

第 7 讲 信源编码

有关编码的几个例子

唯一可译码

即时码

Kraft 不等式

最优码

最优码长的界

惟一可译码的
Kraft 不等式

有关编码的几个例子

唯一可译码

即时码

Kraft 不等式

最优码

最优码长的界

惟一可译码的 Kraft 不等式

有关编码的几个例子

唯一可译码

即时码

Kraft 不等式

最优码

最优码长的界

惟一可译码的

Kraft 不等式

有关编码的几个例子

唯一可译码

即时码

Kraft 不等式

最优码

最优码长的界

惟一可译码的 Kraft 不等式

有关编码的几个例子

唯一可译码

即时码

Kraft 不等式

最优码

最优码长的界

惟一可译码的
Kraft 不等式

有关编码的几个例子

唯一可译码

即时码

Kraft 不等式

最优码

最优码长的界

惟一可译码的
Kraft 不等式

定义 1.1

关于随机变量 X 的信源编码 C 是从 X 的取值空间 $\mathcal{X}$ 到 $\mathcal{D}^*$ 的一个映射, 其中 $\mathcal{D}^*$ 表示 D 元字母表 $\mathcal{D}$ 上有限长度的字符串所构成的集合. 用 $C(x)$ 表示 x 的码字并用 $l(x)$ 表示 $C(x)$ 的长度.

定义 1.2

设随机变量 X 的概率密度函数为 $p(x)$, 定义信源编码 $C(x)$ 的期望长度 $L(C)$ 为

$$L(C) = \sum_{x \in \mathcal{X}} p(x)l(x).$$

其中 $l(x)$ 表示对应于 x 的码字长度.

例 1.3

设随机变量 X 的分布及其码字分配如下：

$$P\{X = 1\} = \frac{1}{2} \quad \text{码字 } C(1) = 0$$

$$P\{X = 2\} = \frac{1}{4} \quad \text{码字 } C(2) = 10$$

$$P\{X = 3\} = \frac{1}{8} \quad \text{码字 } C(3) = 110$$

$$P\{X = 4\} = \frac{1}{8} \quad \text{码字 } C(4) = 111$$

易知 X 的熵 $H(X)$ 为 1.75 比特, 而期望长度 $L(C) = El(X)$ 也是 1.75 比特.

有关编码的几个例子

唯一可译码

即时码

Kraft 不等式

最优码

最优码长的界

惟一可译码的

Kraft 不等式

例 1.4

设随机变量 X 的分布及其码字分配如下：

$$P\{X = 1\} = \frac{1}{3} \quad \text{码字 } C(1) = 0$$

$$P\{X = 2\} = \frac{1}{3} \quad \text{码字 } C(2) = 10$$

$$P\{X = 3\} = \frac{1}{3} \quad \text{码字 } C(3) = 11$$

易知 X 的熵 $H(X)$ 为 $\log 3 = 1.58$ 比特，而期望长度 $L(C) = El(X)$ 是 1.67 比特，此时 $El(X) > H(X)$ 。

定义 1.5

如果编码将 X 的取值空间中的每个元素映射成 $\mathcal{D}^*$ 中不同的字符串, 即

$$x \neq x' \Rightarrow C(x) \neq C(x')$$

则称这个编码是非奇异的.

定义 1.6

编码 C 的扩展 C^* 是从 $\mathcal{X}$ 上的有限长字符串到 $\mathcal{D}$ 上的有限长字符串的映射, 定义为

$$C(x_1x_2\cdots x_n) = C(x_1)C(x_2)\cdots C(x_n).$$

有关编码的几个例子

唯一可译码

即时码

Kraft 不等式

最优码

最优码长的界

唯一可译码的
Kraft 不等式**例 1.7**

如果 $C(x_1) = 00$ 且 $C(x_2) = 11$, 则 $C(x_1x_2) = 0011$.

有关编码的几个例子

唯一可译码

即时码

Kraft 不等式

最优码

最优码长的界

惟一可译码的 Kraft 不等式

有关编码的几个例子

唯一可译码

即时码

Kraft 不等式

最优码

最优码长的界

惟一可译码的
Kraft 不等式

有关编码的几个例子

唯一可译码

即时码

Kraft 不等式

最优码

最优码长的界

唯一可译码的
Kraft 不等式

定义 2.1

如果一个编码的扩展编码是非奇异的, 则称该编码是惟一可译的.

定理 2.2

若码 C 中的所有字码 w_i 都具有相同的长度, 则 C 是唯一可译码。

有关编码的几个例子

唯一可译码

即时码

Kraft 不等式

最优码

最优码长的界

惟一可译码的

Kraft 不等式

例 2.4

由下式给出的二进制编码 C :

$$s_1 \mapsto w_1 = 0, \quad s_2 \mapsto w_2 = 01, \quad s_3 \mapsto w_3 = 011$$

具有变长特征, 但仍然是唯一可译码。在序列 t 中, 每个符号 0 都标志着一个码字 w_i 的开始, 而下标 i 等于 1 加上其后跟随的 1 的个数。例如,

$$t = 0010111010011 = 0.01.011.01.0.011 \Rightarrow s = s_1 s_2 s_3 s_2 s_1 s_3.$$

实际上, 这里我们把符号 $0 \in T$ 当作标点符号来使用。

有关编码的几个例子

唯一可译码

即时码

Kraft 不等式

最优码

最优码长的界

惟一可译码的

Kraft 不等式

有关编码的几个例子

唯一可译码

即时码

Kraft 不等式

最优码

最优码长的界

惟一可译码的

Kraft 不等式

下面我们介绍由 Sardinas 和 Patterson 于 1957 年设计出的一种判断唯一可译码的测试方法.

