PRML（Pattern Recognition And Machine Learning）读书会

第十三章 Sequential Data

主讲人 张巍

（新浪微博：@张巍_ISCAS）

QQ 群 177217565

读书会微信公众平台请扫描下面的二维码

软件所-张巍<zh3f@qq.com> 19:01:27

我们开始吧，十三章是关于序列数据，现实中很多数据是有前后关系的，例如语音或者 DNA 序列，例子就不多举了，对于这类数据我们很自然会想到用马尔科夫链来建模：

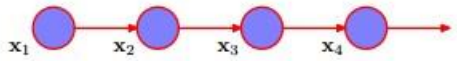

Figure 13.3 A first-order Markov chain of observations $\{\mathbf{x}_n\}$ in which the distribution $p(\mathbf{x}_n|\mathbf{x}_{n-1})$ of a particular observation $\mathbf{x}_n$ is conditioned on the value of the previous observation $\mathbf{x}_{n-1}$.

例如直接假设观测数据之间服从一阶马尔科夫链，这个假设显然太简单了，因为很多数据时明显有高阶相关性的，一个解决方法是用高阶马尔科夫链建模：

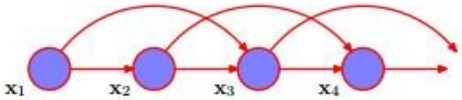

Figure 13.4 A second-order Markov chain, in which the conditional distribution of a particular observation $\mathbf{x}_n$ depends on the values of the two previous observations $\mathbf{x}_{n-1}$ and $\mathbf{x}_{n-2}$.

但这样并不能完全解决问题 ：1、高阶马尔科夫模型参数太多；2、数据间的相关性仍然受阶数限制。一个好的解决方法，是引入一层隐变量，建立如下的模型：

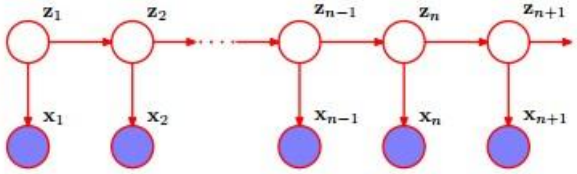

Figure 13.5 We can represent sequential data using a Markov chain of latent variables, with each observation conditioned on the state of the corresponding latent variable. This important graphical structure forms the foundation both for the hidden Markov model and for linear dynamical systems.

这里我们假设隐变量之间服从一阶马尔科夫链，观测变量由其对应的隐变量生成。从上图可以看出，隐变量是一阶的，但是观测变量之间是全相关的，今天我们主要讨论的就是上图中的模型。如果隐变量是离散的，我们称之为 Hidden Markov Models；如果是连续的，我们称之为: Linear Dynamical Systems。现在我们先来看一下 HMM ，从图中可以看出，要完成建模，我们需要指定一下几个分布：

1、转移概率：

$$p(\mathbf{z}_n|\mathbf{z}_{n-1},\mathbf{A}) = \prod_{k=1}^{K}\prod_{j=1}^{K} A_{jk}^{z_{n-1,j}z_{nk}}. \qquad (13.7)$$

2、马尔科夫链的初始概率：

$$p(\mathbf{z}_1|\boldsymbol{\pi}) = \prod_{k=1}^{K} \pi_k^{z_{1k}} \qquad (13.8)$$

3、生成观测变量的概率(emission probabilities)：

$$p(\mathbf{x}_n|\mathbf{z}_n,\boldsymbol{\phi}) = \prod_{k=1}^{K} p(\mathbf{x}_n|\boldsymbol{\phi}_k)^{z_{nk}}. \qquad (13.9)$$

