隐马尔科夫模型学习笔记

李向阳 d1142845997@gmail.com

目录

1 引入

隐马尔可夫模型 (Hidden Markov Model, HMM) 是一个统计模型, 它用来描述一个含有未知参数的马尔科夫过程. 其特点是从可观察到的参数中确定该过程的隐含参数, 然后利用这些参数作进一步的分析.

在正常的马尔可夫模型中, 状态对于观察者来说是直接可见的, 此时状态之间的转移概率便是全部的模型参数. 而在隐马尔可夫模型中, 状态并不是直接可见的, 但受状态影响的某些变量则是可见的. 每一个状态在可能输出的符号上都有一个概率分布, 因此输出符号的序列能够透露出状态序列的一些信息.

本质上, HMM 是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.

2 模型介绍

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), 通常假设状态变量是隐藏的、不可观测的. 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看做一个时刻.

下面用记号来描述一下, 对于有略微强迫症的人来说, 记号一直是一件令人蛋疼的事情, 因为很多资料中的记号都不大相同. HMM 模型也是一样, 这里我采用李航《统计学习方法》中的记号, 并尽量与之前的记号保持一致.

设 I 是长度为 T 的状态序列, O 是对应的观测序列.

$$I = (i_1, i_2, \cdots, i_T), O = (o_1, o_2, \cdots, o_T)$$

状态序列和观测序列都是有各自的取值空间的, 设 Q 是所有可能的状态的集合, 其元素个数为 N, 设 V 是所有可能的观测的集合, 其元素个数为 M.

$$Q = \{q_1, q_2, \cdots, q_N\}, V = \{v_1, v_2, \cdots, v_M\}$$

设矩阵 A 是状态转移概率矩阵, 即 $A = (a_{ij})_{N\times N}$, 其中

$$a_{ij} = P(i_{t+1} = q_j | i_t = q_i), i, j = 1, 2, \cdots, N$$

也就是说 a_{ij} 表示在时刻 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.

设矩阵 B 是观测概率矩阵, 即 $B=(b_{jk})_{N\times M}$, 其中

$$b_{jk}=P(o_t=v_k|i_t=q_j), j=1,2,\cdots,N, k=1,2,\cdots,M$$

也就是说 b_{jk} 表示在时刻 t 处于状态 q_j 的条件下生成观测 v_k 的概率.
设 $\boldsymbol{\pi}$ 是初始状态的概率向量, 即 $\boldsymbol{\pi}=(\pi_i)$, 其中

$$\pi_i=P(i_1=q_i), i=1,2,\cdots,N$$

也就是说 π_i 表示时刻 $t=1$ 时处于状态 q_i 的概率.

隐马尔可夫模型由初始状态概率向量 $\boldsymbol{\pi}$、状态转移概率矩阵 A 和观测概率矩阵 B 完全决定, 其中 $\boldsymbol{\pi}$ 和 A 决定状态序列, B 决定观测序列, 因此, 隐马尔科夫模型 λ 可以用三元符号表示, 即

$$\lambda=(A,B,\boldsymbol{\pi})$$

$A,B,\boldsymbol{\pi}$ 称为隐马尔可夫模型的三要素.

下面举一个具体的例子, 这个知乎上有很多, 可见https://www.zhihu.com/question/20962240. 这里还是用李航《统计学习方法》上摸球的例子.

假设有 4 个盒子, 每个盒子里都装有红白两种颜色的球, 盒子里的红白球数可见表

表 1: 各盒子的红白球数

盒子	**ID1**	**ID2**	**ID3**	**ID4**
红球数	5	3	6	8
白球数	5	7	4	2

按照下面的方法摸球, 产生一个球的颜色的观测序列: 开始, 从 4 个盒子中随机 (等概论) 的选取一个盒子, 从这个盒子中随机摸出一个球, 记录其颜色后, 放回; 然后, 从当前盒子中转移到下一个盒子, 转移规则是: 如果当前盒子是盒子 1, 那么下一个盒子一定是盒子 2, 如果当前盒子是盒子 2 或者 3, 那么分别以概率 0.4 和 0.6 转移到左边或右边的盒子, 如果当前是盒子 4, 那么各以 0.5 的概率停留在盒子 4 或转移到盒子 3; 确定转移的盒子后, 再从这个盒子中随机摸出一个球, 记录其颜色, 放回; 如此下去, 不断重复. 比如进行了 5 次, 得到了一个球的颜色的观测序列:

$$O=\{红, 红, 白, 白, 红\}$$

在这个过程中, 假设观察者只能观测到球的颜色的序列, 观测不到球是从哪个盒子中取出来的, 即观测不到盒子的序列.

