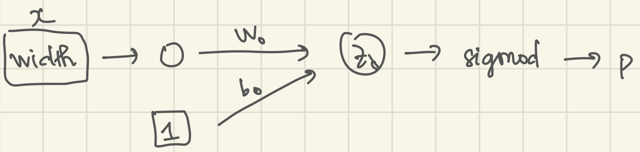
11/10

SOFTMAX REGRESSION

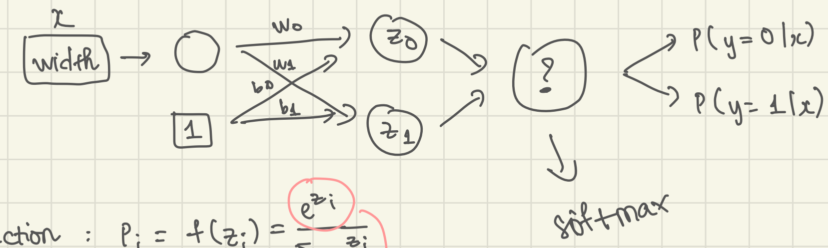
I. MOTIVATION

width	label
1.4	0
1.5	0
4.5	1
5.6	1



1 & 0 riêng thì ?

→ Tại sao phân loại 2 cái mà chỉ có 1 node ? Nếu cho 2 node đại diện



- softmax function : $P_i = f(z_i) = \frac{e^{z_i}}{\sum_j e^{z_j}}$

Tại sao không dùng xs bt $\frac{z_i}{\sum z_i}$

để tránh tuyến tính ⇄ k nhạy cảm

ổn định hơn với những số cực nhỏ hoặc cực lớn

+ Note: khi cài đặt softmax trên máy dễ bị tràn → trừ đi max để scale lại

- Loss function : $-y\log\hat{y}_1 - (1-y)\log\hat{y}_0$

II. MODEL CONSTRUCTION

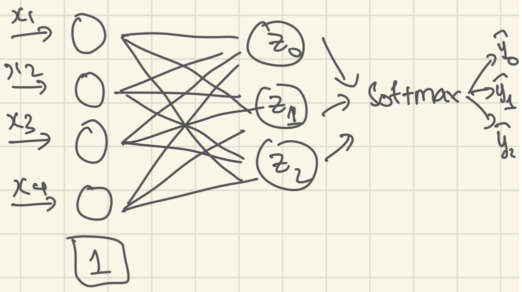
Feature: width	Label: label
1.4	0
1.5	0
4.5	1
5.6	1

class : 2 (0 & 1) → số lượng output node

feature : 1 → số lượng input node

s_length	s_width	P_length	P_width	label
–	—	—		0
—	—			1
–		—		0
—				1
—			—	2
—			—	2

4 features
3 classes



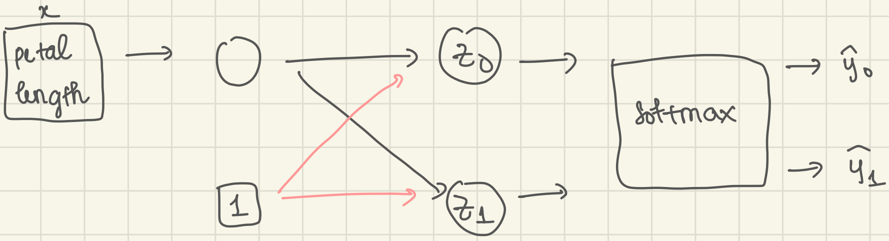
3x5 = 15 parameters

VD:

