第一章随机变量及其信息度量

第四节信息熵

陈兴同

中国矿业大学 数学学院

2021 年 8 月

内容提要

内容提要

内容提要

内容提要

内容提要

内容提要

定义 1.4.1：熵

现在来考虑随机变量的信息度量问题。自信息量是针对随机事件或随机变量某个取值来定义的，可以用于度量随机变量每个取值包含的信息量，随机变量的**平均自信息量**将作为随机变量的整体信息度量。

设离散随机变量 X 具有分布律（1-4）式，则称

1.2.1 例

$$H(X) = \sum_{x\in\mathcal{X}} p(x)\log\frac{1}{p(x)} = -\sum_{x\in\mathcal{X}} p(x)\log p(x), \tag{1.1}$$

为随机变量 X 的熵。

熵的说明：

（1） 熵实际上是自信息量的数学期望

$$H(X) = E\log\frac{1}{p(X)} = -E\log p(X),$$

它表示随机变量的每个取值所包含的平均信息量，因此自信息量可能大于熵也可以小于熵。

（2） 熵的单位与自信息量相同，可用比特，底特，奈特等。

（3） 熵定义中不涉及随机变量的取值，仅涉及到取每个值的概率，因此只要给一个概率分布向量 $p=(p_1,p_2,\cdots,p_N)$，就可以按照（1.1）求出熵

$$H(p) = H(p_1,p_2,\cdots,p_N)) \triangleq -\sum_{i=1}^{N} p_i\log p_i, \quad (1.2)$$

5

从而熵又可以看成是概率分布向量 $p=(p_1,p_2,\cdots,p_N)$ 的多元函数。

练习：

说明上面第（~~4~~3）条的合理性。

例题 1.4.1

设随机变量 X 服从 0-1 分布，求它的熵。

$$X \sim p(x) = \begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix}, 0 \leq p \leq 1,$$

解：

（1） 自信息量：

$$I(x) = \log \frac{1}{p(x)} = -\log p(x) = \begin{cases} -\log p & x = 1 \\ -\log(1-p) & x = 0 \end{cases}.$$

（2） 熵：

$$\begin{aligned} H(X) &= -\sum_{x \in \mathcal{X}} p(x) \log p(x) = -\left[\sum_{x=0} + \sum_{x=1}\right] \\ &= -[p(0)\log p(0) + p(1)\log p(1)] \\ &= -(1-p)\log(1-p) - p\log p. \end{aligned}$$

（3） 熵函数：

$$h(p) = -p\log p - (1-p)\log(1-p), p \in [0,1]. \quad (1.2)$$

采用以 2 为底的对数时称为二进熵函数。

二进熵函数

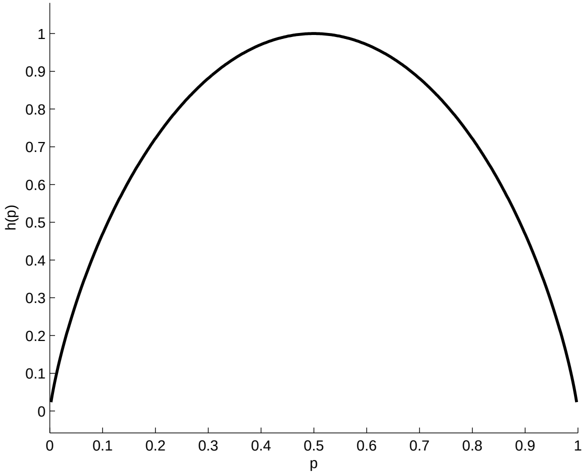

Figure: 图 1-4：二进熵函数的图形

例题 1.4.2

已知离散型随机变量 X 的分布律

$$X \sim p(x) = \begin{pmatrix} -1 & 0 & 2 & 3 \\ 0.2 & 0.3 & 0.2 & 0.3 \end{pmatrix}.$$

