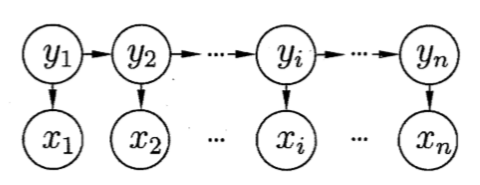
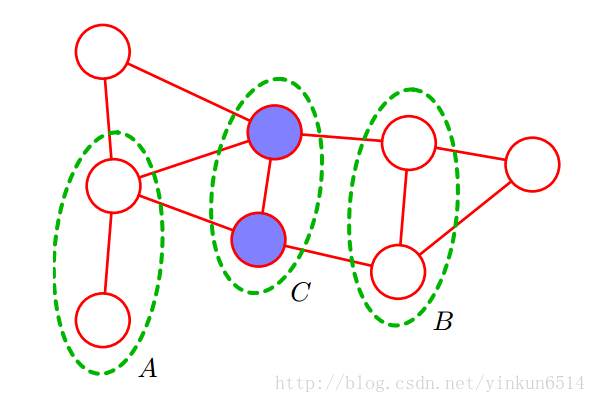
1.概述

概率图模型顾名思义就是把概率的知识与图的知识结合起来，将各种图模型赋予概率的意义，用图的特性来简洁紧凑的表达变量之间的关系。

分为有向图模型和无向图模型，有向图模型用来表示变量之间的依赖关系，无向图用来表示表示变量间的相关关系。





连接词：下面我们来讲一种简单的有向图模型隐马尔可夫模型

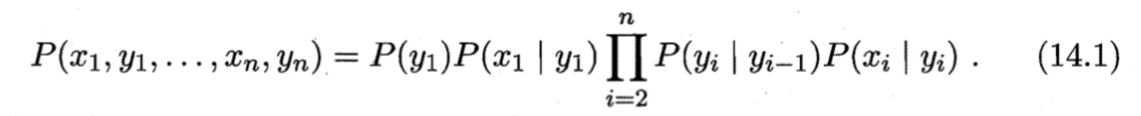
2.隐马尔可夫模型

2.1 HMM简介

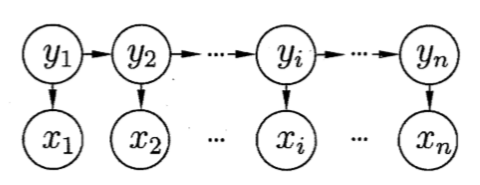
隐马尔科夫模型是一种最简单的动态有向图模型，因为其加上了时间序列所以是一种动态的模型，在NLP方面有比较大的作用

如下图这种就是HMM 那么其中，其中∈У表示第i时刻的系统状态.通常假定状态变量是隐藏的、不可被观测的，因此状态变量亦称隐变量。第二组是观测变量，其中∈χ表示第i时刻的观测值.在隐马尔可夫模型中，系统通常在多个状态之间转换，因此状态变量的取值范围У(称为状态空间)通常是有N个可能取值的离散空间。观测变量可以是离散型也可以是连续型，为便于讨论，我们仅考虑离散型观测变量，并假定其取值范围χ为

从图中可看出变量间的依赖关系，t时刻的观测变量仅依赖于t时刻的状态变量，t>1时刻的状态变量仅依赖于t-1时刻的状态变量，即系统下一时刻的状态仅由当前状态决定，不依赖于以往的任何状态.基于这种依赖关系，所有变量的联合概率分布为（现在决定未来）：



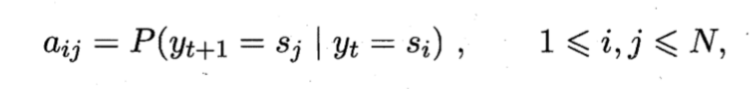
所有的相乘关系都表示慨率的相互独立。三种慨率，，分别表示初始状态概率，输出观测概率和条件转移概率。



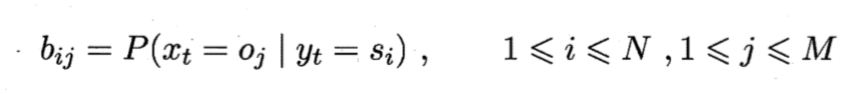
2.2确定一个HMM

那么要确定一个隐马尔可夫模型除了结构信息，还需要知道下面三个参数：

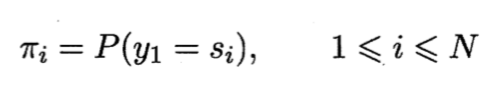
①状态转移概率:记为矩阵



②输出观测概率：记为矩阵



③初始状态概率：记为



通过指定状态空间У、观测空间χ和上述三组参数，就能确定一个隐马尔可夫模型，通常用其参数λ=[A，B，]来指代。给定隐马尔可夫模型λ，可以按照一下的过程来产生观测序列：

（1）设置t=1，并根据初始状态概率选择初始状态;

（2）根据状态和输出观测概率B选择观测变量取值;

（3）根据状态和状态转移矩阵A转移模型状态，即确定;

（4）若t<n，设置t=t＋1，并转到第（2）步，否则停止.

