Maximum Entropy, Bayesian Inference and Evolution

Wu Lingfei 10/06/2009
Clustering Intelligence Club

乡愁

小时候，乡愁是一条命定的进化之路
猴子在那头，我在这头

而现在，乡愁是一条涌现的贝叶斯之河
上帝在那头，我在这头

计算一批离散变量的数据信息熵（以下简称数据熵）$S=-\sum p_i \log_2 p_i$

其中 p_i 代表在集合中随机取一个个体，具有标志值 i 的概率= $\frac{n_i}{N}$

对于连续变量，则数据熵公式变为 $S=-\int_a^b f(x)\ln f(x)dx$（1），$f(x)$ 代表相对密度分布函数

最大熵原理是指积分（1）总是达到最大，在这个条件下，利用拉格朗日方法可以求我们还不知道的 $f(x)$

1. 均匀分布

唯一的约束是 $1=\int_a^b f(x)dx$（2）

依照拉格朗日方法，将式（2）乘以未知常数 C_1，加上（1），构造出

$$F=-\int_a^b f(x)\ln f(x)dx+C_1\left[\int_a^b f(x)dx-1\right]$$

令 F 对 f 的偏微商=0（改变函数 f 的形状但有不变的 x 使 F 极大。就是所谓求泛函数的极值，即变分）

得到 $\ln f(x)=C_1-1$

因此，$f(x)$是一个常数，即均匀分布。且利用（2）可得

$$f=\begin{cases}\dfrac{1}{b-a}, a\le x\le b\\ 0, x<a\cup x>b\end{cases}$$

2.负指数分布

约束有两个，分别是

$1=\int_a^b f(x)dx$（2）和$u=\int_a^b xf(x)dx$（3）（x 算术平均值不变）

依照拉格朗日方法，将式（2）乘以未知常数C_1，将式（3）乘以未知常数C_2，加上（1），构造出

$$F=-\int_a^b f(x)\ln f(x)dx+C_1\left[\int_a^b f(x)dx-1\right]+C_2\left[\int_a^b xf(x)dx-a\right]$$

令 F 对 f 的偏微商=0，

得到

$\ln f(x)=-1+C_1+C_2x$，即指数分布，且利用（2）、（3）可得

$$f(x)=\frac{1}{u}e^{-\frac{x}{u}}$$

3.幂律分布

约束条件有两个，分别是

$1=\int_a^b f(x)dx$（2）和$u=\int_a^b f(x)\ln xdx$（4）（x 几何平均值不变）

依照拉格朗日方法，将式（2）乘以未知常数C_1，将式（4）乘以未知常数C_2，加上（1），构造出

$$F=-\int_a^b f(x)\ln f(x)dx+C_1\left[\int_a^b f(x)dx-1\right]+C_2\left[\int_a^b f(x)\ln xdx-u\right]$$

令 F 对 f 的偏微商=0，

得到

$\ln f(x)=-1+C_1+C_2\ln x=\left[\exp(-1+C_1)\right]x^{C_2}$,即幂律分布

在$1\le x\le b$的情况下，

$$f(x)=\frac{1}{\ln u}x^{-1-\frac{1}{\ln u}}$$

如果 u 很大，$-1-\frac{1}{\ln u}$就接近于-1，此时 f 和 x 的乘积是常数，也就是 f 和 x 是双曲线关系，又称 Zipf 律，与词频和分形等研究有关

