第 1 讲 熵

自信息 (Self-information)
熵
联合熵与条件熵
相对熵与互信息
熵与互信息之间的关系

自信息 (Self-information)

熵

联合熵与条件熵

相对熵与互信息

熵与互信息之间的关系

自信息 (Self-information)

熵

联合熵与条件熵

相对熵与互信息

熵与互信息之间的关系

参考资料

- T. M. Cover and J. A. Thomas, Elements of Information Theory, 2.1-2.5.
- R. W. Yeung, Information and Network Coding, 2.2-2.4.
- F. Alajaji, P. Chen, An Introduction to Single-User Information Theory, 2.1-2.3.
- Y. Polyanskiy, Y. Wu, Information Theory: from Coding to Learning, 1.1, 2.1, 3.1.

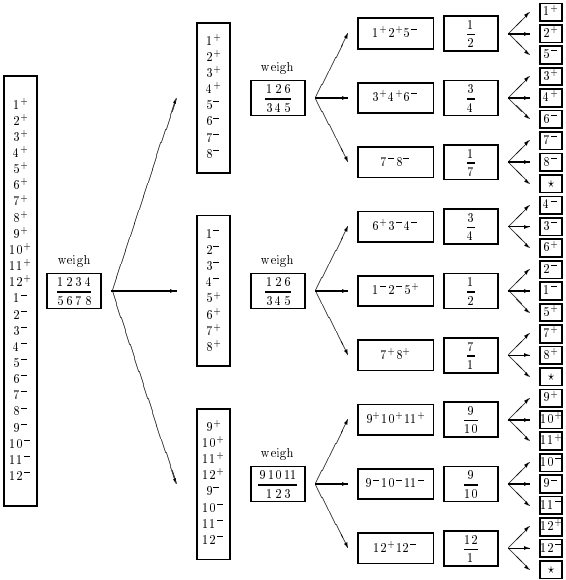
引子

自信息 (Self-information)
熵
联合熵与条件熵
相对熵与互信息
熵与互信息之间的关系

课本 P25 习题 2.7

- 假定有 n 枚硬币, 可能有一枚或者没有假币.
- 如果是假币, 那么它的重量要么重于其它的硬币, 要么轻于其它的硬币.
- 我们有一个没有砝码的天平, 希望借助它能够找出假币.

12 枚硬币的情形 (假设必存在一枚假币)



自信息 (Self-information)
熵
联合熵与条件熵
相对熵与互信息
熵与互信息之间的关系

自信息 (Self-information)
熵
联合熵与条件熵
相对熵与互信息
熵与互信息之间的关系

自信息 (Self-information)

熵

联合熵与条件熵

相对熵与互信息

熵与互信息之间的关系

自信息 (Self-information)
熵
联合熵与条件熵
相对熵与互信息
熵与互信息之间的关系

令 E 为给定事件空间中的一个事件，其概率为 $p_E := \Pr(E)$，其中 $0 \le p_E \le 1$. 令 $\mathcal{I}(E)$——称为事件 E 的自信息——表示当得知事件 E 发生时所获得的信息量（或者等价地，表示在得知 E 发生之前，人们对 E 所抱有的不确定性量）.

一个很自然会问的问题是："$\mathcal{I}(E)$ 应该具备哪些性质?" 尽管这个问题的答案可能因人而异，但以下是人们普遍合理预期 $\mathcal{I}(E)$ 应当具备的几条性质.

自信息 (Self-information)
熵
联合熵与条件熵
相对熵与互信息
熵与互信息之间的关系

1. $\mathcal{I}(E)$ 应当是 p_E 的递减函数.
   换句话说，这条性质表明 $\mathcal{I}(E) = I(p_E)$，其中 $I(\cdot)$ 是定义在 $[0,1]$ 上的实值函数. 此外，人们自然会预期一个越不可能发生的事件 E 发生时，获得的信息量就越多。换句话说，$I(p_E)$ 是 p_E 的递减函数.
2. $I(p_E)$ 应当关于 p_E 连续.
   直观上，人们预期 p_E 的微小变化应该对应于事件 E 所携带信息量的微小变化.
3. 如果 E_1 和 E_2 是独立事件，那么 $\mathcal{I}(E_1 \cap E_2) = \mathcal{I}(E_1) + \mathcal{I}(E_2)$，或者等价地，$I(p_{E_1} \times p_{E_2}) = I(p_{E_1}) + I(p_{E_2})$.
   这条性质表明，当事件 E_1 和 E_2 相互独立时（即它们在概率上互不影响），这两个事件共同发生所获得的信息量，应当等于每个单独事件发生所获得的信息量之和.

