Building Intuition about K-L Divergence

September 22, 2015

1 Definition

The Kullback-Leibler divergence[1] of Q from P is defined as

$$D_{KL}(P\|Q) = \int P(x) \left[\log P(x) - \log Q(x)\right] dx,$$

where P and Q are probability densities with respect to a common measure. (In the case of counting measure, this makes P and Q be probability mass functions.) This is also an expectation over P of a function involving logs of densities:

$$D_{KL}(P\|Q) = \mathbb{E}_P(x \mapsto \log P(x) - \log Q(x)).$$

2 Interpretations

For discrete distributions, KL divergence can be interpreted as the expected extra bits needed to code elements of P with an optimal code for Q, over what would be needed for an optimal code for P.

Freer, Mansinghka, and Roy [2] prove that the KL divergence also gives the performance of a rejection sampler. Specifically,

Proposition 1 (Freer, Mansinghka, Roy, pg. 2)**.** *Let N be the number of attempts before a rejection sampler for P with proposal distribution Q succeeds. N is geometrically distributed with mean* $\exp(D_{KL}(P\|Q))$.

[1] `https://en.wikipedia.org/wiki/Kullback-Leibler_divergence`

[2] `danroy.org/papers/FreerManRoy-NIPSMC-2010.pdf` When are probabilistic programs probably computationally tractable?

This corresponds with the general intuition that proposal distributions should try to be broader than their targets rather than narrower.

3 Properties

KL divergnce is

- Non-negative, and zero only if $P = Q$ almost everywhere
- Invariant to parameterization
- *Not* symmetric
- Does *not* obey the triangle inequality

For non-negativity, consider the function whose expected value is the divergence: $x \mapsto \log P(x) - \log Q(x)$. This function will be (large and) positive in areas that are (much) more likely under P than under Q, and (large and) negative in areas that are (much) more likely under Q than under P. But since the expectation is being taken with respect to P, it should stand to reason that the behavior in areas likely under P dominates.

For parameterization invariance, we reproduce the derivation from Wikipedia.

Proof. If a transformation is made from variable x to variable $y(x)$, then, since $P(x)dx = P(y)dy$ and $Q(x)dx = Q(y)dy$ the KL divergence may be rewritten

$$\begin{aligned} D_{KL}(P\|Q) &= \int P(x) \log\left(\frac{P(x)}{Q(x)}\right) dx \\ &= \int P(y) \log\left(\frac{P(y)dy/dx}{Q(y)dy/dx}\right) dy \\ &= \int P(y) \log\left(\frac{P(y)}{Q(y)}\right) dy \end{aligned}$$

□

4 Example: KL of 1-D Gaussians

As a refresher, the probability density function of the Gaussian distribution with mean μ and standard deviation σ is

$$\mathcal{N}(\mu, \sigma)(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),$$

and its log is correspondingly

$$\log \mathcal{N}(\mu, \sigma)(x) = -\frac{(x-\mu)^2}{2\sigma^2} - \log\sigma - \log\sqrt{2\pi}.$$

Let us examine the behavior of the KL divergence of one Gaussian distribution Q from another, P. By parameterization invariance, we can normalize P to be the standard Gaussian, which we will denote $\mathcal{N}$, dropping the mean and deviation arguments. Then we have

$$\begin{aligned}
D_{KL}(\mathcal{N}\|\mathcal{N}(\mu,\sigma)) &= \mathbb{E}_{\mathcal{N}}\left[\log\mathcal{N}(x) - \log\mathcal{N}(\mu,\sigma)(x)\right] \\
&= \mathbb{E}_{\mathcal{N}}\left[-\frac{x^2}{2} + \frac{(x-\mu)^2}{2\sigma^2} + \log\sigma\right] \\
&= -\frac{1}{2}\mathbb{E}_{\mathcal{N}}[x^2] + \frac{1}{2\sigma^2}\mathbb{E}_{\mathcal{N}}[x^2] - \frac{2\mu}{2\sigma^2}\mathbb{E}_{\mathcal{N}}(x) + \frac{\mu^2}{2\sigma^2} + \log\sigma \\
&= \frac{\mu^2+1}{2\sigma^2} + \log\sigma - \frac{1}{2}.
\end{aligned}$$

where we used that $\mathbb{E}_{\mathcal{N}}(x) = 0$ and $\mathbb{E}_{\mathcal{N}}(x^2) = 1$ are the mean and variance of the standard Gaussian distribution. Using translation and scaling, we get for the general case

$$D_{KL}(\mathcal{N}(\mu_P,\sigma_P)\|\mathcal{N}(\mu_Q,\sigma_Q)) = \frac{(\mu_Q-\mu_P)^2 + \sigma_P^2}{2\sigma_Q^2} + \log\sigma_Q - \log\sigma_P - \frac{1}{2}.$$

Some interpretation:

- If σ_Q is large relative to the other quantities, the $\log\sigma_Q$ term will dominate, so the divergence of a broad Gaussian from a narrow one grows logarithmically.

- For fixed standard deviations that are small relative to the mean difference, the KL divergence of two Gaussians is quadratic in said mean difference.
- If σ_P is large relative to the other quantities, the KL divergence is quadratic in it.

