Softmax

Suriya Chaudary

18 October 2024

My opinion on how and why softmax materialize.
Consider the transformation

$$\mathbf{p} = \operatorname*{argmin}_{\mathbf{w} \in \mathbf{S}} \langle \mathbf{w}, \mathbf{x} \rangle \tag{1}$$

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 $\sum_{i=1}^d \mathbf{p}_i = 1$.
$\hat{\mathbf{p}}$ with minimum entropy regularizer

$$\hat{\mathbf{p}} = \operatorname*{argmin}_{\mathbf{w} \in \mathbf{S}} \langle \mathbf{w}, \mathbf{x} \rangle + \sum_{i=1}^{d} \mathbf{w}_i \log(\mathbf{w}_i) \tag{2}$$

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

$$\hat{\mathbf{p}} = \operatorname*{argmin}_{\mathbf{w} \in \mathbf{S}} \langle \mathbf{w}, \mathbf{x} \rangle + \sum_{i=1}^{d} \mathbf{w}_i \log(\mathbf{w}_i) + \lambda (\langle \mathbf{w}, \mathbf{1} \rangle - 1) \tag{3}$$

$$= \operatorname*{argmin}_{\mathbf{w} \in \mathbf{S}} \sum_{i=1}^{d} \mathbf{w}_i \mathbf{x}_i + \sum_{i=1}^{d} \mathbf{w}_i \log(\mathbf{w}_i) + \lambda \left(\sum_{i=1}^{d} \mathbf{w}_i - 1 \right) \tag{4}$$

$$\tag{5}$$

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

$$\mathbf{x}_i + 1 + \log(\mathbf{w}_i) + \lambda = 0 \tag{6}$$

$$\mathbf{w}_i^\star = \exp(-\mathbf{x}_i) \exp(-1 - \lambda) \tag{7}$$

$$= \frac{\exp(-\mathbf{x}_i)}{\exp(1 + \lambda)} \tag{8}$$

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

$$\mathbf{w}_i^\star = \frac{\exp(-\mathbf{x}_i)}{\sum_{i=1}^d \exp(-\mathbf{x}_i)} \tag{9}$$

$$\hat{\mathbf{p}} = \mathbf{w}^\star \tag{10}$$

Reference

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