

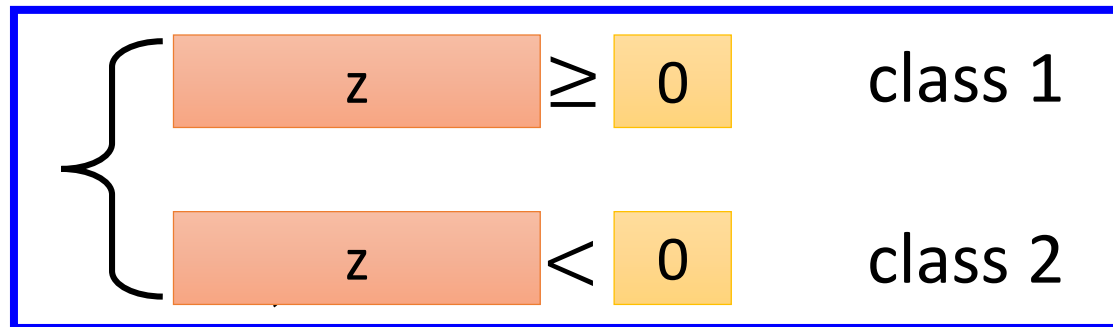
Classification: Logistic Regression

Hung-yi Lee

李宏毅

Step 1: Function Set

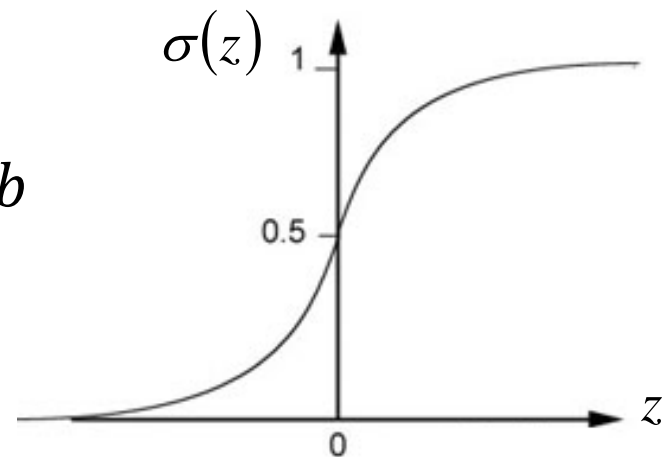
Function set: Including all different w and b



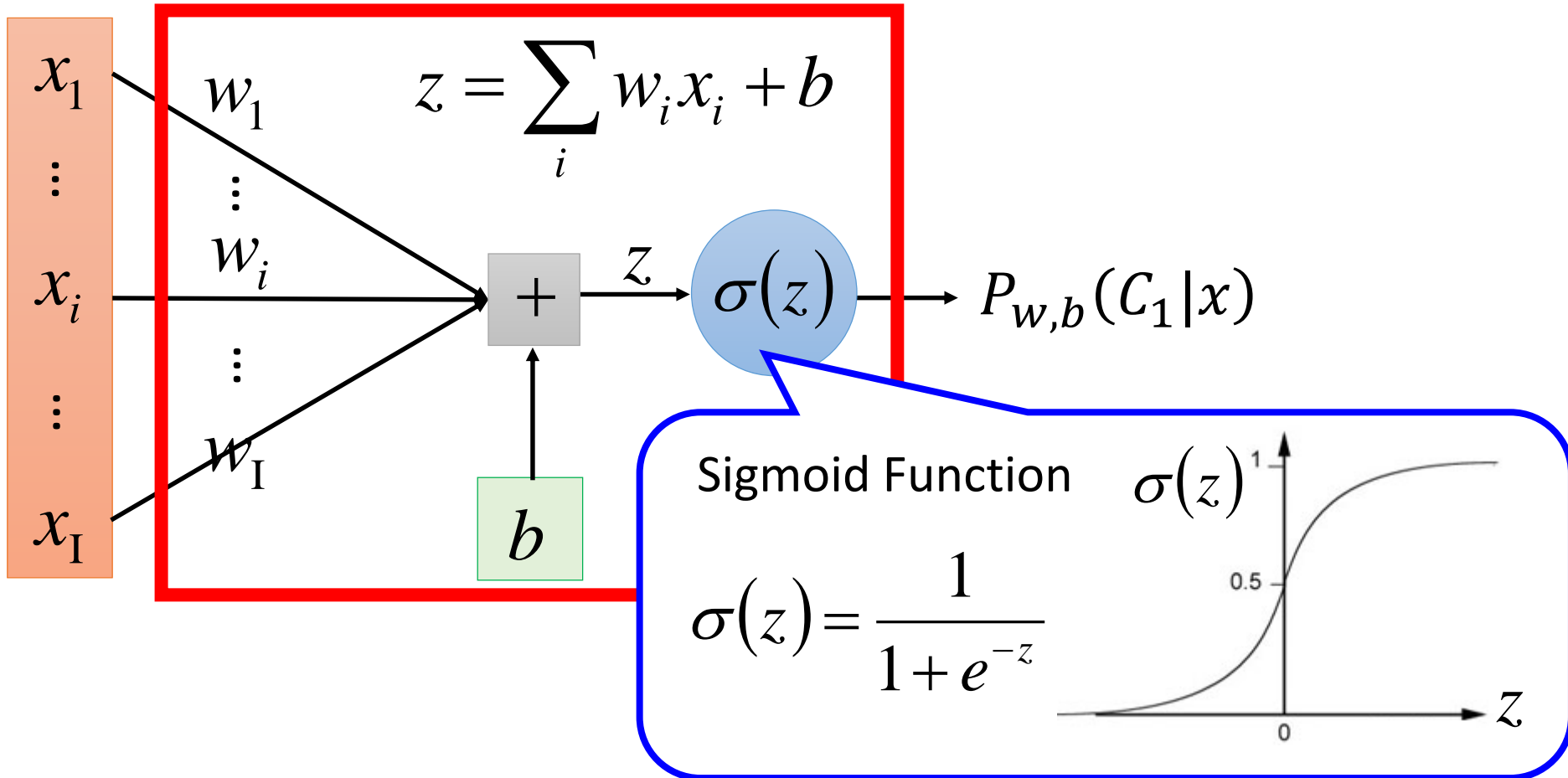
$$P_{w,b}(C_1|x) = \sigma(z)$$

$$z = w \cdot x + b = \sum_i w_i x_i + b$$

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



Step 1: Function Set



Step 2: Goodness of a Function

Training Data	x^1	x^2	x^3	...	x^N
	C_1	C_1	C_2	...	C_1

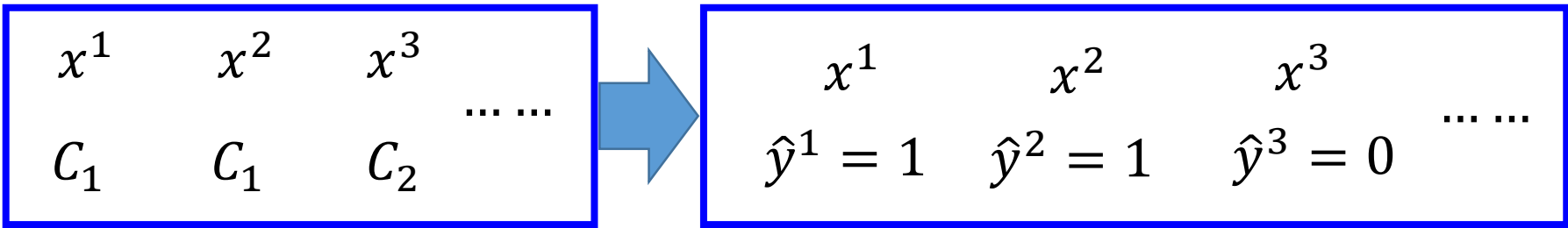
Assume the data is generated based on $f_{w,b}(x) = P_{w,b}(C_1|x)$

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

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

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

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



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

$$L(w, b) = f_{w,b}(x^1) f_{w,b}(x^2) (1 - f_{w,b}(x^3)) \dots$$

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

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

$$= -\ln f_{w,b}(x^1) \quad \Rightarrow \quad -\left[\boxed{1} \ln f(x^1) + \boxed{0} \ln(1 - f(x^1)) \right]$$

$$-\ln f_{w,b}(x^2) \quad \Rightarrow \quad -\left[\boxed{1} \ln f(x^2) + \boxed{0} \ln(1 - f(x^2)) \right]$$

$$-\ln(1 - f_{w,b}(x^3)) \quad \Rightarrow \quad -\left[\boxed{0} \ln f(x^3) + \boxed{1} \ln(1 - f(x^3)) \right]$$

