

Softmax Regression

(Draft)

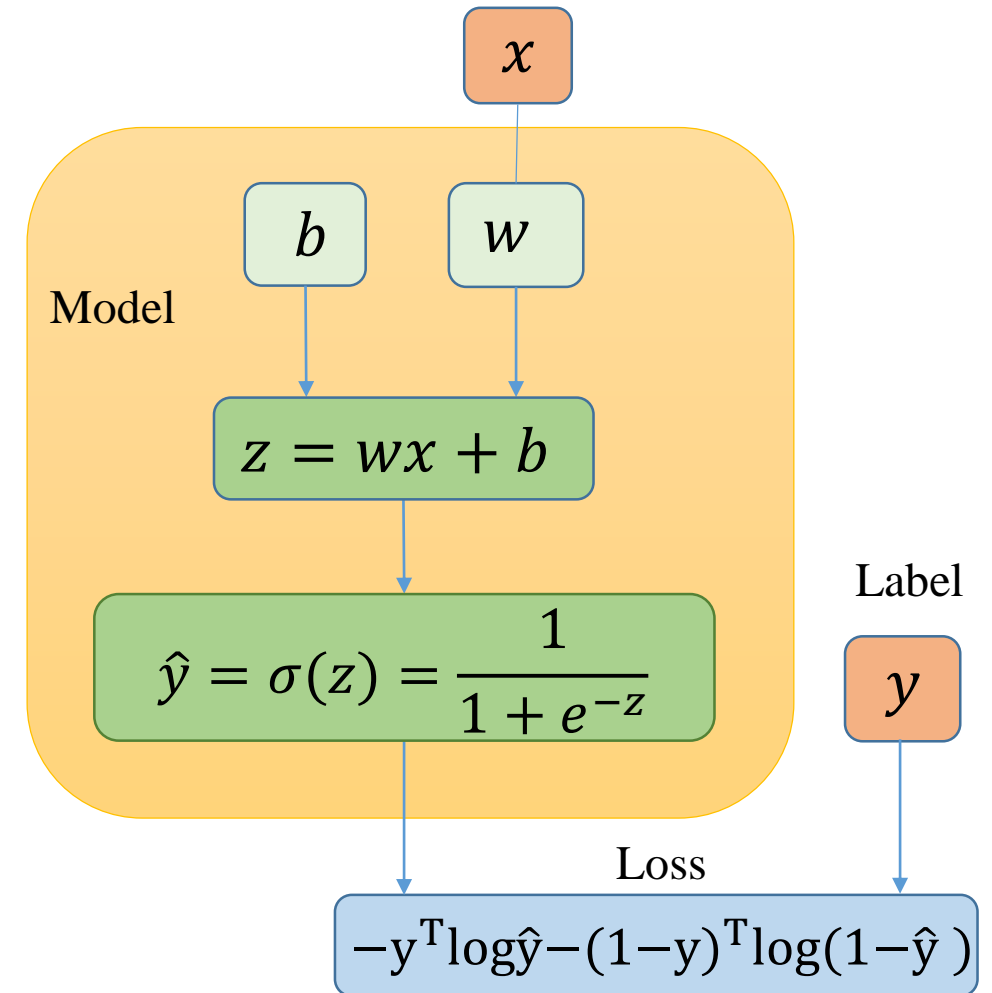
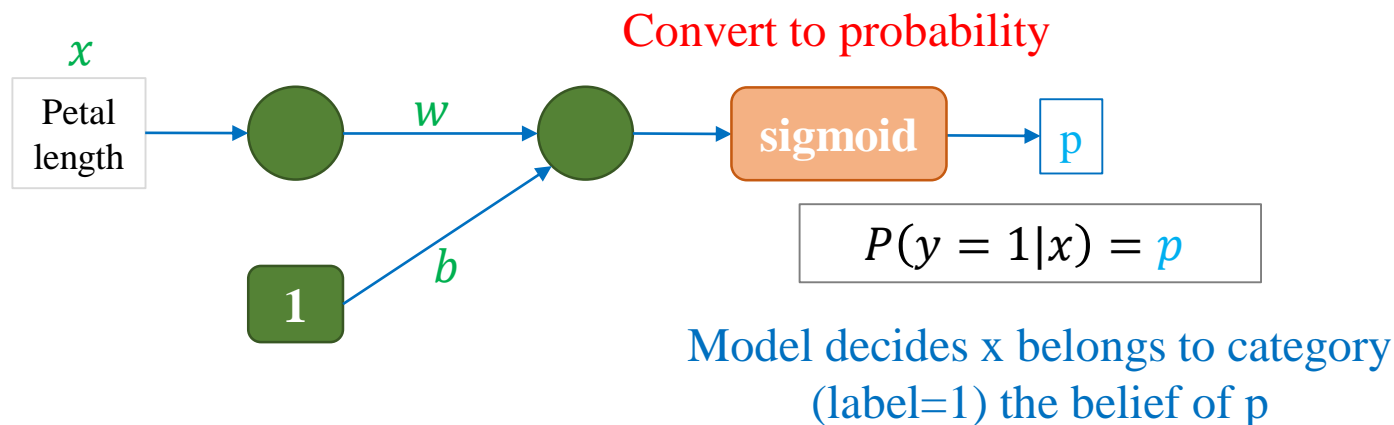
Quang-Vinh Dinh
Ph.D. in Computer Science

Outline

- **Motivation**
- **Model Construction**
- **Loss Function**
- **Simple Example and Generalization**
- **Examples - Stochastic and Batch**
- **Another Approach**

Motivation

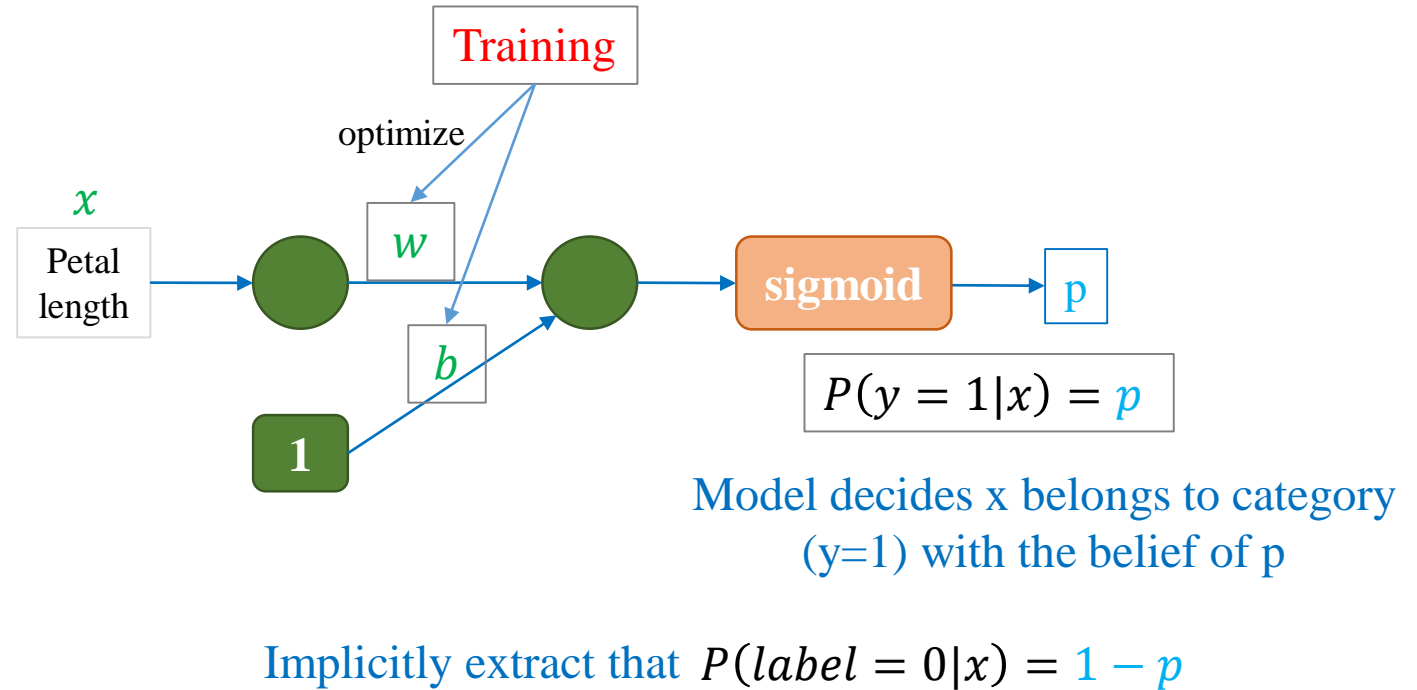
Feature	Label
Petal_Length	Label
1.4	0
1.3	0
1.5	0
4.5	1
4.1	1
4.6	1



Motivation

Problem!

Feature	Label
Petal_Length	Label
1.4	0
1.3	0
1.5	0
4.5	1
4.1	1
4.6	1



Optimize w and b for $P(\text{label} = 1|x)$ affects $P(\text{label} = 0|x)$ and vice versa

How to have explicitly $P(y = 0|x)$?

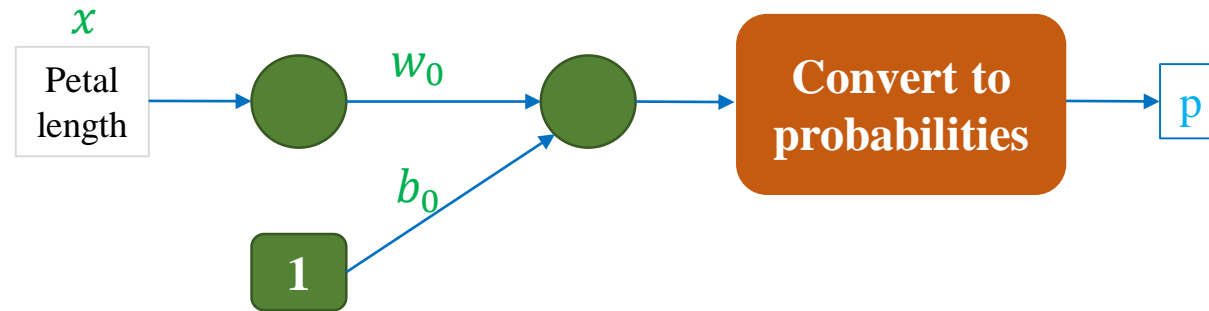
Motivation

Problem!

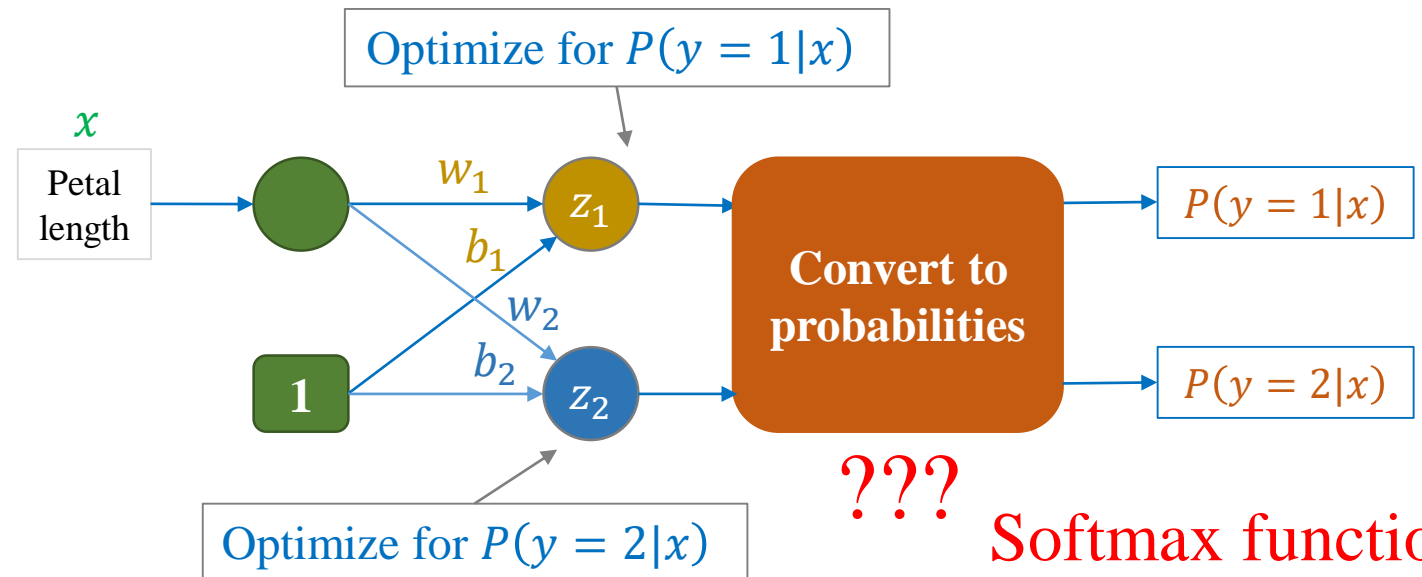
Feature	Label
Petal_Length	Label
1.4	1
1.3	1
1.5	1
4.5	2
4.1	2
4.6	2

* Indices is from 1

Change notation a little bit



Explicitly output $P(y = 1|x)$ and $P(y = 2|x)$



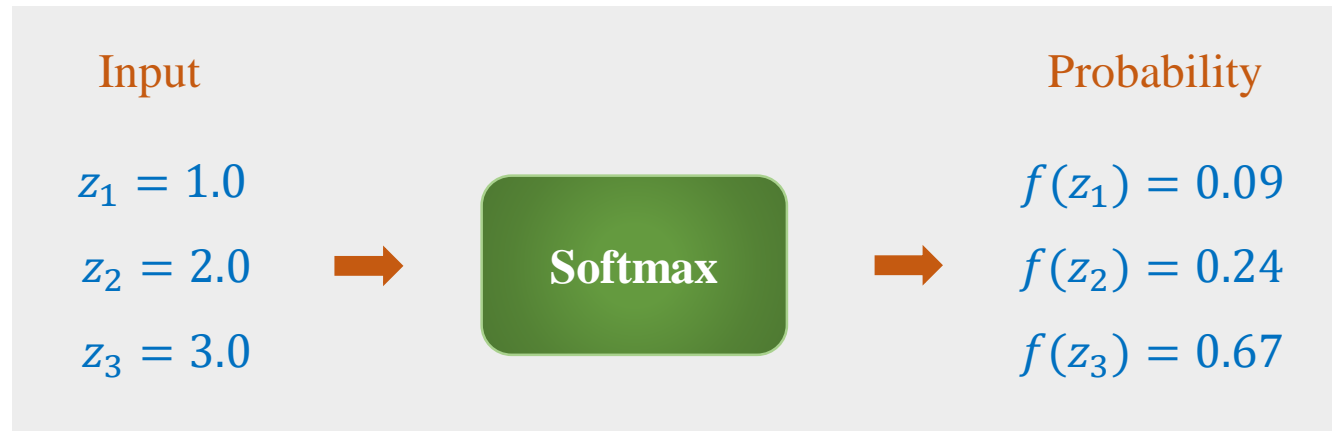
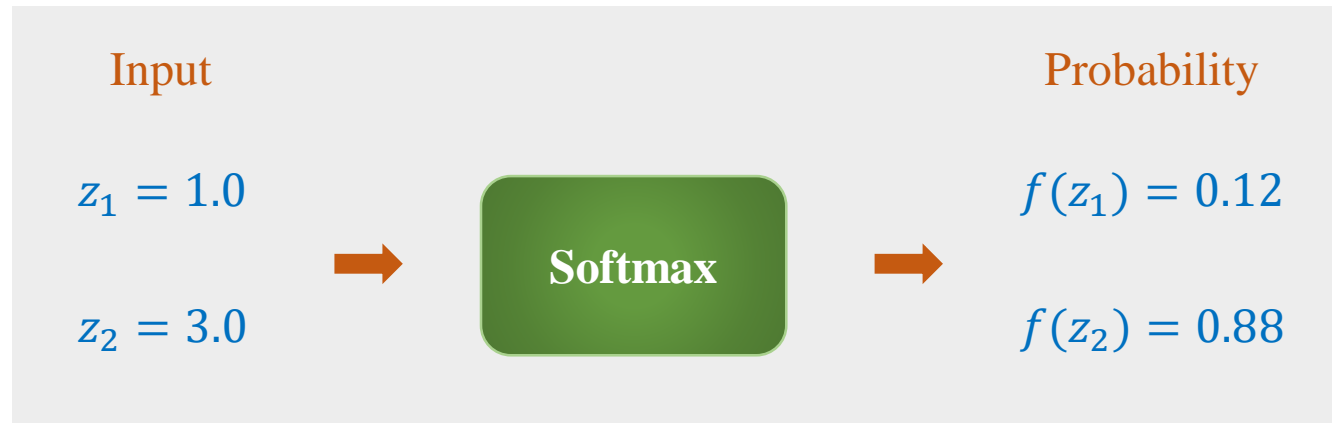
Motivation

Softmax function

$$P_i = f(z_i) = \frac{e^{z_i}}{\sum_j e^{z_j}}$$

$$0 \leq f(z_i) \leq 1$$

$$\sum_i f(z_i) = 1$$



Softmax function

Chuyển các giá trị của một vector thành các giá trị xác suất

Formula

$$f(x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}}$$
$$0 \leq f(x_i) \leq 1$$
$$\sum_i f(x_i) = 1$$

