HMM Algorithm

Shen Hengheng

2017

很早就想写一篇关于**隐马尔可夫模型**的文章了，这次刻意的将模型及其有关算法复习了一下，才有了这个信心去写了这篇文章。这篇文章主要参考了李航老师的统计机器学习部分内容和徐亦达老师在油管上的教程[1]。并且这里的隐马尔可夫模型主要是指**离散形式的动态模型**。

隐马尔可夫模型主要是一种**可用于标注问题的统计学习模型**，描述由隐藏的马尔可夫链随机生成观测序列的过程，属于生成模型。近些年来主要用于语音信号处理，自然语言处理，生物信息，金融分析等领域，该教程涉及很多概率计算问题，所以希望读者能够有概率背景的情况下，阅读更佳。

本部分主要介绍一下几个部分：

- HMM 现象
- 马尔可夫过程/链
- 算法
  - 直接计算法
  - 前向-后向算法
    - 前向算法
    - 后向算法
- 学习算法
- 总结

0.1 HMM 现象

玩股票的朋友都知道，图1右，在市场规则下，股民只知道股票的涨停和不变这三种表面现象，而人们并不希望仅仅知道市场的涨停，而是想知道市场隐藏的信息，比如现在股票市场是熊市还是牛市，因为知道这些隐藏的信息之后，我们可以利用这些信息去长期跟进还是及时退出，以防市场大变。在这里我们称这种人们表面上能观察出来的现象为观测（observation），而那些隐藏的变量被称作状态。

有了上面比较直观的介绍以后，我们就可以定义什么是隐马尔可夫模型了。

定义 10.1（隐马尔可夫模型） 隐马尔可夫模型是关于时序的概率模型，描述由一个隐藏的马尔可夫链随机生成不可观测的状态随机序列，再由各个状态生成一个观测而产生观测随机序列的过程。隐藏的马尔可夫链随机生成的状态的序列，称为状态序列（state sequence）；每个状态生成一个观测，而由此产生的观测的随机序列，称为观测序列（observation sequence）。序列的每一个位置又可以看作是一个时刻。

[1] https://www.youtube.com/watch?v=Ji6KbkyNmk8list=PLyAft-JyjIYoc9LN241WKqLPuggfSBBpt

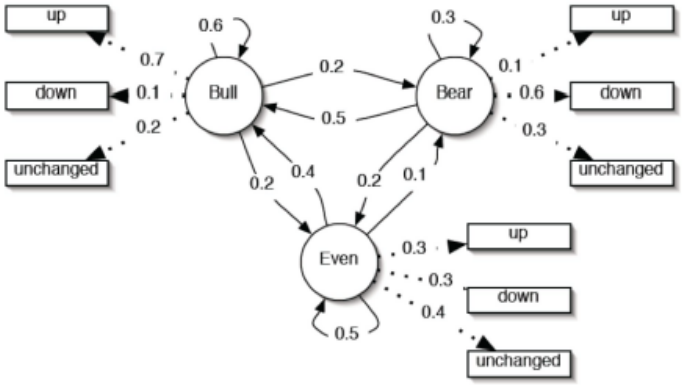

图 1: HMM 现象

比如：上图的 Bull->Bear->Event 就是其中一个状态序列，而其产生的观测形成的序列 up->down->unchange 被称作观测序列，而这些序列在一定假设下，具有非常好的概率性质。

0.2 马尔可夫链

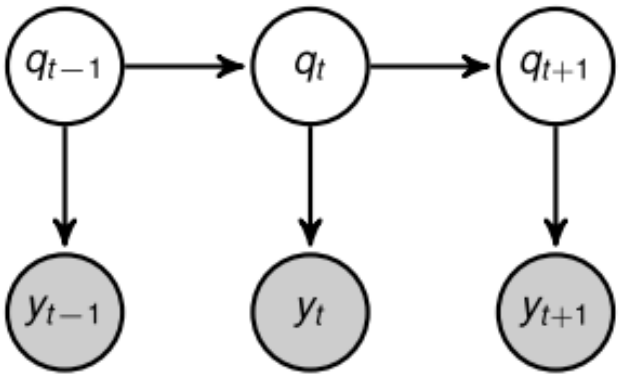

图 2: HMM 的概率图模型

上图2是隐马尔可夫的概率图模型，概率图反应了事件的依赖和独立关系。

首先我们要说明 q 表示状态，不可观测。y 表示可观测的现象。假设状态与状态之间、状态与观测现象之间均满足马尔可夫过程，其中该过程要求具备“无记忆”的性质（马尔可夫性质）：下一状态的概率分布只能由当前状态决定，在时间序列中前面的事件均与之无关。用数学语言表示为：

