Product Recommendations Based on Browsing Trajectories

Liu Chuang

2018-11-13

Update：11-21

Product Recommendations Based on Browsing Trajectories

- Motivation
- Recommendation Alg
- Processing
- Baselines
- Problems

1. Motivation

- **Using an improved LDA(Latent Dirichlet Analysis) topic model which is a statistical method ever used to find the hidden theme in the article but now to find latent features of products.**

- **Using User-browsing tracks(instead of product's property) to mine latent features (semantics) of commodities and users' latent interests or tastes**

- **Considering Time on products which is important to reveal users' interests**

1. Motivation

Pros:

- **Using the user's implicit feedback information (browsing tracks) which is more dense and easier available than the explicit feedback, and ensure a more realistic reflection of users' interests or tastes. Further more , it has a better real-time performance**

- **Mining the latent correlation between products, and customizing the recommended products to users(avoiding continuous repetition and top-selling)**

- **To some extent, solving the cold start problem**

- **Revitalization and utilization of long tail resources. Users are often unfamiliar with long tail content, they cannot actively search. Only through the recommended way can we attract user's attention, and discover user's interest.**

From www.longtail.com

2. Recommendation Alg

- **Content-based**

  Base on the features of the item

- **Collaborative Filtering-based**

  Base on historical usage data to item

Researchers from the Web Search, E-commerce and Marketplace domains have realized that just like one can train word embeddings by treating a sequence of words in a sentence as context, the same can be done for training embeddings of user actions by treating sequence of user actions as context.” —— Mihajlo Grbovic, Airbnb

2. Recommendation Alg

- **Matrix factorization**

  Netflix—— SVD

  - Decompose the matrix, map users and items to k-dimensional space
  - Get some hidden factors behind the item, and users' preferences

Cons :

- Explicit feedback

- Browsing track

2. Recommendation Alg

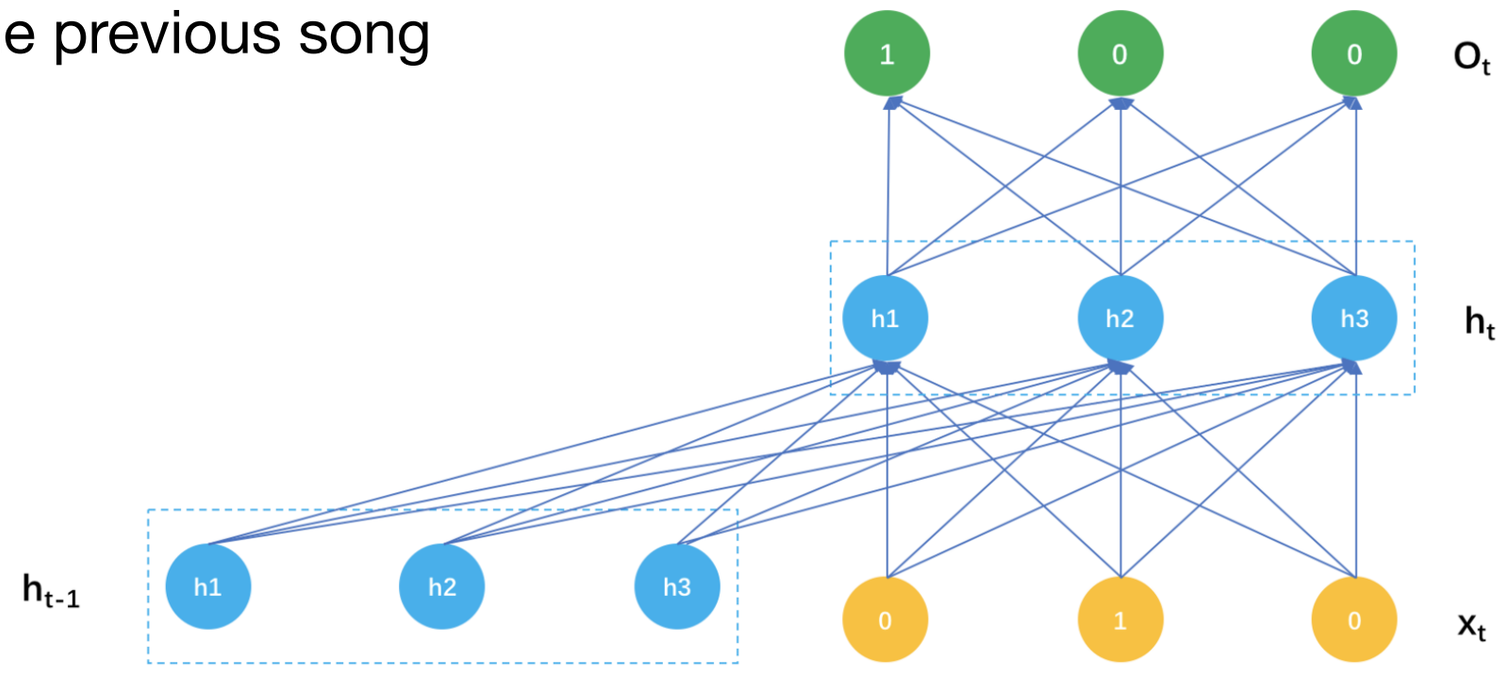
- RNN

Spotify—Music Recommendation

- The generation of music broadcasts is regarded as the generation of time series
- Not only affected by the characteristics, but also affected by the previous song

Cons :

