Assignment 04
Softmax Regression

Nguyen Le Hong Hanh (20C13005)

Question 1: Prove that softmax function map a vector to a probability distribution.

The Softmax function takes an vector of arbitrary real values and produces another vector with real values in range (0, 1). It maps $\mathbb{R}^N \rightarrow \mathbb{R}^N$:

$$\sigma(z_i) = \frac{e^{z_i}}{\sum_{k=1}^{C} e^{z_k}}, \forall i = 1, \ldots N \tag{1}$$

Because of the exponents, the numerator is always positive and the denominator summed up with some other positive numbers, $\sigma(z_i)$ is always positive. Therefore, we can conclude that:

$$\sigma(z_i) = \frac{e^{z_i}}{\sum_{k=1}^{C} e^{z_k}} \leq 1 \tag{2}$$

Moreover, we have:

$$\sum_{i=1}^{N} \frac{e^{z_i}}{\sum_{k=1}^{C} e^{z_k}} = \frac{1}{\sum_{k=1}^{C} e^{z_k}} \sum_{i=1}^{N} e^{z_i} = 1 \tag{3}$$

Therefore, the softmax function maps a vector in $\mathbb{R}^N$ to a probability vector.

Question 2: Find the gradient vector of the loss function in Softmax Regression model.
First, we need to find the gradient of the softmax function:

$$\frac{\partial}{\partial_{z_j}} \sigma_{(z_i)} = \frac{\partial}{\partial_{z_j}} \frac{e^{z_i}}{\sum_{k=1}^{C} e^{z_k}} \tag{4}$$

- if i = j:

$$\begin{aligned}
\frac{\partial}{\partial_{z_j}} \frac{e^{z_i}}{\sum_{k=1}^{C} e^{z_k}} &= \frac{e^{z_i} \sum_{k=1}^{C} e^{z_k} - e^{z_j} e^{z_i}}{[\sum_{k=1}^{C} e^{z_k}]^2} \\
&= \frac{e^{z_i}}{\sum_{k=1}^{C} e^{z_k}} \frac{\sum_{k=1}^{C} e^{z_k} - e^{z_j}}{\sum_{k=1}^{C} e^{z_k}} \\
&= \sigma(z_i)(1 - \sigma(z_j))
\end{aligned} \tag{5}$$

- if i ≠ j:

$$\begin{aligned}
\frac{\partial}{\partial_{z_j}} \frac{e^{z_i}}{\sum_{k=1}^{C} e^{z_k}} &= \frac{0 - e^{z_j} e^{z_i}}{[\sum_{k=1}^{C} e^{z_k}]^2} \\
&= -\frac{e^{z_i}}{\sum_{k=1}^{C} e^{z_k}} \frac{e^{z_j}}{\sum_{k=1}^{C} e^{z_k}} \\
&= -\sigma(z_i)\sigma(z_j)
\end{aligned} \tag{6}$$

Combine the equation (5) and (6), we have:

$$\frac{\partial}{\partial z_j}\sigma(z_i) = \sigma(z_i)(\delta_{ij} - \sigma(z_i)) \tag{7}$$

with $\delta_{ij} = 1$ if i = j, else 0. And, this is also the derivatives of the softmax function.

Second, in the Softmax Regression model, we have $z^{(i)} = W^T x^{(i)} \in \mathbb{R}^C$, with W is the weight matrix $W = [w_1, w_2, ..., w_C]$, and $z^{(i)} = (z_1^{(i)}, z_2^{(i)}, ..., z_C^{(i)})$ is the logits of the i-th sample.

The derivative of the loss function of the softmax regression model:

$$\begin{aligned}\frac{\partial}{\partial w_j^{(i)}} l(y, \hat{y}) &= -\Sigma_{k=1}^{C} \frac{\partial}{\partial w_j^{(i)}} \hat{y}_C log(\sigma(z_k)) \\ &= -\Sigma_{k=1}^{C} \frac{1}{\sigma(z_k)} \hat{y}_k \frac{\partial \sigma(z_k)}{\sigma w_j^{(i)}} \\ &= -\Sigma_{k=1}^{C} \frac{1}{\sigma(z_k)} \hat{y}_k \Sigma_{h=1}^{C} \frac{\partial \sigma(z_k)}{\partial \sigma z_h} \frac{\partial z_k}{\partial w_j^{(i)}} \\ &= -\Sigma_{k=1}^{C} \hat{y}_k \Sigma_{h=1}^{C} [\delta_{k,h} - \sigma(z_h)] \frac{\partial z_k}{\partial w_j^{(i)}}\end{aligned} \tag{8}$$

Since $z_h = \Sigma_{t=1}^{C} W_t^{(h)} x_i$, we have the softmax loss:

$$\frac{\partial}{\partial w_j^{(i)} l(y, \hat{y})} = -\Sigma_{k=1}^{C} \hat{y}_k [\delta_{k,i} - \sigma(z_i)] x_j \tag{9}$$

Reference

1 https://machinelearningcoban.com/2017/02/17/softmax/

2 http://rinterested.github.io/statistics/softmax.html