Input

$$x_1 = 1.0$$

$$x_2 = 2.0$$

$$x_3 = 3.0$$

Softmax

Probability

$$f(x_1) = 0.09$$

$$f(x_2) = 0.24$$

$$f(x_3) = 0.67$$

Implementation
(straightforward)

```
import numpy as np

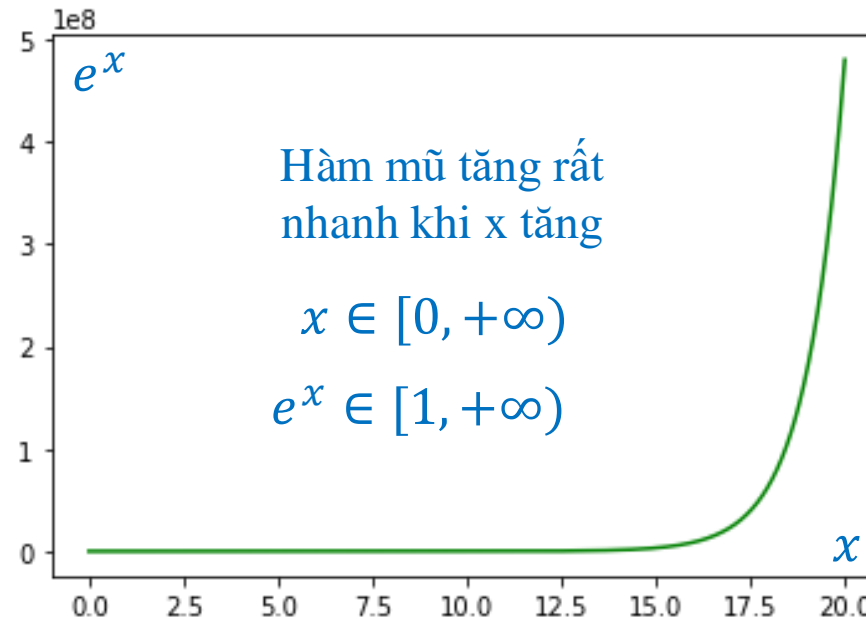
def softmax(X):
    exps = np.exp(X)
    return exps / np.sum(exps)
```

```
1 X = np.array([1.0, 2.0, 3.0])
2 f = softmax(X)
3 print(f)
```

```
[0.09003057 0.24472847 0.66524096]
```

```
1 X = np.array([1000.0, 1001.0, 1002.0])
2 f = softmax(X)
3 print(f)
```

```
[nan nan nan]
```



Giá trị nan vì e^x vượt giới hạn lưu trữ của biến

Softmax function (stable)

(Stable) Formula

$$m = \max(\mathbf{x})$$
$$f(x_i) = \frac{e^{(x_i - m)}}{\sum_j e^{(x_j - m)}}$$

X	X-m		Probability
$x_1 = 1.0$	$x_1 = -2.0$	Softmax	$f(x_1) = 0.09$
$x_2 = 2.0$	$x_2 = -1.0$		$f(x_2) = 0.24$
$x_3 = 3.0$	$x_3 = 0$		$f(x_3) = 0.67$

Implementation (stable)

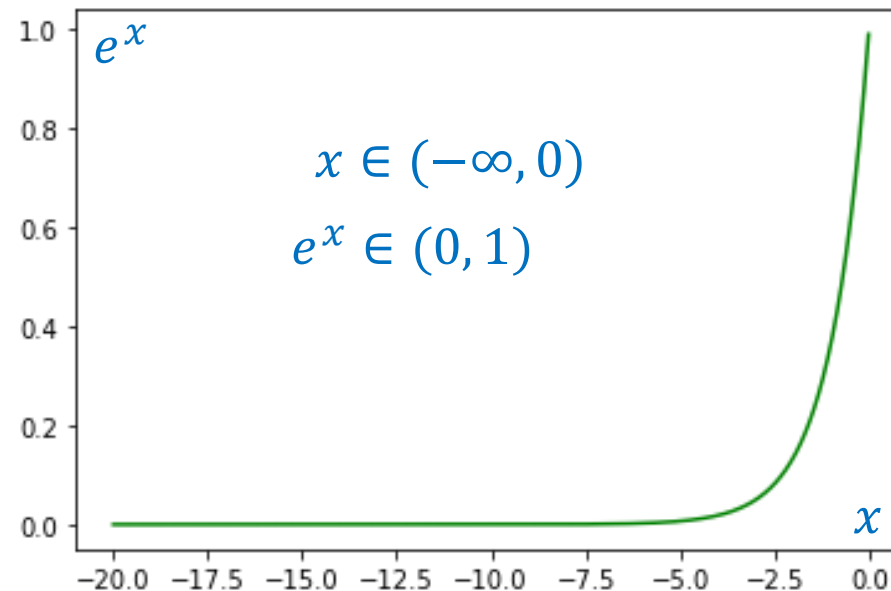
```
1 import numpy as np
2
3 def stable_softmax(X):
4     exps = np.exp(X - np.max(X))
5     return exps / np.sum(exps)
```

```
1 X = np.array([1.0, 2.0, 3.0])
2 f = stable_softmax(X)
3 print(f)
```

```
[0.09003057 0.24472847 0.66524096]
```

```
1 X = np.array([1000.0, 1001.0, 1002.0])
2 f = stable_softmax(X)
3 print(f)
```

```
[0.09003057 0.24472847 0.66524096]
```



```
1 X = np.array([1.0, 1001.0, 1002.0])
2 f = stable_softmax(X)
3 print(f)
```

```
[0.09003057 0.26894142 0.73105858]
```


Motivation

Feature Label

Petal_Length	Label
1.4	1
1.3	1
1.5	1
4.5	2
4.1	2
4.6	2

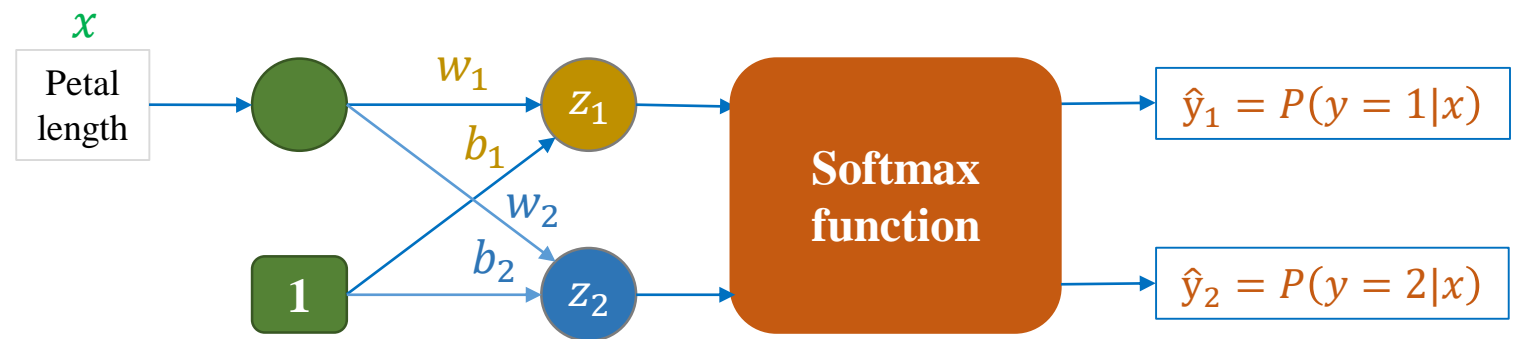
Softmax function

$$P_i = f(z_i) = \frac{e^{z_i}}{\sum_j e^{z_j}}$$

$$0 \leq f(z_i) \leq 1$$

$$\sum_i f(z_i) = 1$$

Explicitly output $P(y = 1|x)$ and $P(y = 0|x)$



How about loss function?

One node

$$L(\theta) = -y \log \hat{y} - (1-y) \log (1-\hat{y})$$

Two nodes

$$L(\theta) = -\delta(y, 1) \log \hat{y}_1 - \delta(y, 2) \log \hat{y}_0$$

$$\delta(i, j) = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

Outline

- **Motivation**
- **Model Construction**
- **Loss Function**
- **Simple Example and Generalization**
- **Examples - Stochastic and Batch**
- **Another Approach**

Model Construction

❖ 1-D Feature and two classes

Feature	Label
Petal_Length	Label
1.4	1
1.3	1
1.5	1
4.5	2
4.1	2
4.6	2

#class=2

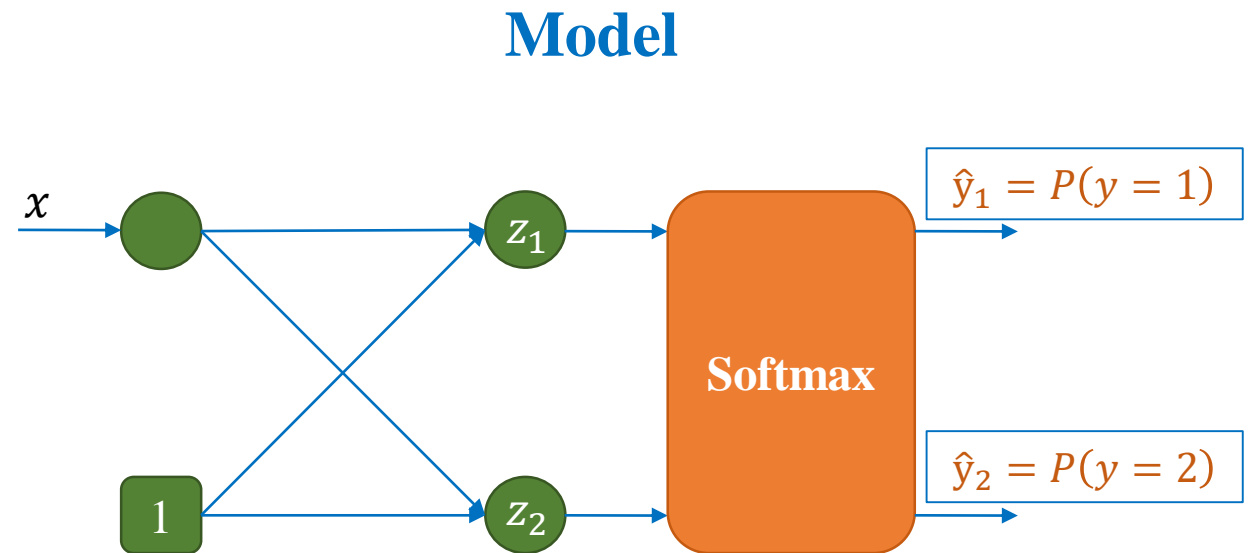
#feature=1

Feature is with one dimension

→ Need one node for input

Three categories

→ Need three node for output



Model Construction

❖ 1-D Feature and three classes

Feature	Label
Petal_Length	Label
1.4	1
1.3	1
1.5	1
4.5	2
4.1	2
4.6	2
5.2	3
5.6	3
5.9	3

#class=3

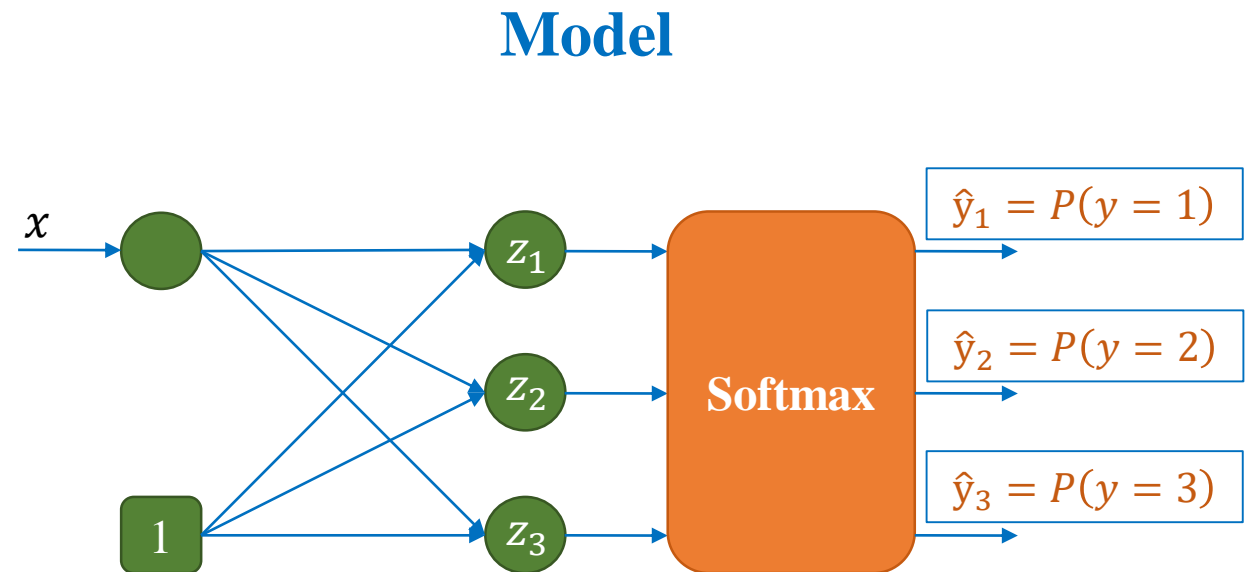
#feature=1

Feature is with one dimension

→ Need one node for input

Three categories

→ Need three nodes for output



Model Construction

❖ 4-D Feature and three classes

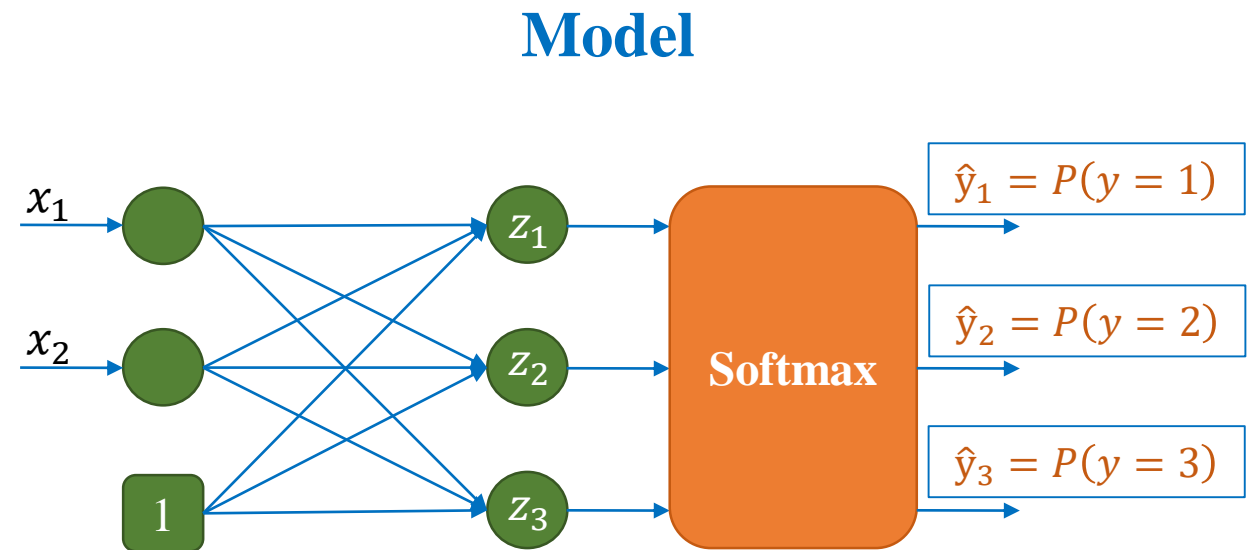
Feature		Label	
Petal_Length	Petal_Width	Label	
1.5	0.2	1	#class=3
1.4	0.2	1	
1.6	0.2	1	
4.7	1.6	2	#feature=2
3.3	1.1	2	
4.6	1.3	2	
5.6	2.2	3	
5.1	1.5	3	
5.6	1.4	3	

Feature is with two dimensions

→ Need two nodes for input

Three categories

→ Need three nodes for output



Model Construction

❖ 4-D Feature and three classes

Feature **Label**

Sepal_Length	Sepal_Width	Petal_Length	Petal_Width	Label
5.2	3.5	1.5	0.2	1
5.2	3.4	1.4	0.2	1
4.7	3.2	1.6	0.2	1
6.3	3.3	4.7	1.6	2
4.9	2.4	3.3	1.1	2
6.6	2.9	4.6	1.3	2
6.4	2.8	5.6	2.2	3
6.3	2.8	5.1	1.5	3
6.1	2.6	5.6	1.4	3

