信息论与编码

马啸
maxiao@mail.sysu.edu.cn

计算机学院
中山大学

2021 年春季学期

1. 自信息

考虑离散随机变量 $X \sim P_X(x)$，$x \in \mathcal{X}$。我们有如下基本概念：

1. 自信息 或 熵密度(Entropy density)

$$I(x) = \log \frac{1}{P_X(x)}$$

我们知道，$\lceil I(x) \rceil$ 是 Shannon 码的码长，因此我们可以近似认为 $I(x)$ 是为了描述 x 所需要的比特数，也可以认为 $I(x)$ 是 $X = x$ 所含的信息量，故称自信息。

2. 熵

$$H(X) = \sum_{x \in \mathcal{X}} P_X(x) \log \frac{1}{P_X(x)}$$

熵是描述随机变量所需比特数的下确界，是熵密度的数学期望。我们可以认为 $H(X)$ 是随机变量的不确定性度量，是模糊度，是揭示 X 所需要的信息量，是得到 X 后所得到的信息量。

3. 相对熵

设 $Q_X(x)$，$x \in \mathcal{X}$ 也是概率向量，即 $Q_X(x) \geq 0$，$\sum_{x\in\mathcal{X}} Q_X(x) = 1$。我们定义

$$D(P||Q) = \sum_{x\in\mathcal{X}} P_X(x) \log \frac{P_X(x)}{Q_X(x)}$$

称之为**相对熵**。通常情况下，$D(P||Q) \neq D(Q||P)$。

相对熵的含义可以粗略解释为，当 Shannon 编码器使用 Q 而不是 P 确定码长时所带来的码长代价。

4. 条件熵

考虑随机向量 $(X, Y) \sim P_{X,Y}(x, y)$，其中$(x, y) \in \mathcal{X} \times \mathcal{Y}$，我们可以定义**条件熵**

$$H(X|Y) = \sum_{y \in \mathcal{Y}} P_Y(y) \sum_{x \in \mathcal{X}} P_{X|Y}(x|y) \log \frac{1}{P_{X|Y}(x|y)}$$

条件熵的含义可以解释为，若已知 Y 的条件下，描述 X 需要的比特数。下面举例说明。

4. 条件熵

例子：现有随机向量 (X, Y) 的联合分布如下

Y\X	a	b	c	d
0	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$
1	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{16}$

(1)求 $H(X)$，$H(X, Y)$和 $H(X|Y)$。

解：计算 X 的分布为 $P_X(x) = \sum_{y\in\mathcal{Y}} P(x, y)$，　Y 分布同理。

$$P(Y) : (1/2,\ 1/2)$$
$$P(X) : (3/8,\ 1/4,\ 3/16,\ 3/16)$$

可计算熵

$$\begin{aligned} H(X) &= \sum_{x\in\mathcal{X}} P_X(x) \log 1/P_X(x) \\ &= 1.9362 \end{aligned}$$

4. 条件熵

计算联合熵和条件熵

$$
\begin{aligned}
H(X,Y) &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P(x,y)\log P(x,y) \\
&= 2.9362 \\
H(X|Y) &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P(x,y)\log P(x|y) \\
&= 1.875
\end{aligned}
$$

4. 条件熵

(2) 构造一个编码方案

$$
\begin{aligned}
\text{若} Y=0, \text{则}\ & a \to 00 \\
& b \to 01 \\
& c \to 10 \\
& d \to 11 \\
\text{若} Y=1, \text{则}\ & a \to 0 \\
& b \to 10 \\
& c \to 110 \\
& d \to 111
\end{aligned}
$$

求该方案的平均码长 $\bar{\ell}$。

解：

$$\bar{\ell} = \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} P(x,y)\ell_{x|y} = 1.875$$

4. 条件熵

(3) 假设收到一段消息序列，试比较已知 Y 取值和不清楚的情况下的编译码过程和相应的平均码长。

若发送者和接收者不知道 Y 的取值，则只能根据 $P(X)$ 对 X 变量进行编码，利用 Shannon 码构造得：

$$a \to 0 \quad b \to 01$$
$$c \to 101 \quad d \to 110$$

译码时根据码表恢复出 X，其平均码长为：$\ell = 2$。

若双方都知道 Y 的取值，则可以根据第(2)小题的编码方式进行恢复，并得到相应的平均码长。若 $Y = 0$，则平均码长为 2。若 $Y = 1$，则平均码长为 1.75。

