第 5 章　隐含马尔可夫模型

隐含马尔可夫模型是一个并不复杂的数学模型，到目前为止，它一直被认为是解决大多数自然语言处理问题最为快速、有效的方法。它成功地解决了复杂的语音识别、机器翻译等问题。当我们看完这些复杂的问题是如何通过简单的模型描述和解决时，会不得不由衷地感叹数学模型之妙。

1　通信模型

我们在第一、二章中介绍了，人类信息交流的发展贯穿了人类的进化和文明的全过程。而自然语言是人类交流信息的工具，语言和通信的联系是天然的。通信的本质就是一个编解码和传输的过程。但是自然语言处理早期的努力都集中在语法、语义和知识表述上，离通信的原理越走越远，而这样离答案也就越来越远。当自然语言处理的问题回归到通信系统中的解码问题时，很多难题都迎刃而解了。

让我们先来看一个典型的通信系统：当一个人（或者机器）发送信息时，他需要采用一种能在媒体中（比如空气、电线）传播的信号，比如语音或者电话线的调制信号，这个过程是广义上的编码。然后通过媒体传播到接收方，这个过程是信道传输。在接收方，收听的人（或者机器）根据事先约定好的方法，将这些信号还原成发送者的信息，这个过程是广

义上的解码。下图表示了一个典型的通信系统，它包含雅格布森（Roman Jakobson）提出的通信的六个要素[1]。

1 雅格布森通信六个要素是：发送者（信息源），信道，接收者，信息，上下文和编码。

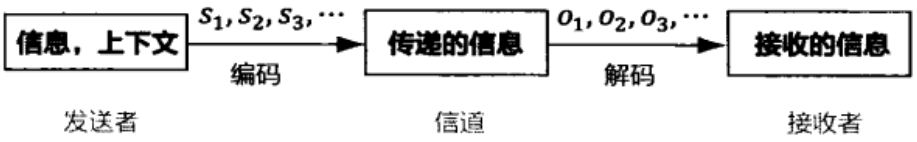

图 5.1 通信模型

其中$s_1, s_2, s_3, \cdots$表示信息源发出的信号，比如手机发送的信号。$o_1, o_2, o_3, \cdots$是接收器（比如另一部手机）接收到的信号。通信中的解码就是根据接收到的信号$o_1, o_2, o_3, \cdots$还原出发送的信号$s_1, s_2, s_3, \cdots$。

那么这与自然语言处理的工作，比如语音识别，又有什么直接的关系呢？不妨换一个角度来考虑这个问题。所谓语音识别，就是听话的人去猜测说话者要表达的意思。这其实就像通信中，根据接收端收到的信号去分析、理解、还原发送端传送过来的信息。我们平时在说话时，脑子就是一个信息源。我们的喉咙（声带）、空气，就是如电线和光缆般的信道。听众的耳朵就是接收器，而听到的声音就是传送过来的信号。根据声学信号来推测说话者的意思，就是语音识别。如果接收端是一台计算机而不是人，那么计算机要做的就是语音的自动识别。

同样，很多自然语言处理的应用也可以这样理解。在从汉语到英语的翻译中，说话者讲的是汉语，但是信道传播编码的方式是英语，如果利用计算机，根据接收到的英语信息，推测说话者的汉语意思，就是机器翻译。同样，如果要根据带有拼写错误的语句推测说话者想表达的正确意思，那就是自动纠错。这样，几乎所有的自然语言处理问题都可以等价成通信的解码问题。

在通信中，如何根据接收端的观测信号$o_1, o_2, o_3, \cdots$来推测信号源发送的信息$s_1, s_2, s_3, \cdots$呢？只需要从所有的源信息中找到最可能产生出观测信

号的那一个信息。用概率论的语言来描述，就是在已知 $o_1, o_2, o_3, \cdots$ 的情况下，求得令条件概率

$P(s_1, s_2, s_3, \cdots | o_1, o_2, o_3, \cdots)$达到最大值的那个信息串$s_1, s_2, s_3, \cdots$，即

$$s_1, s_2, s_3, \cdots = \underset{all\ s_1, s_2, s_3, \cdots}{\text{Arg Max}}\, P(s_1, s_2, s_3, \cdots | o_1, o_2, o_3, \cdots) \tag{5.1}$$

