信息论

概率统计初步

概率：频率的理想化，或频率的极限（频率学派的概率）；由于知识的匮乏，人的主观的认识（Bayes学派的概率）。

随机事件：原子随机事件的组合。

原子随机事件：预先定义的，正交的，且概率和为1的一系列可能发生的事。

随机变量：样本空间到实数的实值函数。随机变量的函数依然是随机变量。

$$
\begin{gathered}
h: R \to R \\
r.v.\ X,\ h(X) \\
r.v.\ Y = f(X)
\end{gathered}
$$

概率分布：随机变量X的概率分布函数

$$
f(x) = p(X = x),\ \sum_k f(k) = 1,\ f(k) \geq 0
$$

期望

$$
\begin{aligned}
r.v.\ X,\ Y &= h(X) \\
EX &= \sum_k k \cdot p(X = k) \\
EY &= \sum_k k \cdot p(h(X) = k) \\
&= \sum_{k'} h(k') \cdot p(X = k')
\end{aligned}
$$

方差

$$
Var(X) = E[(X - EX)^2]
$$

熵和信源编码

编码的基本准则：可恢复，编码后的信息长度的期望尽量得短，完成从源消息到传输消息的映射。

无前缀码：编码中任何编码都不是其余编码的前缀。无前缀码可以唯一解码，但并不是所有唯一解码的编码都是无前缀码。**无前缀码是最优的无误的唯一解码的编码方式。**

Kraft不等式：对于使用0-1进行编码的无前缀码{c1, c2, .., cn}，假设各个码字ci的长度为{l1, l2, ..., ln}，则必然有如下不等式。并且当编码最优时，等号成立。

$$
\sum_{i=1}^{n} 2^{-l_i} \leq 1
$$

证明方法：利用二叉树表示无前缀码，则码字终结必然在叶节点。当且仅当二叉树满时，等号成立；若二叉树不满则编码不是最优。

Huffman编码：根据码字出现的概率进行编码，概率最小的两个码字合为一个“新的码字”，不断重复直至构造出满的Huffman树。

$$p_1 \geq p_2 \geq \cdots \geq p_n \Rightarrow l_1 \leq l_2 \leq \cdots \leq l_n$$
$$Kraft\text{不等式} \Rightarrow \sum_{i=1}^{n} 2^{-l_i} = 1 \Rightarrow \text{满二叉树}$$
$$\text{如上等式} \Rightarrow l_n = l_{n-1} \Rightarrow \text{转动二叉树可以使得}n\text{和}n-1\text{码字称为兄弟}$$
$$\text{熵的加性} \Rightarrow r.v.X, p_1, \cdots, p_n; r.v.Z, p_1, \cdots, p_{n-1}, q_1, q_2; p_n = q_1 + q_2$$
$$Z\text{的最佳编码中，将}q_1\text{与}q_2\text{合并为}p_n\text{即可得到}X\text{的最佳编码}$$

熵：离散的随机变量X，其熵定义如下，单位是bit（若使用ln则单位是nat）。**熵意味着随机信源编码的平均最短长度，出现概率越小的message的信息量越大。**

$$H(X) = \sum_i p_i \log \frac{1}{p_i}$$

问题：随机信源X，分布P={p1, ..., pn}，编码C=(c1, ..., cn)，编码长度L=(l1, ..., ln)，给定X（即给定P），设计C和L使得EL最小，且Kraft不等式取等号。

解法：Lagrange法求解带等式约束的最小化问题。

$$\text{令 } q_i = 2^{-l_i}\text{，则问题转化为 } \min_q \sum_i p_i \log \frac{1}{q_i}$$
$$\sum_i p_i \log \frac{1}{q_i} \geq \sum_i p_i \log \frac{p_i}{q_i} \geq (\sum_i p_i) \log \frac{(\sum_i p_i)}{(\sum_i q_i)} = 0$$
$$\text{当且仅当}p_i = q_i\ \forall i\text{时，等号成立。}$$

结论：li=log(1/pi)时，EL最小，且EL=H(X)