对于 HMM， 这里 1 和 2 我们已经假设成了离散分布，3 我们暂时不做指定。模型建好了，我们接下来主要讨论下面三个问题：

1、学习问题：就是学习模型中的参数；

2、预测问题：即 $p(\mathbf{x}_{N+1}|\mathbf{X})$,给定当前序列预测下一个观测变量；

3、解码问题：即 p(Z|X)，给定观测变量求隐变量，例如语音识别；

游侠(419504839) 19:24:21

什么是解码问题？

软件所-张巍<zh3f@qq.com> 19:25:18

例如观测到了一段语音，要求识别其对应的句子。@游侠 我前面没怎么举例子，不知道这样说清楚没？

游侠(419504839) 19:27:20

这个和一般说的“推断”一样不

软件所-张巍<zh3f@qq.com> 19:28:27

这个也可以叫推断，只是推断是个比较一般的词汇。

我们来看一下 HMM 有多少参数要学，对应于刚才说到的三个分布，我们也有三组参数要学。

球猫(250992259) 19:30:46

其实就是假设东西是一个马尔科夫模型生成的。。然后把参数用某种方法弄出来，最后根据模型的输出来给答案……是这样吧？

软件所-张巍<zh3f@qq.com> 19:32:45

@球猫 对，都是这个思路，先把参数学出来，然后就可以做任何想要的推断了，在这里所谓的解码问题只是大家比较关心。

软件所-张巍<zh3f@qq.com> 19:32:51

好，继续，我们先来看 1、学习问题。这里我们用 EM 算法来学习 HMM 的参数：

1、是转移概率对应的转移矩阵；

2、初始概率对应的离散分布参数；

3、观测变量对应的分布参数（这里暂不指定）。

用 EM 我们要做的就是：

E 步里根据当前参数估计隐变量的后验：

$$p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{\text{old}})$$

M 步里最大化下面的期望：

$$Q(\boldsymbol{\theta},\boldsymbol{\theta}^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta}). \qquad (13.12)$$

先来看 M 步，这里相对简单一点，整个模型的全概率展开为：

$$p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta}) = p(\mathbf{z}_1|\boldsymbol{\pi}) \left[\prod_{n=2}^{N} p(\mathbf{z}_n|\mathbf{z}_{n-1},\mathbf{A})\right] \prod_{m=1}^{N} p(\mathbf{x}_m|\mathbf{z}_m,\boldsymbol{\phi}) \qquad (13.10)$$

把 13.10 代入 13.12,我们会发现计算时需要下面两个式子：

$p(\mathbf{z}_n|\mathbf{X},\boldsymbol{\theta}^{\text{old}})$ 和 $p(\mathbf{z}_{n-1},\mathbf{z}_n|\mathbf{X},\boldsymbol{\theta}^{\text{old}})$.

为了方便，我们就定义：

$$\gamma(\mathbf{z}_n) = p(\mathbf{z}_n|\mathbf{X},\boldsymbol{\theta}^{\text{old}}) \qquad (13.13)$$

$$\xi(\mathbf{z}_{n-1},\mathbf{z}_n) = p(\mathbf{z}_{n-1},\mathbf{z}_n|\mathbf{X},\boldsymbol{\theta}^{\text{old}}). \qquad (13.14)$$

这样我们在 E 步就主要求出这两个式子就行了，当然这也就意味着求出了整个后验 $p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{\text{old}})$ ，利用这

两个式子，13.12 可以化为：

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{k=1}^{K} \gamma(z_{1k}) \ln \pi_k + \sum_{n=2}^{N} \sum_{j=1}^{K} \sum_{k=1}^{K} \xi(z_{n-1,j}, z_{nk}) \ln A_{jk} + \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \ln p(\mathbf{x}_n|\boldsymbol{\phi}_k). \quad (13.17)$$

这个时候就可以用一些通用方法，例如 Lagrange 来求解了，结果也很简单：

$$\pi_k = \frac{\gamma(z_{1k})}{\sum_{j=1}^{K} \gamma(z_{1j})} \quad (13.18)$$

$$A_{jk} = \frac{\sum_{n=2}^{N} \xi(z_{n-1,j}, z_{nk})}{\sum_{l=1}^{K} \sum_{n=2}^{N} \xi(z_{n-1,j}, z_{nl})}. \quad (13.19)$$

对于观测变量的分布参数，与其具体分布形式相关，如果是高斯：$p(\mathbf{x}|\boldsymbol{\phi}_k) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$，对应的最优解为：

$$\boldsymbol{\mu}_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n}{\sum_{n=1}^{N} \gamma(z_{nk})} \quad (13.20)$$

$$\boldsymbol{\Sigma}_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk}) (\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}}{\sum_{n=1}^{N} \gamma(z_{nk})}. \quad (13.21)$$

如果是离散：$p(\mathbf{x}|\mathbf{z}) = \prod_{i=1}^{D} \prod_{k=1}^{K} \mu_{ik}^{x_i z_k}$ 对应的最优解为：

$$\mu_{ik} = \frac{\sum_{n=1}^{N} \gamma(z_{nk}) x_{ni}}{\sum_{n=1}^{N} \gamma(z_{nk})}. \quad (13.23)$$

