# 前言

前段时间为处理查询解析研究了隐马模型，最近又在补充概率模型相关的数学基础，根据自己的理解整理了这篇文章，来详细讨论下隐马尔科夫模型。本文的公式和算法都是笔者查阅资料并根据自己的理解推导整理的，由于水平有限，难免有描述不准确之处，请大家批评指正。

# 内容概要

1. 隐马尔科夫模型数学推导
2. 隐马尔科夫模型建模及概率矩阵的计算
3. 隐马尔科夫模型使用，Viterbi算法
4. Viterbi算法的推导，动态规划法

# 涉及概念

样本空间，随机变量，概率质量函数，条件概率，联合概率，随机过程，独立随机过程，马尔科夫随机过程，动态规划法等；

# 隐马数学推导

一个例子：本文介绍的模型主要用于词性标注，通篇也会以词性标注作为例子。词性标注就是词序列到词性序列的映射，将句子中的每个单词标注成正确词性的过程。不过本文的词性并不是指传统的动词名词形容词，而是自定义的song，artist，album等。

## 词集合和词性集合

词性标注过程和两个实体集合相关，词序列和词性序列，示例如下：

词序列：<周杰伦，十月，演唱会，唱的，七里香>

词性序列： <“artist”，“data”，“place”，“join”，“song”>

本文设所有词集合为A，所有词性集合为B。a属于A为任一具体词，b属于B为任一具体的词性，从概率模型的角度去思考，词a的词性b可近似取使词是a同时词性为b概率最大的词性。即词a的词性b=argmax{p(ab)}，同理词序列A1=a0a1...an词性序列为B1=b0b1...bn满足B1=argmax{P{A1B1}}。

## 样本空间和联合概率

为了计算p(ab)，本文假设一个定义在集合A,B上的样本空间S，S中的每个样本点都是一个词和词性的pair， 则S可定义为：



样本空间S中的每个样本点表示这个词是a同时词性为b。例如<周杰伦，artist>表示这个词是周杰伦同时词性是artist。

## 随机变量X，Y

很显然S是两个样本空间分别为词集合A和词性集合B的随机变量的联合分布的样本空间，设第一个随机变量为X，以单词集合A为样本空间，其概率质量函数为取某个单词的概率：



这个概率质量函数表示：在空间A中任意取一个单词，这个单词是词a的概率；

设第二个随机变量为Y，以词性集合B为样本空间，其概率质量函数为取某个词性的概率。



这个概率质量函数表示，在空间B中凭任意取一个词性，这个词性是b的概率。若是在各自空间中凭空抓出，即两个随机变量X,Y相互独立且和其他外在因素都无关，那么一个单词是a且词性是b的联合概率P(ab)直接由上面两个概率相乘就可得：



同样可以得出词序列A1=a0a1...an和B1=b0b1...bn的联合概率为：

 词序列A1=a0a1...an词性序列为B1=b0b1...bn词序列A1=a0a1...an词性序列为B1=b0b1...bn

但这显然是不可能的…..一个词不会不受任何因素影响而凭空出现，词性更不会和所有条件都无关而凭空产生，词和词性受哪些因素的影响呢？下面逐个分析随机变量X和Y。

## 随机过程X(t)

通过一个例子来分析上述随机变量X，假设在一个训练样例集合中，一共有100个单词，“周杰伦”这个词一共出现过20次，那么根据训练样例很容易可以得出任意取出一个词，该词是“周杰伦”的概率为20%。但现在告诉你，这100个单词中有50个单词被标注为“song”，另外50个被标注为“artist”。而“周杰伦”被标注为“song”的次数为2，被标注为“artist”的次数为18，这时候我们还能不能用Px(a)=0.2来标识“周杰伦”这个词的概率呢？显然是不行的，因为概率的分布受到了另外一个因素（词性）的影响。这时用一个随机变量X已经无法描述词概率的分布，因为训练样本中词性的不同，概率质量函数已经不同了。

这里就引入了随机过程的概念，由上述分析得出取到每个单词a的概率根据词性b的不同而不同，例如周杰伦这个单词，从作为“artist”的所有单词集合A1和作为“song”的所有单词集合A2中分别取一个词，该词是“周杰伦”的概率分别是36%和4%，亦是说X是一组定义在词空间B上的随机变量，随机变量的概率质量函数和词性b相关。

