应用信息论基础——第一次习题课





2022年10月26日

清華大學
Tsinghua University

- 信息论 Information Theory
  - 通信的基础理论

    Shannon C E. A mathematical theory of communication[J]. The Bell system technical journal, 1948, 27(3): 379-423.

    - **基本模型**：信源、信道、信宿
    - **基本问题**：信息的**表示**、**传输**、存储（压缩）……

    从理论上指导了通信系统该如何设计，性能上限为多少

  - 学科交叉与延伸
    - **交叉**：概率论、统计理论、学习理论……
    - **应用**：度量、推断、检测、机器学习……
    - **拓展**：网络、语义……

直观理解
不同视角

重点总结

- **离散熵的定义**

  离散型随机变量X的熵H(X)定义为：

  $$H(X) = -\sum_{(x\in\chi)} p(x)\, logp(x)$$

  对随机变量而言

- **离散熵的性质**
  - $H(X) \geq 0$
  - **条件减小熵，$\boldsymbol{H(X|Y) \leq H(X)}$. 当且仅当X与Y相互独立，等号成立**
  - **$\boldsymbol{H(X_1, X_2, \ldots, X_n) \leq \sum_{i=1}^{n} H(X_i)}$,当且仅当随机变量$\boldsymbol{X_i}$相互独立，等号成立**
  - $H(X) \leq \log|\chi|$，当且仅当X服从χ上的**均匀分布**，等号成立
  - $H(p)$关于p是凹的

- **相对熵的定义**

  对于概率密度函数p和q，则相对熵定义为

  $$D(p||q) = \sum_{x} p(x) \log \frac{p(x)}{q(x)}$$

  对分布而言

- **互信息的定义**

  对于随机变量X和Y，互信息定义为

  $$I(X;Y) = \sum_{x\in\chi}\sum_{y\in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

  需要知道联合分布

重点总结

- 一些表达式
  - $H(X) = E_p \log \frac{1}{p(X)}$， $H(X,Y) = E_p \log \frac{1}{p(X,Y)}$， $H(X|Y) = E_p \log \frac{1}{p(X|Y)}$
  - $I(X;Y) = E_p \log \frac{p(X,Y)}{p(X)p(Y)}$
  - $D(p||q) = E_p \log \frac{p(X)}{q(X)}$

- D和I的性质
  - $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y)$
  - $D(p||q) \geq 0$, 当且仅当对任意$x \in \chi, p(x) = q(x)$，等号成立
  - $I(X;Y) = D\big(p(x,y)||p(x)p(y)\big) \geq 0$，当且仅当$p(x,y) = p(x)p(y)$时，等号成立
  - 若$|\chi| = m$，u是χ上的均匀分布，则$D(p||u) = \log m - H(p)$
  - $D(p||q)$关于二元对(p,q)是凸的

- 链式法则
  - 熵：$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1}, \ldots, X_1)$
  - 互信息：$I(X_1, X_2, \ldots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y|X_1, X_2, \ldots, X_{i-1})$
  - 相对熵：$D\big(p(x,y)||q(x,y)\big) = D\big(p(x)||q(x)\big) + D\big(p(y|x)||q(y|x)\big)$

重点总结

- **Jensen不等式**

  若f为凸函数，则$Ef(X) \geq f(EX)$

- **引理**

  如果X和X'独立同分布，那么

  $$\Pr(X = X') \geq 2^{-H(X)}$$

- **数据处理不等式**

  若$X \to Y \to Z$ 构成马尔科夫链，则$I(X;Y) \geq I(X;Z)$

- **费诺不等式**

  设$P_e = \Pr\{\hat{X}(Y) \neq X\}$, 则

  $$H(P_e) + P_e \log|\chi| \geq H(X|Y)$$

重点总结

- **连续随机变量的熵**
  - 微分熵：$h(X) = -\int f(x)\log f(x)dx$ 不再满足非负
  - 高斯分布：$h(X) = \frac{1}{2}\log 2\pi e\sigma^2$
  - 联合微分熵、条件微分熵类比离散情况
  - 相对微分熵：$D(f||g) = \int f(x)\log\frac{f(x)}{g(x)}dx \geq 0$

