第 5 章　隐马尔可夫模型

学习目标：

（1）明确命名实体识别的含义；

（2）掌握隐马尔可夫模型的模型表示；

（3）掌握用于模型推理的前向—后向算法、维特比算法；

（4）掌握用于参数学习的期望最大化算法；

（5）了解隐马尔可夫模型的模型扩展。

在现实应用中，很多数据并不符合独立同分布假设，它们具有结构化的表示。序列数据是一种典型的较为简单的结构化数据，大家熟知的时间序列就是一种特殊的序列数据。顾名思义，时间序列是指随时间变化的数据，例如语音信号或运动物体的视频。然而，序列数据不仅限于时间序列，所有具有前后关系的数据都是序列数据，如自然语言文本等。

隐马尔可夫模型（hidden Markov model）是用于建模序列数据的概率模型[1]，广泛应用于各种时序数据处理的领域，如语音识别、自然语言建模、在线手写体识别等。下面介绍使用隐马尔可夫模型进行命名实体识别的例子。

命名实体识别的目的是提取句子中的实体块，如人名或动物名、地名、组织机构名。命名实体识别可被形式化为序列标注任务，即给定观测序列，识别出对应的标签序列，然后根据识别出的标签序列得到实体。命名实体识别通常采用 BIO 的标注形式，其中 B 表示实体的开头，I 表示实体的其他部分，O 表示非实体。考查如下例句：多年前，中国向美国赠送了一对大熊

猫“玲玲”和“兴兴”。如果使用 TIM 代表时间实体，PLA 代表地名实体，PER 代表人名（或动物名）实体，要识别出“多年前”为时间实体，理想情况下需要将“多年前”标记为（B-TIM，I-TIM，I-TIM），其他类型的实体类似。该例句的正确标签序列为：B-TIM、I-TIM、I-TIM、B-PLA、I-PLA、O、B-PLA、I-PLA、O、O、O、O、O、O、O、O、B-PER、I-PER、O、B-PER、I-PER，具体对应关系如图 5-1 所示，从标签序列中可以提取出实体块[多年前][中国][美国][玲玲][兴兴]。

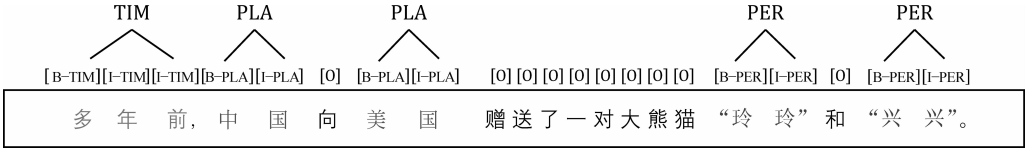

图 5-1　自然语言处理中的命名实体识别示例

进行命名实体识别之前，需要对隐马尔可夫模型进行训练。隐马尔可夫模型假设标签序列是未观测变量，即训练数据可以是没有标签的语料库，此时模型是一种非监督训练，识别时使用贝叶斯推理求取隐状态序列。

5.1　模 型 表 示

隐马尔可夫模型与马尔可夫过程有着密切的联系，从马尔可夫过程入手更有助于理解隐马尔可夫模型。随机变量的集合称为随机过程，隐马尔可夫模型与马尔可夫过程均利用了随机过程中的马尔可夫性质。马尔可夫性质是对序列数据建模的一种常用假设，其含义是：如果在给定现在状态时，随机过程的后续状态与过去状态是条件独立的，那么此随机过程具有马尔可夫性质。最简单的马尔可夫性质是一阶马尔可夫性，意思是未来的状态只条件依赖于当前状态。以此类推，二阶马尔可夫性是指未来的状态只条件依赖于当前状态和上一时刻的状态。二阶及以上的马尔可夫性质统称为高阶马尔可

夫性。

具有马尔可夫性质的随机过程称为马尔可夫过程，它的建模特点是直接假设观测序列具有马尔可夫性质。使用马尔可夫过程进行序列数据建模，可以刻画序列数据内部的依赖关系，但是在许多复杂情况中也表现出局限性。一方面，如果对观测序列进行过强的低阶马尔可夫假设，模型并非适用于某些真实的序列数据；另一方面，如果对观测序列使用高阶马尔可夫过程建模，会引入过多的模型参数。

