HMM

Renhe W.

Contents

List of Figures

List of Tables

1 隐马尔可夫模型

隐马尔可夫模型 (Hidden Markov Model, HMM) 是一个统计模型，用于描述一个隐藏的马尔可夫链产生的观测序列. 系统被假定为一个马尔可夫过程（即无记忆的随机过程）与不可观察（隐藏）的状态.

HMM 有两个序列：一个是观测序列，另一个是隐藏的状态序列. 具体形式如图1：

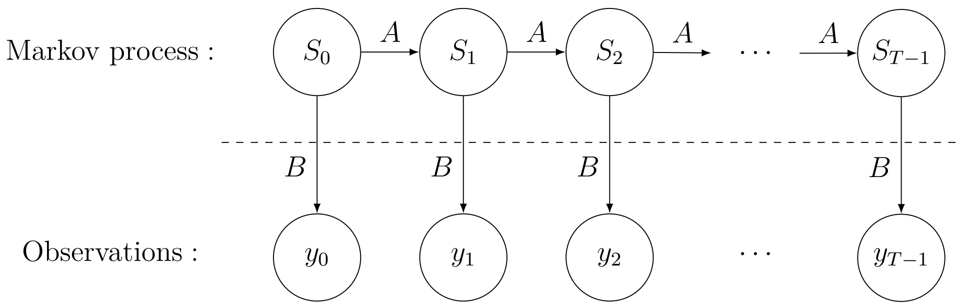

图 1. A hidden Markov model.

HMM 广泛应用于语音识别、自然语言处理、生物信息学（如蛋白质结构预测）等领域.

1.1 HMM 模型组成部分

根据图1，HMM 模型由以下几个部分构成：

- **状态集合**：这是一个有限集合，其中的每个元素称为一个状态。这些状态在模型中是不可观察的，因此称为“隐状态”.
- **观测集合**：每个隐状态可以生成一个观测值，观测集合由这些可能的观测值组成.
- **状态转移概率矩阵**：表示从一个状态转移到另一个状态的概率.
- **观测概率矩阵**：给定某个状态，生成各个观测值的概率.
- **初始状态分布**：系统在开始时各个状态的概率分布.

根据组成部分，用公式表示以上的内容：

- 设 Q 为所有可能的状态的集合，q_i 为一个特定的状态.
- 设 Y 为所有可能的观测的集合，y_t 为一个特定的观测.
- 状态转移概率矩阵 $A = [a_{ij}]$ ，其中 a_{ij} 表示从状态 i 转移到状态 j 的概率.

- 观测概率矩阵 $B=[b_j(y_t)]$ ，其中 $b_j(y_t)$ 表示在状态 j 下观测到 y_t 的概率.
- 初始状态分布 $\pi=[\pi_i]$ ，其中 π_i 表示系统开始时处于状态 i 的概率.

其中 HMM 的参数包括:

- 状态转移概率矩阵 A: 元素 a_{ij} 表示从状态 i 转移到状态 j 的概率.
- 观测概率矩阵 B : 元素 $b_j(y_t)$ 表示在状态 j 下观测到观测值 y_t 的概率.
- 初始状态概率向量 π : 元素 π_i 表示模型在时间 $t=1$ 时处于状态 i 的概率.

HMM 的求解主要包括以下三个基本问题:

1. 估计问题 (Evaluation Problem): 给定模型参数和一个观测序列，计算这个观测序列出现的概率。这个问题通常使用前向算法 (Forward Algorithm) 和后向算法 (Backward Algorithm) 来解决.
2. 解码问题 (Decoding Problem)：给定模型参数和一个观测序列，找到最有可能的隐藏状态序列。这个问题通常使用 Viterbi 算法来解决.
3. 学习问题 (Learning Problem): 给定一个观测序列，如何调整模型参数 (A, B, 和 π) 使得这个观测序列出现的概率最大。这个问题通常使用 Baum-Welch 算法 (一种特殊的 EM 算法) 来解决.

2 求解步骤

在隐马尔可夫模型 (HMM) 中，使用最大似然方法估计模型参数通常涉及到所谓的“完全数据”的概念，完全数据包括观测数据和隐藏数据（即隐藏状态），我们通常用 y_t 表示在时间 t 的观测值，用 s_t 表示在时间 t 的隐藏状态. 构造 Q 函数是期望最大化 (EM) 算法的关键步骤，其中 Baum-Welch 算法是 EM 算法在 HMM 中的特殊应用.

