LDA-123 之一

梁斌

本系列分三个内容因此叫 LDA-123。每段内容凡是我的观点，都用红色标出，不一定正确，批评吸收，属于中医打法，玄学的部分。

假定我们对一个未知世界产生的各种情况，不了解。但我们猜测，产生这些千奇百怪的情况的背后的那个别别窍，其实很简单。或者说复杂是由简单构成的，那么本节，假定银币抛的结果出现正面，或者背面，是由唯一的一个参数作梗而导致的，这个参数就是 P（硬币正面的概率）。但我们都知道这个 P 无论如何也没法精确知道的，只有上帝知道。

假定有一个硬币，我们扔了 N 次，观测到其中α次是正面，β次是背面，N = $\alpha+\beta$

现在问题是，出现这种情况（α 正，β 负）的概率有多大？这个硬币，正背面的出现概率是一半对一半吗？有没有名堂？

如果我们假定出正面的概率为 p，则负面的概率为 1-p。

那么出现我们扔的这个**情况**的概率是$p^{\alpha} * (1-p)^{\beta}$，我们希望出现这个情况的概率最大，也就是我们的这一把扔，是最合乎情理的。

则问题相当于求 f(p)的最大值，即 p 取什么值，可以让$p^{\alpha} * (1-p)^{\beta}$最大。
Log(x) 和 x 具有相同的单调性，即 x 最大值，log（x）也必然取到最大值。
因此对 f(p)取对数，取对数方便求导，而且在实际计算中保存足够的精度。
有 log(f(p)) = $\log\left(p^{\alpha} * (1-p)^{\beta}\right)$
求导数，并令其为 0，得到最值

$$\frac{\alpha}{p}-\frac{\beta}{1-p}=0$$

解得 $p=\frac{\alpha}{\alpha+\beta}$
直观的说，如果我们先验得对一个硬币扔了 100 次，看到证明出现了 40 次，背面出现了 60 次。我们可以认为这个硬币正面出现的概率是 40%，使得我们观测到的这个局面出现的概率最大。

现在问题来了，我们不需要求哪个 p 能导致出现这个局面最大，而是希望计算，每个 p 的概率分布，比如 40 次正面，60 次正面这个情况下，硬币实际正面概率是 20%，这个事情的概率是多大？ 即$P(P|\alpha,\beta)$ =?
我们**一般认为这个概率符合 beta 分布**，beta 分布一个比较好的性质是，beta(α ，β)的期望恰好是α /（α +β ），和上面求导计算的结果一致。而且 beta 的形式和之前公式比较接近

$$P(p|\alpha,\beta) \;=\; \mathrm{Beta}(p|\alpha,\beta) \;= \frac{1}{B(\alpha,\beta)} * p^{\alpha-1} * (1-p)^{\beta-1}$$

也就是任意给定p,α,β，那么出现这个情况的概率就用上面这个公式计算。

$$B(\alpha,\beta) = \frac{\tau(\alpha) * \tau(\beta)}{\tau(\alpha+\beta)}$$
$$\tau(x) = (x-1)!$$

为什么要除以$B(\alpha,\beta)$呢？

这是为了让

$$\int_0^1 \frac{1}{B(\alpha,\beta)} * p^{\alpha-1} * (1-p)^{\beta-1} dp = 1$$

我们可以求一下$B(\alpha,\beta)$

即求一下$\int_0^1 p^{\alpha-1} * (1-p)^{\beta-1} dp$

$$\int_0^1 p^{\alpha-1} * (1-p)^{\beta-1} dp = \frac{1}{\alpha}\int_0^1 (1-p)^{\beta-1} dp^{\alpha}$$
$$= \frac{1}{\alpha}\int_0^1 p^{\alpha} d(1-p)^{\beta-1} = \frac{\beta-1}{\alpha}\int_0^1 p^{\alpha} * (1-p)^{\beta-2} dp$$

即得到下面递推式

$$\int_0^1 p^{\alpha-1} * (1-p)^{\beta-1} dp = \frac{\beta-1}{\alpha}\int_0^1 p^{\alpha} * (1-p)^{\beta-2} dp$$

由这个递推式

得到

$$\int_0^1 p^{\alpha-1} * (1-p)^{\beta-1} dp = \frac{(\beta-1) * (\beta-2) * \ldots * 1}{\alpha * (\alpha+1) * \ldots * (\alpha+\beta-1)} = \frac{(\alpha-1)! * (\beta-1)!}{(\alpha+\beta-1)!}$$