（1）试求 $Y = p(X)$ 的分布律；（2）试求 $Z = \ln p(X)$ 的分布律；（3）求数学期望 $E(Z)$；（4）求熵 $H(X)$。

解：

（1）随机变量 Y 的取值正好是随机变量 X 的概率，故 $Y = 0.2, 0.3$，取每个值的概率为

$$P\{Y = 0.2\} = P\{p(X) = 0.2\} = P\left\{X = -1\text{或}2\right\} = 0.4,$$

所以分布律为

$$Y \sim p_Y(y) = \begin{pmatrix} 0.2 & 0.3 \\ 0.4 & 0.6 \end{pmatrix}.$$

（2）随机变量 Z 实际上是随机变量 Y 的对数函数，故 Z 的取值为 $Z = \ln 0.2, \ln 0.3$，取每个值的概率为

$$P\{Z = \ln 0.2\} = P\{Y = 0.2\} = P\left\{X = -1\text{或}2\right\} = 0.4,$$

所以分布律为

$$Z \sim p_Z(z) = \begin{pmatrix} \ln 0.2 & \ln 0.3 \\ 0.4 & 0.6 \end{pmatrix}.$$

续解：

（3）] 根据数学期望定义得

$$E(Z) = 0.4 * \ln 0.2 + 0.6 * \ln 0.3 = -1.3662.$$

要注意数学期望不是信息度量，它没有信息单位，只有物理单位。（4）根据熵的定义得

$$H(X) = 0.2{*}\ln\frac{1}{0.2} + 0.3{*}\ln\frac{1}{0.3} + 0.2{*}\ln\frac{1}{0.2} + 0.3{*}\ln\frac{1}{0.3} = 1.3662 \quad \text{nats}.$$

练习：

求几种常见分布如二项分布、泊松分布、几何分布的熵。。

自己去查文献

命题 1.4.1

设有限离散型随机变量 X 分布律为 (1-3) 式，那么熵 $H(X)$ 具有如下性质：

（1） （非负性）$H(X) \geq 0$。

（2） $H(X) = 0$ 当且仅当随机变量 X 服从退化分布。

（3） 如果随机变量 X 服从等可能分布，则熵 $H(X) = \log N$。

（4） （**最大熵定理**）$H(X) \leq \log N$；而等号成立当且仅当 X 服从等可能分布。

（5） （对称性）$H(p_1, p_2, \cdots, p_N) = H(p_{i_1}, p_{i_2}, \cdots, p_{i_N})$，其中 $i_1 i_2 \cdots i_N$ 是 $12 \cdots N$ 的任意一个全排列。

命题（续）：

（6） （熵的可加性）如果 $p=(p_1,p_2,\cdots,p_N), q=(q_{11},q_{12},\cdots,q_{1k_1},\cdots,q_{N1},q_{N2},\cdots,q_{Nk_N})$ 是两个概率分布，并且满足 $p_i=\sum_{j=1}^{k_i} q_{ij}, i=1,2,\cdots,N$，则成立

$$H(q)=H(p)+\sum_{i=1} p_i H(q_i),$$

其中

$$q_i=\frac{1}{p_i}(q_{i1},q_{i2},\cdots,q_{ik_i}), i=1,2,\cdot,N.$$

（7） 如果函数 $y=f(x)$ 在集合 $\mathcal{X}$ 上有定义，则 $Y=f(X)$ 的熵不增，即

$$H(Y)=H(f(X))\leq H(X).$$

证明:

第（1）个结论是显然的。现在证明第（2）个结论。事实上：对任何一个概率 p_i 总有 $-p_i \log p_i \geq 0$，因此如果有一个 i 使 $-p_i \log p_i > 0$，从而必有 $H(X) > 0$，这与已知矛盾。故对任一个 i 都有 $-p_i \log p_i = 0$，因此要么 $p_i = 0$ 要么 $p_i = 1$。但是 $\sum_{i=1}^{N} p_i = 1$，因此只能有一个 i 使 $p_i = 1$，从而所给分布是退化分布。

