吉布斯采样和概率图模型介绍

钱烽

qf6101 at gmail.com

References

- Resnik, P., Resnik, P., Hardisty, E., & Hardisty, E. (2009). Gibbs Sampling for the Uninitiated. *Umiacs.Umd.Edu*, (June), 1–23.
- Heinrich, G. (2008). Parameter Estimation for Text Analysis.
- Advanced MCMC Methods (http://mlg.eng.cam.ac.uk/zoubin/tutorials06.html)
- Wikipedia

Part I

MCMC and Gibbs Sampling

Why use MCMC or Gibbs sampling？

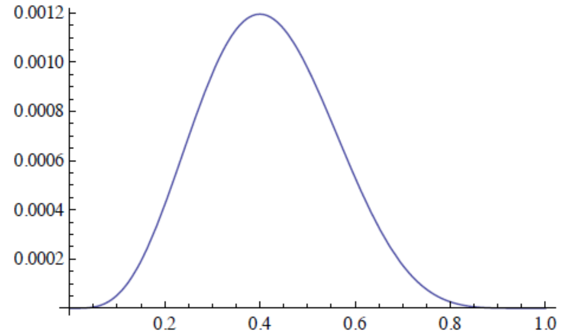
- To approximate the value of an integral

Why integral?

- 离散场景并不需要 (点估计)
  - Maximum likelihood estimate (MLE)

$(\pi, P(X \mid \pi))$

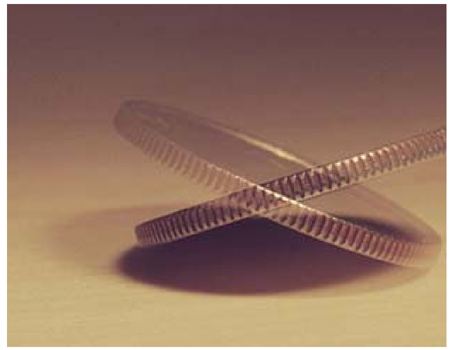
$$
\begin{aligned}
\tilde{\pi}_{MLE} &= \operatorname*{argmax}_{\pi} P(\mathcal{X}|\pi) \\
P(y|\mathcal{X}) &\approx P(y|\tilde{\pi}_{MLE})
\end{aligned}
$$



HHHHTTTTTT

  - *Maximum a posteriori* (MAP)

$$
\begin{aligned}
\tilde{\pi}_{MAP} &= \operatorname*{argmax}_{\pi} P(\pi|\mathcal{X}) \\
&= \operatorname*{argmax}_{\pi} \frac{P(\mathcal{X}|\pi)P(\pi)}{P(\mathcal{X})} \\
&= \operatorname*{argmax}_{\pi} P(\mathcal{X}|\pi)P(\pi) \\
P(y|\mathcal{X}) &\approx P(y|\tilde{\pi}_{MAP})
\end{aligned}
$$



Why integral?

- ：真的不需要吗 (考虑整体分布，例如求期望)？

$$E[f(z)] = \int f(z)p(z)\,\mathrm{d}z.$$

$$P(y|\mathcal{X}) = \int P(y|\pi)P(\pi|\mathcal{X})\,\mathrm{d}\pi$$

$$E_{\pi \sim P(\pi|X)}[P(y|\pi)]$$

$$P(\pi|\mathcal{X}) = \frac{P(\mathcal{X}|\pi)P(\pi)}{P(\mathcal{X})} = \frac{P(\mathcal{X}|\pi)P(\pi)}{\int_\pi P(\mathcal{X}|\pi)P(\pi)\,\mathrm{d}\pi}.$$

Why sampling？

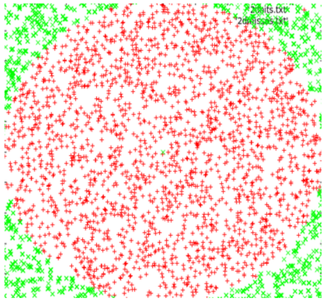
- $f(z)$不可积。。

Monte Carlo Simulation

- Approximate *Pi* (Probabilistic Choice)

$$\frac{C}{S} \approx \frac{\pi(\frac{d}{2})^2}{d^2}$$

$$\pi \approx \frac{4C}{S}$$



撒点即采样

- Properties of Monte Carlo
  - Estimator is unbiased
  - Variance shrinks ∝ 1/N (*N*是撒点数)

"Monte Carlo is an extremely bad method; it should be used only when all alternative methods are worse."

— Alan Sokal, 1996

Markov Chain Monte Carlo (MCMC)

- 回顾目标

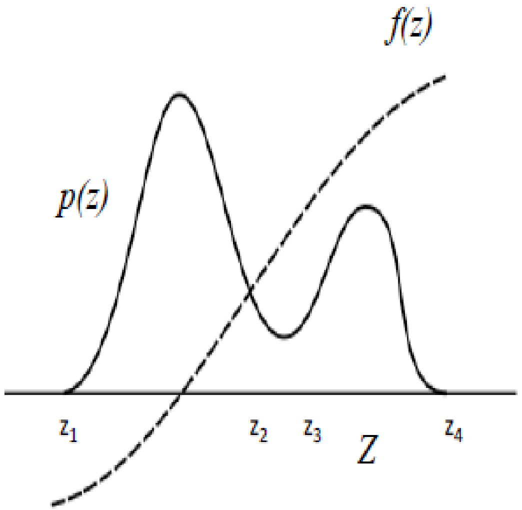
$$E[f(z)] = \int f(z)p(z)\,\mathrm{d}z.$$

- 基于*p*(*z*)采样*z*→代入*f*(*z*)→求均值

Monte Carlo

sample N points $z^{(0)}, z^{(1)}, z^{(2)}, \ldots, z^{(N)}$ at random from the probability density $p(z)$

$$E_{p(z)}[f(z)] = \lim_{N\to\infty} \frac{1}{N}\sum_{t=1}^{N} f(z^{(t)})$$



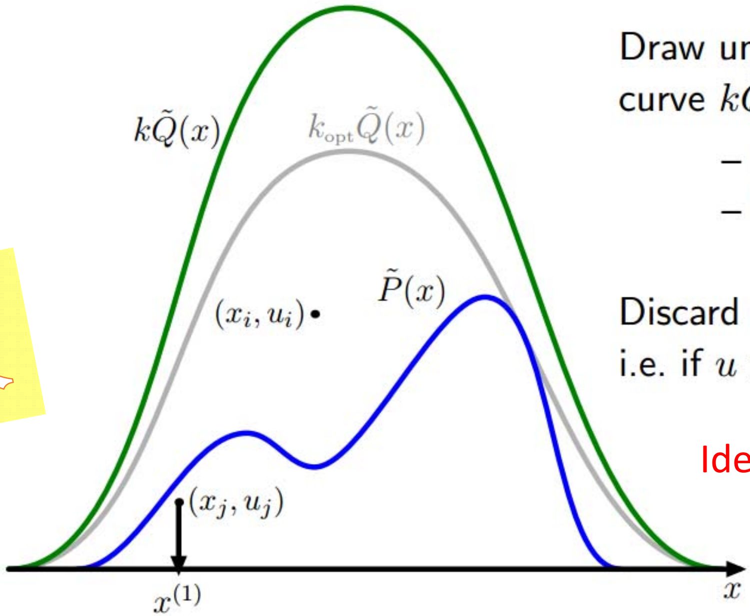
- Walking the right walk (找一个合适的*g*(*z*))

```
z^(0) := a random initial point
for t = 1 to T do
    z^(t+1) := g(z^(t))
end for
```

$$P_{\text{trans}}(z^{(t+1)}|z^{(0)}, z^{(1)}, \ldots, z^{(t)}) = P_{\text{trans}}(z^{(t+1)}|z^{(t)})$$

Markov Chain

Rejection Sampling and Importance Sampling



初级

Draw underneath a simple curve $k\tilde{Q}(x) \geq \tilde{P}(x)$:

- Draw $x \sim Q(x)$
- height $u \sim \text{Uniform}[0, k\tilde{Q}(x)]$

Discard the point if above $\tilde{P}$, i.e. if $u > \tilde{P}(x)$

Idea：找一个容易采样的分布*Q*把*P*包起来

$$\int f(x)P(x)\,\mathrm{d}x = \int f(x)\frac{P(x)}{Q(x)}Q(x)\,\mathrm{d}x, \qquad (Q(x) > 0 \text{ if } P(x) > 0)$$

$$\approx \frac{1}{S}\sum_{s=1}^{S} f(x^{(s)})\frac{P(x^{(s)})}{Q(x^{(s)})}, \qquad x^{(s)} \sim Q(x)$$

