# 最大熵模型：读书笔记

胡江堂，北京大学软件与微电子学院，Jthu@pku.edu.cn

1. 物理学的熵 ..... 2
2. 信息论的熵 ..... 2
3. 熵和主观概率（一个简单注释） ..... 3
4. 熵的性质 ..... 3
4.1. 当所有概率相等时，熵取得最大值 ..... 3
4.2. 小概率事件发生时携带的信息量比大概率事件发生时携带的信息量多 ..... 3
5. 最大熵原理：直觉讨论 ..... 4
6. 最大熵原理：一个手工例子 ..... 4
7. 最大熵原理：正式表述 ..... 5
8. 最大熵模型的训练：GIS 算法 ..... 6
9. 最大熵模型：金融领域内的应用 ..... 7
参考文献 ..... 7

# 最大熵模型：读书笔记

胡江堂，北京大学软件学院，Jthu@pku.edu.cn

## 1. 物理学的熵

熵是一个物理学概念，它是描述事物无序性的参数，熵越大则无序性越强。从宏观方面讲（根据热力学定律），一个体系的熵等于其可逆过程吸收或耗散的热量除以它的绝对温度；从微观讲，熵是大量微观粒子的位置和速度的分布概率的函数。自然界的一个基本规律就是熵递增原理，即，一个孤立系统的熵，自发性地趋于极大，随着熵的增加，有序状态逐步变为混沌状态，不可能自发地产生新的有序结构，这意味着自然界越变越无序。

## 2. 信息论的熵

在物理学中，熵是描述客观事物无序性的参数。信息论的开创者香农认为，信息（知识）是人们对事物了解的不确定性的消除或减少。他把不确定的程度称为信息熵。假设每种可能的状态 $A$ 都有概率 $p(A_i)$，我们用关于被占据状态的未知信息来量化不确定性，这个信息熵 $S$ 即为：

$$S=\sum_i p(A_i)\log\left(\frac{1}{p(A_i)}\right)=-\sum_i p(A_i)\log\left(p(A_i)\right)$$

其中 $\log(\bullet)$ 是以 2 为底的对数，所以这个信息用位衡量。前面说过，在物理学的背景下，这个不确定性被称为熵（在通讯系统中，关于传输的实际信息的不确定性也被称为数据源的熵）。

扩展到连续情形。假设连续变量 $X$ 的概率密度函数是 $f_x(x)$，与离散随机变量的熵的定义类似，信息熵的连续定义为：

$$S(X)=-\int_{-\infty}^{\infty} f_x(x)\log f_x(x)dx=-E[\log f_x(x)]$$

上式 $S(X)$ 就是我们定义的随机变量 $X$ 的微分熵。当 $X$ 被解释为一个随机连续向量时，$f_x(x)$ 就是 $X$ 的联合概率密度函数。

3. 熵和主观概率（一个简单注释）

因为熵用概率表示，所以这涉及到主观概率。概率用于处理知识的缺乏（概率值为 1 表明对知识的完全掌握，这就不需要概率了），而一个人可能比另一个人有着更多的知识，所以两个观察者可能会使用不同的概率分布，也就是说，概率（以及所有基于概率的物理量）都是主观的。在现代的主流概率论教材中，都采用这种主观概率的处理方法。

