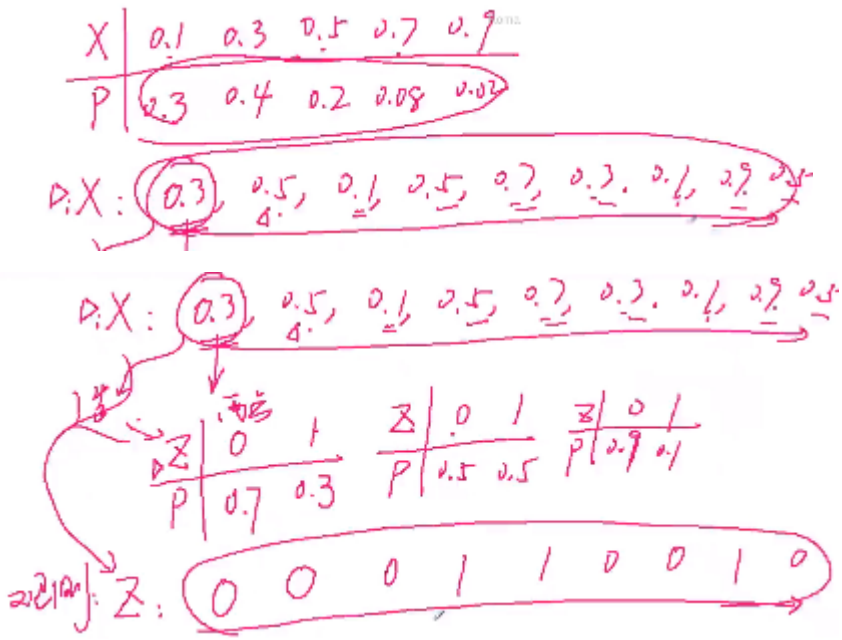
LDA

LDA可以看成是三层的贝叶斯网络。无监督学习／蒋维



如上图所示：

正向：随机变量X，其分布为P。　根据这一分布产生值DiX，然后根据每一个DiX构造变量Z，其分布由Dix决定，然后根据其分布得到一系列的Z值，即000110010.

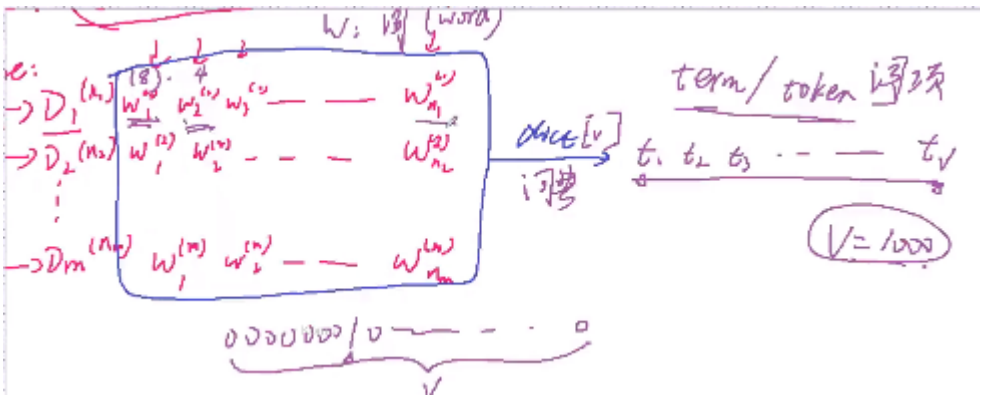
逆向：已经得到了Z的观测值000110010，然后想根据这些观测值推出X的分布P(或者DiX)。

LDA的应用方向（处理隐变量）

- 信息提取和搜索
    - 语义分析
- 文档分类／聚类、文章摘要、社区挖掘
- 基于内容的图像聚类、目标识别
    - 以及其它计算机视觉应用
- 生物信息数据的应用

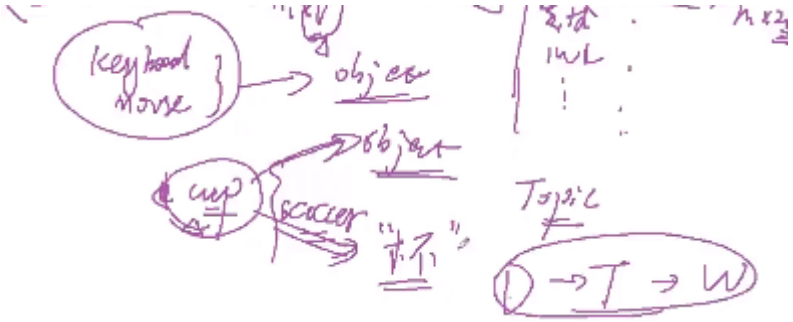
数据表示

一共有m片文档D，第m篇文档中有n_m个单词。　词袋长度为v。



每篇文档的特征都是$n_m * v$维的。要做的就是将左边（词）蒋维到右边（主题）。左边“水杯”这个词对应到右边的爱情的概率为0.2,对应到物体的概率为0.8.