2.3HMM要解决的问题

三个基本问题

①如何评估模型与观测序列之间的匹配程度，例如在已知序列的情况下推测最有可能的

②如何根据观测序列推断出隐藏的模型状态，如在语音识别中语音是观测值，那么文字就是隐藏的状态，通过语音识别出文字就是第二类问题

③如何训练模型使得其能更好的描述观测数据，即如何调整模型参数λ=[A，B，]使得该序列出现的概率最大？有些参数人工指定是很难的，那么如何根据训练样本获得最好的模型参数就是第三个问题

对于这三个问题基于条件独立性均可有效的求解

下面介绍一种无向图模型：

3.马尔可夫随机场（MRF）

3.1马尔科夫网

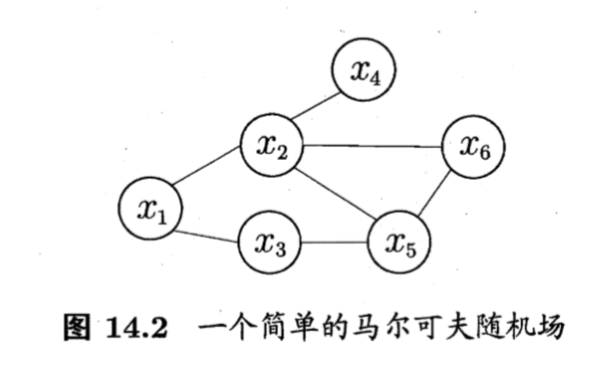
几个概念：①结点代表一个or一组变量

②边代表变量间的依赖关系

③势函数（因子）：定义在变量子集上的非负实函数，用于定义概率分布函数

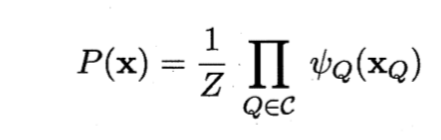
④团：对于下面这幅图中结点的一个子集，如果这个子集中的任意两结点间都有边连接，则称该子集为一个团

⑤极大团：若在一个团中加入其它任何一个结点都不能形成团则称这个团为一个极大团，显然每个结点至少出现在一个极大团中



3.2马尔可夫随机场联合概率

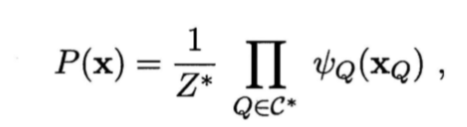
在马尔可夫随机场中，多个变量之间的联合概率分布能基于团分解为多个因子（势函数）的乘积，每个因子（势函数）仅与一个团相关。

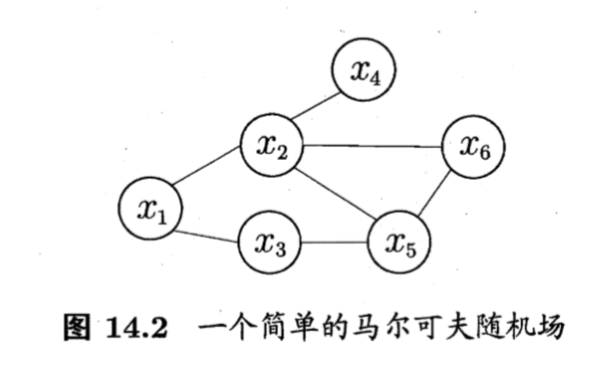


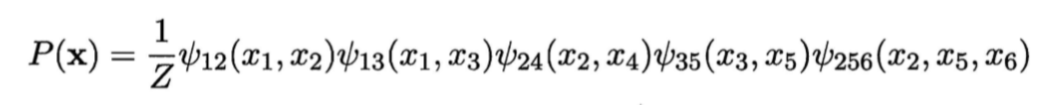
那么在上式之中团与团之间是相互独立的，上面的联合概率计算式中所有团构成的集合为C，与团Q∈C对应的变量集合为

其中为与团Q对应的势函数，用于对团Q中的变量关系进行建模，Z为规范化因子，确保是被正确定义的概率

对于上面的式子来说变量个数太多的话，那么团数也会很多，那么乘积项也会很多，会加大计算量，我们注意到团Q若不是极大团，则必被一个极大团所包括，这说明间的关系不仅在体现了，还在中体现了，会导致重复，所以我们把以极大团来进行定义极大团构成集合为则：

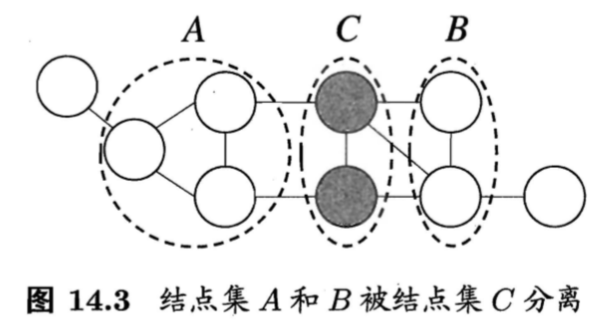


例如这个图中的联合概率密度为：



3.3全局马尔可夫性

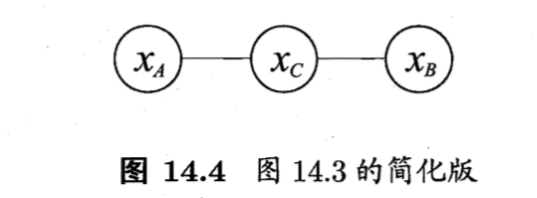
在马尔可夫随机场中如何得到“条件独立性”呢?同样借助“分离”的概念，如图14.3所示，若从结点集A中的结点到B中的结点都必须经过结点集C中的结点，则称结点集A和B被结点集C分离，C称为“分离集”



