第一章随机变量及其信息度量

第七节连续型随机变量的熵

陈兴同

中国矿业大学 数学学院

2021 年 8 月

内容提要

1. 微分熵

内容提要

内容提要

连续随机变量量化

本小节将介绍连续信源、连续信道信息处理过程中涉及的信息度量及其性质。
设 X 是一个取值在有限区间 $\mathcal{X}=(a,b)$ 上的连续型随机变量，并且概率密度函数 $f(x), x\in(a,b)$ 连续。为了确定它的不确定性，采用离散型随机变量来逼近它。
将区间剖分成 n 等份：$a=x_0<x_1<x_2<\cdots<x_n=b$，步长 $h=(b-a)/n$，在每个小区间上的概率记成

$$p_i=P\left\{x_i<X<x_{i+1}\right\}=\int_{x_i}^{x_{i+1}} f(x)dx, i=0,1,2,\cdots,n-1,$$

由积分中值定理可知，必存在一个 $\xi_i\in(x_i,x_{i+1})$ 使得

$$p_i=\int_{x_i}^{x_{i+1}} f(x)dx=f(\xi_i)h, i=0,1,2,\cdots,n-1.$$

续：连续随机变量量化

记 $\mathcal{X}_n = \{\xi_0, \xi_1, \xi_2, \cdots, \xi_{n-1}\}$，构造一个离散随机变量 X_n 使它的分布律为

$$X_n \sim p^{(n)}(x) = \begin{pmatrix} \xi_0 & \xi_1 & \xi_2 & \cdots & \xi_{n-1} \\ p_0 & p_1 & p_2 & \cdots & p_{n-1} \end{pmatrix},$$

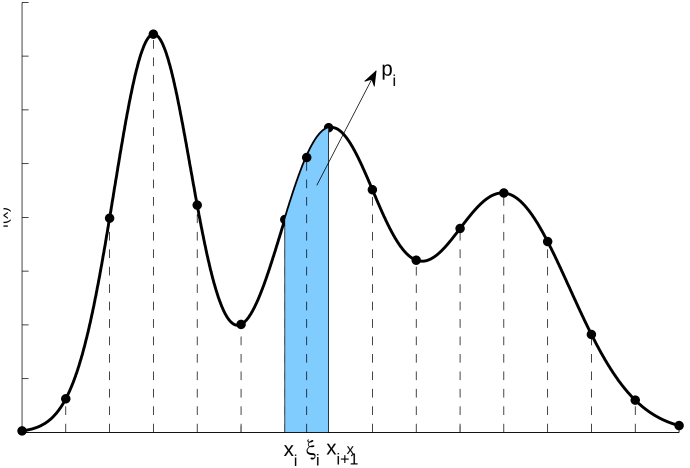
~~(1.1)~~ 1.41

则离散型随机变量 X_n 可以作为连续型随机变量 X 的一个近似。它有离散熵

$$H(X_n) = -\sum_{i=0}^{n-1} p_i \log p_i = \sum_{i=0}^{n-1} f(\xi_i) h \log [f(\xi_i) h],$$

续：连续随机变量量化

如图1.7



续：连续随机变量量化

继续计算这个离散熵得

$$
\begin{aligned}
H(X_n) &= -\sum_{i=0}^{n-1} f(\xi_i)h\log f(\xi_i)) - \sum_{i=0}^{n-1} f(\xi_i)h\log h \\
&= -\sum_{i=0}^{n-1} [f(\xi_i)\log f(\xi_i)]h - \log h \sum_{i=0}^{n-1} f(\xi_i)h \\
&= -\sum_{i=0}^{n-1} [f(\xi_i)\log f(\xi_i)]h - \log h \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} f(x)dx \\
&= -\sum_{i=0}^{n-1} [f(\xi_i)\log f(\xi_i)]h - \log h \int_{a}^{b} f(x)dx \\
&= -\sum_{i=0}^{n-1} [f(\xi_i)\log f(\xi_i)]h - \log h, \qquad \overset{1.42}{\cancel{(1.2)}}
\end{aligned}
$$