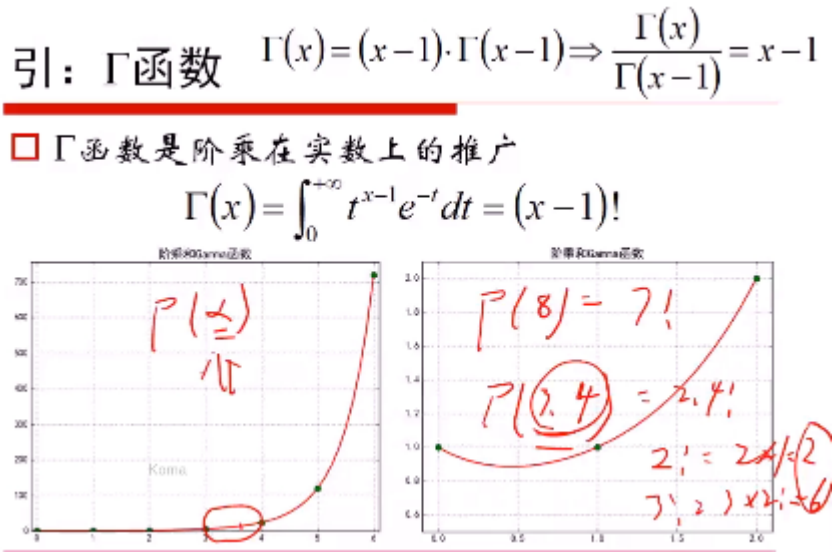
可以将同类的词归为同一类主题；也可以将同一个词归为不同主题（语义不同）。这样就将文档D对应到主题T这个隐变量上。



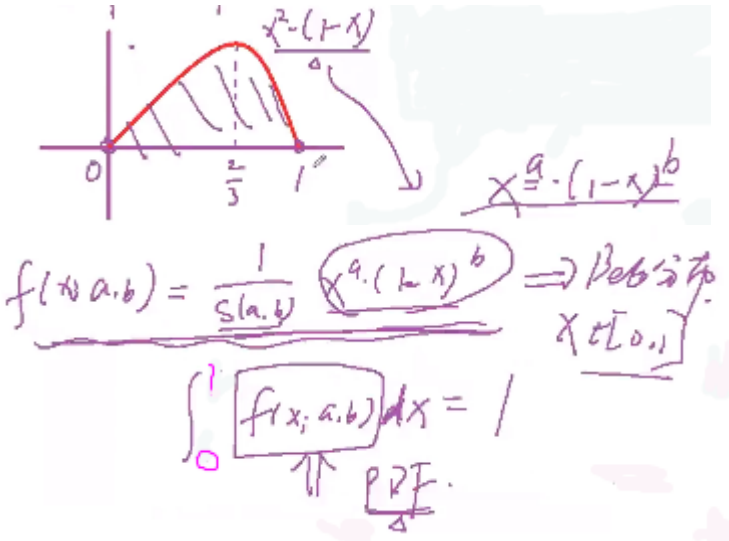
推导

Beta分布

该函数是阶乘在实数上的推广。只需要知道其结果是一个数就可以了。



对于形如$x^a*(1-x)^b$这样一个函数，其总过横轴0,1这两点。参数a,b决定了该函数的峰值往哪偏。该函数跟 x 轴的面积即为S(a,b).a,b为参数。f(x;a,b)即为Beta分布的概率密度函数。



故得到(a记为a-1,b记为b-1)：

- Beta分布的概率密度：

$$f(x)=\begin{cases}\dfrac{1}{B(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}, & x\in[0,1]\\ 0, & \text{其他}\end{cases}$$

- 其中系数B为：

$$B(\alpha,\beta)=\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

- Gamma函数可以看成阶乘的实数域推广：

$$\Gamma(x)=\int_0^{\infty}t^{x-1}e^{-t}dt$$

$$\Rightarrow\Gamma(n)=(n-1)!\Rightarrow B(\alpha,\beta)=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

Koma

$$f(x)=\frac{1}{B(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1},\ x\in[0,1]$$

Beta分布的期望

$$B(\alpha,\beta)=\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

- 根据定义：

$$E(X)=\int_0^1 x\cdot\frac{1}{B(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}dx$$

$$=\frac{1}{B(\alpha,\beta)}\int_0^1 x^{(\alpha+1)-1}(1-x)^{\beta-1}dx$$

$$=\frac{B(\alpha+1,\beta)}{B(\alpha,\beta)}=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\Big/\frac{\Gamma(\alpha+\beta+1)}{\Gamma(\alpha+1)\Gamma(\beta)}$$

$$=\frac{\alpha}{\alpha+\beta}$$

Beta分布的作用

Beta分布可以看作一个概率的概率分布，当我们不知道一个东西的具体概率是多少时，Beta分布可以给出所有概率出现的可能性大小。根据已知的知识，我们可以利用Beta分布将一个东西的具体概率限制在一定范围内。

