Handin 2 - Neural Nets for Multiclass Classification

Rasmus Freund - 201700273

November 9, 2023

Part I: Derivative

Given a one-hot-label vector y with $y_j = 1$, show that:

$$\frac{\partial L}{\partial z_i} = -\delta_{i,j} + \frac{1}{\sum_{a=1}^{k} e^{z_a}} \times e^{z_i} = -\delta_{i,j} + softmax(z)_i$$

where $\delta_{i,j} = 1$ if $i = j$ and zero otherwise.

Solution

Softmax is defined as:

$$softmax(z)_i = \frac{e^{z_i}}{\sum_{a=1}^{k} e^{z_a}}$$

Therefore, the negative log-likelihood for the true class j is:

$$L(z) = -ln(softmax(z)_j) = -ln\left(\frac{e^{z_j}}{\sum_{a=1}^{k} e^{z_a}}\right)$$

The derivate of $L(z)$ w.r.t. z_i when $i = j$ is then:

$$\frac{\partial L}{\partial z_i} = -\frac{1}{softmax(z)_j} \times \frac{\partial softmax(z)_j}{\partial z_i}$$

Calculating the partial derivative of $softmax(z)_j$ w.r.t. z_i

$$\begin{aligned}
\frac{\partial softmax(z)_j}{\partial z_i} &= \frac{\partial}{\partial z_i}\left(\frac{e^{z_i}}{\sum_{a=1}^{k} e^{z_a}}\right) \\
&= \frac{e^{z_j} \times \sum_{a=1}^{k}(e^{z_a}) - e^{z_j} \times e^{z_i}}{(\sum_{a=1}^{k} e^{z_a})^2} \\
&= \frac{e^{z_j}}{\sum_{a=1}^{k} e^{z_a}} - \left(\frac{e^{z_j}}{\sum_{a=1}^{k} e^{z_a}}\right)^2 \\
&= softmax(z)_j - (softmax(z)_j)^2
\end{aligned}$$

Plugging this back into the original partial derivative

$$\begin{aligned}
\frac{\partial L}{\partial z_i} &= -\frac{1}{softmax(z)_j} \times (softmax(z)_j - (softmax(z)_j)^2) \\
&= -1 + softmax(z)_j
\end{aligned}$$

Since $\delta_{i,j} = 1$ if $i = j$

$$\underline{\underline{\frac{\partial L}{\partial z_i} = -\delta_{i,j} + softmax(z)_i}}$$

Part II: Implementation and test