破译矩阵换位

Breaking Rectangular Transposition

刘卓

1 条件概率

条件概率 (Conditional Probability) 是指事件 A 已经发生了，发生事件 B 的概率。

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(B和A)}{\mathbb{P}(A)}$$

例 1

- 已知某病菌在人群中感染率约占 5%
- 某机构研究出一种检测试剂，准确率可达到 95%
- 假设某人使用该试剂，被检测出阳性，问实际感染率有多少?

解:
假设人群有 10000 人，也就是 500 人受到了感染；9500 人是健康的;

- 阳性：500 人受感染的人群中，被检测出阳性的人数为 $500 \times 0.95 = 475$人；被检测出阴性的人数为 $500 \times 0.05 = 25$ 人；
- 阴性 9500 健康人群中，被检测出阳性概率为 $9500 \times (1 - 0.95) = 475$人; 被检测出阴性的人数为 $9500 \times 0.95 = 9025$ 人；

$$\mathbb{P}(\text{真阳性}|\text{被检测出阳性}) = \frac{\mathbb{P}(\text{病人被确诊是真阳性})}{\mathbb{P}(\text{病人被检测出阳性})} = \frac{475}{475 + 475} = \frac{1}{2}$$

□

例 2

1. 给定一串密文，随机选择的字母，λ，是字母 A的概率是多少?

$$\mathbb{P}(\lambda = A) = \mathbb{P}(A) = 0.08399$$

2. 假设我们已知 λ 左边的字母是 μ, 并且知道字母 $\lambda = A$。即密文形式为 $***\mu\lambda****$。求 $\mu =$ "Q"概率是多少。

$$\mathbb{P}(\lambda = A|\mu = Q) = \frac{\mathbb{P}(\mu\lambda = QA)}{\mathbb{P}(\mu = Q)} = \frac{\mathbb{P}(\text{所有字母组合 QA 的总和})}{\mathbb{P}(\text{所有字母 Q 的总和})}$$

3. 假设字母 μ 和字母 λ 相距很远。$\mu = L$ 和 $\lambda = A$的概率是多少?

$$\mathbb{P}(\lambda = A\ \mu = Q) = \mathbb{P}(L)\mathbb{P}(A)$$

2 期望

在概率论和统计学中，一个离散性随机变量的期望值 (mean) 是试验中每次可能的结果乘以其结果概率的总和。

$$\mathbb{E}(X) = x_1\mathbb{P}(X = x_1) + \cdots + x_k\mathbb{P}(X = x_k)$$

例 3

掷出两个六面的色子，求两个正面的值的期望值。

解:

$(1,1) = 2$	$(1,2) = 3$	$(1,3) = 4$	$(1,4) = 5$	$(1,5) = 6$	$(1,6) = 7$
$(2,1) = 3$	$(2,2) = 4$	$(2,3) = 5$	$(2,4) = 6$	$(2,5) = 7$	$(2,6) = 8$
$(3,1) = 4$	$(3,2) = 5$	$(3,3) = 6$	$(3,4) = 7$	$(3,5) = 8$	$(3,6) = 9$
$(4,1) = 5$	$(4,2) = 6$	$(4,3) = 7$	$(4,4) = 8$	$(4,5) = 9$	$(4,6) = 10$
$(5,1) = 6$	$(5,2 = 7)$	$(5,3) = 8$	$(5,4) = 9$	$(5,5) = 10$	$(5,6) = 11$
$(6,1) = 7$	$(6,2) = 8$	$(6,3) = 9$	$(6,4) = 10$	$(6,5) = 11$	$(6,6) = 12$

两面和	2	3	4	$\cdots$	12
概率	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\cdots$	$\frac{1}{36}$

3 凸函数

如果函数 $y = f(x)$ 在区间 $[a, b]$ 满足 $f''(x) \geq 0, a \leq x \leq b$，以及 $f'(x)$ 在该区间是递增的。那么该函数是凸函数 (Convex Function)。

