(2003, Blei, Andrew)

主题模型（Topic Models）

2019.08.11

主题模型

MLE, MAP, Bayesian

- 主题模型属于贝叶斯模型(Fully Bayesian)
- 主题模型是生成模型，以无监督的方式来学习
- 主题模型属于Mixed Membership模型 GMM, MMSB,
- 对于主题模型的推导需要近似算法如MCMC Variational Infer

AE SGD

主题模型概况

输入

Latent Topic

Hyperparameter

文章1

文章2

文章3

LDA

$k = 4, 5, 50, 100$

文章一主题

文章1

文章2

文章3

体 AI 艺 IT

主题—单词分布

	1	2	3	4
apple	0.01	0.001	0.001	0.002
⋮				
soccer	0.1	0.005	0.001	0.001
⋮				
AI	0.01	0.2	0.01	0.05
⋮				
game	0.02	0.01	0.05	0.01

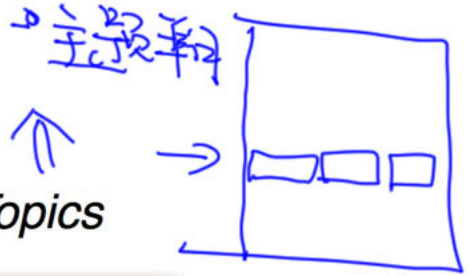
$|V| \times k$

$\phi_k \sim Dir(\beta)$

$\theta_d \sim Dir(\alpha)$

$\Rightarrow |V|$

主题模型介绍



Topics

Documents

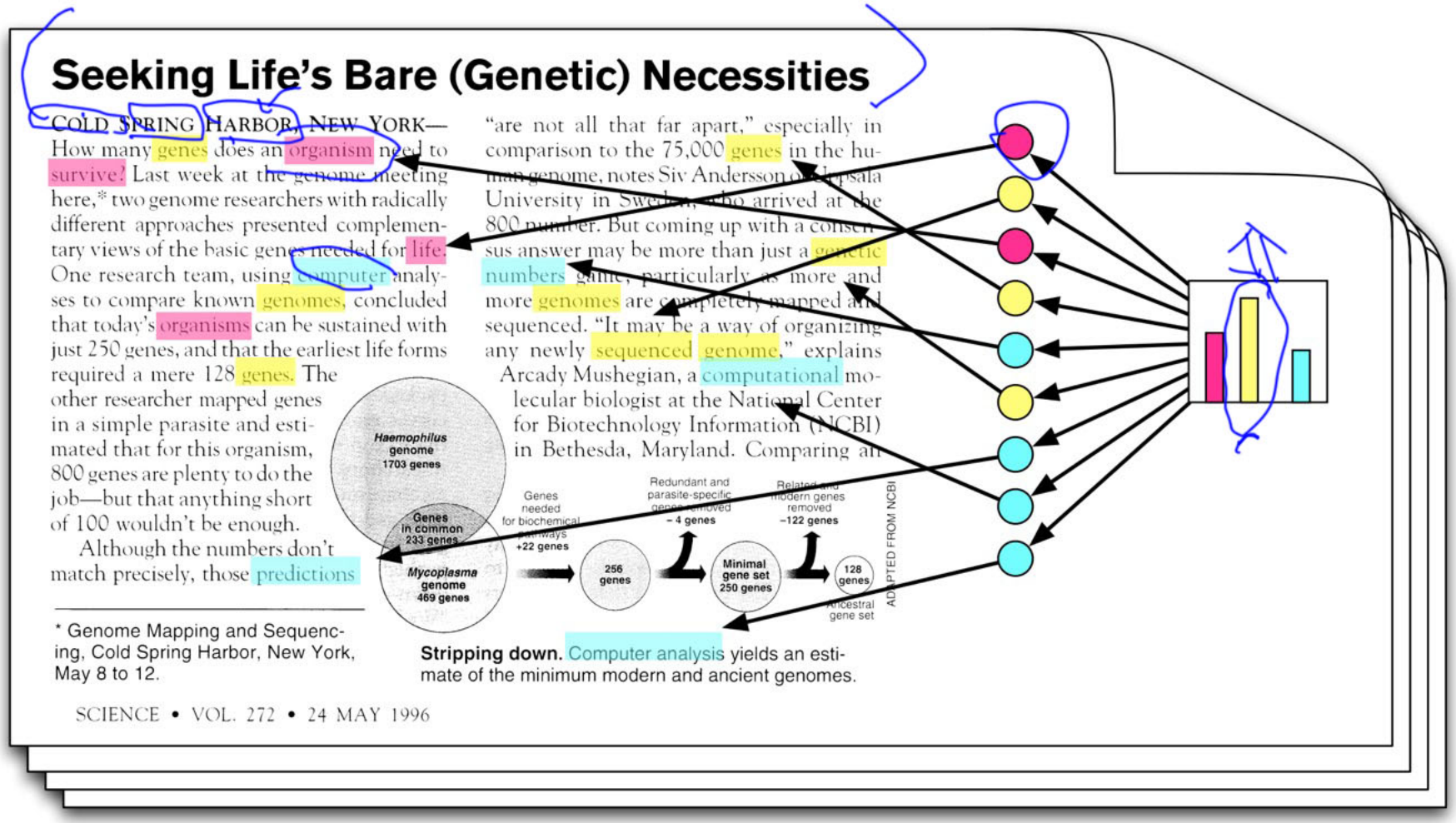
Topic proportions and assignments

```
gene      0.04
dna       0.02
genetic   0.01
...
```

```
life      0.02
evolve    0.01
organism  0.01
...
```

```
brain     0.04
neuron    0.02
nerve     0.01
...
```

```
data      0.02
number    0.02
computer  0.01
...
```



Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

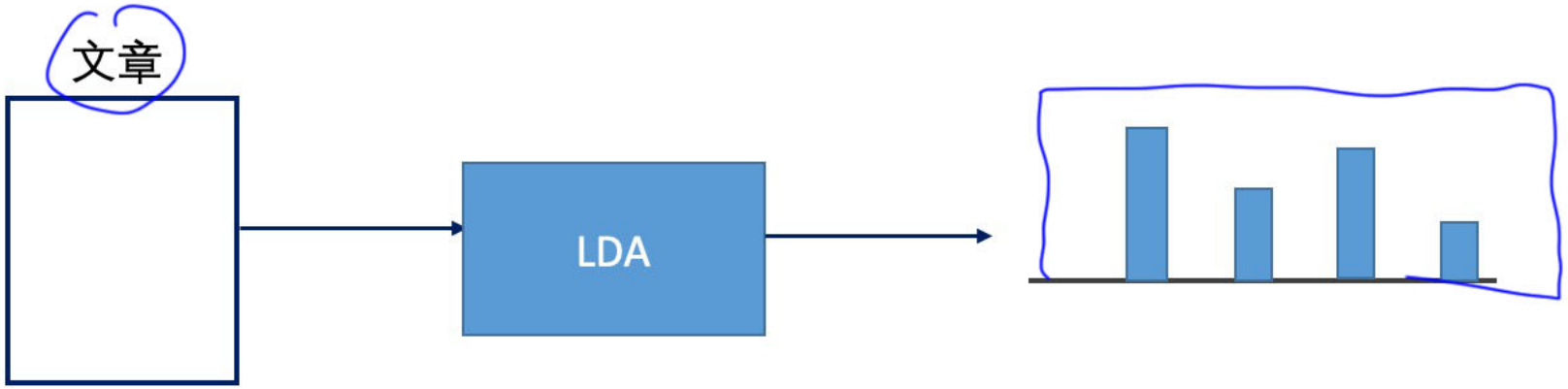
Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

