隐马尔可夫模型

cloud

2017.1.9

1 概述

1.1 基本概念

隐马尔可夫模型 (HMM) 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序列称为状态序列；每个状态生成一个观测由此产生的随机序列称为观测序列。

HMM 由初始概率分布，状态转移概率分布以及观测概率分布确定。设 $Q=\{q_1,q_2,...,q_N\}$ 是所有可能的状态集合，$V=\{v_1,v_2,...,v_M\}$ 是所有可能的观测集合。

对于一个 HMM，它包含一个长度为 T 的状态序列 $I=\{i_1,i_2,...,i_T\}$ 和状态序列对应的观测序列 $O=\{o_1,o_2,...,o_T\}$。$A=[a_{ij}]_{N\times N}$ 是状态转移矩阵，a_{ij} 表示的是在时刻 t 处于状态 q_i 条件下在时刻 $t+1$ 转移到状态 q_j 的概率。$B=[b_j(k)]_{N\times M}$ 是观测概率矩阵，$b_j(k)$ 表示的是在时刻 t 处于状态 q_j 的条件下生成观测 v_k 的概率。$\pi=(\pi_i)$ 是初始状态概率向量，π_i 是 $t=1$ 处于状态 q_i 的概率。

最终 HMM 由初始状态向量 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，HMM 可用三元符号表示即：

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

HMM 作了两个基本假设，即

（1）齐次马尔可夫性假设：假设隐藏的马尔可夫链在任意时刻 t 的状态只依赖于其前一时刻的状态，与其它时刻的状态及观测无关，也与时刻 t 无关。

$$P(i_t|i_{t-1},o_{t-1},...,i_1,o_1)=P(i_t|i_{t-1})$$

（2）观测独立性假设：假设任意时刻的观测只依赖于该时刻的马尔可夫链的状态，与其它观测及状态无关。

$$P(o_t|i_T,o_T,...i_{t+1},o_{t+1},i_t,i_{t-1},o_{t-1},...,i_1,o_1)=P(o_t|i_t)$$

1.2 基本问题

（1）概率计算问题：给定模型 $\lambda=(A,B,\pi)$ 和观测序列 $O=\{o_1,o_2,...,o_T\}$，计算在模型 λ 下观测序列 O 出现的概率 $P(O|\lambda)$。

（2）学习问题：已知观测序列 $O=\{o_1,o_2,...,o_T\}$，估计模型 $\lambda=(A,B,\pi)$ 参数，使得在该模型下观测序列概率 $P(O|\lambda)$ 最大，即使用极大似然估计方法估计参数。

（3）预测问题：又称解码问题，已知模型 $\lambda=(A,B,\pi)$ 和观测序列 $O=\{o_1,o_2,...,o_T\}$，求给定观测序列条件概率 $P(I|O)$ 最大的状态序列 $I=(i_1,i_2,...,i_T)$。即给定观测序列，求最有可能对应的状态序列。

2 概率计算问题

如何解决上节中的概率计算问题？一种方法可以采用直接计算法，即通过列举所有可能长度为 T 的状态序列 $I=(i_1,i_2,...,i_T)$，求各个状态序列 I 与观测序列 $O=\{o_1,o_2,...,o_T\}$ 的联合概率 $P(O,I|\lambda)$，然后对所有可能的状态序列求和得到 $P(O|\lambda)$。这种方法由于计算量过大在实际中是不可行的，所以就有了实际中可以应用的前向和后向算法。

2.1 前向算法

首先定义前向概率：给定 HMM 中的 λ，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，记为 $\alpha_t(i)=P(o_1,o_2,...,o_t,i_t=q_i|\lambda)$。下面给出前向算法：

输入：HMM 中的 λ 和观测序列 O

输出：观测序列概率 $P(O|\lambda)$

（1）初值如下

$$\alpha_1(i)=\pi_i b_i(o_1)\quad i=1,2,..,N$$

（2）递推，对于 $t=1,2,...,T-1$

$$\alpha_{t+1}(i)=[\sum_{j=1}^{N}\alpha_t(j)a_{ji}]b_i(o_{t+1})\quad i=1,2,..,N$$

（3）终止：

$$P(O|\lambda)=\sum_{i=1}^{N}\alpha_T(i)$$

第二步的递推过程图示如下：

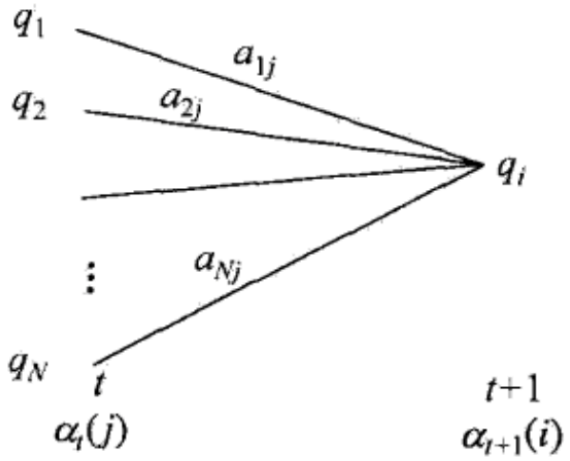

Figure 1: 前向概率的递推公式

2.2 后向算法

首先定义后向概率：给定 HMM 中的 λ，定义在时刻 t 状态为 q_i 条件下，从 $t+1$ 到 T 的部分观测序列 $o_{t+1}, o_{t+2}, ..., o_T$ 的概率为后向概率，记为 $\beta_t(i) = P(o_{t+1}, o_{t+2}, ..., o_T | i_t = q_i, \lambda)$。下面给出前向算法：

