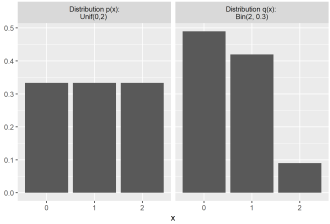
Introduction to Machine Learning

Kullback-Leibler Divergence



x	0	1	2
Distribution p(x)	0.3333333	0.3333333	0.3333333
Distribution q(x)	0.4900000	0.4200000	0.0900000

Learning goals

- Know the KL divergence as distance between distributions
- Understand KL as expected log-difference
- Understand how KL can be used as loss
- Understand that KL is equivalent to the expected likelihood ratio

KULLBACK-LEIBLER DIVERGENCE

We now want to establish a measure of distance between (discrete or continuous) distributions with the same support:

$$D_{KL}(p\|q) = \mathbb{E}_p\left[\log\frac{p(X)}{q(X)}\right] = \sum_{x\in\mathcal{X}} p(x)\cdot\log\frac{p(x)}{q(x)},$$

or:

$$D_{KL}(p\|q) = \mathbb{E}_p\left[\log\frac{p(X)}{q(X)}\right] = \int_{x\in\mathcal{X}} p(x)\cdot\log\frac{p(x)}{q(x)}.$$

In the above definition, we use the convention that $0\log(0/0) = 0$ and the convention (based on continuity arguments) that $0\log(0/q) = 0$ and $p\log(p/0) = \infty$. Thus, if there is any symbol $x \in \mathcal{X}$ such that $p(x) > 0$ and $q(x) = 0$, then $D_{KL}(p\|q) = \infty$.

KULLBACK-LEIBLER DIVERGENCE

$$D_{KL}(p\|q) = \mathbb{E}_p\left[\log \frac{p(X)}{q(X)}\right]$$

- What is the intuition behind this formula?
- We will soon see that KL has quite some value in measuring "differences" but is not a true distance.
- We already see that the formula is not symmetric and it often makes sense to think of *p* as the first or original form of the data, and *q* as something that we want to measure the quality of with reference to *p*.

INFORMATION INEQUALITY

$D_{KL}(p\|q) \geq 0$ holds always true for any pair of distributions, and holds with equality if and only if $p = q$.

We use Jensen's inequality. Let A be the support of p:

$$
\begin{aligned}
-D_{KL}(p\|q) &= -\sum_{x \in A} p(x) \log \frac{p(x)}{q(x)} \\
&= \sum_{x \in A} p(x) \log \frac{q(x)}{p(x)} \\
&\leq \log \sum_{x \in A} p(x) \frac{q(x)}{p(x)} \\
&\leq \log \sum_{x \in \mathcal{X}} q(x) = log(1) = 0
\end{aligned}
$$

As $\log$ is strictly concave, Jensen also tells us that equality can only happen if $q(x)/p(x)$ is constant everywhere. That implies $p = q$.

KL AS LOG-DIFFERENCE

Suppose that data is being generated from an unknown distribution $p(x)$. Suppose we modeled $p(x)$ using an approximating distribution $q(x)$.

First, we could simply see KL as the expected log-difference between $p(x)$ and $q(x)$:

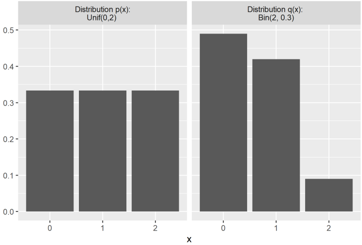
$$D_{KL}(p\|q) = \mathbb{E}_p(\log(p(x)) - \log(q(x)).$$

This is why we integrate out with respect to the data distribution p. A "good" approximation $q(x)$ should minimize the difference to $p(x)$.

KL AS LOG-DIFFERENCE

In machine learning, KL divergence is commonly used to quantify how different one distribution is from another.

