分层狄利克雷过程在文本挖掘中的应用

IIP组暑期讨论班

--李振兴

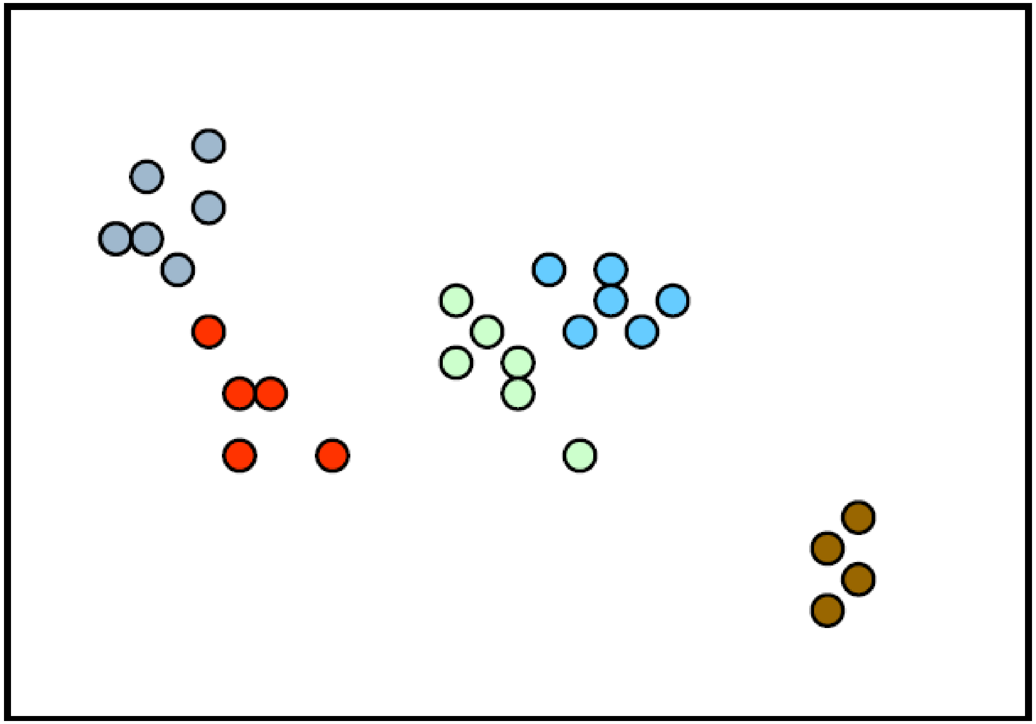
南京大学计算机科学与技术系

NANJING UNIVERSITY

CONTENTS 目录

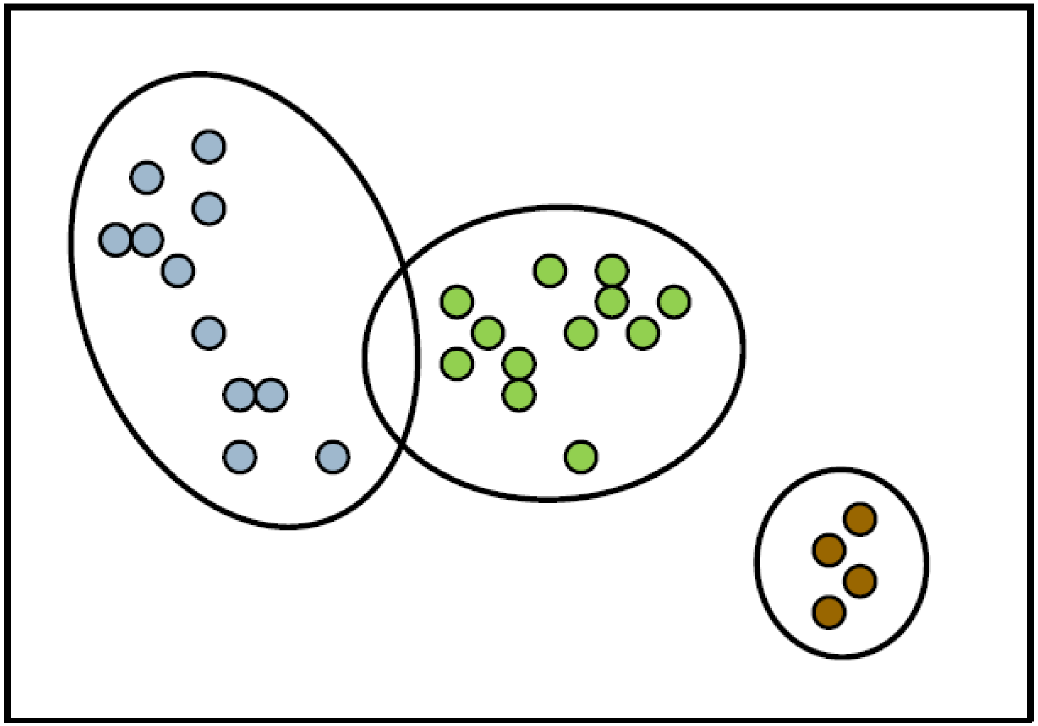
非参数贝叶斯-一个简单的例子

- 考虑一个数据集，我们要将这些数据聚成K类，如下图



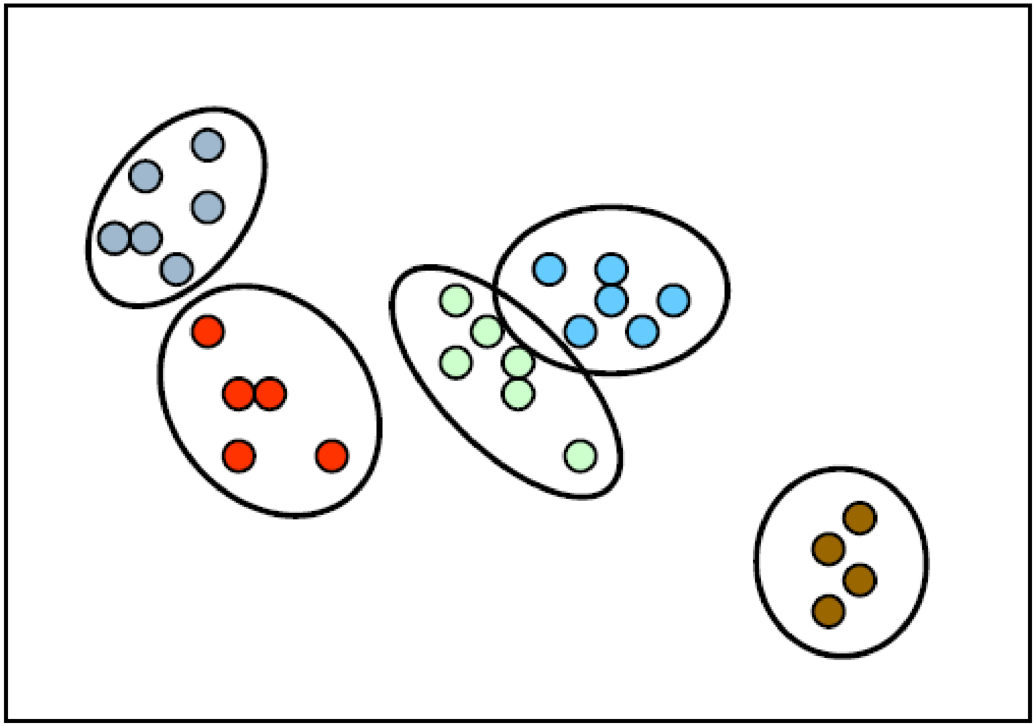
非参数贝叶斯-一个简单的例子

- 考虑一个数据集，我们要将这些数据聚成K类，如下图



非参数贝叶斯-一个简单的例子

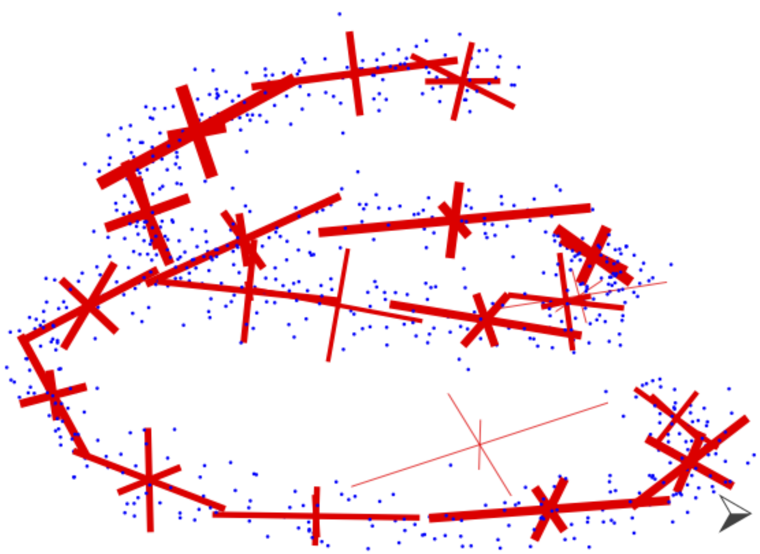
- 考虑一个数据集，我们要将这些数据聚成K类，如下图



- 需要考虑K的取值问题?
- 非参数贝叶斯就是用来解决分类数目不确定性问题。它能够从数据中自动学习到适合的K的值。

狄利克雷过程-实例

无限高斯混合模型



- 对于一个混合模型让：$\mathrm{X} = \mathrm{x}_1, \mathrm{x}_2, \cdots, \mathrm{x}_N$

$$p(x_i\,|\Theta) = \sum_{l=1}^{K} w_l N(\mu_l, \Sigma_l)$$

其中 $\sum_{l=1}^{K} w_l = 1$，得到似然函数

$$\ln P(\mathrm{X}|\Theta) = \sum_{n=1}^{N} \ln\left\{\sum_{l=1}^{K} w_l N(\mu_l, \Sigma_l)\right\}$$

- 当K是常量时，可以采用EM算法对参数进行求解。如果我们让K是一个变量，将会出现下面的情况