续：连续随机变量量化

令 $h \to 0$ 或 $n \to \infty$，对上式求极限得

$$\lim_{n\to\infty} H(X_n) = -\int_a^b f(x)\log f(x)dx + \infty.$$

这说明连续型随机变量的不确定性为无穷大！但是从上式第一部分仍然得到一个确定的积分值，这个积分作为微分熵的定义。

定义 1.7.1：微分熵定义

*设随机变量 X 的概率密度函数 $f(x)$ 定义在区间 $\mathcal{X}$ 上，如果下面积分绝对可积，则称它为随机变量 X 的**微分熵**，记作 $H(X)$，以后也称为熵。*

$$H(X) = -\int_{\mathcal{X}} f(x) \log f(x) dx.$$

根据公式（~~1.2~~ 1.42），若对连续型随机变量 X 进行 $n$bits 量化（$h = (b-a)/2^n$），所得离散随机变量 X_n 熵是 $H(X) + n$bits，它表示连续型随机变量平均信息量近似为 $H(X) + n$bits。

定义 1.7.2：联合微分熵定义

类似地对二维连续随机变量也有微分熵。

设二维随机变量 X，Y 的联合概率密度函数 $f(x,y)$ 定义在区域 $\mathcal{G}$ 上，如果下面二重积分绝对可积，则称它为随机变量 X,Y 的 **联合微分熵**，*记作 $H(X,Y)$，以后也称为联合熵。*

$$H(X,Y) = -\int\int_{\mathcal{G}} f(x,y)\log f(x,y)dxdy.$$

条件概率密度

类似地，对二维连续随机变量也有条件分布，从而也有条件微分熵。设二维随机变量 X, Y 的联合概率密度函数为 $f(x, y)$，则 X, Y 的概率密度函数为

$$f_X(x) = \int_{\mathcal{R}} f(x, y)dy, f_Y(y) = \int_{\mathcal{R}} f(x, y)dx,$$

以及条件概率密度函数为

$$f_{X|Y}(x|y) = \frac{f(x, y)}{f_Y(y)}, f_{Y|X}(y|x) = \frac{f(x, y)}{f_X(x)}.$$

定义 1.7.3：条件微分熵定义

设二维随机变量 X, Y 的联合概率密度函数 $f(x, y), (x, y) \in \mathcal{G}$，如果下面二重积分绝对可积，则称它为随机变量 Y 已知条件下随机变量 X 的**条件微分熵**，记作 $H(X|Y)$，以后也称为条件熵。

$$H(X|Y) = -\int\int_{\mathcal{G}} f_Y(y) f_{X|Y}(x|y) \log f_{X|Y}(x|y) dx dy.$$

微分熵的差有意义

连续随机变量的微分熵其实不具备任何信息含义，并不表示连续型随机变量 X 的信息量，但是两个微分熵的差却具有明确的信息含义！事实上，对两个定义在同一区间中随机变量 X, Y，分别具有概率密度函数 $f_X(x), f_Y(y), x, y \in \mathcal{X}$ ，利用（1.1）中的离散化方法进行相同步长的离散化，得到相应的离散随机变量 X_n, Y_n 及其概率分布 $p^{(n)}(x), q^{(n)}(y), x, y \in \mathcal{X}_n$，则可以证明

$$\lim_{n\to\infty} [H(X_n) - H(Y_n)] = H(X) - H(Y).$$

因为离散熵 $H(X_n)H(Y_n)$ 有确定的信息含义，故这个结论说明两个微分熵的差可以表示连续型随机变量 X 与 Y 包含信息量的差异。

相对熵的极限

在一定条件下还可以证明

1.43

$$\lim_{n\to\infty} D(p^{(n)}||q^{(n)}) = \int_{\mathcal{X}} f_X(x) \log \frac{f_X(x)}{f_Y(x)} dx, \qquad (\cancel{2.1})$$

这说明离散化随机变量 $X^{(n)}, Y^{(n)}$ 的相对熵 $D(p^{(n)}||q^{(n)})$ 极限存在。如果（2.1）中积分存在，则它正好是离散化随机变量 $X^{(n)}, Y^{(n)}$ 的相对熵 $D(p^{(n)}||q^{(n)})$ 的极限，由此可以定义连续型随机变量的相对熵。

