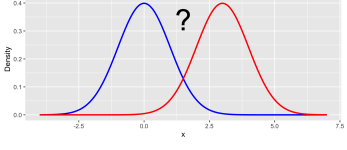
Introduction to Machine Learning

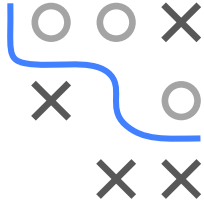
Information Theory
KL for ML



Learning goals

- Understand why measuring distribution similarity is important in ML
- Understand the advantages of forward and reverse KL

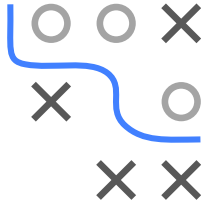
MEASURING DISTRIBUTION SIMILARITY IN ML



- Information theory provides tools (e.g., divergence measures) to quantify the similarity between probability distributions

- The most prominent divergence measure is the KL divergence
- In ML, measuring (and maximizing) the similarity between probability distributions is a ubiquitous concept, which will be shown in the following.

KL DIVERGENCE



Divergences can be used to measure the similarity of distributions.

For distributions p, q they are defined such that

1. $D(p, q) \geq 0$,
2. $D(p, q) = 0$ iff $p = q$.

$\Rightarrow$ divergences can be (and often are) non-symmetrical.

If the same measure dominates the distributions p, q, we can use KL.
For a target distribution p and parametrized distribution q_ϕ, we call

- $D_{KL}(p \| q_\phi)$ forward KL,
- $D_{KL}(q_\phi \| p)$ reverse KL.

In the following, we highlight some properties of the KL that make it attractive from an ML perspective.