Monte Carlo

Metropolis–Hastings Sampling

- 算法过程
  - Propose a move from the current state $Q(x';x)$, e.g. $\mathcal{N}(x,\sigma^2)$
  - Accept with probability $\min\left(1, \frac{P(x')Q(x;x')}{P(x)Q(x';x)}\right)$
  - Otherwise next state in chain is a copy of current state
- 满足Detailed balance condition

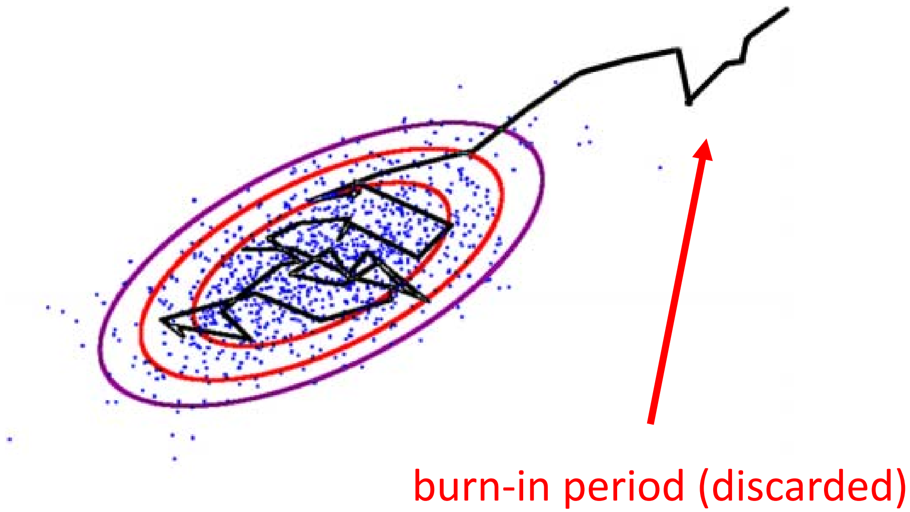
Markov Chain

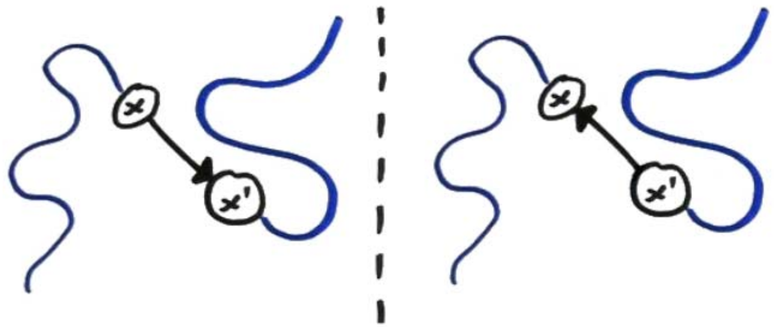
中级

$$
\begin{aligned}
P(x)\cdot T(x' \leftarrow x) = P(x)\cdot Q(x';x)\min\left(1, \frac{P(x')Q(x;x')}{P(x)Q(x';x)}\right) &= \min\left(P(x)Q(x';x),\ P(x')Q(x;x')\right) \\
&= P(x')\cdot Q(x;x')\min\left(1, \frac{P(x)Q(x';x)}{P(x')Q(x;x')}\right) = P(x')\cdot T(x \leftarrow x')
\end{aligned}
$$

Detailed Balance意味着什么？

- Implies the invariant condition
  - 跳转*N*步之后，采样收敛到P(x)





$$T(x' \leftarrow x)P^{\star}(x) = T(x \leftarrow x')P^{\star}(x')$$

还有些问题。。

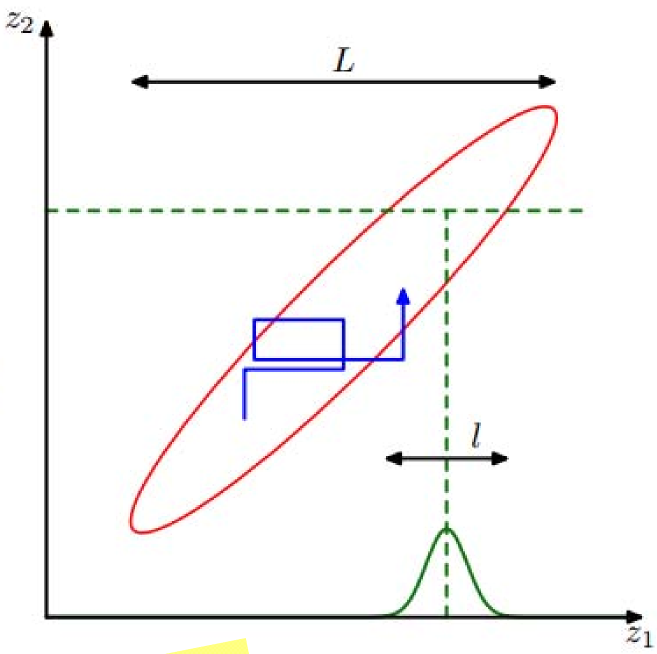
- 选择什么样的*Q*函数
- 采样高维数据很容易被拒绝

Gibbs Sampling

A method with no rejections:

– Initialize $\mathbf{x}$ to some value

– Pick each variable in turn or randomly and resample $P(x_i|\mathbf{x}_{j\neq i})$



Proof of validity: **a)** check detailed balance for component update.
b) Metropolis–Hastings 'proposals' $P(x_i|\mathbf{x}_{j\neq i}) \Rightarrow$ accept with prob. 1
Apply a series of these operators. Don't need to check acceptance.

Gibbs Sampling

- 算法过程

所以。。只需要凑出联合概率形式就阔以啦

$$
\begin{array}{l}
z^{(0)} := \langle z_1^{(0)}, \ldots, z_k^{(0)} \rangle \\
\textbf{for } t = 1 \textbf{ to } T \textbf{ do} \\
\quad \textbf{for } i = 1 \textbf{ to } k \textbf{ do} \\
\quad\quad z_i^{(t+1)} \sim P(Z_i | z_1^{(t+1)}, \ldots, z_{i-1}^{(t+1)}, z_{i+1}^{(t)}, \ldots, z_k^{(t)}) \\
\quad \textbf{end for} \\
\textbf{end for}
\end{array}
$$

$P(z)$

$$
P(Z_i | z_1^{(t+1)}, \ldots, z_{i-1}^{(t+1)}, z_{i+1}^{(t)}, \ldots, z_k^{(t)}) = \frac{P(z_1^{(t+1)}, \ldots, z_{i-1}^{(t+1)}, z_i^{(t)}, z_{i+1}^{(t)}, \ldots, z_k^{(t)})}{P(z_1^{(t+1)}, \ldots, z_{i-1}^{(t+1)}, z_{i+1}^{(t)}, \ldots, z_k^{(t)})}
$$

$P(z)^{(\neg i)}$

Part II

Probabilistic Graphical Model

Notations of Text Classification Task

	V	number of words in the vocabulary.
	N	number of documents in the corpus.
priori params	$\gamma_{\pi 1}, \gamma_{\pi 0}$	hyperparameters of the Beta distribution.
priori params	γ_θ	hyperparameter vector for the multinomial prior.
	$\gamma_{\theta i}$	pseudocount for word i.
set of docs	$\mathbb{C}_x$	set of documents labeled x.
set of docs	$\mathbb{C}$	the set of all documents.
	$C_0 \quad (C_1)$	number of documents labeled 0 (1).
doc = set of words	$\mathbf{W}_j$	document j's frequency distribution.
doc = set of words	W_{ji}	frequency of word i in document j.
doc label	$\mathbf{L}$	vector of document labels.
doc label	L_j	label for document j.
	R_j	number of words in document j.
word probs	θ_i	probability of word i.
word probs	$\theta_{x,i}$	probability of word i from the distribution of class x.
	$\mathcal{N}_{\mathbb{C}_x}(i)$	number of times word i occurs in the set of all documents labeled x.
without j^{th} doc	$\mathbb{C}^{(-j)}$	set of all documents *except* $\mathbf{W}_j$
without j^{th} doc	$\mathbf{L}^{(-j)}$	vector of all document labels *except* L_j
without j^{th} doc	$C_0^{(-j)} \quad (C_1^{(-j)})$	number of documents labeled 0 (1) *except* for $\mathbf{W}_j$
	$\boldsymbol{\mu}$	set of hyperparameters $\langle \gamma_{\pi 1}, \gamma_{\pi 0}, \gamma_\theta \rangle$

