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

性质 $\Gamma(x+1)=x\Gamma(x)$

于是很容易证出 $\Gamma(n)=(n-1)!$???

即阶乘在实数集上的拓展.

推导历史:

1728 哥德巴赫处理阶乘序列 1, 2, 6, 24, 120 …

我们可以算 2!, 3! 那 2.5! 怎么算?

哥与多位情敌丹尼尔·贝努利, 欧拉也在.

1729年 欧拉22岁. 完美解决了此题

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

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

取积分中的函数作概率密度

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

令 $x=\beta t$ 得到更一般形式.

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

α 决定了分布曲线的形状.

β 决定了曲线有多陡.

Gamma分布作为先验分布很强.

指数分布、χ^2分布都是特殊的Gamma分布.

指数分布、泊松分布、卫哥分布都是它的情人.

$\beta=1$ 泊松 $Poisson(X=k|\lambda)=\frac{\lambda^k e^{-\lambda}}{k!}$

↙ 离散的

即 Gamma分布令 $\alpha=k+1$

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

二项分布 $B(n,p)$ 在 $np=\lambda$, $n\to\infty$ 时的极限分布即是泊松分布

共轭分布：先验分布和后验分布的形式一样。

二项分布的共轭分布就是Beta分布。

二项分布 $B(k|n,p)=\binom{n}{k}p^k(1-p)^{n-k}$

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

后验分布

$$P(p|n,k,\alpha,\beta)\propto P(k|n,p)P(p|\alpha,\beta)$$
$$=P(k|n,p)P(p|\alpha,\beta)$$
$$=B(k|n,p)Beta(p|\alpha,\beta)$$

LDA. 文档 ⟶ 主题 ⟶ 主题词.

文档→主题 先验分布是 Dirichlet分布 $\theta_d = D(\vec{\alpha})$ α是K维向量.

主题→主题词 先验分布是 Dirichlet分布 $\beta_k = D(\vec{\eta})$ η是V维向量 V是词表中的所有词数

对于任一文档d中第n个词，从主题分布θ_d中得到它主题编号z_{dn}的分布

$$z_{dn} = multi(\theta_d)$$

而对于该主题编号，词w_{dn}的概率分布为

$$w_{dn} = multi(\beta_{z_{dn}})$$

文本建模 文档 对人来说 有序的词序列 $(d = w_1, w_2 \cdots w_n)$

统计文本建模 对机器 ⟶ 猜测上帝如何抛骰子的 { 什么样的骰子：模型参数，每一面的概率. 如何抛骰子：可能有不同类的骰子，根据规则抛骰

① 最简单无序模型. ①无序模型 ⟶ 假设词之间、文档之间独立可交换，词序不重要 ↘ 像一个袋子，也叫词袋模型

1) 一个骰子，V面，一面对应一个词，概率不同记 $\vec{p} = (p_1, p_2 \cdots p_V)$

2) 一篇文档有n个词，那么就是独立的抛了n次

这个分布服从多项分布：$w \sim Mult(w|\vec{p})$

对一篇文档 $d = \vec{w}(w_1, w_2 \cdots w_n)$

$p(\vec{w}) = p(w_1, w_2 \cdots w_n) = p(w_1)\, p(w_2) \cdots p(w_n)$

对多篇文档 $p(W) = p(\vec{w}_1)\, p(\vec{w}_2) \cdots p(\vec{w}_m)$

假设总词频为N，每个词v_i发生次数n_i

那么 $\vec{n} = (n_1, n_2 \cdots n_V)$ 正好是一个多项分布

$$p(\vec{n}) = Mult(\vec{n}|\vec{p}, N) = \binom{N}{\vec{n}} \prod_{k=1}^{V} p_k^{n_k}$$

语料概率 $p(W) = p(\vec{w}_1)\, p(\vec{w}_2) \cdots p(\vec{w}_m) = \prod_{k=1}^{V} p_k^{n_k}$.

用最大似然估计最大化$p(W)$ $\hat{p}_i = \frac{n_i}{N}$

② 贝叶斯无向模型 ②

1) 无穷个骰子$\vec{p}$每个骰子有V面
2) 上帝抽了一个骰子出来，用这一个骰子不断抛，产生了所有词.

有些类型的骰子多，有些少. 骰子$\vec{p}$服从分布$P(\vec{p})$

这个分布称为参数$\vec{p}$的先验分布.

每个骰子都可能被使用，由先验分布$P(\vec{p})$决定.
对一个具体的骰子$\vec{p}$，产生数据概率是$p(W|\vec{p})$.

