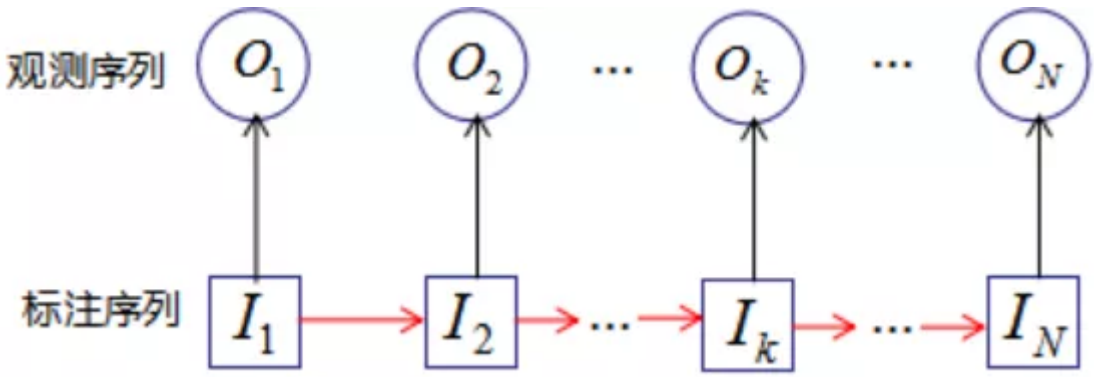
初学者也能看懂的隐马尔科夫模型介绍

原创 石头 机器学习算法那些事 2019-12-27

隐马尔科夫模型是（hidden Markov model，HMM）是可用于标注问题的统计学习模型，描述由隐藏的马尔科夫链随机生成观测序列的过程。

隐马尔可夫模型（hidden Markov model，HMM）是时间序列的概率模型，常用于词性标注，语音识别，文本分析等领域。HMM是基于马尔科夫链进行标注的，我们对已经观察的数据序列O进行标注，标注序列I相当于不可观测序列（隐变量），如何求解概率最大的标注序列 I 是HMM的核心问题，我们以图解的方式形象的描述HMM问题。

已知观测序列 $O=\{O_1,O_2,O_3,...,O_N\}$，标注序列 $I=\{I_1,I_2,I_3,...,I_N\}$，其中观测序列是相互独立的且与对应的标注序列相关，标注序列符合马尔科夫链模型，如下图：



再次强调下隐马尔科夫模型假设的场景，1）观测序列是相互独立的，2）标注序列符合马尔科夫链模型，3）观测序列由标注序列决定。

HMM模型主要处理以下三个问题：

1）已知观测序列O，如何求解HMM模型参数λ。

2）已知观测序列O和模型参数λ，如何求解最有可能的标注序列 I 。

3）已知观测序列O和模型参数λ，如何求解观测序列O出现的概率$P(O|\lambda)$。

下面给出大致的解决方案：

1） 已知观测序列O，如何求解模型参数λ。

$$P(O|\lambda)=\sum_{I}P(O|I,\lambda)P(I|\lambda)$$

其中I为标注序列，在上式中的含义为隐变量，因此可用EM算法求解上式的模型参数λ。

2） 已知观测序列O和模型参数λ，如何求解最有可能的标注序列 I

3）已知观测序列O和模型参数λ，如何求解观测序列O出现的概率概率$P(O|\lambda)$

后续内容会给出具体的算法和实例。

介绍到这里，相信大家对HMM有了初步的理解了吧，举一个例子来说明HMM的应用场景：

小明每天穿的服装为O_i（O_i表示第i天穿的服装），假设小明穿的服装与前一天穿的服装是相互独立的，且服装的选择受精神状态的影响，当天的精神状态受到前一天精神状态的影响，精神状态为S_i（S_i表示第 i 天的精神状态），已知观察序列O_i，求对应的精神状态序列，这类问题我们可以用HMM来解决。

介绍到这里，相信大家对HMM的基本概念和应用场景有了一定的概念，如果对HMM还不是很理解的话，先别慌，下面循环渐进的介绍HMM，数学要求不高的人也能很好的理解。

1.概率与随机过程的区别

概率是反映随机事件发生的可能性，大学课程《概率论与数理统计》全文内容是基于概率介绍的。若随机事件发生的概率随时间而改变，我们在考虑时间因素或空间因素的情况下去分析随机事件发生的概率，称为随机过程。

