第 13 讲 微分熵

定义

连续随机变量的 AEP

微分熵与离散熵的关系

联合微分熵与条件微分熵

相对熵与互信息

微分熵的性质，相对熵和互信息的性质

定义 1.1

设 X 是一个随机变量，其累积分布函数为 $F(x) = P(X \leq x)$. 如果 $F(x)$ 是连续的，则称该随机变量是连续的. 当 $F(x)$ 的导数存在时，令 $f(x) = F'(x)$. 若 $\int_{-\infty}^{\infty} f(x)dx = 1$，则称 $f(x)$ 是 X 的**概率密度函数**. 另外，使 $f(x) > 0$ 的所有 x 构成的集合称为 X 的支撑集.

定义 1.2

一个以 $f(x)$ 为密度函数的连续型随机变量 X 的**微分熵** $h(X)$ 定义为

$$h(X) = -\int_S f(x) \log f(x) dx.$$

其中 S 是这个随机变量的支撑集.

例 1.3: 均匀分布

考虑一个服从 $[0, a]$ 上的均匀分布的随机变量，它的密度函数在 $[0, a]$ 熵为 $1/a$，而在其他地方为 0. 此时，该随机变量的微分熵为

$$h(X) = -\int_0^a \frac{1}{a} \log \frac{1}{a} dx = \log a.$$

例 1.4: 正态分布

设 $X \sim \phi(x) = (1/\sqrt{2\pi\sigma^2})e^{-x^2/2\sigma^2}$. 如果我们以奈特为单位计算微分熵，我们有

$$\begin{aligned} h(\phi) &= -\int \phi \ln \phi \\ &= -\int \phi(x)[-\frac{x^2}{2\sigma^2} - \ln \sqrt{2\pi\sigma^2}] \\ &= \frac{EX^2}{2\sigma^2} + \frac{1}{2}\ln 2\pi\sigma^2 \\ &= \frac{1}{2} + \frac{1}{2}\ln 2\pi\sigma^2 \\ &= \frac{1}{2}\ln 2\pi e\sigma^2 \text{奈特}. \end{aligned}$$

改变对数的底，我们可得

$$h(\phi) = \frac{1}{2}\log 2\pi e\sigma^2 \text{比特}.$$

定义
连续随机变量的 AEP
微分熵与离散熵的关系
联合微分熵与条件微分熵
相对熵与互信息
微分熵的性质，相对熵和互信息的性质

定理 2.1

设 $X_1, X_2, \cdots, X_n$ 使一个服从于密度函数 $f(x)$ 的独立同分布的随机变量序列. 那么下面的极限依概率收敛.

$$-\frac{1}{n}\log f(X_1, X_2, \cdots, X_n) \to E[-\log f(X)] = h(X) \text{ 依概率.}$$

定义 2.2

对 $\epsilon > 0$ 及任意的 n，定义 $f(x)$ 的典型集 $A_\epsilon^{(n)}$ 如下：

$$A_\epsilon^{(n)} = \{(x_1, x_2, \cdots, x_n) \in S^n : |-\frac{1}{n}\log f(X_1, x_2, \cdots, x_n) - h(X)| \leq \epsilon\}.$$

其中 $f(x_1, x_2, \cdots, x_n) = \prod_{i=1}^n f(x_i)$.

定义 2.3

集合 $A \subset \mathbb{R}^n$ 的体积Vol(A) 定义为

$$\text{Vol}(A) = \int_A dx_1 dx_2 \cdots dx_n.$$

定理 2.4

典型集 $A_\epsilon^{(n)}$ 有如下的性质:

1. 对于充分大的 n，$P(A_\epsilon^{(n)}) > 1 - \epsilon$.
2. 对于所有的 n，$\mathrm{Vol}(A_\epsilon^{(n)}) \leq 2^{n(h(X)+\epsilon)}$.
3. 对于充分大的 n，$\mathrm{Vol}(A_\epsilon^{(n)}) \geq (1-\epsilon)2^{n(h(X)-\epsilon)}$.

证明.

根据 AEP，我们有依概率有 $-\frac{1}{n}\log f(X^n) = -\frac{1}{n}\sum \log f(x_i) \to h(X)$，故性质 1 得证. 另外，

$$
\begin{aligned}
1 &= \int_{S^n} f(x_1, x_2, \cdots, x_n) dx_1 dx_2 \cdots dx_n \\
&\geq \int_{A_\epsilon^{(n)}} f(x_1, x_2, \cdots, x_n) dx_1 dx_2 \\
&\geq \int_{A_\epsilon^{(n)}} 2^{-n(h(X)+\epsilon)} dx_1 dx_2 \cdots dx_n \\
&= 2^{-n(h(X)+\epsilon)} \int_{A_\epsilon^{(n)}} dx_1 dx_2 \cdots dx_n \\
&= 2^{-n(h(X)+\epsilon)} \mathrm{Vol}(A_\epsilon^{(n)}).
\end{aligned}
$$

因此，性质 2 获证.

