1 密码体制的概率分布

有关约定:

- 待加密后发送的所有可能消息的集合称为明文空间, 常用 M 来表示
- 所有密文的集合称为密文空间, 常用 C 表示
- 所有密钥的集合称为密钥空间, 常用 K 表示

  实际上, M,C 和 K 都是有限集. 算法确定后, 对于给定的 $m \in M, k \in K$, 则密文 c 唯一确定, 即 $c = E(m, k)$ 或 $c = E_k(m)$, E 是加密变换.

定义 1.1 *假设 X 与 Y 是随机变量, 一般的,*

用 $P(x)$ 表示 X 取值为 x 的概率, 即 $P(x) = PX = x$,

用 $P(y)$ 表示 Y 取值为 y 的概率, 即 $P(y) = PY = y$

用 $P(x, y)$ 表示 X 取值为 x 的概率且 Y 取值为 y 的概率集合, 即

$$P(x, y) = PX = x, Y = y$$

用 $P(x|y)$ 表示当 Y 取值为 y 时 X 取值为 x 的条件概率

- 用 $P(x, y) = P(x)P(y)$ 对所有可能的 X 的取值为 x 和 Y 取值为 y 成立, 则称随机变量 X 和 Y 是相互独立的.
- 联合概率和条件概率的关系: $P(x, y) = P(x)P(y|x) = P(y)P(x|y)$

定理 1.1 *贝叶斯定理 如果 $P(y) > 0$, 那么 $P(x|y) = \frac{P(x)P(y|x)}{P(y)}$*

证明 1.1 *设 x 与 y 是相互独立的随机变量, 当且仅当对所有的 x 和 y 有 $P(x|y) = P(x)$*

- 如果给定一个密码体制, 则关于他的明文, 密文与密钥的联合概率分布为 $P(m, c, k)$
- 由给定的密码体制的联合概率分布可以确定该体制的各种边际分布和条件分布, 并由此确定一些列信息的度量
- 常用边际分布与条件分布如下:

- 明文与密钥的联合概率分布为:

$$P(m,k)=\sum_{c\in C}P(m,c,k),m\in M,k\in K$$

- 明文与密文的联合概率分布为

$$P(m,c)=\sum_{k\in K}P(m,c,k),m\in M,c\in C$$

- 明文的概率分布为

$$P(m)=\sum_{k\in K}P(m,k),m\in M$$

- 密钥的概率分布为

$$P(k)=\sum_{m\in M}P(m,k),k\in K$$

- 密文的概率分布为

$$P(c)=\sum_{m\in M}P(m,c),c\in C$$

- 由联合概率分布与边际分布产生的条件概率分布为
  - 密文关于明文和密钥的条件概率分布为

$$P(c|m,k)=\frac{P(m,k,c)}{P(m,k)}$$

  - 密文关于明文的条件概率分布为

$$P(c|m)=\frac{P(m,c)}{P(m)}$$

  - 明文关于密文的条件概率分布为

$$P(m|c)=\frac{P(m,c)}{P(c)}$$

  - 密钥关于密文的条件概率分布为

$$P(k|c)=\frac{P(k,c)}{P(c)}$$

- 上述分布反映了密码体制中的数据结构关系.

1.1 熵

- 从密码分析者的角度看, 明文无不确定性, 密文则不然. 密文的不确定性程度随着密码分析的进行而逐渐减小, 直至完全确定.
- 不同的密码, 强度也不同; 而使用相同密钥加密的明文越多, 越有利于密码分析者进行" 唯密文攻击".
- 研究密文不确定性为难题的基本工具时熵的思想, " 熵" 是香浓在 1948 年在密码学中引进信息论时用到的概念.
- 熵被认为是信息的数学测度或者不确定性, 可以作为概率分布的函数进行计算, 即假定一个随机变量 X, 根据概率分布在一个有限集合上取值, 即
  $$P\{X=x\}=P_i, i=1,2,3\cdots,n$$
  根据分布发生的事件所获得信息是什么, 或事件还没有发生, 有关这个结果的不确行性是什么, 这个量称为 X 的熵, 表示为 $H(x)$.

