上篇主要介绍了半监督学习，首先从如何利用未标记样本所蕴含的分布信息出发，引入了半监督学习的基本概念，即训练数据同时包含有标记样本和未标记样本的学习方法；接着分别介绍了几种常见的半监督学习方法：生成式方法基于对数据分布的假设，利用未标记样本隐含的分布信息，使得对模型参数的估计更加准确；TSVM给未标记样本赋予伪标记，并通过不断调整易出错样本的标记得到最终输出；基于分歧的方法结合了集成学习的思想，通过多个学习器在不同视图上的协作，有效利用了未标记样本数据 ；最后半监督聚类则是借助已有的监督信息来辅助聚类的过程，带约束k-均值算法需检测当前样本划分是否满足约束关系，带标记k-均值算法则利用有标记样本指定初始类中心。本篇将讨论一种基于图的学习算法--概率图模型。

#**15、概率图模型**

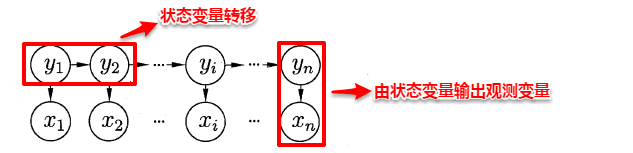
现在再来谈谈机器学习的核心价值观，可以更通俗地理解为：**根据一些已观察到的证据来推断未知**，更具哲学性地可以阐述为：未来的发展总是遵循着历史的规律。其中**基于概率的模型将学习任务归结为计算变量的概率分布**，正如之前已经提到的：生成式模型先对联合分布进行建模，从而再来求解后验概率，例如：贝叶斯分类器先对联合分布进行最大似然估计，从而便可以计算类条件概率；判别式模型则是直接对条件分布进行建模。

**概率图模型**（probabilistic graphical model）是一类用**图结构**来表达各属性之间相关关系的概率模型，一般而言：**图中的一个结点表示一个或一组随机变量，结点之间的边则表示变量间的相关关系**，从而形成了一张“**变量关系图**”。若使用有向的边来表达变量之间的依赖关系，这样的有向关系图称为**贝叶斯网**（Bayesian nerwork）或有向图模型；若使用无向边，则称为**马尔可夫网**（Markov network）或无向图模型。

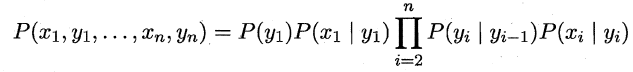
##**15.1 隐马尔可夫模型(HMM)**

隐马尔可夫模型（Hidden Markov Model，简称HMM）是结构最简单的一种贝叶斯网，在语音识别与自然语言处理领域上有着广泛的应用。HMM中的变量分为两组：**状态变量**与**观测变量**，其中状态变量一般是未知的，因此又称为“**隐变量**”，观测变量则是已知的输出值。在隐马尔可夫模型中，变量之间的依赖关系遵循如下两个规则：

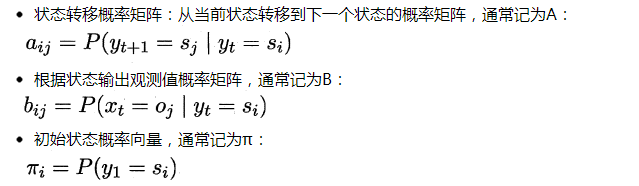
**1. 观测变量的取值仅依赖于状态变量**； **2. 下一个状态的取值仅依赖于当前状态**，通俗来讲：**现在决定未来，未来与过去无关**，这就是著名的**马尔可夫性**。

[](https://camo.githubusercontent.com/a3e7eb2eab3eb9102db9895cdc90cf9180587991/68747470733a2f2f73312e617831782e636f6d2f323031382f31302f31382f697759506d522e706e67)

