07/02/2022

Softmax function

* logistics Regression using Softmax function.
* Multinomial Logistics Regression / Softmax Regression / Maximum Entropy classifier are the names popular.

→ logistics Regression for binary classification.
- used Sigmoid activation function.
- output will be belongs to two classes. (labeled o/p's into two category).

? if we have more than 2 Classes then How to use logistics Regression.

eg

	x_1	x_2	Y
	CGPA	EQ	Placement
Bala	8.4	75	1
Manoj	8.7	60	1
Umesh	8.0	85	2
Aditya	9.0	95	3

label(class)
1) yes
2) NO
3) optout

Multi class $\longrightarrow$ binary class

✡ Softmax function

$$\boxed{\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}}$$

K : no of classes.

$\sigma(z)_i$: Probability of class i

<u>class</u>

1) yes $\Rightarrow$ $\left\{ \sigma(z)_1 = \dfrac{e^{z_1}}{e^{z_1} + e^{z_2} + e^{z_3}} \right\}$

2) NO $\Rightarrow$ $\left\{ \sigma(z)_2 = \dfrac{e^{z_2}}{e^{z_1} + e^{z_2} + e^{z_3}} \right\}$

3) option $\Rightarrow$ $\left\{ \sigma(z)_3 = \dfrac{e^{z_3}}{e^{z_1} + e^{z_2} + e^{z_3}} \right\}$

$$\boxed{\sigma(z)_1 + \sigma(z)_2 + \sigma(z)_3 = 1}$$

$$0 < \sigma(z)_1, \sigma(z)_2, \sigma(z)_3 < 1$$

✡ Intuition

- Training Intuition

CGPA	EQ	Placement
7.1	71	1
8.5	85	2
9.5	95	3

⇓ multiclass to binary classes

one hot encoding

CGPA	EQ	P_1	P_2	P_3
7.1	71	1	0	0
8.5	85	0	1	0
9.5	95	0	0	1

Dataset

D_1 $y^{(1)}$

CG	EQ	P_1
7.1	71	1
8.5	85	0
9.5	95	0

m_1

LR → Train

$\theta_0^{(1)}, \theta_1^{(1)}, \theta_2^{(1)}$

loss function & GD/GA

D_2 $y^{(2)}$

CG	EQ	P_2
7.1	71	0
8.5	85	1
9.5	95	0

m_2

$\theta_0^{(2)}, \theta_1^{(2)}, \theta_2^{(2)}$

Loss fn
GD/GA

D_3 $y^{(3)}$

CG	EQ	P_3
7.1	71	0
8.5	85	0
9.5	95	1

m_3

$\theta_0^{(3)}, \theta_1^{(3)}, \theta_2^{(3)}$

Loss fn
GD/GA

2) Prediction intuition

test data ⇒ CGPA = 7
EQ = 70

Now it will be placed or not?

classes → {yes, No, optout}

$$z_1^{(1)} = \theta_0^{(1)} + \theta_1^{(1)} x_1 + \theta_2^{(1)} x_2$$

$$= \theta_0^{(1)} + \theta_1^{(1)} \times 7 + \theta_2^{(1)} \times 70$$

$$= \underline{\qquad\qquad\qquad\qquad} \checkmark$$

$$z_2^{(2)} = \theta_0^{(2)} + \theta_1^{(2)} \times 7 + \theta_2^{(2)} \times 70$$

$$= \underline{\qquad\qquad\qquad\qquad}$$

$$z_3^{(3)} = \theta_0^{(3)} + \theta_1^{(3)} \times 7 + \theta_2^{(3)} \times 70$$

$$= \underline{\qquad\qquad\qquad\qquad}$$

Probability for 'yes' class.

$$\sigma(z)_1 = \frac{e^{z_1^{(1)}}}{e^{z_1^{(1)}} + e^{z_2^{(2)}} + e^{z_3^{(3)}}} = \underline{0.40}$$

$$\sigma(z)_2 = \frac{e^{z_2}}{e^{z_1}+e^{z_2}+e^{z_3}} = \underline{0.35}$$

$$\sigma(z)_3 = \frac{e^{z_3}}{e^{z_1}+e^{z_2}+e^{z_3}} = \underline{0.25}$$

Probability of "yes" is higher in compare to $\sigma(z)_2$ and $\sigma(z)_3$

$\Rightarrow$ This student who have CGPA=7, and EQ= 70, he will belongs to 'yes' class.

