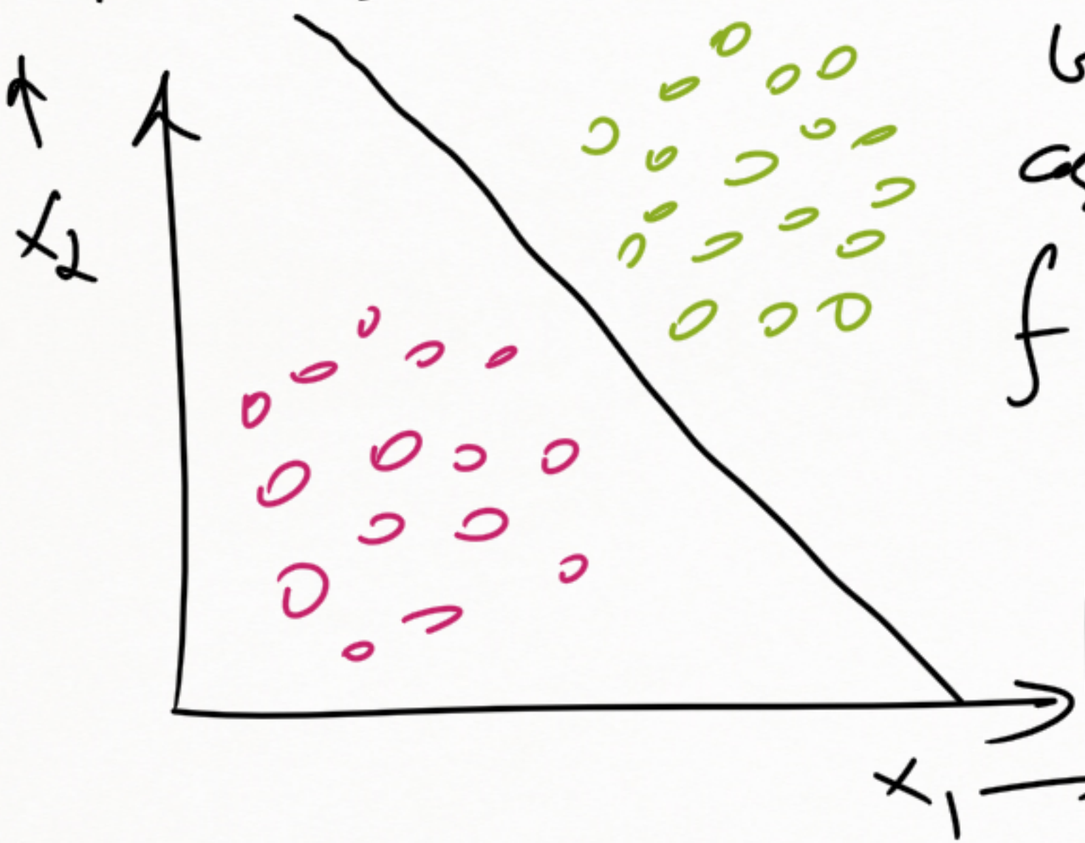
So, what multiclass logistic Regression is actually doing — well we know that logistic regression helps to find the discriminant function. or we can say that equation of the decision boundary which divides the data points of the two classes properly or discriminates between them properly.

In the case of binary classification, the following scenario happens —



Thus the linear discriminant function which we get — and this function is given as —

$$f(X) = \hat{\theta}_0 + \hat{\theta}^T X$$

$$X = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \qquad \hat{\theta} = \begin{bmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{bmatrix}$$

and if we put it equal to zero then it becomes the equation of decision boundary

Feature Vector of an Example

That is –

$$f(x) = \hat{\theta}_0 + \hat{\theta}^T x = 0 \Rightarrow \hat{\theta}_0 + \hat{\theta}_1 x_1 + \hat{\theta}_2 x_2 = 0$$

$$\Rightarrow \hat{\theta}_2 x_2 = -\hat{\theta}_0 - \hat{\theta}_1 x_1$$

Dividing above equation by $\hat{\theta}_2$

$$\Rightarrow \boxed{x_2 = -\frac{\hat{\theta}_0}{\hat{\theta}_2} - \frac{\hat{\theta}_1}{\hat{\theta}_2} x_1} \Rightarrow$$