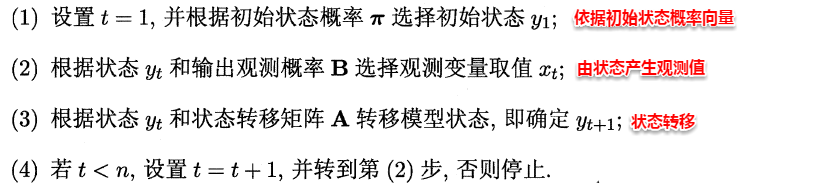
基于上述变量之间的依赖关系，我们很容易写出隐马尔可夫模型中所有变量的联合概率分布：

[](https://camo.githubusercontent.com/80898d76e4112d4bcce06c3ef2d6f68eb508a319/68747470733a2f2f73312e617831782e636f6d2f323031382f31302f31382f6977593958392e706e67)

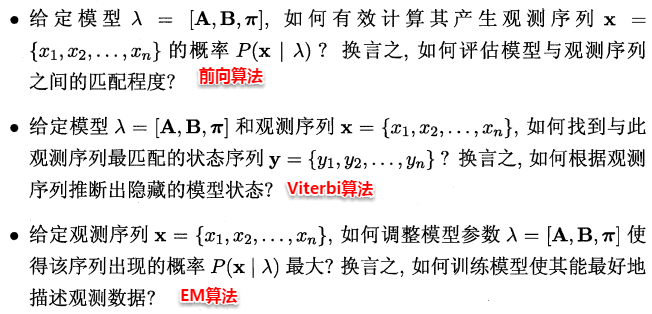
易知：**欲确定一个HMM模型需要以下三组参数**：

[](https://camo.githubusercontent.com/a7e913f5e3be41e4424d52d0e4423bf33cecc45b/68747470733a2f2f73312e617831782e636f6d2f323031382f31302f31382f6977596930312e706e67)

当确定了一个HMM模型的三个参数后，便按照下面的规则来生成观测值序列：

[](https://camo.githubusercontent.com/9540f51fa23a778ca0f02d3ac2d764711d91dfce/68747470733a2f2f73312e617831782e636f6d2f323031382f31302f31382f6977594654782e706e67)

在实际应用中，HMM模型的发力点主要体现在下述三个问题上：

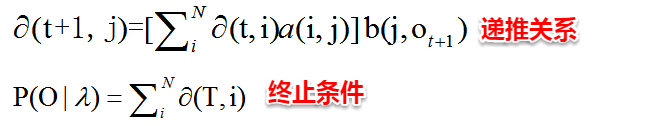
[](https://camo.githubusercontent.com/609d23bae1cee8f2a2d145a40e3438fd7bc297ed/68747470733a2f2f73312e617831782e636f6d2f323031382f31302f31382f69775945744b2e706e67)

###**15.1.1 HMM评估问题**

HMM评估问题指的是：**给定了模型的三个参数与观测值序列，求该观测值序列出现的概率**。例如：对于赌场问题，便可以依据骰子掷出的结果序列来计算该结果序列出现的可能性，若小概率的事件发生了则可认为赌场的骰子有作弊的可能。解决该问题使用的是**前向算法**，即步步为营，自底向上的方式逐步增加序列的长度，直到获得目标概率值。在前向算法中，定义了一个**前向变量**，即给定观察值序列且t时刻的状态为Si的概率：

[iwYVfO.png](https://camo.githubusercontent.com/5db94749651b762c89e7392b2c206c009a3c1683/68747470733a2f2f73312e617831782e636f6d2f323031382f31302f31382f69775956664f2e706e67)

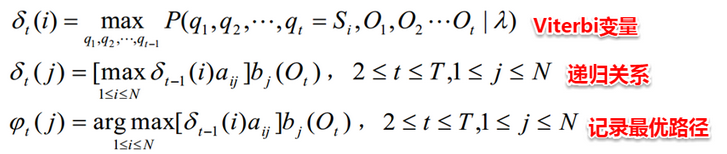
基于前向变量，很容易得到该问题的递推关系及终止条件：

[](https://camo.githubusercontent.com/8c5b66f4875d976ee1dcc14126edfc31c8787825/68747470733a2f2f73312e617831782e636f6d2f323031382f31302f31382f697759416b362e706e67)