- Hard to train
- Cold start



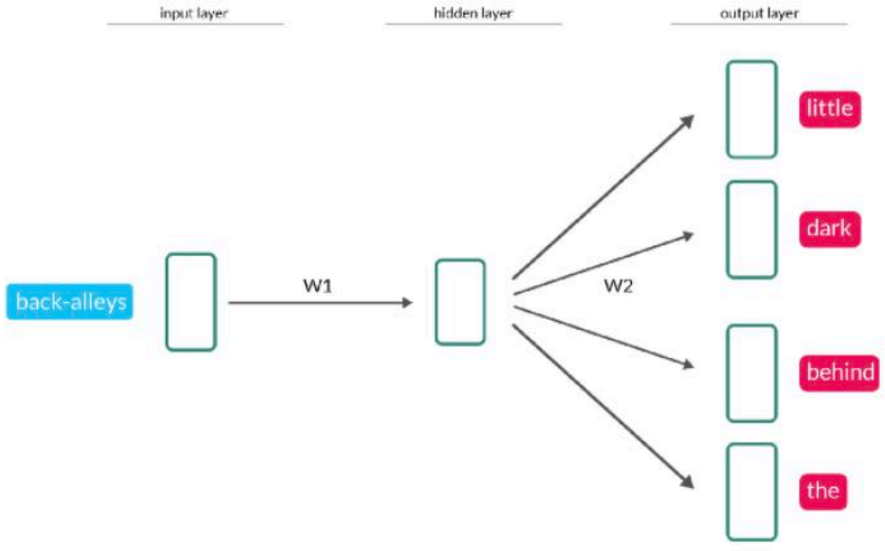
2. Recommendation Alg

- Word2Vec
  - **Music recommendations at Anghami**
  - **Listing recommendations at Airbnb**
  - **Product recommendations in Yahoo Mail**

- **Word2Vec : For each word, generate a set of vectors as its encoding in the context**
  - Similar with CNN
  - Using every word in the window to predict the word in the center of the window ; or vice versa
  - Moving the window ,updating W1,W2

  - If two different words appear in the same context, we expect the output to be similar == having similar matrix

- Using weight as encoding



E-commerce in Your Inbox: Product Recommendations at Scale， Yahoo —2015， KDD

3.Processing

- **Analogy to the problem of discovering the latent topics of a document.**

Vocabulary	⟷	**Transition cuboid**
Documents	⟷	**Products**
Topics	⟷	**Latent features**
Words	⟷	**Browsing patterns**

3. Processing

物品编码，输入每个人的访问序列 a1=(r1,r2,...rn)

↓ LDA k-topics

每个物品对应一个k维隐因子向量(q1,q2...qk)

↓ K-means

物品聚类 m类

再逐一计算用户和每个聚类中心的推荐分数，给用户推荐物品聚类

将个人访问序列 映射为-->访问主题序列 （c1,c2...cm）

加和平均—用户兴趣特点

频繁序列挖掘

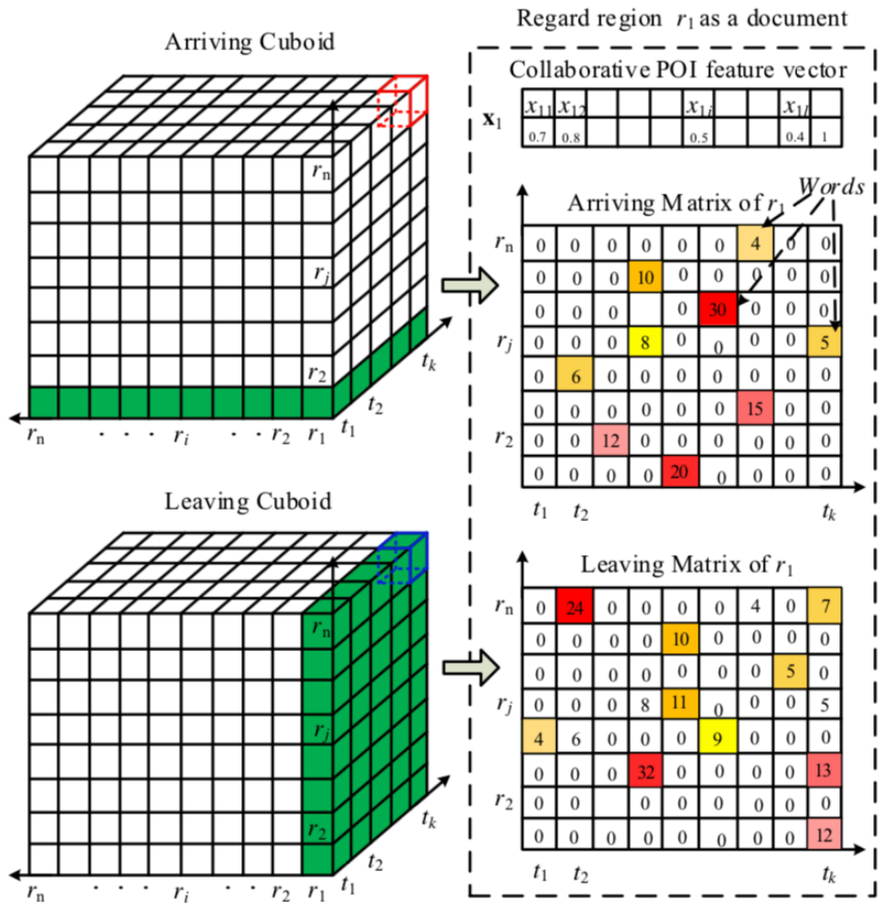
深度学习

按照聚类中 rank 推荐物品

推荐的物品聚类

T-SNF 压缩二维可视

EE



4. Baselines

- **Matrix factorization** → **Lightfm**
- **Word2Vec** → **Gensim**, **Tensorflow-word2vec(Google)**, **FastText(Facebook)**
- **LDA(without time series)** → **Gensim**, **Tensorflow-word2vec(Google)**, **FastText(Facebook)**
- **RNN(TCN)** → **Tensorflow**

Python Multithreading

5. Problems

1. 数据集

有时间的访问序列+ 数据量

2. LDA 输入

3. LDA 其他应用

微博浏览推荐

视频分类

6. LDA 详解

1. LDA想法

- 假设： 一篇文档可以包含多个主题，文档中每一个词都由其中的一个主题生成。

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

如何生成文档?

