信息论

信号传输与处理的理论基础

Markov序列和数据处理不等式

连续随机变量的信息熵

教程阅读：2.8和2.10节；8.1, 8.4, 8.5和8.6节

互信息量的凸性和凹性(1)

* 将条件概率公式p(x,y)=p(y|x)p(x)代入互信息量的定义式得
* $\mathrm{I(X;Y)} \equiv \Sigma_{x,y}\, p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right) = \Sigma_{x,y}\, p(y|x)p(x)\log\left(\frac{p(y|x)}{p(y)}\right)$
* 其中 $p(y) = \Sigma_x p(y|x)p(x)$。

* <u>针对I(X;Y) 的两种函数观点</u>
* (1) 将条件概率参数p(y|x) 视为常数、p(x)视为变量，相应的函数记为 C;
* (2)将条件概率参数p(y|x) 视为变量、p(x)视为常数，相应的函数记为 D;
* **C是凹函数；D是凸函数**

* 注：本讲义在下面所给的证明方法与教程中的不同，请同时阅读教程中的证明，
* 并比较两者的特点。

互信息量的凸性和凹性(2)

- 若将条件概率参数$p(y|x)$视为常数、$p(x)$视为变量，则
- $I(X;Y) = \Sigma_{x,y}\, p(y|x)p(x)\log(\frac{p(y|x)}{p(y)})$
- 是$p(x)$的凹函数，其中$p(y) = \Sigma_x p(y|x)p(x)$。

(Note) $I(X;Y) \equiv \sum_{x,y} P_{xy} \log(\frac{P_{xy}}{P_x P_y})$，其中 $P_y = \sum_x P_x P_{y|x}$，故 $\partial P_y / \partial P_x = P_{y|x}, \forall x, y$.

$= \sum_{x,y} P_{y|x} \cdot P_x \log(P_{y|x}/P_y)$

(1) $I(X;Y)$是(P_x)的凹函数：

$$\frac{\partial I}{\partial P_x} = \Sigma_y P_{y|x} \log(\frac{P_{y|x}}{P_y}) - \sum_{x',y} P_{y|x'} \cdot P_{x'} \underbrace{\frac{\partial}{\partial P_x} \log P_y}_{\frac{1}{P_y}\frac{\partial P_y}{\partial P_{y|x}} = \frac{P_x}{P_y}}$$

$$= \sum_y P_{y|x} \log P_{y|x} - \sum_{y|x} P_{y|x} \log P_y - \underbrace{\sum_{x',y} P_{y|x'} P_{x'} P_x / P_y}_{P_x \cdot \sum_y \frac{1}{P_y} \cdot (\sum_{x'} P_{y|x'} P_{x'}) = P_x \cdot \sum_y \frac{1}{P_y} \cdot P_y = P_x}$$

$$= \underbrace{\Sigma_y P_{y|x} \log P_{y|x}}_{与(P_x)无关的常数} - \sum_{y|x} P_{y|x} \cdot \log P_y - P_x$$

$$\frac{\partial^2 I}{\partial P_x \partial P_{x'}} = -\sum_y P_{y|x} \cdot \frac{P_{y|x'}}{P_y} - \delta_{xx'}$$

$$\sum_{x,x'} \xi_x \xi_{x'} \frac{\partial^2 I}{\partial P_x \partial P_{x'}} = -\sum_y \frac{1}{P_y} \left(\sum_x \xi_x \cdot P_{y|x}\right)^2 - \sum_x \xi_x^2 \leqslant 0 \text{ 且} = 0 \text{ iff } \xi_x = 0, \forall x$$

故 $I(X;Y)$是(P_x)的凸函数且严格凸. [注]这里导出的$\frac{\partial I}{\partial P_x}$和$\frac{\partial^2 I}{\partial P_x \partial P_{x'}}$，表达式对计算Schannon容易的数值算法有用.