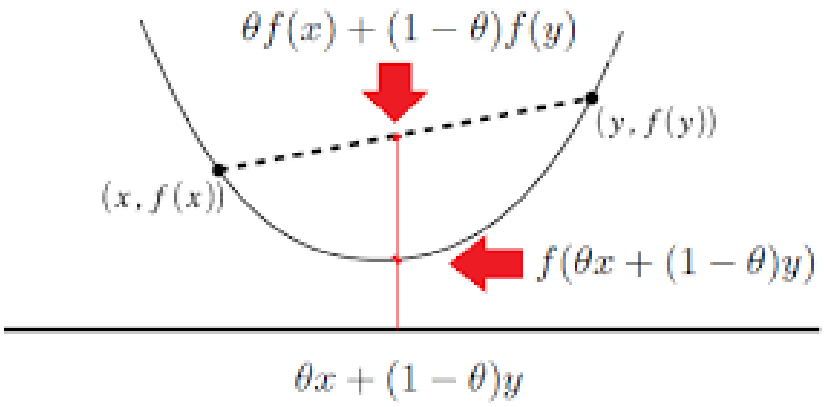

图 1: 凸函数

定理 *1.* 让 $x_1, x_2, \cdots, x_n \in [a, b]$，$p_1, p_2, \cdots, p_n \in R$ 并满足 $p_1 + p_2 + \cdots + p_n = 1$。如果函数 f 是凸函数，并且：

$$f(p_1x_1 + p_2x_2 + \cdots + p_nx_n) \leq p_1f(x_1) + p_2f(x_2) + \cdots + p_nf(x_n)$$

当且仅当 $x_1 = x_2 = \cdots = x_n$ 时成立。

该定理为琴生不等式（Jensen's Inequality），证明过程在这里

让 $f(x) = log(\frac{1}{x})$, 则：

$$log(\frac{1}{p_1x_1 + p_2x_2 + \cdots + p_nx_n}) \leq p_1log(\frac{1}{x_1}) + \cdots + p_nlog(\frac{1}{x_n})$$

定理 *2.* 让 $p_1, p_2, \cdots, p_n$ 是概率并满足 $p_1 + p_2 + \cdots + p_n = 1$。并且对于任意集合概率 $q_1, q_2, \cdots, q_n$ 并满足 $q_1 + \cdots + q_n = 1$, 那么

$$\sum_{i=1}^{n} p_ilog(q_i) \leq \sum_{i=1}^{n} p_ilog(p_i)$$

证明：让 $x_i = \frac{q_i}{p_i}$, 并使用定理 1.

4 破译过程

1. 猜测解密排列长度，比如密钥长度 k。
2. 将密文排列为 k 列，N 行的矩形。
3. 对于 $1 \leq i \neq j \leq k$，提取 i 列和 j 列并计算字母对 $\alpha\beta$ 的出现次数，并将其称为 $n^{ij}_{\alpha,\beta}$

4. 对于字母对 $\alpha\beta$，让 $\mathbb{P}_{\alpha,\beta}$为在英文或者其他语种出现的概率，计算

$$C_{i,j} = \sum_{\alpha\beta} \mathbb{P}_{\alpha,\beta} log(n_{\alpha,\beta}^{ij})$$

例 4

$$k = 10, N = 23, i = 3, j = 7$$

密文排列为：23 行，10 列；

E	C	T	I	H	N	O	H	G	I
O	K	R	O	B	C	A	O	H	F
E	I	N	S	G	N	N	S	A	A
E	T	C	N	I	I	E	C	N	H
O	A	S	R	E	E	H	C	T	L
H	S	A	A	T	E	I	B	N	E
S	F	N	E	U	C	N	O	E	R
R	E	T	I	U	S	S	S	A	A
R	E	O	C	U	W	S	O	I	F
M	N	D	A	O	D	I	D	V	A
T	E	C	H	E	X	O	T	T	E
H	O	F	E	T	C	E	R	L	A
I	I	A	T	S	O	E	S	M	S
M	S	T	E	I	O	N	K	W	N
N	I	C	S	O	S	F	S	O	T
X	Y	S	T	I	U	H	F	R	O
A	R	E	G	X	S	A	A	E	M
S	M	C	Y	H	L	Z	B	I	O
B	A	E	Y	D	R	I	P	T	A
L	R	C	A	U	R	N	A	A	R
M	N	G	E	E	F	I	T	S	O
T	A	X	R	S	H	A	I	T	G
B	O	N	R	D	N	I	K	L	E