1.一定的概率选取上述某个主题

2.以一定的概率选取那个主题下的某个单词

3. 重复1-2步

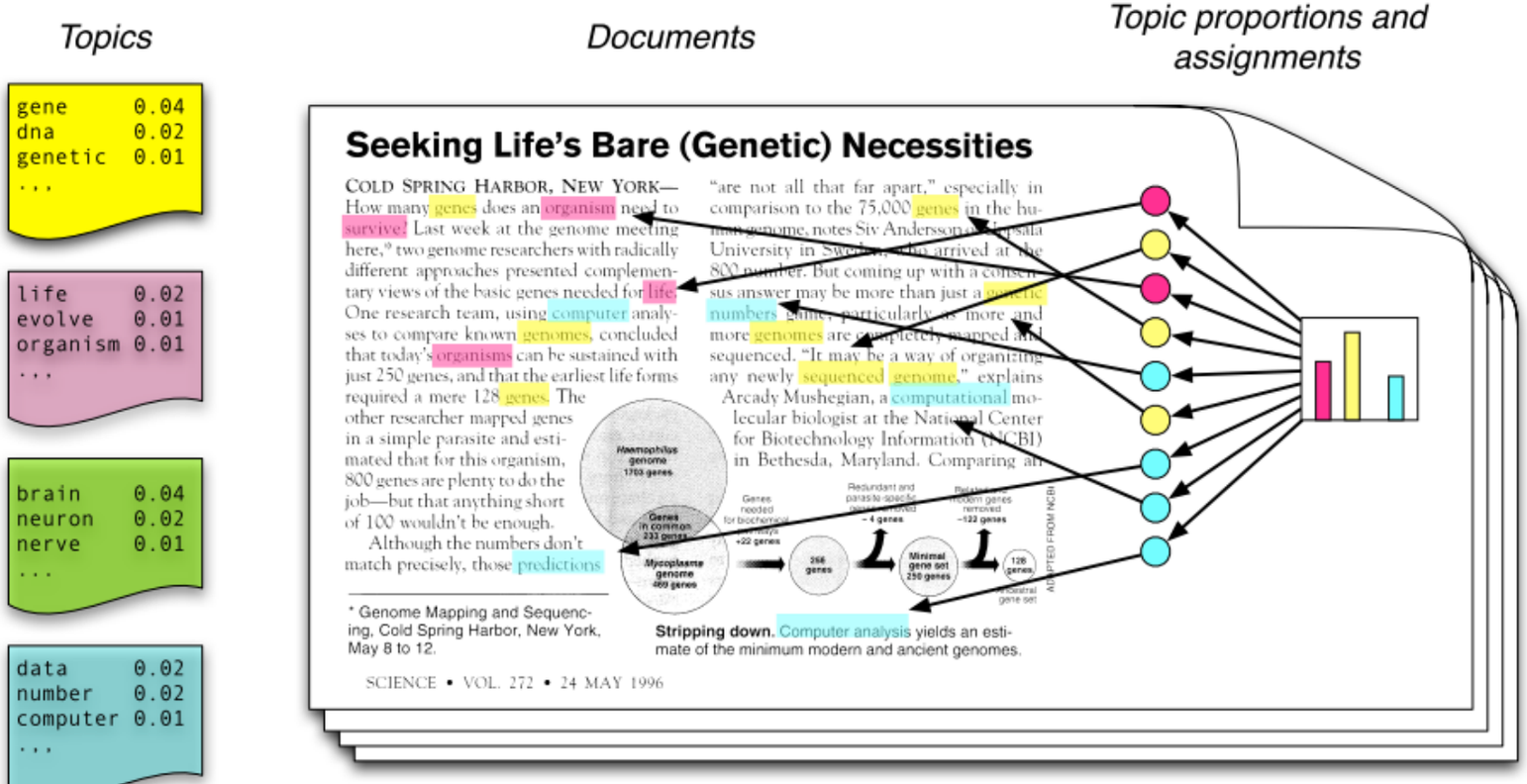
The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Latent Dirichlet Allocation —— David M. Blei, Andrew Y. Ng, Michael I. Jordan; 3(Jan):993–1022, 2003.

