# 信息论导论

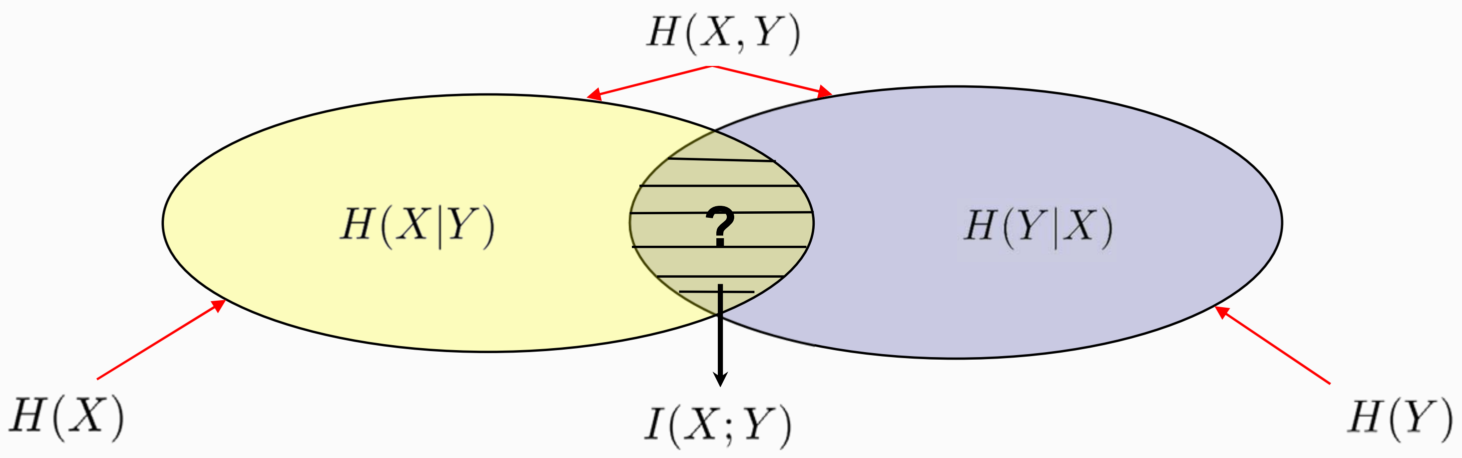
## 第2讲 互信息，相对熵，信息不等式及其推论

[信息论教材中页码范围] 互信息，相对熵：p19~21，
互信息的链式法则：p23~p24，信息不等式及其推论：p25~30

信息学部-信息科学与技术学院 吴绍华
hitwush@hit.edu.cn

# 互信息

哈爾濱工業大學(深圳) HARBIN INSTITUTE OF TECHNOLOGY, SHENZHEN




- 互信息 $I(X;Y)$：观测到 $Y$，所获得的关于 $X$ 的信息

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

**推导自条件熵的变形结果2**

- 互信息具有对称性：

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

**使用分号作为间隔符，以避免多变量情形下指代不明，如：*I(X;Y, Z)*与*I(X,Y;Z)***

# 互信息例题



$p(x, y)$

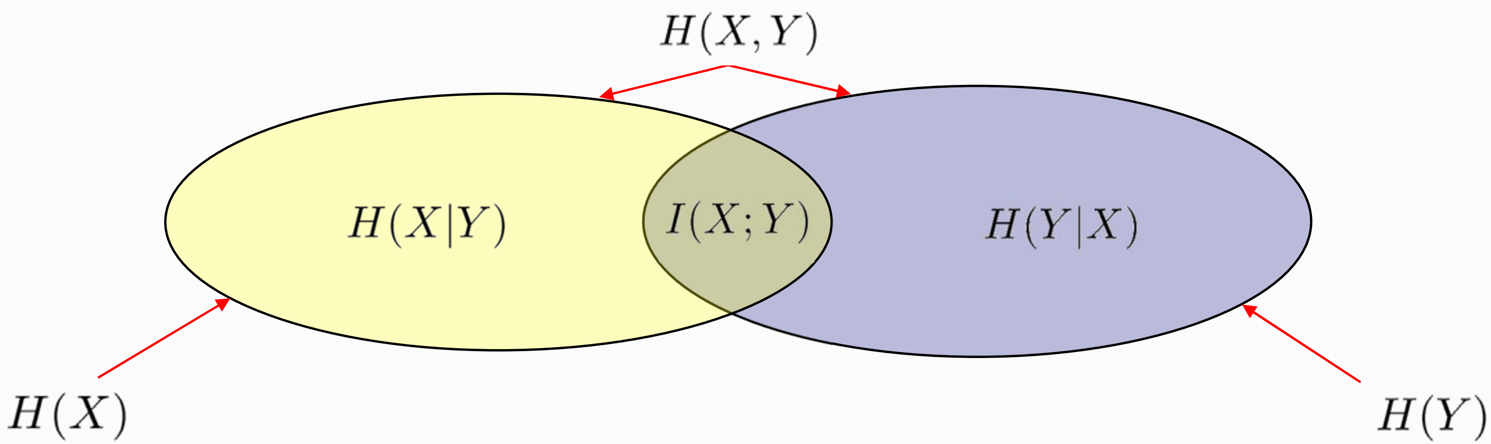
| Y \ X | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | $\frac{1}{8}$ | $\frac{1}{16}$ | $\frac{1}{32}$ | $\frac{1}{32}$ |
| 2 | $\frac{1}{16}$ | $\frac{1}{8}$ | $\frac{1}{32}$ | $\frac{1}{32}$ |
| 3 | $\frac{1}{16}$ | $\frac{1}{16}$ | $\frac{1}{16}$ | $\frac{1}{16}$ |
| 4 | $\frac{1}{4}$ | 0 | 0 | 0 |

- 尝试猜测随机变量 $Y$ 的值时，有 $25\%$ 的概率猜测正确