Feature is with four dimensions

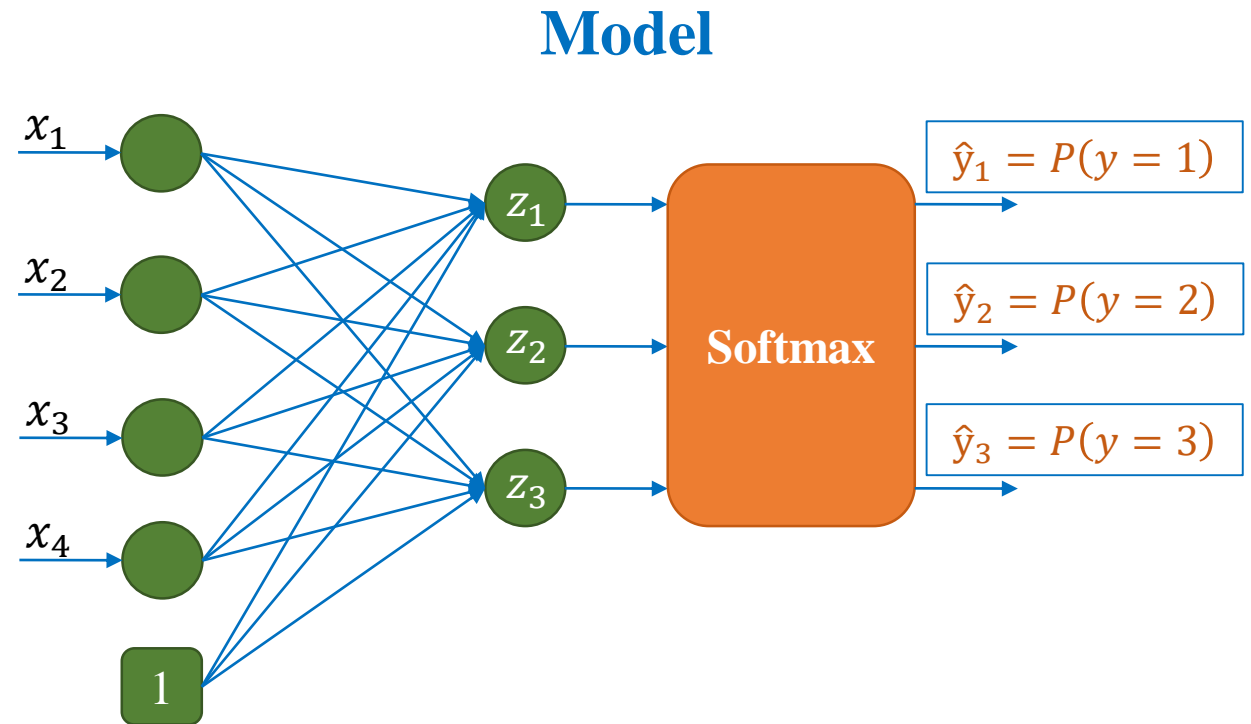
→ Need four nodes for input

Three categories

→ Need three nodes for output

#class=3

#feature=4



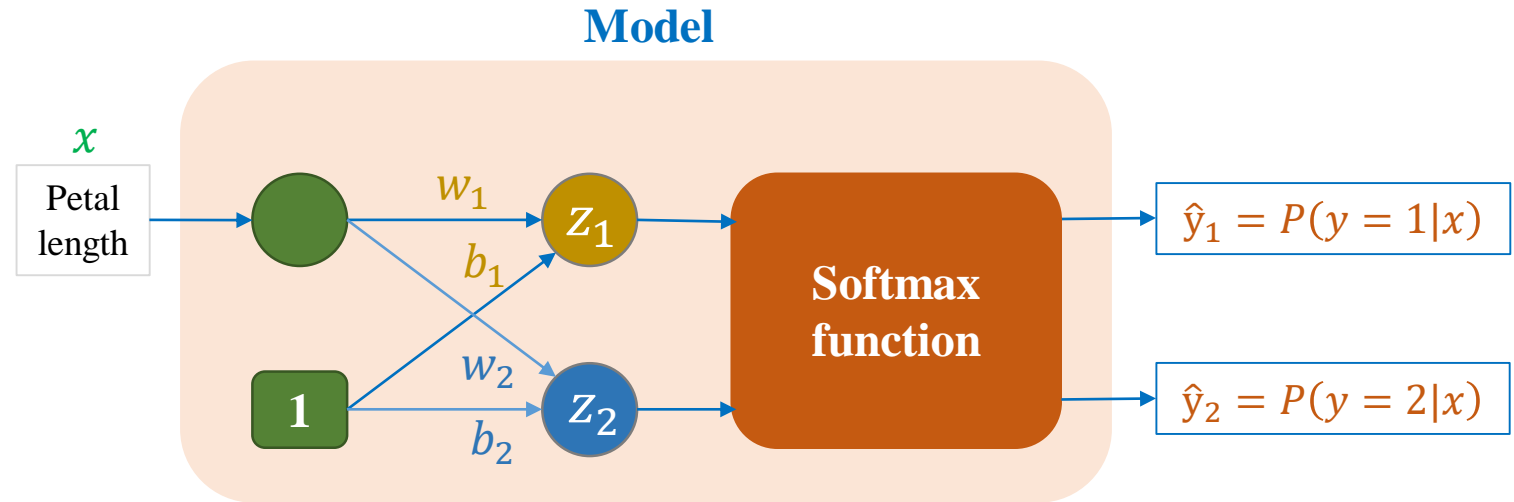
Outline

- **Motivation**
- **Model Construction**
- **Loss Function**
- **Simple Example and Generalization**
- **Examples - Stochastic and Batch**
- **Another Approach**

Loss function

❖ Simple illustration

Feature	Label
Petal_Length	Label
1.4	1
1.3	1
1.5	1
4.5	2
4.1	2
4.6	2



$$z_1 = xw_1 + b_1$$

$$z_2 = xw_2 + b_2$$

$$\hat{y}_1 = \frac{e^{z_1}}{\sum_{j=1}^2 e^{z_j}}$$

$$\hat{y}_2 = \frac{e^{z_2}}{\sum_{j=1}^2 e^{z_j}}$$

$$\mathbf{z} = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} b_1 & w_1 \\ b_2 & w_2 \end{bmatrix} \begin{bmatrix} 1 \\ x \end{bmatrix} = \begin{bmatrix} \theta_1^T \\ \theta_2^T \end{bmatrix} \begin{bmatrix} 1 \\ x \end{bmatrix} = \boldsymbol{\theta}^T \mathbf{x}$$

$$\hat{\mathbf{y}} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \end{bmatrix} = \frac{1}{\sum_{j=1}^2 e^{z_j}} \begin{bmatrix} e^{z_1} \\ e^{z_2} \end{bmatrix} = \frac{e^{\mathbf{z}}}{\sum_{j=1}^2 e^{z_j}}$$

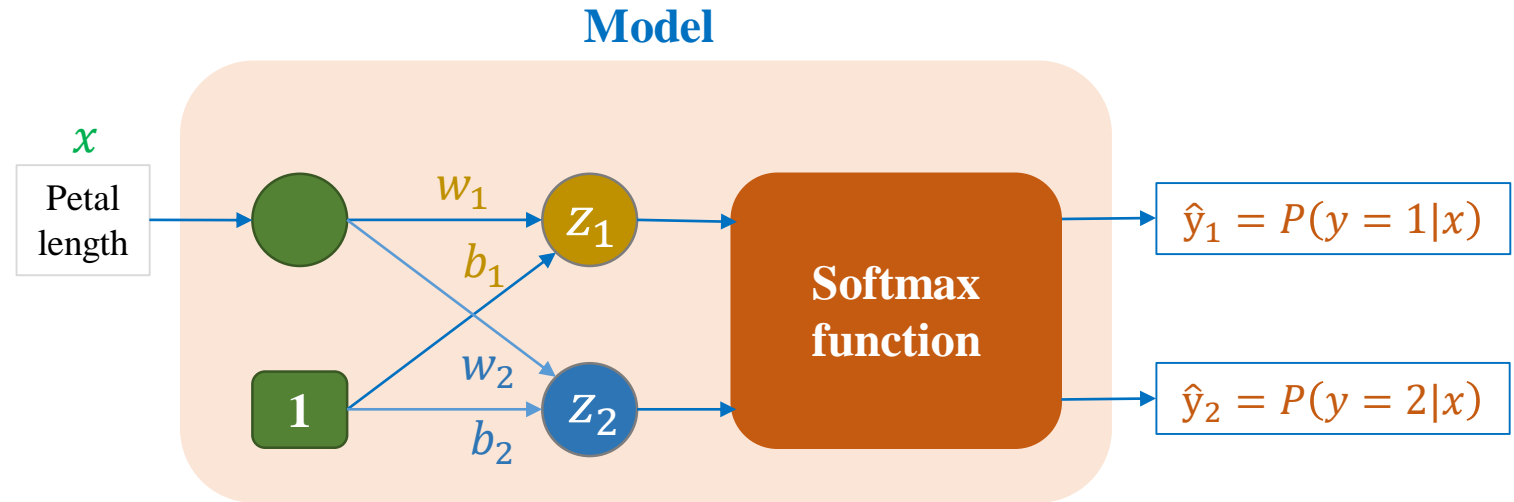
A vector is by default a column vector $\boldsymbol{\theta}_1 = \begin{bmatrix} b_1 \\ w_1 \end{bmatrix}$

vector transpose $\boldsymbol{\theta}_1^T = [b_1 \ w_1]$

Loss function

❖ Simple illustration

Feature	Label
Petal_Length	Label
1.4	1
1.3	1
1.5	1
4.5	2
4.1	2
4.6	2



$$\delta(i, j) = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

$$z_1 = xw_1 + b_1$$

$$z_2 = xw_2 + b_2$$

$$\hat{y}_1 = \frac{e^{z_1}}{\sum_{j=1}^2 e^{z_j}}$$

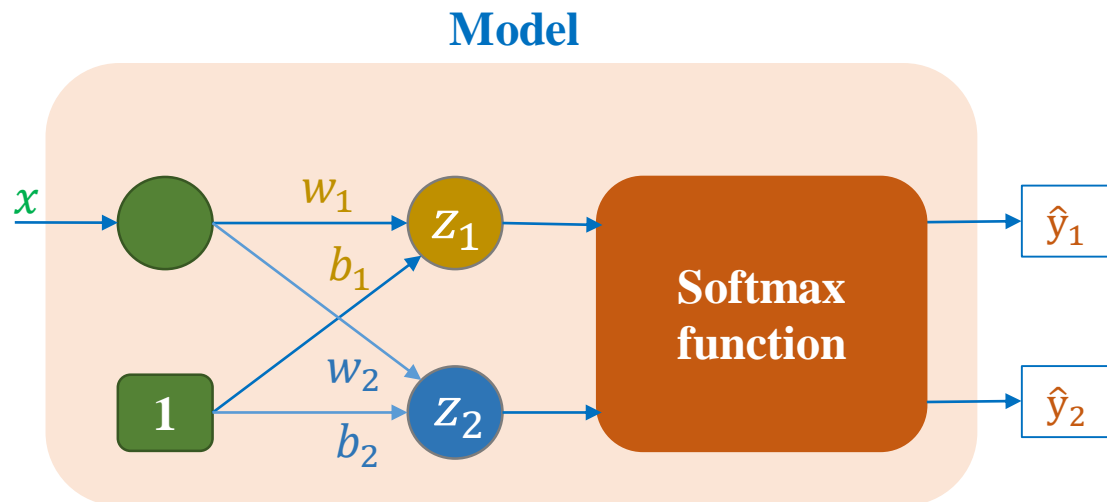
$$\hat{y}_2 = \frac{e^{z_2}}{\sum_{j=1}^2 e^{z_j}}$$

$$\mathbf{z} = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} b_1 & w_1 \\ b_2 & w_2 \end{bmatrix} \begin{bmatrix} 1 \\ x \end{bmatrix} = \begin{bmatrix} \theta_1^T \\ \theta_2^T \end{bmatrix} \begin{bmatrix} 1 \\ x \end{bmatrix} = \boldsymbol{\theta}^T \mathbf{x}$$

$$\hat{\mathbf{y}} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \end{bmatrix} = \frac{1}{\sum_{j=1}^2 e^{z_j}} \begin{bmatrix} e^{z_1} \\ e^{z_2} \end{bmatrix} = \frac{e^{\mathbf{z}}}{\sum_{j=1}^2 e^{z_j}}$$

$$\begin{aligned} L(\boldsymbol{\theta}) &= -\delta(i, j) \log \hat{y}_1 - \delta(i, j) \log \hat{y}_2 \\ &= -\sum_{i=1}^2 \delta(i, y) \log \hat{y}_i \end{aligned}$$

Loss function



$$L(\theta) = - \sum_{i=1}^2 \delta(i, y) \log \hat{y}_i$$

$$\hat{y}_1 = \frac{e^{z_1}}{\sum_{j=1}^2 e^{z_j}}$$

$$\hat{y}_2 = \frac{e^{z_2}}{\sum_{j=1}^2 e^{z_j}}$$

Derivative

$$\frac{\partial \hat{y}_i}{\partial z_j} = \hat{y}_i (\delta(i, j) - \hat{y}_i)$$

$$\frac{\partial L}{\partial z_i} = \hat{y}_i - \delta(i, y)$$

Loss function

One-hot encoding for label

$$y = 0 \rightarrow \mathbf{y} = \begin{bmatrix} 1 & 0 \end{bmatrix} \quad \begin{matrix} y_0 & y_1 \end{matrix}$$

$$y = 1 \rightarrow \mathbf{y} = \begin{bmatrix} 0 & 1 \end{bmatrix}$$



$$z_1 = xw_1 + b_1$$

$$z_2 = xw_2 + b_2$$

$$\hat{y}_1 = \frac{e^{z_1}}{\sum_{j=1}^2 e^{z_j}}$$

$$\hat{y}_2 = \frac{e^{z_2}}{\sum_{j=1}^2 e^{z_j}}$$

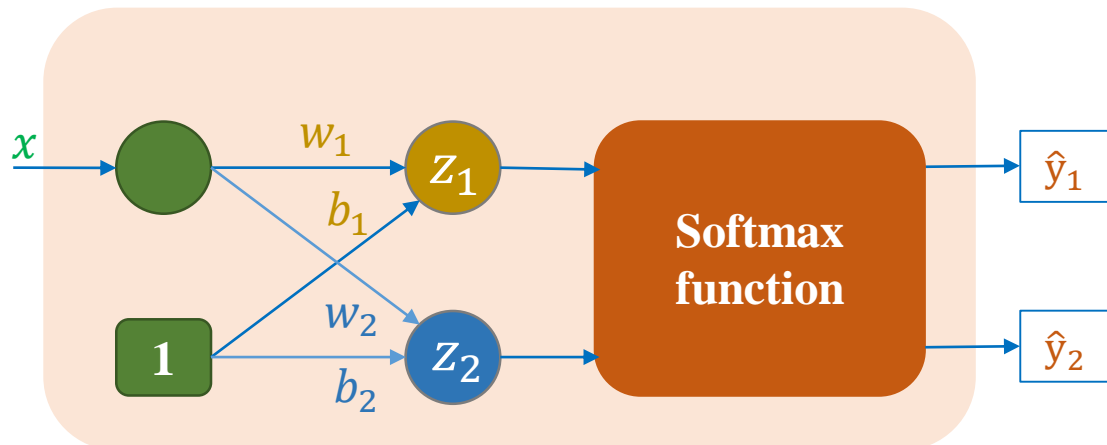
$$\mathbf{z} = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} b_1 & w_1 \\ b_2 & w_2 \end{bmatrix} \begin{bmatrix} 1 \\ x \end{bmatrix} = \begin{bmatrix} \theta_1^T \\ \theta_2^T \end{bmatrix} \begin{bmatrix} 1 \\ x \end{bmatrix} = \boldsymbol{\theta}^T \mathbf{x}$$

$$\hat{\mathbf{y}} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \end{bmatrix} = \frac{1}{\sum_{j=1}^2 e^{z_j}} \begin{bmatrix} e^{z_1} \\ e^{z_2} \end{bmatrix} = \frac{e^{\mathbf{z}}}{\sum_{j=1}^2 e^{z_j}}$$

$$L(\boldsymbol{\theta}) = -\delta(i, j) \log \hat{y}_1 - \delta(i, j) \log \hat{y}_2$$

$$= -\sum_{i=1}^2 \delta(i, y) \log \hat{y}_i$$

Model



Derivative

$$\frac{\partial \hat{y}_i}{\partial z_j} = \hat{y}_i (\delta(i, j) - \hat{y}_i)$$

$$\frac{\partial L}{\partial z_i} = \hat{y}_i - \delta(i, y)$$

$$\frac{\partial L}{\partial w_i} = x(\hat{y}_i - \delta(i, y))$$

$$\frac{\partial L}{\partial b_i} = \hat{y}_i - \delta(i, y)$$

Simple Illustration - Summary

Feature Label

Petal_Length	Label
1.4	1
1.3	1
1.5	1
4.5	2
4.1	2
4.6	2

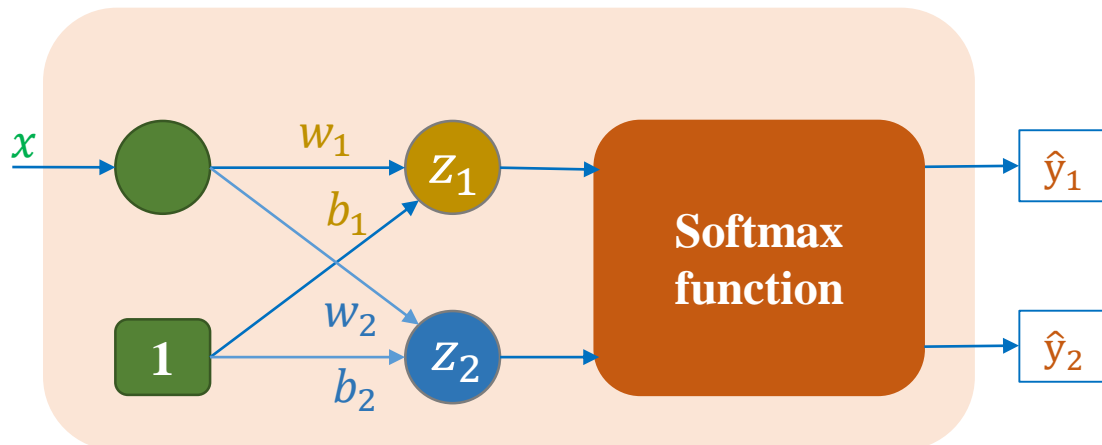
* Label indices are from 1

$$\theta = \begin{bmatrix} b_1 & b_2 \\ w_1 & w_2 \end{bmatrix}$$

$$x = \begin{bmatrix} 1 \\ x \end{bmatrix}$$

Input with one example
(x, y) = (1.4, 1)

Model



Forward computation

$$z = \theta^T x$$

$$\hat{y} = \frac{e^z}{\sum_{j=1}^2 e^{z_j}}$$

Loss function

$$L(\theta) = - \sum_{i=1}^2 \delta(i, y) \log \hat{y}_i$$

$$\delta(i, j) = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

Derivative

$$\frac{\partial \hat{y}_i}{\partial z_j} = \hat{y}_i (\delta(i, j) - \hat{y}_i)$$

$$\frac{\partial L}{\partial z_i} = \hat{y}_i - \delta(i, y)$$

$$\frac{\partial L}{\partial w_i} = x (\hat{y}_i - \delta(i, y))$$

$$\frac{\partial L}{\partial b_i} = \hat{y}_i - \delta(i, y)$$

Outline

- **Motivation**
- **Model Construction**
- **Loss Function**
- **Simple Example and Generalization**
- **Examples - Stochastic and Batch**
- **Another Approach**