"j" indicates doc
and
"i" indicates word

Naïve Bayes Model

- 采用MAP求解(点估计)

$$
\begin{aligned}
\mathrm{L}_j = \operatorname*{argmax}_L P(L|\mathbf{W}_j) &= \operatorname*{argmax}_L \frac{P(\mathbf{W}_j|L)P(L)}{P(\mathbf{W}_j)} \\
&= \operatorname*{argmax}_L P(\mathbf{W}_j|L)P(L),
\end{aligned}
$$

- 属性之间条件独立

$$P(W_j \mid L) = \prod_{i=1}^{n} P(w_i \mid L)$$

$$W_j = \{w_1, w_2, \ldots, w_n\}$$

基于PGM的Naïve Bayes Model

- 优点
  - 贝叶斯方法：引入不确定性，物理过程更平滑
- 理解为生成过程

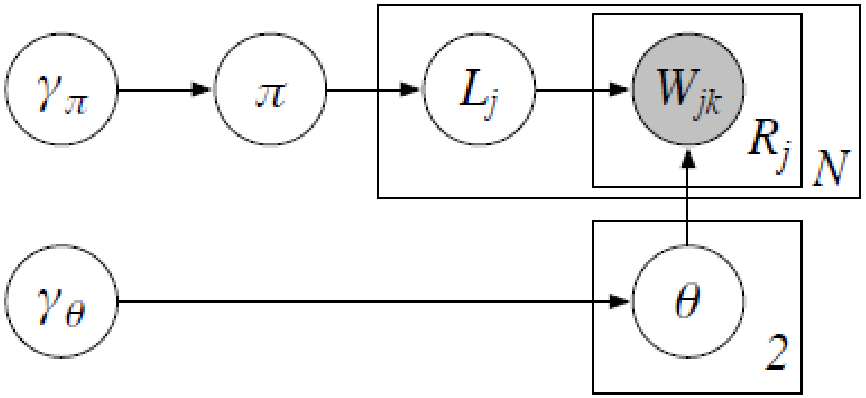
$$\mathbf{W}_j \sim \mathrm{Multinomial}(R_j, \boldsymbol{\theta}_{\mathrm{L}_j})$$

$$\mathrm{L}_j \sim \mathrm{Bernoulli}(\pi)$$

- 先验分布

$$\pi \sim \mathrm{Beta}(\gamma_{\boldsymbol{\pi}})$$

$$\boldsymbol{\theta} \sim \mathrm{Dirichlet}(\gamma_{\boldsymbol{\theta}})$$



等下。。Beta和Dirichlet是什么？

很久很久以前，在欧拉的时代。。

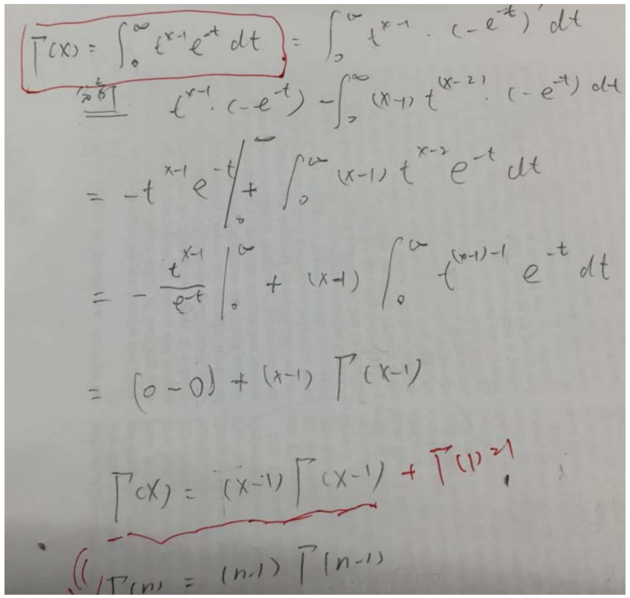
- 从Gamma函数说起

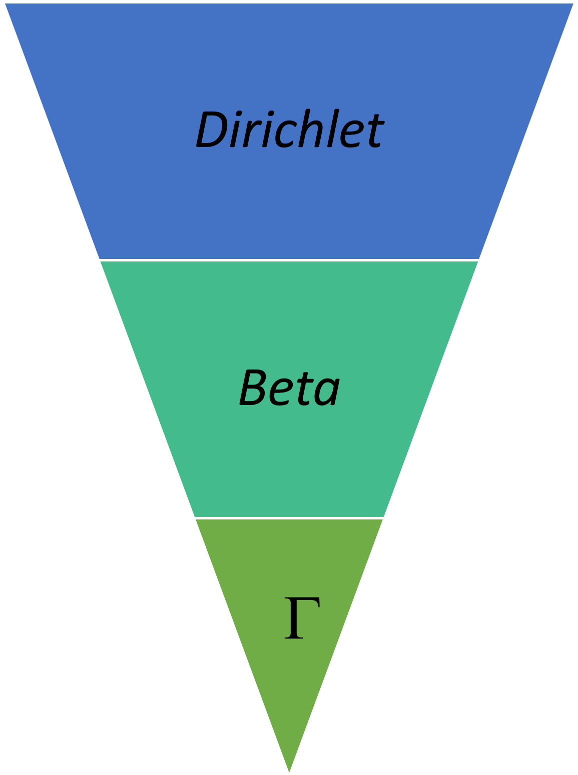
$$\Gamma(t) = \int_0^\infty x^{t-1} e^{-x}\, dx$$

- 将阶乘扩展到实数范围

$$\Gamma(t+1) = t\Gamma(t)$$

$$\begin{aligned} \Gamma(-1) &= (-2)! & \Gamma(-\tfrac{3}{2}) &= \tfrac{4}{3}\sqrt{\pi} \\ \Gamma(0) &= (-1)! & \Gamma(-\tfrac{1}{2}) &= -2\sqrt{\pi} \\ \Gamma(1) &= 0! & \Gamma(\tfrac{1}{2}) &= \sqrt{\pi} \\ \Gamma(2) &= 1! & \Gamma(\tfrac{3}{2}) &= \tfrac{1}{2}\sqrt{\pi} \\ \Gamma(3) &= 2! & \Gamma(\tfrac{5}{2}) &= \tfrac{3}{4}\sqrt{\pi} \\ \Gamma(4) &= 3! & \Gamma(\tfrac{7}{2}) &= \tfrac{15}{8}\sqrt{\pi} \end{aligned}$$





Gamma分布

- 被积项是密度函数

$$\int_0^{\infty} \frac{x^{\alpha-1}e^{-x}}{\Gamma(\alpha)} dx = 1$$

- Gamma分布

$$Gamma(x|\alpha) = \frac{x^{\alpha-1}e^{-x}}{\Gamma(\alpha)}$$

$$\beta = 1 \uparrow$$

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

嗯嗯。。分母就一阶乘

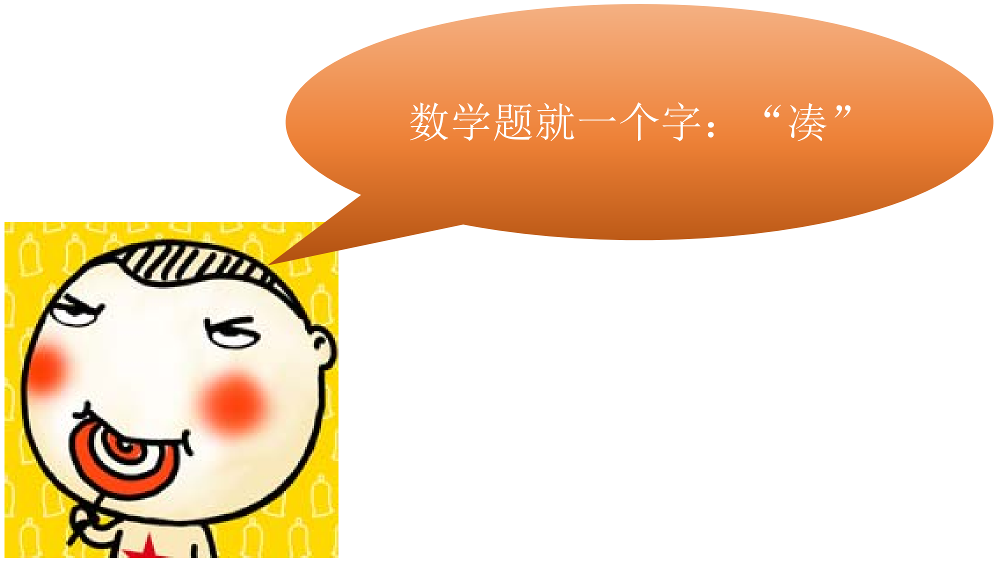
Beta分布

- 凑一个Beta分布

$$
\begin{aligned}
f(x;\alpha,\beta) &= \mathrm{constant}\cdot x^{\alpha-1}(1-x)^{\beta-1} \\
&= \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\int_0^1 u^{\alpha-1}(1-u)^{\beta-1}\,du} \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,x^{\alpha-1}(1-x)^{\beta-1} \\
&= \frac{1}{\mathrm{B}(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}
\end{aligned}
$$

