Machine Learning HW2 Writing

2023.3.1

1. Please explain for the binary classification problem, how can the categorical Cross-Entropy loss (1)

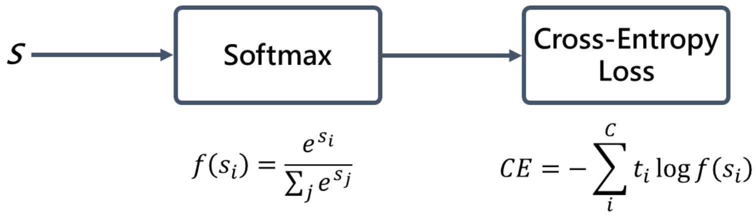
$$CE = -\log\left(\frac{e^{s_p}}{\sum_j^C e^{s_j}}\right)$$

reduce to (2)

$$CE = -t_1 \log(f(s_1)) - (1 - t_1)\log(1 - f(s_1)),$$

where $f(s)$ is the sigmoid function.

Solution



For the binary classification : ($C = 2$)

$$CE = -\sum_i^2 t_i \log f(s_i) = -t_1 \log f(s_1) - t_2 \log f(s_2)$$

$\because t_1 + t_2 = 1\ ;\ f(s_1) + f(s_2) = 1$

$\therefore t_2 = 1 - t_1\ ;\ f(s_2) = 1 - f(s_1)$

$\Rightarrow CE = -t_1 \log f(s_1) - (1 - t_1)\log(1 - f(s_1))$

2. The score of the linear classifier for the i-th sample has the form of $\mathbf{s} = \mathbf{W}^T\mathbf{x}_i$ with $\mathbf{w}_j$ being the j column of $\mathbf{W}$. Let y_i be the label of the i-th sample. The SVM loss function can expressed as

$$L_i = \sum_{j \neq y_i} \max(0, \mathbf{w}_j^T\mathbf{x}_i - \mathbf{w}_{y_i}^T\mathbf{x}_i + 1).$$

Please show that the gradients of L_i w.r.t. $\mathbf{w}_j$ and $\mathbf{w}_{y_i}$ are

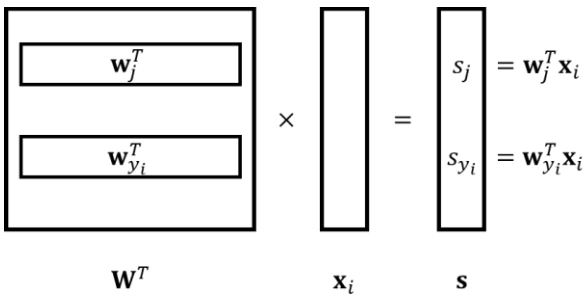
$$\nabla_{\mathbf{w}_j} L_i = 1_{(\mathbf{w}_j^T\mathbf{x}_i - \mathbf{w}_{y_i}^T\mathbf{x}_i + 1 > 0)}\mathbf{x}_i, \tag{3}$$

$$\nabla_{\mathbf{w}_{y_i}} L_i = -\left(\sum_{j \neq y_i} 1_{(\mathbf{w}_j^T\mathbf{x}_i - \mathbf{w}_{y_i}^T\mathbf{x}_i + 1 > 0)}\right)\mathbf{x}_i, \tag{4}$$

where 1_A is the indicator function; $1_A = 1$ if A is true, otherwise $1_A = 0$.

(**Hints**: You can let $\mathbf{w}_j$ and $\mathbf{x}_i$ as scalars to get the same result by using the chain rule. Then, generalize the results to the vector case.

Solution



$$L_i = \sum_{j \neq y_i} \max\left(0, \mathbf{w}_j^T\mathbf{x}_i - \mathbf{w}_{y_i}^T\mathbf{x}_i + 1\right) = \sum_{j \neq y_i} \max\left(0, s_j - s_{y_i} + 1\right)$$

① $\nabla_{\mathbf{w}_j} L_i = \dfrac{\partial L_i}{\partial s_j}\dfrac{\partial s_j}{\partial \mathbf{w}_j} = 1_{(s_j - s_{y_i} + 1 > 0)}\mathbf{x}_i$

Proof

For $j \neq y_i$, if $j = 1$: $\dfrac{\partial L_i}{\partial s_1} = \dfrac{\partial \max\left(0, s_1 - s_{y_i} + 1\right)}{\partial s_1} = 1_{(s_1 - s_{y_i} + 1 > 0)}$

, if $j = 2$: $\dfrac{\partial L_i}{\partial s_2} = \dfrac{\partial \max\left(0, s_2 - s_{y_i} + 1\right)}{\partial s_2} = 1_{\left(s_2 - s_{y_i} + 1 > 0\right)}$

$$\Rightarrow \frac{\partial L_i}{\partial s_j} = 1_{(s_j - s_{y_i} + 1 > 0)}$$

$$\frac{\partial s_j}{\partial \mathbf{w}_j} = \frac{\partial \mathbf{w}_j^T\mathbf{x}_i}{\partial \mathbf{w}_j} = \mathbf{x}_i$$

② $\nabla_{y_i} L_i = \frac{\partial L_i}{\partial s_{y_i}} \frac{\partial s_{y_i}}{\partial \mathbf{w}_{y_i}} = -\sum_{j \neq y_i} 1_{(s_j - s_{y_i} + 1 > 0)} \mathbf{x}_i$

Proof

$$\frac{\partial L_i}{\partial s_{y_i}} = \frac{\partial \sum_{j \neq y_i} \max\left(0, s_j - s_{y_i} + 1\right)}{\partial s_{y_i}} = \sum_{j \neq y_i} 1_{(s_j - s_{y_i} + 1 > 0)} \times (-1)$$

$$\frac{\partial s_j}{\partial \mathbf{w}_{y_i}} = \frac{\partial \mathbf{w}_{y_i}^T \mathbf{x}_i}{\partial \mathbf{w}_{y_i}} = \mathbf{x}_i$$

Note :

$$\frac{\partial \mathbf{w}^T \mathbf{x}}{\partial \mathbf{w}} = \frac{\partial \sum_k w_k x_k}{\partial \mathbf{w}} = \sum_k \frac{\partial w_k x_k}{\partial \mathbf{w}} = \sum_k \frac{\partial w_k}{\partial \mathbf{w}} x_k + \sum_k w_k \frac{\partial x_k}{\partial \mathbf{w}} = \mathbf{x} + \mathbf{0} = \mathbf{x}$$

$$\because \sum_k \frac{\partial w_k}{\partial \mathbf{w}} x_k = \frac{\partial w_1}{\partial \mathbf{w}} x_1 + \frac{\partial w_2}{\partial \mathbf{w}} x_2 + \cdots + \frac{\partial w_k}{\partial \mathbf{w}} x_k = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} x_1 + \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix} x_2 + \cdots + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix} x_k$$

$$= \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix} = \mathbf{x}$$