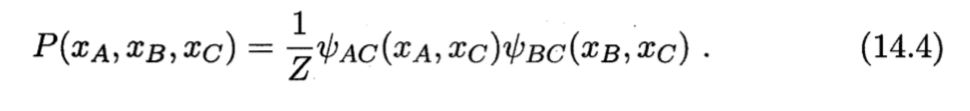
那么我们定义全局马尔科夫性：

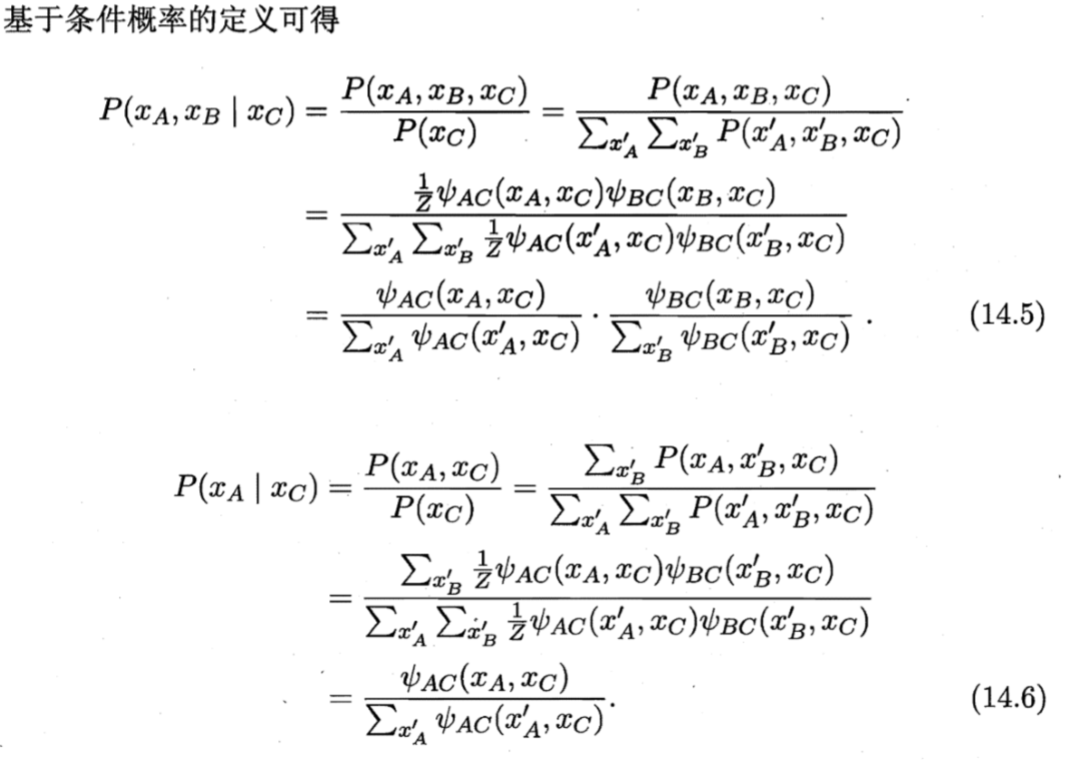
“全局马尔可夫性”:给定两个变量子集的分离集，则这两个变量子集条件独立.将上图记为

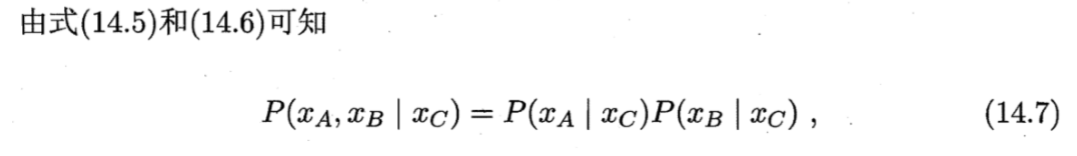
对上图进行简化将结点集简化为单变量



由公式可得联合概率



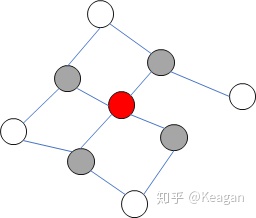






3.4全局马尔可夫性推论

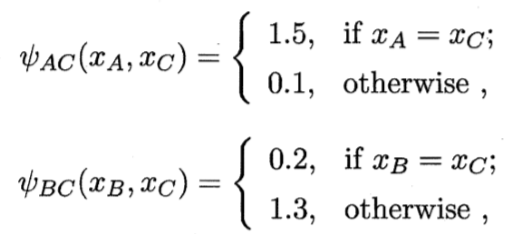
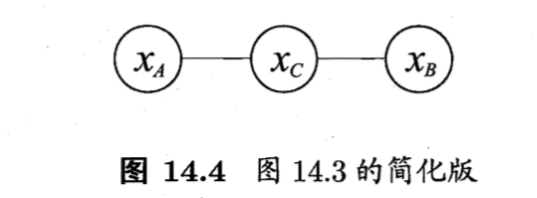
①局部马尔科夫性（马尔可夫毯）：给定某变量的邻接变量，则该变量条件独立于其他变量.形式化地说，令V为图的结点集，n(v)为结点v在图上的邻接结点，n\*(v)=n(v)∪{v}，有



②成对马尔可夫性(pairwise Markov property):给定所有其他变量，两个非邻接变量条件独立.形式化地说，令图的结点集和边集分别为V和E，对图中的两个结点u和v，若，则

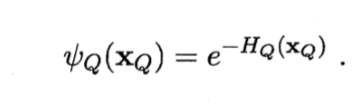
3.5势函数

现在我们来考察马尔可夫随机场中的势函数．显然，势函数的作用是定量刻画变量集中变量之间的相关关系，它应该是非负函数，且在所偏好的变量取值上有较大函数值.例如，假定图14.4中的变量均为二值变量，若势函数为

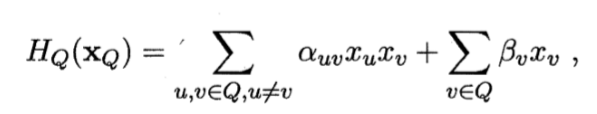


则说明该模型偏好变量与拥有相同的取值，与拥有不同的取值；换言之，在该模型中与正相关，与负相关.结合之前计算联合概率密度的公式易知，令与相同且与不同的变量值指派将取得较高的联合概率。