因此可使用动态规划法，从最小的子问题开始，通过填表格的形式一步一步计算出目标结果。

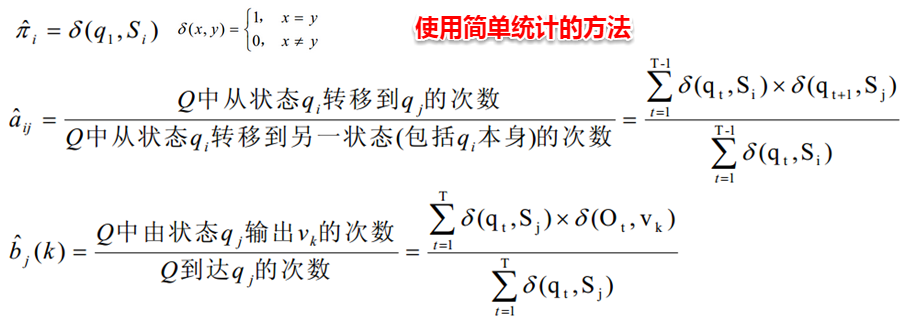
###**15.1.2 HMM解码问题**

HMM解码问题指的是：**给定了模型的三个参数与观测值序列，求可能性最大的状态序列**。例如：在语音识别问题中，人说话形成的数字信号对应着观测值序列，对应的具体文字则是状态序列，从数字信号转化为文字正是对应着根据观测值序列推断最有可能的状态值序列。解决该问题使用的是**Viterbi算法**，与前向算法十分类似地，Viterbi算法定义了一个**Viterbi变量**，也是采用动态规划的方法，自底向上逐步求解。

[](https://camo.githubusercontent.com/daba036736bf9a231291f7202f9d537ccd21372f/68747470733a2f2f73312e617831782e636f6d2f323031382f31302f31382f6977596570442e706e67)

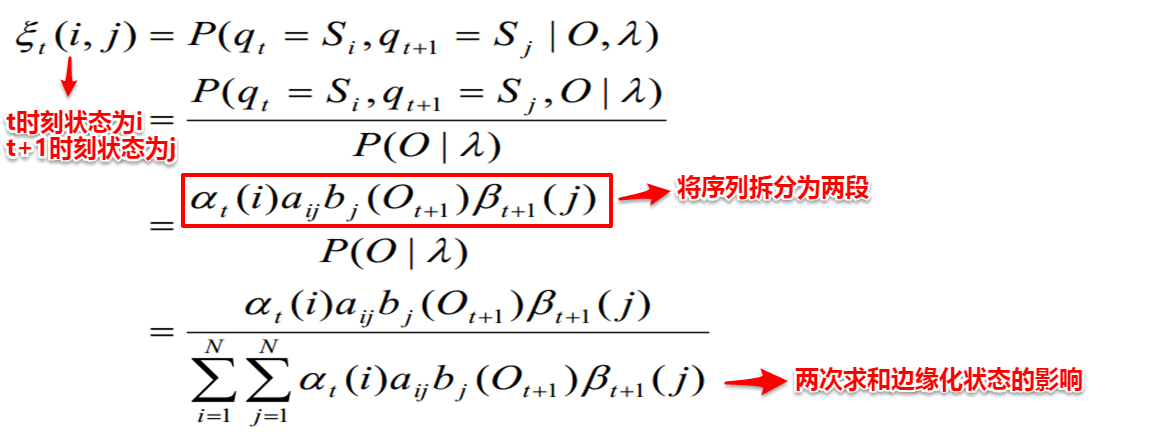
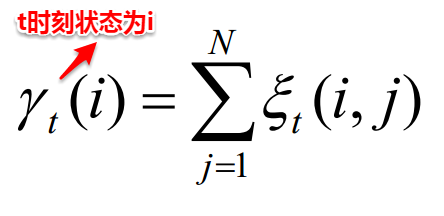
###**15.1.3 HMM学习问题**

