第二章 熵与互信息.

entropy mutual information

2.1 熵

熵,是随机变量的不确定度的度量。

定义: 一个离散型随机变量 X 的熵 $H(X)$:

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

log 用的底是 2, 单位: 比特

e 奈特 (nat)

注: $H(X) = E_p \log\frac{1}{p(X)}$

X 的熵又解释为随机变量 $\log\frac{1}{p(X)}$ 的期望值

$p(x)$ 是 X 的概率密度函数

$$E_p \log\frac{1}{p(X)} = \sum_{x\in\mathcal{X}} \log\frac{1}{p(x)}\cdot p(x)$$

$$= H(X)$$

引理 2.1.1 $H(X)\geqslant 0$

$\because 0\leq p(x)\leq 1 \quad \therefore \log\frac{1}{p(x)}\geqslant 0$

引理 2.1.2 $H_b(X) = (\log_b a) H_a(X)$

定理: 当 $P_1^{\#}=P_2^{\#}=\cdots=P_n^{\#}=\frac{1}{n}$ 时, $H[P_1^{\#},P_2^{\#},\cdots,P_n^{\#}]\geqslant H[P_1,P_2,\cdots,P_n]$

证明: 拉格朗日常数法: $\max H(P_1,P_2,\cdots,P_n)$

$$\sum_{i=1}^{n} P_i = 1 \ , \ P_i\geqslant 0$$

$$L(P_1,P_2,\cdots P_n) = H(P_1,P_2,\cdots,P_n) + \lambda\left(\sum_{i=1}^{n} P_i - 1\right)$$

$$\frac{\partial L}{\partial P_i} = \frac{\partial H}{\partial P_i} + \lambda\cdot\frac{\partial}{\partial P_i}\left(\sum_{i=1}^{n} P_i - 1\right)$$

$$= -\log P_i - 1 + \lambda = 0$$

$\therefore \log P_i = \lambda - 1$

$\therefore P_1 = P_2 = \cdots = P_n$

即 $P_i = \frac{1}{n}$

设 X 是一个离散型随机变量, 其(概率论中的)取值空间为 $\mathcal{X}$, 概率密度函数 $p(x) = P(X=x)$

$p(x)$ 和 $p(y)$ 指两个不同的随机变量

分别表示不同的概率密度函数

引理 2.1.3: $X\sim\{P_\alpha\}_{\alpha=1}^{n}, \ H[X] = -\sum_{\alpha=1}^{n} P_\alpha \log_e P_\alpha \geqslant 0$

且 $H[P_1,P_2,\cdots,P_n]$ 为多元凹函数

证明:

$$\frac{\partial H}{\partial P_\alpha} = -(\log_e P_\alpha + 1)$$

$$\frac{\partial^2 H}{\partial P_\alpha \partial P_\beta} = -\frac{1}{P_\alpha}\delta_{\alpha\beta} \qquad \delta_{\alpha\beta}\begin{cases} =1, & \alpha=\beta \\ =0, & \alpha\neq\beta \end{cases}$$

$$\sum_{\alpha,\beta=1}^{n} \xi_\alpha \xi_\beta \frac{\partial^2 H}{\partial P_\alpha \partial P_\beta} = -\sum_{\alpha=1}^{n} \xi_\alpha^2 \cdot \frac{1}{P_\alpha} < 0$$

$\therefore$ 负定矩阵

$\therefore H$ 为多元凹函数.

2.2 联合熵与条件熵

将 (X, Y) 视为单个向量值随机变量

定义：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 $H(X,Y)$ 为

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

注：$H(X,Y) = -E_p \log p(X,Y)$

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

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

定理 2.2.1（链式法则）

☆ $H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$ ①

一对随机变量的熵等于其中一个随机变量的熵加上另一个随机变量的条件熵。

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

※：由式①得：$H(X) - H(X|Y) = H(Y) - H(Y|X)$

2.3. 互信息 $I(X;Y)$

互信息：mutual information

$\begin{cases}\text{一个随机变量包含另一个随机变量信息量的度量} \\ \text{在给定另一随机变量知识的条件下，原随机变量不确定度的缩减量}\end{cases}$