This is in the form of $y = mx + c$

where $m = -\frac{\hat{\theta}_1}{\hat{\theta}_2}$, $c = -\frac{\hat{\theta}_0}{\hat{\theta}_2}$

Now, the $\hat{\theta}$s in this equation, that is $\hat{\theta}_1$ and $\hat{\theta}_2$ are the same weights which can be found by selecting eigen vector with the highest eigen value after applying SVD to the $S_B \cdot S_W^{-1}$ matrix in the explanation of LDA as dimensionality reduction algorithm. That is $\begin{bmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{bmatrix}$ is a vector such that if we are going to project our data on that vector then we get Maximum Between Variance and minimum within variance

Similarly, in multiclass logistic regression what we are trying to do —

Linear Discriminant function given by —

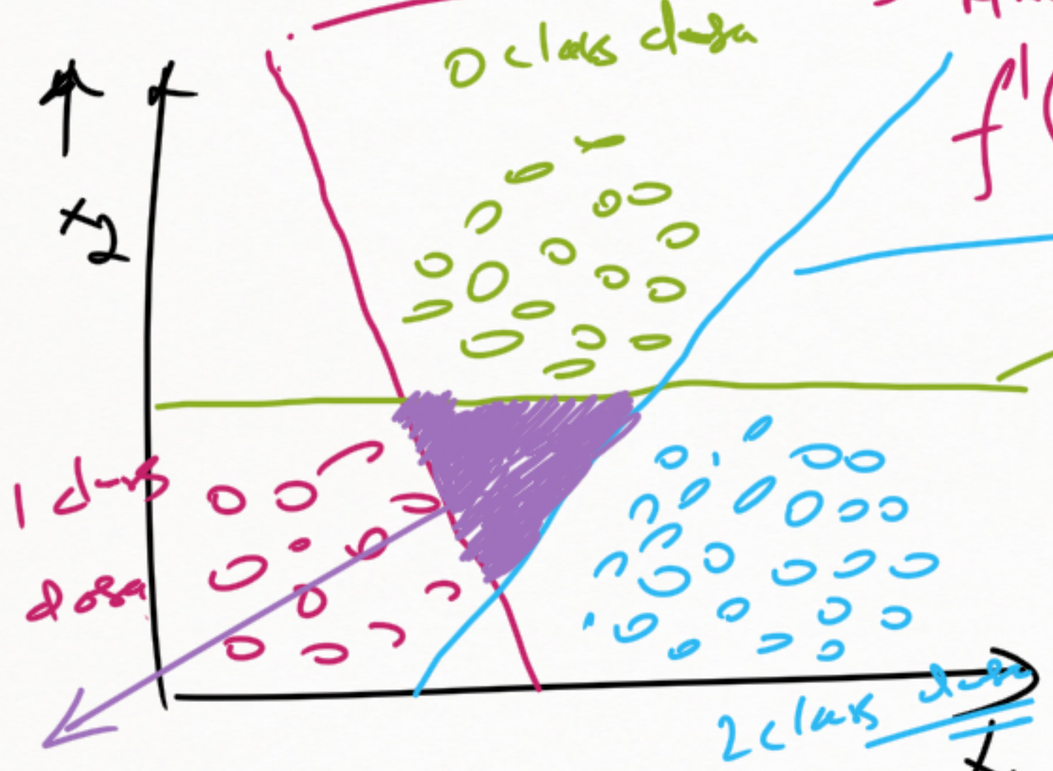
$$f^1(x) = \hat{\theta}_0^1 + \hat{\theta}^{1T} x$$

$$f^2(x) = \hat{\theta}_0^2 + \hat{\theta}^{2T} x$$

Linear Discriminant function given by —

$$\hat{\theta}_0 + \hat{\theta}^T x = f^0(x)$$

The '0' above signifies that it is a discriminant function which helps separate '0' class data from other classes



This is no man's land. If any data point lies here then we may not be able to classify it correctly.

