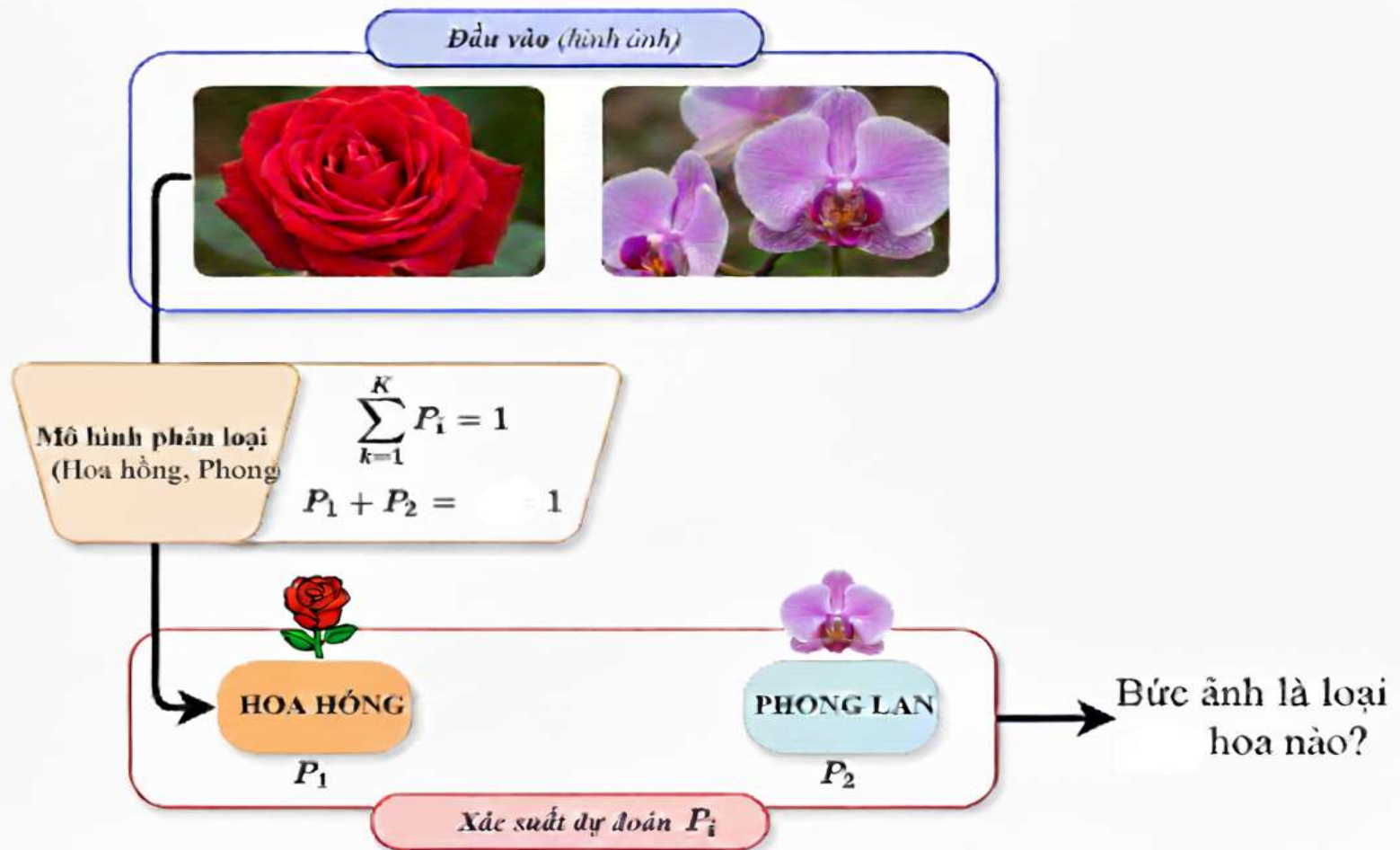
