信息论初探

谭斌

08 August 2019

信息论最早由大名鼎鼎的克劳德·香农提出并发展，信息熵这个概念也出自于他那篇大名鼎鼎的论文 *"A Mathematic Theory of Communication"*。信息论早期用于处理通信中的信号处理、传输、压缩等问题，如今已经发展成为一个更大的学科，广泛应用于计算机、密码学等领域。

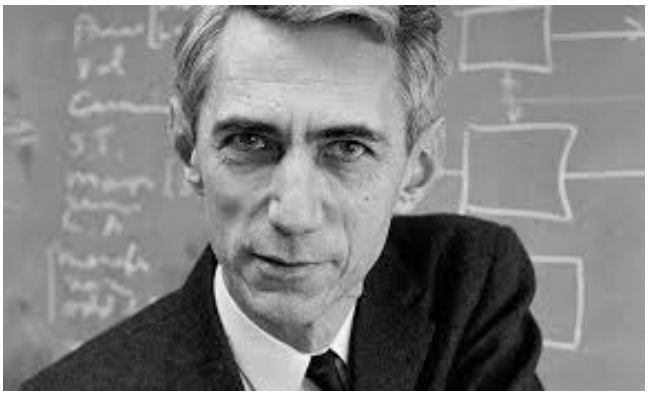

图 1: Claude Elwood Shannon(1916.4.30 － 2001.2.26) 信息论创始人，香农这个名字对我们来说最熟悉的应该是优乐美和计算机网络中的香农定理。

信息论本身并不是面试考察的重点，但是机器学习的重要基础知识之一，多种机器学习理论都可以用信息熵来解释。Shun Watanabe 教授提出过“学习就是一个熵减的过程”，信息熵代表着系统的混乱程度，而优化算法就是不断减少这个“混乱程度”。

1 信息量与信息熵

信息量 (Information Quantity) 和信息熵 (Information Entropy) 是一种非常抽象的概念，我们讨论一件事或者一篇文章包含了多少信息，或者说这条信息有用或者没有用就是最朴素的信息量概念。

比如有人说 2022 国足会踢进世界杯，那么这句话信息量就非常大，从 1930 年开始已经举办了 21 届世界杯，只有一届国足出线，如果国足每年的出线率相互独立，那么可以简单的认为国足下届出线的概率是 1/21，下一次踢进世界杯带来了巨大的信息量；而反过来如果有人说 2022 国足踢不进世界杯，这件事情发生的概率非常大就没有什么信息量。

信息量的计算非常简单，就是

$$\Gamma = -\log p \tag{1}$$

p 是该事件的发生概率，信息量是对事件发生概率的度量，一个事件发生的概率越低，则这个事件包含的信息量越大。国足进下届世界杯这事儿信息量有 $\log 27 \approx 1.4$ 这么大。

信息熵与信息量不同，熵这个概念来源于热力学，热力学熵和信息熵是一样的，熵代表着系统的混乱程度，是表示随机变量不确定性的度量。

我们给定一个分布：

$$P(X = x_i) = p_i, \quad i = 1, 2, ..., n \tag{2}$$

随机变量 X 的熵为：

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i \tag{3}$$

是不是跟信息量有些关系？细心的你肯定能发现，信息熵就是信息量的期望。

一些考题：

1. 给定一个全零的数列，请问方差是多少？信息熵是多少？

熵与方差不同，方差反应的是一组数据的离散程度，虽然看起来一组全部元素都相同的数据方差为 0，熵应该也是 0，但这是错误的理解，熵不是在某一个确定的态下定义的，与数据的取值无关，只反映内容的随机性，必须给定概率分布才能计算。

2. 信息熵越大，信息量是越大还是越小？

信息熵越大，信息量越大。我们考虑熵的公式，如果一个 0-1 分布，取

0 的概率为 1，取 1 得概率为 0，那么整个系统只能有一个态，就是全 0，这时没有不确定性，就没法提供任何信息量，也就是信息量为 0。反过来，比如你不懂 1+1=2，你觉得 1+1 可能等于 (0,100] 任意值，每个取值概率相等，这件事情就有较大信息熵，而后来你学会了，知道了 1+1 只能 =2，你就获得了很大的信息量。