输入：HMM 中的 λ 和观测序列 O

输出：观测序列概率 $P(O|\lambda)$

（1）

$$\beta_T(i) = 1 \quad i = 1, 2, ..., N$$

（2）对 $t = T-1, T-2, ..., 1$

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(o_{t+1}) \beta_{t+1}(j) \quad i = 1, 2, .., N$$

（3）终止：

$$P(O|\lambda) = \sum_{i=1}^{N} \pi_i b_i(o_1) \beta_1(i)$$

第二步的递推过程图示如下：

3 学习问题

HMM 的学习，如果训练数据包括观测序列和对应的状态序列就是监督学习，这时采用极大似然法来估计参数。如果只有观测序列就是非监督学习，这时采用 Baum-Welch 算法也就是 EM 算法。

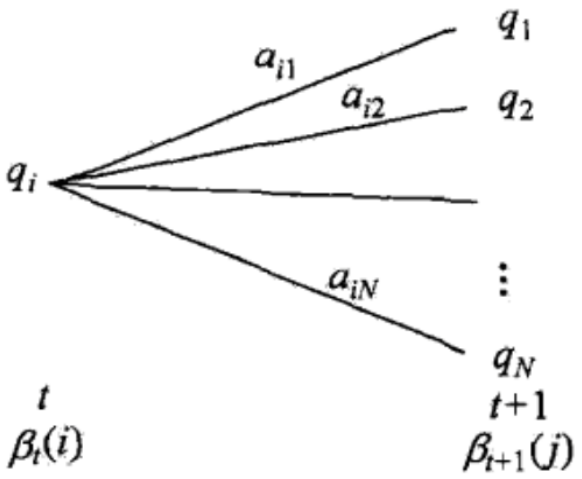

Figure 2: 后向概率的递推公式

3.1 极大似然估计法

假设训练数据包含 S 个长度相同的观测序列和对应的状态序列 $\{(O_1, I_1), (O_2, I_2), ..., (O_S, I_S)\}$，那么使用极大似然估计发来估计 HMM 的参数，具体方法如下：

（1）转移概率 a_{ij} 的估计

设样本中时刻 t 处于状态 i，时刻 $t+1$ 转移到状态 j 的频数为 A_{ij}，那么状态转移概率 a_{ij} 的估计是：

$$\hat{a}_{ij} = \frac{A_{ij}}{\sum_{j=1}^{N} A_{ij}} \quad i = 1, 2, ..., N; j = 1, 2, ..., N$$

（2）观测概率 $b_j(k)$ 的估计

设样本中状态为 j 且观测为 k 的频数是 B_{jk}，那么状态为 j 观测为 k 的概率 $b_j(k)$ 的估计是：

$$\hat{b}_j(k) = \frac{B_{jk}}{\sum_{k=1}^{M} B_{jk}} \quad j = 1, 2, ..., N; k = 1, 2, ..., M$$

（3）初始状态概率 π_i 的估计 $\hat{\pi}_i$ 为 S 个样本中初始状态为 q_i 的频率。

3.2 Baum-Welch 算法

输入：观测数据 $O = (o_1, o_2, ..., o_T)$

输出：HMM 参数

（1）初始化

对 $n = 0$，选取 $a_{ij}^{(0)}, b_j(k)^{(0)}, \pi^{(0)}$，得到模型 $\lambda^{(0)} = (A^{(0)}, B^{(0)}, \pi^{(0)})$

（2）递推，对 $n=1,2,...$，按照观测 $O=(o_1,o_2,...,o_T)$ 和模型 $\lambda^{(n)}=(A^{(n)},B^{(n)},\pi^{(n)})$ 计算。

$$a_{ij}^{(n+1)}=\frac{\sum_{t=1}^{T-1}\varepsilon_t(i,j)}{\sum_{t=1}^{T-1}\gamma_t(i)}$$
$$b_j(k)^{(n+1)}=\frac{\sum_{t=1,o_t=v_k}^{T}\gamma_t(j)}{\sum_{t=1}^{T}\gamma_t(j)}$$
$$\pi_i^{(n+1)}=\gamma_1(i)$$

上式中 $\gamma_t(i)=P(i_t=q_i|O,\lambda)$ 是在给定模型 λ 和观测 O 时，在时刻 t 处于状态 q_i 的概率。$\varepsilon_t(i,j)=P(i_t=q_i,i_{t+1}=q_j|O,\lambda)$ 是在给定模型 λ 和观测 O 时，在时刻 t 处于状态 q_i 且在时刻 $t+1$ 处于状态 q_j 的概率。

