信息论第二讲作业解答

中国科学技术大学《信息论 A》006125.01 班助教组

2024 年 3 月 27 日

第 1 题

Prove the following properties of a probability space $(\Omega, \mathcal{F}, P)$ *:*

a) $P(\emptyset) = 0$.
b) For any $A, B \in \mathcal{F}$, *if* $A \subseteq B$ *then* $P(A) \leq P(B)$.
c) For any $A, B \in \mathcal{F}, P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

我们需要利用到概率测度的三个性质：

1) $\forall A \in \mathcal{F}, P(A) \in [0,1]$
2) $P(\Omega) = 1$
3) $P\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n P(A_i), \quad \forall A_1, A_2, \ldots, A_n \in \mathcal{F}, \text{ s.t. } \forall i \neq j, A_i \cap A_j = \emptyset.$

a) 证明: $\forall A \in \mathcal{F}, P(A) = P(A \cup \emptyset) \overset{3)}{=} P(A) + P(\emptyset), \Rightarrow P(\emptyset) = 0.$ □

b) 证明: $P(A^c \cap B) \overset{3)}{=} P(B) - P(A \cap B) \overset{A \subseteq B}{=} P(B) - P(A) \overset{1)}{\geq} 0$ □

c) 证明: $P(A \cup B) \overset{3)}{=} P(A) + P(B \cap A^c) \overset{3)}{=} P(A) + P(B) - P(A \cap B)$ □

注 1. 从事件集 $\mathcal{F}$ 对补集和可列并的封闭性可以推出对于可列交的封闭性：$\bigcap_{i=1}^\infty A_i = \overline{\bigcup_{i=1}^\infty \overline{A_i}} \in \mathcal{F}$. 在 *b)* 的证明第一步使用了这一点，*c)* 也默认了这一点。

第 2 题

For discrete random variables X *and* Y *over a probability space* $(\Omega, \mathcal{F}, P)$,

a) Prove the law of total expectation,

$$\mathbf{E}[X] = \mathbf{E}[\mathbf{E}[X|Y]]$$

b) Prove the law of total variance,

$$\operatorname{var} X = \mathbf{E}[\operatorname{var}[X|Y]] + \operatorname{var}\mathbf{E}[X|Y]$$

a) 证明:

$$RHS = \sum_{y_j} p(y_j)\left[\sum_{x_i} p(x_i|y_j)x_i\right] = \sum_{x_i} x_i \left[\sum_{y_j} p(x_i|y_j)p(y_j)\right] = \sum_{x_i} x_i p(x_i) = LHS \quad (1)$$

□

b) 证明:

$$\begin{aligned}
\mathbf{E}[\mathrm{Var}(X|Y)] &= \mathbf{E}\left[\mathbf{E}\left[X^2|Y\right] - (\mathbf{E}[X|Y])^2\right] && (2)\\
&= \mathbf{E}\left[\mathbf{E}\left[X^2|Y\right]\right] - \mathbf{E}\left[(\mathbf{E}[X|Y])^2\right] && (3)\\
&= \mathbf{E}\left[X^2\right] - \mathbf{E}\left[(\mathbf{E}[X|Y])^2\right] && (4)
\end{aligned}$$

$$\begin{aligned}
\mathrm{Var}(\mathbf{E}[X|Y]) &= \mathbf{E}\left[(\mathbf{E}[X|Y])^2\right] - \mathbf{E}[\mathbf{E}[X|Y]]^2 && (5)\\
&= \mathbf{E}\left[(\mathbf{E}[X|Y])^2\right] - \mathbf{E}[X]^2 && (6)
\end{aligned}$$

$$\mathbf{E}[\mathrm{Var}(X|Y)] + \mathrm{Var}(\mathbf{E}[X|Y]) = \mathbf{E}\left[X^2\right] - \mathbf{E}[X]^2 = \mathrm{Var}(X)$$

□

注 2. 我们可以认为不同 Y 的取值会导致 X 会遵从不同的分布。$\mathbf{E}[\mathrm{Var}(X|Y)]$ 表示在 Y 分类下 X 不同组间方差的均值，可以刻画组内差异。$\mathrm{Var}(\mathbf{E}[X|Y])$ 表示不同组之间均值的方差，可以刻画组与组间的差异。

第 3 题

Let X_1, X_2, X_3, X_4 be random variables such that $X_1 \leftrightarrow (X_2, X_3) \leftrightarrow X_4$ and $X_1 \leftrightarrow (X_2, X_4) \leftrightarrow X_3$ simultaneously hold.