⋮

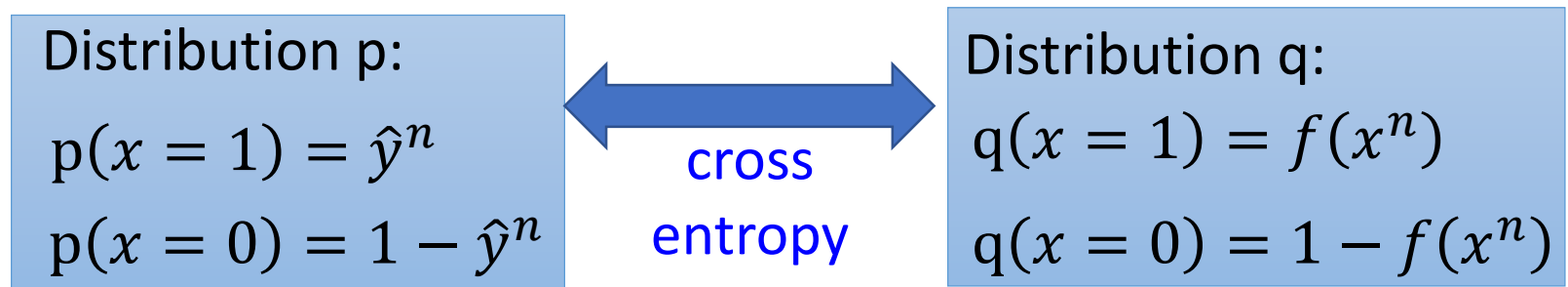
Step 2: Goodness of a Function

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

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

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

$$= \sum_n \underbrace{-\left[\hat{y}^n \ln f_{w,b}(x^n) + (1 - \hat{y}^n) \ln \left(1 - f_{w,b}(x^n)\right)\right]}_{\text{Cross entropy between two Bernoulli distribution}}$$



$$H(p, q) = - \sum_x p(x) \ln(q(x))$$

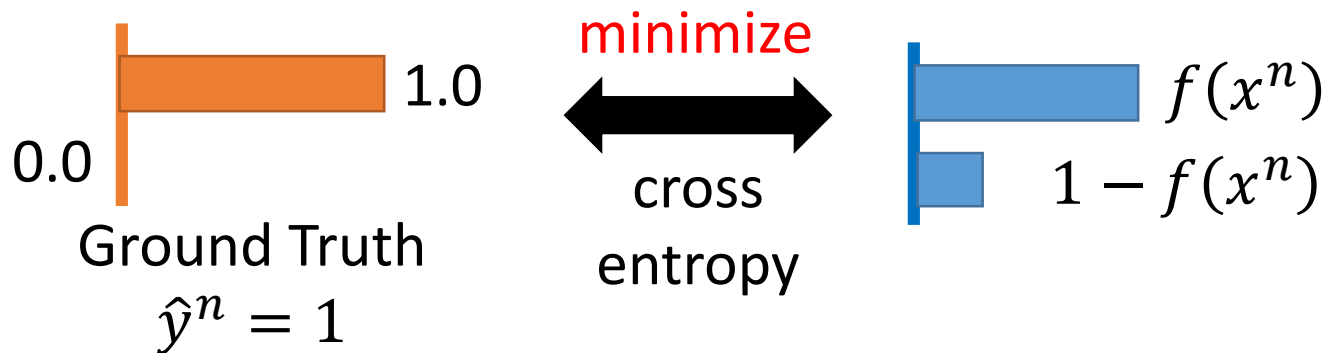
Step 2: Goodness of a Function

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

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

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

$$= \sum_n \underbrace{- \left[\hat{y}^n \ln f_{w,b}(x^n) + (1 - \hat{y}^n) \ln \left(1 - f_{w,b}(x^n)\right) \right]}_{\text{Cross entropy between two Bernoulli distribution}}$$

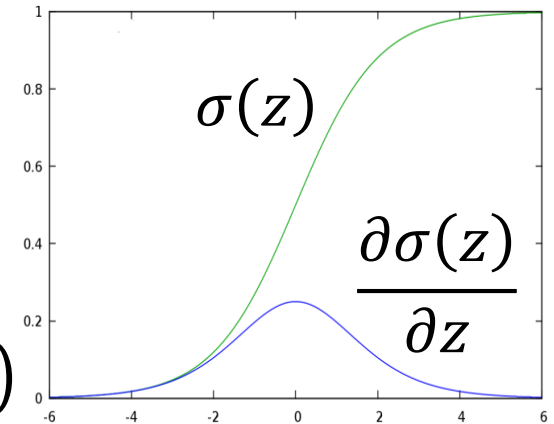


Step 3: Find the best function

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

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

$$\frac{\partial \ln \sigma(z)}{\partial z} = \frac{1}{\sigma(z)} \frac{\partial \sigma(z)}{\partial z} = \frac{1}{\cancel{\sigma(z)}} \cancel{\sigma(z)} (1 - \sigma(z))$$



$$f_{w,b}(x) = \frac{\sigma(z)}{1} = \frac{1}{1 + \exp(-z)}$$

$$z = w \cdot x + b = \sum_i w_i x_i + b$$

Step 3: Find the best function

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

$$\frac{\partial \ln(1 - f_{w,b}(x))}{\partial w_i} = \frac{\partial \ln(1 - f_{w,b}(x))}{\partial z} \frac{\partial z}{\partial w_i} \quad \frac{\partial z}{\partial w_i} = x_i$$

$$\frac{\partial \ln(1 - \sigma(z))}{\partial z} = -\frac{1}{1 - \sigma(z)} \frac{\partial \sigma(z)}{\partial z} = -\frac{1}{1 - \sigma(z)} \sigma(z) (1 - \sigma(z))$$

$$\begin{aligned} f_{w,b}(x) &= \sigma(z) \\ &= 1 / (1 + \exp(-z)) \end{aligned}$$

$$z = w \cdot x + b = \sum_i w_i x_i + b$$

Step 3: Find the best function

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

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

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

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

Larger difference, larger update

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

Logistic Regression + Square Error

Step 1: $f_{w,b}(x) = \sigma \left(\sum_i w_i x_i + b \right)$

Step 2: Training data: (x^n, \hat{y}^n) , \hat{y}^n : **1** for class 1, **0** for class 2

$$L(f) = \frac{1}{2} \sum_n (f_{w,b}(x^n) - \hat{y}^n)^2$$

Step 3:

$$\begin{aligned} \frac{\partial (f_{w,b}(x^n) - \hat{y}^n)^2}{\partial w_i} &= 2(f_{w,b}(x^n) - \hat{y}^n) \frac{\partial f_{w,b}(x^n)}{\partial z} \frac{\partial z}{\partial w_i} \\ &= 2(f_{w,b}(x^n) - \hat{y}^n) f_{w,b}(x^n) (1 - f_{w,b}(x^n)) x_i \end{aligned}$$

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

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

Logistic Regression + Square Error

Step 1: $f_{w,b}(x) = \sigma \left(\sum_i w_i x_i + b \right)$


Step 2: Training data: (x^n, \hat{y}^n) , \hat{y}^n : **1** for class 1, **0** for class 2

$$L(f) = \frac{1}{2} \sum_n (f_{w,b}(x^n) - \hat{y}^n)^2$$

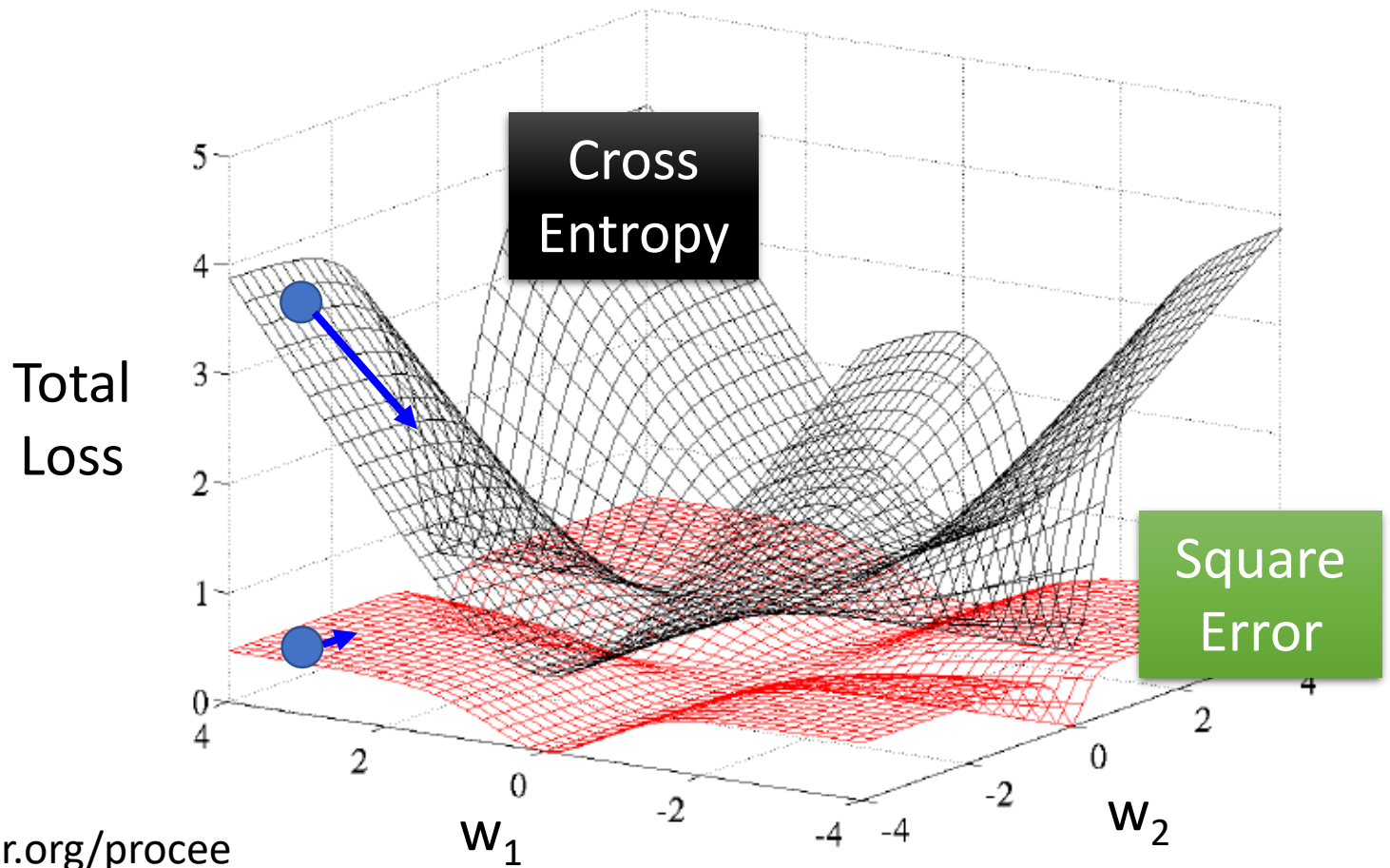
Step 3:

$$\begin{aligned} \frac{\partial (f_{w,b}(x^n) - \hat{y}^n)^2}{\partial w_i} &= 2(f_{w,b}(x^n) - \hat{y}^n) \frac{\partial f_{w,b}(x^n)}{\partial z} \frac{\partial z}{\partial w_i} \\ &= 2(f_{w,b}(x^n) - \hat{y}^n) f_{w,b}(x^n) (1 - f_{w,b}(x^n)) x_i \end{aligned}$$