为了满足非负性，指数函数常被用于定义势函数，即



是一个定义在变量上的实值函数，常见形式为



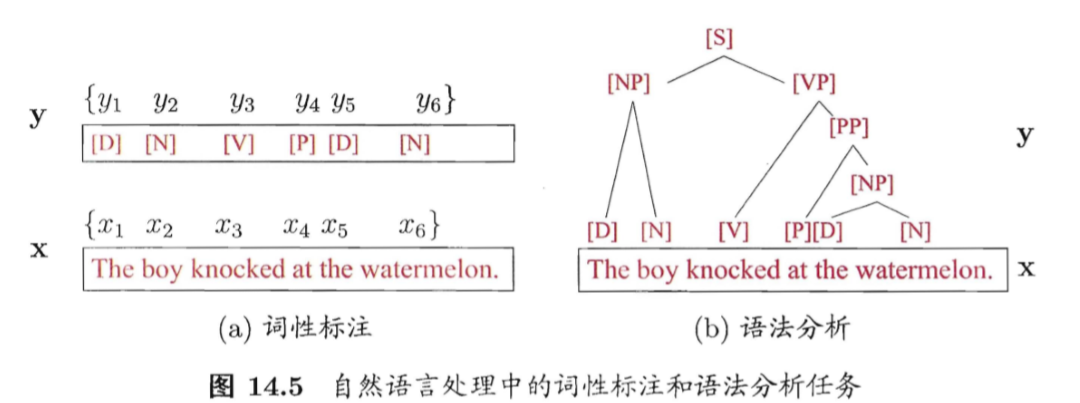
其中和是参数.上式中的第二项仅考虑单结点，第一项则考虑每一对结点的关系。

4.条件随机场

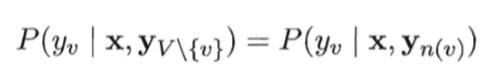
4.1概念

与前面的HMM和马尔可夫随机场模型不同，条件随机场是一种判别式无向图模型，前面两种为生成式模型，模型14.1节提到过，生成式模型是直接对联合分布进行建模，而判别式模型则是对条件分布进行建模

CRF试图对多个变量试图对多个变量在给定观测值后的条件概率进行建模.具体来说，若令x = 为观测序列，y = 为与之相应的标记序列，则条件随机场的目标是构建条件概率模型P(y│x)。需注意的是，标记变量***y***可以是结构型变量，即其分量之间具有某种相关性。例如在自然语言处理的词性标注任务中，观测数据为语句(即单词序列)，标记为相应的词性序列，具有线性序列结构，如图14.5(a)所示;在语法分析任务中，输出标记则是语法树，具有树形结构，如图14.5(b)所示



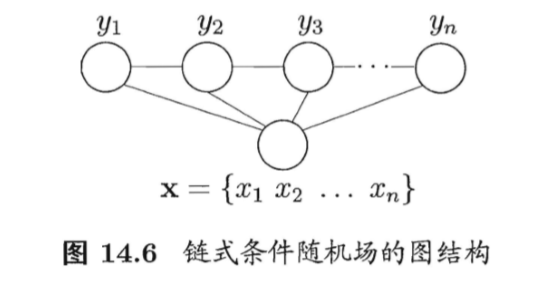
令G=表示结点与标记变量***y***中元素一一对应的无向图，表示与结点v对应的标记变量，n(v)表示结点v的邻接结点，若图G的每个变量都满足马尔科夫性即



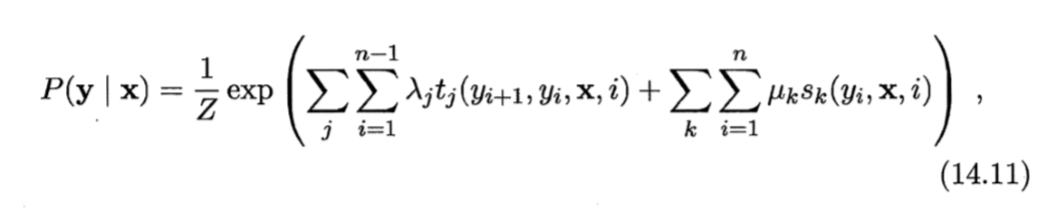
则**（*y，x*）**构成一个条件随机场

4.2CRF的条件概率

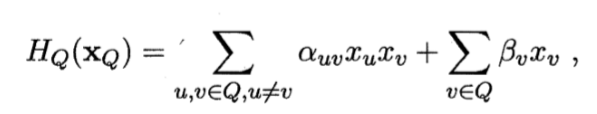
理论上来说，图G可具有任意结构，只要能表示标记变量之间的条件独性关系即可。但在现实应用中，尤其是对标记序列建模时，最常用的仍是图14.6所示的链式结构，即“链式条件随机场”下面我们主要讨论这种条件随机场。



类似于马尔科夫随机场，CRF也使用势函数与图结构上的团来定义。图14.6所示的链式条件随机场主要包含两种关于标记变量的团，即单个标记变量以及相邻的标记变量。选择合适的势函数,即可得到形如式(14.2)的条件概率定义

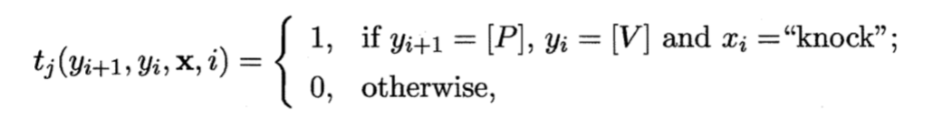


类似于之前的函数

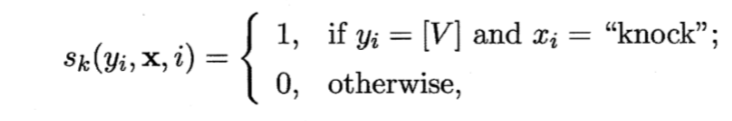


其中是定义在观测序列的两个相邻标记位置上的转移特征函数，用于刻画相邻标记变量之间的相关关系以及观测序列对它们的影响，是定义在观测序列的标记位置i上的状态特征函数，用于刻画观测序列对标记变量的影响，和为参数,Z为规范化因子

显然,要使用条件随机场,还需定义合适的特征函数.特征函数通常是实值函数，以刻画数据的一些很可能成立或期望成立的经验特性.以图14.5(a)的词性标注任务为例，若采用转移特征函数



则表示第i个观测值为单词“knock”时,相应的标记和很可能分别为[V]和[P].若采用状态特征函数



则表示规测值为单词“knock”时,它所对应的标记很可能为[V].