Simple Example

Training data

Feature Label

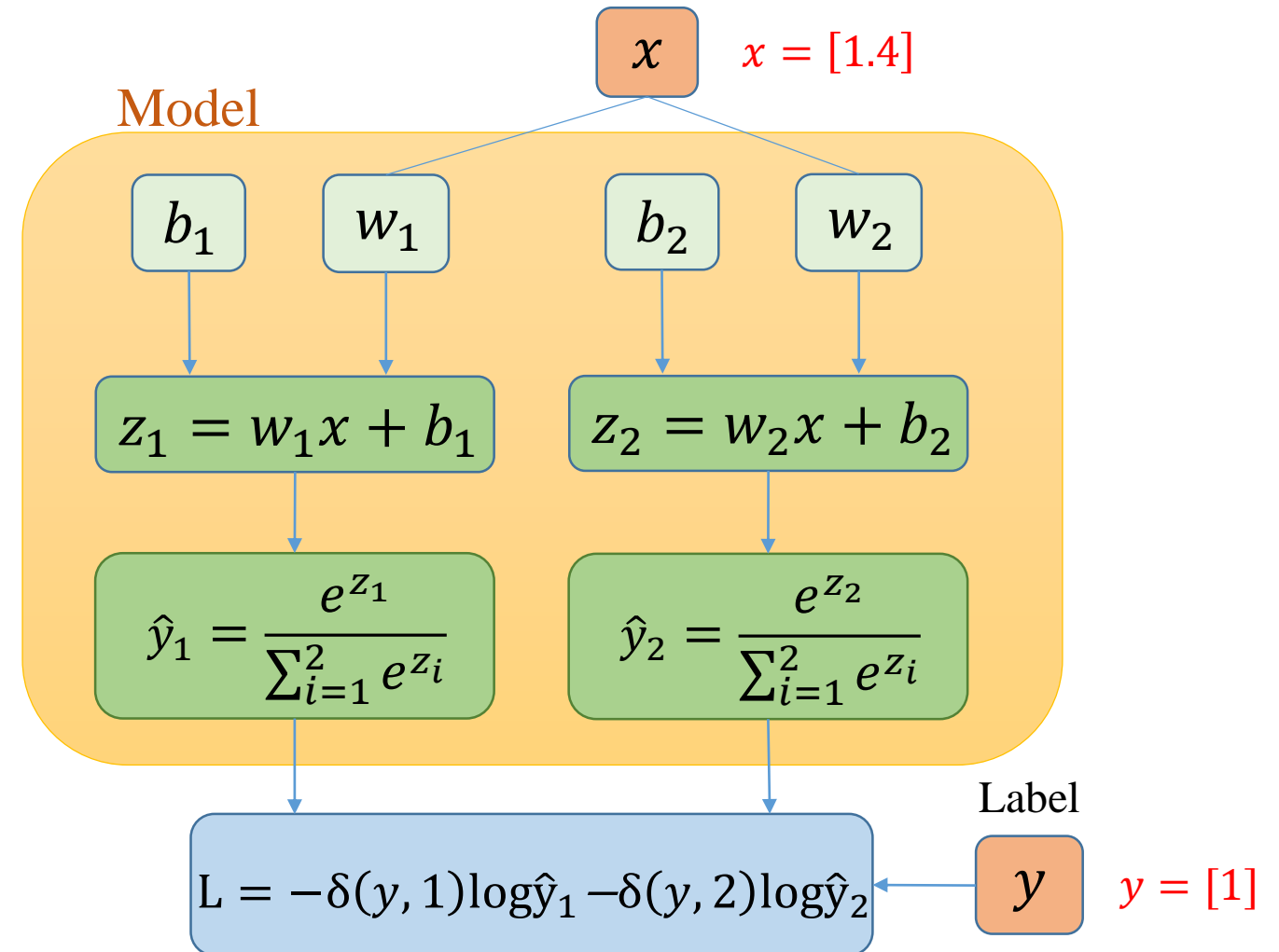
Petal Length	Label
1.4	1
1.3	1
1.5	1
4.5	2
4.1	2
4.6	2

#class=2

#feature=1

Training example

$(x, y) = (1.4, 1)$



Simple Example

Training data

Feature Label

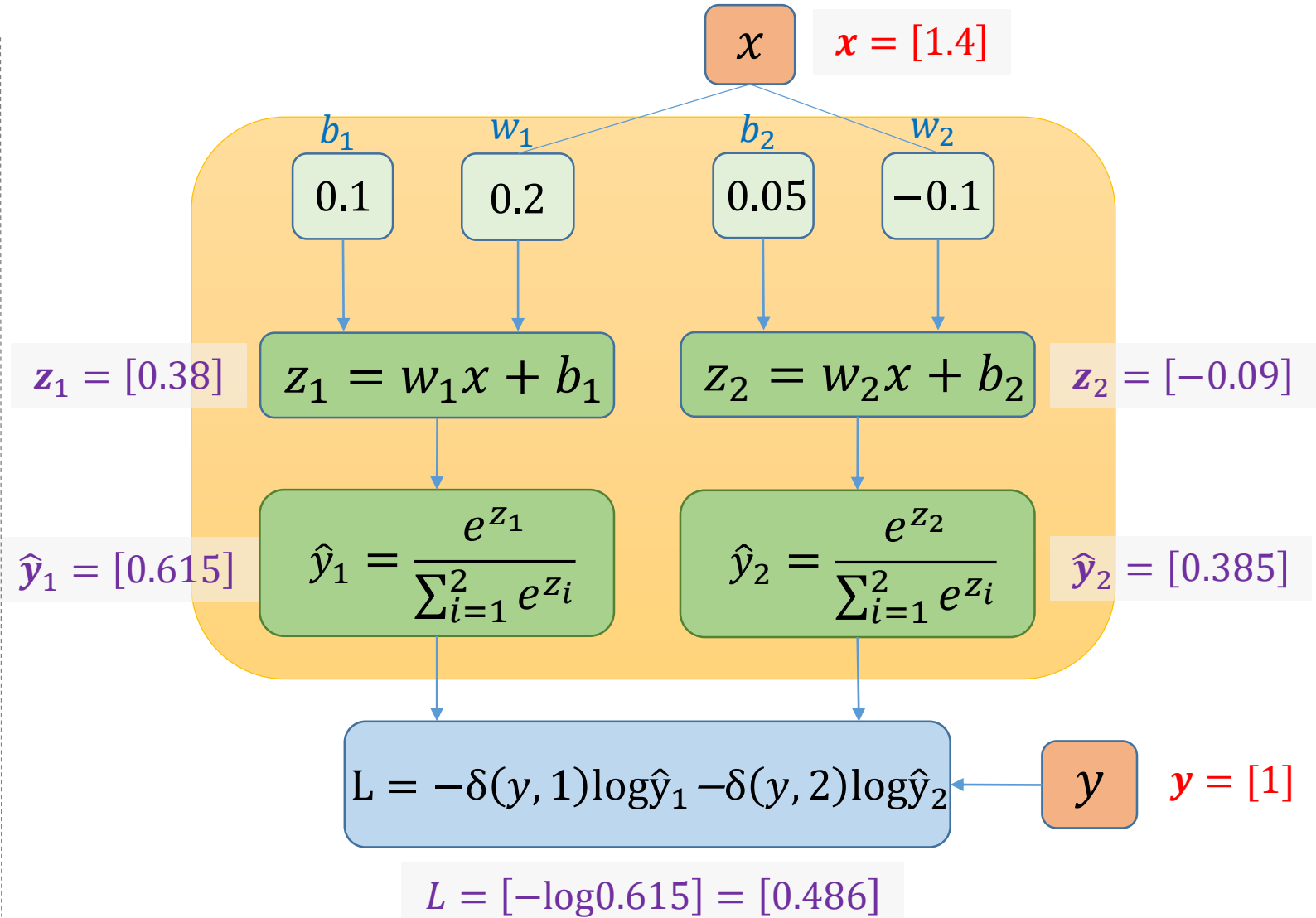
Petal Length	Label
1.4	1
1.3	1
1.5	1
4.5	2
4.1	2
4.6	2

#class=2

#feature=1

Training example

$(x, y) = (1.4, 1)$



Simple Example

Training example

$$(x, y) = (1.4, 1)$$

Derivative

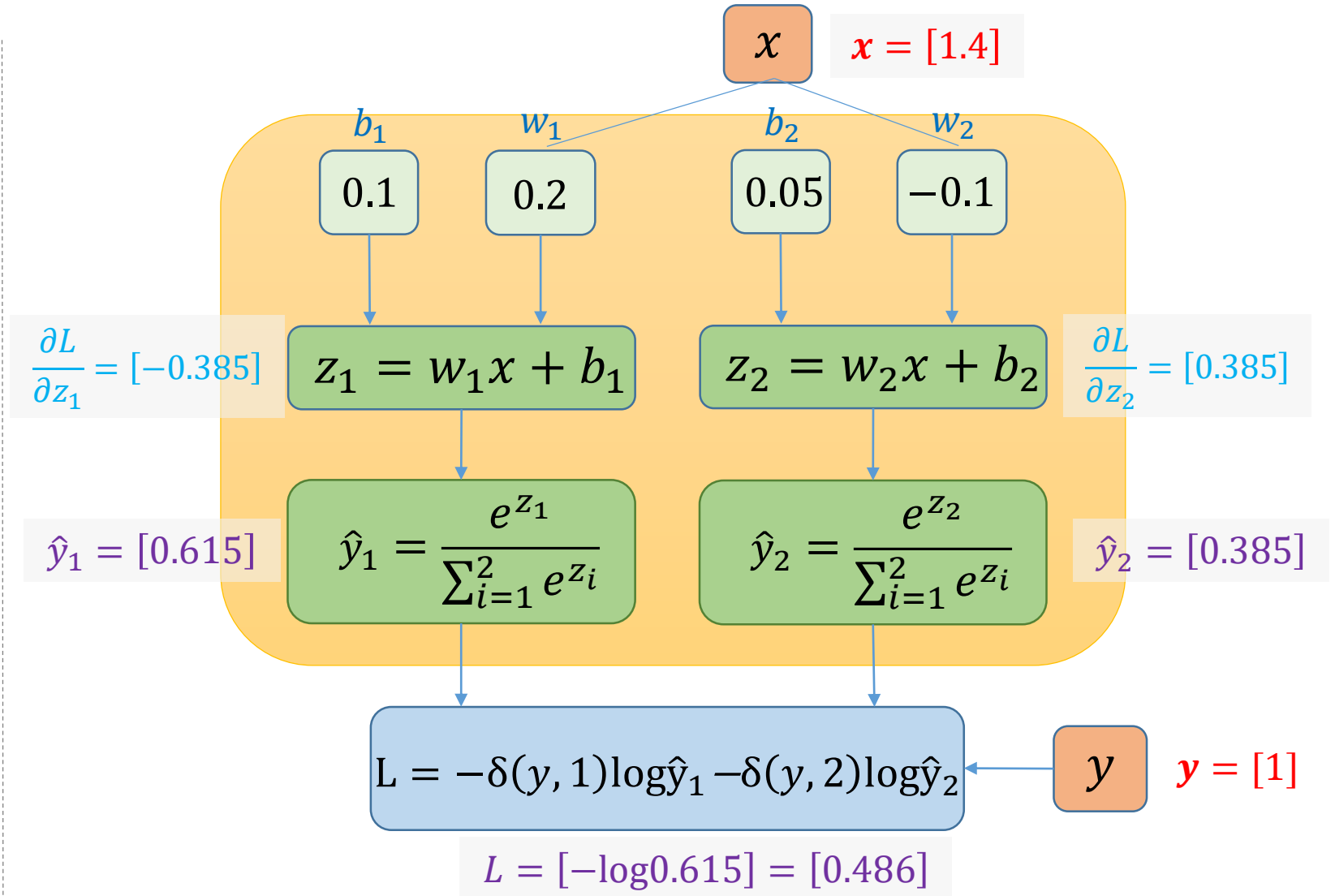
$$\frac{\partial L}{\partial z_i} = \hat{y}_i - \delta(i, y)$$

$$\frac{\partial L}{\partial w_i} = x(\hat{y}_i - \delta(i, y))$$

$$\frac{\partial L}{\partial b_i} = \hat{y}_i - \delta(i, y)$$

$$\begin{aligned} \frac{\partial L}{\partial z_1} &= \hat{y}_1 - \delta(1, y) \\ &= 0.615 - 1 = -0.385 \end{aligned}$$

$$\begin{aligned} \frac{\partial L}{\partial z_2} &= \hat{y}_2 - \delta(2, y) \\ &= 0.385 - 0 = 0.385 \end{aligned}$$



Simple Example

Training example

$$(x, y) = (1.4, 1)$$

Derivative

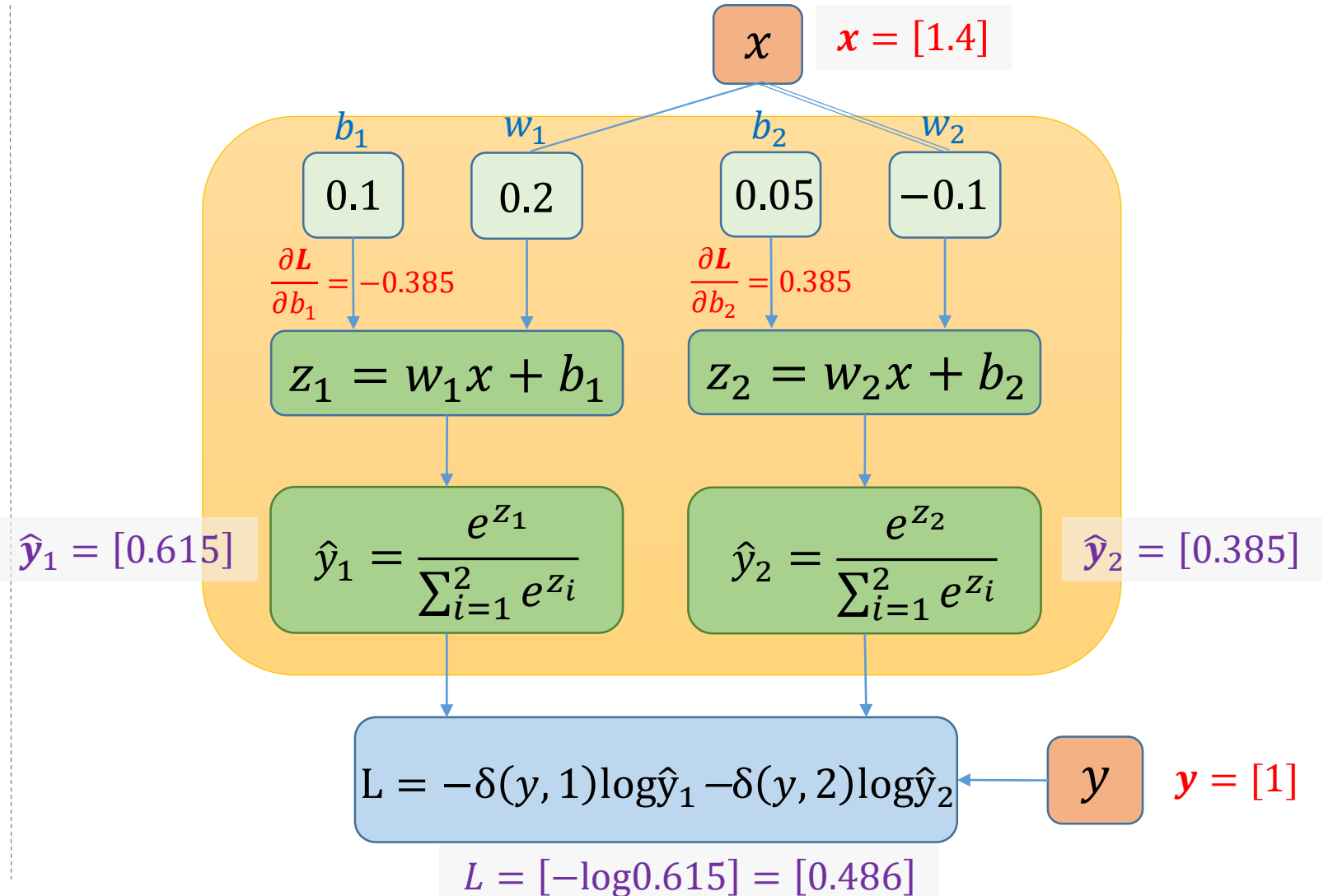
$$\frac{\partial L}{\partial z_i} = \hat{y}_i - \delta(i, y)$$

$$\frac{\partial L}{\partial w_i} = x(\hat{y}_i - \delta(i, y))$$

$$\frac{\partial L}{\partial b_i} = \hat{y}_i - \delta(i, y)$$

$$\begin{aligned} \frac{\partial L}{\partial b_1} &= \hat{y}_1 - \delta(1, y) \\ &= 0.615 - 1 = -0.385 \end{aligned}$$

$$\begin{aligned} \frac{\partial L}{\partial b_2} &= \hat{y}_2 - \delta(2, y) \\ &= 0.385 - 0 = 0.385 \end{aligned}$$



