Softmax

Suriya Chaudary

18 October 2024

My opinion on how and why softmax materializes.
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 entropy regularizer

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

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

$$\begin{aligned}
\hat{\mathbf{p}} &= \operatorname*{argmin}_{\mathbf{w} \in \mathbf{S}} \langle \mathbf{w}, \mathbf{x} \rangle + \sum_{i=1}^{d} \mathbf{w}_i \log\left(\mathbf{w}_i\right) + \lambda\left(\langle \mathbf{w}, \mathbf{1} \rangle - 1\right) && (3) \\
&= \operatorname*{argmin}_{\mathbf{w} \in \mathbf{S}} \sum_{i=1}^{d} \mathbf{w}_i \mathbf{x}_i + \sum_{i=1}^{d} \mathbf{w}_i \log\left(\mathbf{w}_i\right) + \lambda\left(\sum_{i=1}^{d} \mathbf{w}_i - 1\right) && (4) \\
& && (5)
\end{aligned}$$

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

$$\begin{aligned}
\mathbf{x}_i + 1 + \log\left(\mathbf{w}_i\right) + \lambda = 0 && (6) \\
\mathbf{w}_i^\star = \exp\left(-\mathbf{x}_i\right) \exp\left(-1 - \lambda\right) && (7) \\
= \frac{\exp\left(-\mathbf{x}_i\right)}{\exp\left(1 + \lambda\right)} && (8)
\end{aligned}$$

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

$$\begin{aligned}
\mathbf{w}_i^\star &= \frac{\exp\left(-\mathbf{x}_i\right)}{\sum_{i=1}^d \exp\left(-\mathbf{x}_i\right)} && (9) \\
\hat{\mathbf{p}} &= \mathbf{w}^\star && (10)
\end{aligned}$$

Reference

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