和文本处理相关的常用概率分布

学习要点

- Bernoulli - Binomial - Beta
- Categorical - Multinomial - Dirchilet
- 共轭分布

共轭分布

共轭定义

- 通常，给定一个分布$p(x|\theta)$，可以找到一个先验分布$p(\theta)$，与似然函数共轭，使得后验的函数形式与先验相同。

共轭作用

- 利于计算：贝叶斯模型的计算困难
- 解释：后验可以解释为先验的观察加到伪参数α, β上
- 先验-似然对通常使似然参数边缘化到一个封闭形式，使得似然的观察可以直接用超参数表示

二元变量

伯努利分布（bernoulli）

定义（define）

- 0-1分布，是一个离散型概率分布
- 抛硬币问题，1表示正面朝上，0表示反面朝上，则两面出现的概率分别为：

$$p(x=1|\theta)=\theta$$

$$p(x=0|\theta)=1-\theta$$

其中x∈{0,1}，$0\leqslant\theta\leqslant1$

- 则x的分布可以写为：

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

期望（mean）

$$\mathbb{E}[x]=\theta$$

方差（variance）

$$var[x]=\theta(1-\theta)$$

似然函数（likelihood）

对于一系列观察的x组成的集合，D={$x_1,\dots x_n$}，似然函数为

$$p(D|\theta)=\prod_{n=1}^{N}p(x_n|\theta)=\prod_{n=1}^{N}\theta^{x_n}(1-\theta)^{1-x_n}$$

$$\ln p(D|\theta)=\ln\prod_{n=1}^{N}\theta^{x_n}(1-\theta)^{1-x_n}=\sum_{n=1}^{N}[x_n\ln\theta+(1-x_n)\ln(1-\theta)]$$

对θ求导，得$\theta_{ML}=\frac{\sum_{n=1}^{N}x_n}{N}$

二项分布（binomial）

定义（define）

- 二项分布就是重复n次独立的伯努利实验，每次实验中有两种可能的结果，且两种结果发生与

否相互对立，与其它各次试验结果无关。试验次数为1时，二项分布服从0-1分布（伯努利分布）。

- N次抛硬币，其中正面朝上次数为m的概率可以表示为：

$$Bin(m|N,\theta) = \binom{N}{m} \theta^m (1-\theta)^{N-m}$$

其中$\binom{N}{m} \equiv \frac{N!}{(N-m)!m!}$

期望（mean）

$$\mathbb{E}[m] \equiv \sum_{m=0}^{N} mBin(m|N,\theta) = N\theta$$

方差（variance）

$$var[m] \equiv \sum_{m=0}^{N} (m - \mathbb{E}[m])^2 Bin(m|N,\theta) = N\theta(1-\theta)$$

注：对于独立事件，所有事件的期望等于各事件期望的总和，所有事件的方差等于各事件方差的总和。

Beta分布

定义（define）

- 伯努利分布和二项分布的共轭先验分布的密度函数，是一个定义在（0,1）区间的连续概率分布：

$$f(\theta;a,b) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{\int_0^1 u^{a-1}(1-u)^{b-1}du} = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}$$