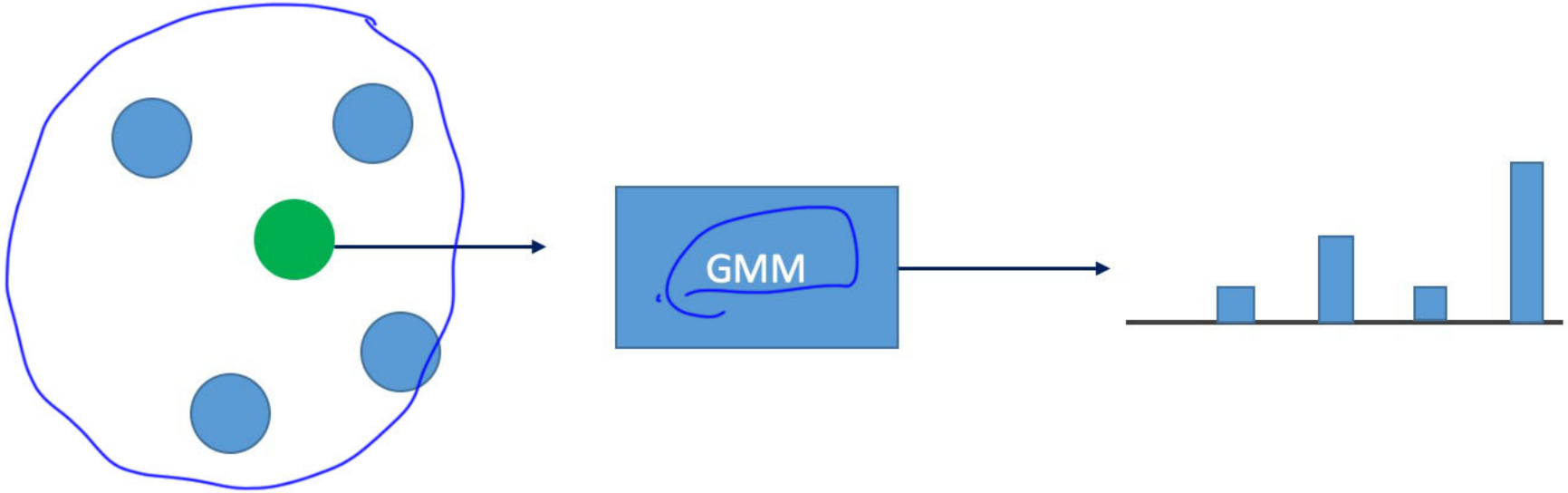
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

Mixed Membership Models





K-means = *

k=4

GMM ⟷ k-means

多个 一

EM算法

问题： 朴素贝叶斯是否属于这一类?

MLE vs MAP vs Bayesian

当样本 $+\infty$

Frequentist　　Bayesian

MLE　　**MAP**　　**Bayesian**

Likelihood　　后验概率

参数估计：$\boxed{\theta^*} = \arg\max_\theta P(D|\theta)$

$\boxed{\theta^*} = \arg\max_\theta P(\theta|D)$

$= \arg\max_\theta P(D|\theta)\cdot P(\theta) / \cancel{P(D)}$

$= \arg\max_\theta \underbrace{P(D|\theta)}_{\text{Likelihood}} \cdot \underbrace{P(\theta)}_{\text{prior}}$

正则

预测：给定新的数据 x'

$y' \sim P(y|x', \theta)$

预测：给定 x'　　后验概率

$P(y|x', D) =$

$\int_\theta P(y'|x', \theta) \cdot \boxed{P(\theta|D)} d\theta$

θ空间 $\in R^p$

θ_1: 0.01	θ_4: 0.1	$+\infty$
θ_2: 0.02	θ_5: 0.03	
θ_3: 0.05	θ_6: 0.01	

θ_1　$P(y'|x', \theta_1) \cdot 0.01$

θ_2　$P(y'|x', \theta_2) \cdot 0.02$

θ_3　$P(y'|x', \theta_3) \cdot 0.05$

$\vdots$

$+\infty$　$P(y'|x', \theta_{+\infty}) \cdot 0.0001$

$\oplus \Rightarrow$ 预测值

估计 $\boxed{一个}$ 最好的参数

贝叶斯模型

贝叶斯模型也可以看作是集成模型，而且一般集成了无穷多的模型。

贝叶斯模型的预测

CRF → 车

预测 $\int p(y'|x',\theta)\,p(\theta|D)\,d\theta$

问题：如何计算$p(\theta|D)$？

不通的 ⟹ $p(\theta|D)=\dfrac{P(D|\theta)\cdot p(\theta)}{P(D)} = \dfrac{P(D|\theta)\cdot p(\theta)}{\int_\theta P(D,\theta)\,d\theta} = \dfrac{P(D|\theta)\,p(\theta)}{\int_{\theta_1}\int_{\theta_2}\int_{\theta_3}\cdots\int_{\theta_{1000}} P(D,\theta)\,d\theta}$ ⊗

