电子科技大学
University of Electronic Science and Technology of China

A Bayesian View into Dirichlet and LDA

Chongming Gao

Data Mining Lab, Big Data Research Center, UESTC
Email：junmshao@uestc.edu.cn
http://staff.uestc.edu.cn/shaojunming

Preamble

关于Dirichlet Distribution, 晨晨、于伯伯都讲过。
然而我却似懂非懂。我问其他人：
方芳：完了，数学是完了，看不懂啊
左爷：我也不懂，求教！

当然我们都知道左爷是装的。但这个问题着实值得仔细分析，因为对于机器学习，Dirichlet和LDA模型只是一个非常基本的坎。

所以，我分析了很久很久，发现了《数学八卦》的缺陷，今天目的是从另一个角度讲解Dirichlet、LDA。

Preamble

➢The Rules of Probability

Sum rule $p(X) = \sum_{Y} p(X, Y)$ $p(x) = \int p(x, y)\,\mathrm{d}y$

Product rule $p(X, Y) = p(Y|X)p(X).$ $p(x, y) = p(y|x)p(x)$

Expectation $\mathbb{E}[f] = \sum_{x} p(x)f(x)$ $\mathbb{E}[f] = \int p(x)f(x)\,\mathrm{d}x.$

Bayes' theorem $p(Y|X) = \dfrac{p(X|Y)p(Y)}{p(X)} = \dfrac{p(X|Y)p(Y)}{\sum_Y p(X|Y)p(Y)}$

Posterior Probability

Prior Probability

Preamble

- Example: Coin Tossing Game

Actually, all theories related to this topic can be explained in the game of **coin tossing.**

1. Bayesian

➢Basic Bayes' Formula

Remember the Bayes' Theorem we learned in the probability course? Toss one of two unfair coins, given the probability:

Coin Number	Head	Tail
1	0.8	0.2
2	0.4	0.6

Suppose the probability we chose the No.1 coin is 0.3

Question: We have tossed a coin and obtain a result of Head, What's the probability of the coin is No.1?

1. Bayesian

➢Basic Bayes' Formula

Event Y: Flipping result is Head

Event Z: The coin used is No.1

$$
\begin{aligned}
p(Z|Y) &= \frac{p(Y|Z)p(Z)}{p(Y)} \\
&= \frac{p(Y|Z)p(Z)}{p(Y|Z)p(Z) + p(Y|\neg Z)p(\neg Z)} \\
&= \frac{0.8 \cdot 0.3}{0.8 \cdot 0.3 + 0.4 \cdot 0.7} \\
&= 0.46
\end{aligned}
$$

Essence: Y is observation variable, Z is latent variable

1. Bayesian

➢Bayesian VS. Frequentist

Bayes' formula is a mathematic theorem.

Bayesian, however, is a special school(学派) on the opposite stance of **Frequentist**

两学派各有其信仰、内在逻辑、解释力和局限性，从20世纪上半页至今，两大学派的辩论从未停歇，但分歧如故。

贝叶斯学派的发展在二十世纪滞后于频率学派，甚至现今主流统计学教材仍然以频率学派的理论框架为主，贝叶斯理论通常一笔带过。

1. Bayesian

➢Bayesian and Frequentist

The divergence of two school up to concepts of **philosophy**. On the issue of **Parameter Estimation**, the divergence lies in that:

Problem Definition:

What's the probability of a coin lands head?