我们设编码的字母表为 T . 我们用 T^+ 表示所有非空字符构成的集合:

$$T^+ = \bigcup_{n>0} T^n.$$

我们利用归纳法定义一个由非空单词集合 $C_0, C_1, \dots$ 构成的序列, 使得对所有 n , 都有 $C_n \subseteq T^+$ 。具体地, 我们定义 $C_0 = C$, 并且对于每个 $n \geq 1$, 定义

$$C_n = \{w \in T^+ \mid uw = v, \text{ 其中 } u \in C, v \in C_{n-1}, \text{ 或者 } u \in C_{n-1}, v \in C\}. \quad (2.1)$$

然后定义

$$C_\infty = \bigcup_{n=1}^{\infty} C_n.$$

这个定义乍看起来可能有些令人望而生畏, 但如果我们一步一步来看, 就会变得更加清楚: 我们从 $C_0 = C$ 开始, 然后利用前一项 C_{n-1} 构造每一个 C_n ($n \geq 1$), 最后取

$$C_\infty = C_1 \cup C_2 \cup \dots$$

有关编码的几个例子

唯一可译码

即时码

Kraft 不等式

最优码

最优码长的界

惟一可译码的

Kraft 不等式

有关编码的几个例子

唯一可译码

即时码

Kraft 不等式

最优码

最优码长的界

惟一可译码的

Kraft 不等式

注意, 当 $n = 1$ 时, C_n 的定义可以进一步简化: 由于 $C_{n-1} = C_0 = C$, 所以在 (2.1) 中的两个条件实际上是相同的, 即

$$C_1 = \{w \in T^+ \mid uw = v, \text{ 其中 } u, v \in C\}.$$

我们还注意到, 如果 $C_{n-1} = \emptyset$, 那么 $C_n = \emptyset$ 。因此不断迭代可以得到

$$C_{n+1} = C_{n+2} = \cdots = \emptyset.$$

例 2.5

设 $C = \{0, 01, 011\}$. 则根据 (2.1) 可得 $C_1 = \{1, 11\}$: 我们有 $1 \in C_1$, 因为 $01.1 = 011$ (其中 $01, 011 \in C = C_0$); 且 $11 \in C_1$, 因为 $0.11 = 011$ (其中 $0, 011 \in C = C_0$). 在下一步中, 当 $n = 2$ 时, 经检验可知不存在 $w \in T^+$ 满足 $uw = v$, 其中 $u \in C, v \in C_1$ (或反过来). 因此 $C_2 = \emptyset$, 从而对所有 $n \geq 2$ 都有 $C_n = \emptyset$, 于是 $C_\infty = C_1 = \{1, 11\}$.

例 2.6

若 $C = \{02, 12, 120, 20, 21\}$, 则 $C_1 = \{0\}, C_2 = \{2\}, C_3 = \{0, 1\}, C_4 = \{2, 20\}, C_5 = \{0, 1\}$; 重复性 $C_3 = C_5$ 意味着对于奇数或偶数 $n \geq 3, C_n = \{0, 1\}$ (奇数) 或 $\{2, 20\}$ (偶数). 因此 $C_\infty = C_1 \cup \dots \cup C_4 = \{0, 1, 2, 20\}$.
 若 $C = \{02, 12, 120, 21\}$, 则 $C_1 = \{0\}, C_2 = \{2\}, C_3 = \{1\}, C_4 = \{2, 20\}, C_5 = \{1\}$; 同样地, $C_3 = C_5$ 意味着对于奇数或偶数 $n \geq 3, C_n = \{1\}$ (奇数) 或 $\{2, 20\}$ (偶数). 因此 $C_\infty = C_1 \cup \dots \cup C_4 = \{0, 1, 2, 20\}$

有关编码的几个例子

唯一可译码

即时码

Kraft 不等式

最优码

最优码长的界

惟一可译码的
Kraft 不等式

命题 2.7

如果 C 的码字长度为 $l_1, \dots, l_q$, 且对于某个 n , 有 $w \in C_n$, 那么 $|w| \leq l = \max(l_1, \dots, l_q)$. 从而每个 C_n 都是有限集, 且集合序列 $C_0, C_1, \dots$ 是最终周期的.

证明.

对 n 使用数学归纳法. 若 $n = 0$, 则 $C_n = C$. 因此 $|w| \leq l$. 若 $n > 0$, 则 $uw = v$ (其中 $v \in C_{n-1}$ 或 C), 因此根据归纳法或 l 的定义可知 $|w| \leq |v| \leq l$. 长度满足 $|w| \leq l$ 的非空 r 进制词的个数为

$$N = r + r^2 + \cdots + r^l = r(r^l - 1)/(r - 1).$$

所以对每个 n , 都有 $|C_n| \leq N$. 此类字 w 的子集最多只有 2^N 个不同的非空集合, 因此在集合序列 $C_0, \dots, C_{2^N}$ 中必然存在重复, 即存在 $i < j \leq 2^N$ 使得 $C_j = C_i$. 根据等式 (1.3), 每一个 C_n 仅取决于 C 和 C_{n-1} . 因此对所有 $k \geq 0$, 都有 $C_{j+k} = C_{i+k}$; 从而每个 C_n 都等于 C_0 或 C_1 或 $\dots$ 或 C_{j-1} . 于是 $C_\infty = C_0 \cup C_1 \cup \cdots \cup C_{j-1}$. 因此, 一旦我们在连续集合序列 $C_0, C_1, \dots$ 中找到重复项, 就构造出了 C_∞ 的所有元素. □

有关编码的几个例子

唯一可译码

即时码

Kraft 不等式

最优码

最优码长的界

惟一可译码的

Kraft 不等式

有关编码的几个例子

唯一可译码

即时码

Kraft 不等式

最优码

最优码长的界

惟一可译码的
Kraft 不等式**定理 2.8: Sardinas-Patterson 定理**

编码 C 是唯一可译码, 当且仅当集合 C 和 C_∞ 的交为空集.