- 然而，如果在猜测之前已知 $X$ 的值呢？—最佳猜测正确概率总体上可达 $50\%$

$$I(X;Y) = H(X) - H(X|Y) \boxed{= H(X) + H(Y) - H(X,Y)} = 0.375 \text{ bits}$$

# 互信息与熵






**定理 2.4.1 (互信息与熵)**

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

# 条件互信息



- 给定 $Z$ 条件下，随机变量 $X$ 和 $Y$ 的条件互信息为:

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

**给定$(Y, Z)$**

**“给定$Z$”条件同时作用于 $X$ 和 $Y$**

**定理 2.5.2 (互信息的链式法则)**

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

## 复习与前瞻

哈爾濱工業大學(深圳)
HARBIN INSTITUTE OF TECHNOLOGY, SHENZHEN

- **熵**: $H(X)=E(-\log_2 p(x))=-\sum_{x\in\mathscr{X}} p(x)\log_2 p(x)$

  $0\le H(X)\boxed{\le \log|\mathscr{X}|}$

- **链式法则**: $H(X,Y)=H(X)+H(Y|X)$

  $\le H(X)+H(Y)$

  $\boxed{H(Y|X)\le H(Y)}$

- **互信息**: $I(Y;X)=H(Y)-H(Y|X)=H(X)+H(Y)-H(X,Y)$

  $I(Y;X)=\boxed{I(X;Y)\ge 0}$

$\boxed{X \text{ 与 } Y \text{ 互相独立} \iff H(X,Y)=H(X)+H(Y) \iff I(X;Y)=0}$

如何证明?