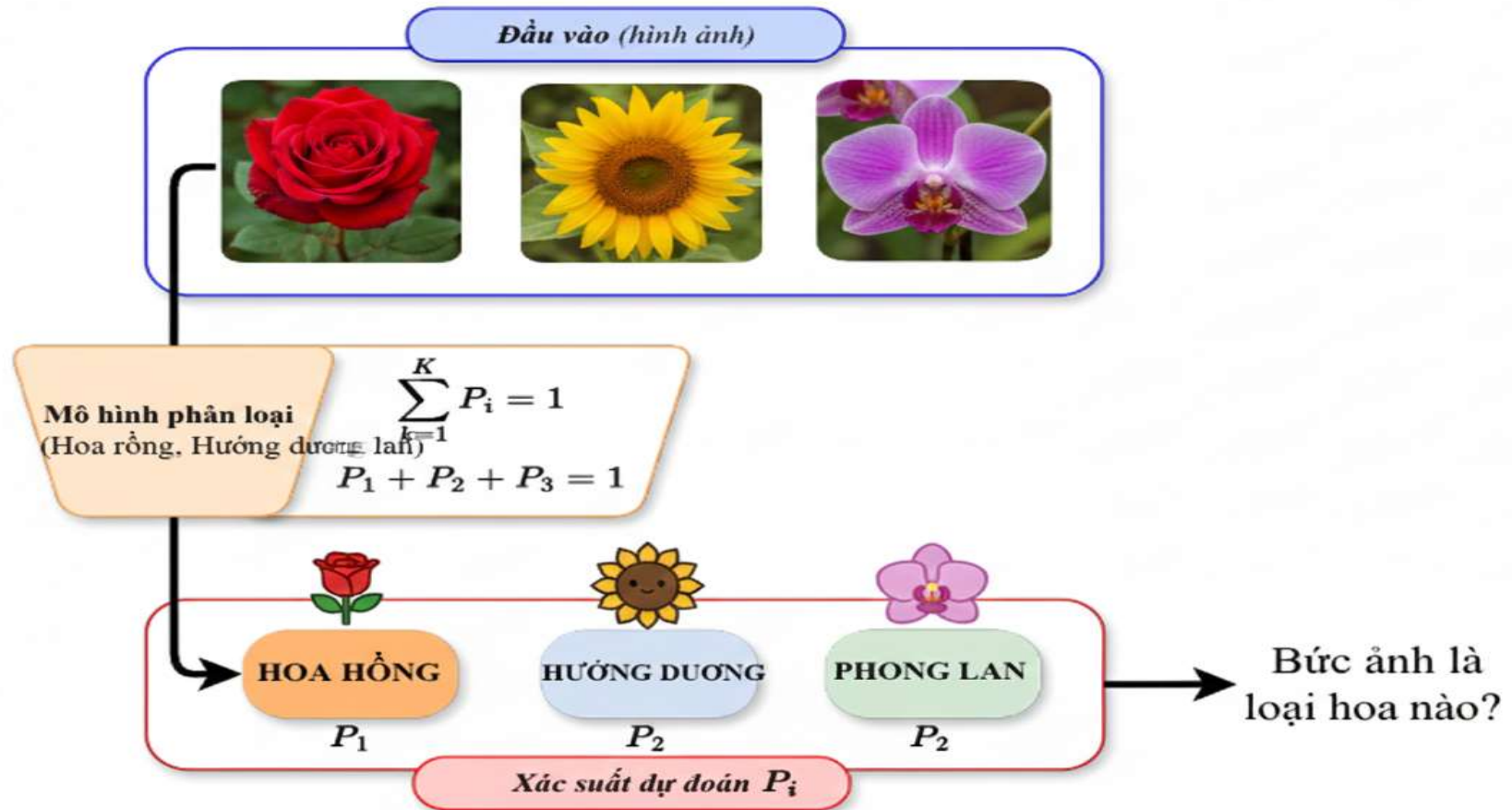