第（3）个结论的证明：当 X 服从等可能分布时有

$$H(X) = -\sum_{i=1}^{N} p_i \log p_i = -\sum_{i=1}^{N} \frac{1}{N} \log \frac{1}{N} = \log N.$$

续证明：

第（4）个结论的证明：现在取另一个分布为等可能分布即 $q_i = 1/N, i = 1, 2, \cdots, N$，则由概率分布不等式（??）可得

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i \leq -\sum_{i=1}^{n} p_i \log \frac{1}{N} = \log N \sum_{i=1}^{n} p_i = \log N.$$

第（5）个结论的证明：当概率重新排列时，$H(p_1, p_2, \cdots, p_N)$ 中各个项也会作相应排列，但和不变。

续证明:

第（6）个结论的证明：由熵的定义可得

$$
\begin{aligned}
H(q) &= -\sum_{i=1}^{N}\sum_{j=1}^{k_i} q_{ij}\log q_{ij} = -\sum_{i=1}^{N}\sum_{j=1}^{k_i} q_{ij}\log\left(\frac{q_{ij}}{p_i}p_i\right)\\
&= -\sum_{i=1}^{N}\sum_{j=1}^{k_i} q_{ij}\log p_i - \sum_{i=1}^{N}\sum_{j=1}^{k_i} q_{ij}\log\frac{q_{ij}}{p_i}\\
&= -\sum_{i=1}^{N}\log p_i\sum_{j=1}^{k_i} q_{ij} - \sum_{i=1}^{N} p_i\sum_{j=1}^{k_i}\frac{q_{ij}}{p_i}\log\frac{q_{ij}}{p_i}\\
&= -\sum_{i=1}^{N} p_i\log p_i + \sum_{i=1}^{N} p_i H(q_i) = H(p) + \sum_{i=1}^{N} p_i H(q_i).
\end{aligned}
$$

续证明：

第（7）个结论的证明：记
$\mathcal{Y}=\{y|y=f(x),x\in\mathcal{X}\}$，$\mathcal{X}_y=\{x|f(x)=y\}$，则随机变量 Y 的概率分布可以看成是由 X 的概率分布合并而成的，即

$$q(y)=P\{Y=y\}=P\{X\in\mathcal{X}_y\}=\sum_{x\in\mathcal{X}_y}p(x),$$

因此

$$\frac{p(x)}{q(y)},x\in\mathcal{X}_y$$

正好是一个概率函数，它的熵是

$$H_y=-\sum_{x\in\mathcal{X}_y}\frac{p(x)}{q(y)}\log\frac{p(x)}{q(y)},$$

故由第（6）个结论得

$$H(X)=H(Y)+\sum_{y\in\mathcal{Y}}q(y)H_y\geq H(Y).$$

定义 1.4.2：二维联合熵

*设随机变量 X, Y 有联合分布律（1-8），则它们的**联合熵**定义为*

$$
\begin{aligned}
H(X,Y) &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{1}{p(x,y)} \\
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y). \quad (1.24)
\end{aligned}
$$

它是联合自信息量的数学期望

$$H(X,Y) = E[\log\frac{1}{p(X,Y)}] = -E[\log p(X,Y)].$$

多维联合熵定义

进一步可以定义更多个随机变量的联合熵。

$$
\begin{aligned}
& H(X_1, X_2, \cdots, X_n) \\
= & -\sum_{x_1}\sum_{x_2}\cdots\sum_{x_n} p(x_1, x_2, \cdots, x_n)\log p(x_1, x_2, \cdots, x_n) \\
= & -E[\log p(X_1, X_2, \cdots, X_n)]
\end{aligned}
$$

命题 1.4.2：简单性质

联合熵具有如下一些简单性质：

（1） $H(X,Y)=H(X)+H(Y)$ 的充要条件是随机变量 X,Y 相互独立。

（2） $H(X_1,X_2,\cdots,X_n)=H(X_1)+H(X_2)+\cdots+H(X_n)$ 的充要条件是 $X_1,X_2,\cdots,X_n$ 相互独立。

