0、概述

在《概率与统计》笔记中我们主要介绍了如何用概率和统计的观点去认识和解释世界，最重要的工具是概率论中的随机变量及其分布，最常用的方法是统计方法中（基于统计量的）参数估计和假设检验。

但这是不够的。现实中需要更复杂的模型——我们需要许多相互之间具有复杂关系的随机变量构成的模型来描述我们关心的现象。这些随机变量整体上形成各种特殊的“结构”。我们需要估计模型的参数，并给定其中某些随机变量的取值，推断另外一些随机变量的分布。

这种复杂的模型无法使用简单的随机变量或者随机向量来刻画，需要基于概率图模型来构建；复杂的模型的参数估计和随机变量的推断也没有特别简单的方法。因此我们在本篇笔记中介绍处理这些问题的方法。

注意，在理解这些方法时，一定要刻意地运用“概率”和“统计”的思维方式。

一、EM算法

**1、EM算法的原理**

EM算法用来解决含有隐变量的概率模型的参数估计问题。

我们知道，对于一般的概率模型（本质上就是一个随机变量X的分布），可以用极大似然等方法对其未知参数进行估计。一般的概率模型的形式为：，对于具体的问题，P有具体的形式，所以当固定时，就能基于P的具体形式，对任意给定的样本观测值x，给出其出现的概率。但是对于含有隐变量的概率模型，即便固定，也无法基于一个具体的形式直接给出任意给定样本观测值x的出现概率——还需要知道模型中隐变量Z的取值z才行。

隐变量Z，与直接观测到的随机变量X相对，是模型中不能直接观测到的随机变量。虽然不能直接观测到，但Z也是模型中必需因素。因为模型需要给出X和Z的联合观测值，才能基于具体形式给出其出现的概率值。

注意，隐变量Z并不是一个任意的随机变量，一般来说，它与X的取值空间是一一对应的。通常，Z和X对应于随机试验中先后执行的两个试验。

在《聚类算法》笔记中，介绍了高斯混合模型，它可以看作是一种含有隐变量的模型。它的隐变量是“某个样本由某个高斯分布生成”。

这里再介绍一个简单的隐变量模型：两硬币模型。话说有A和B两个硬币，它们不均匀，掷它们一次，正面朝上的概率分别是和，我们随机选取其中一枚硬币，掷之，可以观测到其正面朝上还是反面朝上，假设我们不知道选取的是哪个硬币。这个试验对应的概率模型就是典型的含有隐变量的模型，在试验中，我们先后执行了两个试验，一个是随机选一个硬币，它对应于隐变量；一个是随机投掷选中的硬币，它对应于观测变量。我们多次重复这个试验，能够多次观察到“硬币朝上还是朝下”这一结果，然后可以用这些观察结果去估计和。

是的，虽然相比一般的概率模型，隐变量模型中多出了不可观测到的隐变量，但是我们仍然能基于观测值，求解出最优的未知参数（而不会因为多出了未知因素，就使结果变成多解）。这就是统计学的魅力，我们总是依据观测到的值，推断总体的情况，我们给出基于已知认识的最好的推测。即使我们了解到的信息比较少，也要充分利用这些信息，给出这些信息所蕴含的最可能的结论。这是多么理性的方法论啊。

对于隐变量模型，无法使用极大似然法估计未知参数，一般使用EM算法。

EM算法是基于极大似然的思想实现的。它迭代地逐步逼近使似然函数最大的参数，每次迭代分为两步（E步和M步）。

**2、EM算法的迭代公式**

EM的迭代公式如下：



该公式首先对隐变量Z求积分，得到关于的函数，这是E步；然后求的极大，这是M步。可以证明，每次迭代都将使似然函数增加，因此该迭代过程是有效的。

注意，虽然隐变量模型需要X和Z两个变量的观测值，才能基于具体形式直接给出观测值的概率值，但是似然函数也是有意义的，只不过它没有简单的形式而已。

在具体的隐变量概率模型中应用EM算法，需要先给出和的具体形式，然后按照EM算法的步骤迭代求解。

二、随机过程与马尔科夫链

**1、随机过程**

随机过程被用来对一个随时间变化的随机变量进行建模，在每个时刻的取值都是一个随机变量，且不同时刻的随机变量之间可能具有“依赖关系”。随机过程的所有时刻的随机变量的取值空间合起来为该随机过程的状态空间。在实际中，随机过程的各随机变量通常共享同一个取值空间，则该取值空间即为该随机过程的状态空间。

