《信息论基础》
Elements of Information Theory

第一次读书会

第一章、第二章

主讲：Jake

为何要读此书？

感兴趣的主要章节和顺序

* 基础信息论模块
  * 1、绪论与概论；2、熵、相对熵与互信息；5、数据压缩；7、信道容量；9、高斯信道；10、率失真理论；13、通用信源编码；14、科尔莫戈罗夫复杂度；15、网络信息论

感兴趣的主要章节和顺序

* 概率论与数理统计模块
  * 3、渐近均分性；4、随机过程的熵率；8、微分熵；11、信息论与统计学；12、最大熵；17、信息论中的不等式

感兴趣的主要章节和顺序

* 博弈论模块
  * 6、博弈与数据压缩；16、信息论与投资组合理论

建议顺序

* 第一次：1、2；
* 第二次：3；
* 第三次：5
* 第四次：7
* 第五次：9
* 第六次：14(参考文献[1])
* 此谓第一期（计划到2013二月份）

建议顺序

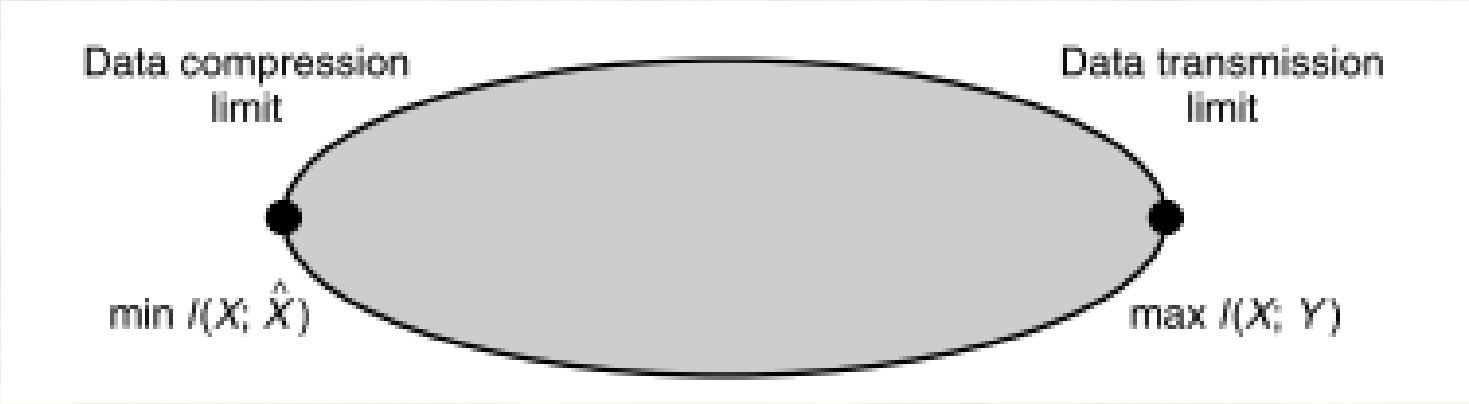
- 第七次：4+参考文献[2]
- 第八次：11
- 第九次：11+参考文献[3]或[4]
- 第十次：12+参考文献[5]或[6]
- 第十一次：6
- 第十二次：16
- 此谓第二期（时间将视情况而定）