In Binary Logistic Regression, the posterior probability is given by -

$$P(c^i = 1 \mid x_i) = \frac{1}{1 + e^{-(\hat{\theta}_0 + \hat{\theta}^T x_i)}}$$

$$x_i = \begin{bmatrix} x_1^i \\ x_2^i \end{bmatrix}$$

Only single decision boundary or single discriminant function is required

and the Posterior Probability of Negative class is given by -

$$P(c^i = 0 \mid x_i) = 1 - P(c^i = 1 \mid x_i) = \left[1 - \frac{1}{1 + e^{-(\hat{\theta}_0 + \hat{\theta}^T x_i)}} \right]$$

But, in the multiclass logistic regression —

$$P(c^i=0 \mid x_i) = \frac{e^{(\hat{\theta}_0^0 + \hat{\theta}^{0T} x_i)}}{e^{(\hat{\theta}_0^0 + \hat{\theta}^{0T} x_i)} + e^{(\hat{\theta}_0^1 + \hat{\theta}^{1T} x_i)} + e^{(\hat{\theta}_0^2 + \hat{\theta}^{2T} x_i)}}$$

Softmax function for class 0

$$P(c^i=1 \mid x_i) = \frac{e^{(\hat{\theta}_0^1 + \hat{\theta}^{1T} x_i)}}{e^{(\hat{\theta}_0^0 + \hat{\theta}^{0T} x_i)} + e^{(\hat{\theta}_0^1 + \hat{\theta}^{1T} x_i)} + e^{(\hat{\theta}_0^2 + \hat{\theta}^{2T} x_i)}}$$

for class 1

Similarly —

$$P(c^i = 2 \mid x_i) = \frac{e^{(\hat{\theta}_0^2 + \hat{\theta}^{2T} x_i)}}{e^{(\hat{\theta}_0^0 + \hat{\theta}^{0T} x_i)} + e^{(\hat{\theta}_0^1 + \hat{\theta}^{1T} x_i)} + e^{(\hat{\theta}_0^2 + \hat{\theta}^{2T} x_i)}}$$

So, in general we can say that if we have M classes in a dataset and N_{train} examples, then for any m^{th} example having feature vector which is N dimensional, we can say that the Posterior probability of an mth example belonging to class k is given by —

$N \times 1$ vector

$$P(c^m = k \mid x_i) = \frac{e^{\left(\hat{\theta}_0^k + \hat{\theta}^{kT} x_i\right)}}{\sum_{k=0}^{M-1} e^{\left(\hat{\theta}_0^k + \hat{\theta}^{kT} x_m\right)}}$$

$N \times 1$ feature vector

Class label of instances

$$\hat{\theta}^k = \begin{bmatrix} \hat{\theta}_1^k \\ \vdots \\ \hat{\theta}_N^k \end{bmatrix} \qquad x_m = \begin{bmatrix} x_1^m \\ \vdots \\ x_N^m \end{bmatrix}$$

$N \times 1$ vector

Since, there are M classes so,

$$K = \underbrace{0, 1, 2, 3, \ldots\ldots, M-1}_{\text{Total M classes}}$$

$\hat{\theta}_0$ values for $f(x)$ of decision boundary for each class

$$\hat{\theta}_0 = \underbrace{\left[\hat{\theta}_0^0, \hat{\theta}_0^1, \hat{\theta}_0^2, \hat{\theta}_0^3, \ldots\ldots \hat{\theta}_0^{M-1}\right]}_{1 \times M \text{ vector}}$$

$$\hat{\theta}^0_0 + \hat{\theta}^{0T} X \rightarrow N\times 1 \text{ vector}$$

$$\hat{\theta}^1 \rightarrow \hat{\theta}^1_0 + \hat{\theta}^{1T} X \rightarrow N\times 1 \text{ vector}$$

$$\hat{\theta}^k \rightarrow \hat{\theta}^k_0 + \hat{\theta}^{kT} X \rightarrow N\times 1 \text{ vector}$$