定义 1.7.4: 相对微分熵定义

*设连续型随机变量 X, Y 均定义在同一个区间 $\mathcal{X}$ 中，其概率密度函数分别为 $f_X(x), f_Y(y), x, y \in \mathcal{X}$。如果下面积分绝对可积，则称它为随机变量 X 对 Y 或概率密度 f_X 对 f_Y 的**相对熵**，记作 $D(f_X||f_Y)$。*

$$D(f_X||f_Y) = \int_{\mathcal{X}} f_X(x) \log \frac{f_X(x)}{f_Y(x)} dx.$$

连续型的相对熵可以作为两个连续型随机变量差异的信息度量，有确定的信息含义。

定义 1.7.5: 互信息定义

类似于离散随机变量互信息，也可以定义连续型随机变量的互信息，它其实就是一种相对熵。

*设二维随机变量 X，Y 的联合概率密度为 $f(x,y),(x,y)\in\mathcal{G}$，如果下面积分绝对可积，则称它为随机变量 X,Y 的**互信息**，记作 $I(X;Y)$。*

$$I(X;Y)=\int\int_{\mathcal{G}} f(x,y)\log\frac{f(x,y)}{f_X(x)f_Y(y)}.$$

定理 1.7.1：微分熵链式法则

类似离散型随机变量熵的链式法则，微分熵也具有链式法则。

（微分熵的链式法则）

（1） $H(X,Y)=H(X)+H(Y|X)$。

（2） $H(X_1,X_2,\cdots,X_n)=$
$H(X_1)+\sum_{i=2}^{n}H(X_i|X_1,\cdots,X_{i-1})$。

（3） $H(X_1,X_2,\cdots,X_n|Y)=$
$H(X_1|Y)+\sum_{i=2}^{n}H(X_i|X_1,\cdots,X_{i-1},Y)$。

定理 1.7.2：互信息计算公式

互信息可以用微分熵来表示，并有如下公式：

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

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

定理 1.7.3：微分熵有关不等式

还可以证明下面一些微分熵，条件微分熵，相对熵，互信息下列不等式成立：

（1） $D(f||g) \geq 0$，并且等号成立的充要条件是 $f = g$ 几乎处处成立。

（2） $I(X;Y) \geq 0$，并且等号成立的充要条件是 X 与 Y 相互独立。

（3） $H(X|Y) \leq H(X)$，并且等号成立的充要条件是 X 与 Y 相互独立。

（4） $H(X_1, X_2, \cdots, X_n) \leq \sum_{i=1}^{n} H(X_i)$，并且等号成立的充要条件是 $X_1, X_2, \cdots, X_n$ 相互独立。

（5） $H(X_1, X_2, \cdots, X_n|Y) \leq \sum_{i=1}^{n} H(X_i|Y)$。

定理 1.7.4：微分熵变换

设 $X=(X_1,X_2,\cdots,X_n)$ 为 n 维随机变量，其概率密度函数为 n 元函数 $f(x)=f(x_1,x_2,\cdots,x_n), x=(x_1,x_2,\cdots,x_n)\in\mathcal{G}$；又设 $y=\varphi(x)=(\varphi_1(x),\varphi_2(x),\cdots,\varphi_n(x))$ 为一一映射，通过这个映射得到 n 维随机变量 $Y=\varphi(X)$，则有

$$H(Y)=H(X)-\int_{\mathcal{G}} f(x)\log|J(\varphi(x))|\,dx,$$

1.44 ~~(2.2)~~

其中 $J(\varphi(x))=J(y)$ 是变换 $y=\varphi(x)$ 逆变换 $x=\psi(y)=(\psi_1(y),\psi_2(y),\cdots,\psi_n(y))$ 的 Jacobian 矩阵的行列式：

$$J(y)=det\left(\frac{\partial\psi_i}{\partial y_j}\right)_{i,j=1,2,\cdots,n}.$$

推论 1.7.1：简单微分熵变换公式

有下列几个简单的微分熵变换公式：

（1） 设 a 为常数，则 $H(X+a)=H(X)$。

（2） 设 a 为非零常数，则 $H(aX)=H(X)+\log|a|$。

（3） 如果采用 n 维线性变换 $Y=AX$，其中 A 是一个 n 阶可逆方阵，则有 $H(Y)=H(X)+\log|det(A)|$。

例题 1.7.1: 一维均匀分布的微分熵

设 $X \sim U(a,b)$，求 $H(X)$。解：

$$f(x) = \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & else \end{cases},$$

$$H(X) = -\int_a^b f(x)\log f(x)dx = \log(b-a).$$

如果 $b-a<1$，则微分熵 $H(X)$ 是负值，因此微分熵与离散熵有本质不同。

例题 1.7.2: 二维均匀分布的微分熵

设 $(X,Y) \sim U(\mathcal{G}), \mathcal{G}: x^2+y^2 \leq 1$，求 $H(X,Y)$。解:

$$f(x,y) = \begin{cases} \frac{1}{\pi} & (x,y) \in \mathcal{G} \\ 0 & else \end{cases},$$