则 最终 $p(W)=\int p(W|\vec{p})p(\vec{p})d\vec{p}$ 积分即加权求和

而注意到 $p(\vec{n})=Mult(\vec{n}|\vec{p},N)$ 实际上是计算一个多项分布的概率

所以先验分布的一个比较好的选择就是多项分布的对应分布.
即 Dirichlet分布. $Dir(\vec{p}|\vec{\alpha})=\frac{1}{\Delta(\vec{\alpha})}\prod_{k=1}^{V}p_k^{\alpha_k-1}$ $\vec{\alpha}=(\alpha_1,\cdots\alpha_V)$

$\Delta(\vec{\alpha})$就是归一化因子 $Dir(\vec{\alpha})$
即 $\Delta(\vec{\alpha})=\int\prod_{k=1}^{V}p_k^{\alpha_k-1}d\vec{p}$

Dirichlet先验 + 多项分布数据 → 后验分布也是Dirichlet分布

$$Dir(\vec{p}|\vec{\alpha})+Mult(\vec{n})=Dir(\vec{p}|\vec{\alpha}+\vec{n})$$

即 在给定参数$\vec{p}$的先验分布为$Dir(\vec{p}|\vec{\alpha})$时，
各个词出现的频次是$\vec{n}$~$Mult(\vec{n}|\vec{p},N)$为多项分布.
则后验分布是 $\vec{p}(\vec{p},W,\vec{\alpha})=Dir(\vec{p}|\vec{n}+\vec{\alpha})$

$$=\frac{1}{\Delta(\vec{n}+\vec{\alpha})}\prod_{k=1}^{V}p_k^{n_k+\alpha_k-1}d\vec{p}$$

$\vec{p}$如何估计呢？ $E(\vec{p})=\left(\frac{n_1+\alpha_1}{\sum_{i=1}^{V}(n_i+\alpha_i)},\frac{n_2+\alpha_2}{\sum_{i=1}^{V}(n_i+\alpha_i)},\cdots\frac{n_V+\alpha_V}{\sum(n_i+\alpha_i)}\right)$

$$\hat{p}_i=\frac{n_i+\alpha_i}{\sum_{i=1}^{V}(n_i+\alpha_i)} \Leftrightarrow \frac{\text{先验的计数}+\text{数据中的计数}}{\text{整体计数}}$$

文本语料产生的概率

$$P(W|\vec{\alpha}) = \int p(W|\vec{p})\, p(\vec{p}|\vec{\alpha})\, d\vec{p}$$

$$= \int \prod_{k=1}^{V} p_k^{n_k} \, Dir(\vec{p}|\vec{\alpha})\, d\vec{p}$$

(同一页)

$$= \int \prod_{k=1}^{V} p_k^{n_k} \frac{1}{\Delta(\vec{\alpha})} \prod_{k=1}^{V} p_k^{\alpha_k - 1}\, d\vec{p}$$

$$= \frac{1}{\Delta(\vec{\alpha})} \int \prod_{k=1}^{V} p_k^{n_k + \alpha_k - 1}\, d\vec{p}$$

$$= \frac{\Delta(\vec{n}+\vec{\alpha})}{\Delta(\vec{\alpha})}$$

③ PLSA Topic 模型

Hoffman 1999年 PLSA. (Probabilistic Latent Semantic Analysis)

1) 有两种骰子 { doc-topic K面 每面一个 topic
topic-word V面 每面一个 word. 有K个. 编号1-K

2) 先有一个特定的 doc-topic 骰子.
当掷得到 topic 编号 z
投掷编号z的骰子得到单词.

即文档之间独立可交换; 一个文档内的词独立可交换.

有K个 topic-word 骰子. 记 $\vec{\varphi}_1 \cdots \vec{\varphi}_K$
语料C有M篇文档 记 $d_1\ d_2 \cdots d_m$ 都对应一个 doc-topic 骰子 $\vec{\theta}_1 \cdots \vec{\theta}_m$

则第m篇文档中 d_m 里每个词生成的概率.

$$P(\vec{w}|d_m) = \prod_{i=1}^{n} \sum_{z=1}^{K} P(w_i|z)\, p(z|d_m) = \prod_{i=1}^{n} \sum_{z=1}^{K} \varphi_{z w_i} \theta_{d z}$$

求解可用 EM 算法. 参考 Hoffman 原文.

④ LDA
Latent Dirichlet Allocation

doc-topic的骰子和topic-word的骰子
都是参数 要有先验分布!!
即 $\vec{\varphi_k}$ 和 $\vec{\theta_m}$ 对应多项分布 先验是Dirichlet分布.