信息不等式

# 凸函数与凹函数

哈爾濱工業大學(深圳)
HARBIN INSTITUTE OF TECHNOLOGY, SHENZHEN

**定义**

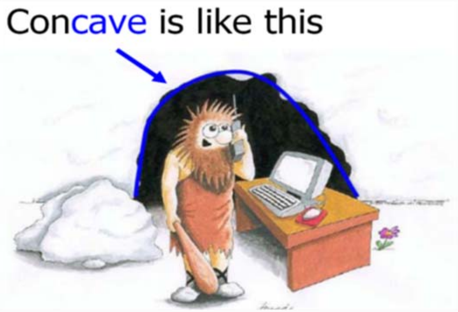
如果 $f(x)$ 在区间 $(a,b)$ 上对所有的 $x_1, x_2 \in (a,b)$ 以及 $0 \le \lambda \le 1$ 都满足:

$$f(\lambda x_1 + (1-\lambda)x_2) \le \lambda f(x_1) + (1-\lambda)f(x_2)$$

那么 $f(x)$ 可以被称为在 $(a,b)$ 区间的凸函数，上式也可以写为:

$$f(x_1 + (1-\lambda)(x_2 - x_1)) \le f(x_1) + (1-\lambda)(f(x_2) - f(x_1))$$

- 严格凸函数：如果上述不等式仅在 $\lambda = 0$ 或 $\lambda = 1$ 时取得等号，那么我们称这个凸函数为严格凸函数，严格凹函数同理。
- 如果 $f(x)$ 是严格凸函数，那么 $f(x)$ 上的每根弦都在 $f(x)$ 函数曲线的上方。
- 如果 $f(x)$ 是凹函数那么 $-f(x)$ 为凸函数，反之同理。




**例**

- 严格凸函数: $x^2, x^4, e^x, x\log x[x \ge 0]$
- 严格凹函数: $\log x, \sqrt{x}[x \ge 0]$
- 既是凸函数也是凹函数: $x$

# 函数凸凹性的判断

哈爾濱工業大學(深圳)
HARBIN INSTITUTE OF TECHNOLOGY, SHENZHEN

**定理 2.6.1**

$f''(x) \geq 0, \forall x \in (a,b) \Rightarrow f(x)$ 为凸函数 (当 $f''(x) > 0$ 时为严格凸函数)

**证明.**

函数 $f(x)$ 在 $x_0$ 点泰勒级数展开：

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \boxed{\frac{f''(x^*)}{2}}(x - x_0)^2 \geq f(x_0) + f'(x_0)(x - x_0)$$

>= 0

令$x_0 = \lambda\, x_1 + (1-\lambda)x_2$ 且令$x = x_1$，我们能得到：

$$f(x_1) \geq f(x_0) + f'(x_0)[(1-\lambda)(x_1 - x_2)] \tag{1}$$

令 $x = x_2$ 我们得到：

$$f(x_2) \geq f(x_0) + f'(x_0)[\lambda(x_2 - x_1)] \tag{2}$$

令 (1) 式乘以 $\lambda$ 加上 (2) 式乘以 $(1-\lambda)$ 即可得证 □

# Jensen不等式



**定理 2.6.2 (Jensen 不等式)**

(a) $f(x)$ 为凸函数 $\Rightarrow Ef(X) \geq f(EX)$

(b) $f(x)$ 严格凸函数 $\Rightarrow Ef(X) > f(EX)$，或仅当 $X$ 为常数时有 $Ef(X) = f(EX)$?

**证明.**

按照 $|\mathscr{X}|$ 的取值，采用数学归纳法：

- $|\mathscr{X}| = 2$：$p_1 f(x_1) + p_2 f(x_2) \geq f(p_1 x_1 + p_2 x_2)$
- $\boxed{|\mathscr{X}| = k-1\text{：假定不等式成立}}$
- $|\mathscr{X}| = k$：

$$Ef(X) = \sum_{i=1}^{k} p_i f(x_i) = p_k f(x_k) + \sum_{i=1}^{k-1} p_i f(x_i)$$

$$= p_k f(x_k) + (1-p_k) \boxed{\sum_{i=1}^{k-1} \frac{p_i}{1-p_k} f(x_i) \geq} p_k f(x_k) + (1-p_k) f \boxed{\left( \sum_{i=1}^{k-1} \frac{p_i}{1-p_k} x_i \right)}$$

$$\geq f\left( p_k x_k + (1-p_k) \sum_{i=1}^{k-1} \frac{p_i}{1-p_k} x_i \right) = f(EX)$$