其中 Arg 是参数 Argument 的缩写，表示能获得最大值的那个信息串。当然，上面的概率不容易直接求出，不过可以间接地计算它。利用贝叶斯公式可以把上述公式等价变换成

$$\frac{P(o_1, o_2, o_3, \cdots | s_1, s_2, s_3, \cdots) \cdot P(s_1, s_2, s_3, \cdots)}{P(o_1, o_2, o_3, \cdots)} \tag{5.2}$$

其中 $P(o_1, o_2, o_3, \cdots | s_1, s_2, s_3, \cdots)$ 表示信息 $s_1, s_2, s_3, \cdots$ 在传输后变成接收的信号 $o_1, o_2, o_3, \cdots$ 的可能性；而 $P(s_1, s_2, s_3, \cdots)$ 表示 $s_1, s_2, s_3, \cdots$ 本身是一个在接收端合乎情理的信号（比如一个合乎情理的句子）的可能性；最后 $P(o_1, o_2, o_3, \cdots)$ 表示在发送端（比如说话的人）产生信息 $o_1, o_2, o_3, \cdots$ 的可能性。

大家读到这里也许会问，你现在是不是把问题变得更复杂了，因为公式越写越长了。别着急，我们现在就来简化这个问题。首先，一旦信息 $o_1, o_2, o_3, \cdots$ 产生了，它就不会改变了，这时 $P(o_1, o_2, o_3, \cdots)$ 就是一个可以忽略的常数。因此，上面的公式可以等价成

$$P(o_1, o_2, o_3, \cdots | s_1, s_2, s_3, \cdots) \cdot P(s_1, s_2, s_3, \cdots) \tag{5.3}$$

当然，这里面还有两项，虽然多过（5.1）的一项，但是这个公式完全可以用隐含马尔可夫模型（Hidden Markov Model）来估计。

2 隐含马尔可夫模型

图 5.2 俄罗斯著名科学家安德烈·马尔可夫

隐含马尔可夫模型（Hidden Markov Model）其实并不是19世纪俄罗斯数学家马尔可夫（Andrey Markov）发明的，而是美国数学家鲍姆（Leonard E. Baum）等人在20世纪六七十年代发表的一系列论文中提出的，隐含马尔可夫模型的训练方法（鲍姆-韦尔奇算法）也是以他的名字命名的。

要介绍隐含马尔可夫模型，还是要从马尔可夫链说起。到了19世纪，概率论的发展从对（相对静态的）随机变量的研究发展到对随机变量的时间序列$s_1, s_2, s_3, \cdots, s_t, \cdots$，即随机过程（动态的）的研究。这在哲学的意义上，是人类认识的一个飞跃。但是，随机过程要比随机变量复杂得多。首先，在任何一个时刻t，对应的状态s_t都是随机的。举一个大家熟悉的例子，我们可以把$s_1, s_2, s_3, \cdots, s_t, \cdots$看成是北京每天的最高气温，这里面每个状态$s_t$都是随机的。第二，任何一个状态$s_t$的取值都可能和周围其他的状态相关。回到上面的例子，任何一天的最高气温，与这段时间以前的最高气温是相关的。这样随机过程就有两个维度的不确定性。马尔可夫为了简化问题，提出了一种简化的假设，即随机过程中各个状态s_t的概率分布，只与它的前一个状态s_{t-1}有关，即$P(s_t|s_1, s_2, s_3, \cdots, s_{t-1}) = P(s_t|s_{t-1})$。比如，对于天气预报，硬性假定今天的气温只与昨天有关而和前天无关。当然这种假设未必适合所有的应用，但是至少对以前很多不好解决的问题给出了近似解。这个假设后来被命名为马尔可夫假设，而符合这个假设的随机过程则称为马尔可夫过程，也称为马尔可夫链。下图表示一个离散的马尔可夫过程。

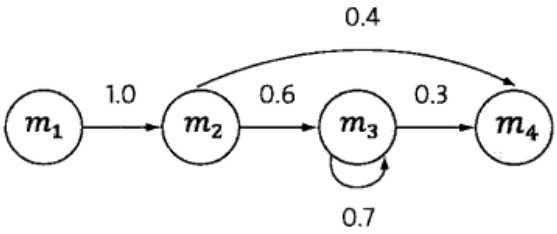

图 5.3　马尔可夫链

在这个马尔可夫链中，四个圈表示四个状态，每条边表示一个可能的状态转换，边上的权值是转移概率。例如，状态 m_1 到 m_2 之间只有一条边，且边上权值为 1.0。这表示从状态 m_1 只可能转换到状态 m_2，转移概率为 1.0。从 m_2 出发的有两条边：到 m_3 和到 m_4。其中权值 0.6 表示：如果某个时刻 t 的状态 s_t 是 m_2，则下一个时刻的状态 $s_{t+1} = m_3$ 的概率（可能性）是 60%。如果用数学符号表示是 $P(s_{t+1} = m_3 | s_t = m_2) = 0.6$。类似的，有 $P(s_{t+1} = m_4 | s_t = m_2) = 0.4$。