$$\underset{\theta_1,\cdots,\theta_K,w_1,\cdots,w_K,K}{\operatorname{argmax}} \quad \ln P(\mathrm{X}\,|\Theta)$$

- 可能会出现$K = N$的情形，显然效果不理性

狄利克雷过程-思考

- 对于数据$\mathrm{x}_1, \mathrm{x}_2, \cdots, \mathrm{x}_N$,每一个$\mathrm{x}_1$对应一个参数$(\mu_i, \Sigma_i)$
- 我们需要一个好的先验知识$\Pr(\Theta|G)$
- 我们也可以考虑K可能是无限的情形
- 具有“聚类”的性质，尽可能的通过单一的参数α来控制
- 此时先验可以写成下面的形式，边缘概率密度为：

$$p(\Theta) = \int_G \Pr(\Theta|G)\, p(G) dG$$

因此我们可以考虑G的有趣的性质

- $\Pr(\Theta|G)$必须是离散的概率分布
- 或者它应该与某个基础分布H有关

狄利克雷过程-定义

狄利克雷过程

假设H是测度空间Θ上的的随机概率分布，参数α是正实数，空间Θ上的概率分布G如果满足以下的条件：对于测度空间Θ上的任意一个有限划分$A_1, A_2, \cdots, A_r$，均有以下关系存在：

$$\big(G(A_1), G(A_2), \cdots, G(A_r)\big) \sim Dir(\alpha H(A_1), H(A_2), \cdots, H(A_r)) \quad (1)$$

则G服从有基分布H和Concentration参数α组成的Dirichlet过程，即：$G \sim DP(\alpha, H)$，反之(1)式成立。

What does this all mean?