（3） （对称性）$H(X,Y)=H(Y,X)$。

（4） （对称性）$H(X_1,X_2,\cdots,X_n)=H(X_{i_1},X_{i_2},\cdots,X_{i_n})$，$i_1i_2\cdots i_n$ 是 $12\cdots n$ 的任意一个全排列。

证明：性质（1），其它留作练习

若 X 与 Y 相互独立，则联合分布概率函数 $(X,Y) \sim p(x,y)$ 与边缘分布概率函数 $X \sim p_X(x), Y \sim p_Y(y)$ 有关系 $p(x,y) = p_X(x)p_Y(y)$，从而由联合熵定义得

$$
\begin{aligned}
H(X,Y) &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y) \\
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p_X(x)p_Y(y)\log\left[p_X(x)p_Y(y)\right] \\
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p_X(x)p_Y(y)\left[\log p_X(x) + \log p_Y(y)\right] \\
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p_X(x)p_Y(y)\log p_X(x) \\
&\quad -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p_X(x)p_Y(y)\log p_Y(y).
\end{aligned}
$$

续证明：性质（1），其它留作练习

$$
\begin{aligned}
&= \left[-\sum_{x\in\mathcal{X}} p_X(x)\log p_X(x)\right]\sum_{y\in\mathcal{Y}} p_Y(y) \\
&\quad + \left[-\sum_{y\in\mathcal{Y}} p_Y(y)\log p_Y(y)\right]\sum_{x\in\mathcal{X}} p_X(x) \\
&= H(X)+H(Y).
\end{aligned}
$$

练习：

证明命题 1.4.2（1）的必要性。必要性是是什么？

定义 1.4.3：条件熵

设随机变量 X, Y 有分布律 $p_X(x), p_Y(y)$，联合分布律 $p(x,y)$，以及条件分布律 $p(y|x)$，$x \in \mathcal{X}, y \in \mathcal{Y}$，则

（1）在给定事件 $\{X = x\}$ 下随机变量 Y 的**条件熵**定义为

$$H(Y|X=x) = \sum_{y\in\mathcal{Y}} p(y|x)\log\frac{1}{p(y|x)} = -\sum_{y\in\mathcal{Y}} p(y|x)\log p(y|x).$$

1-25 ~~(4.1)~~

（2）在给定随机变量 X 下随机变量 Y 的**条件熵**定义为

$$H(Y|X) = \sum_{x\in\mathcal{X}} p(x)H(Y|X=x) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{1}{p(y|x)}.$$

~~(4.2)~~ 1-26

更多条件的条件熵：

条件熵可以进一步推广，比如用 X, Y 作为条件的二元条件熵、用 $X_1, X_2, \cdots, X_n$ 作为条件的 n 元条件熵。

（1）二元条件熵定义

$$H(Z|X,Y) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}\sum_{z\in\mathcal{Z}} p(x,y,z)\log\frac{1}{p(z|x,y)}. \qquad \cancel{(4.3)}$$

127

（2）n 元条件熵定义

$$H(Z|X_1,X_2,\cdots,X_n) = \sum_{x_1\in\mathcal{X}_1}\cdots\sum_{x_n\in\mathcal{X}_n}\sum_{z\in\mathcal{Z}} p(x_1,x_2,\cdots,x_n,z)$$

$$\times\log\frac{1}{p(z|x_1,x_2,\cdots,x_n)}$$

更多条件的条件熵：

（3）在 n 元条件下随机向量 $(Z_1, Z_2, \cdots, Z_m)$ 的条件熵定义为

$$
\begin{aligned}
& H(Z_1, Z_2, \cdots, Z_m | X_1, X_2, \cdots, X_n) \\
= & \sum_{x_1 \in \mathcal{X}_1} \cdots \sum_{x_n \in \mathcal{X}_n} \sum_{z_1 \in \mathcal{Z}_1} \cdots \sum_{z_m \in \mathcal{Z}_m} p(x_1, \cdots, x_n, z_1, \cdots, z_m) \\
& \times \log \frac{1}{p(z_1, \cdots, z_m | x_1, \cdots, x_n)}.
\end{aligned}
$$