定义：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 $p(y)$，则

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

2.4 熵与互信息的关系：

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

$$\begin{cases}=H(X)-H(X|Y)\\=H(Y)-H(Y|X)\end{cases}$$

∴互信息 $I(X;Y)$ 是在给定 Y 知识下 X 的不确定度的缩减量

又∵ $H(X,Y)=H(X)+H(Y|X)$

☆ ∴ $I(X;Y)=H(X)+H(Y)-H(X,Y)$

定理 2.4.1

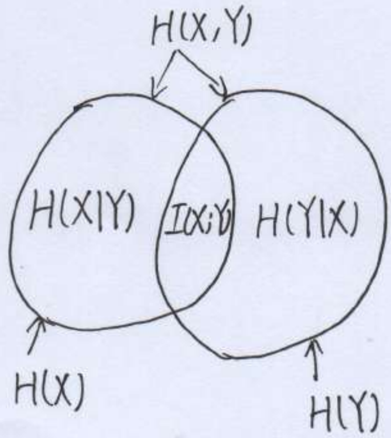
$I(X;Y)=H(X)-H(X|Y)$

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

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

$I(X;Y)=I(Y;X)$

$I(X;X)=H(X)$



互信息量 $I(X;Y)$ 的性质：

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

(i) $I(X;Y)\geqslant 0$

引理：若 $\{p_i\}_{i=1}^{N}$，$\{q_i\}_{i=1}^{N}$ 是概率分布，

则 $\sum_{i=1}^{n}p_i\log\frac{p_i}{q_i}\geqslant 0$

证明：$\sum_{i=1}^{n}p_i\log\frac{p_i}{q_i}$

$$=\sum_{i=1}^{n}q_i\cdot\underbrace{\frac{p_i}{q_i}\log\frac{p_i}{q_i}}_{f(x_i)} \quad ☆$$

$f(x)=x\log x$

$f'(x)=\log x+1 \quad f''(x)=\frac{1}{x}>0$

$\therefore f(x)$ 是凸函数

$\sum_{i=1}^{n}t_i=1$ $\because f$ 是凸函数.

$0\leq t_i\leq 1$ $\therefore \sum_{i=1}^{n}t_if(x_i)\geqslant f(\sum_{i=1}^{n}t_ix_i)$

$\therefore$ ☆式 $\geqslant f(\sum_{i=1}^{n}q_i\cdot\frac{p_i}{q_i})$

$=f(1)$

$=0$ $\therefore$ 引理得证.

$\therefore I(X;Y)\geqslant 0$

由引理证 $I(X;Y)\geqslant 0$

$\because \sum_x P_x=1 \quad \sum_y P_y=1$

$\therefore \sum_{x,y}P_x\cdot P_y=(\sum_x P_x)(\sum_y P_y)=1$

$\therefore$ 令 $P_i=p(x,y)$，$q_i=p(x)\cdot p(y)$

(ii) 互信息量的涵义：② 任何一个变化对另一个蕴含的大小.

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

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

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

$$=H(X)+H(Y)-H(X,Y)\geqslant 0$$

① $\underbrace{I(X;Y)}_{\text{表示不确定度的亏损}}=H[X]+H[Y]-\underbrace{H[X,Y]}_{\text{现实的不确定程度}}$

$\therefore H[X,Y]\leq H[X]+H[Y]$

若 $p(x,y)=p(x)\cdot p(y)$，则

$=H[X]+H[Y]$

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

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

2.5 熵与互信息的链式法则

一组随机变量的熵等于条件熵之和

定理 2.5.1（熵的链式法则）

设随机变量 $X_1, X_2, \cdots, X_n$，服从 $p(x_1, x_2, \cdots, x_n)$ 则

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

证明：法①

$$H(X_1, X_2) = H(X_1) + H(X_2|X_1)$$

$$H(X_1, X_2, X_3) = H(X_1) + H(X_2, X_3|X_1)$$
$$= H(X_1) + H(X_2|X_1) + H(X_3|X_1, X_2)$$
$$\vdots$$
$$H(X_1, X_2, \cdots, X_n) = \sum_{i=1}^{n} H(X_i | X_{i-1} \cdots X_1)$$