性质

(1) $h(X|Y) = h(X,Y) - h(Y)$

(2) $I(X;Y) = h(X) - h(X|Y) = h(Y) - h(Y|X)$

(3) $D(f \| g) \geq 0$，等号成立的条件为 $f(x) = g(x)$ 几乎处处成立。

(4) $I(X;Y) \geq 0$，等号成立的条件 X 与 Y 统计独立。

(5) $h(X|Y) \leq h(X)$，等号成立的条件 X 与 Y 统计独立。

(6) $h(X+a) = h(X)$，其中 a 为一常数。

(7) $h(aX) = h(X) + \log|a|$，$a \neq 0$。

(8) 若 A 为 $n \times n$ 的方阵，$\vec{X}$ 为 n 维随机向量，则

$$h(A\vec{X}) = h(\vec{X}) + \log|A|$$

其中 $|A|$ 为矩阵的行列式的绝对值。

◆ 不同情况下最大微分熵的分布

- 一阶矩受限（大于**0**）
- 二阶矩受限
- 幅度受限

重点总结

- **随机过程的熵率**
  - 熵率：描述随机变量序列的熵随n如何增长

$$H(\chi) = \lim_{n\to\infty} \frac{1}{n} H(X_1, \cdots, X_n)$$

  - 平稳过程下，有：

$$H(\chi) = H'(\chi) \triangleq \lim_{n\to\infty} H(X_n | X_{n-1}, \cdots, X_1)$$

  - Markov过程熵率：平稳分布π，转移矩阵P

$$H(\chi) = H(X_2|X_1) = -\sum_{i,j} \pi_i P_{ij} \log P_{ij} = \sum_i \pi_i H(P_i)$$

平稳分布乘对应行的熵

1-1：事件的信息量

清華大學 Tsinghua University

1. 设有n个球，每个球都以同样的概率落入N个格子（N ≥ n）中。假定：
   A：某指定的 n 个格子各落入一球；
   B：任意 n 个格子各落入一球。
   请计算事件 A、B 发生后所提供的信息量。

- 事件信息量的定义：发生概率的负log

$$P(A) = \frac{n!}{N^n} \qquad I(A) = -\log P(A)$$

$$P(B) = \frac{C_N^n \cdot n!}{N^n} \qquad I(B) = -\log P(B)$$

1-2：二元分布改变后的熵

Tsinghua University

2. 设二元离散随机变量X具有分布P_1和P_2，$P_2 > P_1$，现将其分布变为新的分布$P_1+\varepsilon$和$P_2-\varepsilon$，ε满足$0<2\varepsilon<P_2-P_1$，试分析在新的分布下熵$H(X)$随ε的变化规律，并证明你的结论。

- 二元分布改变后的熵

$H(X) = -(P_1+\epsilon)\log(P_1+\epsilon) - (P_2-\epsilon)\log(P_2-\epsilon)$　　对ϵ求导：

$$\frac{dH(X)}{d\epsilon} = \log\frac{P_2-\epsilon}{P_1+\epsilon}$$

由于$\forall\epsilon \in (0,\frac{P_2-P_1}{2})$，$\frac{dH(X)}{d\epsilon} > 0$，故$H(X)$随$\epsilon$的增大而增大

二元分布P_1越接近0.5熵越大

1-3：事件分类变量的熵

3. 设离散随机变量X_1和X_2分别定义于集合：

$A=\{a_1,a_2,\cdots,a_K\}$和$B=\{a_{K+1},a_{K+2},\cdots,a_M\}$

其概率分布分别为$p(a_i)$和$p(a_j)$，其中$i=1,2,\cdots,K$，$j=K+1,\cdots,M$。现构造随机变量X：

$$X=\begin{cases} X_1 & \text{依概率}\alpha \\ X_2 & \text{依概率}1-\alpha \end{cases}$$