为了建模序列数据内部的依赖关系，并且能够利用马尔可夫性质得到优雅的模型表示，隐马尔可夫模型在马尔可夫过程的基础上引入了隐状态序列。隐马尔可夫模型不再直接假设观测序列具有马尔可夫性，而是通过两个假设构建序列数据内部的依赖关系，即①假设隐状态序列是一个马尔可夫链；②假设每个时刻的观测在给定对应隐状态的条件下互相独立。条件独立假设使得模型的似然表示具有因子化的形式，马尔可夫性质也提供了一种建模隐状态序列内部依赖关系的方法。

假设 $z_1, z_2, \cdots, z_t, \cdots, z_T$ 表示隐状态序列，隐状态是离散变量，具有有限的 K 种取值，$S=\{1,2,\cdots,K\}$ 表示每个时刻所有可能隐状态的集合。假设 $x_1, x_2, \cdots, x_t, \cdots, x_T$ 表示观测序列，观测序列的取值可以是连续值，也可以是离散值。如果观测序列是离散变量，那么同样可以枚举出所有可能的取值以及观测数据在给定当前隐状态条件下的观测概率，例如 $V=\{v_1, v_2, \cdots, v_M\}$ 表示所有可能观测的集合，$\boldsymbol{B}=[b_{ij}]_{KM}$ 表示观测概率矩阵；如果观测序列是连续变量，那么观测数据的概率分布可以使用某种条件概率分布表示，例如假设 $p(x_t \mid z_t=k)=\mathcal{N}(u_k, \sigma_k^2)$。给定上述假设，隐马尔可夫模型可以使用图 5-2 所示的概率图模型表示。

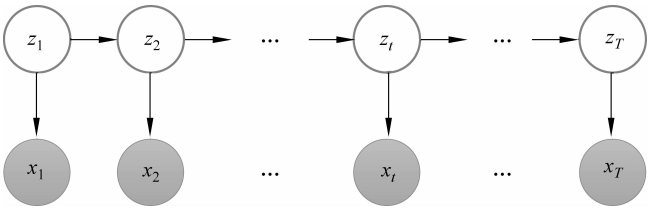

图 5-2　隐马尔可夫模型的概率图模型表示

下面介绍隐马尔可夫模型的具体表示。

首先，隐状态序列是一个马尔可夫链，初始状态的概率可以使用概率向量刻画，相邻时刻的隐状态之间的变化使用状态转移概率矩阵刻画。假设向量 $\boldsymbol{\pi}=[\pi_1,\pi_2,\cdots,\pi_K]^{\top}$ 表示模型中初始时刻状态的概率分布，即

$$p(z_1=k)=\pi_k \tag{5.1}$$

假设 $\boldsymbol{A}=[a_{ij}]_{KK}$ 是状态转移概率矩阵，表示在当前时刻状态为 i 在下一时刻状态为 j 的概率，即

$$p(z_{t+1}=j \mid z_t=i)=a_{ij}, \quad i,j=1,2,\cdots,K \tag{5.2}$$

图 5-3 给出了具有 3 个隐状态的隐马尔可夫模型的隐状态转移图。

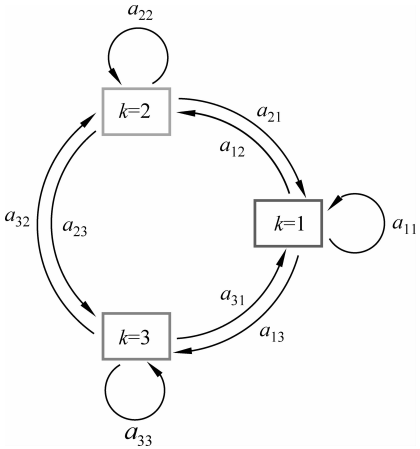

图 5-3　隐马尔可夫模型的状态转移示意图

其次，序列在每一个时刻的观测数据条件独立。如果观测序列是离散

值，假设 $\boldsymbol{B}=[b_{ij}]_{KM}$ 是观测概率矩阵，也叫发射概率矩阵，表示在任意 t 时刻，隐状态是 i 的情况下生成观测 v_j 的概率为

$$p(x_t=v_j \mid z_t=i)=b_{ij}, \quad i=1,2,\cdots,K, \quad j=1,2,\cdots,M \tag{5.3}$$