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

$$\int_0^1 Beta(\theta|a,b)d\theta = 1$$

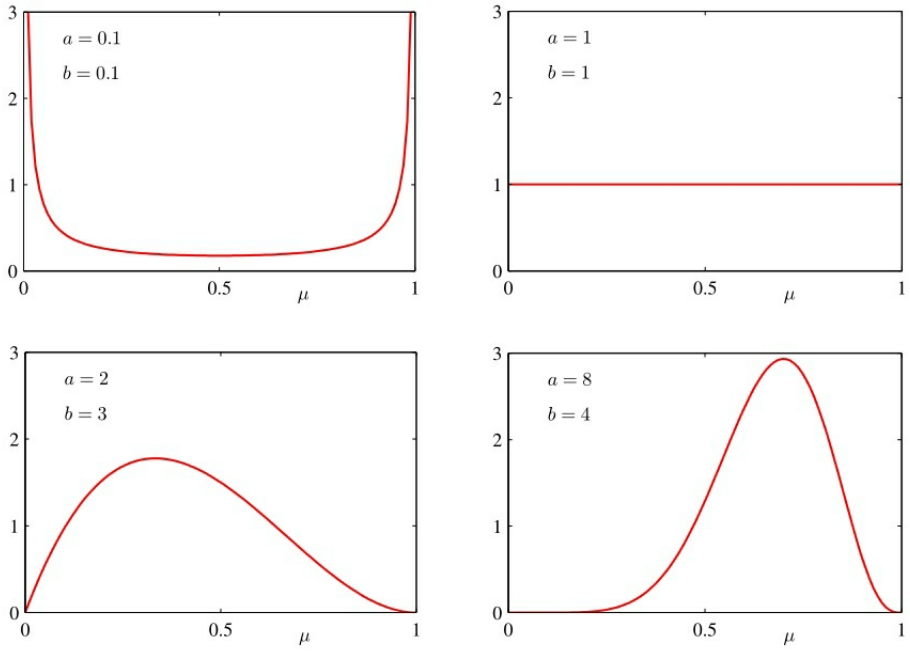
参数 a 和 b 经常被称为超参数（ hyperparameter ），它们控制了参数θ的概率分布

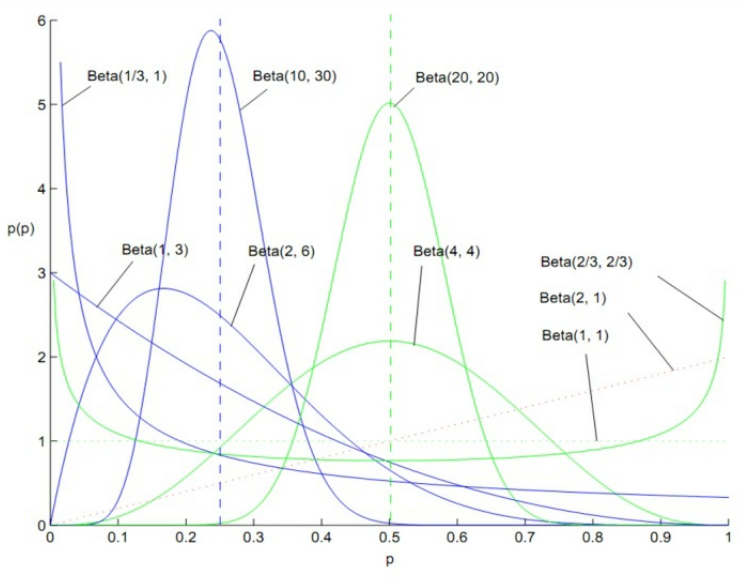
注：Gamma函数：也叫欧拉第二积分，是阶乘函数在实数与复数上扩展的一类函数，写作$\Gamma(x)$，实数域上Gamma函数定义为：

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

递归性：$\Gamma(x+1) = x\Gamma(x)$

对于不同的超参数a和b，Beta （θ|a,b） 关于θ的函数图像如下：





Beta分布的图像可以是凹的、凸的、单调上升的、单调下降的，可以是曲线也可以是直线，而

均匀分布也是特殊的Beta分布。因为Beta分布可以拟合许多形状，因此它在统计数据拟合和贝叶斯分析中被广泛使用。

期望（mean）

$$\mathbb{E}[\theta] = \int \theta \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1-\theta)^{b-1} d\theta$$

$$= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int \theta^{a}(1-\theta)^{b-1} d\theta$$

$$= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)}$$

$$= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{a\Gamma(a)\Gamma(b)}{(a+b)\Gamma(a+b)}$$

$$= \frac{a}{(a+b)}$$

似然函数

$$l(\theta) = \log \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1-\theta)^{b-1}$$

$$= \log \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} + (a-1)\log\theta + (b-1)\log(1-\theta)$$