参考文献

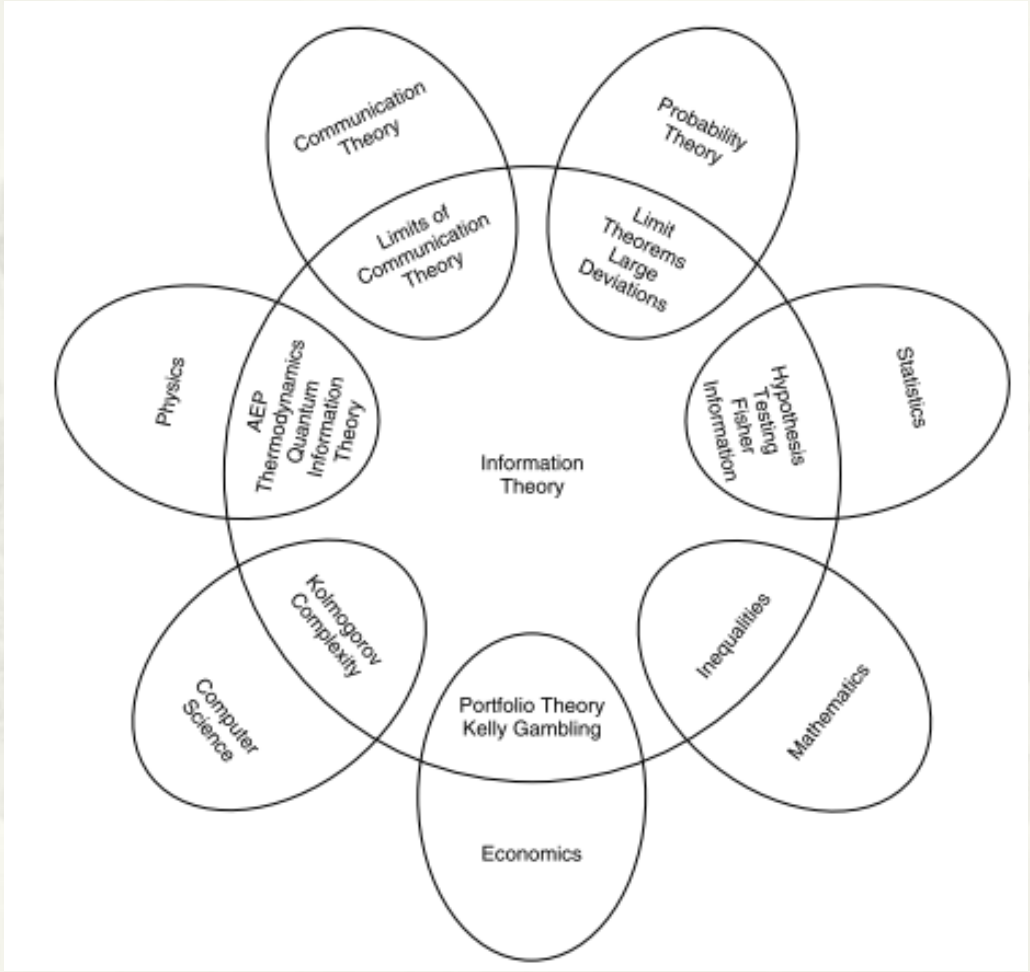
* [1], On universality of Zipf Law
* [2], Least effort and the origins of scaling in human language
* [3], Science from Fisher Information - A unification
* [4], Methods of Information Geometry
* [5], Nonadditive entropy and nonextensive statistical mechanics - An overview after 20 years
* [6], Maximum entropy and the state-variable approach

第一章、绪论与概览

* 信息论的核心：临界数据压缩的值（熵）和临界通信传输速率的值（信道容量C）



第一章、绪论与概览


Communication Theory
Probability Theory
Limits of Communication Theory
Limit Theorems Large Deviations
Physics
AEP Thermodynamics Quantum Information Theory
Information Theory
Hypothesis Testing Fisher Information
Statistics
Kolmogorov Complexity
Computer Science
Portfolio Theory Kelly Gambling
Inequalities
Mathematics
Economics

信息论与概率论

- 概率与各种信息量的关系
  - 概率分布：单变量分布、双变量分布、条件概率分布
  - 信息量（熵）：Shannon熵、互信息、条件熵
  - 信息量是对概率分布弥散性质的一种平均描述
- 新的观点：
  - Information theory must precede probability theory and not be based on it.
  - ---A.N.Kolmogorov

信息论与计算机科学

- Kolmogrov复杂度
- 一个数据段的复杂度可以定义为计算该数据串所需的最短二进制程序的长度。
- 如果序列服从熵为H的分布，则它的复杂度C近似为Shannon熵H
- 算法复杂度与计算复杂度二者之间的互补关系：

