前向传播

x：$shape=(?,784)$，y：$shape=(?,10)$，w：$shape=(784,10)$，b：$shape=(10,)$

$z=xw+bshape(?,10)$

$$z_i=\sum_j x_j w_{ji}+b_i$$

$a=softmax(z)$：$shape=(?,10)$

$$a_i=softmax(z)=\frac{e^{z_i}}{\sum_j e^{z_j}}$$

一个样本的损失：$loss=-\sum_i y_i loga_i$，其中y_i为真实值，如果为第i个类别则为1，否则为0。

损失函数对z求导：

$$\frac{\partial loss}{\partial z_i}=-(\frac{\partial(y_i*loga_i)}{\partial a_i}\frac{\partial \dot{a}_i}{\partial z_i}+\sum_{j\neq i}\frac{\partial(y_j*loga_j)}{\partial a_j}\frac{\partial \dot{a}_j}{\partial z_i}) \tag{1}$$

softmax函数对z求导：

$$\begin{aligned}\frac{\partial a_i}{\partial z_i}&=\frac{e^{z_i}}{\sum_j e^{z_j}}-\frac{e^{z_i}}{\sum_j e^{z_j}}^2=a_i*(1-a_i)\\ \frac{\partial a_i}{\partial z_j}&=-\frac{e^{z_i}}{\sum_j e^{z_j}{}^2}*e^{z_j}=a_i*a_j,\ j\neq i\end{aligned} \tag{2}$$

将式2代入式1得：

$$\begin{aligned}\frac{\partial loss}{\partial z_i}&=-(\frac{y_i}{ai}*a_i*(1-a_i)+\sum_{j\neq i}\frac{y_j}{a_j}*a_i*a_j)\\ &=-(y_i-y_i*a_i+\sum_{j\neq i}y_j*a_i)=a_i-y_i\end{aligned}$$

对w_{ji},b_i求导，

$$\begin{aligned}\frac{\partial loss}{\partial w_{ji}}&=\frac{\partial loss}{\partial z_i}*\frac{\partial z_i}{\partial w_{ji}}=(a_i-y_i)*x_j\\ \frac{\partial loss}{\partial b_i}&=\frac{\partial loss}{\partial z_i}*\frac{\partial z_i}{\partial b_i}=(a_i-y_i)\end{aligned}$$

x很大，z_i可能值较大，当z_i为100时，造成softmax为inf/inf,导致a_i为nan x归一化后，将w初始化为很小的值，这样避免z_i值过大。实际上，当i不是真实类别是，ai趋向于0，最终造成log以0为底数，此时loss会为nan