上周作业解答

Exercise 1. [Cover(2006)]

Minimum entropy. What is the minimum value of $H(p_1, \ldots, p_n) = H(\mathbf{p})$ as $\mathbf{p}$ ranges over the set of n-dimensional probability vectors? Find all $\mathbf{p}$'s that achieve this minimum.

上周作业解答

Exercise 1. [Cover(2006)]

Minimum entropy. What is the minimum value of $H(p_1, \ldots, p_n) = H(\mathbf{p})$ as $\mathbf{p}$ ranges over the set of n-dimensional probability vectors? Find all $\mathbf{p}$'s that achieve this minimum.

Answer:

求$H(p_1, p_2, \cdots, p_n) = H(\mathbf{p})$的最小值，其中$\mathbf{p}$的取值域为$n$维概率向量集合。

根据熵的定义，我们有

$H(\mathbf{p}) = H(p_1, p_2, \cdots, p_n) = -\sum_{i=1}^{n} p_i \log p_i$,

其中$\sum_{i=1}^{n} p_i = 1, p_i \geqslant 0$。根据熵的确定性性质，$H(\mathbf{p})$取最小值**0**，当且仅当$p_i$中有一个取值为**1**，其余为**0**。

所以，$\mathcal{P} = \{(1,0,0,\cdots,0),(0,1,0,\cdots,0),\cdots,(0,0,0,\cdots,1)\}, |\mathcal{P}| = n$。

上周作业解答

Exercise 2. [Cover(2006)]

World Series. The World Series is a seven-game series that terminates as soon as either team wins four games. Let X be the random variable that represents the outcome of a World Series between teams A and B; possible values of X are AAAA, BABABAB, and BBBAAAA. Let Y be the number of games played, which ranges from 4 to 7. Assuming that A and B are equally matched and that the games are independent, calculate $H(X)$, $H(Y)$, $H(Y|X)$, and $H(X|Y)$.

上周作业解答

Answer：

两支队伍AB比赛，直到其中一队赢得4场。随机变量X表示A队和B队比赛的结果，Y表示比赛的场数。假定两支队伍旗鼓相当，则A队和B队每场获胜的概率相等（为1/2）。每场比赛相互独立，我们可以建立以下概率模型：

X	X类型	同类X 数量	同类中每个X的概率$P_X(x)$	Y	$P_Y(y)$
AAAA	4A	1	$\left(\frac{1}{2}\right)^4 = \frac{1}{16}$	4	$2 \times \left(\frac{1}{2}\right)^4 = \frac{1}{8}$
BBBB	4B	1			
□□□□A	4A1B	$\binom{4}{3}$	$\left(\frac{1}{2}\right)^5 = \frac{1}{32}$	5	$2 \times \binom{4}{3} \times \left(\frac{1}{2}\right)^5 = \frac{1}{4}$
□□□□B	4B1A	$\binom{4}{3}$			
□□□□□A	4A2B	$\binom{5}{3}$	$\left(\frac{1}{2}\right)^6 = \frac{1}{64}$	6	$2 \times \binom{5}{3} \times \left(\frac{1}{2}\right)^6 = \frac{5}{16}$
□□□□□B	4B2A	$\binom{5}{3}$			
□□□□□□A	4A3B	$\binom{6}{3}$	$\left(\frac{1}{2}\right)^7 = \frac{1}{128}$	7	$2 \times \binom{6}{3} \times \left(\frac{1}{2}\right)^7 = \frac{5}{16}$
□□□□□□B	4B3A	$\binom{6}{3}$			

上周作业解答

根据熵的定义，可得

$$
\begin{aligned}
H(X) &= -\sum_x p(x)\log p(x) \\
&= -2\times\frac{1}{16}\log\frac{1}{16} - 2\times\binom{4}{3}\times\frac{1}{32}\log\frac{1}{32} \\
&\quad -2\times\binom{5}{3}\times\frac{1}{64}\log\frac{1}{64} - 2\times\binom{6}{3}\times\frac{1}{128}\log\frac{1}{128} \\
&= 5.8125\text{比特/符号}.
\end{aligned}
$$

$$
\begin{aligned}
H(Y) &= -\sum_y p(y)\log p(y) \\
&= -\frac{1}{8}\log\frac{1}{8} - \frac{1}{4}\log\frac{1}{4} - \frac{5}{16}\log\frac{5}{16} - \frac{5}{16}\log\frac{5}{16} \\
&\approx 1.9238\text{比特/符号}.
\end{aligned}
$$

