Categorical cross-entropy loss

$$\mathcal{L}_{CCE} = -\sum_{i=1}^{K} y_i \log(\hat{y}_i)$$

where $\hat{y}_i$ is obtained using

SOFTMAX function

$$\hat{y}_i = \frac{e^{x_i}}{\sum_{j=1}^{K} e^{x_j}} \quad , \quad \hat{y}_1 + \hat{y}_2 + \cdots \hat{y}_K = 1$$

Compute the derivation of the loss function.

$$x_j = [x_1, x_2, x_3, \ldots, x_K]$$

$$\hat{y}_i = [y_1, y_2, y_3, \ldots, y_K]$$

$$\hat{y}_1 = \frac{e^{x_1}}{e^{x_1} + e^{x_2} + \cdots + e^{x_K}}$$

$$\hat{y}_2 = \frac{e^{x_2}}{e^{x_1} + e^{x_2} + \cdots + e^{x_K}}$$

$$\vdots$$

$$\hat{y}_K = \frac{e^{x_K}}{e^{x_1} + e^{x_2} + \cdots + e^{x_K}}$$

$$\Rightarrow \frac{\partial \hat{y}_i}{\partial x_j} = \begin{bmatrix} \frac{\partial \hat{y}_1}{\partial x_1} & \frac{\partial \hat{y}_1}{\partial x_2} & \cdots & \frac{\partial \hat{y}_1}{\partial x_K} \\ \frac{\partial \hat{y}_2}{\partial x_1} & \frac{\partial \hat{y}_2}{\partial x_2} & \cdots & \frac{\partial \hat{y}_2}{\partial x_K} \\ \vdots & \vdots & \ddots & \vdots \end{bmatrix}$$

$$\begin{bmatrix} \partial x_1 & \partial x_2 & \cdots & \partial x_k \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \hat{y}_k}{\partial x_1} & \frac{\partial \hat{y}_k}{\partial x_2} & \cdots & \frac{\partial \hat{y}_k}{\partial x_k} \end{bmatrix}$$

① $$h(x) = \frac{f(x)}{g(x)}$$

$$\frac{\partial h(x)}{\partial x} = \frac{f'(x)g(x) - g'(x)f(x)}{g(x)^2}$$

② $$\frac{\partial \left(f(x) + h(x)\right)}{\partial x} = \frac{\partial f(x)}{\partial x} + \frac{\partial h(x)}{\partial x}.$$

$$\frac{\partial \hat{y}_1}{\partial x_1} = \frac{\partial}{\partial x_1}\left(\frac{e^{x_1}}{e^{x_1} + e^{x_2} + \cdots + e^{x_k}}\right)$$

$$= \frac{e^{x_1}\left(e^{x_1} + e^{x_2} + \cdots + e^{x_k}\right) - e^{x_1} \times e^{x_1}}{\left(e^{x_1} + e^{x_2} + \cdots + e^{x_k}\right)^2}$$

$$= \frac{e^{x_1} \times \left(\Sigma - e^{x_1}\right)}{\Sigma \times \Sigma}$$

$$\boxed{\frac{\partial \hat{y}_1}{\partial x_1} = \hat{y}_1 \left(1 - \hat{y}_1\right)}$$

$$\boxed{\frac{\partial \hat{y}_1}{\partial x_1} = \hat{y}_1(1-\hat{y}_1)}$$

lly

$$\frac{\partial \hat{y}_2}{\partial x_2} = \hat{y}_2 \cdot (1-\hat{y}_2)$$

$$\vdots \qquad\qquad \vdots$$

$$\frac{\partial \hat{y}_k}{\partial x_k} = \hat{y}_k(1-\hat{y}_k)$$

$$\frac{\partial \hat{y}_2}{\partial x_1} = \frac{\partial\left(\frac{e^{x_2}}{e^{x_1}+e^{x_2}+\dots e^{x_k}}\right)}{\partial x_1}$$

$$= 0 - \frac{e^{x_1} e^{x_2}}{\Sigma \times \Sigma}$$

$$\boxed{\frac{\partial \hat{y}_2}{\partial x_1} = -\hat{y}_1 \times \hat{y}_2}$$

$$\boxed{\frac{\partial \hat{y}_k}{\partial x_1} = -\hat{y}_1 \times \hat{y}_k}$$

$$\frac{\partial \hat{y}_i}{\partial x_j} = \begin{bmatrix} \hat{y}_1(1-\hat{y}_1) & -\hat{y}_1 \cdot \hat{y}_2 & \cdots & -\hat{y}_1 \cdot \hat{y}_k \\ -\hat{y}_2 \cdot \hat{y}_1 & \hat{y}_2(1-\hat{y}_2) & \cdots & -\hat{y}_2 \cdot \hat{y}_k \\ \vdots & \vdots & \ddots & \vdots \\ -\hat{y}_k \cdot \hat{y}_1 & -\hat{y}_k \cdot \hat{y}_1 & & \hat{y}_k(1-\hat{y}_k) \end{bmatrix}$$

$\Rightarrow$ In general,

$$\frac{\partial \hat{y}_i}{\partial x_j} = \begin{cases} y_i(1-y_i) \ \forall i=j \\ -y_i \times y_j \ \forall \ i \neq j \end{cases}$$

$$\frac{\partial \mathcal{L}_{CCE}}{\partial x_j} = \frac{\partial}{\partial x_j}\left(-\sum_{i=1}^{k} y_i \log(\hat{y}_i)\right)$$

$$= -\sum_{i=1}^{k} y_i \left(\frac{\partial}{\partial x_j}(\log \hat{y}_i)\right)$$

$$= -\sum_{i=1}^{k} y_i \times \frac{1}{\hat{y}_i} \frac{\partial \hat{y}_i}{\partial x_j}$$

$$= -\sum_{i=1}^{k} \frac{y_i}{\cancel{\hat{y}_i}} \times \left(\cancel{\hat{y}_i}\left(1\{i=j\} - \hat{y}_j\right)\right)$$

$$= -\sum_{i=1}^{k} y_i \left(1\{i=j\} - \hat{y}_j\right)$$

$$= -\sum_{i=1}^{k} y_i \left(1\{i=j\} - \hat{y}_j\right)$$

$$= \sum_{i=1}^{k} y_i \cdot \hat{y}_j - \sum_{i=1}^{k} y_i \cdot 1\{i=j\}$$

$$= \sum_{i=1}^{k} y_i \cdot \hat{y}_j - y_j$$

$$= \hat{y}_j \times \sum_{i=1}^{k} y_i - y_j$$

$$\boxed{\frac{\partial \mathcal{L}_{CCE}}{\partial x_j} = \hat{y}_j - y_j}$$