CRF与马尔科夫随机场形式上差别不大，但是处理的东西不一样，一个是联合概率密度，一个是条件概率密度

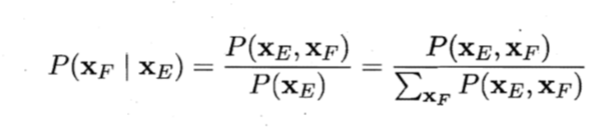
5.学习与推断

5.1概率图模型的推断

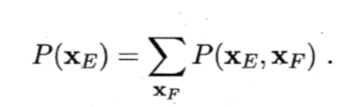
基于概率图模型定义的联合概率分布，我们能对目标变量的边际分布或以某些可观测变量为条件的条件分布进行推断。例如前面的HMM中的，边际分布就是求边缘概率对于离散变量就是对其它变量的求和对于连续就是积分也就是一个边际化的过程

对于分布中的参数贝叶斯学派认为其也是一个随机变量，也有分布因此可以归类到推断的范围之内，因此就不做其它的讨论了。

具体来说，假设图模型所对应的变量集能分为和两个不相交的变量集，推断问题的目标就是计算边际概率P()或条件概率P(|)。由条件概率定义有



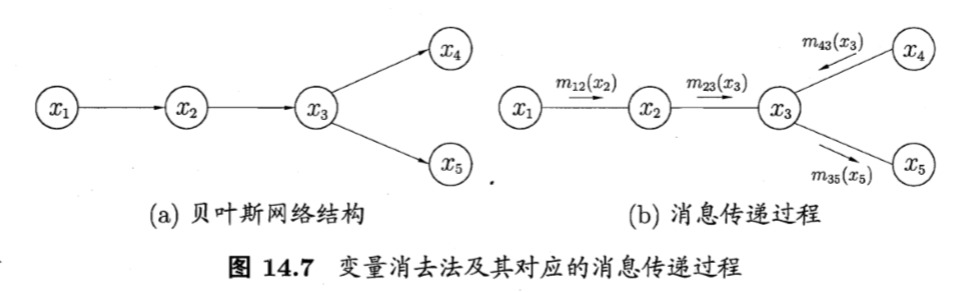
联合概率分布可有概率图模型得出那么重点就在于如何高效的求解



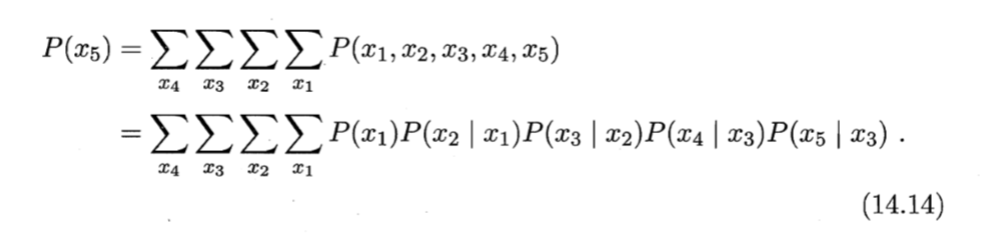
概率图模型的推断方法分为两类，第一类是精确推断过程复杂计算复杂度随团数指数级增长，另一类是近似推断在低时间复杂度情况下获得近似解，下面我们介绍第一类

5.2变量消去法

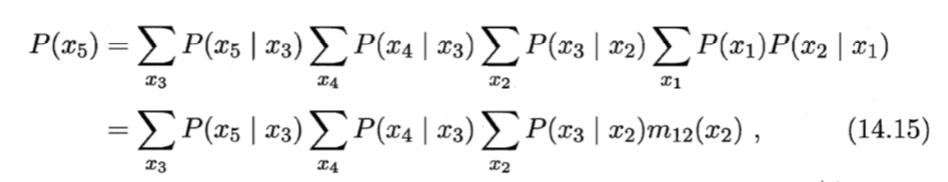
是精确算法的一种基础算法，其它算法是根据这个演变出来的。以下图来描述工作流程



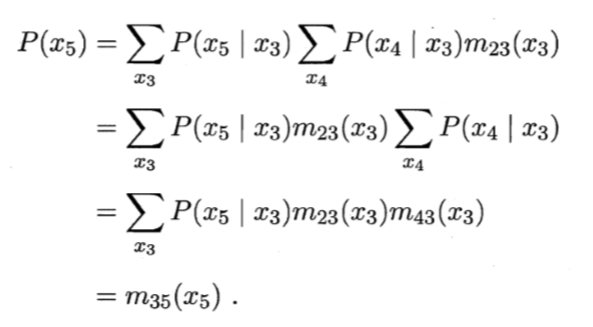
假设我们要推断边际概率，因为我们假设基于离散变量，根据有向图模型我们可以写成下面的形式