(2) $I(X;Y)$是$(P_{y|x})$的凸函数

互信息量的凸性和凹性(3)

* 若将条件概率参数$p(y|x)$视为变量、$p(x)$视为常量，则
* $$\mathrm{I(X;Y)} = \Sigma_{x,y}\, p(y|x)p(x)\log\left(\frac{p(y|x)}{p(y)}\right)$$
* 是$p(y|x)$的凸函数，其中$p(y) = \Sigma_x p(y|x)p(x)$。

(2) $I(X;Y) = \sum_{x,y} P_{xy} \log\left(\frac{P_{xy}}{P_x P_y}\right) = \sum_{x,y} P_{y|x} P_x \log P_{y|x} - \sum_{x,y} P_{y|x} P_x \log P_y$

是$(P_{y|x})$的凸函数.

$P_y = \sum_x P_{y|x} \cdot P_x$, $\partial P_y / \partial P_{y|x} = P_x$, $\partial P_y / \partial P_{y'|x'} = 0$, $\forall y' \neq y$.

计算 $\partial I / \partial P_{y|x} = P_x (1 + \log P_{y|x}) - P_x \log P_y - \underbrace{\sum_{x'} P_x P_{x'} P_{y|x'} / P_y}_{(P_x/P_y) \sum_{x'} P_{x'} P_{y|x'} = \frac{P_x}{P_y} \cdot P_y = P_x}$.

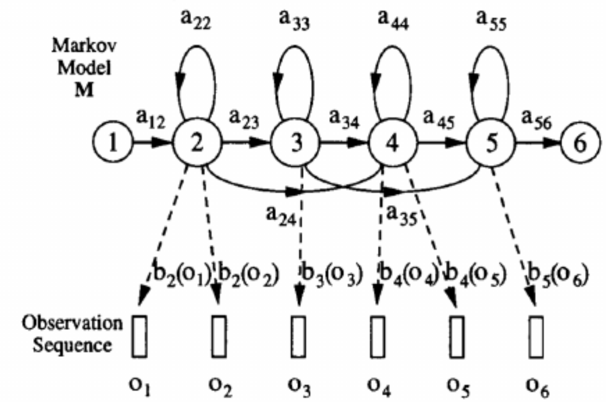
$$\partial^2 I / \partial P_{y|x} \partial P_{y'|x'} = \frac{P_x \delta_{xx'} \delta_{yy'}}{P_{y|x}} - \frac{P_x P_{x'} \delta_{yy'}}{P_y}$$

$$\sum_{x y x' y'} \xi_{y|x} \xi_{y'|x'} \frac{\partial^2 I}{\partial P_{y|x} \partial P_{y'|x'}} = \sum_y \left[\sum_x \frac{P_x \xi_{y|x}^2}{P_{y|x}} - \frac{1}{P_y}\left(\sum_x \xi_{y|x} \cdot P_x\right)^2\right] \geq 0.$$

注：该凸性结论本课程今后用不到，仅用到关于凹性的结论。

习题（选做）：证明该不等式。

Markov序列(1)



- Markov序列
- 随机序列…, X_{n-1}, X_n, X_{n+1}, … 定义为Markov序列，是指
- 对任何$n>n_1>n_2>\ldots>n_p$恒有
- $$P[X_n|X_{n1},X_{n2},\ldots,X_{np}] = P[X_n|X_{n1}]$$
- Markov性质的含义
- 基于任何历史数据(X_{n1},X_{n2},…,X_{np})推断未来状态X_n，其效果总是等价于基于距离未来时刻最近的历史数据X_{n1}推断X_n。
-
- 记号：在上述意义下的Markov序列，记为…$\rightarrow X_{n-1}\rightarrow X_n\rightarrow X_{n+1}\rightarrow$ … 。

Markov序列(2)