- Beta分布的期望

$$
\begin{aligned}
\mu = \mathrm{E}[X] &= \int_0^1 x f(x;\alpha,\beta)\,dx \\
&= \int_0^1 x\,\frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}\,dx \\
&= \frac{\alpha}{\alpha+\beta} \\
&= \frac{1}{1+\frac{\beta}{\alpha}}
\end{aligned}
$$



为什么用Beta做先验分布？

- 共轭先验(形式一样→容易凑)
  - When the posterior probability distribution is of the same family as the prior probability distribution, it is said to be the conjugate prior of the posterior

$$p(\theta|x) = \frac{p(x|\theta)\, p(\theta)}{\int p(x|\theta')\, p(\theta')\, d\theta'}.$$

- Beta-Binomial共轭

Likelihood：Binomial $P(s,f|q=x) = \binom{s+f}{s} x^s (1-x)^f,$

Priori：Beta $P(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha,\beta)},$

$$P(q=x|s,f) = \frac{P(s,f|x)P(x)}{\int P(s,f|x)P(x)dx}$$

$$= \frac{\binom{s+f}{s} x^{s+\alpha-1}(1-x)^{f+\beta-1}/\mathrm{B}(\alpha,\beta)}{\int_{y=0}^{1}\left(\binom{s+f}{s} y^{s+\alpha-1}(1-y)^{f+\beta-1}/\mathrm{B}(\alpha,\beta)\right) dy}$$

Posterior：Beta $= \frac{x^{s+\alpha-1}(1-x)^{f+\beta-1}}{\mathrm{B}(s+\alpha,f+\beta)},$

Dirichlet的故事是一样的~

$$Beta(x|\alpha,\beta)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}\cdot(1-x)^{\beta-1}$$

$$Dir(\vec{p}|\vec{\alpha})=\frac{\Gamma(\sum_{i=1}^{k}\alpha_i)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\prod_{i=1}^{k}p_i^{\alpha_i-1}\quad s.t.\ \sum_{i=1}^{k}p_i=1$$

$$Binomial(x|m_1,m_2)=\binom{m_1+m_2}{m_1,m_2}x^{m_1}(1-x)^{m_2}$$

$$Multinomial(\vec{p}|\vec{m})=\binom{\sum_{i=1}^{k}m_i}{\vec{m}}\prod p_i^{m_i}$$

$$Beta(x|\alpha,\beta)+Binomial(x|m_1,m_2)\propto Beta(x|\alpha+m_1,\beta+m_2)$$

$$Dir(\vec{p}|\vec{\alpha})+Multinomial(\vec{p}|\vec{m})\propto Dir(\vec{p}|\vec{\alpha}+\vec{m})$$

在这里能找到更多的共轭先验
https://en.wikipedia.org/wiki/Conjugate_prior

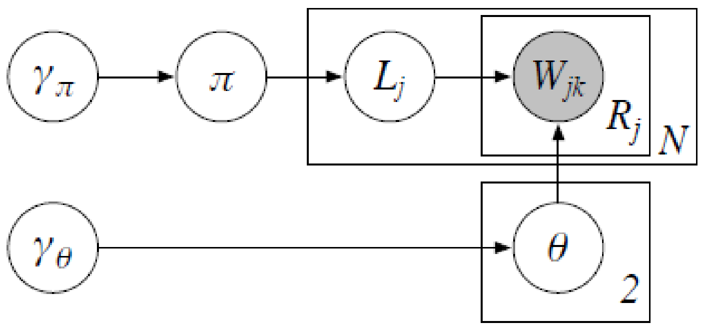
回到Naïve Bayes Model(基于PGM)

- 步骤一：凑出简单形式的联合概率

$$P(\mathbb{C}, \mathbf{L}, \pi, \theta_0, \theta_1; \gamma_{\pi 1}, \gamma_{\pi 0}, \gamma_\theta) = P(\pi|\gamma_{\pi 1}, \gamma_{\pi 0}) P(\mathbf{L}|\pi) P(\theta_0|\gamma_\theta) P(\theta_1|\gamma_\theta) P(\mathbb{C}_0|\theta_0, \mathbf{L}) P(\mathbb{C}_1|\theta_1, \mathbf{L})$$

- 步骤二：用Gibbs sampling估计参数

π, θ, L



步骤一：凑出简单形式的联合概率

- 分别凑每个因子

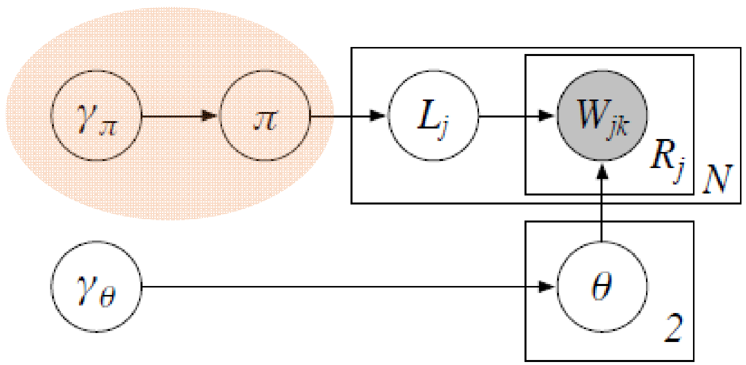
$$P(\mathbb{C}, \mathbf{L}, \pi, \boldsymbol{\theta}_0, \boldsymbol{\theta}_1; \gamma_{\pi 1}, \gamma_{\pi 0}, \gamma_\theta) \cdot = P(\pi|\gamma_{\pi 1}, \gamma_{\pi 0}) P(\mathbf{L}|\pi) P(\boldsymbol{\theta}_0|\gamma_\theta) P(\boldsymbol{\theta}_1|\gamma_\theta) P(\mathbb{C}_0|\boldsymbol{\theta}_0, \mathbf{L}) P(\mathbb{C}_1|\boldsymbol{\theta}_1, \mathbf{L})$$

- 第一个因子(Beta)

$$P(\pi|\gamma_{\pi 1}, \gamma_{\pi 0}) = \frac{\Gamma(\gamma_{\pi 1} + \gamma_{\pi 0})}{\Gamma(\gamma_{\pi 1})\Gamma(\gamma_{\pi 0})} \pi^{\gamma_{\pi 1}-1}(1-\pi)^{\gamma_{\pi 0}-1}$$

$$\downarrow$$

$$P(\pi|\gamma_{\pi 1}, \gamma_{\pi 0}) \propto \pi^{\gamma_{\pi 1}-1}(1-\pi)^{\gamma_{\pi 0}-1}$$