HMM学习问题指的是：**给定观测值序列，如何调整模型的参数使得该序列出现的概率最大**。这便转化成了机器学习问题，即从给定的观测值序列中学习出一个HMM模型，**该问题正是EM算法的经典案例之一**。其思想也十分简单：对于给定的观测值序列，如果我们能够按照该序列潜在的规律来调整模型的三个参数，则可以使得该序列出现的可能性最大。假设状态值序列也已知，则很容易计算出与该序列最契合的模型参数：

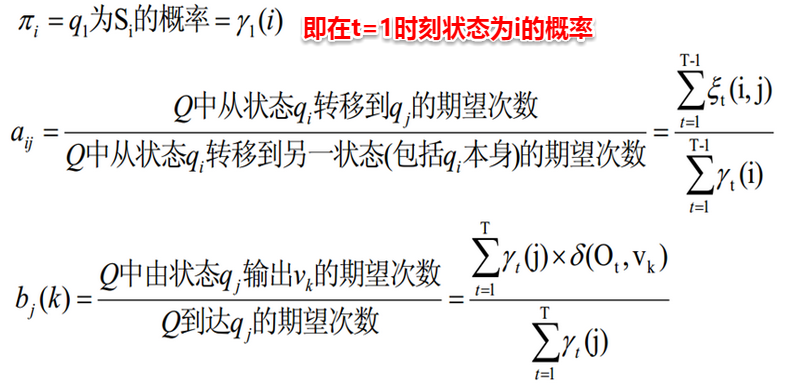
[](https://camo.githubusercontent.com/b9879195ad2ecd929bbe45e84f2951ef59348403/68747470733a2f2f73312e617831782e636f6d2f323031382f31302f31382f6977596d31652e706e67)

但一般状态值序列都是不可观测的，且**即使给定观测值序列与模型参数，状态序列仍然遭遇组合爆炸**。因此上面这种简单的统计方法就行不通了，若将状态值序列看作为隐变量，这时便可以考虑使用EM算法来对该问题进行求解：

【1】首先对HMM模型的三个参数进行随机初始化； 【2】根据模型的参数与观测值序列，计算t时刻状态为i且t+1时刻状态为j的概率以及t时刻状态为i的概率。

[](https://camo.githubusercontent.com/e5cb23970fb8c88a1d6b3e7ebf3c6a82f93672aa/68747470733a2f2f73312e617831782e636f6d2f323031382f31302f31382f6977596e36482e706e67)[](https://camo.githubusercontent.com/e4070afaf9b6cd373866425cc25976b2c9da1c15/68747470733a2f2f73312e617831782e636f6d2f323031382f31302f31382f697759646e732e706e67)

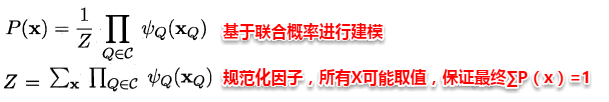
【3】接着便可以对模型的三个参数进行重新估计：

[](https://camo.githubusercontent.com/9ee8c870bf78502b238e54d27e2e551019996ead/68747470733a2f2f73312e617831782e636f6d2f323031382f31302f31382f6977595939532e706e67)

【4】重复步骤2-3，直至三个参数值收敛，便得到了最终的HMM模型。

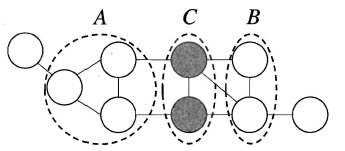
##**15.2 马尔可夫随机场（MRF）**

马尔可夫随机场（Markov Random Field）是一种典型的马尔可夫网，即使用无向边来表达变量间的依赖关系。在马尔可夫随机场中，对于关系图中的一个子集，**若任意两结点间都有边连接，则称该子集为一个团；若再加一个结点便不能形成团，则称该子集为极大团**。MRF使用**势函数**来定义多个变量的概率分布函数，其中**每个（极大）团对应一个势函数**，一般团中的变量关系也体现在它所对应的极大团中，因此常常基于极大团来定义变量的联合概率分布函数。具体而言，若所有变量构成的极大团的集合为C，则MRF的联合概率函数可以定义为：