为了让

$$\int_0^1 \frac{1}{B(\alpha,\beta)} * p^{\alpha-1} * (1-p)^{\beta-1} dp = 1$$

显然

$$B(\alpha,\beta) = \frac{(\alpha-1)! * (\beta-1)!}{(\alpha+\beta-1)!}$$

现在问题又来了，如果我们这个实验是，我们先扔 100 次，得到一个先验概率 40 正，60 背。如果我们又扔了 100 次，45 正，55 背。那么 p 等于多少能让这个**情况**，出现的概率最大呢？

我们用乘法来刻画他们的关系

即：

P(P|40,60)* $p^{45} * (1-p)^{55}$ 可以理解为先验的，40,60 得到了 p 的概率分布，然后这个概率 p 来得到最终的概率：P(likelihood |p) $*$ P(p |α，β）

为了让这个概率最大，即 max P(likelihood |p) $*$ P(p |α，β）

依然用求导，令其为零的方法，得到

$$\frac{45}{p}-\frac{55}{1-p}+\frac{40-1}{p}-\frac{60-1}{1-p}=0$$

$$求得：p=\frac{45+40-1}{100+100-2}=\frac{84}{198}$$

这个含义就是，p = 84/198 可以让先验的这个情况和 likelihood 出现的概率最大。
P(P|40,60)就是先验，$p^{45}*(1-p)^{55}$ 就是 likelihood。

前面我们通过先验和后验，来求得什么 p 可以使得这种情况出现的概率最大，现在问题是我们要求：量化这个概率。也就是 P(p|post,prior)?
α，β 就是先验，C 表示出现一系列正背面的情况（N 次独立的投币实验的结果）。

$$P(p|C,\alpha,\beta)=\frac{\prod_{i=1}^{N}P(C=ci|p)P(p|\alpha,\beta)}{\int_0^1\prod_{i=1}^{N}P(C=ci|p)P(p|\alpha,\beta)dp}$$

$$=\frac{p^{n(1)}*(1-p)^{n(0)}*\frac{1}{B(\alpha,\beta)}*p^{\alpha-1}*(1-p)^{\beta-1}}{Z}$$

$$=\frac{p^{n(1)+\alpha-1}*(1-p)^{n(0)+\beta-1}}{B(n(1)+\alpha,n(0)+\beta)}$$

备注：上面公式推导中，不难得到 Z= $B(n(1)+\alpha,n(0)+\beta)/(B(\alpha,\beta)$ (不细说，可以仔细看我们之前的一个证明)

最后我们在看看 beta 分布

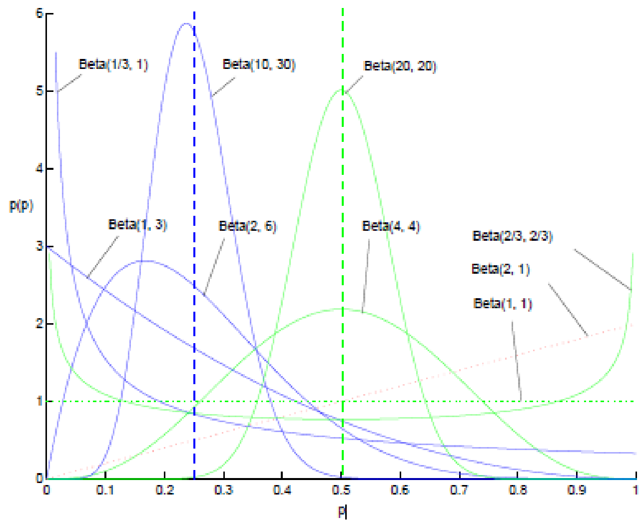


Fig. 1. Density functions of the beta distribution with different symmetric and asymmetric parametrisations.

怎么来直觉的理解这个图形呢？ 我们看到绿线 beta(20,20)，也就是抛了 40 次，20 正，20 背，那么这个概率为 0.5 就比较可信，图形比较高耸，概率密度大。而 beta(4,4)，虽然期望也是 0.5，但这个 0.5 的可信度就没有那么大，图形比较低矮，概率密度小。

现在我们的问题是，如果我们有一个先验<40,60>，又有一个后验<45,55>。我们希望知道由这个先验作条件，产生后验的概率，即求 P(posterior | prior)

$$
\begin{aligned}
\mathrm{P}(\mathrm{post}|\mathrm{prior}) &= \int_0^1 \mathrm{P}(\mathrm{post}|\mathrm{p}) * \mathrm{P}(\mathrm{p}|\alpha,\beta)\mathrm{dp} \\
&= \int_0^1 \mathrm{p}^{\mathrm{n}(1)} * (1-\mathrm{p})^{\mathrm{n}(0)} * \frac{1}{\mathrm{B}(\alpha,\beta)} * \mathrm{p}^{\alpha-1} * (1-\mathrm{p})^{\beta-1}\mathrm{dp}
\end{aligned}
$$

