HMM

Li Liang*

1 HMM

HMM 是结构最简单的动态贝叶斯网，是一种有向图模型。HMM 中的变量可分为两组。第一组为状态变量 (隐变量) $\{y_t\}_{t=1}^n$，y_t 为 t 时刻的系统状态，其取值范围叫做状态空间（$\mathcal{Y}=\{s_i\}_{i=1}^N$）；第二组为观测变量组 $\{x_t\}_{t=1}^n$，x_t 为 t 时刻的观测值，其取值范围叫做观测空间（$\mathcal{X}=\{o_i\}_{i=1}^M$）。

HMM 模型的两个假设为：

- 齐次马尔科夫链假设：当前状态仅依赖于前一时刻的状态；
- 观测独立性假设：当前观测变量仅依赖于当前状态。

基于上述假设，所有变量的联合概率分布为：

$$P(x_1,x_2,...,x_n,y_1,y_2,...,y_n)=P(y_1)P(x_1\mid y_1)P(y_2\mid y_1)...P(x_t\mid y_t)P(y_t\mid y_{t-1}) \quad (1)$$

上式共有三组参数：

- 初始状态概率:初始时刻各状态出现的概率 ($P(y_1)$),记为 $\pi=(\pi_1,\pi_2,...,\pi_N)$，其中 $\pi_i=P(y_1=s_i)$，即初始状态为 s_i 的概率。
- 状态转移概率：当前状态与前一状态的关系 ($P(y_t\mid y_{t-1})$)，记为 $\mathbb{A}=[a_{ij}]_{N\times N}$，其中 $a_{ij}=P(y_t=s_j\mid y_{t-1}=s_i)$，即前一状态为 s_i，当前状态为 s_j 的概率。
- 输出观测概率：根据当前状态获得各个观测值的概率 ($P(x_t\mid y_t)$)，记为 $\mathbb{B}=[b_{ij}]_{N\times M}$，其中 $b_{ij}=P(x_t=o_j\mid y_t=s_i)$，即若状态为 s_i，则观测 o_j 被获取的概率。

通过指定状态空间 $\mathcal{Y}$、观测空间 $\mathcal{X}$ 和上述三组参数 $\lambda=[\mathbb{A},\mathbb{B},\pi]$，就能确定一个 HMM 模型。

*https://github.com/leeliang/

2 观测序列概率求取

利用 HMM 解决的第一类问题为观测序列的概率计算，即已知 HMM 模型的参数 $\lambda = [\mathbb{A}, \mathbb{B}, \pi]$ 和观测序列 $X = \{x_t\}_{t=1}^n$，求取观测序列在模型下出现的条件概率 $p(X \mid \lambda)$。

该问题可以考虑所有状态变量序列 $Y^k = \{y_t^k\}_{t=1}^n$，计算边缘概率进行求解：

$$p(X \mid \lambda) = \sum_{Y^k \in \mathcal{Y}} p(X, Y^k \mid \lambda) \tag{2}$$

对于每个状态变量序列，

$$p(X, Y^k \mid \lambda) = p(Y^k \mid \lambda) p(X \mid Y^k, \lambda) \tag{3}$$

其中，

$$\begin{aligned} p(Y^k \mid \lambda) &= \pi^k a_{12}^k a_{23}^k ... a_{t\ t-1}^k \\ p(X \mid Y^k, \lambda) &= \pi^k b_{1x_1}^k a_{12}^k b_{2x_2}^k a_{23}^k b_{3x_3}^k ... a_{t-1\ t}^k b_{t\ x_t}^k \end{aligned} \tag{4}$$

其中 a_{12}^k 表示状态变量从 y_1^k 变为 y_2^k 的概率，其他类似。若状态变量的取值数为 N，则隐藏变量序列的可能情况有 N^t 种，该算法的计算量为 $O(tN^t)$，若隐藏变量的取值数较多，该算法耗时较大，需要采用简洁的算法。

2.1 前向算法

定义前向概率为：

$$\alpha_t(i) = p(x_1, x_2, ..., x_t, y_t = s_i \mid \lambda) \tag{5}$$

即在给定 HMM 模型下，t 时刻的观测序列为 $x_1, x_2, ..., x_t$ 且隐藏状态为 s_i 的概率。

前向算法本质上属于动态规划的算法，即把多阶段过程转化为一系列单阶段问题，利用各阶段之间的关系，按顺序求解子阶段，前一子问题的解，为后一子问题的求解提供了实用的信息，从而减少了计算量。前向算法的各阶段即为观测序列中的每个观测值。所以求解的第一个子阶段问题为 $p(x_1 \mid \lambda)$：

$$p(x_1 \mid \lambda) = \sum_{i=1}^N \alpha_1(i) = \sum_{i=1}^N \pi_i b_{i\ x_1} \tag{6}$$