若时间t是离散的，则随机过程就是一个随机变量的序列。

具体的随机过程中，有三类比较重要：泊松过程（例如在一个服务窗口前，顾客的到达时刻）、维纳过程（例如布朗运动）、马尔科夫过程。下面仅介绍马尔科夫过程。

**2、马尔科夫链**

马尔科夫过程是指具有马尔科夫性的随机过程。时间离散、状态空间离散的马尔科夫过程称为马尔科夫链。实际中大多数情况下都使用马尔科夫链进行建模，所以下面我们只介绍马尔科夫链。

对于离散的随机过程，马尔科夫性的意义是，任意时刻的随机变量的分布只依赖于前一时刻的随机变量的取值，与再往前的随机变量的取值没有关系。

设状态空间为，则马尔科夫性的形式化表示如下：



注意，马尔科夫性表示某时刻的随机变量的分布依赖于前一时刻的随机变量的取值，而不是前一时刻随机变量的分布。

另外，注意，随机过程中，随机变量的这种依赖关系是有“方向性”的，在马尔科夫链中，后面的依赖前面的，而前面的不依赖后面的。这与概率论中研究的随机变量的“相关关系”有区别，相关关系是两个随机变量之间的一种对称的关系，而依赖关系是不对称的。

（具有依赖关系的随机变量，其相关性是如何的呢？）

可以看出，马尔科夫链只有当前一时刻的状态确定时（即随机变量取到一个观测值），后一时刻的随机变量的分布才确定（也就是说，这个随机变量才确定），如果没有对前一个随机变量进行取值，后一个随机变量就是不确定的。所以不可能一下子确定一个马尔科夫链，而只能从前往后一步一步地取值并确定下一时刻的随机变量。

由于拥有这一特点，马尔科夫链不能简单地看成是“只是一个随机向量”或者“只是一个随机变量的序列”，因为那蕴含着“所有随机变量已经确定”或者“所有随机变量可以同时确定”的意思。

**3、马尔科夫链的状态转移**

根据上面的介绍，我们知道马尔科夫链是一种随机过程，其每个时刻的随机变量的分布依赖于前一时刻随机变量的取值。实际上，我们还可以从另外一种观点——“状态转移”的观点来看待它。

由于马尔科夫链的所有时刻的随机变量都在状态空间中取值，所以它的各随机变量的观测值构成的序列（观测值链）相当于不断地从一个状态转移到下一个状态。根据马尔科夫链的性质，当一个时刻的状态确定时，下一个时刻取各个状态的概率的分布就确定了，这个分布就是，我们从这个分布中采样得到t时刻的状态，就完成了一步状态转移。

对于齐次马尔科夫链，状态转移的概率只与两个状态有关，与时间无关。这时，如果我们知道了状态空间中任意一个状态转移到状态空间中的任意一个状态的概率，那么就能对马尔科夫链给定一个时刻的状态，递推出后面任意时刻的状态。

齐次马尔科夫链的状态空间中各状态之间的转移概率构成一个矩阵：



其中表示第i个状态（一步）转移到第j个状态的概率（也就是某一时刻状态为i时，下一时刻的状态的概率分布中，状态取j的概率），m是状态空间的大小。

除了状态转移概率矩阵外，要想确定一个马尔科夫链，还需要给定初始状态概率分布，因为需要初始概率分布来采样出初始时刻的随机变量的状态，然后才能递推。

关于状态转移，我们还可以计算一个状态经过多步（假设为k步）转移到另一个状态的概率。用条件概率表示这个概率，就是：



以两步转移为例，给定马尔科夫链的状态转移概率矩阵：



则状态i经过两步转移到状态j的概率为：



一般地，状态转移矩阵的k次幂就表示了状态空间中任意一个状态经过k次转移到达任意一个状态的概率。

**4、马尔科夫链的平稳分布**

当马尔科夫链满足一定的条件（不可约、非周期、正常返）时，它有一个重要的性质：无论初始状态概率分布如何，它经过许多次的状态转移，最终将收敛到一个“平稳分布”。也就是说，到达某个时刻以后，无论前一时刻的随机变量取得什么观察值（即取得什么状态），后一时刻的随机变量的分布都不再随着前一时刻的取值发生变化。