a) If $P_{X_1,X_2,X_3,X_4}(x_1, x_2, x_3, x_4) > 0$, $\forall (x_1, x_2, x_3, x_4) \in X_1(\Omega) \times X_2(\Omega) \times X_3(\Omega) \times X_4(\Omega)$, prove that $X_1 \leftrightarrow X_2 \leftrightarrow (X_3, X_4)$ holds.

b) Can you give an example, wherein for some (x_1, x_2, x_3, x_4), $P_{X_1,X_2,X_3,X_4}(x_1, x_2, x_3, x_4) = 0$, and $X_1 \leftrightarrow X_2 \leftrightarrow (X_3, X_4)$ does not hold? This illustrates the delicacy of probability distributions with strictly zero probability [8, Prop. 2.12].

a) 证明: 令 $a_4 \in X_4(\Omega)$. 对所有 $x_1 \in X_1(\Omega)$, $x_2 \in X_2(\Omega)$, $x_3 \in X_3(\Omega)$ 和 $x_4 \in X_4(\Omega)$,

$$\begin{aligned} P_{X_1|X_2,X_3,X_4}(x_1|x_2,x_3,x_4) &= P_{X_1|X_2,X_3}(x_1|x_2,x_3) \\ &= P_{X_1|X_2,X_3,X_4}(x_1|x_2,x_3,a_4) \\ &= P_{X_1|X_2,X_4}(x_1|x_2,a_4). \end{aligned} \tag{7}$$

从 $P_{X_1|X_2,X_3,X_4}(x_1|x_2,x_3,x_4)$ 与 (x_3,x_4) 无关的事实可以看出 $X_1 \leftrightarrow X_2 \leftrightarrow (X_3,X_4)$. 我们也可以用下面的方法验证. 对所有 $x_1 \in X_1(\Omega)$ 和 $x_2 \in X_2(\Omega)$ 用 $P_{X_3,X_4|X_2}(x_3,x_4|x_2)$ 乘 (7) 式两边再对 $x_3 \in X_3(\Omega)$, $x_4 \in X_4(\Omega)$ 求和得

$$P_{X_1|X_2}(x_1|x_2) = P_{X_1|X_2,X_4}(x_1|x_2,a_4).$$

所以对所有 $x_1 \in X_1(\Omega)$, $x_2 \in X_2(\Omega)$, $x_3 \in X_3(\Omega)$ 和 $x_4 \in X_4(\Omega)$ 有

$$P_{X_1|X_2}(x_1|x_2) = P_{X_1|X_2,X_3,X_4}(x_1|x_2,x_3,x_4).$$ □

b) 解: 设 X_2, X_3 独立且都服从 $\{0,1\}$ 上的均匀分布, $X_1 = X_4 = X_2 \oplus X_3$. 这样 $X_3 = X_2 \oplus X_4$ 以概率 1 成立. 可以证明 $X_1 \leftrightarrow (X_2,X_3) \leftrightarrow X_4$, $X_1 \leftrightarrow (X_2,X_4) \leftrightarrow X_3$. 但

$$P_{X_1|X_2}(0|0) = P_{X_3|X_2}(0|0) = \frac{1}{2},$$
$$P_{X_1|X_2,X_3,X_4}(0|0,0,0) = 1,$$

说明 $X_1 \leftrightarrow X_2 \leftrightarrow (X_3,X_4)$ 不成立. □

第 4 题

Prove the following basic inequalities:

a) *Markov's inequality: for a nonnegative random variable X with finite expectation, and any $a > 0$,*

$$P(X \geq a) \leq \frac{\mathbf{E}[X]}{a}.$$

b) *Chebyshev's inequality: for a random variable X with finite expectation and variance, and any $a > 0$,*

$$P(|X - \mathbf{E}[X]| \geq a) \leq \frac{varX}{a^2}.$$

c) *Chernoff's inequality: for a random variable X and any a,*

$$P(X \geq a) \leq \min_{\lambda \geq 0} e^{-\lambda a}\mathbf{E}[e^{\lambda X}].$$

a) 证明: 设随机变量 I 满足：

$$I = \begin{cases} 1 & X \geq a, \\ 0 & \text{else}. \end{cases} \tag{8}$$

则 $I \leq \frac{X}{a}$, 所以 $\mathbf{E}[I] \leq \mathbf{E}\left[\frac{X}{a}\right]$，而 $\mathbf{E}[I] = P(X \geq a)$，因此 $P(X \geq a) \leq \frac{\mathbf{E}[X]}{a}$. $\square$

