Cross Entropy Error with Softmax

Madison Cooley

July 20, 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 $\sum_C = \sum_{c=1}^{C} e^{s_{nc}}$

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

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

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

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

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

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

$$\frac{\partial L_n}{\partial A} = \frac{\partial L_n}{\partial s_n}\frac{\partial s_n}{\partial A} \qquad \text{(where } s_n = BAx_n \text{ and results in a } 2 \times 1 \text{ vector)}$$

$$\frac{\partial L_n}{\partial s_n} = \sum_{j=1}^{C} \frac{\partial L_n}{\partial s_{nj}} \qquad \left(\text{where } \frac{\partial L_n}{\partial s_{nj}} \text{ is derived above, } s_{nj} = B_j.Ax_n\text{, and is a scalar.}\right)$$

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

$$\frac{\partial s_n}{\partial A} = \frac{\partial BAx_n}{\partial s_n} = \sum_{j=1}^{C} \frac{\partial B_j.Ax_n}{\partial A}$$

$$= \sum_{j=1}^{C} \frac{\partial \sum_{n=1}^{d} b_{jn} * \sum_{i=1}^{p} a_{ni}x_i}{\partial a_{ni}}$$

$$= \sum_{n=1}^{d} b_{jn} * \sum_{i=1}^{p} x_i$$

$$= \sum_{j=1}^{C} B_j.^T x_n^T$$

So finally,

$$\boxed{\frac{\partial L_n}{\partial A} = \sum_{j=1}^{C} (\hat{y}_{nj} - y_{nj}) B_j.^T x_n^T}$$