可以看到，若 Y 与 X 有关联，那么在掌握了 Y 的信息之后，便能够减少对随机变量 X 的不确定性，使得 $H(X|Y)$ 能够小于 $H(X)$ 。

5. 信息密度

给定 $P_{X,Y}(x,y)$，我们可以计算 $P_X(x)$，$P_Y(y)$，$P_{Y|X}(y|x)$，$P_{X|Y}(x|y)$。若不知道 X，则描述 Y 需 $H(Y)$ 比特。若已知 X，则描述 Y 需要 $H(Y|X)$ 比特。

给定 $X=x$，我们有 Y 的条件分布律 $P_{Y|x}(y|x)$，$y\in\mathcal{Y}$。我们定义

$$i(x;y)=\log\frac{P(y|x)}{P(y)}$$

为 (x,y) 点对应的**信息密度**。其可以理解为描述 y 的比特数在已知 $Y=y$ 前后的差：$\log\frac{1}{P(y)}-\log\frac{1}{P(y|x)}$，也可以看作自信息量的“减少”或模糊度的“减少”。

在条件熵的例子中，随机变量 $P(X=a)=3/8$，对应 $\log\frac{1}{P(X=a)}=1.4150$，而在已知 $Y=0$ 的条件下，$P(X=a|Y=0)=1/4$。对应 $\log\frac{1}{P(X=a|Y=0)}=2.0$。因此 $i(Y=0;X=a)=\log\frac{P(X=a|Y=0)}{P(X=a)}=-0.585$

6. 平均互信息

定义 $i(X;Y)$ 的均值为**互信息**，即

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

可以证明

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

6. 平均互信息

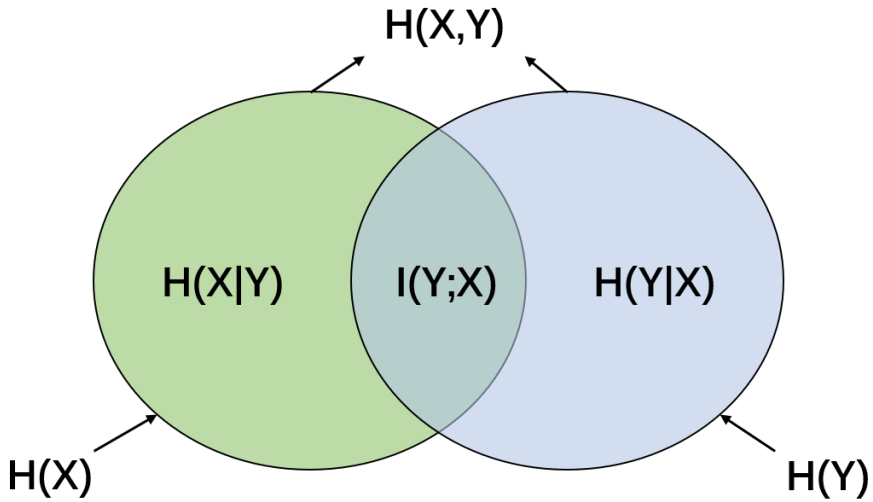

Figure: 互信息

从信源编码的角度，$I(X;Y)$ 可以看作是已知 X 条件下，描述 Y 的平均比特数减少，也可以认为是 X 中包含 Y 的信息量，还可以看作是已知 X的条件下，Y 的模糊度的减少量。

7. 凸函数

记 $\mathbb{R}^n$ 为 n 维实空间，一个集合 D 称为凸集（convex set）是指该集合中任意两点之间的连线仍然在该集合中。准确地说，
若 $x \in D$，$y \in D$，则对于任意的 $0 \leq \alpha \leq 1$，有 $(1-\alpha)x + \alpha y \in D$。

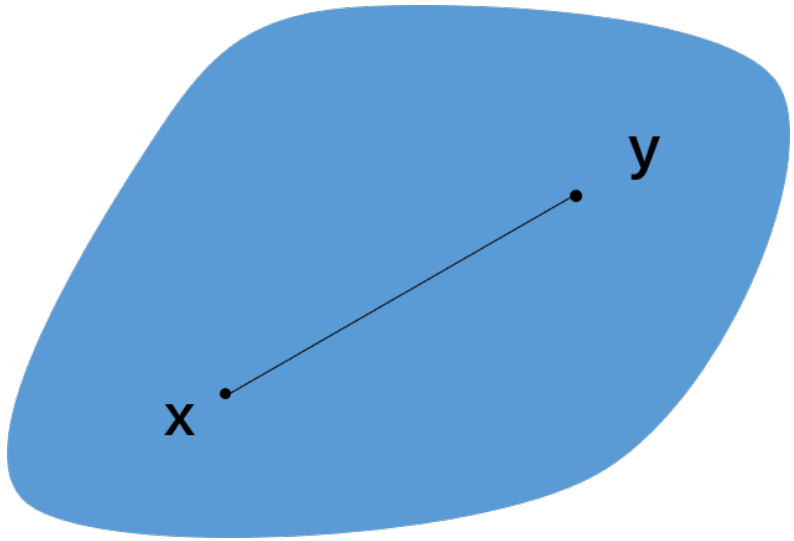

Figure: 凸集

7. 凸函数

在信息论中常见的凸集是 $(p_1, p_2, \cdots, p_n)$，$p_i \geq 0$，$\sum_{i=1}^{n} p_i = 1$。即概率向量的全体。