自信息 (Self-information)
熵
联合熵与条件熵
相对熵与互信息
熵与互信息之间的关系

接下来，我们将说明唯一满足上述性质 1-3 的函数是对数函数.

定理 1.1

存在唯一一个定义在 $[0,1]$ 上的函数 $I(p)$, 满足以下条件:

1. $I(p)$ 关于 p 单调递减;
2. 在 $0 \le p \le 1$ 范围内，$I(p)$ 是关于 p 的连续函数;
3. $I(p_1 \times p_2) = I(p_1) + I(p_2)$. 该函数为 $I(p) = -c \cdot \log_b(p)$，其中 c 是一个正常数，对数的底数 b 是任何大于 1 的数.

自信息
(Self-information)
熵
联合熵与条件熵
相对熵与互信息
熵与互信息之间的关系

证明.

(1) 我们断言: 对于 $n = 1, 2, 3, \cdots$,

$$I\left(\frac{1}{n}\right) = -c \cdot \log_b\left(\frac{1}{n}\right),$$

其中 $c > 0$ 是一个常数.

自信息 (Self-information)
熵
联合熵与条件熵
相对熵与互信息
熵与互信息之间的关系

证明.

断言的证明: 首先注意到对于 $n = 1$, 由条件 3 可知该断言成立. 这是因为它可以推出 $I(1) = I(1) + I(1)$, 因此 $I(1) = 0 = -c\log_b(1)$.
现在令 n 为一个大于 1 的固定正整数. 条件 1 和条件 3 分别意味着

$$n < m \implies I\left(\frac{1}{n}\right) < I\left(\frac{1}{m}\right) \tag{1.1}$$

以及

$$I\left(\frac{1}{mn}\right) = I\left(\frac{1}{m}\right) + I\left(\frac{1}{n}\right) \tag{1.2}$$

其中 $n, m = 1, 2, 3, \ldots$. 利用 (1.2), 我们可以通过（对 k 的）数学归纳法证明: 对所有非负整数 k,

$$I\left(\frac{1}{n^k}\right) = k \cdot I\left(\frac{1}{n}\right) \tag{1.3}$$

自信息 (Self-information)
熵
联合熵与条件熵
相对熵与互信息
熵与互信息之间的关系

证明.

现在，对于任意正整数 r，都存在一个非负整数 k 使得：

$$n^k \leq 2^r < n^{k+1}.$$

由 (1.1) 可得：

$$I\left(\frac{1}{n^k}\right) \leq I\left(\frac{1}{2^r}\right) < I\left(\frac{1}{n^{k+1}}\right),$$

将其与 (1.3) 结合，可得：

$$k \cdot I\left(\frac{1}{n}\right) \leq r \cdot I\left(\frac{1}{2}\right) < (k+1) \cdot I\left(\frac{1}{n}\right).$$

因此，由于 $I(1/n) > I(1) = 0$，我们有：

$$\frac{k}{r} \leq \frac{I(1/2)}{I(1/n)} \leq \frac{k+1}{r}.$$

自信息 (Self-information)
熵
联合熵与条件熵
相对熵与互信息
熵与互信息之间的关系

证明.

另一方面，根据对数函数的单调性，我们可得：

$$\log_b n^k \le \log_b 2^r \le \log_b n^{k+1} \iff \frac{k}{r} \le \frac{\log_b(2)}{\log_b(n)} \le \frac{k+1}{r}.$$

因此，

$$\left| \frac{\log_b(2)}{\log_b(n)} - \frac{I(1/2)}{I(1/n)} \right| < \frac{1}{r}.$$

由于 n 是固定的，且 r 可以取任意大，我们可以令 $r \to \infty$，从而得到：

$$I\left(\frac{1}{n}\right) = c \cdot \log_b(n),$$

其中 $c = I(1/2)/\log_b(2) > 0$。至此完成了该断言的证明.

