前置导学——信息论

计算机祖师爷之一——香农

你不得佩服计算机界的三位祖师爷：艾伦·图灵、冯·诺依曼和克劳德·香农。我们先来大致说说这三位大佬[1]：

艾伦·图灵（1912年—1954年），英国数学家、逻辑学家，被称为“计算机科学之父”“人工智能之父”。他提出了著名的“图灵机”“图灵测试”，其论文《计算机器与智能》（1950年）是人工智能的开山之作。图灵在“二战”中曾协助军方破解德国密码系统Enigma。不幸的是，年轻的图灵死于家中的床上，床头放着被咬了一口泡过氰化物的苹果。为纪念图灵对计算机科学的巨大贡献，美国计算机协会于1966年设立图灵奖，该奖项是“计算机界的诺贝尔奖”。

冯·诺依曼（1903年—1957年），出生于匈牙利的美籍犹太人，是20世纪最著名的数学家之一，理论计算机科学和博弈论的奠基者，常被誉为“电子计算机之父”。他提出了计算机制造的三个基本原则，即采用二进制逻辑、程序存储执行以及计算机由五个部分组成（运算器、控制器、存储器、输入设备、输出设备），该理论被称为冯·诺依曼体系结构。

克劳德·香农（1916年—2001年），美国数学家和密码学家，是“信息论之父”。1948年发表划时代的论文《通信的数学理论》，宣告“信息论”作为一门学科的诞生。香农还被认为是数字计算机理论和数字电路设计理论的创始人。其硕士毕业论文《中继及开关电路的符号分析》，论证了数字计算机及数字线路逻辑设计的可能性，被誉为“有史以来最重要的一篇硕士论文”。

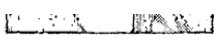
但是我们今天不聊图灵，也不聊冯·诺依曼，我们今天来说说香农。这位大佬曾经手写过一个下棋程序，拉着实验室的科研员和程序下棋；提出的信息论震惊世界，并且直接影响了后来的 NPL（一个句子当中的各个文字出现的次数，后面的决策树，再后面的马尔科夫链，信息论中的互信息、相对熵等）；甚至他的后辈 Leonard Kleinrock（互联网的先驱、奠基人之一）都受他影响：

- 哪些人激发了您的专业灵感？

到目前为止，是麻省理工学院的 Claude Shannon。他是一名卓越的研究者，具有以高度直觉的方式将他的数学理念与物理世界关联起来的能力。他是我的博士论文答辩委员会的成员。

- 您对进入网络/因特网领域的学生们有什么忠告吗？

因特网和由它使能的所有东西是一个巨大的新前沿，充满了令人惊奇的挑战，为众多创新提供了广阔空间。不要受今天技术的束缚，开动大脑，想象能够做些什么，并去实现它。



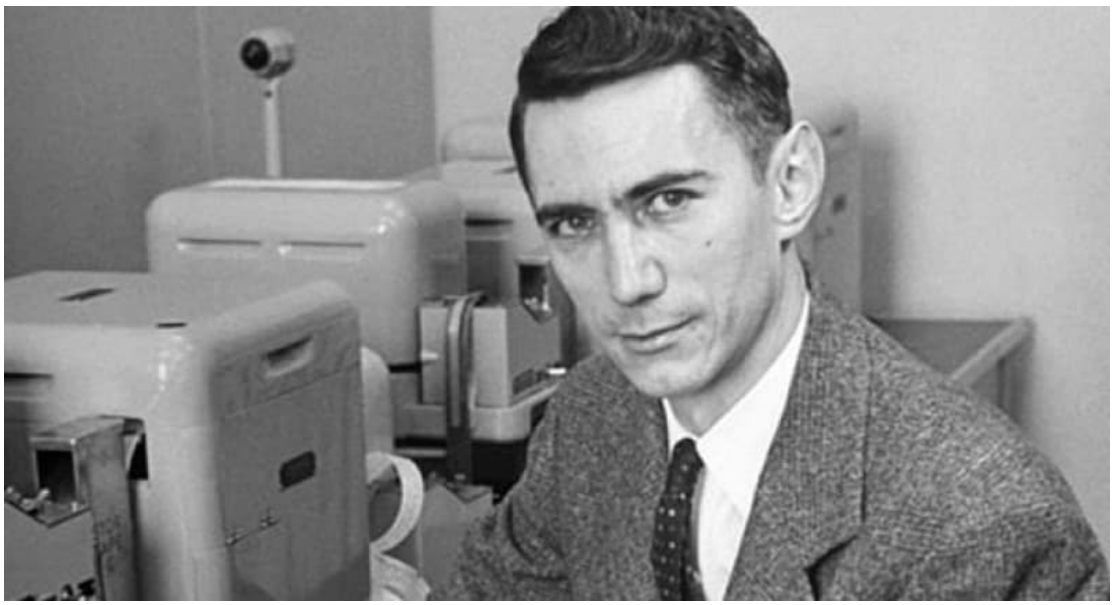
Leonard Kleinrock

- 是什么使得您决定专门研究网络/因特网技术的？

当我于 1959 年在 MIT 读博士时，我发现周围的大多数同学正在信息理论和编码理论领域做研究。在 MIT，那时有伟大的研究者 Claude Shannon，他已经开创这些领域，并且已经解决了许多重要的问题。留下来的研究问题既难又不太重要。因此我决定开始新的研究领域，而该领域其他人还没有想到。回想那时在 MIT 我的周围有许多计算机，我很清楚很快这些计算机将有相互通信的需求。在那时，却没有有效的办法来做到这一点，因此我决定研发能够创建有效的数据网络的技术。

