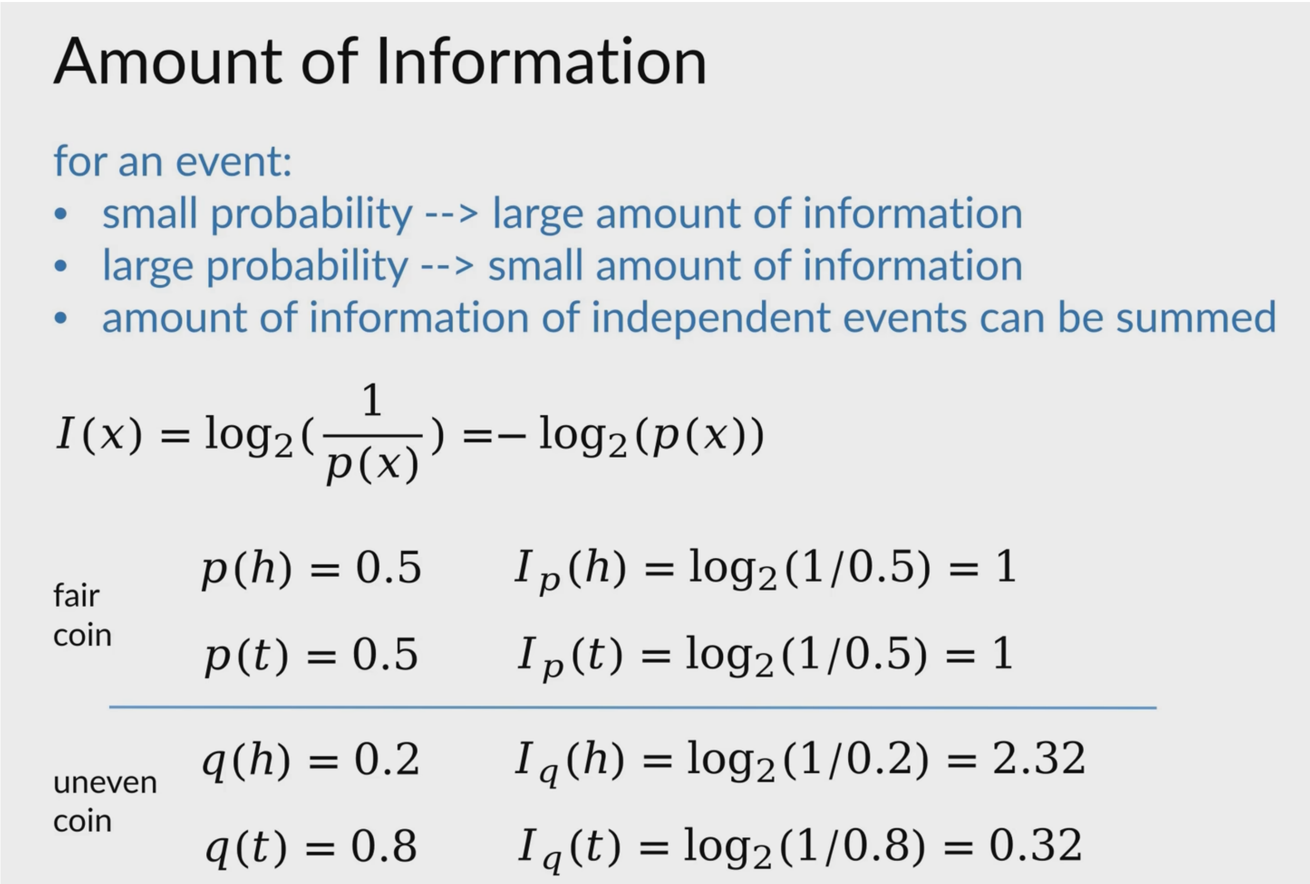
Entronpr, Cross Entropy & KL Divergence

author: doomx
time: 2023.9.22

Amount of Information(fig1)



由此，信息量被定义为和概率负相关，概率越小，信息量越大

$$I(x) = log_2(\frac{1}{p(x)}) = -log_2(p(x))$$

熵这个概念，它不是针对一个事件来说，而是针对一个概率分布来说，它所包含的一个平均信息量，就是代表了这个概率分布的熵。

Shannon Entropy(fig2)

Shannon Entropy

entropy: expected amount of information of a probability distribution

it is also a measurement of uncertainty

$$H(p) = \sum p_i I_i^p = \sum p_i \log_2(\frac{1}{p_i}) = -\sum p_i \log_2(p_i)$$ (assuming discrete, like bernoulli)