求$H(X)$（用$H(X_1)$、$H(X_2)$和α表示。

- 利用熵的定义：

$$\begin{aligned} H(X) &= -\sum_{i=1}^{k}\alpha p_i\log\alpha p_i-\sum_{j=k+1}^{m}(1-\alpha)p_j\log(1-\alpha)p_j \\ &= \alpha H(X_1)-\alpha\log\alpha+(1-\alpha)H(X_2)-(1-\alpha)\log(1-\alpha) \\ &= \alpha H(X_1)+(1-\alpha)H(X_2)+\boldsymbol{H(\alpha)} \end{aligned}$$

清華大學 Tsinghua University

4. 设离散随机变量X_1和X_2具有相同的分布。令

$$\rho = 1 - \frac{H(X_2|X_1)}{H(X_1)}$$

1）证明：$\rho = \frac{I(X_1;X_2)}{H(X_1)}$及$0 \le \rho \le 1$;

2）分别给出$\rho = 0$和$\rho = 1$时X_1和X_2之间的统计关系。

- 互信息表达式及意义：

1） $\rho = 1 - \frac{H(X_2|X_1)}{H(X_1)} = \frac{H(X_1)-H(X_2|X_1)}{H(X_1)} = \frac{H(X_2)-H(X_2|X_1)}{H(X_1)} = \frac{I(X_1;X_2)}{H(X_1)}$

由$0 \le I(X_1;X_2) \le H(X_1)$可得$0 \le \rho = \frac{I(X_1;X_2)}{H(X_1)} \le 1$

2）$\rho = 0$，即$I(X_1;X_2) = 0$，即X_1与X_2独立

$\rho = 1$，即$I(X_1;X_2) = H(X_1) = H(X_2)$. 即 X_1与X_2为一一映射且可逆（相等）

- 分布相同，则熵相同
- 互信息小于等于熵
- 统计独立时，互信息为0
- 一一映射时，互信息等于熵

1-5：互信息计算

5. 设随机变量 X，Y 分别取值于 $\{x_0, x_1\}$ 和 $\{y_0, y_1\}$，已知 $P\{X = x_i\} = 0.5\ (i = 0, 1)$，且联合分布为 $p(x_k, y_k) = \frac{1-\varepsilon}{2}$，$p(x_k, y_{1-k}) = \frac{\varepsilon}{2}$，求$I(X;Y)$。

- 互信息表达式：

Y的边缘分布为$P(Y = y_0) = P(Y = y_1) = \frac{1}{2}$

联合熵$H(X,Y) = \left(-\frac{\epsilon}{2}\log\frac{\epsilon}{2} - \frac{1-\epsilon}{2}\log\frac{1-\epsilon}{2}\right) \times 2 = 1 + H(\epsilon)$

故$I(X,Y) = H(X) + H(Y) - H(X,Y) = 1 - H(\epsilon)$

1-6：条件熵不等式

6. X、Y、Z为离散随机变量，证明如下不等式并说明等号成立条件。

1） $H(XY|Z) \geq H(X|Z)$；

2） $H(XYZ) - H(XY) \leq H(XZ) - H(X)$。

1） $H(X,Y|Z) = H(X|Z) + H(Y|X,Z) \geq H(X|Z)$

等号成立即$H(Y|X,Z) = 0$. 即X,Z获取后，Y完全确定

条件熵链式法则

2） 由条件减少熵，则$H(Z|X,Y) \leq H(Z|X)$

故$H(X,Y,Z) - H(X,Y) \leq H(X,Z) - H(X)$

等号成立即$H(Z|X,Y) = H(Z|X)$，也就是$I(Y;Z|X) = 0$

在给定X的条件下，Y和Z是条件独立的。

增加条件，熵减小

1-7：联合熵/互信息的计算

7. 设随机变量X和Y的联合分布如下所示：

$X \backslash Y$	0	1
0	$\frac{1}{3}$	$\frac{1}{3}$
1	0	$\frac{1}{3}$

随机变量$Z = X \oplus Y$，其中$\oplus$为模2和。试求：

1）$H(X)$，$H(Y)$；

2）$H(XY)$，$H(YX)$，$H(XZ)$；

3）$I(X;Y)$，$H(XYZ)$。

1) $H(X) = -\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3} \approx 0.92\ bit$