定义 1.2 设 X 是一个随机变量, 他根据概率分布在一个有限集合上取值, 即 $PX=x=P_i, i=1,2,\cdots,n$, 那么这个概率分布的熵定义为

$$H(X)=-\sum_{i=1}^{n}P_i lbP_i$$

式中 $lbx=\log_2^x$

说明:

- 尽管 $\lim\limits_{x\to 0} xlbx=0$, 即允许 $x=0$, 但因为在熵的定义中, 当 $P_i=0$ 时,lbP_i 无定义, 所以假设上述定义中所有 $P_i\neq 0$.
- 熵定义中对数的底通常用 2, 因为 $lbP_i=-\log_a P_i\cdot lba$(其中 a 为常数), 所以计算熵时, 若改变对数的底, 熵的值只相差一个常数因子.
- 如果对 $1\leq i\leq n$, 有 $P_i=\frac{1}{n}$, 那么有 $H(x)=lbn$
- $H(x)\geq 0$ 和 $H(x)=0$ 当且仅当对某一个 i 有 $P_i=1$, 并且对所有的 $i\neq j$ 有 $P_j=0$

定理 1.2 对随机变量 X, X 的概率分布 $PX=x=P_i, i=1,2,\cdots,n$, X 的熵的基本性质为 $0\leq H(X)\leq lbn$

由上述定理可知, 当 $P_1 = P_2 = \cdots = P_n = \frac{1}{n}$ 时, 随机变量的熵取最大值, 即当各结果等概率时不确定性达到最大, 最难做出预测.

- 一个密码体制各个组成部分的熵:
  - 密钥概率分布相关的熵为
    $$H(K) = -\sum_{k\in K} P(k)lbP(k)$$
  - 明文概率分布相关的熵为
    $$H(M) = -\sum_{m\in M} P(m)lbP(m)$$
  - 密文概率分布相关的熵为
    $$H(C) = -\sum_{c\in C} P(c)lbP(c)$$
  - 明文和密文联合概率分布相关的熵为
    $$H(M,C) - \sum_{m\in M}\sum_{c\in C} P(m,c)lbP(m,c)$$

例如: 令 $M = \{a,b\}$, 有 $P(a) = \frac{1}{2}, P(b) = \frac{3}{4}; K = \{k_1,k_2,k_3\}$, 有 $P(k_1) = \frac{1}{2}, P(k_2) = \frac{1}{4}, P(k_3) = \frac{1}{4}; C = \{1,2,3,4\}$. 并假设加密函数定义如下:

$$E_{k_1}(a) = 1, E_{k_1}(b) = 2; E_{k_2}(a) = 2, E_{k_2}(b) = 3; E_{k_3}(a) = 3, E_{k_3}(b) = 4$$

这个密码体制可通过如下矩阵表示: $\begin{vmatrix} E_{k_j} & \text{a} & \text{b} \\ k_1 & 1 & 2 \\ k_2 & 2 & 3 \\ k_3 & 3 & 4 \end{vmatrix}$ 计算该密码体制的各组成部分的熵

解: 明文概率分布相关的熵为

$$\begin{aligned}H(M) &= -\sum_{m\in M} P(m)lbP(m)\\ &= -P(a)lbP(a) - P(b)lbP(b)\\ &= -\frac{1}{4}lb\frac{1}{4} - \frac{3}{4}lb\frac{3}{4}\\ &= -\frac{1}{4}\cdot(-2) - \frac{3}{4}(lb3-2)\\ &= 2 - \frac{3}{4}lb3\\ &\approx 0.81\end{aligned}$$

密钥概率分布相关的熵为:

$$\begin{aligned}H(K) &= -\sum_{k\in K} P(k)lbP(k)\\ &= -P(k_1)lbP(k_1) - P(k_2)lbP(k_2) - P(K_3)lbP(k_3)\\ &= -\frac{1}{2}lb\frac{1}{2} - \frac{1}{4}lb\frac{1}{4} - \frac{1}{4}lb\frac{1}{4}\\ &= \frac{1}{2} + 2\times\frac{1}{4} + 2\times\frac{1}{4}\\ &= 1.5\end{aligned}$$

密文概率分布相关的熵为:

因为 $H(C) = -\sum_{c\in C} P(c)lbP(c)$

所以 欲求$H(C)$, 必须先求出$P(1), P(2), P(3), P(4)$

因为 A,B 双方在进行加密通信之前, A 用预先确定的密钥加密明文, 同时在信道上发送产生的密文给 B, 即 A 知道明文之前就选择了密钥

所以 以下假设是合理的: 密钥 k 和明文 m 是相互独立的.