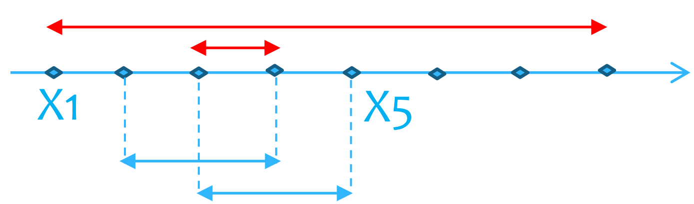
* Markov序列的基本性质
* （1）对称性
* 若随机序列…,X_{n-1}, X_n, X_{n+1}, … 是Markov序列，则沿相反的时间方向也是Markov序列。
* 证明：以三个时间点的情况为例，
* $P[X_{n-1}|X_n, X_{n+1}] = P[X_{n-1}, X_n, X_{n+1}]/P[X_n, X_{n+1}]$（条件概率的定义）
* $= P[X_{n+1}|X_n, X_{n-1}]P[X_n, X_{n-1}]/P[X_n, X_{n+1}]$（根据条件概率公式重写$P[X_{n-1}, X_n, X_{n+1}]$）
* $= P[X_{n+1}|X_n]P[X_n, X_{n-1}]/P[X_n, X_{n+1}]$（$\rightarrow X_{n-1} \rightarrow X_n \rightarrow X_{n+1} \rightarrow$）
* $= P[X_{n+1}|X_n]P[X_{n-1}|X_n]P[X_n]/P[X_{n+1}|X_n]P[X_n]$
* 根据条件概率公式重写$P[X_n, X_{n-1}]$和$P[X_n, X_{n+1}]$）
* $= P[X_{n-1}|X_n]$
* 习题：设$n<n_1<n_2<\ldots<n_p$，按以上思路证明$P[X_n|X_{n1},X_{n2},\ldots,X_{np}] = P[X_n|X_{n1}]$。
* 以上分析表明基于任何未来的数据$(X_{n1},X_{n2},\ldots,X_{np})$推断过去的状态$X_n$，其效果总是等价于基于距离过去最近的未来数据$X_{n1}$推断$X_n$。
* 采用前面的记号，就是$\ldots \leftarrow X_{n-1} \leftarrow X_n \leftarrow X_{n+1} \leftarrow \ldots$。

Markov序列(3)

- Markov序列的基本性质
- （2）继承性
- 若随机序列…,X_{n-1}, X_n, X_{n+1}, … 是<u>Markov序列</u>，则其任何子序列也是Markov序列。
- 思考题：为什么？
- （3）数据处理不等式
- 若 $X \to Y \to Z$，则 $I(X;Z) \leqq I(X;Y)$
- 证明： $I(X;Z) - I(X;Y) = \Sigma_{xz}\, p(x,z)\log[p(x,z)/p(x)p(z)]$
- $\qquad - \Sigma_{xy}\, p(x,y)\log[p(x,y)/p(x)p(y)]$
- $= \Sigma_{xyz}\, p(x,y,z)\log[p(x,z)/p(x)p(z)] - \Sigma_{xy}\, p(x,y,z)\log[p(x,y)/p(x)p(y)]$（为什么？）
- $= \Sigma_{xyz}\, p(x,y,z)\log[p(x,z)p(y)/p(x,y)p(z)]$
- $= \Sigma_{xyz}\, p(x|y,z)p(y,z)\log[p(x|z)/p(x|y)]$ $\quad (p(x,y,z)=p(x|y,z)p(y,z))$
- $= \Sigma_{xyz}\, p(x|y)p(y,z)\log[p(x|z)/p(x|y)]$ $\quad (X \leftarrow Y \leftarrow Z)$
- $= -\Sigma_{xyz}\, p(y,z)\, p(x|y)\log[p(x|y)/p(x|z)] \leqq 0$ （为什么？） 其他的证明参见2.8节。

Markov序列(4)

- Markov序列的基本性质
- (4) 狭义数据处理不等式
- g是任何函数，X和Y是任意的随机变量，则 $X \to Y \to g(Y)$，因此 $I(X; g(Y)) \leqq I(Y; g(Y))$
-
- 小结：对Markov序列$\ldots \to X_{n-1} \to X_n \to X_{n+1} \to \ldots$，"距离近"的状态变量的互信息量，总是不低于"距离远"的状态量之间的互信息量。
- 习题 对以上Markov序列，确定以下互信息量的相对大小，或指出两者不能比较：
- $I(X_2; X_4)$和$I(X_3, X_4)$;
- $I(X_1; X_{100})$和$I(X_3, X_4)$;
- $I(X_2; X_4)$和$I(X_3, X_5)$.