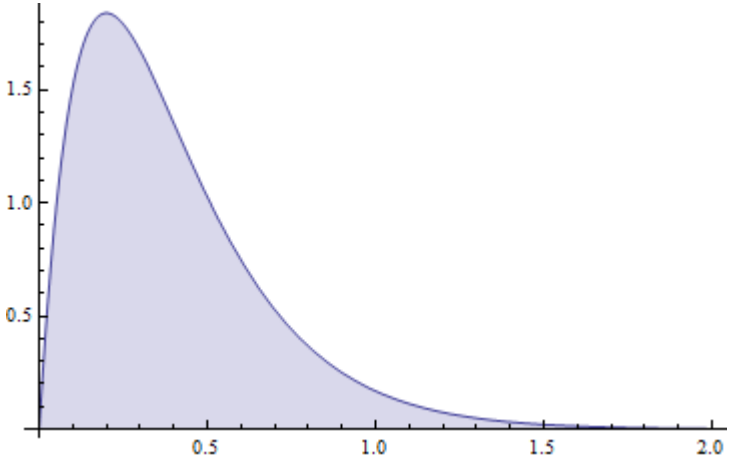
1. Bayesian

What's the probability of a coin lands head?

To **Frequentist**, the probability is a constant C, which can be estimated by numerous experiments, that is

$$C = \frac{n}{N} = \frac{\text{Number of Heads landed}}{\text{Total number of experiments}}$$

To **Bayesian**, there exists no constant, but only the observed data generated uncertainly. Parameters are only the measures of uncertainty!



That is to say, parameters are not constants, they also have probability distributions

1. Bayesian

Bayesian View

$\boldsymbol{p} = (p1, p2)$, is the parameters capture all the uncertainty in this event.

$p1$ denotes the probability of coin lands head
$p2$ denotes the probability of coin lands tail

The random variables denoted by Y is the observation of a toss.

After several experiments, the observation serials are denoted as $\boldsymbol{y} = (y1, y2, y3, \cdots yn)$

$$p(\boldsymbol{p}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{p}) \cdot p(\boldsymbol{p})}{p(\boldsymbol{y})} \propto p(\boldsymbol{y}|\boldsymbol{p}) \cdot p(\boldsymbol{p})$$

$$p(\boldsymbol{p}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{p}) \cdot p(\boldsymbol{p})$$

$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

Maximum Likelihood Estimation(*MLE*) is to maximize the blue part

In the Bayesian framework, we maximize the posterior. This approach named **Maximum posterior**(*MAP*)

1. Bayesian

Bayesian View

$$p(\boldsymbol{p}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{p}) \cdot p(\boldsymbol{p})$$

Suppose, for instance, that a fair-looking coin is tossed 3 times and lands heads each time.

A **classical** maximum likelihood estimate of theprobability of landing heads would give 1, implying that all future tosses will land heads!

By contrast, a **Bayesian** approach with any reasonable prior will lead to a much less extreme conclusion.

