Deep Learning Crash Course

Notes for derivative of CE loss

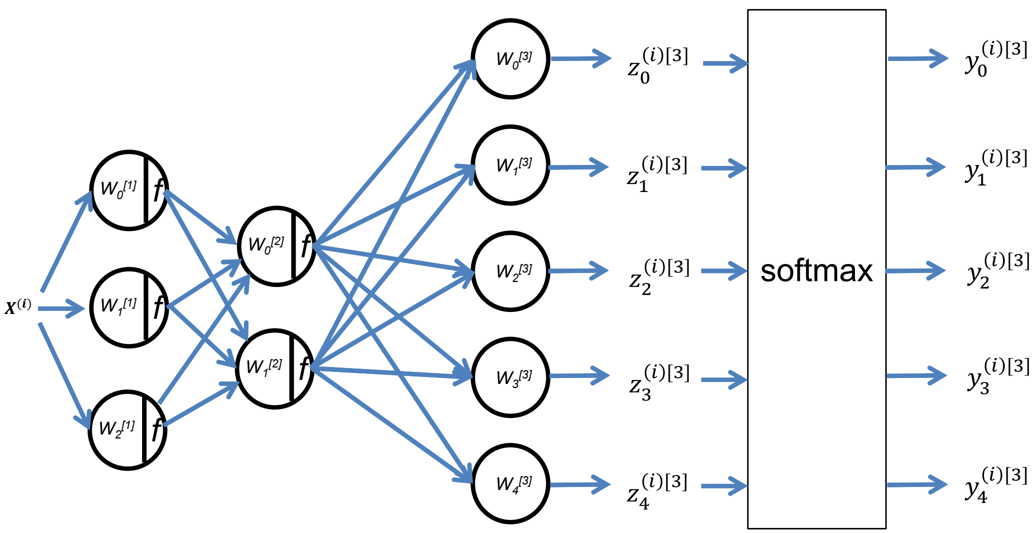
Hui Xue

Fall 2021

www.deeplearningcrashcourse.org

CE-Loss

K=5, 5 class classification



$$loss = \frac{1}{B}\sum_{i=0}^{B-1} L_{CE}(\boldsymbol{y}^{(i)}, \widehat{\boldsymbol{y}}^{(i)})$$

$$\widehat{\boldsymbol{y}}^{(i)} = \widehat{\boldsymbol{y}}(\boldsymbol{X}^{(i)};\ \boldsymbol{\theta})$$

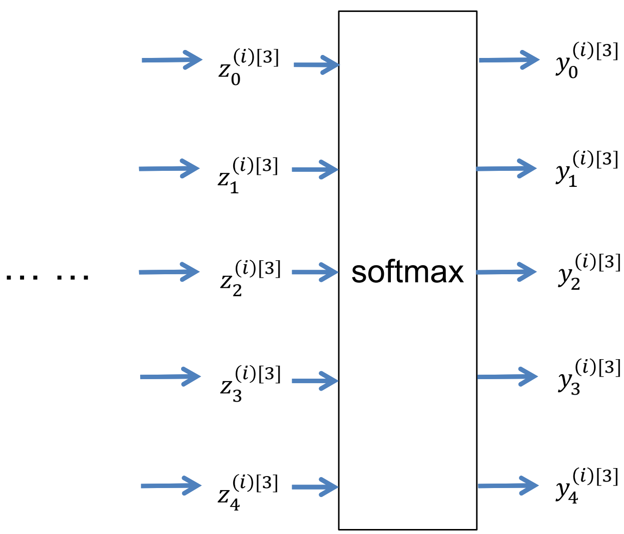
$$L_{CE}\left(\boldsymbol{y}^{(i)}, \widehat{\boldsymbol{y}}^{(i)}\right) = -\sum_{j=0}^{K-1} \boldsymbol{y}_j^{(i)} \log(\widehat{\boldsymbol{y}}_j^{(i)})$$

$$\widehat{\boldsymbol{y}}_j^{(i)} = \frac{e^{z_j}}{\sum_{t=0}^{K-1} e^{z_t}}$$

The goal is to compute $\frac{\partial loss}{\partial \boldsymbol{\theta}}$

CE-Loss

K=5, 5 class classification



$$loss = \frac{1}{B}\sum_{i=0}^{B-1} L_{CE}(\boldsymbol{y}^{(i)}, \widehat{\boldsymbol{y}}^{(i)})$$

