Kullback Matching and Maximum Likelihood Estimation

Seungjin Choi
Department of Computer Science
POSTECH, Korea

Outline

This note shows that maximum likelihood estimation is identical to the minimization of Kullback-Leibler divergence between the empirical distribution and model distribtuion.

Details

The Kullback-Leibler divergence (KL-divergence) is a popular measure for a similarity between two probability distributions, defined by

$$KL[p||q] = \int p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} d\boldsymbol{x}, \tag{1}$$

where $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$ are probability distributions.

Let denote by $\tilde{p}(\boldsymbol{x})$ and $p(\boldsymbol{x}|\boldsymbol{\theta})$, the empirical distribution and model distribution, respectively. Two fundamental properties of KL-divergence are:

- $KL[p||q] \geq 0$(Gibb's inequality) with equality holding if and only if $p(\boldsymbol{x}) = q(\boldsymbol{x})$.
- The KL-divergence is asymmetric, i.e., $KL[p||q] \neq KL[q||p]$.

Given a set of data points, $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ drawn from the underlying distribution $p(\boldsymbol{x})$, let $\tilde{p}(\boldsymbol{x})$ be the empirical distribution which puts probability $\frac{1}{N}$ on each data point, leading to

$$\tilde{p}(\boldsymbol{x}) = \frac{1}{N} \sum_{t=1}^{N} \delta(\boldsymbol{x} - \boldsymbol{x}_t). \tag{2}$$

We consider the KL-divergence from the empirical distribution $\tilde{p}(\boldsymbol{x})$ to the model distribution $p(\boldsymbol{x}|\boldsymbol{\theta})$

$$\begin{aligned} KL[\tilde{p}(\boldsymbol{x})||p(\boldsymbol{x}|\boldsymbol{\theta})] &= \int \tilde{p}(\boldsymbol{x}) \log \frac{\tilde{p}(\boldsymbol{x})}{p(\boldsymbol{x}|\boldsymbol{\theta})} d\boldsymbol{x} \\ &= -H(\tilde{p}) - \int \tilde{p}(\boldsymbol{x}) \log p(\boldsymbol{x}|\boldsymbol{\theta}) d\boldsymbol{x}, \end{aligned} \tag{3}$$

where $H(\tilde{p}) = -\int \tilde{p}(\boldsymbol{x}) \log \tilde{p}(\boldsymbol{x}) d\boldsymbol{x}$ is the entropy of $\tilde{p}$.
It follows from (3) that

$$\arg\min_{\boldsymbol{\theta}} KL[\tilde{p}(\boldsymbol{x})||p(\boldsymbol{x}|\boldsymbol{\theta})] \equiv \arg\max_{\boldsymbol{\theta}} \langle \log p(\boldsymbol{x}|\boldsymbol{\theta}) \rangle_{\tilde{p}} \, , \tag{4}$$

where $\langle \cdot \rangle_{\tilde{p}}$ represents the expectation with respect to the distribution $\tilde{p}$. Plugging (2) into the righthand side of (**??**, leads to

$$\begin{aligned} \langle \log p(\boldsymbol{x}|\boldsymbol{\theta}) \rangle_{\tilde{p}} &= \frac{1}{N} \int \sum_{t=1} N\delta(\boldsymbol{x} - \boldsymbol{x}_t) \log p(\boldsymbol{x}|\boldsymbol{\theta}) d\boldsymbol{x} \\ &= \frac{1}{N} \sum_{t=1}^{N} \log p(\boldsymbol{x}_t|\boldsymbol{\theta}). \end{aligned} \tag{5}$$

Apart from the scaling factor $\frac{1}{N}$, this is just the log-likelihood function. In other words, maximum likelihood estimation is obtained from the minimization of (3) (*Kullback matching*).