狄利克雷过程-均值与方差

- 狄利克雷分布 $p(x_1,x_2,\cdots,x_n) \sim Dir(\alpha_1,\alpha_2,\cdots,\alpha_n)$ 的均值和方差为:

$$E(x_i) = \frac{\alpha_i}{\sum_{k=1}^{n}\alpha_k}$$

$$Var(x_i) = \frac{\alpha_i(\sum_{k=1}^{n}\alpha_k - \alpha_i)}{(\sum_{k=1}^{n}\alpha_k)^2(\sum_{k=1}^{n}\alpha_k + 1)}$$

- 利用上述性质可得狄利克雷过程的均值和方差:

$$E(\mathrm{G}(A_i)) = \frac{\alpha H(A_i)}{\sum_{k=1}^{n}\alpha H(A_k)} = \frac{\alpha H(A_i)}{\alpha\sum_{k=1}^{n}H(A_k)} = H(A_i)$$

$$Var(\mathrm{G}(A_i)) = \frac{\alpha H(A_i)\,(\sum_{k=1}^{n}\alpha H(A_k) - \alpha H(A_i))}{(\sum_{k=1}^{n}\alpha H(A_k)\,)^2(\sum_{k=1}^{n}\alpha H(A_k)\ + 1)}$$

$$= \frac{\alpha H(A_i)\,(\alpha - \alpha H(A_i))}{\alpha^2(\alpha+1)} = \frac{H(A_i)(1-H(A))}{\alpha+1}$$

- α的大小对得狄利克雷过程的影响

当$\alpha \to \infty$时: $Var(\mathrm{G}(A_i)) \to 0$, 此时$\mathrm{G} = \mathrm{H}$

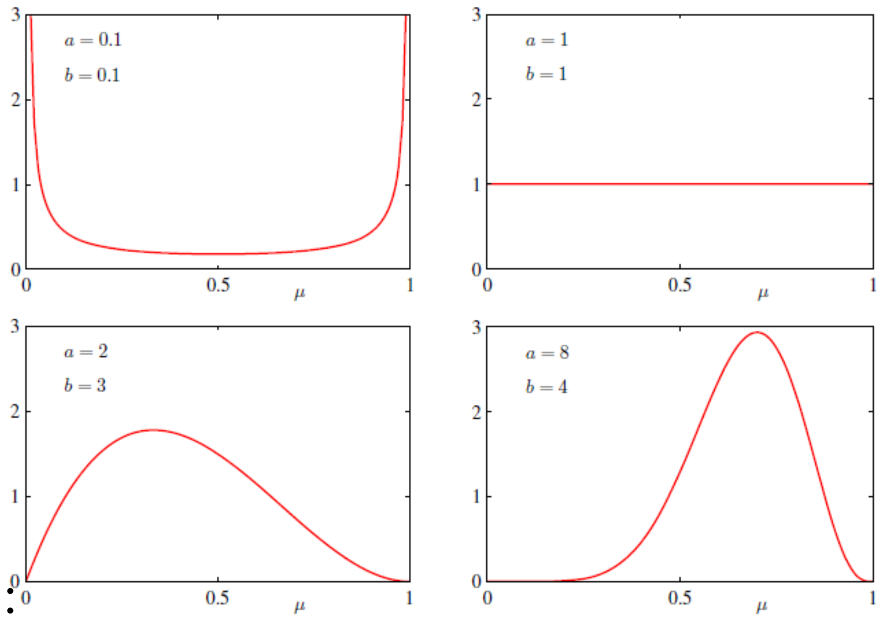
当$\alpha \to 0$时: $Var(\mathrm{G}(A_i)) \to H(A_i)(1 - H(A))$, 类似于伯努利分布

狄利克雷过程-构造

Stick-breaking (截棍)构造

基于相互独立的变量$(\beta_k)_{k=1}^{\infty}$和$(\theta_k)_{k=1}^{\infty}$的Stick-breaking构造

- $\beta_k \sim Beta(1, \alpha)$
- $\theta_k \sim H$
- $\pi_k = \beta_k \prod_{l=1}^{k-1}(1-\beta_l)$
- $G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}$



其中$Beta$分布的形式为（$0<\mu<1$）：

$$Beta(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1}(1-\mu)^{b-1}$$

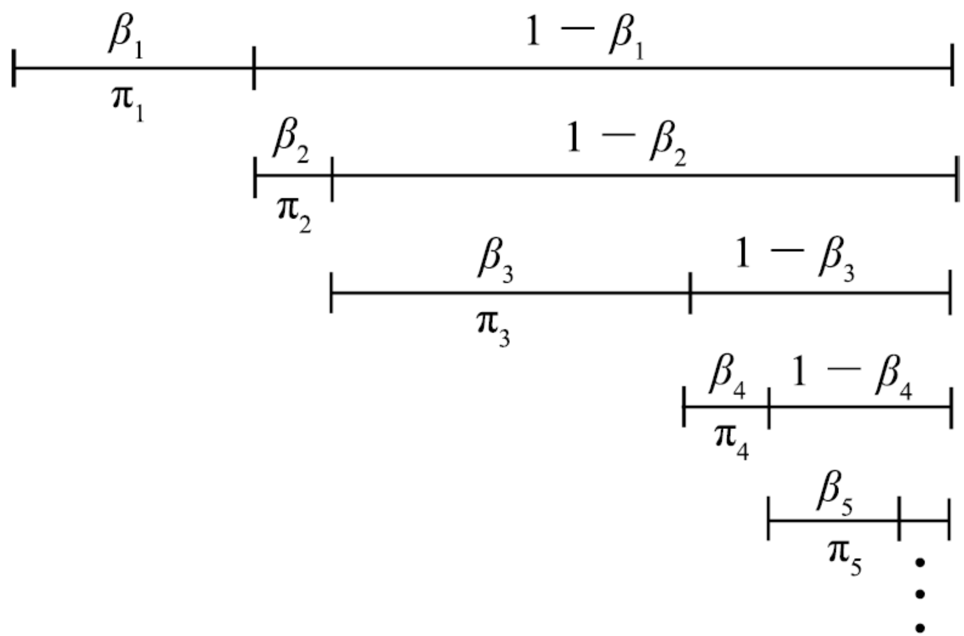
狄利克雷过程-构造

Stick-breaking (截棍)构造

基于相互独立的变量$(\beta_k)_{k=1}^{\infty}$和$(\theta_k)_{k=1}^{\infty}$的Stick-breaking构造

- $\beta_k \sim Beta(1, \alpha)$
- $\theta_k \sim H$
- $\pi_k = \beta_k \prod_{l=1}^{k-1}(1-\beta_l)$
- $G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}$