$\hat{y}^n = 0$ 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)  $\partial L / \partial w_i = 0$

Cross Entropy v.s. Square Error



<http://jmlr.org/proceedings/papers/v9/glorot10a/glorot10a.pdf>

Logistic Regression

Step 1: $f_{w,b}(x) = \sigma \left(\sum_i w_i x_i + b \right)$

Output: between 0 and 1

Linear Regression

$$f_{w,b}(x) = \sum_i w_i x_i + b$$

Output: any value

Step 2:

Step 3:

Logistic Regression

Step 1: $f_{w,b}(x) = \sigma \left(\sum_i w_i x_i + b \right)$

Output: between 0 and 1

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

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

$$L(f) = \sum_n l(f(x^n), \hat{y}^n)$$

Linear Regression

$$f_{w,b}(x) = \sum_i w_i x_i + b$$

Output: any value

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

\hat{y}^n : a real number

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

Cross entropy:

$$l(f(x^n), \hat{y}^n) = -[\hat{y}^n \ln f(x^n) + (1 - \hat{y}^n) \ln(1 - f(x^n))]$$

Logistic Regression

Step 1: $f_{w,b}(x) = \sigma \left(\sum_i w_i x_i + b \right)$

Output: between 0 and 1

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

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

$$L(f) = \sum_n l(f(x^n), \hat{y}^n)$$

Linear Regression

$$f_{w,b}(x) = \sum_i w_i x_i + b$$

Output: any value

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

\hat{y}^n : a real number

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

Logistic regression: $w_i \leftarrow w_i - \eta \sum_n - \left(\hat{y}^n - f_{w,b}(x^n) \right) x_i^n$

Step 3:

Linear regression: $w_i \leftarrow w_i - \eta \sum_n - \left(\hat{y}^n - f_{w,b}(x^n) \right) x_i^n$

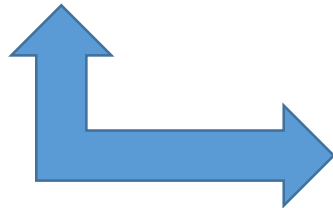
Discriminative v.s. Generative

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



directly find w and b

Find $\mu^1, \mu^2, \Sigma^{-1}$



Will we obtain the same set of w and b ?

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

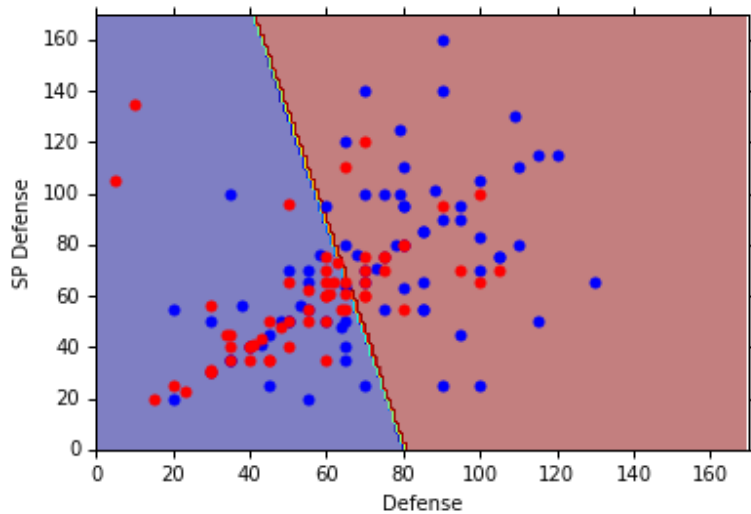
$$b = -\frac{1}{2} (\mu^1)^T (\Sigma^1)^{-1} \mu^1$$

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

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

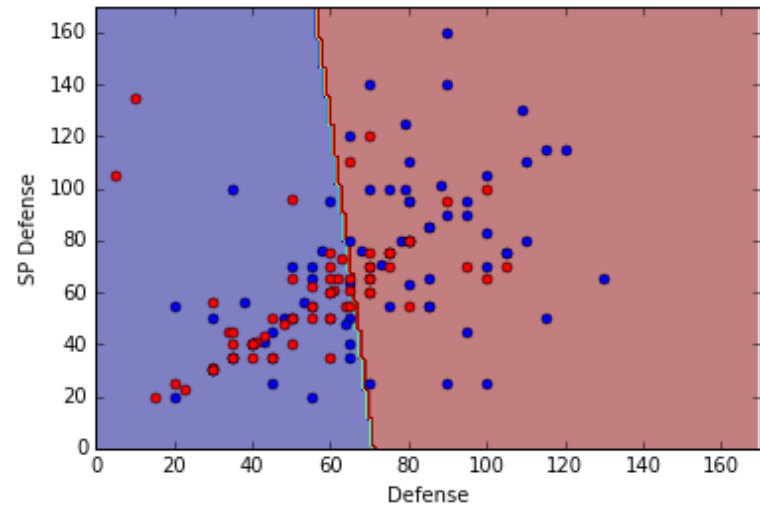
Generative v.s. Discriminative

Generative



73% accuracy

Discriminative

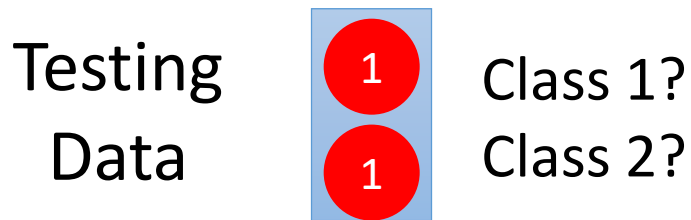
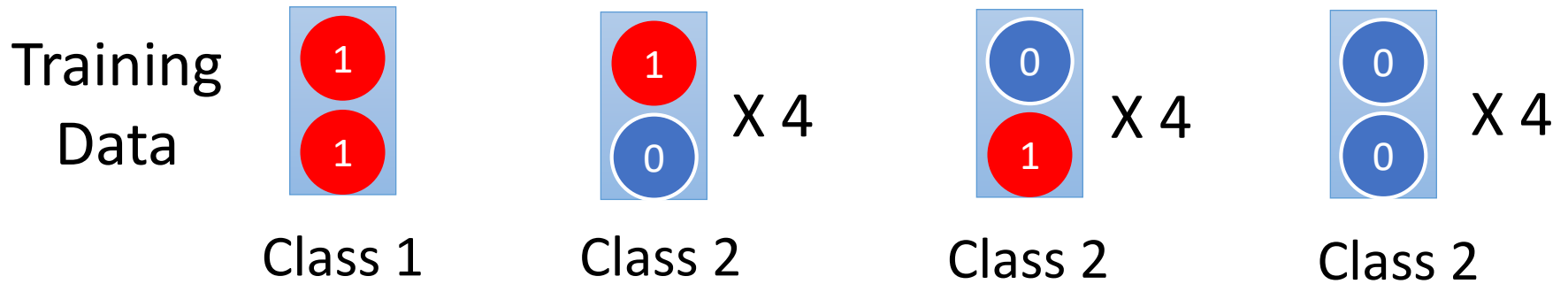


79% accuracy

All: hp, att, sp att, de, sp de, speed

Generative v.s. Discriminative

- Example

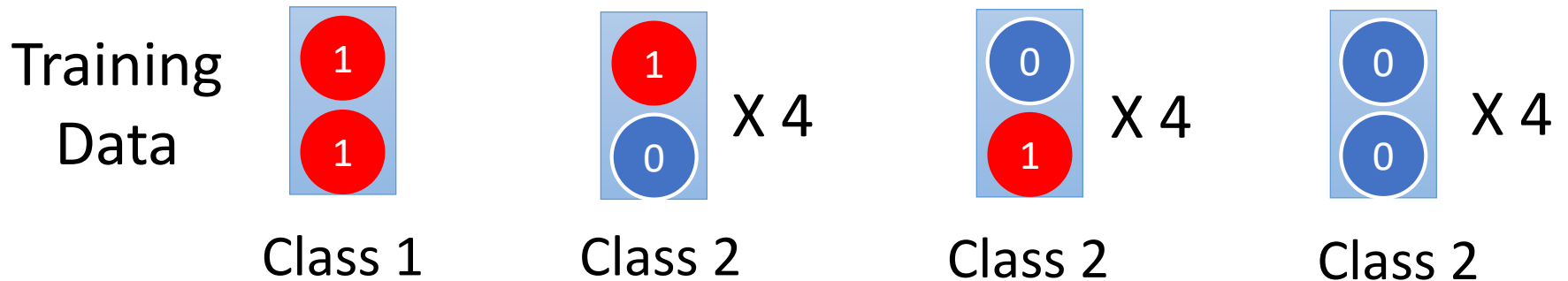


How about Naïve Bayes?