证明.
我们进一步论证该典型集的体积至少有这么大. 如果 n 足够大使得性质 1 成立，那么

$$
\begin{aligned}
1-\epsilon &\leq \int_{A_\epsilon^{(n)}} f(x_1, x_2, \cdots, x_n) dx_1 dx_2 \cdots dx_n \\
&\leq \int_{A_\epsilon^{(n)}} 2^{-n(h(X)-\epsilon)} dx_1 dx_2 \cdots dx_n \\
&= 2^{-n(h(X)-\epsilon)} \int_{A_\epsilon^{(n)}} dx_1 dx_2 \cdots dx_n \\
&= 2^{-n(h(X)-\epsilon)} \mathrm{Vol}(A_\epsilon^{(n)}).
\end{aligned}
$$

从而性质 3 成立. 从而当 n 充分大时，我们有

$$
(1-\epsilon) 2^{n(h(X)-\epsilon)} \leq \mathrm{Vol}(A_\epsilon^{(n)}) \leq 2^{n(h(X)+\epsilon)}. \qquad \square
$$

定理 2.5

在一阶指数意义下，在所有概率 $\geq 1-\epsilon$ 的集合中，$A_\epsilon^{(n)}$ 为使体积最小者.

考虑一个密度函数是 $f(x)$ 的随机变量 X. 假定将 X 的定义域等长度地分割成长为 Δ 的若干小区间，并且假定密度函数在这些小区间内是连续的. 由中值定理可知，在每个小区间内存在一个值 x_i，满足

$$f(x_i)\Delta = \int_{i\Delta}^{(i+1)\Delta} f(x)dx.$$

考虑量化后的随机变量 X^{Δ}，其定义是

$$X^{\Delta} = x_i \text{ 当 } i\Delta \le X < (i+1)\Delta.$$

则 $X^{\Delta} = x_i$ 的概率为

$$p_i = \int_{i\Delta}^{(i+1)\Delta} f(x)dx = f(x_i)\Delta.$$

由于 $\sum f(x_i)\Delta = \int f(x) = 1$，所以量化后的随机变量 X^Δ 的熵为

$$\begin{aligned} H(x^\Delta) &= -\sum_{i=-\infty}^{\infty} p_i \log p_i \\ &= -\sum_{i=-\infty}^{\infty} f(x_i)\Delta \log(f(x_i)\Delta) \\ &= -\sum \Delta f(x_i) \log f(x_i) - \sum f(x_i)\Delta \log \Delta \\ &= -\sum \Delta f(x_i) \log f(x_i) - \log \Delta \end{aligned}$$

如果 $f(x)\log x$ 是黎曼可积的，则我们可以得到下面的定理.

定理 3.1

如果随机变量 X 的密度函数 $f(x)$ 是黎曼可积的，那么

$$H(X^{\Delta}) + \log \Delta \to h(f) = h(X), \text{ 当} \Delta \to 0.$$

于是，连续随机变量 X 经过 n 比特量化处理 (此时分割的小区间长度为 $\frac{1}{2^n}$) 后的熵大约为 $h(X) + n$.

例 3.2

如果 X 服从 $[0, \frac{1}{8}]$ 上的均匀分布，那么在二进制表示中，X 取值的小数点右边取值必定为 0. 因而，在精确到 n 位的意义下，描述 X 仅需 $n-3$ 比特，这与 $h(X) = -3$ 相一致.

一般来讲，在精确到 n 位的意义下，$h(X) + n$ 是为了描述 X 所需要的平均比特数.

定义 4.1

联合密度函数为 $f(x_1, x_2, \cdots, x_n)$ 的一组随机变量 $X_1, X_2, \cdots, X_n$ 的联合微分熵定义为

$$h(X_1, X_2, \cdots, X_n) = -\int f(x^n) \log f(x^n) dx^n.$$

定义 4.2

如果 X, Y 的联合密度函数为 $f(x,y)$，定义条件微分熵 $h(X|Y)$ 为

$$h(X|Y) = -\int f(x,y) \log f(x|y) dx dy.$$

由于一般地 $f(x|y) = f(x,y)/f(y)$，我们可以写为

$$h(X|Y) = h(X,Y) - h(Y).$$

我们要注意微分熵可能为无穷的情况，需要分开讨论.

