Technical Notes on Kullback-Leibler Divergence

Alexander Etz

Sept. 5, 2018
(updated Jan. 2, 2019)

Kullback-Leibler Divergence

The Kullback-Leibler divergence between the true data-generating distribution for random variable X, say $p_1(x)$, and another possible candidate distribution, say $p_2(x)$, is the expected value of the log likelihood ratio in favor of the true model. That is, if $p_1(x)$ is the true model, then

$$\begin{aligned} KL(p_1||p_2) &= \int_{\mathcal{X}} p_1(x) \log \frac{p_1(x)}{p_2(x)} dx & (1) \\ &= \mathbb{E}_{p_1}\left[\log \frac{p_1(x)}{p_2(x)}\right] & (2) \end{aligned}$$

where the $\mathcal{X}$ in the first line is the sample space of the random variable X (i.e., the possible values it can take); if X is discrete then the integral is replaced with a sum.

The log likelihood ratio can be interpreted as the amount of evidence the data provide for one model versus another, so the KL divergence tells us how much evidence we can expect our data to provide in favor of the true model. See Etz (2018) for a refresher on likelihoods and likelihood ratios.

The integral in (1) might be rather complicated, and if we try to derive the KL divergence using brute force integration/summation it can get a little hairy. The representation in (2) makes our life a lot easier, because for many common distributions it reduces the bulk of the derivation to some simple algebra.

There aren't too many derivations for commonly used Kullback-Leibler divergences online, so I have written up some of my notes below.

0.1 Bernoulli

The bernoulli distribution is useful when we want to model a trial with a binary outcome and a certain probability of success. Note: The binomial distribution is a trial consisting of n independent replicates of a single bernoulli trial.

Let $X \sim Bern(\theta)$, $x \in \{0, 1\}$, $0 < \theta < 1$. Then

$$p_\theta(x) = \theta^x (1-\theta)^{1-x},$$

and $\mathbb{E}[X] = \theta$. Then the log likelihood ratio (LLR) in favor of a bernoulli distribution with $\theta = \theta_1$ versus one with $\theta = \theta_2$ is[1]

$$\begin{aligned} LLR(X) &= \log \frac{p_{\theta_1}(X)}{p_{\theta_2}(X)} \\ &= \log \left[\frac{\theta_1^X (1-\theta_1)^{1-X}}{\theta_2^X (1-\theta_2)^{1-X}} \right] \\ &= X \log \left[\frac{\theta_1}{\theta_2} \right] + (1-X) \log \left[\frac{1-\theta_1}{1-\theta_2} \right]. \end{aligned}$$

To get to the Kullback-Leibler divergence we need to take the expectation of this function when truly $\theta = \theta_1$. The log likelihood ratio above is a linear function of X, so to take its expectation we can simply replace X with $\mathbb{E}[X] = \theta_1$, giving

$$\begin{aligned} KL\left(p_{\theta_1}(X) || p_{\theta_2}(X)\right) &= \mathbb{E}_{\theta_1}\left(LLR(X)\right) \\ &= \mathbb{E}[X] \log \left[\frac{\theta_1}{\theta_2} \right] + (1-\mathbb{E}[X]) \log \left[\frac{1-\theta_1}{1-\theta_2} \right] \\ &= \theta_1 \log \left[\frac{\theta_1}{\theta_2} \right] + (1-\theta_1) \log \left[\frac{1-\theta_1}{1-\theta_2} \right]. \end{aligned}$$

Note: If $X_1, X_2, \ldots, X_n$ are independent bernoulli trials with probaility of success θ, then the random variable $Y = \sum X_i$ follows a binomial distribution with probability of success θ and sample size n. In this case the KL divergence for the binomial trial is simply n times the KL divergence of a single bernoulli trial.

0.2 Geometric

The geometric distribution is useful when we want to know the failure rate of a process using a design that continues collecting data until the first failure occurs.

Let $X \sim Geo(\theta)$, $x \in \{0, 1, \ldots\}$, $0 < \theta < 1$. Then

$$p_\theta(x) = \theta(1-\theta)^x,$$

and $\mathbb{E}[X] = \frac{1-\theta}{\theta}$. Then the log likelihood ratio between θ_1 and θ_2 is

$$\begin{aligned} LLR(X) &= \log \frac{p_{\theta_1}(X)}{p_{\theta_2}(X)} \\ &= \log \left[\frac{\theta_1(1-\theta_1)^X}{\theta_2(1-\theta_2)^X} \right] \\ &= \log \left[\frac{\theta_1}{\theta_2} \right] + X \log \left[\frac{1-\theta_1}{1-\theta_2} \right]. \end{aligned}$$

[1] From now on I'll just use the shorthand "the log likelihood ratio between θ_1 and θ_2" but remember we are really talking about the *distributions* indexed by those parameters.