命题 1.4.3 性质：

条件熵具有如下性质：（1）$H(Y|X) = H(X,Y) - H(X), H(X|Y) = H(X,Y) - H(Y)$。

（2）*如果随机变量* X, Y *相互独立，则* $H(Y) = H(Y|X)$ *或* $H(X) = H(X|Y)$。

（3）*如果* (X,Y) *与* Z *独立，则* $H(Z|X,Y) = H(Z)$。

（4）*如果* $(Z_1, Z_2, \cdots, Z_m)$ *与* $(X_1, X_2, \cdots, X_n)$ *相互独立，则*

$$H(Z_1, Z_2, \cdots, Z_m | X_1, X_2, \cdots, X_n) = H(Z_1, Z_2, \cdots, Z_m).$$

（5）*如果由* X *可以确定* Y *即有函数关系* $Y = f(X)$*，则* $H(Y|X) = 0$。

问题：

证明上面条件熵性质（1），（2），（3），（5）；探索独立性条件是否可以是充要条件。

定义 1.4.4：相对熵

设两个取值在同一字符集 $\mathcal{X}$ 上的概率分布 $p(x), q(x), x \in \mathcal{X}$，定义

$$D(p||q) = \sum_{x\in\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \tag{5.1}$$

1.28

为分布 $p(x)$ 对分布 $q(x)$ 的**相对熵**。

命题 1.4.4：性质

相对熵具有如下性质：

（1） $D(p||q) \geq 0$。

（2） $D(p||q) = 0$ 的充要条件是 $p(x) = q(x), \forall x \in \mathcal{X}$。

由概率分布不等式容易证明。(1.12)

例题 1.4.3

设随机变量 (X,Y) *具有例题 1.3.1 中的分布律* $p(x,y)$，

（1） 试求随机变量 $Z=p(X,Y), W=\ln p(X,Y)$ 分布律。

（2） 求熵 $H(X), H(Y), H(Z)$。

（3） 求联合熵 $H(X,Y)$。

（4） 求条件熵 $H(Y|X), H(X|Y)$。

解：

第（1）问：随机变量 Z 的取值为 $Z = 0, 1/4, 1/8, 1/12, 1/16$，而且 Z 的分布率为

$$Z \sim p_Z(z) = \begin{pmatrix} 1/4 & 1/8 & 1/12 & 1/16 \\ 1/4 & 1/4 & 1/4 & 1/4 \end{pmatrix},$$

于是随机变量 W 的分布律为

$$W \sim p_W(w) = \begin{pmatrix} -\ln 4 & -\ln 8 & -\ln 12 & -\ln 16 \\ 1/4 & 1/4 & 1/4 & 1/4 \end{pmatrix}.$$

续解：

第（2）问：易求得边缘分布律

$$X \sim p_X(x) = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1/4 & 1/4 & 1/4 & 1/4 \end{pmatrix}$$

$$Y \sim p_Y(y) = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 25/48 & 13/48 & 7/48 & 3/48 \end{pmatrix}$$

从而所求熵为

$$\begin{aligned} H(X) &= \ln 4 = 1.3863 \quad \text{nats}, \\ H(Y) &= -\sum_y p_Y(y) \ln p_Y(y) \\ &= \frac{25}{48} \ln \frac{48}{25} + \frac{13}{48} \ln \frac{48}{13} + \frac{7}{48} \ln \frac{48}{7} + \frac{3}{48} \ln \frac{48}{3} \\ &= 1.1475 \quad \text{nats}, \\ H(Z) &= H(W) = \ln 4 = 1.3863 \quad \text{nats}. \end{aligned}$$