举例说明：

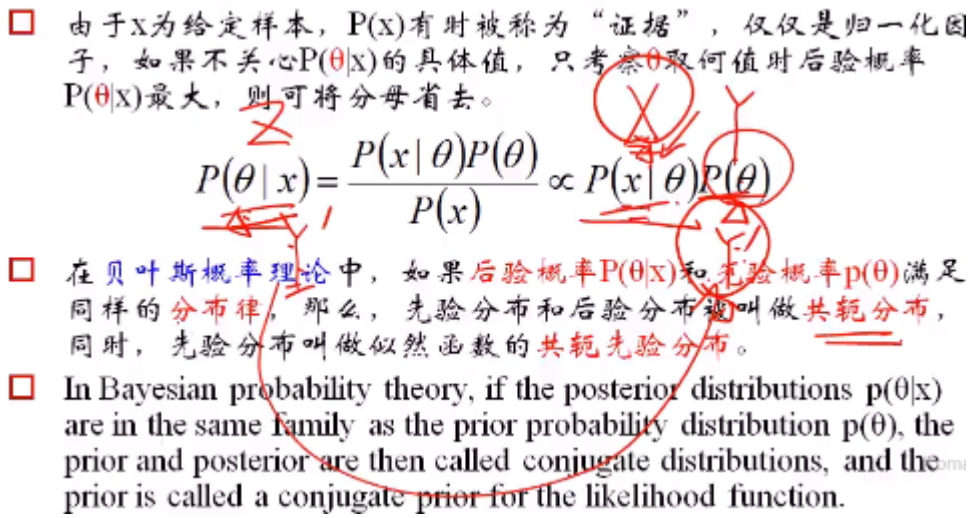
熟悉棒球运动的都知道有一个指标就是棒球击球率(batting average)，就是用一个运动员击中的球数除以击球的总数，我们一般认为0.266是正常水平的击球率，而如果击球率高达0.3就被认为是非常优秀的。现在有一个棒球运动员，我们希望能够预测他在这一赛季中的棒球击球率是多少。你可能就会直接计算棒球击球率，用击中的数除以击球数，但是如果这个棒球运动员只打了一次，而且还命中了，那么他就击球率就是100%了，这显然是不合理的，因为根据棒球的历史信息，我们知道这个击球率应该是0.215到0.36之间才对啊。对于这个问题，我们可以用一个二项分布表示（一系列成功或失败），一个最好的方法来表示这些经验（在统计中称为先验信息）就是用beta分布，这表示在我们没有看到这个运动员打球之前，我们就有了一个大概的范围。beta分布的定义域是(0,1)这就跟概率的范围是一样的。接下来我们将这些先验信息转换为beta分布的参数，我们知道一个击球率应该是平均0.27左右，而他的范围是0.21到0.35，那么根据这个信息，我们可以取α=81,β=219。之所以取这两个参数是因为beta分布的均值是$\frac{\alpha}{\alpha+\beta}=\frac{81}{81+219}=0.27$.即Beta(81,219).所以，击中一次之后为Beta(81+1,219).而击中100,未击中200，此时为Beta(81+100, 219+200),可以得到此时的数学期望为:$\frac{\alpha}{\alpha+\beta}=\frac{82+100}{82+100+219+200}=.303$。我们事实上就是在这个运动员在击球之前可以理解为他已经成功了81次，失败了219次这样一个先验信息。因此，对于一个我们不知道概率是什么，而又有一些合理的猜测时，beta分布能很好的作为一个表示概率的概率分布。

共轭先验分布

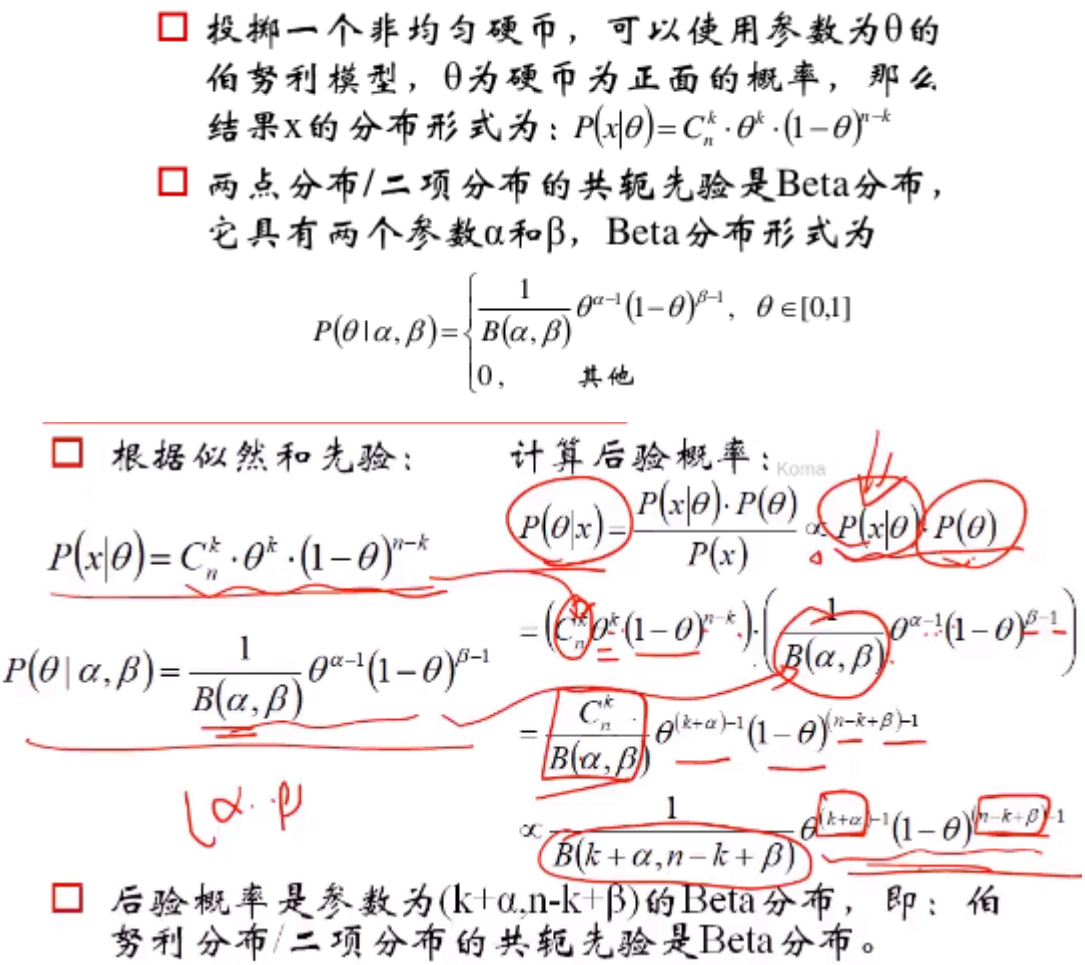
- 先验概率$P(\theta)$:未给定任何样本时，参数θ的分布

- 后验概率$P(\theta \mid x)$:给定样本x时，参数θ的分布
- 似然概率$P(x|\theta)$:有了参数θ时，计算的x的概率

实际的样本x的分布可以根据实际的场景来选择分布进行建模，例如高斯分布(x为身高，房价)、六点分布（筛子）、泊松分布（数个数）、周期（指数分布）。即先确定了X的分布。后验概率(Z)＝似然概率(X)×先验概率(Y)，若得到的Z与Y有相同的分布(X,Y有一定的关系)，则为共轭分布。然后，使用得到的Z作为Y继续乘以X进行迭代，即可更新参数，使模型收敛。　即：先验概率根据样本得到后验概率。



