熵 (entropy)

2018 年 10 月 19 日

信息量 信息的衡量, 也可以理解为不确定度. 当某件事的不确定度很高时, 其包含的信息量也越大.

如果 x 是 0 到 100 之间的某个数字, 采用二分法猜的话, 需要 $log_2 100$ 次猜测才能保证猜中. 因此 0 到 100 中的某个数字 x 的信息量是 $log_2 100$. 如果改为 0 到 1000, 不确定增加的话, 信息量也从 $log_2 100$ 增加至 $log_2 1000$.

如果从每个数字等于 x 的概率为 $\frac{1}{100}$ 的角度考虑, 上述例子的信息量可以表达为 $-log_2 \frac{1}{100}$, 和 $-log_2 \frac{1}{1000}$.

这里考虑每个数字等于 x 的概率不相等的情况. 将 x 是 0 到 100 之间的某数字, 改为 x 的可能取值是 A,B,C,D 中的某个字母, 且其对应概率为 $\frac{1}{2}$,$\frac{1}{4}$,$\frac{1}{8}$,$\frac{1}{8}$. ABCD 的信息量分别为 $-log_2 \frac{1}{2}$,$-log_2 \frac{1}{4}$,$-log_2 \frac{1}{8}$,$-log_2 \frac{1}{8}$.

信息熵 用来衡量平均而言某个事件的信息量, 也就是某事件信息量的期望. 上述的例子的熵分别为 $\sum_0^{100} \frac{1}{100} \log_2 100$ 和 $-(\frac{1}{2} \log_2 \frac{1}{2} + \frac{1}{4} \log_2 \frac{1}{4} + \frac{1}{8} \log_2 \frac{1}{8} + \frac{1}{8} \log_2 \frac{1}{8})$. 所以信息熵的定义为:

$$H(X) = -\sum P(x) \log P(x)$$

最大熵原理 当面对未知情况时, 应当对其概率分布作最通用的假设, 而不假设其为某种特殊情况[1]. 比如当预测一个骰子掷出的点数时, 应当假设各个数字出现的概率一致 ($\frac{1}{6}$), 而非假设某个数字出现的概率是一个特殊的概率 (如 $\frac{1}{3}$, 而其他数字概率为 $\frac{2}{15}$).

此时, 概率分布最为均匀, 而预测风险最小, 并且此时信息熵最大, 所以这种模型称为“最大熵模型”[1].

[1] 特殊情况的信息熵较小, 而一般情况的信息熵较大 (一般情况指各个事件出现概率均等).

参考文献

[1] 吴军. 数学之美. 2008.