逻辑（对数几率）回归

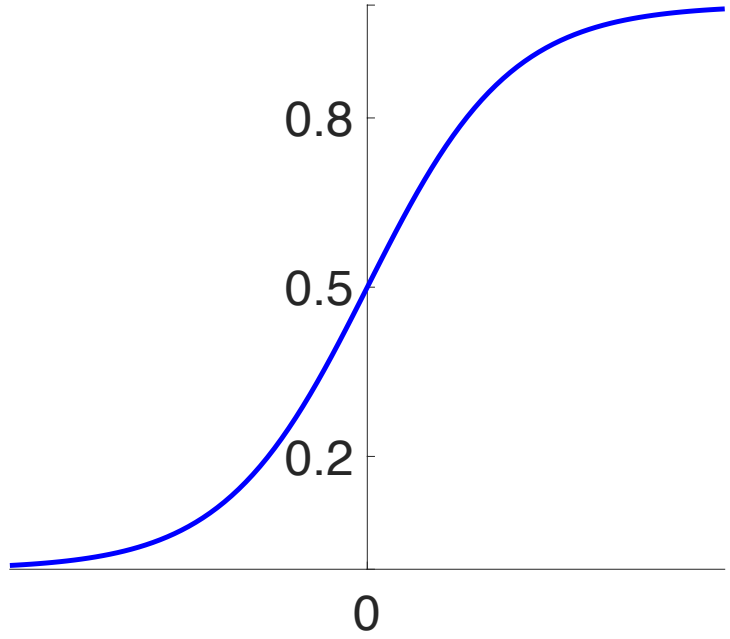
Logistic Regression

郝　奇
南京大学 天文与空间科学学院

Application of Machine Learning in Astronomy

机器学习在天文中的应用

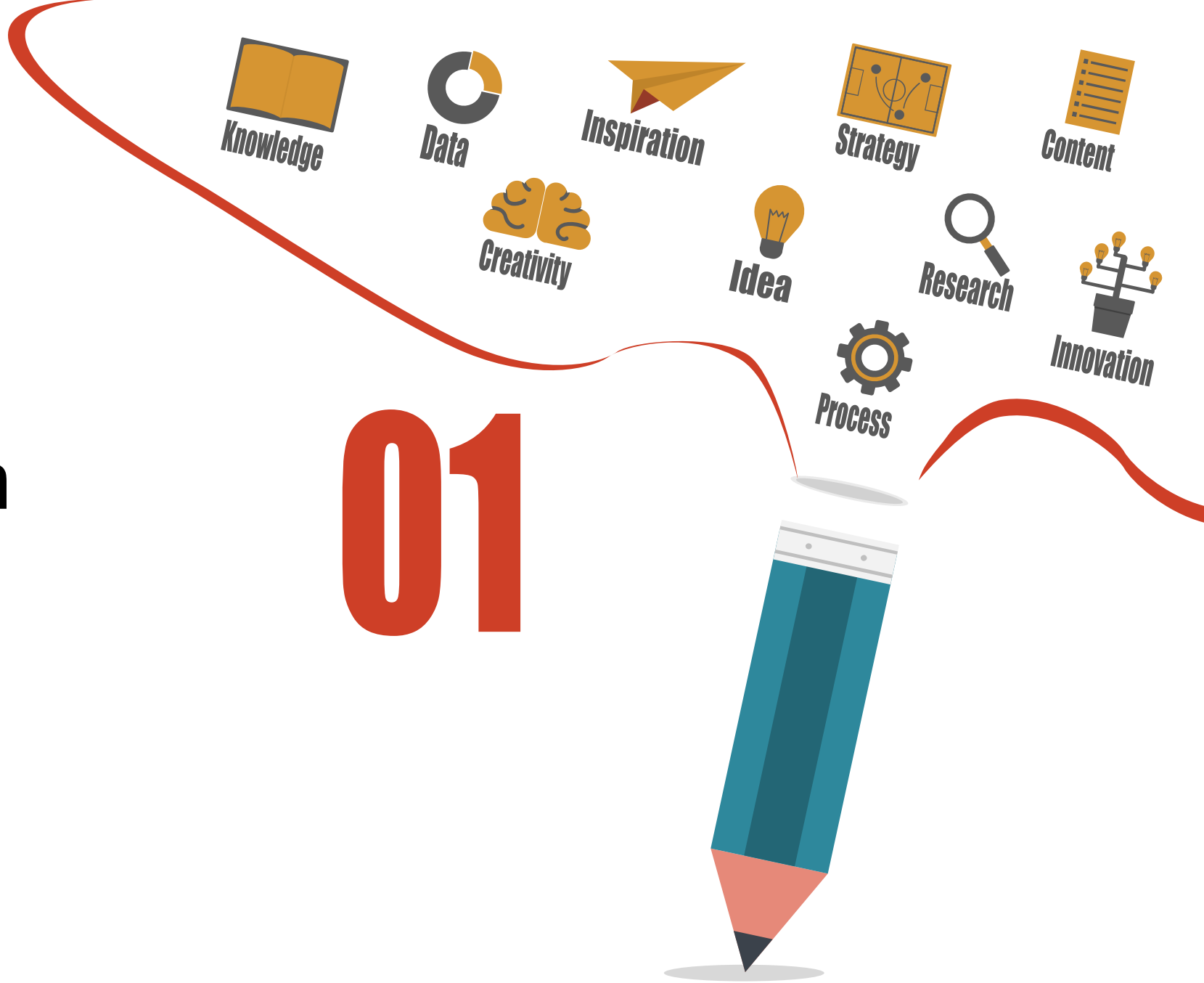
Contents



Cite: Hung-yi Lee, Machine Learning

Logistic Regression

01



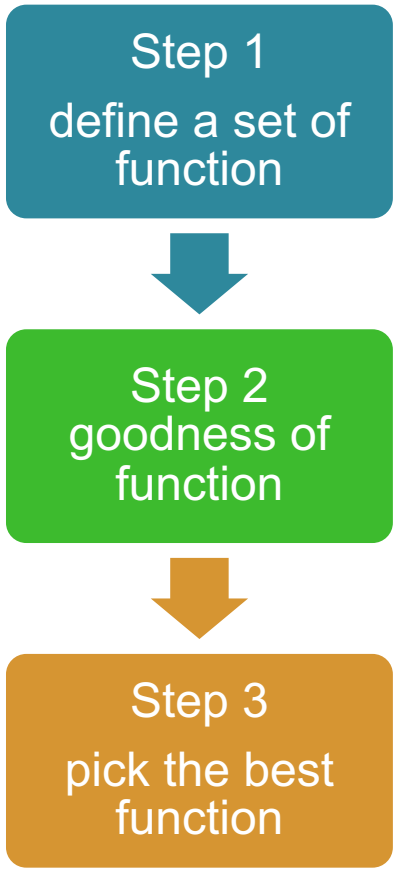
1. Logistic Regression

Logistic regression V.S. Linear Regression

Previous Lecture

1. Logistic Regression

Logistic regression V.S. Linear Regression



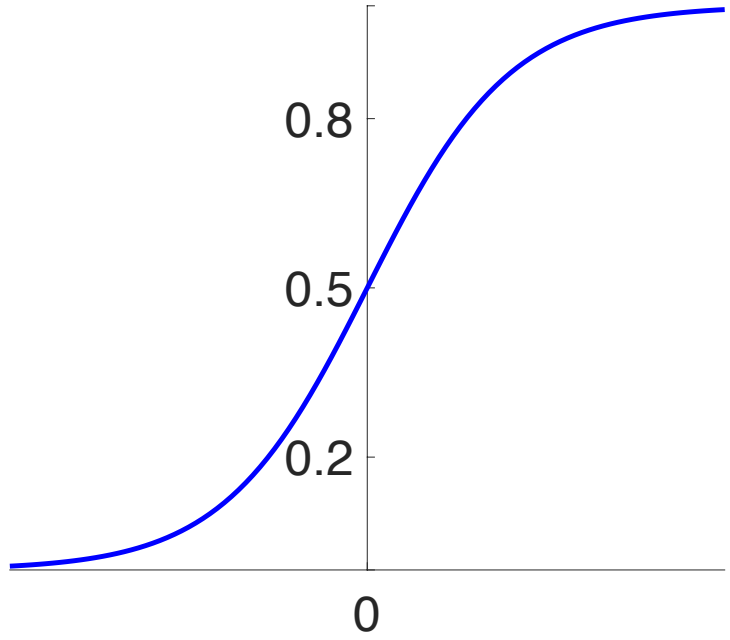
	Logistic regression	Linear regression
Step 1	?	$f_{w,b}(x) = b + \sum w_i x_i$ Output: any value
Step 2	?	Training data: $(x^n, \hat{y}^n)$ $\hat{y}^n$: a real number $L(f) = \frac{1}{2}\sum_n \left(\hat{y}^n - f(x^n)\right)^2$
Step 3	?	$w_{i+1} \leftarrow w_i - \eta \sum_n -\left(\hat{y}^n - f_{w,b}(x^n)\right) x_i^n$