6. LDA 详解

1. LDA想法

根据给定的一篇文档，反推其主题分布

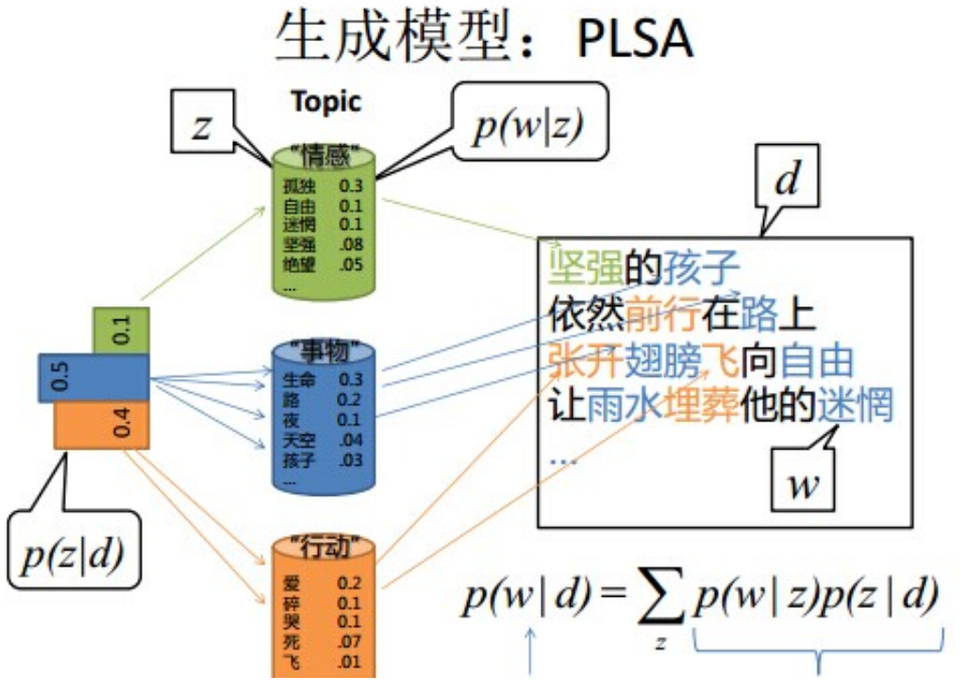


主题分布是隐藏的

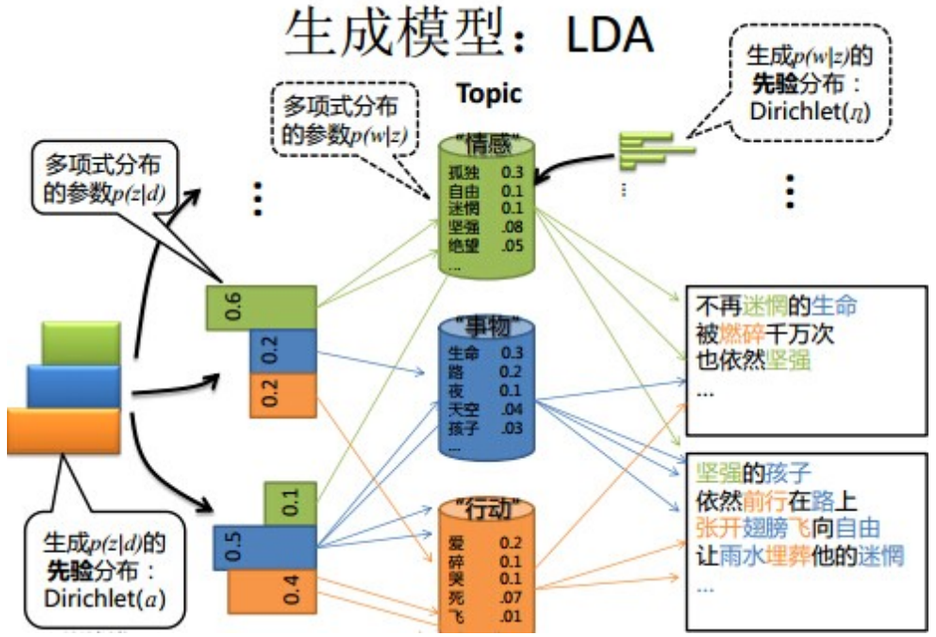
Introduction to Probabilistic Topic Models---David M. Blei

6. LDA 详解

1. LDA想法



概率派



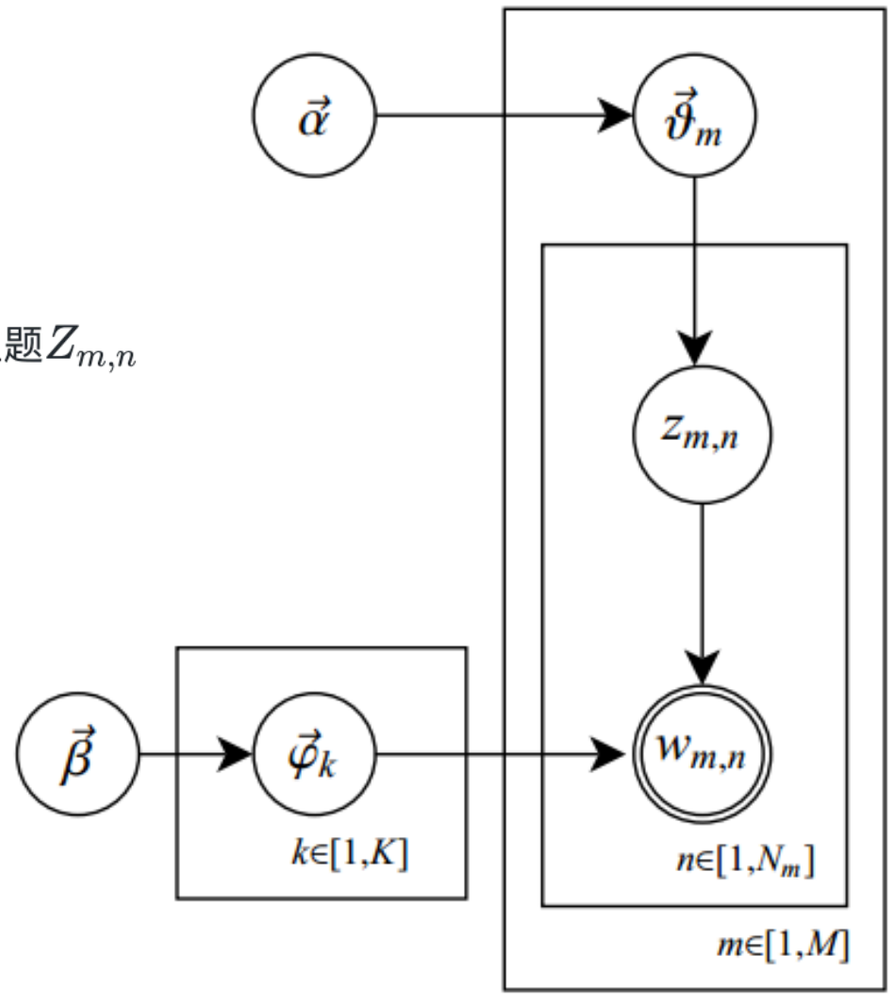
贝叶斯

任何一个未知量都可以看作随机变量，使用概率分布描述，分布为先验分布

6. LDA 详解

2. LDA 生成文档

- 从 Dirichlet 分布 α 中取样生成文档m 的 主题分布 θ_m
- 从主题的 Multinominal 分布 θ_m 中取样生成文档 m 的第 n 个词的主题$Z_{m,n}$
- 从 Dirichlet 分布 β 中取样生成主题$Z_{m,n}$ 的对应的词语分布 $\phi_{z_{m,n}}$
- 从词语的多项式分布$\phi_{z_{m,n}}$中采样最终生成词语$W_{m,n}$



单圆圈表示隐变量；双圆圈表示观察到的变量；把节点用方框(plate)圈起来，表示其中的节点有多种选择。plate notation —— PRML 8.0 Graphical Models

Introduction to Probabilistic Topic Models---David M. Blei

详解：

- **图中 K 为主题个数，M 为文档总数，N_m 是第 m 个文档的单词总数**
- **α 是每个文档下 Topic 的多项分布的 Dirichlet 先验参数，β 是每个 Topic 下词的多项分布的 Dirichlet 先验参数**
- **$Z_{m,n}$ 是第 m 个文档中第 n 个词的主题，$W_{m,n}$ 是 m 个文档中的第 n 个词**
- **θ(k维) 表示第 m 个文档下的 Topic 分布**
  φ(v维) 表示第 k个 Topic 下词的分布

6. LDA 详解

3. LDA 参数确定——Gibbs Sampling 算法

- 给定一个文档集合，W 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，其他的变量z，θ和φ都是未知的隐含变量，需要根据观察到的变量来学习估计的