$$\Pr(X_{n+1} = x \mid X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \Pr(X_{n+1} = x \mid X_n = x_n)$$

那么根据概率图模型，我们得出了两个重要的公式：

离散状态转移概率：

$$p(q_t|q_1, ..., q_{t-1}, y_1, ..., y_{t-1}) = p(q_t|q_{t-1})$$

离散/连续观测概率：

$$p(y_t|q_1, ..., q_{t-1}, y_1, ..., y_{t-1}) = p(y_t|q_t)$$

那么有了这个公式我们可以计算出图2的转移概率矩阵，计算过程见图3

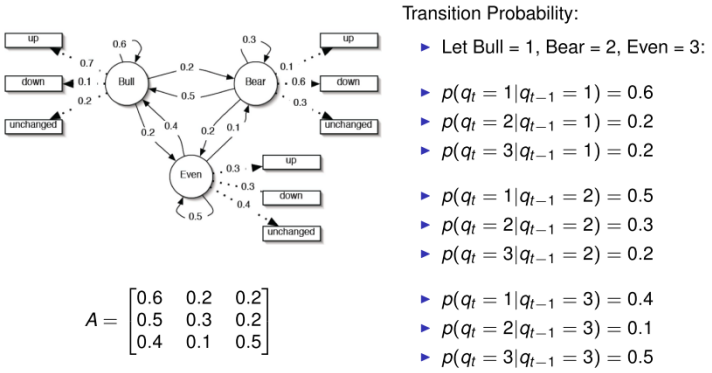

图 3: 状态转移概率计算

A 是状态转移概率矩阵,A 描述了状态之间转换关系及其分布。

其中 $a_{ij} = P(i_{t+1} = q_j|i_t = q_i),\ \ i = 1, 2, ..., N; j = 1, 2, ..., N;$，是在时刻 t 处于状态 q_i 的条件下在时刻 $t + 1$ 转移到状态 q_j 的概率。

同样地，我们可以计算出观测概率矩阵，具体计算过程见4

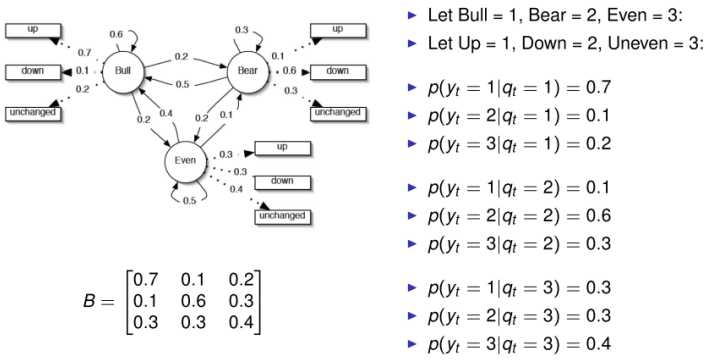

图 4: 观测概率的计算

B 是观测概率矩阵, 其中 $b_j(k) = P(o_t = v_k|i_t = q_j)\ \ k = 1, 2, ..., M; j = 1, 2, ..., N;$，是在时刻 t 处于状态 q_j 的条件下生成观测 v_k 的概率。

有了这些性质之后我们就可以计算观测序列的概率了。比如利用图3和图4计算 $Pr(y_1 = up, y_2 = up, y_3 = down)$ 的概率.