若采用的消去顺序计算加法因为值为1且与独立所以可与交换位置

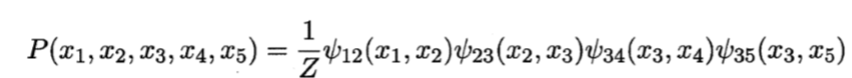


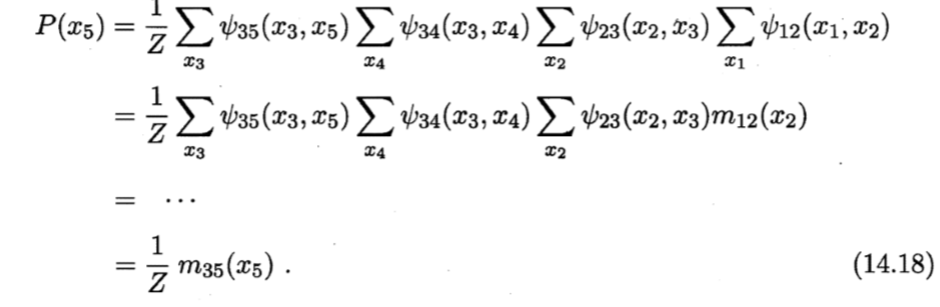
其中是求加过程的中间结果，下标i表示此项是对求加的结果，下标j表示此项中剩下的其他变量.显然是关于的函数.不断执行此过程可得



显然是关于的函数仅仅与有关

对于无向图模型有



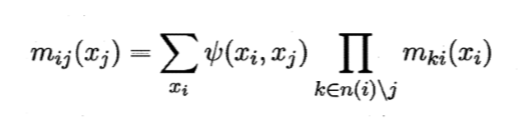


显然，通过利用乘法对加法的分配律，变量消去法把多个变量的积的求和问题，转化为对部分变量交替进行求积与求和的问题。这种转化使得每次的求和与求积运算限制在局部，仅与部分变量有关，从而简化了计算。

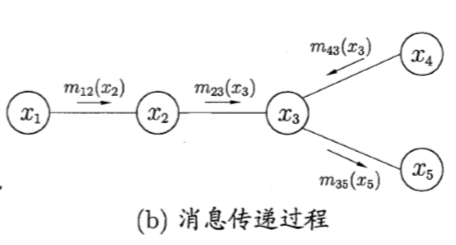
缺点在于在计算多个边际分布时会有重复的冗余计算

5.3信念传播

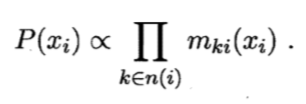
信念传播算法将变量消去法中的求和操作看作一个消息传递过程。较好地解决了求解多个边际分布时的重复计算问题。具体来说变量消去法通过求和操作



消去变量,其中n(i)表示结点的邻接结点.在信念传播算法中,这个操作被看作从向传递了一个消息。这样，式(14.15)和(14.16)所描述的。变量消去过程就能描述为图14.7(b)所示的消息传递过程。不难发现，每次消息传递操作仅与变量及其邻接结点直接相关，换言之，消息传递相关的计算被限制在局部进行



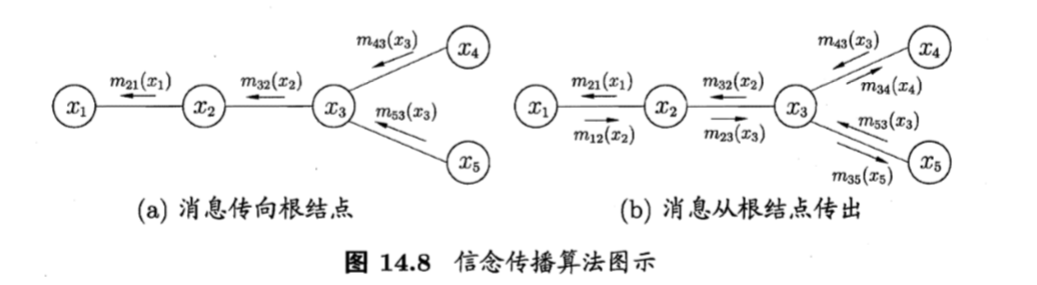
在信念传播算法中，一个结点仅在接收到来自其他所有结点的消息后才能向另一个结点发送消息,且结点的边际分布正比于它所接收的消息的乘积，即



若图结构中没有环,则信念传播算法经过两个步骤即可完成所有消息传递，进而能计算所有变量上的边际分布:

①指定一个根结点,从所有叶结点开始向根结点传递消息，直到根结点收到所有邻接结点的消息;

②从根结点开始向叶结点传递消息，直到所有叶结点均收到消息.



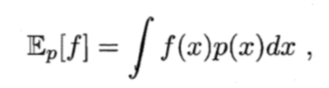
6.近似推断

近似推断主要可以分为两大类：第一是采样，第二是是使用确定性近似完成近似推断

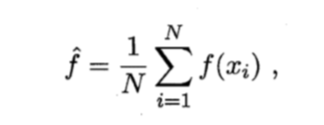
6.1MCMC采样（马尔科夫链蒙特卡罗方法）

采样法的基本思路就是绕过复杂的推断概率分布的过程，直接对目的进行计算推断，比如为了计算某些期望，直接计算逼近期望比先推出分布再计算要简单的多。

例如我们的目标是计算函数f(x)在概率密度函数p(x)下的期望

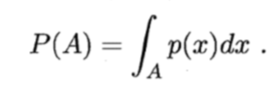


我们可以根据大数定律用样本均值来近似求解f的期望，对于n个独立的样本

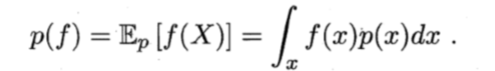


对于概率图模型，那么现在的问题就集中在根据概率图模型描述的概率分布来进行高效的采样了

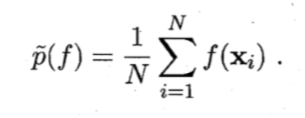
MCMC就是一种高效的采样方法给定属于总体X的一个变量x的概率密度函数，x在区间A中的概率可计算为



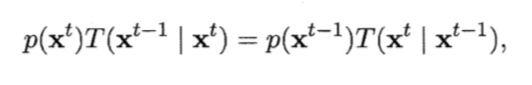
那么对于一个随机变量的函数的期望如下



但是对于高维的多元变量，且服从复杂的分布，则对上式求积分会很困难。为了解决积分复杂的问题，MCMC采用先构造服从泊松分布的独立同分布的随机变量再得到上式的无偏估计