法② 由 $p(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} p(x_i | x_{i-1}, \cdots, x_1)$ 可得

$$H(X_1, X_2, \cdots, X_n)$$
$$= -\sum_{x_1, \cdots, x_n} p(x_1, x_2, \cdots, x_n) \log p(x_1, x_2, \cdots, x_n)$$
$$= -\sum_{x_1 \cdots x_n} p(x_1, x_2, \cdots, x_n) \log \prod_{i=1}^{n} p(x_i | x_{i-1}, \cdots, x_1)$$
$$= -\sum_{x_1, \cdots, x_n} \sum_{i=1}^{n} p(x_1, x_2, \cdots, x_n) \log p(x_i | x_{i-1}, \cdots, x_1)$$
$$= -\sum_{i=1}^{n} \sum_{x_1, \cdots, x_n} p(x_1, x_2, \cdots, x_n) \log p(x_i | x_{i-1}, \cdots, x_1)$$
$$= -\sum_{i=1}^{n} \sum_{x_1, \cdots, x_i} p(x_1, x_2, \cdots, x_i) \log p(x_i | x_{i-1}, \cdots, x_1)$$
$$= \sum_{i=1}^{n} H(X_i | X_{i-1}, \cdots, X_1)$$

条件互信息：在给定 Z 时由于 Y 的知识而引起关于 X 的不确定度的缩减量

定义：随机变量 X 和 Y 在给定随机变量 Z 时的条件互信息：

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

2.6 Jensen 不等式及其结果

定义 若对于任意的 $x_1, x_2 \in (a,b)$ 及 $0 \le \lambda \le 1$，满足

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

则称函数 $f(x)$ 在区间 (a,b) 上是凸的.(convex)

定义：如果 $-f$ 为凸函数，则称函数 f 是凹的。如果函数总是位于任何一条弦的下面，则函数是凸的；如果函数总是位于任何一条弦上面，则该函数是凹的.

Ex：

凸函数：x^2, $|x|$, e^x, $x\log x$

凹函数：$\log x$, $\sqrt{x}$

定理2.6.1 如果函数 f 在某个区间上存在非负(正)的二阶导数，则 f 为该区间的凸函数.

证明：$f(x) = f(x_0) + (x-x_0)\cdot f'(x_0) + \underbrace{\frac{f''(x^*)}{2}(x-x_0)^2}_{由假设 f''(x) \geq 0 可知，☆ \geq 0}$ ☆ 泰勒级数展开. x^* 位于 x_0 与 x 之间

设 $x_0 = \lambda x_1 + (1-\lambda)x_2$，取 $x = x_1$，可得：$f(x_1) \geq f(\lambda x_1 + (1-\lambda)x_2) + (1-\lambda)(x_1 + x_2) f'(\lambda x_1 + (1-\lambda)x_2)$

$$= f(x_0) + (1-\lambda)(x_1 - x_2) f'(x_0)$$

取 $x = x_2$，可得. $f(x_2) \geq f(x_0) + \lambda(x_2 - x_1) f'(x_0)$

$$\therefore \lambda f(x_1) + (1-\lambda) f(x_2) \geq \lambda f(x_0) + (1-\lambda) f(x_0)$$

$$= f(x_0)$$

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

$\therefore f$ 为凸函数.

E:数学期望.
离散：$EX = \sum_{x \in \mathcal{X}} x \cdot p(x)$
连续：$EX = \int x f(x) dx$

定理 2.6.2 (Jensen 不等式) 若给定凸函数 f 和一个随机变量 X，则

$$Ef(X) \geq f(EX)$$

对于两点分布，不等式变为：$p_1 f(x_1) + p_2 f(x_2) \geq f(p_1 x_1 + p_2 x_2)$，由凸函数定义可直接得到.

假定当分布点个数为 $k-1$ 时，定理成立，此时记 $p_i' = \frac{p_i}{1-p_k}$ $(i=1, 2, \cdots, k-1)$

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

$$\geq p_k f(x_k) + (1-p_k) f\left(\sum_{i=1}^{k-1} p_i' x_i\right)$$ 归纳假设

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

$$= f\left(\sum_{i=1}^{k} p_i x_i\right)$$

推论(互信息的非负性):

对任意两个随机变量X和Y, $I(X;Y)\geq 0$.

当且仅当X与Y相互独立,等号成立.

推论: $I(X;Y|Z)\geq 0$

当且仅当对给定随机变量Z,X和Y是条件独立的,等号成立

定理2.6.5 条件作用使熵减小(信息不会有负面影响)

$H(X|Y)\leq H(X)$. 当且仅当X与Y相互独立,等号成立.

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

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

定理2.6.6.(熵的独立界)

设$X_1,X_2,\cdots,X_n$服从$p(x_1,x_2,\cdots,x_n)$,则

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

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

2.7 对数和不等式及其应用.