4. 熵的性质

4.1. 当所有概率相等时，熵取得最大值

上面关于熵的公式有一个性质：假设可能状态的数量有限，当所有概率相等时，熵取得最大值。证明如下：

假设一个有限集$\{1,2,...\mathrm{n}\}$下的一个概率分布$p=\{p_1,p_2,...p_n\}$,

我们要最大化$S(p)$,约束条件为$\sum_{i=1}^{n} p_i=1$

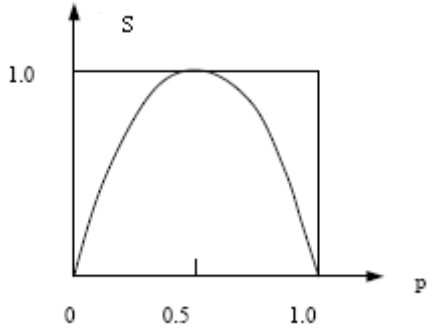
采用拉格朗日方法，有

$$L(p)=-\sum_{i=1}^{n} p_i\log(p_i)-\delta(\sum_{i=1}^{n} p_i-1)$$

$$\frac{\partial L}{\partial p_i}=-\log(ep_i)-\delta=0$$

$$\Rightarrow p_i=\frac{1}{n}$$

在只有两个状态的例子中，要使熵最大，每个状态发生的概率都是 1/2，如下图所示：




4.2. 小概率事件发生时携带的信息量比大概率事件发生时携带的信息量多

证明略，可以简要说明一下，也挺直观的。如果事件 $A$ 发生的概率为 1，在这种情况下，事件 $A$ 发生就没有什么“惊奇”了，并且不传达任何“信息”，因为我们已经知道这“信息”是什么，没有任何的“不确定”；反之，如果事件 $A$ 发生的概率很小，这就有更大的“惊奇”和有“信息”了。这里，“不确定”、“惊奇”和“信息”是相关的，信息量与事件发生的概率成反比。

## 5. 最大熵原理：直觉讨论

最大熵原理是根据样本信息对某个未知分布做出推断的一种方法。日常生活中，很多事情的发生表现出一定的随机性，试验的结果往往是不确定的，而且也不知道这个随机现象所服从的概率分布，所有的只有一些试验样本或样本特征，统计学常常关心的一个问题，在这种情况下如何对分布作出一个合理的推断？最大熵采取的原则就是：保留全部的不确定性，将风险降到最小。在金融理论中，一个类似的教训是，为了降低风险，投资应该多样化，不要把所有的鸡蛋都放在一个篮子里。

吴军（2006）举了一个例子。对一个均匀的骰子，问它每个面朝上的概率分别是多少。所有人都会说是 1/6。这种“猜测”当然是对的，因为对这个“一无所知”的色子，假定它每一个朝上概率均等是最安全的做法，你不应该假设它被做了手脚。从信息论的角度讲，就是保留了最大的不确定性，让熵达到最大（从投资的角度来看，这就是风险最小的做法）。但是，如果这个骰子被灌过铅，已知四点朝上的概率是 1/3，在这种情况下，每个面朝上的概率是多少？当然，根据简单的条件概率计算，除去四点的概率是 1/3 外，其余的概率都是 2/15。也就是说，除已知的条件（四点概率为 1/3）必须满足外，对其它各点的概率，我们仍然无从知道，也只好认为它们相等。这种基于直觉的猜测之所以准确，是因为它恰好符合了最大熵原理。

回到物理学例子中。在涉及物理系统的情形中，一般要确定该系统可以存在的多种状态，需要了解约束下的所有参数。比如能量、电荷和其他与每个状态相关的物理量都假设为已知。为了完成这个任务常常需要量子力学。我们不假设在这个步骤系统处于特定状态；事实上我们假定我们不知道也不可能知道这一点，所以我们反而可以处理被占据的每个状态的概率。这样把概率当作应对知识缺乏的一种方法。我们很自然地想避免假定了比我们实际有的更多的知识，最大熵原理就是完成这个的方法。

这里可以总结出最大熵对待已知事物和未知事物的原则：承认已知事物（知识）；对未知事物不做任何假设，没有任何偏见。最大熵原理指出，当我们需要对一个随机事件的概率分布进行预测时，我们的预测应当满足全部已知的条件，而对未知的情况不要做任何主观假设（不做主观假设，这点很重要。）在这种情况下，概率分布最均匀，预测的风险最小。因为这时概率分布的信息熵最大，所以人们称这种模型叫“最大熵模型”。我们常说，不要把所有的鸡蛋放在一个篮子里，其实就是最大熵原理的一个朴素的说法，因为当我们遇到不确定性时，就要保留各种可能性。

## 6. 最大熵原理：一个手工例子

举个例子，一个快餐店提供 3 种食品：汉堡(B)、鸡肉(C)、鱼(F)。价格分别是 1 元、2 元、3 元。已知人们在这家店的平均消费是 1.75 元，求顾客购买这 3 种食品的概率。如果你假设一半人买鱼另一半人买鸡肉，那么根据熵公式，这不确定性就是 1 位（熵等于 1）。但是这个假设很不合适，因为它超过了你所知道的事情。我们已知的信息是：

$$p(B)+p(C)+p(F)=1$$

$$1p(B)+2p(C)+3p(F)=1.75$$

以及关于对概率分布的不确定性度量，熵：

$$S=-p(B)\log(p(B))-p(C)\log(p(C))-p(F)\log(p(F))$$

对前两个约束，两个未知概率可以由第三个量来表示，可以得到：

$$p(C) = 0.75 - 2p(F)$$

$$p(B) = 0.25 + p(F)$$

把上式代入熵的表达式中，熵就可以用单个概率 $p(F)$ 来表示：

$$S = -(0.25 + p(F))\log((0.25 + p(F))) - (0.75 - 2p(F))\log((0.75 - 2p(F))) - p(F)\log(p(F))$$