热力学与信息论

- 卡诺热机→克劳修斯熵 $dS = \int_{reversible} dQ/T$
- 玻尔兹曼熵 $S = \ln W$
- 玻尔兹曼的H定理，第一次提出了熵的现代表达式： $H = -\int p(x)\log p(x)$
- 吉布斯将统计力学中的熵统一形式
- Shannon熵
- E. T. Jaynes的统计物理

科学的哲学观

- 奥克姆剃刀原理：“因不宜超出果之所需”
- Solomonoff和Chaitin猜想：
  - 谁能获得适合处理数据的所有程序的加权组合，并能观察到下一步的输出值，谁就能得到万能的预测程序。
  - 当然，这样的程序及其不切实际，因为清除所有不适合生成现有数据的程序需要花费的时间是不可接受的
- 何为好的科学理论：利用尽可能少的输入信息，能够得到最多的输出信息的模型。

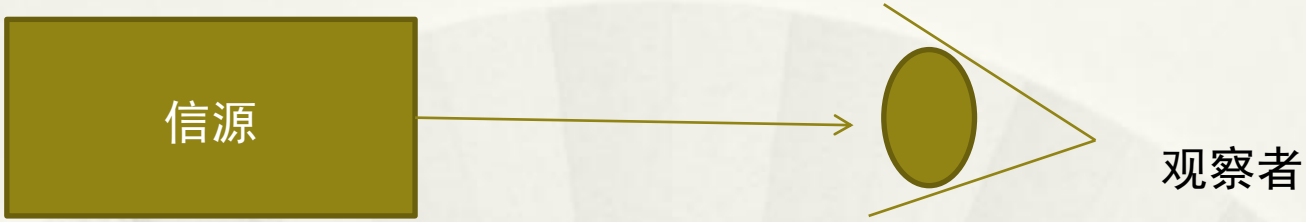
信息论与经济学（投资）

- 最优投资第一原理：
  - 不要把鸡蛋放到同一个篮子里
  - 这就是熵原理，如果把分在一个篮子里的鸡蛋比例看作概率，那么最大化熵就得到最合理的投资组合。
- 赛马问题：有n匹马，每匹马获胜的概率为pi，那么你应该如何持续地对这n匹马下注？答案是，你下注在第i匹马上的赌资比例为：log(pi)

信息论与复杂系统

- 信息论为“老三论”的重要分支
- 生命是横跨在物质与信息边界上的系统
- 复杂系统的研究应该奠定在现代信息论基础之上
  - 信息力
  - 统计标度律与信息论

第二章：熵、相对熵与互信息



- 信源发射信号（字母），信号的可能集合为：$\{a1, a2, . . , an\}$
- 每一个元素出现的概率$\{p1, p2, \cdots, pn\}$
- 那么，定义第i个元素所包含的信息量为$\log(pi)$
- 则，信源所容纳的信息为所有n个元素的平均信息量

