METL

TP 1

Tientso Ning

Softmax

1. Define softmax.

Softmax: $\begin{cases} \mathbb{R}^n \rightarrow [0,1]^n \\ x \mapsto \frac{e^x}{\sum_{j=1}^{n} e^{x_j}} \end{cases}$

With $x = (x_1, \ldots, x_n)\,, n \in \mathbb{N}\ \sum_{k=1}^{n} \operatorname{softmax}(x) = 1$

$\operatorname{softmax}(x)_i = \frac{e^{x_i}}{\sum_{i=1}^{n} x^j}$

2. Show softmax is invariant to constants.

Let $x \in \mathbb{R}^n, c \in \mathbb{R}^n, n \in \mathbb{N}$

n = dim(x), where c is a constant vector.

$\operatorname{softmax}(x + c)i = \frac{e^{\left(x_i + c_i\right)}}{\sum_{i=1}^{n} e^{\left(x_i + c_j\right)}}$

$= \frac{e^{x_i} e^{c_i}}{\sum_{i=1}^{n} e^{x_j + c_j}}$

$= \frac{e^{c_i} e^{x_i}}{e^{c_j} \sum_{j=1}^{n} e^{x_j}}$

Since c is a constant.

$= \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_i}}$

$= \operatorname{softmax}(x)_i, \forall i \in [1, n]$

$\forall x \in \mathbb{R}^n, \forall c \in \mathbb{R}^n, \operatorname{softmax}(x + c) = \operatorname{softmax}(x)$

Sigmoid

1. Define sigmoid and gradient.

sigmoid: $\begin{cases} \mathbb{R}^n \rightarrow [-1,1]^n \\ x \mapsto \frac{1}{1+e^{-x}} \end{cases}$ with $x = (x_1, \cdots, x_n)\,, n \in \mathbb{N}\ \delta(x) = \frac{1}{1+e^{-x_i}}$

$$\begin{aligned}
\frac{d}{dx}\delta(x) &= \frac{d}{dx}\left(\frac{1}{1+e^{-x}}\right) \\
&= \frac{d}{dx}\left(1+e^{-x}\right)^{-1} \\
&= \frac{d}{dx}\left(1+e^{-x}\right)^{-1} \\
&= -\left(1+e^{-x}\right)^{-2}\left(-e^{-x}\right) \\
&= \frac{e^{-x}}{\left(1+e^{-x}\right)^2} \\
&= \frac{1}{1+e^{-x}} \cdot \frac{e^{-x}}{1+e^{-x}}
\end{aligned}$$

$= \frac{1}{1+e^{-x}} \cdot \frac{\left(1+e^{-x}\right)-1}{1+e^{-x}} = \frac{1}{1+e^{-x}} \cdot \left(1 - \frac{1}{1+e^{-x}}\right) = \delta(x) \cdot (1 - \delta(x))$

2. Derive gradient with respect to CE loss.

$\frac{\partial CE}{\partial \theta} = \hat{y} - y \; CE = -\sum_{i=1}^{n} y_i \log\left(\hat{y}_i\right)$

We have a one-hot representation, where we have a 1 at the k-th position, turning the sum into just one line (at the k-th line).

$= -y_i \log\left(\hat{y}_i\right) = -\log(\mathrm{softmax}(\theta))$

$\frac{\partial CE}{\partial \theta} = \partial \left(-\log\left(\frac{e^{\theta}}{\sum_{i=1}^{n} e^{\theta_i}}\right)\right)$

$= \partial \left(-\left(\log e^{\theta} - \log \sum e^{\theta_i}\right)\right) = \partial \left(-\theta + \log \sum e^{\theta_i}\right) = \partial \left(\log \sum e^{\theta} - \theta_i\right)$

$= \frac{1}{\sum e^{\theta}} \cdot \partial \left(\sum e^{\theta_i}\right) - y$

$= \frac{e^{\theta_i}}{\sum e^{\theta}} - y = \hat{y} - y$

3. Derive the gradients.

$z_1 = xw_1 + b_1 \; z_2 = hw_2 + b_2$

$\frac{\partial CE}{\partial z_2} = \hat{y} - y$, from the previous derivations.

$\frac{\partial CE}{\partial h} = (\hat{y} - y) \frac{\partial z_2}{\partial h} = (\hat{y} - y) w_2^{\top}$

where $\frac{\partial z_i}{\partial x_i} = W_i^{\top}$ since the relationship is that for each layer, we have the weights multiplied by the inputs with a bias. But the bias doesn't affect the derivative calculation, so we just evaluate the relationship with the weight.

$\frac{\partial CE}{\partial z_1} = (\hat{y} - y) w_2^{T} \frac{\partial h}{\partial z_1} = (\hat{y} - y) w_2^{T} \odot \theta'\left(z_1\right)$ since the layer is the application of the sigmoid, in element-wise fashion.

$\frac{\partial CE}{\partial x} = (\hat{y} - y) \omega_2^{\top} \odot \theta'\left(z_1\right) \frac{\partial z_1}{\partial x}$

$= (\hat{y} - y) \omega_2^{\top} \odot \theta^{2}\left(z_1\right) W_1^{\top}$

4. The number of parameters is `(Dx + 1)H + (H+1)Dy` Since Dx is the dimension of x, and the layer `xW1 + b` requires the bias to be of the same shape H, and HDx is the dimension of xW1. The second part is from the requirement that it matches the output dimensions, since its a one layer NN. Dy corresponds to the output dimension. The bias has to be the same shape Dy, and the layer produces output dimension HDy.