二项分布/伯努利分布的共轭先验是Beta分布，证明如下：



举例：

复习：二项分布的最大似然估计

- 投硬币试验中，进行N次独立试验，n次朝上，N-n次朝下。
- 假定朝上的概率为p，使用对数似然函数作为目标函数：

$$f(n \mid p) = \log\left(p^n (1-p)^{N-n}\right) \xrightarrow{\Delta} h(p)$$

$$\frac{\partial h(p)}{\partial p} = \frac{n}{p} - \frac{N-n}{1-p} \xrightarrow{\Delta} 0 \Rightarrow p = \frac{n}{N}$$

- 在校门口统计一定时间段内出入的男女生数目分别为N_B和N_G，估算该校男女生比例。

$$\begin{cases} P_B = \dfrac{N_B}{N_B + N_G} \\ P_G = \dfrac{N_G}{N_B + N_G} \end{cases}$$

- 若观察到4个女生和1个男生，可以得出该校女生比例是80%吗？
- 修正公式：

$$\begin{cases} P_B = \dfrac{N_B + 5}{N_B + N_G + 10} \\ P_G = \dfrac{N_G + 5}{N_B + N_G + 10} \end{cases} \Rightarrow \begin{cases} P_B = \dfrac{1+5}{1+4+10} = 40\% \\ P_G = \dfrac{4+5}{1+4+10} = 60\% \end{cases}$$

Beta分布. Beta(5,5)

$$P(\theta) = \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1}$$

$$P(\theta \mid x) = \frac{1}{B(k+\alpha, n-k+\beta)} \theta^{(k+\alpha)-1} (1-\theta)^{(n-k+\beta)-1}$$

伪计数

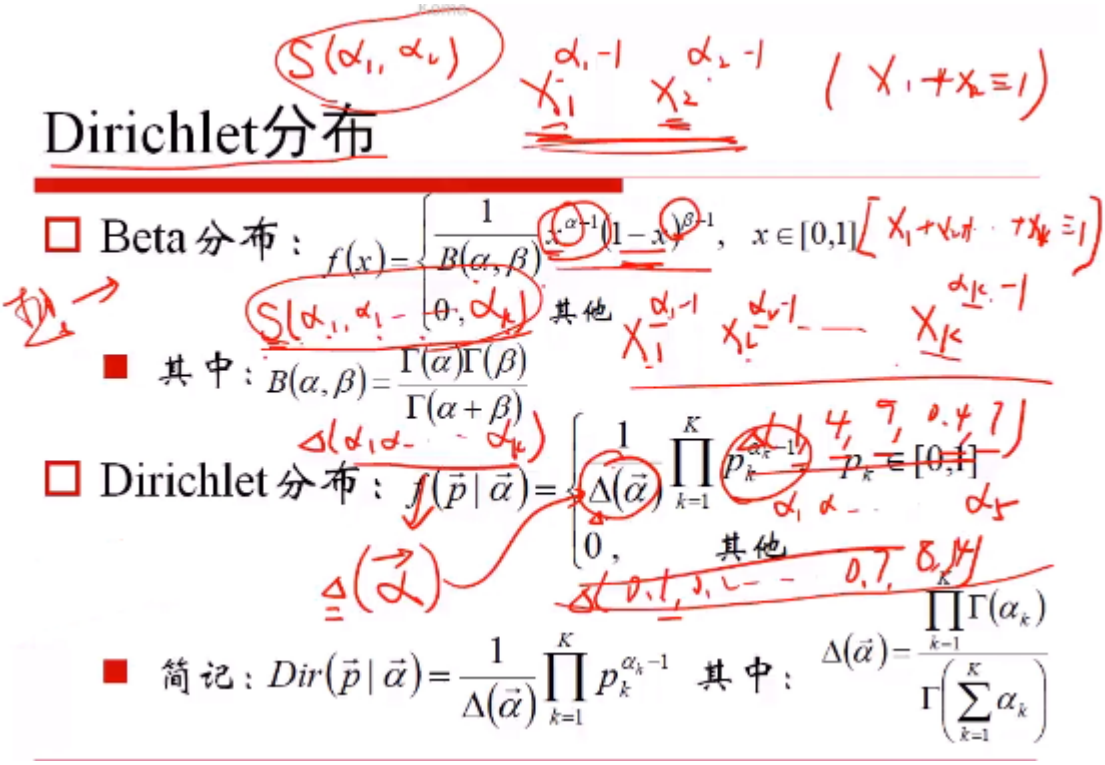
- 参数α、β是决定参数θ的参数，即超参数。
- 在后验概率的最终表达式中，参数α、β和k、n-k一起作为参数θ的指数——后验概率的参数为(k+α,n-k+β)。
- 根据这个指数的实践意义：投币过程中，正面朝上的次数，α和β先验性的给出了在没有任何实验的前提下，硬币朝上的概率分配；因此，α和β可被称作“伪计数”。