$$
\begin{aligned}
P(y_1, y_2, y_3) &= \sum_{q_1=1}^{k}\sum_{q_2=1}^{k}\sum_{q_3=1}^{k} P(y_1, y_2, y_3, q_1, q_2, q_3) \\
&= \sum_{q_1=1}^{k}\sum_{q_2=1}^{k}\sum_{q_3=1}^{k} P(y_3|y_1, y_2, q_1, q_2, q_3)P(y_1, y_2, q_1, q_2, q_3) \\
&= \sum_{q_1=1}^{k}\sum_{q_2=1}^{k}\sum_{q_3=1}^{k} P(y_3|q_3)P(y_1, y_2, q_1, q_2, q_3) \\
&= \sum_{q_1=1}^{k}\sum_{q_2=1}^{k}\sum_{q_3=1}^{k} P(y_3|q_3)P(q_3|y_1, y_2, q_1, q_2)P(y_1, y_2, q_1, q_2) \\
&= \sum_{q_1=1}^{k}\sum_{q_2=1}^{k}\sum_{q_3=1}^{k} P(y_3|q_3)P(q_3|q_2)P(y_1, y_2, q_1, q_2) \\
&= \sum_{q_1=1}^{k}\sum_{q_2=1}^{k}\sum_{q_3=1}^{k} P(y_3|q_3)P(q_3|q_2)P(y_2|y_1, q_1, q_2)P(y_1, q_1, q_2) \\
&= \sum_{q_1=1}^{k}\sum_{q_2=1}^{k}\sum_{q_3=1}^{k} P(y_3|q_3)P(q_3|q_2)P(y_2|q_2)P(y_1, q_1, q_2) \\
&= \sum_{q_1=1}^{k}\sum_{q_2=1}^{k}\sum_{q_3=1}^{k} P(y_3|q_3)P(q_3|q_2)P(y_2|q_2)P(q_2|q_1)P(y_1|q_1)P(q_1)
\end{aligned}
$$

但是发现除了 A 和 B 我们已知，还无法计算 $P(y_1, y_2, y_3)$，还需要知道 $P(q_1)$ 的分布情况，所以在这里引出了另外一个条件，$P(q_1)$ 的分布。也就是初始状态概率。为了符号化，这里 π 是初始状态概率向量 $\pi = (\pi_i)$：其中 $\pi_i = P(i_1 = q_i)$，是时刻 t = 1 处于状态 q_i 的概率。

总结： 隐马尔可夫模型由初始状态概率向量 π、状态转移概率矩阵 A 和观测概率矩阵 B 决定。π 和 A 决定状态序列，B 决定观测序列。因此，隐马尔可夫模型 λ 可以用三元符号表示，即

$$\lambda = (A, B, \pi)$$

A,B,π 称为隐马尔可夫模型的三要素。

0.3 算法

HMM 的三个基本计算问题：

$$
\begin{aligned}
&\text{Evaluate } \ p(Y|\lambda) \\
&\lambda_{MLE} = \text{argmax}_{\lambda} p(Y|\lambda) \\
&\text{argmax}_{Q} p(Y|Q, \lambda)
\end{aligned}
$$

0.4 直接计算法

首先解决第一个问题，Evaluate $p(Y|\lambda)$,

$$\begin{aligned}
P(Y|\lambda) &= \sum_{Q}[p(Y,Q|\lambda)] = \sum_{q_1=1}^{k}...,\sum_{q_T=1}^{k}[p(y_1,...,y_T,q_1,...,q_T|\lambda)] \\
&= \sum_{q_1=1}^{k}...,\sum_{q_T=1}^{k}[p(y_1,...,y_T,q_1,...,q_T|\lambda)] \\
&= \sum_{q_1=1}^{k}...,\sum_{q_3=1}^{k}p(q_1)p(y_1|q_1)p(q_2|q_1)...p(q_t|q_{t-1}p(y_t|q_t)) \\
&= \sum_{q_1=1}^{k}...,\sum_{q_3=1}^{k}\pi(p_1)\ \amalg_{t=2}^{T}\ a_{q_{t-1},q_t}b_{q_t}(y_t)
\end{aligned}$$

其中:

$$\begin{aligned}
p(q_t = j|q_{t-1} = i) &\equiv a_{i,j} \\
p(y_t|q_t = j) &\equiv bj(y_t)
\end{aligned}$$