Discussion: Advantages and Drawbacks of Bayesian

$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

Advantages

- Reasonable prior
- Unrepeatable experiment

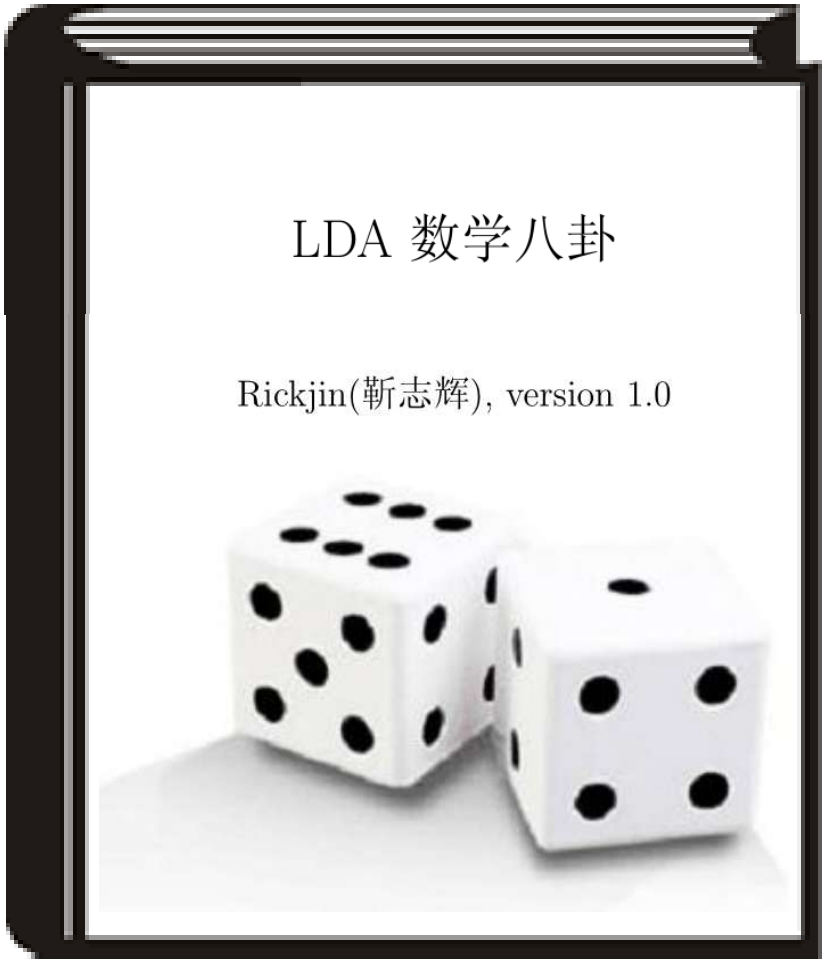
Drawbacks

- Hard to define the prior

More discussion can be found on 知 https://www.zhihu.com/question/20587681

Choose of Prior

谈到选择先验分布，是时候再回到我们的启蒙经典了！



2. Beta &Dirichlet

游戏：猜数字

问题1：

“随机产生10个(0,1)之间的数字，猜一猜第7大的数字X_7是多少？”

Algorithm 1 游戏1

1: $X_1, X_2, \cdots, X_n \overset{\text{iid}}{\sim} Uniform(0,1)$,
2: 把这n 个随机变量排序后得到顺序统计量$X_{(1)}, X_{(2)} \cdots, X_{(n)}$,
3: 问$X_{(k)}$ 的分布是什么

问题2：

“再随机产生5个(0,1)之间的数字，告诉你这五个数字与刚刚的X_7相比谁大谁小，你再猜一猜X_7是多少？”

Algorithm 2 游戏2

1: $X_1, X_2, \cdots, X_n \overset{\text{iid}}{\sim} Uniform(0,1)$，排序后对应的顺序统计量$X_{(1)}, X_{(2)} \cdots, X_{(n)}$, 我们要猜测$p = X_{(k)}$;
2: $Y_1, Y_2, \cdots, Y_m \overset{\text{iid}}{\sim} Uniform(0,1)$, Y_i中有m_1个比p小，m_2个比p大;
3: 问$P(p|Y_1, Y_2, \cdots, Y_m)$ 的分布是什么。

2. Beta &Dirichlet

类比游戏：抛硬币

问题1：
有一个均匀度为知的硬币，抛了9次，其中有6次为正面，3次为反面。请问，硬币正面向上的概率p？

问题2
再用这个硬币抛5次，告诉你这五次正反面的结果，再问你这个硬币正面向上的概率p？

2. Beta &Dirichlet

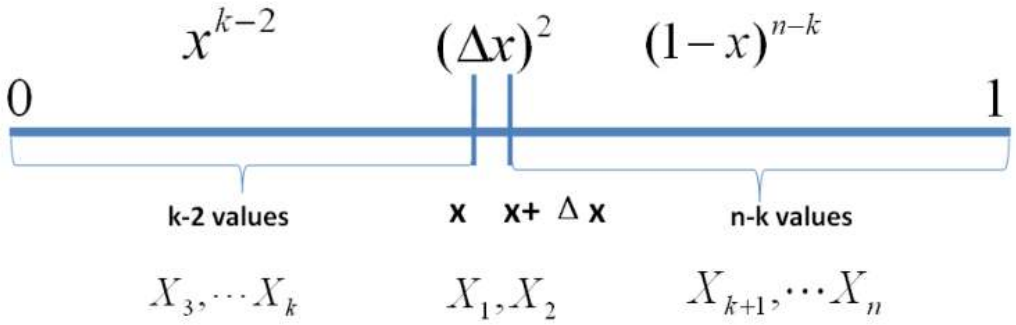
猜数字 VS. 抛硬币

“随机产生10个(0,1)之间的数字，猜一猜第7大的数字X_7是多少？”