有关编码的几个例子

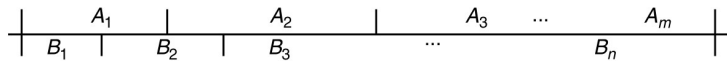
唯一可译码

即时码

Kraft 不等式

最优码

最优码长的界

惟一可译码的
Kraft 不等式

参考文献: A. A. Sardinas and C. W. Patterson, A necessary and sufficient condition for the unique decomposition of coded messages, IRE. Internat. Conv. Rec. 8 (1953), 104-108.

例 2.9

如果 $C = \{0, 01, 011\}$ (如例 1.8 和例 1.9 所示), 那么 $C_\infty = \{1, 11\}$, 它与 C 互不相交。正如我们之前已经看到的, 由定理 2.8 可知 C 是唯一可译码。

例 2.10

设 C 为三进制编码 $\{01, 1, 2, 210\}$. 我们求得 $C_1 = \{10\}$, $C_2 = \{0\}$, 以及 $C_3 = \{1\}$, 因此 $1 \in C \cap C_\infty$, 从而 C 不是唯一可译码 (没有必要再计算 $n > 3$ 时的 C_n). 为了找到一个非唯一译码的例子, 注意到 $10 \in C_1$ 是因为 $2 \in C$ 且 $2.10 = 210 \in C$; 接着 $0 \in C_2$ 是因为 $1 \in C$ 且 $1.0 = 10 \in C_1$; 然后 $1 \in C_3$ 是因为 $0 \in C_2$ 且 $0.1 = 01 \in C$. 将这些等式结合起来, 我们得到:

$$210.1 = 2.10.1 = 2.1.0.1 = 2.1.01$$

因此, 码序列 $t = 2101$ 既可以被解码为 210.1 , 也可以被解码为 $2.1.01$.

由于 Sardinas-Patterson 定理的证明相当长. 在此处, 我们将给出两个典型的推导过程来说明其中涉及的思想。

($\Rightarrow$) 假设 $C \cap C_\infty \neq \emptyset$, 不妨设 $w \in C \cap C_2$; 因此 $uw = v$, 其中 $u \in C$ 且 $v \in C_1$ (反之亦然, 为了简单起见, 我们假设第一种情况成立, 第二种情况类似). 那么 $u'v = v'$, 其中 $u', v' \in C$, 因此序列 $t = u'uw \in T^*$ 可以表示三个信源符号构成的序列 (因为 $u', u, w \in C$), 或者表示一个信源符号 (因为 $u'uw = u'v = v' \in C$)。因此解码不是唯一的。

有关编码的几个例子

唯一可译码

即时码

Kraft 不等式

最优码

最优码长的界

惟一可译码的 Kraft 不等式

($\Leftarrow$) 假设我们存在形式为 $t = u_1 u_2 = v_1 v_2$ 的非唯一解码实例, 其中 $u_1, u_2, v_1, v_2 \in C$. 我们不可能有 $|u_1| = |v_1|$, 因为这会导致 $u_1 = v_1$, 进而得到 $u_2 = v_2$. 因此, 如有必要可以通过重新编号, 不妨假设 $|u_1| > |v_1|$, 即 $u_1 = v_1 w$, 其中 $|w| > 0$. 于是 $w \in C_1$, 又因为 $w u_2 = v_2$, 所以 $u_2 \in C_2$. 因此 $u_2 \in C \cap C_\infty$, 这就说明 C 与 C_∞ 不相交.

例 2.11

考虑二元码 $C = \{0, 01, 011, 111\}$. 我们有 $C_1 = \{1, 11\}$ 且 $C_2 = \{1, 11\}$, 因此对于所有 $n \geq 1$ 有 $C_n = \{1, 11\}$; 从而 $C_\infty = \{1, 11\}$, 它与 C 不相交, 故根据 Sardinas-Patterson 定理, C 是唯一可译的。等待直到数字 1 的序列结束; 如果共有 k 个 1, 且满足 $k \equiv 0, 1$ 或 $2 \pmod{3}$, 则对应地 (唯一地) 解码为 $0.(111)^{k/3}$ 、 $01.(111)^{(k-1)/3}$ 或 $011.(111)^{(k-2)/3}$.

例 2.12

若 $\mathcal{C} = \{02, 12, 120, 20, 21\}$, 则由例 2.6 可得 $\mathcal{C}_\infty = \{0, 1, 2, 20\}$, 其中包含码字 20, 因此根据 Sardinas-Patterson 定理, $\mathcal{C}$ 不是唯一可译的. 例如, 序列 1202120 可以解码为 120.21.20 或 12.02.120.

例 2.13

若 $\mathcal{C} = \{02, 12, 120, 21\}$, 例 2.6 给出 $\mathcal{C}_\infty = \{0, 1, 2, 20\}$, 它与 $\mathcal{C}$ 不相交, 因此 $\mathcal{C}$ 是唯一可译的.

有关编码的几个例子

唯一可译码

即时码

Kraft 不等式

最优码

最优码长的界

惟一可译码的
Kraft 不等式

我们已经将唯一可译性定义为所有有限码序列 t 都可以被唯一解码, 但我们也可以考虑更强的要求, 即对于所有码序列 (无论有限还是无限) 这一结论都成立.

Even, Levenshtein 和 Riley 证明的一个定理表明, 当且仅当 $C \cap C_\infty = \emptyset$ 同时对于某个 $n \geq 1$ 有 $C_n = \emptyset$ 时, 该情况才会发生. 这些也是 C 成为有界延迟唯一可译码 (uniquely decodable with bounded delay) 的充要条件, 这意味着存在一个常数 d , 使得如果两个码序列的前 d 个符号相同, 则它们具有相同的第一个码字; 因此解码可以在延迟最多 d 个符号后开始.

有关编码的几个例子

唯一可译码

即时码

Kraft 不等式

最优码

最优码长的界

惟一可译码的 Kraft 不等式

有关编码的几个例子

唯一可译码

即时码

Kraft 不等式

最优码

最优码长的界

惟一可译码的
Kraft 不等式

在给出即时码的定义之前，我们先来看几个例子.