定理2.7.3(熵的凹性)

$H(p)$是关于p的凹函数.

熵作为分布的函数时,具有凹性.

定理2.7.4

设$(X,Y)\sim p(x,y)=p(x)p(y|x)$。

如果固定$p(y|x)$,则互信息$I(X;Y)$是关于$p(x)$的凹函数

而如果固定$p(x)$,则互信息$I(X;Y)$是关于$p(y|x)$的凸函数.

$$\underset{\text{因}}{x} \xrightarrow{\quad P_{y|x} \quad} \underset{\text{果}}{y}$$

$P_{y|x}$：噪声转移概率.

$$P_y = \sum_{x\in\mathcal{X}} p(x,y) = \sum_{x\in\mathcal{X}} \underbrace{p(y|x)}_{\text{参量 → 信道}} \cdot \underbrace{p(x)}_{\text{变量 → 信源的随机变化}}$$

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

$$= \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(y|x)\cdot p(x) \log \frac{p(y|x)}{p(y)}$$

在通信领域，固定$p(y|x)$，则$I(X;Y)$是关于$p(x)$的凹函数

证明：$I(X;Y)$是关于$p(x)$的凹函数.

$$\frac{\partial I}{\partial p(x)} = \frac{\partial}{\partial p(x)} \cdot \sum_{x'\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(y|x')\cdot p(x') \log \frac{p(y|x')}{p(y)}$$

$$= \frac{\partial}{\partial p(x)} \cdot \sum_{x'\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(y|x')\cdot p(x')\left(\log p(y|x') - \log p(y)\right)$$

$$= \cancel{\sum_{x'\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(y|x') \log p(y|x')}$$

$$p(y) = \sum_{x\in\mathcal{X}} p(y|x)\cdot p(x)$$

$$\sum_{y\in\mathcal{Y}} p(y|x) \log p(y|x) - \frac{\partial}{\partial p(x)} \cdot \sum_{x'\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(y|x')\cdot p(x') \cdot \log P(y)$$

$$= \sum_{y\in\mathcal{Y}} p(y|x) \log p(y) + \sum_{x'\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(y|x')\cdot p(x') \frac{p(y|x)}{p(y)}$$

$$= \underbrace{\sum_{y\in\mathcal{Y}} p(y|x) \log p(y|x)}_{C_1\ (\text{参数})} - \sum_{y\in\mathcal{Y}} p(y|x) \log p(y) - \underbrace{\sum_{y\in\mathcal{Y}} p(y|x)}_{C_2\ (\text{参数})}$$

$$= C_1 - \sum_{y\in\mathcal{Y}} p(y|x) \log p(y) - C_2$$

$$\frac{\partial I^2}{\partial p(x)\,\partial p(x')} = -\frac{\partial}{\partial p(x')} \sum_{y\in\mathcal{Y}} p(y|x)\cdot \log p(y)$$

$$= -\sum_{y\in\mathcal{Y}} p(y|x) \frac{\partial}{\partial p(x')} \log p(y)$$

$$= -\sum_{y\in\mathcal{Y}} p(y|x) \cdot \frac{p(y|x')}{p(y)}$$

负定矩阵：

$$\sum_{x,x'} \xi_x \xi_{x'} \frac{\partial I^2}{\partial p(x')\,\partial p(x)} = -\sum_{y\in\mathcal{Y}} \frac{1}{p(y)} \left(\sum_{x} \xi_x P_{y|x}\right)^2 \le 0$$

2.8 数据处理不等式

马尔可夫过程定义:

随机序列中的每个随机变量仅依赖于它的前一个随机变量,而条件独立于其他前面的所有随机变量.

定义: 如果对 $n=1,2,\cdots$, 及所有的 $x_1, x_2, \cdots, x_n \in \mathcal{X}$, 有

$$\Pr\{\cancel{X_{n+}} X_{n+1} = x_{n+1} \mid X_n = x_n, X_{n-1} = x_{n-1}, \cdots, X_1 = x_1)$$
$$= \Pr(X_{n+1} = x_{n+1} \mid X_n = x_n)$$

则称离散随机过程 $X_1, X_2, \cdots$ 为马尔可夫链或马尔可夫过程.