自信息 (Self-information)
熵
联合熵与条件熵
相对熵与互信息
熵与互信息之间的关系

证明.

(2) 我们断言: 对于正有理数 p，$I(p) = -c \cdot \log_b(p)$，其中 $c > 0$ 是一个常数. 断言的证明: 一个正有理数 p 可以表示为两个整数的比值，即 $p = r/s$，其中 r 和 s 均为正整数. 那么由条件 3 可得:

$$I\left(\frac{1}{s}\right) = I\left(\frac{r}{s \cdot r}\right) = I\left(\frac{r}{s}\right) + I\left(\frac{1}{r}\right),$$

于是

$$I(p) = I\left(\frac{r}{s}\right) = I\left(\frac{1}{s}\right) - I\left(\frac{1}{r}\right) = c \cdot \log_b s - c \cdot \log_b r = -c \cdot \log_b p.$$

自信息
(Self-information)
熵
联合熵与条件熵
相对熵与互信息
熵与互信息之间的关系

证明.

(3) 对于任意 $p \in [0,1]$，根据有理数在实数集中的连续性与稠密性，可得：

$$I(p) = \lim_{a \uparrow p,\, a \text{ 为有理数}} I(a) = \lim_{b \downarrow p,\, b \text{ 为有理数}} I(b) = -c \cdot \log_b(p). \quad \square$$

自信息 (Self-information)
熵
联合熵与条件熵
相对熵与互信息
熵与互信息之间的关系

上述常数 c 按照惯例通常归一化为 $c = 1$。此外，对数的底数 b 决定了衡量信息量时所使用的单位类型。当底数 $b = 2$ 时，信息量以比特（bits，即二进制位）为单位表示。当底数 $b = e$ 时——即使用自然对数 $(\ln)$——信息量以纳特（nats，即自然单位或自然位）为单位测量。例如，如果事件 E 代表抛掷一枚均匀硬币结果为“正面朝上”，那么它的自信息量（self-information）为：$\mathcal{I}(E) = -\log_2(1/2) = 1$ bit 或者 $-\ln(1/2) = 0.693$ nats.

自信息 (Self-information)

熵

联合熵与条件熵

相对熵与互信息

熵与互信息之间的关系

自信息 (Self-information)

熵

联合熵与条件熵

相对熵与互信息

熵与互信息之间的关系

熵

- 熵是随机变量不确定度的度量.
- 设 X 是一个离散型的随机变量，其字母表 (即概率论中的取值空间) 为 $\mathcal{X}$，概率质量函数 $p(x) = \Pr(X = x)$，$x \in \mathcal{X}$.
- 为方便起见，记概率质量函数为 $p(x)$ 以代替 $p_X(x)$.
- 由此，$p(x)$ 和 $p(y)$ 指两个不同的随机变量，实际上分别表示两个不同的概率质量函数 $p_X(x)$ 和 $p_Y(y)$.

自信息 (Self-information)
熵
联合熵与条件熵
相对熵与互信息
熵与互信息之间的关系

熵

自信息 (Self-information)
熵
联合熵与条件熵
相对熵与互信息
熵与互信息之间的关系

定义 2.1

一个离散型随机变量 X 的**熵** $H(X)$ 定义为

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x).$$

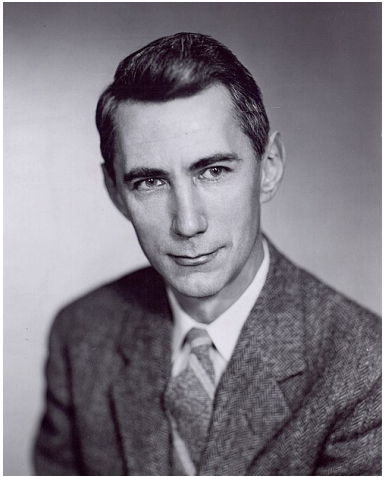
有时也将上面的量记为 $H(p)$. 其中对数的底是 2，熵的单位用比特表示. 由于当 $x \to 0$ 时，$x \log x \to 0$，今后我们约定 $0 \log 0 = 0$.

