[我们在投资时常常讲不要把所有的鸡蛋放在一个篮子里，这样可以降低风险。在信息处理中，这个原理同样适用。在数学上，这个原理称为最大熵原理(the maximum entropy principle)。这是一个非常有意思的题目，但是把它讲清楚要用两个系列的篇幅。]

前段时间，Google 中国研究院的刘骏总监谈到在网络搜索排名中，用到的信息有上百种。更普遍地讲，在自然语言处理中，我们常常知道各种各样的但是又不完全确定的信息，我们需要用一个统一的模型将这些信息综合起来。如何综合得好，是一门很大的学问。

让 我们看一个拼音转汉字的简单的例子。假如输入的拼音是"wang-xiao-bo"，利用语言模型，根据有限的上下文(比如前两个词)，我们能给出两个最 常见的名字"王小波"和"王晓波"。至于要唯一确定是哪个名字就难了，即使利用较长的上下文也做不到。当然，我们知道如果通篇文章是介绍文学的，作家王小 波的可能性就较大；而在讨论两岸关系时，台湾学者王晓波的可能性会较大。在上面的例子中，我们只需要综合两类不同的信息，即主题信息和上下文信息。虽然有 不少凑合的办法，比如：分成成千上万种的不同的主题单独处理，或者对每种信息的作用加权平均等等，但都不能准确而圆满地解决问题，这样好比以前我们谈到的 行星运动模型中的小圆套大圆打补丁的方法。在很多应用中，我们需要综合几十甚至上百种不同的信息，这种小圆套大圆的方法显然行不通。

数学上最漂亮的办法是最大熵(maximum entropy)模型，它相当于行星运动的椭圆模型。"最大熵"这个名词听起来很深奥，但是它的原理很简单，我们每天都在用。说白了，就是要保留全部的不确定性，将风险降到最小。让我们来看一个实际例子。

有 一次，我去 AT&T 实验室作关于最大熵模型的报告，我带去了一个色子。我问听众"每个面朝上的概率分别是多少"，所有人都说是等概率，即各点的概率均为1/6。这种猜测当然 是对的。我问听众们为什么，得到的回答是一致的：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。（你不应该主观假设它象韦小宝的色 子一样灌了铅。）从投资的角度看，就是风险最小的做法。从信息论的角度讲，就是保留了最大的不确定性，也就是说让熵达到最大。接着，我又告诉听众，我的这 个色子被我特殊处理过，已知四点朝上的概率是三分之一，在这种情况下，每个面朝上的概率是多少？这次，大部分人认为除去四点的概率是 1/3，其余的均是 2/15，也就是说已知的条件（四点概率为 1/3）必须满足，而对其余各点的概率因为仍然无从知道，因此只好认为它们均等。注意，在猜测这两种不同情况下的概率分布时，大家都没有添加任何主观的假 设，诸如四点的反面一定是三点等等。（事实上，有的色子四点反面不是三点而是一点。）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。

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。

回到我们刚才谈到的拼音转 汉字的例子，我们已知两种信息，第一，根据语言模型，wang-xiao-bo 可以被转换成王晓波和王小波；第二，根据主题，王小波是作家，《黄金时代》的作者等等，而王晓波是台湾研究两岸关系的学者。因此，我们就可以建立一个最大 熵模型，同时满足这两种信息。现在的问题是，这样一个模型是否存在。匈牙利著名数学家、信息论最高奖香农奖得主希萨（Csiszar）证明，对任何一组不 自相矛盾的信息，这个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。下面公式是根据

上下文（前两个词）和主题预测下一个词的最大熵模型，其中 w3 是要预测的词（王晓波或者王小波）w1 和 w2 是它的前两个字（比如说它们分别是"出版"，和""），也就是其上下文的一个大致估计，subject 表示主题。

$$P(w_3 \mid w_1, w_2, subject) = \frac{e^{\{\lambda_1(w_1,w_2,w_3)+\lambda_2(subject,w_3)\}}}{Z(w_1, w_2, subject)}$$

我们看到，在上面的公式中，有几个参数 lambda 和 Z ，他们需要通过观测数据训练出来。

最大熵模型在形式上是最漂亮的统计模型，而在实现上是最复杂的模型之一。我们在将下一个系列中介绍如何训练最大熵模型的诸多参数，以及最大熵模型在自然语言处理和金融方面很多有趣的应用。

我们上次谈到用最大熵模型可以将各种信息综合在一起。我们留下一个问题没有回答，就是如何构造最大熵模型。我们已经所有的最大熵模型都是指数函数的形式，现在只需要确定指数函数的参数就可以了，这个过程称为模型的训练。

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。GIS 的原理并不复杂，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。

GIS 最早是由 Darroch 和 Ratcliff 在七十年代提出的。但是，这两人没有能对这种算法的物理含义进行很好地解释。后来是由数学家希萨（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 GIS。大家只是通过它来了解最大熵模型的算法。

八十年代，很有天才的孪生兄弟的达拉皮垂(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，是宾夕法尼亚大学马库斯的另一个高徒原 IBM 现微软的研究员拉纳帕提(Adwait Ratnaparkhi)。拉纳帕提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。拉纳帕提的论文发表后让人们耳目一新。拉纳帕提的词性标注系统，至今仍然是使用单一方法最好的系统。科学家们从拉纳帕提 的成就中，又看到了用最大熵模型解决复杂的文字信息处理的希望。

但是，最大熵模型的计算量仍然是个拦路虎。我在学校时花了很长时间考虑如 何简化最大熵模型的计算量。终于有一天，我对我的导师说，我发现一种数学变换，可以将大部分最大熵模型的训练时间在 IIS 的基础上减少两个数量级。我在黑板上推导了一个多小时，他没有找出我的推导中的任何破绽，接着他又回去想了两天，然后告诉我我的算法是对的。从此，我们就 建造了一些很大的最大熵模型。这些模型比修修补补的凑合的方法好不少。即使在我找到了快速训练算法以后，为了训练一个包含上下文信息，主题信息和语法信息 的文法模型(language model)，我并行使用了 20 台当时最快的 SUN 工作站，仍然计算了三个月。由此可见最大熵模型的复杂的一面。最大熵模型快速算法的实现很复杂，到今天为止，世界上能有效实现这些算法的人也不到一百人。 有兴趣实现一个最大熵模型的读者可以阅读我的论文。

最大熵模型，可以说是集简与繁于一体，形式简单，实现复杂。值得一提的是，在Google的很多产品中，比如机器翻译，都直接或间接地用到了最大熵模型。

讲 到这里，读者也许会问，当年最早改进最大熵模型算法的达拉皮垂兄弟这些年难道没有做任何事吗？他们在九十年代初贾里尼克离开 IBM 后，也退出了学术界，而到在金融界大显身手。他们两人和很多 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(hedge fund)公司----文艺复兴技术公司(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。达拉皮 垂兄弟等科学家在那里，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伯克夏哈撒韦（ Berkshire Hathaway)。同期，伯克夏哈撒韦的总回报是 16 倍。

值得一提的是，信息处理的很多数学手段，包括隐含马尔可夫模型、子波变换、贝叶斯网络等等，在华尔街多有直接的应用。由此可见，数学模型的作用。