$P(D)$ ← 边(Marginal)

Exact Inference

$\int p(y'|x',\theta)\,p(\theta|D)\,d\theta$

Approximate Inference

$\approx \frac{1}{S}\sum_{s=1}^{S} P(y'|x',\theta_s)$ $\quad \theta_s \sim P(\theta|D)$

$\theta_1 \cdot \theta_1\,\theta_2\,\theta_1 \cdots \theta_2$

10

$\theta_1 - 0.8$

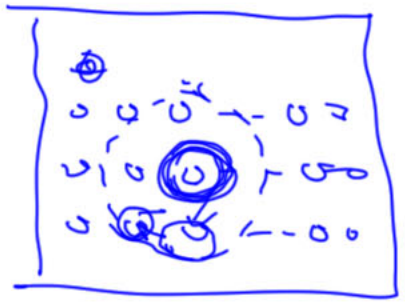
$\theta_2 - 0.1$

θ

$\left(P(y'|x',\theta_1)+P(y'|x',\theta_2)+P(y'|x',\theta_3)+P(y'|x',\theta_4)\right)/4 \Rightarrow$

蒙特卡洛采样

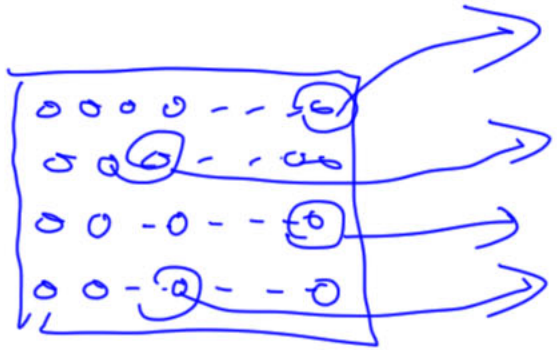
$$\int p(y'|x',\theta)p(\theta|D)d\theta \approx \frac{1}{S}\sum_{s=1}^{S} p(y'|x',\theta^s) \quad \boxed{\theta^s \sim p(\theta|D)}$$



MC

1. Independent Sampling

~~Monte Carlo Sampling~~



for $t=1$ ...

$\theta_t \sim p(\theta_t|D)$

2. Sequential Sampling

Markov Chain Monte Carlo Sampling (MCMC)

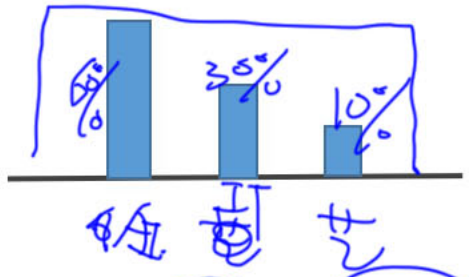
MCMC

- Metropolis Hasting
- Gibbs Sampling
- Collapsed Gibbs Sampling
- Importance Sampling
- Rejection Sampling
- Langevin Dynamic

SGLD ← SGD

从生成的角度来看LDA

步骤1： 指定文章的主题



步骤2： 根据主题采样单词

	1	2	3
apple			
is			
the	0.001		
AI	0.05		
technique			
game			
key	ϕ_1	ϕ_2	ϕ_3
soccer			

主题 单词

AI is the key techniques

文章

for $i=1 \dots N_d$ N_d: # of words in d

主题$_i$ ← (0.6, 0.3, 0.1) → w_i ∈ AI

单词$_i$ ← $\phi_{主题_i}$

PGM (Probabilistic Graphical Models)