续解：

第（3）问：

$$
\begin{aligned}
H(X,Y) &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y) \\
&= \frac{1}{4}\ln 4 + \frac{1}{8}\ln 8 + \frac{1}{12}\ln 12 + \frac{1}{16}\ln 16 \\
&\quad + \frac{1}{8}\ln 8 + \frac{1}{12}\ln 12 + \frac{1}{16}\ln 16 \\
&\quad + \frac{1}{12}\ln 12 + \frac{1}{16}\ln 16 \\
&\quad + \frac{1}{16}\ln 16 \\
&= 2.181 \quad \text{nats.}
\end{aligned}
$$

续解：

第（4）问：为了求条件熵需要求条件分布矩阵。事实上，由条件概率公式

$$p(y|x)=\frac{p(x,y)}{p_X(x)},p(x|y)=\frac{p(x,y)}{p_Y(y)},$$

即可求得

$$p(y|x)=\begin{pmatrix}1&0&0&0\\1/2&1/2&0&0\\1/3&1/3&1/3&0\\1/4&1/4&1/4&1/4\end{pmatrix}$$

$$p(x|y)=\begin{pmatrix}12/25&6/25&4/25&3/25\\0&6/13&4/13&3/13\\0&0&4/7&3/7\\0&0&0&1\end{pmatrix},$$

续解：

从而

$$H(Y|X=1)=0, H(Y|X=2)=\ln 2, H(Y|X=3)=\ln 3,$$
$$H(Y|X=4)=\ln 4 \text{ nats},$$
$$H(X|Y=1)=\frac{12}{25}\ln\frac{25}{12}+\frac{6}{25}\ln\frac{25}{6}+\frac{4}{25}\ln\frac{25}{4}+\frac{3}{25}\ln\frac{25}{3}$$
$$=1.24246 \quad \text{nats},$$
$$H(X|Y=2)=\frac{6}{13}\ln\frac{13}{6}+\frac{4}{13}\ln\frac{13}{4}+\frac{3}{13}\ln\frac{13}{3}=1.0579 \quad \text{nats},$$
$$H(X|Y=3)=\frac{4}{7}\ln\frac{7}{4}+\frac{3}{7}\ln\frac{7}{3}=0.6829 \quad \text{nats},$$
$$H(X|Y=4)=0,$$

续解：

因此得条件熵

$$H(Y|X) = \sum_{x} p_X(x) H(Y|X=x) = 0.7945 \quad \text{nats},$$

$$H(X|Y) = \sum_{y} p_Y(y) H(X|Y=y) = 1.03322 \quad \text{nats}.$$

另外，也可以用命题 1.4.3（1）来求条件熵。

例题 1.4.4

设二维联合分布

$$(X,Y) \sim p(x,y), (x,y) \in \mathcal{X} \times \mathcal{Y}, \quad 1.29 \quad \cancel{(5.2)}$$

它所对应的边缘分布为$p_X(x), x \in \mathcal{X}, p_Y(y) \in \mathcal{Y}$。

（1） 证明

$$p_X(x)p_Y(y), (x,y) \in \mathcal{X} \times \mathcal{Y} \quad 1.30 \quad \cancel{(5.3)}$$

是一个二维概率分布。

（2） 求概率分布（1-29）对（1-30）的相对熵。

解：

（1） 由 $p_X(x), p_Y(y)$ 是边缘分布得 $p_X(x)p_Y(y) \geq 0$； 又

$$\sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} p_X(x)p_Y(y) = \sum_{x\in\mathcal{X}} p_X(x) \sum_{y\in\mathcal{Y}} p_Y(y) = 1,$$

由概率分布的定义可知，（1-30）是一个二维概率分布。

续解:

第（2）问: 根据以上证明可知，（1-29）和（1-30）是两个定义在同一个字符集 $\mathcal{X}\times\mathcal{Y}$ 上的概率分布，故可以求相对熵。

$$\begin{aligned}
& D(p\|p_X p_Y) && \cancel{(5.4)}\\
=& \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{p_X(x)p_Y(y)}\\
=& \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{p(y|x)}{p_Y(y)}\\
=& -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p_Y(y)+\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(y|x)\\
=& H(Y)-H(Y|X). && \overset{1.31}{\cancel{(5.5)}}
\end{aligned}$$