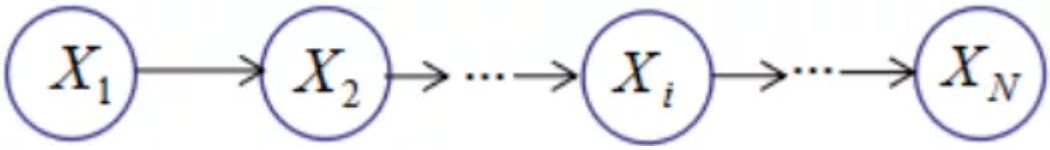
若我们每次抛掷硬币且正面向上的概率为0.5，正面向上的概率不随抛掷次数而改变，我们可以用概率来描述这一事件，如P(X)，其中X表示硬币正面向上的随机事件。

若我们每次抛掷硬币，硬币落在地上会导致形状的改变，正面向上的概率随抛掷次数的改变而改变，我们用随机过程来描述这一事件，如$P(X_1, X_2, \ldots, X_i)$，其中X_i表示第i次抛掷硬币正面向上的随机事件。笼统的讲，概率是分析一个随机变量，而随机过程是分析一组随机变量。

2.马尔科夫过程的概念

当随机过程的变量满足马尔科夫性的无记忆性质时，则称为马尔科夫过程。

我们定义X_i为第i个时刻的随机变量，如下图的随机过程：



若随机变量满足：

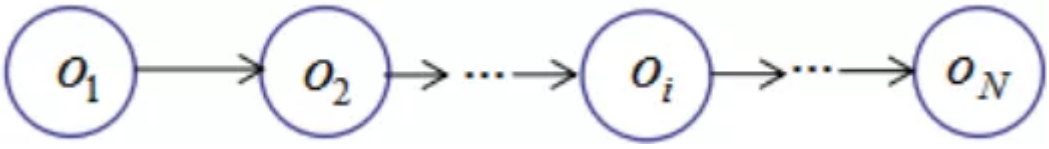
$$P(X_t \mid X_{t-1}, X_{t-2}, \ldots, X_1) = P(X_t \mid X_{t-1})$$

上式的含义为t时刻的随机变量只依赖于t-1时刻的随机变量，与其他时刻无关，则称该过程为马尔科夫过程。

3.马尔科夫模型与隐马尔科夫模型的区别

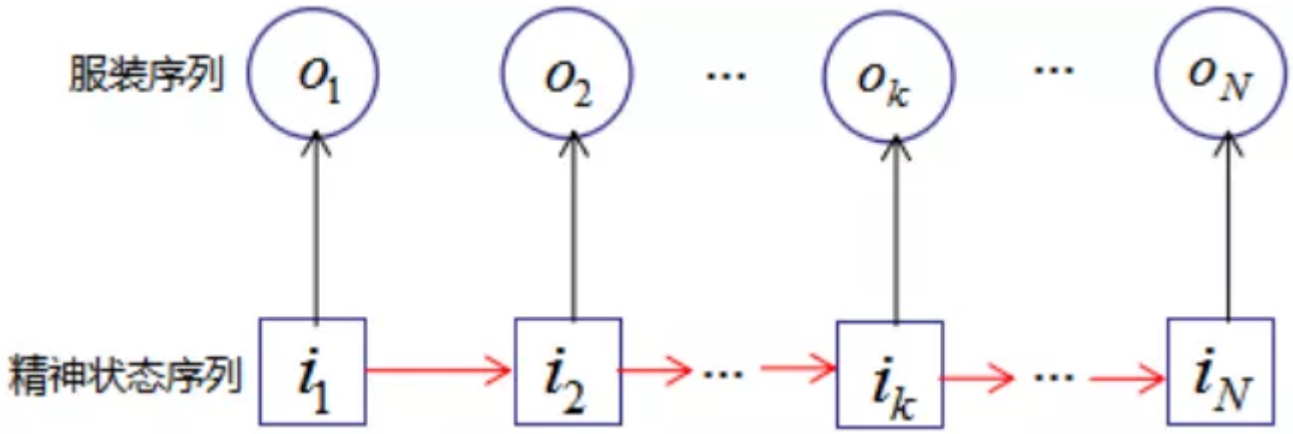
马尔科夫模型与隐马尔科夫模型的区别在于是否含有隐变量，本节还是用小明穿服装的例子来形象的说明马尔科夫模型与隐马尔可夫模型的区别：