那么这么做还是可能遇到困难，如果概率密度函数很复杂那么构造服从p分布的独立同分布样本也很困难，MCMC的方法的关键就在，关键就在于通过构造“平稳分布为*p*的马尔可夫链”来产生样本:若马尔可夫链运行时间足够长(即收敛到平稳状态)，则此时产出的样本x近似服从于分布*p*.如何判断马尔可夫链到达平稳状态呢?假定平稳马尔可夫链T的状态转移概率(即从状态x转移到状态的概率)为T(x'|x),t时刻状态的分布为p(xt)，则若在某个时刻马尔可夫链满足平稳条件（找到一个p使得下面这个等式成立，那么这个p就是该马尔科夫链的平稳分布）



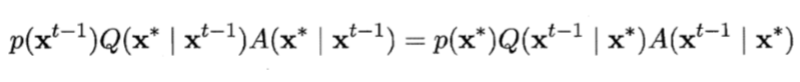
那么是该马尔科夫链的平稳分布，马尔科夫链也收敛到平稳状态

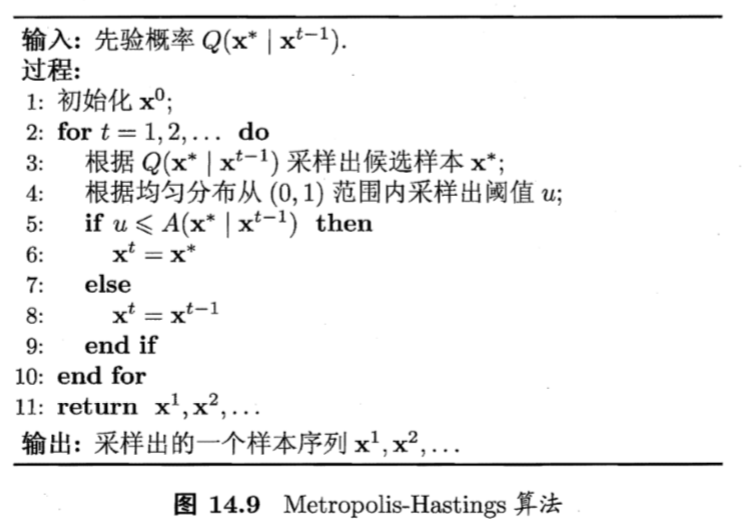
简而言之MCMC就是要构造一个马尔科夫链使得这个马尔科夫链收敛至平稳分布恰为待估计参数的后验分布，然后用马尔科夫链产生符合后验分布的样本，用样本进行对参数进行估计。

那么问题就转变为了对马尔科夫链转移概率的构造，不同方法就会产生不同的采样方法

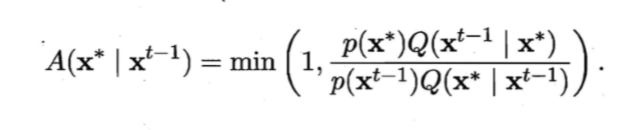
6.2MH算法

这是一种基于拒绝采样的方法来逼近平稳分布p.如图14.9所示，算法每次据上一轮采样结果xt-1来采样获得候选状态样本x\*，但这个候选样本会以一定的概率被“拒绝”掉.假定从状态xt-1到状态x\*的转移概率为，其中是用户给定的先验概率，是x\*被接受的概率.若x\*最终收敛到平稳状态,则根据式(14.26)有

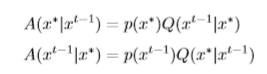




其中接受率为



实际上其实只要满足一下条件那么就可满足平稳状态的等式



但是等式右边的值可能比较小，那么样本利用率就太低了可将接受率设为0.5和1也就是上面的公式来提高样本的接受率

吉布斯采样是MH算法的特例，不同的地方在于例如在采样低n维变量的时候固定其它维度，其步骤如下：

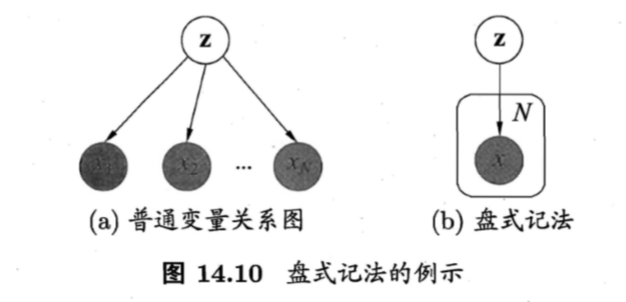
①随机或以某个次序选取某变量;

②根据x中除外的变量的现有取值，计算条件概率，其中

③根据对变量α采样,用采样值代替原值.

6.3变分推断

变分推断通过使用已知简单分布来逼近需推断的复杂分布,并通过限制近似分布的类型,从而得到一种局部最优、但具有确定解的近似后验分布.