因为 $P(C) = \sum_{k\in K}\sum_{m\in M} P(m,k,c)$

所以

$$\begin{aligned}
P(1) &= P(a,k_1,1) = P(a)\cdot P(k_1) = \frac{1}{4}\times\frac{1}{2} = \frac{1}{8}\\
P(2) &= P(a,k_2,2) + P(b,k_1,2) = P(a)\cdot P(k_2) + P(b)\cdot P(k_1)\\
&\qquad\qquad\qquad\qquad\qquad\quad = \frac{1}{4}\times\frac{1}{4}+\frac{3}{4}\times\frac{1}{2}\\
&\qquad\qquad\qquad\qquad\qquad\quad = \frac{7}{16}\\
P(3) &= P(a,k_3,3) + P(b,k_2,3) = P(a)\cdot P(k_3) + P(b)\cdot P(k_2)\\
&\qquad\qquad\qquad\qquad\qquad\quad = \frac{1}{4}\times\frac{1}{4}+\frac{3}{4}\times\frac{1}{4}\\
&\qquad\qquad\qquad\qquad\qquad\quad = \frac{1}{4}\\
P(4) &= P(b,k_3,4) = P(b)\cdot P(k_3) = \frac{3}{4}\times\frac{1}{2} = \frac{3}{16}
\end{aligned}$$

最后得到

$$\begin{aligned}
H(C) &= -\sum_{c\in C} P(c)lbP(c)\\
&= -P(1)lbP(1) - P(2)lbP(2) - P(3)lbP(3) - P(4)lbP(4)\\
&= -\frac{1}{8}lb\frac{1}{8} - \frac{7}{16}lb\frac{7}{16} - \frac{1}{4}lb\frac{1}{4} - \frac{3}{16}lb\frac{3}{16}\\
&\approx 1.85
\end{aligned}$$

1.2 条件熵

- 密码学研究中感兴趣额时在获得某些密文的条件下, 对发送某些消息或使用某一密钥的不确定性测定. 为此定义暧昧度 (即条件熵) 如下:

定义 1.3 *设 X 和 Y 是两个随机变量, 则对 Y 的任何一个固定值 y, 都可达到一个 (条件) 概率分布 $P(x|y)$. 显然*

$$H(X|y) = -\sum_x P(x|y)lbP(x|y)$$

- 若定义条件熵 $H(X|Y)$ 是所有可能值 y 得熵 $H(X|y)$ 的加权平均 (关于概率 $P(y)$), 即

$$H(X|Y) = -\sum_y P(y)H(X|y) = -\sum_y\sum_x P(y)P(x|y)lbP(x|y)$$

- 条件熵测度通过 Y 来泄露有关 X 的信息的平均数

定理 1.3

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

证明 1.2 *若 X 与 Y 是相互独立的, 则*

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

证明 1.3 *若 X,Y 和 Z 是相互独立的, 则*

$$\begin{aligned} H(X,Y,Z) &= H(X,Y) + H(Z|X,Y) \\ &= H(X) + H(Y) + H(Z|X,Y) \\ &= H(X) + H(Y) + H(Z|X,Y) \end{aligned}$$

1.3 多余度和唯一解码量

- 把熵的结果应用到密码体制, 可以证明在密码体制的组成部分之间存在一个基本关系, 称为密钥暧昧度.

定理 1.4 *设 (M,C,K,E,D) 是一个密码体制, 那么*

$$H(K|C) = H(K) + H(M) - H(C)$$

例题续: 前面已计算 $H(M) \approx 0.81, H(K) \approx 1.5, H(C) \approx 1.85$, 于是有

$$H(K|C) = H(M) + H(K) - H(C) \quad \approx 0.81 + 1.5 - 1.85 = 0.46$$

- 假设 (M,C,K,E,D) 是一个正在使用的密码体制, 明文串 $m_1m_2\cdots m_n$, 用一个密钥加密, 产生一个密文串 $c_1c_2\cdots c_n$. 同时假设攻击者有无限的计算资源, 并知道明文为一个" 自然" 语言, 如英语.