即求取边缘概率。第二个子问题为 $p(x_1, x_2 \mid \lambda)$，同样地即为：

$$p(x_1, x_2 \mid \lambda) = \sum_{i=1}^N \alpha_2(i) \tag{7}$$

根据动态规划算法的特点，后一子问题可以采用前一子问题的结果进行计算，从而减少计算量，现在的问题为要推导上述两个问题的关系，上两个问题的关系即 $\alpha_2(i)$ 与 $\alpha_1(i)$ 的关系。更一般地，需要知道 $\alpha_{t-1}(i)$ 和 $\alpha_t(i)$ 的关系。

$$\begin{aligned}\alpha_{t-1}(i) &= p(x_1, x_2, ..., x_{t-1}, y_{t-1} = s_i \mid \lambda) \\ \alpha_t(i) &= p(x_1, x_2, ..., x_{t-1}, x_t, y_t = s_i \mid \lambda)\end{aligned} \tag{8}$$

下面根据 HMM 模型的性质进行推导。

$$\begin{aligned}\alpha_t(i) &= p(x_1, x_2, ..., x_{t-1}, x_t, y_t = s_i \mid \lambda) \\ &= p(x_1, x_2, ..., x_{t-1}, y_t = s_i \mid \lambda) b_{ix_t} \\ &= \sum_{j=1}^{N} p(x_1, x_2, ..., x_{t-1}, y_{t-1} = s_j \mid \lambda) a_{ji} b_{ix_t} \\ &= \sum_{j=1}^{N} \alpha_{t-1}(j) a_{ji} b_{ix_t}\end{aligned} \tag{9}$$

根据上述推导，每一子阶段的计算为：

$$p(x_1, x_2, ..., x_t \mid \lambda) = \sum_{i=1}^{N} \alpha_t(i) = \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_{t-1}(j) a_{ji} b_{ix_t} \tag{10}$$

最后一个子阶段的结果即为最终需要计算的 $p(X \mid \lambda) = p(x_1, x_2, ..., x_t \mid \lambda)$。每个子阶段的计算量为 $O(N^2)$，共 t 个子阶段，前向算法的复杂度为 $O(N^2 t)$，计算量减少的原因为后一子问题采用了前一子问题的结果进行计算。

3 后向算法

定义后向概率为：

$$\beta_t(i) = p(x_{t+1}, x_{t+2}, ..., x_T \mid y_t = s_i \lambda) \tag{11}$$

即在给定 HMM 模型，t 时刻的隐藏状态为 s_i 的条件下，从时刻 $t+1$ 到最后时刻 T 的观测序列为 $(x_{t+1}, x_{t+2}, ..., x_T)$ 的概率。

与前向算法推导类似，区别为从后往前推导后向概率。最后时刻 T 的后向概率为：

$$\beta_T(i) = 1, i = 1, 2, ..., N \tag{12}$$

从 $t+1$ 推导到 t 时刻的关系为：

$$\begin{aligned}
\beta_t(i) &= p(x_{t+1}, x_{t+2}, ..., x_T \mid y_t = s_i, \lambda) \\
&= \sum_{j=1}^{N} p(x_{t+1}, x_{t+2}, ..., x_T \mid y_{t+1} = s_j, \lambda) a_{ji} \\
&= \sum_{j=1}^{N} [p(x_{t+2}, x_{t+3}, ..., x_T \mid y_{t+1} = s_j, \lambda) b_{j\ x_{t+1}}] a_{ji} \\
&= \sum_{j=1}^{N} \beta_{t+1}(j) b_{j\ x_{t+1}} a_{ji}
\end{aligned} \tag{13}$$

从 $T-1$ 一直递推到 1 时刻，最终结果为：

$$p(X \mid \lambda) = \sum_{i=1}^{N} \pi_i b_{i\ x_1} \beta_1(i) \tag{14}$$

4 模型参数求解

4.1 监督学习方法

已知 L 个长度为 T 的观测序列和隐藏状态序列（$\{(X_1, Y_1), ..., (X_L, Y_L), \}$），求解 HMM 模型参数。该问题可以直接采用极大似然法估计。

假设样本从 s_i 转移至 s_j 的频数为 $\mathcal{A}_{ij}$，则状态转移概率：

$$a_{ij} = \frac{\mathcal{A}_{ij}}{\sum\limits_{n=1}^{N} \mathcal{A}_{in}} \tag{15}$$

