1. 熵的性质

   离散熵的定义为

   定义 1. 随机变量 $X \sim p(x), p(x_i) = p_i, i = 1, \ldots, n, x_i \in \mathcal{X} = \{x_1, \ldots, x_n\}$,

   $$H(p_1, \ldots, p_n) \triangleq -\sum_{i=1}^{n} p_i \log p_i = \mathbb{E}[\log p(X)]$$

   其中 log 以 *2* 为底，熵的单位是比特。

   熵函数的性质:

   - 非负性
   - 是 n 的增函数
   - 可加性

     $$H(p_1, p_2, \ldots, p_n) = H(\sum_{i=1}^{k} p_i, p_{k+1}, \ldots, p_n) + \sum_{i=1}^{k} p_i H(p'_1, \ldots, p'_k)$$

     其中 $p'_i = p_i / \sum_{i=1}^{k} p_i$
   - 对称性，若 σ 为 $1, \ldots, n$ 上的一个置换，则:

     $$H(p_1, \ldots, p_n) = H(p_{\sigma(1)}, \ldots, p_{\sigma(n)})$$

   - 称 $h(p) \triangleq H(p, 1-p)$ 为二元熵函数，易证 $h(p)$ 是上凸函数。一般地，设 $\mathbf{P} = (p_1, \ldots, p_n)$, 所有长为 n 的概率向量 $\mathbf{P}$ 组成一个凸域 D。利用 $-p \log p$ 函数的上凸性可以证明 $\forall \mathbf{P}, \mathbf{P}' \in D$ 有

     $$H(\lambda \mathbf{P} + (1-\lambda)\mathbf{P}) \geq \lambda H(\mathbf{P}) + (1-\lambda) H(\mathbf{P}')$$

     即 $H(\mathbf{P})$ 是 $\mathbf{P}$ 的上凸函数。

2. 联合熵和条件熵

   定义 2. 一对离散型随机变量 (X, Y)，联合分布为 $p(x, y)$, 它们的联合熵为

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

定义 3. Y 对于 X 的条件熵定义为

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

关系式:

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

  证明.

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

$\square$

- 记 $H(X_1|X_0) = H(X_1)$, 则有

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

3. 相对熵

定义 4. 设 $p(x), q(x)$ 是 $\mathcal{X}$ 中字母表相同的两个概率分布，则它们的相对熵定义为:

$$D(p||q) \triangleq \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)} = \mathbb{E}_X[\log\frac{p(X)}{q(X)}]$$

定理 1. $D(p||q) \geq 0$，等号成立当且仅当 $p = q$

证明. 由 $-\log p$ 函数的下凸性质，

$$
\begin{aligned}
D(p||q) =& \mathbb{E}_X[-\log \frac{q(X)}{p(X)}], \text{by Jensen's Inequality} \\
\geq & -\log \left(\mathbb{E}_X[\frac{q(X)}{p(X)}] \right) \\
=& 0
\end{aligned}
$$

由 Jensen 不等式的取等条件，$\frac{p}{q}$ 应为常数 □

利用相对熵的非负性，我们可以证明

推论 1. *设 X 是在字母表 $\mathcal{X}$ 上取值的随机变量，则 $H(X) \leq \log |\mathcal{X}|$，等号成立当且仅当 X 是均匀分布。*

证明. 设 u 是 $\mathcal{X}$ 上的均匀分布，则有 $D(X|u) \geq 0 \Rightarrow H(X) \leq \log |\mathcal{X}|$ □

相对熵 $D(p||q)$ 的下凸性可以总结为以下三点:

- q 固定，由 $t \log t$ 的下凸性可以得到

$$D(\lambda p + (1-\lambda)p'||q) \leq \lambda D(p||q) + (1-\lambda) D(p'||q)$$

- p 固定，由 $-\log t$ 的下凸性可以得到

$$D(p||\lambda q + (1-\lambda)q') \leq \lambda D(p||q) + (1-\lambda) D(p||q')$$

- 二元凸性

$$D(\lambda p + (1-\lambda)p'||\lambda q + (1-\lambda)q') \leq \lambda D(p||q) + (1-\lambda) D(p'||q')$$

证明. 由对数和不等式

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq (\sum_{i=1}^{n} a_i) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i} \tag{1}$$

$$
\begin{aligned}
D(\lambda p + (1-\lambda)p'||\lambda q + (1-\lambda)q') =& \sum_{x \in \mathcal{X}} (\lambda p(x) + (1-\lambda)p'(x)) \log \frac{\lambda p(x) + (1-\lambda)p'(x)}{\lambda q(x) + (1-\lambda)q'(x)} \\
\leq & \sum_{x \in \mathcal{X}} [\lambda p(x) \log \frac{p(x)}{q(x)} + (1-\lambda)p'(x) \log \frac{p'(x)}{q'(x)}] \\
=& \lambda D(p||q) + (1-\lambda) D(p'||q')
\end{aligned}
$$

□

4. 互信息

定义 5. 设 $(X,Y)\sim p(x,y), x\in\mathcal{X}, y\in\mathcal{Y}$, X,Y 的互信息定义为

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

互信息量有如下的性质:

- $I(X;Y)=D(p(x,y)||p(x)p(y))\Rightarrow I(X,Y)\geq 0$ 且 $I(X;Y)=0\iff p(x,y)=p(x)p(y)$
- $I(X;Y)=H(X)-H(X|Y)=H(Y)-H(Y|X)=H(X)+H(Y)-H(X,Y)$
- $I(X;Y|Z)=0\iff p(x,y|z)=p(x|z)p(y|z)$($X,Y$ 关于 Z 条件独立) $\iff X\to Z\to Y$ 构成马氏链
- 互信息的链式法则

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

  证明.