- Gibbs Sampling 算法的运行方式是每次选取概率向量的一个维度，给定其他维度的变量值采样当前维度的值。不断迭代，直到收敛输出待估计的参数

- 初始时随机给文本中的每个单词分配主题$z^{(0)}$，然后统计每个主题 z 下出现词的数量以及每个文档 m 下出现主题 z 中的词的数量，每一轮计算$p(z_i|z_{-i},d,w)$，即排除当前词的主题分配：根据其他所有词的主题分配估计当前词分配各个主题的概率。当得到当前词属于所有主题z的概率分布后，根据这个概率分布为该词采样一个新的主题。然后用同样的方法不断更新下一个词的主题，直到发现每个文档下 Topic分布 θ 和每个 Topic下词的分布 φ 收敛，算法停止，输出待估计的参数θ和φ，最终每个单词的主题z也同时得出

Input : word vector *w* ; *hyper parameter α and β; topic number K*

Output : topic association z, multinomial θ and φ

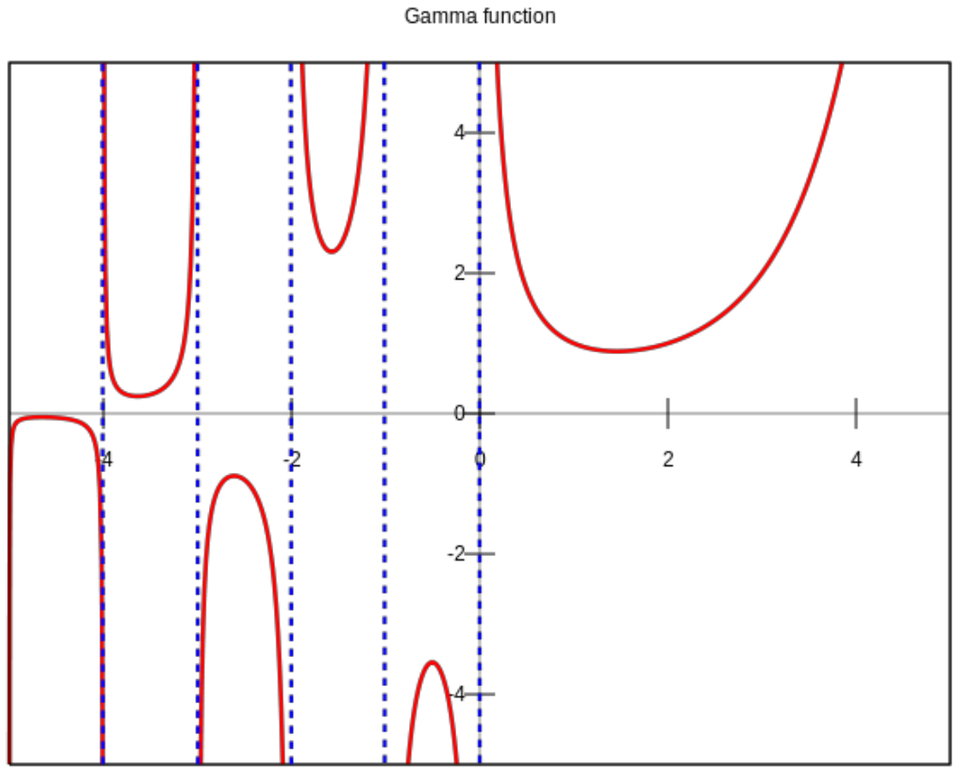
6. LDA 详解

4. LDA—— 数学原理

1-1.Gamma 函数

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

$$\Gamma(n) = (n-1)!$$



- Gamma 函数将导数的定义拓展到实数

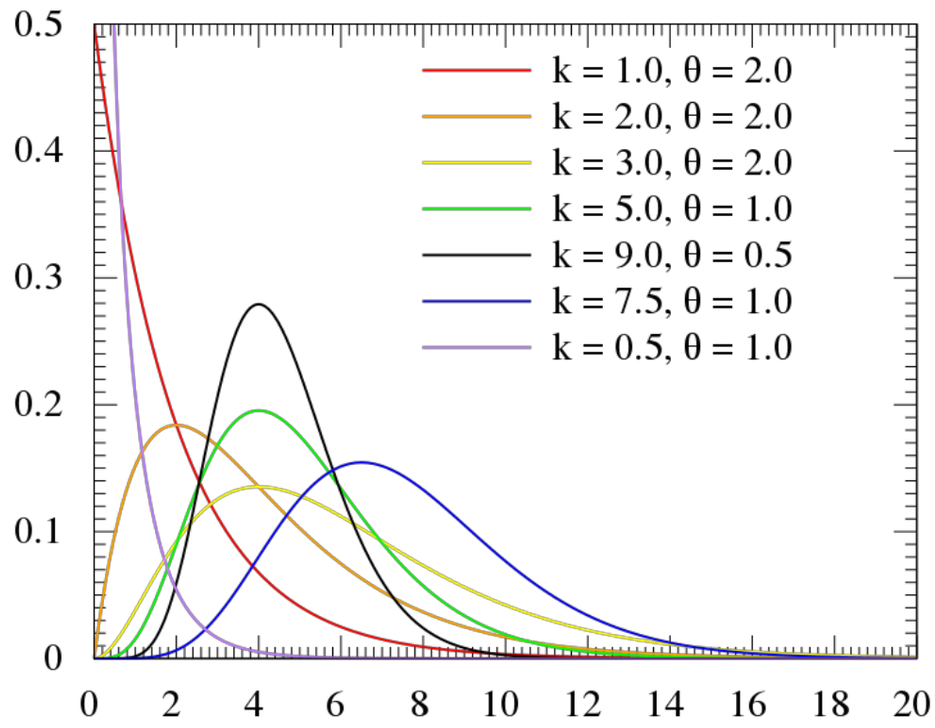
- Gamma 函数是凸函数，log gamma也是凸函数

6. LDA 详解

4. LDA—— 数学原理

1-2. Gamma 分布

$$Gamma(t|\alpha,\beta) = \frac{\beta^{\alpha}t^{\alpha-1}e^{-\beta t}}{\Gamma(\alpha)}$$