To get to the KL divergence we need the expectation of this function when truly $\theta = \theta_1$. This is again a linear function of X, so we again simply replace X with its expectation, giving

$$KL(p_{\theta_1}||p_{\theta_2}) = \log\left[\frac{\theta_1}{\theta_2}\right] + \left(\frac{1-\theta_1}{\theta_1}\right)\log\left[\frac{1-\theta_1}{1-\theta_2}\right]$$

Note: If $X_1, X_2, \ldots, X_n$ are independent geometric trials with probability of success θ, then the random variable $Y = \sum X_i$ follows a negative binomial distribution with probability of success θ. In this case the KL divergence for a negative binomial trial is simply n times the KL divergence of a single geometric trial.

0.3 Poisson

Let $X \sim Pois(\lambda)$, $x \in \{0, 1, \ldots\}$, $\lambda > 0$. Then

$$p_\lambda(x) = \lambda^x e^{-\lambda}/x!$$

and $\mathbb{E}[X] = \lambda$. Then the log likelihood ratio between λ_1 and λ_2 is

$$\begin{aligned} LLR(X) &= \log\frac{p_{\lambda_1}(X)}{p_{\lambda_2}(X)} \\ &= \log\left[\frac{\lambda_1^X e^{-\lambda_1}/X!}{\lambda_2^X e^{-\lambda_2}/X!}\right] \\ &= X\log\left[\frac{\lambda_1}{\lambda_2}\right] - (\lambda_1 - \lambda_2). \end{aligned}$$

To get to the KL divergence we need the expectation of this function when truly $\lambda = \lambda_1$. Again we have a linear function of X, so we just replace X with its expectation, giving

$$KL(p_{\lambda_1}||p_{\lambda_2}) = \lambda_1\log\left[\frac{\lambda_1}{\lambda_2}\right] - (\lambda_1 - \lambda_2)$$

0.4 Exponential

Let $X \sim Exp(\theta)$, $x > 0$, $\theta > 0$. Then

$$p_\theta(x) = \theta e^{-\theta x},$$

and $\mathbb{E}[X] = 1/\theta$. Then the log likelihood ratio between θ_1 and θ_2 is

$$\begin{aligned} LLR(X) &= \log \frac{p_{\theta_1}}{p_{\theta_2}} \\ &= \log \left[\frac{\theta_1 e^{-\theta_1 X}}{\theta_2 e^{-\theta_2 X}} \right] \\ &= \log \left[\frac{\theta_1}{\theta_2} \right] - X\left(\theta_1 - \theta_2\right). \end{aligned}$$

To get to the KL divergence we need the expectation of this function when truly $\theta = \theta_1$. Again we have a linear function of X, so we just replace X with its expectation, giving

$$KL(p_{\theta_1} || p_{\theta_2}) = \log \left[\frac{\theta_1}{\theta_2} \right] - \frac{\theta_1 - \theta_2}{\theta_1}.$$

0.5 Normal (part 1)

Let $X \sim N(\mu, \sigma^2)$, $-\infty < x < \infty$, $-\infty < \mu < \infty$, $\sigma^2 > 0$. Then if we let $\theta = (\mu, \sigma^2)$, we have

$$p_\theta(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2} \right],$$

and $\mathbb{E}[X] = \mu$. Then the log likelihood ratio between two normal distributions with different means, μ_1 versus μ_2, but the same variance σ^2, is

$$\begin{aligned} LLR(X) &= \log \frac{p_{\mu_1(X)}}{p_{\mu_2}(X)} \\ &= \log \frac{1/\sqrt{2\pi\sigma^2} \exp\left[-\frac{(x-\mu_1)^2}{2\sigma^2} \right]}{1/\sqrt{2\pi\sigma^2} \exp\left[-\frac{(x-\mu_2)^2}{2\sigma^2} \right]} \\ &= -\frac{1}{2\sigma^2} \left[(X-\mu_1)^2 - (X-\mu_2)^2 \right]. \end{aligned}$$

