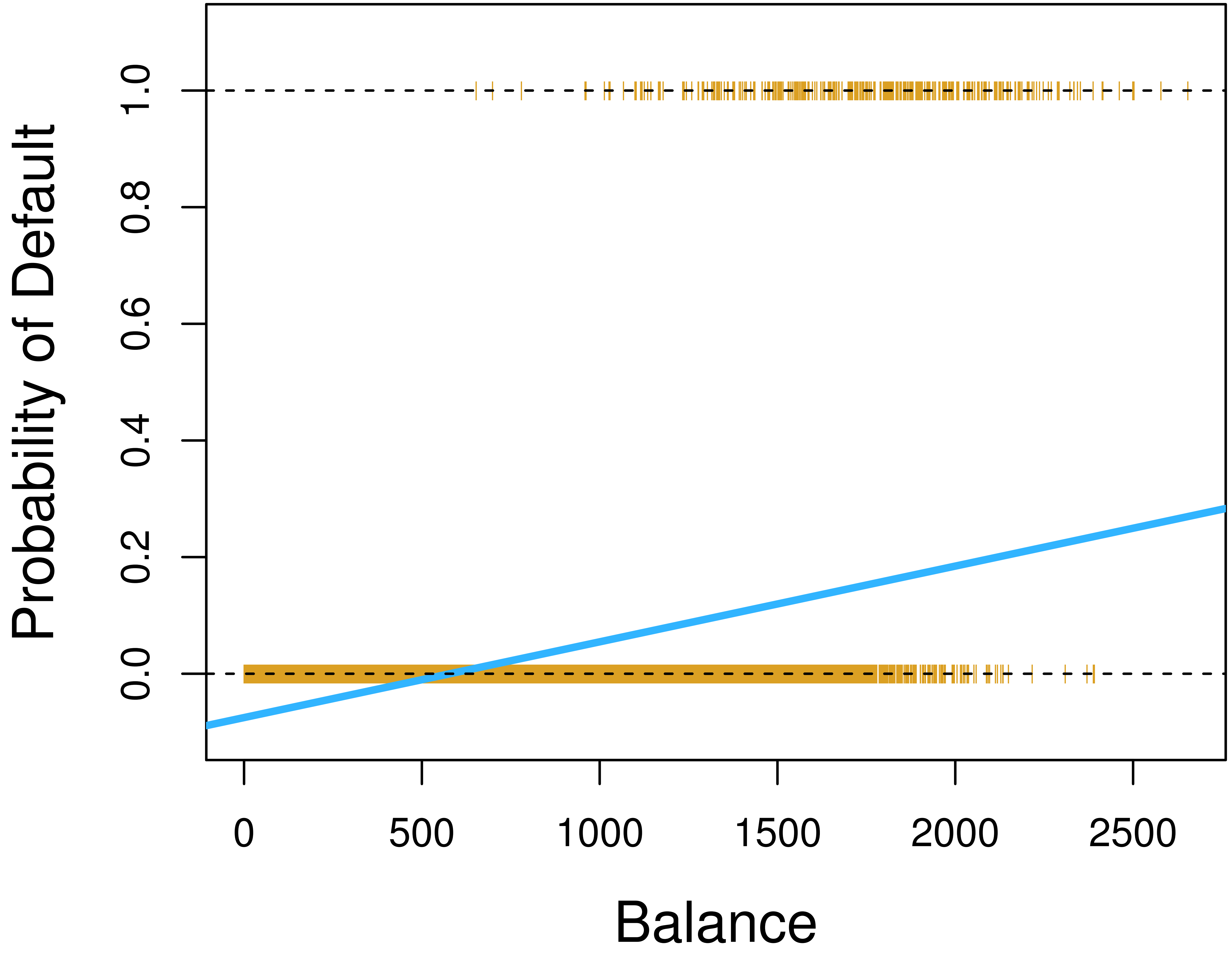
Output $\{y_i\}_{i=1}^{s}$

$$y_i = \begin{cases} 1 & \text{Customer } i \text{ did default} \\ 0 & \text{Customer } i \text{ did not default} \end{cases}$$