Simple Example

Training example

$$(x, y) = (1.4, 1)$$

Derivative

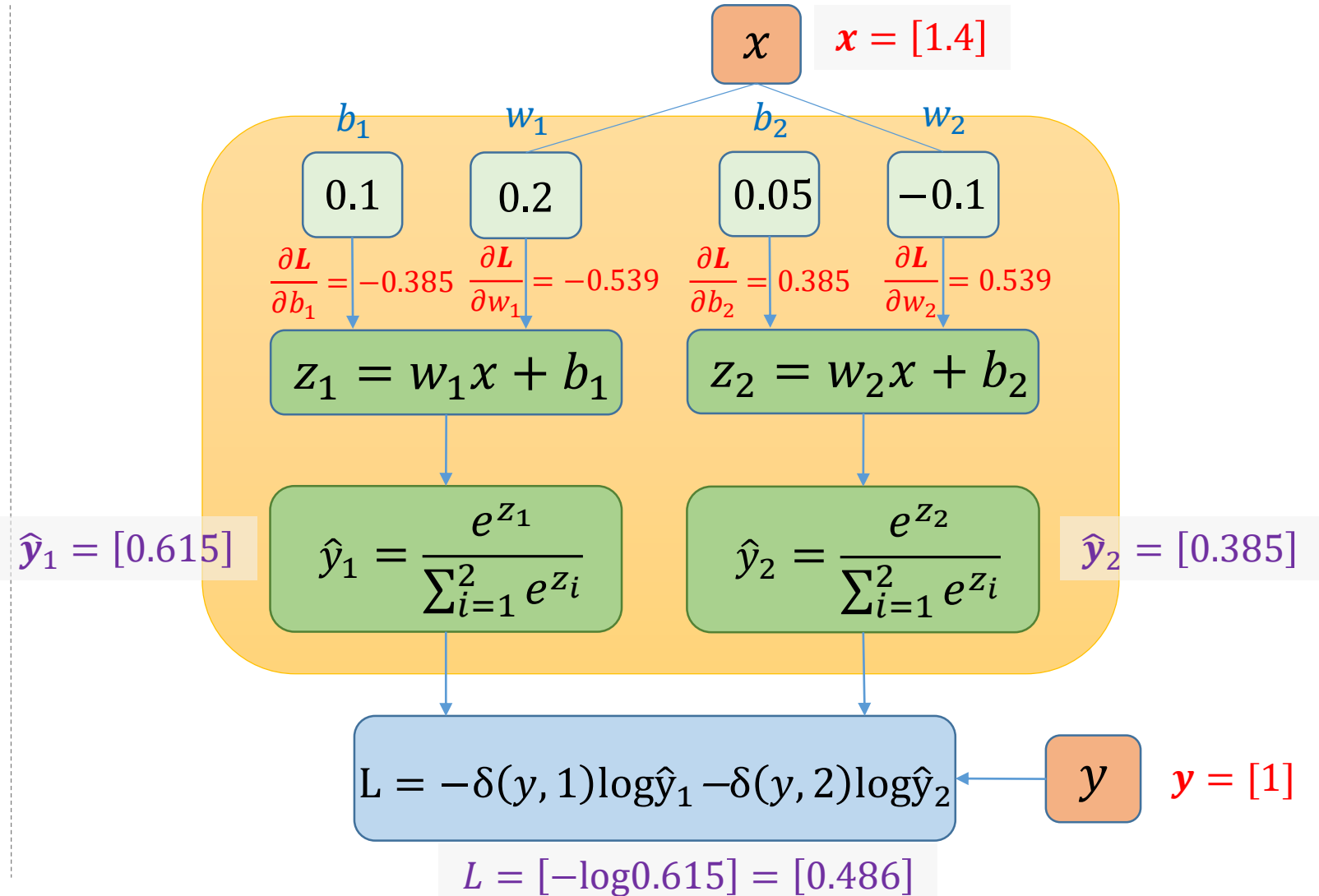
$$\frac{\partial L}{\partial z_i} = \hat{y}_i - \delta(i, y)$$

$$\frac{\partial L}{\partial w_i} = x(\hat{y}_i - \delta(i, y))$$

$$\frac{\partial L}{\partial b_i} = \hat{y}_i - \delta(i, y)$$

$$\begin{aligned} \frac{\partial L}{\partial w_1} &= x(\hat{y}_1 - \delta(1, y)) \\ &= -0.385 * 1.4 = -0.539 \end{aligned}$$

$$\begin{aligned} \frac{\partial L}{\partial w_2} &= x(\hat{y}_2 - \delta(2, y)) \\ &= 0.385 * 1.4 = 0.539 \end{aligned}$$



Simple Example

Update parameters

$$\theta = \theta - \eta L'_\theta$$

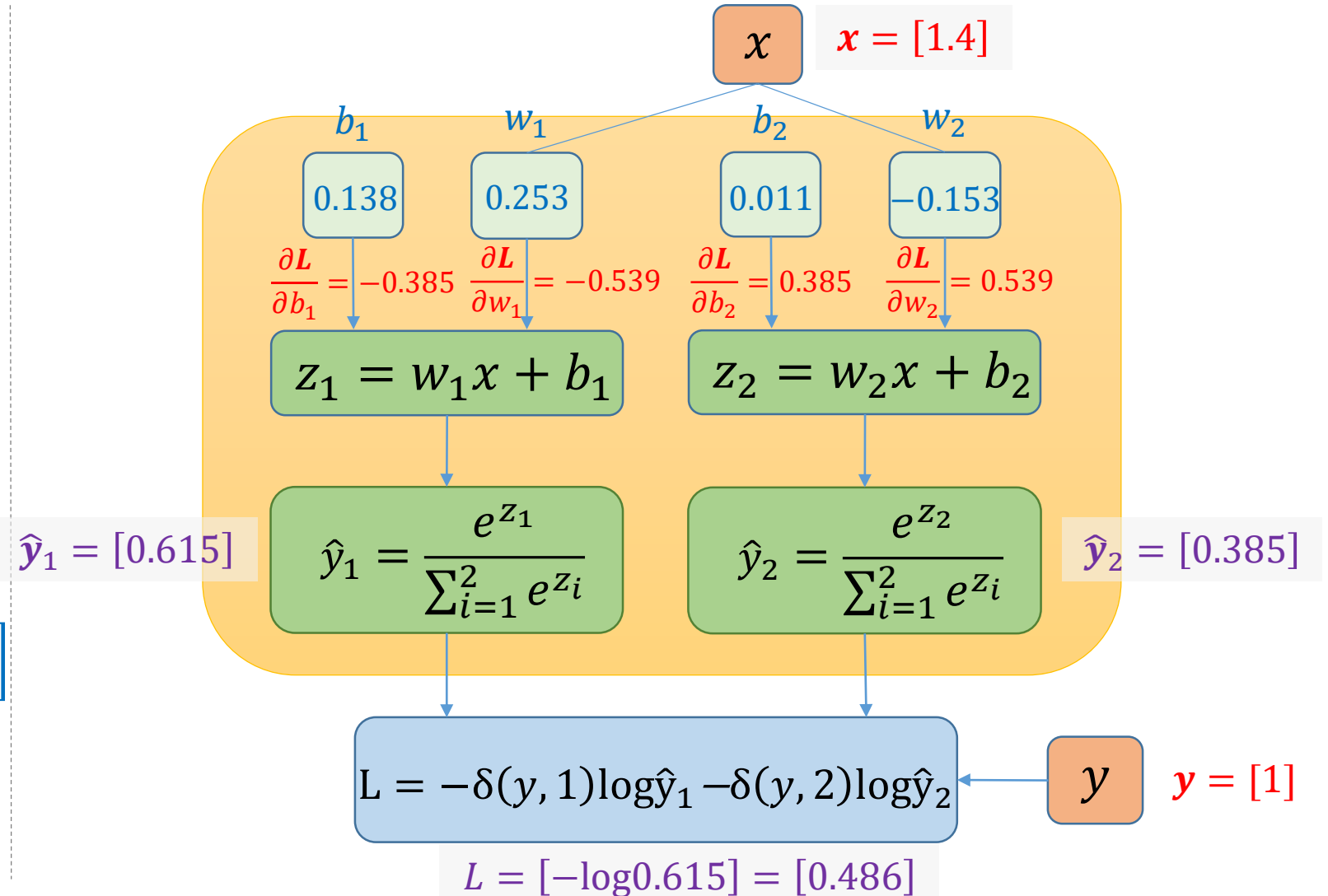
η is learning rate

$$\theta = \begin{bmatrix} b_1 & b_2 \\ w_1 & w_2 \end{bmatrix} \quad L'_\theta = \begin{bmatrix} \frac{\partial L}{\partial b_1} & \frac{\partial L}{\partial b_2} \\ \frac{\partial L}{\partial w_1} & \frac{\partial L}{\partial w_2} \end{bmatrix}$$

$\eta = 0.1$

$$\theta = \begin{bmatrix} 0.1 & 0.05 \\ 0.2 & -0.1 \end{bmatrix} - 0.1 \begin{bmatrix} -0.385 & 0.385 \\ -0.539 & 0.539 \end{bmatrix}$$

$$= \begin{bmatrix} 0.138 & 0.011 \\ 0.253 & -0.153 \end{bmatrix}$$



Generalization

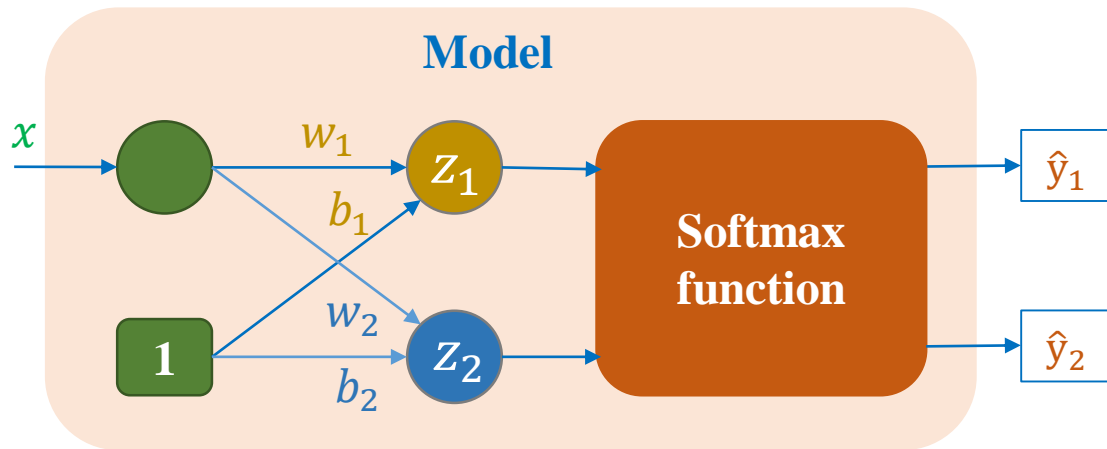
Feature Label

Petal_Length	Label
1.4	1
1.3	1
1.5	1
4.5	2
4.1	2
4.6	2

$$\theta = [\theta_1 \quad \theta_2] = \begin{bmatrix} b_1 & b_2 \\ w_1 & w_2 \end{bmatrix}$$

$$x = \begin{bmatrix} 1 \\ x \end{bmatrix}$$

#feature n=1 #class k=2



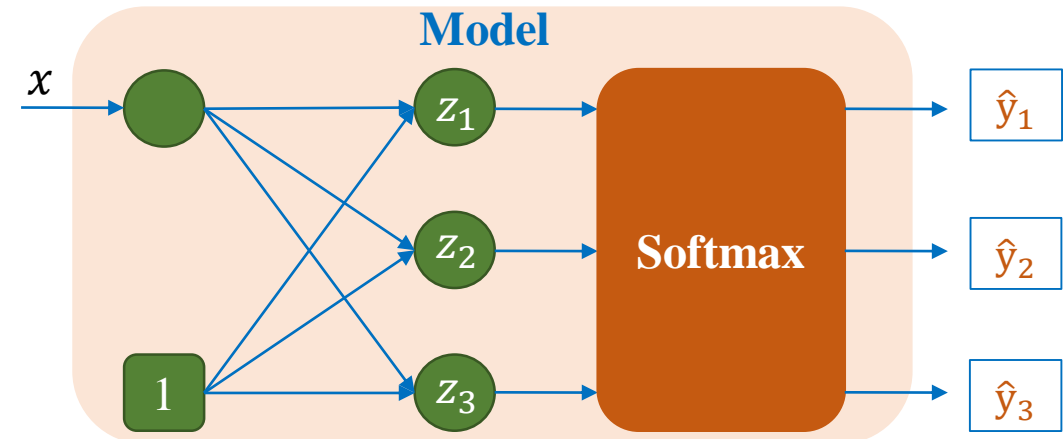
Petal_Length	Label
1.4	1
1.3	1
1.5	1
4.5	2
4.1	2
4.6	2
5.2	3
5.6	3
5.9	3

$$\theta = [\theta_1 \quad \theta_2 \quad \theta_3]$$

$$= \begin{bmatrix} b_1 & b_2 & b_3 \\ w_1 & w_2 & w_3 \end{bmatrix}$$

$$x = \begin{bmatrix} 1 \\ x \end{bmatrix}$$

#feature n=1 #class K=3



Generalization - Stochastic

Feature				Label
Sepal_Length	Sepal_Width	Petal_Length	Petal_Width	Label
5.2	3.5	1.5	0.2	1
5.2	3.4	1.4	0.2	1
4.7	3.2	1.6	0.2	1
6.3	3.3	4.7	1.6	2
4.9	2.4	3.3	1.1	2
6.6	2.9	4.6	1.3	2
6.4	2.8	5.6	2.2	3
6.3	2.8	5.1	1.5	3
6.1	2.6	5.6	1.4	3

#feature n=4

#class k=3

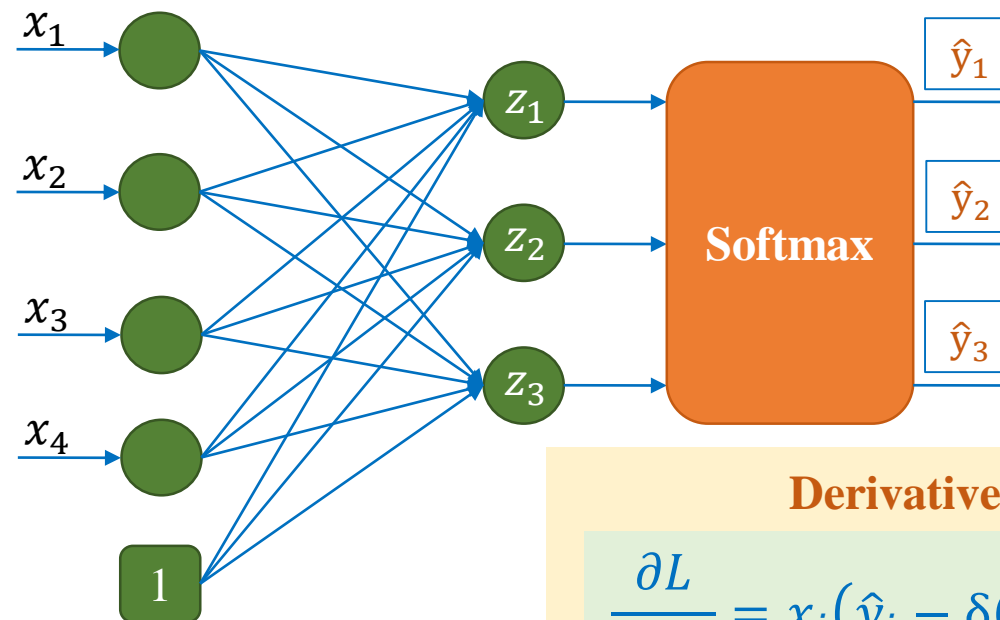


$$\theta = [\theta_1 \quad \dots \quad \theta_k]$$

$$= \begin{bmatrix} b_1 & \dots & b_k \\ w_{11} & \dots & w_{k1} \\ \dots & \dots & \dots \\ w_{1n} & \dots & w_{kn} \end{bmatrix}$$

$$x = \begin{bmatrix} x_0 \\ x_1 \\ \dots \\ x_n \end{bmatrix}$$

$x_0 = 1$



Forward computation

$$z = \theta^T x \quad \hat{y} = \frac{e^z}{\sum_{i=1}^k e^{z_i}}$$

Loss function

$$L(\theta) = - \sum_{i=1}^k \delta(i, y) \log \hat{y}_i$$

Derivative

$$\frac{\partial L}{\partial w_{ij}} = x_j (\hat{y}_i - \delta(i, y))$$

$$\frac{\partial L}{\partial b_i} = \hat{y}_i - \delta(i, y)$$



$$\frac{\partial L}{\partial \theta_i} = x (\hat{y}_i - \delta(i, y))$$

Generalization - Batch

Feature Label

Sepal_Length	Sepal_Width	Petal_Length	Petal_Width	Label
5.2	3.5	1.5	0.2	1
5.2	3.4	1.4	0.2	1
4.7	3.2	1.6	0.2	1
6.3	3.3	4.7	1.6	2
4.9	2.4	3.3	1.1	2
6.6	2.9	4.6	1.3	2
6.4	2.8	5.6	2.2	3
6.3	2.8	5.1	1.5	3
6.1	2.6	5.6	1.4	3

#feature n=4 #class k=3 #example m=9

$$\theta = [\theta_1 \quad \dots \quad \theta_k]$$

$$= \begin{bmatrix} b_1 & \dots & b_k \\ w_{11} & \dots & w_{k1} \\ \vdots & & \vdots \\ w_{1n} & \dots & w_{kn} \end{bmatrix}$$

$$x = [x^{(1)} \quad \dots \quad x^{(m)}]$$

$$x = \begin{bmatrix} x_0^{(1)} & \dots & x_0^{(m)} \\ x_1^{(1)} & \dots & x_1^{(m)} \\ \vdots & & \vdots \\ x_n^{(1)} & \dots & x_n^{(m)} \end{bmatrix}$$

Forward computation

$$z = \theta^T x = \begin{bmatrix} z_0^{(1)} & \dots & z_0^{(m)} \\ z_1^{(1)} & \dots & z_1^{(m)} \\ \vdots & & \vdots \\ z_k^{(1)} & \dots & z_k^{(m)} \end{bmatrix}$$

$$\hat{y} = \frac{e^z}{\sum_{i=1}^k e^{z_i}} = \begin{bmatrix} \hat{y}_0^{(1)} & \dots & \hat{y}_0^{(m)} \\ \hat{y}_1^{(1)} & \dots & \hat{y}_1^{(m)} \\ \vdots & & \vdots \\ \hat{y}_k^{(1)} & \dots & \hat{y}_k^{(m)} \end{bmatrix}$$

Loss function

$$L(\theta) = -\frac{1}{m} \sum_{u=1}^m \sum_{i=1}^k \delta(i, y^{(u)}) \log \hat{y}_i^{(u)}$$

Derivative

$$\frac{\partial L^{(u)}}{\partial w_{ij}} = x_j \left(\hat{y}_i^{(u)} - \delta(i, y^{(u)}) \right)$$

$$\frac{\partial L^{(u)}}{\partial b_i} = \hat{y}_i^{(u)} - \delta(i, y^{(u)})$$



$$\frac{\partial L}{\partial \theta_i} = \frac{1}{m} \sum_{u=1}^m x \left(\hat{y}_i^{(u)} - \delta(i, y^{(u)}) \right)$$

Outline

- **Motivation**
- **Model Construction**
- **Loss Function**
- **Simple Example and Generalization**
- **Examples - Stochastic and Batch**
- **Another Approach**

Softmax Regression - Stochastic

Petal_Length	Petal_Width	Label
1.5	0.2	1
1.4	0.2	1
1.6	0.2	1
4.7	1.6	2
3.3	1.1	2
4.6	1.3	2
5.6	2.2	3
5.1	1.5	3
5.6	1.4	3

