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，估计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；对参数的估计和随机变量的预测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

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。

我们前面介绍了参数估计的几种方法：最大似然法、最大后验法、贝叶斯估计法。

它们的目标其实都是求出后验分布，其中D为样本的观测值集合。

在简单的情况下，

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。你可能觉得只要给定一个分布的形式，那么从该分布中采样就是一个很简单的事，实际上这是不对的。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。

四、概率图模型

**1、隐马尔可夫模型（HMM）**

**2、条件随机场（CRF）**