练习：

问相对熵 $D(p||p_X p_Y) = H(X) - H(X|Y)$ *?*

例题 1.4.5

使用例题 1.3.1 中的二维联合分布。

（1）求相对熵 $D(p_X||p_Y), D(p_Y||p_X)$。

（2）求二维概率分布 $q(x,y) = p_X(x)p_Y(y)$。

（3）求相对熵 $D(p||q), D(q||p)$。

解：

显然随机变量 X,Y 的取值在相同的字符空间 $\mathcal{X}=\{1,2,3,4\}$ 中，并且

$$X \sim p_X(x) = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1/4 & 1/4 & 1/4 & 1/4 \end{pmatrix}$$

$$Y \sim p_Y(x) = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 25/48 & 13/48 & 7/48 & 3/48 \end{pmatrix}.$$

续解：

第（1）问：由相对熵的公式得

$$\begin{aligned}
D(p_X||p_Y) &= \sum_x p_X(x)\ln\frac{p_X(x)}{p_Y(x)} \\
&= \frac{1}{4}\ln\frac{1/4}{25/48}+\frac{1}{4}\ln\frac{1/4}{13/48}+\frac{1}{4}\ln\frac{1/4}{7/48}+\frac{1}{4}\ln\frac{1/4}{3/48} \\
&= \frac{1}{4}\left[\ln\frac{12}{25}+\ln\frac{12}{13}+\ln\frac{12}{7}+\ln\frac{12}{3}\right] \\
&= 0.27782 \quad \text{nats.}
\end{aligned}$$

续解：

第（1）问：由相对熵的公式得

$$
\begin{aligned}
D(p_Y \| p_X) &= \sum_x p_Y(x) \ln \frac{p_Y(x)}{p_X(x)} \\
&= \frac{25}{48} \ln \frac{25/48}{1/4} + \frac{13}{48} \ln \frac{13/48}{1/4} + \frac{7}{48} \ln \frac{7/48}{1/4} + \frac{3}{48} \ln \frac{3/48}{1/4} \\
&= \frac{25}{48} \ln \frac{25}{12} + \frac{13}{48} \ln \frac{13}{12} + \frac{7}{48} \ln \frac{7}{12} + \frac{3}{48} \ln \frac{3}{12} \\
&= 0.2387 \quad \text{nats.}
\end{aligned}
$$

续解：

第（2）问：由边缘分布律可得二维概率分布

$$q(x,y)=p_X(x)p_Y(y)=\frac{1}{4}p_Y(y),$$

写成表格形式为

$Y \setminus X$	1	2	3	4
1	$\frac{25}{192}$	$\frac{25}{192}$	$\frac{25}{192}$	$\frac{25}{192}$
2	$\frac{13}{192}$	$\frac{13}{192}$	$\frac{13}{192}$	$\frac{13}{192}$
3	$\frac{7}{192}$	$\frac{7}{192}$	$\frac{7}{192}$	$\frac{7}{192}$
4	$\frac{3}{192}$	$\frac{3}{192}$	$\frac{3}{192}$	$\frac{3}{192}$

续解：

第（3）问：由相对熵公式得

$$
\begin{aligned}
& D(p||q) \\
= \;& \sum_x \sum_y p(x,y) \ln \frac{p(x,y)}{q(x,y)} \\
= \;& \frac{1}{4} \ln \frac{1/4}{25/192} + \frac{1}{8} \ln \frac{1/8}{25/192} + \frac{1}{12} \ln \frac{1/12}{25/192} + \frac{1}{16} \ln \frac{1/16}{25/192} \\
& + \frac{1}{8} \ln \frac{1/8}{13/192} + \frac{1}{12} \ln \frac{1/12}{13/192} + \frac{1}{16} \ln \frac{1/16}{13/192} \\
& + \frac{1}{12} \ln \frac{1/12}{7/192} + \frac{1}{16} \ln \frac{1/16}{7/192} \\
& + \frac{1}{16} \ln \frac{1/16}{3/192} \\
= \;& 0.353 \quad \text{nats.}
\end{aligned}
$$