5 Mixture distributions

Consider the definition of KL divergence again

$$D_{KL}(P\|Q) = \mathbb{E}_P(x \mapsto \log P(x) - \log Q(x)),$$

this time in the context of P and Q being mixture distributions

$$P(x) = \sum_i w_i P_i(x) \qquad Q(x) = \sum_j w_j Q_j(x) \qquad \sum_i w_i = \sum_j w_j = 1.$$

If the mixture components P_i and widely separated, one term will dominate each sum. Call its index $i(x)$. Then

$$D_{KL}(P\|Q) \approx \mathbb{E}_P \left[\log P_{i(x)}(x) + \log w_{i(x)} - \log Q(x)\right].$$

The approximateness comes from the smaller terms in each sum, which can be viewed as bumping up the weight of the best component a bit if there is another component that's amost as good.[3] Moreover, the expectation over P is the weighted sum of expectations over P_i, and we may assume that for x drawn from P_i, P_i stands to dominate the density function. This brings us to

$$\begin{aligned} D_{KL}(P\|Q) &\approx \sum_i w_i \log w_i + \sum_i w_i \mathbb{E}_{P_i}\left[\log P_i(x) - \log Q(x)\right] \\ &= \sum_i w_i D_{KL}(P_i\|Q) - H(w_i). \end{aligned}$$

This is the weighted sum of divergences of Q from the mixture components P_i, less the entropy of the mixture weights w_i.

If the mixture components Q_j are also widely separated, the same reasoning applies to the $\log Q(x)$ term. If we further assume that for samples

[3]For instance, if $P_1(x) = P_2(x) \gg P_i(x)$, then $\log P(x) \approx \log P_1(x) + \log w_1 + \log w_2$.

from any given component P_i, the contribution of a single component $Q_{j(i)}$ always dominates, we find

$$\begin{aligned} D_{KL}(P\|Q) &\approx \sum_i w_i \log w_i + \sum_i w_i \mathbb{E}_{P_i}\left[\log P_i(x) - \log Q_{j(i)}(x) - \log w_{j(i)}\right] \\ &= \sum_i w_i D_{KL}(P_i\|Q_{j(i)}) - H(w_i) + H(w_i, w_{j(i)}), \end{aligned}$$

which is the weighted (by the weights in P) sum of the KL divergences of corresponding mixture components, less the entropy of the weights of P, plus the cross-entropy of the weights $w_{j(i)}$ with respect to the w_i.[4]

6 Computing KL divergence

KL divergence of Q from P is an expectation over P of a function determined by the density functions of P and Q. Therefore, if P is samplable and P and Q are assessable (resp. approximately assessable), we can approximate the KL (resp. the KL of approximate distributions) by Monte Carlo integration:

$$\begin{aligned} x_i &\sim P \qquad N \text{ samples from } P \\ \widehat{D_{KL}} &= \frac{1}{N}\sum_i \log P(x_i) - \log Q(x_i). \end{aligned}$$

The Central Limit Theorem implies that if the random variable obtained by computing $\log P(x_i) - \log Q(x_i)$ for x_i sampled from P has finite mean M and variance Σ^2, then as N rises the distribution of the above estimate will converge to the Gaussian $\mathcal{N}(M, \Sigma/\sqrt{N})$. We can therefore estimate the error of any particular computed mean as the sample variance of $\{\log P(x_i) - \log Q(x_i)\}_1^N$.

Let's check this for the Gaussian case. From Section 4, the true mean M is

$$M = D_{KL}(\mathcal{N}\|\mathcal{N}(\mu,\sigma)) = \frac{\mu^2+1}{2\sigma^2} + \log\sigma - \frac{1}{2}.$$

The variance of our estimator is

[4] Well, not exactly, because the $w_{j(i)}$ needn't sum to 1, but the notation serves as a mnemonic.

$$\begin{aligned}
\Sigma^2 &= \mathbb{E}_{\mathcal{N}}\left[\log \mathcal{N}(x) - \log \mathcal{N}(\mu,\sigma)(x) - M\right]^2 \\
&= \mathbb{E}_{\mathcal{N}}\left[-\frac{x^2}{2} + \frac{(x-\mu)^2}{2\sigma^2} + \log\sigma - M\right]^2.
\end{aligned}$$

To perform this calculation, we are going to first expand both squares and group the powers of x. Letting

$$\begin{aligned}
A &= -\frac{1}{2} + \frac{1}{2\sigma^2} \\
B &= \frac{\mu}{\sigma^2} \\
C &= \frac{\mu^2}{2\sigma^2} + \log\sigma
\end{aligned}$$

we have

$$\begin{aligned}
M &= A + C \\
\Sigma^2 &= \mathbb{E}_{\mathcal{N}}\left[Ax^2 - Bx + C - (A+C)\right]^2 \\
&= \mathbb{E}_{\mathcal{N}}\left[A^2x^4 - 2ABx^3 + \left(-2A^2 + B^2\right)x^2 + 2ABx + A^2\right] \\
&= 3A^2 - 2A^2 + B^2 + A^2 \\
&= 2A^2 + B^2 \\
&= \frac{1}{2} - \frac{1}{\sigma^2} + \frac{1+2\mu^2}{2\sigma^4},
\end{aligned}$$

where line 4 relies on the known values for the moments of the standard Gaussian distribution: $\mathbb{E}_{\mathcal{N}}(1) = 1$, $\mathbb{E}_{\mathcal{N}}(x) = 0$, $\mathbb{E}_{\mathcal{N}}(x^2) = 1$, $\mathbb{E}_{\mathcal{N}}(x^3) = 0$, $\mathbb{E}_{\mathcal{N}}(x^4) = 3$.

I read the above formulas as good news and bad news. The good news is that the variance of our KL estimator is finite for all divergences of non-degenerate Gaussians, so the CLT applies and the estimate will eventually converge to having a predictable error. The bad news is that this may take a *long* time: the variance is quadratic in the difference of means (when that difference begins to exceed the standard deviation of P), and *quartic* in the ratio of standard deviations. Comparing this to the formula for the mean estimate itself, getting down to 10% relative error means taking $O(100/\sigma^2)$ samples in the small-σ regime.