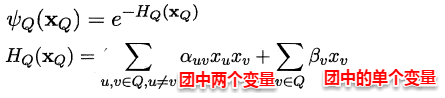
[](https://camo.githubusercontent.com/4bd466858399445c22581eac7f753466c460cee6/68747470733a2f2f73312e617831782e636f6d2f323031382f31302f31382f6977594768382e706e67)

对于条件独立性，**马尔可夫随机场通过分离集来实现条件独立**，若A结点集必须经过C结点集才能到达B结点集，则称C为分离集。书上给出了一个简单情形下的条件独立证明过程，十分贴切易懂，此处不再展开。基于分离集的概念，得到了MRF的三个性质：

**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。 **局部马尔可夫性**：给定某变量的邻接变量，则该变量与其它变量条件独立。 **成对马尔可夫性**：给定所有其他变量，两个非邻接变量条件独立。

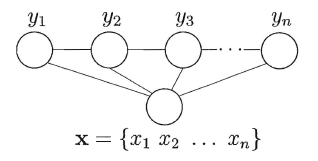
[](https://camo.githubusercontent.com/ded99aefd2cc697359c4937d419bc42f16a254ad/68747470733a2f2f73312e617831782e636f6d2f323031382f31302f31382f6977593037712e706e67)

对于MRF中的势函数，势函数主要用于描述团中变量之间的相关关系，且要求为非负函数，直观来看：势函数需要在偏好的变量取值上函数值较大，例如：若x1与x2成正相关，则需要将这种关系反映在势函数的函数值中。一般我们常使用指数函数来定义势函数：

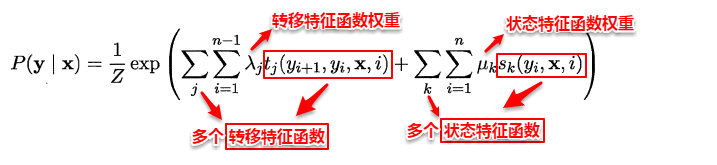
[](https://camo.githubusercontent.com/13e66306b762be738a38c87af005f0d9bea83342/68747470733a2f2f73312e617831782e636f6d2f323031382f31302f31382f6977593874662e706e67)

##**15.3 条件随机场（CRF）**

前面所讲到的**隐马尔可夫模型和马尔可夫随机场都属于生成式模型，即对联合概率进行建模，条件随机场则是对条件分布进行建模**。CRF试图在给定观测值序列后，对状态序列的概率分布进行建模，即P(y | x)。直观上看：CRF与HMM的解码问题十分类似，都是在给定观测值序列后，研究状态序列可能的取值。CRF可以有多种结构，只需保证状态序列满足马尔可夫性即可，一般我们常使用的是**链式条件随机场**：

[](https://camo.githubusercontent.com/759c3ab1fc1cdc8b086031767e84b8f6011216bc/68747470733a2f2f73312e617831782e636f6d2f323031382f31302f31382f6977597431672e706e67)

与马尔可夫随机场定义联合概率类似地，CRF也通过团以及势函数的概念来定义条件概率P(y | x)。在给定观测值序列的条件下，链式条件随机场主要包含两种团结构：单个状态团及相邻状态团，通过引入两类特征函数便可以定义出目标条件概率：

[](https://camo.githubusercontent.com/e4534d39aa384f9444772d682e190898357dcea5/68747470733a2f2f73312e617831782e636f6d2f323031382f31302f31382f6977594e63512e706e67)

