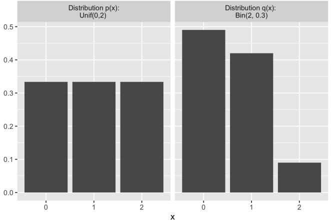
Introduction to Machine Learning

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

$\Rightarrow$ 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

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

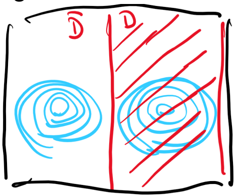
with $D \in \mathcal{F}$. The constraint function $a(\mathbf{x})$ is analogous to moment condition functions $g(\cdot)$ in the discrete case. We want to update the prior distribution q to a posterior distribution p that fulfills the constraint and is maximal w.r.t. $S(p)$.
For this maximization to make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



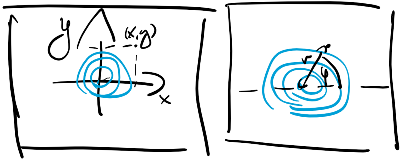
For this, it can be shown that the non-overlapping domains of $\mathcal{X}$ must contribute additively to the entropy, i.e.,

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

where F is an unknown function.

CONSTRUCTING THE KL / 2

2) Invariance to coordinate system



Enforcing 2) results in

$$S(p) = \int \Phi\left(\frac{dp}{dm}(\mathbf{x})\right) dm(\mathbf{x})$$

where Φ is an unknown function, m is another measure on $\mathcal{X}$ dominated by μ and $\frac{dp}{dm}$ the Radon–Nikodym derivative which becomes

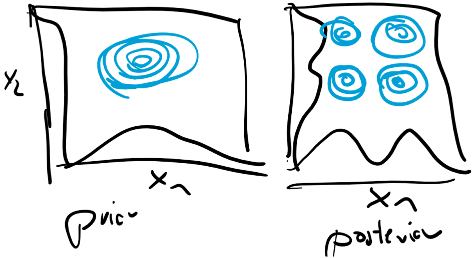
- the quotient of the respective pmfs for discrete measures,
- the quotient of respective pdfs (if they exist) for cont. measures.

CONSTRUCTING THE KL / 3

1 + 2)

$\Rightarrow$ m must be the prior distribution q, and our entropy measure must be understood relatively to this prior, so $S(p)$ becomes, in fact, $S(p\|q)$.

3) Independent subsystems



If the prior distribution defines a subsystem of $\mathcal{X}$ to be independent, then the priors can be independently updated, and the resulting posterior is just their product density.

CONSTRUCTING THE KL / 4

1 + 2 + 3)

Up to constants that do not change our entropy ranking, it follows that

$$S(p\|q) = -\int \log\left(\frac{dp}{dq}(\mathbf{x})\right) dp(\mathbf{x})$$

which is just the negative KL, i.e., $-D_{KL}(p\|q)$.

- With our desired properties, we ended up with KL minimization
- This is called the principle of minimum discrimination information, i.e., the posterior should differ from the prior as least as possible
- This principle is meaningful for continuous and discrete RVs
- The maximum entropy principle is just a special case when $\mathcal{X}$ is discrete and q is the uniform distribution.
- Analogously, Shannon entropy can always be treated as negative KL with uniform reference distribution.