高斯分布的均值的共轭分布还是高斯分布。

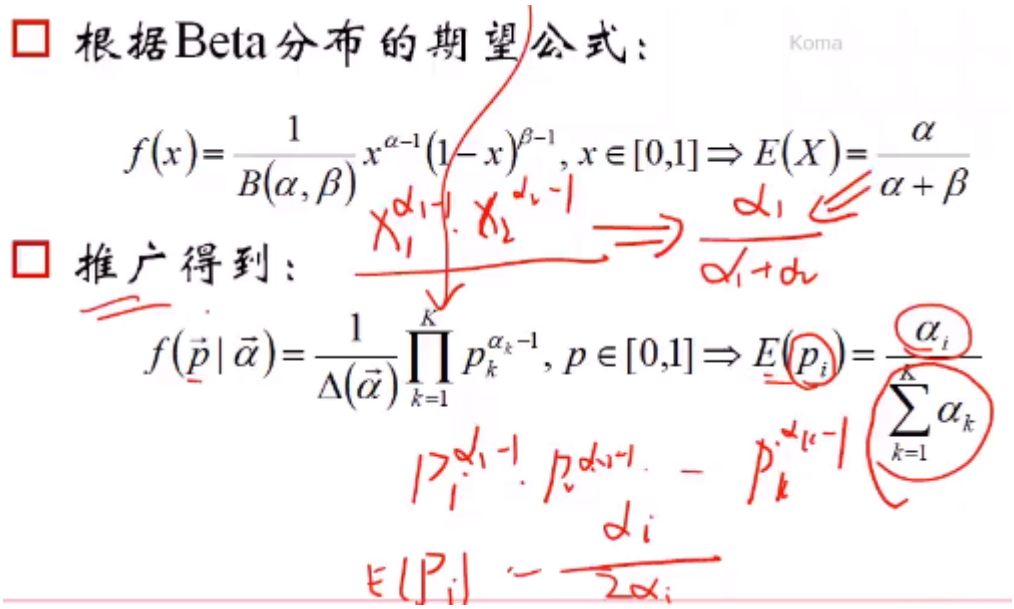
高斯分布的方差的共轭分布是伽马分布。

共轭先验的直接推广，从二元到多元：

- 二项分布 -> 多项分布
- Beta分布 -> Dirichlet分布

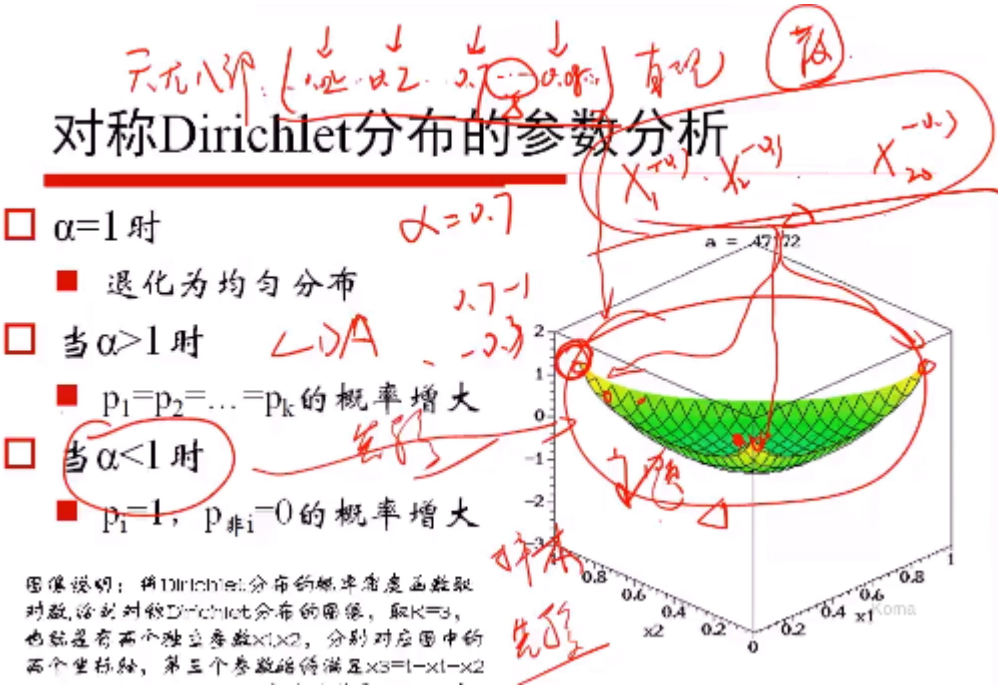


期望：



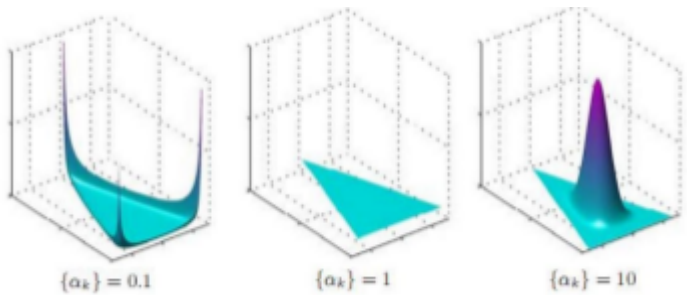
α_i参数相同时，为对称Dirivhlet分布。对称Dirichlet分布性质：

- $\alpha = 1$时，退化为均匀分布
- $\alpha > 1$时，函数向上凸起，P1=p2=...pk的概率增大。
- $\alpha < 1$时，函数向下凹，Pi=1,P非i=0的概率增大。即取边上的角角概率最大。