#feature n=2

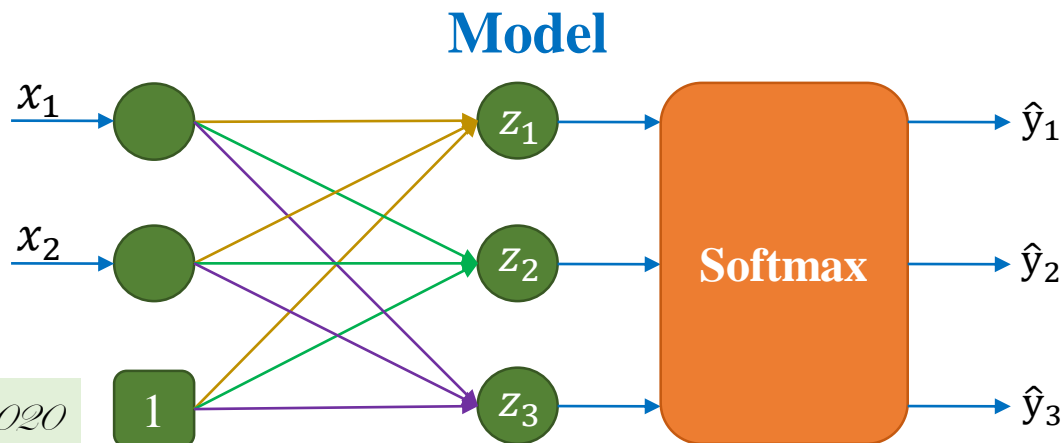
#example m=9

#class k=3

$$\theta = [\theta_1 \quad \theta_2 \quad \theta_3]$$

$$= \begin{bmatrix} b_1 & b_2 & b_3 \\ w_{11} & w_{21} & w_{31} \\ w_{12} & w_{22} & w_{32} \end{bmatrix}$$

$$x = \begin{bmatrix} x_0 \\ x_1 \\ x_2 \end{bmatrix} \quad x_0 = 1$$



1) Pick a sample (x, y) from training data

2) Tính output \hat{y}

$$z = \theta^T x$$

$$\hat{y} = \frac{e^z}{\sum_{i=1}^k e^{z_i}}$$

3) Tính loss (cross-entropy)

$$L(\theta) = - \sum_{i=1}^k \delta(i, y) \log \hat{y}_i$$

4) Tính đạo hàm

$$\frac{\partial L}{\partial \theta_i} = x(\hat{y}_i - \delta(i, y))$$

5) Cập nhật tham số

$$\theta = \theta - \eta L'_\theta$$

η is learning rate

Softmax Regression - Stochastic

Petal_Length	Petal_Width	Label
1.5	0.2	1
1.4	0.2	1
1.6	0.2	1
4.7	1.6	2
3.3	1.1	2
4.6	1.3	2
5.6	2.2	3
5.1	1.5	3
5.6	1.4	3

$$\theta = [\theta_1 \quad \theta_2 \quad \theta_3]$$

$$= \begin{bmatrix} 0.1 & 0.05 & -0.1 \\ 0.1 & -0.1 & 0.1 \\ 0.2 & 0.2 & -0.1 \end{bmatrix}$$

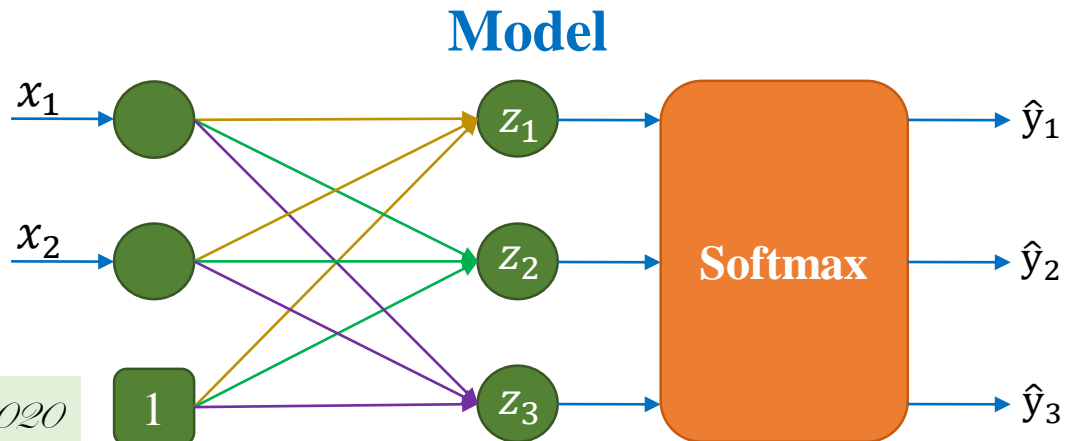
$$x = \begin{bmatrix} 1 \\ 1.4 \\ 0.2 \end{bmatrix} \quad y = 1$$

1) Pick a sample (x, y) from training data

#feature n=2

#example m=9

#class k=3



Softmax Regression - Stochastic

$$\begin{aligned}\theta &= [\theta_1 \quad \theta_2 \quad \theta_3] \\ &= \begin{bmatrix} 0.1 & 0.05 & -0.1 \\ 0.1 & -0.1 & 0.1 \\ 0.2 & 0.2 & -0.1 \end{bmatrix} \\ \mathbf{x} &= \begin{bmatrix} 1 \\ 1.4 \\ 0.2 \end{bmatrix} \quad y = 1\end{aligned}$$

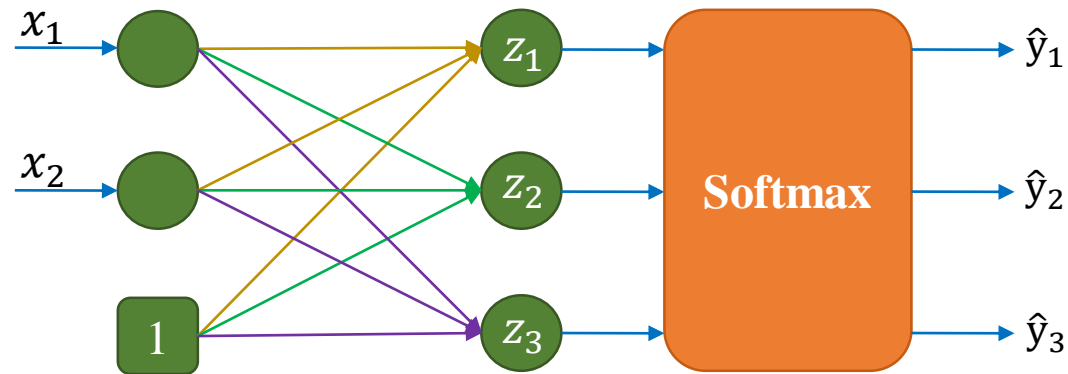
2) Tính output \hat{y}

$$\begin{aligned}\mathbf{z} &= \theta^T \mathbf{x} \\ \hat{y} &= \frac{e^z}{\sum_{i=1}^k e^{z_i}}\end{aligned}$$

3) Tính loss (cross-entropy)

$$L(\theta) = - \sum_{i=1}^k \delta(i, y) \log \hat{y}_i$$

$$\mathbf{z} = \begin{bmatrix} 0.1 & 0.1 & 0.2 \\ 0.05 & -0.1 & 0.2 \\ -0.1 & 0.1 & -0.1 \end{bmatrix} \begin{bmatrix} 1 \\ 1.4 \\ 0.2 \end{bmatrix} = \begin{bmatrix} 0.28 \\ -0.05 \\ 0.02 \end{bmatrix} \quad \hat{\mathbf{y}} = \begin{bmatrix} 0.4016 \\ 0.2887 \\ 0.3096 \end{bmatrix}$$



$$\begin{aligned}\text{\#class } k &= 3 \\ L(\theta) &= - \sum_{i=1}^3 \delta(i, 1) \log \hat{y}_i \\ &= -\delta(1, 1) \log \hat{y}_1 \\ &= -\log 0.4016 = 0.9122\end{aligned}$$

Softmax Regression - Stochastic

$$\theta = [\theta_1 \quad \theta_2 \quad \theta_3]$$

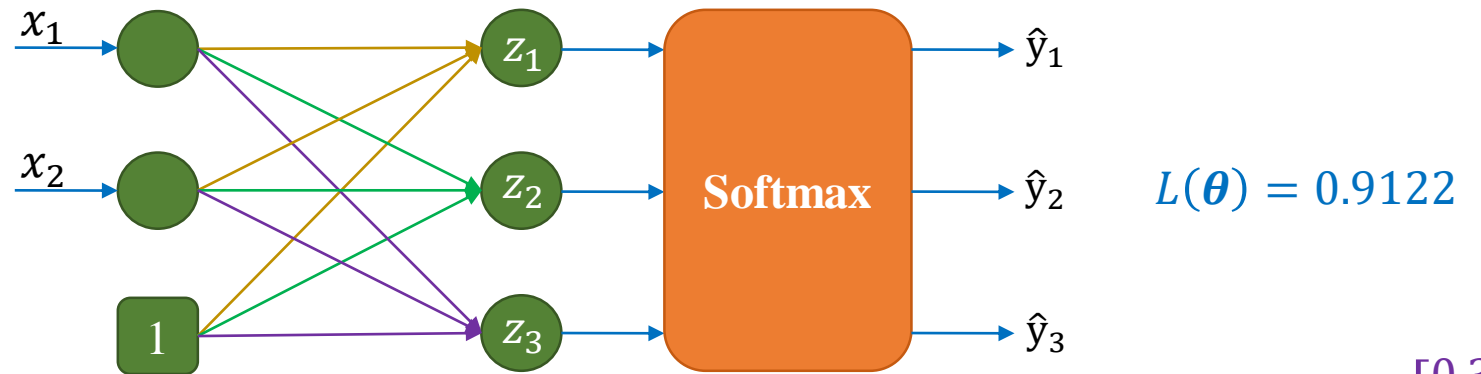
$$= \begin{bmatrix} 0.1 & 0.05 & -0.1 \\ 0.1 & -0.1 & 0.1 \\ 0.2 & 0.2 & -0.1 \end{bmatrix}$$

$$\mathbf{x} = \begin{bmatrix} 1 \\ 1.4 \\ 0.2 \end{bmatrix} \quad y = 1$$

4) Tính đạo hàm

$$\frac{\partial L}{\partial \theta_i} = \mathbf{x}(\hat{y}_i - \delta(i, y))$$

$$\mathbf{z} = \begin{bmatrix} 0.1 & 0.1 & 0.2 \\ 0.05 & -0.1 & 0.2 \\ -0.1 & 0.1 & -0.1 \end{bmatrix} \begin{bmatrix} 1 \\ 1.4 \\ 0.2 \end{bmatrix} = \begin{bmatrix} 0.28 \\ -0.05 \\ 0.02 \end{bmatrix} \quad \hat{\mathbf{y}} = \begin{bmatrix} 0.4016 \\ 0.2887 \\ 0.3096 \end{bmatrix}$$



$$\frac{\partial L}{\partial \theta_3} = \begin{bmatrix} 0.309 \\ 0.433 \\ 0.061 \end{bmatrix}$$

$$\frac{\partial L}{\partial \theta_1} = \mathbf{x}(\hat{y}_1 - \delta(1, y))$$

$$= \begin{bmatrix} 1 \\ 1.4 \\ 0.2 \end{bmatrix} (0.4016 - 1) = \begin{bmatrix} -0.598 \\ -0.837 \\ -0.119 \end{bmatrix}$$

$$\frac{\partial L}{\partial \theta_2} = \mathbf{x}(\hat{y}_2 - \delta(2, y))$$

$$= \begin{bmatrix} 1 \\ 1.4 \\ 0.2 \end{bmatrix} 0.2887 = \begin{bmatrix} 0.288 \\ 0.404 \\ 0.057 \end{bmatrix}$$

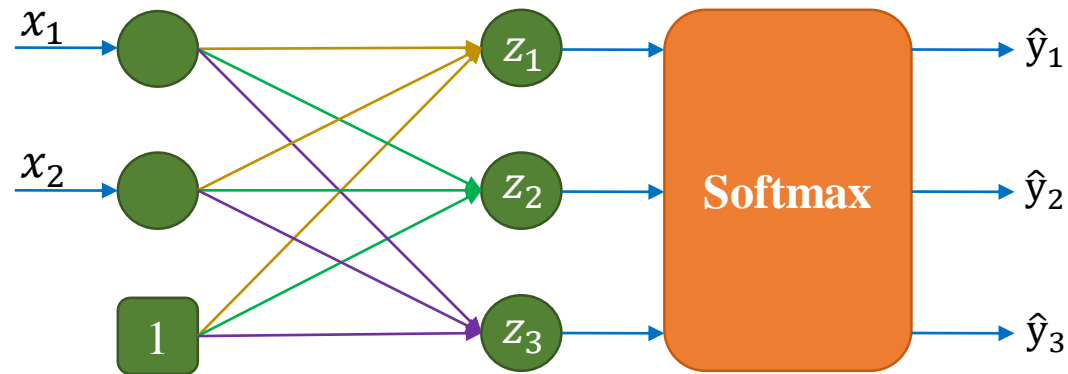
Softmax Regression - Stochastic

$$\begin{aligned}\theta &= [\theta_1 \quad \theta_2 \quad \theta_3] \\ &= \begin{bmatrix} 0.1 & 0.05 & -0.1 \\ 0.1 & -0.1 & 0.1 \\ 0.2 & 0.2 & -0.1 \end{bmatrix} \\ \mathbf{x} &= \begin{bmatrix} 1 \\ 1.4 \\ 0.2 \end{bmatrix} \quad y = 1\end{aligned}$$

5) Cập nhật tham số

$$\begin{aligned}\theta &= \theta - \eta L'_\theta \\ \eta &= 0.1\end{aligned}$$

$$\mathbf{z} = \begin{bmatrix} 0.1 & 0.1 & 0.2 \\ 0.05 & -0.1 & 0.2 \\ -0.1 & 0.1 & -0.1 \end{bmatrix} \begin{bmatrix} 1 \\ 1.4 \\ 0.2 \end{bmatrix} = \begin{bmatrix} 0.28 \\ -0.05 \\ 0.02 \end{bmatrix} \quad \hat{\mathbf{y}} = \begin{bmatrix} 0.4016 \\ 0.2887 \\ 0.3096 \end{bmatrix}$$



$$L(\theta) = 0.9122$$

$$\theta = \begin{bmatrix} 0.1 & 0.05 & -0.1 \\ 0.1 & -0.1 & 0.1 \\ 0.2 & 0.2 & -0.1 \end{bmatrix} - 0.1 \begin{bmatrix} -0.598 & 0.288 & 0.309 \\ -0.837 & 0.404 & 0.433 \\ -0.119 & 0.057 & 0.061 \end{bmatrix}$$

$$= \begin{bmatrix} 0.159 & 0.021 & -0.13 \\ 0.183 & -0.14 & 0.056 \\ 0.211 & 0.194 & -0.11 \end{bmatrix}$$

$$\frac{\partial L}{\partial \theta_1} = \begin{bmatrix} -0.598 \\ -0.837 \\ -0.119 \end{bmatrix}$$

$$\frac{\partial L}{\partial \theta_2} = \begin{bmatrix} 0.288 \\ 0.404 \\ 0.057 \end{bmatrix}$$

$$\frac{\partial L}{\partial \theta_3} = \begin{bmatrix} 0.309 \\ 0.433 \\ 0.061 \end{bmatrix}$$

Outline

- **Motivation**
- **Model Construction**
- **Loss Function**
- **Simple Example and Generalization**
- **Examples - Stochastic and Batch**
- **Another Approach**

Softmax Regression - Minibatch

Petal_Length	Petal_Width	Label
1.5	0.2	1
1.4	0.2	1
1.6	0.2	1
4.7	1.6	2
3.3	1.1	2
4.6	1.3	2
5.6	2.2	3
5.1	1.5	3
5.6	1.4	3

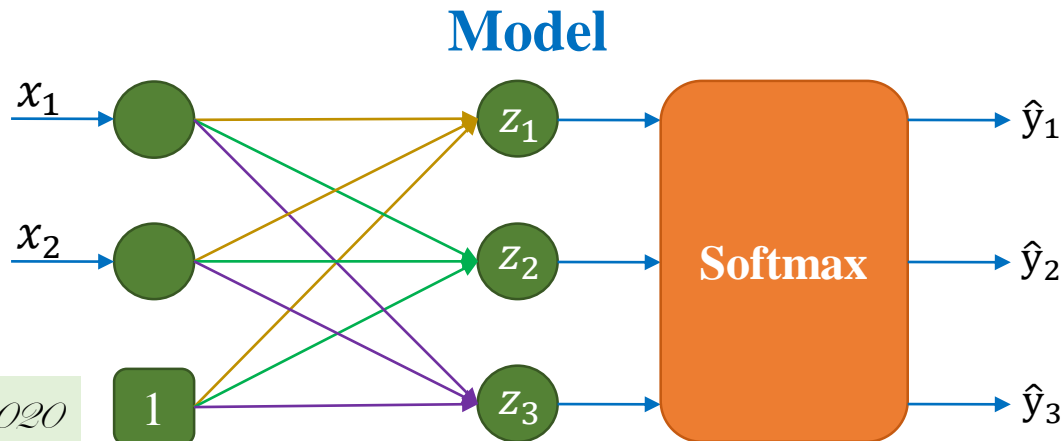
#feature n=2 #class k=3
#example m=9 #minibatch s=3

$$\theta = [\theta_1 \quad \theta_2 \quad \theta_3]$$

$$= \begin{bmatrix} b_1 & b_2 & b_3 \\ w_{11} & w_{21} & w_{31} \\ w_{12} & w_{22} & w_{32} \end{bmatrix}$$

$$x = \begin{bmatrix} x_0^{(1)} & x_0^{(2)} & x_0^{(3)} \\ x_1^{(1)} & x_1^{(2)} & x_1^{(3)} \\ x_2^{(1)} & x_2^{(2)} & x_2^{(3)} \end{bmatrix}$$

$$x_0^{(1)} = 1$$



1) Pick s samples (\mathbf{x}, \mathbf{y})