若小明每天的服装只受到前一天所穿服装的影响， 下图为小明最近N天所穿服装的序列：



由上节介绍可知服装序列为马尔科夫过程，基于马尔科夫过程的统计模型则称为马尔科夫模型。

若小明每天的服装与前一天所穿服装是相互独立的，且每天所穿的服装是受到当天的精神状态影响，精神状态符合马尔科夫链原理，令服装序列为 $O=\{o_1,o_2,\ldots,o_N\}$，精神状态序列为 $I=\{i_1,i_2,\ldots,i_N\}$，用下图描述这一过程：



其中服装序列是可观测序列且观测变量是相互独立的，精神状态序列是不可观测序列（隐变量）且状态变量符合马尔科夫模型，基于上述过程的统计模型称为隐马尔科夫模型。

4.隐马尔科夫模型参数介绍

由上节介绍可知，隐马尔科夫模型包含了观测序列和未观测的状态序列，其中状态序列由初始状态概率向量π和状态转移概率矩阵A决定，观测序列由观测概率矩阵B决定。

还是用小明作为例子，假设小明可选的服装为三件，分别为 v_1，v_2，v_3。小明的精神状态有两种，分别为 q_1，q_2。

因此状态转移概率矩阵A表示为：

$$A=[a_{ij}]_{2\times 2}$$

其中，

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

是在时刻t处于状态 q_i 的条件下在时刻t+1转移到 q_j 的概率。

观测概率矩阵B表示为：

$$B=[b_j(k)]_{2\times3}$$

其中，

$$b_j(k)=P(o_t=v_k\mid i_t=q_j),k=1,2,3;j=1,2$$

是在时刻t处于状态 q_j 的条件下生成观测 v_k 的概率。

初始状态概率向量 π 表示为：

$$\pi=(\pi_i)$$

其中，

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

是时刻t=1处于状态 q_i 的概率。

我们用参数λ表示隐马尔科夫模型参数，即：

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

A，B，π称为隐马尔可夫模型的三要素。

5. 隐马尔可夫模型参数求解算法

上节我们介绍了隐马尔可夫模型参数的含义，如何仅通过观测序列 $O=\{o_1,o_2,\ldots,o_T\}$，求解模型参数 λ。

我们知道隐马尔可夫过程包含了隐变量 I，那么隐马尔可夫模型事实上是一个含有隐变量的概率模型：

$$P(O\mid\lambda)=\sum_I P(O,I\mid\lambda)=\sum_I P(O\mid I,\lambda)P(I\mid\lambda)$$

当构建包含隐变量的模型参数时，我们首先想到用EM（期望最大值）算法实现，算法可参考文章（一文让你完全入门EM算法）

模型参数求解步骤：

1）初始化隐马尔可夫模型参数$\overline{\lambda}$

2）确定模型完全数据的对数似然函数，隐马尔可夫模型的完全数据包含了观测数据$O=\{o_1,o_2,\ldots,o_T\}$和隐数据$I=\{i_1,i_2,\ldots,i_T\}$，完全数据是$(O,I)=(o_1,o_2,\ldots,o_T,i_1,i_2,\ldots,i_T)$

因此完全数据的对数似然函数为：$\log P(O,I\mid\lambda)$

3）EM算法的E步，即求完全对数似然函数下隐变量的期望，称为Q函数$Q(\lambda,\overline{\lambda})$

有：

$$Q(\lambda,\overline{\lambda})=\sum_I \log P(O,I\mid\lambda)P(O,I\mid\overline{\lambda})$$

其中$\overline{\lambda}$是隐马尔科夫模型的当前估计值，λ是马尔科夫模型参数。

根据上节介绍的隐马尔可夫模型参数λ的含义，完全数据的似然函数可展开为：

$$P(O,I\mid\lambda)=\pi_{i_1}b_{i_1}(o_1)a_{i_1i_2}b_{i_2}(o_2)\cdots a_{i_{T-1}i_T}b_{i_T}(o_T)$$

因此Q函数$Q(\lambda,\overline{\lambda})$可写成：

$$Q(\lambda,\overline{\lambda})=\sum_I \log\pi_{i_1}P(O,I\mid\overline{\lambda})+\sum_I\Big(\sum_{t=1}^{T-1}\log a_{i_ti_{t+1}}\Big)P(O,I\mid\overline{\lambda})+$$
$$\sum_I\Big(\sum_{t=1}^{T}\log b_{i_t}(o_t)\Big)P(O,I\mid\overline{\lambda})$$