4.一个推导分布的通用函数

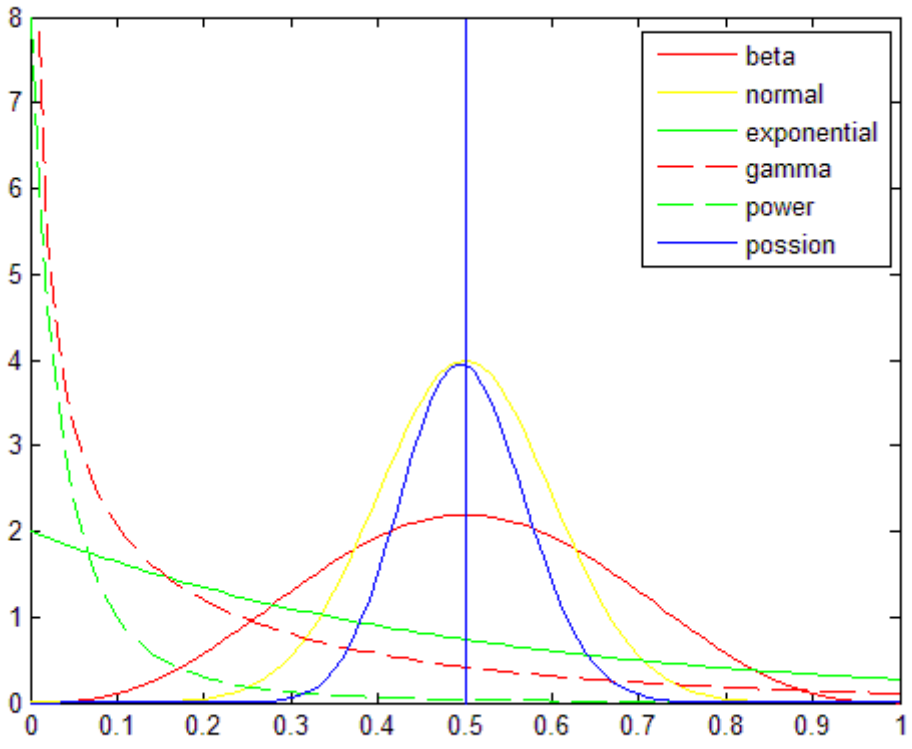
我们根据这几次使用拉格朗日方法推导的经验，可以总结出满足最大熵的相对密度分布函数为

$$f(x)=\exp\left[-1+\sum C_iu_i(x)\right]$$

其中C_i是第 i 个未知常数，$u_i(x)$为第 i 个已经知道的函数，且该函数与分布函数的乘积的积分为常数k_i

约束：$k_i = \int u_i(x) f(x) dx$，　$i = 1$，2，…，m

Distribution name	**Constraints (other than $1 = \int_0^\infty f(x)dx$)**	**Formula**	**Notes**
Uniform distribution	$f(x) = \frac{1}{b-a}$		
Normal distribution	$\sigma^2 = \int_{-\infty}^{\infty} (x-a)^2 f(x)dx$，	$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-a)^2}{2\sigma^2}\right]$	
Lognormal distribution	$a = \int_0^\infty (\ln x) f(x) dx$	$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(\ln x - a)^2}{2\sigma^2}\right]$	
	$\sigma^2 = \int_0^\infty (\ln x - a)^2 f(x) dx$		
γ distribution	$u = \int_0^\infty x f(x) dx$	$f(x) = \frac{\beta^n}{(n-1)!} x^{n-1} e^{-\beta x}$	
	$u = \int_0^\infty \ln x f(x) dx$		
β distribution	$u = \int_0^1 \ln x f(x) dx$	$f(x) = \frac{1}{B(p,q)} x^{p-1}(1-x)^{q-1}$	
	$v = \int_0^1 \ln(1-x) f(x) dx$，		
Weibull distribution	$u = \int_0^\infty x^n f(x) dx$，	$f(x) = \frac{n}{u} x^{n-1} \exp\left[-\frac{x^n}{u}\right]$	
	$v = \int_0^\infty \ln x(x) dx$		
	(Negative exponential distribution (n=1 in Weibull)	$f(x) = \frac{1}{a} \exp\left[-\frac{x}{a}\right]$	
	Rayleigh distribution	(n=2 in Weibull)	
$f(x) = \frac{2(x-a)}{u} \exp\left[-\frac{(x-a)^2}{u}\right]$)			
Power law distribution	$u = \int_0^\infty \ln x f(x) dx$	$f(x) = \frac{1}{u} x^{-1-\frac{1}{u}}$	


beta
normal
exponential
gamma
power
possion
0
1
2
3
4
5
6
7
8
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

