Kullback-Leibler Divergence

2019年5月9日 11:50

Good essays about Kullback-Leibler Divergence:

Kullback-Leibler Divergence Explained: https://www.countbayesie.com/blog/2017/5/9/kullback-leibler-divergence-explained

如何理解K-L散度（相对熵）: https://www.jianshu.com/p/43318a3dc715

Intuitive Guide to Understanding KL Divergence: https://towardsdatascience.com/light-on-math-machine-learning-intuitive-guide-to-understanding-kl-divergence-2b382ca2b2a8

KL-divergence

Kullback-Leibler Divergence Definition

To measure the difference between two probability distributions over the same variable x, a measure, called the *Kullback-Leibler divergence*, or simply, the *KL divergence*, has been popularly used in the data mining literature. The concept was originated in probability theory and information theory.

The KL divergence, which is closely related to *relative entropy*, *information divergence*, and *information for discrimination*, is a non-symmetric measure of the difference between two probability distributions $p(x)$ and $q(x)$. Specifically, the Kullback-Leibler (KL) divergence of $q(x)$ from $p(x)$, denoted $D_{KL}(p(x), q(x))$, is a measure of the information lost when $q(x)$ is used to approximate $p(x)$.

Let $p(x)$ and $q(x)$ are two probability distributions of a discrete random variable x. That is, both $p(x)$ and $q(x)$ sum up to 1, and $p(x) > 0$ and $q(x) > 0$ for any x in X. $D_{KL}(p(x), q(x))$ is defined in Equation (2.1).

$$D_{KL}(p(x)||q(x)) = \sum_{x \in X} p(x) \ln \frac{p(x)}{q(x)} \tag{2.1}$$

The entropy of our distribution

KL Divergence has its origins in information theory. The primary goal of information theory is to quantify how much information is in data. The most important metric in information theory is called Entropy, typically denoted as H. The definition of Entropy for a probability distribution is:

$$H = -\sum_{i=1}^{N} p(x_i) \cdot \log p(x_i)$$

If we use log_2 for our calculation we can interpret entropy as "the minimum number of bits it would take us to encode our information". In this case, the information would be each observation of teeth counts given our empirical distribution. Given the data that we have observed, our probability distribution has an entropy of 3.12 bits. The number of bits tells us the lower bound for how many bits we would need, on average, to encode the number of teeth we would observe in a single case.

What entropy doesn't tell us is the optimal encoding scheme to help us achieve this compression. Optimal encoding of information is a very interesting topic, but not necessary for understanding KL divergence. The key thing with Entropy is that, simply knowing the theoretical lower bound on the number of bits we need, we have a way to quantify exactly how much information is in our data. Now that we can quantify this, we want to quantify how much information is lost when we substitute our observed distribution for a parameterized approximation.

Measuring information lost using Kullback-Leibler Divergence

Kullback-Leibler Divergence is just a slight modification of our formula for entropy. Rather than just having our probability distribution p we add in our approximating distribution q. Then we look at the difference of the log values for each:

$$D_{KL}(p||q) = \sum_{i=1}^{N} p(x_i) \cdot (\log p(x_i) - \log q(x_i))$$

Essentially, what we're looking at with the KL divergence is the expectation of the log difference between the probability of data in the original distribution with the approximating distribution. Again, if we think in terms of log_2 we can interpret this as "how many bits of information we expect to lose". We could rewrite our formula in terms of expectation:

$$D_{KL}(p||q) = E[\log p(x) - \log q(x)]$$

The more common way to see KL divergence written is as follows:

$$D_{KL}(p||q) = \sum_{i=1}^{N} p(x_i) \cdot log\frac{p(x_i)}{q(x_i)}$$

since $\log a - \log \mathrm{b} = \log \frac{a}{b}$.

With KL divergence we can calculate exactly how much information is lost when we approximate one distribution with another. Let's go back to our data and see what the results look like.