- 对于任意密文, 用不同的密钥脱密可能得到一种以上有意义的译文, 有意义译文的数量越多, 判断哪一个是原文的难度就越大.

1.4 唯一解码量

1. K 中的密钥都是等概率的, 设 $K = \{k_1, \cdots, k_N\}$, N 是密钥数量, 所以

$$p = \frac{1}{N}, H(K) = -\sum_{k \in K} P(k) lbP(k) = -N \times \frac{1}{N} lb \frac{1}{N} = lbN$$

2. 长度为 n 的字符串共有 $t_n = 2^{r \cdot n}$, 其中有意义的明文数量 $S_n = 2^{r_n \cdot n}$. 所以

$$\begin{aligned} H(M) &= r_n \cdot n \\ H(C) &= r \cdot n \end{aligned} \tag{1}$$

3. 若 $H(K|C) = H(K) + H(M) - H(C) = 0$, 则表明对给定的密文, 密钥不存在不确定性.

   即字符数 n 使

$$H(K) + H(M) - H(C) = 0 \tag{2}$$

   将式 (1) 代入 (2) 得到

$$H(K) = (r - r_n)n \tag{3}$$

   (3) 式中的解便是唯一解码量, 用 u_d 表示令 $r^* = r - r_n$ 为语言的多余度, 则

$$H(K) = r^* \cdot u_d, u_d = \frac{H(K)}{r^*}$$

4. u_d 给出了破译一种密码所需要的最少字符串, 也就是确定密钥的最少密文字符数目

   例如: 对于单表置换密码, 密钥的数量为 26!

$$H(K) = lb26! \approx 88.38(bit)$$

设长度为 n 的明文, 密文串都取自英文字母表 $A = a, b, \cdots, z$

1. 则 $t_n = w^{r \cdot n} = 2^{4.7004n}$ 注: $t_n = |A|^n$, 而 $|A| = 26$, 令 $r = lb26 = 4.7004, |A| = 2^r$

2. 对于长度为 n 的有意义明文的数目, 有不同的估计值

- 当明文的长度充分大时, 设字母 $a, b, \cdots, z$ 出现的频数分别用 $n_a, n_b, \cdots, n_z$ 表示. 则明文的概率 p 为

$$p \approx p_a^{n_a} p_b^{n_b} \cdots p_z^{n_z}$$

其中 $p_a, p_b, \cdots, p_z$ 分别是字母 $a, b \cdots, z$ 出现的概率

- 令长度为 n 的有意义的明文数目为 S_n, 假设它们是等概率的, 即

$$p = \frac{1}{s} 或 S_n = \frac{1}{p}$$

同时假定 n 充分大时有

$$\begin{cases} n_a = n \cdot p_a \\ n_b = n \cdot p_b \\ \cdots\cdots \\ n_z = n \cdot p_z \end{cases}$$

则

$$\begin{aligned} lbS_n &= -lbp \\ &= -n(p_a lbp_a + p_b lbp_b + \cdots + p_z lbp_z) \\ &= -n(\sum_{\alpha=a}^{z} p_\alpha lbp_\alpha) \end{aligned}$$

令 $r_n = -\sum|\alpha = a^z p_\alpha lbp_\alpha$

根据英文字母的频率计算得到 $r_n = 4.19bit$

$$\begin{aligned} lbS_n &= r_n \cdot n = 4.19 \times 26 = 108.16 \\ S_n &= 2^{r_n \cdot n} = 2^{108.16} \\ r^* &= r - r_n = 4.7004 - 4.19 = 0.5104 \end{aligned}$$

所以

$$u_d = \frac{H(K)}{r^*} = \frac{88.38}{0.5104} = 173$$

即对单表置换密码, 唯一解码量为 173 个字符

3. 唯一解码量依赖于对语言多余度的估计, 归根到底是基于对有意义的报文概率的计算.