$$\widehat{\boldsymbol{y}}^{(i)} = \widehat{\boldsymbol{y}}(\boldsymbol{X}^{(i)};\ \boldsymbol{\theta})$$

$$L_{CE}\left(\boldsymbol{y}^{(i)}, \widehat{\boldsymbol{y}}^{(i)}\right) = -\sum_{j=0}^{K-1} \boldsymbol{y}_j^{(i)} \log(\widehat{\boldsymbol{y}}_j^{(i)})$$

$$\widehat{\boldsymbol{y}}_j^{(i)} = \frac{e^{z_j}}{\sum_{t=0}^{K-1} e^{z_t}}$$

The goal is to compute $\frac{\partial loss}{\partial \boldsymbol{\theta}}$

$$\frac{\partial loss}{\partial \boldsymbol{\theta}} = \frac{1}{B}\sum_{i=0}^{B-1} \frac{\partial L_{CE}(\boldsymbol{y}^{(i)}, \widehat{\boldsymbol{y}}^{(i)})}{\partial \boldsymbol{\theta}}$$

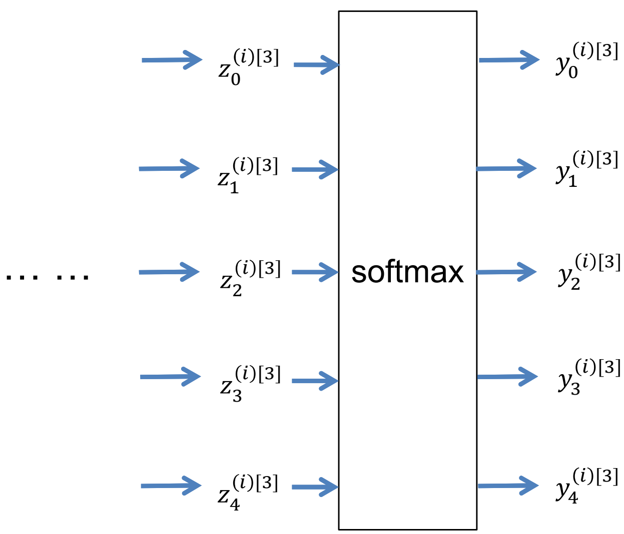
$$\frac{\partial L_{CE}(\boldsymbol{y}^{(i)}, \widehat{\boldsymbol{y}}^{(i)})}{\partial \boldsymbol{\theta}} = \left(\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{\theta}}\right)^T \frac{\partial L}{\partial \boldsymbol{z}}$$

We will compute

$$\frac{\partial L}{\partial \boldsymbol{z}} = \left[\frac{\partial L}{\partial z_0}, \frac{\partial L}{\partial z_1}, \dots, \frac{\partial L}{\partial z_{K-1}}\right]^T$$

CE-Loss

K=5, 5 class classification



$$loss = \frac{1}{B}\sum_{i=0}^{B-1} L_{CE}(\boldsymbol{y}^{(i)}, \hat{\boldsymbol{y}}^{(i)})$$

$$\hat{\boldsymbol{y}}^{(i)} = \hat{\boldsymbol{y}}(\boldsymbol{X}^{(i)};\ \boldsymbol{\theta})$$

$$L_{CE}\left(\boldsymbol{y}^{(i)}, \hat{\boldsymbol{y}}^{(i)}\right) = -\sum_{j=0}^{K-1} y_j^{(i)} \log(\hat{y}_j^{(i)})$$

$$\hat{y}_j^{(i)} = \frac{e^{z_j}}{\sum_{t=0}^{K-1} e^{z_t}}$$

The goal is to compute $\frac{\partial loss}{\partial \boldsymbol{\theta}}$

$$\frac{\partial L}{\partial z_k} = \frac{\partial[-\sum_{j=0}^{K-1} y_j \log(\hat{y}_j)]}{\partial z_k}$$

$$= -\sum_{j=0}^{K-1} y_j \frac{\partial \log(\hat{y}_j)}{\partial z_k}$$

$$= -\sum_{j=0}^{K-1} \frac{y_j}{\hat{y}_j} \frac{\partial \hat{y}_j}{\partial z_k}$$

$$\frac{\partial \hat{y}_j}{\partial z_k} = \frac{\partial\left[\frac{e^{z_j}}{\sum_{t=0}^{K-1} e^{z_t}}\right]}{\partial z_k}$$