# 相对熵

哈爾濱工業大學(深圳) HARBIN INSTITUTE OF TECHNOLOGY, SHENZHEN

**定义**

两个概率质量向量 $\mathbf{p}$ 和 $\mathbf{q}$ 之间的相对熵或 Kullback-Leibler 距离（也叫 KL 散度）定义为：

$$D(\mathbf{p}||\mathbf{q}) = \sum_{x\in\mathscr{X}} p(x)\log\frac{p(x)}{q(x)} = \underbrace{E_{\mathbf{p}}}_{E_{\mathrm{p}}\text{表示以分布}\mathbf{p}\text{计算期望}}\log\frac{p(x)}{q(x)} = \underbrace{\boxed{E_{\mathbf{p}}(-\log q(x))}}_{\text{交叉熵，记为}H(\mathrm{p},\,\mathrm{q})} - H(\mathbf{p})$$

- 相对熵 $D(\mathbf{p}||\mathbf{q})$ 用于衡量两种分布（$\mathbf{p}$ 和 $\mathbf{q}$）之间的差异程度，并非真实的“距离”（不满足对称性、不满足三角不等式）；
- 相对熵与交叉熵在机器学习中被广泛应用，例如交叉熵常被用作分类任务的损失函数，相对熵则常用于生成模型（如生成对抗网络 GAN、变分自编码器 VAE）、强化学习中。

# 相对熵例题



**例**

$\mathscr{X} = [1;2;3;4;5;6]$

$\mathbf{p} = [\frac{1}{6};\frac{1}{6};\frac{1}{6};\frac{1}{6};\frac{1}{6};\frac{1}{6}] \Rightarrow H(\mathbf{p}) = 2.585$

$\mathbf{q} = [\frac{1}{10};\frac{1}{10};\frac{1}{10};\frac{1}{10};\frac{1}{10};\frac{1}{2}] \Rightarrow H(\mathbf{q}) = 2.161$

$D(\mathbf{p}||\mathbf{q}) = E_{\mathbf{p}}(-\log q(x)) - H(\mathbf{p}) = 2.935 - 2.585 = \boxed{0.35}$

$D(\mathbf{q}||\mathbf{p}) = E_{\mathbf{q}}(-\log p(x)) - H(\mathbf{q}) = 2.585 - 2.161 = \boxed{0.424}$

$D(\mathbf{p}||\mathbf{q}) \neq D(\mathbf{q}||\mathbf{p})$

# 信息不等式

哈爾濱工業大學(深圳) HARBIN INSTITUTE OF TECHNOLOGY, SHENZHEN

**定理 2.6.3 (信息不等式)**

$D(\mathbf{q}||\mathbf{p}) \geq 0$，当且仅当 $\mathbf{p} \equiv \mathbf{q}$，即 $\mathbf{p}$ 和 $\mathbf{q}$ 为同一分布时，等号成立

**证明.**

设 $\mathscr{A} = \{x : p(x) > 0\} \subseteq \mathscr{X}$，则

$$-D(\mathbf{p}||\mathbf{q}) = -\sum_{x\in\mathscr{A}} p(x)\log\frac{p(x)}{q(x)} = \sum_{x\in\mathscr{A}} p(x)\log\frac{q(x)}{p(x)}$$

$$\boxed{\leq} \log\left(\sum_{x\in\mathscr{A}} p(x)\frac{q(x)}{p(x)}\right) = \log\left(\sum_{x\in\mathscr{A}} q(x)\right) \leq \log\left(\sum_{x\in\mathscr{X}} q(x)\right) = \log 1 = 0$$

若 $D(\mathbf{p}||\mathbf{q}) = 0$，考虑到 log() 是严格凹函数，则需其括号内参量部分 $\frac{q(x)}{p(x)}$ 恒为常数时才能成立。又由于 $\sum_{x\in\mathscr{X}} p(x) = \sum_{x\in\mathscr{X}} q(x) = 1$，所以该常数必定为 1，进而可得取等条件为 $\mathbf{p} \equiv \mathbf{q}$。 □