我们一般将马尔科夫链达到稳态时的分布记为。设状态转移矩阵为，则满足



注意，前面说过，马尔科夫链只有当前一个随机变量取值确定时，后一个随机变量的分布才确定，所以严格来讲，是一个条件分布，只不过在稳态时，这个条件分布跟前一个随机变量的取值已经无关了，所以可以写作无条件分布。

三、统计推断与随机采样

**1、统计推断中的问题**

统计推断就是利用样本的信息推断总体的信息。最常见的统计推断任务就是参数估计，即基于给定一组样本的观测值，推断总体的均值、方差等参数，或者参数的分布。另外，在概率图模型中，往往需要根据模型中一部分随机变量的联合观察值，去推断另一部分随机变量的联合分布。这些推断任务常常面临着同一个困难。下面列举三种常见的情形。

**（1）贝叶斯估计**

我们在《概率与统计》笔记的附录里介绍了贝叶斯估计方法。它的目标是求出的后验分布（其中D为样本的观测值集合），然后对新的样本出现的概率进行预测。这个过程中，后验分布的求取是关键。

的后验分布的求取需要基于贝叶斯定理，由先验分布和似然分布，以及观测值的概率求出：



式中，D表示样本观测值的集合，是先验分布，是似然分布。注意，上式中最右边的分子中的是一个特定的值，而分母中的是作为积分变量，它跟分子中的并不是同一个变量。

求出后验分布后，我们就可以用后验分布预测新的样本的概率（贝叶斯决策）：



实际中，由于后验分布的形式可能比较复杂，所以积分可能不易计算出来。

**（2）EM算法估计**

前面介绍了EM算法，它用来估计含隐变量模型的参数。它的迭代公式为：



这个式子中，由于分布和的形式可能比较复杂，甚至很难确定解析形式，这个积分同样也不易计算出来。

**（3）概率图模型中的推断**

概率图模型的推断问题为：已知联合概率分布，估计。其中称为“问题变量”，称为“证据变量”，它们都是随机向量。。

当和维数较高，且它们中的随机变量之间关系复杂时，的形式非常复杂，其计算可以使用信念传播算法，但是相当的费时间。

上面三个问题是我们在统计推断中常会遇到的情况。前两个是积分计算的困难，后一个是条件概率计算的困难，它们其实都是由于概率分布形式复杂造成的。除此之外，还有其他类似的情况，这些问题都可以通过近似推断的思路来解决。

近似推断通过对复杂形式的概率分布的近似来简化基于该概率分布的积分或者概率的计算。近似推断包括确定性近似和随机近似两种方法。

确定性近似的代表方法是变分近似，随机近似的主要思路则是通过对概率分布进行随机采样来间接地计算基于概率分布的积分或者概率值。下面主要介绍随机采样。

**2、随机采样**

随机采样，又称蒙特卡洛采样。它的基本思想是对复杂形式的概率分布进行采样，再用样本观察值计算基于该概率分布的积分或者概率。

对于计算积分（如上面介绍的问题（1）和问题（2）），通常是把该积分看作该分布的期望，随机采样后通过估计对应的平均值来近似这个期望。积分的通用形式为：



从p中随机采样m个样本：，然后计算样本均值，作为对s的近似。有人可能会问，既然和的形式都已知，那为什么不直接使用数值积分的方法？这是因为数值积分法不如随机采样法计算效率高。

另外一种情况是计算概率分布上的一个概率，如上面介绍的问题（3），要计算，我们可以先从分布中采m个样本：，然后计算的样本均值，来近似：

。显然，在这个问题中，随机采样的过程中需要避开直接计算（因为如果这个值能很容易计算出来，也用不着费劲拔力地用随机采样了）（Gibbs采样可以实现这个需求）。

上述两种情况是随机采样的两种常见的应用场景。它们的关键都在于从一个形式复杂的分布中采样。常用的蒙特卡洛采样方法是“接受-拒绝采样”和“重要性采样”，蒙特卡洛采样与马尔科夫链结合，又形成了“MCMC采样”。

对统计学理解不深的同学，可能有一种错觉，认为只要能确定一个概率分布的形式，那么对其采样也是很容易的事。

这是不对的。对于形式复杂的分布，即使能够给出其解析的形式，但是从该分布中采样也是很难的。比如你可以想一下对标准正态分布进行采样，应该使用什么方法？该方法是否很简单？

