信息的度量

May 28, 2023

Yimin Zhao
ym-zhao.com

记号

- X, Y 等表示离散随机变量。
- x, y 等表示离散随机变量取得的确定值。
- $\mathcal{X}, \mathcal{Y}$ 表示有限集，其包含X, Y的所有可能取值，也被称作字母表（alphabet）。
- 给定离散随机变量X, $g(X)$是以X为定义域的一个函数，$\mathbb{E}(g(X))$ 表示$g(X)$的数学期望。
- $p(\cdot)$表示事件$\cdot$的概率, $p(X=x)$可以简写为$p(x)$， $p(X)$指代X的概率密度函数。

熵

X是离散随机变量，概率质量函数 (pmf) 是 $p(x)$，即 $X \sim p(x)$。X 的 **不确定性** 可以由熵 (entropy) 来衡量。其中，熵被定义为：

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

如果把 $\log p(X)$ 视为一个离散随机变量，那么显然：

$$H(X) = -\mathbb{E}(\log p(X))$$

熵的最大值 一个变量的熵的最大值被字母表的大小所限制，当X均匀随机分布时，等号成立：

$$H(X) \leq \log |\mathcal{X}|$$

证明： 由于$H(X)$是凹函数，因此根据Jensen不等式：

$$\begin{aligned} H(X) &= \sum_{x \in \mathcal{X}} p(x) \log \frac{1}{p(x)} \\ &\leq \log \sum_{x \in \mathcal{X}} p(x) \frac{1}{p(x)} \\ &= \log |\mathcal{X}| \end{aligned}$$

特别地，在通信场景下，$\mathcal{X} = \{0, 1\}$, 因此 $H(X) \leq \log |X| = 1$

熵的函数作用 假设 $g(X)$ 是 X 的任意确定性函数，那么：

$$H(X) \geq H(g(X))$$

其中等号成立当且仅当$g(X)$是单射。 证明如下：

$$H(X, g(X)) = H(g(X), X)$$

$$H(X) + \underbrace{H(g(X)|X)}_{=0} = H(g(X)) + \underbrace{H(X|g(X))}_{\geq 0, \text{ 取等当且仅当} g(X) \text{是单射}}$$

联合熵

X, Y是离散随机变量，概率质量函数是$p(x, y)$，联合熵定义为：

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

条件熵

条件熵被定义为：

$$\begin{aligned}
H(X|Y) &= \sum_{y\in\mathcal{Y}} p(Y=y)H(X|Y=y) \\
&= \sum_{y\in\mathcal{Y}} \underbrace{p(Y=y)}_{\text{可以简写为}p(y)} \left(-\sum_{x\in\mathcal{X}} \underbrace{p(X=x|Y=y)}_{\text{可以简写为}p(x|y)} \log \underbrace{p(X=x|Y=y)}_{\text{可以简写为}p(x|y)} \right) \\
&= \sum_{y\in\mathcal{Y}} p(y) \left(-\sum_{x\in\mathcal{X}} p(x|y)\log p(x|y) \right) \\
&= -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y)\log p(x|y)
\end{aligned}$$

性质：

- $H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$ （链式法则，推广可以参考后文）
- 如果X,Y相互独立，那么$H(X,Y) = H(X) = H(Y)$，且$H(X|Y) = H(Y|X) = 0$, 直观理解：给了X对Y的不确定性没有影响（反之亦然）

值得理清的是，条件符号“ | ”优先级低于逗号。因此:

1. $p(x|y,z,k)$ 表示“确定 $(Y=y,Z=z,K=k)$ 后 $X=x$”的条件概率。
2. $p(x,y,z|k)$ 表示“确定 $K=k$ 后 $(X=x,Y=y,Z=z)$”的条件概率。
3. $H(X|Y,Z)$ 表示在“Y, Z”条件下的X的条件熵。
4. $x|y$ 根本就不是一个变量。

链式法则（chain rule）

$$\underbrace{H(X_1,X_2,...,X_n,)}_{X_1\sim X_n\text{的不确定性}} = \underbrace{H(X_1)}_{X_1\text{的不确定性}} + \underbrace{H(X_2|X_1)}_{\text{确定}X_1\text{后}X_2\text{的不确定性}} + \underbrace{H(X_3|X_2,X_1)}_{。。。} + ... + H(X_n|X_{n-1}...X_1)$$

此外，可以对链式法则下的公式进行条件增广。比如已知$H(X,Y) = H(X) + H(Y|X)$, 那么，两边同时加上某条件仍成立，比如：

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

上式也是一个重要公式，通过离散随机变量的熵的非负性可以进一步得到：

$$H(X,Y|Z) \geq H(X|Z)$$

这个不等式的性质又被称为“information never hurts”。这个性质可以进一步推广，得到条件熵的链式法则：

$$H(X_1,X_2,...,X_n \mid Y) = \sum_{i=1}^{n} H(X_i \mid Y,X_1,X_2,...,X_{i-1})$$

相对熵

给定随机变量X的两个概率质量函数$p(x)$与$q(x)$, 相对熵（又称KL距离）被用来衡量这两者的差异，被定义为：

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

性质：

1. $p(x)=0$ 时，$D(p\|q)=0$
2. $q(x)=0\wedge p(x)>0$ 时，$D(p\|q)=\infty$
3. $q(x)=0\wedge p(x)=0$ 时，$D(p\|q)=0$
4. $D(p\|q)\geq 0$, 等号成立当且仅当$p=q$
5. $D(p\|q)=\mathbb{E}_p\left(\log\frac{p(x)}{q(x)}\right)$
   $=\mathbb{E}_p(-\log q(x))-\mathbb{E}_p(-\log p(x))$
   $=\mathbb{E}_p(-\log q(x))-H(X),(X\sim p(x))$