<u>Conclusion -</u>

Real implementation

logistic Regression (in binary classification)

Loss function (sigmoid)

$$L = -\left[\sum_{i=1}^{m} y^{i} \log h_\theta(x)^{i} + (1-y^{i}) \log(1-h(x^{i}))\right]$$

$$L = -\left[\sum_{i=1}^{m} Y_i \log \hat{Y}_i + (1-Y_i) \log(1-\hat{Y}_i)\right]$$

Softmax function (loss function)

$$= -\sum_{i=1}^{m} \sum_{k=1}^{K} Y_k^{(i)} \log(\hat{Y}_k^{(i)})$$

no. of rows

no. of classes

	CGPA	EQ	P	P_1 ($Y_{K=1}$)	P_2 ($Y_{K=2}$)	P_3 ($Y_{K=3}$)
i=1	7.1	71	1	1	0	0
i=2	8.5	85	2	0	1	0
i=3	9.1	91	3	0	0	1

$$= \underbrace{Y_1^{(1)} \log(\hat{Y}_1^{(1)}) + Y_2^{(1)} \log \hat{Y}_2^{(1)} + Y_3^{(1)} \log \hat{Y}_3^{(1)}}_{i=1,\ K=1,2,3} +$$

$$\underbrace{Y_1^{(2)} \log \hat{Y}_1^{(2)} + Y_2^{(2)} \log \hat{Y}_2^{(2)} + Y_3^{(2)} \log \hat{Y}_3^{(2)}}_{i=2,\ K=1,2,3} +$$

$$\underbrace{Y_1^{(3)} \log \hat{Y}_1^{(3)} + Y_2^{(3)} \log \hat{Y}_2^{(3)} + Y_3^{(3)} \log \hat{Y}_3^{(3)}}_{i=3,\ K=1,2,3}$$

$$\boxed{L = y_1^{(1)} \log \hat{y}_1^{(1)} + y_2^{(2)} \log \hat{y}_2^{(2)} + y_3^{(3)} \log \hat{y}_3^{(3)}}$$

$\underbrace{y_1^{(1)}}_{1}, \underbrace{y_2^{(2)}}_{1}, \underbrace{y_3^{(3)}}_{1}$: it is available known as dataset

but we don't know

$\hat{y}_1^{(1)}, \hat{y}_2^{(2)}, \hat{y}_3^{(3)} = ?$

$$\hat{y}_1^{(1)} = \sigma(\theta_0^{(1)} + \theta_1^{(1)} x_1 + \theta_2^{(1)} x_2)$$

$$= \sigma(\theta_0^{(1)} + \theta_1^{(1)} \times 7.1 + \theta_2^{(1)} \times 71)$$

$$= \underline{\hspace{6cm}}$$

$$\hat{y}_2^{(2)} = \sigma(\theta_0^{(2)} + \theta_1^{(2)} \times 8.5 + \theta_2^{(2)} \times 8.5)$$

$$\hat{y}_3^{(3)} = \sigma\left(\theta_0^{(3)} + \theta_1^{(3)} \times 9.5 + \theta_2^{(3)} \times 95\right)$$

again

coefficient $= \begin{bmatrix} \theta_0^{(1)} & \theta_1^{(1)} & \theta_2^{(1)} \\ \theta_0^{(2)} & \theta_1^{(2)} & \theta_2^{(2)} \\ \theta_0^{(3)} & \theta_1^{(3)} & \theta_2^{(3)} \end{bmatrix}$

Gradient descent

$$\boxed{\theta_j^{(i)} = \theta_j^{(i)} + \alpha * \frac{\partial}{\partial \theta_j^{(i)}}(L)}$$