# SOFTMAX REGRESSION

## Giới thiệu



## Giới thiệu



Một số khái niệm

- Phân loại nhị phân:
  - Một bài toán học máy nhằm dự đoán đầu ra thuộc về một trong hai nhóm, thường được gán nhãn là 0 hoặc 1.
- Phân loại đa lớp (Multi-class classification):
  - Một bài toán học máy nhằm dự đoán đầu ra thuộc về một trong  $k$  ( $k > 2$ ) nhóm.
  - Ví dụ:
    - Phân loại chữ số viết tay (10 lớp)
    - Phân loại màu da (4 lớp)

Một số khái niệm

- Softmax regression:
  - Mô hình tổng quát của Logistic regression
  - Dùng cho bài toán phân loại đa lớp
  - Mô hình này dự đoán một phân phối xác suất trên  $k$  lớp
- Hàm Softmax:
  - Một hàm dùng để nén một vector  $k$  chiều chứa các điểm số bất kỳ thành một vector  $k$  chiều chứa các xác suất. Tổng các xác suất này bằng 1.

Một số khái niệm

- One-hot encoding:
  - Một phương pháp biểu diễn nhãn y dạng category thành một vector dạng 0,1.
  - Ví dụ:
    - Bài toán có 3 nhãn là “chó, mèo, chuột”
    - Tạo vector có độ dài bằng 3
    - Chó: [1,0,0]
    - Mèo: [0,1,0]
    - Chuột: [0,0,1]

Một số khái niệm

- Categorical Cross Entropy:
  - Hàm mất mát dùng cho Softmax regression
  - Đo lường sự khác biệt giữa phân phối xác suất dự đoán ( $\hat{y}_i$ ) và nhãn thực tế dạng one-hot ( $y_i$ )

## Hàm Softmax

- Được thiết kế để đảm bảo:
  - Tính K điểm số tuyến tính riêng biệt (một điểm  $z_k$  cho mỗi lớp k).
  - Biến đổi K điểm số này thành K xác suất riêng biệt ( $P(y = k|x)$ ).
  - Đảm bảo K xác suất này có tổng bằng 1
    - (ví dụ:  $P(\text{Hoa Hồng}) + P(\text{Hoa Lan}) + P(\text{Hoa Phượng}) = 1$ ).



## Hàm Softmax

- Quá trình tính Softmax bao gồm 2 bước:
  - Tính toán K điểm số tuyến tính
  - Chuyển đổi K điểm số tuyến tính thành K xác suất:

## Hàm Softmax

- Quá trình tính Softmax bao gồm 2 bước:
  - Tính toán K điểm số tuyến tính
    - $z_i = Wx_i + b$
    - $W$  là  $[K \times D]$  với K là số lớp đầu ra, D là số features
    - $x_i$  là  $[D \times 1]$ , b là  $[K \times 1]$

## Hàm Softmax

- Quá trình tính Softmax bao gồm 2 bước:

- Tính toán K điểm số tuyến tính

- $z_{i,chó} = w_{chó}x_i + b_{chó}$

- $z_{i,mèo} = w_{mèo}x_i + b_{mèo}$

- $z_{i,chuột} = w_{chuột}x_i + b_{chuột}$

Như vậy  $z_i = \begin{bmatrix} z_{i,chó} \\ z_{i,mèo} \\ z_{i,chuột} \end{bmatrix}$

## Hàm Softmax

- Quá trình tính Softmax bao gồm 2 bước:

- Tính toán K điểm số tuyến tính

$$z_i = Wx_i + b$$

- Chuyển đổi K điểm số tuyến tính thành K xác suất

- $\hat{y}_{ij} = \frac{e^{z_{ij}}}{\sum_{k=1}^K e^{z_{ik}}}$

- Kết quả: vector xác suất dự đoán  $\hat{y}_i = [\hat{y}_{ij}]$  và tổng của chúng là 1.0

## Hàm Softmax

- Quá trình tính Softmax bao gồm 2 bước:

- Tính toán K điểm số tuyến tính

$$\text{Như vậy } z_i = \begin{bmatrix} z_{i,chó} \\ z_{i,mèo} \\ z_{i,chuột} \end{bmatrix} = \begin{bmatrix} 2.0 \\ 1.0 \\ -1.0 \end{bmatrix}$$

- Chuyển đổi K điểm số tuyến tính thành K xác suất

$$\hat{y}_{i,chó} = P(y_i = chó|x_i) = \frac{e^{2.0}}{e^{2.0} + e^{1.0} + e^{-1.0}}$$

$$\hat{y}_{i,mèo} = P(y_i = mèo|x_i) = \frac{e^{1.0}}{e^{2.0} + e^{1.0} + e^{-1.0}}$$

$$\hat{y}_{i,chuột} = P(y_i = chuột|x_i) = \frac{e^{-1.0}}{e^{2.0} + e^{1.0} + e^{-1.0}}$$

- Kết quả: vector xác suất dự đoán  $\hat{y}_i = [0.705, 0.260, 0.035]$  và tổng của chúng là 1.0

Hàm mất mát

- Categorical Cross-Entropy:
  - Tính cho từng sample:
    - $L_i = - \sum_{j=1}^K y_{ij} \log(\hat{y}_{ij})$
  - Tổng quát
    - $L = \frac{1}{N} \sum_{i=1}^N L_i$

## Đạo hàm

- $z_i = Wx_i + b$
- $\hat{y}_{ij} = \frac{e^{z_{ij}}}{\sum_{k=1}^K e^{z_{ik}}}$
- $L_i = -\sum_{j=1}^K y_{ij} \log(\hat{y}_{ij})$
- Mục tiêu 1:
  - $\frac{\partial L_i}{\partial w}$
  - $\frac{\partial L}{\partial w}$
- Mục tiêu 2:
  - $\frac{\partial L_i}{\partial b}$
  - $\frac{\partial L}{\partial b}$