b) 证明: $P(|X - \mathbf{E}[X]| \geq a) = P((X - \mathbf{E}[X])^2 \geq a^2)$, 记随机变量 $S = (X - \mathbf{E}[X])^2$, 则 $\mathbf{E}[S] = varX$, 由 a) 得:$P(S \geq a^2) \leq \frac{\mathbf{E}[S]}{a^2}$, 即:$P(|X - \mathbf{E}[X]| \geq a) \leq \frac{\mathrm{var}X}{a^2}$. $\square$

c) 证明:

$$\begin{aligned} P(X \geq a) &= \int_a^{+\infty} f(x)dx \\ &\leq \int_a^{+\infty} e^{\lambda(x-a)} f(x)dx, \quad (\lambda \geq 0) \\ &\leq \int_{-\infty}^{+\infty} e^{\lambda(x-a)} f(x)dx \\ &= e^{-\lambda a}\mathbf{E}[e^{\lambda X}]. \end{aligned} \tag{9}$$

即对任意 $\lambda \geq 0$, 均有 $P(X \geq a) \leq e^{-\lambda a}\mathbf{E}[e^{\lambda X}]$, 故

$$P(X \geq a) \leq \min_{\lambda \geq 0} e^{-\lambda a}\mathbf{E}[e^{\lambda X}].$$

$\square$

第 5 题

If we model a pair of random variables X and Y with weak dependence as $P_{X,Y}(x,y) = P_X(x)P_Y(y)(1+\epsilon(x,y))$, such that there exists $\delta < 1$ satisfying $|\epsilon(x,y)| \leq \delta$, $\forall (x,y) \in X(\Omega) \times Y(\Omega)$, can you provide an upper bound on the difference between $H(X,Y)$ and $H(X)+H(Y)$?

解: 因为 $H(X) + H(Y) - H(X,Y) = I(X;Y)$, $I(X;Y) \geq 0$,

$$I(X;Y) = \mathbf{E}\left[\log\left(\frac{P_{X,Y}(X,Y)}{P_X(X)P_Y(Y)}\right)\right] = \mathbf{E}[\log(1+\varepsilon(X,Y))] \leq \log(1+\delta),$$

所以 $0 \leq H(X) + H(Y) - H(X,Y) \leq \log(1+\delta)$.

还可以换一种方法求出 $H(X) + H(Y) - H(X,Y)$ 一个更紧的上界. 求 $P_{X,Y}(x,y) = P_X(x)P_Y(y)(1+\varepsilon(x,y))$ 式两边对 $x \in X(\Omega)$ 和 $y \in Y(\Omega)$ 的和得

$$\sum_{x \in X(\Omega)} \sum_{y \in Y(\Omega)} P_{X,Y}(x,y) = \sum_{x \in X(\Omega)} P_X(x) \cdot \sum_{y \in Y(\Omega)} P_Y(y) + \sum_{x \in X(\Omega)} \sum_{y \in Y(\Omega)} P_X(x)P_Y(y)\epsilon(x,y).$$

所以 $\sum_{x\in X(\Omega)}\sum_{y\in Y(\Omega)} P_X(x)P_Y(y)\epsilon(x,y)=0.$

$$
\begin{aligned}
&\mathbf{E}[\log(1+\epsilon(X,Y))]\\
=&\ \log(e)\mathbf{E}[\ln(1+\epsilon(X,Y))]\\
\le&\ \log(e)\mathbf{E}[\epsilon(X,Y)]\\
=&\ \log(e)\sum_{x\in X(\Omega)}\sum_{y\in Y(\Omega)} P_X(x)P_Y(y)(1+\epsilon(x,y))\epsilon(x,y)\\
=&\ \log(e)\sum_{x\in X(\Omega)}\sum_{y\in Y(\Omega)} P_X(x)P_Y(y)\epsilon(x,y)+\log(e)\sum_{x\in X(\Omega)}\sum_{y\in Y(\Omega)} P_X(x)P_Y(y)\epsilon^2(x,y)\\
\le&\ \log(e)\sum_{x\in X(\Omega)}\sum_{y\in Y(\Omega)} P_X(x)P_Y(y)\delta^2\\
=&\ \delta^2\log(e).
\end{aligned}
$$

因此 $H(X)+H(Y)-H(X,Y)\le\delta^2\log(e).$ □

此外，有不少同学在展开计算中出现

$$
-\sum_{x\in X(\Omega)}\sum_{y\in Y(\Omega)}\epsilon(x,y)P_X(x)P_Y(y)\log(P_X(x)) \tag{10}
$$