$$I = E(I_i) = \sum_i p_i I_i = \sum_i p_i \log(p_i) = -H$$

$$H = -\sum_i p_i \log(p_i)$$

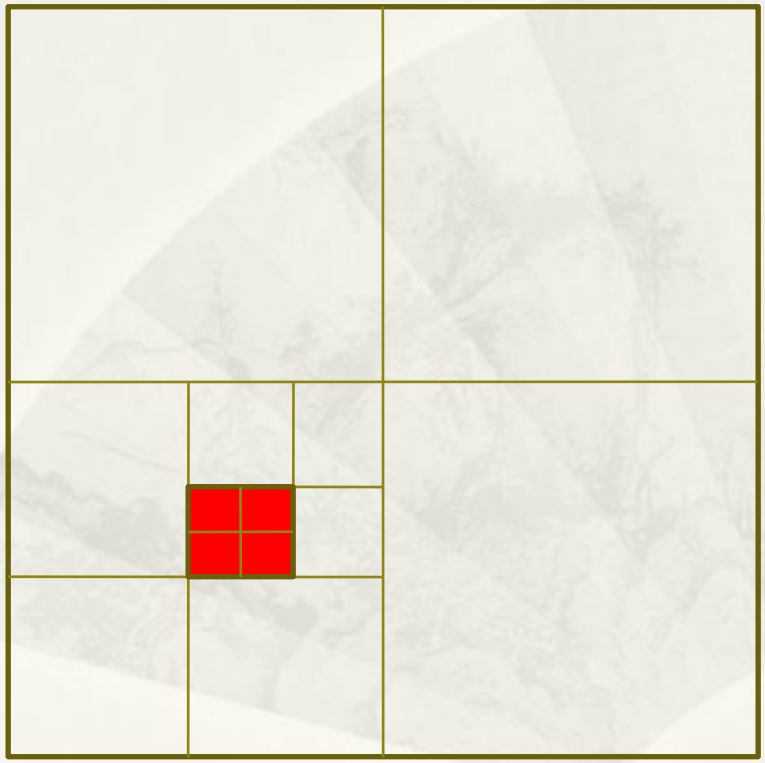
最短描述长度

- 信源发射信号（字母），信号的可能集合为：{a1, a2, . . , an}
- 每一个元素出现的概率{p1, p2, …, pn}
- 用二进制串给这个信源编码，例如：
  a1→00, a2→01, a3→10, a4→11
- 但是这种编码不经济（描述该信源的一串信号的编码长度比较长），因为它们出现的概率不相等。
- 应该让每个字母编码的长度正比于它出现的概率：li=-log(pi)
- 最短的编码长度就是Shannon熵
- 参见P8页，以及第5章

如何理解Shannon熵的定义

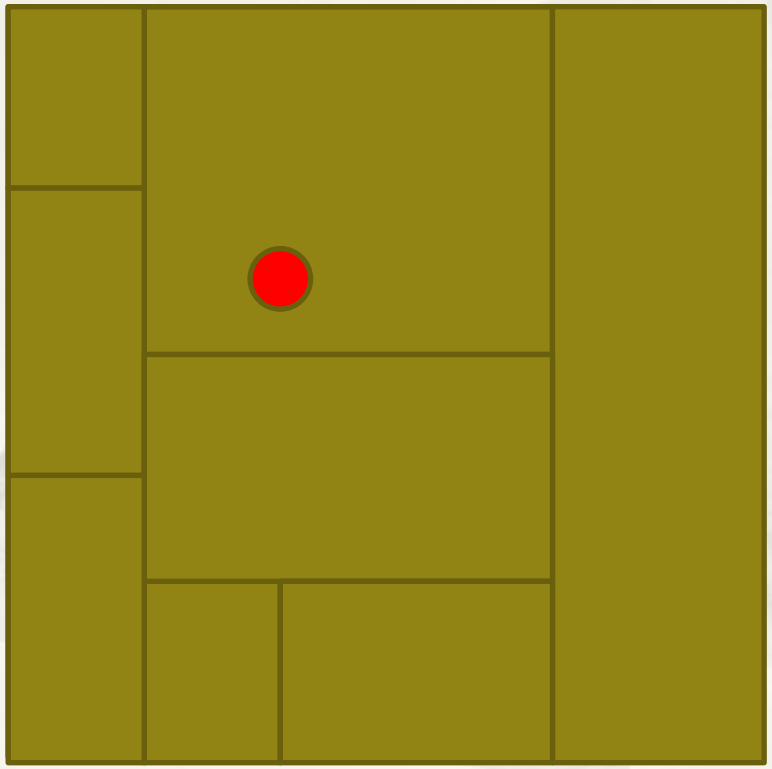
- 21问游戏
- 平均问问题的次数就是这个游戏所包含的信息量

21问的简化版本——找到红色方框



- 将方框按照横平竖直划分为大大小小的方框
- 将其中一个面积为Ai的方框染成红色
- 要找到Ai属于哪一个子集需要问大概log(Ai)个划分的问题
- 设某个方块被选中成为红色方块的概率正比于Ai
- 则对于这个游戏，你平均需要问∑Ailog(Ai)个问题