解：

将 $i = 3$ 列，$j = 7$ 列提出来，变成两列

E	C	T	I	H	N	O	H	G	I		T	O
O	K	R	O	B	C	A	O	H	F		R	A
E	I	N	S	G	N	N	S	A	A		N	N
E	T	C	N	I	I	E	C	N	H		C	E
O	A	S	R	E	E	H	C	T	L		S	H
H	S	A	A	T	E	I	B	N	E		A	I
S	F	N	E	U	C	N	O	E	R		N	N
R	E	T	I	U	S	S	S	A	A		T	S
R	E	O	C	U	W	S	O	I	F		O	S
M	N	D	A	O	D	I	D	V	A		D	I
T	E	C	H	E	X	O	T	T	E		C	O
H	O	F	E	T	C	E	R	L	A		F	E
I	I	A	T	S	O	E	S	M	S	$\Rightarrow$	A	E
M	S	T	E	I	O	N	K	W	N		T	N
N	I	C	S	O	S	F	S	O	T		C	F
X	Y	S	T	I	U	H	F	R	O		S	H
A	R	E	G	X	S	A	A	E	M		E	A
S	M	C	Y	H	L	Z	B	I	O		C	Z
B	A	E	Y	D	R	I	P	T	A		E	I
L	R	C	A	U	R	N	A	A	R		C	N
M	N	G	E	E	F	I	T	S	O		G	I
T	A	X	R	S	H	A	I	T	G		X	A
B	O	N	R	D	N	I	K	L	E		N	I

统计 $TO, RA, NN, CE, \cdots, NI$ 出现的次数，所以 $n_{TO}^{3,7} = 1, n_{RA}^{3,7} = 1, n_{NN}^{3,7} = 2, \cdots, n_{NI}^{3,7} = 1$

$$\begin{aligned} C_{3,7} &= \mathbb{P}_{TO} log(n_{TO}^{3,7}) + \mathbb{P}_{RA} log(n_{RA}^{3,7}) + \cdots + \mathbb{P}_{NI} log(n_{NI}^{3,7}) \\ &= \mathbb{P}_{TO} \cdot log(1) + \mathbb{P}_{RA} \cdot log(1) + \mathbb{P}_{NN} \cdot log(2) \cdots + \mathbb{P}_{NI} \cdot log(1) \\ &= \mathbb{P}_{TO} \cdot 0 + \mathbb{P}_{RA} \cdot 0 + \mathbb{P}_{NN} \cdot log(2) \cdots + \mathbb{P}_{NI} \cdot 0 \end{aligned}$$

然后计算所有的 $C_{i,j}, i \neq j$ □

定义 $f_{\alpha,\beta}^{(i,h)} = \frac{n_{\alpha,\beta}^{(i,h)}}{N}$. 当 i 和 j 在明文中不是连续的, 则:

$$\begin{aligned} C_{ij} &= \sum_{\alpha,\beta} p_{\alpha\beta} \log\left(N \cdot f_{\alpha\beta}^{(ij)}\right) \\ &= \log(N) + \sum_{\alpha,\beta} p_{\alpha\beta} \log\left(f_{\alpha\beta}^{(ij)}\right) \end{aligned}$$

$$\leq \sum_{\alpha,\beta} p_{\alpha\beta} \log\left(p_{\alpha\beta}\right) \tag{1}$$

当明文中两列 i 和 j 连续时 C_{ij} 非常得小。因此，如果我们猜测 k 是正确的，则矩阵 $C_{ij}, 1 \leq i \neq j \leq k$ 除了第一行外，每一行的数量都大得多。

- 如果 C_{ij} 是第 i 行上的较大数字，则 j 在 i 后面跟随 i 解密排列。
- 如果第 k 行是唯一没有实质性条目的唯一行，则 k 是解密排列中的第一个条目。

密钥空间有 $26! \approx 2^{88}$

5 破译单字母替代

例 5

假设抛 1000 次色子，并记录每面朝上的值。能否确定是正确的?