2 KL 散度和交叉熵

以提出人命名总会给人一种屌屌的感觉，实际上 KL 散度 (Kullback—Leibler divergence) 就是相对熵 (relative entropy)
分布是离散情况时:

$$D_{KL}(P||Q) = -\sum_i P(i) \log \frac{Q(i)}{P(i)} \qquad Q(i) > 0, P(i) > 0 \tag{4}$$

连续时:

$$D_{KL}(P||Q) = \int_{-\infty}^{\infty} P(x) \log \frac{Q(x)}{P(x)} \, \mathrm{d}x \qquad Q(x) > 0, P(x) > 0 \tag{5}$$

其中 P 表示数据的真实分布，Q 表示理论分布，KL 散度衡量的就是两个分布间的不相似性。信息论角度来说，KL 散度是用来度量使用基于 Q 的编码来编码来自 P 的样本平均所需的额外的比特数。

交叉熵是我们的老朋友了，作为做常用的损失函数来说，交叉熵不仅仅存在于各种论文和网络结构之中，也存在与各家公司的面试题里，交叉熵本质的含义是：真实分布是 P，Q 是理论分布，那么交叉熵就是基于 Q 分布进行编码时，在事件集合中唯一标识一个事件所需要的平均比特数。在机器学习中交叉熵用来表示目标和预测值之间的差距。
离散情况下:

$$H(P, Q) = -\sum_x P(x) \log Q(x) \tag{6}$$

连续时:

$$H(P, Q) = -\int_{-\infty}^{\infty} P(x) \log Q(x) \, \mathrm{d}x \tag{7}$$

信息熵就是信息量的期望, 我们也可以将交叉熵表示为:

$$H(P, Q) = E_p[-\log Q] \tag{8}$$

交叉熵作为代价函数时是面试最常考内容之一，具体推导和讨论我们放在下一节，本节主要探讨一下信息论中的交叉熵。但是再讲之前我们需要回顾一下本科的一门课程“计算机网络”，这门课中我们学过编码。编码，即对信源输出的符号按一定的数学规则映射到码字的过程，也是信息论的重要课题，前缀码、最优前缀码（哈夫曼编码）、海明码、曼彻斯特编码，都是我们熟知的编码。

，又可以分为唯一可译码和非唯一可译码；定长码中每个码字长度相等，显然只要定长码非奇异 (信源码字一一对应) 就唯一可译，但是变长码却不一定，怎么判断变长码是否唯一可译呢？这就要请出大名鼎鼎的 Kraft-McMillan 不等式了：

$$\sum_{i=1}^{N} 2^{-l_i} \leq 1 \tag{9}$$

其中，N 是编码的码字数，$l_1, l_2, ..., l_N$ 是其长度。

公式的含义是要求编码长度比较长，即各码字不能过短。如果我们把编码过程看成一个 X 上的隐式概率分布 (implicit probability distribution)：

$$q(x_i) = (\frac{1}{2})^{l_i} \tag{10}$$

其中 l_i 是 x_i 的编码长度，那么交叉熵就是在理论分布 Q 下，每个数据编码长度的期望，即：

$$H(p, q) = E_p[l] = -E_p[\log q(x)/\log 2] = -E_p[\log_2 q(x)] \tag{11}$$

一些考题：

1. KL 散度表示的是两个分布的距离这句话对吗？

可以大概这么理解，一般没什么错，但严格来说 KL 散度并不是“距离”或者“度量”，因为它并没有对称性，即：

$$D_{KL}(P||Q) \neq D_{KL}(Q||P) \tag{12}$$

2. 交叉熵和 KL 散度都可以表示目标和预测之间的差距，两者有什么联系？

从定义中我们可以简单推导出两者的关系，由公式 (4)(6) 得：

$$H(P, Q) = H(P) + D_{KL}(P||Q) = D_{KL}(P||Q) + constant \tag{13}$$

由于 P 是真实分布，$H(P)$ 是固定值，交叉熵就是 KL 散度加上一个常数，有时也可以理解为是同样的东西。