“再随机产生5个(0,1)之间的数字，告诉你这五个数字与刚刚的X_7相比谁大谁小，你再猜一猜X_7是多少？”

有一个均匀度为知的硬币，抛了9次，其中有6次为正面，3次为反面。请问，硬币正面向上的概率p？

再用这个硬币抛5次，告诉你这五次正反面的结果。再问你这个硬币正面向上的概率p？



由于产生随机数的次序是可以交换的，猜数字问题其实就是抛硬币问题！

2. Beta &Dirichlet

“峰芳会谈”

对于猜硬币正面向上的概率*p*这个问题：Frequentist与Bayesian各自派出代表，展开了**幼稚的争论**，历史上称为“峰芳会谈”。

Frequentist代表：Fengfeng

“这个好简单，硬币正面向上的概率不就是$\frac{6}{9}$嘛，再做5次试验，假如正面向上3次，反面2次，那硬币正面向上的概率不就是$\frac{9}{14}$嘛！”

Bayesian代表：Fangfang

“别幼稚了，太naïve了！这样考虑太，应该考虑一个分布才对！把硬币正面朝上的概率设为p，**我们应该考虑一个关于p的概率密度函数！**”

峰芳会谈：结果

由于峰峰的**思维方式**太过幼稚，
我们采用方芳的思路！即贝叶斯理论。

2. Beta &Dirichlet

猜数字 & 抛硬币

Now，终于发现它们的**实质：求参数的分布！**

$$p(\boldsymbol{p}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{p}) \cdot p(\boldsymbol{p})$$

两个游戏的第一个问题，求的是 $p(\boldsymbol{p})$ 即参数 $\boldsymbol{p}$ 的**prior分布**，这个时候我们认为还没有做实验（没有观测数据），6个正面，3个反面是**超参数(计数)**，是我们强行定义的。也可以说是根据专家知识(knowledge)赋予的值。

两个游戏的第二个问题，求的是 $p(\boldsymbol{p}|\boldsymbol{y})$ 即参数 $\boldsymbol{p}$ 的**posterior分布**，这时候我们得到观测数据**(计数)**：3个正面和2个反面。此时加上先验的超参数，总共就有9个正面和5个反面了！

2. Beta &Dirichlet

进一步理解抛硬币

对于我这种初接触Posterior的人，总感觉哪里怪怪的。为什么现在抛硬币游戏和以前感觉完全不一样了？

通常的：

给定参数**p**，求数据**y**产生的概率：

如峰峰给定**p**=(p(正面向上), p(反面向上))=(0.4,0.6)，或者方芳给一个**p**的概率密度p(**p**)，求硬币序列为**y** = {正、正、反、正...}的概率。

现在的：

1. 给定某种先验知识，也就是超参数α和β，求参数**p**的Prior分布
2. 做实验给定实验数据**y**，结合超参数α和β，求参数**p**的Posterior分布

如刚刚我们讨论的，其中$\alpha = 6$个正面，$\beta = 3$个反面，可以得到**p**的Prior分布，再加上后面5个实验的3个正面，2个反面，又可以得到一个**p**分布