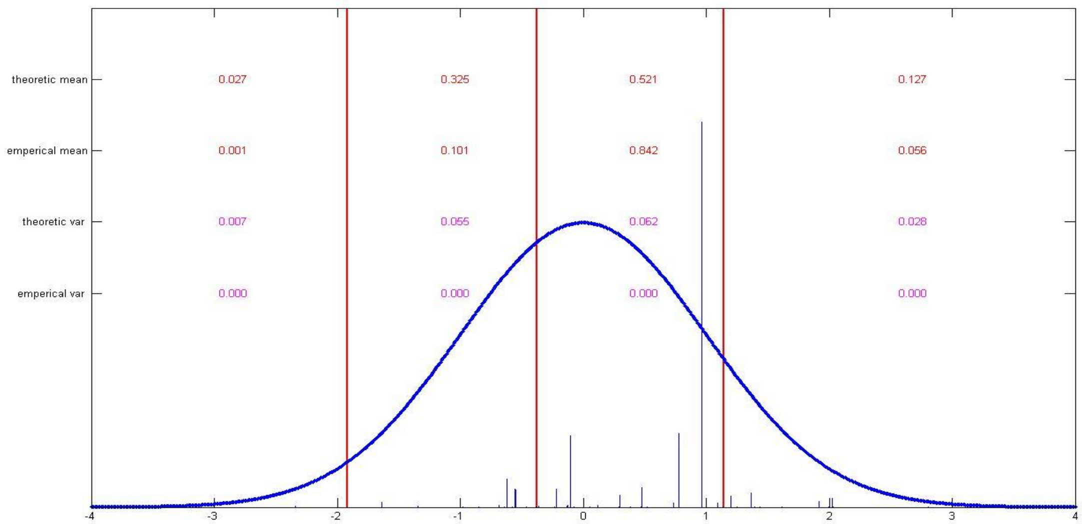
Stick-breaking构造过程如右图:



狄利克雷过程-构造

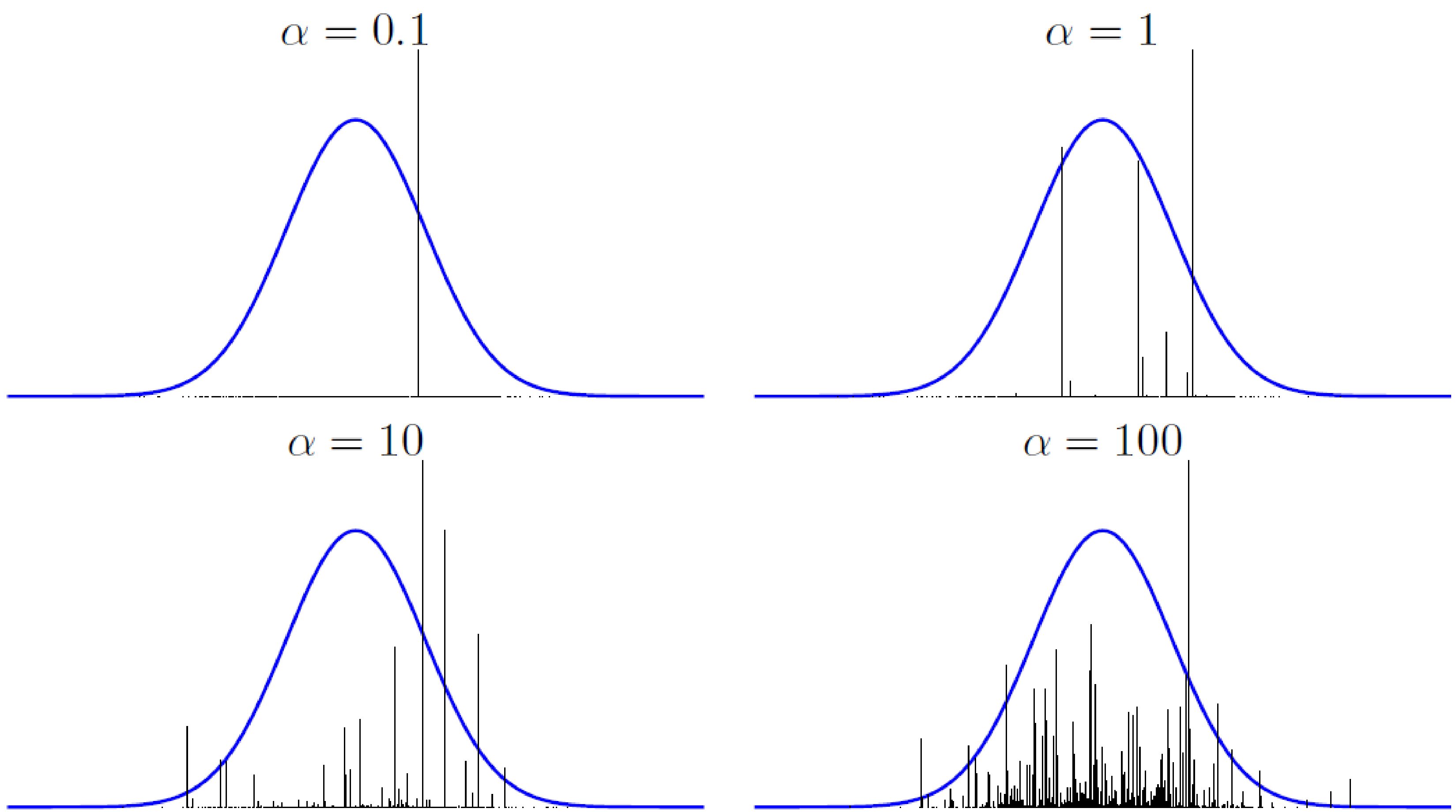
- 选取基础分布H为高斯分布
- 根据$G \sim DP(\alpha, H)$，利用截棍过程进行采样，然后随机将区间分成r个互不相交的区间$A_1, A_2, \cdots A_r$，根据DP的定义在此处的直观理解为：

$$\left(\sum_{A_1} \pi_{l_1}, \cdots, \sum_{A_r} \pi_{l_r}\right) \sim Dir(\alpha H(A_1), \cdots, \alpha H(A_r))$$



狄利克雷过程-构造

- DP过程$G\sim DP(\alpha,H)$中参数α对分布函数的影响



狄利克雷过程-后验

- 设$G\sim DP(\alpha,H)$，我们从分布G中采样$\theta\sim G$，则对测度空间Θ的有限划分的后验分布也是Dirichlet过程，观测数据只影响其所在划分区域的分布参数，即：

$$(G(A_1),\cdots,G(A_r)\,|\theta\in A_k)\sim$$
$$Dir(\alpha H(A_1),\cdots,\alpha H(A_k)+1,\cdots,\alpha H(A_r))$$