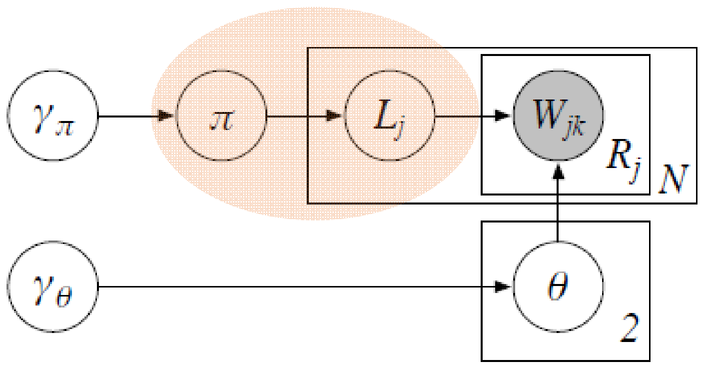
步骤一：凑出简单形式的联合概率

- 分别凑每个因子

$$P(\mathbb{C}, \mathbf{L}, \pi, \theta_0, \theta_1; \gamma_{\pi 1}, \gamma_{\pi 0}, \gamma_\theta) = P(\pi|\gamma_{\pi 1}, \gamma_{\pi 0}) P(\mathbf{L}|\pi) P(\theta_0|\gamma_\theta) P(\theta_1|\gamma_\theta) P(\mathbb{C}_0|\theta_0, \mathbf{L}) P(\mathbb{C}_1|\theta_1, \mathbf{L})$$

- 第二个因子(Binomial)

$$\begin{aligned} P(\mathbf{L}|\pi) &= \prod_{n=1}^{N} \pi^{\mathbf{L}_n}(1-\pi)^{(1-\mathbf{L}_n)} \\ &= \pi^{C_1}(1-\pi)^{C_0} \end{aligned}$$



步骤一：凑出简单形式的联合概率

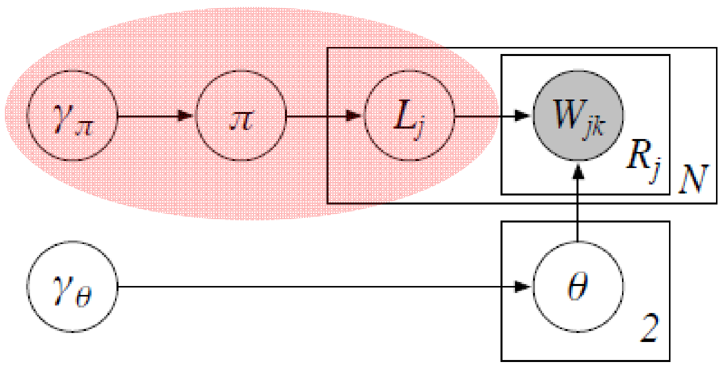
- 分别凑每个因子

$$P(\mathbb{C},\mathbf{L},\pi,\theta_0,\theta_1;\gamma_{\pi 1},\gamma_{\pi 0},\gamma_\theta) = P(\pi|\gamma_{\pi 1},\gamma_{\pi 0})P(\mathbf{L}|\pi)P(\theta_0|\gamma_\theta)P(\theta_1|\gamma_\theta)P(\mathbb{C}_0|\theta_0,\mathbf{L})P(\mathbb{C}_1|\theta_1,\mathbf{L})$$

- 第一个因子(Beta)*第二个因子(Binomial)

$$\begin{aligned}
P(\pi|\mathbf{L};\gamma_{\pi 1},\gamma_{\pi 0}) &\overset{\propto}{\cancel{=}} P(\mathbf{L}|\pi)P(\pi|\gamma_{\pi 1},\gamma_{\pi 0}) \\
&\propto \left[\pi^{C_1}(1-\pi)^{C_0}\right]\left[\pi^{\gamma_{\pi 1}-1}(1-\pi)^{\gamma_{\pi 0}-1}\right] \\
&\propto \pi^{C_1+\gamma_{\pi 1}-1}(1-\pi)^{C_0+\gamma_{\pi 0}-1}
\end{aligned}$$

$$\begin{aligned}
P(\pi|L,\gamma) &= \frac{P(\pi,L,\gamma)}{P(L,\gamma)} \\
&= \frac{P(L|\pi,\gamma)\,P(\pi|\gamma)\,\cancel{P(\gamma)}}{\sum_\pi P(L|\pi,\gamma)\,P(\pi|\gamma)\cdot\cancel{P(\gamma)}}
\end{aligned}$$



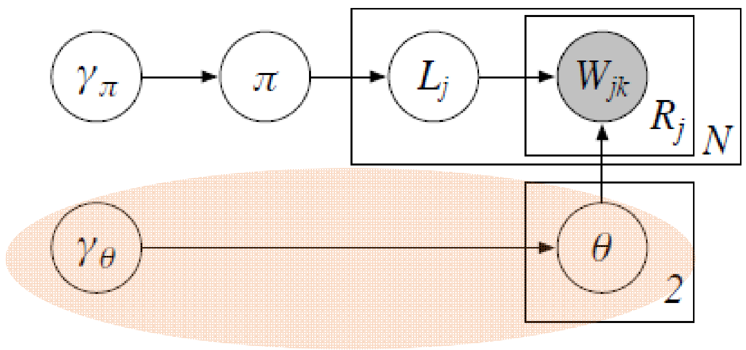
步骤一：凑出简单形式的联合概率

- 分别凑每个因子

$$P(\mathbb{C},\mathbf{L},\pi,\boldsymbol{\theta}_0,\boldsymbol{\theta}_1;\gamma_{\pi 1},\gamma_{\pi 0},\gamma_\theta). = P(\pi|\gamma_{\pi 1},\gamma_{\pi 0})P(\mathbf{L}|\pi)P(\boldsymbol{\theta}_0|\gamma_\theta)P(\boldsymbol{\theta}_1|\gamma_\theta)P(\mathbb{C}_0|\boldsymbol{\theta}_0,\mathbf{L})P(\mathbb{C}_1|\boldsymbol{\theta}_1,\mathbf{L})$$

- 第三、四个因子(Dirichlet)

$$\begin{aligned} P(\theta|\gamma_\theta) &= \frac{\Gamma(\sum_{i=1}^{V}\gamma_{\theta i})}{\prod_{i=1}^{V}\Gamma(\gamma_{\theta i})}\prod_{i=1}^{V}\theta_i^{\gamma_{\theta i}-1} \\ &= c'\prod_{i=1}^{V}\theta_i^{\gamma_{\theta i}-1} \\ &\propto \prod_{i=1}^{V}\theta_i^{\gamma_{\theta i}-1} \end{aligned}$$



步骤一：凑出简单形式的联合概率

- 分别凑每个因子

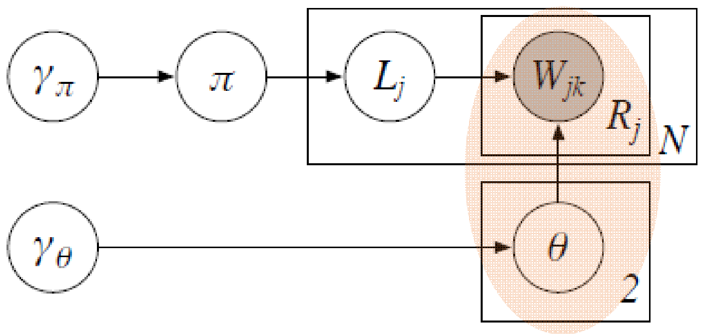
$$P(\mathbb{C},\mathbf{L},\pi,\theta_0,\theta_1;\gamma_{\pi 1},\gamma_{\pi 0},\gamma_\theta) = P(\pi|\gamma_{\pi 1},\gamma_{\pi 0})P(\mathbf{L}|\pi)P(\theta_0|\gamma_\theta)P(\theta_1|\gamma_\theta)P(\mathbb{C}_0|\theta_0,\mathbf{L})P(\mathbb{C}_1|\theta_1,\mathbf{L})$$

- 第五、六个因子(Multinomial)

$$P(\mathbf{W}_n|\mathbf{L},\theta_{\mathbf{L}_n}) = \prod_{i=1}^{V}\theta_i^{\mathbf{W}_{ni}}$$

$$\downarrow$$

$$P(\mathbb{C}_x|\mathbf{L},\theta_x) = \prod_{n\in\mathbb{C}_x}\prod_{i=1}^{V}\theta_{x,i}^{\mathbf{W}_{ni}} = \prod_{i=1}^{V}\theta_{x,i}^{\mathcal{N}_{\mathbb{C}_x}(i)}$$



