Jinchao Xu

Deep Learning Algorithms and Analysis

Summer 2020

Contributors:

This set of notes are based on contributions from many of graduate students, post-doctoral fellows and other collaborators. Here is a partial list:

Juncai He, Qingguo Hong, Li Jiang.....

Contents

0.1 KL divergence, cross-entropy and logistic regression

KL divergence and cross-entropy

Given two discrete probability distributions,

$$p = \begin{pmatrix} p_1 \\ \vdots \\ p_k \end{pmatrix}, q = \begin{pmatrix} q_1 \\ \vdots \\ q_k \end{pmatrix}$$

namely $0 \le p_i, q_i \le 1$ and $\sum_{i=1}^k p_i = \sum_{i=1}^k q_i = 1$. The KL divergence defines a special distance between p and q:

$$D_{\mathrm{KL}}(q, p) = \sum_{i=1}^{k} q_i \log \frac{q_i}{p_i}. \tag{0.1}$$

$D_{\mathrm{KL}}(q, p)$ works like a "distance" without the symmetry:

Lemma 1.

1. $D_{\mathrm{KL}}(q, p) \ge 0$;
2. $D_{\mathrm{KL}}(q, p) = 0$ *if and only if* $p = q$;

Proof. We first note that the elementart inequality

$$\log x \le x - 1, \quad \text{for any } x \ge 0, \tag{0.2}$$

and the equality holds if and only if $x = 1$.

$$-D_{\mathrm{KL}}(q,p) = -\sum_{i=1}^{c} q_i \log \frac{q_i}{p_i} = \sum_{i=1}^{k} q_i \log \frac{p_i}{q_i} \leq \sum_{i=1}^{k} q_i (\frac{p_i}{q_i} - 1) = 0. \tag{0.3}$$

And the equality holds if and only if

$$\frac{p_i}{q_i} = 1 \quad \forall i = 1 : k. \tag{0.4}$$

□

Note that

$$D_{\mathrm{KL}}(q,p) = \sum_{i=1}^{k} q_i \log \frac{q_i}{p_i} = \sum_{i=1}^{k} q_i \log q_i - \sum_{i=1}^{k} q_i \log p_i \tag{0.5}$$

We write

$$H(q,p) = H(q) + D_{\mathrm{KL}}(q,p), \tag{0.6}$$

where

$$H(q) = -\sum_{i=1}^{k} q_i \log q_i, \tag{0.7}$$

which is called entropy for distribution p and

$$H(q,p) = -\sum_{i=1}^{k} q_i \log p_i. \tag{0.8}$$

which is called cross-entropy for distribution p and q.

It follows from the relation (0.6) that

$$\arg\min_{p} D_{\mathrm{KL}}(q,p) = \arg\min_{p} H(q,p). \tag{0.9}$$

Cross-entropy

In §0.1 we introduced the concept of cross-entropy, which can be used to define a loss function in machine learning and optimization. Let us assume y_i is the true label for x_i, for example $y_i = e_{k_i}$ if $x_i \in A_{k_i}$. Then ,consider the predicted distribution

$$\boldsymbol{p}(x;\boldsymbol{\theta}) = \frac{1}{\sum_{i=1}^{k} e^{w_i x + b_i}} \begin{pmatrix} e^{w_1 x + b_1} \\ e^{w_2 x + b_2} \\ \vdots \\ e^{w_k x + b_k} \end{pmatrix} = \begin{pmatrix} p_1(x;\boldsymbol{\theta}) \\ p_2(x;\boldsymbol{\theta}) \\ \vdots \\ p_k(x;\boldsymbol{\theta}) \end{pmatrix} \tag{0.10}$$

for any data $x \in A$. By (0.9), we have

$$\arg\min_{\theta} \sum_{i=1}^{N} D_{\mathrm{KL}}(y_i, \boldsymbol{p}(x_i;\boldsymbol{\theta})) = \arg\min_{\theta} \sum_{i=1}^{N} H(y_i, \boldsymbol{p}(x_i;\boldsymbol{\theta})), \tag{0.11}$$

Recall that we have all data $D = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$. Then, it is natural to consider the loss function as following:

$$\sum_{j=1}^{N} H(y_i, \boldsymbol{p}(x_i;\boldsymbol{\theta})), \tag{0.12}$$

which measures the distance between the real label and predicted one for all data. In the meantime, we can check that

$$\begin{aligned}
\sum_{j=1}^{N} H(y_j, \boldsymbol{p}(x_j;\boldsymbol{\theta})) &= -\sum_{j=1}^{N} y_j \cdot \log \boldsymbol{p}(x_j;\boldsymbol{\theta}) \\
&= -\sum_{j=1}^{N} \log p_{i_j}(x_i;\boldsymbol{\theta}) \quad \text{(because } y_j = e_{i_j} \text{ for } x_j \in A_{i_j}\text{)} \\
&= -\sum_{i=1}^{k} \sum_{x\in A_i} \log p_i(x;\boldsymbol{\theta}) \\
&= -\log \prod_{i=1}^{k} \prod_{x\in A_i} p_i(x;\boldsymbol{\theta}) \\
&= L(\theta)
\end{aligned} \tag{0.13}$$

That is to say, the logistic regression loss function defined by likelihood is exact the loss function defined by measuring the distance between real label and predicted one via cross-entropy. Or we can note as

$$\min_{\boldsymbol{\theta}} L_\lambda(\boldsymbol{\theta}) \Leftrightarrow \min_{\boldsymbol{\theta}} \sum_{j=1}^{N} H(y_i, \boldsymbol{p}(x_i;\boldsymbol{\theta})) + \lambda R(\|\boldsymbol{\theta}\|) \Leftrightarrow \min_{\boldsymbol{\theta}} \sum_{j=1}^{N} D_{\mathrm{KL}}(y_i, \boldsymbol{p}(x_i;\boldsymbol{\theta})) + \lambda R(\|\boldsymbol{\theta}\|). \tag{0.14}$$