熵的性质

Properties of the Entropy

刘卓

1 信息论的一些基本要素

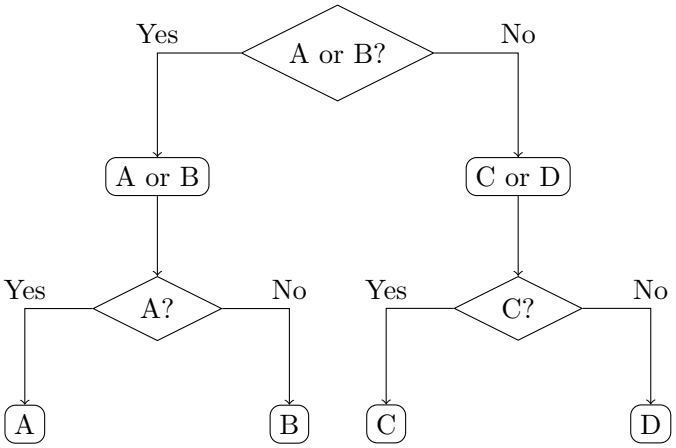
例 1 考虑两个字母表各个字符的概率：

字母	A	B	C	D
字母表 1 的频数	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$
字母表 2 的频数	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$

请问多少个**是/否 (Yes or No) 问题**可以判断出每一个字母?

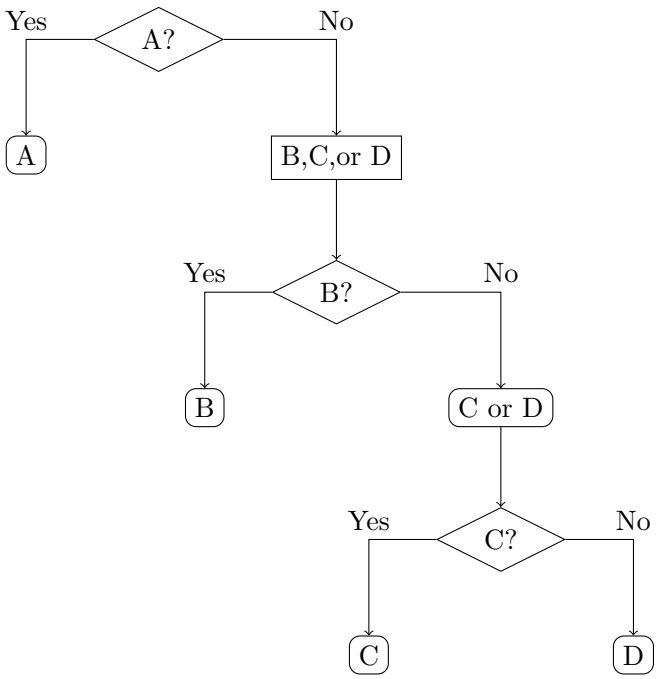
解：

字母表 1 判断方法：



平均需要 2次询问才可以确定一个字符；

字母表 2 判断方法：



平均需要 $\underbrace{1\cdot\frac{1}{2}}_{A}+\underbrace{2\cdot\frac{1}{4}}_{B}+\underbrace{3\cdot\frac{1}{4}}_{C}+\underbrace{3\cdot\frac{1}{4}}_{D}=1.75$次询问就可以确定一个字符；

用 1 代表 Yes, 0 代表 No。就可以使用一个二进制数字表示是/否问题的结果。一般来说，如果实验有 N 个可能的**均等**结果，那么需要：

$$log_2(N)$$

信息大小 (bit) 以存储实验结果。

例 2

1. 一个六面色子，需要 $log_2(6)\approx 2.585$bit 单位。向上取整，需要 3 个 bit 来储存信息。
2. 拉丁字母表，需要 $log_2(26)\approx 4.7$bit 单位。向上取整，需要 5 个 bit 来储存信息。
3. ASCII 表，需要 $log_2(128)=7$bit 单位来储存信息。

如果实验结果是的**不均等**的。如果事件 A 发生概率为 $p=\mathbb{P}[Z=A]$, 那么每个字符需要大小为

$$log_2(\frac{1}{p})$$

的储存空间。

所有事件合在一起所需要的空间一共是:

$$\sum_a p\cdot log_2\frac{1}{p}$$

定义 *1*: 事件 A 的熵是对我们对事件 A 发生的不确定性的度量。随机变量 X 的熵为

$$H(X) = \sum_a \mathbb{P}(X = a) \cdot \log_2 \left(\frac{1}{\mathbb{P}(X = a)} \right)$$

定义 *2*: 两个随机变量 X 和 Y 的熵为

$$H(X, Y) = \sum_{a,b} \mathbb{P}(X = a, Y = b) \cdot \log_2 \left(\frac{1}{\mathbb{P}(X = a, Y = b)} \right)$$