Let's define $Y = X - \mu_1$ and $\delta = \mu_2 - \mu_1$. Then we have

$$\begin{aligned} LLR(X) &= -\frac{1}{2\sigma^2} \left[Y^2 - (Y-\delta)^2 \right] \\ &= -\frac{1}{2\sigma^2} \left[Y^2 - Y^2 + 2Y\delta - \delta^2 \right] \\ &= -\frac{Y\delta}{\sigma^2} + \frac{\delta^2}{2\sigma^2}, \end{aligned}$$

and if we change back to our original variables we obtain

$$LLR(X) = -\frac{(X-\mu_1)(\mu_2-\mu_1)}{\sigma^2} + \frac{(\mu_2-\mu_1)^2}{2\sigma^2}.$$

Again we have a linear function of X. Taking the expectation when truly $\mu = \mu_1$, the first term becomes zero, and thus we obtain

$$KL(p_{\mu_1}||p_{\mu_2}) = \frac{(\mu_2 - \mu_1)^2}{2\sigma^2}.$$

If we write $\Delta = (\mu_2 - \mu_1)/\sigma$ for the standardized mean difference, we can see that the KL divergence is $\Delta^2/2$, i.e., half the squared standardized mean difference between the two distributions.

0.6 Normal (part 2)

Let $X \sim N(\mu, \sigma^2)$ as before. The log likelihood ratio between two normal distributions with different means, μ_1 versus μ_2, and different variances, σ_1^2 versus σ_2^2, is

$$\begin{aligned} LLR(X) &= \log \frac{1/\sqrt{2\pi\sigma_1^2} \exp\left[-\frac{(X-\mu_1)^2}{2\sigma_1^2}\right]}{1/\sqrt{2\pi\sigma_2^2} \exp\left[-\frac{(X-\mu_2)^2}{2\sigma_2^2}\right]} \\ &= \frac{1}{2}\log\left[\frac{\sigma_2^2}{\sigma_1^2}\right] - \left[\frac{(X-\mu_1)^2}{2\sigma_1^2} - \frac{(X-\mu_2)^2}{2\sigma_2^2}\right]. \end{aligned}$$

Let's again define $Y = X - \mu_1$ and $\delta = \mu_2 - \mu_1$, and let's also define $\tau_i = 1/\sigma_i^2$, $i = 1, 2$. Then

$$\begin{aligned} LLR(X) &= \frac{1}{2}\log\left[\frac{\tau_1}{\tau_2}\right] - \left[\frac{1}{2}\tau_1 Y^2 - \frac{1}{2}\tau_2 (Y-\delta)^2\right] \\ &= \frac{1}{2}\log\left[\frac{\tau_1}{\tau_2}\right] - \frac{1}{2}\left[\tau_1 Y^2 - \tau_2 Y^2 + 2\tau_2 Y\delta - \tau_2\delta^2\right] \\ &= \frac{1}{2}\log\left[\frac{\tau_1}{\tau_2}\right] - \frac{1}{2}(\tau_1 - \tau_2)Y^2 - \tau_2 Y\delta + \frac{1}{2}\tau_2\delta^2. \end{aligned}$$

We need to find the expectation of the above function to obtain the KL divergence. Unfortunately, it is not just a linear function of Y, but of Y^2 as well. Recall that in general $\mathbb{E}[Y^2] = Var[Y] + \mathbb{E}[Y]^2$. When $\mu = \mu_1$ and $\sigma^2 = \sigma_1^2$, $\mathbb{E}[Y] = \mathbb{E}[X - \mu_1] = 0$ and $Var[Y] = Var[X] = \sigma_1^2 = 1/\tau_1$. Thus,

$$\begin{aligned} KL(p_{\theta_1}||p_{\theta_2}) &= \frac{1}{2}\log\left[\frac{\tau_1}{\tau_2}\right] - \frac{(\tau_1 - \tau_2)}{2\tau_1} + \frac{1}{2}\tau_2\delta^2 \\ &= \frac{1}{2}\log\left[\frac{\tau_1}{\tau_2}\right] - \frac{1}{2} + \frac{\tau_2}{2\tau_1} + \frac{1}{2}\tau_2\delta^2 \\ &= \frac{1}{2}\left(\log\left[\frac{\sigma_2^2}{\sigma_1^2}\right] + \frac{\sigma_1^2}{\sigma_2^2} + \frac{(\mu_2 - \mu_1)^2}{\sigma_2^2} - 1\right). \end{aligned}$$

Note that the KL divergence for normal distributions with the same variance but different means (section 0.5) is a special case of the above result, where

$\sigma_1^2 = \sigma_2^2 = \sigma^2$. Likewise, the KL divergence for two normal distributions with the same mean but different variances is also a special case of the above where $\mu_1 = \mu_2 = \mu$.

References

Etz, A. (2018). Introduction to the concept of likelihood and its applications. *Advances in Methods and Practices in Psychological Science*, 1(1):60–69.