这一项无法处理，选择放缩处理为 $\delta H(X)$，但是此处的界和上面给出的结果便有了本质上的不同，当 $H(X),H(Y)\to\infty$ 时，此处的界也会趋于无穷失去控制，原证明中依旧很紧。从而这一项应当为 0，下面我们来证明：

$$
\begin{aligned}
&P_{X,Y}(x,y)=P_X(x)P_Y(y)(1+\epsilon(x,y))\\
\xRightarrow{\text{对 y 求和}}&\sum_{y\in Y(\Omega)}P_{X,Y}(x,y)=\sum_{y\in Y(\Omega)}P_X(x)P_Y(y)(1+\epsilon(x,y))\\
\xRightarrow{\text{边缘分布}}&P_X(x)=P_X(x)\sum_{y\in Y(\Omega)}P_Y(y)(1+\epsilon(x,y))\\
\xRightarrow{\text{归一化}}&\sum_{y\in Y(\Omega)}P_Y(y)\epsilon(x,y)=0.
\end{aligned}
$$

从而 10转化为：

$$
-\sum_{x\in X(\Omega)}\sum_{y\in Y(\Omega)}\epsilon(x,y)P_X(x)P_Y(y)\log(P_X(x)) \tag{11}
$$

$$
=-\sum_{x\in X(\Omega)}P_X(x)\log(P_X(x))\sum_{y\in Y(\Omega)}\epsilon(x,y)P_Y(y)=-\sum_{x\in X(\Omega)}P_X(x)\log(P_X(x))\cdot 0=0. \tag{12}
$$

第 6 题

Cross entropy is an important concept in machine learning, usually used as objective function when training neural networks for classification tasks. For two probability distributions $P(x)$ and $Q(x)$ with domain $\mathcal{X}$, the cross entropy of $Q(x)$ relative to $P(x)$ is defined as

$$H_{\mathrm{c}}(P,Q) = -\sum_{x\in\mathcal{X}} P(x)\log Q(x).$$

What is the cross entropy $H_{\mathrm{c}}(P,Q)$ when $P(x)$ and $Q(x)$ are geometric distributions with parameters ϵ_P and ϵ_Q, respectively?

解: 由几何分布的定义可得 $P(x)=\epsilon_P(1-\epsilon_P)^{x-1}$，$Q(x)=\epsilon_Q(1-\epsilon_Q)^{x-1}$，带入相对熵计算公式有

$$\begin{aligned}
H_{\mathrm{c}}(P,Q) &= -\sum_{x\in\mathcal{X}} P(x)\log Q(x) && (13)\\
&= -\sum_{x\in\mathcal{X}} P(x)\log \epsilon_Q(1-\epsilon_Q)^{x-1} && (14)\\
&= -\sum_{x\in\mathcal{X}} P(x)\left(\log\epsilon_Q + (x-1)\log(1-\epsilon_Q)\right) && (15)\\
&= -\log\epsilon_Q\sum_{x\in\mathcal{X}} P(x) - \log(1-\epsilon_Q)\sum_{x\in\mathcal{X}} P(x)(x-1) && (16)\\
&= -\log\epsilon_Q - \log(1-\epsilon_Q)\left(\mathbf{E}_P\left[X-1\right]\right) && (17)\\
&= -\log\epsilon_Q - \log(1-\epsilon_Q)\left(\frac{1}{\epsilon_P}-1\right) && (18)\\
&= \frac{-\epsilon_P\log\epsilon_Q-(1-\epsilon_P)\log(1-\epsilon_Q)}{\epsilon_P}. && (19)
\end{aligned}$$

□

第 7 题

For a random variable whose range has size m, we may denote its pmf as a vector of elements $\{p_i\}_{i=1}^m$, and denote its entropy as $H(p_1,p_2,\ldots,p_m) = -\sum_{i=1}^m p_i\log p_i$. Verify the following properties of entropy:

a) *expansibility: $H(p_1,p_2,\ldots,p_m,0) = H(p_1,p_2,\ldots,p_m)$.*
b) *additivity:*

$$\begin{aligned}
&H(p_1,p_2,\ldots,p_m) + H(q_1,q_2,\ldots,q_n)\\
=&H(p_1q_1,\ldots,p_1q_n,p_2q_1,\ldots,p_mq_1,\ldots,p_mq_n)
\end{aligned}$$

c) grouping:

$$H(p_1, p_2, \ldots, p_m, q_1, q_2, \ldots, q_n) = H\left(\sum_{i=1}^{m} p_i, \sum_{j=1}^{n} q_j\right)$$
$$+\left(\sum_{i=1}^{m} p_i\right) H\left(\frac{p_1}{\sum_{i=1}^{m} p_i}, \frac{p_2}{\sum_{i=1}^{m} p_i}, \ldots, \frac{p_m}{\sum_{i=1}^{m} p_i}\right)$$
$$+\left(\sum_{j=1}^{n} q_j\right) H\left(\frac{q_1}{\sum_{j=1}^{n} q_j}, \frac{q_2}{\sum_{j=1}^{n} q_j}, \ldots, \frac{q_n}{\sum_{j=1}^{n} q_j}\right).$$

a) 证明:

$$\begin{aligned}H(p_1, p_2, \ldots, p_m, 0) &= -\sum_{i=1}^{m} p_i \log p_i - 0 \cdot \log 0 \\ &= -\sum_{i=1}^{m} p_i \log p_i \\ &= H(p_1, p_2, \cdots, p_m).\end{aligned}$$