把这个马尔可夫链想象成一台机器，它随机地选择一个状态作为初始状态，随后按照上述规则随机选择后续状态。这样运行一段时间 T 之后，就会产生一个状态序列：$s_1, s_2, s_3, \cdots s_T$。看到这个序列的人，不难数出某个状态 m_i 的出现次数 $\#(m_i)$，以及从 m_i 转换到 m_j 的次数 $\#(m_i, m_j)$，从而估计出从 m_i 到 m_j 的转移概率 $\#(m_i, m_j)/\#(m_i)$。每一个状态只和前面一个有关，比如从状态 3 到状态 4，不论在此之前是如何进入到状态 3 的（是从状态 2 进入，还是在状态 3 本身转了几个圈子），这个概率都是 0.3。

隐含马尔可夫模型是上述马尔可夫链的一个扩展：任一时刻 t 的状态 s_t 是不可见的。所以观察者没法通过观察到一个状态序列 $s_1, s_2, s_3, \cdots s_T$ 来推测转移概率等参数。但是，隐含马尔可夫模型在每个时刻 t 会输出一个符号 o_t，而且 o_t 和 s_t 相关且仅和 s_t 相关。这个被称为独立输出假设。隐含马尔可夫模型的结构如下：其中隐含的状态 $s_1, s_2, s_3, \cdots$ 是一个典型的马尔可夫链。鲍姆把这种模型称为“隐含”马尔可夫模型。

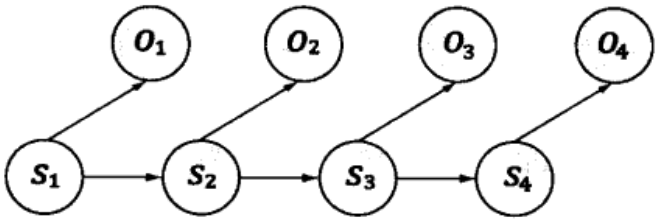

图 5.4　隐含马尔可夫模型

基于马尔可夫假设和独立输出假设，我们可以计算出某个特定的状态序列$s_1, s_2, s_3, \cdots$产生出输出符号$o_1, o_2, o_3, \cdots$的概率。

$$P(s_1, s_2, s_3, \cdots, o_1, o_2, o_3, \cdots) = \prod_t P(s_t|s_{t-1}) \cdot P(o_t|s_t) \qquad (5.4)$$

读者可能已经看出，公式（5.4）在形态上和公式（5.3）非常相似。现在我们把马尔可夫假设和独立输出假设用于通信的解码问题（5.3），即把

$$\begin{aligned} &P(o_1, o_2, o_3, \cdots \mid s_1, s_2, s_3, \cdots) = \prod_t P(o_t|s_t) \\ &P(s_1, s_2, s_3, \cdots) = \prod_t P(s_t|s_{t-1}) \end{aligned} \qquad (5.5)$$

代入（5.3），这时正好得到（5.4）。这样通信的解码问题就可以用隐含马尔可夫模型来解决了。而很多自然语言处理问题是和通信的解码问题等价的，因此它们完全可以由隐含马尔可夫模型来解决。至于如何找出上面式子的最大值，进而找出要识别的句子$s_1, s_2, s_3, \cdots$，可以利用维特比算法（Viterbi Algorithm），这里面的细节我们会在后面的章节中介绍。

在公式（5.3）中，$P(s_1, s_2, s_3, \cdots)$是语言模型，我们在前面的一章已经介绍过了。

$P(s_1, s_2, s_3, \cdots | o_1, o_2, o_3, \cdots)$根据应用的不同而有不同的名称，在语音识别中它被称为“声学模型”（Acoustic Model），在机器翻译中是“翻译模型”（Translation Model），而在拼写校正中是“纠错模型”（Correction Model）。