- 若$\theta_1,\theta_2,\cdots,\theta_n$独立服从$G$，即：$\theta_1,\theta_2,\cdots,\theta_n\sim$G，则：

$$(G(A_1),\cdots,G(A_r)\,|\theta_1,\cdots,\theta_n)\sim$$
$$Dir(\alpha H(A_1)+n_1,\cdots,\alpha H(A_r)+n_r)$$

- 根据DP的定义可知：

$$G'=G|\theta_1,\theta_2,\cdots,\theta_n\sim\mathrm{DP}\left(\alpha+n,\frac{\alpha H+\sum_{i=1}^{n}\delta_{\theta_i}}{\alpha+n}\right)$$
$$\sim\mathrm{DP}\left(\alpha+n,\frac{\alpha}{\alpha+n}H+\frac{\sum_{i=1}^{n}\delta_{\theta_i}}{\alpha+n}\right)$$

狄利克雷过程-预测

- 现在我们从G采取样本$\theta_{\mathrm{n+1}}$，即$\theta_{\mathrm{n+1}} \sim G$，对$\theta_{\mathrm{n+1}}$进行预测：

$$\begin{aligned}
\mathrm{P}(\theta_{\mathrm{n+1}} \in A|\theta_1,\theta_2,\cdots,\theta_n) &= \int_G P(\theta_{\mathrm{n+1}} \in A|G)\, P(G|\theta_1,\theta_2,\cdots,\theta_n)dG \\
&= \mathrm{E}(\mathrm{G(A)}|\theta_1,\theta_2,\cdots,\theta_n) \\
&= \mathrm{E}\big(G'(A)\big) \\
&= \frac{\alpha}{\alpha+n}H + \frac{\sum_{i=1}^{n}\delta_{\theta_i}}{\alpha+n}
\end{aligned}$$

其中$\left(\mathrm{E}\big(G(A)\big) = H(A) \implies E\big(G'(A)\big) = \frac{\alpha}{\alpha+n}H + \frac{\sum_{i=1}^{n}\delta_{\theta_i}}{\alpha+n}\right)$

上述公式即为中国餐馆过程（Chinese Restaurant Process）。

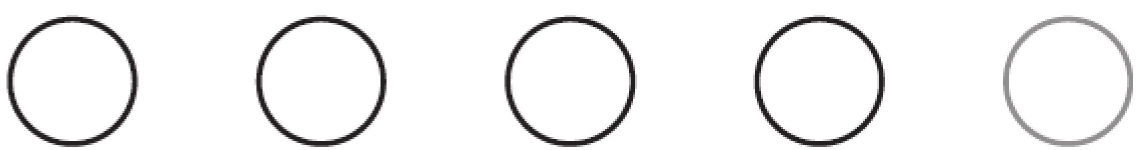
狄利克雷过程-构造

Chinese Restaurant Process

考虑一中国菜餐馆，可以容纳无限多张桌子。θ_i被比作进入餐厅的顾客，而不同的值ϕ_k对应顾客就做的桌子，第一个顾客就座于第一张桌子，第i个顾客以正比于已经就坐于第k张桌子ϕ_k的顾客数η_k的概率就坐于第k张桌子，以正比于α的概率就坐于一张新桌子。

$$p(\text{就坐于第}k\text{张桌子}) = \frac{\eta_k}{\alpha + \sum_k \eta_k}$$

$$p(\text{就坐于一张新桌子}) = \frac{\alpha}{\alpha + \sum_k \eta_k}$$



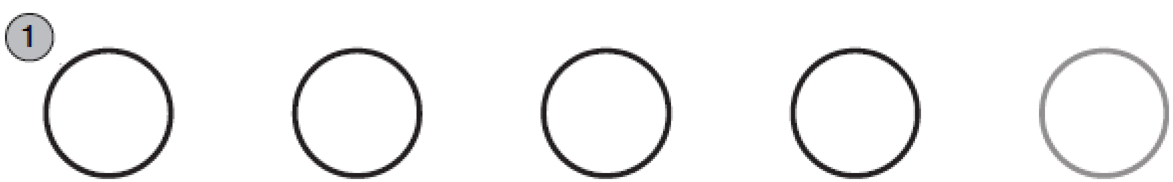
狄利克雷过程-构造

Chinese Restaurant Process

考虑一中国菜餐馆，可以容纳无限多张桌子。θ_i被比作进入餐厅的顾客，而不同的值ϕ_k对应顾客就做的桌子，第一个顾客就座于第一张桌子，第i个顾客以正比于已经就坐于第k张桌子ϕ_k的顾客数η_k的概率就坐于第k张桌子，以正比于α的概率就坐于一张新桌子。

$$p(\text{就坐于第}k\text{张桌子}) = \frac{\eta_k}{\alpha + \sum_k \eta_k}$$

$$p(\text{就坐于一张新桌子}) = \frac{\alpha}{\alpha + \sum_k \eta_k}$$



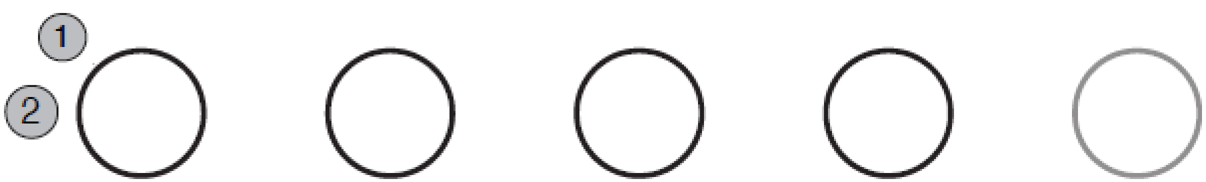
狄利克雷过程-构造

Chinese Restaurant Process

考虑一中国菜餐馆，可以容纳无限多张桌子。θ_i被比作进入餐厅的顾客，而不同的值ϕ_k对应顾客就做的桌子，第一个顾客就座于第一张桌子，第i个顾客以正比于已经就坐于第k张桌子ϕ_k的顾客数η_k的概率就坐于第k张桌子，以正比于α的概率就坐于一张新桌子。

