Homework - November 19, 2024

Student: Ayrton Chilibeck, achilibe@ualberta.ca
Lecturer: Lili Mou, UoA.F24.466566@gmail.com

> **Problem 1: Parial Derivative of the Softmax Regression**
>
> If we consider the softmax regression $y = \text{softmax}(Wx + b)$ where $x \in R^d, b, y \in R^k$ Then we know that the cross entropy loss for a single sample is $J = -\sum_{k=1}^{K} t_k \log(y_k)$ where t_k denotes whether or not the sample is in the kth category.
> How can we derive the gradients $\frac{\partial J}{\partial w_{k,i}}$ and $\frac{\partial J}{\partial b_k}$

First, we recall the definition of the softmax function:

$$y_k = \frac{\exp(z_k)}{\sum_{j=1}^{K} \exp(z_j)}$$

where we set $z_k = w_k^\top x + b_k$. Now, recalling the chain rule, we can see that our partial derivatives can be computed as follows:

$$\frac{\partial J}{\partial w_{k,i}} = \frac{\partial J}{\partial y_k}\frac{\partial y_k}{\partial z_k}\frac{\partial z_k}{\partial w_k}$$
$$\frac{\partial J}{\partial b_k} = \frac{\partial J}{\partial y_k}\frac{\partial y_k}{\partial z_k}\frac{\partial z_k}{\partial b_k}$$

Thus we proceed by finding $\frac{\partial J}{\partial z_k}$ as follows:

$$\begin{aligned}
\frac{\partial J}{\partial z_k} &= \frac{\partial J}{\partial y_k}\frac{\partial y_k}{\partial z_k} \\
\frac{\partial J}{\partial y_k} &= \frac{\partial}{\partial y_k} t_k \log(y_k) && \text{generalizes across multiple elements} \\
&= \frac{t_k}{y_k}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial y_k}{\partial z_k} &= \frac{\partial}{\partial z_k}\frac{\exp(z_k)}{\sum_{j=1}^{K} \exp(z_j)} \\
&= \frac{\frac{\partial}{\partial z_k}\exp(z_k) \cdot \left(\sum_{j=1}^{K} \exp(z_j)\right) - \exp(z_k) \cdot \frac{\partial}{\partial z_k}\left(\sum_{j=1}^{K} \exp(z_j)\right)}{\left(\sum_{j=1}^{K} \exp(z_j)\right)^2} \\
&= \frac{\exp(z_k) \cdot \sum_{j=1}^{K} \exp(z_k) - \exp(z_k)\exp(z_k)}{\left(\sum_{j=1}^{K} \exp(z_j)\right)^2} \\
&= \frac{\exp(z_k)\left(\sum_{j=1}^{K} \exp(z_j) - \exp(z_k)\right)}{\left(\sum_{j=1}^{K} \exp(z_j)\right)^2}
\end{aligned}$$

$$
\begin{aligned}
=&y_k(1-y_k) && \text{since } y_k = \frac{\exp(z_k)}{\sum_{j=1}^{K}\exp(z_j)} \text{for } y_k, z_k = && -y_k y_w && \text{for} y_k, z_i \\
=&y_k(\delta - y_i) = y_k - t_k
\end{aligned}
$$

We can then easily find the derivative w.r.t. b and w:

$$\frac{\partial z}{\partial b} = 1 \therefore \frac{\partial J}{\partial b_k} = y_k - t_k$$

$$\frac{\partial z}{\partial w} = x \therefore \frac{\partial J}{\partial w_k} = x(y_k - t_k)$$

> **Problem 2: Partial Derivatives in Matrix form**
>
> Rewrite your solution to the above in matrix form.

We can rewrite both of these gradients in matrix form:

$$
\begin{aligned}
\frac{\partial J}{\partial b} =& Y - \vec{y} \\
\frac{\partial J}{\partial w} =& X^\top (Y - \vec{t})
\end{aligned}
$$