上周作业解答

$H(Y|X)=0$，因为Y的值取决于X，即当知道X的取值时，Y即可确定。所以已知X的情况下，Y的熵为**0**，即$H(Y|X)=0$。由联合熵的链式法则，有
$H(X)+H(Y|X)=H(X,Y)=H(Y)+H(X|Y)$。
所以，

$$\begin{aligned} H(X|Y) &= H(X)+H(Y|X)-H(Y) \\ &\approx 5.8125+0-1.9238 \\ &= 3.8887\text{比特/符号}. \end{aligned}$$

也可以选择代入公式求$H(X|Y)$，

$$\begin{aligned} H(X|Y) &= \sum_y p(y)H(X|Y=y) \\ &= -\sum_y p(y)\sum_x p(x|y)\log p(x|y) \\ &= -\frac{1}{8}(2\times\frac{1}{2}\log\frac{1}{2})-\frac{1}{4}(8\times\frac{1}{8}\log\frac{1}{8}) \\ &\quad -\frac{5}{16}(20\times\frac{1}{20}\log\frac{1}{20})-\frac{5}{16}(40\times\frac{1}{40}\log\frac{1}{40}) \\ &\approx 3.8887\text{比特/符号}. \end{aligned}$$

上周作业解答

Exercise 3. [王育民(2013)]

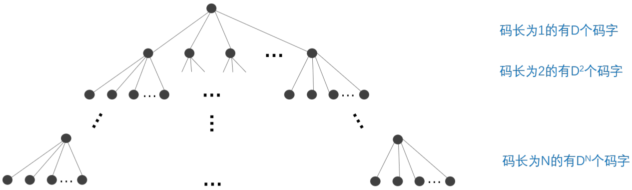
3.1 试证明长为 N 的 D 元不等长码至多有 $D(D^N-1)/(D-1)$ 个码字。

依据教材（王育民等，2013，pp.49)中的定义：

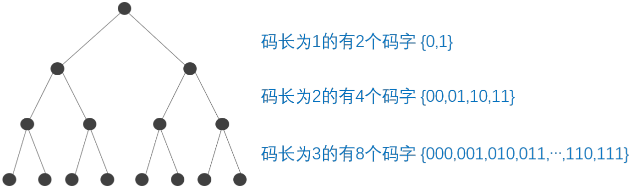
"设有一含 D个字母的集合 $B=\{b_1,b_2,\cdots,b_D\}$，称 B 为码的字母（或符号）表。可用从 B中选出的符号序列表示信源的输出，例如若 $B=\{0,1\}$，则 $\{01,011,0111,01111\}$可以表示4个不同的信源序列，其中每个码符号序列称做码字。同样 $\{000,011,110,101\}$也可表示信源输出的4个不同序列。这两个码字集合都称做二元码，前者为不等长码，后者为等长码。一般为 D元码。对于等长 D元码，若长度为 N，则至多有 D^N个码字。对于不等长 D元码，若码字最长限定为 N，则至多有 $D(D^N-1)/(D-1)$个码字。"

$$\sum_{i=1}^{N} D_i = D + D^2 + \cdots + D^N = \frac{D(D^N-1)}{D-1}.$$

上周作业解答



(a) D元码树（最长码字为N）



(b) 2元码树（最长码字为3）

上周作业解答

Exercise 4. [王育民(2013)]

3.2 设有一离散无记忆信源 $U=\begin{Bmatrix} a_1 & a_2 \\ 0.004 & 0.996 \end{Bmatrix}$。若对其输出的长为 100 的事件序列中含有两个或少于两个 a_1 的序列提供不同的码字。

(a) 在等长编码下，求二元码的最短码长。

(b) 求错误概率（误组率）。

上周作业解答

Exercise 4. [王育民(2013)]

3.2 设有一离散无记忆信源 $U=\begin{Bmatrix} a_1 & a_2 \\ 0.004 & 0.996 \end{Bmatrix}$。若对其输出的长为 100 的事件序列中含有两个或少于两个 a_1 的序列提供不同的码字。

(a) 在等长编码下,求二元码的最短码长。

(b) 求错误概率(误组率)。

(a)设长为100的事件序列中少于等于两个 a_1的序列有 M个，

$$M=\binom{100}{0}+\binom{100}{1}+\binom{100}{2}=1+100+4950=5051,$$

所以在等长编码下，二元码的最短码长为，

$$N=\lceil \log M \rceil = 13.$$

(b)正确概率为：

$P_c=\binom{100}{0}P(a_1)^0P(a_1)^{100}+\binom{100}{1}P(a_1)^1P(a_1)^{99}+\binom{100}{2}P(a_1)^2P(a_1)^{98}=$
$1\times 0.004^0\times 0.996^{100}+100\times 0.004^1\times 0.996^{99}+4950\times 0.004^2\times 0.996^{98}\approx 0.9922,$

错误概率为：

$P_e=1-P_c\approx 0.0078.$

上周作业解答

Exercise 5.