1. Logistic Regression

1.1 Function set

$$P(C_1|x) = \frac{P(x|C_1)P(C_1)}{P(x|C_1)P(C_1) + P(x|C_2)P(C_2)}$$

$$\sigma(z) = \frac{1}{1 + exp(-z)}$$



If $P(C_1|x) > 0.5$ Output = class 1

$else$ Output = class 2

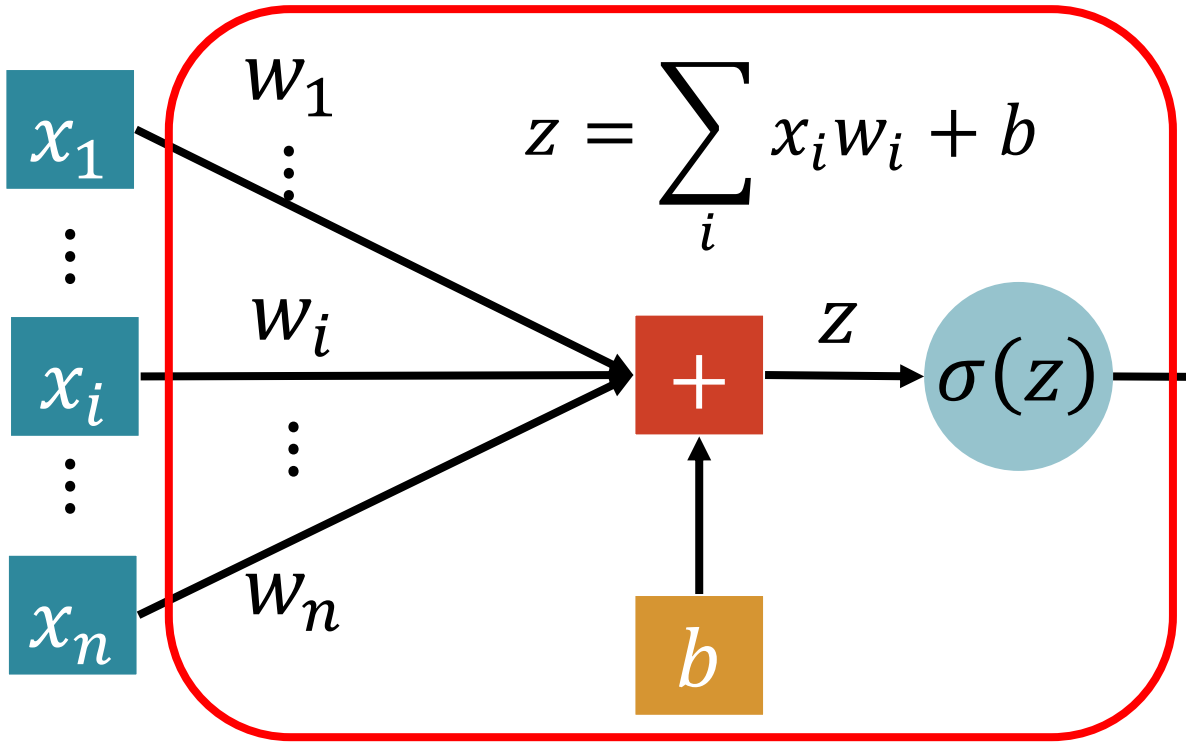
$$z = wx + b$$

$$P(C_1|x) = \sigma(z) = \sigma(wx + b)$$

Function set: $f_{w,b}(x) = P(C_1|x)$ including all different w and b

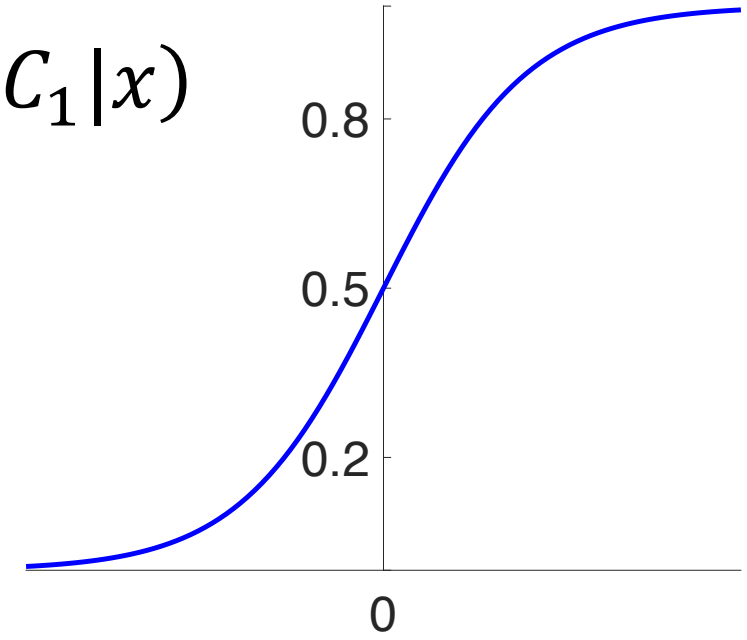
1. Logistic Regression

1.1 Function set



$$\sigma(z) = \frac{1}{1 + exp(-z)}$$

$$P_{w,b}(C_1|x)$$



$$f_{w,b}(x)$$

1. Logistic Regression

Logistic regression V.S. Linear Regression

Step 1
define a set of function

Step 2
goodness of function

Step 3
pick the best function

	Logistic regression	Linear regression
Step 1	$f_{w,b}(x) = \sigma\left(b + \sum w_i x_i\right)$ Output: between 0 and 1	$f_{w,b}(x) = b + \sum w_i x_i$ Output: any value
Step 2	?	Training data: $(x^n, \hat{y}^n)$ $\hat{y}^n$: a real number $L(f) = \frac{1}{2}\sum_n \left(\hat{y}^n - f(x^n)\right)^2$
Step 3	?	$w_{i+1} \leftarrow w_i - \eta \sum_n -\left(\hat{y}^n - f_{w,b}(x^n)\right) x_i^n$

1. Logistic Regression

1.2 Goodness of function

Training Data	x^1	x^2	x^3	……	x^n
	C_1	C_1	C_2		C_1

Assume the data is generated based on $f_{w,b}(x) = P(C_1|x)$;

Given a set of w and b, what is its probability of generating the data?

$$L(w,b) = f_{w,b}(x^1)\ f_{w,b}(x^2)\left(1 - f_{w,b}(x^3)\right)\cdots f_{w,b}\ (x^n)$$

The most likely w^* and b^* is the one with the largest $L(w,b)$:

$$w^*, b^* = \arg\max_{w,b} L(w,b)$$

1. Logistic Regression

1.2 Goodness of function

$$L(w,b) = f_{w,b}(x^1)\, f_{w,b}(x^2) \left(1 - f_{w,b}(x^3)\right) \cdots f_{w,b}(x^n)$$

Training Data

x^1	x^2	x^3	……	x^n
C_1	C_1	C_2		C_1

x^1	x^2	x^3	……
$\hat{y}^1 = 1$	$\hat{y}^2 = 1$	$\hat{y}^3 = 0$	

$$w^*, b^* = \arg\max_{w,b} L(w,b) \quad \Longleftrightarrow \quad w^*, b^* = \arg\min_{w,b} -\ln L(w,b)$$

$$-\ln L(w,b) =$$

$$-\ln f_{w,b}(x^1) \Rightarrow -\left[\hat{y}^1 \ln f_{w,b}(x^1) + (1-\hat{y}^1) \ln\left(1 - f_{w,b}(x^1)\right)\right]$$

$$-\ln f_{w,b}(x^2) \Rightarrow -\left[\hat{y}^2 \ln f_{w,b}(x^2) + (1-\hat{y}^2) \ln\left(1 - f_{w,b}(x^2)\right)\right]$$

$$-\ln\left(1 - f_{w,b}(x^3)\right) \Rightarrow -\left[\hat{y}^3 \ln f_{w,b}(x^3) + (1-\hat{y}^3) \ln\left(1 - f_{w,b}(x^3)\right)\right]$$

$$\vdots$$

1. Logistic Regression

1.2 Goodness of function

$$L(w,b) = f_{w,b}(x^1)\, f_{w,b}(x^2) \left(1 - f_{w,b}(x^3)\right) \cdots f_{w,b}\,(x^n)$$

$$-\ln L(w,b) = \ln f_{w,b}(x^1) + \ln f_{w,b}(x^2) + \ln \left(1 - f_{w,b}(x^3)\right) \cdots + \ln f_{w,b}(x^n)$$

$$= \sum_n -\left[\hat{y}^n \ln f_{w,b}(x^n) + (1-\hat{y}^n) \ln \left(1 - f_{w,b}(x^n)\right)\right]$$

Cross entropy between two Bernoulli distribution

Distribution p:

$p(x=1) = \hat{y}^n$

$p(x=0) = 1 - \hat{y}^n$

cross entropy

$$\boldsymbol{H(p,q) = -\sum_x p(x) \ln\left(q(x)\right)}$$

Distribution q:

$q(x=1) = f(x^n)$

$q(x=0) = 1 - f(x^n)$

1. Logistic Regression

Logistic regression V.S. Linear Regression

Step 1
define a set of function

Step 2
goodness of function

Step 3
pick the best function

	Logistic regression	Linear regression
Step 1	$f_{w,b}(x) = \sigma\left(b + \sum w_i x_i\right)$ Output: between 0 and 1	$f_{w,b}(x) = b + \sum w_i x_i$ Output: any value
Step 2	Training data: $(x^n, \hat{y}^n)$ $\hat{y}^n$: 1 for class 1, 0 for class 2 $L(f) = \sum_n C\left(\hat{y}^n, f(x^n)\right)$	Training data: $(x^n, \hat{y}^n)$ $\hat{y}^n$: a real number $L(f) = \frac{1}{2}\sum_n \left(\hat{y}^n - f(x^n)\right)^2$
Step 3	?	$w_{i+1} \leftarrow w_i - \eta \sum_n -\left(\hat{y}^n - f_{w,b}(x^n)\right) x_i^n$

1. Logistic Regression

1.2 Goodness of function

Cross entropy

$$C\left(\hat{y}^n, f(x^n)\right) = \left[\hat{y}^n \ln f_{w,b}(x^n) + (1-\hat{y}^n)\ln\left(1-f_{w,b}(x^n)\right)\right]$$

$$\mathrm{L}(f) = \sum_n C\left(\hat{y}^n, f(x^n)\right) \qquad \mathrm{L}(f) = \frac{1}{2}\sum_n \left(\hat{y}^n - f(x^n)\right)^2$$

Why don't we simply use square error as linear regression?

1. Logistic Regression

Step 1
define a set of function

Step 2
goodness of function

Step 3
pick the best function

$$f_{w,b}(x) = \sigma\left(b + \sum w_i x_i\right) \qquad z = \sum_i x_i w_i + b \qquad \sigma(z) = \frac{1}{1 + exp(-z)}$$

Training data: $(x^n, \hat{y}^n)$

$\hat{y}^n$: 1 for class 1, 0 for class 2

$$L(f) = \frac{1}{2}\sum_n \left(\hat{y}^n - f(x^n)\right)^2$$

$$\frac{\partial\left(\hat{y}^n - f(x^n)\right)^2}{\partial w_i} = -2\left(\hat{y}^n - f(x^n)\right)\frac{\partial f_{w,b}(x)}{\partial z}\frac{\partial z}{\partial w_i}$$

$$= -2\left(\hat{y}^n - f(x^n)\right) f_{w,b}(x)\left(1 - f_{w,b}(x)\right) x_i$$

$\hat{y}^n = 1$

If $f_{w,b}(x^n) = 1$ (close to target) → $\partial L/\partial w_i = 0$

If $f_{w,b}(x^n) = 0$ (far from target) → $\partial L/\partial w_i = 0$

1. Logistic Regression

Step 1
define a set of function

Step 2
goodness of function

Step 3
pick the best function

$$f_{w,b}(x) = \sigma\left(b + \sum w_i x_i\right) \qquad z = \sum_i x_i w_i + b \qquad \sigma(z) = \frac{1}{1 + exp(-z)}$$

Training data: $(x^n, \hat{y}^n)$

$\hat{y}^n$: 1 for class 1, 0 for class 2

$$L(f) = \frac{1}{2}\sum_n \left(\hat{y}^n - f(x^n)\right)^2$$

$$\frac{\partial\left(\hat{y}^n - f(x^n)\right)^2}{\partial w_i} = -2\left(\hat{y}^n - f(x^n)\right)\frac{\partial f_{w,b}(x)}{\partial z}\frac{\partial z}{\partial w_i}$$

$$= -2\left(\hat{y}^n - f(x^n)\right) f_{w,b}(x)\left(1 - f_{w,b}(x)\right) x_i$$

$\hat{y}^n = 0$

If $f_{w,b}(x^n) = 1$ (far from target) → $\partial L/\partial w_i = 0$

If $f_{w,b}(x^n) = 0$ (close to target) → $\partial L/\partial w_i = 0$

1. Logistic Regression

1.3 Find the best function

$$z = \sum_i x_i w_i + b \qquad \sigma(z) = \frac{1}{1 + exp(-z)}$$

$$-\ln L(w,b) = \sum_n -\left[\hat{y}^n \ln f_{w,b}(x^n) + (1-\hat{y}^n) \ln\left(1 - f_{w,b}(x^n)\right)\right]$$