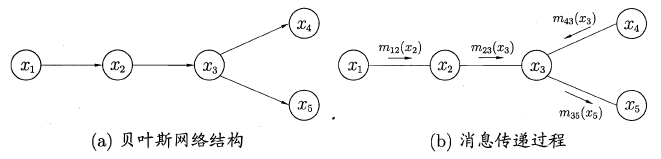
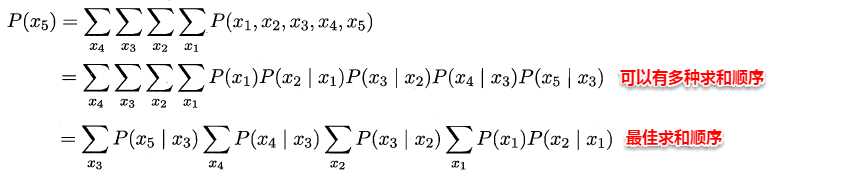
以词性标注为例，如何判断给出的一个标注序列靠谱不靠谱呢？**转移特征函数主要判定两个相邻的标注是否合理**，例如：动词+动词显然语法不通；**状态特征函数则判定观测值与对应的标注是否合理**，例如： ly结尾的词-->副词较合理。因此我们可以定义一个特征函数集合，用这个特征函数集合来为一个标注序列打分，并据此选出最靠谱的标注序列。也就是说，每一个特征函数（对应一种规则）都可以用来为一个标注序列评分，把集合中所有特征函数对同一个标注序列的评分综合起来，就是这个标注序列最终的评分值。可以看出：**特征函数是一些经验的特性**。

##**15.4 学习与推断**

对于生成式模型，通常我们都是先对变量的联合概率分布进行建模，接着再求出目标变量的**边际分布**（marginal distribution），那如何从联合概率得到边际分布呢？这便是学习与推断。下面主要介绍两种精确推断的方法：**变量消去**与**信念传播**。

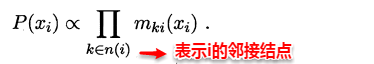
###**15.4.1 变量消去**

变量消去利用条件独立性来消减计算目标概率值所需的计算量，它通过运用**乘法与加法的分配率**，将对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，从而将每次的**运算控制在局部**，达到简化运算的目的。

[](https://camo.githubusercontent.com/4ef9cea95f3fc878e1f11351ae07ffbf09c50178/68747470733a2f2f73312e617831782e636f6d2f323031382f31302f31382f69775955586a2e706e67) [](https://camo.githubusercontent.com/e4fd2c9c29dc8fff76ed17f0bed280f946cb09e4/68747470733a2f2f73312e617831782e636f6d2f323031382f31302f31382f69775977426e2e706e67)

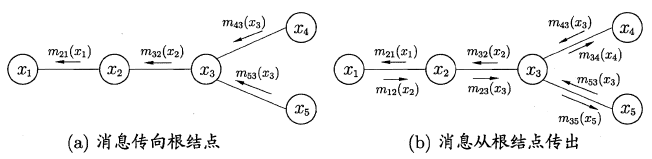
###**15.4.2 信念传播**

若将变量求和操作看作是一种消息的传递过程，信念传播可以理解成：**一个节点在接收到所有其它节点的消息后才向另一个节点发送消息**，同时当前节点的边际概率正比于他所接收的消息的乘积：

[](https://camo.githubusercontent.com/704b346ab79dcc580fb7e4c7eddce6bbcc13fbc8/68747470733a2f2f73312e617831782e636f6d2f323031382f31302f31382f6977594441302e706e67)

因此只需要经过下面两个步骤，便可以完成所有的消息传递过程。利用动态规划法的思想记录传递过程中的所有消息，当计算某个结点的边际概率分布时，只需直接取出传到该结点的消息即可，从而避免了计算多个边际分布时的冗余计算问题。

1.指定一个根节点，从所有的叶节点开始向根节点传递消息，直到根节点收到所有邻接结点的消息\*\*（从叶到根）**； 2.从根节点开始向叶节点传递消息，直到所有叶节点均收到消息**（从根到叶）\*\*。