但是，利用上面公式计算量很大，是 $O(TN^T)$ 阶的，这种算法不可行。

0.5 前向-后向算法

0.6 前向算法

定义 10.2(前向概率)给定隐马尔可夫模型,定义到时刻 t 部分观测序列为 $o_1, o_2, ..., o_t$ 且状态为 q_i 的概率为前向概率，记作

$$\alpha_t(i) = P(o_1, o_2, ..., o_t, i_t = q_t|\lambda)$$

下面是前向算法的概率图模型。

前向过程：

$$\alpha_t(i) = P(o_1, o_2, ..., o_t, i_t = q_t|\lambda) \Longrightarrow p(Y|\lambda) = \sum_{i=1}^{k}\alpha_i(T)$$

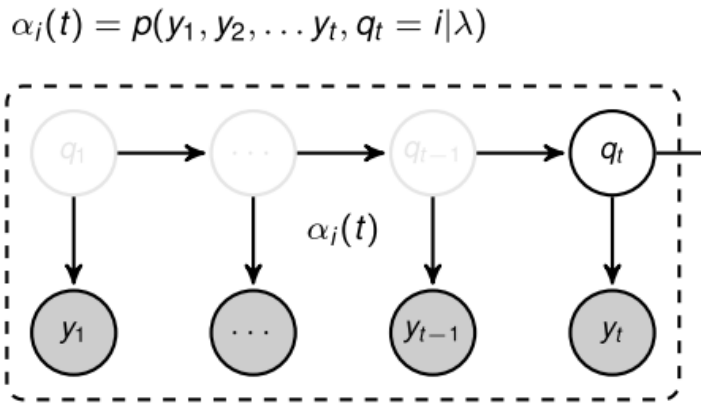

图 5: 前向算法的概率图模型

递推: 对 $t = 1, 2, ..., T-1$，

$$\alpha_i(1) = p(y_1, q_1 = i|\lambda) = p(q_1)p(y_1|q_1) = \pi_i b_i(y_1)$$

$$\alpha_j(2) = p(y_1, y_2, q_2 = j|\lambda) = \sum_{i=1}^{k} p(q_1 = i)p(y1|q_1 = i)p(q_2 = i|q_1 = i)p(y_2|q_2 = j)$$

$$= [\sum_{i=1}^{k} \alpha_i(1)\alpha_{i,j}]b_j(y_2) = p(q_1)p(y_1|q_1) = \pi_i b_i(y_1)$$

...

$$\alpha_j(t+1) = [\sum_{i=1}^{k} \alpha_i(t)\alpha_{i,j}]b_j(y_{t+1})$$

...

$$\alpha_j(T) = [\sum_{i=1}^{k} \alpha_i(T-1)\alpha_{i,j}]b_j(y_T)$$

0.7 后向算法

定义 10.3（后向概率）给定隐马尔可夫模型，定义在时刻 t 状态为 q_i 的条件下，从 t+1 到 T 的部分观测序列为 o_{t+1},o_{t+2},$\cdots$,o_T 的概率为后向概率，记作

$$\beta_t(i) = P(o_{t+1}, o_{t+2}, o_{t+3}, ..., o_T|i_t = q_i, \lambda)$$

可以用递推的方法求得后向概率 $\beta_t(i)$ 及观测序列概率 $P(O|\lambda)$。

$$\beta_i(t) = p(y_{t+1}, ..., y_T|q_t = i, \lambda)$$

下面是后向算法的概率图模型。

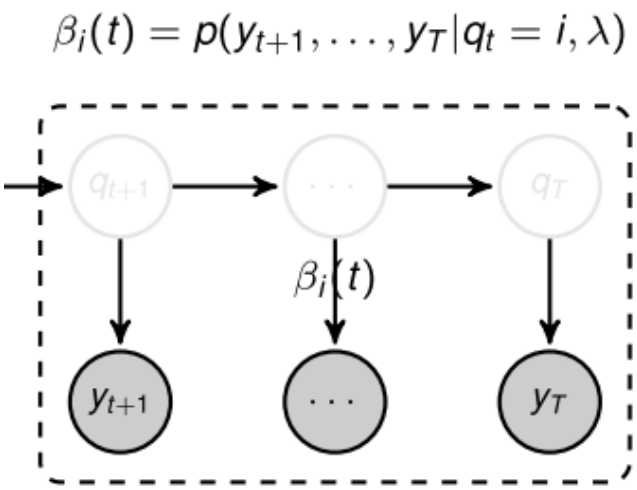

图 6: 后向算法的概率图模型

后向过程:

$$\beta_i(t) = p(y_{t+1}, ..., y_T | q_t = i, \lambda) \Longrightarrow \sum_{i=1}^{k} \beta_i(1)\pi_i b_i(y_1) = p(Y|\lambda)$$

迭代过程:

$$\beta_i(T) = 1$$

$$\beta_i(T-1) = p(y_T | q_{T-1} = i) = \sum_{j=1}^{k} p(q_T = j | q_{T-1} = i) p(y_T | q_T = j) = \sum_{j=1}^{k} a_{i,j} b_j(T)$$

$$\begin{aligned}
\beta_i(T-2) &= p(y_T, y_{T-1} | q_{T-2} = i) \\
&= \sum_{j=1}^{k} \sum_{l=1}^{k} p(q_T = l | q_{T-1} = j) p(y_T | q_T = l) p(q_{T-1} = j | q_{T-2} = i) p(y_{T-1} | q_{T-1} = j) \\
&= \sum_{j=1}^{k} a_{i,j} b_j(y_{T-1}) \beta_j(T-1)
\end{aligned}$$

...

$$\beta_i(t) = \sum_{j=1}^{k} \alpha_{i,j} b_j(y_t + 1) \beta_j(t+1)$$

...

$$\beta_i(1) = \sum_{j=1}^{k} \alpha_{i,j} b_j(y_2) \beta_j(2)$$

在时刻 t 处位于序列 Y 状态 q_i 时的概率：

$$p(q_t = i|Y,\lambda) = \frac{p(Y,q_t=i|\lambda)}{p(Y|\lambda)} = \frac{p(Y,q_t=i|\lambda)}{\sum_{j=1}^{k} p(Y,q_t=j|\lambda)} = \frac{\alpha_i(t)\beta_i t}{\sum_{j=1}^{k}\alpha_j(t)\beta_j(t)}$$

$$\begin{aligned} p(Y,q_t=i|\lambda) &= p(Y|q_t=i)p(q_t=i) \\ &= p(y_1,...y_t|q_t=i)p(y_{t+1},...y_T|q_t=i)p(q_t=i) \\ &= p(y_1,...y_t,q_t=i)p(y_{t+1},...y_T|q_t=i) \\ &= \alpha_i(t)\beta_i(t) \end{aligned}$$

0.8 学习算法

隐马尔可夫模型的学习，根据训练数据是包括观测序列和对应的状态序列还是只有观测序列，可以分别由监督学习与非监督学习实现。主要讲解非监督学习算法——Baum-Welch 算法（也就是 EM 算法）。

复习下 EM 算法：

$$\theta^{(g+1)} = \arg\max_{\theta}[Q(\theta,\theta^{(g)})] = \arg\max_{\theta}(\int_Z \log(p(X,Z|\theta))P(Z|X,\theta^{(g)})dz)$$

在 HMM 中，我们把要求解的问题写成 E-M 的形式，如下所示：

$$\lambda^{(g+1)} = \arg\max_{\lambda}\left(\underbrace{\int_{q\in Q}\ln(p(Y,q|\lambda))P(q,Y|\lambda^{(g)})dq}_{Q(\lambda,\lambda^{(g)})}\right)$$

其中，

$$\begin{aligned} Q(\lambda,\lambda^{(g)}) &= \int_{q\in Q}\ln(p(Y,q|\lambda))P(q,Y|\lambda^{(g)})dq \\ &= \sum_{q_0=1}^{k}\cdots\sum_{q_T=1}^{k}\left(\ln\pi_0 + \sum_{t=1}^{T}\ln a_{q_{t-1},q_t} + \sum_{t=1}^{T}\ln b_{q_t}(y_t)\right)p(q,Y|\lambda^{(g)}) \end{aligned}$$