$$\hat{\theta}^{M-1}$$

$$\hat{\theta} = \underbrace{\begin{bmatrix} \hat{\theta}^0_1 & \hat{\theta}^1_1 & \cdots & \hat{\theta}^k_1 & \cdots & \hat{\theta}^{M-1}_1 \\ \hat{\theta}^0_2 & \hat{\theta}^1_2 & \cdots & \hat{\theta}^k_2 & \cdots & \hat{\theta}^{M-1}_2 \\ \vdots & \vdots & & \vdots & & \vdots \\ \hat{\theta}^0_N & \hat{\theta}^1_N & \cdots & \hat{\theta}^k_N & \cdots & \hat{\theta}^{M-1}_N \end{bmatrix}}_{N\times M \text{ matrix}}$$

$N\times 1$ vector

$N\times 1$ vector

$N\times 1$ vector

$N\times 1$ vector

$N\times 1$ vector

Similarly, for the last class $M-1$, there is also a discriminant function, $f^{M-1}(x)$

$$= \hat{\theta}^{M-1}_0 + \hat{\theta}^{M-1T} X$$

$$\begin{bmatrix} \hat{\theta}^{M-1}_1 \\ \hat{\theta}^{M-1}_2 \\ \vdots \\ \hat{\theta}^{M-1}_N \end{bmatrix} N\times 1 \text{ vector}$$

The $-\log_e L(\hat{\theta}_0, \hat{\theta})$ for Multiclass Logistic Regression is given by —

$\hat{\theta}_0$: $1 \times M$ vector

$\hat{\theta}$: $N \times M$ matrix

$$\frac{-1}{N_{train}} \sum_{m=1}^{N_{train}} \sum_{k=0}^{M-1} C_m^k \log_e \left[\frac{e^{(\hat{\theta}_0^k + \hat{\theta}^{kT} X_m)}}{\sum_{i=0}^{M-1} e^{(\hat{\theta}_0^i + \hat{\theta}^{iT} X_m)}} \right]$$

C_m^k = k^{th} bit of M dimensional one hot encoded row vector for the class label of mth Training Example

$$C^m \Rightarrow \underbrace{\left[C_m^0 \; C_m^1 \; C_m^2 \; C_m^3 \; \cdots \; C_m^k \; \cdots \; C_m^{M-1} \right]}_{1 \times M \text{ vector}}$$

↳ An integer value between 0 to $M-1$

If $c^m = k$ then —

$$c^m \Rightarrow \begin{bmatrix} c_m^0 & c_m^1 & c_m^2 & c_m^3 & \cdots & c_m^k & \cdots & c_m^{M-1} \end{bmatrix}$$

$$\Rightarrow \underbrace{\begin{bmatrix} 0 & 0 & 0 & 0 & \cdots & 1 & \cdots & 0 \end{bmatrix}}_{1 \times M \text{ row vector}}$$

In Gradient Descent algorithm, in order to minimize $-\log_e L(\hat{\theta}_0, \hat{\theta})$, we have to perform following two updates —

$$\left[\frac{\partial}{\partial \hat{\theta}_0^0}\Big|_{\substack{\hat{\theta}_0^{initial} \\ \hat{\theta}^{initial}}} \cdots \cdots \frac{\partial}{\partial \hat{\theta}_0^{M-1}}\Big|_{\substack{\hat{\theta}_0^{initial} \\ \hat{\theta}^{initial}}}\right]$$

1 × M vector

$$\hat{\theta}_0^{final} = \hat{\theta}_0^{initial} + \epsilon \cdot \boxed{\nabla_{\hat{\theta}_0}\left[\log_e L(\hat{\theta}_0, \hat{\theta})\right]\Big|_{\substack{\hat{\theta}_0^{initial} \\ \hat{\theta}^{initial}}}}$$

1 × M vector

$$\left[\hat{\theta}_0^{0\,final} \cdots \cdots \hat{\theta}_0^{M-1\,final}\right]$$

1 × M vector

$$\left[\hat{\theta}_0^{0\,initial} \cdots \cdots \hat{\theta}_0^{M-1\,final}\right]$$