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

Generative v.s. Discriminative

- Example



$$P(C_1) = \frac{1}{13}$$

$$P(x_1 = 1|C_1) = 1$$

$$P(x_2 = 1|C_1) = 1$$

$$P(C_2) = \frac{12}{13}$$

$$P(x_1 = 1|C_2) = \frac{1}{3}$$

$$P(x_2 = 1|C_2) = \frac{1}{3}$$

Training Data



Class 1



Class 2

X 4



Class 2

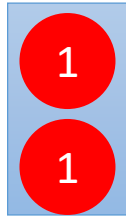
X 4



Class 2

X 4

Testing Data



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

< 0.5

Diagram illustrating the calculation of $P(C_1|x)$ for testing data $x = (1, 1)$. The numerator is 1×1 (pointing to $P(x|C_1)P(C_1)$) and the denominator is $1 \times 1 + \frac{1}{3} \times \frac{1}{3}$ (pointing to $P(x|C_1)P(C_1) + P(x|C_2)P(C_2)$). The denominator terms are further broken down: 1×1 (pointing to $P(x_1=1|C_1)P(C_1)$) and $\frac{1}{3} \times \frac{1}{3}$ (pointing to $P(x_1=1|C_2)P(C_2)$).

$$P(C_1) = \frac{1}{13}$$

$$P(x_1 = 1|C_1) = 1$$

$$P(x_2 = 1|C_1) = 1$$

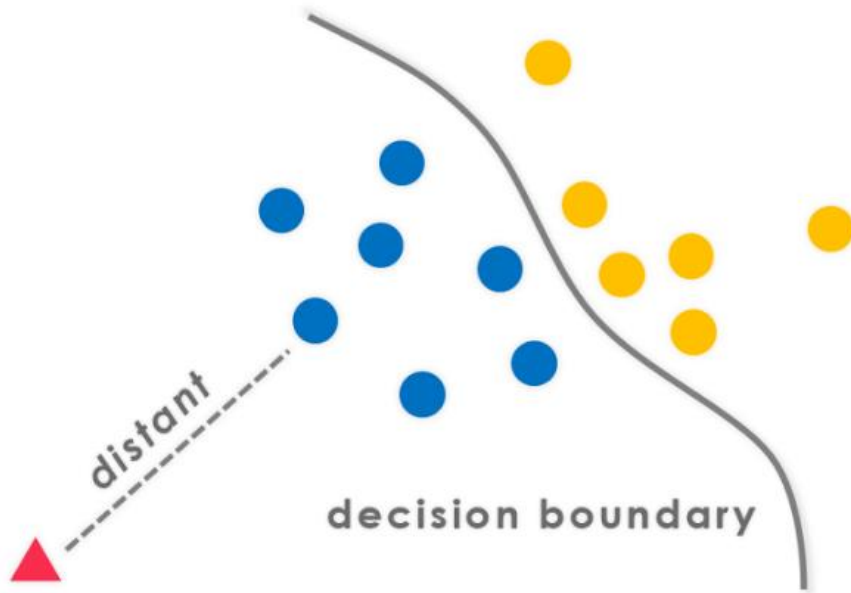
$$P(C_2) = \frac{12}{13}$$

$$P(x_1 = 1|C_2) = \frac{1}{3}$$

$$P(x_2 = 1|C_2) = \frac{1}{3}$$

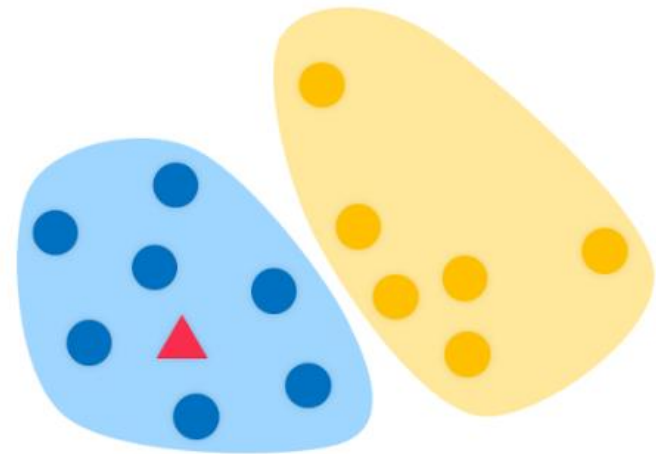
Generative v.s. Discriminative

Discriminative



Model the **decision boundary between classes.**

Generative



Models the **actual distribution of each class.**

Generative v.s. Discriminative

- Usually people believe discriminative model is better
- Benefit of generative model
 - With the assumption of probability distribution
 - less training data is needed
 - more robust to the noise
 - Priors and class-dependent probabilities can be estimated from different sources.

Multi-class Classification (3 classes as example)

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

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

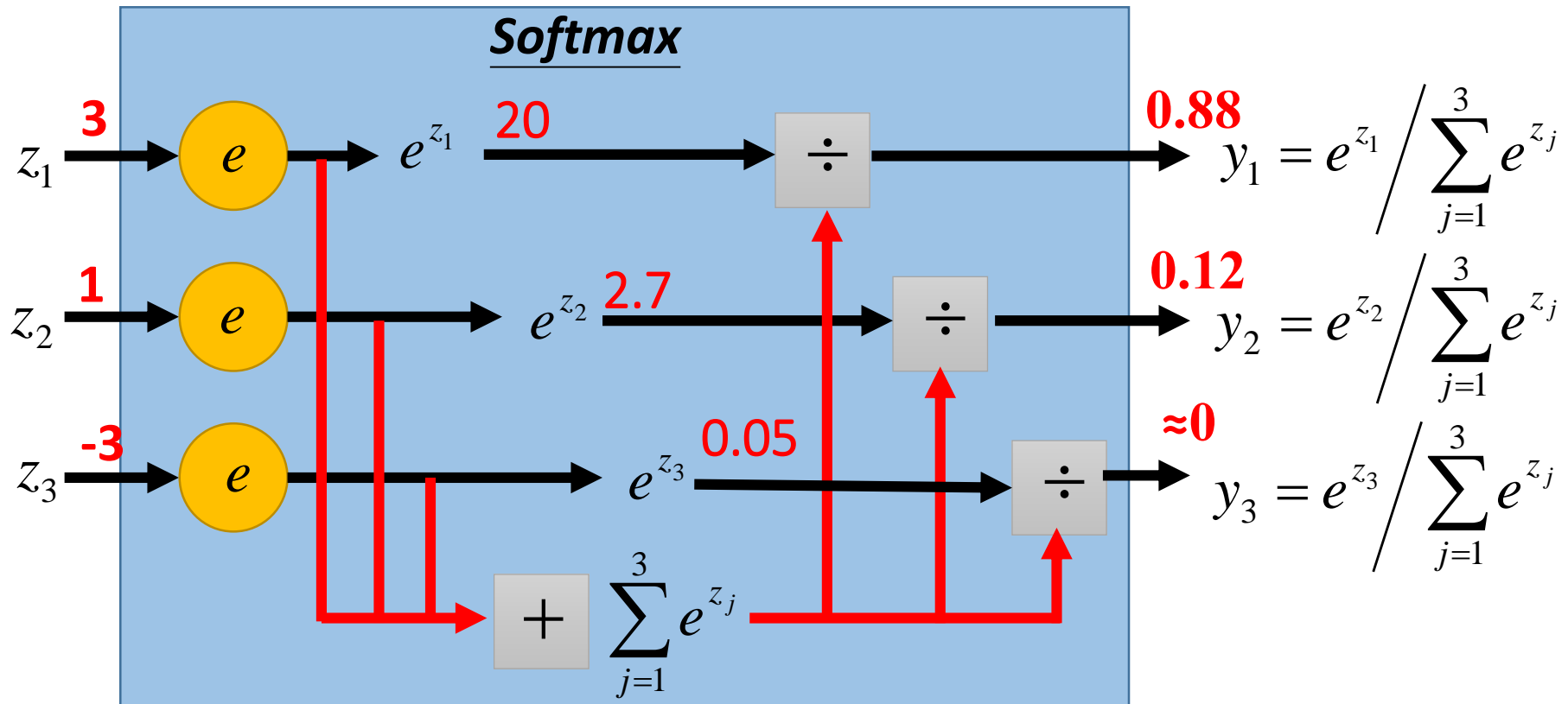
$$C_3: w^3, b_3 \quad z_3 = w^3 \cdot x + b_3$$

Probability:

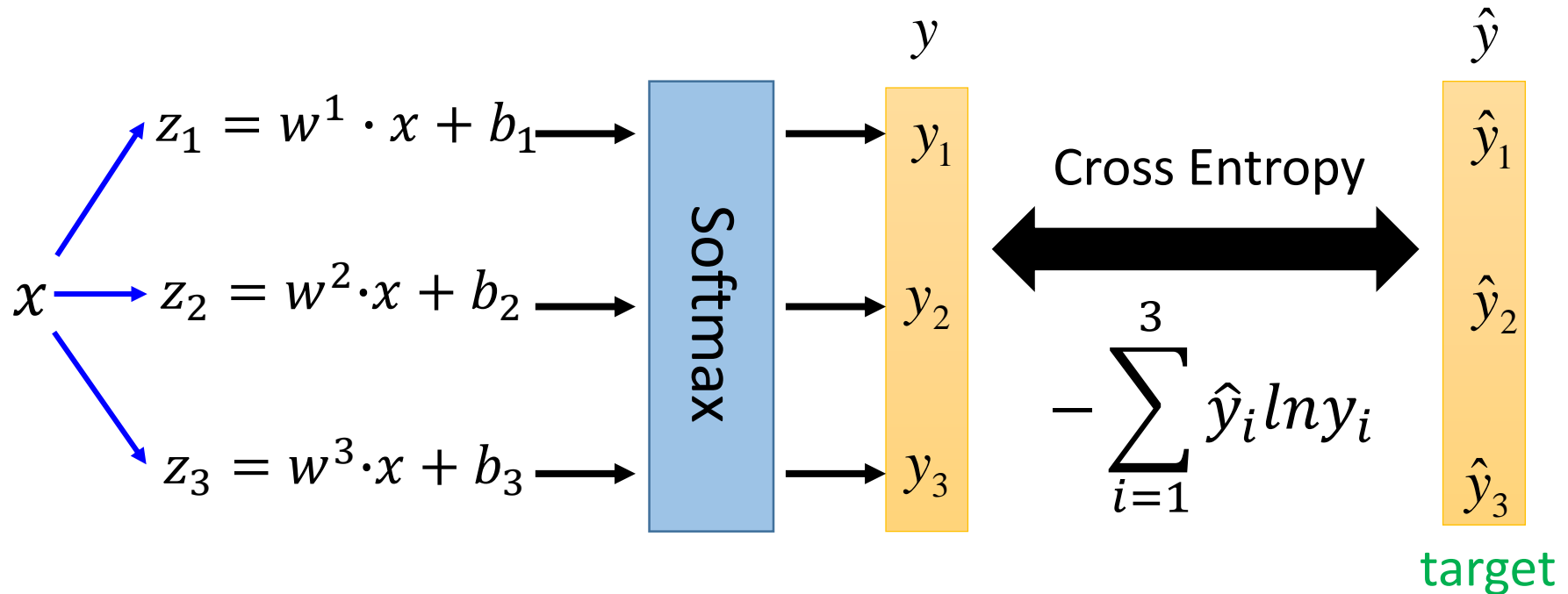
■ $1 > y_i > 0$

■ $\sum_i y_i = 1$

$$y_i = P(C_i | x)$$



Multi-class Classification (3 classes as example)



If $x \in$ class 1

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

$$-\ln y_1$$

If $x \in$ class 2

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

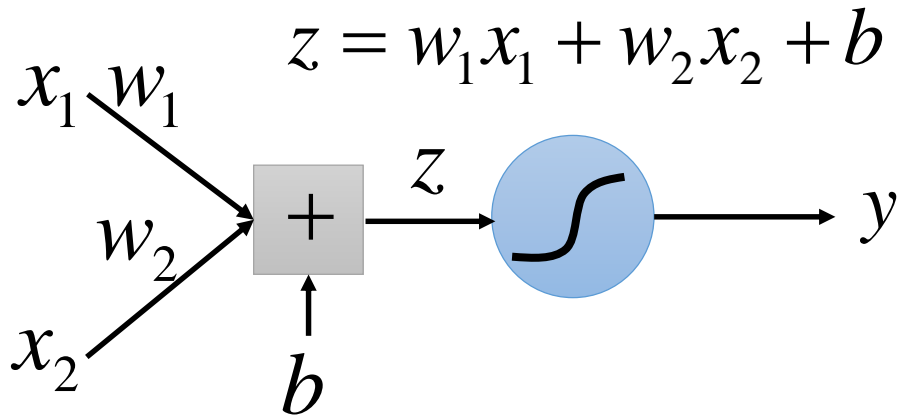
$$-\ln y_2$$

If $x \in$ class 3

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

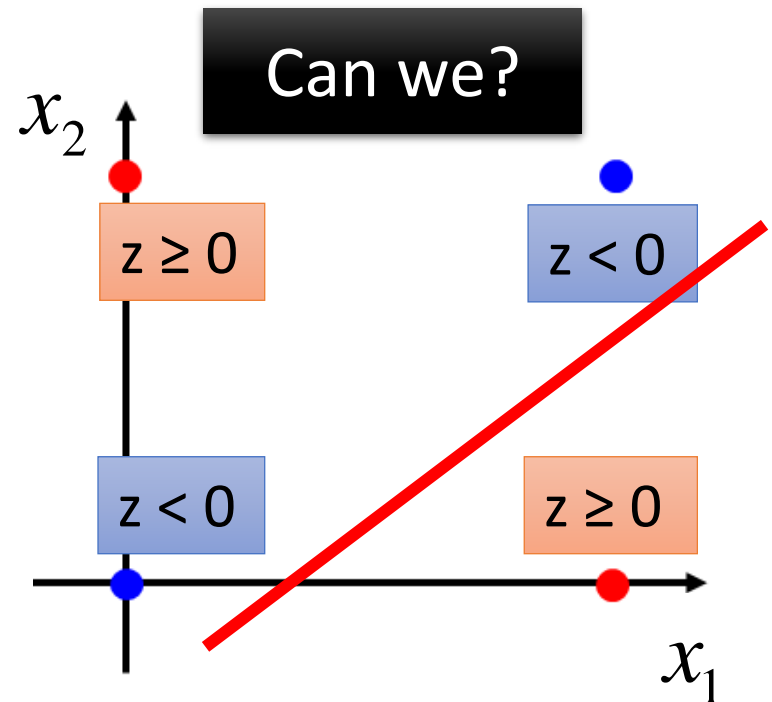
$$-\ln y_3$$

Limitation of Logistic Regression



$\begin{cases} \text{Class 1} & y \geq 0.5 & (z \geq 0) \\ \text{Class 2} & y < 0.5 & (z < 0) \end{cases}$

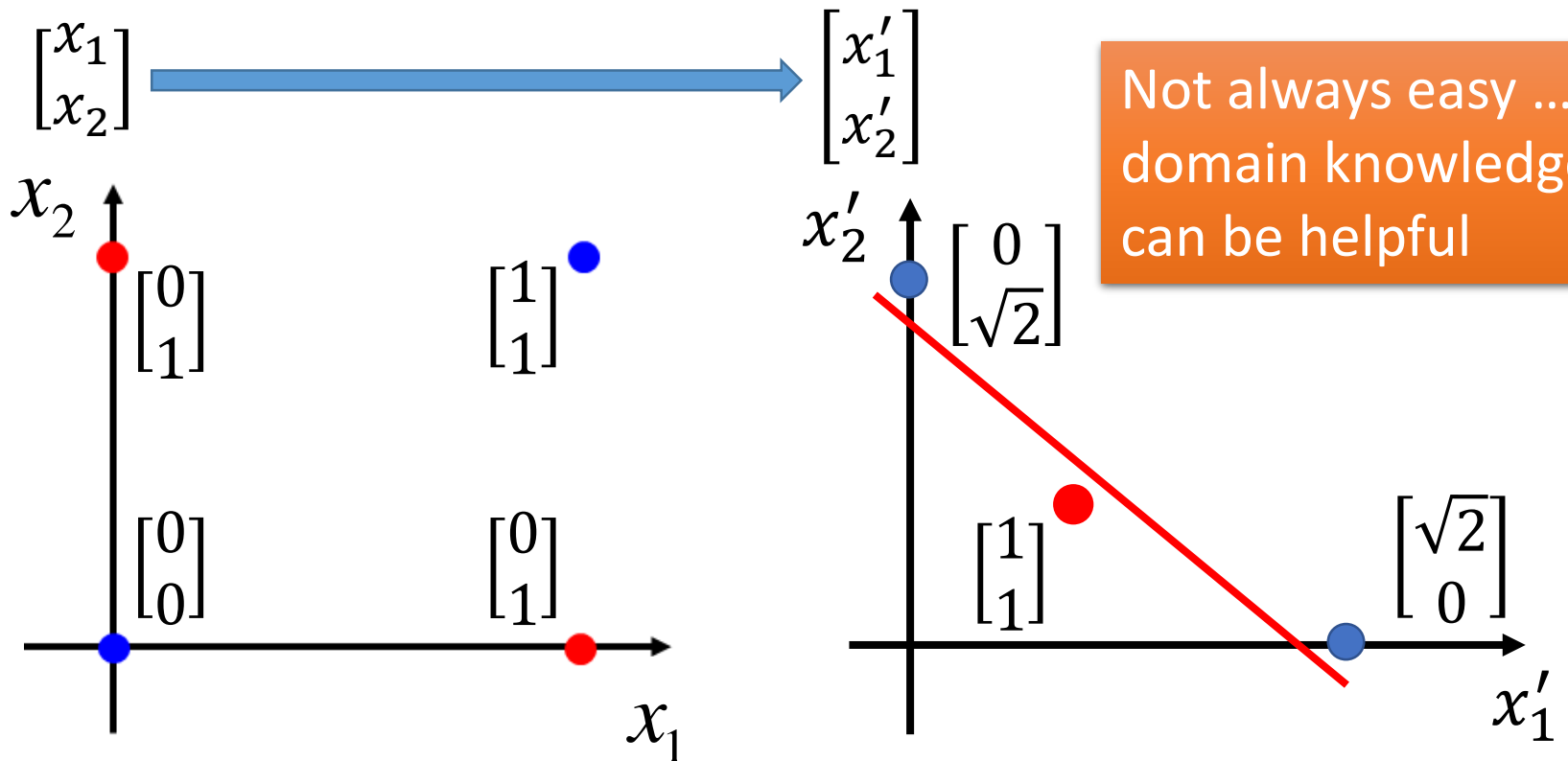
Input Feature		Label
x_1	x_2	
0	0	Class 2
0	1	Class 1
1	0	Class 1
1	1	Class 2



