In mathematical statistics, the Kullback–Leibler divergence (also called relative entropy) is a measure of how one probability distribution is different from a second, reference probability 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

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

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

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

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

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

With KL divergence we can calculate exactly how much information is lost when we approximate one distribution with another.

Now we can go ahead and calculate the KL divergence for our two approximating distributions. For the uniform distribution we find:

$$D_{kl}(\text{Observed} \,||\, \text{Uniform}) = 0.338$$

And for our Binomial approximation:

$$D_{kl}(\text{Observed} \,||\, \text{Binomial}) = 0.477$$

As we can see the *information lost* by using the Binomial approximation is greater than using the uniform approximation. If we have to choose one to represent our observations, we're better off sticking with the Uniform approximation.

Cross entropy and KL divergence

$$D_{KL}(P||Q) = \sum_i P(i) log \frac{P(i)}{Q(i)}$$

$$= \sum_i P(i) log P(i) - \sum_i P(i) log Q(i)$$

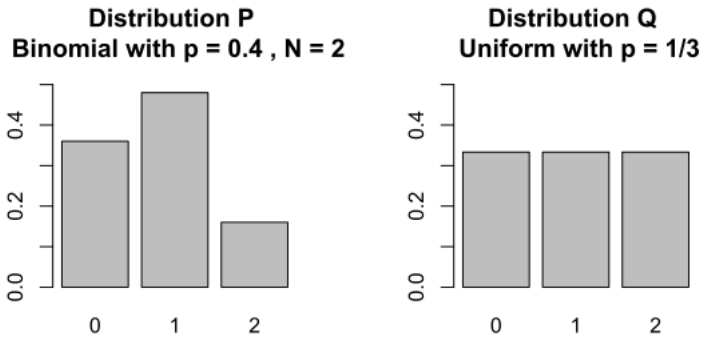
$$= -H(P) + H(P, Q)$$

The cross entropy is a combination of the entropy of the 'true' distribution P and the KL divergence between P and Q:

$$H(p, q) = H(p) + D_{KL}(p \parallel q)$$

Basic example

x	0	1	2
Distribution *P*(x)	0.36	0.48	0.16
Distribution *Q*(x)	0.333	0.333	0.333



The KL divergences $D_{KL}(p||q)$ and $D_{KL}(q||p)$ are calculated as follows. This example uses the natural log with base e, designated ln to get results in nats (see units of information).

$$\begin{aligned} D_{\mathrm{KL}}(P \parallel Q) &= \sum_{x \in \mathcal{X}} P(x) \ln\left(\frac{P(x)}{Q(x)}\right) \\ &= 0.36 \ln\left(\frac{0.36}{0.333}\right) + 0.48 \ln\left(\frac{0.48}{0.333}\right) + 0.16 \ln\left(\frac{0.16}{0.333}\right) \\ &= 0.0852996 \end{aligned}$$

$$
\begin{aligned}
D_{\mathrm{KL}}(Q \parallel P) &= \sum_{x \in \mathcal{X}} Q(x) \ln\left(\frac{Q(x)}{P(x)}\right) \\
&= 0.333 \ln\left(\frac{0.333}{0.36}\right) + 0.333 \ln\left(\frac{0.333}{0.48}\right) + 0.333 \ln\left(\frac{0.333}{0.16}\right) \\
&= 0.097455
\end{aligned}
$$

References

- Wikipedia https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence (https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence)
- Kullback-Leibler Divergence Explained https://www.countbayesie.com/blog/2017/5/9/kullback-leibler-divergence-explained (https://www.countbayesie.com/blog/2017/5/9/kullback-leibler-divergence-explained)
- An introduction to entropy, cross entropy and KL divergence in machine learning https://adventuresinmachinelearning.com/cross-entropy-kl-divergence/ (https://adventuresinmachinelearning.com/cross-entropy-kl-divergence/)
- 정보이론2편,: KL-Divergence, https://brunch.co.kr/@chris-song/69 (https://brunch.co.kr/@chris-song/69)