(x, y)

$$H(X) = -\int\int_{\mathcal{G}} f(x,y) \log f(x,y) dxdy = \log \pi.$$

例题 1.7.3: 一维指数分布的微分熵

设 $X \sim E(\theta)$，求 $H(X)$。解：

$$f(x) = \begin{cases} \frac{1}{\theta} e^{-\frac{x}{\theta}} & x > 0 \\ 0 & else \end{cases},$$

$$H(X) = -\int_{-\infty}^{\infty} f(x) \ln f(x) dx = 1 + \ln \theta \ \ \text{nats.}$$

例题 1.7.4: 一维正态分布的微分熵

设 $X \sim N(\mu, \sigma^2)$，求 $H(X)$。解：

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, x \in \mathcal{R},$$

$$\begin{aligned}
H(X) &= -\int_{-\infty}^{\infty} f(x) \ln f(x) dx \\
&= \int_{-\infty}^{\infty} f(x) \left[\frac{1}{2} \ln(2\pi\sigma^2) + \frac{(x-\mu)^2}{2\sigma^2} \right] dx \\
&= \frac{1}{2} \ln(2\pi\sigma^2) \int_{-\infty}^{\infty} f(x) dx + \frac{1}{2\sigma^2} \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx \\
&= \frac{1}{2} \ln(2\pi\sigma^2) + \frac{1}{2\sigma^2} \sigma^2 \\
&= \frac{1}{2} \ln(2e\pi\sigma^2) \quad \text{nats.}
\end{aligned}$$

例题 1.7.5: 二维正态分布的微分熵

设 $(X,Y)\sim N(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)$，求 $H(X,Y)$。解：

$$f(x,y)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}e^{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_1)^2}{\sigma_1^2}-2\rho\frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2}+\frac{(y-\mu_2)^2}{\sigma_2^2}\right]},$$

$$H(X,Y)=-\int\int_{\mathcal{R}^2}f(x,y)\ln f(x,y)dxdy$$

$$=\int\int_{\mathcal{R}^2}f(x,y)\left\{\ln\left(2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}\right)\right.$$
$$\left.+\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_1)^2}{\sigma_1^2}-2\rho\frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2}+\frac{(y-\mu_2)^2}{\sigma_2^2}\right]\right\}dxd$$

续例题 1.7.5: 二维正态分布的微分熵

$$
\begin{aligned}
= \;& \ln\left(2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}\right)\int\int_{\mathcal{R}^2} f(x,y)dxdy \\
& +\frac{1}{2(1-\rho^2)}\left[\frac{1}{\sigma_1^2}\int\int_{\mathcal{R}^2}(x-\mu_1)^2 f(x,y)dxdy\right. \\
& -2\rho\frac{1}{\sigma_1\sigma_2}\int\int_{\mathcal{R}^2}(x-\mu_1)(y-\mu_2)f(x,y)dxdy \\
& \left.+\frac{1}{\sigma_2^2}\int\int_{\mathcal{R}^2}(y-\mu_2)^2 f(x,y)dxdy\right] \\
= \;& \ln\left(2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}\right) \\
& +\frac{1}{2(1-\rho^2)}\left[\frac{1}{\sigma_1^2}\sigma_1^2-2\rho\frac{1}{\sigma_1\sigma_2}Var(X,Y)+\frac{1}{\sigma_2^2}\sigma_2^2\right] \\
= \;& \ln(2e\pi\sigma_1\sigma_2\sqrt{1-\rho^2}) \quad \text{nats.}
\end{aligned}
$$

命题 1.7.1：高维正态分布的微分熵

设 $(X_1, X_2, \cdots, X_n) \sim N(\mu, \Sigma)$ 即概率密度函数为

$$f(x) = \frac{1}{\sqrt{(2\pi)^n \det \Sigma}} e^{(x-\mu)\Sigma^{-1}(x-\mu)^T},$$

其中 $x = (x_1, x_2, \cdots, x_n), \mu = (\mu_1, \mu_2, \cdots, \mu_n)$，并且 $\Sigma = (\sigma_{ij}), \sigma_{ij} = \mathrm{Var}(X_i, X_j), i, j = 1, 2, \cdots, n$ 是协方差矩阵，它是对称正定的，则

$$H(X_1, X_2, \cdots, X_n) = \frac{n}{2} \ln\left[2e\pi(\det \Sigma)^{\frac{1}{n}}\right] \text{ nats.} \qquad \cancel{(2.3)}\ 1.45$$