（3）终止

得到模型参数 $\lambda^{(n+1)}=(A^{(n+1)},B^{(n+1)},\pi^{(n+1)})$

4 预测问题

4.1 近似算法

近似算法在每个时刻 t 选择在该时刻最有可能出现的状态 i_t^*，从而得到一个状态序列 $I^*=(i_1^*,i_2^*,...,i_T^*)$ 将其作为预测的结果。

在给定模型 λ 和观测 O 时，在时刻 t 处于状态 q_i 的概率是：

$$\begin{aligned}\gamma_t(i)&=P(i_t=q_i|O,\lambda)\\&=\frac{P(i_t=q_i,O|\lambda)}{P(O|\lambda)}\\&=\frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{N}\alpha_t(i)\beta_t(i)}\end{aligned}$$

在每一时刻 t 最优可能的状态是 $i_t^*=arg\max\limits_{1\leqslant i\leqslant N}[\gamma_t(i)]\quad t=1,2,...,T$，从而得到状态序列 $I^*=(i_1^*,i_2^*,...,i_T^*)$。近似算法的优点是计算简单，缺点就是不能保证预测的状态序列是整体最有可能的状态序列。

4.2 维特比算法

维特比算法实际是利用动态规划解 HMM 预测问题，利用动态规划球概率最大路径，这时一条路径对应着一个状态序列。根据动态规划原理，只需要从时刻 $t=1$ 开始递推的计算在时刻 t 状态为 i

的各条部分路径的最大概率，直至得到时刻 $t=T$ 状态为 i 的各条路径的最大概率。时刻 $t=T$ 的最大概率路径即为最优路径的概率 P^*，最优路径的终结点 i_T^* 也同时得到。之后从 i_T^* 开始，由后向前逐步求得节点 $i_{T-1}^*,...,i_1^*$，得到最优路径 $I^*=(i_1^*,i_2^*,...,i_T^*)$, 这就是维特比算法。

定义在时刻 t 状态为 i 的所有单个路径 $(i_1,i_2,...,i_t)$ 中概率最大值为：

$$\delta_t(i)=\max_{i_1,i_2,...,i_{t-1}} P(i_t=i,i_{t-1},...,i_1,o_t,...,o_1|\lambda) \quad i=1,2,...,N$$

由此可得变量 δ 的递推公式：

$$\begin{aligned}\delta_t(i)&=\max_{i_1,i_2,...,i_{t-1}} P(i_t=i,i_{t-1},...,i_1,o_t,...,o_1|\lambda)\\&=\max_{1\leqslant j\leqslant N}[\delta_t(j)a_{ji}]b_i(o_{t+1}) \quad i=1,2,...,N;t=1,2,...,T-1\end{aligned}$$

定义在时刻 t 状态为 i 的所有单个路径 $(i_1,i_2,...,i_{t-1},i_t)$ 中概率最大路径的第 $t-1$ 个结点为：

$$\psi_t(i)=arg\max_{1\leqslant j\leqslant N}[\delta_{t-1}(j)a_{ji}] \quad i=1,2,...,N$$

下面给出维特比算法的流程：

输入：模型 $\lambda=(A,B,\pi)$ 和观测 $O=(o_1,o_2,...,o_T)$

输出：最优路径 $I^*=(i_1^*,i_2^*,...,i_T^*)$

（1）初始化

$$\delta_1(i)=\pi_i b_i(o_1) \quad i=1,2,...,N$$

$$\psi_1(i)=0 \quad i=1,2,...,N$$

（2）递推：对 $t=2,3,...,T$

$$\delta_t(i)=\max_{1\leqslant j\leqslant N}[\delta_{t-1}(j)a_{ji}]b_i(o_t) \quad i=1,2,...,N$$

$$\psi_t(i)=arg\max_{1\leqslant j\leqslant N}[\delta_{t-1}(j)a_{ji}] \quad i=1,2,...,N$$

（3）终止

$$P_*=\max_{1\leqslant i\leqslant N}\delta_T(i)$$

$$i_T^*=arg\max_{1\leqslant i\leqslant N}[\delta_T(i)]$$

（4）最优路径回溯，对 $t=T-1,T-2,...,1$

$$i_t^*=\psi_{t+1}(i_{t+1}^*)$$

求得最优路径 $I^*=(i_1^*,i_2^*,...,i_T^*)$。

5 一点感想

写到这一章的时候，感觉自己的理解力不是很够，笔记内容基本都是将书上的核心内容摘抄下来。自己希望用口语化的文字去解释但是感觉不够精确，感觉 HMM 算法通过书上的例题可以很好理解，但是总结成公式之后就是看了就忘，过一阵甚至都说不上来 HMM 是什么了。在学习机器学习算法的时候，如果看的文章太直观或者都是通过实例来理解，会感觉理解的不透彻，因为实际问题最终还是要转化成数学问题的。找了很多资料，有些是很好理解，但是过了一阵似乎又随风飘散了，所以对我来说吃透数学公式，经历这种来回往复比较纠结的过程，会理解的更为深刻吧。

6 参考文献

1. 李航博士的统计学习方法