一个是给定参数分布求观测变量序列分布概率，一个给定观测变量序列分布求参数分布概率！

2. Beta &Dirichlet

抛硬币中的先验、后延分布

再看先验分布和后验分布，从直觉上我们发现他们完全一样，是同一个硬币产生的结果，只不过数目不一样！分别是6个正面和3个反面，以及9个正面和5个反面。

所以，我们在这里就断定：
Posterior 与 Prior 的**分布类型**是相同的！

Generally Speaking， Posterior与Prior不要求其分布类型是相同的，但如果是相同的，将会给我们的求解过程带来非常大的方便！

2. Beta &Dirichlet

抛硬币&猜数字中的先验、后验分布

1.先求先验分布：

猜硬币概率

已知硬币的超参数为$\alpha = 6$个正面，$\beta = 3$个反面，求其正面向上的概率p1分布。

$$p(p_1|\alpha,\beta) = \binom{\alpha+\beta}{\alpha,\beta} p_1^{\alpha}(1-p_1)^{\beta}$$
$$= \frac{(\alpha+\beta)!}{\alpha!\beta!} p_1^{\alpha}(1-p_1)^{\beta}$$

猜数字

同理，对于猜数字，一开始有10个(0,1)之间的数字，求第7大的数字x_7的分布，认为超参数为$\alpha = 6$个比x_7小，$\beta = 3$个比x_7大。则x_7的分布为：

$$p(x_7|\alpha,\beta) = \binom{\alpha+\beta+1}{\alpha,\beta} x_7^{\alpha}(1-x_7)^{\beta}$$
$$= \frac{(\alpha+\beta+1)!}{\alpha!\beta!} x_7^{\alpha}(1-x_7)^{\beta}$$

2. Beta &Dirichlet

抛硬币中的先验分布

有一个非常神奇的函数，Gamma函数，有这么一个性质：

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

对于猜数字，刚才有$\alpha = 6$，$\beta = 3$。现在令$\alpha = 6+1$，$\beta = 3+1$。
则原式子可以写成：

$$\begin{aligned} p(x_7|\alpha,\beta) &= \binom{\alpha+\beta-1}{\alpha-1,\beta-1} x_7^{\alpha-1}(1-x_7)^{\beta-1} \\ &= \frac{(\alpha+\beta-1)!}{(\alpha-1)!(\beta-1)!} x_7^{\alpha-1}(1-x_7)^{\beta-1} \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x_7^{\alpha-1}(1-x_7)^{\beta-1} \end{aligned}$$

$$Beta(x_7|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x_7^{\alpha-1}(1-x_7)^{\beta-1}$$

2. Beta &Dirichlet

先验、后验分布

1.先验分布：

对于猜数字，有$\alpha = 6 + 1$，$\beta = 3 + 1$。x_7分布为：

$$Beta(x_7|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x_7^{\alpha-1}(1-x_7)^{\beta-1}$$

2.后验分布：

对于猜数字，有$\alpha + m_1 = 6 + 3 + 1$，$\beta + m_1 = 3 + 2 + 1$。x_7分布为：

$$Beta(x_7|\alpha+m_1,\beta+m_2) = \frac{\Gamma(\alpha+m_1+\beta+m_2)}{\Gamma(\alpha+m_1)\Gamma(\beta+m_2)}x_7^{\alpha+m_1-1}(1-x_7)^{\beta+m_2-1}$$

其中$\Gamma(x) = (x-1)!$

2. Beta &Dirichlet

数据分布

现在给定参数x_7分布，则求数据的分布为一个普通的二项分布问题：

求后五次试验三个比x_7小，两个比x_7大的$m_1 = 3$，$m_2 = 2$的数据分布：

$$\begin{aligned} BinomCount(m_1, m_2|x_7) &= \binom{m_1+m_2}{m_1, m_2} x_7^{m_1}(1-x_7)^{m_2} \\ &= \frac{(m_1+m_2)!}{m_1!m_2!} x_7^{m_1}(1-x_7)^{m_2} \end{aligned}$$