folder
- f_1 : cat
- f_2 : dog
- f_3 : duck
- f_4 : tiger
- f_5 : goat

image : 10x10
- feature : 100
- class : 5

III. LOSS FUNCTION

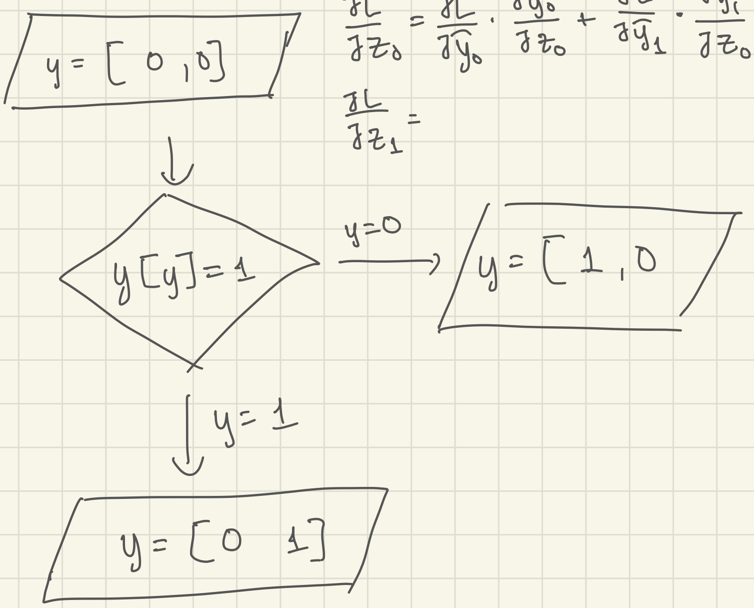


$$\left.\begin{aligned} z_0 &= x w_0 + b_0 \\ z_1 &= x w_1 + b_1 \end{aligned}\right\} z = \theta^T x \left(z = \begin{bmatrix} b_0 & w_0 \\ b_0 & w_1 \end{bmatrix} \begin{bmatrix} 1 \\ x \end{bmatrix} = \begin{bmatrix} \theta_0^T \\ \theta_1^T \end{bmatrix} \begin{bmatrix} 1 \\ x \end{bmatrix} = \theta^T x \right)$$

$$\left.\begin{aligned} y_0 &= \frac{e^{z_0}}{\sum_j e^{z_j}} \\ y_1 &= \frac{e^{z_1}}{\sum_j e^{z_j}} \end{aligned}\right\} y = \frac{e^{z}}{\sum_j e^{z_j}} \qquad L(\theta) = -y_0 \log \hat{y}_0 - y_1 \log \hat{y}_1$$

- One-hot encoding :

code :



$$\frac{\partial L}{\partial z_0} = \frac{\partial L}{\partial \hat{y}_0} \cdot \frac{\partial y_0}{\partial z_0} + \frac{\partial L}{\partial \hat{y}_1} \cdot \frac{\partial \hat{y}_1}{\partial z_0}$$

$$\frac{\partial L}{\partial z_1} =$$

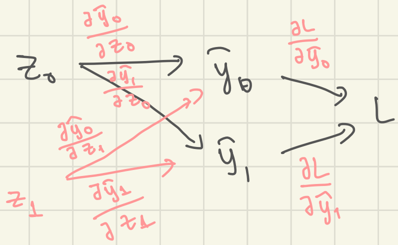
$$L(\theta) = -y \log \hat{y}_0 - y_1 \log \hat{y}_1 = \sum_{i=0}^{1} y_i \log \hat{y}_i = -y^T \log \hat{y}$$

$$y_0 = \frac{e^{z_0}}{\sum_{j=0}^{1} e^{z_j}} \qquad y_1 = \frac{e^{z_1}}{\sum_{j=0}^{1} e^{z_j}}$$

$$\left(\frac{u}{v}\right)' = \frac{u'v - v'u}{v^2}$$

$$\left(\frac{1}{u}\right)' = -\frac{1}{u^2} \cdot u'$$

$$(e^u)' = u' \cdot e^u$$



$$\frac{\partial L}{\partial \hat{y}_0} = \frac{-y_0}{\hat{y}_0} \qquad \frac{\partial L}{\partial \hat{y}_1} = \frac{-y_1}{\hat{y}_1}$$

