Noname manuscript No.
(will be inserted by the editor)

Sigsoftmax

Kanai et al's Sigsoftmax properties

TK

the date of receipt and acceptance should be inserted later

Abstract Sigsoftmax properties

Keywords softmax · softmax bottleneck · sigsoftmax

1 Introduction

Kamai et al, NIPS 2018: "Sigsoftmax: Reanalysis of the Softmax Bottleneck"

It is different than having a set of multiple binary outputs and then normalizing them with softmax, as follows.

Sigsoftmax:

$$y_i = \frac{\exp(a_i)\sigma(a_i)}{\sum_{j=1}^{K} \exp(a_j)\sigma(a_j)} \tag{1}$$

Binary with softmax:

$$y_i = \frac{\exp(\sigma(a_i))}{\sum_{j=1}^{K} \exp(\sigma(a_j))} \tag{2}$$

2 Desired properties

"As the alternative function to softmax, a new output function $f(z)$ and its $g(z)$ should have all of the following properties ..."

Address(es) of author(s) should be given

Theorem 5. Sigsoftmax has the following properties:

1. Nonlinearity of $\log(g(a))$: $\log(g(a)) = 2a - \log(1+\exp(a))$.
2. Numerically stable:
$$\frac{\partial \log y_i}{\partial a_j} = \begin{cases} (1-y_j)\cdot(2-\sigma(a_j)) & i=j, \\ -y_j\cdot(2-\sigma(a_j)) & i\neq j. \end{cases} \tag{3}$$
3. Non-negative: $g(a_i) = \exp(a_i)\sigma(a_i) \geq 0$.
4. Monotonically increasing: $a_1 \leq a_2 \Rightarrow \exp(a_1)\sigma(a_1) \leq \exp(a_2)\sigma(a_2)$.

What we have. (1.) Kamai et al's sigsoftmax:

$$g(a) = \exp(a)\cdot\sigma(a) = \frac{\exp(a)}{1+\exp(-a)} = \frac{\exp(2a)}{\exp(a)+1} \tag{4}$$

(2.) Sigmoid:

$$\sigma(a) = \frac{1}{1+\exp(-a)} = \frac{\exp(a)}{\exp(a)+1} \tag{5}$$

(3.) Property 1:

$$\log(g(a)) = 2a - \log(1+\exp(a)) \tag{6}$$

(4.) Normalization:

$$y_i = \frac{g(a_i)}{\sum_k g(a_k)} \tag{7}$$

(5.) Analyse partial derivative of log output:

$$\begin{aligned} \frac{\partial \log(y_i)}{\partial z_j} &= \frac{\partial \log(g(z_i))}{\partial z_j} - \frac{\partial \log(\sum_k g(z_k))}{\partial z_j} \\ &= \frac{1}{g(z_i)}\frac{\partial g(z_i)}{\partial z_j} - \frac{1}{\sum_k g(z_k)}\frac{\partial g(z_j)}{\partial z_j} \\ &= \begin{cases} -\frac{1}{\sum_k g(z_k)}\frac{\partial g(z_j)}{\partial z_j} & ,j\neq i, \\ \left(\frac{1}{g(z_j)} - \frac{1}{\sum_k g(z_k)}\right)\frac{\partial g(z_j)}{\partial z_j} & ,j=i. \end{cases} \\ &= \begin{cases} -\frac{g(z_j)}{\sum_k g(z_k)}\frac{1}{g(z_j)}\frac{\partial g(z_j)}{\partial z_j} & ,j\neq i, \\ \left(\frac{g(z_j)}{g(z_j)} - \frac{g(z_j)}{\sum_k g(z_k)}\right)\frac{1}{g(z_j)}\frac{\partial g(z_j)}{\partial z_j} & ,j=i. \end{cases} \\ &= \begin{cases} -y_j\cdot\frac{\partial \log(g(z_j))}{\partial z_j} & ,j\neq i, \\ (1-y_j)\cdot\frac{\partial \log(g(z_j))}{\partial z_j} & ,j=i \end{cases} \end{aligned} \tag{8}$$

From property 1 (Equation 6) and $\frac{d\log(1+\exp(a))}{da} = \frac{\exp(a)}{1+\exp(a)} = \sigma(a)$, we get:

$$\frac{\partial \log(y_i)}{\partial z_j} = \begin{cases} -y_j\cdot(2-\sigma(a_j)) & ,j\neq i, \\ (1-y_j)\cdot(2-\sigma(a_j)) & ,j=i \end{cases} \tag{9}$$