Cross Entropy Error with Softmax

Madison Cooley

July 24, 2018

Cost Function:

$$L = -\frac{1}{N}\sum_{n=1}^{N} y_n log(f(BAx_n))$$

Loss Function:

$$L_n = -\sum_{j=1}^{C} y_{nj} log(\hat{y}_{nj})$$

$$\hat{y}_{nj} = \frac{e^{s_{nj}}}{\sum_{c=1}^{C} e^{s_{nc}}}$$

$$s_{nj} = B_j.Ax_n$$

1 Background

$A \in \mathbb{R}^{D\times P}$
$B \in \mathbb{R}^{C\times D}$
$x_n \in \mathbb{R}^{P\times 1}$
$y_n \in \mathbb{R}^{1\times C}$
$\hat{y}_n \in \mathbb{R}^{1\times C}$

P: number of words
D: dimension
C: classes

2 Derivative of Loss Function w.r.t. $B_{j\cdot}$ (gradient of top layer weights, where j denotes the row)

$$\frac{\partial L_n}{\partial B_{j\cdot}} = \frac{\partial L_n}{\partial s_{nj}} \frac{\partial s_{nj}}{\partial B_{j\cdot}} \qquad \text{(chain rule)}$$

So, examining each factor separately:

$$\frac{\partial L_n}{\partial s_{nj}} = \sum_{j=1}^{C} \frac{\partial L_n}{\partial \hat{y}_{nj}} \frac{\partial \hat{y}_{nj}}{\partial s_{nj}}$$

$$\frac{\partial L_n}{\partial \hat{y}_{nj}} = -\frac{y_{nj}}{\hat{y}_{nj}}$$

If $i = j$: , and $\sum_C = \sum_{c=1}^{C} e^{s_{nc}}$

$$\frac{\partial \hat{y}_{nj}}{\partial s_{nj}} = \frac{\frac{\partial e^{s_{nj}}}{\partial s_{nj}} \sum_C - \frac{\partial \sum_C}{\partial s_{nj}} e^{s_{nj}}}{[\sum_C]^2} \qquad \text{(quotient rule)}$$

$$= \frac{e^{s_{nj}} \sum_C - e^{s_{nj}} e^{s_{nj}}}{[\sum_C]^2}$$

$$= \frac{e^{s_{nj}}}{\sum_C} \frac{\sum_C - e^{s_{nj}}}{\sum_C}$$

$$= \frac{e^{s_{nj}}}{\sum_C} (1 - \frac{e^{s_{nj}}}{\sum_C})$$

$$= \hat{y}_{nj}(1 - \hat{y}_{nj})$$

If $i \neq j$:

$$\frac{\partial \hat{y}_{ni}}{\partial s_{nj}} = \frac{\frac{\partial e^{s_{ni}}}{\partial s_{nj}} \sum_C - \frac{\partial \sum_C}{\partial s_{nj}} e^{s_{nj}}}{[\sum_C]^2} \qquad \text{(quotient rule)}$$

$$= \frac{0 \sum_C - e^{s_{ni}} e^{s_{nj}}}{[\sum_C]^2}$$

$$= \frac{-e^{s_{ni}}}{\sum_C} \frac{e^{s_{nj}}}{\sum_C}$$

$$= -\hat{y}_{ni}\hat{y}_{nj}$$

So,

$$\frac{\partial L_n}{\partial s_{nj}} = \sum_{j=1}^{C} \frac{\partial L_n}{\partial \hat{y}_{nj}} \frac{\partial \hat{y}_{nj}}{\partial s_{nj}}$$

$$= \frac{\partial L_n}{\partial \hat{y}_{nj}} \frac{\partial \hat{y}_{nj}}{\partial s_{nj}} - \sum_{j \neq i} \frac{\partial L_n}{\partial \hat{y}_{nj}} \frac{\partial \hat{y}_{nj}}{\partial s_{nj}} \qquad \text{(pull out case where } i = j \text{)}$$

$$= -\frac{y_{nj}}{\hat{y}_{nj}} \hat{y}_{nj}(1 - \hat{y}_{nj}) + \sum_{j \neq i} \frac{y_{nj}}{\hat{y}_{nj}} \hat{y}_{nj}\hat{y}_{nj}$$

$$= -y_{nj}(1 - \hat{y}_{nj}) + \sum_{j \neq i} y_{nj}\hat{y}_{nj}$$

$$= -y_{nj} + y_{nj}\hat{y}_{nj} + \sum_{j \neq i} y_{nj}\hat{y}_{nj}$$

$$= \hat{y}_{nj}(y_{nj} + \sum_{j \neq i} y_{nj}) - y_{nj} \qquad \text{(rewrite)}$$

$$= \hat{y}_{nj} - y_{nj} \qquad \text{(since, } y_{nj} + \sum_{j \neq i} y_{nj} = 1 \text{)}$$

$$\frac{\partial B_j.Ax_n}{\partial B_j.} = (Ax_n)^T$$

And so finally,

$$\boxed{\frac{\partial L_n}{\partial B_j.} = (\hat{y}_{nj} - y_{nj})(Ax_n)^T}$$

3 Derivative of Loss Function w.r.t. $A_{i.}$, where $A_{i.}$ is one row vector of A, and $i = 1, 2..., d$.

$h_i = \sum_{p=1}^{P} A_{ip} x_{np}$

$s_{nj} = \sum_{i=1}^{N} B_{ji} h_i$

$$\frac{\partial L_n}{\partial A_{i.}} = \sum_{j=1}^{C} [\frac{\partial L_n}{\partial \hat{y}_{nj}} \frac{\partial \hat{y}_{nj}}{\partial s_{nj}} \frac{\partial s_{nj}}{\partial h_i}] \frac{\partial h_i}{\partial A_{i.}} = \sum_{j=1}^{C} [\frac{\partial L_n}{\partial s_{nj}} \frac{\partial s_{nj}}{\partial h_i}] \frac{\partial h_i}{\partial A_{i.}}$$

(where $\frac{\partial L_n}{\partial s_{nj}}$ is derived above.)

So, examining each factor separately...

$$\frac{\partial s_{nj}}{\partial h_i} = \frac{\partial \sum_{i=1}^{N} B_{ji} h_i}{\partial h_i} = B_{ji}$$

$$\frac{\partial h_i}{\partial A_{i.}} = \frac{\partial \sum_{p=1}^{P} A_{ip} x_{np}}{\partial A_{i.}} = \frac{\partial A_i x_n}{\partial A_{i.}} = x_n$$

So finally,

$$\boxed{\frac{\partial L_n}{\partial A_{i.}} = \sum_{j=1}^{C} [(\hat{y}_{nj} - y_{nj}) B_{ji}] x_n}$$