步骤一：凑出简单形式的联合概率

- 分别凑每个因子

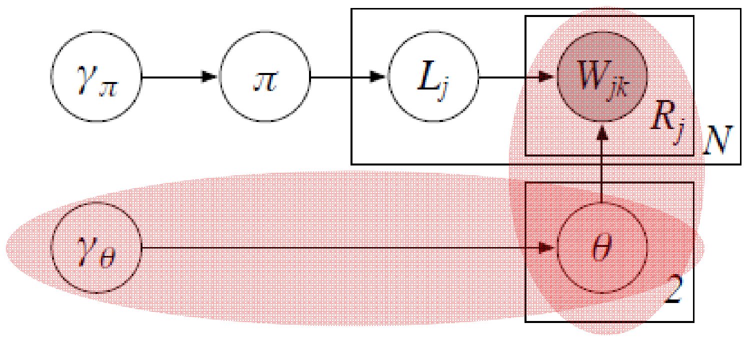
$$P(\mathbb{C},\mathbf{L},\pi,\theta_0,\theta_1;\gamma_{\pi 1},\gamma_{\pi 0},\gamma_\theta) = P(\pi|\gamma_{\pi 1},\gamma_{\pi 0})P(\mathbf{L}|\pi)P(\theta_0|\gamma_\theta)P(\theta_1|\gamma_\theta)P(\mathbb{C}_0|\theta_0,\mathbf{L})P(\mathbb{C}_1|\theta_1,\mathbf{L})$$

- 第三、四个因子(Dirichlet)*第五、六个因子(Multinomial)

$$\begin{aligned} P(\theta|\mathbf{W}_n;\gamma_\theta) &= P(\mathbf{W}_n|\theta)P(\theta|\gamma_\theta) \\ &\propto \prod_{i=1}^{V}\theta_i^{W_{ni}}\prod_{i=1}^{V}\theta_i^{\gamma_{\theta i}-1} \\ &\propto \prod_{i=1}^{V}\theta_i^{W_{ni}+\gamma_{\theta i}-1} \end{aligned}$$

$$\downarrow$$

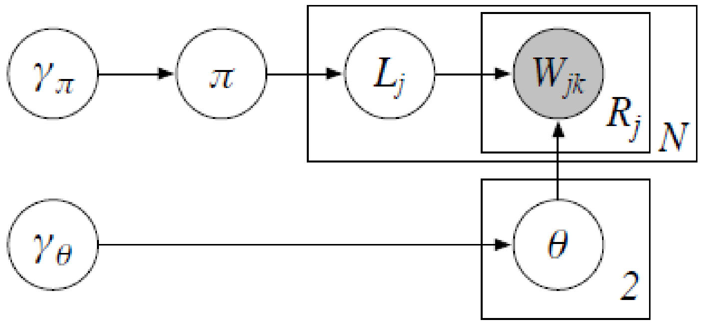
$$P(\theta_x|\mathbb{C}_x;\gamma_\theta) \propto \prod_{i=1}^{V}\theta_{x,i}^{\mathcal{N}_{\mathbb{C}_x}(i)+\gamma_{\theta i}-1}$$



步骤一：凑出简单形式的联合概率

- 凑出的结果

$$
\begin{aligned}
P(\mathbb{C}, \mathbf{L}, \pi, \theta_0, \theta_1; \gamma_{\pi 1}, \gamma_{\pi 0}, \gamma_\theta) &= P(\pi|\gamma_{\pi 1}, \gamma_{\pi 0}) P(\mathbf{L}|\pi) P(\theta_0|\gamma_\theta) P(\theta_1|\gamma_\theta) P(\mathbb{C}_0|\theta_0, \mathbf{L}) P(\mathbb{C}_1|\theta_1, \mathbf{L}) \\
&\propto \pi^{C_1+\gamma_{\pi 1}-1}(1-\pi)^{C_0+\gamma_{\pi 0}-1} \prod_{i=1}^{V} \theta_{0,i}^{\mathcal{N}_{\mathbb{C}_0}(i)+\gamma_{\theta i}-1} \theta_{1,i}^{\mathcal{N}_{\mathbb{C}_1}(i)+\gamma_{\theta i}-1}
\end{aligned}
$$



步骤一：凑出简单形式的联合概率

- 还有更简单的形式吗(*COLLAPSED* Gibbs sampling：积掉π)

$$
\begin{aligned}
P(\mathbf{L}, \mathbb{C}, \boldsymbol{\theta}_0, \boldsymbol{\theta}_1; \boldsymbol{\mu}) &= \int_\pi P(\mathbf{L}, \mathbb{C}, \boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \pi; \boldsymbol{\mu})\, \mathrm{d}\pi \\
&= P(\boldsymbol{\theta}_0|\boldsymbol{\gamma}_\theta)P(\boldsymbol{\theta}_1|\boldsymbol{\gamma}_\theta)P(\mathbb{C}_0|\boldsymbol{\theta}_0, \mathbf{L})P(\mathbb{C}_1|\boldsymbol{\theta}_1, \mathbf{L}) \int_\pi P(\pi|\gamma_{\pi 1}, \gamma_{\pi 0})P(\mathbf{L}|\pi)\, \mathrm{d}\pi
\end{aligned}
$$

$$
P(\mathbf{L}, \mathbb{C}, \boldsymbol{\theta}_0, \boldsymbol{\theta}_1; \boldsymbol{\mu}) \propto \underline{\frac{\Gamma(\gamma_{\pi 1} + \gamma_{\pi 0})}{\Gamma(\gamma_{\pi 1})\Gamma(\gamma_{\pi 0})} \frac{\Gamma(C_1 + \gamma_{\pi 1})\Gamma(C_0 + \gamma_{\pi 0})}{\Gamma(N + \gamma_{\pi 1} + \gamma_{\pi 0})} \prod_{i=1}^{V} \theta_{0,i}^{\mathcal{N}_{\mathbb{C}_0}(i)+\gamma_{\theta i}-1} \theta_{1,i}^{\mathcal{N}_{\mathbb{C}_1}(i)+\gamma_{\theta i}-1}}
$$

$$
\uparrow
$$

$$
\begin{aligned}
\int_\pi P(\pi|\gamma_{\pi 1}, \gamma_{\pi 0})P(\mathbf{L}|\pi)\, \mathrm{d}\pi &= \int_\pi \frac{\Gamma(\gamma_{\pi 1} + \gamma_{\pi 0})}{\Gamma(\gamma_{\pi 1})\Gamma(\gamma_{\pi 0})} \pi^{\gamma_{\pi 1}-1}(1-\pi)^{\gamma_{\pi 0}-1} \pi^{C_1}(1-\pi)^{C_0}\, \mathrm{d}\pi \\
&= \frac{\Gamma(\gamma_{\pi 1} + \gamma_{\pi 0})}{\Gamma(\gamma_{\pi 1})\Gamma(\gamma_{\pi 0})} \int_\pi \pi^{C_1+\gamma_{\pi 1}-1}(1-\pi)^{C_0+\gamma_{\pi 0}-1}\, \mathrm{d}\pi
\end{aligned}
$$

步骤二：用Gibbs sampling估计参数

- Sampling for Document Labels

采呀采呀快采样。。。
采样采得心花开。。。

① $$P(\mathrm{L}_j|\mathbf{L}^{(-j)},\mathbb{C}^{(-j)},\boldsymbol{\theta}_0,\boldsymbol{\theta}_1;\boldsymbol{\mu}) = \frac{P(\mathrm{L}_j,\mathbf{W}_j,\mathbf{L}^{(-j)},\mathbb{C}^{(-j)},\boldsymbol{\theta}_0,\boldsymbol{\theta}_1;\boldsymbol{\mu})}{P(\mathbf{L}^{(-j)},\mathbb{C}^{(-j)},\boldsymbol{\theta}_0,\boldsymbol{\theta}_1;\boldsymbol{\mu})} = \frac{P(\mathbf{L},\mathbb{C},\boldsymbol{\theta}_0,\boldsymbol{\theta}_1;\boldsymbol{\mu})}{P(\mathbf{L}^{(-j)},\mathbb{C}^{(-j)},\boldsymbol{\theta}_0,\boldsymbol{\theta}_1;\boldsymbol{\mu})}$$

② $$P(\mathbf{L},\mathbb{C},\boldsymbol{\theta}_0,\boldsymbol{\theta}_1;\boldsymbol{\mu}) \propto \frac{\Gamma(\gamma_{\pi1}+\gamma_{\pi0})}{\Gamma(\gamma_{\pi1})\Gamma(\gamma_{\pi0})}\frac{\Gamma(C_1+\gamma_{\pi1})\Gamma(C_0+\gamma_{\pi0})}{\Gamma(N+\gamma_{\pi1}+\gamma_{\pi0})}\prod_{i=1}^{V}\theta_{0,i}^{\mathcal{N}_{\mathbb{C}_0}(i)+\gamma_{\theta i}-1}\theta_{1,i}^{\mathcal{N}_{\mathbb{C}_1}(i)+\gamma_{\theta i}-1}$$