如果观测序列可以是连续值或离散值，可以统一通过概率分布描述 $p(x_t \mid z_t)$。这里使用更一般化的表示，使用 $\boldsymbol{\phi}$ 表示分布的参数。例如，对于离散观测，可以假设离散分布，且参数为 $\boldsymbol{\phi}=\boldsymbol{B}$，对于连续观测，可以假设高斯分布，且参数为 $\boldsymbol{\phi}=\{\mu_k,\sigma_k^2\}_{k=1}^K$。

根据隐马尔可夫模型的先验分布和似然概率，可以写出模型关于观测变量与隐变量的联合分布为

$$p(\boldsymbol{x},\boldsymbol{z} \mid \boldsymbol{\pi},\boldsymbol{A},\boldsymbol{\phi})=p(z_1 \mid \boldsymbol{\pi})\Big[\prod_{t=2}^{T} p(z_t \mid z_{t-1},\boldsymbol{A})\Big]\prod_{t=1}^{T} p(x_t \mid z_t,\boldsymbol{\phi}) \tag{5.4}$$

使用隐马尔可夫模型进行中文命名实体识别任务时，观测序列是由汉字组成的句子，隐状态序列是每个汉字对应的标签。用隐状态的初始概率向量、状态转移概率矩阵及观测概率矩阵即可确定一个隐马尔可夫模型。在该任务中，初始概率向量为第一个汉字属于某个标签的概率，状态转移概率矩阵为从某一个标签转移到下一个标签的概率（对于状态转移矩阵 $\boldsymbol{A}$，若当前词的标签为 i，则下一个词的标签为 j 的概率为 a_{ij}）；观测概率矩阵即发射概率矩阵，指某个标签生成某个汉字的概率。模型隐状态的数目是标签的类别数，观测的数目是汉字的个数。

5.2 模型推理

通常情况下，训练一个隐马尔可夫模型包括模型推理与参数学习，使用隐马尔可夫模型完成具体的预测任务时也需要模型推理。这里介绍的模型推理包括三个具体的推理目标：边缘似然的推理、隐状态序列的推理、隐状态边缘后验的推理。这些推理有些是为了完成具体任务，有些则是在进行诸如最大似然

估计时计算目标函数所需。进行模型推理时，均假设模型参数$\{\boldsymbol{\pi},\boldsymbol{A},\boldsymbol{\phi}\}$已知。

首先，目前模型表示可以给出观测变量与隐变量的联合分布，即公式(5.4)。然而，在使用隐马尔可夫模型对序列数据进行建模时，人们希望能够评价模型对数据的拟合程度。例如，隐马尔可夫模型可以用来进行序列数据分类，其方法是将模型作为类条件分布模型，然后通过贝叶斯决策对新的测试序列进行分类。这时需要通过计算模型的边缘似然进行评估。因此，一个重要的推理任务是对边缘似然的推理，即给定观测序列 $\boldsymbol{x}$ 和模型参数$\{\boldsymbol{\pi},\boldsymbol{A},\boldsymbol{\phi}\}$，计算某观测序列的概率 $p(\boldsymbol{x}|\boldsymbol{\pi},\boldsymbol{A},\boldsymbol{\phi})$。

其次，隐马尔可夫模型还可以进行最优隐状态序列的求解。例如，在应用示例中介绍的命名实体识别问题，其实是一个序列标注问题。对于隐马尔可夫模型而言，就是求解对应某个观测序列的最可能的隐状态序列，即给定观测序列 $\boldsymbol{x}$ 和模型参数$\{\boldsymbol{\pi},\boldsymbol{A},\boldsymbol{\phi}\}$，找到最优的隐状态序列 $\boldsymbol{z}$。