Limitation of Logistic Regression

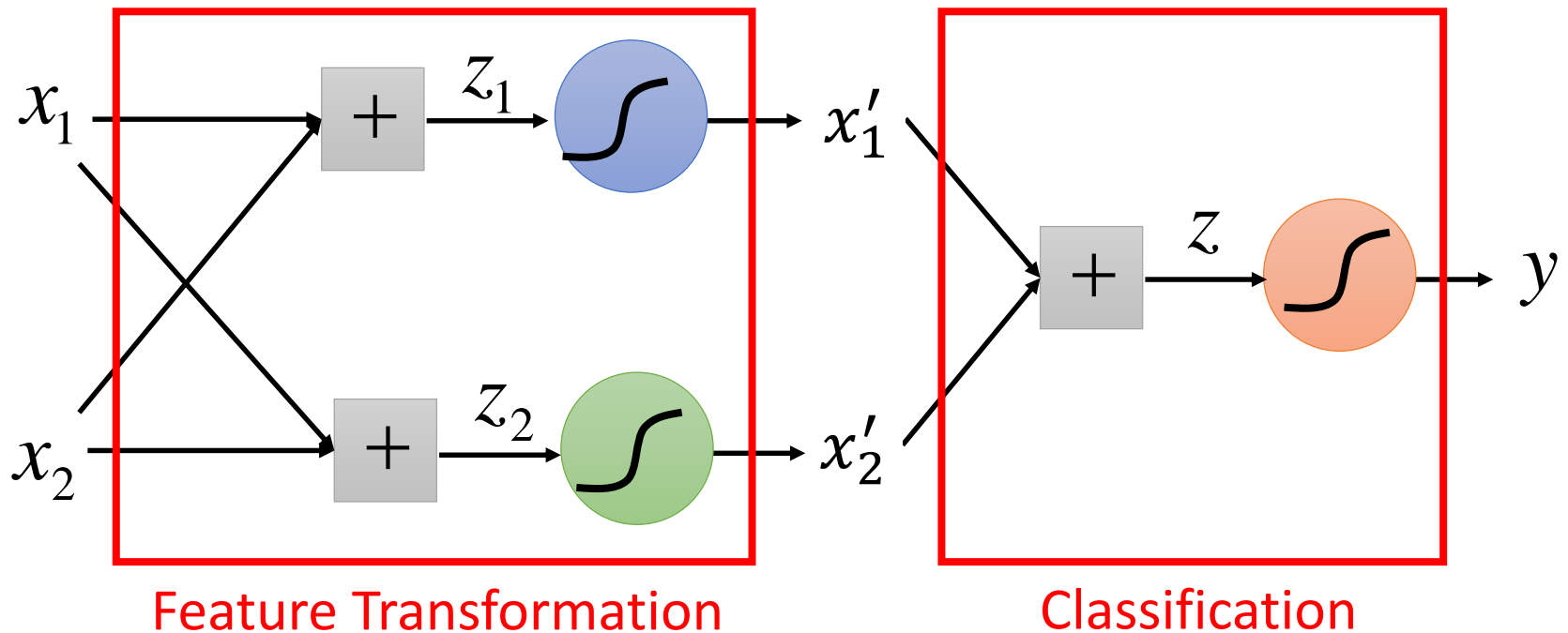
- Feature transformation

x'_1 : distance to $\begin{bmatrix} 0 \\ 0 \end{bmatrix}$
 x'_2 : distance to $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$

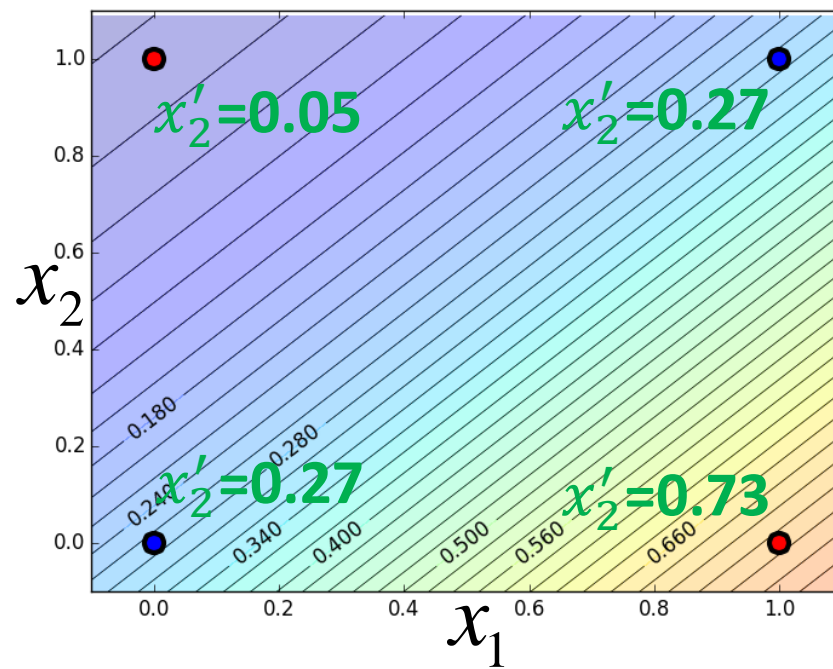
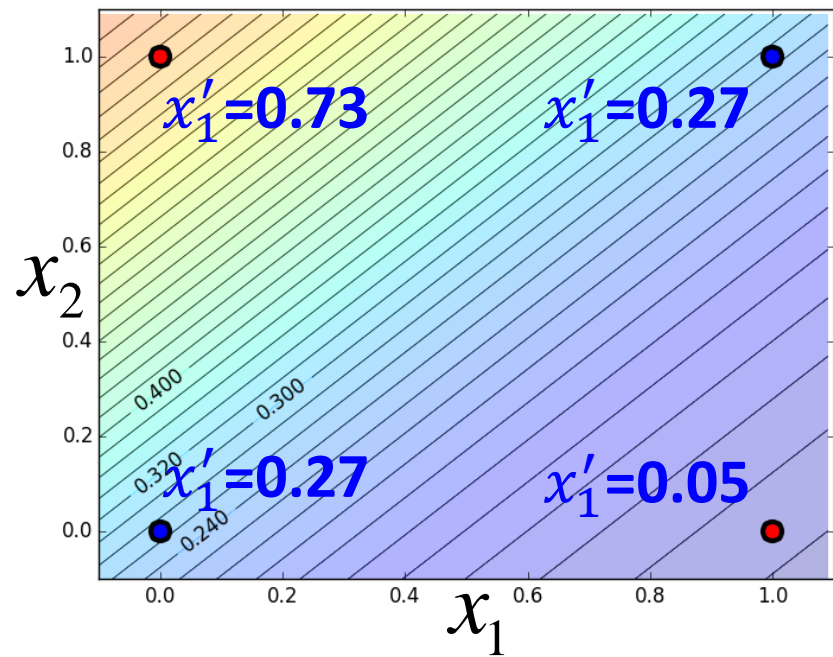
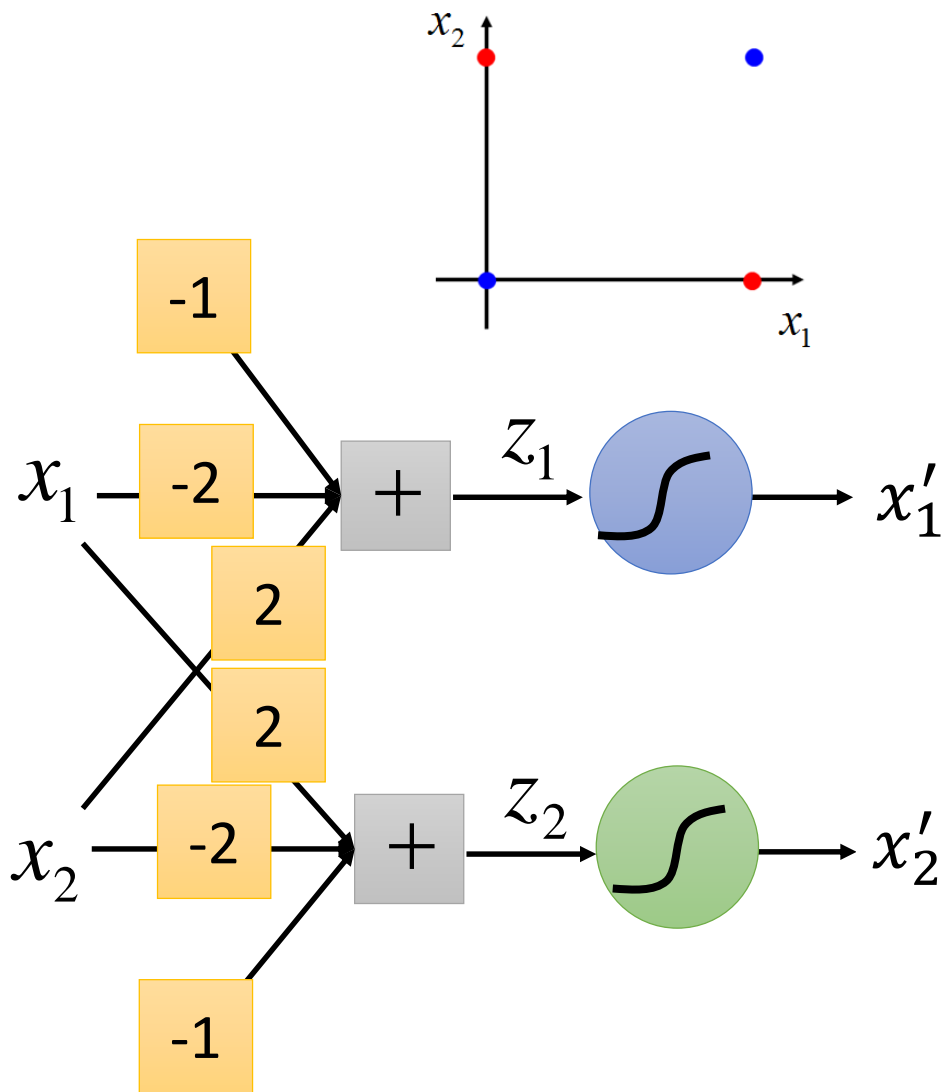