# 信息不等式推论 (1)

哈爾濱工業大學(深圳)
HARBIN INSTITUTE OF TECHNOLOGY, SHENZHEN

**定理 2.6.4 (熵的均匀分布界——离散随机变量的最大熵)**

$H(X) \leq \log|\mathscr{X}|$，当且仅当 $X$ 服从 $\mathscr{X}$ 上的均匀分布时等号成立。

**证明.**

设 $\mathbf{q}$ 为均匀分布，即 $\mathbf{q} = [|\mathscr{X}^{-1}|, ..., |\mathscr{X}^{-1}|]^T$，则 $H(\mathbf{q}) = \log|\mathscr{X}|$。设 $\mathbf{p}$ 为 $X$ 的任意分布，考查其与 $\mathbf{q}$ 的相对熵，由信息不等式有：

$$D(\mathbf{p}||\mathbf{q}) = E_{\mathbf{p}}(-\log q(x)) - H(\mathbf{p}) = \log|\mathscr{X}| - H(\mathbf{p}) \geq 0$$

□

# 信息不等式推论 (2)



**推论（互信息非负性）**

$I(X;Y) \geq 0$，当且仅当 $X$ 与 $Y$ 相互独立时等号成立。

**证明.**

$$I(X;Y) = H(X) + H(Y) - H(X,Y) = E\log\frac{p(x,y)}{p(x)p(y)} = D(\mathbf{p}_{x,y}||\mathbf{p}_x\mathbf{p}_y^T) \geq 0$$

当且仅当 $p(x,y) \equiv p(x)p(y)$，即 $X$ 与 $Y$ 相互独立时，等号成立。 □

# 信息不等式推论 (3)

哈爾濱工業大學(深圳)
HARBIN INSTITUTE OF TECHNOLOGY, SHENZHEN

**定理 2.6.5（条件作用使熵减小）**

$H(Y|X) \leq H(Y)$，当且仅当 $X$ 与 $Y$ 相互独立时等号成立。

**证明.**

$$I(X;Y) = H(Y) - H(Y|X) \geq 0$$

- 定理 2.6.5 表明：$Y$ 的不确定性在获知另一个随机变量 $X$ 的值的情况下会减小
- 但要注意这只在平均意义（统计意义）上成立，若只获知 $X$ 的某个特定值，则 $Y$ 的不确定性并不一定减小。

举例说明

# 信息不等式推论 (4)

哈爾濱工業大學(深圳)
HARBIN INSTITUTE OF TECHNOLOGY, SHENZHEN

**定理 2.6.6 （熵的独立界）**

$H(X_1, X_2, ..., X_n) \leq \sum_{i=1}^{n} H(X_i)$，当且仅当 $X_1, X_2, ..., X_n$ 相互独立时等号成立。

**证明.**

条件使熵减小

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

当且仅当对所有的 $i$，$X_i$ 与 $X_{i-1}, ..., X_1$ 独立，即当且仅当 $X_1, X_2, ..., X_n$ 相互独立时，等号成立。 □

# 阶段小结

哈爾濱工業大學(深圳) HARBIN INSTITUTE OF TECHNOLOGY, SHENZHEN

- **互信息**：$I(X;Y) = H(X) - H(X|Y)$
- **Jensen 不等式**：如果 $f(x)$ 为凸函数，则 $Ef(X) \geq f(EX)$
- **相对熵**：$D(\mathbf{p}||\mathbf{q}) = E_{\mathbf{p}} \log \frac{p(x)}{q(x)}$
- **信息不等式**：$D(\mathbf{p}||\mathbf{q}) \geq 0$ 当且仅当 $\mathbf{p} \equiv \mathbf{q}$，等号成立
- **信息不等式推论**：
  - 熵的均匀分布界：$H(\mathbf{p})$ 在 $\mathbf{p}$ 为均匀分布时取得最大值
  - $I(X;Y) \geq 0$，并由此可推出“条件作用使熵减小”
  - 熵的独立界：$H(X_1, X_2, \ldots, X_n) \leq \sum_{i=1}^{n} H(X_i)$

# 结束