例 3.1

考虑由下式给出的二进制编码 C :

$$s_1 \mapsto 0, \quad s_2 \mapsto 01, \quad s_3 \mapsto 11.$$

我们有 $C_1 = C_2 = \dots = \{1\}$, 因此 $C_\infty = \{1\}$; 由此可得 $C \cap C_\infty = \emptyset$, 根据定理编码 C 是唯一可译码。现在假设我们接收到一个以 $t = 0111\dots$ 开头的有限消息。尽管我们知道该消息可以被唯一译码，但在读到这串连续 1 的末尾之前，我们无法开始译码：如果这块连续 1 的个数为偶数，则 t 的分解必然为 $0.11.11.11\dots$ ，译码后的消息必然以 $s = s_1 s_3 s_3 s_3 \dots$ 开头；然而，如果 1 的个数为奇数，则分解必然为 $01.11.11.11\dots$ ，因此 $s = s_2 s_3 s_3 s_3 \dots$ 。在实际应用中，这种译码延迟可能会带来困难。我们称编码 C 是非即时的。

有关编码的几个例子

唯一可译码

即时码

Kraft 不等式

最优码

最优码长的界

惟一可译码的

Kraft 不等式

有关编码的几个例子

唯一可译码

即时码

Kraft 不等式

最优码

最优码长的界

惟一可译码的
Kraft 不等式

现在给出即时码的正式定义.

定义 3.2

如果对于码字序列 $w_{i_1}, w_{i_2}, \dots, w_{i_n}$, 无论 t 后续的符号是什么, 任何以该序列开头的码序列 $t = w_{i_1} w_{i_2} \dots w_{i_n} \dots$ 都能被唯一解码为 $s = s_{i_1} s_{i_2} \dots s_{i_n} \dots$, 则称码 C 是**即时码**.

例 3.3

考虑由以下映射给出的二进制码 $\mathcal{D}$:

$$s_1 \mapsto 0, s_2 \mapsto 10, s_3 \mapsto 11,$$

它是例 3.1 中码 $\mathcal{C}$ 的反向码 (reverse code). 我们可以看出它是唯一可译的, 这既可以通过 Sardinas-Patterson 定理得出, 也可以因为 $\mathcal{C}$ 是唯一可译的而得出. 它同时也是即时码, 这意味着我们可以一边接收消息 t 一边进行解码: 数字 0 表示码字 w_1 (我们将其解码为 s_1); 数字 1 则表示 $w_2 = 10$ 或 $w_3 = 11$ 的开头, 一旦知道下一个符号, 就可以立即解码为 s_2 或 s_3 . 因此, 消息 t 中的任何码字一旦到达就可以被立即解码, 没有任何延迟.

有关编码的几个例子

唯一可译码

即时码

Kraft 不等式

最优码

最优码长的界

唯一可译码的 Kraft 不等式

有关编码的几个例子

唯一可译码

即时码

Kraft 不等式

最优码

最优码长的界

惟一可译码的
Kraft 不等式

定义 3.4

若码中无任何码字是其它码字的前缀，则称该编码为前缀码。

定理 3.5

编码 C 是即时码，当且仅当它是前缀码。

证明.

- ($\Rightarrow$) 如果 C 不是前缀码，设 w_i 是 w_j 的前缀，那么以 $t = w_i \dots$ 开头的代码序列在译码时既可以译为 $s = s_i \dots$ ，也可以译为 $s = s_j \dots$ ，因此 C 不是即时码。
- ($\Leftarrow$) 如果 C 是前缀码，且 t 以 $w_i \dots$ 开头，那么 s 必须以 s_i 开头，因为没有其他码字 w_j ($j \neq i$) 是 w_i 的前缀或以 w_i 为前缀。我们可以像这样继续下去，在接收到 t 中的后续码字时依次进行译码，因此 C 是即时码。 $\square$

有关编码的几个例子

唯一可译码

即时码

Kraft 不等式

最优码

最优码长的界

惟一可译码的
Kraft 不等式

为了理解即时码的构造, 将 T 中的字构成的集合 T^* 视为一个图是很有帮助的. 图是一个由点 (称为顶点) 构成的集合, 其中某些顶点对由边相连. 在这种情况下, 顶点就是字 $w \in T^*$, 并且每个字 w 都与由在 w 末尾添加单个符号 $t_i \in T$ 所形成的 r 个字 $wt_1, \dots, wt_r$ 通过边相连. 我们可以将这个图可视化为向上生长的树, 其中空字 ε 在底部, 长度为 l 的字位于 ε 上方第 l 层; 在图论中, 这样的图被称为 r 元有根树. (树是没有圈的连这里根是 ε .) 下图展示了当 $T = \mathbb{Z}_2$ 时, 直到第 $l = 3$ 层的二叉树 T^* .

有关编码的几个例子

唯一可译码

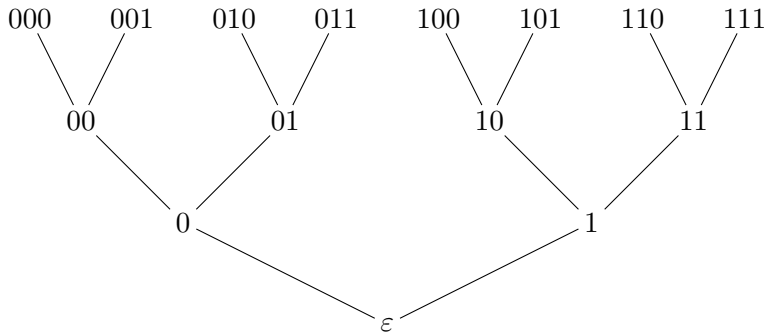
即时码

Kraft 不等式

最优码

最优码长的界

惟一可译码的 Kraft 不等式



有关编码的几个例子

唯一可译码

即时码

Kraft 不等式

最优码

最优码长的界

惟一可译码的
Kraft 不等式

码 C 可以被看作是树 T^* 的顶点的有限集合。当且仅当顶点 w_i 被顶点 w_j 支配（也就是说，在 T^* 中存在一条从 w_i 到 w_j 的向上路径）时，字 w_i 才是 w_j 的前缀。我们知 C 是即时码当且仅当 C 中没有任何顶点 w_i 被 C 中的另一个顶点 w_j ($i \neq j$) 所支配。我们可以利用这个准则来构造即时码，即在 T^* 中逐个选择顶点，使得没有任何顶点支配（或被支配于）任何先前选择的顶点。