这里定义一个随机过程X(t)来表述单词a在空间A中的分布。X(t)的样本空间是词集合A，集合B是随机过程X(t)的指标集合，即参数t属于词性集合B。随机过程X(t)意味着对于每个词性b属于集合B，X(b)都是样本空间在词集合A上有着不同概率分布的随机变量。随机变量X(b)的概率质量函数为：从词性为b的所有词中随机选取一个词，该词是a的概率：



由上述分析可得出：词a在作为不同的词性时有不同的概率且其概率只和所属词性相关。随机过程X(t)是一个独立的随机过程。

## 随机过程Y(t)

通常情况下我们考虑的不是将一个孤零零的单词a标注为某个词性b，而是将一个词序列a0a1…an标注为b0b1…bn。假设在一个单词序列中，若当前单词的前一个单词词性为artist，那么当前单词被标注为song的概率是要大于被标注为artist的。即当前位置词性的概率受到前序词性的影响。  
 这里设在一个词序列中词性的概率具有马尔科夫性。那么就不能简单的用一个随机变量Y来表示描述词性的概率了。用一个马尔科夫随机过程Y(t)来表示一个词性序列中词性的概率分布，其中t是当前词在词序列中的位置。根据马尔科夫链的性质：



当前随机变量给定Y(t),…,Y(1)的状态的情况下，Y(t+1)取值为b的概率只受当前随机变量Y(t)状态的影响。

一个词性序列中，我们在确定当前词性时，其前序词性已经确定，而当前词性又受到且仅受到前序词性的影响，因此在马尔科夫链中并不需要考虑在每个位置取不同词性的概率。只需要确定首个位置取各种状态的概率，及前一状态到当前状态的转移概率。如此一来词性序列为b0b1…bn的概率P(b0b1…bn)，可计算为：



## 联合概率

由上文叙述可知，词序列A1 = a0a1a2a3a4和词性序列B1=b0b1b2b3b4，当B1满足B1=argmax {P(A1B1)}时，则B1是词序列A1 的词性序列。因此为A1标注词性的过程就变成了求A和所有词性序列Bi的联合概率的过程。而联合概率P(A1B1)如何求呢？有了两个随机过程X(t),Y(t)的定义再来看看词序列A1和词性序列B1的联合概率P(A1B1)。

当序列A1和B1长度为1时联合概率P(A1B1)，即求a0被标注为词性b0的概率用联合概率p(a0b0)为：



当序列A1和B1长度为1时联合概率P(A1B1)，即求a0a1被标注为词性b0b1的联合概率p(a0a1b0b1)为：



由上面公式可推广到A1A2为任意长度n+1时，得出联合概率p(a0…an,b0…bn)的递推式为：



由上述公式计算任意长度n的词序列A1和所有可能词性序列Bi的联合概率，最后磁性标注序列B1的取值为：



# 隐马模型介绍

上一章描述了隐马模型的数学推导，这里总结以下上面讲了什么：

一个词空间A，一个词性空间B。

一个样本空间为词性空间B，指标集为序列位置的马尔可夫随机过程Y(t)

一个样本空间为词空间A，指标集为词性集合的独立随机过程X(t)

给定显性状态序列A1，求出隐性状态序列B1使联合概率P(A1B1)最大，即B1=argmax{P(A1Bi)}。

## 隐马五元组

本文从这里开始将数学模型的概念转变一个称呼：

1. 将词空间A称之为显性状态空间A；
2. 将词性空间B称之为隐性状态空间B；
3. 将随机过程Y(t)首位置的概率分布放入一个矩阵中称之为初始概率矩阵I；
4. 将随机过程Y(t)的状态之间的转移概率放入一个矩阵中并称之为转移概率矩阵S；
5. 将独立随机过程X(t)在所有随机变量X(b)概率分布统一放入一个矩阵中并称之为输出概率矩阵O。

如此一来就有了传统的亦是我们百度hmm时都会搜到的隐马五元组：

1. 隐性状态集合B：词性集合
2. 显性状态集合A：词集合
3. 初始概率矩阵I：词序列的第一个单词是某个词性的概率
4. 转移概率矩阵S：词性之间相互转移的概率
5. 输出概率矩阵O：词性向单词转换的概率