注意调参：LDA，(0,1)为先验值，参数尽量不要太大，这样就可以更多地考虑样本的影响。如果参数过大，则考虑先验的影响交大。

参数α对Direchlet分布的影响：



α越小，文档之间的主题越鲜明（在每个轴上），主题之间的概率越不均等。

α越大，文档中所涉及的主题越不鲜明（所有主题都可能，即每个主题的概率都相等）。

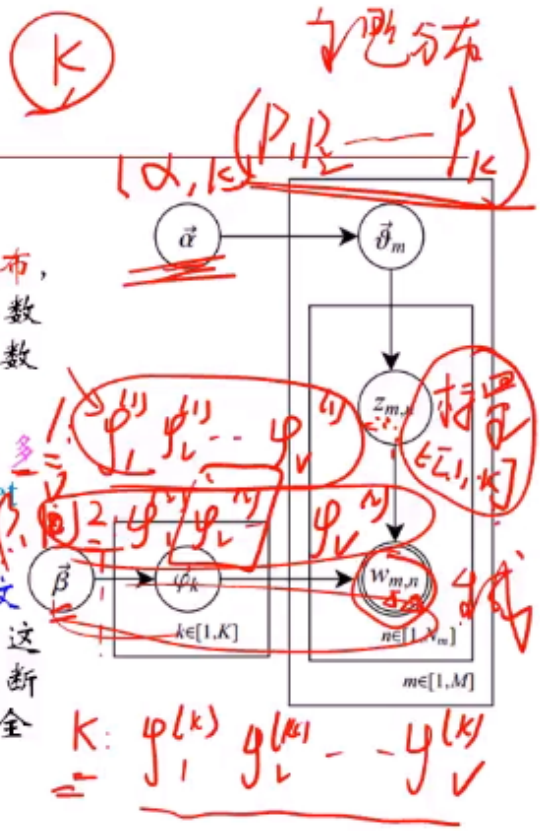
生成模型

即用该模型生成样本。

1. α参数的Dirichlet分布决定了第m篇文章的主题分布（多项分布θ_m）(P1P2...Pk)
2. 从主题分布(θ_m)中采样第n个词的主题$Z_{m,n}$
3. β参数的Dirichlet分布决定了每个主题的词分布φ_k。共k个主题，则有k个词分布(多项分布)。
4. 从主题$Z_{m,n}$对应的词分布φ_k中采样最终生成的词$W_{m,n}$。即第m篇文章的第n个词。

LDA的解释

- 共有m篇文章，一共涉及了K个主题；
- 每篇文章(长度为N_m)都有各自的主题分布，主题分布是多项分布，该多项分布的参数服从Dirichlet分布，该Dirichlet分布的参数为α；
- 每个主题都有各自的词分布，词分布为多项分布，该多项分布的参数服从Dirichlet分布，该Dirichlet分布的参数为β；
- 对于某篇文章中的第n个词，首先从该文章的主题分布中采样一个主题，然后在这个主题对应的词分布中采样一个词。不断重复这个随机生成过程，直到m篇文章全部完成上述过程。



认识事物的方式：

- 生成模型：`y -> model -> x` 例如：`LDA, NB, HMM`
- 判别模型：`x -> model -> y` 例如：`LR,DT,RF,SVM,CNN,CRF`

参数学习：

给定一个文档集合，$W_{m,n}$是可以观察到的已知变量，α和β是根据经验给定的先验参数，其它的变量$z_{m,n}, \theta, \phi$都是未知的隐变量，需要根据观察到的变量来学习估计。所有变量的联合分布：

$$p(\vec{w}_m, \vec{z}_m, \vec{\vartheta}_m, \underline{\Phi}|\vec{\alpha}, \vec{\beta}) = \prod_{n=1}^{N_m} p(w_{m,n}|\vec{\varphi}_{z_{m,n}}) p(z_{m,n}|\vec{\vartheta}_m) \cdot p(\vec{\vartheta}_m|\vec{\alpha}) \cdot p(\underline{\Phi}|\vec{\beta})$$

似然概率：一个词初始化为词t的概率为：

$$p(w_{m,n}=t|\vec{\vartheta}_m, \underline{\Phi}) = \sum_{k=1}^{K} p(w_{m,n}=t|\vec{\varphi}_k) p(z_{m,n}=k|\vec{\vartheta}_m)$$

词出现的概率，可以看成每个文档中出现主题K的概率乘以主题k下出现词t的概率，然后枚举所有主题求和得到。整个文档集合的似然函数为：

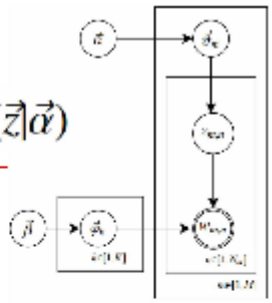
$$p(\mathcal{W}|\underline{\Theta}, \underline{\Phi}) = \prod_{m=1}^{M} p(\vec{w}_m|\vec{\vartheta}_m, \underline{\Phi}) = \prod_{m=1}^{M} \prod_{n=1}^{N_m} p(w_{m,n}|\vec{\vartheta}_m, \underline{\Phi})$$

联合分布：$p(\vec{w}, \vec{z}|\vec{\alpha}, \vec{\beta}) = p(\vec{w}|\vec{z}, \vec{\beta}) p(\vec{z}|\vec{\alpha})$

即为：　给定主题下采样词的过程　×　采样主题的过程

- n_z^t：表示主题z中出现词t的次数
- n_m^k：表示文档m出现主题k的次数

计算因子 $p(\vec{w},\vec{z}|\vec{\alpha},\vec{\beta}) = p(\vec{w}|\vec{z},\vec{\beta})p(\vec{z}|\vec{\alpha})$



$$p(\vec{w}|\vec{z},\vec{\beta}) = \int p(\vec{w}|\vec{z},\underline{\Phi})\ p(\underline{\Phi}|\vec{\beta})\ \mathrm{d}\underline{\Phi}$$