4） EM算法的M步，求Q函数的最大值，得到模型参数λ，由上节可知模型参数

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

即令：

$$\frac{\partial Q(\lambda,\overline{\lambda})}{\partial \pi_i}=0$$

$$\frac{\partial Q(\lambda,\overline{\lambda})}{\partial a_{ij}}=0$$

$$\frac{\partial Q(\lambda,\overline{\lambda})}{\partial b_j(k)}=0$$

约束条件为初始状态概率分布的和等于1，即：

$$\sum_{i=1}^{N}\pi_i=1$$

状态已知的情况下，观测概率分布的和等于1，即：

$$\sum_{k=1}^{M}b_j(k)=1$$

由上面的等式得到模型参数π_i，a_{ij}和$b_j(k)$的值，即更新了模型参数λ的值。具体计算过程这里不再详细描述了，具体可参考李航老师的《统计学习方法》，若有不懂欢迎交流。

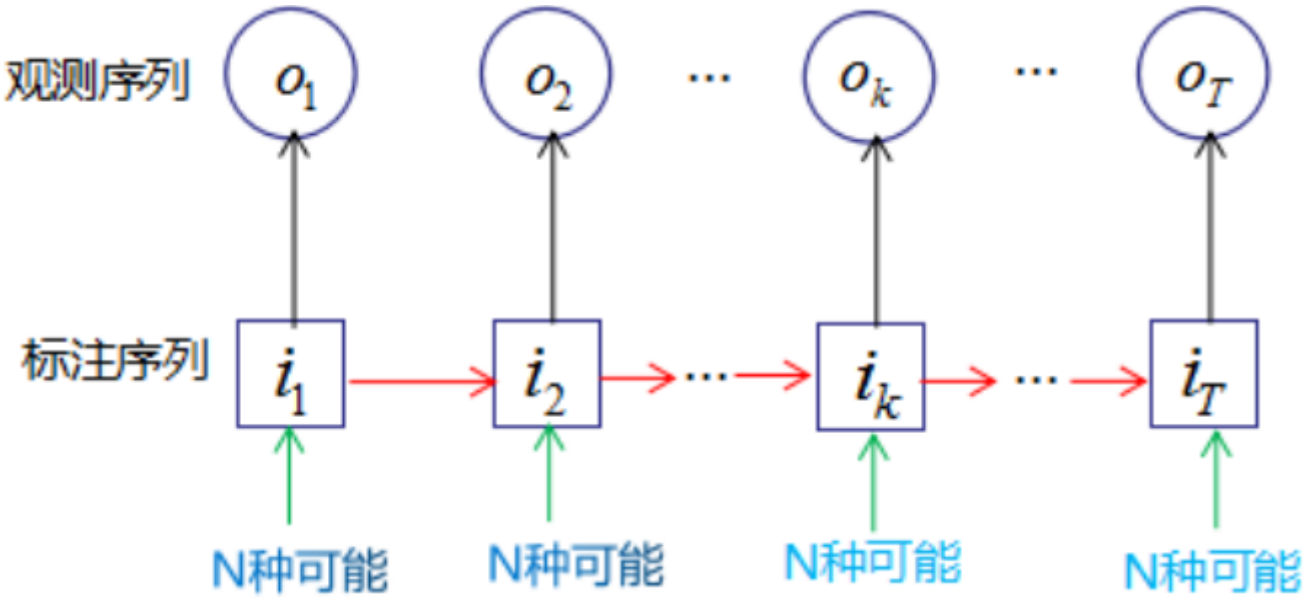
5）重复步骤（3）（4），直到函数$Q(\lambda,\overline{\lambda})$收敛，得到最终的模型参数$\lambda$

6.观测序列概率计算算法

上一节介绍了如何通过观测序列去估计模型参数λ，当模型参数λ已知时，隐马尔可夫模型也相应的确定了，观测序列的概率可以通过模型参数计算出来。下面介绍两种观测序列概率计算算法：

1）枚举法

给定模型$\lambda=(A,B,\pi)$和观测序列$O=\{o_1,o_2,\ldots,o_T\}$，计算观测序列O出现的概率$P(O\mid\lambda)$。假设状态序列$I=\{i_1,i_2,\ldots,i_T\}$，可能的状态数是N，因此每个观测变量都可能由N个状态生成，且模型参数已知，我们就能算出每个可能的状态序列生成观测序列的概率。如下图：