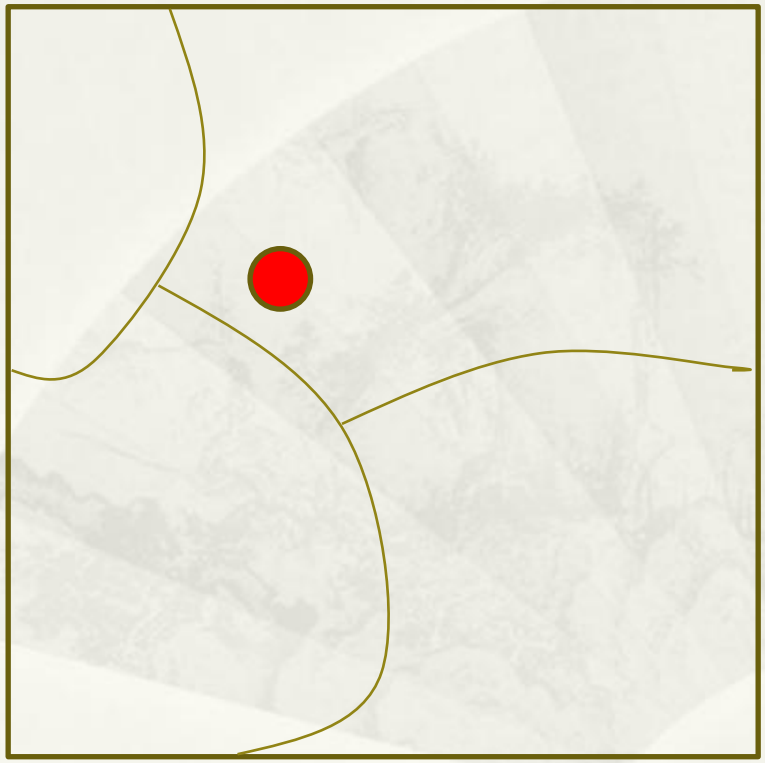
Renyi版本的熵游戏



- 在盒子中寻找一个未知的红点
- 玩家只能按照固定的面积（概率）p1,p2,…,pk把盒子中的空间进行任意的划分
- 庄家告诉玩家红点在哪一个集合之中
- 庄家可以多次对盒子进行划分，直到精确找到红点为止。

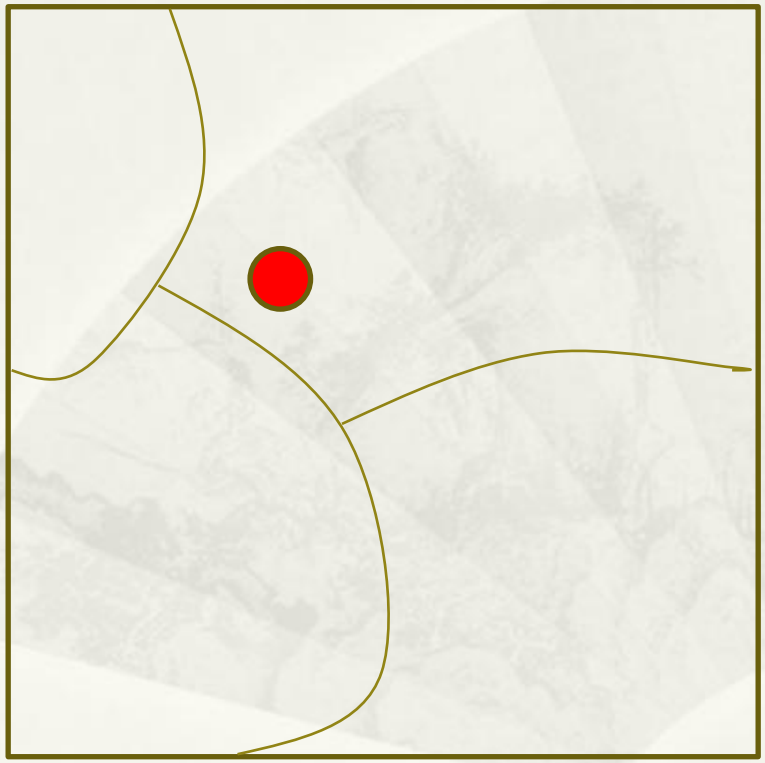
A. Rényi: On the Foundations of Information Theory, Review of the International Statistical Institute, Vol. 33, No. 1 (1965), pp. 1-14