其中 n(1)表示证明出现的次数，n(0)表示背面出现的次数
最终上式子等于

$$
= \frac{\mathrm{B}(\mathrm{n}^{(1)}+\alpha, \mathrm{n}^{(0)}+\beta)}{\mathrm{B}(\alpha,\beta)}
$$

现在我们预测马上就要抛的这个硬币证明的概率，即计算在先验，后验后，再看到一个硬币是正面还是背面的概率，用贝叶斯公式

$$
\mathrm{P}（\mathrm{coin}=+|\mathrm{post}，\mathrm{porior}） = \frac{\mathrm{P}（\mathrm{coin}=+,\mathrm{post}|\mathrm{orior}）}{\mathrm{P}（\mathrm{post}|\mathrm{porior}）}
$$

$$
= \frac{\mathrm{B}(\mathrm{n}^{(1)}+1+\alpha, \mathrm{n}^{(0)}+\beta)}{\mathrm{B}(\mathrm{n}^{(1)}+\alpha, \mathrm{n}^{(0)}+\beta)}
$$

$$
= \frac{\mathrm{n}^{(1)}+\alpha}{\mathrm{n}^{(1)}+\mathrm{n}^{(0)}+\alpha+\beta)}
$$

这种预测，可以看做窗口是 post+proior 的预测。

总结

1） 用先验来估计 P
2） 用先验和后验来估计 p
3） 用先验和后验了计算 P
4） 用先验来计算后验的概率
5） 用先验和后验，来预测下一次出现的概率

LDA-123 之二

假定我们看到一个句子“我是梁斌，是博士”，且句子中的词恰好是我们词典中的词，且我们的词典只有下面 4 个词。

W1 = 我，W2 = 是，W3 = 梁斌，W4 = 博士

但是我们不知道每个词出现的概率 P=$\begin{Bmatrix} p1 \\ p2 \\ p3 \\ p4 \end{Bmatrix}$，我们希望用已经出现的情况来最大化 likelihood

$$p(W|\vec{p}) = \prod_{t=1}^{V} p_t^{n(t)}, \sum_{t=1}^{V} n(t) = N, \sum_{t=1}^{V} P_t = 1$$

$$p\left(\begin{Bmatrix} p1 \\ p2 \\ p3 \\ p4 \end{Bmatrix} \middle| \begin{Bmatrix} 我，1 \\ 是，2 \\ 梁斌，1 \\ 博士，1 \end{Bmatrix}\right) = p_1^1 * p_2^2 * p_3^1 * p_4^1, \sum_{t=1}^{V} P_t = 1$$

我们对上面概率求 log，并加上后面的约束条件得到

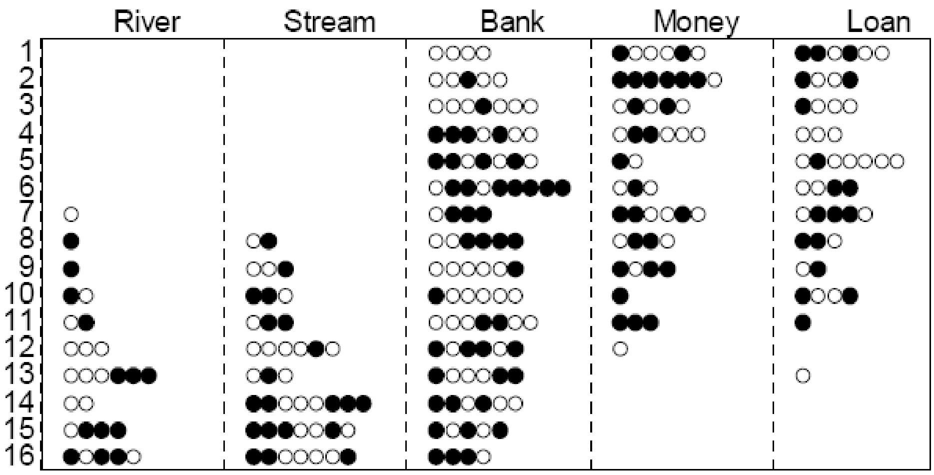
$$\log(p_1^1 * p_2^2 * p_3^1 * p_4^1) + \lambda(p1 + p2 + p3 + p3 - 1)$$