这种列举所有可能的状态数来计算观测序列的概率理论上是可行的，但是计算量非常大，如上图对于长度为T的观测序列，复杂度达到了$O(TN^T)$。

2）递推法

隐马尔科夫具有时间序列的特点，因此我们可以用递推的方法去计算观测序列出现的概率。给定马尔科夫模型λ，定义到时刻t部分观测序列为$o_1,o_1,\ldots,o_t$且状态为q_i的前向概率为$\alpha_t(i)$。

$$\alpha_t(i)=P(o_1,o_2,\ldots,o_t,i_t=q_i\mid\lambda)$$

1）时刻t=1时刻的观测概率为：

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

$$P(o_1 \mid \lambda) = \sum_{i=1}^{N} \alpha_1(i)$$

2）递推，对于$t = 1,2,\ldots,T-1$，有：

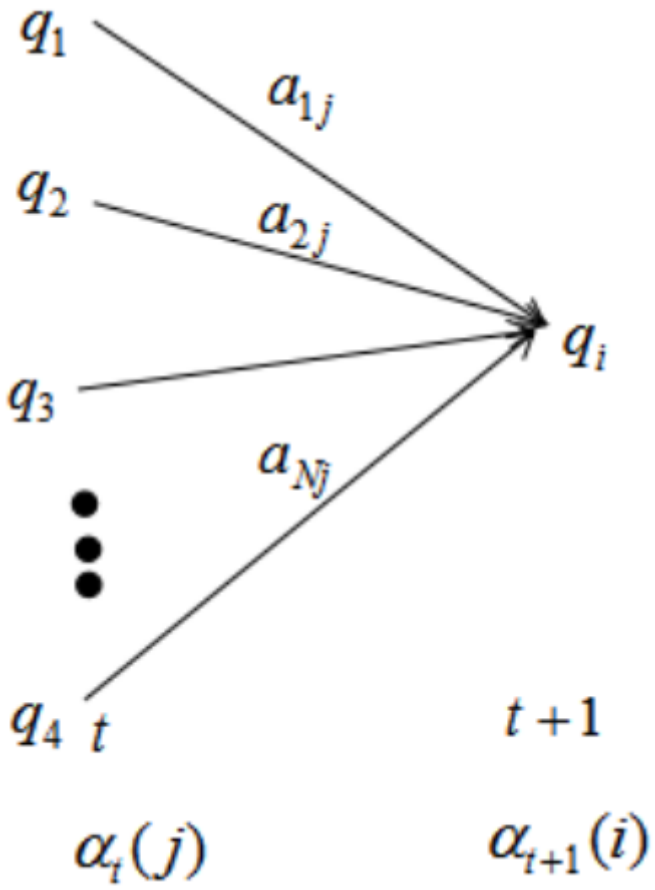
$$\alpha_{t+1}(i) = \left[\sum_{j=1}^{N} \alpha_t(j) a_{ji}\right] b_i(o_{t+1}), \quad i = 1,2,\ldots,N$$

$$P(o_1, o_2, \ldots, o_{t+1} \mid \lambda) = \sum_{i=1}^{N} \alpha_{t+1}(i)$$

3）当t=T时，

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

算法复杂度研究：因为该算法考虑的是用递推关系求观测序列的概率，递推过程如下图：



方程表示为：

$$\alpha_{t+1}(i) = \sum_{j=1}^{N} \alpha_t(j) a_{ji}$$

由上式可知，计算t+1时刻状态为 i 的计算量为N次相乘求和，且t+1时刻状态的可能数为N，因此由t时刻递推到t+1时刻的计算量为$O(N^2)$阶，观测序列的长度为T，那么计算量为$O(N^2T)$阶，与之前枚举法相比，计算量大大降低了。

7.状态序列预测算法

给定模型参数$\lambda=(A,B,\pi)$和观测序列$O=\{o_1,o_2,\ldots,o_T\}$，如何预测最有可能的状态序列，这是隐马尔科夫模型的应用最广的场景，如词性标注，给定一个句子，标注每个单词的特性；

本节介绍维特比算法（Viterbi algorithm）来预测状态序列。维特比算法是一种动态规划算法，用于寻找最有可能产生的状态序列。

通过节点

算法的核心思想是：若t时刻最有可能的状态序列$I=(i_1,i_2,\ldots,i_t^*)$通过节点i_t^*（i_t^*为t时刻的状态），那么从t时刻到T时刻的最优路径一定包括i_t^*。我们利用这一思想确定了最优状态序列的最后一个时刻的状态i_T，然后利用该状态回溯时刻$t=1,2,\ldots,T-1$的最优状态。用一个例子说明维特比算法：

