Softmax Regression / Multinomial Regression

→ for multiclass (more then 2) containing dataset.

CGPA	IQ	Placement
		Yes
		No
		Optout (Not sit in a placement)

No of class = 3

For above case simple logistic Regression is not wooks, make it wooks on this we have to applied some logic,

1. One not encoding.
2. Devide & conquer.

We transform above dateset into below →

CGPA	IQ	Yes	NO	Optout
5	80	0	1	0
10	70	1	0	0
9	80	0	0	1

This is nothing but one hot encoding / Nominal encoding.

For that We use Softmax function,

$$\sigma(t)_i = \frac{e^{t_i}}{\sum_{j=1}^{K} e^{t_j}}$$

let,

$Y_{es} \Rightarrow 1$

$No \Rightarrow 2$

$Optout \Rightarrow 3$

For yes,

$$\sigma(t)_1 = \frac{e^{t_1}}{e^{t_1} + e^{t_2} + e^{t_3}}$$

For No,

$$\sigma(t)_2 = \frac{e^{t_2}}{e^{t_1} + e^{t_2} + e^{t_3}}$$

For optout,

$$\sigma(t)_3 = \frac{e^{t_3}}{e^{t_1} + e^{t_2} + e^{t_3}}$$

Here, K is the NO of classes.

Yes + ×10 + optoutE

$$= \frac{e^{t_1}}{e^{t_1}+e^{t_2}+e^{t_3}} + \frac{e^{t_2}}{e^{t_1}+e^{t_2}+e^{t_3}} + \frac{e^{t_3}}{e^{t_1}+e^{t_2}+e^{t_3}}$$

$$= \frac{e^{t_1}+e^{t_2}+e^{t_3}}{e^{t_1}+e^{t_2}+e^{t_3}}$$

$$= 1$$

Note:-

$$\therefore \sigma(t)_1 + \sigma(t)_2 + \sigma(t)_3 + \cdots + \sigma(t)_k = 1$$

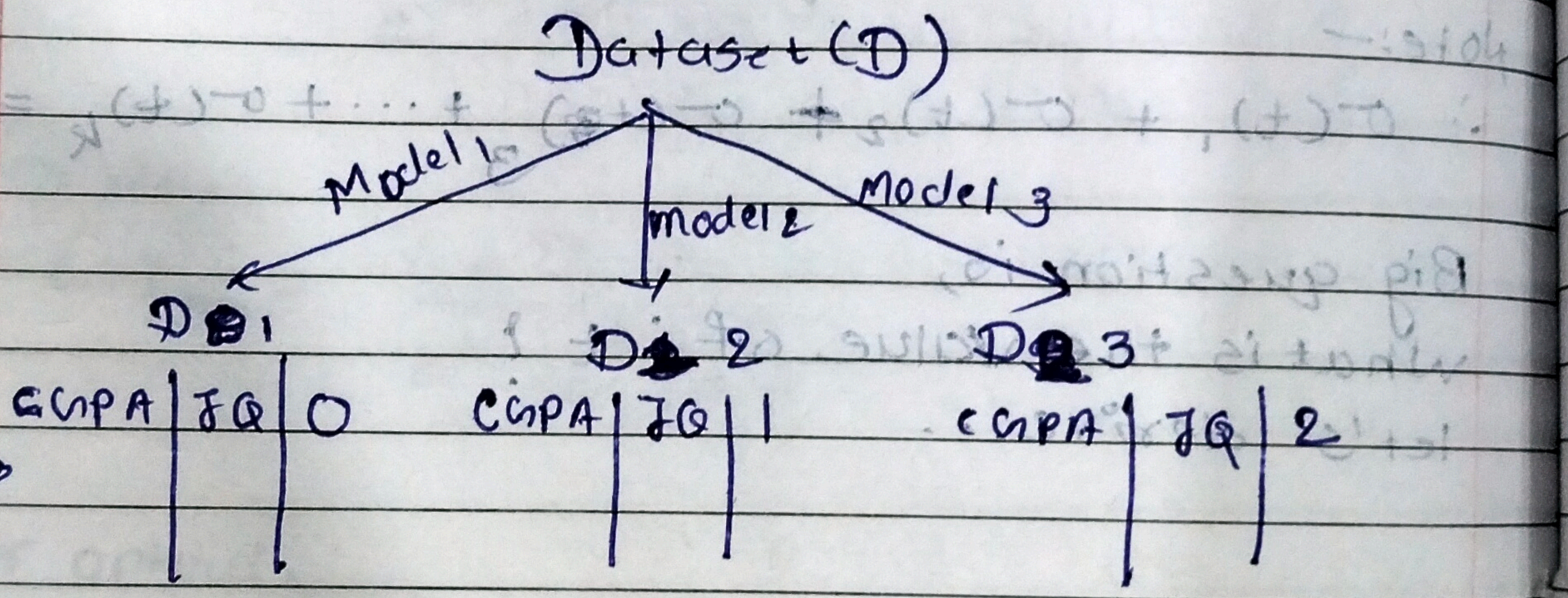
Big question is,
what is the value of t?
let's derive...