结论

比对之前参数的分布，可以得到《数学八卦》上的结论。注意那个加号不是做加法运算，而是一种表示方法：

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

先验分布+ 数据的知识= 后验分布

这就是定义的所谓的Beta-Binomial 共轭，**即数据知识的分布使得参数的先验、后验分布类型一样。**

2. Beta &Dirichlet

从Beta-Binumial共轭到Dirichlet-Multinomial共轭

问题变得非常简单，原问题只要转化一下:

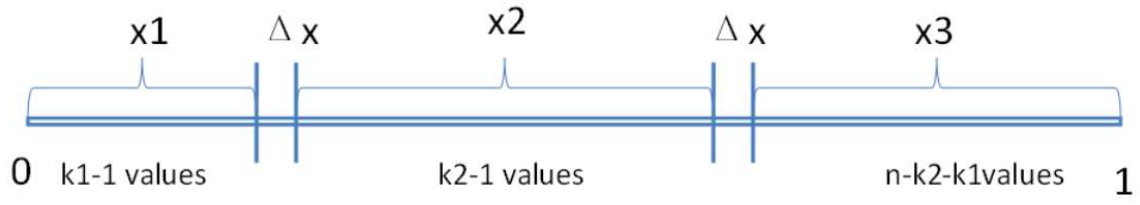
猜数字 VS. 抛硬币

“随机产生10个(0,1)之间的数字，猜一猜第3大和第7大的数字X_3和X_7分别是多少？”

“再随机产生5个(0,1)之间的数字，告诉你这五个数字与刚刚的X_3和X_7相比谁大谁小，你再猜一猜X_3和X_7是多少？”

有一个均匀度为知的筛子，有三个面！抛了9次，其中有2次为第一面，3次为第二个面，2次为第三面。请问，筛子三个面各自概率$\boldsymbol{p}=(p_1,p_2,p_2)$？

再用这个筛子抛5次，告诉你这五次三个面的结果。再问你这个筛子三个面的概率$\boldsymbol{p}=(p_1,p_2,p_2)$？



2. Beta &Dirichlet

从Beta到Dirichlet的参数分布

刚刚的Beta函数关于x_7的先验、后验分布

先验

$$Beta(x_7|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x_7^{\alpha-1}(1-x_7)^{\beta-1}$$

其中$\Gamma(x) = (x-1)!$

后验

$$Beta(x_7|\alpha+m_1,\beta+m_2) = \frac{\Gamma(\alpha+m_1+\beta+m_2)}{\Gamma(\alpha+m_1)\Gamma(\beta+m_2)}x_7^{\alpha+m_1-1}(1-x_7)^{\beta+m_2-1}$$

现在的Dirichlet函数关于x_3，x_7的先验、后验分布

先验

$$Dir(x_3,x_7|\alpha1,\alpha2,\alpha3) = \frac{\Gamma(\alpha1+\alpha2+\alpha3)}{\Gamma(\alpha1)\Gamma(\alpha2)\Gamma(\alpha3)}x_3^{\alpha1-1}(x_7-x_3)^{\alpha2-1}(1-x_7)^{\alpha3-1}$$

其中$\Gamma(x) = (x-1)!$

后验

$$Dir(x_3,x_7|\alpha1+m_1,\alpha2+m_2,\alpha3+m_3) = \frac{\Gamma(\alpha1+m_1+\alpha2+m_2+\alpha3+m_3)}{\Gamma(\alpha1+m_1)\Gamma(\alpha2+m_2)\Gamma(\alpha3+m_3)}x_3^{\alpha1+m_1-1}(x_7-x_3)^{\alpha2+m_2-1}(1-x_7)^{\alpha3+m_3-1}$$

2. Beta &Dirichlet

从Binomial到MultiNomial的数据分布

对比地得出观测数据的分布

二项分布

$$BinomCount(m_1, m_2|x_7) = \binom{m_1+m_2}{m_1, m_2} x_7^{m_1}(1-x_7)^{m_2} = \frac{(m_1+m_2)!}{m_1!m_2!} x_7^{m_1}(1-x_7)^{m_2}$$