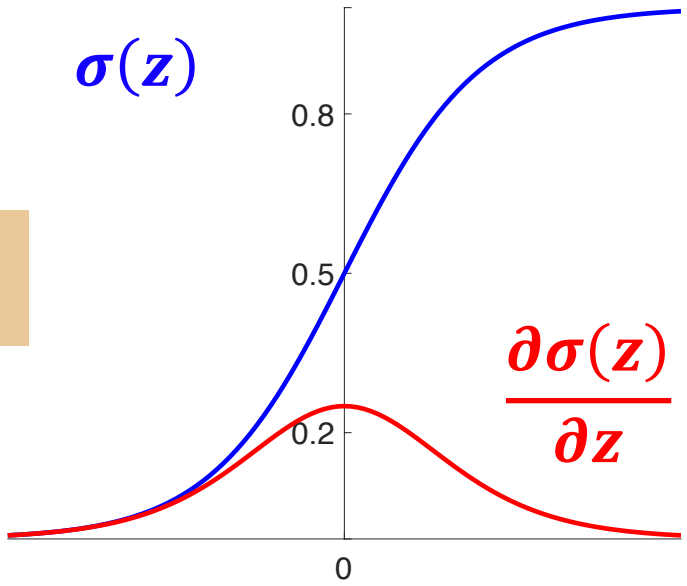
$$\frac{\partial(-\ln L(w,b))}{\partial w_i} \qquad \frac{\partial \ln f_{w,b}(x^n)}{\partial w_i} \qquad \frac{\partial \ln\left(1 - f_{w,b}(x^n)\right)}{\partial w_i}$$

$$\frac{\partial \ln f_{w,b}(x)}{\partial w_i} = \frac{\partial \ln f_{w,b}(x)}{\partial z}\frac{\partial z}{\partial w_i}$$

$$\frac{\partial z}{\partial w_i} = x_i$$

$$\left(1 - f_{w,b}(x^n)\right) x_i^n$$

$$\frac{\partial \ln f_{w,b}(x)}{\partial z} = \frac{1}{\sigma(z)}\frac{\partial \sigma(z)}{\partial z} = \frac{1}{\sigma(z)}\sigma(z)\left(1-\sigma(z)\right)$$



1. Logistic Regression

1.3 Find the best function

$$z = \sum_i x_i w_i + b \qquad \sigma(z) = \frac{1}{1 + exp(-z)}$$

$$-\ln L(w,b) = \sum_n -\left[\hat{y}^n \ln f_{w,b}(x^n) + (1-\hat{y}^n) \ln\left(1 - f_{w,b}(x^n)\right)\right]$$

$$\frac{-\ln L(w,b)}{\partial w_i} \qquad \frac{\partial \ln f_{w,b}(x^n)}{\partial w_i} \qquad \frac{\partial \ln\left(1 - f_{w,b}(x^n)\right)}{\partial w_i}$$

$$\frac{\partial \ln\left(1 - f_{w,b}(x^n)\right)}{\partial w_i} = \frac{\partial \ln\left(1 - f_{w,b}(x^n)\right)}{\partial z} \frac{\partial z}{\partial w_i}$$

$$\boldsymbol{f_{w,b}(x^n)\, x_i^n}$$

$$\frac{\partial z}{\partial w_i} = x_i$$

$$\frac{\partial \ln\left(1 - f_{w,b}(x^n)\right)}{\partial z} = -\frac{1}{1-\sigma(z)} \frac{\partial \sigma(z)}{\partial z} = \frac{1}{1-\sigma(z)} \sigma(z)\left(1-\sigma(z)\right)$$

1. Logistic Regression

1.3 Find the best function

$$-\ln L(w,b) = \sum_n -\left[\hat{y}^n \ln f_{w,b}(x^n) + (1-\hat{y}^n)\ln\left(1-f_{w,b}(x^n)\right)\right]$$

$$\frac{\partial -\ln L(w,b)}{\partial w_i} = \sum_n -\left[\hat{y}^n \frac{\partial \ln f_{w,b}(x^n)}{\partial w_i} + (1-\hat{y}^n)\frac{\partial \ln\left(1-f_{w,b}(x^n)\right)}{\partial w_i}\right]$$

$$\frac{\partial \ln f_{w,b}(x^n)}{\partial w_i} = \left(1-f_{w,b}(x^n)\right)x_i^n \qquad \frac{\partial \ln\left(1-f_{w,b}(x^n)\right)}{\partial w_i} = f_{w,b}(x^n)\,x_i^n$$

$$= \sum_n -\left[\hat{y}^n\left(1-f_{w,b}(x^n)\right)x_i^n + (1-\hat{y}^n)f_{w,b}(x^n)\,x_i^n\right]$$

$$= \sum_n -\left[\hat{y}^n - \cancel{\hat{y}^n f_{w,b}(x^n)} - f_{w,b}(x^n) + \cancel{\hat{y}^n f_{w,b}(x^n)}\right]x_i^n$$

Larger difference, larger update

$$= \sum_n -\left(\hat{y}^n - f_{w,b}(x^n)\right)x_i^n$$

$$w_{i+1} \leftarrow w_i - \eta \sum_n -\left(\hat{y}^n - f_{w,b}(x^n)\right)x_i^n$$

1. Logistic Regression

Logistic regression V.S. Linear Regression

Step 1
define a set of function

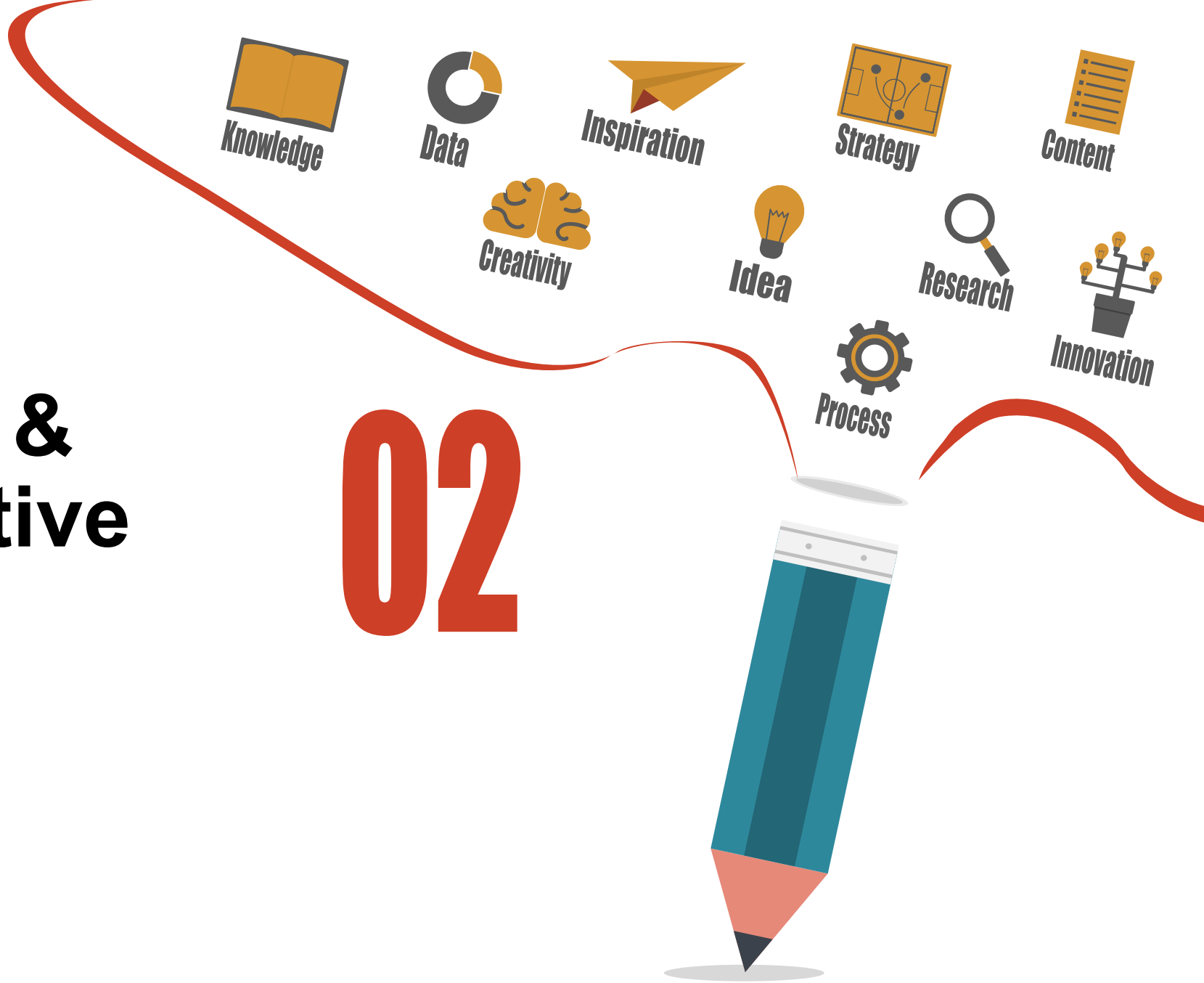
Step 2
goodness of function

Step 3
pick the best function

	Logistic regression	Linear regression
Step 1	$f_{w,b}(x) = \sigma\left(b + \sum w_i x_i\right)$ Output: between 0 and 1	$f_{w,b}(x) = b + \sum w_i x_i$ Output: any value
Step 2	Training data: $(x^n, \hat{y}^n)$ $\hat{y}^n$: 1 for class 1, 0 for class 2 $L(f) = \sum_n C\left(\hat{y}^n, f(x^n)\right)$	Training data: $(x^n, \hat{y}^n)$ $\hat{y}^n$: a real number $L(f) = \frac{1}{2}\sum_n \left(\hat{y}^n - f(x^n)\right)^2$
Step 3	$w_{i+1} \leftarrow w_i - \eta \sum_n -\left(\hat{y}^n - f_{w,b}(x^n)\right) x_i^n$	$w_{i+1} \leftarrow w_i - \eta \sum_n -\left(\hat{y}^n - f_{w,b}(x^n)\right) x_i^n$

Generative & Discriminative

02



2. Generative & Discriminative

$$P(C_1|x) = \sigma(wx + b)$$

Find w and b directly

Find μ^1, μ^2, Σ

$$\boldsymbol{w^T} = (\mu^1 - \mu^2)^T (\Sigma^1)^{-1}$$

$$\boldsymbol{b} = -\frac{1}{2}(\mu^1)^T (\Sigma^1)^{-1} \mu^1 + \frac{1}{2}(\mu^2)^T (\Sigma^2)^{-1} \mu^2 + \ln \frac{N_1}{N_2}$$

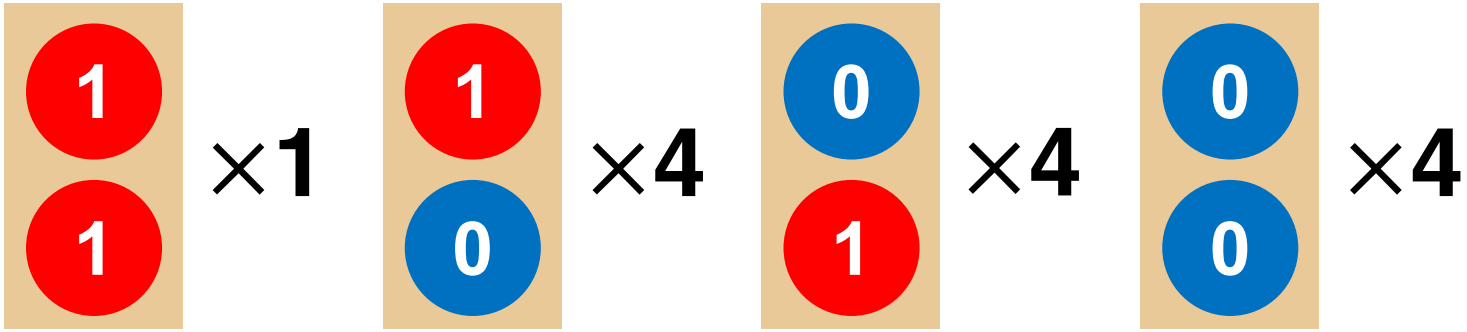
Shall we obtain the same set of w and b ?