对上面式子，对 P1，P2，P3，P4，λ分别求偏导并令其为零

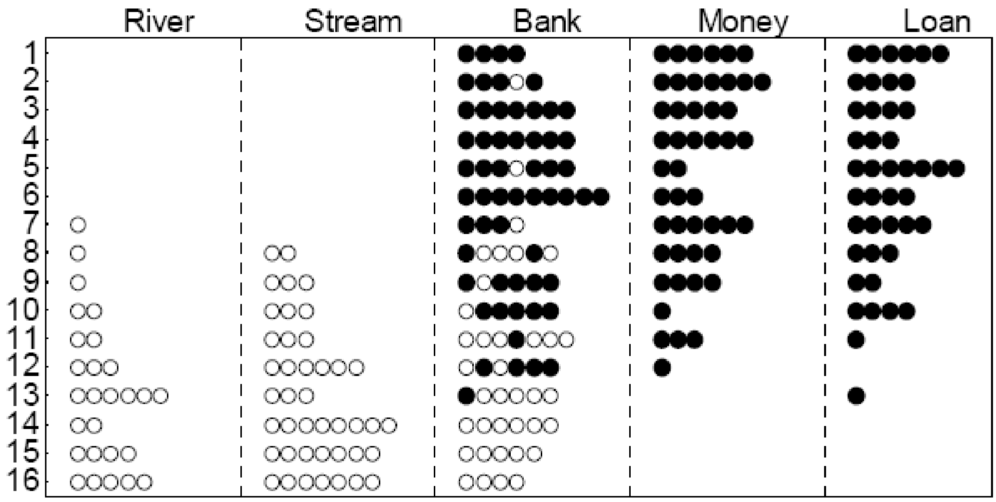
得到

P1= 1/4,p2 = 2/4,p3=1/4,p4=1/4

假定有这样一个聚类任务，有 16 个文档，文档中只有 5 个词不同的出现，每个文档恰好 16 个词。假定这 16 个文档我们认为有 2 个 topic，白色的为 topic_0，黑色的为 topic_1。



通过某种聚类方法，得到下图，含 River 和 Stream 多的文档 14,15,16 的词都是白点，为 topic0，含 Money 和 Load 比较多的文档 1,3,4,等都是黑点为，为 topic1。而 Bank 具有一定歧义，黑白均有分布。



怎么得到这个效果呢？首先需要了解一个 Dirichlet 分布。

之前提到过 Beta 分布

$$\mathrm{Beta}(p|\alpha,\beta)\ =\frac{1}{B(\alpha,\beta)}*p^{\alpha-1}*(1-p)^{\beta-1}$$

如果我们改一个形式

$$\mathrm{Beta}\left(\begin{Bmatrix}p1\\p2\end{Bmatrix}\middle|\alpha,\beta\right)=\frac{1}{B(\alpha,\beta)}*p_1^{\alpha-1}*p_2^{\beta-1},p1+p2=1$$

则上面这个公式的含义是，如果一个硬币有正背面，如果抛了 N 次，α 次正面，β 次背面，那么出正面的概率 p1 和出概率 p2 的程度是多大，或者说概率密度是多大

而且这个概率密度，如果求积分：

$$\iint_{p\in P}\frac{1}{B(\alpha,\beta)}*p_1^{\alpha-1}*p_2^{\beta-1}\,dp_1dp_2{=}1$$

$$p\in P\ 表示，p\ 取一切满足\ p1+p2=1\ 的条件$$

上面式子还可以写成

$$\iint_{\vec{p}\in P}\frac{1}{B(\alpha,\beta)}*p_1^{\alpha-1}*p_2^{\beta-1}\,d\vec{p}\ =1$$

那么 Dirichlet 分布就是 Beta 分布的高维形式

$$\mathrm{Dir}(\alpha_1,\alpha_2,\ldots\alpha_K)=\frac{\tau(\sum_{j=1}^{K}\alpha_j)}{\prod_{j=1}^{K}\tau(\alpha_j)}\prod_{j=1}^{K}p_j^{(\alpha_j)-1}$$

我们通过一个例子来理解 Dirichlet 分布：

例子中和衡量的 Dir（4,4,2）在不同$\vec{P}$的组合下的概率密度：

即 $\mathrm{Dir}\begin{Bmatrix}p1\\p2\\p3\end{Bmatrix}|\begin{Bmatrix}\alpha_1\\\alpha_2\\\alpha_3\end{Bmatrix}$

https://github.com/pennyliang/MachineLearning-C---code/blob/master/dirichlet_distribution/main.cpp

