softmax 与相关梯度推导

1. 平移不变性证明

$$设：\ y_1 = softmax(z);\ y_2 = softmax(z + c)$$

即证明，$y_1 = y_2$。对于y_1、y_2中的任意元素i:

$$y_1^i = \frac{e^{z_i}}{\sum_{j=1}^{k} e^{z_j}}$$

而

$$y_2^i = \frac{e^{z_{i+c}}}{\sum_{j=1}^{k} e^{z_j+c}} = \frac{e^{z_i}e^{c}}{\sum_{j=1}^{k} e^{z_j}e^{c}} = \frac{e^{z_i}}{\sum_{j=1}^{k} e^{z_j}}$$

因此，对于y_1、y_2中任意元素，均有：

$$y_1^i = y_2^i$$

所以，

$$y_1 = y_2$$

即$softmax(z)$具有平移不变性得证。

2．z_i的梯度$\frac{\partial y}{\partial z_i}$计算

考虑两种情况，

1. $i = j$，当处于该情况时:

$$\frac{\partial y_j}{\partial z_i} = \frac{\partial \frac{e^{z_j}}{\sum_{n=1}^{k} e^{z_n}}}{\partial z_i} = \frac{e^{z_j}\left(\sum_{n=1}^{k} e^{z_n}\right) - e^{z_i}e^{z_j}}{\left(\sum_{n=1}^{k} e^{z_n}\right)^2} = \frac{e^{z_j}}{\left(\sum_{n=1}^{k} e^{z_n}\right)}\frac{\left(\sum_{n=1}^{k} e^{z_n}\right) - e^{z_i}}{\left(\sum_{n=1}^{k} e^{z_n}\right)}$$

$$= y_j(1 - y_i) = y_i(1 - y_i)$$

2. $i \neq j$，当处于该情况时:

$$\frac{\partial y_j}{\partial z_i} = \frac{\partial \frac{e^{z_j}}{\sum_{n=1}^{k} e^{z_n}}}{\partial z_i} = \frac{0 * \left(\sum_{n=1}^{k} e^{z_n}\right) - e^{z_i}e^{z_j}}{(\sum_{n=1}^{k} e^{z_n})^2} = \frac{-e^{z_j}e^{z_i}}{\left(\sum_{n=1}^{k} e^{z_n}\right)} = -y_i y_j$$

因此，最终可得，

$$\frac{\partial y}{\partial z_i} = \left[\frac{\partial y_1}{\partial z_i}, \frac{\partial y_2}{\partial z_i}, \dots, \frac{\partial y_i}{\partial z_i}, \dots, \frac{\partial y_k}{\partial z_i}\right]^T = [-y_1 y_i, -y_2 y_i, \dots, y_i - y_i y_i, -y_k y_i]^T$$

3、各层的后向传播公式

（1）已知$\frac{\partial l}{\partial y}$，推导$\frac{\partial l}{\partial z}$

$$\frac{\partial l}{\partial z}=\frac{\partial l}{\partial y}\frac{\partial y}{\partial z}$$

由上一题结果，有

$$\frac{\partial y}{\partial z}=[\frac{\partial y}{\partial z_1},\frac{\partial y}{\partial z_2},\dots,\frac{\partial y}{\partial z_k}]^T=\begin{bmatrix} y_1-y_1y_1 & -y_1y_2 & \dots & -y_1y_k \\ -y_1y_2 & y_2-y_2y_2 & \dots & -y_2y_k \\ \dots & \dots & \dots & \dots \\ -y_1y_k & -y_2y_k & \dots & y_k-y_ky_k \end{bmatrix}$$

所以，

$$\frac{\partial l}{\partial z}=\frac{\partial l}{\partial y}\frac{\partial y}{\partial z}=\frac{\partial l}{\partial y}\begin{bmatrix} y_1-y_1y_1 & -y_1y_2 & \dots & -y_1y_k \\ -y_1y_2 & y_2-y_2y_2 & \dots & -y_2y_k \\ \dots & \dots & \dots & \dots \\ -y_1y_k & -y_2y_k & \dots & y_k-y_ky_k \end{bmatrix}$$

（2）已知$\frac{\partial l}{\partial z}$，推导$\frac{\partial l}{\partial r}$

$$\frac{\partial l}{\partial r}=\frac{\partial l}{\partial z}\frac{\partial z}{\partial r}=\frac{\partial l}{\partial z}W_2^T$$

（3）已知$\frac{\partial l}{\partial r}$，推导$\frac{\partial l}{\partial p}$

$$\frac{\partial l}{\partial p}=\frac{\partial l}{\partial r}\frac{\partial r}{\partial p}=\frac{\partial l}{\partial r}f(p)$$

其中，

$$f(p)=\begin{cases}1, & if\ p_i>0\\ 0, & if\ p_i\le 0\end{cases}$$

（4）已知$\frac{\partial l}{\partial p}$，推导$\frac{\partial l}{\partial x}$

$$\frac{\partial l}{\partial x}=\frac{\partial l}{\partial p}\frac{\partial p}{\partial x}=\frac{\partial l}{\partial p}W_1^T$$