Renyi版本的熵游戏



- 在盒子中寻找一个未知的红点
- 玩家只能按照固定的面积（概率）p1, p2, …, pk把盒子中的空间进行任意的划分
- 庄家告诉玩家红点在哪一个集合之中
- 庄家可以多次对盒子进行划分，直到精确找到红点为止。

Renyi版本的熵游戏



- 假设平均需要进行d次游戏就能找到红点，那么
- Log(A)/d
- 就是Shannon熵：

$$\frac{\log A}{d} = H = -\sum_i p_i \log p_i$$

- 其中A是整个盒子的面积

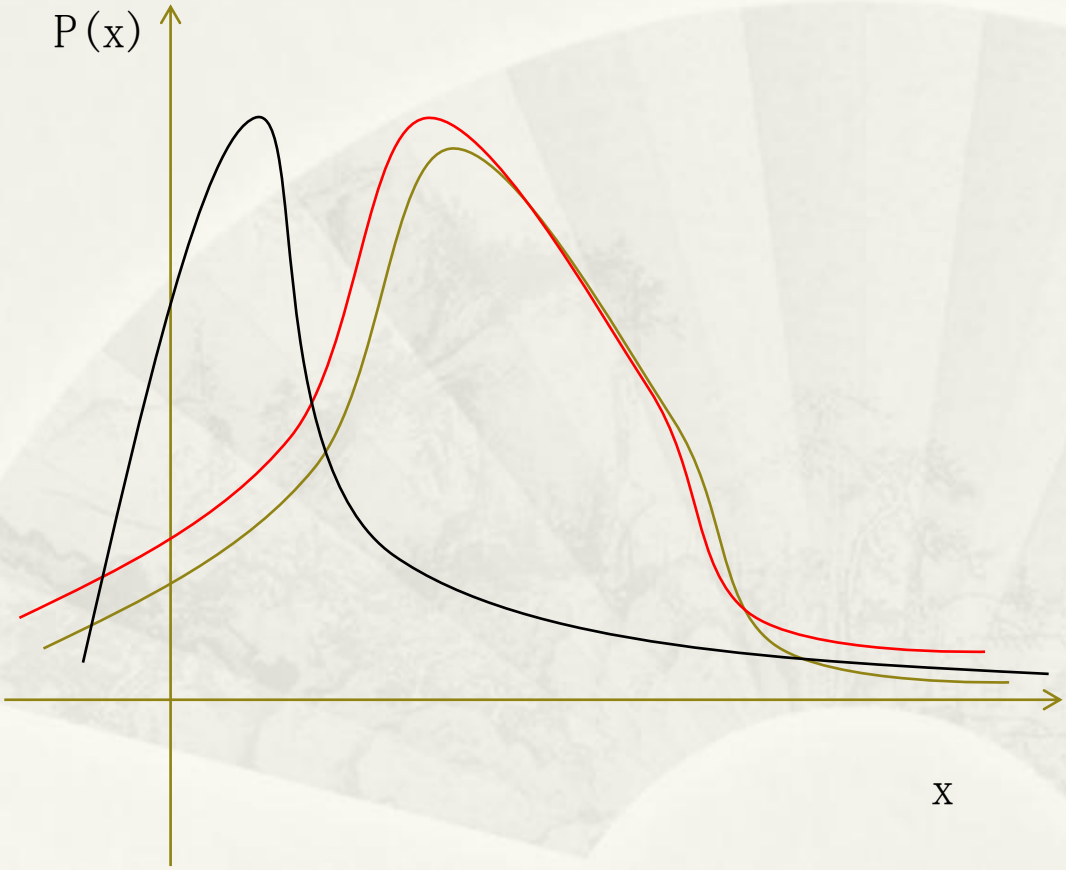
联合熵与条件熵

	概率论	信息论（熵）
单变量分布	$p(x)$	$H(X)=-\sum p(x)\log p(x)$
双变量联合分布	$p(x,y)$	$H(X,Y)=-\sum_{x,y} p(x,y)\log p(x,y)$
条件概率分布	$p(x\|Y=y)$	$H(X\|Y)=\sum_y p(y)H(X\|Y=y)=\sum_{x,y} p(x,y)\log p(x\|y)$

例题：已知联合分布$p(X,Y)$，计算$H(X), H(X,Y), H(X|Y=1), H(X|Y)$