Update #2 :-

$N \times M$ vector

$$\hat{\theta}^{\text{final}} = \hat{\theta}^{\text{initial}} + \epsilon \cdot \boxed{\nabla_{\hat{\theta}} \left[\log L(\hat{\theta}_0, \hat{\theta}) \right] \Big|_{\substack{\hat{\theta}_0^{\text{initial}} \\ \hat{\theta}^{\text{initial}}}}}$$

$$\begin{bmatrix} \hat{\theta}_1^{0\,\text{final}} & \cdots & \hat{\theta}_1^{M-1\,\text{final}} \\ \vdots & & \vdots \\ \hat{\theta}_N^{0\,\text{final}} & \cdots & \hat{\theta}_N^{M-1\,\text{final}} \end{bmatrix}$$

$N \times M$ vector

$$\begin{bmatrix} \hat{\theta}_1^{0\,\text{initial}} & \cdots & \hat{\theta}_1^{M-1\,\text{initial}} \\ \vdots & & \vdots \\ \hat{\theta}_N^{0\,\text{initial}} & \cdots & \hat{\theta}^{M-1\,\text{initial}} \end{bmatrix}$$

$N \times M$ vector

$$\begin{bmatrix} \frac{\partial}{\partial \hat{\theta}_1^0} \Big|_{\substack{\hat{\theta}_0^{\text{initial}} \\ \hat{\theta}^{\text{initial}}}} & \cdots & \frac{\partial}{\partial \hat{\theta}_1^{M-1}} \Big|_{\substack{\hat{\theta}_0^{\text{initial}} \\ \hat{\theta}^{\text{initial}}}} \\ \vdots & & \\ \frac{\partial}{\partial \hat{\theta}_N^0} \Big|_{\substack{\hat{\theta}_0^{\text{initial}} \\ \hat{\theta}^{\text{initial}}}} & \cdots & \frac{\partial}{\partial \hat{\theta}_N^{M-1}} \Big|_{\text{initial}} \end{bmatrix}$$

Therefore now our focus will be on computing following two gradients —

$$\nabla_{\hat{\theta}_0} \log_e L(\hat{\theta}_0, \hat{\theta}) \quad \& \quad \nabla_{\hat{\theta}} \log_e L(\hat{\theta}_0, \hat{\theta})$$

1 X M row vector N X M vector

Lets first try computing $\nabla_{\hat{\theta}_0} \log_e L(\hat{\theta}_0, \hat{\theta})$ –

$$\nabla_{\hat{\theta}_0} \log_e L(\hat{\theta}_0, \hat{\theta}) = \left[\boxed{\frac{\partial}{\partial \hat{\theta}_0^0}}, \frac{\partial}{\partial \hat{\theta}_0^1}, \ldots \ldots \ldots \frac{\partial}{\partial \hat{\theta}_0^{M-1}} \right]$$

$1 \times M$ vector

Lets compute this first

$$\frac{\partial}{\partial \hat{\theta}_0^0}\left[\cancel{\frac{-1}{N_{train}}\sum_{m=1}^{N_{train}}}\sum_{k=0}^{M-1} c_m^k \cdot \log_e\left[\frac{e^{(\hat{\theta}_0^k + \hat{\theta}^{kT} x_m)}}{\sum_{i=0}^{M-1} e^{(\hat{\theta}_0^i + \hat{\theta}^{iT} x_i)}}\right]\right]$$

Let's compute this derivative for a single example and remove the outer summation

$$\frac{\partial}{\partial \hat{\theta}_0^{*}} \underbrace{\left[\sum_{k=0}^{M-1} c_m^k \log_e \frac{e^{(\hat{\theta}_0^k + \theta^{kT} x_m)}}{\sum_{i=0}^{M-1} e^{(\hat{\theta}_0^i + \theta^{iT} x_m)}} \right]}$$

This sum can be written as —

$$\frac{\partial}{\partial \hat{\theta}_0^0}\left[C_m^0 \cdot \log_c \frac{e^{(\hat{\theta}_0^0 + \hat{\theta}^{0T} x_m)}}{\sum_{k=0}^{M-1}} + \cdots\cdots C_m^{M-1} \cdot \log_c \frac{e^{(\hat{\theta}_0^{M-1} + \theta^{M-1T} x_m)}}{\sum_{k=0}^{M-1}} \right]$$