在旭东的帮助下，我们终于把各类分布（J 熵？）画到一个图上了（possion 本来是离散的，为便于观察，在这里显示为连续的太。此外，抽象不好，我们可以把这个图就看做是网络度分布的图。不过这个图不精确，因为这些分布的参数是我们自己调的）。中间那个 x=0.5 就是 J 熵最小（正如和尚所说，这也是 S 熵最大的情况，即在“原来”的以人为 X 轴的图里是一条水平线）的情况，平行于 X 轴的线是 J 熵最大的情况。所有的分布都应该是这一横一竖“夹”出来的。

为何不在这个图中画出一条 Y=1 的线呢？我认为，在这个图里，越“平”的线，熵越大，而这些分布与 X 轴夹的面积大小，可以用来衡量“平”的程度。所以 X 轴可以用来代替 Y=1。

旭东提出，其实各种分布的熵究竟是多少，一算就知道。但我们当时没有精确计算。同时我回忆起《组成论》随想录里，冯向军为张学文老师做的一个分析证明，方差和熵是正比的，我找到了这个模拟图：

那这么说的话有如下命题待证

1.正态分布的熵和泊松分布的熵比指数和幂律小，因为前者靠近“竖线”，后者靠近“横线”

2.同理，正态分布的熵和泊松分布的方差比指数和幂律小。问题在于在我们画的图里，方差可是自己调的。但我仍然坚持这个判断。并且我认为这个和 Scale free 的问题相关。

正态分布有一个均值，按照 barabasi 说法，以网络为例，有 charactor node 可以代表大多数点，因此自然方差较小，而幂律这样 scale free 的，几乎每个 node 都在不同的量级上，没有什么可以代表大部分点的 node，方差自然大。