熵的性质：

对称性：H(p1, ..., pn)=H(pi1, ...,pin)，i1, ..., in是1, ..., n的排列。

最大值：均匀分布（平均分布）的熵值最大。

熵的加性：随机变量X，(p1, p2, ..., pn)，pn=q1+q2；随机变量Y，(r1, r2)，r1=q1/(q1+q2)，r2=q2/(q1+q2)；随机变量Z，(p1, p2, ..., p{n-1}, q1, q2)，则有下式。X是“粗粒度”的分布，Y是按照Z对X中pn的精细刻画，Z是“细粒度”的分布。

$$H(X) + p_n H(Y) = H(Z)$$

联合熵：随机变量X，Y，P(X,Y)=(pij)

$$H(X,Y) = \sum_{i,j} p_{ij} \log \frac{1}{p_{ij}}$$
$$\text{独立的}X\text{和}Y \Rightarrow H(X) + H(Y) = H(X,Y)$$

条件熵：随机变量X，Y，条件分布P(Y|X=i)。条件熵H(Y|X)衡量给出X后Y的信息量。

$$H(Y|X=i)=\sum_j P(Y=j|X=i)\log\frac{1}{P(Y=j|X=i)}$$

$$H(Y|X)=\sum_i P(X=i)H(Y|X=i)=H(Y)-H(X,Y)$$

$$\text{独立的}X\text{和}Y\Rightarrow H(Y|X)=H(Y)$$

$$Y=X\Rightarrow H(Y|X)=0$$

$$\text{链式法则：}\begin{matrix}H(X,Y)=H(X)+H(Y|X)=H(Y)+H(X|Y)\\ H(X_1,\cdots,X_n)=H(X_1)+H(X_2|X_1)+\cdots+H(X_n|X_1,\cdots,X_{n-1})\end{matrix}$$

互信息：随机变量X，Y，I(X;Y)=H(X)-H(X|Y)=H(Y)-H(Y|X)。互信息衡量X中包含Y的信息量，或Y中包含了X的信息量。

$$I(X;Y)=H(X)-H(X|Y)=H(Y)-H(Y|X)=H(X)+H(Y)-H(X,Y)$$

$$I(X;Y)=\sum_{i,j}P(X=i,Y=j)\log\frac{P(X=i,Y=j)}{P(X=i)P(Y=j)}$$

$$\text{独立的}X\text{和}Y\Rightarrow I(X;Y)=0$$

$$X=Y\Rightarrow I(X;Y)=H(X)=H(Y)=\max I(X;Y)$$

KL散度：概率分布P=(p1, ..., pn)和Q=(q1, ..., qn)。KL散度衡量两个分布的相似度。熵只是在离散变量上定义，而KL散度在任何分布上都可以使用。

$$D(P\|Q)=\sum_i p_i\log\frac{p_i}{q_i}\geq 0\text{，当}P=Q\text{时等号成立}$$

$$D(P\|U)=\log n-H(X),\ r.v.X,P;U\sim U(n)$$

$$I(X;Y)=D(P(X,Y)\|P(X)\cdot P(Y))$$

Markov性质：p(Z|X,Y)=p(Z|Y)

Markov过程：x1, x2, ..., xt为随机过程，满足。

$$\forall t,p(x_t|x_1,x_2,\cdots,x_{t-1})=p(x_t|x_{t-1})$$

Post-processing不等式：若X，Y，Z满足Markov性质，则有

$$I(X;Z)\leq I(X;Y)$$

微分熵和熵率

Huffman码长的界：H(X)+1 > L(Huffman) >= H(X)

$$\lim_{k\to\infty}\frac{H(X_1,\cdots,X_k)}{k}=H(X),\quad X_1,\cdots,X_k,\ i.i.d.$$

熵率

$$Entropy\ rate=\lim_{t\to\infty}\frac{H(X_1,\cdots,X_t)}{t}$$

$$\lim_{t\to\infty}\frac{H(X_1,\cdots,X_t)}{t}=\lim_{t\to\infty}H(X_t|X_1,\cdots,X_{t-1}),\ when\ limit\ exists$$

微分熵：连续变量上，对熵的推广。h(X)与连续随机变量中包含的信息量有关。

$$h(X)=-\int f_X(x)\log f_X(x)dx$$

连续变量的离散化：连续随机变量X，按△的粒度进行离散化，得到离散的随机变量X'。△越小，H(X')越大，最终趋于无穷。**离散化会给连续变量带来额外的信息。**

$$H(X'_{\Delta})\approx h(X)+\log\frac{1}{\Delta},\ when\ \Delta\ is\ small.$$

离散随机变量X，则Y=aX (a>0)也是随机变量，并且H(Y)=H(X)。

连续随机变量X，则Y=aX (a>0)也是随机变量，并且h(Y)=h(X) - loga.