例 3.6

让我们为含有五个符号 $s_1, \dots, s_5$ 的信源 S 寻找一个二元即时码 C . 首先, 让我们尝试 $s_1 \mapsto w_1 = 0$, 因此 0 是 C 中的一个顶点. 如果 C 要成为一个前缀码, 那么 C 中的其他任何顶点都不能支配 0, 因此它们必须全都支配 1 (即其他码字必须以 1 开头). 如果我们尝试 $s_2 \mapsto w_2 = 1$, 那么无法添加更多的码字, 因为它们会支配 w_1 或 w_2 . 相反, 让我们尝试 $s_2 \mapsto w_2 = 10$. 接着 $s_3 \mapsto w_3 = 11$ 是不可行的, 因为它不允许有进一步的码字, 所以让我们尝试 $s_3 \mapsto w_3 = 110$. 继续下去, 我们发现有以下可能性: $s_4 \mapsto w_4 = 1110$, $s_5 \mapsto w_5 = 1111$. 这就给出了一个二元即时码 $C = \{0, 10, 110, 1110, 1111\}$, 其字长分别为 $l_i = 1, 2, 3, 4, 4$, 如下图所示. (这并不是唯一的可能方案. 例如, 二进制码 $\{00, 01, 10, 110, 111\}$ 同样是即时的.)

有关编码的几个例子

唯一可译码

即时码

Kraft 不等式

最优码

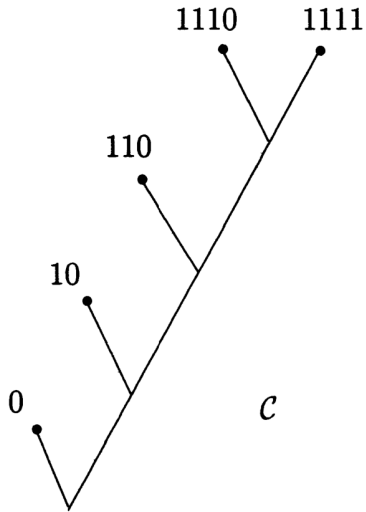
最优码长的界

惟一可译码的

Kraft 不等式

即时码

最优码长的界



例 3.7

是否存在一个针对该信源 S 、码长分别为 1, 2, 3, 3, 4 的即时二码? 同样地, 我们使用二叉树 T^* 。任何对长度 $l_1 = 1$ 的码字 w_1 的选择 (即高度为 1 的顶点), 都会排除 T^* 中一半的顶点作为后续码字 $w_2, \dots, w_5$ 的候选 (即所有支配/位于 w_1 子树下的顶点), 因此剩余的比例为 $1 - \frac{1}{2} = \frac{1}{2}$ 。在高度 2 处选择码字 w_2 , 会进一步排除 T^* 中 $1/2^2 = 1/4$ 的顶点, 剩余比例为 $1 - \frac{1}{2} - \frac{1}{4} = \frac{1}{4}$ 。在高度 3 处选择码字 w_3 和 w_4 , 会排除 T^* 中 $1/2^3 + 1/2^3 = 1/4$ 的顶点, 这样就没有剩余的顶点可供选择作为 w_5 。这里的困难在于, 每个 w_i 之上所占 T^* 比例的总和超过了 1:

$$\frac{1}{2} + \frac{1}{2^2} + \frac{1}{2^3} + \frac{1}{2^3} + \frac{1}{2^4} > 1$$

因此, 针对信源 S , 不存在码长为 $1, 2, 3, 3, 4$ 的即时二元码.

然而, 存在具有这些码长的即时三元码: 如果 $r = 3$, 那么对 $w_1, \dots, w_5$ 的选择将分别排除三叉树 T^* (其中 $|T| = 3$) 中比例为 $1/3, 1/3^2, 1/3^3, 1/3^3, 1/3^4$ 的顶点。由于:

$$\frac{1}{3} + \frac{1}{3^2} + \frac{1}{3^3} + \frac{1}{3^3} + \frac{1}{3^4} \leq 1$$

因此这样的选择是可能的.

有关编码的几个例子

唯一可译码

即时码

Kraft 不等式

最优码

最优码长的界

惟一可译码的 Kraft 不等式

有关编码的几个例子

唯一可译码

即时码

Kraft 不等式

最优码

最优码长的界

惟一可译码的
Kraft 不等式

有关编码的几个例子

唯一可译码

即时码

Kraft 不等式

最优码

最优码长的界

惟一可译码的
Kraft 不等式

定理 4.1: Kraft 不等式

对于 D 元字母表上的即时码 (前缀码), 码字长度 $l_1, l_2, \dots, l_m$ 必定满足不等式

$$\sum_i D^{-l_i} \leq 1$$

反之, 若给定满足以上不等式的一组码字长度, 则存在一个相应的即时码, 其码字长度就是给定的长度.

证明.

考虑每一节点均含 D 个子节点的 D 叉树. 假定树枝代表码字的字符. 例如, 源于根节点的 D 条树枝代表着码字第一个字符的 D 个可能值. 另外, 每个码字均由树的一片叶子表示. 因此, 始于根节点的路径可描绘出码字中的所有字符. 码字的前缀条件表明树中无一码字是其他任一码字的祖先. 因而, 在这样的编码树中, 每一码字都去除了其可能成为码字的所有后代.