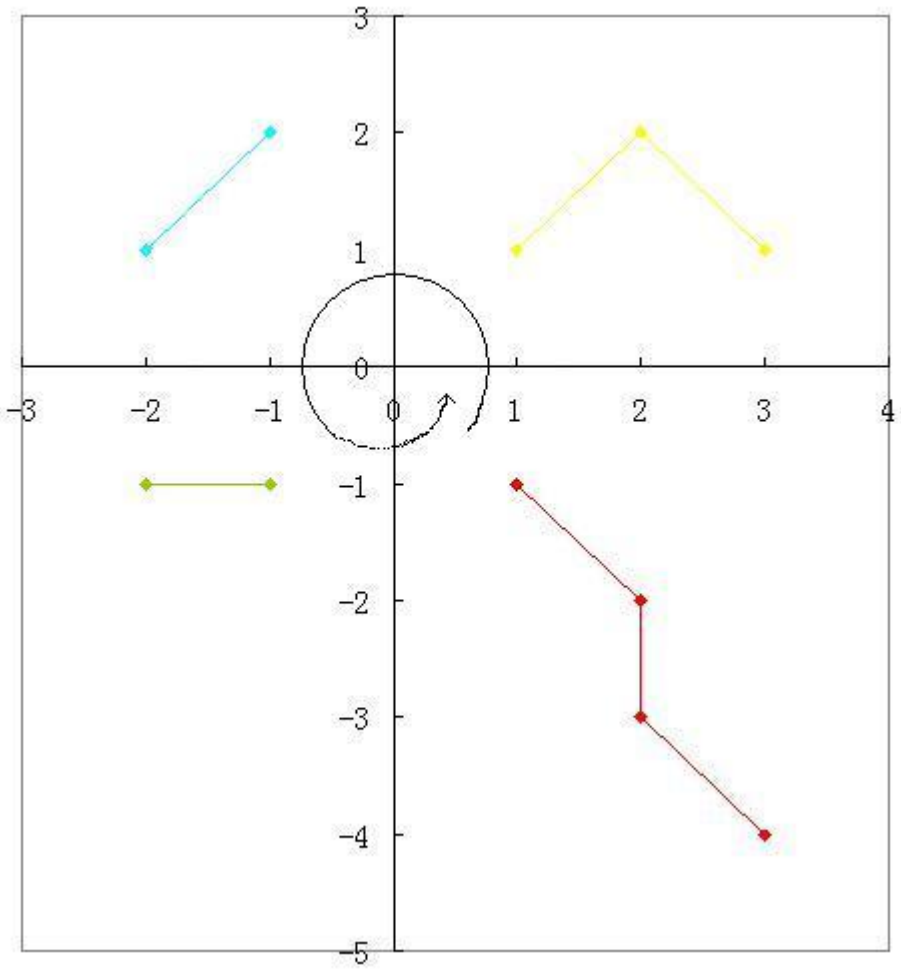
（附冯对 MSD 的定义：

假设广义集合仅有 2 个标志值: v1 = 1; v2 = 2. 令 v1,v2 的概率分别为 p1, p2, 则有标志值均值为
M = p1v1 + p2v2 (1)

我们定义标志值相对其均值的均方距 MSD 为标志值分散度.则有
MSD = p1(v1 - M)2 + p2 (v2 - M)2 (2)

实际上， 倘若视标志值为随机变量，则 MSD 即为此标志值随机变量的方差。）

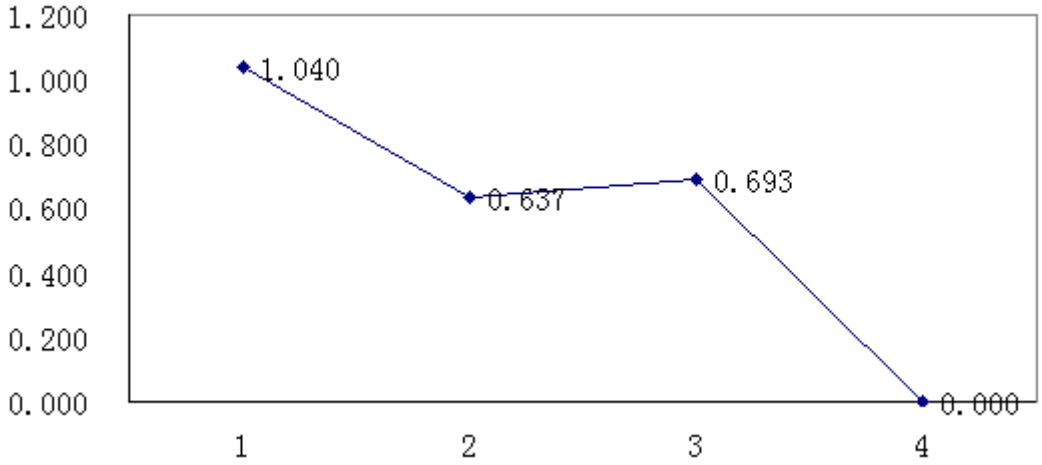
可不可以这样看多重熵，就是一个直角坐标系，比如原来的节点是 X 轴，度是 Y 轴的，把它顺时针转 90°，得到的新的 X '轴，也就是原来的 Y 轴，是度，而此时的 Y'轴，是去数（离散的情况下）原来那个曲线在每个 Y 上有多少个 X，也就是有这个度的节点有几个（旭东说这个叫测度）。这样，每次顺时针旋转 90°，就得到越来越小的熵，最后熵降低为 0，所有的系统差异导致的无序都被忽视了。