实际上是不简单的。一般来说，只有均匀分布是比较容易采样的，其他形式的分布的采样都得动点脑筋。

MCMC采样是近似推断中最常用的采样方法，它结合了蒙特卡洛采样（接受-拒绝采样和重要性采样）和马尔科夫链两种方法。本篇笔记不介绍接受-拒绝采样和重要性采样，只介绍MCMC采样。

**3、MCMC**

MCMC采样，即马尔科夫链-蒙特卡洛采样。它的基本思想是构造一个马尔科夫链（其实就是构造其状态转移矩阵），使得其平稳分布等于我们的目标分布。然后从一个初始状态出发，使用拒绝采样的思路，沿着马尔科夫链进行状态转移，每次转移相当于一次采样。当马尔科夫链收敛到稳态以后，后面采集到的样本（即马尔科夫链达到的状态）就相当于在平稳分布（也就是目标分布）上采集的样本。

从上面的介绍中可以看出，马尔科夫链的状态转移矩阵的构造很重要，它决定了其平稳分布是否能逼近目标分布。不同的构造方法将产生不同的MCMC算法。

**（1）M-H采样**

M-H采样是最常见的MCMC算法。在介绍它之前，需要介绍一下马尔科夫链的细致平稳条件。设定义在马尔科夫链的状态转移矩阵P，以及马尔科夫链的状态空间S上的分布，若对S中的任意两个状态i和j，有，其中i和j是状态空间中的任意两个状态，则称分布满足细致平稳条件。

可以证明，若满足细致平稳条件，则是马尔科夫链的平稳分布。

M-H采样基于细致平稳条件，构造如下的马尔科夫链的转移矩阵：



其中和分别表示在状态i和状态j上的采样。实际上，（）和i（j）是等价的，这里使用前者来表示只不过是为了突出“采样”的意义。

上式表示均值为，方差为1的正态分布在处的概率值。

然后引入一个接受率因子：



其中，p是我们的目标分布，Q是刚才构造的状态转移矩阵。

可以证明：



这与细致平稳条件的形式非常类似，但并不是细致平稳条件，因此不能认为目标分布p就是Q对应的马尔科夫链的平稳分布。也就是说，以Q为状态转移矩阵的马尔科夫链，最终并不能收敛到目标分布p。

但是上式指明了一个基于Q的，且叠加的状态转移过程，该过程最终能收敛到p。这样，我们就间接地达到了我们的目的。这就形成了M-H采样算法的流程，其每一步的操作如下：

（a）从均匀分布中采一个数；

（b）依Q，根据上一轮采样值，确定本轮的状态分布，并从分布中采一个样本；

（c）计算接受率；

（d）如果，则我们把作为本轮的采样值，如果，则抛弃，而沿用上一轮的采样值作为本轮的采样值。

**（2）Gibbs采样**

吉布斯采样也属于MCMC采样方法，它是一种针对高维随机变量的分布的采样方法。

四、概率图模型

**1、概率图模型简介**

概率图模型，又称结构化概率模型，指的是用图结构来表示一组随机变量，图中的每个节点代表一个随机变量，节点之间的边代表随机变量之间的关系。

概率图模型主要有两类：一类是有向无环的概率图模型，称为贝叶斯网，它包括朴素贝叶斯、半朴素贝叶斯、马尔科夫链、隐马尔科夫模型等；另一类是无向的且具有马尔科夫性的概率图模型，称为马尔科夫网（也叫马尔科夫随机场）。

概率图模型的三个关键问题是建模、学习和推断。“建模”和“学习”比较容易理解，“推断”则是指在确定的概率图模型的基础上，用图中一部分随机变量的观测值，去预测另一部分随机变量的分布，本质上其实就是计算条件概率。在实际问题中，概率图通常包含一些“可观测变量”和一些“不可观测变量”，这时，推断任务一般是用可观测变量的观测值，去预测一些不可观测的随机变量的分布。