最后，在训练隐马尔可夫模型时，通常使用最大似然估计中的期望最大化方法，这种参数估计问题也需要推理进行辅助。例如，在期望步骤，需要计算的目标函数中包含隐状态序列的两种边缘后验分布，一个是单一时刻的边缘后验 $p(z_t|\boldsymbol{x})$，一个是相邻时刻的边缘后验 $p(z_t,z_{t+1}|\boldsymbol{x})$。

隐马尔可夫模型是一种链状的贝叶斯网络，其推理可以使用概率图模型推理的一般方法，虽然针对特定结构会有不同的表现形式。边缘似然的推理需要对隐变量进行积分，可以使用一般的和积算法，在隐马尔可夫模型中，常用的是前向-后向算法。隐状态序列的推理目标是求解最优隐变量，使用的是最大和算法，在隐马尔可夫模型中称为维特比解码算法。隐状态边缘后验的推理同样可以使用和积算法。下面将详细介绍隐马尔可夫模型的推理。

5.2.1　边缘似然的推理

边缘似然的推理是指给定某观测序列 $\boldsymbol{x}=(x_1,x_2,\cdots,x_T)$和模型参数$\{\boldsymbol{\pi},\boldsymbol{A},\boldsymbol{\phi}\}$，计算观测序列的概率 $p(\boldsymbol{x}|\boldsymbol{\pi},\boldsymbol{A},\boldsymbol{\phi})$。给定模型的联合分布以后，

求边缘似然的方法是对隐变量进行积分。由于隐变量是离散变量，一个直接的算法就是对状态的取值进行枚举。首先，列举所有可能的隐状态序列 $\boldsymbol{z}$，求出各个状态序列与观测序列的联合分布 $p(\boldsymbol{x},\boldsymbol{z}|\boldsymbol{\pi},\boldsymbol{A},\boldsymbol{\phi})$。然后，对所有可能的状态序列求和，得到 $p(\boldsymbol{x}|\boldsymbol{\pi},\boldsymbol{A},\boldsymbol{\phi})$。该方法虽然在理论上可行，但其高昂的复杂度使其在计算上不可行，因此需要使用和积算法，在隐马尔可夫模型中等价于前向—后向算法。

下面具体分析为何直接枚举求和不可行。给定模型参数$\{\boldsymbol{\pi},\boldsymbol{A},\boldsymbol{\phi}\}$，隐状态序列 $\boldsymbol{z}=(z_1,z_2,\cdots,z_T)$和观测序列 $\boldsymbol{x}=(x_1,x_2,\cdots,x_T)$的联合分布如公式(5.4)所示。对于所有的状态序列 $\boldsymbol{z}$ 求和，得到观测序列 $\boldsymbol{x}$ 的概率分布 $p(\boldsymbol{x})$，即

$$\begin{aligned} p(\boldsymbol{x}) &= \sum_{\boldsymbol{z}} p(\boldsymbol{x},\boldsymbol{z}) \\ &= \sum_{z_1,z_2,\cdots,z_T} p(z_1|\boldsymbol{\pi})\Big[\prod_{t=2}^{T} p(z_t|z_{t-1},\boldsymbol{A})\Big]\prod_{t=1}^{T} p(x_t|z_t,\boldsymbol{\phi}) \end{aligned} \tag{5.5}$$

公式(5.5)的计算复杂度是 $O(TK^T)$，使得该算法在链长 T 较大时计算不可行。

下面分别介绍两种计算边缘分布的算法，即和积算法和前向—后向算法。

(1) 和积算法。

为了降低计算复杂度，在求和的过程中，一些公共运算可以被提出，从而只进行一次运算。正如第 4 章介绍的，可以借用乘法分配律实现消息传递算法，即和积算法。在链式结构中，和积算法可以用前向—后向算法的形式实现。首先构建图 5-4 所示的隐马尔可夫模型的因子图。