$H(Y) = -\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3} \approx 0.92\ bit$

2) $H(X,Y) = H(Y,X) = \log 3 \approx 1.59\ bit$

$\boldsymbol{(X,Y)}$**与**$\boldsymbol{(X,Z)}$**等价**，故$H(X,Z) = H(X,Y)$

3) $I(X,Y) = H(X) + H(Y) - H(X,Y)$

由于$Z = X \oplus Y$，故$\boldsymbol{H(Z|X,Y) = 0}$

故$H(X,Y,Z) = H(X,Y) + H(Z|X,Y) = H(X,Y) \approx 1.59\ bit$

利用熵、联合熵、条件熵、互信息之间的变换关系及链式法则

Tsinghua University

8. 设离散随机变量X，Y，Z的值均取自集合$\{0,1\}$，试给出实例，满足：$I(X;Y)=0\text{bit}$，$I(X;Y|Z)=1\text{bit}$。

要求同时满足！

- 二元离散变量最大熵为1bit，则XY为独立等概分布，且已知Z和XY中任意一个，另一个可被确定。

设X,Y是独立同分布的随机变量，且$P(X=0)=P(X=1)=\frac{1}{2}$，$Z=X\oplus Y$

则$I(X;Y)=0$

$$I(X;Y|Z)=H(X|Z)-H(X|Y,Z)=H(X|Z)$$
$$=P(Z=0)H(X|Z=0)+P(Z=1)H(X|Z=1)=1\ bit$$

独立和互信息的关系

1-9：条件互信息的不等式及含义

清華大學 Tsinghua University

9. X、Y、Z为离散随机变量，证明如下不等式并借助通信系统的例子说明其物理含义：

1）$I(XY;Z) \geq I(X;Z)$;

2）若X与Y独立，则$I(Y;Z|X) \geq I(Y;Z)$;

3）若X与Y独立，则$I(XY;Z) \geq I(X;Z) + I(Y;Z)$。

1) $I(X,Y;Z) = I(X;Z) + I(Z;Y|X) \geq I(X;Z)$

条件互信息链式法则展开

2) X与Y独立，则$I(X;Y) = 0$

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

3) 使用(2)的结论

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

物理解释：多种角度
- 联合传输
- 联合译码
- 推断/数据处理

不要涉及信道容量

1-10：变换后的熵

10. X为离散随机变量，$g(X)$为X的函数，证明：$H(g(X)) \leq H(X)$，并给出等号成立条件。

- 两种方式展开联合熵：

$$H(X, g(X)) = H(X) + H(g(X)|X) = \mathrm{H}(\mathrm{g}(\mathrm{X})) + \mathrm{H}(\mathrm{X}|\mathrm{g}(\mathrm{X}))$$

$g(X)$是X的函数

由于$H(g(X)|X) = 0$，则

$$H(X) = H(g(X)) + H(X|g(X)) \geq H(g(X))$$

等号成立的条件是$g(X)$可逆（一一对应）。

1-11：三元组的互信息

清華大學 Tsinghua University

11. 随机变量 X、Y、Z 联合分布与边际分布乘积之间的KL散度为 $D\big(p(x,y,z)\|p(x)p(y)p(z)\big)$，用熵的形式将其展开，并说明何时该散度为0.