- **Gamma 分布是很多分布的先验概率 分布**

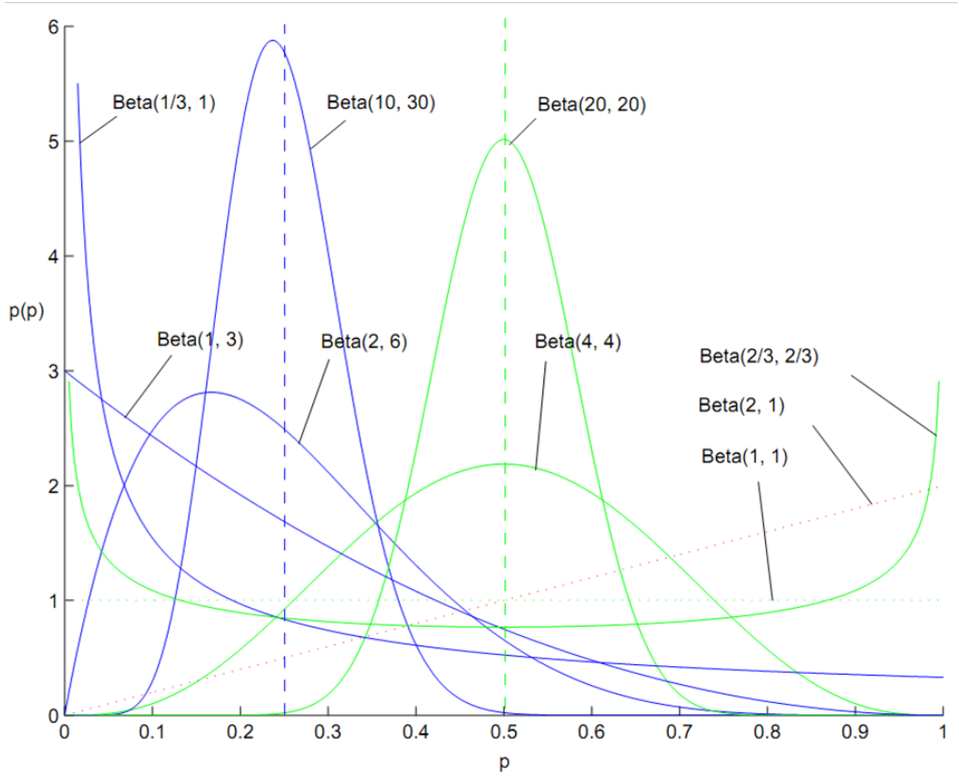
poisson分布，正态分布

6. LDA 详解

4. LDA—— 数学原理

2. Beta 分布

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$



Beta-Binoimial

$$Beta(p|\alpha,\beta) + BinomCount(m_1,m_2) = Beta(p|\alpha+m_1,\beta+m_2)$$

先验概率 + 样本信息 = 后验概率

- **新观察到的样本信息将修正人们以前对事物的认知**
- **如果二项分布的参数 p 的先验分布是 Beta 分布，那么以p为参数的二项分布用贝叶斯估计得到的后验分布仍然服从 Beta 分布。**

PS：叶斯概率理论中，如果后验概率*P(θ|x)*和先验概率*p(θ)*满足同样的分布律，那么，先验分布和后验分布被叫做共轭分布，同时，先验分布叫做似然函数的共轭先验分布

6. LDA 详解

4. LDA—— 数学原理

2. Beta 分布 — 例子

- 投掷一个非均匀硬币，可以使用参数为θ的伯努利模型，θ为硬币为正面的概率，那么结果x的分布形式为

$$P(x|\theta) = \theta^x \cdot (1-\theta)^{1-x}$$

- 其共轭先验为beta分布，具有两个参数和，称为超参数（hyperparameters）。
  且这两个参数决定了θ参数，其Beta分布形式为

$$P(\theta|\alpha,\beta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta}$$

- 计算后验概率

$$\begin{aligned} & P(\theta|x) \\ & \propto P(x|\theta)\cdot P(\theta) \\ & \propto \left(\theta^x(1-\theta)^x\right)\left(\theta^{\alpha-1}(1-\theta)^{\beta-1}\right) \\ & = \theta^{x+a-1}(1-\theta)^{1-x+\beta-1} \end{aligned}$$

- 后验概率依然是 beta 分布

6. LDA 详解

4. LDA—— 数学原理

3-1. Dirichlet 分布

- **3维 dirichlet 分布**

$$f(x_1, x_2, x_3) = \frac{\Gamma(\alpha_1 + \alpha_2 + \alpha_3)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)} x_1^{\alpha_1 - 1} x_2^{\alpha_2 - 1} x_3^{\alpha_3 - 1}$$

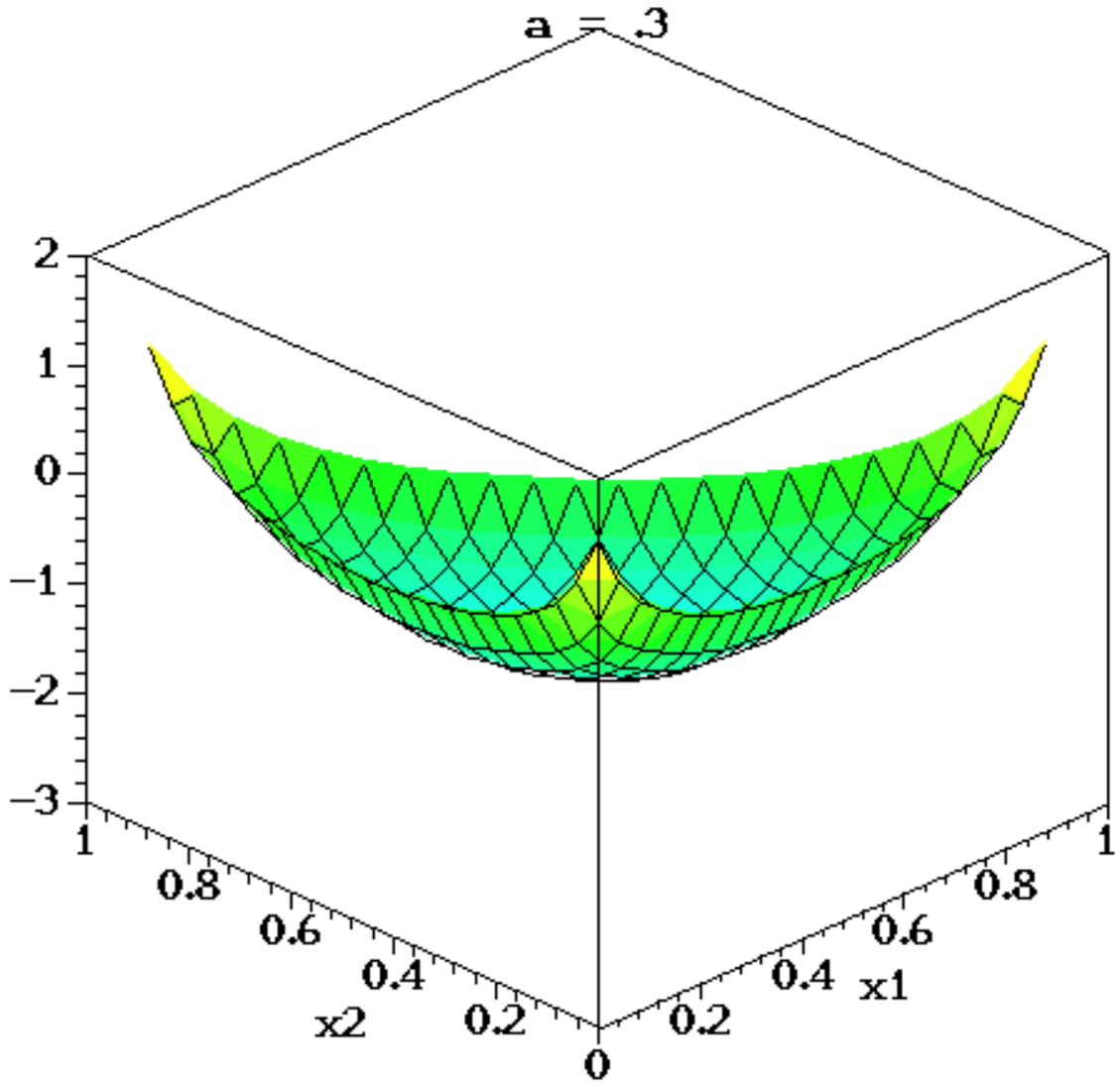
$$f(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha - 1}(1 - x)^{\beta - 1}$$