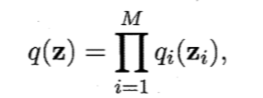


例如这个里面的z是一个隐变量我们要根据观测值x来对z的分布进行估计，以及求解对应参数θ。

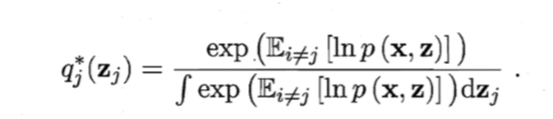
那么概率模型的参数估计在有隐变量的条件下的求解我么用的是EM算法，通过EM算法可以获得稳定的参数θ，隐变量也可由参数来获得。

那么在EM算法的E步中对于变量z的后验分布的近似分布的推断过程中可能会出现困难，那么我们可以用变分推断。

变分推断通常假设z服从分布，这个分布分布是z后验分布的近似分布



所以这个假设就是把分成M个独立的组来方便进行计算，然后经过书上336页一系列复杂的推导最终可以得出一个式子



结论就是在之前的假设之下，变量子集最接近真实情形的分布可由上式给出，也就是说能用这个结论的式子对隐变量z进行高效的推断。

7.话题模型

话题模型是生成式有向图模型，主要用于处理离散型数据

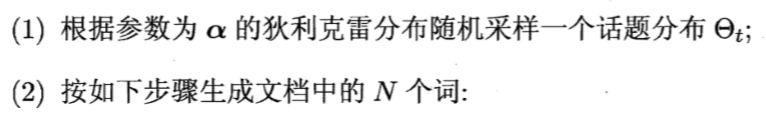
隐狄利克雷分配模型(LDA)是话题模型的杰出代表。在话题模型中，有以下几个基本概念：词（word）、文档（document）、话题（topic）

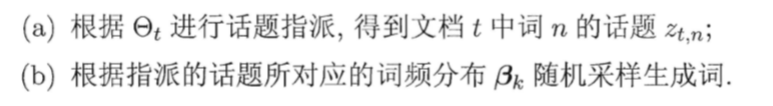
词：最基本的离散单元；

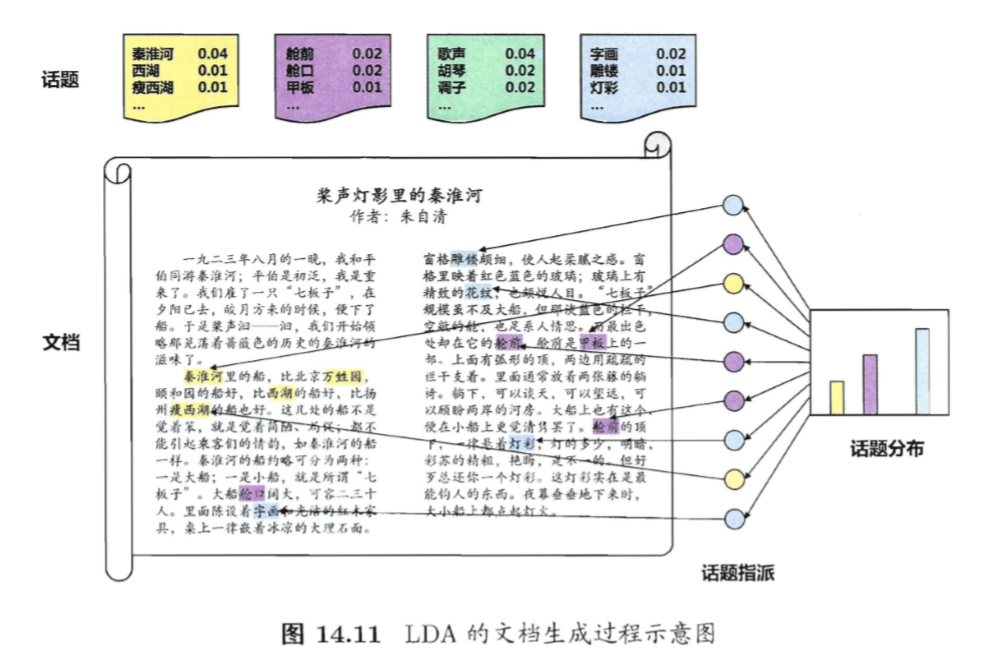
文档：由一组词组成，词在文档中不计顺序；

话题：由一组特定的词组成，这组词具有较强的相关关系。

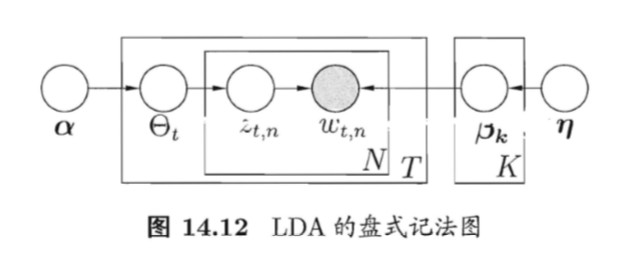
在现实任务中，一般我们可以得出一个文档的词频分布，但不知道该文档对应着哪些话题，LDA话题模型正是为了解决这个问题。具体来说：LDA认为每篇文档包含多个话题，且其中每一个词都对应着一个话题。因此可以假设文档是通过如下方式生成：



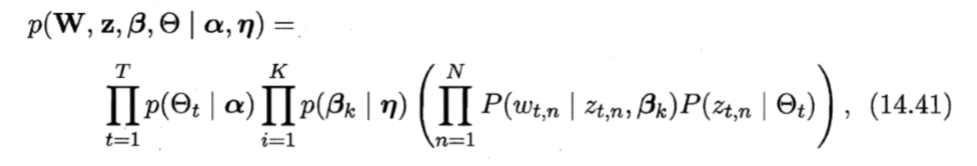




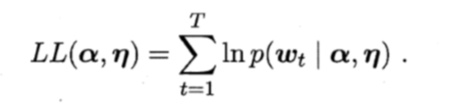
LDA的变量关系可以用下面这个有向图进行描述，其中唯一可以观测到的是



那么根据这个有向图模型我们可以写出LDA对应的概率分布

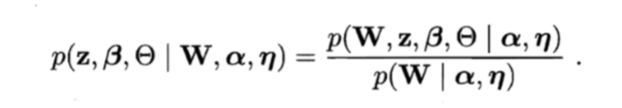


那么α,η是我们要确定的参数，我们可以用极大似然的方法对参数进行估计，对于给定数据，寻找最大对数似然



求解比较困难，所以会采用变分法求近似解

若已知模型就是已知参数α,η，那么根据词频来推断文档话题结构可求解



分母可以通过变分法近似推断或吉布斯采样