February 27, 2021

2.15 MLE minimizes KL Divergence to the Empirical Distribution

I learned a lot about probability and statistics from this question!

We are concerned with the value of $KL(p_{emp}||q(\theta))$.

First of all, what is the empirical distribution? The empirical distribution function is the distribution function associated with the empirical measure of a sample. For example, if we had a dataset sample from a Gaussian distribution, we could estimate the mean and the variance. Those estimates are parameters of the empirical Gaussian distribution. FYI, the empirical distribution function converges with probability 1 to the underlying distribution under the Gilvenko-Cantelli theorem as the dataset size grows.

$$\begin{aligned} argmin_\theta KL(p_{emp}||q) &= argmin_\theta \mathbb{E}_{x\sim p_{emp}}[\log p_{emp} - \log q(x;\theta)] \\ &= argmax_\theta \mathbb{E}_{x\sim p_{emp}}[\log q(x;\theta)] \end{aligned}$$

Let X be your dataset of size n. I.e, $X = \{x_1, ..., x_n\}$. Then the MLE is

$$\hat{\theta} = argmax_\theta q(X;\theta) \tag{1}$$

Where q is the likelihood of your data. In particular, q may encode dependencies in your data. If the data is iid, then q can be decomposed into products of the likelihoods of individual data points, and then the MLE is exactly the right hand side of our KL expression, as the empirical distribution converges to the real distribution in the limit of the sample size.

Combined with the fact that the KL divergence is always non-negative, we've made an argument that the MLE converges in KL distance to the true distribution in the limit of sample size.