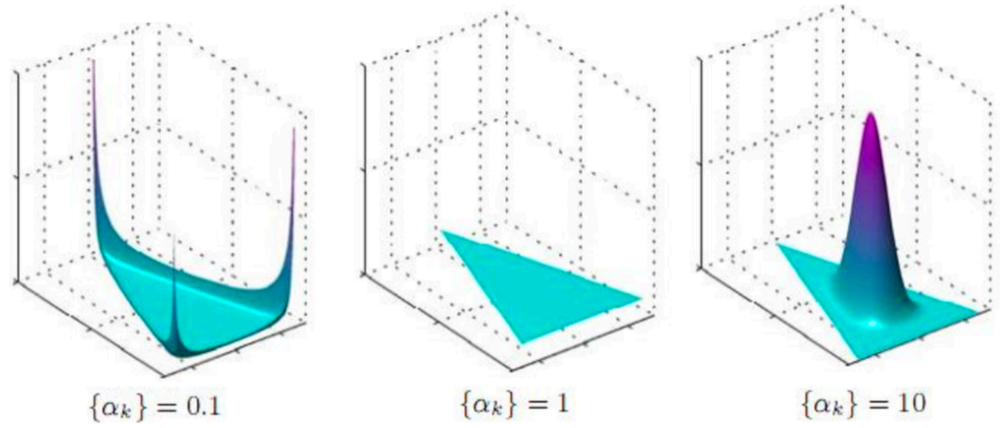
- **看出： 形式上是 beta 函数的高纬度拓展**

When α=1, the symmetric Dirichlet distribution is equivalent to a uniform distribution over the open standard (K – 1)-simplex, i.e. it is uniform over all points in its support.

Values of the concentration parameter above 1 prefer variates that are dense, evenly distributed distributions, i.e. all the values within a single sample are similar to each other.

Values of the concentration parameter below 1 prefer sparse distributions, i.e. most of the values within a single sample will be close to 0, and the vast majority of the mass will be concentrated in a few of the values.





From Wikipedia

6. LDA 详解

4. LDA—— 数学原理

3-2. Dirichlet 分布——Beta分布的推广

- 对于 Beta 分布期望

$$E(p) = \frac{\alpha}{\alpha + \beta}$$

- Dirichlet 分布期望

$$E(\vec{p}) = \left(\frac{\alpha_1}{\sum_{i=1}^{K} \alpha_i}, \frac{\alpha_2}{\sum_{i=1}^{K} \alpha_i}, \cdots, \frac{\alpha_K}{\sum_{i=1}^{K} \alpha_i} \right)$$

Dirichlet - Multinoimail 共轭

$$Dir(\vec{p}|\vec{\alpha}) + MultCount(\vec{m}) = Dir(\vec{p}|\vec{\alpha} + \vec{m})$$

- PS： A very common special case is the symmetric Dirichlet distribution, where all of the elements making up the parameter vector have the same value. Symmetric Dirichlet distributions are often used when a Dirichlet prior is called for, since there typically is no prior knowledge favoring one component over another. Since all elements of the parameter vector have the same value, the distribution alternatively can be parametrized by a single scalar value α, called the concentration parameter

6. LDA 详解

4. LDA—— 数学原理

4-1.马尔科夫链蒙特卡洛（MCMC）方法

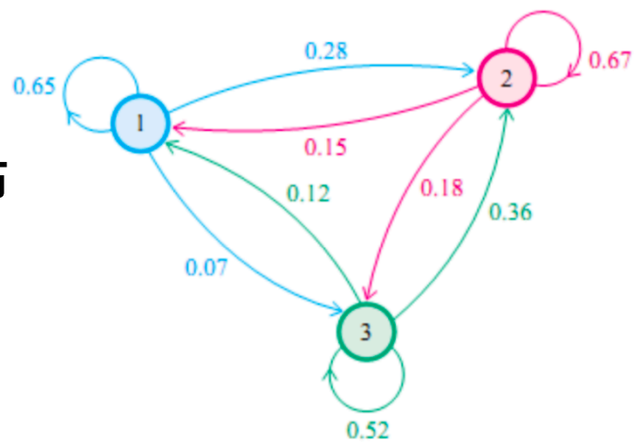
- 马尔可夫链—— 状态只取决于前一个状态

$$P(X_{t+1} = x|X_t, X_{t-1}, \cdots) = P(X_{t+1} = x|X_t)$$

例如：

- 各阶层收入转移概率

		子代		
	State	1	2	3
父代	1	0.65	0.28	0.07
	2	0.15	0.67	0.18
	3	0.12	0.36	0.52



- 经计算，n足够大，马尔可夫链收敛，收敛到平稳分布

6. LDA 详解

4. LDA—— 数学原理

4-1.马尔科夫链蒙特卡洛（MCMC）方法

- 给定概率分布p(x)，生成对应的样本
- 关键问题是构造转移矩阵 P，使得平稳分布恰好是我们要的分布p(x)