The same model (function set), but different function is selected by the same training data.

2. Generative & Discriminative

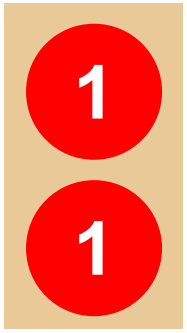
2.1 Toy example

Training Data



Class 1 **Class 2** **Class 2** **Class 2**

Testing Data



Class 1?
Class 2?

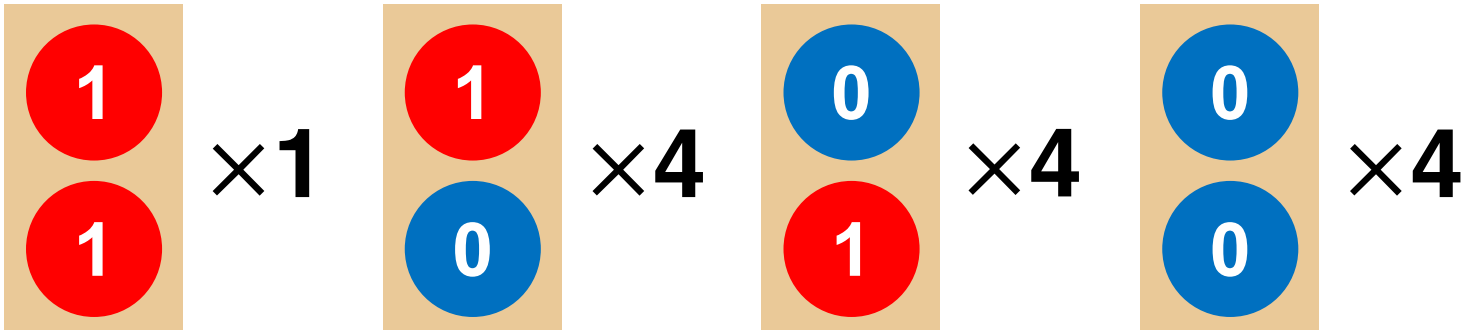
How about Naïve Bayes?

$$P(X|C_i) = P(x_1|C_i)P(x_2|C_i)$$

2. Generative & Discriminative

2.1 Toy example

Training Data



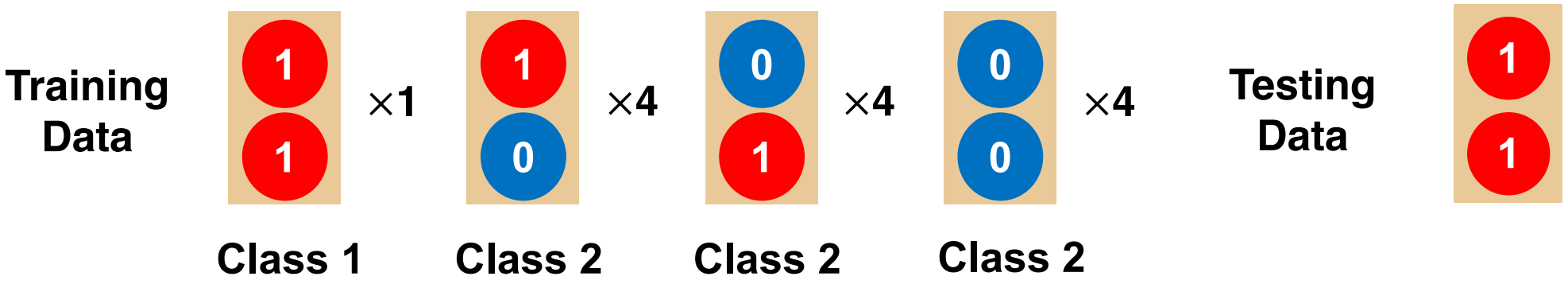
Class 1 **Class 2** **Class 2** **Class 2**

$$P(C_1) = \frac{1}{13} \qquad P(x_1 = 1|C_1) = 1 \qquad P(x_2 = 1|C_2) = 1$$

$$P(C_2) = \frac{12}{13} \qquad P(x_1 = 1|C_2) = \frac{4}{12} \qquad P(x_2 = 1|C_2) = \frac{4}{12}$$

2. Generative & Discriminative

2.1 Toy example



$$P(C_1) = \frac{1}{13} \quad P(x_1 = 1|C_1) = 1 \quad P(x_2 = 1|C_2) = 1 \quad P(x|C_1) = P(x_1 = 1|C_1)P(x_2 = 1|C_1) = 1 \times 1 = 1$$

$$P(C_2) = \frac{12}{13} \quad P(x_1 = 1|C_2) = \frac{4}{12} \quad P(x_2 = 1|C_2) = \frac{4}{12} \quad P(x|C_2) = P(x_1 = 1|C_2)P(x_2 = 1|C_2) = \frac{4}{12} \times \frac{4}{12} = \frac{1}{9}$$

$$P(C_1|x) = \frac{P(x|C_1)P(C_1)}{P(x|C_1)P(C_1) + P(x|C_2)P(C_2)} = \frac{1 \times \frac{1}{13}}{1 \times \frac{1}{13} + \frac{1}{9} \times \frac{12}{13}} = \frac{3}{7} \quad < \mathbf{0.5}$$

2. Generative & Discriminative

2.2 Characteristics of Generative models

- **Focus on probability distributions of the data**
  - With the assumption of probability distribution, they can generate samples;
- **More powerful with less of examples**
  - With the assumption of probability distribution, less training data is needed.
- **Can generate new samples**
  - Can be used in semi-unsupervised learning and unsupervised learning
- **Priors and class-dependent probabilities can be estimated from different sources**

2. Generative & Discriminative

Characteristics of Discriminative models

Previous Lecture

- **Focus on decision boundary**
  - they do not allow one to generate samples from the joint distribution $\mathbf{P}(\boldsymbol{x}, \boldsymbol{y})$;
- **More powerful with lot of examples**
  - Whenever the training data is big ,the accuracy for future data will be good.
- **Not designed to use unlabeled data**
  - Most discriminative models are inherently supervised and cannot easily be extended to unsupervised learning.
- **Outperform generative models at conditional prediction tasks**
  - When the number of parameters is limited, for tasks such as classification and regression that do not require the joint distribution.

2. Generative & Discriminative

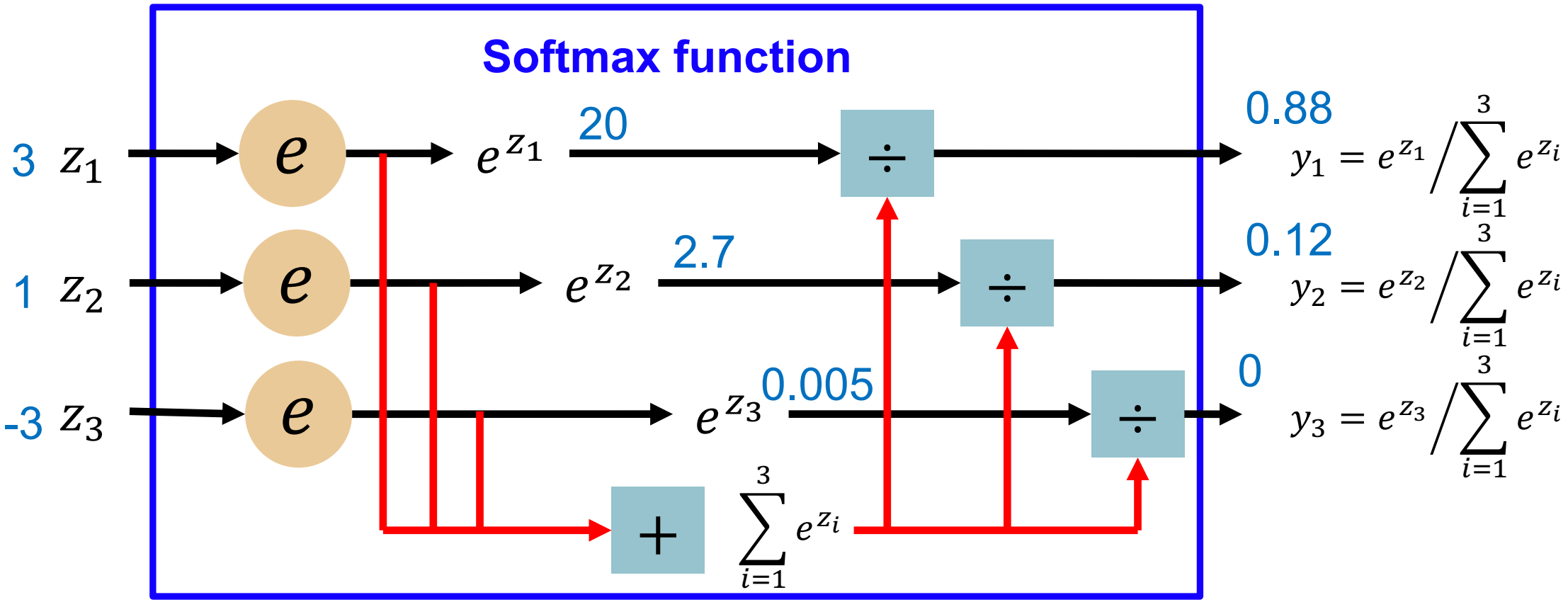
2.3 Multi-class classification

Probability: $0 < y_i < 1; \sum_i y_i = 1.$

$C_1: w^1, b_1 \quad z_1 = w^1 \cdot x + b_1$

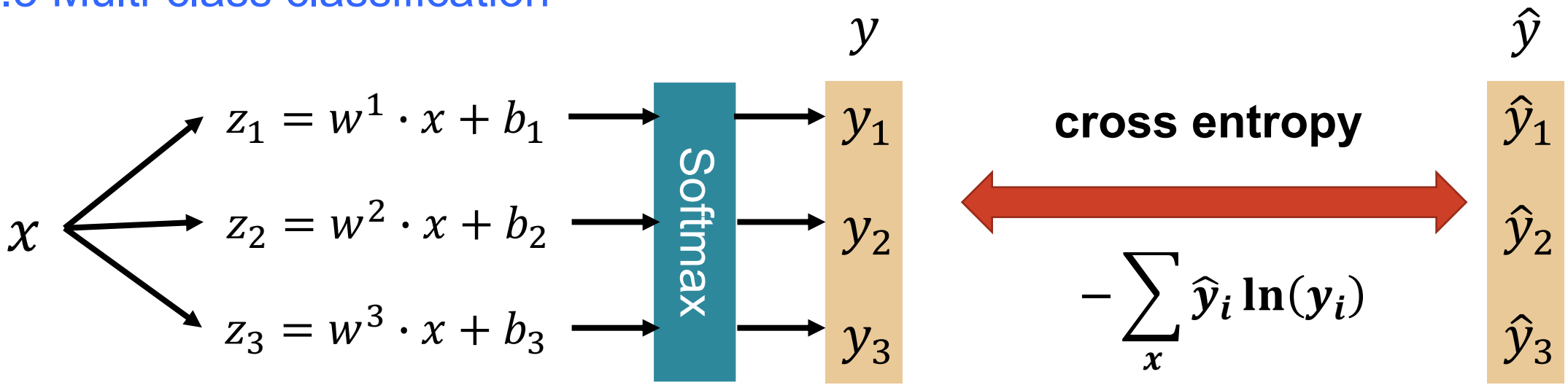
$C_2: w^2, b_2 \quad z_2 = w^2 \cdot x + b_2$