好，M 步就这样， 现在来看 E 步，也是 HMM 比较核心的地方。刚才我们看到，E 步需要求的是：

$$\gamma(\mathbf{z}_n) = p(\mathbf{z}_n|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \quad (13.13)$$

$$\xi(\mathbf{z}_{n-1}, \mathbf{z}_n) = p(\mathbf{z}_{n-1}, \mathbf{z}_n|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}). \quad (13.14)$$

由马尔科夫的性质，我们可以推出：

$$\gamma(\mathbf{z}_n) = \frac{p(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{z}_n)p(\mathbf{x}_{n+1}, \ldots, \mathbf{x}_N|\mathbf{z}_n)}{p(\mathbf{X})} = \frac{\alpha(\mathbf{z}_n)\beta(\mathbf{z}_n)}{p(\mathbf{X})} \tag{13.33}$$

其中：

$$\alpha(\mathbf{z}_n) \equiv p(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{z}_n) \tag{13.34}$$
$$\beta(\mathbf{z}_n) \equiv p(\mathbf{x}_{n+1}, \ldots, \mathbf{x}_N|\mathbf{z}_n). \tag{13.35}$$

接下来我们就建立 $\alpha(\mathbf{z}_n)$ 和 $\beta(\mathbf{z}_n)$ 的递推公式

$$\begin{aligned}
\alpha(\mathbf{z}_n) &= p(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{z}_n) \\
&= p(\mathbf{x}_1, \ldots, \mathbf{x}_n|\mathbf{z}_n)p(\mathbf{z}_n) \\
&= p(\mathbf{x}_n|\mathbf{z}_n)p(\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}|\mathbf{z}_n)p(\mathbf{z}_n) \\
&= p(\mathbf{x}_n|\mathbf{z}_n)p(\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}, \mathbf{z}_n) \\
&= p(\mathbf{x}_n|\mathbf{z}_n)\sum_{\mathbf{z}_{n-1}} p(\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}, \mathbf{z}_{n-1}, \mathbf{z}_n) \\
&= p(\mathbf{x}_n|\mathbf{z}_n)\sum_{\mathbf{z}_{n-1}} p(\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}, \mathbf{z}_n|\mathbf{z}_{n-1})p(\mathbf{z}_{n-1}) \\
&= p(\mathbf{x}_n|\mathbf{z}_n)\sum_{\mathbf{z}_{n-1}} p(\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}|\mathbf{z}_{n-1})p(\mathbf{z}_n|\mathbf{z}_{n-1})p(\mathbf{z}_{n-1}) \\
&= p(\mathbf{x}_n|\mathbf{z}_n)\sum_{\mathbf{z}_{n-1}} p(\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}, \mathbf{z}_{n-1})p(\mathbf{z}_n|\mathbf{z}_{n-1})
\end{aligned}$$

Making use of the definition (13.34) for $\alpha(\mathbf{z}_n)$, we then obtain

$$\alpha(\mathbf{z}_n) = p(\mathbf{x}_n|\mathbf{z}_n)\sum_{\mathbf{z}_{n-1}} \alpha(\mathbf{z}_{n-1})p(\mathbf{z}_n|\mathbf{z}_{n-1}). \tag{13.36}$$

其中：

$$\alpha(\mathbf{z}_1) = p(\mathbf{x}_1, \mathbf{z}_1) = p(\mathbf{z}_1)p(\mathbf{x}_1|\mathbf{z}_1) = \prod_{k=1}^{K} \{\pi_k p(\mathbf{x}_1|\boldsymbol{\phi}_k)\}^{z_{1k}} \tag{13.37}$$

这样我们从 $\alpha(\mathbf{z}_1)$ 开始，可以递推出所有的 $\alpha(\mathbf{z}_n)$，对于 $\beta(\mathbf{z}_n)$，也进行类似的推导：

$$\begin{aligned}
\beta(\mathbf{z}_n) &= p(\mathbf{x}_{n+1}, \ldots, \mathbf{x}_N|\mathbf{z}_n) \\
&= \sum_{\mathbf{z}_{n+1}} p(\mathbf{x}_{n+1}, \ldots, \mathbf{x}_N, \mathbf{z}_{n+1}|\mathbf{z}_n) \\
&= \sum_{\mathbf{z}_{n+1}} p(\mathbf{x}_{n+1}, \ldots, \mathbf{x}_N|\mathbf{z}_n, \mathbf{z}_{n+1})p(\mathbf{z}_{n+1}|\mathbf{z}_n) \\
&= \sum_{\mathbf{z}_{n+1}} p(\mathbf{x}_{n+1}, \ldots, \mathbf{x}_N|\mathbf{z}_{n+1})p(\mathbf{z}_{n+1}|\mathbf{z}_n) \\
&= \sum_{\mathbf{z}_{n+1}} p(\mathbf{x}_{n+2}, \ldots, \mathbf{x}_N|\mathbf{z}_{n+1})p(\mathbf{x}_{n+1}|\mathbf{z}_{n+1})p(\mathbf{z}_{n+1}|\mathbf{z}_n).
\end{aligned}$$