与L无关→cancel out

$$\frac{\frac{\Gamma(C_1+\gamma_{\pi1})\Gamma(C_0+\gamma_{\pi0})}{\Gamma(N+\gamma_{\pi1}+\gamma_{\pi0})}}{\frac{\Gamma(C_0^{(-j)}+\gamma_{\pi0})\Gamma(C_1^{(-j)}+\gamma_{\pi1})}{\Gamma(N+\gamma_{\pi1}+\gamma_{\pi0}-1)}}$$

$$\frac{C_x+\gamma_{\pi x}-1}{N+\gamma_{\pi1}+\gamma_{\pi0}-1}$$

$$\prod_{i=1}^{V}\frac{\theta_{x,i}^{\mathcal{N}_{\mathbb{C}_x}(i)+\gamma_{\theta i}-1}}{\theta_{x,i}^{\mathcal{N}_{\mathbb{C}_x^{(-j)}}(i)+\gamma_{\theta i}-1}} = \prod_{i=1}^{V}\theta_{x,i}^{\mathrm{W}_{ji}}$$

③ $$\Pr(\mathrm{L}_j = x|\mathbf{L}^{(-j)},\mathbb{C}^{(-j)},\boldsymbol{\theta}_0,\boldsymbol{\theta}_1;\boldsymbol{\mu}) = \frac{C_x+\gamma_{\pi x}-1}{N+\gamma_{\pi1}+\gamma_{\pi0}-1}\prod_{i=1}^{V}\theta_{x,i}^{\mathrm{W}_{ji}}$$

步骤二：用Gibbs sampling估计参数

- Sampling for Document Labels

$$\Pr(\mathrm{L}_j = x|\mathbf{L}^{(-j)}, \mathbb{C}^{(-j)}, \boldsymbol{\theta}_0, \boldsymbol{\theta}_1; \boldsymbol{\mu}) = \frac{C_x + \gamma_{\pi x} - 1}{N + \gamma_{\pi 1} + \gamma_{\pi 0} - 1} \prod_{i=1}^{V} \theta_{x,i}^{\mathrm{W}_{ji}}$$

已知Label的Docs，
按真实Label计数。

1. Let value0 = expression (49) with $x = 0$
2. Let value1 = expression (49) with $x = 1$
3. Let the distribution be $\langle \frac{value0}{value0+value1}, \frac{value1}{value0+value1} \rangle$
4. Select the value of $\mathrm{L}_j^{(t+1)}$ as the result of a Bernoulli trial (weighted coin flip) according to this distribution.

步骤二：用Gibbs sampling估计参数

- Sampling for θ
  - *C*, *L*与μ独立

$$P(\theta_x|\mathbb{C}_x;\gamma_\theta) \propto \prod_{i=1}^{V} \theta_{x,i}^{N_{\mathbb{C}_x}(i)+\gamma_{\theta i}-1}$$

$$P(C,L\,|\,\theta;\mu) = P(C,L\,|\,\theta)$$

$$\underbrace{P(\theta|\mathbb{C},\mathbf{L};\mu)}_{\text{Dirichlet}} \propto \underbrace{P(\mathbb{C},\mathbf{L}|\theta)}_{\text{Multinomial}}\underbrace{P(\theta|\mu)}_{\text{Dirichlet}}$$

  - So，直接按狄利克雷分布采样

$$\theta \sim Dir(N_{C_x}(i) + \gamma_{\theta i})$$

问：怎么采样Dirichlet分布？

with $\mathbf{a} = \langle a_1, \ldots, a_V \rangle$

答：分别采样每个因子

$$\mathrm{Gamma}(\alpha_i, 1) = \frac{y_i^{\alpha_i - 1}\, e^{-y_i}}{\Gamma(\alpha_i)}$$

$$a_i = y_i / \textstyle\sum_{j=1}^{V} y_j$$

PGM based Naïve Bayes Model

- Gibbs sampling algorithm

for $t := 1$ to T **do**
    for $j := 1$ to N **do**
        if j is not a training document **then**
            Subtract j's word counts from the total word counts of whatever class it's currently a member of
            Subtract 1 from the count of documents with label L_j
            Assign a new label $\mathrm{L}_j^{(t+1)}$ to document j as described at the end of Section 2.5.1
            Add 1 to the count of documents with label $\mathrm{L}_j^{(t+1)}$
            Add j's word counts to the total word counts for class $\mathrm{L}_j^{(t+1)}$
        end if
    end for
    $\mathbf{t_0}$:= vector of total word counts from class 0, including pseudocounts
    $\theta_0 \sim \mathrm{Dirichlet}(\mathbf{t_0})$, as described in Section 2.5.2
    $\mathbf{t_1}$:= vector of total word counts from class 1, including pseudocounts
    $\theta_1 \sim \mathrm{Dirichlet}(\mathbf{t_1})$, as described in Section 2.5.2
end for

Part III

PLSA and LDA

Probabilistic latent semantic analysis (PLSA)

pLSA: $d \rightarrow z \rightarrow w$

EM 算法:

X: 观测变量

Z: 隐藏变量

θ: 参数.

E: $q(z) = p(z|x,\theta^{old})$

$L(\theta) = \sum_z q(z) \ln p(x,z|\theta)$

M: $\theta^{new} = \arg\max_{\theta} L(\theta, \theta^{old})$

PLSA的EM求解过程（E-Step）

E: $p(z_k|w_j,d_i) = \dfrac{P(d_i,w_j,z_k)}{P(w_j,d_i)}$

$$= \frac{P(w_j|d_i,z_k)\cdot P(z_k|d_i)P(d_i)}{\sum_k P(w_j|d_i,z_k)\cdot P(z_k|d_i)\cdot P(d_i)}$$

$$= \frac{P(w_j|z_k)P(z_k|d_i)}{\sum_k P(w_j|z_k)P(z_k|d_i)}$$

令 $\phi_{kj} = P(w_j|z_k)$ 则参数 $\theta = \begin{pmatrix}\phi_{kj}\\ \theta_{ik}\end{pmatrix}$

$\theta_{ik} = P(z_k|d_i)$

参数化

令 $\phi_{kj} = P(w_j|z_k)$ 则参数 $\theta = \begin{pmatrix}\phi_{kj}\\ \theta_{ik}\end{pmatrix}$

$\theta_{ik} = P(z_k|d_i)$

(注：当文档的总数为M，词个数为N，隐含主题的总数为K时

ϕ_{kj} 为 $K\times N$ 个

θ_{ik} 为 $M\times K$ 个)

根据条件概率，有以下的约束

$$\sum_j \phi_{kj} = 1$$

$$\sum_k \theta_{ik} = 1$$

PLSA的EM求解过程（M-Step）

M:

令 $q(z_k) = p(z_k | w_j, d_i)$

$$L(\theta) = \sum_k q(z_k) \ln \prod_{\text{样本}} P(d_i, w_j, z_k | \theta)$$

$$= \sum_k q(z_k) \ln \prod_i \prod_j P(d_i, w_j, z_k | \theta)^{n(d_i, w_j)}$$

$$= \sum_k q(z_k) \sum_i \sum_j n(d_i, w_j) \ln P(d_i, w_j, z_k | \theta)$$

$$= \cancel{\sum_k q(z_k) \sum_i \sum}$$

$$= \sum_i \sum_j n(d_i, w_j) \sum_k q(z_k) \ln P(d_i, w_j, z_k | \theta)$$

$$= \sum_i \sum_j n(d_i, w_j) \sum_k q(z_k) \ln P(w_j | z_k) P(z_k | d_i) P(d_i)$$

$$= \sum_i \sum_j n(d_i, w_j) \sum_k q(z_k) \ln \phi_{kj} \theta_{ik} \cdot P(d_i)$$