$$p(\text{就坐于第}k\text{张桌子}) = \frac{\eta_k}{\alpha + \sum_k \eta_k}$$

$$p(\text{就坐于一张新桌子}) = \frac{\alpha}{\alpha + \sum_k \eta_k}$$



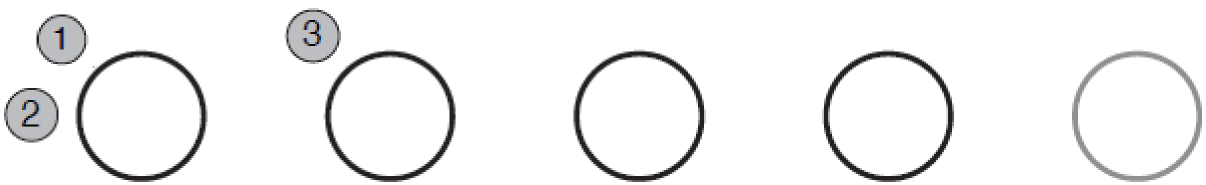
狄利克雷过程-构造

Chinese Restaurant Process

考虑一中国菜餐馆，可以容纳无限多张桌子。θ_i被比作进入餐厅的顾客，而不同的值ϕ_k对应顾客就做的桌子，第一个顾客就座于第一张桌子，第i个顾客以正比于已经就坐于第k张桌子ϕ_k的顾客数η_k的概率就坐于第k张桌子，以正比于α的概率就坐于一张新桌子。

$$p(\text{就坐于第}k\text{张桌子}) = \frac{\eta_k}{\alpha + \sum_k \eta_k}$$

$$p(\text{就坐于一张新桌子}) = \frac{\alpha}{\alpha + \sum_k \eta_k}$$



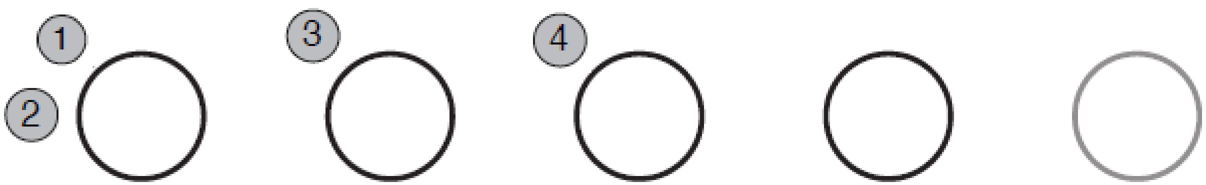
狄利克雷过程-构造

Chinese Restaurant Process构造

考虑一中国菜餐馆，可以容纳无限多张桌子。θ_i被比作进入餐厅的顾客，而不同的值ϕ_k对应顾客就做的桌子，第一个顾客就座于第一张桌子，第i个顾客以正比于已经就坐于第k张桌子ϕ_k的顾客数η_k的概率就坐于第k张桌子，以正比于α的概率就坐于一张新桌子。

$$p(\text{就坐于第}k\text{张桌子}) = \frac{\eta_k}{\alpha + \sum_k \eta_k}$$

$$p(\text{就坐于一张新桌子}) = \frac{\alpha}{\alpha + \sum_k \eta_k}$$



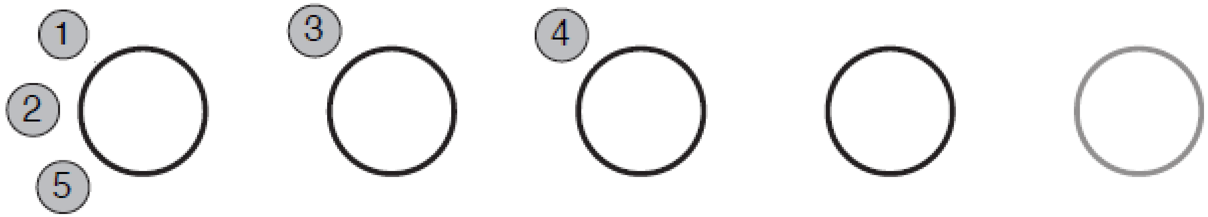
狄利克雷过程-构造

Chinese Restaurant Process

考虑一中国菜餐馆，可以容纳无限多张桌子。θ_i被比作进入餐厅的顾客，而不同的值ϕ_k对应顾客就做的桌子，第一个顾客就座于第一张桌子，第i个顾客以正比于已经就坐于第k张桌子ϕ_k的顾客数η_k的概率就坐于第k张桌子，以正比于α的概率就坐于一张新桌子。

$$p(\text{就坐于第}k\text{张桌子}) = \frac{\eta_k}{\alpha + \sum_k \eta_k}$$

$$p(\text{就坐于一张新桌子}) = \frac{\alpha}{\alpha + \sum_k \eta_k}$$



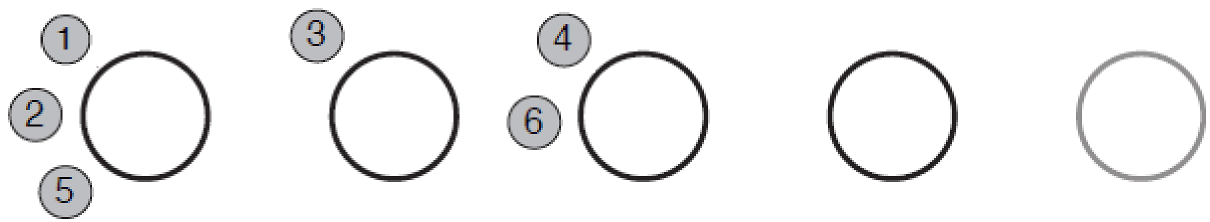
狄利克雷过程-构造

Chinese Restaurant Process

考虑一中国菜餐馆，可以容纳无限多张桌子。θ_i被比作进入餐厅的顾客，而不同的值ϕ_k对应顾客就做的桌子，第一个顾客就座于第一张桌子，第i个顾客以正比于已经就坐于第k张桌子ϕ_k的顾客数η_k的概率就坐于第k张桌子，以正比于α的概率就坐于一张新桌子。

