EE142
Fundamentals of Information Theory

Lecture notes

Zhou Shouchen

2024 年 9 月 30 日

目录

第一章　Entropy, Relative Entropy, Mutual Information

$\log x$ 若无特殊说明, 默认为 $\log_2 x$, $0 \log 0 = 0$.

1.1　Entropy

定义 1.1.1. 事件 x 发生的概率为 $p(x)$, 则 x 的信息量为 $\log \frac{1}{p(x)}$.

定义 1.1.2. 离散型随机变量 X 的熵 *(entropy)* $H(X)$:

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

$H(X)$ 物理意义: 事件发生的概率 $p(x)$, 信息量为 $\log \frac{1}{p(x)}$. 所有事件发生的期望信息量.

命题 1.1.3. $H(X) \geq 0$, 当且仅当 $p(x) = 1$ 时, $H(X) = 0$.
$p(x) = 1$ 时, 事件是确定的 *(deterministic)*, 信息量为 *0*.

定义 1.1.4. $X \in \mathcal{X}, Y \in \mathcal{Y}; |\mathcal{X}|, |\mathcal{Y}| < \infty$(离散型随机变量).

X 和 Y 的联合熵 *(joint entropy)* $H(X,Y)$:

$$\begin{aligned} H(X,Y) &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y) \\ &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{1}{p(x,y)} \\ &= \mathbb{E}\left[\log\frac{1}{p(x,y)}\right] \end{aligned}$$

命题 1.1.5. $0 \le H(X) \le \log|\mathcal{X}|$.

X 为冲激函数时取 0 *(deterministic)*, X 为均匀分布时取 $\log|\mathcal{X}|$.

定义 1.1.6. 条件熵 *(conditional entropy)* $H(Y|X)$:

$$\begin{aligned} H(Y|X) &= \sum_{x\in\mathcal{X}} p(x)H(Y|X=x) \\ &= -\sum_{x\in\mathcal{X}} p(x)\sum_{y\in\mathcal{Y}} p(y|x)\log p(y|x) \\ &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{1}{p(y|x)} \\ &= \mathbb{E}\left[\log\frac{1}{p(y|x)}\right] \end{aligned}$$

定理 1.1.7. *chain rule* 剥洋葱:

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

例 1.1.8. *Find* $H(X), H(Y), H(X|Y), H(X,Y)$.

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

$$H(X) = H\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}\right) = \frac{7}{4} bits$$

$$H(Y) = H\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right) = \log 4 = 2 bits \ (\textcolor{red}{uniform\ distribution})$$

$$\begin{aligned}
H(X|Y) &= \sum_{y\in\mathcal{Y}} p(y) H(X|Y=y) \\
&= \frac{1}{4}\left(H\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}\right) + H\left(\frac{1}{4}, \frac{1}{2}, \frac{1}{8}, \frac{1}{8}\right) + H\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right) + H(1)\right) \\
&= \frac{11}{8} bits
\end{aligned}$$

$$H(X,Y) = H(Y) + H(X|Y) = 2 + \frac{11}{8} = \frac{27}{8}\ bits.$$

命题 1.1.9.

$$\begin{aligned}
H(X,Y) &= H(X) + H(Y) \quad && (X \perp Y) \\
H(X,Y) &= H(X) && (Y = X)
\end{aligned}$$

1.2 Relative Entropy and Mutual Information

概率论衡量两个变量相关程度 (概率论方法):

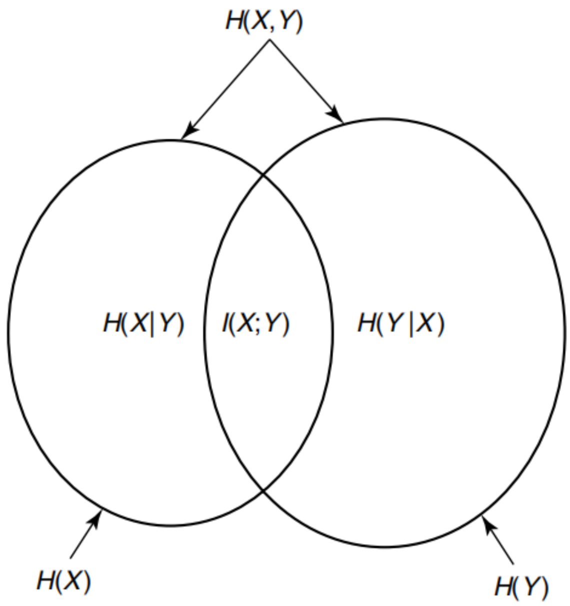
$$\rho_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}} \in [-1, 1]$$

只能刻画**线性**相关性, 且正负相关程度相同 (正负相关).