- 答案：小于；小于；不能比较（相对大小不确定）。

小结与习题（教程第一版第二章）

讲义未包含/阅读教程的内容:

* 条件互信息量$I(X:Y|Z)$及其基本性质，参阅2.5节定义式（2.60-61）。该量在后续应用很少，属于一个有用但不是本质性的工具。凡是应用$I(X:Y|Z)$分析和论证的问题，都可以等效地采用熵和条件熵的已经学过的关系来论证。
* 2.9节：充分统计量，可略去。
* 2.10节：核心内容是*Fano*不等式，该不等式重要但在后续应用很少。该节没有新概念，可作为前述已学知识的一个有趣的应用练习。

* 习题： 2.2，2.3，2.4，2.5，2.10，2.15，2.21，2.26，2.27，
* 2.28，2.29，2.30（较难/选做）.

* 部分参考答案或简要提示：
* 2.2 暂时将X作为取值离散的随机变量，下同：(a)$H(X) = H(2^X)$；(b) $H(X) \geq H(\cos X)$；
* 思考：如果Y是X的函数，什么情况下$H(X)=H(Y)$?

部分习题选解（概要）

* 习题2.10
* 因为 X_1、X_2 的取值范围不相交，可以 定义一个随机变量X，其取值
* 范围是X_1、X_2值域 的并集，概率$P[X=X_1] = a$，概率$P[X=X_2] = 1 - a$.
* 再定义一个函数 $y=f(X)$，$f(X)=1$，若$X=X_1$；$f(X) = 2$，若$X=X_2$。
* 于是 $H(X) = H(X, f(X)) = H(y) + H(X|y)$
* $= H(y) + p[y = 1]H(X|y = 1) + p[y = 2]H(X|y = 2)$
* $= H(a,1-a) + aH(X_1) + (1 - a)H(X_2)$
* 其中$H(a,1-a) = -a \log a - (1 - a) \log(1 - a)$. 【补充完整的细节】

部分习题选解（概要）

* 习题2.15
* 对Markov序列 $X_1 \to X_2 \to X_3 \to \cdots \to X_n$
* 有 $P[X_1,\ldots,X_n]=P[X_n|X_{n-1},\ldots,X_1]P[X_{n-1},\ldots,X_1]$
* $= P[X_n|X_{n-1}]P[X_{n-1}|X_{n-2},\ldots,X_1]P[X_{n-2},\ldots,X_1]$
* $= P[X_n|X_{n-1}]P[X_{n-1}|X_{n-2}]P[X_{n-2}|X_{n-3},\ldots,X_1]$
* $=\ldots=P[X_n|X_{n-1}]P[X_{n-1}|X_{n-2}]\ldots P[X_2|X_1]P[X_1]$
* 代入$I(X_1; X_2,\ldots,X_n)$
* $= \Sigma P[X_n,\ldots,X_1]\log\{P[X_n,\ldots,X_1]/P[X_1]P[X_2,\ldots,X_n]$
* 计算化简得
* $I(X_1; X_2,\ldots,X_n) = I(X_1; X_2)$
* 【完成完整的计算，并思考：该结论的含义是什么？】

部分习题选解（概要）

* 习题2.21

* 证明概要

$$
\begin{aligned}
P(p(X) < d) \log \frac{1}{d} &= \sum_{x:p(x)<d} p(x) \log \frac{1}{d} \\
&\leq \sum_{x:p(x)<d} p(x) \log \frac{1}{p(x)} \\
&\leq \sum_{x} p(x) \log \frac{1}{p(x)} \\
&= H(X)
\end{aligned}
$$

部分习题选解（概要）

* 习题2.29 推导提示

* (a) $H(X, Y|Z) = H(X|Z) + H(Y|X,Z) > H(X|Z)$
* (b) $I(X, Y;Z) = I(X;Z) + I(Y;Z|X) > I(X;Z)$
* (c) $H(X, Y,Z) - H(X, Y) = H(Z|X, Y) = H(Z|X) - I(Y;Z|X)$
* $< H(Z|X) = H(X,Z) - H(X)$
* (d) $I(X;Z|Y) + I(Z; Y) = I(X, Y;Z) = I(Z; Y|X) + I(X;Z)$
* 和
* $I(X;Z|Y) = I(Z; Y|X) - I(Z; Y) + I(X;Z)$
* 注意该题的结论实际上是一个等式。

