## Introduction to Machine Learning

# Information Theory Kullback-Leibler Divergence

× 0 0 × × ×



#### 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 for  $X \sim p(X)$ :

$$D_{\mathit{KL}}(p\|q) = \mathbb{E}_{X \sim 
ho}\left[\log rac{p(X)}{q(X)}
ight] = \sum_{x \in \mathcal{X}} p(x) \cdot \log rac{p(x)}{q(x)},$$

or:

$$D_{\mathit{KL}}(p\|q) = \mathbb{E}_{X \sim 
ho}\left[\log rac{p(X)}{q(X)}
ight] = \int_{x \in \mathcal{X}} p(x) \cdot \log rac{p(x)}{q(x)} \mathrm{d}x.$$

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



#### KULLBACK-LEIBLER DIVERGENCE / 2

$$\mathcal{D}_{ extsf{KL}}(p\|q) = \mathbb{E}_{X \sim 
ho}\left[\log rac{p(X)}{q(X)}
ight]$$

- 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*.

× 0 0 × 0 × × ×

#### **KL-DIVERGENCE EXAMPLE**

KL divergence between p(x) = N(0, 1) and q(x) = LP(0, 1.5) given by

$$D_{\mathcal{KL}}(p\|q) = \int_{x\in\mathcal{X}} p(x) \cdot \log \frac{p(x)}{q(x)}.$$





#### **KL-DIVERGENCE EXAMPLE**

KL divergence between p(x) = LP(0, 1.5) and q(x) = N(0, 1) is different since KL not symmetric



× 0 0 × 0 × ×

#### **KL-DIVERGENCE EXAMPLE**

KL divergence of p(x) = N(0, 1) and  $q(x) = LP(0, \sigma)$  for varying  $\sigma$ 





### INFORMATION INEQUALITY

 $D_{KL}(p||q) \ge 0$  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:

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

× × 0 × ×

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) and we model 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):

$$D_{\mathcal{KL}}(p\|q) = \mathbb{E}_{X \sim p}[\log(p(X)) - \log(q(X))].$$

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).

$$x \sim p(x) \qquad x \qquad \log p(x) - \log q(x)$$

× < 0 × × ×

#### KL AS LOG-DIFFERENCE / 2

Let p(x) = N(0, 1) and q(x) = LP(0, 3). Observe

$$egin{aligned} \mathcal{D}_{\mathcal{KL}}(m{
ho}\|m{q}) &= \mathbb{E}_{X\sim 
ho}[\log(m{
ho}(X)) - \log(m{q}(X))] \ &= \mathbb{E}_{X\sim 
ho}[\log(m{
ho}(X))] - \mathbb{E}_{X\sim 
ho}[\log(m{q}(X))] \end{aligned}$$

× 0 0 × 0 × × ×



.

#### **KL IN FITTING**

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

Because KL quantifies the difference between distributions, it can be used as a loss function between distributions.

In our example, we investigated the KL between p = N(0, 1) and  $q = LP(0, \sigma)$ . Now, we identify an optimal  $\sigma$  which minimizes the KL.





× < 0 × × ×

#### **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!

$$LR = \prod_{i} \frac{p(\mathbf{x}^{(i)})}{q(\mathbf{x}^{(i)})}$$
  $LLR = \sum_{i} \log \frac{p(\mathbf{x}^{(i)})}{q(\mathbf{x}^{(i)})}$ 

If for  $\mathbf{x}^{(i)}$  we have  $p(\mathbf{x}^{(i)})/q(\mathbf{x}^{(i)}) > 1$ , then *p* seems better, for  $p(\mathbf{x}^{(i)})/q(\mathbf{x}^{(i)}) < 1$  *q* seems better.

- Now assume that the data is generated by *p*. Can also ask:
- "How to quantify how much better does *p* fit than *q*, on average?"

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

That expected LLR is really KL!

× 0 0 × ×