对这个单变量优化问题，很容易求出 $p(F) = 0.216$ 时熵最大，有 $p(B) = 0.466$，

$p(C) = 0.318$ 和 $S = 1.517$。

总结一下。以上，我们根据未知的概率分布表示了约束条件，又用这些约束条件消去了两个变量，用剩下的变量表示熵，最后求出了熵最大时剩余变量的值，结果就求出了一个符合约束条件的概率分布，它有最大不确定性，我们在概率估计中没有引入任何偏差。

7. 最大熵原理：正式表述

假设有一个随机系统，已知一组状态，但不知道其概率，而且我们知道这些状态的概率分布的一些限制条件。这些限制条件或者是已知一定的总体平均值，或者是它们的一些界限。在给定关于模型的先验知识的条件下，问题是选择一个在某种意义下最佳的概率模型。Jaynes(1957)提出了一个最大熵原则：当根据不完整的信息作为依据进行推断时，应该由满足分布限制条件的具有最大熵的概率分布推得。也就是说，熵的概念在概率分布空间定义一种度量，使得具有较高熵的分布比其它的分布具有更大的值。显然，“最大熵问题”是一个带约束的最优化问题。

为方便叙述，考虑最大微分熵

$$S(X) = -\int_{-\infty}^{\infty} f_x(x)\log f_x(x)dx$$

对所有随机变量 $X$ 的概率密度函数 $f_x(x)$，满足以下约束条件：

$$1)\ f_x(x) \geq 0, \text{在} x \text{为确定值时等式成立}$$

$$2)\int_{-\infty}^{\infty} f_x(x)dx = 1$$

$$3)\int_{-\infty}^{\infty} f_x(x)g_i(x)dx = \alpha_i, \text{对} i = 1,2,...m$$

其中，$g_i(x)$ 是 $x$ 的一个函数。约束 1 和约束 2 描述的是概率密度函数的基本属性，约束 3 定义变量 $X$ 的矩，它随函数 $g_i(x)$ 的表达式不同而发生变化，它综合了随机变量 $X$ 的所有可用的先验知识。为了解这个约束最优化问题，利用拉格朗日乘子法，目标函数为：

$$L(f)=\int_{-\infty}^{\infty}[-f_x(x)\log f_x(x)+\delta_0 f_x(x)+\sum_{i=1}^{m}\delta_i g_i(x) f_x(x)]dx$$

其中， $\delta_0,\delta_1,...\delta_m$ 是拉格朗日乘子。对被积函数求 $f_x(x)$ 的微分，并令其为 0，有：

$$-1-\log f_x(x)+\delta_0+\sum_{i=1}^{m}\delta_i g_i(x)=0$$

解得：

$$f_x(x)=e^{-1+\delta_0+\sum_{i=1}^{m}\delta_i g_i(x)}$$

我们看到这个概率密度函数具有指数形式。匈牙利数学家 Csiszar 曾经证明，对任何一组不自相矛盾的信息，最大熵模型不仅存在，而且是唯一的。而且它们都有同一个非常简单的形式 -- 指数函数。我们还可以得到，在所有零均值随机向量可达到的微分熵中，多元正态分布具有最大的微分熵。最大熵的解，同时是最吻合样本数据分布的解。

8. 最大熵模型的训练：GIS 算法和其他

上节我们得到，一个最大熵模型可以有效地把各种信息综合在一起（无偏见地对待不确定性），而且具有指数函数的形式，下面模型的训练就要确定这个指数函数的各个参数。最原始的最大熵模型的训练方法是一种称为通用迭代算法 GIS(generalized iterative scaling) 的迭代算法，由 Darroch 和 Ratcliff 在七十年代提出，大致可以概括为以下几个步骤：

1. 假定第零次迭代的初始模型为等概率的均匀分布。
2. 用第 N 次迭代的模型来估算每种信息特征在训练数据中的分布，如果超过了实际的，就把相应的模型参数变小；否则，将它们便大。
3. 重复步骤 2 直到收敛。