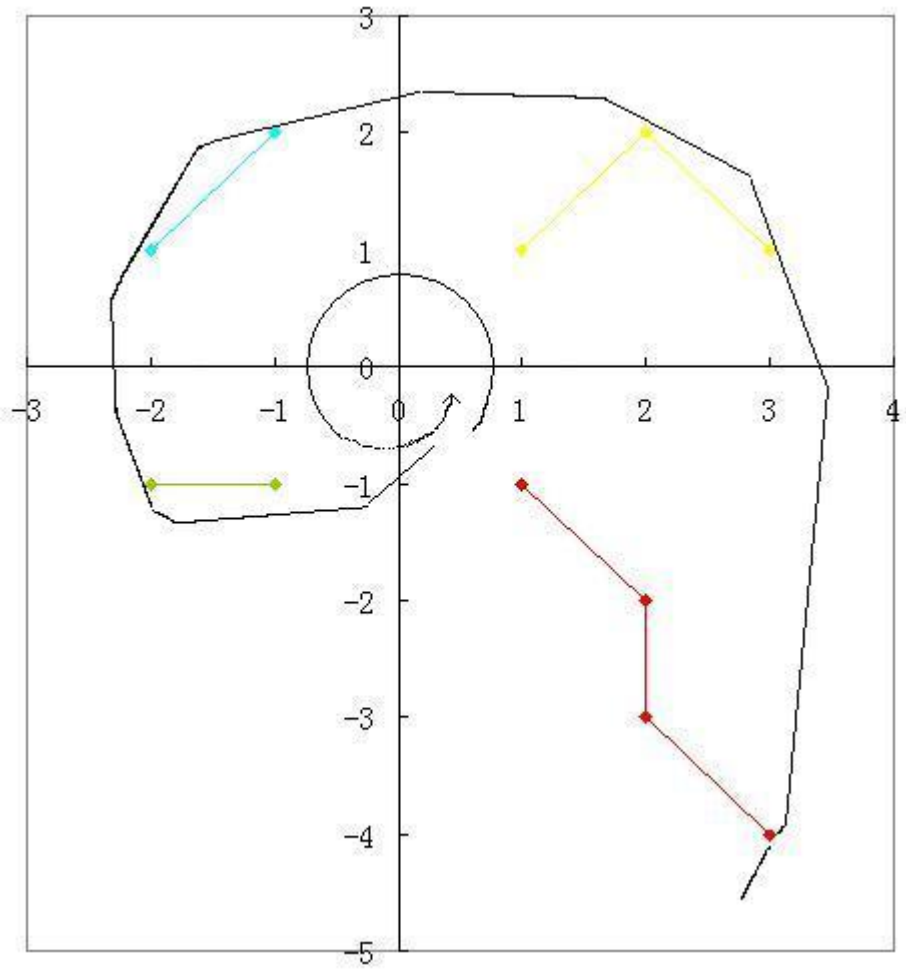


上图中，从右下角到左下角，不断进行旋转和“投影”，熵不断减小（奇怪的是，第三步比二步熵增加了，是不是我算错了？）

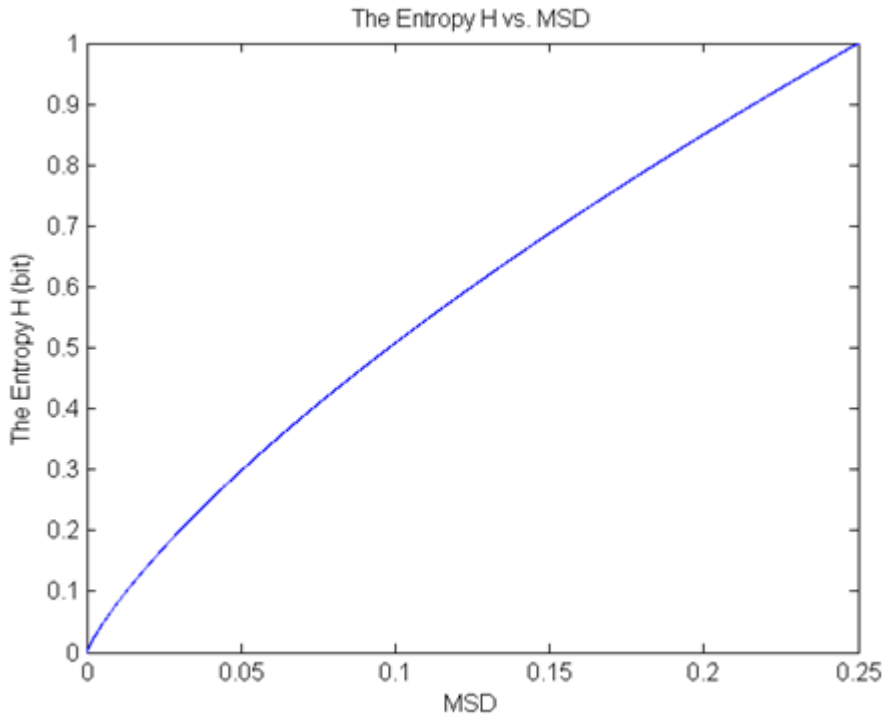
旭东指出一个很重要的问题，在上述坐标系中，如果是离散变量自然没关系，但如果是连续概率密度函数，比如一个正态分布函数，可能顺时针转 90 °后，除了均值点，在每个 Y 上都只有“两个”X，如果是指数，则在每个 Y 上都只有“一个”X，那怎么办？而对于那种一会儿是水平线一会儿是曲线的分布，这么搞的话有的 Y 对应无数个 X，有的 Y 对于一个 X，又怎么办？我的提议是将连续变量最小单位规定后分解为离散变量，但好像并不能完全解决问题。

	x	y	S熵		S熵的值
第一步：右下方	1	-1	[LN(2)+LN(4)]/2		1.039721
	2	-2			
	2	-3			
	3	-4			
第二步：右上方	1	1			
	2	2	[2LN(3/2)+LN3]/3		0.636514
	3	1			
第三步：左上方	-2	1			0.693147
	-1	2	LN(2)		
	-2	-1			
第四步：左下方	-1	-1	LN(1)		0





我有一个奇妙的设想，这个图好像一个大海螺（如果我们用的是连续分布会不会更像？）o(∩_∩)o...，万物的信息就这样被不断精简，最后归零 ····



