So lets see what would happen if doing such a thing.

Softmax和交叉熵的求导过程

Softmax

$$y_i = \frac{e^{a_i}}{\sum_k e^{a_k}}, \quad \text{where } k = 1, \dots, K$$

Softmax的性质（代码时使用）

由于softmax中使用了指数运算，即 e^{a_i} 运算，则在使用softmax时，若 a_i 或 a_k 过大，则在numpy中可能会出现inf的结果。inf是我们不想看到的，因为会丢失很多信息。所以要想办法解决这一问题。

softmax有一条重要性质，即 $\text{softmax}(a_i) = \text{softmax}(a_i + c)$，其中c为任意常数。

$$\frac{e^{a_i+c}}{\sum_k e^{a_k+c}} = \frac{e^{a_i} \cdot e^c}{\sum_k (e^{a_k} \cdot e^c)} = \frac{e^{a_i} \cdot \cancel{e^c}}{\cancel{e^c} \cdot \sum_k e^{a_k}} = \frac{e^{a_i}}{\sum_k e^{a_k}}$$

相等

coding时可以减去最大值以达到优化过程，即 $c = \max(a)$

$\text{softmax}(a_i) = \text{softmax}(a_i - \max(a))$

Softmax的求导过程

$$\frac{\partial y_i}{\partial a_j} = \frac{\partial}{\partial a_j} \cdot \frac{e^{a_i}}{\sum_k e^{a_k}}$$

$e^{a_i} \rightarrow g(a_i) = g_i$

$\sum_k e^{a_k} \rightarrow h(a) = h$

求导公式

$$\boxed{f' = \frac{g_i' \cdot h - g_i \cdot h'}{h^2}}$$

此时会出现两种情况，即当 $i=j$ 和当 $i \neq j$ 时。

① 当 $i=j$ 时：

$$\frac{\partial y_i}{\partial a_j} = \frac{e^{a_i} \cdot \sum_k e^{a_k} - e^{a_i} \cdot e^{a_j}}{\left(\sum_k e^{a_k}\right)^2}$$

$$= \underbrace{\frac{e^{a_i} \cdot \cancel{\sum_k e^{a_k}}}{\left(\sum_k e^{a_k}\right)^{\cancel{2}}}}_{y_i} - \frac{e^{a_i} \cdot e^{a_j}}{\left(\sum_k e^{a_k}\right)^2}$$

$$\frac{(e^{a_j})^2}{\left(\sum_k e^{a_k}\right)^2} = y_j^2$$

$$= y_i - y_j^2$$

$$= y_i(1 - y_j)$$

② 当 $i \neq j$ 时

$$\frac{\partial y_i}{\partial a_j} = \frac{0 \cdot e^{a_j} - e^{a_i} \cdot e^{a_j}}{\left(\sum_k e^{a_k}\right)^2} = -y_i \cdot y_j$$

$e^{a_i} \rightarrow y_i$, $e^{a_j} \rightarrow y_j$

Cross Entropy（交叉熵）的求导过程

什么是交叉熵？

$$L = -\sum_i label_i \cdot \log(y_i)$$

其中，y_i 为 softmax 后预测的结果，
label 为真实结果（ground truth），
log 的底数为 e，为了方便求导。

对于 softmax 的输入 O_k 而言，（这里的 O_k 是为了进行区分）

$$\frac{\partial L}{\partial O_k} = \frac{\partial L}{\partial y_k} \cdot \frac{\partial y_k}{O_k}$$

$$\frac{\partial L}{\partial y_k} = \frac{\partial - \sum_k label_k \cdot \log(y_k)}{\partial y_k}$$

$$= -\sum_k label_k \cdot \frac{\partial \log(y_k)}{\partial y_k}$$

$$= -\sum_k label_k \cdot \frac{1}{y_k}$$

由前面得到的 softmax 求导结果可知

$$\frac{\partial L}{\partial O_k} = -label_i(1-y_i) - \sum_{k\neq i} label_k \cdot \frac{1}{y_k}(-y_k \cdot y_i)$$

$$= -label_i + label_i \cdot y_i + \sum_{k\neq i} label_k \cdot y_i$$

$$= -label_i + y_i\left(label_i + \sum_{k\neq i} label_k\right)$$

$$= -label_i + y_i \boxed{\sum_k label_k}$$

→ label 是一个 one-hot encoding
如 [0, 0, 1, 0]，所以求和结果为 1.

$$= y_i - label_i$$

最终结果是 cross-entropy + softmax 的求导结果.

Signoid 求导

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

$\to u(x) \to u$

$$\sigma'(x) = \frac{\partial \sigma(x)}{\partial x} = \frac{\partial \sigma}{\partial u} \cdot \frac{\partial u}{\partial x}$$

$$= -\frac{1}{u^2} \times (-e^{-x})$$

$$= \frac{e^{-x}}{(1+e^{-x})^2}$$

$\sigma(x)$

$$= \frac{1+e^{-x}-1}{(1+e^{-x})^2}$$

$-\sigma^2(x)$

$$= \sigma(x) - \sigma^2(x)$$

$$= \sigma(x)(1-\sigma(x))$$

$$= \sigma(1-\sigma)$$