from *Elements of Statistical Learning*
by Hastie, Tibshirani and Friedman

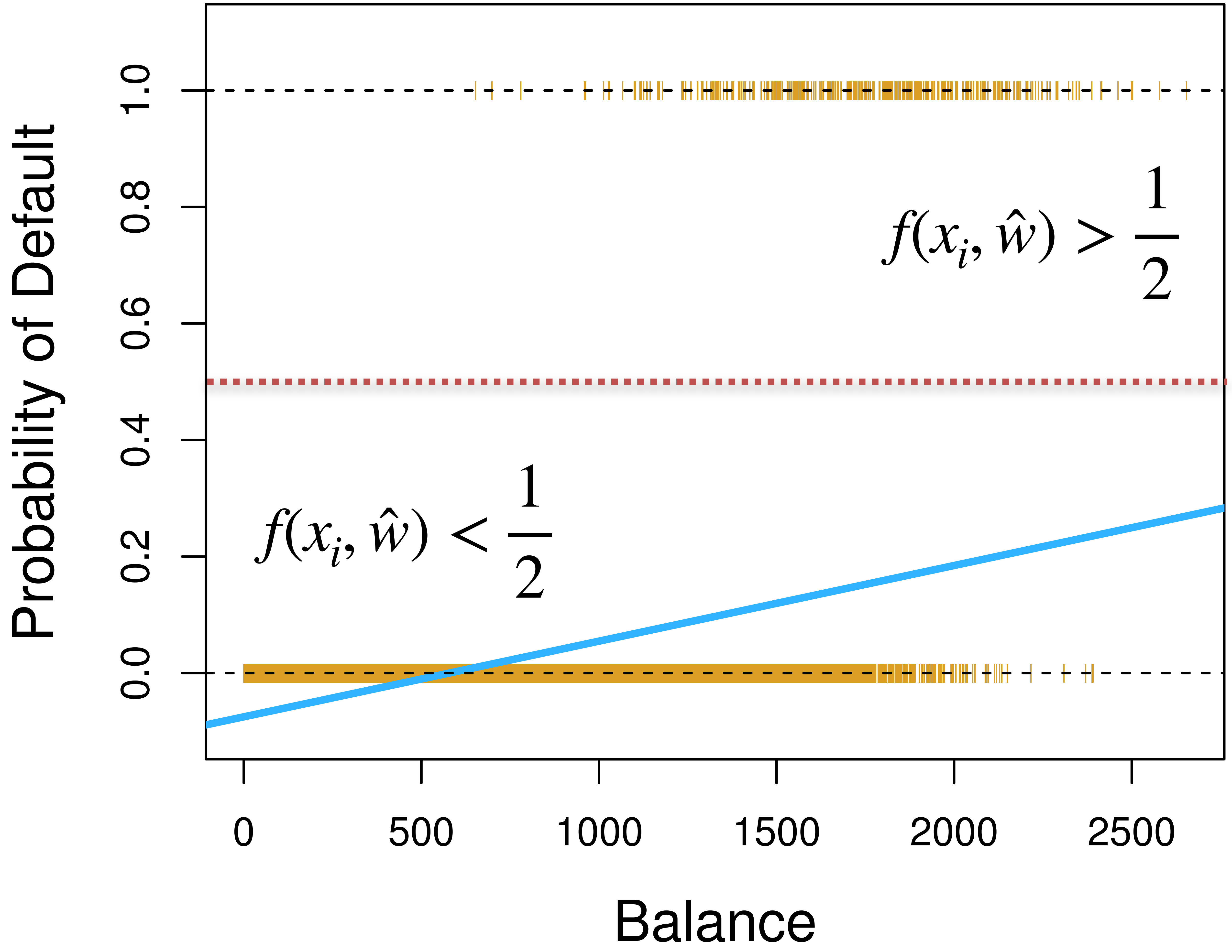
Classification as a special case of regression

Regression outcome is as follows:



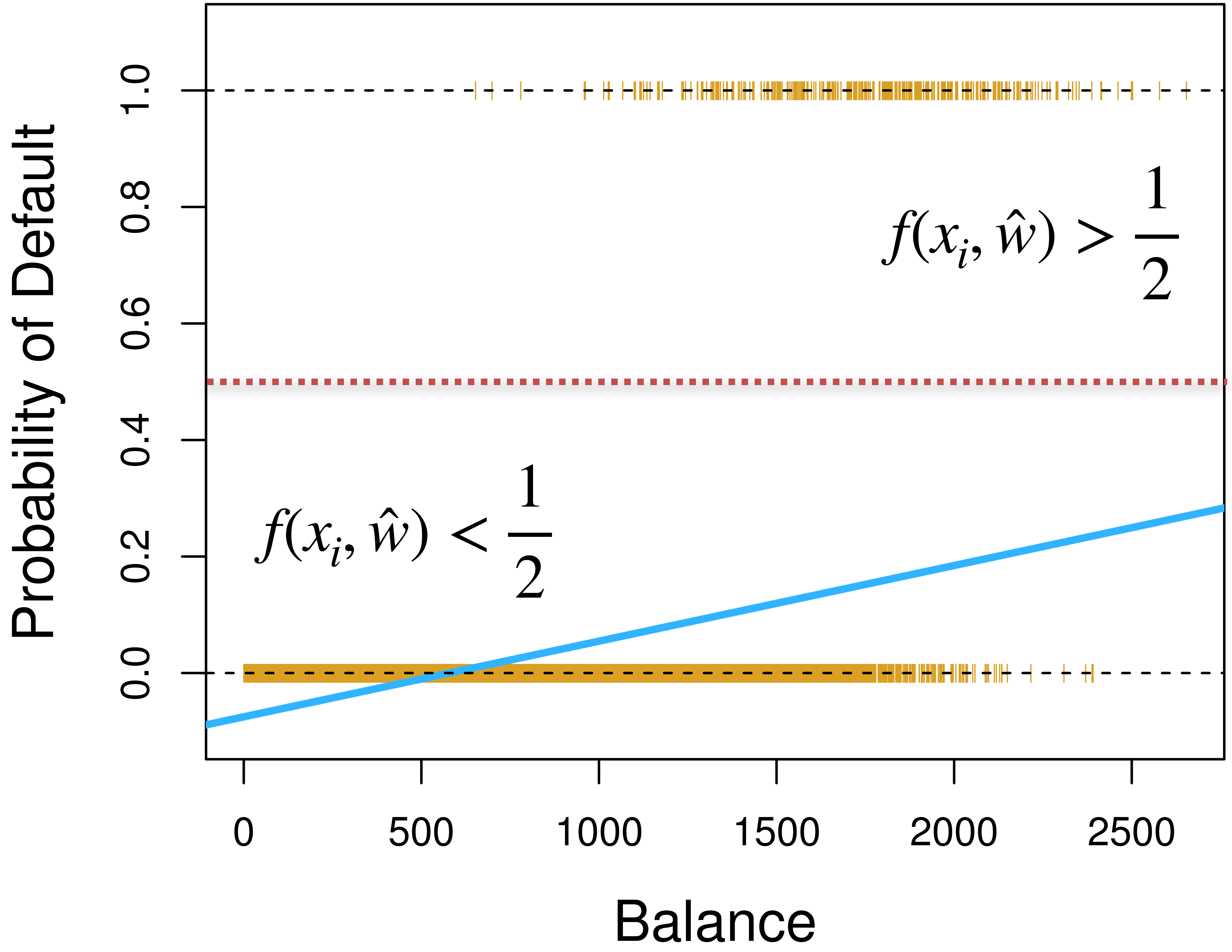
Classification as a special case of regression

Regression outcome is as follows:



Classification as a special case of regression

Regression outcome is as follows:



That didn't go according to plan. What went wrong?

Classification as a special case of regression

‘Position’ of line will crucially depend

Classification as a special case of regression

'Position' of line will crucially depend

- on how many points are in each class
- and where these points lie

Classification as a special case of regression

'Position' of line will crucially depend

- on how many points are in each class
- and where these points lie

Example: if we add a few points with $y = 1$ and very high balance, the line will be shifted/tilted, although only few points have changed.

Classification as a special case of regression

‘Position’ of line will crucially depend

- on how many points are in each class
- and where these points lie

Example: if we add a few points with $y = 1$ and very high balance, the line will be shifted/tilted, although only few points have changed.

This is not a desirable property!

Why does this happen?

Classification as a special case of regression

We would like the fraction of misclassified cases to be small

But: MSE is very loosely related to this objective!

Classification as a special case of regression

We would like the fraction of misclassified cases to be small

But: MSE is very loosely related to this objective!

Example: MSE treats positive and negative deviations from class label equally

But only one can lead to misclassification

Classification as a special case of regression

We would like the fraction of misclassified cases to be small

But: MSE is very loosely related to this objective!

Example: MSE treats positive and negative deviations from class label equally

But only one can lead to misclassification

Small MSE $\Rightarrow$ small classification error

Classification as a special case of regression

We would like the fraction of misclassified cases to be small

But: MSE is very loosely related to this objective!

Example: MSE treats positive and negative deviations from class label equally

But only one can lead to misclassification

Small MSE $\Rightarrow$ small classification error

But the opposite is not necessarily true!

Classification as a special case of regression

We would like the fraction of misclassified cases to be small

But: MSE is very loosely related to this objective!

Example: MSE treats positive and negative deviations from class label equally

But only one can lead to misclassification

Small MSE $\Rightarrow$ small classification error

But the opposite is not necessarily true!

MSE is not a good metric for these types of problems!