对连续随机变量的推广(1)

* 前面建立的全部概念和关系，都可以推广到连续型随机变量。

【在学习这部分时，注意连续随机变量情形特有的一些性质，例如信息熵的数值可以为负】

【阅读教程第8章 微分熵 8.1、8.2（在学习第三章后阅读）、8.4、8.5、8.6（略去定理8.6.6）】

1　连续随机变量X，取值范围记为S，概率密度为$f(x)$, 定义其信息熵

$$\boxed{h(X) = -\int_S f(x)\log f(x)\,dx,}$$

2　连续随机变量X和Y，其联合概率密度是$f(x,y)$, 定义其联合熵

$$\boxed{h(X,Y) = -\int dxdy\ f(x,y)\log f(x,y)}$$

3　连续随机变量X和Y，定义其互信息量

$$\boxed{I(X;Y) = -\int dxdy\ f(x,y)\log \{f(x,y)/f(x)f(y)\}}$$

对连续随机变量的推广(2)

4 基本性质

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

$$h(X|Y) \le h(X)$$

$$h(X,Y) \le h(X) + h(Y)$$并且上界在X和Y概率独立时达到

$$I(X;Y) \ge 0$$

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

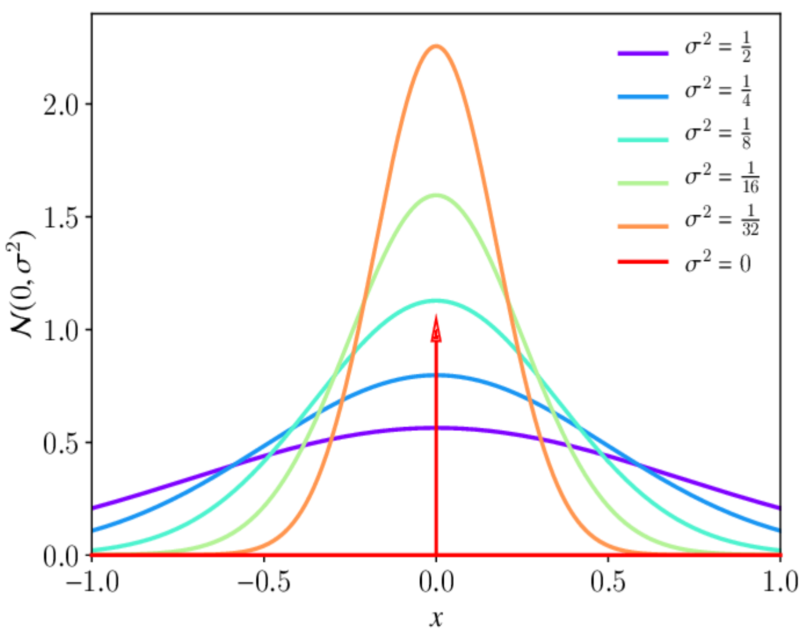
【习题】仿照离散变量的情形检验上述关系。

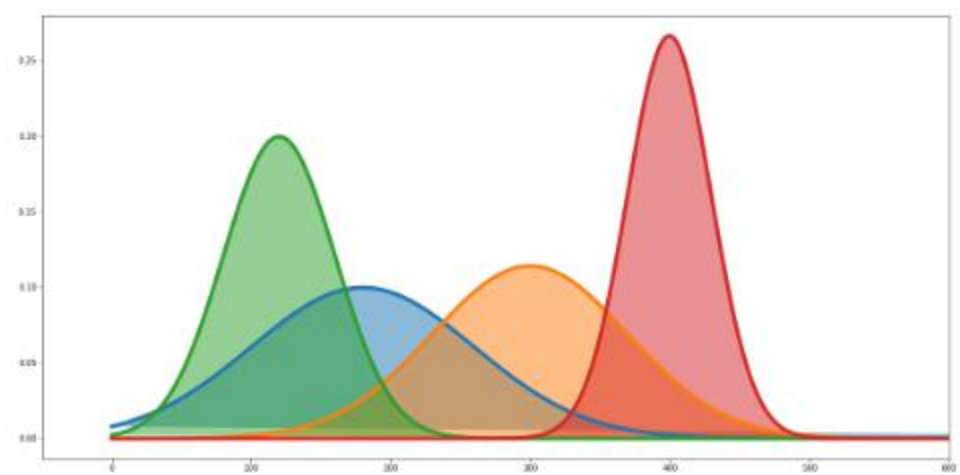
对连续随机变量的推广(3)