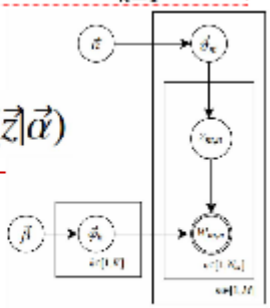
$$= \int \prod_{z=1}^{K} \frac{1}{\Delta(\vec{\beta})} \prod_{t=1}^{V} \varphi_{z,t}^{n_z^{(t)}+\beta_t-1}\ \mathrm{d}\vec{\varphi}_z$$

$$= \prod_{z=1}^{K} \frac{\Delta(\vec{n}_z+\vec{\beta})}{\Delta(\vec{\beta})}, \quad \vec{n}_z = \{n_z^{(t)}\}_{t=1}^{V}$$

$$\int_{\vec{p}} \prod_{k=1}^{K} p_k^{\alpha_k-1} d\vec{p} = \Delta(\vec{\alpha})$$

$$p(\vec{p}|\vec{\alpha}) = \mathrm{Dir}(\vec{p}|\vec{\alpha}) \triangleq \frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)} \prod_{k=1}^{K} p_k^{\alpha_k-1} \triangleq \frac{1}{\Delta(\vec{\alpha})} \prod_{k=1}^{K} p_k^{\alpha_k-1}$$

计算因子 $p(\vec{w},\vec{z}|\vec{\alpha},\vec{\beta}) = p(\vec{w}|\vec{z},\vec{\beta})p(\vec{z}|\vec{\alpha})$



$$p(\vec{z}|\vec{\alpha}) = \int p(\vec{z}|\underline{\Theta})\ p(\underline{\Theta}|\vec{\alpha})\ \mathrm{d}\underline{\Theta}$$

$$= \int \prod_{m=1}^{M} \frac{1}{\Delta(\vec{\alpha})} \prod_{k=1}^{K} \vartheta_{m,k}^{n_m^{(k)}+\alpha_k-1}\ \mathrm{d}\vec{\vartheta}_m$$

$$= \prod_{m=1}^{M} \frac{\Delta(\vec{n}_m+\vec{\alpha})}{\Delta(\vec{\alpha})}, \quad \vec{n}_m = \{n_m^{(k)}\}_{k=1}^{K}$$

$$\int_{\vec{p}} \prod_{k=1}^{K} p_k^{\alpha_k-1} d\vec{p} = \Delta(\vec{\alpha})$$

$$p(\vec{p}|\vec{\alpha}) = \mathrm{Dir}(\vec{p}|\vec{\alpha}) \triangleq \frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)} \prod_{k=1}^{K} p_k^{\alpha_k-1} \triangleq \frac{1}{\Delta(\vec{\alpha})} \prod_{k=1}^{K} p_k^{\alpha_k-1}$$

Gibbs updating rule

$$p(z_i{=}k|\vec{z}_{\neg i},\vec{w}) = \frac{p(\vec{w},\vec{z})}{p(\vec{w},\vec{z}_{\neg i})} = \frac{p(\vec{w}|\vec{z})}{p(\vec{w}_{\neg i}|\vec{z}_{\neg i})p(w_i)} \cdot \frac{p(\vec{z})}{p(\vec{z}_{\neg i})}$$

$$\propto \frac{\Delta(\vec{n}_z+\vec{\beta})}{\Delta(\vec{n}_{z,\neg i}+\vec{\beta})} \cdot \frac{\Delta(\vec{n}_m+\vec{\alpha})}{\Delta(\vec{n}_{m,\neg i}+\vec{\alpha})}$$

$$= \frac{\Gamma(n_k^{(t)}+\beta_t)\,\Gamma(\sum_{t=1}^{V} n_{k,\neg i}^{(t)}+\beta_t)}{\Gamma(n_{k,\neg i}^{(t)}+\beta_t)\,\Gamma(\sum_{t=1}^{V} n_k^{(t)}+\beta_t)} \cdot \frac{\Gamma(n_m^{(k)}+\alpha_k)\,\Gamma(\sum_{k=1}^{K} n_{m,\neg i}^{(k)}+\alpha_k)}{\Gamma(n_{m,\neg i}^{(k)}+\alpha_k)\,\Gamma(\sum_{k=1}^{K} n_m^{(k)}+\alpha_k)}$$

$$= \frac{n_{k,\neg i}^{(t)}+\beta_t}{\sum_{t=1}^{V} n_{k,\neg i}^{(t)}+\beta_t} \cdot \frac{n_{m,\neg i}^{(k)}+\alpha_k}{[\sum_{k=1}^{K} n_m^{(k)}+\alpha_k]-1}$$

$$\propto \frac{n^{(w)}_{k,\neg i}+\beta_t}{\sum_{t=1}^{V} n^{(t)}_{k,\neg i}+\beta_t}(n^{(k)}_{m,\neg i}+\alpha_k)$$

得到词分布和主题分布（用期望来估计）：

$$\varphi_{k,t} = \frac{n^{(t)}_k + \beta_t}{\sum_{t=1}^{V} n^{(t)}_k + \beta_t}$$

$$\vartheta_{m,k} = \frac{n^{(k)}_m + \alpha_k}{\sum_{k=1}^{K} n^{(k)}_m + \alpha_k}$$

超参数的确定

α表达了不同文档间主题是否鲜明，β度量了有多少近义词能够属于同一个类别。

给定主题数目k，可以使用：$\alpha = 50/k; \beta = 0.01$，然后在验证集上看效果。交叉验证。

一种迭代求超参数的方法：

- Digamma函数：$\Psi(x) = \frac{d \ln \Gamma(x)}{dx} = \frac{\Gamma'(x)}{\Gamma(x)}$
- 迭代公式：(T. Minka)

$$\alpha_k = \frac{\left(\left(\sum_{m=1}^{M} \Psi\left(n^{(k)}_m + \alpha_k\right)\right) - M \cdot \Psi(\alpha_k)\right)}{\left(\sum_{m=1}^{M} \Psi\left(n_m + \sum_{j=1}^{K} \alpha_j\right)\right) - M \cdot \Psi\left(\sum_{j=1}^{K} \alpha_j\right)} \cdot \alpha_k$$