概率转移矩阵的凸组合也是概率转移矩阵。

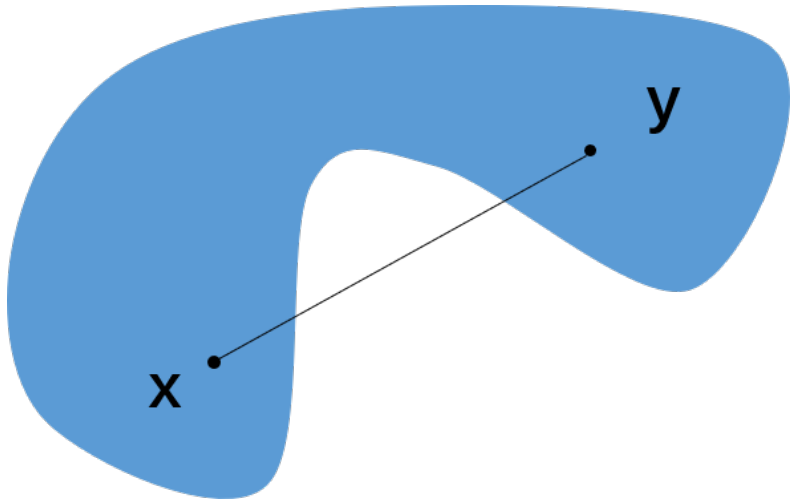

Figure: 非凸集

7. 凸函数

设 $f(x)$，$X \in D \subseteq \mathbb{R}^n$，其中 D 是一个凸集，则 $f(x)$ 称为凸函数，若

$$f((1-\alpha x)+\alpha y) \leq (1-\alpha)f(x)+\alpha f(y)$$

即“线在弦下”。

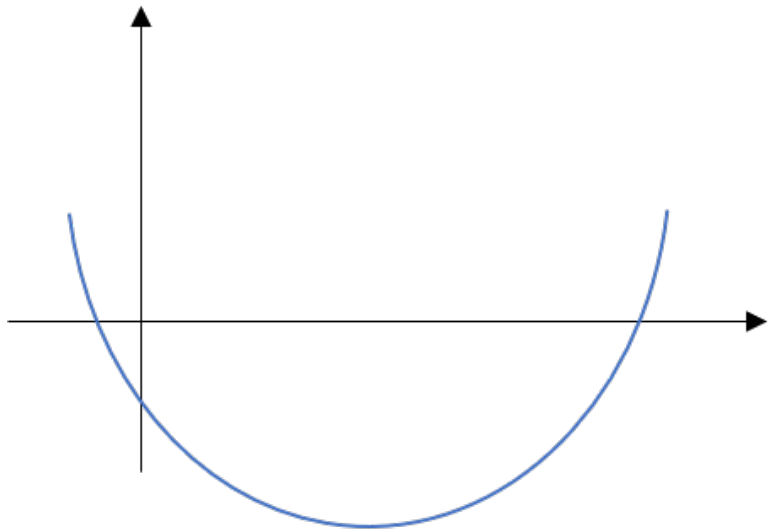

Figure: 凸函数图例

7. 凸函数

若 $-f$ 是凸函数（convex），则称 $f(x)$ 为凹函数（concave）。

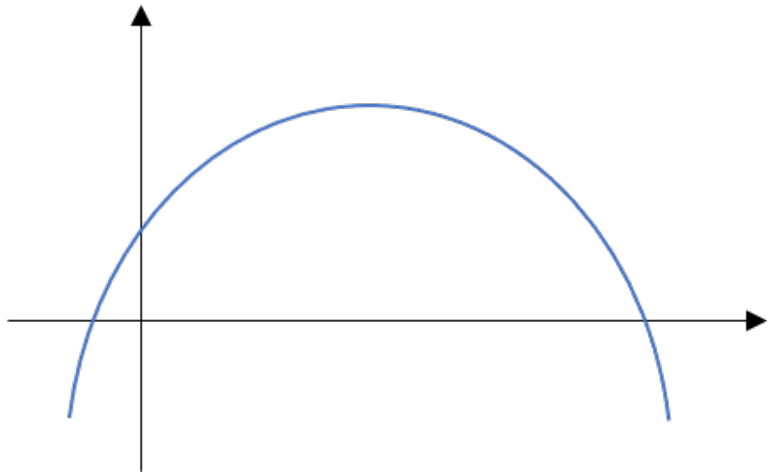

Figure: 凹函数图例

Jensen's 不等式： 如果 f 是convex函数，X 是随机变量，
则 $f(E(X)) \leq E(f(X))$，即数学期望的函数不超过函数的数学期望。

7. 凸函数

例子： 下面哪些函数是convex的，那些是concave的？

(1) $e^x, x \in \mathbb{R}$。

(2)$\ln x, x \in \mathbb{R}_+$。

(3)$x^2, x \in \mathbb{R}$。

(4)$\sqrt{x}, x \in \mathbb{R}_+$。

(5)$|x|, x \in \mathbb{R}$。

(6) $x \ln x, x \in \mathbb{R}_+$。

(7) $H(\mathbf{p}) = -\sum_i p_i \log p_i$，其中 $\mathbf{p}$ 是概率向量。

(8) $I(X;Y) = \sum_{x,y} P(x,y) \log \frac{P(x|y)}{P(x)}$，其中 X 的分布为 $P_X(x)$，X 到 Y 的转移概率矩阵为 $P_{Y|X}(y|x)$。若固定转移概率矩阵，则 $I(X;Y)$ 是 P_X 的concave函数；固定 P_X，则 $I(X;Y)$ 是 $P_{Y|X}$ 的convex函数。

8. 信息不等式

1. $D(P||Q) \geq 0$，当且仅当 $p(x) = q(x)$ 对于所有 $x \in \mathcal{X}$ 成立时，等号成立
2. $I(X;Y) \geq 0$，当且仅当 X 与 Y 独立时等号成立