多项分布

$$MultiCount(m_1, m_2, m_3|x_4, x_7) = \frac{(m_1+m_2+m_3)!}{(m_1)!(m_2)!(m_3)!} x_3^{m_1}(x_7-x_3)^{m_2}(1-x_7)^{m_3}$$

结论

同理可得结论，注意下式的加号不是加法运算！

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

先验分布+ 数据的知识= 后验分布

这就是定义的所谓的Dirichlet-Multinomial 共轭，即数据知识的分布使得参数的先验、后验分布类型一样。

2. Beta &Dirichlet

参数$\boldsymbol{p}$的具体取值问题

游戏结论：

猜测$\boldsymbol{p}$取$\boldsymbol{p}$的posterior概率达到最大的峰值，即猜数字、投硬币游戏最可能的答案

现实中：

我们不要求猜数字，我们要求某个数据序列的概率，给定$\boldsymbol{p}$的分布即可，不用给出一个具体数字。

对于刚刚的多项分布MultiNomial，三维的，有：

$$MultiCount(m_1, m_2, m_3|\boldsymbol{p}) = \frac{(m_1+m_2+m_3)!}{(m_1)!(m_2)!(m_3)!}\iint_{p_1,p_2} p_1^{m_1}(p_2-p_1)^{m_2}(1-p_2)^{m_3}\mathrm{d}p_1\mathrm{d}p_2$$

2. Beta &Dirichlet

参数$\boldsymbol{p}$的具体取值问题

$$MultiCount(m_1, m_2, m_3|\boldsymbol{p}) = \frac{(m_1+m_2+m_3)!}{(m_1)!(m_2)!(m_3)!}\iint_{p_1,p_2} p_1^{m_1}(p_2-p_1)^{m_2}(1-p_2)^{m_3}\mathrm{d}p_1\mathrm{d}p_2$$

但是！随着参数空间$\boldsymbol{p}$的维度增大，积分将变得不可实现！故我们必须确定$\boldsymbol{p}$的一组取值，而非用分布函数。

退一步

最好的取值当然是$\boldsymbol{p}$的后验概率最大值（MAP）。

再退一步

然而这还是不方便求的，于是退而求其次，取 $\boldsymbol{p}$ 的平均值作为估计值 $\hat{\boldsymbol{p}}$ 。

$$MultiCount(m_1, m_2, m_3|\hat{\boldsymbol{p}}) = \frac{(m_1+m_2+m_3)!}{(m_1)!(m_2)!(m_3)!}\hat{p_1}^{m_1}(\hat{p_2}-\hat{p_1})^{m_2}(1-\hat{p_2})^{m_3}$$

2. Beta &Dirichlet

参数$\boldsymbol{p}$的具体取值问题

如果$p \sim Beta(t|\alpha,\beta)$, 则

$$\begin{aligned} E(p) &= \int_0^1 t * Beta(t|\alpha,\beta)dt \\ &= \int_0^1 t * \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} t^{\alpha-1}(1-t)^{\beta-1}dt \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 t^{\alpha}(1-t)^{\beta-1}dt \end{aligned}$$

上式右边的积分对应到概率分布$Beta(t|\alpha+1,\beta)$，对于这个分布，我们有

$$\int_0^1 \frac{\Gamma(\alpha+\beta+1)}{\Gamma(\alpha+1)\Gamma(\beta)} t^{\alpha}(1-t)^{\beta-1}dt = 1$$

把上式带入$E(p)$的计算式，得到

$$\begin{aligned} E(p) &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot \frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+\beta+1)} \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha+\beta+1)} \frac{\Gamma(\alpha+1)}{\Gamma(\alpha)} \\ &= \frac{\alpha}{\alpha+\beta} \end{aligned}$$

2. Beta &Dirichlet

参数$\boldsymbol{p}$的具体取值问题

这说明，对于Beta 分布的随机变量，其均值可以用$\frac{\alpha}{\alpha+\beta}$来估计。Dirichlet 分布也有类似的结论，如果$\vec{p} \sim Dir(\vec{t}|\vec{\alpha})$，同样可以证明