Ref,

CGPA	IQ	placement
8	59	0
9	75	1
8.8	80	2

So,

CGPA	IQ	placement_0 ~~Placement~~ 0	placement_1	placement_2
8	59	1	0	0
9	75	0	1	0
8.8	80	0	0	1



coef_1 : w_1^0, w_2^0, w_3^0 coef_2 : $w_1^1 w_2^1 w_3^1$ coef_3 : $w_1^2 w_2^2 w_3^2$

① $t_1 = w_1^0 \times 8 + w_2^0 \times 59 + w_3^0$ $t_2 = w_1^1 \times 8 + w_2^1 \times 59 + w_3^1$ $t_3 = w_1^2 \times 8 + w_2^2 \times 59 + w_3^2$

② $t_1 = w_1^0 \times 9 + w_2^0 \times 75 + w_3^0$ t_2 t_3

③ $t_1 = w_1^0 \times 8.8 + w_2^0 \times 80 + w_3^0$

→ In above we devide our dataset & considering one dataset at time we calculate weights & bias.

→ This is how we get the value of t_1, t_2 etc.

→ Now we can apply gradient descent to obtain $w_1^0, w_2^0, w_3^0, w_1^1, w_2^1, w_3^1, w_1^2, w_2^2, w_3^2$

→ evaluate them seperate we can also merge them.

We have loss function,

$$L = \frac{1}{n}\left(-\sum_{i=1}^{n} \cdot y_i (\log \hat{y}_i) - \sum_{i=1}^{n} (1-y_i)(\log(1-\hat{y}_i))\right)$$

For logistic regression, but we can modify it such a way that it can handle combined senario also,

$$L = \frac{-1}{n}\left(\sum_{i=1}^{n}\sum_{j=1}^{K} y_j^i \log(\hat{y}_j^i)\right)$$

Means,

For. i=0

1st Row, Prediction of model 1 +
Prediction of model 2 +
Prediction of model 3.

i=1

2nd Row, Prediction of model 1
Prediction of model 2
Prediction of model 3

K = No of class / model

⇒ Remember $\hat{y}$ is never been 0 so $\log(\hat{y})$ never been undefined because $\hat{y}$, the prediction is pass throw the sigmoid function then stor into $\hat{y}$.

→ In loss function of logistic Regression, if in two classes ~~prediction~~ (actual value of dependent variable) is 1 then $-\frac{1}{m} y_i \cdot (\log \hat{y}_i)$ is calculated & if actual value of dependent variable is 0 then $-\frac{1}{m}(1-y_i)(\log \hat{y}_i)$ is calculated.

→ But here as we conquer all the classes we only calculate $-\frac{1}{m} y_i \cdot (\log \hat{y}_i)$ as another class is in another model.

$$L = -\frac{1}{n}\left(\sum_{i=1}^{n}\sum_{j=1}^{K} y_j^i \left(\log\left(\hat{y}_j^i\right)\right)\right)$$

$n = 2$
$K = 3$

means, row = 2 & model = 3 / class = 3

x_1	x_2	y (0/1/2)
5	6	

↓ conversion,

x_1	x_2	0
5	6	1

m_1

x_1	x_2	1
5	6	0

m_2

x_1	x_2	2
5	6	0

m_3

$$L = -\frac{1}{1}\left(\sum_{i=1}^{1}\sum_{j=1}^{3} y_j^i \left(\log\left(\hat{y}_j^i\right)\right)\right)$$

$$= -1\left(y_1^1\left(\log \hat{y}_1^1\right) + y_2^1\left(\log \hat{y}_2^1\right) + y_3^1\left(\log \hat{y}_3^1\right)\right)$$

$$= -1\left(1\left(\log \hat{y}_1^1\right) + 0\left(\log \hat{y}_2^1\right) + 0\left(\log \hat{y}_3^1\right)\right)$$

$$\boxed{L = -\log \hat{y}_i}$$

Above is just an example for understanding how the loss function work!

⇒ Now for $w_1^0, w_2^0, w_3^0, w_1^1, w_2^1, w_3^1, \ldots$ we differentiate L with $w_1^0, w_2^0, w_3^0, w_1^1$

Because in gradient decent

$$\boxed{w = w - \eta \frac{dL}{dw}}$$

$$w = \begin{bmatrix} w_0^{(0)} & w_1^0 & w_2^0 \\ w_0^{(1)} & w_2^{(1)} & w_2^{(1)} \\ w_0^{(2)} & w_1^{(2)} & w_2^{(2)} \end{bmatrix}$$