$$\beta(\mathbf{z}_n) = \sum_{\mathbf{z}_{n+1}} \beta(\mathbf{z}_{n+1})p(\mathbf{x}_{n+1}|\mathbf{z}_{n+1})p(\mathbf{z}_{n+1}|\mathbf{z}_n). \tag{13.38}$$

从上式可以看到 $\beta(\mathbf{z}_n)$ 是一个逆推过程，所以我们需要初始值 $\beta(\mathbf{z}_N)$，定义 13.35 并没有明确 $\beta(\mathbf{z}_N)$ 的定义：

$$\alpha(\mathbf{z}_n) \equiv p(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{z}_n) \tag{13.34}$$
$$\beta(\mathbf{z}_n) \equiv p(\mathbf{x}_{n+1}, \ldots, \mathbf{x}_N | \mathbf{z}_n). \tag{13.35}$$

因为 z_N 后没有观测数据，不过我们可以从

$$\gamma(\mathbf{z}_n) = \frac{p(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{z}_n) p(\mathbf{x}_{n+1}, \ldots, \mathbf{x}_N | \mathbf{z}_n)}{p(\mathbf{X})} = \frac{\alpha(\mathbf{z}_n)\beta(\mathbf{z}_n)}{p(\mathbf{X})} \tag{13.33}$$

，得出:

$$p(\mathbf{z}_N | \mathbf{X}) = \frac{p(\mathbf{X}, \mathbf{z}_N)\beta(\mathbf{z}_N)}{p(\mathbf{X})} \tag{13.39}$$

这样 $\beta(\mathbf{z}_N)$ 就等于 1，现在我们可以方便的求出所有的 $\alpha(\mathbf{z}_n)$ 和 $\beta(\mathbf{z}_n)$ 了，利用 13.13 也就可以求出所有的 $\gamma(\mathbf{z}_n)$。类似的，我们可以求出 $\xi(\mathbf{z}_{n-1}, \mathbf{z}_n)$：

$$\begin{aligned}
\xi(\mathbf{z}_{n-1}, \mathbf{z}_n) &= p(\mathbf{z}_{n-1}, \mathbf{z}_n | \mathbf{X}) \\
&= \frac{p(\mathbf{X} | \mathbf{z}_{n-1}, \mathbf{z}_n) p(\mathbf{z}_{n-1}, \mathbf{z}_n)}{p(\mathbf{X})} \\
&= \frac{p(\mathbf{x}_1, \ldots, \mathbf{x}_{n-1} | \mathbf{z}_{n-1}) p(\mathbf{x}_n | \mathbf{z}_n) p(\mathbf{x}_{n+1}, \ldots, \mathbf{x}_N | \mathbf{z}_n) p(\mathbf{z}_n | \mathbf{z}_{n-1}) p(\mathbf{z}_{n-1})}{p(\mathbf{X})} \\
&= \frac{\alpha(\mathbf{z}_{n-1}) p(\mathbf{x}_n | \mathbf{z}_n) p(\mathbf{z}_n | \mathbf{z}_{n-1}) \beta(\mathbf{z}_n)}{p(\mathbf{X})}
\end{aligned} \tag{13.43}$$