□

b) 证明:

$$\begin{aligned}&H(p_1q_1, \cdots, p_1q_n, p_2q_1 \cdots, p_mq_1, \cdots, p_mq_n) \\ =& -\sum_{i}^{m}\sum_{j}^{n} p_iq_j \log p_iq_j \\ =& -\sum_{i}^{m}\sum_{j}^{n} p_iq_j (\log p_i + \log q_j) \\ =& -\sum_{j}^{n} q_j \sum_{i}^{m} p_i \log p_i - \sum_{i}^{m} p_i \sum_{j}^{n} q_j \log q_j \\ =& H(p_1, p_2, \cdots, p_m) + H(q_1, q_2, \cdots q_n).\end{aligned}$$

□

c) 证明: 不妨令 $\sum_{i=1}^{m} p_i = \alpha$，$\sum_{j=1}^{n} q_j = \beta$ 右边各项可以分别化简为

$$\begin{aligned}H\left(\sum_{i=1}^{m} p_i, \sum_{j=1}^{n} q_j\right) &= H(\alpha, \beta) \\ &= -\alpha \log \alpha - \beta \log \beta\end{aligned}$$

$$\begin{aligned}
&\left(\sum_{i=1}^m p_i\right) H\left(\frac{p_1}{\sum_{i=1}^m p_i}, \frac{p_2}{\sum_{i=1}^m p_i}, \cdots, \frac{p_m}{\sum_{i=1}^m p_i}\right)\\
=&-\alpha \sum_{i=1}^m \frac{p_i}{\alpha} \log \frac{p_i}{\alpha}\\
=&-\alpha\left(\sum_{i=1}^m \frac{p_i}{\alpha} \log p_i - \sum_{i=1}^m \frac{p_i}{\alpha} \log \alpha\right)\\
=&-\sum_i^m p_i \log p_i + \alpha \log \alpha.
\end{aligned}$$

同理下式成立

$$\left(\sum_{j=1}^n q_j\right) H\left(\frac{q_1}{\sum_{j=1}^n q_j}, \frac{q_2}{\sum_{j=1}^n q_j}, \ldots, \frac{q_n}{\sum_{j=1}^n q_j}\right) = -\sum_j^n q_j \log q_j + \beta \log \beta.$$

累加后可得

$$\mathrm{RHS} = -\sum_i^m p_i \log p_i - \sum_j^n q_j \log q_j = \mathrm{LHS}. \tag{20}$$

□

第 8 题

Consider independent random variables X and Y, each uniformly distributed over $1, 2, ..., n$

a) *Use computer to numerically study $H(X+Y)$ and plot its growth with n.*

b) *Use computer to numerically study $H(X \cdot Y)$ and plot its growth with n.*

```
import numpy as np
from scipy.stats import entropy
import matplotlib.pyplot as plt

def calc_entropy_plus(n):
    x = np.arange(1, n+1)
    y = np.arange(1, n+1)
    xy = np.add.outer(x, y)
    flat_xy = xy.flatten()
    counts = np.bincount(flat_xy, minlength=2*n)
    pmf = counts / np.sum(counts)
```

```
    return entropy(pmf)

def calc_entropy_multiply(n):
    x = np.arange(1, n+1)
    y = np.arange(1, n+1)
    xy = np.multiply.outer(x, y)
    flat_xy = xy.flatten()
    counts = np.bincount(flat_xy, minlength=n**2)
    pmf = counts / np.sum(counts)
    return entropy(pmf)

n = 100
n_vals = np.arange(1, n+1)
entropies_plus = [calc_entropy_plus(n) for n in n_vals]
entropies_multiply = [calc_entropy_multiply(n) for n in n_vals]

plt.plot(n_vals, entropies_plus,label = 'H(X+Y)')
plt.plot(n_vals, entropies_multiply,label = 'H(X*Y)')
plt.xlabel('n')
plt.ylabel('Entropy')
plt.legend()
plt.show()
```