定理 4.3: 多元正态分布的熵

设 $X_1, X_2, \cdots, X_n$ 满足多元正态分布，平均值为 μ，协方差矩阵为 K. 则

$$h(X_1, X_2, \cdots, X_n) = h(\mathcal{N}_n(\mu, K)) = \frac{1}{2}\log(2\pi e)^n|K| \text{ 比特},$$

其中 $|K|$ 记为 K 的行列式.

证明.

$X_1, X_2, \cdots, X_n$ 的概率密度函数为 $f(\mathbf{x}) = \frac{1}{(\sqrt{2\pi})^n |K|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T K^{-1}(\mathbf{x}-\mu)}$. 则

$$
\begin{aligned}
h(f) &= -\int f(x)[-\frac{1}{2}(\mathbf{x}-\mu)^T K^{-1}(\mathbf{x}-\mu) - \ln(\sqrt{2\pi})^n |K|^{\frac{1}{2}}]d\mathbf{x}. \\
&= = \frac{1}{2}E[\sum_{i,j}(X_i-\mu_i)(K^{-1})_{ij}(X_j-\mu_j)] + \frac{1}{2}\ln(2\pi)^n|K| \\
&= = \frac{1}{2}E[\sum_{i,j}(X_i-\mu_i)(X_j-\mu_j)(K^{-1})_{ij}] + \frac{1}{2}\ln(2\pi)^n|K| \\
&= = \frac{1}{2}\sum_{i,j}E[(X_i-\mu_i)(X_j-\mu_j)](K^{-1})_{ij} + \frac{1}{2}\ln(2\pi)^n|K| \\
&= = \frac{1}{2}\sum_{i,j}K_{ji}(K^{-1})_{ij} + \frac{1}{2}\ln(2\pi)^n|K| \\
&= = \frac{1}{2}\sum_{j}I_{jj} + \frac{1}{2}\ln(2\pi)^n|K| \\
&= \frac{n}{2} + \frac{1}{2}\ln(2\pi)^n|K| \\
&= \frac{1}{2}\ln(2\pi e)^n|K|
\end{aligned}
$$

奈特

定义
连续随机变量的 AEP
微分熵与离散熵的关系
联合微分熵与条件微分熵
相对熵与互信息
微分熵的性质，相对熵和互信息的性质

定义 5.1

两个概率密度函数 f 和 g 的相对熵 (或 Kullback-Leibler 距离)$D(f\|g)$ 定义为

$$D(f\|g) = \int f \log \frac{f}{g}.$$

注意到 $D(f\|g)$ 有限当且仅当 f 的支集包含在 g 的支集中. (和以前一样我们默认 $0 \log \frac{0}{0} = 0$.)

定义 5.2

两个联合分布为 $f(x,y)$ 的随机变量 X 和 Y 的互信息定义为

$$I(X;Y) = \int f(x,y)\log\frac{f(x,y)}{f(x)f(y)}dxdy.$$

由定义我们知和离散的情形类似，

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

以及

$$I(X;Y) = D(f(x,y)\|f(x)f(y)).$$

设 $\mathcal{X}$ 位随机变量 X 的值域，$\mathcal{P}$ 为 $\mathcal{X}$ 的一个分割是指存在有限个不相交的集合 P_i 使得 $\cup_i P_i = \mathcal{X}$. X 关于 $\mathcal{P}$ 的量化 (记为 $[X]_{\mathcal{P}}$) 是定义如下的离散随机变量：

$$P([X]_{\mathcal{P}} = i) = P(X \in P_i) = \int_{P_i} dF(x).$$

对于任意两个分割分别为 $\mathcal{P}$ 和 $\mathcal{Q}$ 的随机变量 X 和 Y，可以利用 (??) 来计算它们对应的量化随机变量的互信息. 于是，对任意成对的随机变量，我们可以如下定义其互信息：

定义 5.3

任何随机变量 X 与 Y 间的互信息如下：

$$I(X;Y) = \sup_{\mathcal{P},\mathcal{Q}} I([X]_{\mathcal{P}};[Y]_{\mathcal{Q}}),$$

其中上确界遍历所有可能的有限剖分 $\mathcal{P}$ 和 $\mathcal{Q}$.

