From Maximum Entropy to Softmax

Yanhua Huang

Jan 2017

In this post, we will review the principle of maximum entropy, by which we can get some structural assumptions of the Softmax layer.

Consider two random variables x and y with a collection of constraints $f_i(x, y)$, the principle of maximum entropy propose to estimate the posterior probability $\hat{p}(y|x)$, such that the conditional entropy of $\hat{p}(y|x)$ with the prior $p(x)$ is maximum. Furthermore, the constraints $\mathbb{E}_{p(x,y)}[f_i] = \mathbb{E}_{\hat{p}(y|x)p(x)}[f_i]$ and $\sum_y \hat{p}(y|x) = 1$ must be satisfied.

With Lagrange multipliers, we can get the following min-max problem

$$\min_{\hat{p}} \max_{\alpha_i,\beta} \mathcal{L} = \min_{\hat{p}} \max_{\alpha_i,\beta} \sum_{x,y} p(x)\hat{p} \log \hat{p} + \sum_i \alpha_i \sum_{x,y} (\hat{p}p(x) - p(x, y))f_i + \beta(1 - \sum_y \hat{p}). \tag{1}$$

With the solution of its duel problem

$$\frac{\partial \mathcal{L}}{\partial \hat{p}} = \sum_{x,y} p(x)(\log \hat{p} + 1) + \sum_i \alpha_i \sum_{x,y} p(x)f_i = 0, \tag{2}$$

we can get

$$\hat{p}(y|x) = \frac{1}{Z} \exp^{-\sum_i \alpha_i f_i}, \tag{3}$$

where Z is the normalization factor. It points out that the logits learned by the neural network are the linear combinations of joint constraints.