- 由KL散度的定义：

$$D\big(p(x,y,z)||p(x)p(y)p(z)\big) = \sum_{x,y,z} p(x,y,z)\log\frac{p(x,y,z)}{p(x)p(y)p(z)}$$

$$= -H(X,Y,Z) - \sum_{x,y,z} p(x,y,z)\log p(x)p(y)p(z)$$

$$= H(X) + H(Y) + H(Z) - H(X,Y,Z)$$

类比两个分布间的互信息表达式

2-1：微分熵的计算

1. 连续型随机变量X的概率密度函数为$p(x)=exp(-a|x|)$，其中$a>0$，求其微分熵。

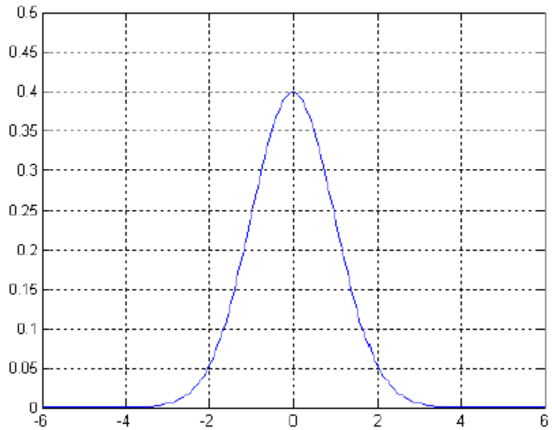
- 根据微分熵的定义：

$$h(X)=\int_{-\infty}^{+\infty}-p(x)\ln p(x)\,dx=2\underbrace{\int_{0}^{+\infty}x\cdot ae^{-ax}dx}_{\text{负指数分布的期望}}=\frac{2}{a}\text{ nat}$$

又由$p(x)$积分为1可得$a=2$。

2-2：高斯分布的运用

2. 连续型随机变量 X 和 Y 的联合分布为下图所示的区域内（x 轴与曲线 $y=\frac{1}{\sqrt{2\pi}}e^{-x^2/2}$，$-\infty<x<+\infty$）的均匀分布。



1）求 $h(X,Y)$，$h(X)$；

2）证明：$-\frac{1}{2}\ln 2\pi-\frac{1}{2}<h(Y)<-\frac{1}{2}\ln 2\pi$。

- 观察到曲线为一高斯分布

1）积分归一可得 $p(x,y)=1$，则 $h(X,Y)=0$，$h(X)=\frac{1}{2}\log 2\pi e$

连续随机变量互信息仍非负

2）由 $h(X)+h(Y)>h(X,Y)$ 可得左半边不等式；由于 $Y\in\left[0,\frac{1}{\sqrt{2\pi}}\right]$，幅度受限的连续随机变量取均匀分布时熵最大，可得右半边。

2-3：英文熵率的估计

- 1)

$$H(X) = -\Sigma_i p_i \log p_i$$

熵值的零阶估计：均匀分布

熵值的一阶估计：将频率看作概率计算熵

- 2) 3)

二阶估计：1阶Markov链，统计频率$p(X_i|X_{i-1})$

三阶估计：2阶Markov链，统计频率$p(X_i|X_{i-2}X_{i-1})$

m阶Markov信源熵率：

$$H_m(X) = H(X_i|X_{i-m} \cdots X_{i-1})$$
$$= -\Sigma p(X_{i-m} \cdots X_i) \log p(X_i|X_{i-m} \cdots X_{i-1})$$

4. 设无记忆稳恒信源产生由0和1构成i.i.d.的$\{X_n\}$，$p(X_n=0)=0.4$，$p(X_n=1)=0.6$。

1）计算2次扩展信源熵$H(X^2)$，$H(X_3|X_1X_2)$和$\lim_{N\to\infty}\frac{1}{N}H(X_1X_2\ldots X_N)$；

2）计算4次扩展信源熵$H(X^4)$并给出信源中所有的符号序列。

- 无记忆稳恒信源构成i.i.d.序列
- N次信源扩展，有

$$H(X^N)=NH(X)$$

- 条件熵：由独立性

$$H(X_3|X_1X_2)=H(X_3)=H(X)$$

2-5：平稳随机过程熵率

5. 设$\{X_n\}_{n=-\infty}^{\infty}$为一平稳随机过程。证明：

1）$H(X_0|X_{-1}X_{-2}\cdots X_{-n}) = H(X_0|X_1X_2\ldots X_n)$，即给定过去，当前时刻的条件熵与给定将来、当前时刻的条件熵相等；