令 l_{max} 为码字集中最长码字长度. 考虑在树中 l_{max} 层的所有节点, 可知其中有些是码字, 有些是码字的后代, 而另外的节点不是码字, 也不是码字的后代. 在树中 l_i 层的码字拥有 l_{max} 层中的 $D^{l_{max}-l_i}$ 个后代. 所有这些后代互不相交. 而且, 这些集合中的总结点数必定不超过 $D^{l_{max}}$. 因此, 对所有码字求和, 可得

$$\sum D^{l_{max}-l_i} \leq D^{l_{max}}.$$

于是

$$\sum D^{-l_i} \leq 1,$$

这就是 Kraft 不等式.

有关编码的几个例子

唯一可译码

即时码

Kraft 不等式

最优码

最优码长的界

惟一可译码的

Kraft 不等式

有关编码的几个例子

唯一可译码

即时码

Kraft 不等式

最优码

最优码长的界

惟一可译码的 Kraft 不等式

证明.

反之, 若给定任意一组满足 Kraft 不等式的码字长度 $l_1, l_2, \dots, l_m$, 总可以构造出如下的一个编码树: 将第一个深度为 l_1 的节点 (依字典序) 标为码字 1, 同时去除其所有后代. 然后在剩余的节点中找出第一个深度为 l_2 的节点, 将其标为码字 2, 同时去除树中所有属于它的所有后代, 等等. 按此方法继续下去, 则可构造出一个码字长度为 $l_1, l_2, \dots, l_m$ 的前缀码. $\square$

有关编码的几个例子

唯一可译码

即时码

Kraft 不等式

最优码

最优码长的界

惟一可译码的
Kraft 不等式

定理 4.2: 推广的 Kraft 不等式

对任意构成前缀码的可数无限码字集, 码字长度也满足推广的 Kraft 不等式:

$$\sum_{i=1}^{\infty} D^{-l_i} \leq 1.$$

反之, 若给定任意满足推广的 Kraft 不等式的 $l_1, l_2, \dots$, 则可构造出具有相应码长的前缀码.

证明.

我们不妨设 D 元字母表为 $\{0, 1, \dots, D-1\}$, 第 i 个码字为 $y_1 y_2 \cdots y_{l_i}$. 记 $0.y_1 y_2 \cdots y_{l_i}$ 是以 D 进制表示的实值小数, 即

$$0.y_1 y_2 \cdots y_{l_i} = \sum_{j=1}^{l_i} y_j D^{-j}.$$

由此, 这个码字对应一个区间

$$[0.y_1 y_2 \cdots y_{l_i}, 0.y_1 y_2 \cdots y_{l_i} + \frac{1}{D^{l_i}}),$$

这是一个实数集合, 集合中的所有数的 D 进制展开都以 $0.y_1 y_2 \cdots y_{l_i}$ 开始. 由前缀条件我们知所有这些区间互不相交. 因此, 它们的区间长度总和小于或等于 1. 因此我们证明了

$$\sum_{i=1}^{\infty} D^{-l_i} \leq 1.$$

有关编码的几个例子

唯一可译码

即时码

Kraft 不等式

最优码

最优码长的界

惟一可译码的 Kraft 不等式

有关编码的几个例子

唯一可译码

即时码

Kraft 不等式

最优码

最优码长的界

惟一可译码的
Kraft 不等式

证明.

反之, 和有限情形类似。首先将长度下表重新排列, 使得 $l_1 \leq l_2 \leq \dots$. 然后从单位区间的左端开始, 依次进行分配, 即可获得满足条件的码字集. 例如, 如果想构造一个二元编码使得 $l_i = i$, $i = 1, 2, \dots$, 那么将区间 $[0, \frac{1}{2})$, $[\frac{1}{2}, \frac{3}{4})$, $\dots$ 分配给字符, 使其对应码字为 $0, 10, \dots$. □

有关编码的几个例子

唯一可译码

即时码

Kraft 不等式

最优码

最优码长的界

惟一可译码的 Kraft 不等式

有关编码的几个例子

唯一可译码

即时码

Kraft 不等式

最优码

最优码长的界

惟一可译码的
Kraft 不等式

前面我们已经证明了满足前缀条件的任何一个码字集合满足 Kraft 不等式, 并且当一组码字长度集合满足 Kraft 不等式时, 存在这样的码字集, 它们的长度集合正好就是给定的长度集合. 下面我们来考虑求解前缀码的最小期望长度问题.

由前面的讨论我们知该问题等价于寻找一个前缀码, 它的码字长度集合满足 Kraft 不等式, 其码字期望长度 $l = \sum p_i l_i$ 达到最小. 这就化为了一个最优化的问题: 在所有整数 $l_1, l_2, \dots, l_m$ 上, 最小化

$$L = \sum p_i l_i$$

其约束条件为

$$\sum D^{-l_i} \leq 1.$$

有关编码的几个例子

唯一可译码

即时码

Kraft 不等式

最优码

最优码长的界

惟一可译码的 Kraft 不等式

先利用微积分的知识做一个简单的分析, 以此说明达到最小值时 l^* 应具有的形式. 取消 l_i 必须是整数的限制, 并假定约束条件中的等号成立. 于是, 利用拉格朗日乘子法, 将带约束的最小化问题转化为求

$$J = \sum p_i l_i + \lambda (\sum D^{-l_i})$$

的最小化问题. 关于 l_i 求微分, 可得

$$\frac{\partial J}{\partial l_i} = p_i - \lambda D^{-l_i} \ln D.$$

令偏导数为 0, 得

$$D^{-l_i} = \frac{p_i}{\lambda \ln D}.$$

将此代入约束条件中以求得合适的 λ , 可得 $\lambda = 1/\ln D$, 因而 $p_i = D^{-l_i}$. 即最优码长为

$$l_i^* = -\log_D p_i.$$

若可以取码字长度为非整数, 则此时的期望码字长度为

$$L^* = \sum p_i l_i^* = -\sum p_i \log_D p_i = H_D(X).$$

定理 5.1

随机变量 X 的任一 D 元即时码的期望长度必定大于等于熵 $H_D(X)$, 即

$$L \geq H_D(X)$$

当且仅当 $D^{-l_i} = p_i$, 等号成立.