自信息 (Self-information)
熵
联合熵与条件熵
相对熵与互信息
熵与互信息之间的关系

- 熵 (entropy) 一词源自于希腊语中的 entropia, 其意思为转折点, 变化. 熵的概念是由德国物理学家克劳修斯于 1865 年所提出.
- 物理学家玻尔兹曼于 1870 年将熵的概念引入统计力学.
- 1923 年, 德国科学家普朗克来中国讲学. 讲学时用到 “entropy” 一词, 胡刚复先生在翻译时灵机一动, 把 “商” 字加火旁来意译此词, 创造了 “熵” 字.
- 胡刚复先生说, 此名词的物理含意深奥须多字翻译, 又不合汉字规范, 于是他建议译成一个汉字 “熵”. 火代表与热学有关, 含义为温度 T 与热量 Q 之商, 熵读如商.
- 在 1948 年, 克劳德·艾尔伍德·香农将热力学的熵, 引入到信息论, 因此它又被称为香农熵 (Shannon entropy).

克劳德·艾尔伍德·香农（英语：Claude Elwood Shannon, 1916 年 4 月 30 日—2001 年 2 月 24 日）. 美国数学家、电子工程师和密码学家。香农 1936 年于密歇根大学获学士学位, 1940 年于麻省理工学院获得博士学位. 其在 1948 年发表了划时代的论文——《通信的数学理论》, 这篇论文奠定了现代信息论的基础. 香农还被认为是数字计算机理论和数字电路设计理论的创始人.

推荐读物: Jimmy Soni and Rob Goodman, A Mind at Play: How Claude Shannon Invented the Information Age, Simon and Schuster, 2017. (有中译本, 香农传, 中信出版社, 2019.)



自信息 (Self-information)
熵
联合熵与条件熵
相对熵与互信息
熵与互信息之间的关系

自信息 (Self-information)
熵
联合熵与条件熵
相对熵与互信息
熵与互信息之间的关系

- 如果使用底为 b 的对数，则相应的熵记为 $H_b(X)$. 当对数底为 e 时，熵的单位用**奈特** (nat) 表示.
- 如无特别声明，一般选取对数底为 2.
- 注意，熵实际上是随机变量 X 的分布的泛函，并不依赖于 X 的实际取值，而仅依赖于其概率分布.
- 用 E 表示数学期望. 如果 $X \sim p(x)$，则随机变量 $g(X)$ 的期望值可以表示为

$$E_p g(X) = \sum_{x \in \mathcal{X}} g(x)p(x).$$

- 特别地，当 $g(X) = \log \frac{1}{p(X)}$ 时，$E_p g(X)$ 就是 X 的熵.

自信息 (Self-information)
熵
联合熵与条件熵
相对熵与互信息
熵与互信息之间的关系

引理 2.2

$H(X) \geq 0.$

证明.

由 $0 \leq p(x) \leq 1$ 知 $\log\left(\frac{1}{p(x)}\right) \geq 0.$ □

自信息 (Self-information)
熵
联合熵与条件熵
相对熵与互信息
熵与互信息之间的关系

引理 2.3

$$H_b(X) = (\log_b a) H_a(X).$$

证明.