对 θ求导

$$\frac{dl(\theta)}{d\theta} = (a-1)\frac{1}{\theta} - (b-1)\frac{1}{1-\theta} = 0$$

得 $\theta_{max} = \frac{a-1}{a+b-2}$

方差（variance）

$$var[\theta] = \mathbb{E}[\theta^2] - [\mathbb{E}[\theta]]^2$$

与 $\mathbb{E}[\theta]$同理，可得

$$\mathbb{E}[\theta^2] = \int \theta^2 \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1-\theta)^{b-1} d\theta$$

$$= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int \theta^{a+1}(1-\theta)^{b-1} d\theta$$

$$= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+2)\Gamma(b)}{\Gamma(a+b+2)}$$

$$= \frac{a(a+1)}{(a+b)(a+b+1)}$$

则方差为：

$$var[\theta] = \mathbb{E}[\theta^2] - [\mathbb{E}[\theta]]^2 = \frac{ab}{(a+b)^2(a+b+1)}$$

beta-bernoulli

先验（prior）

$$\theta \sim Beta(\theta; a, b)$$

似然函数（likelihood）

$$D = (X_1, \ldots, X_i, \ldots, X_N)$$

$$X_i \sim Bern(\theta)$$

$$p(D|\theta) = \theta^{N_1}(1-\theta)^{N_0}$$

其中$N_1 = \sum_{i=1}^{N} \mathbb{I}(x_i = 1)$，$N_0 = \sum_{i=1}^{N} \mathbb{I}(x_i = 0)$，$N_1 + N_0 = N$

后验（posterior）

$$p(\theta|D) = \frac{p(D,\theta)}{p(D)} = \frac{p(D|\theta)p(\theta)}{\int p(D|\theta)p(\theta)d\theta} \propto p(D|\theta)p(\theta)$$

$$p(\theta|D) \propto \theta^{N_1}(1-\theta)^{N_0}\theta^{a-1}(1-\theta)^{b-1} = \theta^{N_1+a-1}(1-\theta)^{N_0+b-1}$$

$$\int \theta^{N_1+a-1}(1-\theta)^{N_0+b-1}d\theta = \frac{\Gamma(N_1+a)\Gamma(N_0+b)}{\Gamma(a+b+N_1+N_0)}$$

由$\int p(\theta|D)d\theta = 1$得归一化参数

$$C = \frac{\Gamma(a+b+N_1+N_0)}{\Gamma(a+N_1)\Gamma(b+N_0)}$$

所以后验为：

$$p(\theta|D) = \frac{\Gamma(a+b+N_1+N_0)}{\Gamma(a+N_1)\Gamma(b+N_0)}\theta^{N_1+a-1}(1-\theta)^{N_0+b-1}$$

即 $Beta(\theta; N_1 + a, N_0 + b)$

由Beta分布可得后验的期望为：

$$\mathbb{E}[\theta|D] = \frac{a + N_1}{a + b + N_1 + N_0}$$

方差为：

$$var[\theta|D] = \frac{(a + N_1)(b + N_0)}{(a + N_1 + b + N_0)^2 (a + N_1 + b + N_0 + 1)}$$

令 $N = N_1 + N_0$ ， $M = a + b$ ， $\lambda = \frac{M}{M+N}$ 则

$$\mathbb{E}[\theta|D] = \frac{\frac{N_1}{N} N + \frac{a}{M} M}{M + N} = \frac{M}{M + N} \frac{N_1}{N} + \frac{N}{M + N} \frac{a}{M} = \lambda \frac{a}{M} + (1 - \lambda) \frac{N_1}{N}$$

- 后验均值是先验均值和MLE的凸组合，也就是后验是我们之前所认为的与数据告诉我们之间的折衷。
- the posterior mean is convex combination of the prior mean and the MLE,which captures the notion that the posterior is a compromise between what we previously believed and what the data is telling us.
- 先验越弱，λ越小，因此后验均值越接近MLE。可以类似地表明后验模式是先验模式和MLE的凸组合，并且它也收敛到 MLE。
- the weaker the prior,the smaller is λ,and hence the closer the posterior mean is to the MLE.One can show similarly that the posterior mode is convex combination of the prior mode and the MLE,and that it too converges to the MLE.

