Sigmoid and Softmax

Suriya Chaudary

18 November 2024

My opinion on how and why sigmoid and softmax materialize.

Sigmoid

Consider the transformation

$$p = \arg\min_{w \in [0,1]} wx \tag{1}$$

where $x \in \mathbf{R}$, $w \in [0,1]$ and $p \in [0,1]$.
p with entropy regularizer

$$\hat{p} = \arg\min_{w \in [0,1]} wx - w\log(w) - (1-w)\log(1-w) \tag{2}$$

Differentiate the objective function with respect to w and equate to 0

$$x - 1 - \log(w) + 1 + \log(1-w) = 0 \tag{3}$$

$$\log(1-w) - \log(w) = -x \tag{4}$$

$$\log\left(\frac{1-w}{w}\right) = -x \tag{5}$$

$$\log\left(\frac{1}{w} - 1\right) = -x \tag{6}$$

$$\frac{1}{w} = 1 + \exp(-x) \tag{7}$$

$$w^{\star} = \frac{1}{1+\exp(-x)} \tag{8}$$

$$\hat{p} = w^{\star} \tag{9}$$

Softmax

Consider the transformation

$$\mathbf{p} = \operatorname*{argmin}_{\mathbf{w} \in \mathbf{S}} \langle \mathbf{w}, \mathbf{x} \rangle \tag{10}$$

where $\mathbf{x} \in \mathbf{R}^d, \mathbf{w} \in \mathbf{S} \subset [0,1]^d$ such that $\mathbf{S} = \left\{\mathbf{w} \mid \sum_i^d \mathbf{w}_i = 1\right\}$, $\mathbf{p} \in [0,1]^d$ and $\sum_{i=1}^d \mathbf{p}_i = 1$.
$\mathbf{p}$ with negative entropy regularizer

$$\hat{\mathbf{p}} = \underset{\mathbf{w}\in\mathbf{S}}{\operatorname{argmin}} \langle \mathbf{w}, \mathbf{x}\rangle + \sum_{i=1}^d \mathbf{w}_i \log(\mathbf{w}_i) \tag{11}$$

Since $\mathbf{w} \in \mathbf{S}$, add a Lagrange multiplier $\lambda(\langle \mathbf{w}, \mathbf{1}\rangle - 1)$ to the objective function.

$$\begin{aligned}
\hat{\mathbf{p}} &= \underset{\mathbf{w}\in\mathbf{S}}{\operatorname{argmin}} \langle \mathbf{w}, \mathbf{x}\rangle + \sum_{i=1}^d \mathbf{w}_i \log(\mathbf{w}_i) + \lambda(\langle \mathbf{w}, \mathbf{1}\rangle - 1) && (12)\\
&= \underset{\mathbf{w}\in\mathbf{S}}{\operatorname{argmin}} \sum_{i=1}^d \mathbf{w}_i \mathbf{x}_i + \sum_{i=1}^d \mathbf{w}_i \log(\mathbf{w}_i) + \lambda\left(\sum_{i=1}^d \mathbf{w}_i - 1\right) && (13)\\
& && (14)
\end{aligned}$$

Differentiate the objective function with respect to $\mathbf{w}_i$ and equate to 0

$$\begin{aligned}
\mathbf{x}_i + 1 + \log(\mathbf{w}_i) + \lambda = 0 && (15)\\
\mathbf{w}_i^\star = \exp(-\mathbf{x}_i)\exp(-1-\lambda) && (16)\\
= \frac{\exp(-\mathbf{x}_i)}{\exp(1+\lambda)} && (17)
\end{aligned}$$

Set λ such that $\sum_i^d \mathbf{w}_i^\star = 1$

$$\begin{aligned}
\mathbf{w}_i^\star &= \frac{\exp(-\mathbf{x}_i)}{\sum_{i=1}^d \exp(-\mathbf{x}_i)} && (18)\\
\hat{\mathbf{p}} &= \mathbf{w}^\star && (19)
\end{aligned}$$

Reference

Luca Trevison. The "Follow-the-Regularized-Leader" algorithm. Topics in computer science and optimization (Fall 2019).