$C_3: w^3, b_3 \quad z_3 = w^3 \cdot x + b_3$



2. Generative & Discriminative

2.3 Multi-class classification



If $x \in$ class 1

$$\hat{y} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

If $x \in$ class 2

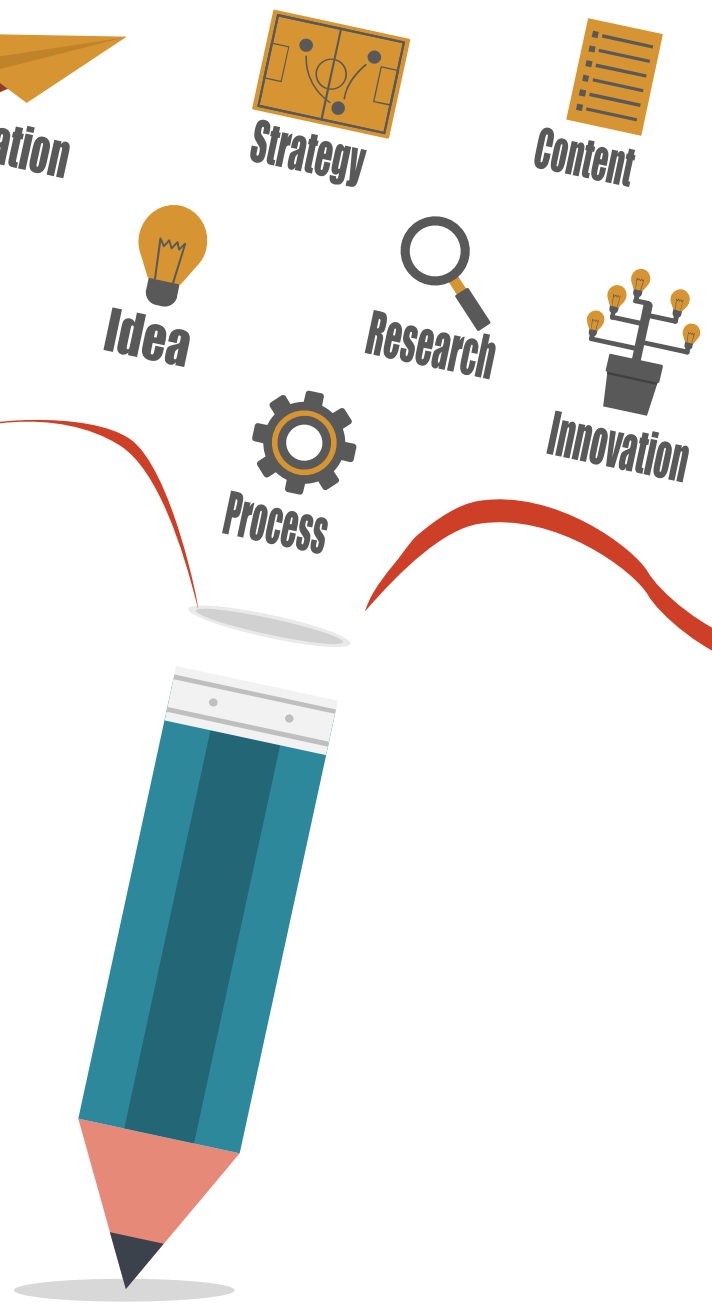
$$\hat{y} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

If $x \in$ class 3

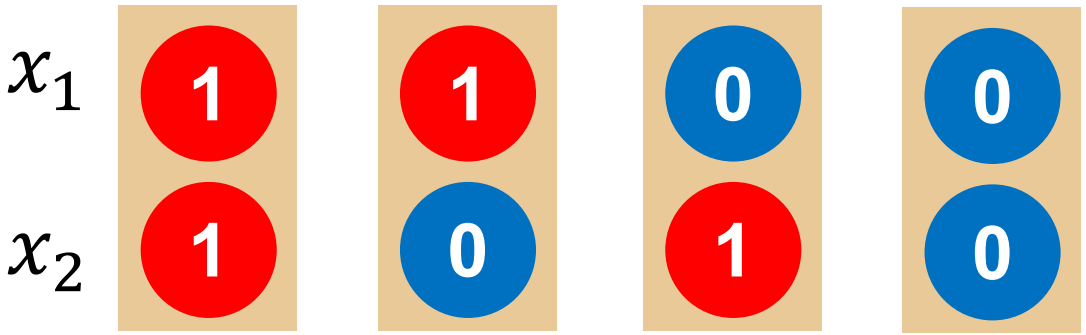
$$\hat{y} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