隐含马尔可夫模型成功的应用最早是语音识别。20 世纪 70 年代，当时

IBM 华生实验室的贾里尼克领导的科学家们，主要是刚刚从卡内基－梅隆大学毕业的贝克夫妇（James and Janet Baker）[2]，提出用隐含马尔可夫模型来识别语音，语音识别的错误率相比人工智能和模式匹配等方法降低了三分之二（从 30% 到 10%）。20 世纪 80 年代末李开复博士坚持采用隐含马尔可夫模型的框架，成功地开发了世界上第一个大词汇量连续语音识别系统 Sphinx。接下来，隐含马尔可夫模型陆续成功地应用于机器翻译、拼写纠错、手写体识别、图像处理、基因序列分析等很多 IT 领域，近 20 年来，它还广泛应用于股票预测和投资。

2 李开复的师兄和师姐，后来共同创立了 Dragon 语言公司，现已离异。

我最早接触到隐含马尔可夫模型是 20 多年前的事。那时在“随机过程”（清华过去“臭名昭著”的一门课）里学到这个模型，但当时实在想不出它有什么实际用途。几年后，我在清华跟随王作英教授学习、研究语音识别时，他给了我几十篇文献。我印象最深的就是贾里尼克和李开复的文章，它们的核心思想就是隐含马尔可夫模型。复杂的语音识别问题居然能如此简单地表述、解决，我由衷地感叹数学模型之妙。

3　延伸阅读：隐含马尔可夫模型的训练

读者知识背景：概率论。

围绕着隐含马尔可夫模型有三个基本问题：

1. 给定一个模型，如何计算某个特定的输出序列的概率；
2. 给定一个模型和某个特定的输出序列，如何找到最可能产生这个输出的状态序列；
3. 给定足够量的观测数据，如何估计隐含马尔可夫模型的参数。

第一个问题相对简单，对应的算法是 Forward-Backward 算法，在此略过，有兴趣的读者可以参看弗里德里克·贾里尼克（Frederick Jelinek）的 *Statistical Methods for Speech Recognition*（*Language, Speech, and Communication*）一书[3]。第二个问题可以用著名的维特比算法解决，我们

3 The MIT Press (January 16, 1998)

在以后的章节中会介绍。第三个问题就是我们这一节要讨论的模型训练问题。

在利用隐含马尔可夫模型解决实际问题中，需要事先知道从前一个状态 s_{t-1} 进入当前状态 s_t 的概率 $P(s_t|s_{t-1})$，也称为转移概率（Transition Probability），和每个状态 s_t 产生相应输出符号 o_t 的概率 $P(o_t|s_t)$，也称为生成概率（Generation Probability）。这些概率被称为隐含马尔可夫模型的参数，而计算或者估计这些参数的过程称为模型的训练。

我们从条件概率的定义出发，知道：

$$P(o_t|s_t) = \frac{P(o_t, s_t)}{P(s_t)} \tag{5.6}$$

$$P(s_t|s_{t-1}) = \frac{P(s_{t-1}, s_t)}{P(s_{t-1})} \tag{5.7}$$

对于公式（5.6）的状态输出概率，如果有足够多人工标记（Human Annotated）的数据，知道经过状态 s_t 有多少次 $\#(s_t)$，每次经过这个状态时，分别产生的输出 o_t 是什么，而且分别有多少次 $\#(o_t, s_t)$ 就可以用两者的比值

$$P(o_t|s_t) \approx \frac{\#(o_t, s_t)}{\#(s_t)} \tag{5.8}$$

直接算出（估计出）模型的参数。因为数据是人工标注的，因此这种方法称为有监督的训练方法（Supervised Training）。对于公式（5.7）的转移概率，其实和前面提到的训练统计语言模型的条件概率是完全相同的，因此可以依照统计语言模型的训练方法

$$P(w_i|w_{i-1}) \approx \frac{\#(w_{i-1}, w_i)}{\#(w_{i-1})} \tag{5.9}$$

直接得到。有监督的训练的前提是需要大量人工标注的数据。很遗憾的是，很多应用都不可能做到这件事，比如在语言识别中的声学模型训练，人是无法确定产生某个语音的状态序列的，因此也就无法标注训练模型的数据。而在另外一些应用中，虽然标注数据是可行的，但是成本非常高。比如训练中英机器翻译的模型，需要大量中英对照的语料，还要把中英