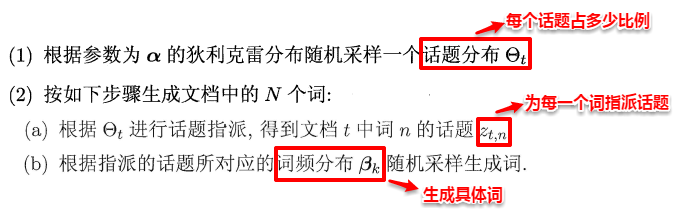
[](https://camo.githubusercontent.com/3ec13d6f538d8244f6f4aad6bdd7f038ec3b231d/68747470733a2f2f73312e617831782e636f6d2f323031382f31302f31382f6977596763342e706e67)

##**15.5 LDA话题模型**

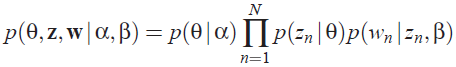
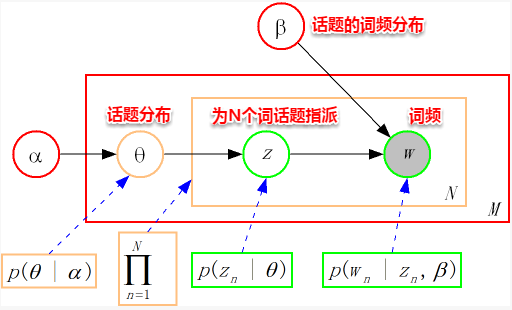
话题模型主要用于处理文本类数据，其中**隐狄利克雷分配模型**（Latent Dirichlet Allocation，简称LDA）是话题模型的杰出代表。在话题模型中，有以下几个基本概念：词（word）、文档（document）、话题（topic）。

**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：

[](https://camo.githubusercontent.com/9975a2f6adeb4597c969e27eec65683b8223b33d/68747470733a2f2f73312e617831782e636f6d2f323031382f31302f31382f697759326a4a2e706e67)

这样一个文档中的所有词都可以认为是通过话题模型来生成的，当已知一个文档的词频分布后（即一个N维向量，N为词库大小），则可以认为：**每一个词频元素都对应着一个话题，而话题对应的词频分布则影响着该词频元素的大小**。因此很容易写出LDA模型对应的联合概率函数：

[](https://camo.githubusercontent.com/5c321a2f6c3536b612129aa0845b8ca1bd22cfda/68747470733a2f2f73312e617831782e636f6d2f323031382f31302f31382f6977596333462e706e67) [](https://camo.githubusercontent.com/02f0c24927fc4711e0ff1212b08e29195e428a9b/68747470733a2f2f73312e617831782e636f6d2f323031382f31302f31382f6977595775392e706e67)

从上图可以看出，LDA的三个表示层被三种颜色表示出来：

**corpus-level（红色）：** α和β表示语料级别的参数，也就是每个文档都一样，因此生成过程只采样一次。 **document-level（橙色）：** θ是文档级别的变量，每个文档对应一个θ。 **word-level（绿色）：** z和w都是单词级别变量，z由θ生成，w由z和β共同生成，一个单词w对应一个主题z。

通过上面对LDA生成模型的讨论，可以知道**LDA模型主要是想从给定的输入语料中学习训练出两个控制参数α和β**，当学习出了这两个控制参数就确定了模型，便可以用来生成文档。其中α和β分别对应以下各个信息：

**α**：分布p(θ)需要一个向量参数，即Dirichlet分布的参数，用于生成一个主题θ向量； **β**：各个主题对应的单词概率分布矩阵p(w|z)。

把w当做观察变量，θ和z当做隐藏变量，就可以通过EM算法学习出α和β，求解过程中遇到后验概率p(θ,z|w)无法直接求解，需要找一个似然函数下界来近似求解，原作者使用基于分解（factorization）假设的变分法（varialtional inference）进行计算，用到了EM算法。每次E-step输入α和β，计算似然函数，M-step最大化这个似然函数，算出α和β，不断迭代直到收敛。

在此，概率图模型就介绍完毕。上周受到协同训练的启发，让实验的小伙伴做了一个HMM的slides，结果扩充了好多知识，所以完成这篇笔记还是花费了不少功夫，还刚好赶上实验室没空调回到解放前的日子，可谓汗流之作...