* 5 基本实例
* 5.1 一维Gauss变量X
* 概率密度 $N(m,\sigma^2) = (2\pi\sigma^2)^{-1/2} exp(-\frac{1}{2\sigma^2}(x-m)^2)$
* 信息熵H(X)
* $= -\int_{-\infty}^{+\infty} dx\, (2\pi\sigma^2)^{-1/2} exp(-\frac{1}{2\sigma^2}(x-m)^2) log_e[(2\pi\sigma^2)^{-1/2} exp(-\frac{1}{2\sigma^2}(x-m)^2)]$
* $= -\int_{-\infty}^{+\infty} dx\, (2\pi\sigma^2)^{-1/2} exp(-\frac{1}{2\sigma^2}(x-m)^2) log_e[(2\pi\sigma^2)^{-1/2}]$
* $+ \frac{1}{2\sigma^2}\int_{-\infty}^{+\infty} dx\, (x-m)^2 (2\pi\sigma^2)^{-1/2} exp(-\frac{1}{2\sigma^2}(x-m)^2)$
* $= log_e[(2\pi\sigma^2)^{1/2}] + ½ = log_e[(2\pi e\sigma^2)^{1/2}]$

对连续随机变量的推广(4)

* 5 基本实例
* 5.1（续）一维Gauss变量X~N(m,σ^2) 的信息熵
* $$H(X)=\frac{1}{2}log_e(2\pi e\sigma^2)$$
* 注：Gauss变量的熵H(X)同均值m无关，仅与方差σ^2有关，并是方差的增函数。
* 你能定性地解释为什么如此吗?





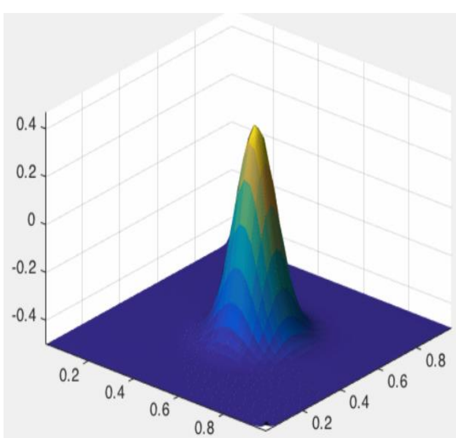
对连续随机变量的推广(5)

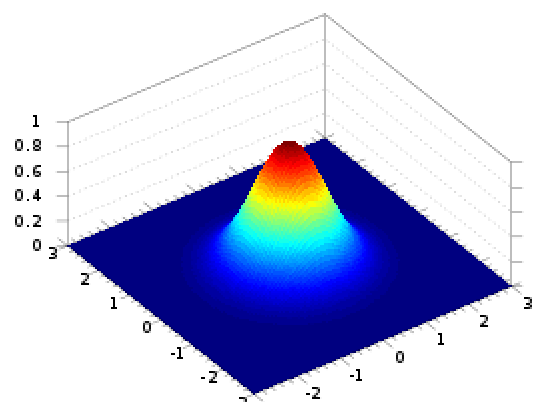
* 5 基本实例
* 5.2 n维Gauss随机向量$\mathbf{X}=(X_1,\ldots,X_n)\sim N(\boldsymbol{m},\mathbf{Q})$
* $$N(\boldsymbol{m},\mathbf{Q})=(2\pi|Q|)^{-1/2}exp(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{m})^{T}Q^{-1}(\boldsymbol{x}-\boldsymbol{m}))$$
* **m**是均值向量，|Q|表示Q的行列式，Q是正定对称的协方差矩阵：
* $$Q_{ij}=E[(x_i-m_i)(x_j-m_j)]$$
* 信息熵 $$H(X)=\frac{1}{2}\log_e((2\pi e)^n|Q|)$$

