漫谈隐马尔科夫模型

Michael Leifield

1 Introduction

HMM包含三个基本任务：Scoring、Decoding和Training。Scoring在给定模型下，计算某特定观察序列的概率。这个问题相对简单，可通过Forward算法解决。Decoding问题是说如何根据建立好的模型以及某特定的观察序列来计算最可能的隐含状态转换序列。Training是Scoring与Decoding两个任务的基础，它运用期望最大化的Baum-Welch算法训练模型参数。

首先形式化描述HMM，它包含两组状态（隐含状态和观察状态）集合和三组概率集合。三组概率集合分别是：从一个隐含状态到另一个隐含状态的概率矩阵，也被称作状态转移矩阵$A = (a_{ij})$；在某特定隐藏状态下，观察到状态的概率矩阵$B = (b_{ij})$；再加上初始时刻每个隐含状态的概率向量$\pi = (\pi_i)$。模型中矩阵A和B的每一概率都是时间无关的，也就是说，在系统演化过程中这些矩阵并不随时间改变。

2 Scoring

解决Scoring的方法有很多，最简单是穷举，前提是代价很高。采用递归的Forward算法是解决这个任务的一种方法。首先定义一个局部概率α，它是到达模型中某个状态的概率。对一个长度为T的观察序列$Y^{(k)} = y_{k_1}, y_{k_2}, ..., y_{k_T}$，时刻$t$位于状态$j$的局部概率$\alpha_t(j)$可计算为：

$$\alpha_t(j) = Pr(observation\ |hidden\ state\ is\ j) \times Pr(all\ paths\ to\ state\ j\ at\ time\ t)$$

特别地，$t = 1$时没有任何通向当前节点的路径。因而$t = 1$的局部概率等于当前状态的初始概率乘以相关的观察概率$\alpha_1(j) = \pi(j) * b_{jk_1}$。我们采用递归方式，用$t - 1$时刻的各$\alpha$来定义$t$时刻的$\alpha$：

$$\alpha_t(j) = b_{jk_t} \sum_{i=1}^{n} \alpha_{t-1}(i) a_{ij}$$

因而，最后的观察状态，其局部概率包含了所有可能路径达到达这些状态的概率，再对这些局部概率求和便可得到给定模型下的观察序列的概率

$$Pr(Y^{(k)}) = \sum_{j=1}^{n} \alpha_T(j)$$

3 Decoding

Viterbi算法解决这个问题。首先定义局部概率$\delta(i,t)$，它表示到达时刻t状态i的最大概率。对应于最大局部概率，最佳局部路径是得此最大概率的隐含状态序列。需要注意的是，δ和Forwad算法中的局部概率α不同，因为当前表示的是时刻t时到达某个状态最可能路径的概率，而不是所有路径的概率。

时刻1，到达某状态的最可能路径明显不存在。我们用$t = 1$时所处状态的初始概率及对应观察状态$k1$的概率计算$\delta_1(i) = \pi(i)b_{ik_1}$。$t > 1$时，递归地使用前一刻状态的局部概率来计算状态$X$的最可能路径的概率

$$Pr(X\ at\ time\ t) = \max_i Pr(i\ at\ time\ (t-1)) \times Pr(X|i) \times Pr(obs\ at\ time\ t \mid X)$$

这里，利用状态转移概率和相应的观察概率来计算每一个中间状态和终止状态的局部概率

$$\delta_t(i) = \max_j(\delta_{t-1}(j)a_{ji}b_{ik_t})\Theta$$

然而我们的目标是最可能的状态序列。在这里，通过设置一个反向标记指针$\phi_t(i) = \arg\max_j(\delta_{t-1}(j)a_{ji})$来记录引发当前状态的前一时刻的最优状态。

4 Training

上述两个任务都依赖HMM的先验参数。求解模型参数的Baum-Welch算法并不直接计算，而是通过学习的方式来进行估计。它首先对模型参数进行一个初步的设定，然后通过训练数据来评估这些参数的价值进而更新这些参数。

要理解Baum-Welch，先看Backward算法。这里重新定义一下前向算法中的局部概率$\alpha_t(i)$，称其为前向变量。

$$\alpha_t(i) = P(O_1O2...O_t, q_t = S_i|\lambda)$$

相似地，这里定义一个后向变量

$$\beta_t(i) = P(O_{t+1}O_{t+2}..., O_T|q_t = S_i, \lambda)$$

它表示的是已知模型λ以及t时刻位于隐含状态S_i，从$t+1$时刻到终止时刻的局部观察序列的概率。$t=T$时令$\beta_T(i)=1$。递归地，计算每个时间片$t=T-1,T-2,...,1$时的后向变量

$$\beta_t(i)=\sum_{j=1}^{N}\alpha_{ij}b_j(O_{t+1})\beta_{t+1}(j)$$

我们知道，估计模型参数的标准在于最大化给定观察序列O的概率$P(O|\lambda)$。Baum-Welch采用EM思想，首先定义t时刻位于状态S_i的概率变量

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

然后定义t时刻位于隐含状态S_i以及$t+1$时刻位于隐含状态S_j的概率变量

$$\begin{aligned}\xi_t(i,j)&=P(q_t=S_j,q_{t+1}=S_j|O,\lambda)\\&=\frac{\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)}{P(O|\lambda)}\\&=\frac{\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)}\end{aligned}$$

有了上述两个变量，Baum-Welch方法用它们来重新估计HMM的参数

$$\begin{aligned}\overline{\pi}_i&=\gamma_1(i)\\\overline{a}_{ij}&=\frac{\sum_{t=1}^{T-1}\xi_t(i,j)}{\sum_{t=1}^{T-1}\gamma_t(i)}\\\overline{b}_{jk}&=\frac{\sum_{t=1,O_t=v_k}^{T}\gamma_t(j)}{\sum_{t=1}^{T}\gamma_t(j)}\end{aligned}$$

最后再用计算出的新参数来重新估计。如此反复，直到达到指定迭代次数或者误差小于一定值时停止。