已知模型$\lambda=(A,B,\pi)$，观测集合V={红，白}，状态集合Q={1，2，3}，其中

$$A=\begin{bmatrix}0.5,0.2,0.3\\0.3,0.5,0.2\\0.2,0.3,0.5\end{bmatrix}\quad B=\begin{bmatrix}0.5,0.5\\0.4,0.6\\0.7,0.3\end{bmatrix}\quad \pi=(0.2,0.4,0.2)^T$$

若观测序列O=(红，白，红)，求最优状态序列$I=(i_1^*,i_2^*,i_3^*)$

解：定义$\delta_t(i)$为时刻t状态为 i 的所有单个路径$(i_1,i_2,..,i_t)$中概率的最大值，含义为给定前t-1时刻的状态和前t时刻的观测序列，求最优路径t时刻的状态，即：

$$\delta_t(i)=\max_{i_1,i_2,\ldots,i_{t-1}} P(i_t=i,i_{t-1},\ldots,i_1,o_t,\ldots,o_1\mid\lambda)$$

定义$\Psi_t(i)$为时刻t状态为 i 的所有单个路径$(i_1,i_2,..,i_{t-1},i)$中概率最大路径的第t-1个节点为：

$$\Psi_t(i)=\arg\max_{1<i<N}[\delta_{t-1}(j)a_{ji}]$$

根据维特比算法的核心思想，我们计算观测序列下的最优路径：

1）t=1时，令$\delta_1(i)$是观测为o_1状态为i的概率，由题目可知o_1为红色，有：

$$\delta_1(i) = \pi_i b_i(o_1) = \pi_i b_i(红), \quad i = 1,2,3$$

代入已知条件得：

$$\delta_1(1) = 0.1 \ , \ \delta_1(2) = 0.16 \ , \ \delta_1(3) = 0.28$$

记

$$\Psi_t(i) = 0 \ , i = 1,2,3$$

2）t=2时，

$$\delta_2(i) = \max_{1 \le j \le 3}[\delta_1(j) a_{ji}] b_i(o_2)$$

$$\Psi_2(i) = \arg \max_{1 \le j \le 3}[\delta_1(j) a_{ji}] \ , i = 1,2,3$$

由上面两式得：

$$\begin{aligned} \delta_2(1) &= \max_{1 \le j \le 3}[\delta_1(j) a_{j1}] b_1(o_2) \\ &= \max_j \{0.10 \times 0.5, 0.16 \times 0.3, 0.28 \times 0.2\} \\ &= 0.028 \end{aligned}$$

$$\Psi_2(1)=3$$

$$\delta_2(2)=0.0504,\quad \Psi_2(2)=3$$

$$\delta_3(3)=0.042,\quad \Psi_3(3)=3$$

2）t=3时，

$$\delta_3(i)=\max_{1\le j\le 3}[\delta_2(j)a_{ji}]b_i(o_3)$$

$$\Psi_3(i)=\arg\max_{1\le j\le 3}[\delta_2(j)a_{ji}],i=1,2,3$$

得：

$$\delta_3(1)=0.00756,\quad \Psi_3(1)=2$$

$$\delta_3(2)=0.01008,\quad \Psi_3(2)=2$$

$$\delta_3(3)=0.0147,\quad \Psi_3(3)=3$$

以P^*表示最优路径的概率，最优路径的终点i_3^*：

$$i_3^*=\arg\max_i[\delta_3(i)]=3$$

开始回溯找到其他时刻的最优节点i_2^*，i_1^*

t=2 时，$i_2^*=\Psi_3(i_3^*)=\Psi_3(3)=3$

t=1 时，$i_1^*=\Psi_2(i_2^*)=\Psi_2(3)=3$

因此最优状态序列：

$$I^* = (i_1^*, i_2^*, i_3^*) = (3,3,3)$$

8.小结

本文首先用一个例子通俗的讲解了隐马尔可夫模型的适用场景以及隐马尔可夫模型的三个主要处理问题，然后循序渐进的介绍了隐马尔可夫模型的概念和相关问题的解决方法，所用的例子选取了《统计学习方法》的内容。后续文章会介绍另一种相似的算法——条件随机场以及两者的区别，希望对初学者能有所帮助。

推荐阅读

520 页机器学习笔记！图文并茂可能更适合你，文末附下载方法

李航老师《统计学习方法》(第2版）课件分享，文末附下载

Github | 吴恩达新书《Machine Learning Yearning》完整中文版开源

经典好书 | 141页的《Deep Learning with PyTorch》开源书籍

欢迎扫码关注：