后验预测（predict）

给定数据集D的情况下，可以预测x的分布，根据加和规则和乘积规则可得

$$p(X=1|D)=\int_0^1 p(X=1|\theta)p(\theta|D)d\theta$$

$$=\int_0^1 \theta p(\theta|D)d\theta$$

$$=\mathbb{E}[\theta|D]$$

$$=\frac{a+N_1}{a+N_1+b+N_0}$$

多元变量

multinoulli

定义（define）

将二项式中的两种状态推广至多种状态，即可得到多项式分布

二元变量变为K元变量，此时x为k维向量，其中仅有一个元素为1，其余元素为0，x分布为：

$$p(x|\theta)=\prod_{k=1}^{K}\theta_k^{x_k}$$

其中$x=(x_1,\ldots,x_K)^T$，$\theta=(\theta_1,\ldots,\theta_K)^T$，$\theta_k\geqslant 0$，$\sum_k\theta_k=1$，$\sum_k x_k=1$

期望（mean）

$$\mathbb{E}[x|\theta]=\sum_x p(x|\theta)x=(\theta_1,\ldots,\theta_M)^T=\theta$$

似然函数（likelihood）

$$p(D|\theta) = \prod_{n=1}^{N}\prod_{k=1}^{K}\theta_k^{x_{nk}} = \prod_{k=1}^{K}\theta_k^{\sum_n x_{nk}} = \prod_{k=1}^{K}\theta_k^{m_k}$$

其中$m_k = \sum_n x_{nk}$

为了找到θ的最大似然解，我们需要关于θ_k最化$\ln p(D|\theta)$，并且要限制θ_k的和必须等于1。这可以通过拉格朗乘数λ实现，即最大化

$$\sum_{k=1}^{K} m_k \ln \theta_k + \lambda(\sum_{k=1}^{K}\theta_k - 1)$$

对θ_k求导，得到$\theta_k = -m_k/\lambda$

将$\sum_{k=1}^{K}\theta_k = 1$代入，得到$\lambda = -N$

最大似然解$\theta_k^{ML} = \frac{m_k}{N}$

multinomial

定义（define）

$m_1, \ldots m_k$在θ、N下的联合分布

$$Mult(m_1, m_2, \ldots, m_K|\theta, N) = \binom{N}{m_1, m_2, \ldots, m_K}\prod_{k=1}^{K}\theta_k^{m_k}$$

其中$\binom{N}{m_1, m_2, \ldots, m_K} = \frac{N!}{m_1!m_2!\ldots m_K!}$

狄利克雷分布（dirichlet）

定义（define）

连续多变量的概率分布，可做多项分布的先验概率

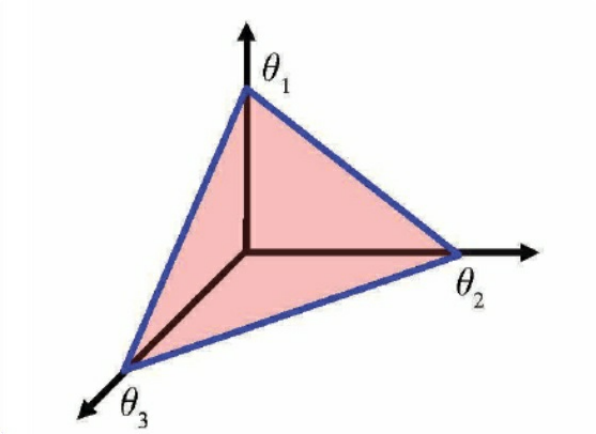
$$\alpha = (\alpha_1, \ldots, \alpha_K)$$

$$\theta = (\theta_1, \ldots, \theta_K)$$

$$p(\theta;\alpha) = Dir(\theta;\alpha) = \frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\prod_{k=1}^{K}\theta_k^{\alpha_k-1}$$