连续变量的KL散度（相对熵）：连续随机变量X，f；Y，g，其KL散度定义如下。**KL散度不受离散化的影响，即连续变量进行相同的离散化，KL散度不变。**

$$D(f\|g)=\int_{-\infty}^{\infty}f(x)\log\frac{f(x)}{g(x)}dx$$

互信息：连续随机变量X，Y，其互信息定义如下。**互信息不受离散化的影响，即连续变量进行相同的离散化，互信息不变**（因为互信息是特殊的KL散度）。

$I(X;Y)=h(X)-h(X|Y)=h(Y)-h(Y|X)=h(X)+h(Y)-h(X,Y)$

Kolmogorov复杂度

图灵机：视为一种计算模型，其本身就是一个“程序”或“算法”。

通用图灵机：可以模拟其它任何一种图灵机的图灵机。

可计算性：[0,1)之间所有实数是否可数？否，证明方法为反证法。

$$\text{假设可数，则将所有实数列成一个}0-1\text{编码的矩阵}\\ \text{取对角线序列后，逐位取反，得到的新序列不会出现在矩阵中。}$$

可判定性：对所有的串x属于语言L，判断x是否属于L的子集L0，是否存在程序P可以完成该判断？否，证明方法如下。

$$L_0\subset L\text{不可数，即语言无限不可数（对角线法可证），}\\ \text{但}P\text{可数，因此不可能由可数的程序判定不可数的语言。}$$

0-1语言：符号表为{0,1}，语言为{0,1}*，其中每一个串都是0-1串。

Kolmogorov复杂度：通用图灵机u，描述0-1串s的程序的最短程序的长度。使用不同的u，K(s)均不同，但由于通用图灵机可以相互转换，因此不同的s的K(s)的相对大小不会发生改变。

$$K_u(s)=\min_P|P|,\ \text{其中}P\text{为}u\text{上程序且}u(P)=s$$

使用通用图灵机的原因：对任意的通用图灵机u和u'，对任意的0-1串s，有

$$K_u(s)\le K_{u'}(s)+C,\ C\text{为常数且只依赖于}u\text{和}u'$$

停机问题：给定程序P和输入I，是否存在通用图灵机H，可以判定P在I输入时能否停机？否，反证法证明。

$$\begin{aligned}&\text{假设存在}H(P,I)\text{，则令}I=P\text{，有}\\&\quad H'(P)=H(P,P)=1\text{，如果}P\text{在}P\text{输入停机}\\&\quad H‘(P)=H(P,P)=0\text{，如果}P\text{在}P\text{输入不停机}\\&\text{那么构造另一图灵机}\bar{H}\text{如下}\\&\quad \bar{H}'(P)=\bar{H}(P,P)=\text{死循环，如果}P\text{在}P\text{输入停机}\\&\quad \bar{H}'(P)=\bar{H}(P,P)=0\text{，如果}P\text{在}P\text{输入不停机}\\&\text{最后，考虑}\bar{H}'(\bar{H}')\text{的输出，出现矛盾。}\end{aligned}$$

