Maximum Likelihood Estimation and Kullback-Leibler Divergence

Albert Dorador

UW-Madison

March 5, 2025

Introduction

- Can view the negative log-likelihood $-\sum_{i=1}^{n} \log p_\theta(x_i)$ as the sum of loss functions measuring the loss incurred when using p_θ to model the true distribution of x_i, which is given by q.
- Risk: the expected value of a loss function. The risk of a model using the negative log-likelihood as loss function is

$$R(q, p_\theta) = \mathbb{E}_{x\sim q}(-\log p_\theta(x)) = -\int q(x) \log p_\theta(x)\, dx$$

- Excess risk: the risk of a given model in excess of that of the true model. In our context,

$$R(q, p_\theta) - R(q, q) = \mathbb{E}_{x\sim q}\left(\log \frac{q(x)}{p_\theta(x)}\right) = \int q(x) \log \frac{q(x)}{p_\theta(x)}\, dx$$

  Note that this is precisely the KL divergence of p_θ from q, termed $D(q||p_\theta)$. As such, the excess risk is always non-negative.

Minimum KL divergence and MLE

- Consider $\theta^* \in \arg\min_\theta D(q||p_\theta)$
- The density p_{θ^*} is the member of the parametric family of distributions p_θ that is closest in KL divergence to the true data-generating distribution q.
- Does MLE get you θ^*? In general, no, but it gets arbitrarily close asymptotically (under mild assumptions).

Convergence of MLE to minimum-KL distribution

- TL;DR: Finding the MLE is finding the parameter value that brings the assumed distribution p_θ the closest to the true distribution q based on a set of i.i.d. samples drawn from q. As the number of samples increases, the MLE tends to the parameter value that is optimal in the KL sense.
- Note: there is no guarantee that p_{θ^*} will match q even if you got infinite samples (why?). What we do have is

$$\arg\max_\theta \prod_{i=1}^n p_\theta(x_i) = \arg\min_\theta -\sum_{i=1}^n \log p_\theta(x_i)$$

$$= \arg\min_\theta \sum_{i=1}^n \log q(x_i) - \log p_\theta(x_i) = \arg\min_\theta \sum_{i=1}^n \log \frac{q(x_i)}{p_\theta(x_i)}$$

Now, by the Strong Law of Large Numbers (SLLN)*, for any $\theta \in \Theta$,

$$\frac{1}{n}\sum_{i=1}^n \log \frac{q(x_i)}{p_\theta(x_i)} \to D(q||p_\theta) \quad w.p.1 \text{ as } n \to \infty$$

Caveat

- Note*: $\hat{\theta}_n$ is a random variable that depends on the sample $\{x_1, \ldots, x_n\}$ observed, so the independence assumption in the standard SLLN does not hold. However, convergence still holds under mild regularity conditions on the likelihood function. For more details, see Aad W van der Vaart and Jon A Wellner. Weak convergence and empirical processes with applications to statistics. Journal of the Royal Statistical Society-Series A Statistics in Society, 160(3):596–608, 1997.