运行实验后，可以看到在$p1=\frac{\alpha_1-1}{\alpha_1-1+\alpha_2-1+\alpha_3-1}, p2=\frac{\alpha_2-1}{\alpha_1-1+\alpha_2-1+\alpha_3-1}, \frac{\alpha_3-1}{\alpha_1-1+\alpha_2-1+\alpha_3-1}$时概率密度达到最大。精确计算的方法类似本节一开始的求导取 0 的例子。

换句话说，如果一个硬币有 3 面，A 面，B 面，C 面。A 面出现 4 次，B 面出现 4 次，C 面出现 2 次。那么如果这个过程符合 Dirichlet 分布，我们认为 A 面出现的最大可能是 3/7，B 面出现最大可能是 3/7，C 面出现的最大可能是 1/7。这也是符合直觉的。一开始 A 面出的多，那么背后的神秘的不可得到的真正 A 出现的概率也可能较大。

Gibbs sampling 的方法实质是给定文档，一个 term 求这个 term 最大期望是什么 topic 的一个过程。

$$p(\vec{\vartheta}_m|\mathcal{M},\vec{\alpha}) = \frac{1}{Z_{\vartheta_m}} \prod_{n=1}^{N_m} p(z_{m,n}|\vec{\vartheta}_m)\ p(\vec{\vartheta}_m|\vec{\alpha}) = \mathrm{Dir}(\vec{\vartheta}_m|\vec{n}_m + \vec{\alpha}),$$

$$p(\vec{\varphi}_k|\mathcal{M},\vec{\beta}) = \frac{1}{Z_{\varphi_k}} \prod_{\{i:z_i=k\}} p(w_i|\vec{\varphi}_k)\ p(\vec{\varphi}_k|\vec{\beta}) = \mathrm{Dir}(\vec{\varphi}_k|\vec{n}_k + \vec{\beta})$$

这个公式的含义就是：
从已经达到的状态来推测#文档 m 属于哪个 topic 的概率#这个事情，符合 Dir 分布。
从已经达到的状态来推测#每个 topic 产生 wj 的概率#这个事情，也符合 Dir 分布。

拿本节一开始的例子来说：

我们对文档 1 的第一个 word，即 bank 这个词来评价 P(z=topic？|w=bank,doc=1)。
这个过程分两步走

第一步看投哪个 topic，我们有 Dir 分布 Dir（$\begin{Bmatrix}P0\\P1\end{Bmatrix}|\begin{Bmatrix}10\\5\end{Bmatrix}$），按照 Dirichlet 的性质，P0 最大可能性是 4/13（投 topic0 的概率），P1 最大可能性是 9/13。
10：表示除了第一个点以外，有 10 个白点
5：表示除了第一个点以外，有 5 个黑点

第二步看有了 topic 以后，产生每个 term 的能力，我们有 Topic0 的 Dir 分布

Dir{$\begin{Bmatrix}P_{River}\\P_{Stream}\\P_{bank}\\P_{Money}\\P_{Bank}\end{Bmatrix}|\begin{Bmatrix}14\\22\\51\\22\\26\end{Bmatrix}$}，按照 Dirichlet 性质，$P_{Bank}=51/(14+22+51+22+26)$

其中 14，表示 River 有 14 次被标记为 topic0（14 个白点的 River）。余下雷同。
同理 topic1 的 Dir 分布

Dir{$\begin{Bmatrix}P_{River}\\P_{Stream}\\P_{bank}\\P_{Money}\\P_{Bank}\end{Bmatrix}|\begin{Bmatrix}13\\20\\44\\25\\18\end{Bmatrix}$}，按照 Dirichelet 性质，$P_{Bank}=44/(13+20+44+25+18)$

则计算 P（topic0| w=bank,doc=0) = $4/13*51/(14+22+51+22+26)=0.1162$
P（topic1| w=bank,doc=1) = $9/13*44/(13+20+44+25+18)=0.2538$
P（topic1| w=bank,doc=1) > （topic0| w=bank,doc=1)
因此 doc1 的第一个 word：bank，应该给黑点。
余下对每个 doc 中的 word 都按这种方法计算，迭代足够多的轮次，就可以出现收敛的现象。

实验例子来源：

https://github.com/pennyliang/MachineLearning-C---code/blob/master/gibbs_sampling/SteyversGriffithsLSABookFormatted.pdf

实验代码 1：

https://github.com/pennyliang/MachineLearning-C---code/blob/master/gibbs_sampling/main.cpp

另一种实现 2：

https://github.com/pennyliang/MachineLearning-C---code/blob/master/gibbs_sampling/main2.cpp