Kolmogorov复杂度的不可计算性：类似停机问题的反证法证明。

$$\begin{aligned}&\text{假设}Kolmogorov\text{复杂度可计算，}\\&\quad\text{则存在函数}ComputeK\text{接收}0-1\text{串输入，输出其}Kolmogorov\text{复杂度}\\&\quad\text{令}L=ComputeK(ComputeK)=\|ComputeK\|\\&\text{考虑另一程序}FindFirstK\text{，输入为常数}K\text{，}\\&\quad\text{输出按序排列的第一个}Kolmogorov\text{复杂度大于}K\text{的}0-1\text{串}\\&\quad FindFirstK(K):\\&\quad\quad for\ i\ from\ 0\ to\ \infty:\\&\quad\quad\quad for\ s\ in\ 'OrderedSequencesWithLeni':\\&\quad\quad\quad\quad if\ ComputeK(s)\geq K:\\&\quad\quad\quad\quad\quad return\ s\\&\quad\text{当}K>L+\|FindFirstK\|\text{时，}\\&\quad\quad\text{输出的串是第一个}Kolmogorov\text{复杂度大于}K\text{的串}\\&\quad\quad\text{而整个}FindFirstK\text{长度却为}L+\|FindFirstK\|<K\text{，出现矛盾}\end{aligned}$$

最大熵原理

信道编码

有噪声信道：无噪声信道是不可能的，有噪声信道时是必然的，噪声会有概率引起消息出错，可以采用多数投票的方式进行解码，那么解码后消息出错的概率为

$$p(error)=\sum_{k>\lfloor\frac{n}{2}\rfloor}^{n}\binom{n}{k}p^k(1-p)^{n-k}\text{，}n\text{为编码长度}$$

信道编码：信源编码到信道上消息的映射。

$$\{0,1\}^n\to\{0,1\}^m(m>n)$$

Hamming距离：码字不同的位的数目。编码可以容忍的最多的错误位数为Hamming距离的一半（向下取整）。

$$d_{Hamming}=(CodeWord_1,CodeWord_2)=l$$
$$ErrorTorlerance=\lfloor\frac{l}{2}\rfloor$$

矩生成函数：用于生成随机变量X的各阶矩，是t的函数。

$$f(t)=Ee^{tX}=1+t\cdot EX+\frac{t^2}{2}\cdot EX^2+\frac{t^3}{3!}\cdot EX^3+\cdots$$

Markov不等式：对任意ϵ>0，有如下不等式。

$$P(e^{tX} \geq e^{t\cdot\epsilon}) \leq \frac{Ee^{tX}}{e^{t\cdot\epsilon}}$$

Chernoff边界：(X, X1, X2, ..., Xn)输入[0,1] i.i.d. EX = p。Chernoff边界说明了编码的错误率有明确的上界。

$$\text{相对熵}Chernoff\text{边界：}P(\frac{1}{n}\sum_{i=1}^{n}X_i \geq p+\delta) \leq e^{-nD_B^{(e)}(p+\delta\|p)},$$
$$\text{加性}Chernoff\text{边界：}P(\frac{1}{n}\sum_{i=1}^{n}X_i - p \geq \delta) \leq e^{-2n\delta^2}$$

证明如下：

$$\begin{aligned}
&\text{假设所有}X_i\text{为独立同分布二项分布}\\
&\forall t>0, P(\sum_{i=1}^{n}X_i \geq n(p+\delta)) \leq e^{-n(p+\delta)}\cdot Ee^{t\sum_{i=1}^{n}X_i}\\
&\quad = e^{-n(p+\delta)}(Ee^{tX})^n\\
&\quad Ee^{tX} = p\cdot e^t + (1-p) \Rightarrow\\
&\quad P(\sum_{i=1}^{n}X_i \geq n(p+\delta)) \leq e^{-n(p+\delta)}\cdot(p\cdot e^t+(1-p))^n\\
&Define\ D_B^{(e)}(p\|q) = p\ln\frac{p}{q} + (1-p)\ln\frac{1-p}{1-q},\\
&\qquad where\ P=(p,1-p), Q=(q,q-1)\\
&\quad P(\frac{1}{n}\sum_{i=1}^{n}X_i \geq p+\delta) \leq e^{-nD_B^{(e)}(p+\delta\|p)},\ Relative\ entropy\ Chernoff\ bound\\
&\quad D_B^{(e)}(p+\delta\|p) \geq 2\delta^2 \Rightarrow\\
&\quad P(\frac{1}{n}\sum_{i=1}^{n}X_i - p \geq \delta) \leq e^{-2n\delta^2},\ Additive\ Chernoff bound
\end{aligned}$$