(continuous case use integral)

example: a coin with $p(h) = 0.5 \;\; p(t) = 0.5$

$$H(p) = p(h) \times \log_2(1/p(h)) + p(t) \times \log_2(1/p(t)) = 0.5 \times 1 + 0.5 \times 1 = 1$$

example: a coin with $q(h) = 0.2 \;\; q(t) = 0.8$

$$H(q) = q(h) \times \log_2(1/q(h)) + q(t) \times \log_2(1/q(t)) = 0.2 \times 2.32 + 0.8 \times 0.32 = 0.72$$

for a probability distribution:

- pdf more uniform --> more random --> larger entropy
- pdf more condensed --> more certain --> smaller entropy

图2中写的是离散概率分布的Shannon Entropy，如果要求的连续概率分布的shannon entropy，应该把求和符号改为积分符号。可以解释Shannon entropy为概率分布的每种“事件发生的概率值”分别乘以对应“事件所包含的信息量”的和。

Cross Entropy(fig3)

Cross Entropy

a coin with ground truth probability $p(h) = 0.5 \; p(t) = 0.5$

its estimated probability $q(h) = 0.2 \;\; q(t) = 0.8$

given estimated probability distribution, the estimation of expected amount of information of ground truth probability distribution:

$$H(p,q) = \sum p_i I_i^q = \sum p_i \log_2(\frac{1}{q_i}) = -\sum p_i \log_2(q_i)$$

- expectation taking over ground truth probability distribution as data always appear according to ground truth probability distribution
- amount of information using estimated probability distribution as that’s what we estimated

$q(h) = 0.2 \; q(t) = 0.8$

$$H(p,q) = p(h) \times \log_2(1/q(h)) + p(t) \times \log_2(1/q(t)) = 0.5 \times 2.32 + 0.5 \times 0.32 = 1.32$$

$q(h) = 0.4 \; q(t) = 0.6$

$$H(p,q) = p(h) \times \log_2(1/q(h)) + p(t) \times \log_2(1/q(t)) = 0.5 \times 1.32 + 0.5 \times 0.74 = 1.03$$

现在一个事件有一个真实分布，而我们不知道这个分布，我们对于这个事件有一个估计的概率分布。假设我们估计的概率分布为q，真实的概率分布为p，我们对真实概率分布的信息量的估计就叫做cross entropy。

Kullback-Leibler Divergence(fig4)

Kullback-Leibler Divergence (Relative Entropy)

a quantitative way to measure difference between two probability distributions
difference between cross entropy and entropy

$$D(p\|q) = H(p,q) - H(p) = \sum p_i I_i^q - \sum p_i I_i^p$$
$$= \sum p_i \log_2(\frac{1}{q_i}) - \sum p_i \log_2(\frac{1}{p_i})$$
$$= \sum p_i \log_2(\frac{p_i}{q_i})$$

$D(p\|q) \geq 0$ gibbs inequality equals 0 only when two distributions are the same

$D(p\|q) \neq D(q\|p)$ not a distance metric

minimizing kl divergence sometimes equivilant to minimizing cross entropy

$$\nabla_\theta D(p\|q_\theta) = \nabla_\theta H(p, q_\theta) - \nabla_\theta H(p) = \nabla_\theta H(p, q_\theta)$$

p和q之间的KL散度的计算就是p和q的cross entropy减去p的Shannon entropy，因为它们的数学形式很像，通过对数形式可以将加减转化为乘除。但是要注意p和q的位置不能换。

KL散度是大于等于0的，只有当p和q两个概率分布完全一致的时候，KL散度才会等于0。p和q的KL散度和q和p的KL散度是不同的，所以它不是一个距离值。最小化KL散度可以转化为：**cross entropy 对 \theta; 求梯度 减去 shannon entropy 对 \theta; 求梯度**，我们一般认为真实分布的shannon entropy是一个常量，求导之后就是0，所以只需要认为**求导KL散度等于求导cross entropy。**

1. amount of information is the inversion of log probability of an event.（信息量是事件的对数概率的倒数。）
2. entropy measures expected amount of information of a probability distribution. （熵是概率分布的信息量。）
3. cross entropy measures estimation of expected amount of information given estimated probability distribution and is always larger than entropy. （交叉熵度量给定估计概率分布的期望信息量的估计，并且总是大于熵。）
4. KL divergence measures difference between two probability distributions and can be understood as the probability difference of a same sequence given two different probability distributions. （KL散度测量两个概率分布之间的差异，可以理解为给定两个不同概率分布的同一序列的概率差异。）
5. KL divergence is a key concept of probablistic models in machine learning & deep learning and is closely linked to cross entropy loss function. （KL散度是机器学习和深度学习中概率模型的一个关键概念，与交叉熵损失函数密切相关。）