【习题】

(1)根据熵的定义推导该结果；

(2)当$X_1,\ldots,X_n$是独立的Gauss变量，方差分别为σ_1^2、$\ldots\sigma_n^2$，这时H(X)的表达式是什么？结合普遍的关系式$h(X_1,\ldots,X_n)\leq h(X_1)+\ldots+h(X_n)$，能由此导出著名的Hardamard不等式(8.64)。





对连续随机变量的推广(6)

* 6 $I(X;Y) = -\int dxdy\, f(x,y)\log\{f(x,y)/f(x)f(y)\}$
* $= -\int dxdy\, f(y|x)f(x)\log\{f(y|x)/f(y)\}$
* $f(y) = \int dx\, f(y|x)f(x)$
* $\boxed{I(X;Y)\text{是}f(y|x)\text{的凸泛函、}f(x)\text{的凹泛函}}$

* 补充知识：泛函(*functional*)

* (1) 所谓泛函，就是一个随不同函数而变化的实数或复数，即函数的函数。
* (2) 典型的实例：h(X),和I(X;Y), 两者是概率密度的函数；
* Dirac泛函：$\delta_a[f(x)] = f(a)$； $\delta_a^{(m)}[f(x)] = (-1)^m f^{(m)}(a)$
* (3) 当函数变化时，可以考虑泛函的变分(variation)，即泛函微分:
* $F[f(x)+\Delta f(x)] \approx F[f(x)]+L_f[\Delta f(x)]$
* 其中$\Delta f(x)$是幅度很小的函数、L_f是$\Delta f(x)$的所谓线性泛函。则L_f定义做泛函F在$f(x)$处的变分。
* (4) 如果在$f^*(x)$处$L_{f^*}[\Delta f(x)] = 0$对任何$\Delta f(x)$成立，则$f^*(x)$是使泛函达到某种极值的函数。
* 【以上概念参见下面的例子。在Gauss信道容量一章将继续出现泛函及其优化的例子】

对连续随机变量的推广(7)

* 7 关于Gauss概率分布的一个极值性质（定理8.6.5）

* 在均值为给定的向量**m**、协方差矩阵为给定的正定对称矩阵**Q**的各种n维随机变量X中，Gauss变量的熵h(X)最大。
* 证明：第一步：考虑带约束的极值问题
* $\max -\int d\boldsymbol{x}\, f(\boldsymbol{x})\log f(\boldsymbol{x})$
* $\text{s.t.} \int d\boldsymbol{x}\boldsymbol{x}f(\boldsymbol{x}) = \boldsymbol{m}, \quad \int d\boldsymbol{x}(\boldsymbol{x}-\boldsymbol{m})(\boldsymbol{x}-\boldsymbol{m})^{\mathrm{T}}f(\boldsymbol{x}) = \boldsymbol{Q}$
* 第二步：针对约束引进乘子向量$\boldsymbol{a}$和乘子矩阵**A**，计算广义目标泛函
* $\mathrm{F}[f(\boldsymbol{x})] = -\int d\boldsymbol{x}\, f(\boldsymbol{x})\log f(\boldsymbol{x}) + \boldsymbol{a}^{\mathrm{T}}(\int d\boldsymbol{x}\boldsymbol{x}f(\boldsymbol{x}) - \boldsymbol{m})$
* $+ tr(\mathbf{A}(\int d\boldsymbol{x}(\boldsymbol{x}-\boldsymbol{m})(\boldsymbol{x}-\boldsymbol{m})^{\mathrm{T}}f(\boldsymbol{x}) - \boldsymbol{Q}))$
* 对$f(\boldsymbol{x})$的变分。
* 【符号】 $\int d\boldsymbol{x}$是n重积分$\int \mathrm{dx}_1 \ldots dx_n$, tr(M)表示矩阵的迹，即对角线元素之和。
* 【注】这里对定理8.6.5的推导与教程不同，请阅读教程上的推导，并比较两种方法的特点。

对连续随机变量的推广(8)