由 $\log_b p = (\log_b a) \log_a p$ 即可得到. □

自信息 (Self-information)
熵
联合熵与条件熵
相对熵与互信息
熵与互信息之间的关系

例 2.4

设

$$X = \begin{cases} 1 & \text{概率为 } p \\ 0 & \text{概率为 } 1-p. \end{cases}$$

于是

$$H(X) = -p\log p - (1-p)\log(1-p) =: H(p).$$

自信息 (Self-information)
熵
联合熵与条件熵
相对熵与互信息
熵与互信息之间的关系

例 2.5

设

$$X = \begin{cases} a & \text{概率为}\frac{1}{2} \\ b & \text{概率为}\frac{1}{4} \\ c & \text{概率为}\frac{1}{8} \\ d & \text{概率为}\frac{1}{8}. \end{cases}$$

则 X 的熵为

$$H(X) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{4}\log\frac{1}{4} - \frac{1}{8}\log\frac{1}{8} - \frac{1}{8}\log\frac{1}{8} = \frac{7}{4}\text{比特}.$$

自信息 (Self-information)
熵
联合熵与条件熵
相对熵与互信息
熵与互信息之间的关系

下面例子说明随机变量的熵可以是无穷的.

例 2.6: 课本习题 2.19

设 $A = \sum_{n=2}^{\infty}(n\log^2 n)^{-1}$. 定义随机变量 X 满足 $\Pr(X = n) = (An\log^2 n)^{-1}$, 其中 $n = 2, 3, \cdots$, 则 $H(X) = +\infty$.

自信息 (Self-information)
熵
联合熵与条件熵
相对熵与互信息
熵与互信息之间的关系

证明.

由定义, $p_n = \mathsf{Pr}(X = n) = 1/An\log^2 n$, $n \geq 2$. 于是

$$
\begin{aligned}
H(X) &= -\sum_{n=2}^{\infty} p(n)\log p(n) \\
&= -\sum_{n=2}^{\infty} (1/An\log^2 n)\log(1/An\log^2 n) \\
&= \sum_{n=2}^{\infty} \frac{\log(An\log^2 n)}{An\log^2 n} \\
&= \sum_{n=2}^{\infty} \frac{\log A + \log n + 2\log\log n}{An\log^2 n} \\
&= \log A + \sum_{n=2}^{\infty} \frac{1}{An\log n} + \sum_{n=2}^{\infty} \frac{2\log\log n}{An\log^2 n}.
\end{aligned}
$$

自信息 (Self-information)
熵
联合熵与条件熵
相对熵与互信息
熵与互信息之间的关系

证明.

和式中第一项是有限的. 由于我们考虑对数以 2 为底, 最后一项中每一项均是非负的. 对于中间项, 我们有

$$\sum_{n=2}^{+\infty} \frac{1}{An\log n} > \int_2^{\infty} \frac{1}{Ax\log x} dx = K \ln\ln 2|_2^{\infty} = +\infty,$$

其中 $K = \frac{ln2}{A}$. 从而我们有 $H(X) = +\infty$. □

自信息 (Self-information)

熵

联合熵与条件熵

相对熵与互信息

熵与互信息之间的关系

自信息 (Self-information)

熵

联合熵与条件熵

相对熵与互信息

熵与互信息之间的关系

自信息 (Self-information)
熵
联合熵与条件熵
相对熵与互信息
熵与互信息之间的关系

定义 3.1

对于服从联合分布为 $p(x,y)$ 的一对离散随机变量 (X,Y)，其**联合熵** $H(X,Y)$ (joint entropy) 定义为

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

上式亦可表示为

$$H(X,Y) = -E\log p(X,Y).$$

一般地, 对于 n 个随机变量 $X_1,\cdots,X_n$, 它们的联合熵定义为

$$H(X^n) = -\sum_{x_1}\cdots\sum_{x_n} p(x_1,\cdots,x_n)\log\frac{1}{p(x_1,\cdots,x_n)}$$

自信息 (Self-information)
熵
联合熵与条件熵
相对熵与互信息
熵与互信息之间的关系

定义 3.2

若 $(X,Y) \sim p(x,y)$，条件熵 $H(Y|X)$ 定义为：

$$H(Y|X) = \sum_{x \in \mathcal{X}} p(x) H(Y|X=x)$$

自信息 (Self-information)
熵
联合熵与条件熵
相对熵与互信息
熵与互信息之间的关系

定义 3.3

若 $(X,Y) \sim p(x,y)$，条件熵 $H(Y|X)$ 定义为：

$$\begin{aligned} H(Y|X) &= \sum_{x\in\mathcal{X}} p(x) H(Y|X=x) \\ &= -\sum_{x\in\mathcal{X}} p(x) \sum_{y\in\mathcal{Y}} p(y|x) \log p(y|x) \end{aligned}$$