Example: Let $q(x)$ be a binomial distribution with $N = 2$ and $p = 0.3$ and let $p(x)$ be a discrete uniform distribution. Both distributions have the same support $\mathcal{X} = \{0, 1, 2\}$.



x	0	1	2
Distribution p(x)	0.3333333	0.3333333	0.3333333
Distribution q(x)	0.4900000	0.4200000	0.0900000

KL AS LOG-DIFFERENCE

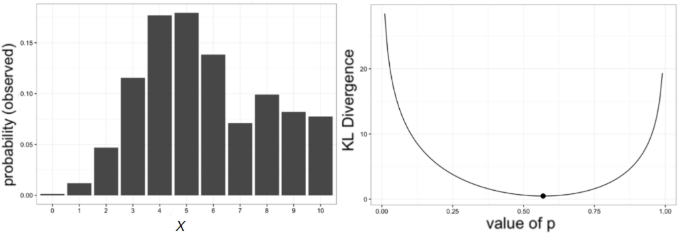
$$
\begin{aligned}
D_{KL}(p\|q) &= \sum_{x\in\mathcal{X}} p(x)\log\left(\frac{p(x)}{q(x)}\right) \\
&= 0.333\log\left(\frac{0.333}{0.49}\right) + 0.333\log\left(\frac{0.333}{0.42}\right) + 0.333\log\left(\frac{0.333}{0.09}\right) \\
&= 0.23099 \text{ (nats)}
\end{aligned}
$$

$$
\begin{aligned}
D_{KL}(q\|p) &= \sum_{x\in\mathcal{X}} q(x)\log\left(\frac{q(x)}{p(x)}\right) \\
&= 0.49\log\left(\frac{0.49}{0.333}\right) + 0.42\log\left(\frac{0.42}{0.333}\right) + 0.09\log\left(\frac{0.09}{0.333}\right) \\
&= 0.16801 \text{ (nats)}
\end{aligned}
$$

Again, note that $D_{KL}(p\|q) \neq D_{KL}(q\|p)$.

KL IN FITTING

Because KL quantifies the difference between distributions, it can be used as a loss function to find a good fit for the observed data.



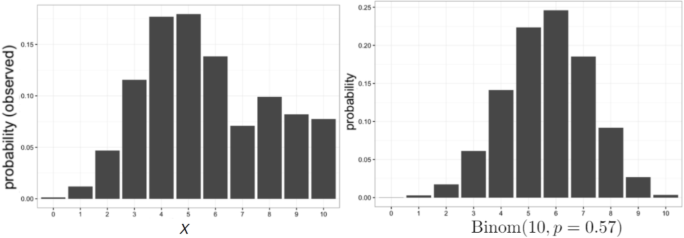
Credit: Will Kurt

Figure: *Left*: Histogram of observed frequencies of a random variable X which takes values between 0 and 10. *Right*: The KL divergence between the observed data and Binom(10,p) is minimized when $p \approx 0.57$.

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

KL IN FITTING

Because KL quantifies the difference between distributions, it can be used as a loss function to find a good fit for the observed data.



Credit: Will Kurt

Figure: *Left*: Histogram of observed frequencies of a random variable which takes values between 0 and 10. *Right*: Fitted Binomial distribution ($p \approx 0.57$).

On the right is the Binomial distribution that minimizes the KL divergence.

KL AS LIKELIHOOD RATIO

- Let us assume we have some data and want to figure out whether $p(x)$ or $q(x)$ matches it better.
- How do we usually do that in stats? Likelihood ratio!

$$LR = \frac{p(x)}{q(x)}$$

In the above, if for x we have $LR > 1$, then p seems better, for $LR < 1$ q seems better.

KL AS LIKELIHOOD RATIO

Or we can compute LR for a complete set of data (as always, logs make our life easier):

$$LR = \prod_i \frac{p(\mathbf{x}^{(i)})}{q(\mathbf{x}^{(i)})} \qquad LLR = \sum_i \log \frac{p(\mathbf{x}^{(i)})}{q(\mathbf{x}^{(i)})}$$

Now let us assume that our data already come from p. It does not really make sense anymore to ask whether p or q fit the data better.
But maybe we want to pose the question "How different is q from p?" by formulating it as: "If we sample many data from p, how easily can we see that p is better than q through LR, on average?"

$$\mathbb{E}_p \left[\log \frac{p(x)}{q(x)} \right]$$

That expected LR is really KL!

KL AS LIKELIHOOD RATIO

In summary we could say for KL:

- It measures how much "evidence" each sample provides on average to distinguish p from q, if you sample from p.
- If p and q are very similar, most samples will not help much, and vice versa for very different distributions.
- In practice, we often want to make the approximation q as indistinguishable from the real p (our data) as possible. We already did that when we fitted (in our log-difference perspective).