LDA的生成过程

w_{ij}: 文档i的第j个单词

K: 主题个数

N: 文档个数

N_i: 文档i中单词的个数

θ_i: 文档i的主题分布

z_{ij} 文档i的第j个单词的主题

θ 文档主题: $\theta_i \in R^K$

主题-单词: $\phi_k \in R^{|V|}$

全局

生成过程:

for k=1 ... K

sample ϕ_k

for i=1 ... N

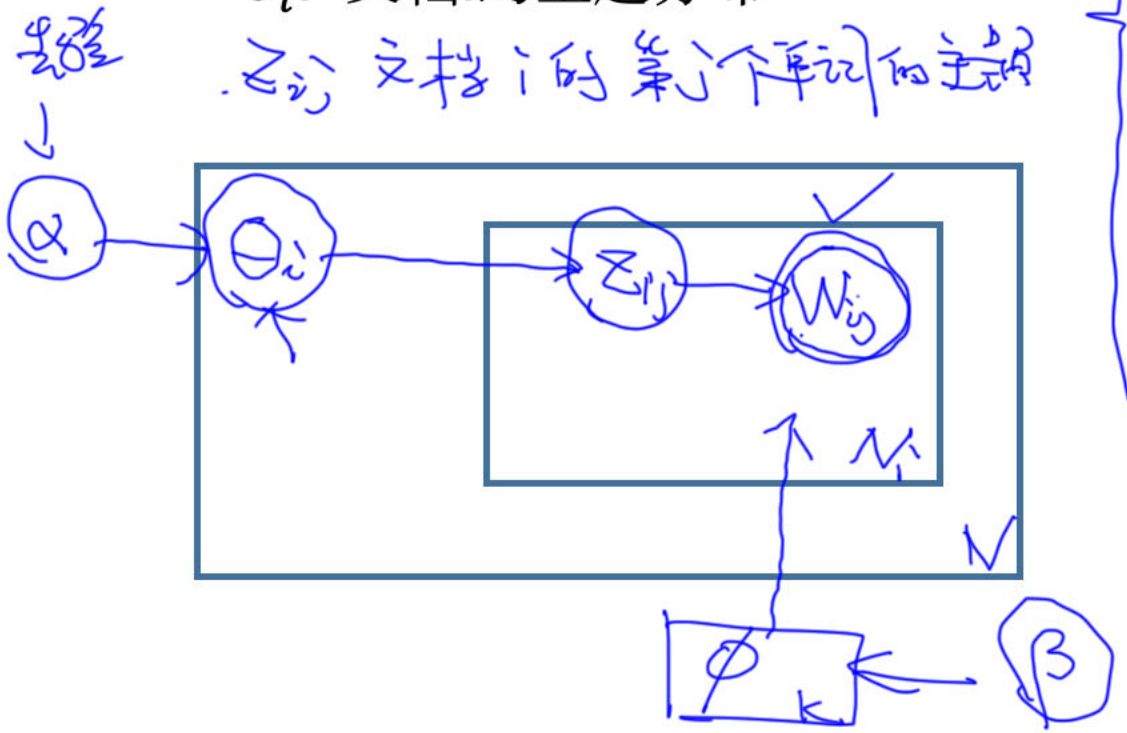
$\theta_i \leftarrow$ Sample(α)

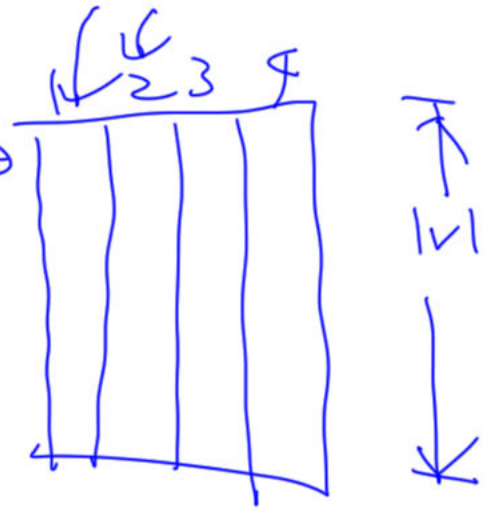
for j=1 ... N_i

$z_{ij} \sim$ Sample(θ_i)

$w_{ij} \sim$ Sample($\phi_{z_{ij}}$)

$\{\phi, \theta\}$





近似算法

Gibbs Sampler

Metropolish Hasting

Collapsed Gibbs Sampler

Variational Inference

Langevin Dynamics

Importance Sampling

……

生成文档 i

① 文档主题 θ_i

$\theta_i \sim \text{sample}(-) \Rightarrow \theta_i = (0.7, 0.2, 0.1)$

② 文档主题—单词分布

	AI	IT	艺	体
AI	0.4	0.2	0.05	0.05
技术	0.2	0.5	0.05	0.05
机器	0.3	0.1	0.1	0.1
唱	0.05	0.05	0.7	0.1
球	0.05	0.05	0.1	0.7