$n(d_i)$

加入约束

$$L(\theta) = \sum_i \sum_j n(d_i, w_j) \sum_k q(z_k) \ln \phi_{kj} \theta_{ik} + \lambda_1 (1 - \sum_j \phi_{kj}) + \lambda_2 (1 - \sum_k \theta_{ik})$$

$$\frac{\partial L(\theta)}{\partial \phi_{kj}} = \sum_i n(d_i, w_j) q(z_k) \cdot \frac{1}{\phi_{kj}} - \lambda_1 = 0$$

$$\frac{\partial L(\theta)}{\partial \theta_{ik}} = \sum_j n(d_i, w_j) q(z_k) \cdot \frac{1}{\theta_{ik}} - \lambda_2 = 0$$

$$\sum_j \phi_{kj} = \frac{1}{\lambda_1} \sum_i \sum_j n(d_i, w_j) q(z_k) = 1$$

$$\sum_k \theta_{ik} = \frac{1}{\lambda_2} \sum_j \sum_k n(d_i, w_j) q(z_k) = 1$$

得到: $\lambda_1 = \sum_i \sum_j n(d_i, w_j) q(z_k)$

$\lambda_2 = \sum_j \sum_k n(d_i, w_j) q(z_k) = \sum_j n(d_i, w_j) = n(d_i)$

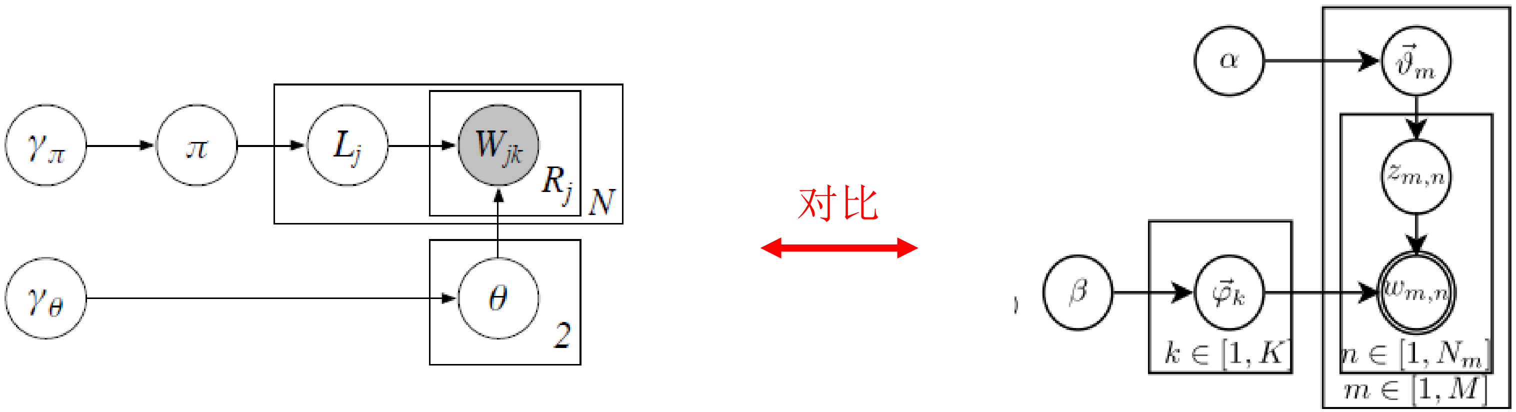
求解参数

又得到: $\phi_{kj} = \dfrac{\sum_i n(d_i, w_j) q(z_k)}{\sum_i \sum_j n(d_i, w_j) q(z_k)}$

$\theta_{ik} = \dfrac{\sum_j n(d_i, w_j) q(z_k)}{\sum_j n(d_i, w_j)}$ $n(d_i)$

Latent Dirichlet allocation（LDA）

- 对比差异（ Label的个数从2扩展到K ）
  - π从Beta分布扩展到Dirichlet分布
  - θ被生成的次数从2扩展到K次



Naive Bayes Model

LDA Model

LDA Notations

M number of documents to generate (const scalar).
K number of topics / mixture components (const scalar).
V number of terms t in vocabulary (const scalar).
$\vec{\alpha}$ hyperparameter on the mixing proportions (K-vector or scalar if symmetric).
$\vec{\beta}$ hyperparameter on the mixture components (V-vector or scalar if symmetric).
$\vec{\vartheta}_m$ parameter notation for $p(z|d{=}m)$, the topic mixture proportion for document m. One proportion for each document, $\underline{\Theta} = \{\vec{\vartheta}_m\}_{m=1}^{M}$ ($M \times K$ matrix).
$\vec{\varphi}_k$ parameter notation for $p(t|z{=}k)$, the mixture component of topic k. One component for each topic, $\underline{\Phi} = \{\vec{\varphi}_k\}_{k=1}^{K}$ ($K \times V$ matrix).
N_m document length (document-specific), here modelled with a Poisson distribution [BNJ02] with constant parameter ξ.
$z_{m,n}$ mixture indicator that chooses the topic for the nth word in document m.
$w_{m,n}$ term indicator for the nth word in document m.

Solving LDA using Gibbs Sampling

- 凑出联合概率

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

$$p(\vec{z},\vec{w}|\vec{\alpha},\vec{\beta}) = \prod_{z=1}^{K}\frac{\Delta(\vec{n}_z+\vec{\beta})}{\Delta(\vec{\beta})}\cdot\prod_{m=1}^{M}\frac{\Delta(\vec{n}_m+\vec{\alpha})}{\Delta(\vec{\alpha})}.$$

- 凑出条件概率

$$\begin{aligned}
p(z_i=k|\vec{z}_{\neg i},\vec{w}) &= \frac{p(\vec{w},\vec{z})}{p(\vec{w},\vec{z}_{\neg i})} = \frac{p(\vec{w}|\vec{z})}{p(\vec{w}_{\neg i}|\vec{z}_{\neg i})p(w_i)}\cdot\frac{p(\vec{z})}{p(\vec{z}_{\neg i})} \\
&\propto \frac{\Delta(\vec{n}_z+\vec{\beta})}{\Delta(\vec{n}_{z,\neg i}+\vec{\beta})}\cdot\frac{\Delta(\vec{n}_m+\vec{\alpha})}{\Delta(\vec{n}_{m,\neg i}+\vec{\alpha})} \\
&\propto \frac{\Gamma(n_k^{(t)}+\beta_t)\,\Gamma(\sum_{t=1}^{V} n_{k,\neg i}^{(t)}+\beta_t)}{\Gamma(n_{k,\neg i}^{(t)}+\beta_t)\,\Gamma(\sum_{t=1}^{V} n_k^{(t)}+\beta_t)}\cdot\frac{\Gamma(n_m^{(k)}+\alpha_k)\,\Gamma(\sum_{k=1}^{K} n_{m,\neg i}^{(k)}+\alpha_k)}{\Gamma(n_{m,\neg i}^{(k)}+\alpha_k)\,\Gamma(\sum_{k=1}^{K} n_m^{(k)}+\alpha_k)} \\
&\propto \frac{n_{k,\neg i}^{(t)}+\beta_t}{\sum_{t=1}^{V} n_{k,\neg i}^{(t)}+\beta_t}\cdot\frac{n_{m,\neg i}^{(t)}+\alpha_k}{[\sum_{k=1}^{K} n_m^{(k)}+\alpha_k]-1}
\end{aligned}$$

topic part document part

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

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

Remaining Problems in LDA

- hyper parameters
  - symmetric Dirichlet priors
  - $\alpha = 50/K$ and $\beta = 0.01$
- Querying (波浪表示查询文档)

$$p(\tilde{z}_i = k | \tilde{w}_i = t, \tilde{\vec{z}}_{\neg i}, \tilde{\vec{w}}_{\neg i}; \mathcal{M}) = \frac{n_k^{(t)} + \tilde{n}_{k,\neg i}^{(t)} + \beta_t}{\sum_{t=1}^{V} n_k^{(t)} + \tilde{n}_{k,\neg i}^{(t)} + \beta_t} \cdot \frac{n_{\tilde{m},\neg i}^{(k)} + \alpha_k}{[\sum_{k=1}^{K} n_{\tilde{m}}^{(k)} + \alpha_k] - 1}$$

大规模语料下数值很稳定

$$\vartheta_{\tilde{m},k} = \frac{n_{\tilde{m}}^{(k)} + \alpha_k}{\sum_{k=1}^{K} n_{\tilde{m}}^{(k)} + \alpha_k}.$$

Further Reading

- Gershman, S. J., & Blei, D. M. (2012). A tutorial on Bayesian nonparametric models. *Journal of Mathematical Psychology*, *56*(1), 1–12.
- Knight, K. (2009). Bayesian Inference with Tears. *A Tutorial Workbook for Natural Language Researchers*, *4*(September), 388–395.