* 7 关于Gauss概率分布的一个极值性质（定理8.6.5）

* 证明（续）计算泛函 $F[f(\mathbf{x})] = -\int d\mathbf{x}\, f(\mathbf{x})\log f(\mathbf{x}) + \mathbf{a}^{\mathrm{T}}(\int d\mathbf{x}\mathbf{x}f(\mathbf{x}) - \mathbf{m})$
* $+ tr(\mathbf{A}(\int d\mathbf{x}(\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})^{\mathrm{T}}f(\mathbf{x}) - \mathbf{Q}))$ 对$f(\mathbf{x})$的变分：
* 仿照微分算法计算$\Delta f(\mathbf{x})$无穷小时的差，得到
* $F[f(\mathbf{x})+\Delta f(\mathbf{x})] - F[f(\mathbf{x})] \approx -\int d\mathbf{x}\, \Delta f(\mathbf{x})(1+\log f(\mathbf{x})) + \mathbf{a}^{\mathrm{T}}(\int d\mathbf{x}\mathbf{x}\, \Delta f(\mathbf{x})$
* $+ tr(\mathbf{A}(\int d\mathbf{x}(\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})^{\mathrm{T}}\Delta f(\mathbf{x}))$
* 【习题】计算上述结果。注意将第一项对比普通的微分公式 $d(u\log u)=(1+\log u)du$.
* <u>第三步</u>：令以上表达式为零，得到一个关于最优分布的积分等式。由于该等式需要对其中任意的函数$\Delta f(\mathbf{x})$被满足，因此被积函数中$\Delta f(\mathbf{x})$的权因子必须为0：
* $1+\log f^*(\mathbf{x}) + \mathbf{a}^{\mathrm{T}}\mathbf{x}+tr(\mathbf{A}(\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})^{\mathrm{T}}) = 0$
* 这就是使得熵达到极大值的概率分布$f^*(\mathbf{x})$应满足的方程。
* <u>第四步</u>：容易解出$f^*(\mathbf{x})=exp(\mathbf{x}$的二次型)，这种函数形式的概率分布正是
* Gauss分布。证毕。

【习题】应用初等概率轮的知识，完成最后一步的论证。

对连续随机变量的推广(9)

* 7 小结（定理8.6.5）

* 在均值为给定的向量**m**、协方差矩阵为给定的正定对称矩阵**Q**的各种n维随机变量X中，<u>Gauss变量</u>的熵h(X)最大。
* 特例：如果一个标量随机信号x的均值为零、功率P给定，则Gauss信号的熵最大。

* 【习题】证明上述结论。注意一维信号若均值为0，则方差恰是其功率。
* 【注】改组结论将在学习Gauss信道（第9章）和MIMO信道的容量性能时用到。
* 【习题】用类似的方法证明：如果一维随机信号X的均值为已知的正数 ***m***，
* （协方差任意）则指数分布的随机信号具有最大熵h(X)。

习题（第8章 微分熵）

* 习题8.1、8.5
* 参考答案：
* 8.1(a) 指数随机变量的熵

$$\begin{aligned} h(f) &= -\int_0^\infty \lambda e^{-\lambda x}[\ln\lambda - \lambda x]dx \\ &= -\ln\lambda + 1 \text{ nats.} \\ &= \log\frac{e}{\lambda} \text{ bits.} \end{aligned}$$

* (b) Laplace随机变量的熵h(X) = log(2e/λ)
* (c) 独立Gauss变量之和的熵h(X1+X2) = $\frac{1}{2}\log_e(2\pi e(\sigma_1^2+\sigma_2^2))$
* 8.5 应用积分变量变换公式计算得：

$$\begin{aligned} h(A\mathbf{X}) &= -\int g(\mathbf{y})\ln g(\mathbf{y})\,d\mathbf{y} \\ &= -\int \frac{1}{|A|}f(A^{-1}\mathbf{y})\left[\ln f(A^{-1}\mathbf{y}) - \log|A|\right]\,d\mathbf{y} \\ &= -\int \frac{1}{|A|}f(\mathbf{x})\left[\ln f(\mathbf{x}) - \log|A|\right]|A|\,d\mathbf{x} \\ &= h(\mathbf{X}) + \log|A|. \end{aligned}$$

下一讲：典型集合与渐进均分性质

* 预习：教程第三章 渐进均分性质
* 基础：概率的大数定律