Đạo hàm

- $z_i = Wx_i + b$
- $\hat{y}_{ij} = \frac{e^{z_{ij}}}{\sum_{k=1}^K e^{z_{ik}}}$
- $L_i = -\sum_{j=1}^K y_{ij} \log(\hat{y}_{ij})$
- Mục tiêu 1:
  - $\frac{\partial L_i}{\partial w} = (\hat{y}_i - y_i) x_i^T$
  - $\frac{\partial L}{\partial w} = \frac{1}{N} \sum_{i=1}^N (\hat{y}_i - y_i) x_i^T$
- Mục tiêu 2:
  - $\frac{\partial L_i}{\partial b} = (\hat{y}_i - y_i)$
  - $\frac{\partial L}{\partial b} = \frac{1}{N} \sum_{i=1}^N (\hat{y}_i - y_i)$



## Quy trình thực hiện

- Bước 1: Khởi tạo  $W$  ( $[K \times D]$ ),  $b$  ( $[K \times 1]$ ),  $LR$ , ...
- Bước 2: Duyệt qua từng epoch:
  - Duyệt qua từng dòng dữ liệu:
    - Tính giá trị dự đoán:
      - Tính đầu vào tuyến tính:  $z_{ij} = w_j x_i + b_j \Rightarrow z_i = [z_{ij}]$
      - Tính xác suất dự đoán:  $\hat{y}_{ij} = \frac{e^{z_{ij}}}{\sum_{k=1}^K e^{z_{ik}}} \Rightarrow \hat{y}_i = [\hat{y}_{ij}]$
    - Tính hàm mất mát:  $L_i = - \sum_{j=1}^K y_{ij} \log(\hat{y}_{ij})$
    - Tính đạo hàm:
      - $\frac{\partial L_i}{\partial w} = (\hat{y}_i - y_i) x_i^T$
      - $\frac{\partial L_i}{\partial b} = (\hat{y}_i - y_i)$
    - Cập nhật  $W$ ,  $b$

$$W = W - \eta \frac{\partial L_i}{\partial W}$$

$$b = b - \eta \frac{\partial L_i}{\partial b}$$

- Bước 3: Sau khi hoàn thành, chúng ta thu được ma trận  $W$  và vector  $b$  đã được tối ưu.

## Quy trình thực hiện

- Bước 1: Khởi tạo  $W$  ( $[K \times D]$ ),  $b$  ( $[K \times 1]$ ),  $LR$ , ...
- Bước 2: Duyệt qua từng epoch:
  - Tính giá trị dự đoán cho  $N$  sample:
    - Tính đầu vào tuyến tính:  $z_i = W \times x_i + b$
    - Tính xác suất dự đoán:  $\hat{y}_i = \frac{e^{z_k}}{\sum_{j=1}^K e^{z_j}}$
  - Tính hàm mất mát:  $L_i = -\sum_{j=1}^K y_{ij} \log(\hat{y}_{ij})$
  - Tính đạo hàm:  $\frac{\partial L}{\partial w} = \frac{1}{N} \sum_{i=1}^N (\hat{y}_i - y_i) x_i^T$  và  $\frac{\partial L}{\partial b} = \frac{1}{N} \sum_{i=1}^N (\hat{y}_i - y_i)$
  - Cập nhật  $W$ ,  $b$

$$W = W - \eta \frac{\partial L}{\partial w}$$

$$b = b - \eta \frac{\partial L}{\partial b}$$

- Bước 3: Sau khi hoàn thành, chúng ta thu được ma trận  $W$  và vector  $b$  đã được tối ưu.

## THỰC HÀNH

- Xét bài toán ví dụ: dự đoán khách hàng mua (1) hay không mua (0) dựa trên thu nhập của họ. Chúng ta sẽ lấy ra 6 điểm dữ liệu ( $N = 6$ ) để thực hành tính toán:
- Mục tiêu:
  - 1. Thực hiện cập nhật  $W$  và  $b$  theo one-sample
  - 2. Thực hiện cập nhật  $W$  và  $b$  theo full-sample
- Gợi ý:  $W = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$ ,  $b = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$ ,  $\eta=0.2$

Chiều rộng cánh hoa	Giống hoa
1.0	0 (hoa A)
2.5	0
4.0	1 (hoa B)
5.5	1
7.0	2 (hoa C)
8.0	2