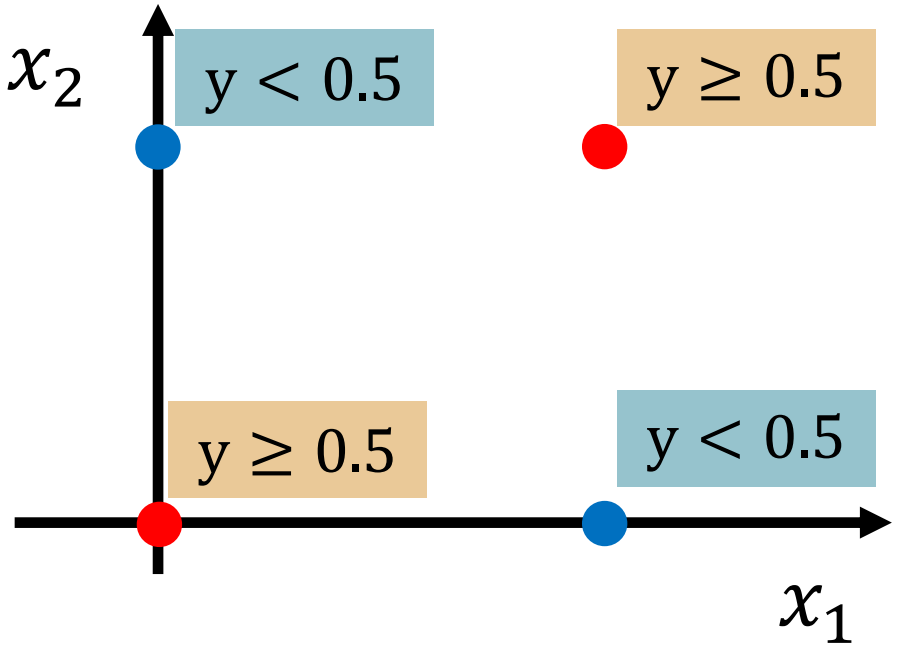
Limitation of Logistic Regression

03

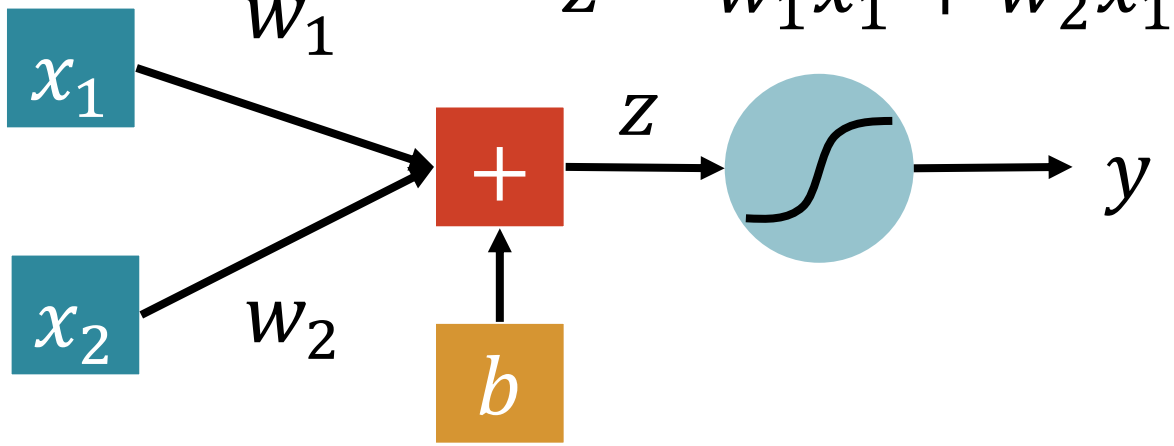


3. Limitation of Logistic Regression



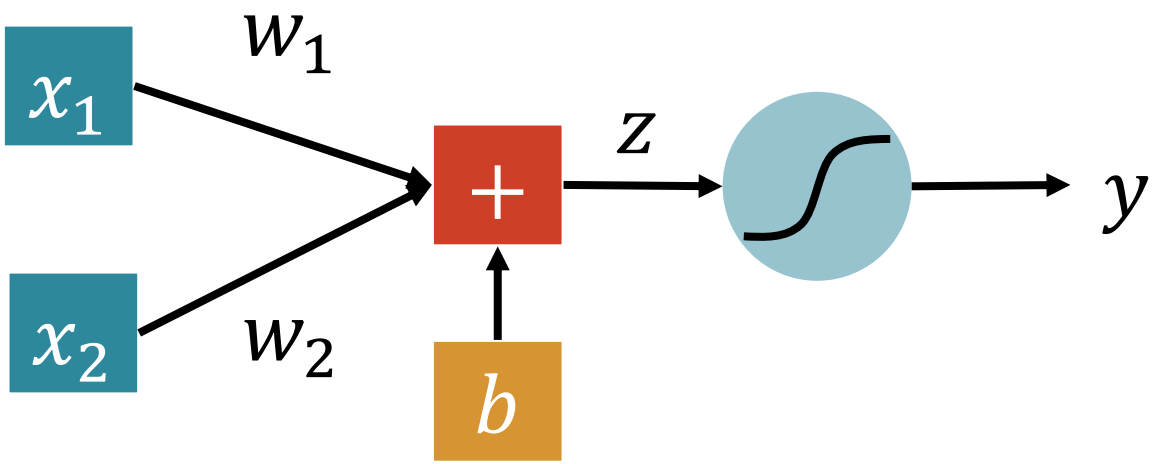


$$z = w_1 x_1 + w_2 x_1 + b$$

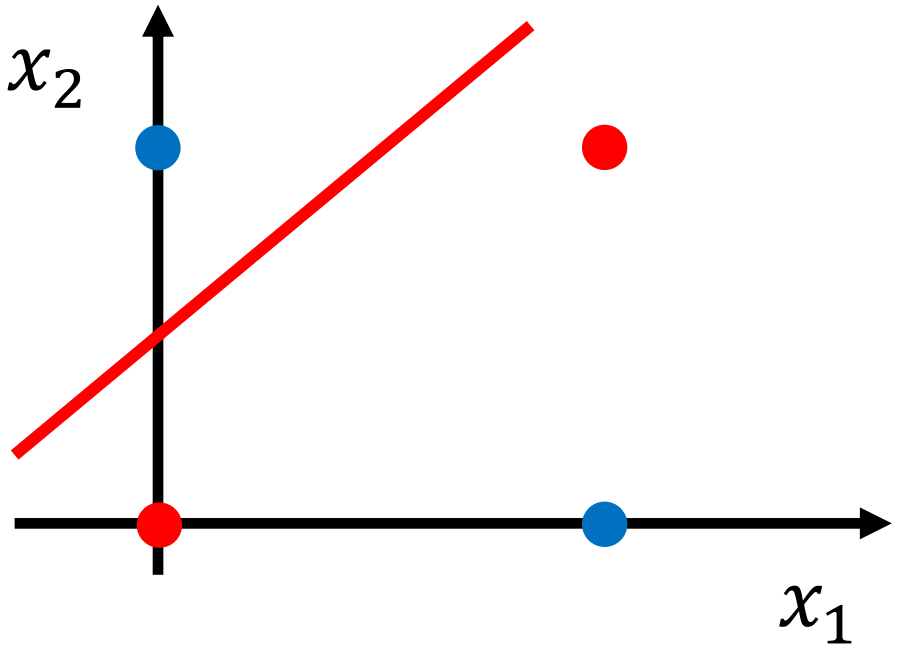


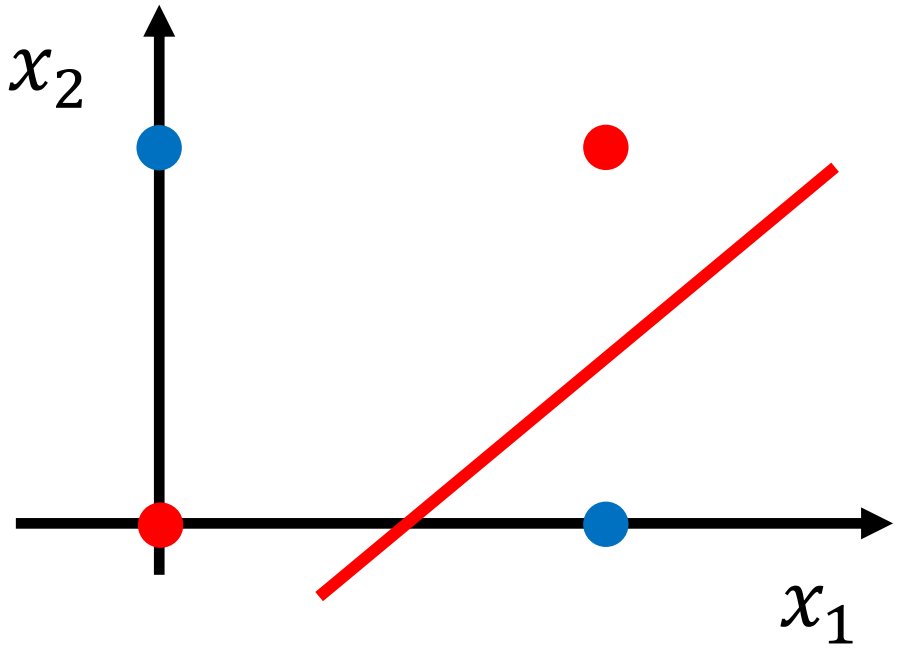
$$\begin{cases} y \geq 0.5 & \text{Output} = \text{class 1} \\ y < 0.5 & \text{Output} = \text{class 2} \end{cases}$$

3. Limitation of Logistic Regression



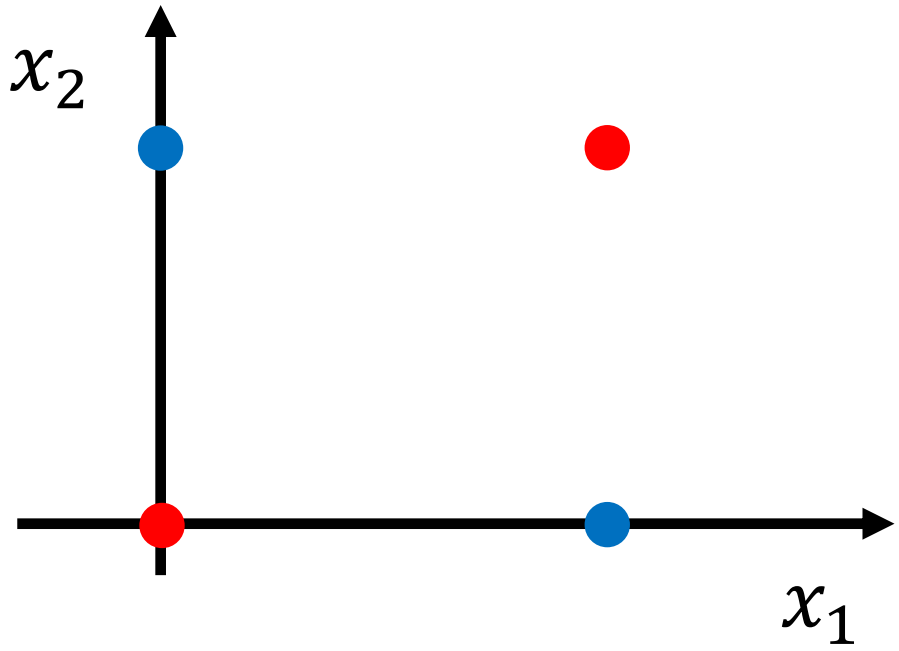
Sorry, we can't…





3. Limitation of Logistic Regression

How to classify this ?



3. Limitation of Logistic Regression

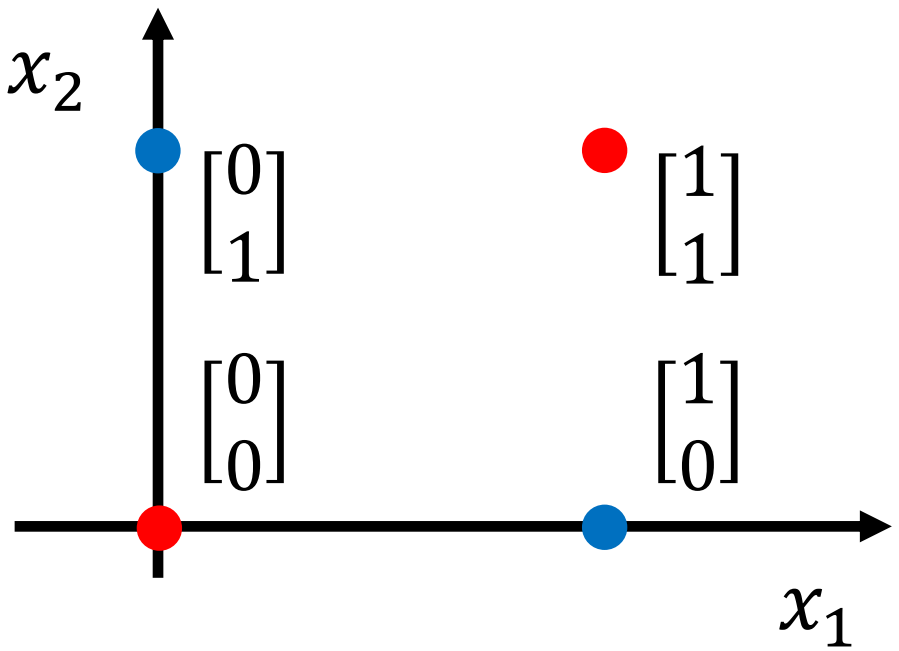
Solution 1: Feature Transformation

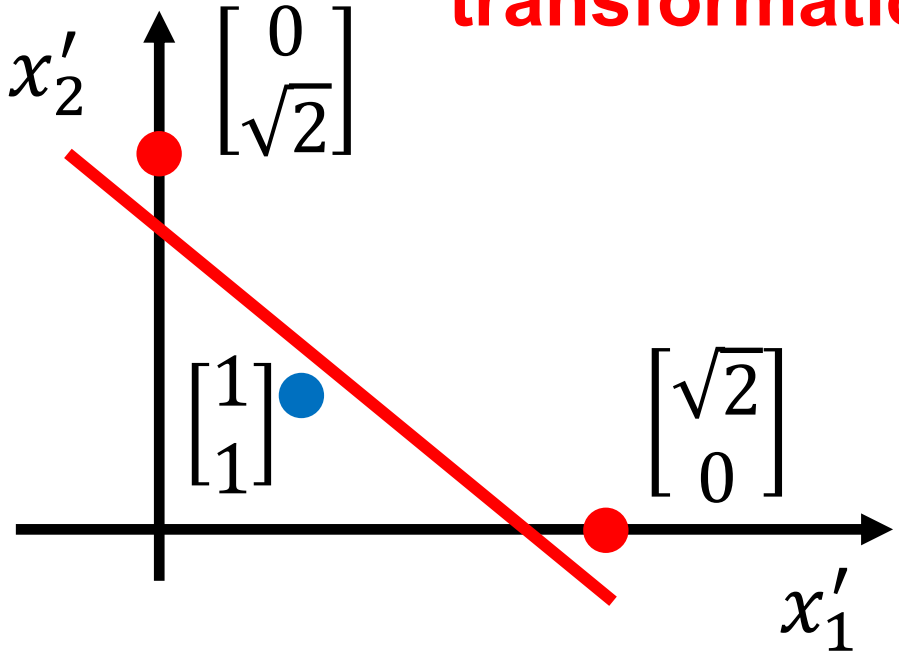
$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \rightarrow \begin{bmatrix} x_1' \\ x_2' \end{bmatrix}$$