非负性的证明

通过结论：当 $x>0$ 时，$\log x\leq x-1$，可以进行放缩：

$$-D(p\|q)=\sum p\log\frac{q}{p}\leq\sum p\left(\frac{q}{p}-1\right)=\sum q-\sum p\leq 0$$

条件相对熵

给定随机变量X,Y的两个联合概率质量函数$p(x,y)$与$q(x,y)$, 那么可以定义$p(y|x)$与$q(y|x)$之间的条件相对熵：

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

链式法则（chain rule）

$$D(p(x,y)\|q(x,y))=D(p(x)\|q(x))+D(p(y|x)\|q(y|x))$$

互信息

给定字母域$\mathcal{X},\mathcal{Y}$上的两个随机变量X,Y, 它们的联合概率质量函数是$p_{X,Y}(x,y)$, 互信息被定义为$p_{X,Y}(x,y)$和$p_X(x),p_Y(y)$之间的相对熵：

$$\begin{aligned}I(X;Y)&=D\big(p_{X,Y}(x,y)\|p_X(x)p_Y(y)\big)\\&=\sum_{x\in\mathcal{X},y\in\mathcal{Y}}p_{X,Y}(x,y)\log\frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)}\end{aligned}$$

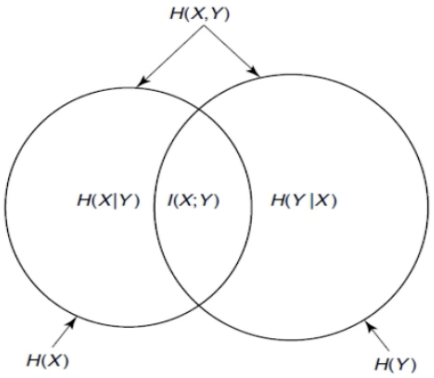
性质：

1. $I(X;Y)=I(Y;X)$
2. $I(X;X)=H(X)$
3. 如果X,Y互相独立，那么由于$p_{X,Y}(x,y)=p_X(x)p_Y(y)$, $I(X;Y)=0$
4. 由于互信息本质是相对熵，因此具有非负性。取0当且仅当X,Y互相独立。

特别的，$I(X;Y)$与$H(X),H(Y)$等满足以下等式关系（等效为Venn图所示）：

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

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



条件互信息

定义随机变量X, Y, Z。给定Z时，X, Y之间的条件互信息被定义为：

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

链式法则（chain rule）

$$I(X_1, X_2, ..., X_n; Y) = \sum_{i=1}^{n} I(X_i; Y|X_{i-1}, X_{i-2}, ..., X_1)$$

直观理解：左式可以分解为每一个X_i与Y的（在给定其他X_i的值时）条件互信息之和。

此外，条件互信息也有链式法则：

$$I(X_1, X_2, ..., X_n; Y|Z) = \sum_{i=1}^{n} I(X_i; Y|\ X_1, X_2, ..., X_{i-1}, Z)$$

应用

假设存在马尔科夫链 $X \to Y \to Z$, 可以进行以下推算：

- 根据互信息链式法则：

$$\begin{aligned} I(X;Y,Z) &= I(X;Z) + I(X;Y|Z) \\ &= I(X;Y) + \underbrace{I(X;Z|Y)}_{=0} \end{aligned}$$

- 根据马尔科夫链的性质，给定Y时，X, Z之间不存在关联，即 $I(X;Z|Y) = 0$
- 因此：
  - $I(X;Y) \geq I(X;Y|Z)$, 物理含意：加了条件Z之后，X, Y之间的相关性反而减小了。
  - $I(X;Y) \geq I(X;Z)$, 物理含意: 原信息X处理的越多，损失的越多。

信息处理不等式

$I(X;Y,Z) \geq I(X;Y)$ 恒成立，而取等号当且仅当$X \to Y \to Z$构成马尔科夫链（即当且仅当$I(X;Z|Y) = 0$）。此外，该马尔科夫链蕴含以下性质：

1. $I(X;Y) \geq I(X;Z)$
2. $I(Z;Y) \geq I(X;Z)$

Reference

1. https://www.info612.ece.mcgill.ca/lecture_02.pdf

2. https://gr.xjtu.edu.cn/c/document_library/get_file?folderId=2422385&name=DLFE-128719.pdf
3. https://moser-isi.ethz.ch/scripts.html
4. http://blog.huangjunqin.com/2018/10/13/information-theory-basis