定理 6.1

我们有

$$D(f\|g) \geq 0.$$

等号成立当且仅当 $f = g,\ a.e..$

证明.
设 f 的支撑集为 S. 则

$$\begin{aligned}
-D(f\|g) &= \int_S f\log\frac{g}{f} \\
&\le \log\int_S f\frac{g}{f} \\
&= \log\int_S g \\
&\le \log 1 = 0
\end{aligned}$$

当且仅当 Jensen 不等式中的等号成立，即当且仅当 $f=g$ a.e. 等号成立. $\square$

推论 6.2

$I(X;Y) \geq 0$，等号成立当且仅当 X 和 Y 独立.

定理 6.3: 微分熵的链式法则

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

推论 6.4

$$h(X_1, X_2, \cdots, X_n) \leq \sum h(X_i),$$

等号成立当且仅当 $X_1, X_2, \ldots, X_n$ 为独立的.

推论 6.5: 阿达玛不等式

设 $\mathbf{x} \sim \mathcal{N}(0, K)$ 是一个多元正态分布，那么将熵的定义公式代入上面的不等式中，我们可以得到

$$|K| \leq \Pi_{i=1}^{n} K_{ii}.$$

定理 6.6

$$h(X+c)=h(X).$$

定理 6.7

$$h(aX) = h(X) + \log |a|.$$

证明.

令 $Y - aX$. 则 $f_Y(y) = \frac{1}{|a|} f_X(\frac{y}{a})$，且经过积分变量替换，有

$$
\begin{aligned}
h(aX) &= -\int f_Y(y) \log f_Y(y) dy \\
&= -\int \frac{1}{|a|} f_X(\frac{y}{a}) \log(\frac{1}{|a|} f_X(\frac{y}{a})) dy \\
&= -\int f_X(x) \log f_X(x) dx + \log|a| \\
&= h(X) + \log|a|
\end{aligned}
$$

从而得证. $\square$

类似地，对于取值为向量的随机变量，可以证明下面的推论.

推论 6.8

$$h(A\mathbf{X}) = h(\mathbf{X}) + \log|\det(A)|.$$

定理 6.9

设随机向量 $\mathbf{X} \in \mathbb{R}^n$ 的均值为零，协方差矩阵为 $K = E\mathbf{X}\mathbf{X}^t$，则 $h(\mathbf{X}) \leq \frac{1}{2}\log(2\pi e)^n|K|$，当且仅当 $\mathbf{X} \sim \mathcal{N}(0, K)$ 等号成立.

证明.

设 $g(\mathbf{x})$ 是对任意的 i 和 j 均满足 $\int g(\mathbf{x})x_i x_j d\mathbf{x} = K_{ij}$ 的密度函数. 令 ϕ_K 是服从高斯分布 $\mathcal{N}(0, K)$ 随机向量的密度函数. 注意到 $\log\phi_K(\mathbf{x})$ 是一次二次型，并且 $\int x_i x_j \phi_K(\mathbf{x})d\mathbf{x} = K_{ij}$，则

$$
\begin{aligned}
0 &\le D(g\|\phi_K) \\
&= \int g\log(g/\phi_K) \\
&= -h(g) - \int g\log\phi_K \\
&= -h(g) - \int \phi_K\log\phi_K \\
&= -h(g) + h(\phi_K)
\end{aligned}
$$

其中所作的替换 $\int g\log\phi_K = \int \phi_K\log\phi_K$ 是由于二次型 $\log\phi_K(\mathbf{x})$ 关于 g 和 ϕ_K 具有相同的矩. □

定理 6.10: 估计误差与微分熵

对任意随机变量 X 及其估计 $\hat{X}$，

$$E(X-\hat{X})^2 \geq \frac{1}{2\pi e} e^{2h(X)}.$$

其中等号成立的充分必要条件是 X 为高斯分布而 $\hat{X}$ 为其均值. 这里 $h(X)$ 的量纲为比特.

证明.
令 $\hat{X}$ 为 X 的一个估计，此时

$$
\begin{aligned}
E(X-\hat{X})^2 &\geq \min_{\hat{X}} E(X-\hat{X})^2 \\
&= E(X-E(X))^2 \\
&= \mathrm{var}(X) \\
&\geq \frac{1}{2\pi e} e^{2h(X)}
\end{aligned}
$$

不等式成立当且仅当 X 为高斯分布而 $\hat{X}$ 为其均值. $\square$

推论 6.11

当边信息 Y 以及估计 $\hat{X}(Y)$ 已知时，可以推出

$$E(X-\hat{X}(Y))^2 \geq \frac{1}{2\pi e} e^{2h(X|Y)}.$$