x_1': distance to $\begin{bmatrix} 0 \\ 0 \end{bmatrix}$

x_2': distance to $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$

Not always easy to find a good transformation

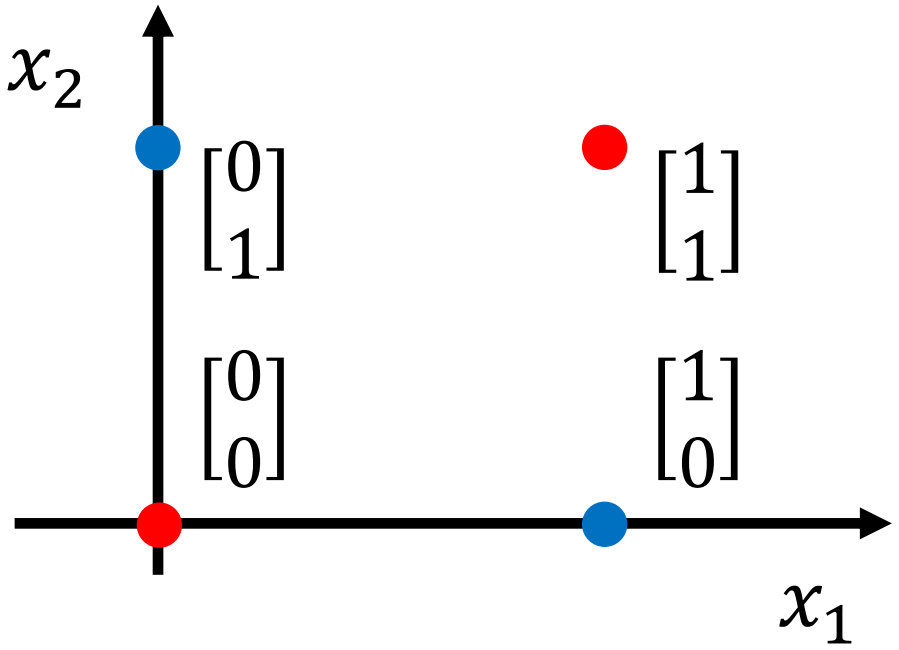


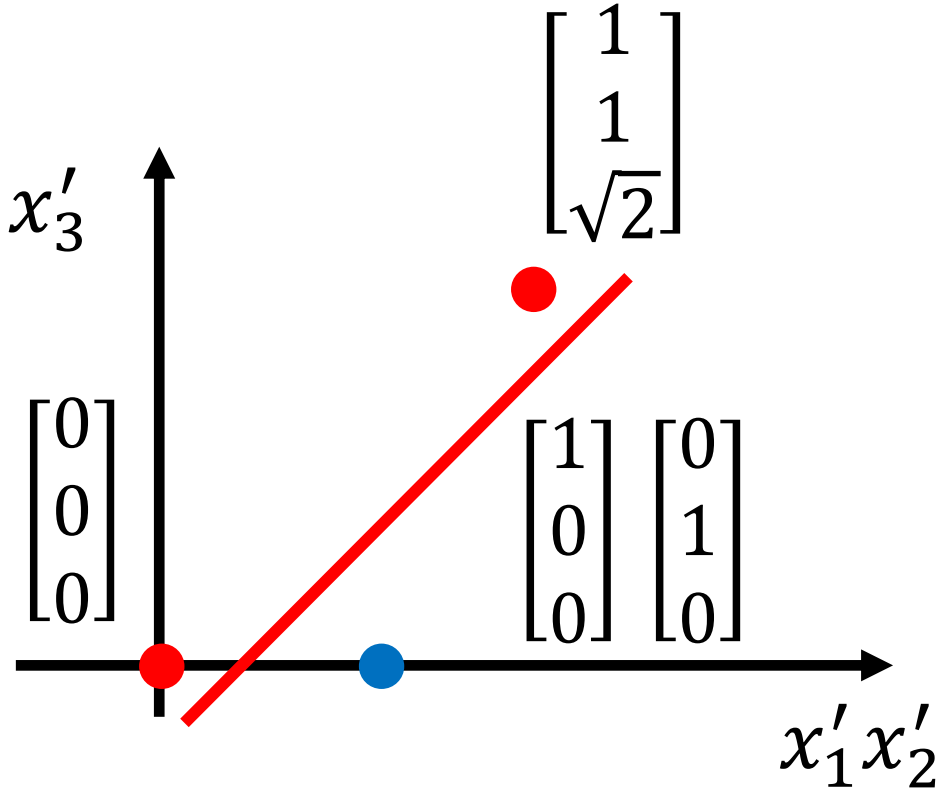


3. Limitation of Logistic Regression

Solution 2: Map data to higher dimension: SVM

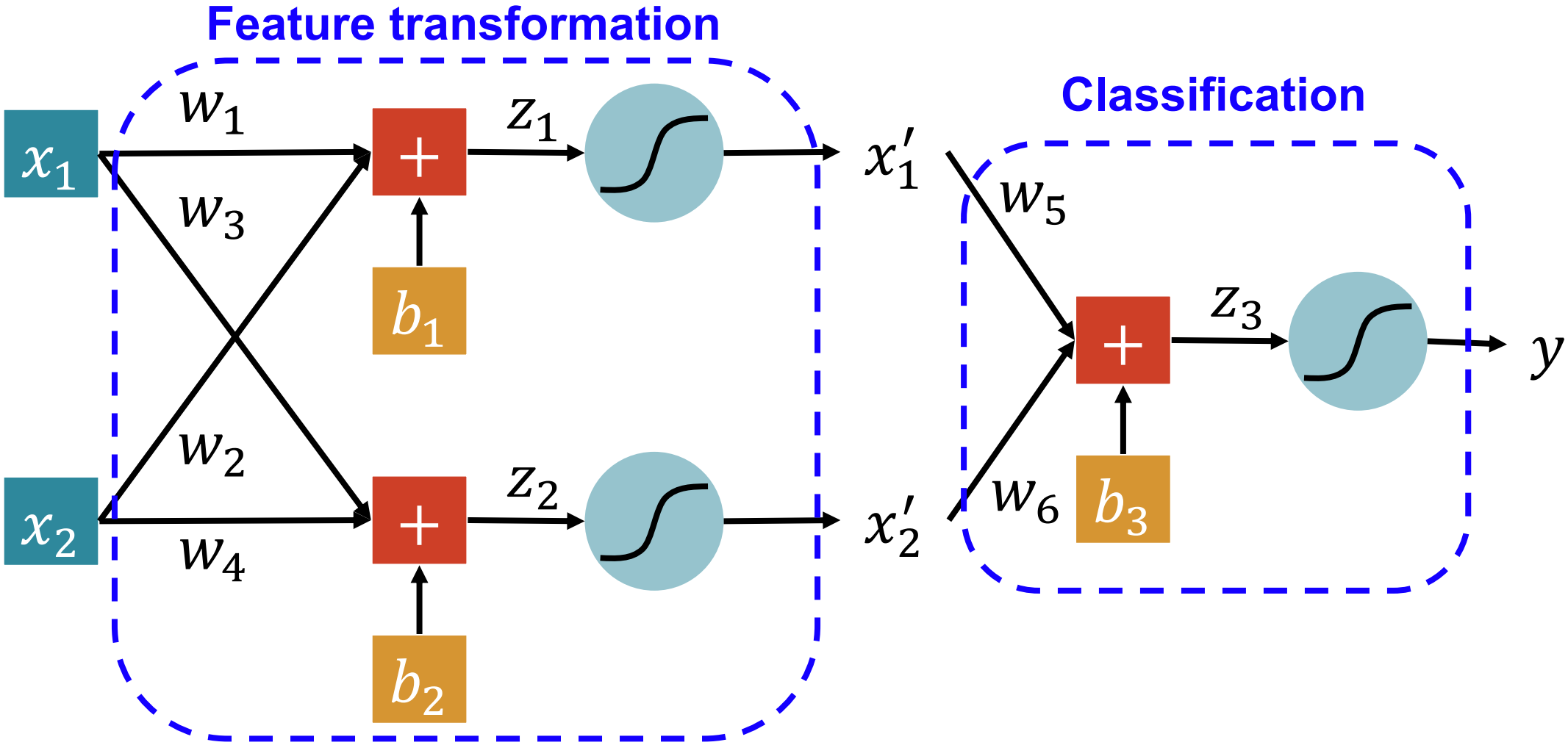
$$\Phi: \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \rightarrow \begin{bmatrix} x_1' \\ x_2' \\ x_3' \end{bmatrix} = \begin{bmatrix} x_1^2 \\ x_2^2 \\ \sqrt{2}x_1x_2 \end{bmatrix} \quad \mathbb{R}^2 \rightarrow \mathbb{R}^3$$





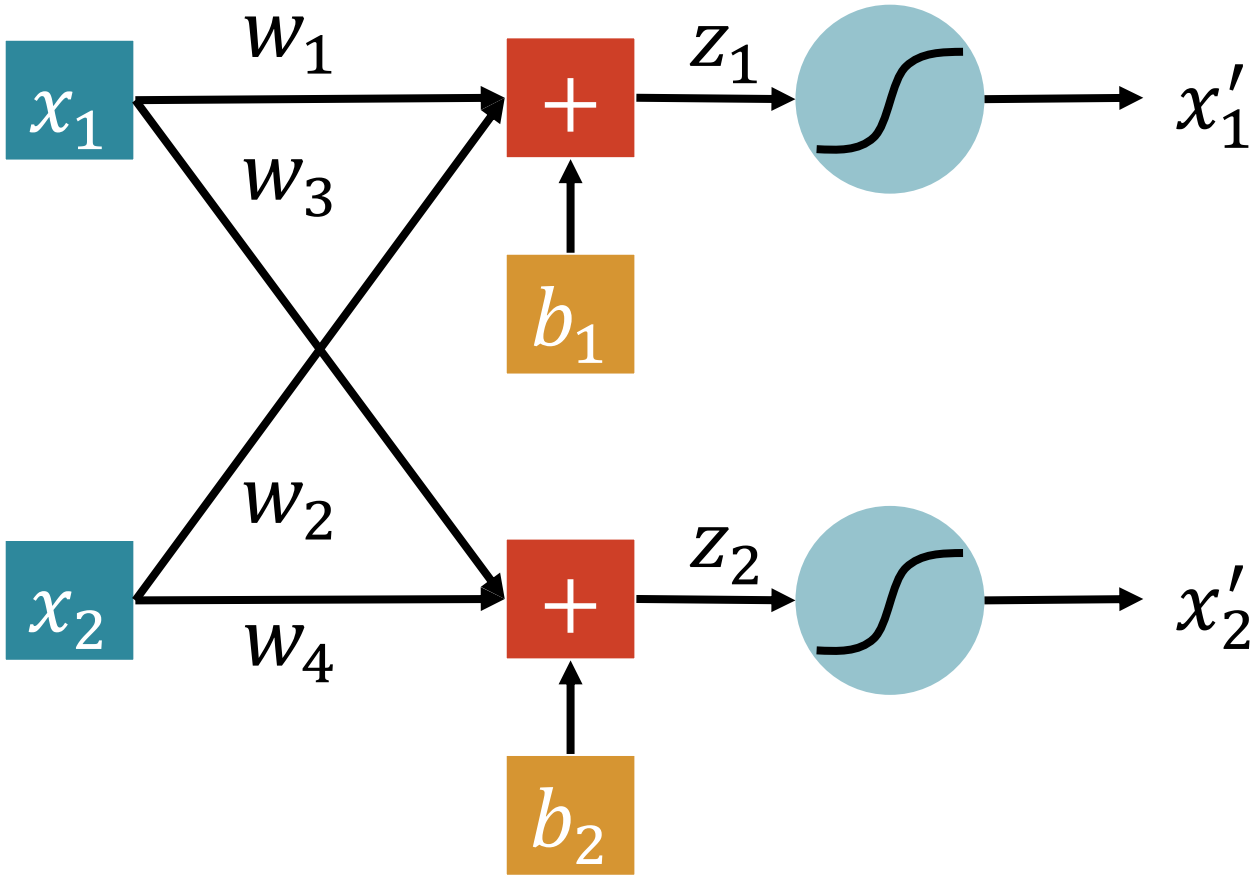
3. Limitation of Logistic Regression

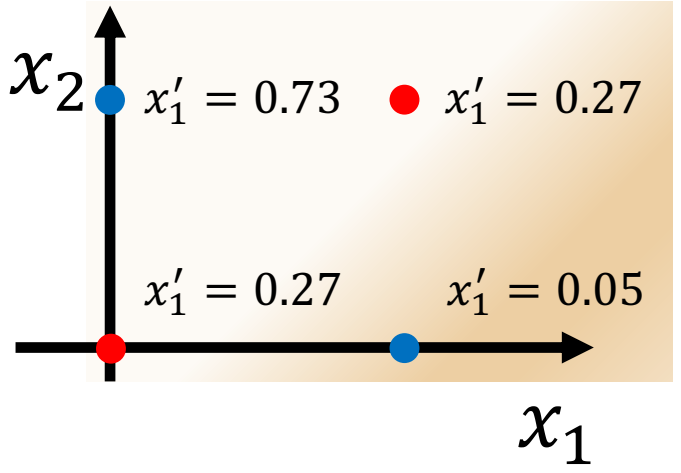
Solution 3: Cascading logistic regression models