- 利用细致平稳条件

$$\pi(i)P_{ij} = \pi(j)P_{ji} \qquad \text{for all} \quad i, j$$

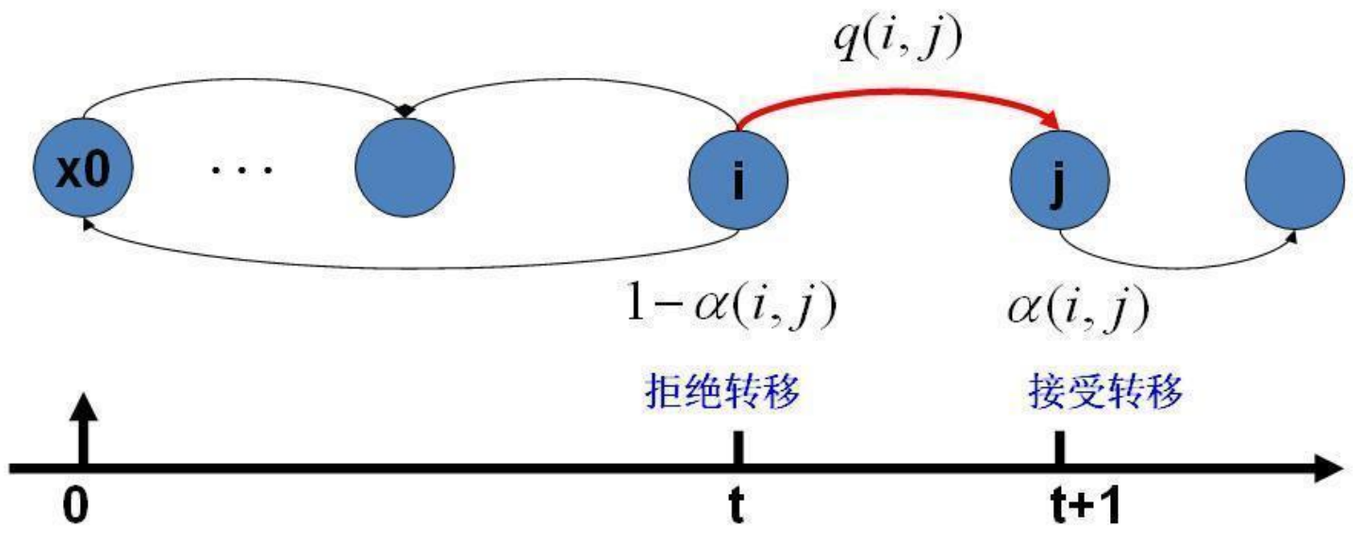
- 一般情况不满足

$$p(i)q(i,j) \neq p(j)q(j,i)$$

  - 引入接受率

$$p(i)q(i,j)\alpha(i,j) = p(j)q(j,i)\alpha(j,i)$$

$$\alpha(i,j) = p(j)q(j,i) \quad \alpha(j,i) = p(i)q(i,j)$$



6. LDA 详解

4. LDA—— 数学原理

4-2.Metropolis-Hastings 算法

- 解决接受率 α(i,j) 可能偏小问题

- α(i,j),α(j,i) 同比例放大，使得两数中最大的一个放大到1，提高采样中的跳转接受率

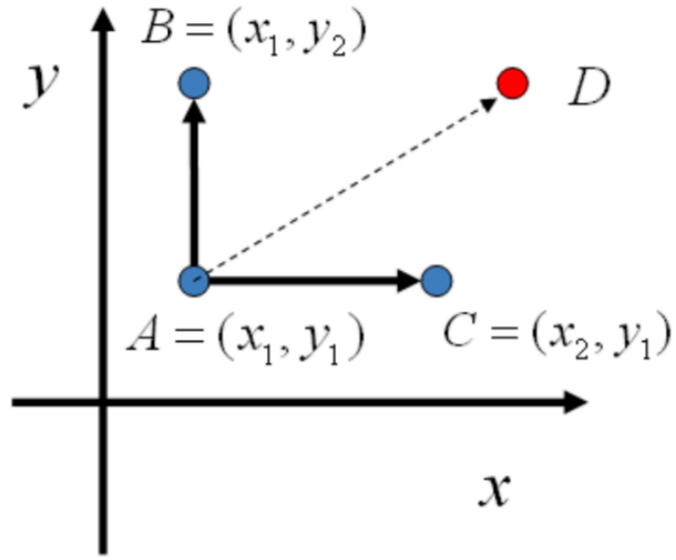
$$\alpha(i,j) = \min\left\{\frac{p(j)q(j,i)}{p(i)q(i,j)}, 1\right\}$$

6. LDA 详解

4. LDA——数学原理

4-3. Gibbs Sampling 算法

- 解决高维效率不高问题



$$Q(A \to B) = p(y_B|x_1) \qquad \text{如果} \quad x_A = x_B = x_1$$

$$Q(A \to C) = p(x_C|y_1) \qquad \text{如果} \quad y_A = y_C = y_1$$

$$Q(A \to D) = 0 \qquad \text{其它}$$

- 1. 二维平面——右图

在 x=x1 这条平行于 y 轴的直线上，如果使用条件分布 p(y|x1) 做为任何两个点之间的转移概率，那么任何两个点之间的转移满足细致平稳条件

2. 高维

- 如果当前状态为 (x1,x2,⋯,xn)，马氏链转移的过程中，只能沿着坐标轴做转移。沿着 xi 这根坐标轴做转移的时候，转移概率由条件概率 p(xi|x1,⋯,xi−1,xi+1,⋯,xn) 定义

- 其它无法沿着单根坐标轴进行的跳转，转移概率都设置为 0