例题 1.7.6：二维正态分布的互信息

已知 $(X,Y) \sim N(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)$，求 $I(X;Y)$。解：
由例题 ??、~~??~~ 可得：
1.7.4 1.7.5

$$H(X) = \frac{1}{2}\ln\left(2e\pi\sigma_1^2\right), H(Y) = \frac{1}{2}\ln\left(2e\pi\sigma_2^2\right), H(X,Y) = \ln(2e\pi\sigma_1\sigma_2\sqrt{1$$

所以

$$I(X;Y) = H(X) + H(Y) - H(X;Y) = -\frac{1}{2}\ln\left(1-\rho^2\right) \text{ nats.}$$

定义 1.7.6：最大熵定义

有限离散型随机变量熵有最大值，对连续型随机变量的微分熵有没有最大值？即是否有某个概率密度函数 $f(x)$ 的微分熵不小于任何其它概率密度函数 $f(x)$ 的微分熵？本小节只对一维随机变量来讨论这种问题。

设 $\mathcal{F}$ 是定义在同一区间 $\mathcal{X}$ 上的概率密度函数的集合，如果存在概率密度 $f_0 \in \mathcal{F}$ 使得

$$H(f_0) = \max_{f \in \mathcal{F}} H(f),$$

*则称概率密度函数 $f_0(x)$ 为 $\mathcal{F}$ 上的**最大熵分布**，并称 $H(f_0)$ 为 $\mathcal{F}$ 上的**最大熵**。*

定理 1.7.5：最大熵存在条件

设 $\mathcal{F}$ 是定义在同一区间 $\mathcal{X}$ 上的概率密度函数的集合，如果存在 $f_0 \in \mathcal{F}$ 使得对任何 $f \in \mathcal{F}$ 下面积分是一个与 f 无关的常数 H_0，

$$-\int_{\mathcal{X}} f(x) \log f_0(x) dx = H_0,$$

则 f_0 是 $\mathcal{F}$ 中的最大熵分布，而 H_0 就是最大熵。

证明：

事实上：

$$
\begin{aligned}
H(f) &= -\int_{\mathcal{X}} f(x)\log f(x)dx = -\int_{\mathcal{X}} f(x)\log f_0(x)\frac{f(x)}{f_0(x)}dx \\
&= -\int_{\mathcal{X}} f(x)\log f_0(x)dx - \int_{\mathcal{X}} f(x)\log\frac{f(x)}{f_0(x)}dx \\
&= H_0 - D(f||f_0) \le H_0.
\end{aligned}
$$

利用这个定理可以证明三个最大熵定理，作为三个例子。

例题 1.7.7：均匀分布是最大熵

设 $\mathcal{F}=\left\{f(x)\geq 0 | \int_a^b f(x)=1\right\}$，它表示定义在有限区间 $[a,b]$ 上的概率密度函数的集合，则它的最大熵分布为均匀分布，因此最大熵由例题 1.7.1 给出。

例题 1.7.8：指数分布是最大熵

设 $\theta > 0$ 为常数，记

$$\mathcal{F} = \left\{ f(x) \geq 0 | \int_0^{\infty} f(x) = 1, \int_0^{\infty} x f(x) dx = \theta \right\},$$

它表示定义在半直线 $(0, \infty)$ 上并且期望为常数的概率密度函数的集合，则最大熵分布为参数 θ 的指数分布，因此最大熵由例题 1.7.3 给出。

例题 1.7.7：正态分布是最大熵

设 $\mu, \sigma > 0$ 为已知常数，记

$$\mathcal{F} = \left\{ f(x) \geq 0 \middle| \int_{-\infty}^{\infty} f(x) = 1, \mu = \int_{0}^{\infty} x f(x) dx, \sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(\right.$$

它表示定义在全直线 $(-\infty, \infty)$ 上并且期望与方差都为常数的概率密度函数的集合，则最大熵分布为参数 μ, σ^2 的正态分布，因此最大熵由例题 1.7.~~5~~4 给出。

证明例题 1.7.7:

取均匀分布密度函数 $f_0(x)$，则有

$$-\int_a^b f(x)\log f_0(x)dx = -\log\frac{1}{b-a}\int_a^b f(x)dx = \log(b-a) = H_0.$$

由定理 3.2，例题 1.7.7 得证。

练习：

证明后两个例子。