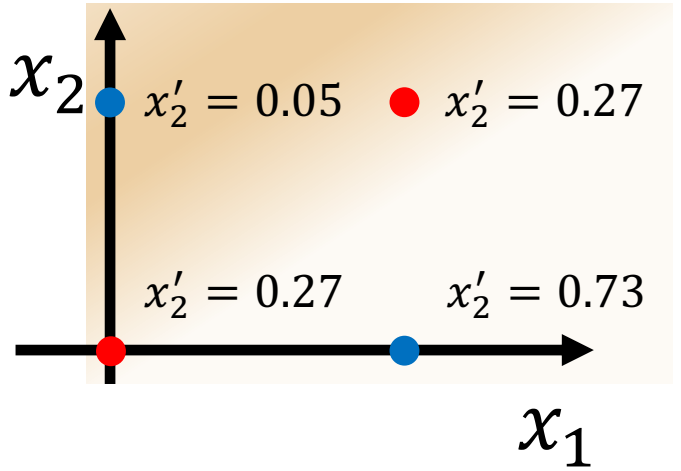


3. Limitation of Logistic Regression

Solution 3: Cascading logistic regression models

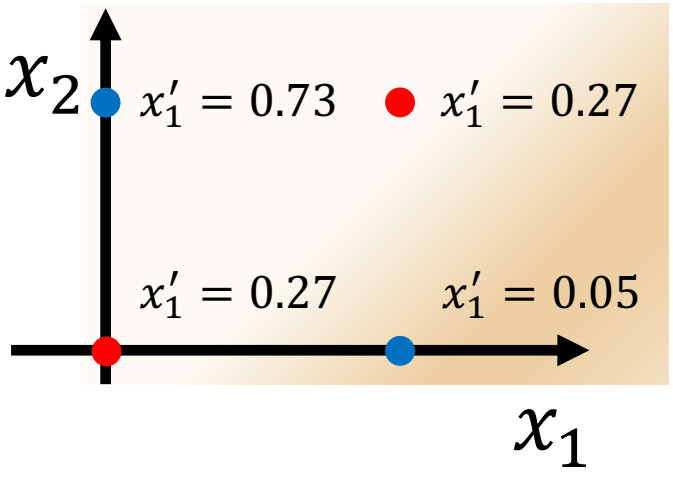


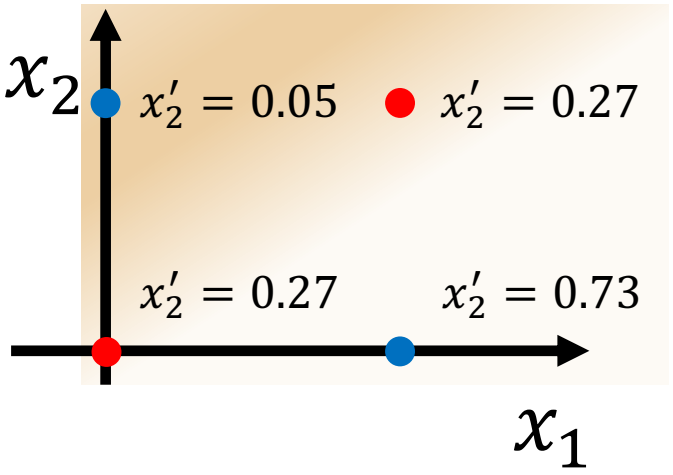


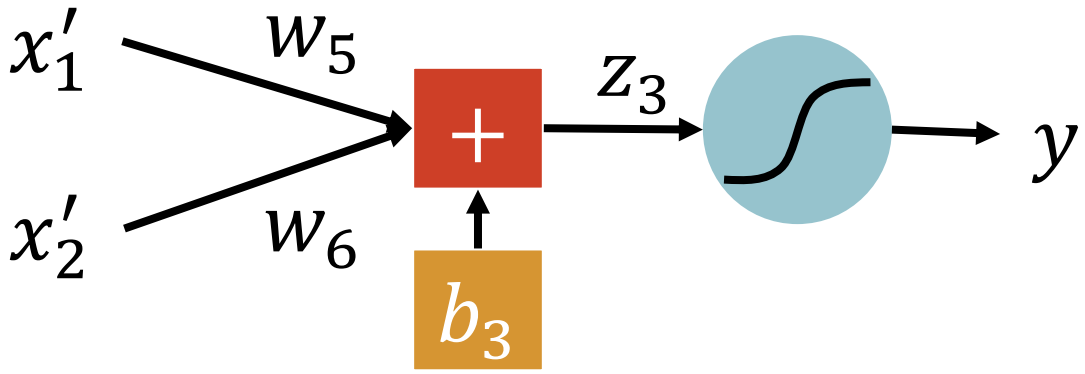


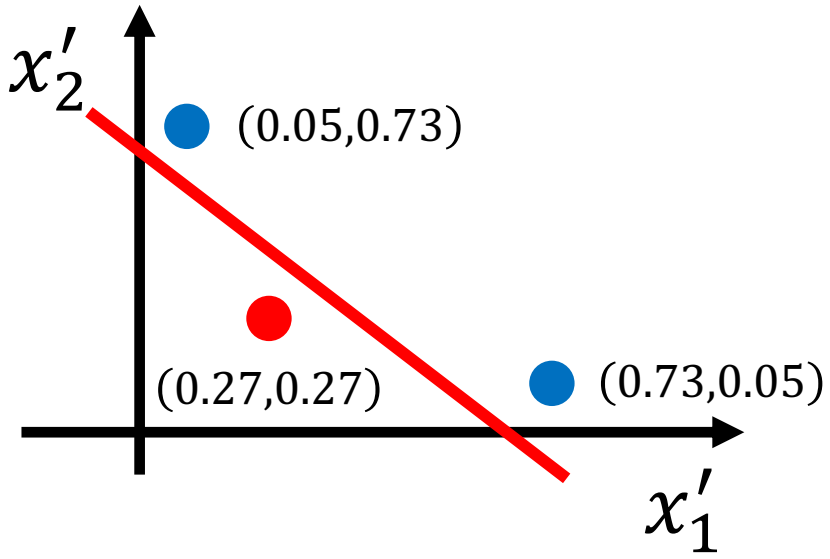
3. Limitation of Logistic Regression

Solution 3: Cascading logistic regression models

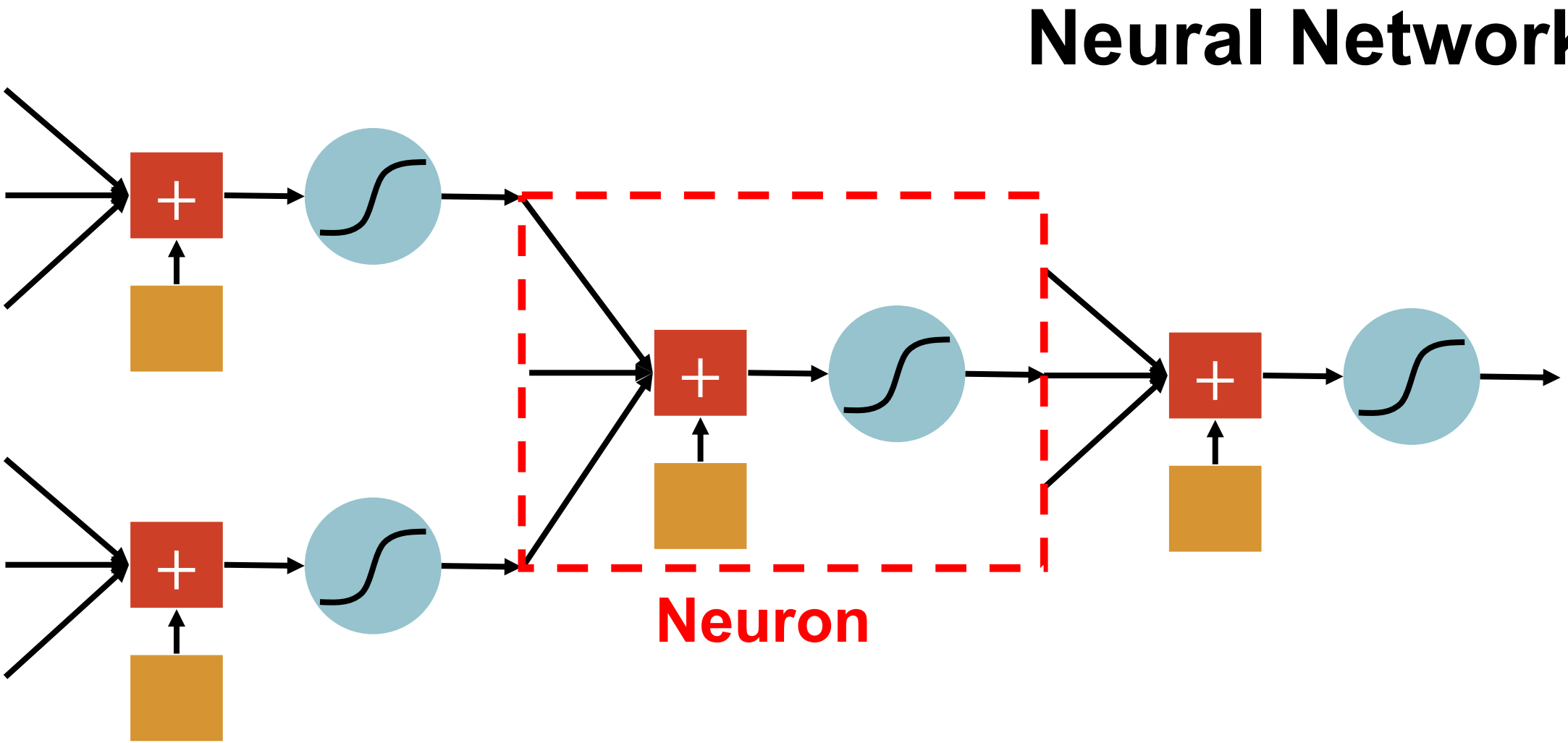








3. Limitation of Logistic Regression



Logistic Regression

Hao, Qi
School of Astronomy and Space Science

THANKS