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

  □

互信息 $I(X;Y)$ 可以看成是 $p(x),p(y|x)$ 的泛函数:

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

关于互信息的凸性可以总结为以下两点:

- $I(X;Y)$ 关于 $p(x)$ 上凸
- $I(X;Y)$ 关于 $p(y|x)$ 下凸。

数据处理不等式:设 $P_Y = P_{Y|X}\circ P_X, Q_Y = P_{Y|X}\circ Q_X$ 则 $D(P_X||Q_X) \geq D(P_Y||Q_Y)$。

证明. $\mathbb{E}_{XY}[\log \frac{P_{XY}}{Q_{XY}}] = \mathbb{E}_{XY}[\log \frac{P_{Y|X}}{Q_{Y|X}} + \log \frac{P_X}{Q_X}]$ 因为 $Q_{Y|X} = P_{Y|X} \Rightarrow \mathbb{E}_{XY}[\log \frac{P_{XY}}{Q_{XY}}] = D(P_X||Q_X)$ 又

$$\begin{aligned}\mathbb{E}_{XY}[\log \frac{P_{XY}}{Q_{XY}}] =& \mathbb{E}_{XY}[\log \frac{P_{X|Y}}{Q_{X|Y}} + \log \frac{P_Y}{Q_Y}] \\ &= \mathbb{E}_Y[D(P_{X|Y}||Q_{X|Y})] + D(P_Y||Q_Y) \\ &\geq D(P_Y||Q_Y)\end{aligned}$$

所以有 $D(P_X||Q_X) \geq D(P_Y||Q_Y)$。 □

5. 微分熵设 X 是连续型随机变量，X 的微分熵定义为:

$$h(X) \triangleq -\int_{\mathbb{R}} p(x)\log p(x)dx \tag{2}$$

$X^{(\Delta)}$ 是 X 按区间长度为 Δ 离散的结果。则 $H(X^{(\Delta)}) + \log\Delta \to h(X)$

微分熵可正可负

常见分布的微分熵

命题 1. *(a) 区间 $(0,a)$ 上的均匀分布的微分熵为 $\log a$，可正可负。*

(b) 方差为 σ^2 的高斯分布微分熵的大小为 $\frac{1}{2}\log(2\pi e\sigma^2)$

(c) 均值为 $\frac{1}{\lambda}$ 的指数分布的微分熵为 $\log e - \log\lambda$

证明. (a) 不妨设高斯分布 X 的均值为 0，概率密度函数为 $p(x)$ 则

$$\begin{aligned}h(X) &= -\int_{\mathbb{R}} p(x)\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{x^2}{2\sigma^2})\right)dx \\ &= \frac{1}{2}\log(2\pi\sigma^2) + \frac{\log e}{2\sigma^2}\int_{\mathbb{R}} x^2 p(x)dx \\ &= \frac{1}{2}\log(2\pi e\sigma^2)\end{aligned}$$

(b)

$$h(X) = -\int_0^\infty \lambda e^{-\lambda x}(\log\lambda - \lambda x\log e)dx = \log e - \log\lambda$$

□

6. 相对熵和互信息（连续情形）

定义 6. 若 $X \sim p(x), Y \sim q(y)$ 连续，则 X 和 Y 的相对熵为

$$D(p||q) \triangleq \int_{\mathbb{R}} p(x) \log \frac{p(x)}{q(x)} dx \tag{3}$$

X 和 Y 的互信息为：$I(X;Y) = D(p(x,y)||p(x)q(y))$

常见的连续型概率分布可以看成是某种条件下的最大熵分布：

定理 2. 对概率密度族 $\mathcal{P}$，若存在 $p_0(x) \in \mathcal{P}$，使得

$$\forall p(x) \in \mathcal{P}, -\int_{\mathbb{R}} p(x) \log p_0(x) dx = h_0$$

是一个与 $p(x)$ 无关的常数，则 $p_0(x)$ 为最大熵分布，h_0 为最大熵。

证明. 由相对熵的非负性得：

$$-\int_{\mathbb{R}} p(x) \log p(x) dx \le -\int_{\mathbb{R}} p(x) \log p_0(x) dx = h_0$$

等号成立当且仅当 $p(x)$ 与 $p_0(x)$ 几乎处处相等。 □