这样在 M 步里求解所需要的分布就都求出来了，也就可以用 EM 来学习 HMM 的参数了，这里式子比较多，大家自己推一下会比较好理解。

第一个学习问题就这样了，接下来是预测问题，预测问题可以直接推导：

$$\begin{aligned}
p(\mathbf{x}_{N+1} | \mathbf{X}) &= \sum_{\mathbf{z}_{N+1}} p(\mathbf{x}_{N+1}, \mathbf{z}_{N+1} | \mathbf{X}) \\
&= \sum_{\mathbf{z}_{N+1}} p(\mathbf{x}_{N+1} | \mathbf{z}_{N+1}) p(\mathbf{z}_{N+1} | \mathbf{X}) \\
&= \sum_{\mathbf{z}_{N+1}} p(\mathbf{x}_{N+1} | \mathbf{z}_{N+1}) \sum_{\mathbf{z}_N} p(\mathbf{z}_{N+1}, \mathbf{z}_N | \mathbf{X}) \\
&= \sum_{\mathbf{z}_{N+1}} p(\mathbf{x}_{N+1} | \mathbf{z}_{N+1}) \sum_{\mathbf{z}_N} p(\mathbf{z}_{N+1} | \mathbf{z}_N) p(\mathbf{z}_N | \mathbf{X}) \\
&= \sum_{\mathbf{z}_{N+1}} p(\mathbf{x}_{N+1} | \mathbf{z}_{N+1}) \sum_{\mathbf{z}_N} p(\mathbf{z}_{N+1} | \mathbf{z}_N) \frac{p(\mathbf{z}_N, \mathbf{X})}{p(\mathbf{X})} \\
&= \frac{1}{p(\mathbf{X})} \sum_{\mathbf{z}_{N+1}} p(\mathbf{x}_{N+1} | \mathbf{z}_{N+1}) \sum_{\mathbf{z}_N} p(\mathbf{z}_{N+1} | \mathbf{z}_N) \alpha(\mathbf{z}_N)
\end{aligned} \tag{13.44}$$

现在就剩最后一个解码问题，也就是 argmax_Z{p(Z|X)}，刚才我们在 E 步已经求出了：

$$\gamma(\mathbf{z}_n) = p(\mathbf{z}_n|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \quad (13.13)$$

$$\xi(\mathbf{z}_{n-1}, \mathbf{z}_n) = p(\mathbf{z}_{n-1}, \mathbf{z}_n|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}). \quad (13.14)$$

但是现在的问题要复杂一点，因为我们要求概率最大的隐变量序列，用 13.13 可以求出单个隐变量，但是他们连在一起形成的整个序列可能概率很小，这个问题可以归结为一个动态规划：

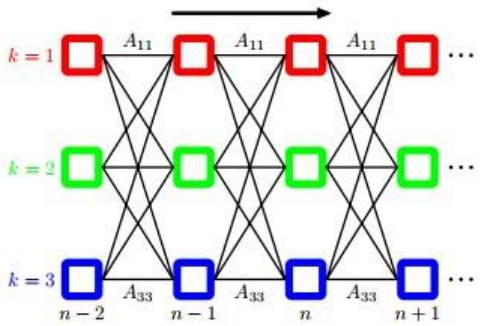

Figure 13.7 If we unfold the state transition diagram of Figure 13.6 over time, we obtain a lattice, or trellis, representation of the latent states. Each column of this diagram corresponds to one of the latent variables $\mathbf{z}_n$.

我们把 HMM 化成如上图的样子，最大化后验等价于最大化全概率，对于上图中的边，我们赋值为：

log ($p(\mathbf{z}_n|\mathbf{z}_{n-1}, \mathbf{A}) = \prod_{k=1}^{K}\prod_{j=1}^{K} A_{jk}^{z_{n-1,j}z_{nk}}$)

初始节点赋值为：

log($p(\mathbf{z}_1|\boldsymbol{\pi}) = \prod_{k=1}^{K} \pi_k^{z_{1k}}$ *p(x_1|z_1))

其余节点赋值为：

log($p(\mathbf{x}_n|\mathbf{z}_n, \boldsymbol{\phi}) = \prod_{k=1}^{K} p(\mathbf{x}_n|\boldsymbol{\phi}_k)^{z_{nk}}$)

这样任何一个序列 Z，其全概率等于 exp（Z 对应路径上节点和边的值求和），这样，解码问题就转化为一个最长路径问题，用动态规划可以直接求解。大家看这里有没有问题，HMM 的主要内容就是这些

接下来的 Linear Dynamical Systems 其实和 HMM 大同小异，只是把离散分布换成了高斯，然后就是第二章公式的反复应用，都是细节问题，就不在这里讲了，大家看看有问题我们可以讨论。这一章还是式子主导的，略过了不少式子，大家推的时候有问题我们可以随时讨论。

天涯游(872352128) 21:09:21

我对 hmm 的理解，觉得这麻烦的是概率的理解的了，概率分解才是 hmm 的核心，当然了还有动态规划了。概率分解其实是实验事件的分解，如前向 和后向了，还有就是 EM 算法了。