Ex:

马尔可夫过程:

$$\cdots \to X_{n-1} \to X_n \to X_{n+1} \to X_{n+2} \to \cdots$$

$$P(X_{n+k_1}, \cdots, X_{n+k_t} \mid X_{n-j_1}, X_{n-j_2}, \cdots, X_{n-j_s}) \quad (j_1 < j_2 < \cdots < j_s)$$
$$= P(X_{n+k_1}, \cdots, X_{n+k_t} \mid X_{n-j_1})$$

性质: (1) 马尔可夫的任何子序列仍是马尔可夫序列
(2) 逆向序列仍是马尔可夫序列

$$P(X_{n-j_1}, X_{n-j_2}, \cdots X_{n-j_s} \mid X_{n+k_1}, X_{n+k_2}, \cdots, X_{n+k_t}) \quad (k_1 < k_2 < \cdots < k_t)$$
$$= P(X_{n-j_1}, X_{n-j_2}, \cdots, X_{n-j_s} \mid X_{n+k_1})$$

Ex: 若马尔可夫链 $X_1 \to X_2 \to X_3$, ~~即~~

$$P(X_3 \mid X_1, X_2) = P(X_3 \mid X_2),$$

那么 $X_1 \leftarrow\text{---} X_2 \leftarrow\text{---} X_3$; $P(X_1 \mid X_2, X_3) = P(X_1 \mid X_2)$

证明:

$$P(X_1 \mid X_2, X_3) = \frac{P(X_1, X_2, X_3)}{P(X_2, X_3)} = \frac{P(X_3 \mid X_2, X_1) \cdot P(X_2, X_1)}{P(X_2, X_3)}$$
$$= \frac{P(X_3 \mid X_2) \cdot P(X_1 \mid X_2) \cdot P(X_2)}{P(X_3 \mid X_2) \cdot P(X_2)}$$
$$= P(X_1 \mid X_2)$$

性质(3)　$X_1 \to X_2 \to X_3$ 是马尔可夫链，则 $I(X_1;X_2) \geq I(X_1;X_3)$

$$I(X_2;X_3) \geq I(X_1;X_3)$$

证明：

$$I(X_1;X_2) - I(X_1;X_3) = \sum_{x_1 \in \mathcal{X}} \sum_{x_2 \in \mathcal{X}} p(x_1,x_2) \log \frac{p(x_1,x_2)}{p(x_1)p(x_2)} - \sum_{x_1 \in \mathcal{X}} \sum_{x_3 \in \mathcal{X}} p(x_1,x_3) \log \frac{p(x_1,x_3)}{p(x_1)p(x_3)}$$

$$= \sum_{x_1} \sum_{x_2} \sum_{x_3} p(x_1,x_2,x_3) \log \frac{p(x_1,x_2) \cdot p(x_1) \cdot p(x_3)}{p(x_1)\, p(x_2)\, p(x_1,x_3)}$$

$$= \sum_{x_1} \sum_{x_2} \sum_{x_3} p(x_1,x_2,x_3) \log \frac{p(x_1|x_2)}{p(x_1|x_3)}$$

$$= \sum_{x_1} \sum_{x_2} \sum_{x_3} p(x_2,x_3) \cdot \underbrace{p(x_1|x_2,x_3)}_{p(x_1|x_2)} \cdot \log \frac{p(x_1|x_2)}{p(x_1|x_3)}$$

$$= \sum_{x_2} \sum_{x_3} p(x_2,x_3) \Big(\underbrace{\sum_{x_1} p(x_1|x_2) \log \frac{p(x_1|x_2)}{p(x_1|x_3)}}_{\sum_{i=1}^{n} p_i \log \frac{p_i}{q_i} \geq 0} \Big) \geq 0$$

$$\begin{cases} \sum_{i=1}^{n} p_i = 1 \\ \sum_{i=1}^{n} q_i = 1 \end{cases}$$

≥ 0

总结：

$$\cdots \to \underbrace{X_{j2} \to \underbrace{\overbrace{X_{j1} \to X_{k1}}^{\alpha} \to X_{k2}}_{I_{j1,k2}}}_{I_{j2,k2}} \to X_{k3} \to \cdots$$

β

$\because$ ~~$\alpha < I$~~　$\alpha \geq I_{j1,k2} \geq I_{j2,k2} \geq \beta$

$$\cdots \to \overbrace{X_{-1} \to X_1}^{\gamma'} \to \overbrace{X_2 \to X_3}^{\gamma''} \to$$

不是嵌套关系没法比.