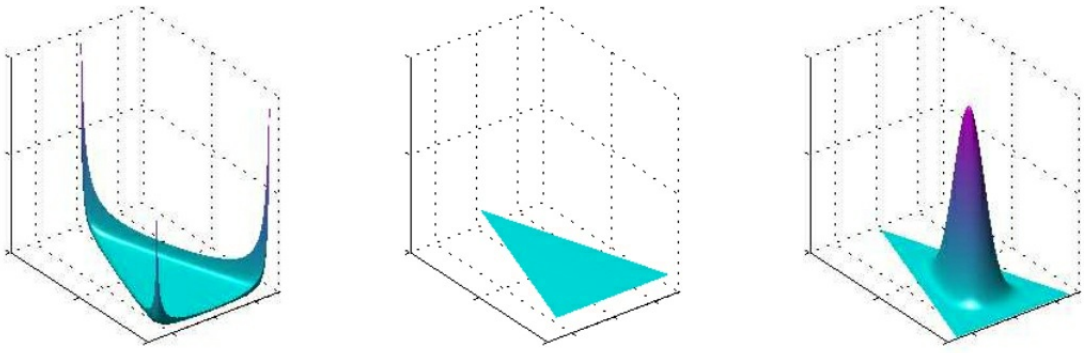
simplex

- 定义在θ_1、θ_2、θ_3上的狄利克雷分布被限制在一个单纯形中，这是由于$0 \leqslant \theta_k \leqslant 1$和$\sum_k \theta_k = 1$



造成的。

不同的参数θ_k的情况下，单纯形上的狄利克雷分布的图像。（下面$\{\theta_k\}=0.1$对应于左图，$\{\theta_k\}=1$对应于中图，$\{\theta_k\}=10$对应于右图。）



狄利克雷的概率是多项分布的概率的概率，所以狄利克雷是分布的分布。共轭先验是一种概率密度，使后验概率的密度函数和先验概率的密度函数有相同函数形式。Dirichlet分布就是多项分布的共轭先验分布。先验概率取为共轭先验的好处就在于：每当有新的观测数据，就把上次的后验概率作为先验概率，乘以新数据的likelihood，然后就得到新的后验概率，而不必用先验

概率乘以所有数据的likelihood得到后验概率。

期望（mean）

$$\mathbb{E}[\theta_k] = \int \theta_k Dir(\theta;\alpha)d\theta = \int \theta_k \frac{\Gamma(\sum_{i=1}^K \alpha_i)}{\prod_{i=1}^K \Gamma(\alpha_i)} \prod_{i=1}^K \theta_i^{\alpha_i - 1} d\theta$$

$$= \frac{\Gamma(\sum_{i=1}^K \alpha_i)}{\prod_{i=1}^K \Gamma(\alpha_i)} \int \theta_k^{(\alpha_k - 1)+1} \prod_{i \neq k} \theta_i^{\alpha_i - 1} d\theta$$

$$= \frac{\Gamma(\sum_{i=1}^K \alpha_i)}{\prod_{i=1}^K \Gamma(\alpha_i)} \int \theta_k^{(\alpha_k + 1)-1} \prod_{i \neq k} \theta_i^{\alpha_i - 1} d\theta$$

$$\int \theta_k^{(\alpha_k + 1)-1} \prod_{i \neq k} \theta_i^{\alpha_i - 1} d\theta = \frac{\prod_{i \neq k} \Gamma(\alpha_i)\Gamma(\alpha_k + 1)}{\Gamma(1 + \sum_{i=1}^K \alpha_i)}$$

$$= \frac{\alpha_k \Gamma(\alpha_k) \prod_{i \neq k} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^K \alpha_i) \sum_{i=1}^K \alpha_i}$$

$$= \frac{\alpha_k \prod_{i=1}^K \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^K \alpha_i) \sum_{i=1}^K \alpha_i}$$