Y \ X	1	2	3	4
1	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{32}$
2	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{32}$	$\frac{1}{32}$
3	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$
4	$\frac{1}{4}$	0	0	0

两个概率分布之间的相对熵



* Kullback-Leibler距离：

$$D(p\|q)=\sum_x p(x)\log\frac{p(x)}{q(x)}$$

* 思考：
* D(p||q)为何不满足三角不等式

互信息

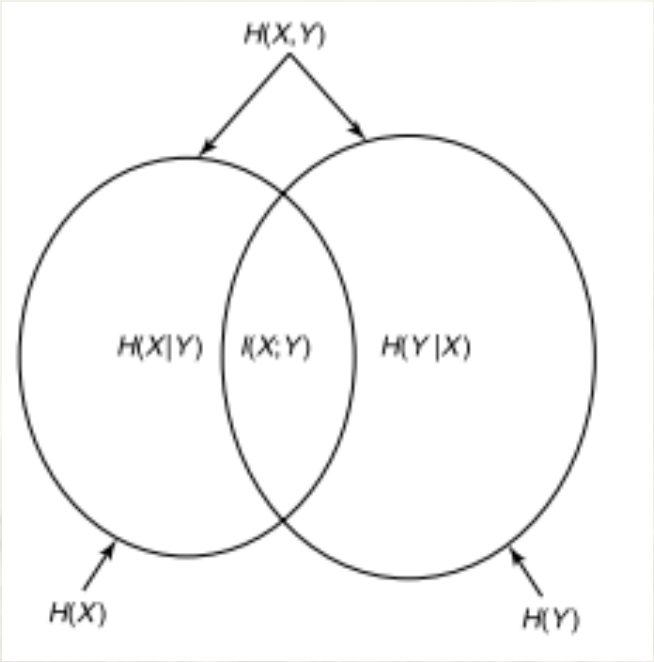
- 经典概率论中，我们用相关系数来刻画两个随机变量之间相互独立的程度

$$\rho = \frac{E((X - E(X))(Y - E(Y)))}{\sqrt{E((X - E(X))^2}\sqrt{E((Y - E(Y))^2}}$$