定义 *3*: 如果**事件 Y** 发生，随机变量 X 的条件熵为

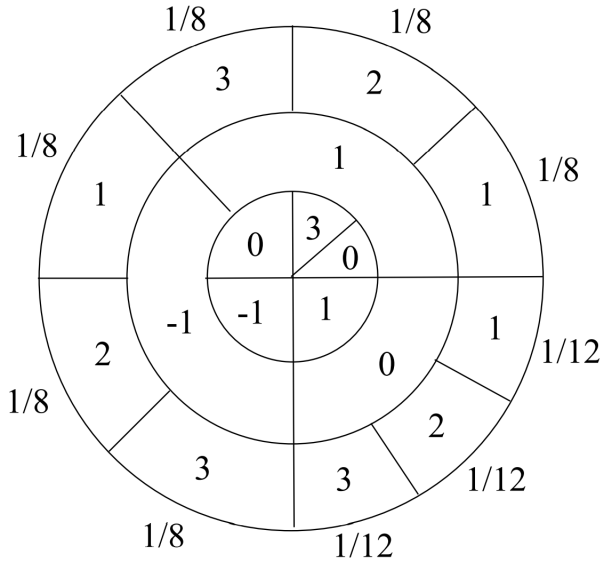
$$H(X|Y) = \sum_a \mathbb{P}(X = a|Y) \cdot \log_2 \left(\frac{1}{\mathbb{P}(X = a|Y)} \right)$$

定义 *4*: 给定**随机变量 Y** 的随机变量 X 的条件熵为

$$H(X|Y) = \sum_b \mathbb{P}(Y = b) \cdot \ H(X|Y = b))$$

例 3

如图所示，假设随机变量 X，Y，Z 是通过旋转获得的。其中 X 由**最里面**的圆给定的，Y 由**中间圆**给定的，Z 由**最外面**的圆给定的。



(1) 计算 $H(X)$

a	-1	0	1	3
$P(X = a)$	$\frac{1}{4}$	$\frac{3}{8}$	$\frac{1}{4}$	$\frac{1}{8}$

$$H(X) = \frac{1}{4} \cdot log_2(\frac{1}{1/4}) + \frac{3}{8} \cdot log_2(\frac{1}{3/8}) + \frac{1}{4} \cdot log_2(\frac{1}{1/4}) + \frac{1}{8} \cdot log_2(\frac{1}{1/8}) \approx 1.906$$

(2) 需要多少 bit 来存储结果 100,000 次 Z 旋转?

a	1	2	3
$P(Z=a)$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

$$H(Z)=\frac{1}{3}\cdot log_2(\frac{1}{1/3})+\frac{1}{3}\cdot log_2(\frac{1}{1/3})+\frac{1}{3}\cdot log_2(\frac{1}{1/3})=log_2(3)\approx 1.585$$

$$H(Z)\cdot 100000=158500\ bit$$

(3) 给定 X = 0，计算 Z 的熵

$$H(Z|X=0)=\sum_a P(Z=a|X=0)\cdot\ log_2(\frac{1}{P(Z=a|X=0)})=\frac{2}{3}\cdot log_2(\frac{2}{3})+\frac{1}{3}\cdot log_2(\frac{1}{3})\approx 0.9183$$

(4) 计算 $H(Z|Y)$

$$H(Z|Y)=\sum_y P(Y=y)\cdot H(Z|Y=y)=\frac{3}{8}\cdot H(Z|Y=-1)+\frac{3}{8}\cdot H(Z|Y=1)+\frac{1}{4}\cdot H(Z|Y=0)$$

$H(Z|Y=y)$ 计算方法同上；

注意：如果 X,Y 是不相关的 (independent)，那么 $H(Y|X)=H(Y)$

(5) 计算 $H(X|Y,Z)$

$H(X|Y,Z)=0$ 如果 X 可以用 Y,Z 代替。即：一旦 $X=aY^n+bZ^m$,n,m 为整数，那么 $H(X|Y,Z)=0$。本题满足该条件。

假设我们知道 X 的值，然后知道 Y 的值，则

$$H(X,Y)=H(X)+H(Y|X)$$

$$H(X_1,X_2,\cdots,X_n)=H(X_1)+H(X_2|X_1)+H(X_3|X_1,X_2)+\cdots+H(X_n|X_1,X_2,\cdots,X_{n-1})$$

证明过程请参考这里

□

一些重要不等式

1. 对于仅取 k 个值的随机变量 X，总是满足：

$$H(X)\leq log_2(k)$$

当且仅当 X 以相等的概率取其所有值时才相等。

证明：

$$\begin{aligned} H(X) &= \sum \mathbb{P}(X=a)\cdot \log_2(\frac{1}{\mathbb{P}(X=a)}) \\ &\leq \log_2(\sum \mathbb{P}(X=a)\cdot \frac{1}{\mathbb{P}(X=a)}) \\ &= \log_2(k) \end{aligned}$$