## THỰC HÀNH

Bước 1: Khởi tạo:  $W = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$ ,  $b = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$ ,

$\eta=0.2$

Bước 2: Duyệt từng epoch:

Duyệt từng sample:

2.1. Tính điểm tuyến tính

$$z_{1,0} = (0 \cdot x_i) + 0 = 0.$$

$$z_{1,1} = (0 \cdot x_i) + 0 = 0.$$

$$z_{1,2} = (0 \cdot x_i) + 0 = 0.$$

$$\Rightarrow z_1 = [0, 0, 0].$$

Chiều rộng cánh hoa	Giống hoa
1.0	0 (hoa A)
2.5	0
4.0	1 (hoa B)
5.5	1
7.0	2 (hoa C)
8.0	2

## THỰC HÀNH

Bước 1: Khởi tạo:  $W = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$ ,  $b = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$ ,  $\eta=0.2$

Bước 2: Duyệt từng epoch:

Duyệt từng sample:

2.1. Tính xác suất dự đoán:

$$\hat{y}_{ij} = \frac{e^{z_{ij}}}{\sum_{k=1}^K e^{z_{ik}}}$$
$$\hat{y}_{10} = \frac{e^0}{e^0 + e_0^0 + e^0} = \frac{1}{3} = 0.333$$
$$\hat{y}_{11} = \frac{e_0^0}{e^0 + e_0^0 + e^0} = \frac{1}{3} = 0.333$$
$$\hat{y}_{12} = \frac{e^0}{e^0 + e^0 + e^0} = \frac{1}{3} = 0.333$$

$\Rightarrow \hat{y}_i = [0.333, 0.333, 0.333]$ .

Chiều rộng cánh hoa	Giống hoa
1.0	0 (hoa A)
2.5	0
4.0	1 (hoa B)
5.5	1
7.0	2 (hoa C)
8.0	2

## THỰC HÀNH

Bước 1: Khởi tạo:  $W = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$ ,  $b = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$ ,  $\eta=0.2$

Bước 2: Duyệt từng epoch:

Duyệt từng sample:

2.2. Tính hàm mất mát:

$$L_i = - \sum_{j=1}^K y_{ij} \log(\hat{y}_{ij})$$

$$L_1 = -[y_{10} \log(\hat{y}_{10}) + y_{11} \log(\hat{y}_{11}) + y_{12} \log(\hat{y}_{12})] = -\log(1/3) = \log(3) = 1.0986$$

Chiều rộng cánh hoa	Giống hoa
1.0	0 (hoa A)
2.5	0
4.0	1 (hoa B)
5.5	1
7.0	2 (hoa C)
8.0	2

## THỰC HÀNH

Bước 1: Khởi tạo:  $W = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$ ,  $b = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$ ,  $\eta=0.2$

Bước 2: Duyệt từng epoch:

Duyệt từng sample:

2.3. Tính đạo hàm:

- $\frac{\partial L_i}{\partial w} = (\hat{y}_i - y_i) x_i^T$
- $\frac{\partial L_i}{\partial b} = (\hat{y}_i - y_i)$

$$\frac{\partial L_1}{\partial w} = ([0.333, 0.333, 0.333] - [1, 0, 0]) * 1.0 = [-0.667, 0.333, 0.333]$$

$$\frac{\partial L_1}{\partial b} = ([0.333, 0.333, 0.333] - [1, 0, 0]) = [-0.667, 0.333, 0.333]$$

Chiều rộng cánh hoa	Giống hoa
1.0	0 (hoa A)
2.5	0
4.0	1 (hoa B)
5.5	1
7.0	2 (hoa C)
8.0	2

## THỰC HÀNH

Bước 1: Khởi tạo:  $W = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$ ,  $b = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$ ,  $\eta=0.2$

Bước 2: Duyệt từng epoch:

Duyệt từng sample:

2.4. Cập nhật W và b:

$$\begin{aligned} \bullet W &= \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} - 0.2 \times [-0.667, 0.333, 0.333]^T = \begin{bmatrix} 0,1334 \\ -0,0666 \\ -0,0666 \end{bmatrix} \\ \bullet b &= \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} - 0.2 \times [-0.667, 0.333, 0.333]^T = \begin{bmatrix} 0,1334 \\ -0,0666 \\ -0,0666 \end{bmatrix} \end{aligned}$$

Chiều rộng	Giống hoa
1.0	0 (hoa A)
2.5	0
4.0	1 (hoa B)
5.5	1
7.0	2 (hoa C)
8.0	2