For sometime, I am not writing huge sum completely as it takes lot of time

Lets only solve this term

$$\left[\boxed{C_m^0 \cdot \frac{\partial}{\partial \hat{\theta}_0^0} \log_c \frac{e^{\left(\hat{\theta}_0^0 + \hat{\theta}^{0T} x_m\right)}}{\sum_{k=0}^{M-1}}} + \cdots C_m^{M-1} \cdot \frac{\partial}{\partial \hat{\theta}_0^0} \log_c \frac{e^{\left(\hat{\theta}_0^{M-1} + \hat{\theta}^{M-1T} x_m\right)}}{\sum_{k=0}^{M-1}} \right]$$

$$\frac{\partial}{\partial \hat{\theta}_0^0} \log\left[\frac{e^{(\hat{\theta}_0^0 + \hat{\theta}^{0T} x_m)}}{\sum_{k=0}^{M-1}}\right] =$$

I am solving them directly, you guys please cross verify my results =

$$\left[\frac{\sum_{k=0}^{M-1} \cdot e^{(\hat{\theta}_0^0 + \hat{\theta}^{0T} x_m)} - e^{(\hat{\theta}_0^0 + \hat{\theta}^{0T} x_m)} \cdot e^{(\hat{\theta}_0^0 + \hat{\theta}^{0T} x_m)}}{\left(\sum_{k=0}^{M-1}\right)^2} - \right]$$

$$\frac{\sum_{k=0}^{M-1}}{e^{(\hat{\theta}_0^0 + \hat{\theta}^{0T} x_m)}}$$

$$= \left[\frac{\sum_{k=0}^{M-1} - e^{(\hat{\theta}_0^0 + \hat{\theta}^{0T} x_m)}}{\sum_{k=0}^{M-1}} \right] \longrightarrow (1)$$

$$\boxed{= \left[1 - \frac{e^{(\hat{\theta}_0^0 + \hat{\theta}^{0T} x_m)}}{\sum_{k=0}^{M-1}} \right] = \frac{\partial}{\partial \hat{\theta}_0^0} \frac{e^{(\hat{\theta}_0^0 + \hat{\theta}^{0T} x_m)}}{\sum_{k=0}^{M-1}}}$$

Lets compute for any other class -

$$\frac{\partial}{\partial \hat{\theta}_0^0} \log_e \frac{e^{\left(\hat{\theta}_0^k + \hat{\theta}^{k^T} X_{tr}\right)}}{\sum_{k=0}^{M-1}}$$

where

$k \neq 0$ class

$$= \frac{\sum_{k=0}^{M-1}}{e^{\left(\hat{\theta}_0^k + \hat{\theta}^{kT} X_m\right)}} \left[\frac{\sum_{k=0}^{M-1} (0) - e^{\left(\hat{\theta}_0^k + \hat{\theta}^{kT} X_m\right)} \cdot e^{\left(\hat{\theta}_0^0 + \hat{\theta}^{0T} X_m\right)}}{\left(\sum_{k=0}^{M-1}\right)^2} \right]$$

$$= \frac{-\left(\sum_{k=0}^{M-1}\right) \cancel{e^{(\hat{\theta}_0^k + \hat{\theta}^{kT} X_m)}} \cdot e^{(\hat{\theta}_0^0 + \hat{\theta}^{0T} X_m)}}{\cancel{e^{(\hat{\theta}_0^k + \hat{\theta}^{kT} X_m)}} \cdot \left(\sum_{k=0}^{M-1}\right)^{\cancel{2}}}$$

$$\frac{\partial}{\partial \hat{\theta}_0^0}\left[\log_e \frac{e^{(\hat{\theta}_0^k + \hat{\theta}^{kT} x_m)}}{\sum}\right] = \frac{-e^{(\hat{\theta}_0^0 + \hat{\theta}^{0T} x_m)}}{\sum_{k=0}^{M-1}}$$