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

结论：

Bayesian方芳的方法，在现实中(如LDA算法中)无奈只能取平均值，取平均值的结果，恰恰是Freqentist峰峰的结果！！

虽然这样，但我们在推理的时候(LDA算法)，还是要用方芳的思路推，只在最后部分地方取用峰峰的结果，这样更具有普适意义。

新游戏:

刚刚我们仅考虑一枚硬币（或者筛子），现在我们来考虑多个硬币，我们这么玩：

游戏：

1. 给硬币A、B、C，先抛硬币A，得到结果，记为随机变量$\boldsymbol{Z}$。
2. 结果正面的话就抛B，反面抛C，得到结果记为$\boldsymbol{Y}$

问题：

我们做了一堆实验，得到了一串$\boldsymbol{y} = (y_1, y_2, y_3, \cdots)$，问A、B、C各自正面的概率$(\theta, \varphi_1, \varphi_2)$？

联系旧知识:

分析这个问题，Z是隐变量，Y是观测变量，$(\theta, \varphi_1, \varphi_2)$是参数。

这就是一个标准的**隐马尔科夫**问题！但很特殊，因为马尔科夫的隐变量序列和观测变量序列不是独立的，是存在转移矩阵的。但这里所有试验都是独立同分布的！是(0,1)之间的随机产生的！

马尔科夫似然函数，用**最大似然法(MLE)的思想**，由于有隐变量无法直接解析求解，于是用迭代的思想求解用EM算法求解！

$$p(\boldsymbol{y}|\theta, \varphi_1, \varphi_2) = \prod_{i=1}^{n} \sum_{j=1}^{2} p(z_j|\theta) p(y_i|z_j, \varphi_1, \varphi_2)$$

3. LDA

联系旧知识:

分析这个问题，Z是隐变量，Y是观测变量，$(\theta, \varphi_1, \varphi_2)$是参数。

Frequentist

这就是一个标准的**隐马尔科夫**问题！但很特殊，因为马尔科夫的隐变量序列和观测变量序列不是独立的，是存在转移矩阵的。但这里所有试验都是独立同分布的！是(0,1)之间的随机产生的！

马尔科夫似然函数，用**最大似然法(MLE)的思想**，由于有隐变量无法直接解析求解，于是用迭代的思想求解用EM算法求解！

$$p(\boldsymbol{y}|\theta, \varphi_1, \varphi_2) = \prod_{i=1}^{n} \sum_{j=1}^{2} p(z_j|\theta) p(y_i|z_j, \varphi_1, \varphi_2)$$

Bayesian View: LDA Model

Bayesian

凭什么认为$(\theta, \varphi_1, \varphi_2)$是常数，我们要将其作为**分布**代入！

Bayesian View: LDA Model

硬币A的隐变量序列似然函数

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

$$Dir(\vec{p}|\vec{\alpha}) = \frac{1}{\Delta(\vec{\alpha})} \prod_{k=1}^{V} p_k^{\alpha_k - 1} \quad \vec{\alpha} = (\alpha_1, \cdots, \alpha_V)$$

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

硬币B,C的观测变量序列似然函数

$$p(\vec{\mathbf{w}}|\vec{\mathbf{z}}, \vec{\beta}) = p(\vec{\mathbf{w}}'|\vec{\mathbf{z}}', \vec{\beta})$$
$$= \prod_{k=1}^{K} p(\vec{w}_{(k)}|\vec{z}_{(k)}, \vec{\beta})$$
$$= \prod_{k=1}^{K} \frac{\Delta(\vec{n}_k + \vec{\beta})}{\Delta(\vec{\beta})}$$

A，B，C联合概率

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

Gibbs采样

一种分布采样方法。

Thanks

Chongming Gao
Yingcai Experimental School
gchongming@126.com