$$\frac{\partial \hat{y}_0}{\partial z_1} = \frac{-e^{z_0}}{(e^{z_0}+e^{z_1})^2} \cdot e^{z_1} = \frac{-e^{z_0}}{(e^{z_0}+e^{z_1})} \cdot \frac{e^{z_1}}{(e^{z_0}+e^{z_1})}$$

$$= -\hat{y}_0 \cdot \hat{y}_1$$

$$\frac{\partial \hat{y}_0}{\partial z_0} = \left(\frac{e^{z_0}}{e^{z_0}+e^{z_1}}\right)' = \frac{e^{z_0} \cdot \Sigma - e^{z_0} \cdot e^{z_0}}{(e^{z_0}+e^{z_1})^2} = \frac{e^{z_0}(\Sigma - e^{z_0})}{(\Sigma^2)} = \frac{e^{z_0}}{\Sigma} \cdot \left(\frac{\Sigma}{\Sigma} - \frac{e^{z_0}}{\Sigma}\right)$$

$$= \hat{y}_0 \cdot (1 - \hat{y}_0)$$

$$\frac{\partial y_1}{\partial z_1} = \hat{y}_1 (1 - \hat{y}_1)$$

$$\frac{\partial y_1}{\partial z_0} = -\hat{y}_0 \cdot \hat{y}_1$$

Derivative

$$\frac{\partial \hat{y}_i}{\partial z_j} = \begin{cases} y_i(1-y_i) & i=j \\ -\hat{y}_i\, y_j & i \neq j \end{cases}$$

$$\frac{\partial L}{\partial z_0} = \frac{\partial L}{\partial \hat{y}_0} \cdot \frac{\partial \hat{y}_0}{\partial z_0} + \frac{\partial L}{\partial \hat{y}_1} \cdot \frac{\partial \hat{y}_1}{\partial z_0} = -\frac{y_0}{\cancel{\hat{y}_0}} \cdot \cancel{\hat{y}_0}(1-\hat{y}_0) + \frac{y_1}{\cancel{\hat{y}_1}} \cdot (\hat{y}_0 \cdot \cancel{\hat{y}_1})$$

$$= -y_0(1-\hat{y}_0) + y_1 \cdot \hat{y}_0 = -y_0 + y\hat{y}_0 + y_1\hat{y}_0 = \hat{y}_0(y_0+y_1) - y_0$$

$$= \hat{y}_0 - y_0$$ (do one-hot thì $y_1 = 0$)

$$\frac{\partial L}{\partial z_1} = \hat{y}_1 - y_1$$

$$\frac{\partial L}{\partial w_0} = x(\hat{y}_0 - y_0)$$

IV. SUMMARY

1. Forward computation

$$z = \theta^T x$$

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

2. loss function

$$L(\theta) = -y^T \log y$$

3. Derivative

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

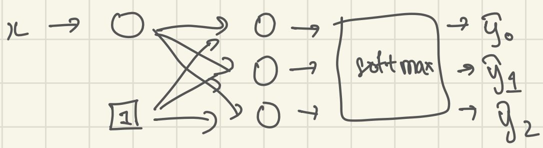
$$\nabla_\theta L = x(\hat{y} - y)^T$$

4. Update

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

V. A SUGGEST FUNCTION

data → # feature : 1
→ # classes : 3



$$L(\theta) = \underbrace{-\frac{y(1-y)}{-2}\log(y_2)}_{y_2} \underbrace{- y(2-y)\log(\hat{y}_1)}_{y_1} \underbrace{-(1-y)\left(\frac{2-y}{2}\right)\log(\hat{y}_0)}_{y_0}$$

→ Khá phức tạp nếu nhiều label → one-hot encoding

$$L(\theta) = -y_2 \log(\hat{y}_2) - y_1 \log(\hat{y}_1) - y_0 \log(\hat{y}_0)$$

VI. MODEL