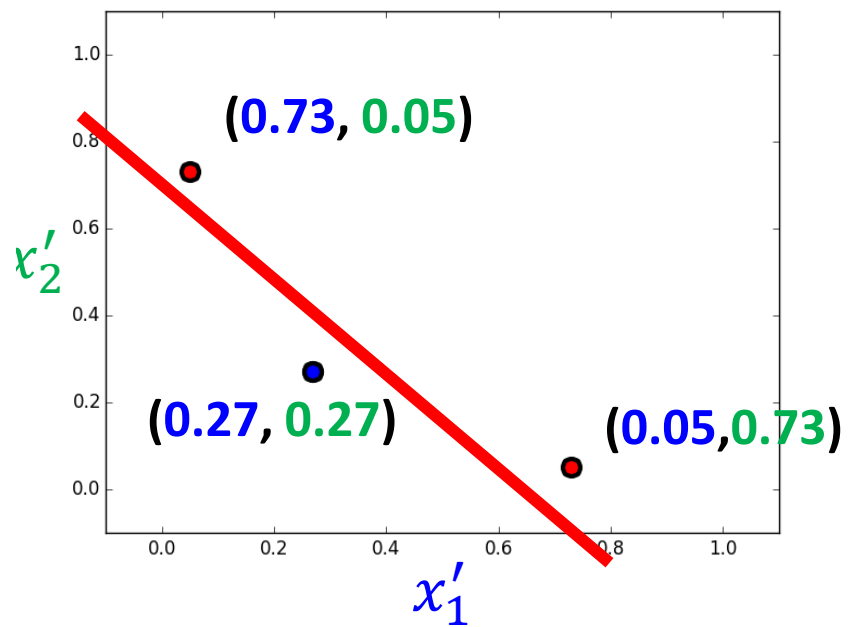
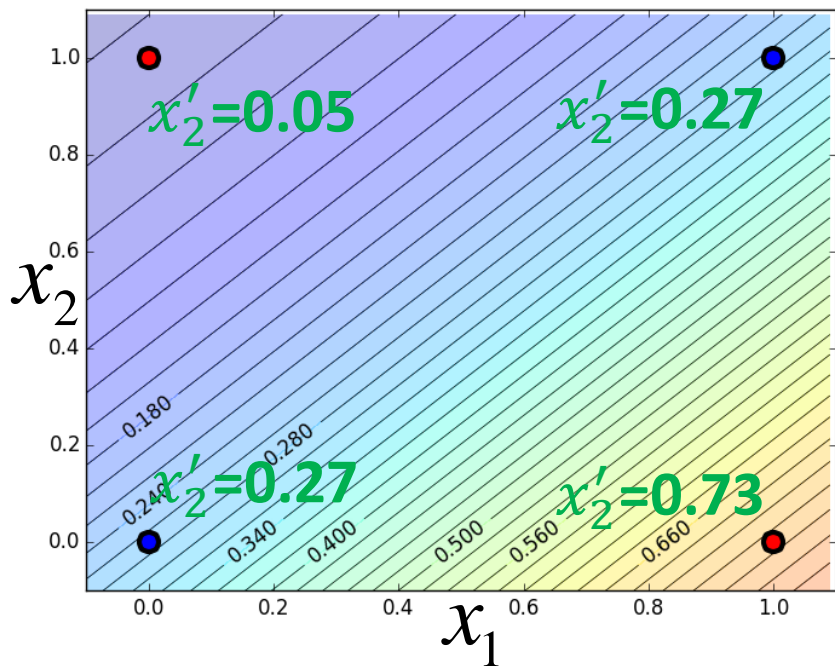
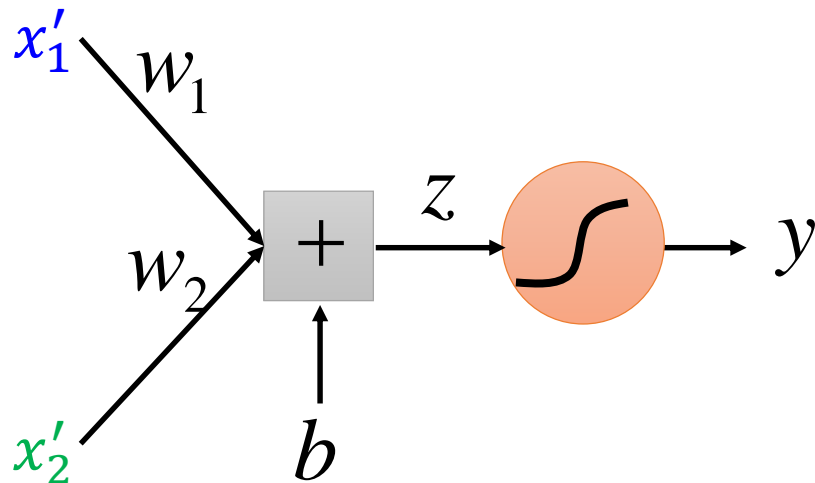
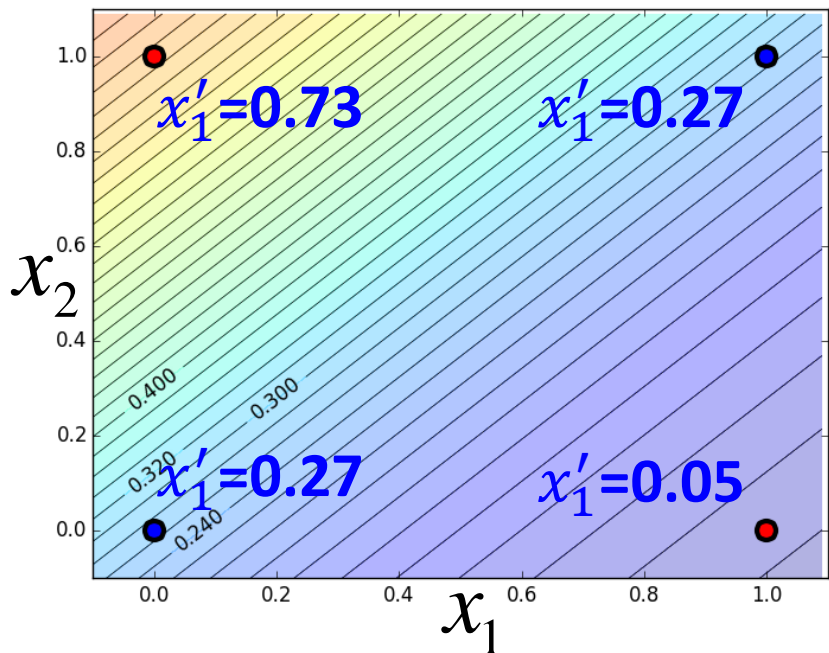
Limitation of Logistic Regression

