Softmax

$$X \in R^{k \times n}$$

k = batch size
n = input dim

$$Y = \text{Softmax}(X)$$

$$Y \in R^{k \times n}$$

Forward Propagation

Suppose $X \in R^{1 \times n}$, $Y \in R^{1 \times n}$ then

$$Y = \text{Softmax}(X)$$

$$Y_i = \frac{e^{X_i}}{\sum_j e^{X_j}}, \quad \text{here } 1 \leq i \leq n, \; 1 \leq j \leq n$$

1

Now suppose $X \in R^{k \times n}$, $Y \in R^{k \times n}$

here k = batch size

n = input dim

then by equation 1

$$\boxed{Y_{ij} = \frac{e^{x_{ij}}}{\sum_{d} e^{x_{id}}}, \quad \text{where } \begin{array}{l} 1 \leq i \leq k \\ 1 \leq j \leq n \\ 1 \leq d \leq n \end{array}} \qquad (2)$$

Derivative

Suppose $X \in R^{1 \times n}$, $Y \in R^{1 \times n}$

then
$$\frac{\partial L}{\partial X} = \frac{\partial L}{\partial Y} \frac{\partial Y}{\partial X}$$

where $\frac{\partial L}{\partial Y} \in R^{1 \times n}$

Now by equation (1)

$$Y_i = \frac{e^{x_i}}{\sum_j e^{x_j}}$$

$$\left(\frac{\partial Y_i}{\partial x}\right)_j = \begin{cases} \end{cases}$$

To solve above suppose $n=3$ and

$$X = [x_1, x_2, x_3]$$

$$Y = [Y_1, Y_2, Y_3]$$

then,

$$Y_i = \frac{e^{x_i}}{e^{x_1}+e^{x_2}+e^{x_3}}$$

Now there are two cases when $\frac{\partial Y_i}{\partial x_i}$ and $\frac{\partial Y_i}{\partial x_j}$

So for first

$$\frac{dY_1}{dx_1} = \frac{d}{dx_1}\left(\frac{e^{x_1}}{e^{x_1}+e^{x_2}+e^{x_3}}\right)$$

$$\frac{dY_1}{dx_1} = e^{x_1}\frac{d}{dx_1}\left(\frac{1}{e^{x_1}+e^{x_2}+e^{x_3}}\right) +$$

$$\frac{de^{x_1}}{dx_1}\left(\frac{1}{e^{x_1}+e^{x_2}+e^{x_3}}\right)$$

$$\frac{dY_1}{dx_1} = \frac{-e^{x_1}\,e^{x_1}}{(e^{x_1}+e^{x_2}+e^{x_3})^{-2}} + \frac{e^{x_1}}{e^{x_1}+e^{x_2}+e^{x_3}}$$

Now Suppose

$$p_1 = \frac{e^{x_1}}{(e^{x_1}+e^{x_2}+e^{x_3})^{-2}}$$

then

$$\frac{dY_1}{dx_1} = -P_1^2 + P_1$$

$$\frac{dY_1}{dx_1} = P_1(1-P_1) \qquad \text{③}$$

now for $\dfrac{dY_i}{dx_j}$

$$\frac{dY_1}{dx_2} = \frac{d}{dx_2}\left(\frac{e^{x_1}}{e^{x_1}+e^{x_2}+e^{x_3}}\right)$$

$$\frac{dY_1}{dx_2} = \frac{-e^{x_1}e^{x_2}}{(e^{x_1}+e^{x_2}+e^{x_3})^2}$$

Now Suppose

$$P_1 = \frac{e^{x_1}}{(e^{x_1}+e^{x_2}+e^{x_3})}$$

$$P_2 = \left(\frac{e^{x_2}}{e^{x_1}+e^{x_2}+e^{x_3}}\right)$$

then

$$\frac{\partial Y_1}{\partial x_2} = -P_1 P_2 \qquad \text{——— (4)}$$

Now we can generalize (3) and (4)

$$\left(\frac{\partial Y}{\partial x}\right)_{i,j} = \left(\frac{\partial Y_i}{\partial x}\right)_j = \begin{cases} P_i(1-P_i) & i=j \\ -P_i P_j & i \neq j \end{cases}$$

$$\text{where, } \frac{\partial Y}{\partial x} \in R^{n \times n} \qquad \text{——— (5)}$$

Now

$$\frac{\partial L}{\partial x} = \frac{\partial L}{\partial Y} \frac{\partial Y}{\partial x}$$

here $\frac{\partial L}{\partial Y} \in R^{1 \times n}$, $\frac{\partial Y}{\partial x} \in R^{n \times n}$

then by equation 5

$$\boxed{\begin{gathered}\left(\frac{\partial L}{\partial \vec{x}}\right)_i = \left(\frac{\partial L}{\partial \vec{y}}\right)_1 \left(\frac{\partial \vec{y}}{\partial \vec{x}}\right)_{i1} + \cdots + \left(\frac{\partial L}{\partial \vec{y}}\right)_n \left(\frac{\partial \vec{y}}{\partial \vec{x}}\right)_{in} \\ \frac{\partial L}{\partial \vec{x}} \in R^{1\times n}\end{gathered}}$$

6

Now suppose $X \in R^{k\times n}$, $Y \in R^{k\times n}$

=) Here take each input independent and once find put it in matrix so the output from 6 will be

$$\frac{\partial L}{\partial Y} \in R^{k\times n}, \quad \frac{\partial Y}{\partial X} \in R^{k\times n\times n}, \quad \frac{\partial L}{\partial \vec{x}} \in R^{k\times n}$$

=) Means we have to apply 6 k times.