有关编码的几个例子

唯一可译码

即时码

Kraft 不等式

最优码

最优码长的界

惟一可译码的
Kraft 不等式

定义 5.2

对于某个 n , 如果概率分布的每一个概率值均等于 D^{-n} . 则称这个概率分布是 D 进制的.

因此, 当且仅当 X 的分布是 D 进制的, 上述定理等号成立.

有关编码的几个例子

唯一可译码

即时码

Kraft 不等式

最优码

最优码长的界

惟一可译码的 Kraft 不等式

有关编码的几个例子

唯一可译码

即时码

Kraft 不等式

最优码

最优码长的界

惟一可译码的
Kraft 不等式

定理 5.3

设 $l_1^*, l_2^*, \dots, l_m^*$ 是关于信源分布 $\mathbf{p}$ 和一个 D 元字母表的一组最优码长, L^* 为最优码的相应期望长度 ($L^* = \sum p_i l_i^*$), 则

$$H_D(X) \leq L^* < H_D(X) + 1.$$

有关编码的几个例子

唯一可译码

即时码

Kraft 不等式

最优码

最优码长的界

惟一可译码的
Kraft 不等式

证明.

设 $l_i = \lceil \log_D \frac{1}{p_i} \rceil$, 则 l_i 满足 Kraft 不等式且

$$H_D(X) \leq L = \sum p_i l_i < H_D(X) + 1.$$

但由于 L^* 是最优码的期望长度, 它不大于 $L = \sum p_i l_i$. 再由定理5.1可知 $L^* \geq H_D$, 从而得证. □

定义 L_n 为每输入字符期望码字长度, 也就是说, 如果设 $l(x_1, x_2, \dots, x_n)$ 是与 $(x_1, x_2, \dots, x_n)$ 相应的二进制码字长度 (为简便起见, 在本小节余下的部分, 假定 $D = 2$), 则

$$L_n = \frac{1}{n} p(x_1, x_2, \dots, x_n) l(x_1, x_2, \dots, x_n) = \frac{1}{n} El(X_1, X_2, \dots, X_n).$$

将前面推导的界应用于此时的编码, 有

$$H(X_1, X_2, \dots, X_n) \leq El(X_1, X_2, \dots, X_n) < H(X_1, X_2, \dots, X_n) + 1.$$

若 $X_1, X_2, \dots, X_n$ 是独立同分布的, 因此

$H(X_1, X_2, \dots, X_n) = \sum H(X_i) = nH(X)$. 上式两边同除以 n , 得

$$H(X) \leq L_n < H(X) + \frac{1}{n}.$$

因此, 通过足够大的分组长度, 可以获得一个编码, 可以使其每字符期望码长任意地接近熵.

有关编码的几个例子

唯一可译码

即时码

Kraft 不等式

最优码

最优码长的界

惟一可译码的

Kraft 不等式

若此随机过程不是独立同分布的, 我们亦然可以类似讨论, 我们有估计

$$H(X_1, X_2, \dots, X_n) \leq El(X_1, X_2, \dots, X_n) < H(X_1, X_2, \dots, X_n) + 1.$$

上式两边同时除以 n , 且定义 L_n 为每字符期望描述长度, 可得

$$\frac{H(X_1, X_2, \dots, X_n)}{n} \leq L_n < \frac{H(X_1, X_2, \dots, X_n)}{n} + \frac{1}{n}.$$

如果随机过程是平稳的, 则 $H(X_1, X_2, \dots, X_n)/n \rightarrow H(\mathcal{X})$. 于是我们有如下定理.

定理 5.4

每字符最小期望码字长满足

$$\frac{H(X_1, X_2, \dots, X_n)}{n} \leq L_n^* < \frac{H(X_1, X_2, \dots, X_n)}{n} + \frac{1}{n}.$$

进一步, 若 $X_1, X_2, \dots, X_n$ 是平稳随机过程, 则

$$L_n^* \rightarrow H(\mathcal{X}).$$

其中 $H(\mathcal{X})$ 为随机过程的熵率.

有关编码的几个例子

唯一可译码

即时码

Kraft 不等式

最优码

最优码长的界

惟一可译码的

Kraft 不等式

最后我们讨论当面对的对象非真实分布时，期望描述长度会变得怎么样？下面考虑概率密度函数 $q(x)$ 的香农编码，相应的码长为 $l(x) = \lceil \log \frac{1}{q(x)} \rceil$. 假定真实分布的概率密度函数是 $p(x)$.

有关编码的几个例子

唯一可译码

即时码

Kraft 不等式

最优码

最优码长的界

惟一可译码的 Kraft 不等式

最后我们讨论当面对的对象非真实分布时，期望描述长度会变得怎么样？下面考虑概率密度函数 $q(x)$ 的香农编码，相应的码长为 $l(x) = \lceil \log \frac{1}{q(x)} \rceil$. 假定真实分布的概率密度函数是 $p(x)$.

定理 5.6: 偏码

码字长度分配 $l(x) = \lceil \log \frac{1}{q(x)} \rceil$ 关于 $p(x)$ 的期望码长满足

$$H(p) + D(p\|q) \leq E_p l(X) < H(p) + D(p\|q) + 1.$$

证明.
期望码长为

$$\begin{aligned}
 El(x) &= \sum_x p(x) \lceil \log \frac{1}{q(x)} \rceil \\
 &< \sum_x p(x) (\log \frac{1}{q(x)} + 1) \\
 &= \sum_x p(x) \log \frac{p(x)}{q(x)} \frac{1}{p(x)} + 1 \\
 &= \sum_x p(x) \log \frac{p(x)}{q(x)} + \sum_x p(x) \log \frac{1}{p(x)} + 1 \\
 &= D(p\|q) + H(p) + 1.
 \end{aligned}$$

类似地，可以得到期望码长的下界.

