The KL Divergence: From Information to Density Estimation

The KL divergence, also known as "relative entropy", is a commonly used metric for density estimation. I re-derive the relationships between probabilities, entropy, and relative entropy for quantifying similarity between distributions.

PUBLISHED

22 January 2019

In statistics, the Kullback–Leibler (KL) divergence is a metric for how similar two probability distributions are. A standard formulation—and the one I encountered first—is the following. Given two probability distributions P and Q, the KL divergence is the integral

$$D_{\mathrm{KL}}[P\|Q] = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} dx \tag{1}$$

In this post, I want to show how Eq. 1 is both a measure of *relative information entropy* and a reasonable way to compare densities.

Relative information

Let us reconstruct the notion of *information entropy* (abbr. *information*) from first principles. The *information* received from a random variable taking a particular value can be viewed as a measure of our "surprise" about that value. A value with low information is not surprising; a value with high information is. Now imagine you didn't know or couldn't remember the equation for information. What would be a sensible formulation? First, it makes sense that information is a monotonic function of probability. Higher probability means strictly lower information. For example, rolling a die and getting an even number should be less surprising than rolling a die and getting a 2.

It would also make sense that information should be additive for independent events. If I roll a die and flip a coin, my total information should be some additive combination of the two probabilities. This thought exercise is useful because, at least for me, it makes clear *why* the information about a random variable X taking on a value x, denoted $h(x)$, is defined as it is:

$$h(x) = -\log p(x) \qquad H(X) = \mathbb{E}[h(x)]$$

The negative sign is because higher probabilities result in less information. And the $\log$ is a monotonically increasing function of probabilities that has the useful property that two independent random events have *additive information*:

$$\begin{aligned} h(x,y) &= -\log p(x,y) \\ &= -\log\{p(x)p(y)\} \\ &= -\log p(x) - \log p(y) \\ &= h(x) + h(y) \end{aligned}$$

What does that mean for the KL divergence? Let's consider Eq. 1 again, but now write it in terms of information:

$$\begin{aligned} D_{\mathrm{KL}}[P\|Q] &= \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} dx \\ &= \mathbb{E}_{p(x)}\big[\log \frac{p(x)}{q(x)}\big] \\ &= \mathbb{E}_{p(x)}\big[\log p(x) - \log q(x)\big] \\ &= \mathbb{E}_{p(x)}\big[-\log q(x)\big] - \mathbb{E}_{p(x)}\big[-\log p(x)\big] \\ &= H(Q) - H(P) \end{aligned} \tag{2}$$

In other words, one interpretation of the KL divergence is that it captures the *relative information* or *relative entropy* between two distributions P and Q. Also note that the KL divergence is not symmetric, i.e. $D_{\mathrm{KL}}[P\|Q] \neq D_{\mathrm{KL}}[Q\|P]$ in general.

At this point, it makes sense that the KL divergence might be a good metric for understanding how similar two distributions are. But why does *minimizing* the KL divergence between two densities—as in variational inference—guarantee that our optimization objective is performing density estimation? The answer to this relies on the convexity of logarithms and Jensen's inequality.

Nonnegativity of the KL divergence

First, the big picture. We want to use the notion of convexity to prove Jensen's inequality. Jensen's inequality will allow us to move the logarithm in Eq. 1 outside the integral. Since the integral of a density is 1, the log of the integral is 0. This will provide a lower bound on the KL divergence or formally: $D_{\mathrm{KL}} \geq 0$ with equality when $p(x) = q(x)$. With that in mind, let's move forward.

A function f is *convex* if the following holds

$$f(\lambda a + (1-\lambda)b) \leq \lambda f(a) + (1-\lambda) f(b)$$

for some $0 \leq \lambda \leq 1$. This is a common formulation, and the reader can find numerous explanations and visualizations for why this is true. Intuitively, the function of any point between a and b inclusive is less than or equal to any point between $f(a)$ and $f(b)$. Draw a few functions on a piece of paper and see which ones are convex.