所以

$$\mathbb{E}[\theta_k] = \frac{\Gamma(\sum_{i=1}^K \alpha_i)}{\prod_{i=1}^K \Gamma(\alpha_i)} \frac{\alpha_k \prod_{i=1}^K \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^K \alpha_i) \sum_{i=1}^K \alpha_i}$$

$$= \frac{\alpha_k}{\sum_{i=1}^K \alpha_i}$$

似然

$$l(\theta) = \log p(\theta; a, b) = \log \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}$$

$$= \log \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} + \sum_{k=1}^{K} (\alpha_k - 1) \log \theta_k$$

因为$\sum_{k=1}^{K} \theta_k = 1$，可得到：

$$l(\theta) = \log \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} + \sum_{k=1}^{K} (\alpha_k - 1) \log \theta_k + \lambda(1 - \sum_{k=1}^{K} \theta_k)$$

对 θ_k 求导

$$\frac{\partial l(\theta)}{\partial \theta_k} = \frac{\alpha_k - 1}{\theta_k} - \lambda = 0$$

得 $\theta_k = \frac{\alpha_k - 1}{\lambda}$

将 $\sum_{k=1}^{K} \theta_k = 1$ 代入

$$\sum_{k=1}^{K} \frac{\alpha_k - 1}{\lambda} = 1$$

$$\lambda = \sum_{k=1}^{K} (\alpha_k - 1)$$

所以

$$\theta_{k(max)} = \frac{\alpha_k - 1}{(\sum_{k=1}^{K} \alpha_k) - K}$$

方差（variance）

由 $\mathbb{E}[\theta_k]$，同理可得

$$\mathbb{E}[\theta_k^2] = \int \theta_k^2 Dir(\theta;\alpha)d\theta$$

$$= \int \theta_k^2 \frac{\Gamma(\sum_{i=1}^K \alpha_k)}{\prod_{i=1}^K \Gamma(\alpha_i)} \prod_{i=1}^K \theta_i^{\alpha_i - 1} d\theta$$

$$= \frac{\Gamma(\sum_{i=1}^K \alpha_k)}{\prod_{i=1}^K \Gamma(\alpha_i)} \int \theta_k^{(\alpha_k - 1)+2} \prod_{i \neq k} \theta_i^{\alpha_i - 1} d\theta$$

$$= \frac{\Gamma(\sum_{i=1}^K \alpha_k)}{\prod_{i=1}^K \Gamma(\alpha_i)} \int \theta_k^{(\alpha_k + 2)-1} \prod_{i \neq k} \theta_i^{\alpha_i - 1} d\theta$$

$$= \frac{\Gamma(\sum_{i=1}^K \alpha_k)}{\prod_{i=1}^K \Gamma(\alpha_i)} \frac{\prod_{i \neq k} \Gamma(\alpha_i)\Gamma(\alpha_k + 2)}{\Gamma(2 + \sum_{i=1}^K \alpha_i)}$$

$$= \frac{\Gamma(\sum_{i=1}^K \alpha_k)}{\prod_{i=1}^K \Gamma(\alpha_i)} \frac{\alpha_k(\alpha_k + 1)\Gamma(\alpha_k) \prod_{i \neq k} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^K \alpha_i) \sum_{i=1}^K \alpha_i (1 + \sum_{i=1}^K \alpha_i)}$$

$$= \frac{\Gamma(\sum_{i=1}^K \alpha_k)}{\prod_{i=1}^K \Gamma(\alpha_i)} \frac{\alpha_k(\alpha_k + 1) \prod_{i=1}^K \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^K \alpha_i) \sum_{i=1}^K \alpha_i (1 + \sum_{i=1}^K \alpha_i)}$$

$$= \frac{\alpha_k(\alpha_k + 1)}{\sum_{i=1}^K \alpha_i (1 + \sum_{i=1}^K \alpha_i)}$$