命题 1.2.1. *X,Y 独立, 则 $\rho_{X,Y} = 0$. 但是 $\rho_{X,Y} = 0$* **不一定***独立.*

Gaussian 分布独立 $\Leftrightarrow$ 不相关.

信息论衡量方法 (用 bit 衡量):



Relationship between entropy and mutual information.

定义 1.2.2. *$I(X;Y)$: X,Y 之间的互信息 (mutual information).*

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

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

proof:

$$\begin{aligned}
I(X;Y) &= \sum_{x,y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \\
&= \sum_{x,y} p(x,y)\log\frac{p(x|y)p(y)}{p(x)p(y)} \\
&= \sum_{x,y} p(x,y)\log\frac{p(x|y)}{p(x)} \\
&= \sum_{x,y} p(x,y)\log\frac{1}{p(x)} - \sum_{x,y} p(x,y)\log\frac{1}{p(x|y)} \\
&= \sum_{x} p(x)\log\frac{1}{p(x)} - H(X|Y) \quad (p(x,y) \text{ 对 } y \text{ 求和, 求出 margin distribution } p(x)) \\
&= H(X) - H(X|Y)
\end{aligned}$$

定义 1.2.3. 两个分布 $p(x), q(x)$ 的相对熵 *Relative Entropy(KL-Divergence):*

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

$x \in \mathcal{X}$: 不考虑两个 *support* 不同的分布.

物理意义: 两个分布之间的距离.

命题 1.2.4. $D\left(p(x)\|q(x)\right) \geq 0$.

proof:

$$\begin{aligned}
-D\left(p(x)\|q(x)\right) &= \sum_{x} p(x)\log\frac{q(x)}{p(x)} \\
&= \mathbb{E}_{x\sim p(x)}\left[\log\frac{q(x)}{p(x)}\right] \\
&\leq \log\mathbb{E}_{x\sim p(x)}\left[\frac{q(x)}{p(x)}\right] \qquad \textit{(Jensen's Inequality)} \\
&= \log\sum_{x} p(x)\frac{q(x)}{p(x)} \\
&= 0
\end{aligned}$$

i.e. $D\left(p(x)\|q(x)\right) \geq 0$.

当且仅当 $p(x) = q(x)$ 时等号成立 (*Jensen's Inequality* 成立条件: 函数是线性的).

命题 1.2.5. $I(X;Y) = I(Y;X)$
$D\left(p(x)\|q(x)\right) \neq D\left(q(x)\|p(x)\right)$
$I(X;Y) = D\left(p(x,y)\|p(x)p(y)\right)$

命题 1.2.6. *1.* $I(X;Y) \geq 0$;
2. $I(X;Y) \leq \min H(X), H(Y)$

1. $I(X;Y) \geq 0$: 当且仅当 X,Y 独立时等号成立.

$$\begin{aligned} I(X;Y) &= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\ &= D\left(p(x,y)\|p(x)p(y)\right) \geq 0 \end{aligned}$$

当且仅当 $p(x,y) = p(x)p(y)$ 时等号成立, 即 X,Y 独立.

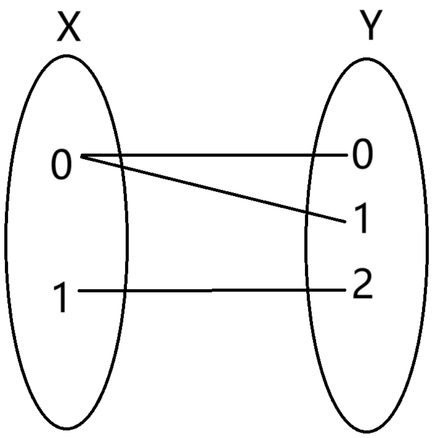
2. $I(X;Y) \leq \min\{H(X), H(Y)\}$:

Since $H(X) \geq 0$, similarly, $H(X|Y) \geq 0$.

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

当且仅当 $H(X|Y)$=0 时等号成立.
同理 $I(X;Y) \leq H(Y)$, 当且仅当 $H(Y|X)$=0 时等号成立.

命题 1.2.7. *即使* $H(X|Y) = 0$, *也无法得出* X, Y *有关系.*
如图 $H(X|Y) = 0, H(Y|X) \neq 0$.



命题 1.2.8. *Conditioning Reduces Entropy(Information can't hurt):*

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

proof:

$$\begin{aligned} & I(X;Y) = H(X) - H(X|Y) \geq 0 \\ \Rightarrow & H(X) \geq H(X|Y) \end{aligned}$$

当且仅当 $I(X;Y) = 0$, 即 X, Y 独立时取等.

将 X, Y 两个分布的性质拓展到 n 个分布:

多元 KL 散度:

$$D\left(p(x,y) \| q(x,y)\right) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{q(x,y)}$$

$$D\left(p(y|x) \| q(y|x)\right) = \sum_{x,y} p(x,y) \log \frac{p(y|x)}{q(y|x)}$$

无论 KL 散度的形式如何, log 前都是 $p(x,y)$!!!