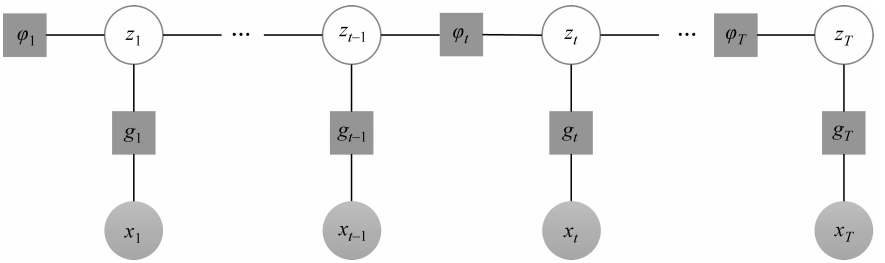

图 5-4　隐马尔可夫模型的因子图

其中，因子 ϕ_t 和 g_t 的表达为

$$\phi_1(z_1)=p(z_1) \tag{5.6}$$

$$\phi_t(z_{t-1},z_t)=p(z_t \mid z_{t-1}),\quad t=2,3,\cdots,T \tag{5.7}$$

$$g_t(x_t,z_t)=p(x_t \mid z_t),\quad t=1,2,\cdots,T \tag{5.8}$$

由于 x_n 是观测变量，不需要积分，在消息传递的过程中，该变量可以一直保留，因此因子图可以通过合并因子进一步简化，得到图 5-5 所示的简化版的因子图。其中，因子的表达为

$$f_1(z_1)=p(z_1)p(x_1 \mid z_1) \tag{5.9}$$

$$f_t(z_{t-1},z_t)=p(z_t \mid z_{t-1})p(x_t \mid z_t),\quad t=2,3,\cdots,T \tag{5.10}$$

在消息传递中，需要计算变量节点到因子节点的消息和因子节点到变量节点的消息。

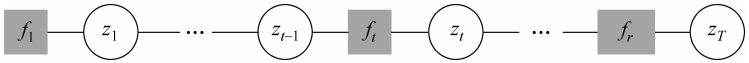

图 5-5　简化的隐马尔可夫模型因子图

这一节的目标是对所有的隐变量进行积分。可以首先得到某个时刻隐变量与观测变量的联合分布，然后进行一次单变量的求和。该隐变量的选择非常灵活，可以选择起始时刻或终止时刻的隐变量，也可以选择任意中间时刻的隐变量。如果选择起始时刻或终止时刻，因子图只需要执行单向的消息传递；如果选择中间时刻，则需要进行双向的消息传递。

首先，考虑前向传递，将最后一个节点 z_T 定为根节点，从叶子节点 f_1 到根节点 z_T 的消息传递可以写成

$$\mu_{z_{t-1}\to f_t}(z_{t-1})=\mu_{f_{t-1}\to z_{t-1}}(z_{t-1}) \tag{5.11}$$

$$\mu_{f_t\to z_t}(z_t)=\sum_{z_{t-1}} f_t(z_{t-1},z_t)\mu_{z_{t-1}\to f_t}(z_{t-1}) \tag{5.12}$$

递推关系（公式(5.11)和公式(5.12)）还可以化简为一个递推式，即

$$\mu_{f_t\to z_t}(z_t)=\sum_{z_{t-1}} f_t(z_{t-1},z_t)\mu_{f_{t-1}\to z_{t-1}}(z_{t-1}) \tag{5.13}$$

其中，初始消息为

$$\mu_{f_1\to z_1}(z_1)=f_1(z_1)=p(z_1)p(x_1\mid z_1) \tag{5.14}$$

由此可见，在链式结构中，消息传递可以只使用一种消息。执行了一次向前的消息传递之后，可以得到最后时刻隐变量的边缘分布为

$$p(z_T,\boldsymbol{x})=\mu_{f_T\to z_T}(z_T) \tag{5.15}$$

公式(5.15)之所以是隐变量和观测变量的联合，是因为消息传递使用了简化的因子图，将关于观测变量的因子合并到了新的因子中。因此，可以得到边缘似然概率分布为

$$p(\boldsymbol{x})=\sum_{z_T}p(z_T,\boldsymbol{x})=\sum_{z_T}\mu_{f_T\to z_T}(z_T) \tag{5.16}$$

这种前向消息传递在前向—后向算法中也称为前向算法，单独执行一次前向算法可以计算出边缘似然的概率分布。

然后，考虑后向传递，从根节点 z_T 到叶子节点 f_1 的消息传递可以写成

$$\mu_{f_{t+1}\to z_t}(z_t)=\sum_{z_{t+1}}f_{t+1}(z_t,z_{t+1})\mu_{f_{t+2}\to z_{t+1}}(z_{t+1}) \tag{5.17}$$