文的词组一一对应起来，这个成本非常高。因此，训练隐含马尔可夫模型更实用的方式是仅仅通过大量观测到的信号 $o_1, o_2, o_3, \cdots$ 就能推算模型参数的 $P(s_t|s_{t-1})$ 和 $P(o_t|s_t)$ 的方法，这类方法称为无监督的训练算法，其中主要使用的是鲍姆 - 韦尔奇算法（Baum-Welch Algorithm）。

两个不同的隐含马尔可夫模型可以产生同样的输出信号，因此，仅仅通过观察到的输出信号来倒推产生它的隐含马尔可夫模型可能会得到很多个合适的。但是总会是一个模型 $M_{\theta 2}$ 比另一个 $M_{\theta 1}$ 更有可能产生观测到的输出，其中 θ_2 和 θ_1 是隐含马尔可夫模型的参数。鲍姆 - 韦尔奇算法就是寻找这个最可能的模型 $M_{\hat{\theta}}$。

在鲍姆 - 韦尔奇算法的思想是这样的：

首先找到一组能够产生输出序列 O 的模型参数。（显然它们是一定存在的，因为转移概率 P 和输出概率 Q 为均匀分布时，模型可以产生任何输出，当然包括我们观察到的输出 O。）现在，有了这样一个初始的模型，我们称为 $M_{\theta 0}$，需要在此基础上找到一个更好的模型。假定解决了第一个问题和第二个问题，不但可以算出这个模型产生 O 的概率 $P(O|M_{\theta 0})$，而且能够找到这个模型产生 O 的所有可能的路径以及这些路径的概率。这些可能的路径，实际上记录了每个状态经历的多少次，到达了哪些状态，输出了哪些符号，因此可以将它们看做是“标注的训练数据”，并且根据公式（5.6）和（5.7）计算出一组新的模型参数 θ_1，从 $M_{\theta 0}$ 到 $M_{\theta 1}$ 的过程称为一次迭代。可以证明

$$P(O|M_{\theta 1}) > P(O|M_{\theta 0}) \tag{5.10}$$

接下来，我们从 $M_{\theta 1}$ 出发，可以找到一个更好的模型 $M_{\theta 2}$，并且不断地找下去，直到模型的质量没有明显提高为止。这就是鲍姆 - 韦尔奇算法的原理，对于具体算法的公式，有兴趣的读者可以阅读参考文献 [5.2]，我就不再赘述了。

鲍姆 - 韦尔奇算法的每一次迭代都是不断地估计（Expectation）新的模

型参数，使得输出的概率（我们的目标函数）达到最大化（Maximization），因此这个过程被称为期望值最大化（Expectation-Maximization），简称EM过程。EM过程保证算法一定能收敛到一个局部最优点，很遗憾它一般不能保证找到全局最优点。因此，在一些自然语言处理的应用中，比如词性标注，这种无监督的鲍姆-韦尔奇算法训练出的模型比有监督的训练得到的模型效果略差，因为前者未必能收敛到全局最优点。但是如果目标函数是凸函数（比如信息熵），则只有一个最优点，在这种情况下EM过程可以找到最佳值。

4 小结

隐含马尔可夫模型最初应用于通信领域，继而推广到语音和语言处理中，成为连接自然语言处理和通信的桥梁。同时，隐含马尔可夫模型也是机器学习主要工具之一。和几乎所有的机器学习的模型工具一样，它需要一个训练算法（鲍姆-韦尔奇算法）和使用时的解码算法（维特比算法），掌握了这两类算法，就基本上可以使用隐含马尔可夫模型这个工具了。

参考文献：

1. Baum, L. E.; Petrie, T. (1966). "Statistical Inference for Probabilistic Functions of Finite State Markov Chains". The Annals of Mathematical Statistics 37 (6): 1554-1563.

2. Baum, L. E.; Eagon, J. A. (1967). "An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology". Bulletin of the American Mathematical Society 73 (3):

3. Baum, L. E.; Sell, G. R. (1968). "Growth transformations for functions on manifolds". Pacific Journal of Mathematics 27 (2): 211-227.

4. Baum, L. E.; Petrie, T.; Soules, G.; Weiss, N. (1970). "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains". The Annals of Mathematical Statistics 41:

5. Jelinek, F.; Bahl, L.; Mercer, R. (1975). "Design of a linguistic statistical decoder for the recognition of continuous speech". IEEE Transactions on Information Theory 21 (3): 250.