CE-Loss

$$\frac{\partial \hat{y}_j}{\partial z_k} = \frac{\partial [\frac{e^{z_j}}{\sum_{t=0}^{K-1} e^{z_t}}]}{\partial z_k}$$

If $k = j$

$$\frac{\partial \hat{y}_j}{\partial z_j} = \frac{\partial [\frac{e^{z_j}}{\sum_{t=0}^{K-1} e^{z_t}}]}{\partial z_j}$$

$$= \frac{e^{z_j}}{\sum_{t=0}^{K-1} e^{z_t}} \cdot - \frac{e^{z_j}}{[\sum_{t=0}^{K-1} e^{z_t}]^2} e^{z_j}$$

$$= \hat{y}_j - \hat{y}_j^{\,2}$$

If $k \neq j$

$$\frac{\partial \hat{y}_j}{\partial z_k} = \frac{\partial [\frac{e^{z_j}}{\sum_{t=0}^{K-1} e^{z_t}}]}{\partial z_k}$$

$$= - \frac{e^{z_j}}{[\sum_{t=0}^{K-1} e^{z_t}]^2} e^{z_k}$$

$$= -\hat{y}_j \hat{y}_k$$

Jacobian

$$\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z}} = \begin{bmatrix} \frac{\partial \hat{y}_0}{\partial z_0} & \cdots & \frac{\partial \hat{y}_0}{\partial z_{k-1}} \\ \vdots & \ddots & \vdots \\ \frac{\partial \hat{y}_{K-1}}{\partial z_0} & \cdots & \frac{\partial \hat{y}_{K-1}}{\partial z_{K-1}} \end{bmatrix} = \begin{bmatrix} \hat{y}_0 - \hat{y}_0{}^2 & \cdots & -\hat{y}_0\hat{y}_{K-1} \\ \vdots & \ddots & \vdots \\ -\hat{y}_{k-1}\hat{y}_0 & \cdots & \hat{y}_{K-1} - \hat{y}_{K-1}{}^2 \end{bmatrix}$$

A symmetric jacobian $\hat{y}_j = \frac{e^{z_j}}{\sum_{t=0}^{K-1} e^{z_t}}$

CE-Loss

$$\frac{\partial L}{\partial z_k} = -\sum_{j=0}^{K-1} \frac{y_j}{\hat{y}_j} \frac{\partial \hat{y}_j}{\partial z_k}$$

$$= -\frac{y_k}{\hat{y}_k}\left(\hat{y}_k - \hat{y}_k{}^2\right) - \sum_{j=0, j\neq k}^{K-1} \frac{y_j}{\hat{y}_j}\left(-\hat{y}_j \hat{y}_k\right)$$

$$= -y_k(1 - \hat{y}_k) + \sum_{j=0, j\neq k}^{K-1} y_j \hat{y}_k$$

$$= -y_k + y_k\, \hat{y}_k + \sum_{j=0, j\neq k}^{K-1} y_j \hat{y}_k$$

$$= -y_k + \sum_{j=0}^{K-1} y_j \hat{y}_k$$

$$= -y_k + \hat{y}_k \sum_{j=0}^{K-1} y_j = \hat{y}_k - y_k$$

So we get:

$$\frac{\partial L}{\partial \boldsymbol{z}} = [\hat{y}_0 - y_0, \hat{y}_1 - y_1, \ldots, \hat{y}_{k-1} - y_{k1}]^T$$

$$\frac{\partial L_{CE}(\boldsymbol{y}^{(i)}, \hat{\boldsymbol{y}}^{(i)})}{\partial \boldsymbol{\theta}} = \left(\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{\theta}}\right)^T \frac{\partial L}{\partial \boldsymbol{z}}$$

We will not need to explicitly compute $\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{\theta}}$

But compute $\frac{\partial L_{CE}(\boldsymbol{y}^{(i)}, \hat{\boldsymbol{y}}^{(i)})}{\partial \boldsymbol{W}^{[3]}}$, $\frac{\partial L_{CE}(\boldsymbol{y}^{(i)}, \hat{\boldsymbol{y}}^{(i)})}{\partial \boldsymbol{W}^{[2]}}$, $\frac{\partial L_{CE}(\boldsymbol{y}^{(i)}, \hat{\boldsymbol{y}}^{(i)})}{\partial \boldsymbol{W}^{[1]}}$ layer by layer using backprop