自信息 (Self-information)
熵
联合熵与条件熵
相对熵与互信息
熵与互信息之间的关系

定义 3.4

若 $(X,Y) \sim p(x,y)$，条件熵 $H(Y|X)$ 定义为：

$$
\begin{aligned}
H(Y|X) &= \sum_{x\in\mathcal{X}} p(x) H(Y|X=x) \\
&= -\sum_{x\in\mathcal{X}} p(x) \sum_{y\in\mathcal{Y}} p(y|x) \log p(y|x) \\
&= -\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x,y) \log p(y|x)
\end{aligned}
$$

自信息 (Self-information)
熵
联合熵与条件熵
相对熵与互信息
熵与互信息之间的关系

定义 3.5

若 $(X,Y) \sim p(x,y)$，条件熵 $H(Y|X)$ 定义为：

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

自信息 (Self-information)
熵
联合熵与条件熵
相对熵与互信息
熵与互信息之间的关系

定理 3.6: 链式法则

$$H(X,Y) = H(X) + H(Y|X).$$

自信息 (Self-information)
熵
联合熵与条件熵
相对熵与互信息
熵与互信息之间的关系

证明.

我们有

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

自信息 (Self-information)
熵
联合熵与条件熵
相对熵与互信息
熵与互信息之间的关系

证明.

我们有

$$
\begin{aligned}
H(X,Y) &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y) \\
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log(p(x)p(y|x))
\end{aligned}
$$

自信息 (Self-information)
熵
联合熵与条件熵
相对熵与互信息
熵与互信息之间的关系

证明.

我们有

$$
\begin{aligned}
H(X,Y) &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y) \\
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log(p(x)p(y|x)) \\
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x) - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(y|x)
\end{aligned}
$$

自信息 (Self-information)
熵
联合熵与条件熵
相对熵与互信息
熵与互信息之间的关系

证明.

我们有

$$
\begin{aligned}
H(X,Y) &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y) \\
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log(p(x)p(y|x)) \\
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x) - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(y|x) \\
&= -\sum_{x\in\mathcal{X}} p(x)\log p(x) - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(y|x)
\end{aligned}
$$

自信息 (Self-information)
熵
联合熵与条件熵
相对熵与互信息
熵与互信息之间的关系

证明.

我们有

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

从而得证. □

自信息 (Self-information)
熵
联合熵与条件熵
相对熵与互信息
熵与互信息之间的关系

例 3.7

设 (X, Y) 服从如下的联合分布：

Y \ X	1	2	3	4
1	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{32}$
2	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{32}$	$\frac{1}{32}$
3	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$
4	$\frac{1}{4}$	0	0	0

X 的边际分布为 $(1/2, 1/4, 1/8, 1/8)$，Y 的边际分布为 $(1/4, 1/4, 1/4, 1/4)$，因而 $H(X) = \frac{7}{4}$比特，$H(Y) = 2$比特. 同样地我们可以求得 $H(X|Y) = \sum_{i=1}^{4} p(Y = i)H(X|Y = i) = \frac{11}{8}$比特，$H(Y|X) = \frac{13}{8}$比特，以及 $H(X, Y) = \frac{27}{8}$比特.

自信息 (Self-information)
熵
联合熵与条件熵
相对熵与互信息
熵与互信息之间的关系

注记 3.8

注意 $H(Y|X) \neq H(X|Y)$，但 $H(X) - H(X|Y) = H(Y) - H(Y|X)$，稍后会用到这个性质.