命题 1.2.9. *1. Chain Rule:*

<1> Entropy's Chain Rule:

$$\begin{aligned} H(X_1, X_2, \cdots, X_n) &= H(X_1) + H(X_2|X_1) + \cdots + H(X_n|X_{n-1}, \cdots, X_1) \\ &= \sum_{i=1}^{n} H(X_i|X_1, \cdots, X_{i-1}) \\ &= \sum_{i=1}^{n} H(X_i|X_{i+1}, \cdots, X_n) \end{aligned}$$

<2> Mutual Information's Chain Rule:

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

proof:

$$\begin{aligned}
I(X_1, X_2, \cdots, X_n; Y) &= H(X_1, X_2, \cdots, X_n) - H(X_1, X_2, \cdots, X_n|Y) \\
&= \sum_{i=1}^{n} H(X_i|X_1, \cdots, X_{i-1}) - \sum_{i=1}^{n} H(X_i|X_1, \cdots, X_{i-1}, Y) \\
&= \sum_{i=1}^{n} [H(X_i|X_1, \cdots, X_{i-1}) - H(X_i|X_1, \cdots, X_{i-1}, Y)] \\
&= \sum_{i=1}^{n} I(X_i; Y|X_1, \cdots, X_{i-1})
\end{aligned}$$

<3> KL-Divergence's Chain Rule:

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

proof:

$$\begin{aligned}
D\left(p(x,y)\|q(x,y)\right) &= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{q(x,y)} \\
&= \sum_{x,y} p(x,y) \log \frac{p(x)p(y|x)}{q(x)q(y|x)} \\
&= \sum_{x,y} p(x,y) \log \frac{p(x)}{q(x)} + \sum_{x,y} p(x,y) \log \frac{p(y|x)}{q(y|x)} \\
&= D\left(p(x)\|q(x)\right) + D\left(p(y|x)\|q(y|x)\right)
\end{aligned}$$

$$\Rightarrow D\left(p(x_1, x_2, \cdots, x_n)\|q(x_1, x_2, \cdots, x_n)\right) = \sum_{i=1}^{n} D\left(p(x_i)\|q(x_i)\right)$$

2. Mutual Information $\Rightarrow$ Conditional Mutual Information:

$I(X;Y|Z)$: *given* Z, X,Y 的互信息.

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

已知 $H(X) \geq H(X|Y)$, 但是 $I(X;Y)$ 和 $I(X;Y|Z)$ 大小关系不确定.

例 1.2.10. $I(X;Y|Z) > I(X;Y)$
$X, Y \overset{i.i.d.}{\sim} \text{Bern}\left(\frac{1}{2}\right)$, $Z = X + Y$
$X \perp Y \Rightarrow I(X;Y) = 0$

$$
\begin{aligned}
I(X;Y|Z) &= H(X|Z) - H(X|Y,Z) \\
&= H(X|Z) \quad \textit{(X=Z-Y, deterministic)} \\
&= P(Z=0)H(X|Z=0) + P(Z=1)H(X|Z=1) + P(Z=2)H(X|Z=2) \\
&= 0 + \frac{1}{2}H(X|Z=1) + 0 \\
&> 0
\end{aligned}
$$

$I(X;Y|Z) > I(X;Y)$

例 1.2.11. $I(X;Y|Z) \leq I(X;Y)$ *Construct Markov Chain:* $X \to Y \to Z$

$$
\begin{aligned}
p(x,y,z) &= p(x)p(y|x)p(z|y) \\
I(X;Y,Z) &= I(Y,Z;X) \\
&= I(Y;X) + I(Z;X|Y) \\
&= I(Z;X) + I(Y;X|Z)
\end{aligned}
$$

Since $Z \perp X|Y \Rightarrow I(Z;X|Y) = 0$
So $I(Y;X) = I(Z;X) + I(Y;X|Z) \geq I(Y;X|Z)$
所以 $I(X;Y) \geq I(X;Y|Z)$

proof $Z \perp X|Y \Rightarrow I(Z;X|Y) = 0$:

$$
\begin{aligned}
I(Z;X|Y) &= H(Z|Y) - H(Z|X,Y) \\
H(Z|X,Y) &= \sum_{x,y,z} p(x,y,z) \log \frac{1}{p(z|x,y)} \\
&= \sum_{x,y,z} p(x,y,z) \log \frac{1}{p(z|y)} = \sum_{y,z} p(y,z) \log \frac{1}{p(z|y)} \\
&= H(Z|Y)
\end{aligned}
$$

$\Rightarrow I(Z;X|Y) = 0$

$Y \backslash X$	0	1
0	0	$\frac{3}{4}$
1	$\frac{1}{8}$	$\frac{1}{8}$