An aside: proof of Jensen's inequality

But at this point, I think many explanations of the KL divergence skip a step. They say something like, "And by Jensen's inequality…" without *proving* Jensen's inequality. Let's actually do that. Jensen's inequality is

$$f\Big(\sum_{i=1}^{N} \lambda_i x_i\Big) \leq \sum_{i=1}^{N} \lambda_i f(x_i) \tag{3}$$

where $\lambda_i \geq 0$ and $\sum_i \lambda_i = 1$. The proof is by induction. Let f be a convex function. Now consider the base case:

$$f(\lambda_1 x_1 + \lambda_2 x_2) \leq \lambda_x f(x_1) + \lambda_2 f(x_2)$$

This clearly holds because f is convex and $x_1 + x_2 = 1 \iff (1 - x_1) = x_2$. Now for the inductive case, we want to show that

$$f\Big(\sum_{i=1}^{K} \lambda_i x_i\Big) \leq \sum_{i=1}^{K} \lambda_i f(x_i) \implies f\Big(\sum_{i=1}^{K+1} \lambda_i x_i\Big) \leq \sum_{i=1}^{K+1} \lambda_i f(x_i)$$

First, let's start with our inductive hypothesis and add $\lambda_{K+1} f(x_{K+1})$ to both sides:

$$f\Big(\sum_{i=1}^{K} \lambda_i x_i\Big) + \lambda_{K+1} f(x_{K+1}) \leq \sum_{i=1}^{K+1} \lambda_i f(x_i)$$

Now the λs on the right-hand-side no longer sum to 1. Let's normalize both sides of the equation by multiplying by $\frac{1}{1+\lambda_{K+1}}$:

$$\overbrace{\frac{1}{1+\lambda_{K+1}}}^{A} f\Big(\sum_{i=1}^{K} \lambda_i x_i\Big) + \overbrace{\frac{\lambda_{K+1}}{1+\lambda_{K+1}}}^{B} f(x_{K+1}) \leq \frac{1}{1+\lambda_{K+1}} \sum_{i=1}^{K+1} \lambda_i f(x_i)$$

This normalization constant makes sense because $\sum_{i=1}^{K} \lambda_i = 1 \iff \sum_{i=1}^{K+1} \lambda_i = 1 + \lambda_{K+1}$. Now note that the terms labeled A and B above sum to 1. And since f is convex, we can say

$$f\Big(\frac{1}{1+\lambda_{K+1}} \sum_{i=1}^{K} \lambda_i x_i + \frac{\lambda_{K+1}}{1+\lambda_{K+1}} x_{K+1}\Big) \leq \frac{1}{1+\lambda_{K+1}} f\Big(\sum_{i=1}^{K} \lambda_i x_i\Big) + \frac{\lambda_{K+1}}{1+\lambda_{K+1}} f(x_{K+1})$$

At this point, we're basically done. The left-hand-side of the above inequality can be simplified to

$$f\Big(\frac{1}{1+\lambda_{K+1}} \sum_{i=1}^{K+1} \lambda_i x_i\Big)$$

which we have already shown is less than or equal to $\frac{1}{1+\lambda_{K+1}} \sum_{i=1}^{K+1} \lambda_i f(x_i)$ as desired.

Jensen's inequality for distributions

Now consider this: since $\lambda_i \geq 0$ and $\sum_i \lambda_i = 1$, we can interpret λ_i as the probability of our random variable X taking on a specific value x_i, giving us

$$f\big(\mathbb{E}[X]\big) \leq \mathbb{E}\big[f(X)\big]$$

which for continuous densities is equivalent to

$$f\Big(\int xp(x)dx\Big) \leq \int f(x)p(x)dx$$

Since logarithms are convex functions, we can apply Jensen's inequality to the KL divergence to prove a lower bound:

$$\begin{aligned} D_{\mathrm{KL}}[P\|Q] &= \int_{-\infty}^{\infty} p(x)\log\frac{p(x)}{q(x)}dx \\ &= -\int_{-\infty}^{\infty} p(x)\log\frac{q(x)}{p(x)}dx \\ &\geq -\log\int_{-\infty}^{\infty} q(x)dx \\ &= 0 \end{aligned}$$

We first flip the fraction so that the $p(x)$ terms cancel, then apply Jensen's inequality, and finally use the fact that $\log(1) = 0$.

Conclusion

In summary, we have used convexity to prove Jensen's inequality to prove that the KL divergence is always nonnegative. If we minimize the KL divergence between two densities, we are minimizing the relative information between the two distributions.