其中，初始消息为

$$\mu_{z_T\to f_T}(z_T)=1 \tag{5.18}$$

执行完所需要的向后消息传递后，可以得到与初始时刻隐变量有关的边缘分布为

$$p(z_1,\boldsymbol{x})=\mu_{f_2\to z_1}(z_1)f_1(z_1) \tag{5.19}$$

因此，可以得到边缘似然概率分布为

$$p(\boldsymbol{x})=\sum_{z_1}p(z_1,\boldsymbol{x})=\sum_{z_1}\mu_{f_2\to z_1}(z_1)p(z_1)p(x_1\mid z_1) \tag{5.20}$$

这种后向消息传递在前向—后向算法中也称为后向算法，单独执行一次后向算法同样可以计算出边缘似然的概率分布。

最后，考虑双向传递，即同时使用从叶子节点和根节点向中间传递的消息，运用的递推关系如公式(5.13)和公式(5.17)所示。那么，可以得到任意时刻的隐变量与观测变量的联合边缘分布为

$$p(z_t,\boldsymbol{x})=\mu_{f_t\to z_t}(z_t)\mu_{f_{t+1}\to z_t}(z_t) \tag{5.21}$$

边缘似然概率分布为

$$p(\boldsymbol{x})=\sum_{z_t}p(z_t,\boldsymbol{x})=\sum_{z_t}\mu_{f_t\to z_t}(z_t)\mu_{f_{t+1}\to z_t}(z_t) \tag{5.22}$$

(2) 前向—后向算法。

前向—后向算法(forward-backward algorithm)可以从和积算法的角度描述,也可以通过直接定义前向和后向概率分布,并利用概率论的运算法则得到递推关系。前向—后向算法包括前向算法和后向算法,使用单独的前向算法或后向算法,或者二者共同协作,都可以实现边缘概率分布的推理,最终计算出的边缘概率分布是一致的。下面分别介绍如何单独使用前向算法或后向算法,或同时使用二者得到隐马尔可夫模型的边缘似然概率分布。

前向算法是指从初始时刻到终止时刻进行递推运算,当计算到终止时刻时,可以获得目标概率分布。前向算法的关键是前向概率分布。

给定隐马尔可夫模型参数$\{\boldsymbol{\pi},\boldsymbol{A},\boldsymbol{\phi}\}$,定义初始时刻到时刻 t 的部分观测序列为$x_1,x_2,\cdots,x_t$,时刻 t 的隐状态为z_t 的概率分布为前向概率分布,记为

$$\alpha_t(z_t)=p(x_1,x_2,\cdots,x_t,z_t\mid\boldsymbol{\pi},\boldsymbol{A},\boldsymbol{\phi}) \tag{5.23}$$

利用概率论的运算法则可以得到前向概率分布 $\alpha_t(z_t)$的递推关系式为

$$\begin{aligned}
\alpha_t(z_t)&=p(x_1,x_2,\cdots,x_t,z_t)\\
&=p(x_1,x_2,\cdots,x_t\mid z_t)p(z_t)\\
&=p(x_t\mid z_t)p(x_1,x_2,\cdots,x_{t-1}\mid z_t)p(z_t)\\
&=p(x_t\mid z_t)p(x_1,x_2,\cdots,x_{t-1},z_t)\\
&=p(x_t\mid z_t)\sum_{z_{t-1}}p(x_1,x_2,\cdots,x_{t-1},z_{t-1},z_t)\\
&=p(x_t\mid z_t)\sum_{z_{t-1}}p(x_1,x_2,\cdots,x_{t-1},z_t\mid z_{t-1})p(z_{t-1})\\
&=p(x_t\mid z_t)\sum_{z_{t-1}}p(x_1,x_2,\cdots,x_{t-1}\mid z_{t-1})p(z_t\mid z_{t-1})p(z_{t-1})\\
&=p(x_t\mid z_t)\sum_{z_{t-1}}p(x_1,x_2,\cdots,x_{t-1},z_{t-1})p(z_t\mid z_{t-1})
\end{aligned}$$