$$p(\text{就坐于第}k\text{张桌子}) = \frac{\eta_k}{\alpha + \sum_k \eta_k}$$

$$p(\text{就坐于一张新桌子}) = \frac{\alpha}{\alpha + \sum_k \eta_k}$$



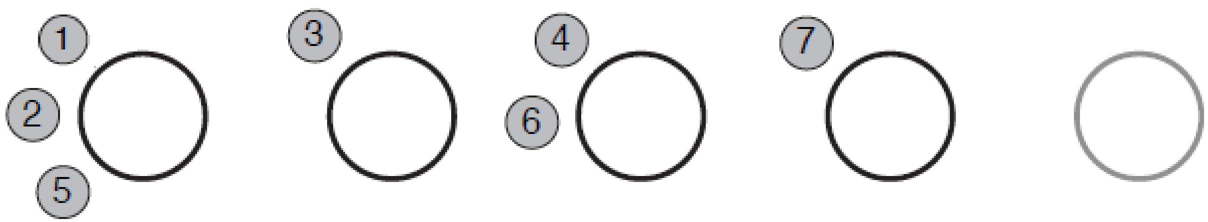
狄利克雷过程-构造

Chinese Restaurant Process

考虑一中国菜餐馆，可以容纳无限多张桌子。θ_i被比作进入餐厅的顾客，而不同的值ϕ_k对应顾客就做的桌子，第一个顾客就座于第一张桌子，第i个顾客以正比于已经就坐于第k张桌子ϕ_k的顾客数η_k的概率就坐于第k张桌子，以正比于α的概率就坐于一张新桌子。

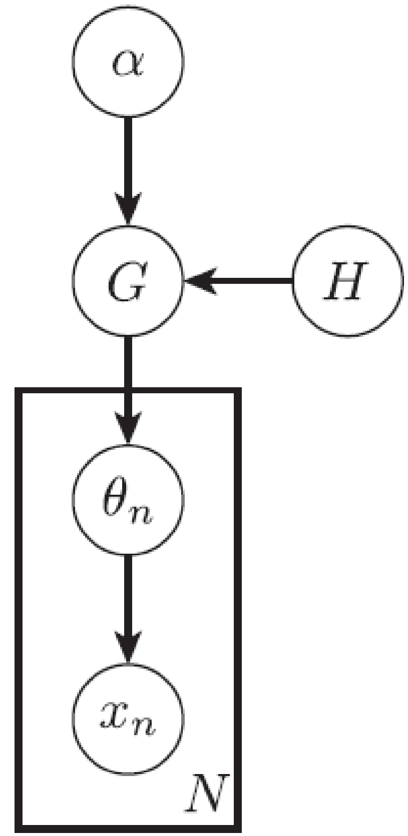
$$p(\text{就坐于第}k\text{张桌子}) = \frac{\eta_k}{\alpha + \sum_k \eta_k}$$

$$p(\text{就坐于一张新桌子}) = \frac{\alpha}{\alpha + \sum_k \eta_k}$$



狄利克雷混合模型

图模型:



在Dirichlet过程混合模型中，Dirichlet过程作为数据的先验分布存在，假设观测数据是x_i，其分布服从:

$$G \mid \alpha, H \sim DP(\alpha, H)$$
$$\theta_n \mid G \sim G$$
$$x_n \mid \theta_n \sim F(\theta_n)$$

利用此混合模型可以解决前面的无限高斯混合模型中，参数K不确定性问题。

分层狄利克雷过程

生成模型:

$$G_0 \mid \gamma, H \sim DP(\gamma, H)$$

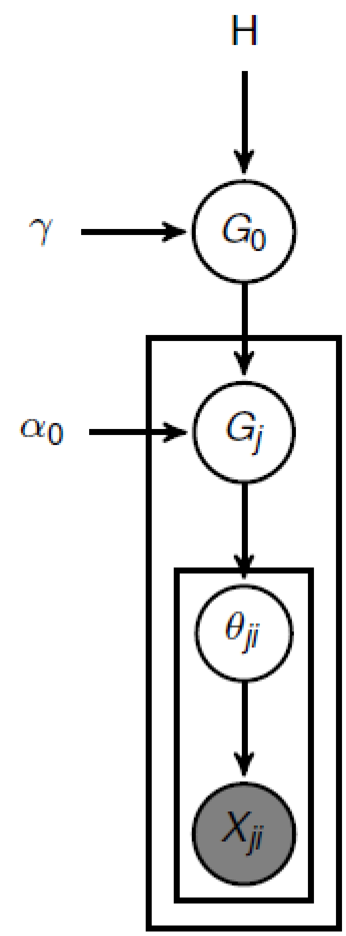
$$G_j \sim DP(\alpha_0, G_0)$$

$$\theta_{ji} \sim G_j$$

$$X_{ji} \sim F(x|\theta_{ji})$$

从$G_0 \sim DP(\cdot)$中采样是困难的，因此我们采用前面构造的方法生成G_0。

图模型:



分层狄利克雷过程-Stick breaking构造

生成模型:

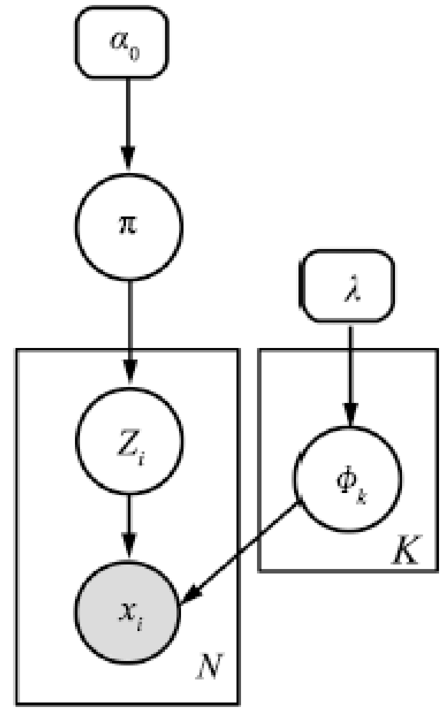
$$\beta \sim GEM(\gamma) \qquad G_0 \sim \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k}$$

$$\pi_j \sim DP(\alpha_0, \beta) \qquad G_j \sim \sum_{k=1}^{\infty} \pi_{jk} \delta_{\phi_k}$$

$$z_{ji} \sim \pi_j \qquad \phi_k \sim H \qquad X_{ji} \sim F(x|\phi_{z_{ji}})$$

从$G_0 \sim DP(\cdot)$中采样是困难的，因此我们采用前面构造的方法生成G_0。

图模型:



分层狄利克雷过程

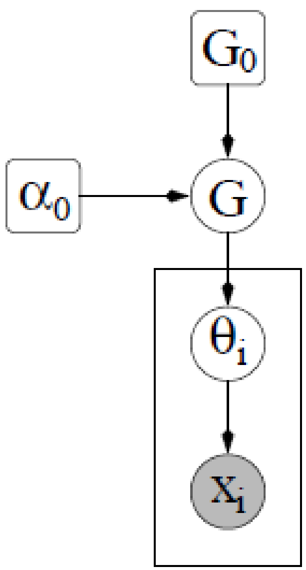
HDP-LDA

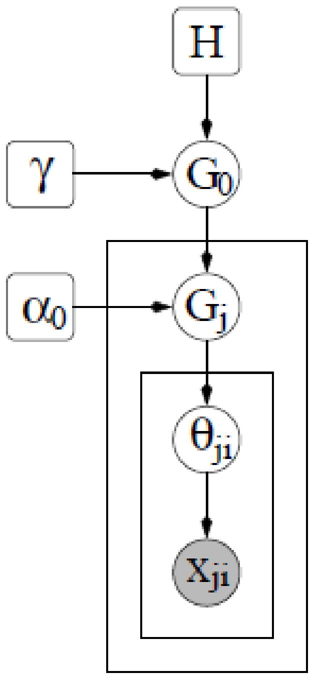
- 相同点:
  ① 均实现文档数据内部的主题挖掘，实现多文档之间的主题共享，同时对文档的主题和主题内部的单词建立概率描述。
  ② 均是非监督模型学习模型，均假设文档是服从某个概率分布的主题组成，每个主题服从某种概率分布的单词组成。
  ③ 均将文档数据看成Bag of word，文档中的Word满足可交换性。
- 不同点:
  ① HDP不但能够实现聚类和推断功能，而且能够自动生成聚类数目，因此大大增强了算法的鲁棒性;
  ② HDP模型不再是一种单纯的主题提取算法，而且广泛应用于HDP-HMM模型等其他算法的构造中。

分层狄利克雷过程

HDP-DP

- HDP为什么能够工作？为什么不直接使用DP？





狄利克雷过程-扩展

狄利克雷过程混合模型(*DP*)

$x_i \mid \theta_i \sim F(\theta_i); \theta_i \mid G \sim G; G \mid \alpha, H \sim DP(\alpha, H)$

说明：原子θ_i由基础分布 *H* 抽样，共享相同原子的观测样本分配到同一聚类中

层次狄利克雷过程混合模型(*HDP*)

$x_{ji}|\theta_{ji} \sim F(\theta_{ji}); \theta_{ji}|G_j \sim G_j;$

$G_j|\alpha, G_0 \sim DP(\alpha, G_0);$

$G_0|\gamma, H \sim DP(\gamma, H)$

说明：将狄利克雷过程的基础分 *H* 扩展为随机测度 G_0，则多个数据源（任务）之间共享随机测度 G_0，相关任务之间通过调整共享原子，实现信息迁移

嵌套狄利克雷过程混合模型(*nDP*)

$x_{ij} \sim p(x_{ij} \mid \theta_{ij}); \theta_{ij} \sim G_j;$

$\{G_1, G_2, ..., G_J\} \sim DP(\alpha, DP(\beta, H))$

说明：将狄利克雷过程的基础分 *H* 扩展为狄利克雷过程 *DP*(*β*, *H*)，则多个数据源(任务)之间共享随机测度空间，相关任务之间通过调整共享原子，实现分布(任务)聚类

关联狄利克雷过程(πDDP)

$p(y_i|\psi_i) = f(y_i|\psi_i); \psi_i \overset{\text{iid}}{\sim} F_{x_i};$

$F_{x_i} \sim \pi DDP(\alpha, H, \lambda)$

说明：将狄利克雷过程应用到观测样本 y_i 所在位置 x_i，则可以通过两个相关位置 x_i, x_j 之间原子共享，推断两个样本 y_i, y_j 时间空间相关性

矩阵、核截棍过程混合模型

说明：将狄利克雷过程扩展到特征级(非样本级)，可以实现多个相关任务(数据源)之间特征级的细粒度信息共享，可以消除冗余无关特征

其他相关算法

- Models for time series
  ① Infinite Hidden Markov Model / HDP-HMM (Beal, Ghahramani, Rasmussen, 2002; Teh, Jordan, Beal, Blei, 2006; Fox, Sudderth, Jordan, Willsky, 2008)
  ② infinite factorial HMM / Markov IBP (van Gael, Teh, Ghahramani, 2009)
  ③ beta process HMM (Fox, Sudderth, Jordan, Willsky, 2009)

- Hierarchical models for sharing structure
  ① hierarchical Dirichlet processes (Teh, Jordan, Beal, Blei, 2006)
  ② hierarchical beta processes (Thibaux, Jordan, 2007)

参考文献

- David Blackwell and James B. MacQueen. ―Ferguson Distributions via PolyaUrn Schemes.‖ *Annals of Statistics* 1(2), 1973, 353-355.
- David M. Bleiand Michael I. Jordan. ―Variational inference for Dirichlet process mixtures.‖ *Bayesian Analysis*1(1), 2006.
- Thomas S. Ferguson. ―A Bayesian Analysis of Some Nonparametric Problems‖ *Annals of Statistics*1(2), 1973, 209-230.
- Zoubin Gharamani. ―Non-parametric Bayesian Methods.‖ UAI Tutorial July 2005.
- Teg Grenager. ―Chinese Restaurants and Stick Breaking: An Introduction to the Dirichlet Process
- R.M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9:249-265, 2000.
- C.E. Rasmussen. The Infinite Gaussian Mixture Model. *Advances in Neural Information Processing Systems*12, 554-560. (Eds.) Solla, S. A., T. K. Leenand K. R. Müller, MIT Press (2000).
- Y.W. Teh. ―Dirichlet Processes.‖ Machine Learning Summer School 2007 Tutorial and Practical Course.
- Y.W. Teh, M.I. Jordan, M.J. Beal and D.M. Blei. ―Hierarchical Dirichlet Processes.‖ *J. American Statistical Association*101(476):1566-1581, 2006.

Thank you!

分层狄利克雷过程在文本挖掘中的应用

IIP组暑期讨论班
南京大学计算机科学与技术系