**2、概率图的分解**

概率图的分解是指将概率图中所有随机变量的联合分布分解成边缘分布和条件分布的乘积的形式。对有向图的分解和对无向图的分解的思路不同。

对与有向图的分解，由于其指明了节点之间的依赖关系，所以就根据条件概率公式，从节点之间的关系导出分解式。比如下图：

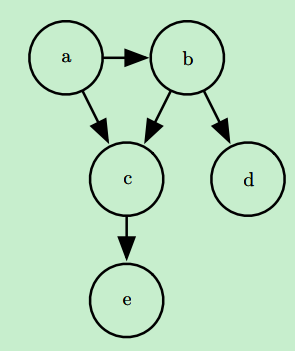


图1 一个有向概率图模型的例子

该图的分解式为：



对于无向图的分解，则使用完全不同的思路。

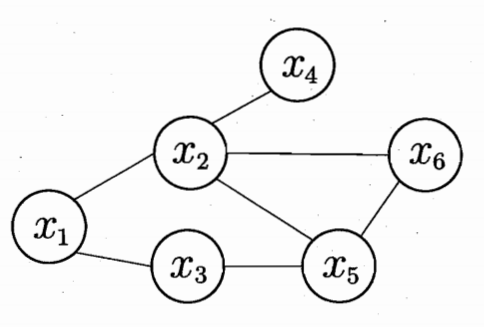


图2 一个无向概率图模型的例子

无向图基于“极大团”及定义于极大团上的“势函数”对联合分布进行分解。

无向概率图模型的“团”是指图的一个子图，该子图中任意两个节点之间都有边连接。而“极大团”是指这样的一个团：向它加入任何一个节点，都不再形成团。而定义在团上的“势函数”类似于该团包含的随机变量的联合分布，势函数是非负的，但不像真正的概率分布函数那样，是归一化的。

设无向概率图模型包含的所有随机变量为：，则联合概率可以分解为：



其中C是概率图中的所有极大团的集合，Q是其中的一个极大团，是极大团中的所有随机变量构成的随机向量，Z是归一化因子，以保证是一个概率。实际应用中，精确计算Z通常很困难，但是许多任务往往并不需要获得Z的精确值。

**1、隐马尔可夫模型（HMM）**

隐马尔科夫模型是结构最简单的动态贝叶斯网。其结构如下：

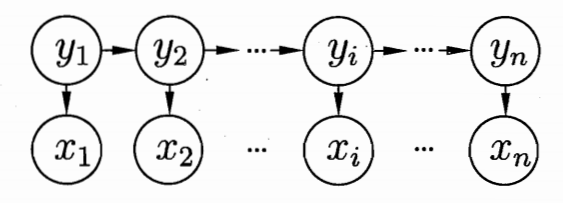


图3 隐马尔科夫模型的结构

如图3，上面是状态变量序列，构成马尔科夫链；下面是观测变量。每个观测变量的分布仅由相应的状态变量的取值决定。

隐马尔科夫模型的参数为：，其中A是状态转移概率矩阵；B是输出观测概率矩阵；是初始状态概率分布。

隐马尔科夫模型的三个重要问题为：

（1）观测值序列的概率计算问题：给定模型，计算其产生一个观测序列的概率（注意，此处x表示观测值序列，而不是随机变量序列）；

（2）状态序列的预测问题：给定模型和观测值序列，推断出状态序列；

（3）模型的学习问题：给出观测值序列，求解最优的模型，使得该模型产生观测序列观测值的概率最大。

其中观测值序列的概率计算问题通常使用前向-后向算法；状态序列的预测问题通常使用维特比算法（是一种动态规划算法）；模型的学习问题通常使用Baum-Welch算法。

**2、条件随机场（CRF）**

条件随机场是判别模型，它对条件概率分布进行建模。它是给定随机向量X的条件下，随机向量Y的马尔科夫随机场。在具体的问题中，X是观测序列，是模型的输入，Y是标记序列（也成为状态序列），是模型的输出。注意Y是一个马尔科夫随机场。

我们最常用到的条件随机场是线性链条件随机场。它的结构如下：

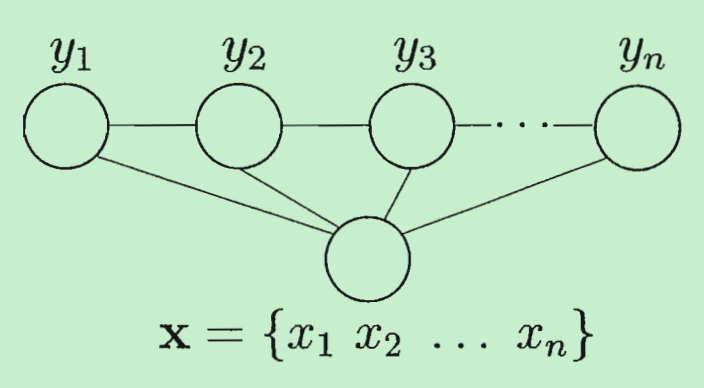


图4 线性链条件随机场结构

线性链条件随机场用于标记问题。