when $k \neq 0$

2

Therefore, by ① and ②, please remember—

$$\frac{\partial}{\partial \hat{\theta}_0^i}\left[\log_e \frac{e^{(\hat{\theta}_0^k + \hat{\theta}^{kT} x_m)}}{\sum_{k=0}^{M-1} e^{(\hat{\theta}_0^k + \hat{\theta}^{kT} x_m)}}\right] = 1 - \frac{e^{(\hat{\theta}_0^i + \hat{\theta}^{iT} x_m)}}{\sum_{k=0}^{M-1} e^{(\hat{\theta}_0^k + \hat{\theta}^{kT} x_m)}} \longrightarrow ③$$

$\boxed{\text{if } i = k}$

$$\frac{\partial}{\partial \hat{\theta}_0^i}\left[\log_e \frac{e^{(\hat{\theta}_0^k + \hat{\theta}^{kT} X_m)}}{\sum_{k=0}^{M-1} e^{(\hat{\theta}_0^k + \hat{\theta}^{kT} X_m)}}\right] = -\frac{e^{(\hat{\theta}_0^i + \hat{\theta}^{iT} X_m)}}{\sum_{k=0}^{M-1} e^{(\hat{\theta}_0^k + \hat{\theta}^{kT} X_m)}} \longrightarrow (4)$$

$\boxed{\text{if } i \neq k}$

Similarly, we can say that — n^{th} feature of mth example

5

$$\boxed{\frac{\partial}{\partial \hat{\theta}_n^i}\left[\log_e \frac{e^{\left(\hat{\theta}_0^k + \hat{\theta}^{kT} x_m\right)}}{\sum_{k=0}^{M-1} e^{\left(\hat{\theta}_0^k + \hat{\theta}^{kT} x_m\right)}}\right] = \left[1 - \frac{e^{\left(\hat{\theta}_0^i + \hat{\theta}^{iT} x_m\right)}}{\sum_{k=0}^{M-1} e^{\left(\hat{\theta}_0^k + \hat{\theta}^{kT} x_m\right)}}\right] \cdot x_n^m \quad \text{if } i = k}$$

PARAMETER CORRESPONDING TO n^{th} FEATURE OF DATASET AND IT IS ALSO n^{th} COMPONENT OF $\hat{\theta}^i$ PARAMETER VECTOR OF DISCRIMINANT FUNCTION FOR CLASS i =

Similarly –

6

$$\frac{\partial}{\partial \hat{\theta}_n^i}\left[\log_e \frac{e^{(\hat{\theta}_0^k + \hat{\theta}^{kT} x_m)}}{\sum_{k=0}^{M-1} e^{(\hat{\theta}_0^k + \hat{\theta}^{kT} x_m)}}\right] = -\frac{e^{(\hat{\theta}_0^i + \hat{\theta}^{iT} x_m)}}{\sum_{k=0}^{M-1} e^{(\hat{\theta}_0^k + \hat{\theta}^{kT} x_m)}} \cdot x_n^m$$

$\boxed{\text{if } i \neq k}$

So, now going back —

$$\frac{\partial}{\partial \hat{\theta}_0^0}\left[C_m^0 \log_e \frac{e^{(\hat{\theta}_0^0 + \hat{\theta}^{0T} x_m)}}{\Sigma} + \cdots + C_m^{M-1} \cdot \log_e \frac{e^{(\hat{\theta}_0^{M-1} + \hat{\theta}^{M-1T} x_m)}}{\Sigma} \right]$$

$=$

By ③ and ④

$$\left[c_m^0 \left(1 - \frac{e^{(\hat{\theta}_0^0 + \hat{\theta}^{0T} x_m)}}{\Sigma} \right) + c_m^1 \left(0 - \frac{e^{(\hat{\theta}_0^0 + \hat{\theta}^{0T} x_m)}}{\Sigma} \right) + \cdots + c_m^{M-1} \left(0 - \frac{e^{(\hat{\theta}_0^{M-1} + \hat{\theta}^{M-1T} x_m)}}{\Sigma} \right) \right]$$

This can be written as —