2) Tính output $\hat{\mathbf{y}}$

$$\mathbf{z} = \boldsymbol{\theta}^T \mathbf{x}$$

$$\hat{y} = \frac{e^z}{\sum_{i=1}^k e^{z_i}}$$

3) Tính loss (cross-entropy)

$$L(\boldsymbol{\theta}) = -\frac{1}{s} \sum_{u=1}^s \sum_{i=1}^k \delta(i, y^{(u)}) \log \hat{y}_i^{(u)}$$

4) Tính đạo hàm

$$\frac{\partial L}{\partial \theta_i} = \frac{1}{s} \sum_{u=1}^s x^{(i)} \left(\hat{y}_i^{(u)} - \delta(i, y^{(u)}) \right)$$

5) Cập nhật tham số

$$\boldsymbol{\theta} = \boldsymbol{\theta} - \eta L'_{\boldsymbol{\theta}}$$

η is learning rate

Softmax Regression - Minibatch

Petal_Length	Petal_Width	Label
1.5	0.2	1
1.4	0.2	1
1.6	0.2	1
4.7	1.6	2
3.3	1.1	2
4.6	1.3	2
5.6	2.2	3
5.1	1.5	3
5.6	1.4	3

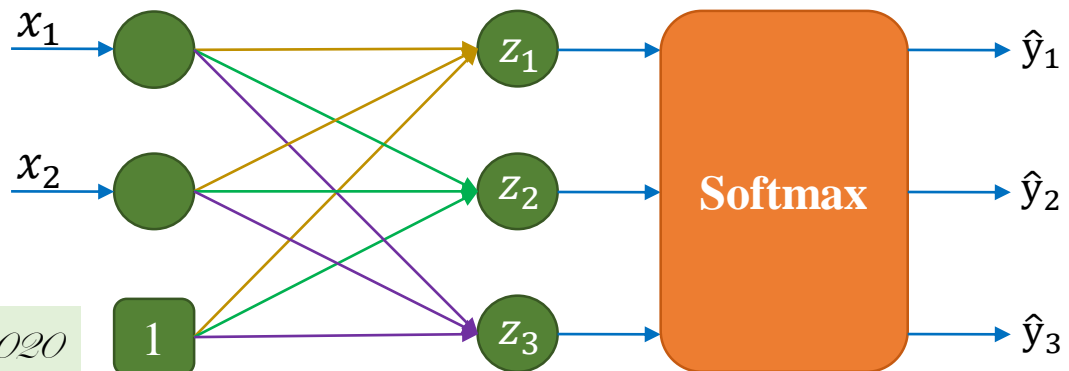
$$\theta = [\theta_1 \quad \theta_2 \quad \theta_3]$$

$$= \begin{bmatrix} 0.1 & 0.05 & -0.1 \\ 0.1 & -0.1 & 0.1 \\ 0.2 & 0.2 & -0.1 \end{bmatrix}$$

$$x = \begin{bmatrix} 1 & 1 & 1 \\ 1.5 & 4.7 & 5.6 \\ 0.2 & 1.6 & 2.2 \end{bmatrix}$$

$$y = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

1) Pick s samples (x, y)



Softmax Regression - Minibatch

$$\theta = [\theta_1 \quad \theta_2 \quad \theta_3]$$

$$= \begin{bmatrix} 0.1 & 0.05 & -0.1 \\ 0.1 & -0.1 & 0.1 \\ 0.2 & 0.2 & -0.1 \end{bmatrix}$$

$$x = \begin{bmatrix} 1 & 1 & 1 \\ 1.5 & 4.7 & 5.6 \\ 0.2 & 1.6 & 2.2 \end{bmatrix} \quad y = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

2) Tính output \hat{y}

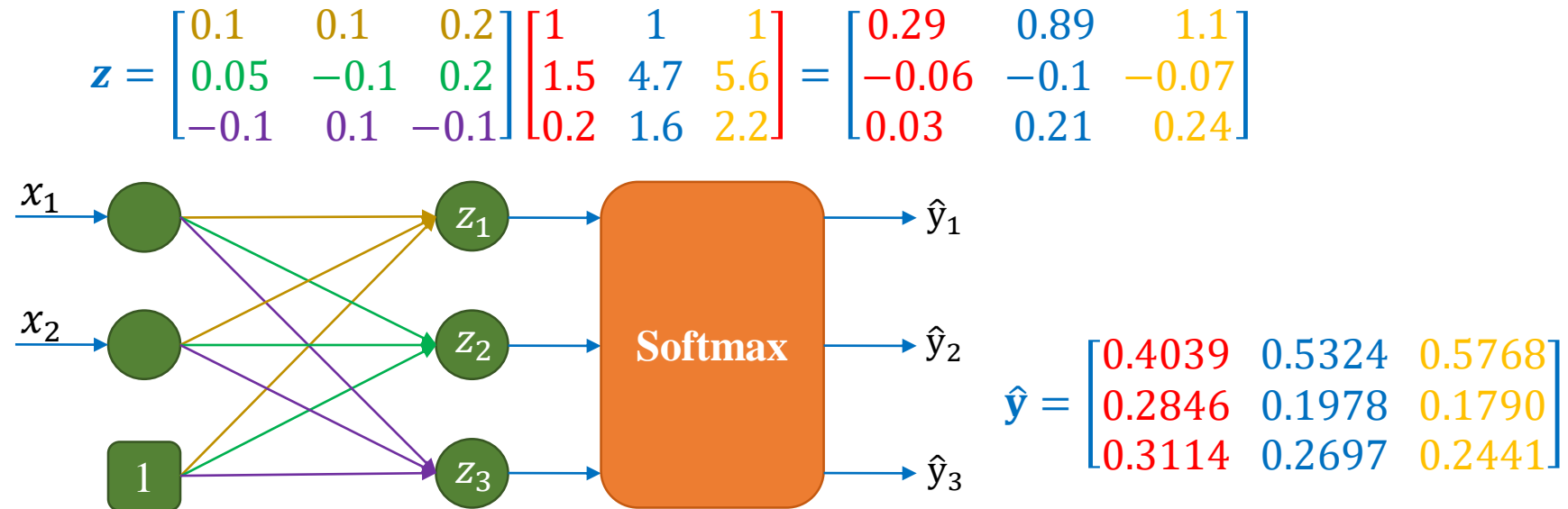
$$z = \theta^T x$$

$$\hat{y} = \frac{e^z}{\sum_{i=1}^k e^{z_i}}$$

3) Tính loss (cross-entropy)

$$L(\theta) = -\frac{1}{s} \sum_{u=1}^s \sum_{i=1}^k \delta(i, y^{(u)}) \log \hat{y}_i^{(u)}$$

#class k=3 #minibatch s=3



$$L(\theta) = -\frac{1}{3} \left[\sum_{i=1}^k \delta(i, y^{(1)}) \log \hat{y}_i^{(1)} + \sum_{i=1}^k \delta(i, y^{(2)}) \log \hat{y}_i^{(2)} + \sum_{i=1}^k \delta(i, y^{(3)}) \log \hat{y}_i^{(3)} \right]$$

$$= -\frac{1}{3} \left[\delta(1, y^{(1)}) \log \hat{y}_1^{(1)} + \delta(2, y^{(2)}) \log \hat{y}_2^{(2)} + \delta(3, y^{(3)}) \log \hat{y}_3^{(3)} \right]$$

$$= -\frac{1}{3} [\log 0.4039 + \log 0.1978 + \log 0.2441]$$

$$= -\frac{1}{3} [-0.90 - 1.62 - 1.41] = 1.31$$

Softmax Regression - Minibatch

$$\theta = [\theta_1 \quad \theta_2 \quad \theta_3]$$

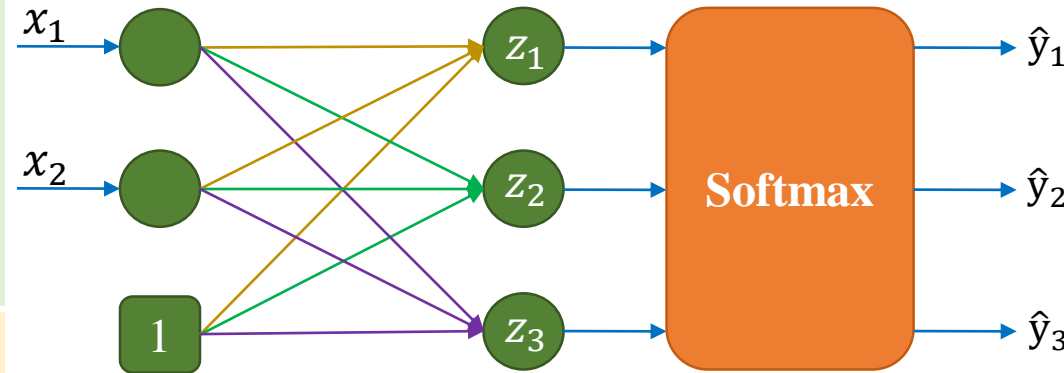
$$= \begin{bmatrix} 0.1 & 0.05 & -0.1 \\ 0.1 & -0.1 & 0.1 \\ 0.2 & 0.2 & -0.1 \end{bmatrix}$$

$$\mathbf{x} = \begin{bmatrix} 1 & 1 & 1 \\ 1.5 & 4.7 & 5.6 \\ 0.2 & 1.6 & 2.2 \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

4) Tính đạo hàm

$$\frac{\partial L}{\partial \theta_i} = \frac{1}{s} \sum_{u=1}^s \mathbf{x}^{(i)} \left(\hat{y}_i^{(u)} - \delta(i, y^{(u)}) \right)$$

$$\mathbf{z} = \begin{bmatrix} 0.1 & 0.1 & 0.2 \\ 0.05 & -0.1 & 0.2 \\ -0.1 & 0.1 & -0.1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 1.5 & 4.7 & 5.6 \\ 0.2 & 1.6 & 2.2 \end{bmatrix} = \begin{bmatrix} 0.29 & 0.89 & 1.1 \\ -0.06 & -0.1 & -0.07 \\ 0.03 & 0.21 & 0.24 \end{bmatrix}$$



$$\hat{\mathbf{y}} = \begin{bmatrix} 0.4039 & 0.5324 & 0.5768 \\ 0.2846 & 0.1978 & 0.1790 \\ 0.3114 & 0.2697 & 0.2441 \end{bmatrix}$$

$$L(\theta) = 1.31$$

$$\begin{aligned} \frac{\partial L}{\partial \theta_1} &= -\frac{1}{3} \left[\mathbf{x}^{(1)} (\hat{y}_1^{(1)} - 1) + \mathbf{x}^{(1)} (\hat{y}_1^{(2)}) + \mathbf{x}^{(1)} (\hat{y}_1^{(3)}) \right] \\ &= -\frac{1}{3} \left[\begin{bmatrix} 1 \\ 1.5 \\ 0.2 \end{bmatrix} (0.4039 - 1) + \begin{bmatrix} 1 \\ 1.5 \\ 0.2 \end{bmatrix} 0.1978 + \begin{bmatrix} 1 \\ 1.5 \\ 0.2 \end{bmatrix} 0.2441 \right] = \begin{bmatrix} 0.171 \\ 1.612 \\ 0.667 \end{bmatrix} \end{aligned}$$

$$\frac{\partial L}{\partial \theta_2} = \begin{bmatrix} -0.112 \\ -0.780 \\ -0.277 \end{bmatrix}$$

$$\frac{\partial L}{\partial \theta_3} = \begin{bmatrix} -0.058 \\ -0.832 \\ -0.389 \end{bmatrix}$$

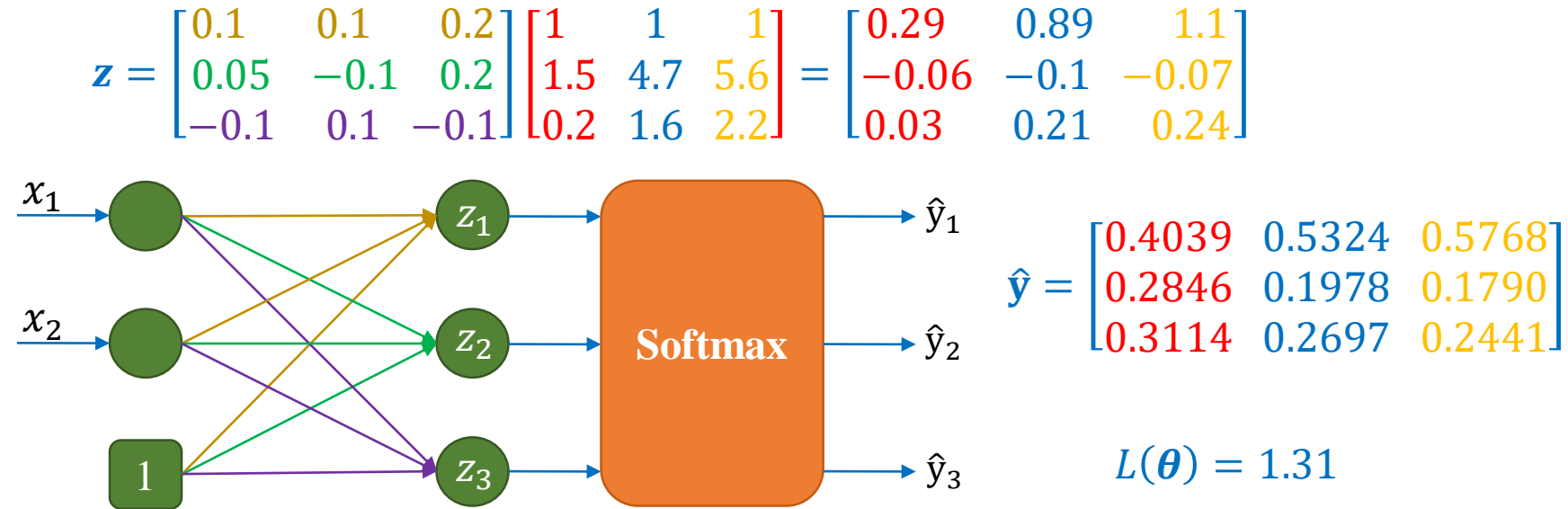
Softmax Regression - Minibatch

$$\begin{aligned}\theta &= [\theta_1 \quad \theta_2 \quad \theta_3] \\ &= \begin{bmatrix} 0.1 & 0.05 & -0.1 \\ 0.1 & -0.1 & 0.1 \\ 0.2 & 0.2 & -0.1 \end{bmatrix} \\ x &= \begin{bmatrix} 1 & 1 & 1 \\ 1.5 & 4.7 & 5.6 \\ 0.2 & 1.6 & 2.2 \end{bmatrix} \quad y = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}\end{aligned}$$

5) Cập nhật tham số

$$\theta = \theta - \eta L'_\theta$$

$$\eta = 0.1$$



$$\begin{aligned}\theta &= \begin{bmatrix} 0.1 & 0.05 & -0.1 \\ 0.1 & -0.1 & 0.1 \\ 0.2 & 0.2 & -0.1 \end{bmatrix} - 0.1 \begin{bmatrix} 0.171 & -0.112 & -0.058 \\ 1.612 & -0.780 & -0.832 \\ 0.667 & -0.277 & -0.389 \end{bmatrix} \\ &= \begin{bmatrix} 0.083 & 0.061 & -0.094 \\ -0.061 & -0.022 & 0.183 \\ 0.133 & 0.228 & -0.061 \end{bmatrix}\end{aligned}$$

$$\frac{\partial L}{\partial \theta_1} = \begin{bmatrix} 0.171 \\ 1.612 \\ 0.667 \end{bmatrix}$$

$$\frac{\partial L}{\partial \theta_2} = \begin{bmatrix} -0.112 \\ -0.780 \\ -0.277 \end{bmatrix}$$

$$\frac{\partial L}{\partial \theta_3} = \begin{bmatrix} -0.058 \\ -0.832 \\ -0.389 \end{bmatrix}$$

Softmax Regression

❖ Demo

```
Python 3.7.3 (default, Apr 24 2019, 15:29:51) [MSC v.1915 64 bit (AMD64)] ::  
Type "help", "copyright", "credits" or "license" for more information.  
>>>  
>>>  
>>>  
>>>  
>>>  
>>>  
>>>  
>>>  
>>> for epoch in range(n_epochs):  
...     sum_of_losses = 0  
...     gradients = np.zeros((2,1))  
...  
...     for index in range(4):  
...         xi = X_b[index:index+1]  
...         yi = y[index:index+1]
```

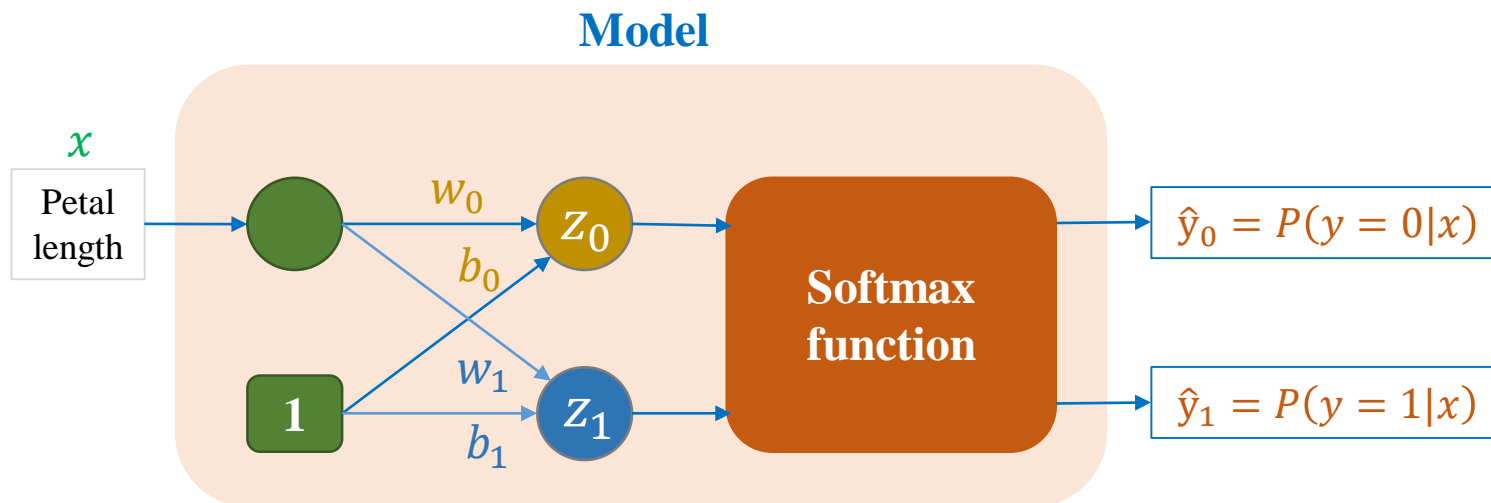
Outline

- **Motivation**
- **Model Construction**
- **Loss Function**
- **Simple Example and Generalization**
- **Examples - Stochastic and Batch**
- **Another Approach**