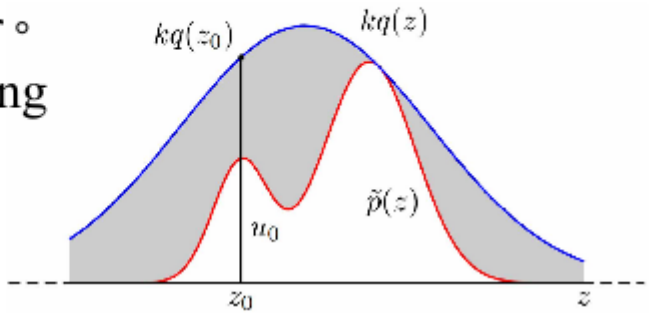
LDA总结

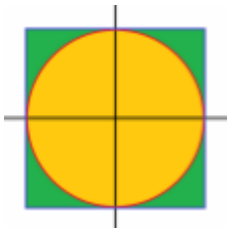
- 由于在词和文档之间加入的主题的概念，可以较好的解决一词多义和多词一义的问题。
- 在实践中发现，LDA用于短文档往往效果不明显——这是可以解释的：因为一个词被分配给某个主题的次数和一个主题包括的词数目尚未敛。往往需要通过其他方案“连接”成长文档。
  - 用户评论/Twitter/微博
- LDA可以和其他算法相结合。首先使用LDA将长度为Ni的文档降维到K维(主题的数目)，同时给出每个主题的概率(主题分布)，从而可以使用if-idf继续分析或者直接作为文档的特征进入聚类或者标签传播算法——用于社区发现等问题。

LDA也可以用来处理图像。LDA运算时间很长，每一次的采样，都会发生微妙的变化。

带拒绝的采样

- 上述抽样问题能否用来解决一般概率分布函数的抽样问题？如：根据均匀分布函数得到正态分布的抽样。
- Rejection sampling





对某概率分布函数进行采样的意义：

根据抽样结果估算该分布函数的参数，从而完成参 数的学习。 前提：系统已经存在，但参数未知； 方法：通过采样的方式，获得一定数量的样本，从而学习该系统的参数。 例：投硬币试验中，进行N次试验，n次朝上，N-n次朝下—可以认为，是进行了N次(独立)抽样。 假定朝上的概率为p，使用对数似然函数作为目标函数：

$$f(n\mid p)=\log\left(p^{n}(1-p)^{N-n}\right)\xrightarrow{\Delta}h(p)=\log\left(p^{n}(1-p)^{N-n}\right)$$

$$\frac{\partial h(p)}{\partial p}=\frac{n}{p}-\frac{N-n}{1-p}\xrightarrow{\Delta}0\Rightarrow p=\frac{n}{N}$$

一般的说，上述结论可以直接推广：频率的极限为概率： p=n/N 将上述二项分布扩展成多项分布，如K项分布：pi=ni/N 从而得到K项分布的参数：Pi 在主体模型LDA中，每个文档的主题分布和每个主题的词分布都是多项分布，如果能够通过采样的方式获得它们的一定数量的样本，即可估算主题分布和词分布的参数，从而完成参数学习！ 贝叶斯网络的另一种重要参数学习手段是EM算法，参见GMM、pLSA、HMM的推导过程。

TF-IDF

TF-IDF(Term Frequency-Inverse Document Frequency) 词频－逆文件频率

作用：用于评估一字词对于一个文件集或一个语料库中的其中一份文件的重要程度。字词重要性随着它在文件中出现的次数成正比增加，随着它在语料库中出现的频率成反比下降。

含义：一个词语在一篇文章中出现的次数越多，同时在所有文档中出现的次数越少，越能够代表这篇文章。

词频 **(term frequency, TF)** 指的是某一个给定的词语在该文件中出现的次数。这个数字通常会被归一化(一般是词频除以文章总词数)， 以防止它偏向长的文件。（同一个词语在长文件里可能会比短文件有更高的词频，而不管该词语重要与否。

> 但是， 需要注意， 一些通用的词语对于主题并没有太大的作用， 反倒是一些出现频率较少的词才能够表达文章的主题， 所以单纯使用是TF不合适的。权重的设计必须满足：一个词预测主题的能力越强，权重越大，反之，权重越小。所有统计的文章中，一些词只是在其中很少几篇文章中出现，那么这样的词对文章的主题的作用很大，这些词的权重应该设计的较大。IDF

> 就是在完成这样的工作.

公式:

$$TF_w = \frac{在某一类中词条w出现的次数}{该类中所有的词条数目}$$

逆向文件频率 **(inverse document frequency, IDF)**IDF的主要思想是：如果包含词条t的文档越少, IDF越大，则说明词条具有很好的类别区分能力。某一特定词语的IDF，可以由总文件数目除以包含该词语之文件的数目，再将得到的商取对数得到。

公式:

$$IDF = log(\frac{语料库的文档总数}{包含词条w的文档数+1}),分母之所以要加1，是为了避免分母为0$$

某一特定文件内的高词语频率，以及该词语在整个文件集合中的低文件频率，可以产生出高权重的TF-IDF。因此，TF-IDF倾向于过滤掉常见的词语，保留重要的词语。

TF-IDF=TF*IDF

相似度