Gilbert-Vashamov边界：对于0<δ<1/2，在空间$\{0,1\}^m$中存在2^n个码字，C1, C2, ..., C{2^n}，使得dHamming(Ci, Cj)>=δm，若m，n，δ满足如下条件。

$$m \geq \frac{2n}{H(\delta)},\ H(\delta) = \delta\log\frac{1}{\delta} + (1-\delta)\log\frac{1}{1-\delta}$$

证明如下：使用概率的方法证明。

$$\begin{aligned}
&\text{随机生成}C_1, C_2, \cdots, C_{2^n},\\
&\text{考虑}P(\forall i\neq j, d_H(C_i,C_j)\geq\delta m) > 0\text{是否成立}\\
&\text{固定}i\text{和}j,\\
&P(d_H(C_i,C_j)\leq\delta\cdot m) = \frac{\sum_{k=0}^{\delta\cdot m}\binom{m}{k}}{2^m}\\
&\quad \leq 2^{-m[1-H(\delta)]},\ Chernoff\ bound\\
&\text{对所有的}C_i\text{和}C_j,\\
&P(\exists i\neq j, d_H(C_i,C_j)\leq\delta\cdot m) \leq \binom{2^n}{2}\cdot 2^{-m[1-H(\delta)]}\\
&\text{令}\binom{2^n}{2}\cdot 2^{-m[1-H(\delta)]} < 1,\ P(\forall i\neq j, d_H(C_i,C_j)\geq\delta m) > 0\text{成立}
\end{aligned}$$

Hamming编码：各个码字间Hamming距离均大于等于3bit，编解码效率很高。

7-4Hamming码：定义H如下，H=[P, I]，P的列的顺序无关；之后根据P构造G如下，G=[I,P^T]。G和H满足G·H^T=0**。

$$\text{定义矩阵}H=\begin{bmatrix}1&1&0&1&1&0&0\\1&0&1&1&0&1&0\\0&1&1&1&0&0&1\end{bmatrix}_{3\times 7}$$

$$dim(Ker(H))=4,\text{其核空间中有}16\text{个元素}$$

$$\min_{i,C_i\neq 0}\|C_i\|=3\Rightarrow\min_{i\neq j}d_H(C_i,C_j)\geq 3\ bits$$

$$\text{根据}H\text{构造矩阵}G=\begin{bmatrix}1&0&0&0&1&1&0\\0&1&0&0&1&0&1\\0&0&1&0&0&1&1\\0&0&0&1&1&1&1\end{bmatrix}_{4\times 7}$$

7-4Hamming码编码过程：GF(2)^2→Ker(4)

$$s\in GF(2)^4\text{，编码}s\cdot G=Concat(s,s\cdot P^T)$$

7-4Hamming码解码过程：先判断编码是否正确，若不正确则纠错，完成后直接截取前半部分。

$$\begin{aligned}x&=C_i+Noise\\Hx&=0\Rightarrow Match\text{！}\\Hx\neq 0\Rightarrow Hx&=HC_i+H\cdot Noise=H_i,\ if\ \|Noise\|\leq 1\end{aligned}$$

通信复杂度

Protocol设计：Alice，$x\in\{0,1\}^n$；Bob，$y\in\{0,1\}^n$。Alice与Bob彼此不知道对方的串，设计流程用于计算 $f(x,y)\in\{0,1\}$