2.1 Q 函数的构造

对于完全数据的似然，在 HMM 中，完全数据的似然由观测序列和相应的隐藏状态序列共同确定。完全数据的似然函数表示为:

$$P(Y, S \mid \theta),$$

其中，$Y=\{y_1,y_2,\ldots,y_T\}$ 是观测序列，$S=\{s_1,s_2,\ldots,s_T\}$ 是隐藏状态序列，θ 是模型参数(状态转移概率、观测概率、初始状态概率).

Q 函数是在 EM 算法中用来估计参数的关键函数（具体可以参考 EM 算法过程，为什么 Q 函数更新可以使得似然最大）。它计算了给定观测数据和当前参数估计下，参数的新估计值. Q 函数的定义为隐藏数据的条件期望下的完全数据对数似然:

$$Q\left(\theta,\theta^{(old)}\right)=E\left[\log P(Y,S\mid\theta)\mid Y,\theta^{(old)}\right], \tag{2.1}$$

其中，$\theta^{(old)}$ 是当前参数估计，θ 是新的参数估计. 下面将进一步完善 Q 函数计算的细节，根据 HMM 的定义，完全数据对数似然可以写作:

$$\log P(Y,S\mid\theta)=\log P\left(y_1,s_1\mid\theta\right)+\sum_{t=2}^{T}\log P\left(y_t,s_t\mid s_{t-1},\theta\right), \tag{2.2}$$

以上(2.2)可以进一步分解为:

$$\begin{aligned}\log P(Y,S\mid\theta)&=\log P\left(y_1,s_1\mid\theta\right)+\sum_{t=2}^{T}\log P\left(y_t,s_t\mid s_{t-1},\theta\right),\\&=\log P\left(y_1\mid s_1,\theta\right)P\left(s_1\mid\theta\right)+\sum_{t=2}^{T}\log P\left(y_t\mid s_t,s_{t-1},\theta\right)P\left(s_t\mid s_{t-1},\theta\right),\\&=\log P\left(s_1\mid\theta\right)+\log P\left(y_1\mid s_1,\theta\right)+\sum_{t=2}^{T}\log P\left(y_t\mid s_t,s_{t-1},\theta\right)+\sum_{t=2}^{T}\log P\left(s_t\mid s_{t-1},\theta\right).\end{aligned}$$

最后得到：

$$\log P(Y,S\mid\theta)=\log\pi_{s_1}+\sum_{t=1}^{T}\log b_{s_t}\left(y_t\right)+\sum_{t=2}^{T}\log a_{s_{t-1},s_t}, \tag{2.3}$$

这里，π_{s_1} 是初始状态概率，$b_{s_t}\left(y_t\right)$ 是在状态 s_t 下观测到 y_t 的概率，a_{s_{t-1},s_t} 是从状态 s_{t-1} 转移到状态 s_t 的概率. Q 函数是(2.3)

2.2 估计问题 (Evaluation Problem)

给定模型参数和一个观测序列，我们希望计算得到观测序列的似然，运用最大似然的想法求解模型的参数.

前向算法 (Forward Algorithm): 使用前向概率 $\alpha_t(i)$ 表示到时间 t 为止，系统处于状态 i 并且观测到序列 $O_1,O_2,\ldots O_t$ 的概率.

递推公式为:

$$\alpha_1(i) = \pi_i b_i (O_1)$$

$$\alpha_{t+1}(j) = \left(\sum_{i=1}^{N} \alpha_t(i) a_{ij} \right) b_j (O_{t+1})$$

其中，N 是状态数，a_{ij} 是从状态 i 到状态 j 的转移概率，$b_j(O_t)$ 是在状态 j 下观测到 O_t 的概率.

2.3 解码问题 (Decoding Problem)

给定模型参数和观测序列，我们希望找到最有可能的隐藏状态序列.

Viterbi 算法: 定义 $\delta_t(i)$ 为时刻 t 系统处于状态 i 并且最有可能的状态序列路径的概率.

递推公式为:

$$\delta_1(i) = \pi_i b_i (O_1)$$

$$\delta_{t+1}(j) = \max_i (\delta_t(i) a_{ij}) b_j (O_{t+1})$$

2.4 学习问题 (Learning Problem)

给定观测序列，我们希望调整模型参数使得观测序列概率最大.

Baum-Welch 算法 (一种 EM 算法):

定义前向概率 $\alpha_t(i)$ 和后向概率 $\beta_t(i)$. 后向概率表示从时刻 $t+1$ 到最终时刻的部分观测序列和状态序列的概率，给定在时刻 t 的状态是 i.

递推公式为:

$$\beta_T(i) = 1$$

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

使用 α 和 β 值，我们可以估计模型参数 A 和 B.