自信息 (Self-information)

熵

联合熵与条件熵

相对熵与互信息

熵与互信息之间的关系

自信息 (Self-information)
熵
联合熵与条件熵
相对熵与互信息
熵与互信息之间的关系

- 相对熵是两个随机分布之间距离的度量.
- 在统计学中，它对应的是似然比的对数期望.
- 相对熵 $D(p\|q)$ 度量当真实分布为 p 而假定分布为 q 时的无效性.
- 例如，已知随机变量的真实分布为 p，可以构造平均长度为 $H(p)$ 的码.
- 但是，如果使用针对分布 q 的编码，那么在平均意义上就需要 $H(p)+D(p\|q)$ 比特来描述这个随机变量.

自信息 (Self-information)
熵
联合熵与条件熵
相对熵与互信息
熵与互信息之间的关系

定义 4.1

两个概率质量函数 $p(x)$ 和 $q(x)$ 之间的**相对熵**或 **Kullback-Leibler 距离**定义为

$$D(p \parallel q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} = E_p \log \frac{p(X)}{q(X)}. \tag{4.1}$$

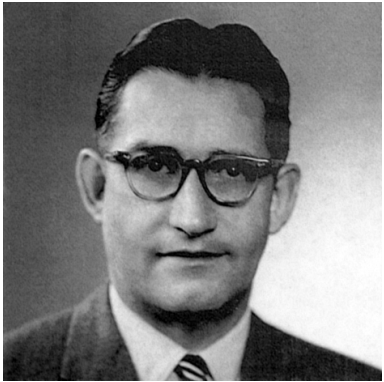
在上述定义中，我们采用约定 $0 \log \frac{0}{0} = 0$，约定 $0 \log \frac{0}{q} = 0$，$p \log \frac{p}{0} = \infty$(基于连续性). 因此，若存在字符 $x \in \mathcal{X}$ 使得 $p(x) > 0$，$q(x) = 0$，则有 $D(p\|q) = \infty$.

所罗门·库尔巴克（Solomon Kullback，1907 年 4 月 3 日 - 1994 年 8 月 5 日）是一位美国密码学家和数学家，在信息论和统计学领域做出了重要贡献. 他最为人熟知的是与理查德·莱布勒（Richard Leibler）共同提出了库尔巴克-莱布勒散度（Kullback–Leibler divergence，简称 KL 散度）. 这一概念在信息论、统计学、机器学习和数据科学中被广泛使用.

自信息 (Self-information)
熵
联合熵与条件熵
相对熵与互信息
熵与互信息之间的关系

自信息 (Self-information)
熵
联合熵与条件熵
相对熵与互信息
熵与互信息之间的关系

理查德·莱布勒（Richard A. Leibler，1914 年 3 月 18 日 - 2003 年 10 月 25 日）是一位美国密码学家和数学家，在信息论和统计学领域做出了重要贡献. 1962-1977 年，他是美国国防分析研究所通信研究部的负责人.



自信息 (Self-information)
熵
联合熵与条件熵
相对熵与互信息
熵与互信息之间的关系

- 稍后我们将证明相对熵总是非负的，并且当且仅当 $p = q$ 时其为零.
- 但是，由于相对熵并不对称，也不满足三角不等式，因此相对熵实际上并非两个分布之间的真正距离.
- 然而，将相对熵视作分布之间的“距离”往往会很有用.

自信息 (Self-information)
熵
联合熵与条件熵
相对熵与互信息
熵与互信息之间的关系

定义 4.2

考虑两个随机变量 X 和 Y，它们的联合概率密度函数为 $p(x,y)$，其边际概率密度函数分别是 $p(x)$ 和 $q(x)$. 互信息 $I(X;Y)$ 为联合分布 $p(x,y)$ 和乘积分布 $p(x)p(y)$ 之间的相对熵，即：

$$\begin{aligned} I(X;Y) &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \\ &= D(p(x,y)\parallel p(x)p(y)) \\ &= E_{p(x,y)}\log\frac{p(X,Y)}{p(X)p(Y)}. \end{aligned} \tag{4.2}$$