- Cascading logistic regression models



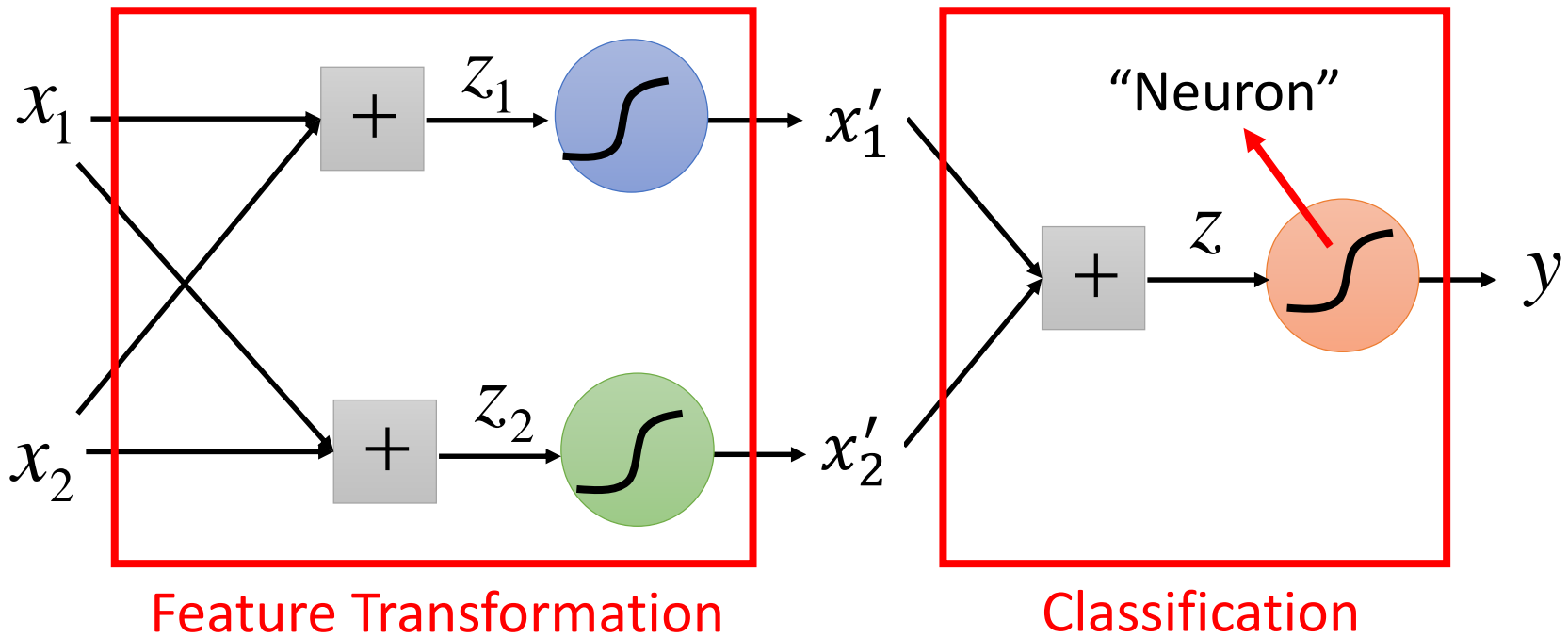
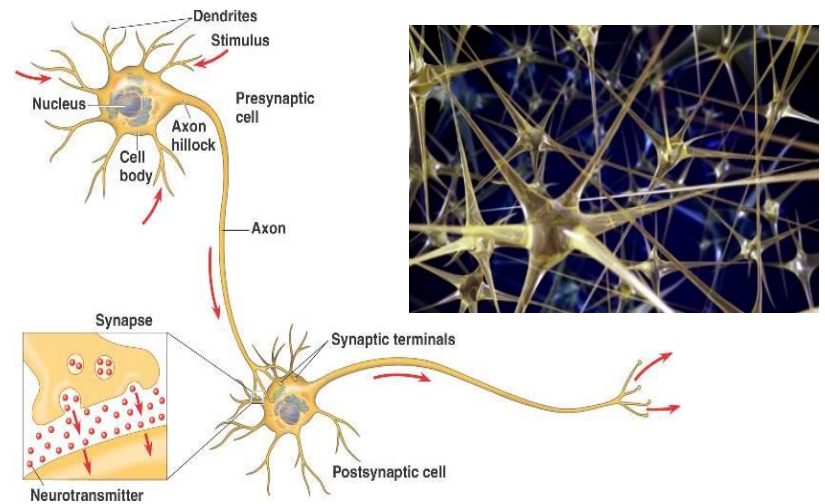
(ignore bias in this figure)





Deep Learning!

All the parameters of the logistic regressions are jointly learned.



Neural Network

Reference

- Bishop: Chapter 4.3

Appendix

Step 2: Loss function

$$f_{w,b}(x) = \begin{cases} z \geq 0 & +1 \\ z < 0 & -1 \end{cases}$$

Ideal loss:

$$L(f) = \sum_n \delta(f(x^n) \neq \hat{y}^n)$$

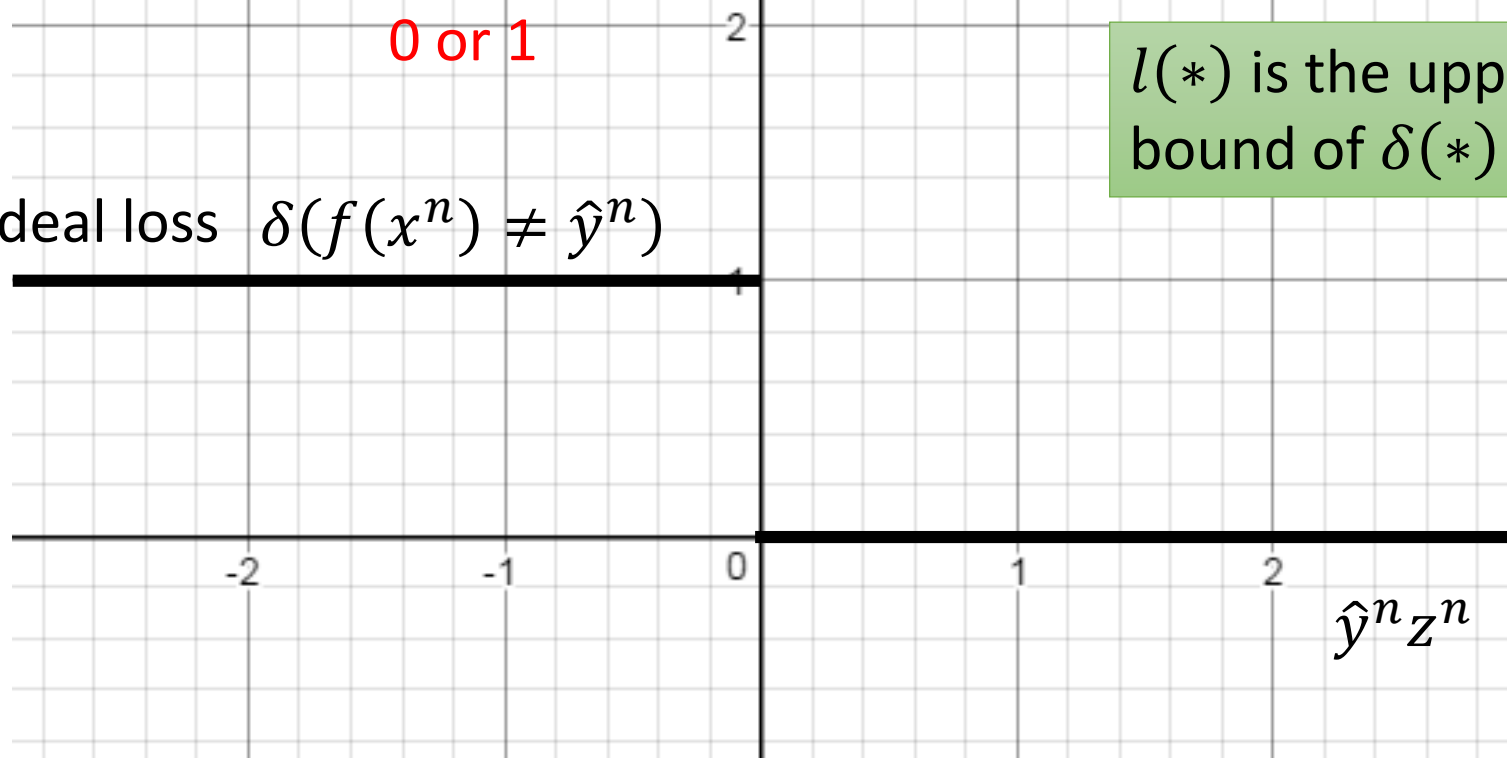
0 or 1

Approximation:

$$L(f) = \sum_n l(f(x^n), \hat{y}^n)$$

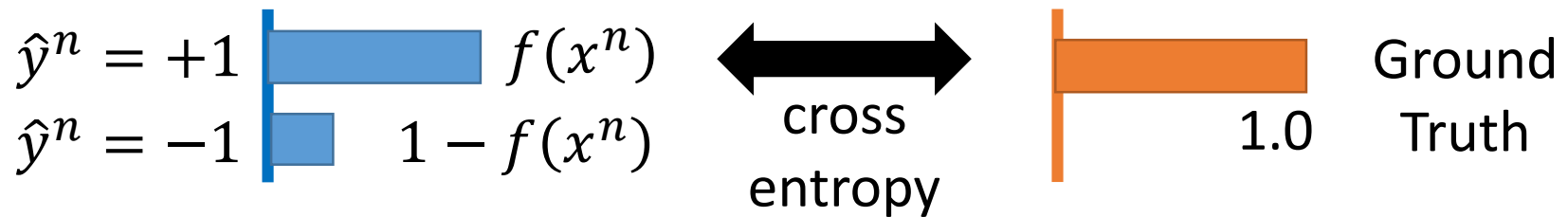
$l(*)$ is the upper bound of $\delta(*)$

Ideal loss $\delta(f(x^n) \neq \hat{y}^n)$



Step 2: Loss function

$l(f(x^n), \hat{y}^n)$: cross entropy



If $\hat{y}^n = +1$:

$$\begin{aligned} l(f(x^n), \hat{y}^n) &= -\ln f(x^n) = -\ln \sigma(z^n) = -\ln \frac{1}{1 + \exp(-z^n)} \\ &= \ln(1 + \exp(-z^n)) = \underline{\ln(1 + \exp(-\hat{y}^n z^n))} \end{aligned}$$

If $\hat{y}^n = -1$:

$$\begin{aligned} l(f(x^n), \hat{y}^n) &= -\ln(1 - f(x^n)) \\ &= -\ln(1 - \sigma(z^n)) = -\ln \frac{\exp(-z^n)}{1 + \exp(-z^n)} = -\ln \frac{1}{1 + \exp(z^n)} \\ &= \ln(1 + \exp(z^n)) = \underline{\ln(1 + \exp(-\hat{y}^n z^n))} \end{aligned}$$

Step 2: Loss function

$l(f(x^n), \hat{y}^n)$: cross entropy

$$l(f(x^n), \hat{y}^n) = \ln(1 + \exp(-\hat{y}^n z^n))$$