□

有关编码的几个例子

唯一可译码

即时码

Kraft 不等式

最优码

最优码长的界

惟一可译码的

Kraft 不等式

有关编码的几个例子

唯一可译码

即时码

Kraft 不等式

最优码

最优码长的界

惟一可译码的 Kraft 不等式

有关编码的几个例子

唯一可译码

即时码

Kraft 不等式

最优码

最优码长的界

惟一可译码的
Kraft 不等式

有关编码的几个例子

唯一可译码

即时码

Kraft 不等式

最优码

最优码长的界

唯一可译码的
Kraft 不等式**定理 6.1: McMillan**

任意唯一可译的 D 元码的码字长度必然满足 Kraft 不等式

$$\sum D^{-l_i} \leq 1.$$

反之, 若给定满足上述不等式的一组码字长度, 则可以构造出具有同样码字长度的唯一可译码.

证明.

考虑编码 C 的 k 次扩展 C^k (即原先惟一可译码 C 的 k 次串联所形成的码). 由惟一可译的定义, 该码的 k 次扩展是非奇异的. 由于所有长度为 n 的不同 D 元串的数目仅为 D^n , 故由惟一可译性可知, 在码的 k 次扩展中, 长度为 n 的码序列数目必定不超过 D^n . 由此讨论我们来证明 Kraft 不等式. 设字符 $x \in \mathcal{X}$ 所对应的码字长度记为 $l(x)$. 对于扩展码, 码序列的长度为

$$l(x_1, x_2, \dots, x_k) = \sum_{i=1}^k l(x_i).$$

我们的目标是要证明

$$\sum_{x \in \mathcal{X}} D^{-l(x)} \leq 1.$$

有关编码的几个例子

惟一可译码

即时码

Kraft 不等式

最优码

最优码长的界

惟一可译码的 Kraft 不等式

证明.

我们考虑上式左边的 k 次方, 从而有

$$\begin{aligned}
 \left(\sum_{x \in \mathcal{X}} D^{-l(x)} \right)^k &= \sum_{x_1 \in \mathcal{X}} \sum_{x_2 \in \mathcal{X}} \cdots \sum_{x_k \in \mathcal{X}} D^{-l(x_1)} D^{-l(x_2)} \cdots D^{-l(x_k)} \\
 &= \sum_{x_1, x_2, \dots, x_k \in \mathcal{X}^k} D^{-l(x_1)} D^{-l(x_2)} \cdots D^{-l(x_k)} \\
 &= \sum_{x^k \in \mathcal{X}^k} D^{(x^k)}.
 \end{aligned}$$

我们将上式右边按字符长度重新求和, 有

$$\sum_{x^k \in \mathcal{X}^k} D^{(x^k)} = \sum_{m=1}^{kl_{max}} a(m) D^{-m},$$

其中 l_{max} 是最大词长, $a(m)$ 为编码后码长为 m 的 x^k 的个数.

有关编码的几个例子

唯一可译码

即时码

Kraft 不等式

最优码

最优码长的界

唯一可译码的

Kraft 不等式

证明.

由码的惟一可译性, 我们有 $a(m) \leq D^m$, 从而我们有

$$\begin{aligned} \left(\sum_{x \in \mathcal{X}} D^{-l(x)} \right)^k &= \sum_{m=1}^{kl_{max}} a(m) D^{-m} \\ &\leq \sum_{m=1}^{kl_{max}} D^m D^{-m} \\ &= kl_{max}, \end{aligned}$$

从而

$$\sum_j D^{-l_j} \leq (kl_{max})^{1/k}.$$

由于上式对任意 k 成立, 当 $k \rightarrow \infty$ 时上式也成立. 由于 $(kl_{max})^{1/k} \rightarrow 1$, 我们有

$$\sum_{x \in \mathcal{X}} D^{-l(x)} \leq 1.$$

从而我们有 Kraft 不等式.

有关编码的几个例子

惟一可译码

即时码

Kraft 不等式

最优码

最优码长的界

惟一可译码的

Kraft 不等式



有关编码的几个例子

唯一可译码

即时码

Kraft 不等式

最优码

最优码长的界

唯一可译码的 Kraft 不等式

推论 6.2

可数无限字母表 $\mathcal{X}$ 上的一个唯一可译码也满足 Kraft 不等式.

有关编码的几个例子

唯一可译码

即时码

Kraft 不等式

最优码

最优码长的界

唯一可译码的 Kraft 不等式

证明.

由于唯一可译码的任一子集也是唯一可译码，因此无限码字集的有限子集亦满足 Kraft 不等式，故

$$\sum_{i=1}^{\infty} D^{-l_i} = \lim_{N \rightarrow \infty} \sum_{i=1}^N D^{-l_i} \leq 1.$$

反过来，给定满足 Kraft 的一组码字长度我们可以构造出相应的即时码，由于即时码是唯一可译的，所以我们就得到了所需要的唯一可译码。因而，McMillan 定理对于无限字母表情形同样成立。 □

注记 6.3

上面结论并没有告诉我们：码长为 $l_1, \dots, l_q$ 的 r 进制编码是即时码或唯一可译码的充要条件是 $\sum r^{-l_i} \leq 1$. 例如，二进制码 $C = \{0, 01, 011\}$ 的码长为 $l_i = 1, 2, 3$ ，因此 $\sum r^{-l_i} = \frac{7}{8} \leq 1$ ；然而， C 不是前缀码，因此它不是即时码。类似地，我们可以找到一个具有这些码长的二进制码，它不是唯一可译的： $\{0, 01, 001\}$ 就是一个明显的例子。