2）$\lim\limits_{n\to\infty} H(X_nX_{n-1}|X_1X_2\cdots X_{n-2}) = 2H_\infty$。

- 严平稳：任意维联合分布函数/概率密度与时间起点无关

$$p(X_0 \cdots X_n) = p(X_t \cdots X_{t+n})$$

$$\begin{aligned}1)\ H(X_0|X_{-1}\cdots X_{-n}) &= H(X_0\cdots X_{-n}) - H(X_{-1}\cdots X_{-n}) \\ &= H(X_0\cdots X_n) - H(X_1\cdots X_n) = H(X_0|X_1\cdots X_n)\end{aligned}$$

或利用条件概率公式

2）由条件熵的链式法则和第一问结论及熵率定义可证

$$H(X_nX_{n-1}|X_1\cdots X_{n-2}) = H(X_n|X_1\cdots X_{n-1}) + H(X_{n-1}|X_1\cdots X_{n-2})$$

6. 一个离散稳恒遍历的Markov过程转移概率矩阵如下：

$$\bar{P}=\begin{bmatrix}1-p_{01} & p_{01}\\ p_{10} & 1-p_{10}\end{bmatrix}$$

1）求此Markov信源熵率；

2）p_{01}和p_{10}分别为多少时，熵率最大？并求最大熵率；

3）另一个离散稳恒遍历Markov过程转移概率矩阵如下，求其熵率；

$$\bar{P}'=\begin{bmatrix}1-p & p\\ 1 & 0\end{bmatrix}$$

4）p为多少时上一问中熵率最大，最大熵率为多少？此外，解释p应小于0.5这一结果。

- 1) 3) 通过Markov熵率定义即可得到
- 2）通过二元熵性质可得
- 4）$p^*=\frac{3-\sqrt{5}}{2}<0.5$

数值解释：熵率表达式的两个乘项。

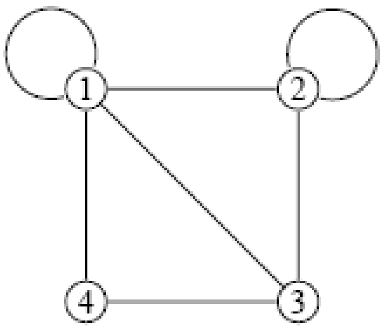
定性解释：平稳分布下状态1的概率和第一行的熵。

2-7：随机游走的熵率

7. 考虑下图所示的无向图中的随机走动。在每一步，当前节点都会以相同的概率选择一条移动路径。

1）求图中随机走动的稳态分布；

2）求图中随机走动的熵率。



- 转移概率：

$$P_{ij}=\begin{bmatrix}\frac{1}{4}&\frac{1}{4}&\frac{1}{4}&\frac{1}{4}\\ \frac{1}{3}&\frac{1}{3}&\frac{1}{3}&0\\ \frac{1}{3}&\frac{1}{3}&0&\frac{1}{3}\\ \frac{1}{2}&0&\frac{1}{2}&0\end{bmatrix}$$

其中 P_{ij} 表示从状态 i 到状态 j 的转移概率。设随机走动的稳态分布为 π，解得 $\pi=\begin{bmatrix}\frac{1}{3}&\frac{1}{4}&\frac{1}{4}&\frac{1}{6}\end{bmatrix}$。因此随机走动的熵为

$$H_\infty(U)=\sum_{j=1}^{4}\pi_j H(U\,|\,s=j)=-\sum_{j=1}^{4}\pi_j\sum_{i=1}^{4}P_{ij}\log_2 P_{ij}=1.6258\ \text{bits/symbol}$$