可以看到公式里面有三个元素，现在依次求解每一项。

首先求解第一项 $\sum_{q_0=1}^{k}\cdots\sum_{q_T=1}^{k}\ln\pi_0 p(q,Y|\lambda^{(g)})$，

$$\sum_{q_0=1}^{k}\cdots\sum_{q_T=1}^{k}\ln\pi_0 p(q,Y|\lambda^{(g)}) = \sum_{i=1}^{k}\ln\pi_i p(q_0=i,Y|\lambda^{(g)})$$

其中约束条件为 $\sum_{i=1}^{k}\pi_i = 1$。

那么这种带有约束的最优化问题常用拉格朗日乘数法进行求解，即：

$$\mathbb{LM}^{\text{term1}}=\sum_{i=1}^{k}\ln\pi_i p(q_0=i,Y|\lambda^{(g)})+\tau\left(\sum_{i=1}^{k}\pi_i-1\right)$$

$$\frac{\partial^{\text{term1}}}{\partial\pi_i}=\frac{p(q,Y|\lambda^{(g)})}{\pi_i}+\tau=0\qquad\frac{\partial^{\text{term1}}}{\partial\tau}=\sum_{i=1}^{k}\pi_i-1=0$$

$$p(q_0=i,Y|\lambda^{(g)})=-\tau\pi_i\Longrightarrow\sum_{i=1}^{k}p(q_0=i,Y|\lambda^{(g)})=-\tau\sum_{i=1}^{k}\pi_i=-\tau$$

因此有

$$\pi_i=\frac{p(q_0=i,Y|\lambda^{(g)})}{-\tau}\Longrightarrow\pi_i=\frac{p(q_0=i,Y|\lambda^{(g)})}{\sum\limits_{i=1}^{k}p(q_0=i,Y|\lambda^{(g)})}$$

下面求解第二项 $\sum\limits_{q_0=1}^{k}\cdots\sum\limits_{q_T=1}^{k}\sum\limits_{t=1}^{T}\ln a_{q_{t-1},q_t}p(q,Y|\lambda^{(g)})=\sum\limits_{i=1}^{k}\sum\limits_{j=1}^{k}\sum\limits_{t=1}^{T}\ln a_{i,j}p(q_{t-1}=i,q_t=j,Y|\lambda^{(g)})$。同样的是拉格朗日优化问题，此时的拉格朗日函数为

$$\mathbb{LM}^{\text{term2}}=\sum_{i=1}^{k}\sum_{j=1}^{k}\sum_{t=1}^{T}\ln a_{i,j}p(q_{t-1}=i,q_t=j,Y|\lambda^{(g)})+\sum_{i=1}^{k}\tau_i\left(\sum_{i=1}^{k}\pi_i-1\right)$$

第三项 $\sum\limits_{q_0=1}^{k}\cdots\sum\limits_{q_T=1}^{k}\sum\limits_{t=1}^{T}\ln b_{q_t}(y_t)p(q,Y|\lambda^{(g)})=\sum\limits_{i=1}^{k}\sum\limits_{j=1}^{k}\sum\limits_{t=1}^{T}\ln b_j(y_t)p(q_t,Y|\lambda^{(g)})$ 的拉格朗日优化目标函数为

$$\mathbb{LM}^{\text{term3}}=\sum_{i=1}^{k}\sum_{j=1}^{k}\sum_{t=1}^{T}\ln b_j(y_t)p(q_t,Y|\lambda^{(g)})+\sum_{i=1}^{k}\tau_i\left(\sum_{i=1}^{k}b_j(y_t)-1\right)$$

具体解法和第一项的求解类似！

0.9 总结

隐马尔可夫模型作为动态模型的一种，主要对时间序列的数据进行建模，在学习算法上主要是利用 E-M 算法进行求解，但是总体上大大降低了计算效率，目前只介绍了离散型的动态模型，后期如果有时间，会具体介绍有关连续性的动态模型算法。