自信息 (Self-information)
熵
联合熵与条件熵
相对熵与互信息
熵与互信息之间的关系

例 4.3

设 $\mathcal{X}=\{0,1\}$，考虑 $\mathcal{X}$ 上的两个分布 p 和 q. 设 $p(0)=1-r,\ p(1)=r$ 及 $q(0)=1-s,\ q(1)=s$，则有

$$D(p\|q)=(1-r)\log\frac{1-r}{1-s}+r\log\frac{r}{s}$$

以及

$$D(q\|p)=(1-s)\log\frac{1-s}{1-r}+r\log\frac{s}{r}.$$

如果 $r=s$，则 $D(p\|q)=D(q\|p)=0$. 一般地，$D(p\|q)\neq D(q\|p)$. 例如，若 $r=\frac{1}{2},\ s=\frac{1}{4}$，我们有

$$D(p\|q)=0.2075\text{比特},\ D(q\|p)=0.1887\text{比特}.$$

自信息 (Self-information)

熵

联合熵与条件熵

相对熵与互信息

熵与互信息之间的关系

自信息 (Self-information)
熵
联合熵与条件熵
相对熵与互信息
熵与互信息之间的关系

可将互信息 $I(X;Y)$ 重新写为：

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

自信息 (Self-information)
熵
联合熵与条件熵
相对熵与互信息
熵与互信息之间的关系

可将互信息 $I(X;Y)$ 重新写为:

$$
\begin{aligned}
I(X;Y) &= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\
&= \sum_{x,y} p(x,y) \log \frac{p(x|y)}{p(x)}
\end{aligned}
$$

自信息 (Self-information)
熵
联合熵与条件熵
相对熵与互信息
熵与互信息之间的关系

可将互信息 $I(X;Y)$ 重新写为:

$$
\begin{aligned}
I(X;Y) &= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\
&= \sum_{x,y} p(x,y) \log \frac{p(x|y)}{p(x)} \\
&= -\sum_{x,y} p(x,y) \log p(x) + \sum_{x,y} p(x,y) \log p(x|y)
\end{aligned}
$$

自信息 (Self-information)
熵
联合熵与条件熵
相对熵与互信息
熵与互信息之间的关系

可将互信息 $I(X;Y)$ 重新写为:

$$
\begin{aligned}
I(X;Y) &= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\
&= \sum_{x,y} p(x,y) \log \frac{p(x|y)}{p(x)} \\
&= -\sum_{x,y} p(x,y) \log p(x) + \sum_{x,y} p(x,y) \log p(x|y) \\
&= -\sum_{x} p(x) \log p(x) - (-\sum_{x,y} p(x,y) \log p(x|y))
\end{aligned}
$$

自信息 (Self-information)
熵
联合熵与条件熵
相对熵与互信息
熵与互信息之间的关系

可将互信息 $I(X;Y)$ 重新写为:

$$
\begin{aligned}
I(X;Y) &= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\
&= \sum_{x,y} p(x,y) \log \frac{p(x|y)}{p(x)} \\
&= -\sum_{x,y} p(x,y) \log p(x) + \sum_{x,y} p(x,y) \log p(x|y) \\
&= -\sum_{x} p(x) \log p(x) - (-\sum_{x,y} p(x,y) \log p(x|y)) \\
&= H(X) - H(X|Y)
\end{aligned}
$$

由此知互信息 $I(X;Y)$ 是在给定 Y 知识的条件下 X 的不确定度的缩减量.

自信息 (Self-information)
熵
联合熵与条件熵
相对熵与互信息
熵与互信息之间的关系

对称地，我们亦可以得到

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

因而，X 含有 Y 的信息量等同于 Y 含有 X 的信息量. 又由定理 3.6，可得

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

最后，注意到

$$I(X;X) = H(X) - H(X|X) = H(X).$$

因此，随机变量与自身的互信息为该随机变量的熵.

自信息 (Self-information)
熵
联合熵与条件熵
相对熵与互信息
熵与互信息之间的关系

我们将前面关于熵和互信息的关系做一个总结如下.

定理 5.1

$$I(X;Y) = H(X) - H(X|Y) \tag{5.1}$$

$$I(X;Y) = H(Y) - H(Y|X) \tag{5.2}$$

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \tag{5.3}$$

$$I(X;Y) = I(Y;X) \tag{5.4}$$

$$I(X;X) = H(X) \tag{5.5}$$