2. 对于两个随便变量 X 和 Y 来说，总是满足：

$$H(X,Y) \leq H(X)+H(Y)$$

当且仅当 X 和 Y 是不相关 (independent) 时，等式成立。

证明：

$$H(X|Y)H(X)+H(Y|X) \leq H(X)+H(Y)$$

3. 对于两个随便变量 X 和 Y 来说，总是满足：

$$H(X|Y) \leq H(X)$$

当且仅当 X 和 Y 是不相关 (independent) 时，等式成立。

证明：

$$\begin{aligned} H(X,Y) &= \sum_b \mathbb{P}(Y=b)H(X|Y=b) \\ &= \sum_b \mathbb{P}(Y=b)\sum_a \mathbb{P}(X=a|Y=b)\log_2\left(\frac{1}{\mathbb{P}(X=a|Y=b)}\right) \\ &= \sum_b \mathbb{P}(Y=b)\sum_a \frac{\mathbb{P}(X=a\cap Y=b)}{\mathbb{P}(Y=b)}\times\log_2\left(\frac{1}{\mathbb{P}(X=a|Y=b)}\right) \\ &= \sum_b\sum_a \mathbb{P}(X=a\cap Y=b)\log_2\left(\frac{1}{\mathbb{P}(X=a|Y=b)}\right) \end{aligned}$$

$$
\begin{aligned}
&= \sum_a \mathbb{P}(X=a) \sum_b \frac{\mathbb{P}(Y=b \cap X=a)}{\mathbb{P}(X=a)} \log_2 \left(\frac{1}{\mathbb{P}(X=a|Y=b)} \right) \\
&= \sum_a \mathbb{P}(X=a) \sum_b \mathbb{P}(Y=b|X=a) \log_2 \left(\frac{1}{\mathbb{P}(X=a|Y=b)} \right) \\
&\leq \sum_a \mathbb{P}(X=a) \log_2 \left(\sum_b \frac{\mathbb{P}(X=a|Y=b)}{\mathbb{P}(X=a|Y=b)} \right) \\
&= \sum_a \mathbb{P}(X=a) \log_2 \left(\sum_b \frac{\mathbb{P}(Y=b \cap X=a)}{\mathbb{P}(X=a)} \times \frac{\mathbb{P}(Y=b)}{\mathbb{P}(X=a \cap Y=b)} \right) \\
&= \sum_a \mathbb{P}(X=a) \log_2 \left(\sum_b \frac{\mathbb{P}(Y=b)}{\mathbb{P}(X=a)} \right) \\
&= \sum_a \mathbb{P}(X=a) \log_2 \left(\frac{}{\mathbb{P}(X=a)} \right) \\
&= H(X)
\end{aligned}
$$

2 英语中的熵

以下是英文写作中大约 1000 个字母频率表：

字母	a	b	c	d	e	f	g	h	i	j	k	l	m
频数	73	9	30	44	130	28	16	35	74	2	3	35	25

n	o	p	q	r	s	t	u	v	w	x	y	z
78	74	27	3	77	63	93	27	13	16	5	19	1

计算英语中的熵公式为：

$$
\begin{aligned}
H(X) &= \sum_{\alpha=A}^{Z} \mathbb{P}[X=\alpha] \cdot \log_2 \left(\frac{1}{\mathbb{P}[X=\alpha]} \right) \\
&= \mathbb{P}(A) \log_2 \left(\frac{1}{\mathbb{P}(A)} \right) + \cdots + \mathbb{P}(Z) \log_2 \left(\frac{1}{\mathbb{P}(Z)} \right) \\
&= 0.073 \cdot \log_2 \left(\frac{1}{0.073} \right) + \cdots + 0.001 \cdot \log_2 \left(\frac{1}{0.001} \right) \\
&\approx 4.1621
\end{aligned}
$$

移位平均需要 4.1621 个 bit 来储存一个拉丁字母。

3 摩斯密码

摩斯密码 (Morse code) 是一种时通时断的信号代码通过不同的排列顺序来表达不同的英文字母、数字和标点符号。是由美国人艾尔菲德 · 维尔与萨缪尔 · 摩尔斯在 1836 年发明。

1. 点（·）：1 （读 “滴” dit ， 时间占据 1t ）

2. 划（—）：111 （读 “嗒” dah ， 时间占据 3t ）

3. 字符内部的停顿（在点和划之间）：0 （时间占据 1t ）

4. 字符间停顿：000 （时间占据 3t ）

5. 单词间的停顿：0000000 （时间占据 7t ）

点的长度（也就是上面的时间长度 t）决定了发报的速度。

图 1: 国际摩斯密码

其熵约为: 7.039，意味着平均需要 7 个 bit 来储存一个字符。