3. $H(X) \leq \log |\mathcal{X}|$
4. $H(X|Y) \leq H(X)$
5. Fano 不等式：若 $X \to Y \to \hat{X}$，
   则 $H(X|Y) \leq H(P_e) + P_e \log(|\mathcal{X}| - 1)$，其中 $P_e = \Pr\{X \neq \hat{X}\}$。
6. 信息处理定理：对于马尔科夫链 $X \to Z \to Y$，对集 Z 到集 Y 的任意变换 $Y = f(Z)$有

$$H(X|Z) \leq H(X|f(Z)) = H(X|Y)$$
$$I(X;Z) \geq I(X;f(Z)) = I(X;Y)$$

当且仅当 $y = f(z)$ 为可逆函数时等号成立。

作业

Exercise 1.

说明convex函数 f，g 的和 $f+g$ 也是convex的。

Exercise 2.[田宝玉(2008)]

设一个二元信源的符号集为 $\{0,1\}$，有两个概率分布 p 和 q，并且 $p(0)=1-r$，$p(1)=r$，$q(0)=1-s$，$q(1)=s$，求 $D(p||q)$ 和 $D(q||p)$，并分别求当 $r=s$ 和 $r=2s=1/2$ 时两种相对熵的值。

作业

Exercise 3.[田宝玉(2008)]

某城市天气情况与气象预报分别看成包含 {雨, 无雨} 的随机变量集合 X 和 Y，且 X 与 Y 的联合概率为: $P(\text{雨}, \text{雨}) = 1/8$，$P(\text{雨}, \text{无雨}) = 1/16$，$P(\text{无雨}, \text{雨}) = 3/16$，$P(\text{无雨}, \text{无雨}) = 10/16$。求:

(1)气象预报的准确率;

(2)气象预报所提供的关于天气情况的信息量 $I(X; Y)$;

(3)如果天气预报总是预报“无雨”，求此时气象预报的准确率以及气象预报所提供的关于天气情况的信息量 $I(X; Y)$;

(4)以上两种情况相比，哪种情况天气预报准确率高?从信息论的观点看，哪种情况下的天气预报有意义?

作业

Exercise 4.[王育民(2013)]

随机掷 3 颗骰子,以 X 表示第一颗骰子抛掷的结果,以 Y 表示第一和第二颗骰子抛掷的点数之和,以 Z 表示 3 颗骰子的点数之和。试求 $I(Y;Z)$、$I(X;Z)$、$I(X,Y;Z)$、$I(Y;Z|X)$和$I(X;Z|Y)$。

谢谢！