## 概率模型的构建

概率模型构建的过程也就是隐马尔科夫模型训练的过程，也就是指通过训练样例集合来计算概率矩阵I，S，O的过程。下面介绍通过训练样例训练隐马模型的过程

### 训练样例的转换

虾米搜索的hmm模型的训练样例格式为：

（周杰伦|atrist)(唱的|join) (青花瓷|song)

则训练样例集合中所有的样例首先要对训练样例进行分解，将训练样例集合分解成3个集合如下：

初始概率矩阵I训练集TI

< “artist”>

转移概率矩阵S训练集TS

< “artist”, “join”> <“join”, “song”>

输出概率矩阵O训练集TO

< “artist”,“周杰伦”> <“join”,“唱的”> <“song”,“青花瓷”>

### 概率矩阵的构造

初始概率矩阵I：



隐性状态b作为初始隐性状态的概率为：初始概率矩阵训练集TI中<b>的数目除以<\*>的数目。

转移概率矩阵：



隐性状态bi向隐性状态bj转移的概率为：转移概率矩阵训练集TS中二元组<bi,bj>的数目除以二元组<bi,\*>的数目。

输出概率矩阵：



隐性状态bi输出到显性状态aj的概率为：输出概率矩阵训练集TO中二元组<bi,aj>的数目除以二元组<bi,\*>的数目。

# 隐马模型的使用

隐马尔科夫模型使用就是根据词序列A1，找出词性序列B1使P(A1B1)的概率最大。最简单最容易思考的方法就是上文中提到的计算出所有Bi的可能序列时各自P(A1Bi)的大小。但是其复杂度是N的[B]次方，[B]为隐性状态空间中b的数目。

解决计算复杂性问题有一个Viterbi算法。Viterbi算法是一个动态规划法，基本思路是从前到后计算序列的每一个位置，计算出到该位置每一个状态下最大概率值并保存概率值和该值下的前序状态，用于计算后续节点。计算到最后一个位置时，取概率最大的路径，然后从后往前的寻址回去就找到了联合概率最大的隐性状态序列。

## Viterbi算法的递推式

既然是动态规划法就需要找到递推式和递推式的初始值。设R [ai，b]为序列中第i个单词ai标注为词性b时，从0到i整个序列联合概率的最大值。

可以得出R[a,b]满足以下递推式：



将概率转变为矩阵的值之后，递推公式转变为：



则最终长度为N的序列的M和N的联合概率最大值为：



## Viterbi算法递推式的推导



如上图所示，a0，a1，a2是一个词序列，每个词可能被标注为b0，b1，b2。下面根据Viterbi算法解析每一步的计算

当i=0时标注a0，对词a0可能的标注值b0，b1，b2分别计算其概率

R[a0, b0]=P(a1|b0)\*P(b0|init) 将该值保存到某个位置

R[a0, b1]=P(a1|b1)\*P(b1|init) 将该值保存到某个位置

R[a0, b2]=P(a1|b2)\*P(b2|init) 将该值保存到某个位置

当i=1时标注a1，对词a1可能的标注值b0，b1，b2根据上一步保存的R[a0, b0]，R[a0, b1]，R[a0, b2]分别计算R[a1,b0],R[a1,b1],R[a1,b2]。

当a1被标注为b0时，由上图得知可能由三条路径得来给别是a0取b0，b1，b2时，而a0取b0，b1，b2时联合概率的最大值分别为R[a0, b0]，R[a0, b1]，R[a0, b2]并由上一步保存，得出：

 该结果保存在某个位置

当a1被标注为b1时，由上图得知可能由三条路径得来给别是a0取b0，b1，b2时，而a0取b0，b1，b2时联合概率的最大值分别为R[a0, b0]，R[a0, b1]，R[a0, b2]并由上一步保存，得出

 该结果保存在某个位置

当a1被标注为b2时，由上图得知可能由三条路径得来给别是a0取b0，b1，b2时，而a0取b0，b1，b2时联合概率的最大值分别为R[a0, b0]，R[a0, b1]，R[a0, b2]并由上一步保存，得出

 该结果保存在某个位置

其中P(b|b)指前一个单词被标注为b时当前单词被标注成b的概率。右上边相同的公式可带入求出所有的S2，直到求出所有的Sn-1，通过求所有Sn-1中的最大值便可求出最大的联合概率，然后按照从后往前回溯便可求出整个标注序列。