ϕ

$\phi_1 \sim \text{sample}(z)$

$\phi_2 \sim \text{sample}(\sim)$

$\phi_3 \sim \text{sample}(x$

for $w_{ij} = 1 \ldots N_i$ (生成多少个单词)

$z_{ij} \leftarrow$ 主题(第 j 个单词)

$z_{i1} \sim (0.7, 0.2, 0.1) \Rightarrow z_{i1} = 1 \Leftarrow$ AI

$w_{i1} \sim (0.4, 0.2, 0.3, 0.05, 0.05)$

$z_{i2} \sim (0.7, 0.2, 0.1) \Rightarrow z_{i2} = 2$ IT

$w_{i2} \sim (0.2, 0.5, 0.2, 0.05, 0.05)$

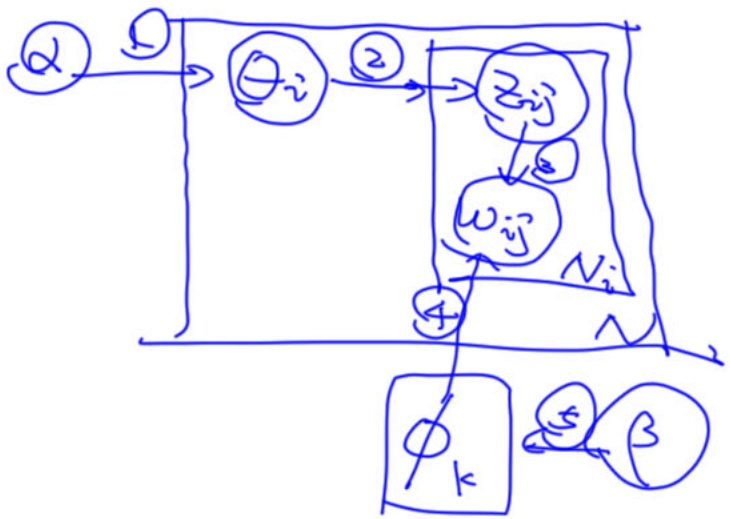
$z_{i3} \sim (0.7, 0.2, 0.1) \Rightarrow z_{i3} = 1 \Leftarrow$ AI

