15.

Intro

In this question, we will prove that the MLE (maximum likelihood estimator) minimizes KL divergence between some model and the empirical distribuion. Before the solution, is important to understand what this means. First of all, let's give some context. Suppose we get access to some data $\mathcal{D}$ generated by some unknown distribution and we are asked to estimate this distribution based on the data. In this situation, there is two approaches we could take: the nonparametric and the parametric. On the former approach, we could make an estimation based purely on the data we have at hands. This estimation could be the empirical distribution $p_{emp}(x)$, which was defined in this chapter of the book. On the latter approach, we could make some reasonable assumptions about the data pattern and create a paSound familiar?rametric model $q(x|\theta)$ for it. The model is based on the prior that we might have about the process behind the data and it dictates the general shape of the distribution(e.g. gaussian). In this case, we only need to estimate the parameters of our distribution (μ and σ for a gaussian). One of the most popular ways (for several reasons) of estimating the parameters is the MLE. The MLE is the parameter value θ^* which solve the following optimization problem:

$$\theta^* = \arg\max_{\theta} \prod_{x \in \mathcal{D}} p(x|\theta) \tag{1}$$

The right side of (1) is called the likelihood of the dataset and it is a function of θ. Is worth noting that most of the time it is most convenient to optimize the log-likelihood, which gives the same result as optimizing the likelihood.

$$\theta^* = \arg\max_{\theta} \prod_{x \in \mathcal{D}} p(x|\theta) = \arg\max_{\theta} \sum_{x \in \mathcal{D}} log(p(x|\theta)) \tag{2}$$

Given two such different scenarios, this exercise propose that we find a bridge between them. We will find the parametric model $q(x|\theta)$ that is the most similar to the empirical distribution $p_{emp}(x)$. As you can suspect, one great way to do this is to express the similarity by the KL divergence and minimize this value. So, let's begin the solution.

Solution

$$\begin{gathered} KL(p_{emp}||q(\theta)) = \sum p_{emp} log\left(\frac{p_{emp}}{q}\right) = \\ \sum p_{emp} log(p_{emp}) - \sum p_{emp} log(q) \end{gathered} \tag{3}$$

Since p_{emp} is fixed given the dataset $\mathcal{D}$, the fist term of (3) is a constant and we only need to focus on the second term.

$$\begin{gathered} \arg\min_{\theta} KL(p_{emp}||q(\theta)) = \arg\min_{\theta} -\sum p_{emp} log(q) = \arg\max_{\theta} \sum log\left(q^{p_{emp}}\right) = \\ \arg\max_{\theta} log\left(\prod_{x \in \mathcal{D}} q^{p_{emp}}\right) \end{gathered} \tag{4}$$

For the final step, note that the empirical distribution is defined as $p_{emp}(x_i) = \frac{N_i}{N}$, where N_i is the number of occurences of x_i and N is the size of our dataset. Thus:

$$\begin{aligned} \arg\min_{\theta} KL(p_{emp}||q(\theta)) = \arg\max_{\theta} log\left(\prod_{x\in\mathcal{D}} q^{p_{emp}}\right) = \\ \arg\max_{\theta} log\left(\prod_{x\in\mathcal{D}} q^{\frac{N_i}{N}}\right) = \\ \arg\max_{\theta} \frac{1}{N} log\left(\prod_{x\in\mathcal{D}} q(x_i|\theta)^{N_i}\right) \end{aligned} \tag{5}$$

This is the same expression of the log-likelihood, with exception of the $\frac{1}{N}$ term. Therefore, the solution for this problem is the MLE parameter θ^*.

Conclusion

In this exercise, we saw that the MLE of a parametric model is the most similar distribution to the empirical distribuion, among all models considered ($\theta \in \Theta$). This make a lot of sense if you think about it. As the name tells, the maximum likelihood estimator is the parameter that maximizes the likelihood that our given dataset could occur. So, its optimazition is strongly based on the data (do not take into account a probability prior, like Bayesian models. The only prior is the shape of the function, like a gaussian, which is given by the model). Therefore, as both the MLE and p_{emp} have such a strong relationship with the empirical data, it makes sense that they have a relationship among themselves.