假设样本隐藏状态为 s_i 并观测为 o_j 的频数为 $\mathcal{B}_{ij}$，则状态转移概率：

$$b_{ij} = \frac{\mathcal{B}_{ij}}{\sum\limits_{n=1}^{N} \mathcal{B}_{in}} \tag{16}$$

假设样本初始隐藏状态为 s_i 的频数为 $\mathcal{C}_i$，则状态转移概率：

$$\pi = \frac{\mathcal{C}_i}{\sum\limits_{n=1}^{N} \mathcal{C}_n} \tag{17}$$

4.2 Baum-Welch 算法

上述情况是隐藏状态已知，若隐藏状态未知，则采用 Baum-Welch 算法求解参数，该算法实际上就是 EM 算法。根据 EM 算法，该问题为：

- E-step：固定模型参数为 $\hat{\lambda}$，计算：

$$\mathcal{L}(\lambda \mid \hat{\lambda}) = \sum_{y_i} p(y_i \mid X, \hat{\lambda}) \ \log p(X, y_i \mid \hat{\lambda})$$

- M-step：最大化 $\mathcal{L}(\lambda \mid \hat{\lambda})$：

$$\lambda = \arg\max_{\lambda} \mathcal{L}(\lambda \mid \hat{\lambda})$$

5 隐藏状态序列求解

5.1 近似算法

求取每个时刻 t 最有可能的状态 $s_i(t)$，从而得到一个隐藏状态序列 $(s_i(1), ..., s_i(T))$。给定 HMM 模型参数和观测序列，在时刻 t 处于状态 s_i 的概率：

$$\gamma_t(i) = p(y_t = s_i \mid \lambda, X) = \frac{p(y_t = s_i, X \mid \lambda)}{p(X \mid \lambda)} \tag{18}$$

根据前向和后向概率的定义：

$$\gamma_t(i) = \frac{p(y_t = s_i, X \mid \lambda)}{p(X \mid \lambda)} = \frac{\alpha_t(i)\beta_t(i)}{\sum\limits_{j=1}^{N} \alpha_t(j)\beta_t(j)} \tag{19}$$

则每个时刻 t 最有可能的状态 $s_i(t) = \arg\max \gamma_t(i)$。近似算法的优点是计算简单，但得到的隐藏状态序列可能是不存在的序列，因为得到的隐藏序列中，有可能存在两个相邻状态的转移概率为 0，即不可能转移过去。

5.2 维特比算法

维特比算法实际上是用动态规划解决问题，即用动态规划求解最大概率隐藏序列。假设隐藏状态有 N 个取值，则隐藏序列有 N^T 种，我们将一个隐藏序列对应着一条从开始时刻到 T 时刻的路径。

- 如果概率最大的路径 (P) 经过某个点（如 t 时刻的隐藏状态 $s_k(t)$），那么这条路径上从起始点到 $s_k(t)$ 的这一段子路径，一定是起始点到 $s_k(t)$ 的最大概率路径，否则这个子路径可以被其他子路径替代，得到更大的 P。

- 如果记录了从起始点到 t 时刻的所有 N 个隐藏状态的最大概率路径（N 条 P_t），最终的最大概率路径必经过其中的一条。这样，在任何时刻，只需要考虑非常有限条最大概率路径即可。

- 结合上述两点，假定当我们从 t 时刻到 $t+1$ 时刻，只需要在 N 条 P_t 的基础上，计算 t 时刻的 N 个节点到 $t+1$ 时刻某个状态的最大概率。

根据上述基础，定义两个变量，记从起始点到 t 时刻 s_i 隐藏状态的最大概率为 $\delta_t i$ (对应的路径为 $\mathrm{P}_t(i)$)，记 $\mathrm{P}_t(i)$ 上 $t-1$ 时刻的隐藏状态为 $\Psi_t(i)$。根据定义，两个变量的递推关系为：

$$
\begin{aligned}
\delta_{t+1}(i) &= \max_{1 \le j \le N} [\delta_t(j) a_{ji}] b_{i\ x_{t+1}} \\
\Psi_t(i) &= \arg \max_{1 \le j \le N} [\delta_{t-1}(j) a_{ji}]
\end{aligned}
\tag{20}
$$

有了这两个变量，可以从起始时刻递推到最终时刻，然后利用 $\Psi_t(i)$ 记录的前一个最可能隐藏状态节点回溯，直到回溯到起始时刻。