$w_{i3} \sim (0.4, 0.2, 0.3, 0.05, 0.05)$

$D = \{(x_1, y_1), \ldots\}$

θ

点 $\xrightarrow{w_{i1}}$ 技术 $\xrightarrow{w_{i2}}$ 机器 $\xrightarrow{w_{i3}}$ ---



① $\alpha \rightarrow \theta_i$ $\sum = 1$

→ simplex

$\theta_i \sim \beta(\alpha)$

Dirichlet Distribution

$\theta_i \sim Dir(\alpha)$

--

⑤ for $k=1 \dots K$

$\phi_k \sim sample(\beta)$

→ 相加之和等于1

$\sum = 1$

$\phi_k \sim Dir(\beta)$

② $z_{ij} \leftarrow \theta_i$

$(0.7, 0.2, 0.1)$

①

$z_{ij} \sim Multinomial(\theta_i)$

④ ③→④

$w_{ij} \sim Multinomial(\phi_{z_{ij}})$

for $k=1 \dots K$

$\phi_k \sim Dir(\beta)$

for $i=1 \dots M$

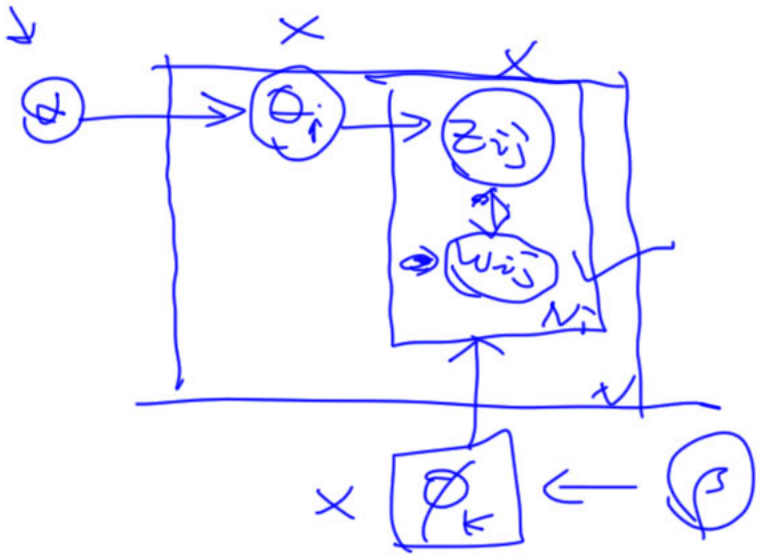
$\theta_i \sim Dir(\alpha)$

for $j=1 \dots N_i$

$z_{ij} \sim Multinomial(\theta_i)$

$w_{ij} \sim Multinomial(\phi_{z_{ij}})$

z_{ij}



超参数：$\{\alpha, \beta\}$

模型参数：$\{\theta, \phi\}$ ⇐

Latent variable：$\{z\}$ ⇐

θ, ϕ, z

$p(\theta \mid D)$ → 后验

$$\underbrace{p(\theta, \phi, z \mid D, \alpha, \beta)}_{\text{后验概率}}$$

MCMC

variable variational Inference

Gibbs Sampling ← Coordinate Descent

优化：θ

$\theta_1, \theta_2 \cdots \theta_n$

$$\Rightarrow \theta_i \sim p(\theta_i \mid \underbrace{\alpha, \beta, D, \phi, z}_{\text{已知}})$$

$$z_{ij} \sim p(z_{ij} \mid \alpha, \beta, D, \phi, \theta)$$

$$\phi_k \sim p(\phi_k \mid \alpha, \beta, D, \phi, z)$$

$\theta_1 \ \theta_2 \ \theta_3 \cdots \theta_n$

$\phi_1 \ \phi_2 \cdots \phi_k$

$z_{11}, z_{12} \cdots z_{n1} \cdots z_{nn_j}$

文本 → 提取特征 → 分类器

↓

tf-idf, word2vec, ngram, LDA特征

文本 → LDA → (0.2 0.3 0.3 0.2)

LDA特征：$(\underbrace{0.5, 0.1, 0.05, 0.05, 0.1, 0.1, 0.1}_{k=7\text{维}})$

文本：(tf-idf | Ngram | 词性 | LDA) → 分类器

↓

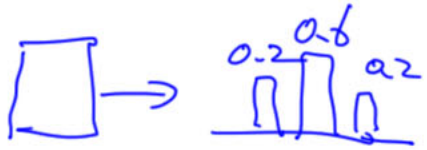
预测

模型: ⇐ Categories

① Frequentist vs Bayesian. (Fully Bayesian)
(MLE, MAP)

② Generative vs Discriminative
(Naive Bayes) (LR, CRF, SVM, NN..)
LDA
HMM, LM, VAE, GAN,

③ Mixted Membership vs 非 Mixed Membership



k-means → hard

soft clustering ← GMM, MMSB

Naive, LSA

DBSCAN
Spectral clustering
Hierarchical clustering

④ Convex vs Non-convex
LR, NN,

为什么平均？：

构造一个模型 ⇒ 表现好的，同时这点能力

过拟合 → 正则
→ + 数据量
→ 选择简单的

模型的不稳定性

统计：方差/偏差

$X_1 \sim N(\mu_1, \sigma_1^2)$

$X_2 \sim N(\mu_2, \sigma_2^2)$

$X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$

分散 ⇒ 多样 ⇒ 稳定

集成模型：

[随机森林
XGBoost] ← # of finite models

Bayesian ← # of infinit models

- Dropout
- Denoising Autoencoder

→ Bert

集成

Bayesian Approximate Inference

- Sampling
  - MCMC
    - Metropolis
    - ⋮
- Optimization
  - Variational Inference
    - Mean-Variational Infer
    - Neural Variational
    - Block Variational Infer
    - Collapsed Variational Infer

问题 → P → ✓

→ NP-hard → Approximate Algorithms