No.

Date.

Q1.1

let x' be $x+C$, $x_i' = x_i + C$. (x_i' is value in x', so does x_i).

$$\text{softmax}(x_i') = \frac{e^{x_i'}}{\sum_j e^{x_j'}} = \frac{e^{x_i+C}}{\sum_j e^{x_j+C}} = \frac{e^{x_i}}{\sum_j e^{x_j}} = \text{softmax}(x_i)$$

If we use $C = -\max x_i$, we can make sure $x_i' \le 0$, and $e^{x_i'} \in [0,1]$

In this way we can prevent overflow.

Q1.2.

(1). the range of each element is $(0,1)$, the sum is 1.

(2). probability.

(3). The first step calcute the weight of each element, the second step sum the weights and the third step normalize the weights.

Q1.3.

First we express a 2-layer NN into LR.

$$y = W_2(W_1X + b_1) + b_2$$

$$\Leftrightarrow y = W_2W_1X + W_2b_1 + b_2$$

$$\Leftrightarrow y = W'x + b' \quad (W' = W_2W_1, \quad b' = W_2b_1 + b_2).$$

We can recursively reduce a multi-layer NN to LR using the above way.

Q1.4.

$$\sigma'(x) = -\frac{(1+e^{-x})'}{(1+e^{-x})^2} = \frac{e^{-x}}{(1+e^{-x})^2} = \frac{1}{1+e^{-x}} \times \left(1 - \frac{1}{1+e^{-x}}\right) = \sigma(x)(1-\sigma(x))$$

notebook

No.

Date.

Q 1.5.

$$\underset{k\times1}{y} = \underset{k\times d}{W^T}\underset{d\times1}{x} + \underset{k\times1.}{b}. \qquad y_j = \sum_{i=1}^{d} x_i W_{ij} + b_j$$

I'll use the above notation.

① $$\frac{\partial J}{\partial W_{ij}} = \sum_{n=1}^{k} \left(\frac{\partial J}{\partial y_n}\right) \times \left(\frac{\partial y_n}{\partial W_{ij}}\right) = \sum_{n=1}^{k} \delta_n \times \frac{\partial(\sum_{i=1}^{d} x_i W_{in} + b_n)}{\partial W_{ij}}$$

$$= \delta_j \times \frac{\partial(\sum_{i=1}^{d} x_i W_{ij} + b_j)}{\partial W_{ij}} = \delta_j x_i$$

So $$\underset{k\times d}{\frac{\partial J}{\partial W}} = \underset{k\times1}{\delta} \cdot \underset{1\times d.}{x^T}$$

②. $$\frac{\partial J}{\partial x_i} = \sum_{j=1}^{k} \left(\frac{\partial J}{\partial y_j}\right) \times \left(\frac{\partial y_j}{\partial x_i}\right) = \sum_{j=1}^{k} \delta_j \times \frac{\partial \sum_{i=1}^{d} x_i W_{ij} + b_j}{\partial x_i}$$

$$= \sum_{j=1}^{k} \delta_j W_{ij}$$

So $$\underset{1\times d}{\partial J/\partial x} = (\underset{d\times k,\, k\times1.}{W\delta})^T$$

③. $$\frac{\partial J}{\partial b_n} = \sum_{j=1}^{k} \left(\frac{\partial J}{\partial y_j}\right) \times \left(\frac{\partial y_j}{\partial b_n}\right) = \sum_{j=1}^{k} \delta_j \times \frac{\partial \sum_{i=1}^{d} x_i W_{ij} + b_j}{\partial b_n} = \delta_n.$$

So $$\underset{1\times k}{\partial J/\partial b} = \underset{k\times1}{\delta^T}$$

No.

Date.

Q1.6.

①. take a 2-layer NN as example.

$z_1 = W_1 x + b_1 \quad y_1 = \sigma(z_1) \quad z_2 = W_2 y_1 + b_2 \quad y_2 = \sigma(z_2)$

then $\partial y_2 / \partial W_1 = \partial y_2 / \partial z_2 \cdot \partial z_2 / \partial y_1 \cdot \partial y_1 / \partial z_1 \cdot \partial z_1 / \partial W_1$

In the above equation $\partial y_2 / \partial z_2$ and $\partial y_1 / \partial z_1$ follows $\partial y / \partial z = \sigma(z)(1 - \sigma(z))$

When z goes smaller, $\sigma(z)$ will be close to 0, making gradient close to 0.

②. $\tanh \in (-1, 1)$ sigmoid $\in (0, 1)$.

tanh has a better gradient when initializing around 0.

③. When initialized close to 0, tanh has a larger gradient than sigmoid.

④. $\tanh(x) = \frac{1 - e^{-2x}}{1 + e^{-2x}} = \frac{2}{1 + e^{-2x}} - 1 = 2\,\mathrm{sigmoid}(2x) - 1$.

~~Q1.7~~

notebook