有一离散无记忆信源X，对于该信源中的序列$(x_1, x_2, \cdots, x_n) \in \mathcal{X}^n$，如果它满足

$$2^{-n(H(X)+\epsilon)} \leq P_{X^n}(x^n) \leq 2^{-n(H(X)-\epsilon)},$$

则称它是$\epsilon-$典型 的。投掷一枚不均匀硬币，其正面朝上的概率为1/4，反面朝上的概率为3/4。对应有随机变量$X \in \{0, 1\}$，其概率质量函数为$P(1) = 1/4$ 和$P(0) = 3/4$。现在独立投掷该硬币n 次，求：

(a) 熵$H(X)$。

(b) 设$n = 5$，则当ϵ 等于0.1 时，哪些序列是$\epsilon-$典型 的？

(c)$n = 5, \epsilon = 0.01$ 呢？

上周作业解答

Answer:

投掷一枚不均匀硬币，正面朝上的概率为 $P(X=0)=1/4$，反面朝上的概率为 $P(X=1)=3/4$。

(a)根据熵的定义，

$$\begin{aligned} H(X) &= -\sum_{X} p(x)\log p(x) \\ &= -\frac{1}{4}\times\log\frac{1}{4}-\frac{3}{4}\times\log\frac{3}{4} \\ &\approx 0.8113\text{比特/符号}. \end{aligned}$$

(b)$n=5,\epsilon=0.1$，根据题中给出的 $\epsilon-$ 典型 序列的定义，有
$2^{-5(0.8113+0.1)}\le P_{X^5}\left(x^5\right)\le 2^{-5(0.8113-0.1)}$，
即满足概率为 $0.0425\le P_{X^5}(x^5)\le 0.0850$ 的序列x^5为 $\epsilon-$ 典型 序列。

上周作业解答

独立投掷该硬币5次，可以产生2^5种序列。由于每次投掷相互独立，所以这32种序列可以分为6类，同一类中的序列概率相同。令$X^5 = X_1X_2X_3X_4X_5$
(1)若$x_i, 1 \leqslant i \leqslant 5$全为**1**，则$P_{X^5}(x^5) = (\frac{3}{4})^5 \approx 0.2373$,
(2)若$x_i, 1 \leqslant i \leqslant 5$中有**1**个为**0**，则$P_{X^5}(x^5) = (\frac{1}{4})(\frac{3}{4})^4 \approx 0.0791$,
(3)若$x_i, 1 \leqslant i \leqslant 5$中有**2**个为**0**，则$P_{X^5}(x^5) = (\frac{1}{4})^2(\frac{3}{4})^3 \approx 0.0264$,
(4)若$x_i, 1 \leqslant i \leqslant 5$中有**3**个为**0**，则$P_{X^5}(x^5) = (\frac{1}{4})^3(\frac{3}{4})^2 \approx 0.0088$,
(5)若$x_i, 1 \leqslant i \leqslant 5$中有**4**个为**0**，则$P_{X^5}(x^5) = (\frac{1}{4})^4(\frac{3}{4}) \approx 0.0029$,
(6)若$x_i, 1 \leqslant i \leqslant 5$全为**0**，则$P_{X^5}(x^5) = (\frac{1}{4})^5 \approx 0.00098$，
其中，只有当$x_i, 1 \leqslant i \leqslant 5$中有**1**个为**0**时，$P_{X^5}(x^5)$落在$[0.0425, 0.0850]$区间内。所以，投掷**5**次硬币，其中只有**1**个是正面的序列是 $\epsilon -$ 典型 序列。

(c)当$n = 5, \epsilon = 0.01$时，满足$2^{-5(0.8113+0.01)} \leq P_{X^5}\left(x^5\right) \leq 2^{-5(0.8113-0.01)}$，即序列概率落在$[0.0581, 0.0622]$区间的序列是 $\epsilon -$ 典型 序列。由(b)中算出的**6**类序列概率可以看到，没有一类序列的概率落在该区间内，所以，当$n = 5, \epsilon = 0.01$时，没有序列是 $\epsilon -$ 典型 序列。