则方差为

$$var(\theta_k) = \frac{\alpha_k(\alpha_k+1)}{\sum_{i=1}^{K}\alpha_i(1+\sum_{i=1}^{K}\alpha_i)} - (\frac{\alpha_k}{\sum_{i=1}^{K}\alpha_i})^2$$

$$= \frac{\alpha_k((\sum_{i=1}^{K}\alpha_i)-\alpha_k)}{(\sum_{i=1}^{K}\alpha_i)^2(1+\sum_{i=1}^{K}\alpha_i)}$$

dirichlet-multinoulli

先验（prior）

$$\theta \sim Dir(\theta;\alpha)$$

似然函数（likelihood）

$$D = (X_1,\ldots,X_i,\ldots,X_N)$$

$$X_i \sim Multinoulli(\theta)$$

$$p(D|\theta) = \prod_{k=1}^{K}\theta_k^{N_k}$$

其中$\sum_{k=1}^{K}N_k = N$

后验（posterior）

$$p(\theta|D) = \frac{p(D,\theta)}{p(D)} = \frac{p(D|\theta)p(\theta)}{\int p(D|\theta)p(\theta)} \propto p(D|\theta)p(\theta)$$

$$p(\theta|D) \propto \prod_{k=1}^{K} \theta_k^{N_k} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} = \prod_{k=1}^{K} \theta_k^{N_k + \alpha_k - 1}$$

$$\int \prod_{k=1}^{K} \theta_k^{N_k + \alpha_k - 1} = \frac{\prod_{k=1}^{K} \Gamma(N_k + \alpha_k)}{\Gamma(\sum_{k=1}^{K} (N_k + \alpha_k))}$$

由 $\int p(\theta|D) d\theta = 1 = \int C \prod_{k=1}^{K} \theta_k^{N_k + \alpha_k - 1} d\theta$，得C

$$C = \frac{\Gamma(\sum_{k=1}^{K} (N_k + \alpha_k))}{\prod_{k=1}^{K} \Gamma(N_k + \alpha_k)}$$

从而得到后验

$$p(\theta|D) = \frac{\Gamma(\sum_{k=1}^{K} (N_k + \alpha_k))}{\prod_{k=1}^{K} \Gamma(N_k + \alpha_k)} \prod_{k=1}^{K} \theta_k^{N_k + \alpha_k - 1}$$

即 $Dir(\theta; N_1 + \alpha_1, \ldots, N_K + \alpha_K)$，由狄利克雷分布，可得后验的期望为：

$$\mathbb{E}[\theta_k|D] = \frac{\alpha_k + N_k}{\sum_{i=1}^{K} (\alpha_i + N_i)}$$

设 $M = \sum_{k=1}^{K} \alpha_k$， $N = \sum_{k=1}^{K} N_k$， $\lambda = \frac{M}{M+N}$，同理可得

$$\mathbb{E}[\theta_k|D] = \frac{\frac{\alpha_k}{M} \times M + \frac{N_k}{N} \times N}{M + N}$$

$$= \frac{M}{M+N} \times \frac{\alpha_k}{M} + \frac{N}{M+N} \times \frac{N_k}{N}$$

$$= \lambda \frac{\alpha_k}{M} + (1 - \lambda) \frac{N_k}{N}$$

后验预测（predict）

$$p(X=j|D)=\int_0^1 p(X=j,\theta|D)d\theta$$

$$=\int p(X=j|\theta,D)p(\theta|D)d\theta$$

$$=\int p(X=j|\theta)p(\theta|D)d\theta$$

$$=\int \theta_j Dir(\theta;N_1+\alpha_1,\dots N_k+\alpha_k)d\theta$$

$$=\mathbb{E}\left[\theta_j|D\right]=\frac{\alpha_j+N_j}{\sum_{k=1}^K(\alpha_k+N_k)}$$

参考文献

- PRML 2.1/2.2/2.4.2
- Parameter estimation for text analysis 中的 **3 Conjugate distributions** http://www.arbylon.net/publications/text-est2.pdf
- 王俊博士的笔记 https://www.scribd.com/document/388021540/Notes-on-Beta-and-Dirchilet-Distribution
- MLaPP 3.3/3.4