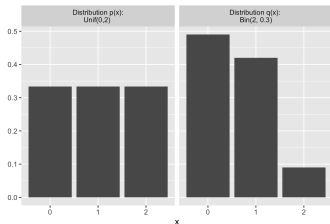
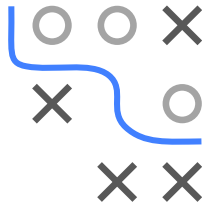


Introduction to Machine Learning

Information Theory

KL and Maximum Entropy



x	x0	x1	x2
Distribution p(x)	0.33	0.33	0.33
Distribution q(x)	0.49	0.42	0.09

Learning goals

- Know the defining properties of the KL
- Understand the relationship between the maximum entropy principle and minimum discrimination information
- Understand the relationship between Shannon entropy and relative entropy

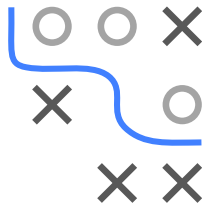
PROBLEMS WITH DIFFERENTIAL ENTROPY

Differential entropy compared to the Shannon entropy:

- Differential entropy can be negative
- Differential entropy is not invariant to coordinate transformations

⇒ Differential entropy is not an uncertainty measure and can not be meaningfully used in a maximum entropy framework.

In the following, we derive an alternative measure, namely the KL divergence (relative entropy), that fixes these shortcomings by taking an inductive inference viewpoint. [► Caticha 2004](#)



INDUCTIVE INFERENCE

We construct a "new" entropy measure $S(p)$ just by desired properties.

Let \mathcal{X} be a measurable space with σ -algebra \mathcal{F} and measure μ that can be continuous or discrete.

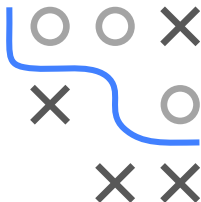
We start with a prior distribution q over \mathcal{X} dominated by μ and a constraint of the form

$$\int_D a(\mathbf{x}) dq(\mathbf{x}) = c \in \mathbb{R}$$

with $D \in \mathcal{F}$. The constraint function $a(\mathbf{x})$ is analogous to moment condition functions $g(\cdot)$ in the discrete case. We want to update the prior distribution q to a posterior distribution p that fulfills the constraint and is maximal w.r.t. $S(p)$.

For this maximization to make sense, S must be transitive, i.e.,

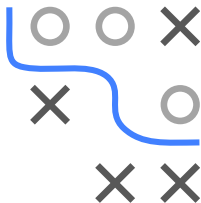
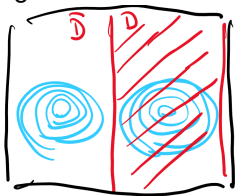
$$S(p_1) < S(p_2), S(p_2) < S(p_3) \Rightarrow S(p_1) < S(p_3).$$



CONSTRUCTING THE KL

1) Locality

The constraint must only update the prior distribution in D , *i.e.*, the region where it is active.



For this, it can be shown that the non-overlapping domains of \mathcal{X} must contribute additively to the entropy, i.e.,

$$S(p) = \int F(p(\mathbf{x}), \mathbf{x}) d\mu(\mathbf{x})$$

where F is an unknown function.