香农与他的信息论

1948年，美国数学家克劳德·香农发表论文《通信的数学理论》（A Mathematical Theory of Communication），奠定了信息论的基础。



今天，信息论在信号处理、数据压缩、自然语言等许多领域，起着关键作用。虽然，它的数学形式很复杂，但是核心思想非常简单，只需要中学数学就能理解。

信息与不确定性

1. 信息的直观理解

- 信息与“不确定性”紧密相关：一个事件越不确定，它带来的信息就越大。
- 举例：
  - 小明说“明天太阳从东边升起” → 信息量接近 0，因为确定性极高。
  - 小明说“掷硬币会出现正面” → 信息量较大，因为结果不确定。
  - 如果信息内容存在大量冗余，重复内容越多，可以压缩的余地就越大。日常生活的经验也是如此，一篇文章翻来覆去都是讲同样的内容，摘要就会很短。反倒是，每句话意思都不一样的文章，很难提炼出摘要。

2. 信息量的基本公式

设某事件 x 的概率为 $p(x)$，则该事件携带的信息量定义为：

$$I(x) = -\log p(x)$$

特点：

- 事件越罕见，一句话中某个单词出现次数越少（出现频率），概率 $p(x)$ 越小，信息量 $I(x)$ 越大。

- 如果 $p(x) = 1$，则 $I(x) = 0$。

例子：

- 掷硬币：$p(\text{正面}) = 0.5$，信息量 $= -\log_2 0.5 = 1$比特。
- 掷骰子：$p(\text{掷出6}) = 1/6$，信息量 $= -\log_2(1/6) \approx 2.585$比特。

概率与信息量成反比

信息熵（Entropy）

1. 定义

熵是信息量的期望值，度量一个系统平均的不确定性：

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i)$$

2. 性质

- 熵最大：在均匀分布时达到最大值 $\log n$。
- 熵最小：完全确定时（某一类概率为 1），熵为 0。

3. 举例

1. **硬币**：

$$H(X) = -[0.5 \log_2 0.5 + 0.5 \log_2 0.5] = 1$$

→ 表示一次掷硬币平均带来 1 比特信息。

1. **骰子**：

$$H(X) = -\sum_{i=1}^{6} \frac{1}{6} \log_2 \frac{1}{6} = \log_2 6 \approx 2.585$$

→ 表示骰子平均信息量约 2.6 比特。

条件熵（Conditional Entropy）

1. 定义

已知随机变量 X，另一个随机变量 Y 的剩余不确定性：

$$H(Y|X) = \sum_{x} p(x) H(Y|X = x)$$

2. 举例

假设预测“是否打网球（Y）”，属性是“天气（X）”：

- 如果天气晴天 → 打球概率 0.8
- 如果天气下雨 → 打球概率 0.2

那么：

$$H(Y|X=\text{晴}) = -[0.8\log_2 0.8 + 0.2\log_2 0.2] \approx 0.72$$
$$H(Y|X=\text{雨}) = -[0.2\log_2 0.2 + 0.8\log_2 0.8] \approx 0.72$$

整体条件熵：

$$H(Y|X) = p(\text{晴}) \cdot H(Y|\text{晴}) + p(\text{雨}) \cdot H(Y|\text{雨})$$
$$\text{不下雨打球} + \text{下雨打球}$$

解释：如果知道天气 (X)，预测是否打球 (Y) 的不确定性会减少。

互信息（Mutual Information）

1. 定义

两个变量的关联程度：

$$I(X;Y) = H(Y) - H(Y|X)$$
$$I(X;Y) = H(Y) - \sum_x p(x)H(Y|X=x)$$

解释：知道 X 后，Y 的不确定性减少了多少。

2. 与决策树的联系

- 决策树中“信息增益”本质上就是 **互信息**。
- 在划分数据决策时，选择让 Y 的不确定性减少最多的属性。

相对熵（KL散度）【选学】

1. 定义

衡量两个分布 P 和 Q 的差异：

$$D_{KL}(P\|Q) = \sum_x P(x)\log\frac{P(x)}{Q(x)}$$

2. 应用

- 在机器学习优化中用来***衡量“预测分布”与“真实分布”的差异***。
- 在决策树中不直接用，但在后续深度学习、生成模型里很重要。

信息论在决策树中的应用（过一遍，决策树再细说）

1. 信息增益（ID3）

信息增益 = 熵的减少量：

$$Gain(D, A) = H(D) - H(D|A)$$

- $H(D)$：样本集的熵
- $H(D|A)$：按属性 A 划分后的条件熵

2. 信息增益率（C4.5）

修正了偏好取值多的缺点：

$$Gain\_ratio(D, A) = \frac{Gain(D, A)}{H(A)}$$

3. 基尼指数（CART）

另一种“不纯度”度量：

$$Gini(D) = 1 - \sum_{k=1}^{K} p_k^2$$

其中 p_k 是第 k 类的概率。

References

[1] 数字技术简史：三位奠基人、三个阶段、五大定律、十项发明

[2] 信息论入门教程