确定性通信：最坏情况下，判定相等需要n bit。

$$D(EQ)=n,\ EQ\text{中}f(x,y)=1\text{当且仅当}x=y$$

随机性通信：选取素数p∈ $[n^3,n^4]$作为Alice和Bob共有的“种子”，由于素数定理，该区间不重要，保证其中存在素数即可。

$$\begin{aligned}Alice&\text{：}a_0,a_1,\cdots,a_{n-1}\\Bob&\text{：}b_0,b_1,\cdots,b_{n-1}\\Alice&\text{：随机选取}x\in\{0,\cdots,p-1\},\\&\quad\text{计算}y=(a_0+a_1x+\cdots+a_{n-1}x^{n-1})\mod p\\&\quad\quad\text{发送}x\text{和}y\text{给}Bob\\Bob&\text{：检查}(b_{n-1}x^{n-1}+\cdots+b_1x+b_0)\mod p=_?y\\&\quad\text{若不成立，则}Bob\text{确认}a\neq b\text{，并且不会发生错误}\\&\quad\text{若成立，则}Bob\text{认为}a=b\text{，该情况下有可能发生错误}\\&\quad\quad\text{若发生错误，则}a\neq b\text{，但}\\&\quad\quad(a_{n-1}-b_{n-1})x^{n-1}+\cdots+(a_1-b_1)x+(a_0-b_0)=0\\&\quad\quad\text{以上}x\text{的方程至多存在}n-1\text{个解（}n-1\text{阶方程），而}x\text{可选数目为}p\\&\quad\quad\text{因此，}p(error)\leq\frac{n-1}{p}<\frac{1}{n^2}\\&\text{综上，}Bob\text{知道}Alice\text{的串的概率}P>1-n^{-2}\end{aligned}$$

通信复杂度的下界：通信过程中，Alice和Bob实际上是在一个$2^n \times 2^n$方阵M中不断缩小范围，确认取值。当“取值块”中取值唯一时便可以确定f的结果结束通信了。将方阵中取值相同的点组成的矩形称作“块”，方阵中正交且覆盖方阵的块的数目的最小值记作$x(f)$。则通信复杂度最坏情况下的下界如下。

$$CC(f) \geq \log_2 x(f) \leq rank(M(f))$$

通信复杂度的上界

$$CC(f) \leq O(\log^2 x(f))$$

信道容量

$$SrcMsg \xrightarrow[coding]{src} SrcCode \xrightarrow[coding]{channel} CodeWord \xRightarrow[channel]{noisy} \xrightarrow[decoding]{channel\ code} \xrightarrow[decoding]{src\ code}$$

信道容量：给定信道P(Y|X)，信道容量C定义如下。

$$C = \max_{P(X)} I(X;Y)$$

信道编码定理：对于有噪声信道P(Y|X)，信道容量为C。若输入的平均bit数小于等于C，那么存在信道编码方法，可以使得解码错误可以任意小；若输入的平均bit数大于C，则不存在信道编码方法，使得解码错误任意小。

渐近等分性（AEP）：随机变量X，$f:\mathbb{R} \to \mathbb{R}$，概率分布函数p(x)，AEP表述如下。AEP说明了当随机变量的序列足够长时，其中一部分序列中各个符号的出现频数非常接近于各自的出现概率，而这些序列的概率则趋近于相等，且它们的和非常接近于1 。

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n}\{-\log p(X_i)\} = E\{-\log p(X)\} = H(X)$$
$$\log p(X_1,\cdots,X_n) \in -n[H(X)\pm\epsilon]\text{，大概率情况下}$$

典型序列：x1, ..., xn是典型序列，当且仅当其满足如下条件。

$$-\frac{1}{n}\log p(x_1,\cdots,x_n) \in [H(X)\pm\epsilon]$$
$$\text{或 } p(x_1,\cdots,x_n) \in 2^{-n[H(X)\pm\epsilon]}$$
$$\text{或 } p(x_1,\cdots,x_n) \approx 2^{-nH(X)}$$

典型序列集合的性质：典型序列集合$A_\epsilon^{(n)}$，其满足如下性质。

$$P(A_\epsilon^{(n)}) \approx 1$$
$$\forall(x_1,\cdots,x_n) \in A_\epsilon^{(n)}, P(x_1,\cdots,x_n) \approx 2^{-nH(X)}$$
$$\|A_\epsilon^{(n)}\| \approx 2^{nH(X)}\text{，用以上两式可以推出}$$

联合典型序列：(x1, y1), ..., (xn, yn) i.i.d.服从P(X,Y)，(x1, y1), ..., (xn, yn)是联合典型序列的条件为x1, ..., xn和y1, ..., yn都是典型序列，并且满足下式。

$$p(x_1,\cdots,x_n,y_1,\cdots,y_n) \in 2^{-n[H(X,Y)+\epsilon]}$$