续解：

同理，可求另一个相对熵：

$$
\begin{aligned}
& D(q\|p) \\
= & \sum_x \sum_y q(x,y) \ln \frac{q(x,y)}{p(x,y)} \\
= & \frac{25}{192} \ln \frac{25/192}{1/4} + \frac{25}{192} \ln \frac{25/192}{1/8} + \frac{25}{192} \ln \frac{25/192}{1/12} + \frac{25}{192} \ln \frac{25/192}{1/16} \\
& + \frac{13}{192} \ln \frac{13/192}{0} + \frac{13}{192} \ln \frac{13/192}{1/8} + \cdots \\
= & \infty.
\end{aligned}
$$

这个例子说明相对熵可能不存在，也不具有对称性。

命题 1.4.5：熵的链式法则

熵、联合熵、条件熵之间可建立联系。
（1）$H(X,Y)=H(X)+H(Y|X)=H(Y)+H(X|Y)$。
（2）可推广到一般情况：

$$H(X_1,X_2,\cdots,X_n)=H(X_1)+\sum_{i=2}^{n}H(X_i|X_{i-1},\cdots,X_2,X_1).$$

（3）$H(X,Y|Z)=H(X|Z)+H(Y|X,Z)=H(Y|Z)+H(X|Y,Z)$.（4）可推广到更一般情况：

$$H(X_1,X_2,\cdots,X_n|Y)=H(X_1|Y)+\sum_{i=2}^{n}H(X_i|X_1,\cdots,X_{i-1},Y),$$

如果 Y 是一个随机向量，该式也成立。

第（1）条证明：

$$
\begin{aligned}
H(X,Y) &= -\sum_x\sum_y p(x,y)\log p(x,y)\\
&= -\sum_x\sum_y p(x,y)\log p(x)p(y|x)\\
&= -\sum_x\sum_y p(x,y)\log p(x) - \sum_x\sum_y p(x,y)\log p(y|x)\\
&= H(X)+H(Y|X).
\end{aligned}
$$

第（3）条证明类似，第（2）、（4）条证明可由归纳法完成。

命题 1.4.6：熵的不等式

（1）$H(Y|X) \leq H(Y)$。
（2）$H(X,Y) \leq H(X) + H(Y)$。
（3）$H(Z|X,Y) \leq H(Z|X)$或$H(Z|Y)$。
（4）推广第（1）条：$H(Y|X_1, X_2, \cdots, X_n) \leq H(Y)$。
（5）推广第（2）条：$H(X_1 X_2 \cdots X_n) \leq H(X_1) + H(X_2) + \cdots + H(X_n)$。
第（1）、（2）、（3）条可以由熵的定义或链式法等来证明；

命题 1.4.6：熵的不等式

（6）推广第（3）
条：$H(Z|X_1, X_2, \cdots, X_n) \leq H(Z|X_1, X_2, \cdots, X_{n-1})$。
（7）推广第（4）
条：$H(Z_1, Z_2, \cdots, Z_m|X_1, X_2, \cdots, X_n) \leq H(Z_1, Z_2, \cdots, Z_m)$。
（8）推广第（7）条：

$$H(Z_1, \cdots, Z_m|X_1, \cdots, X_n) \leq H(Z_1, Z_2, \cdots, Z_m|X_1, X_2, \cdots, X_{n-1}).$$

第（1）、（2）、（3）条可以由熵的定义或链式法等来证明；
其余各条可由归纳法完成，留作练习。
第（1）（3）（4）（6）（7）（8）条性质表明，增加条件，条件熵不增加。