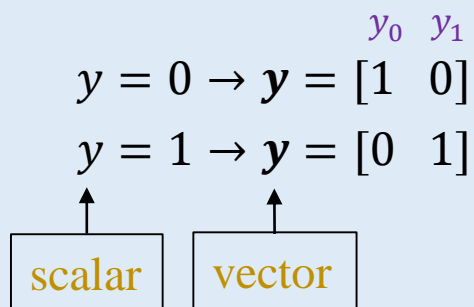
Another Approach

❖ Simple illustration

Feature	Label
Petal_Length	Category
1.4	0
1	0
1.5	0
3	1
3.8	1
4.1	1



One-hot encoding for label



$$z_0 = xw_0 + b_0$$

$$z_1 = xw_1 + b_1$$

$$\hat{y}_0 = \frac{e^{z_0}}{\sum_{j=0}^1 e^{z_j}}$$

$$\hat{y}_1 = \frac{e^{z_1}}{\sum_{j=0}^1 e^{z_j}}$$

$$\mathbf{z} = \begin{bmatrix} z_0 \\ z_1 \end{bmatrix} = \begin{bmatrix} b_0 & w_0 \\ b_1 & w_1 \end{bmatrix} \begin{bmatrix} 1 \\ x \end{bmatrix} = \begin{bmatrix} \theta_0^T \\ \theta_1^T \end{bmatrix} \begin{bmatrix} 1 \\ x \end{bmatrix} = \boldsymbol{\theta}^T \mathbf{x}$$

$$\hat{\mathbf{y}} = \begin{bmatrix} \hat{y}_0 \\ \hat{y}_1 \end{bmatrix} = \frac{1}{\sum_{j=0}^1 e^{z_j}} \begin{bmatrix} e^{z_0} \\ e^{z_1} \end{bmatrix} = \frac{e^{\mathbf{z}}}{\sum_{j=0}^1 e^{z_j}}$$

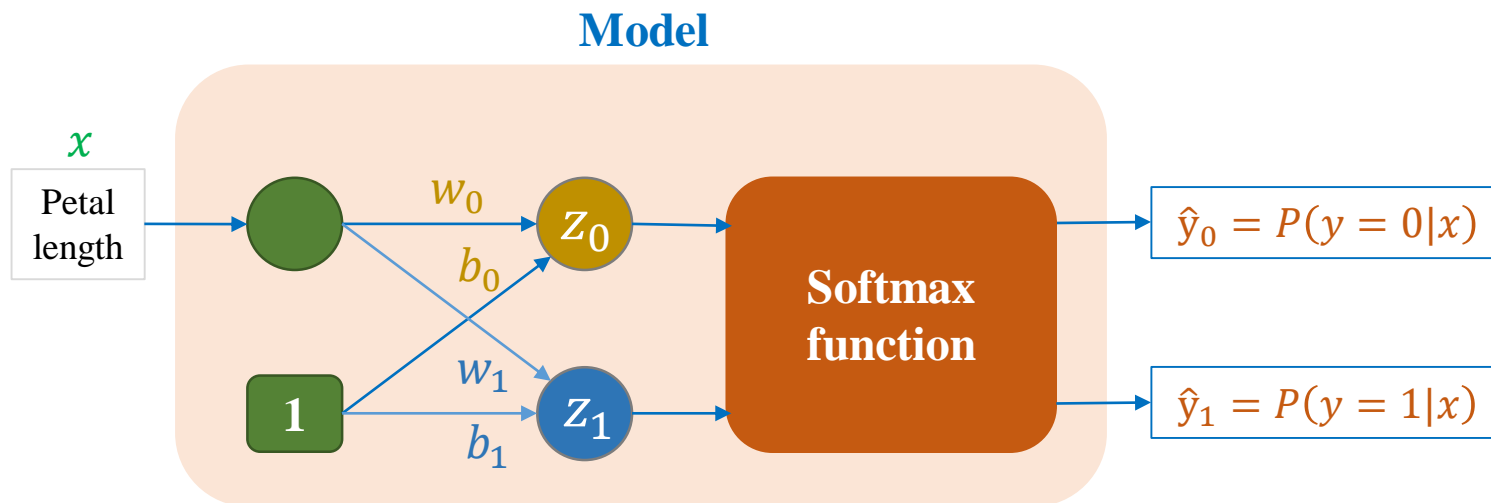
A vector is by default a column vector $\boldsymbol{\theta}_0 = \begin{bmatrix} b_0 \\ w_0 \end{bmatrix}$

vector transpose $\boldsymbol{\theta}_0^T = [b_0 \ w_0]$

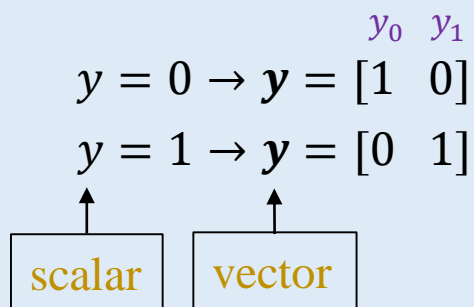
Another Approach

❖ Simple illustration

Feature	Label
Petal Length	Category
1.4	0
1	0
1.5	0
3	1
3.8	1
4.1	1



One-hot encoding for label



$$z_0 = xw_0 + b_0$$

$$z_1 = xw_1 + b_1$$

$$\hat{y}_0 = \frac{e^{z_0}}{\sum_{j=0}^1 e^{z_j}}$$

$$\hat{y}_1 = \frac{e^{z_1}}{\sum_{j=0}^1 e^{z_j}}$$

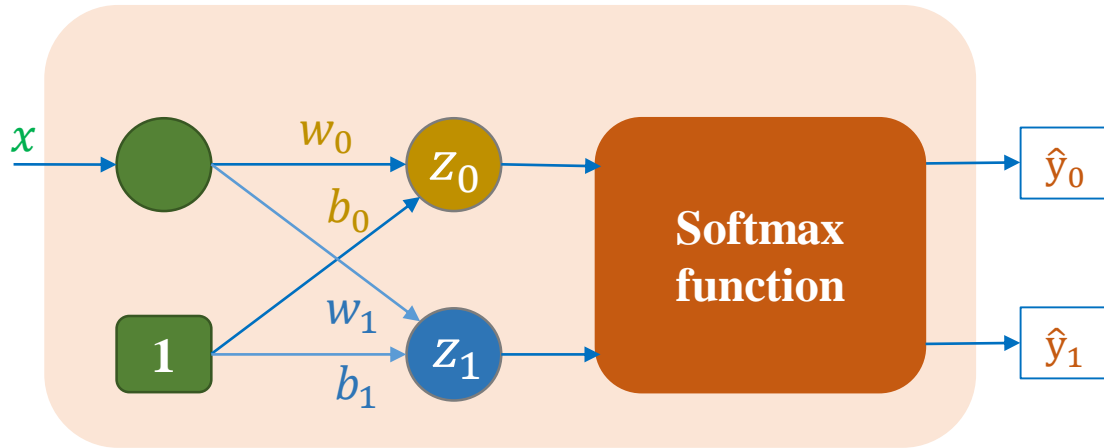
$$\mathbf{z} = \begin{bmatrix} z_0 \\ z_1 \end{bmatrix} = \begin{bmatrix} b_0 & w_0 \\ b_1 & w_1 \end{bmatrix} \begin{bmatrix} 1 \\ x \end{bmatrix} = \begin{bmatrix} \theta_0^T \\ \theta_1^T \end{bmatrix} \begin{bmatrix} 1 \\ x \end{bmatrix} = \boldsymbol{\theta}^T \mathbf{x}$$

$$\hat{\mathbf{y}} = \begin{bmatrix} \hat{y}_0 \\ \hat{y}_1 \end{bmatrix} = \frac{1}{\sum_{j=0}^1 e^{z_j}} \begin{bmatrix} e^{z_0} \\ e^{z_1} \end{bmatrix} = \frac{e^{\mathbf{z}}}{\sum_{j=0}^1 e^{z_j}}$$

$$L(\boldsymbol{\theta}) = -y_0 \log \hat{y}_0 - y_1 \log \hat{y}_1 = - \sum_{i=0}^1 y_i \log \hat{y}_i$$

Another Approach

Model



$$L(\theta) = -y_0 \log \hat{y}_0 - y_1 \log \hat{y}_1 = -\sum_{i=0}^1 y_i \log \hat{y}_i$$

$$\hat{y}_0 = \frac{e^{z_0}}{\sum_{j=0}^1 e^{z_j}}$$

$$\hat{y}_1 = \frac{e^{z_1}}{\sum_{j=0}^1 e^{z_j}}$$

Derivative

$$\frac{\partial \hat{y}_i}{\partial z_i} = \begin{cases} \hat{y}_i(1 - \hat{y}_i) & \text{if } i = j \\ -\hat{y}_i \hat{y}_j & \text{if } i \neq j \end{cases}$$

$$\begin{aligned} \frac{\partial L}{\partial z_i} &= -\sum_k y_k \frac{\partial \log(\hat{y}_k)}{\partial z_i} \\ &= -\sum_k y_k \frac{\partial \log(\hat{y}_k)}{\partial \hat{y}_k} \frac{\partial \hat{y}_k}{\partial z_i} \\ &= -\sum_k y_k \frac{1}{\hat{y}_k} \frac{\partial \hat{y}_k}{\partial z_i} \end{aligned}$$

$$\begin{aligned} \frac{\partial L}{\partial z_i} &= -y_i(1 - \hat{y}_i) - \sum_{k \neq i} y_k \frac{1}{\hat{y}_k} - \hat{y}_k \hat{y}_i \\ &= -y_i(1 - \hat{y}_i) - \sum_{k \neq i} y_k \hat{y}_i \\ &= -y_i + y_i \hat{y}_i - \sum_{k \neq i} y_k \hat{y}_i \\ &= \hat{y}_i \left(y_i - \sum_{k \neq i} y_k \right) - y_i \\ &= \hat{y}_i - y_i \end{aligned}$$

Another Approach

One-hot encoding for label

$$y = 0 \rightarrow \mathbf{y} = \begin{bmatrix} 1 & 0 \end{bmatrix} \quad \begin{matrix} y_0 & y_1 \end{matrix}$$

$$y = 1 \rightarrow \mathbf{y} = \begin{bmatrix} 0 & 1 \end{bmatrix}$$



$$z_0 = xw_0 + b_0$$

$$z_1 = xw_1 + b_1$$

$$\hat{y}_0 = \frac{e^{z_0}}{\sum_{j=0}^1 e^{z_j}}$$

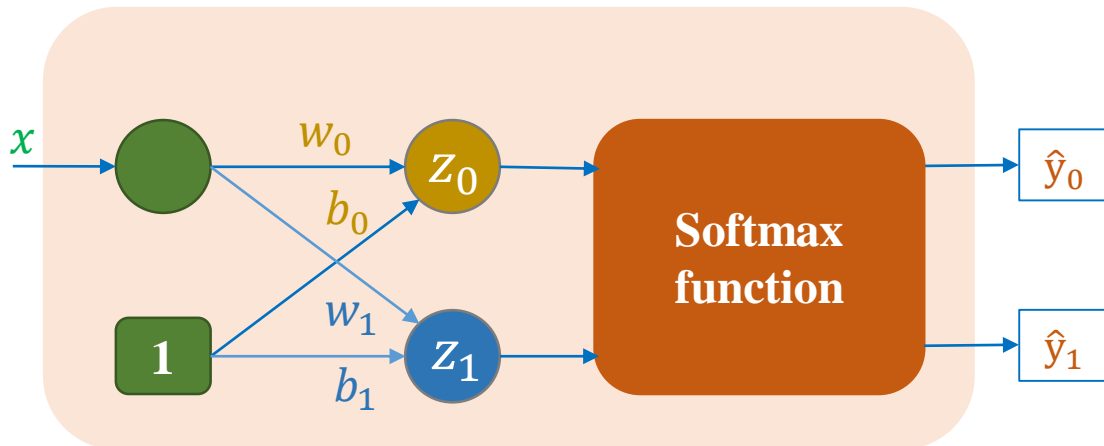
$$\hat{y}_1 = \frac{e^{z_1}}{\sum_{j=0}^1 e^{z_j}}$$

$$\mathbf{z} = \begin{bmatrix} z_0 \\ z_1 \end{bmatrix} = \begin{bmatrix} b_0 & w_0 \\ b_1 & w_1 \end{bmatrix} \begin{bmatrix} 1 \\ x \end{bmatrix} = \begin{bmatrix} \theta_0^T \\ \theta_1^T \end{bmatrix} \begin{bmatrix} 1 \\ x \end{bmatrix} = \boldsymbol{\theta}^T \mathbf{x}$$

$$\hat{\mathbf{y}} = \begin{bmatrix} \hat{y}_0 \\ \hat{y}_1 \end{bmatrix} = \frac{1}{\sum_{j=0}^1 e^{z_j}} \begin{bmatrix} e^{z_0} \\ e^{z_1} \end{bmatrix} = \frac{e^{\mathbf{z}}}{\sum_{j=0}^1 e^{z_j}}$$

$$L(\boldsymbol{\theta}) = -y_0 \log \hat{y}_0 - y_1 \log \hat{y}_1 = - \sum_{i=0}^1 y_i \log \hat{y}_i$$

Model



Derivative

$$\frac{\partial L}{\partial \hat{y}_i} = \frac{y_i}{\hat{y}_i}$$

$$\frac{\partial \hat{y}_i}{\partial z_j} = \begin{cases} \hat{y}_i(1 - \hat{y}_i) & \text{if } i = j \\ -\hat{y}_i \hat{y}_j & \text{if } i \neq j \end{cases}$$

$$\frac{\partial L}{\partial z_i} = \hat{y}_i - y_i$$

$$\frac{\partial L}{\partial w_i} = x(\hat{y}_i - y_i)$$

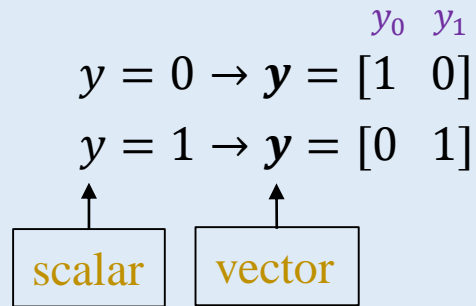
$$\frac{\partial L}{\partial b_i} = \hat{y}_i - y_i$$

Summary

Feature Label

Petal_Length	Category
1.4	0
1	0
1.5	0
3	1
3.8	1
4.1	1

One-hot encoding for label



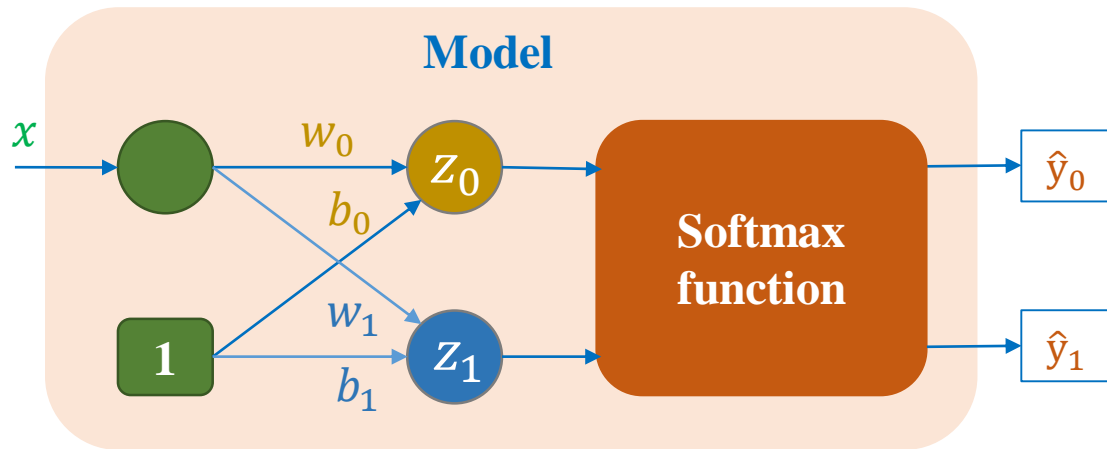
Forward computation

$$\mathbf{z} = \boldsymbol{\theta}^T \mathbf{x}$$

$$\hat{y} = \frac{e^z}{\sum_{j=0}^1 e^{z_j}}$$

Loss function

$$L(\boldsymbol{\theta}) = - \sum_{i=0}^1 y_i \log \hat{y}_i$$



$$\boldsymbol{\theta} = \begin{bmatrix} b_0 & b_1 \\ w_0 & w_1 \end{bmatrix}$$

$$\mathbf{x} = \begin{bmatrix} 1 & x \end{bmatrix}$$

Derivative

$$\frac{\partial L}{\partial \hat{y}_i} = \frac{y_i}{\hat{y}_i}$$

$$\frac{\partial \hat{y}_i}{\partial z_j} = \begin{cases} \hat{y}_i(1 - \hat{y}_i) & \text{if } i = j \\ -\hat{y}_i \hat{y}_j & \text{if } i \neq j \end{cases}$$

$$\frac{\partial L}{\partial z_i} = \hat{y}_i - y_i$$

$$\frac{\partial L}{\partial w_i} = x(\hat{y}_i - y_i)$$

$$\frac{\partial L}{\partial b_i} = \hat{y}_i - y_i$$

References

Softmax Regression

<http://deeplearning.stanford.edu/tutorial/supervised/SoftmaxRegression/>