图六：熵 H 与标志值分散度 MSD 的关系.

贝叶斯公式的本质是什么？

后验分布 $\propto$ 似然函数*先验分布

Min|数据熵-信息熵|的过程，就是用似然函数（对样本数据进行加工）不断修改先验分布（源于历史知识积累），取得后验分布去逼近真实分布的过程。

我有一个比较好玩的想法：区分“数据熵”和“信念熵”（我原来叫“信息熵”，旭东说容易乱，不如结合贝叶斯叫信念熵），贝叶斯判断，就是一个不断修改自己的信念分布（熵），去抓真实分布（数据熵）的过程。**一开始，在没有任何先验知识的情况下，肯定是无约束最大熵分布，也就是均匀分布，慢慢根据知识去修改自己的分布。这个过程的难度还要取决于数据熵，如果数据熵是 0，一个缸里装的全是黑石头，那我抓几次就明白了，如果石头颜色各不相同，那抓好多次也不明白。**

咱们可以用一个比喻的提法：在前文那个好多分布的图里，X 轴就是风平浪静的大海，就是最大熵先验分布，后来起了风，就出现了各种分布，仿佛波浪。风从何来？也许是从心而来，正是我心制造的后验分布啊。

我认为，张学文推导方法里，幂律分布那个几何平均值的不变的约束，或者说 Sigmaln（x）=m 的约束，其实应该也等价于最大熵（公式的形式是类似的），究竟怎么等价，还有待证明。

做一点哲学思考，如果这个 m 指的是“心智内存”的上限呢？