The Kullback Leibler divergence 6

COMSM0075 Information Processing and Brain

comsm0075.github.io

October 2020

The KL divergence

The **Kullback Leibler (KL) divergence** differs from the other information theory quantities in that it deals with two probability distributions $p(x)$ and $q(x)$ on the same set of outcomes $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$.

The KL divergence

The **Kullback Leibler (KL) divergence**, also called the **relative entropy** is

$$d(p \| q) = \sum_i p(x_i) \log_2 \frac{p(x_i)}{q(x_i)}$$

The KL divergence

The **Kullback Leibler (KL) divergence**, also called the **relative entropy** is

$$d(p\|q) = E_p(\log_2 p(X)) - E_p(\log q(X))$$

The KL divergence - coding example

The Kullback Leibler (KL) divergence *is the expected value of the number of extra bits required to encode data with distribution* $p(x)$ *compared to* $q(x)$ *if the code is the optimal code for* $q(x)$.

The KL divergence - coding example

	A	B	C	D
q	1/2	1/4	1/8	1/8
p	1/4	1/8	1/2	1/8
	0	10	110	111

$$L(p) = 1.75$$

where here we mean coding p with the code optimal for p.

The KL divergence - coding example

	A	B	C	D
q	1/2	1/4	1/8	1/8
p	1/4	1/8	1/2	1/8
	0	10	110	111

$$L(q) = \frac{1}{4} + \frac{1}{8} \times 2 + \frac{1}{2} \times 3 + \frac{1}{8} \times 3 = 2.375$$

where here we are coding p with the code optimal for q.

The KL divergence - coding example

	A	B	C	D
q	1/2	1/4	1/8	1/8
p	1/4	1/8	1/2	1/8
	0	10	110	111

$$L(q) - L(p) = 2.375 - 1.75 = 0.625$$

The KL divergence - coding example

	A	B	C	D
q	1/2	1/4	1/8	1/8
p	1/4	1/8	1/2	1/8
	0	10	110	111

$$d(p||q) = \frac{1}{4}\log_2\frac{1}{2} + \frac{1}{8}\log_2\frac{1}{2} + \frac{1}{2}\log_2 4 + \frac{1}{8}\log_2 1$$

The KL divergence - coding example

	A	B	C	D
q	1/2	1/4	1/8	1/8
p	1/4	1/8	1/2	1/8
	0	10	110	111

$$d(p\|q) = 1 - \frac{3}{8} = 0.625$$

The information inequality

The **information inequality**, also called the **Gibbs inequality** says

$$d(p \| q) \geq 0$$

with equality if and only if $p(x) = q(x)$ for all x.

The information inequality

The **information inequality**, also called the **Gibbs inequality** says

$$d(p\|q) \geq 0$$

with equality if and only if $p(x) = q(x)$ for all x. Follows from Jensen's inequality.

Two tasks

1. Use the information processing inequality show that $I(X, Y) \geq 0$; to do this relate $I(X, Y)$ to $d(p\|q)$ by treating $p(x, y)$ and $p(x)p(y)$ as two distributions on the same set of outcomes.
2. Use the information processing inequality show that $H(X) \leq \log_2 n$ where $n = |\mathcal{X}|$. To do this use the uniform distribution as $q(x)$.

Another coding example

	A	B	C	D
q	$1/2$	$1/4$	$1/8$	$1/8$
p	$1/4$	$1/4$	$1/4$	$1/4$
q-code	0	10	110	111
p-code	00	01	10	11

Check the relationship between the divergence and the difference in code lenghts, both using the code optimized to p and q.