Darroch 和 Ratcliff 没有能对这种算法的物理含义进行很好地解释，后来是由 Csiszar 解释清楚的，因此，人们在谈到这个算法 时，总是同时引用 Darroch 和 Ratcliff 以及希萨的两篇论文。GIS 算法每次迭代的时间都很长，需要迭代很多次才能收敛，而且不太稳定，即使在 64 位计算机上都会出现溢出。因此，在实际应用中很少有人真正使用，大家只是通过它来了解最大熵模型的算法。

八十年代，Della Pietra 在 IBM 对 GIS 算法进行了两方面的改进，提出了改进迭代算法 IIS（improved iterative scaling）。这使得最大熵模型的训练时间缩短了一到两个数量级。这样最大熵模型才有可能变得实用。即使如此，在当时也只有 IBM 有条件是用最大熵模型。

由于最大熵模型在数学上十分完美，对科学家们有很大的诱惑力，因此不少研究者试图把自己的问题用一个类似最大熵的近似模型去套。谁知这一近似，最大熵模型就变得不完美了，结果可想而知，比打补丁的凑合的方法也好不了多少。于是，不少热心人又放弃了这种方法。第一个在实际信息处理应用中验证了最大熵模型的优势的，是原 IBM 现微软的研究员 Adwait Ratnaparkhi。Ratnaparkhi 的聪明之处在于他没有对最大熵模型进行近似，而是找到了几个最适合用最大熵模型、而计算量相对不太大的自然语言处理问 题，比如词性标注

和句法分析。拉纳帕提成功地将上下文信息、词性（名词、动词和形容词等）、句子成分（主谓宾）通过最大熵模型结合起来，做出了当时世界上 最好的词性标识系统和句法分析器。

9. 最大熵模型：金融领域内的应用

最大熵模型在自然语言处理领域内得到了广泛的应用，在金融界，也能见到它的影子。当年最早改进最大熵模型算法的 Della Pietra 在九十年代初退出了学术界，而到在金融界大显身手。他和很多 IBM 语音识别的同事一同到了一家当时还不大，但现在是世界上最成功对冲基金公司----(Renaissance Technologies。我们知道，决定股票涨落的因素可能有几十甚至上百种，而最大熵方法恰恰能找到一个同时满足成千上万种不同条件的模型。Della Pietra 等科学家在那里，用于最大熵模型和其他一些先进的数学工具对股票预测，获得了巨大的成功。从该基金 1988 年创立至今，它的净回报率高达平均每年 34%。也就是说，如果 1988 年你在该基金投入一块钱，今天你能得到 200 块钱。这个业绩，远远超过股神巴菲特的旗舰公司 Berkshire Hathaway（同期，Berkshire Hathaway 的总回报是 16 倍）。

参考文献


1. 吴军《数学之美系列十六(上)-不要把所有的鸡蛋放在一个篮子里 -- 谈谈最大熵模型》，http://googlechinablog.com/2006/10/blog-post.html
2. 吴军《数学之美系列十六(下)-不要把所有的鸡蛋放在一个篮子里 -- 谈谈最大熵模型》，http://googlechinablog.com/2006/11/blog-post.html
3. Jaynes, E.T., 1957. "Information Theory and Statistical Mechanics", Physical Review, vol.106, pp.620-630. http://bayes.wustl.edu/etj/articles/theory.1.pdf
4. Haykin, Simon《神经网络原理》（第 10 章 信息论模型，叶世伟等译，北京：机械工业出版社，2004）
5. 王厚峰. 机器学习课程讲义之六 MEM (Maximum Entropy Model).北京大学软件与微电子学院，2007 年春季学期
6. Penfield, Paul. Information and Entrop. MIT Open Course, Spring 2003. http://ocw.mit.edu/OcwWeb/Electrical-Engineering-and-Computer-Science/6-050JInformation-and-EntropySpring2003/CourseHome/index.htm
7. Wei, Xiaoliang《最大熵模型与自然语言处理》 www.cs.caltech.edu/~weixl/research/read/summary/MaxEnt2.ppt
8. 常宝宝《自然语言处理的最大熵模型》 www.icl.pku.cn/WebData_http-dir-listable/ICLseminars/2003spring/最大熵模型.pdf
9. 廖先桃《最大熵理论及其应用》 http://ir.hit.edu.cn/phpwebsite/index.php?module=documents&JAS_DocumentManager_op=downloadFile&JAS_File_id=196