结果	1	2	3	4	5	6
频数	171	186	174	170	192	107

使用卡方检验。

卡方检验（*Chi-Squared Test*）是一种统计量的分布在零假设成立时近似服从卡方分布的假设检验。在没有其他的限定条件或说明时，卡方检验一般指代的是皮尔森卡方检验。

$$X^2 = \sum_{i=1}^{k} \frac{(n_i - n \cdot_i)^2}{n \cdot p_i}$$

其中:

- k 是指将会出现多少种结果
- n_i 是被每个结果的频数
- p_i 是每个结果的概率
- n 是操作次数

在例 5 中，k=6,n_i 是频数，$p_i = \frac{1}{6}, n = 1000$;

$$X^2 = \sum_{i=1}^{6} \frac{(n_i - n \cdot p_i)^2}{n \cdot p_i} = \frac{\left(171 - 1000 \times \frac{1}{6}\right)^2}{1000 \times \frac{1}{6}} + \frac{\left(186 - 1000 \times \frac{1}{6}\right)^2}{1000 \times \frac{1}{6}} + \ldots = 27.95$$

对照表格,$X^2 \geq 27.95$ 是小于 0.001% 的，即表示拒绝原假设，有 99.999% 的把握说明这件事不可能发生！即这个表格可能是造假的。 □

例 6

已知是单字母替代算法，使用明文攻击。给定特定密文的字母频率如下:

密文	l	h	a	w	d	q	O	n	f	s	z
频数	80	61	55	46	44	40	39	35	33	26	22

k	p	i	t	v	y	r	x	u	m	c	g	j	b	e
26	22	18	17	12	11	9	9	8	7	5	3	1	0	0

并且已经知道单词 $WHERE$ 是明文，在密文中，找到两组字符串分别是 $HDFKF$和 $PDLHL$与 $WHERE$ 的结构一样。如何确定哪个字符串是与 $WHERE$ 匹配?

解:

第一步，计算每个字母出现的概率;

$Letter$	$Relative frequency(\%)$	$Letter$	$Relative frequency(\%)$
A	8.399	N	6.778
B	1.442	O	7.493
C	2.527	P	1.991
D	4.800	Q	0.077
E	12.150	R	6.063
F	2.132	S	6.319
G	2.323	T	8.999
H	6.025	U	2.783
I	6.485	V	0.996
J	0.102	W	2.464
K	0.689	X	0.204
L	4.008	Y	2.157
M	2.566	Z	0.025

$$\mathbb{P}_W = \mathbb{P}(W|W或H或E或R) = \frac{0.02464}{0.02464 + 0.06025 + 0.1215 + 0.06063} = 0.0923$$

$$\mathbb{P}_H = 0.226$$

$$\mathbb{P}_E = 0.455$$

$$\mathbb{P}_R = 0.227$$

下一步计算 $HDFKF$的卡方检验结果。

	$W = H$	$H = D$	$E = F$	$R = K$
P_i	0.0923	0.226	0.455	0.227
n_i	61	44	33	26

n = 164,k=4,n_i 是频数，p_i 是每个字母出现的概率
代入公式

$$X^2 = \frac{(61 - 169 \cdot 0.0923)^2}{169.0.0923} + \frac{(44 - 169 \cdot 0.226)^2}{169 \cdot 0.226} + \frac{(33 - 169.0.955)^2}{169.0.455} + \frac{(26 - 169 - 0.227)}{169.0.227} \approx 181.88$$

同理，计算 $PDLHL$:

$$X^2 \approx 6.59$$

对照表格，取 X^2 值小的字符串作为假设正确。P 值稍大，不拒绝原假设。即 $PDLHL$正确。