- 在信息论中，我们用实际的联合分布p(X,Y)与他们相互独立的时候两个变量的联合分布的Kullback-Leibler距离来刻画相互独立程度，即X与Y的互信息

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

信息论文氏图



看图，可以证明如下等式：

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

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

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

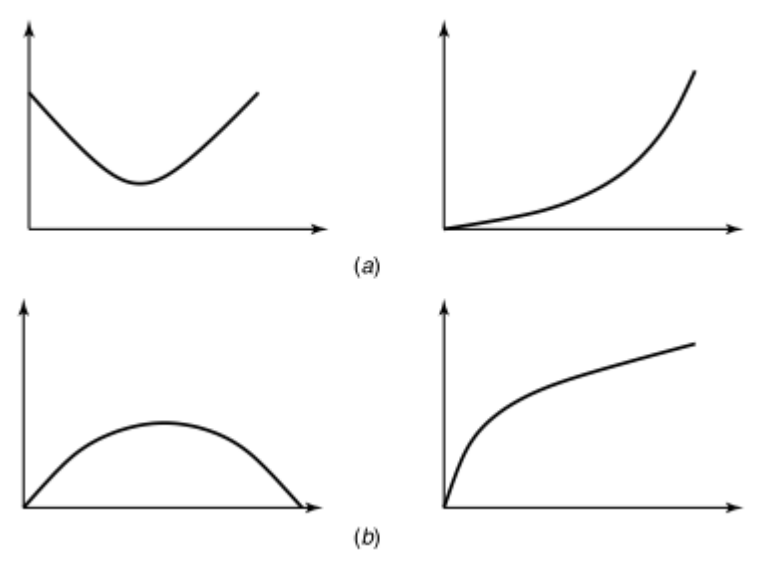
试证明如下更复杂的等式：

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

$$I(X_1, X_2, \ldots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y|X_{i-1}, X_{i-2}, \ldots, X_1).$$

高等数学——函数的凹凸性

凸函数(convex)



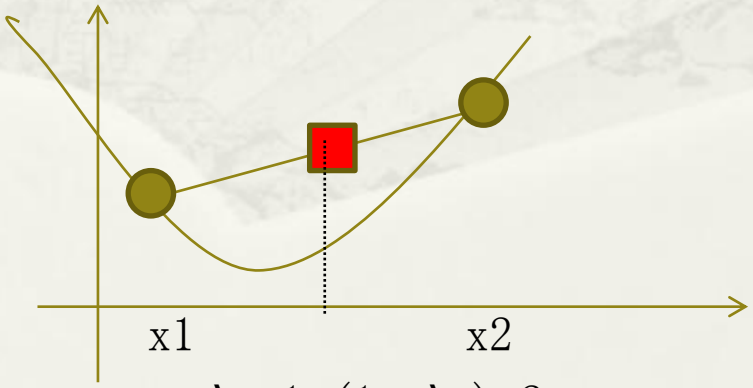
凹函数(concave)

函数凸性定义：对于任意两点x1, x2，以及任意[0, 1]内的实数λ ，如果函数f满足：

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

则，f(x)为凸函数

定理：如果f''(x)>0，则f凸



Jensen不等式

* 如果f(x)为凸函数，则有下列不等式：

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

* 由此可以证明D(p||q)>=0

$$\begin{aligned}
-D(p||q) &= -\sum_{x\in A} p(x)\log\frac{p(x)}{q(x)} \\
&= \sum_{x\in A} p(x)\log\frac{q(x)}{p(x)} \\
&\leq \log\sum_{x\in A} p(x)\frac{q(x)}{p(x)} \\
&= \log\sum_{x\in A} q(x) \\
&\leq \log\sum_{x\in \mathcal{X}} q(x) \\
&= \log 1 \\
&= 0,
\end{aligned}$$

马尔科夫链

- 定义：设三个随机变量X, Y, Z，如果它们的联合分布满足下列等式，则称它们形成了马尔可夫链X→Y→Z

$$p(x, y, z) = p(x)p(y|x)p(z|y).$$

- 注意，一般地，按照贝叶斯等式：

$$p(x,y,z) = p(x)p(y,z \mid x) = p(x)p(y \mid x)p(z \mid x,y)$$

- 这就意味着：

$$p(z \mid x,y) = p(z \mid y)$$

数据处理不等式

- 如果X→Y→Z，那么
- I(X;Y)>=I(X;Z)
- 推论：如果Z=f(Y)，即Z这个随机变量是Y的某一个确定性的函数，那么：X→Y→f(Y)
- 所以：I(X;Y)>=I(X;f(Y))
- 推论：加条件使得互信息减少

Corollary *If* $X \to Y \to Z$, *then* $I(X;Y|Z) \leq I(X;Y)$.

重点重述

- 熵与其它学科之间的关系；
- Shannon熵的定义及理解
- 信息论的文氏图