本例子中, 盒子的序列便是状态序列, 而球的颜色的序列便是观测序列, 前者是隐藏的, 只有后者是可观测的. 根据所给条件, 可以明确状态集合、观测集合、序列长度和模型的三要素.

盒子对应状态, 状态的集合是

$$Q = \{\text{盒子 } 1, \text{盒子 } 2, \text{盒子 } 3, \text{盒子 } 4\}, N = 4$$

球的颜色对应观测序列, 观测的集合是

$$V = \{\text{红}, \text{白}\}, M = 2$$

本次具体的状态序列和观测序列的长度为 $T = 5$, 其中状态序列未知, 而观测序列为

$$O = \{\text{红}, \text{红}, \text{白}, \text{白}, \text{红}\}$$

初始概率分布为

$$\boldsymbol{\pi} = (0.25, 0.25, 0.25, 0.25)^T$$

状态转移概率矩阵为

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0.4 & 0 & 0.6 & 0 \\ 0 & 0.4 & 0 & 0.6 \\ 0 & 0 & 0.5 & 0.5 \end{bmatrix}$$

观测概率矩阵为

$$B = \begin{bmatrix} 0.5 & 0.5 \\ 0.3 & 0.7 \\ 0.6 & 0.4 \\ 0.8 & 0.2 \end{bmatrix}$$

这便是本例子的模型三要素.

那么隐马尔可夫模型有什么用呢? 它主要解决 3 类基本问题.

(1) 解码问题, 也称为预测问题. 给定模型 $\lambda = (A, B, \boldsymbol{\pi})$ 和观测序列, 求最有可能的状态序列.

这个问题有两种解法, 会给出两个不同的答案. 每个答案都合理, 因为答案的意义不太一样. 第一种解法是求最大似然路径, 就是求一串状态序列, 使得其产生观测结果序列的概率最大. 第二种解法, 就不是求一组状态序列了, 而是求出每次的观测结果分别是某种状态产生的概率.

(2) 概率计算问题. 给定模型 $\lambda = (A, B, \boldsymbol{\pi})$ 和观测序列, 计算在模型 λ 下观测序列出现的概率.

这个问题的意义看似不大, 毕竟我们的观测结果很多时候都对应了一个比较大的概率. 这个问题的目的在于, 检测观测到的结果和已知的模型是否吻合. 如果很多次结果都对应了比较小的概率, 那么就说明我们已知的模型很有可能是错的.

(3) 学习问题. 已知观测序列, 要估计出模型 $\lambda = (A, B, \boldsymbol{\pi})$ 的参数.

这个问题就很重要了, 因为这是最常见的情况. 很多时候我们只有可见结果, 不知道 HMM 模型里面的参数, 我们需要从可见结果中估计出这些参数.

3 关于 HMM 的补充

4 总结

4.1 参考资料

(1) 李航的《统计学习方法》

(2) 周志华的《机器学习》

(3) 维基: https://zh.wikipedia.org/wiki/%E9%9A%90%E9%A9%AC%E5%B0%94%E5%8F%AF%E5%A4%AB%E6%A8%A1%E5%9E%8B

(4) 知乎: https://www.zhihu.com/question/20962240, 有非常形象的例子.

参考文献

[1] 李荣华. *偏微分方程数值解法*. 高等教育出版社 (2010)

[2] Zhilin Li,Zhonghua Qiao,Tao Tang.*Numerical Solutions of Partial Differential Equations-An Introduction to Finite Difference and Finite Element Methods*.(2011)

[3] 孙志忠. *偏微分方程数值解法*. 科学出版社 (2011)

[4] 陆金甫关治. *偏微分方程数值解法*. 清华大学出版社 (2004)

附录