

持之以恒

Every thing that has a beginning has an end.

【JMLR'03】Latent Dirichlet Allocation (LDA) - David M.Blei

发表日期: 2012 年 3 月 22 日

分类: **Academics** 标签: **LDA, Paper Comments** 作者: 管理员. 4,251 views

若公式显示有问题请复制链接到新TAB重新打开

听说国外大牛都认为LDA只是很简单的模型，吾辈一听这话，只能加油了~

另外这个大牛写的LDA导读很不错: http://bbs.byrcn.cn/#!article/PR_AI/2530?p=1

一、预备知识:

1. 概率密度和二项分布、多项分布，在[这里](#)

2. 狄利克雷分布，在[这里](#)，主要内容摘自《Pattern Recognition and Machine Learning》第二章

3. 概率图模型，在PRML第九章有很好的介绍

二、变量表示:

1. **word**: **word**是最基本的离散概念，在自然语言处理的应用中，就是词。我觉得比较泛化的定义应该是观察数据的最基本的离散单元。**word**的表示可以是一个V维向量v，V是所有**word**的个数。这个向量v只有一个值等于1，其他等于0。呵呵，这种数学表示好浪费，我以前做过的项目里一般中文词在200-300w左右，每一个都表示成300w维向量的话就不用活了。哈哈，所以真正应用中**word**只要一个编号表示就成了。

2. **document**: 一个**document**就是多个**word**的合体。假设一篇文档有N个词，这些**word**是不计顺序的，也就是**exchangeable**的，LDA论文 3.1有说这个

概念。论文中document的个数是M。

3. topic: 就是主题啦，比如“钱”的主题可能是“经济”，也可能是“犯罪”~ LDA中主题的表示是隐含的，即只预先确定主题的个数，而不知道具体的主题是什么。论文中表示主题个数的字母是k，表示主题的随机变量是z。

好了，总结一下所有的变量的意思，V是所有单词的个数（固定值），N是单篇文档词的个数（随机变量），M是总的文档的个数（固定值），k是主题的个数（需要预先根据先验知识指定，固定值）。

三、基础模型：

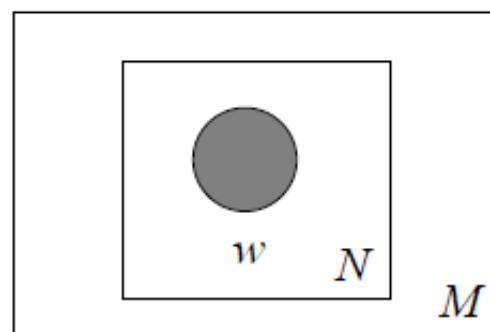
先从两个基础模型说起：

1. Unigram model (LDA 4.1)

一个文档的概率就是组成它的所有词的概率的乘积，这个一目了然，无需多说：

$$p(\mathbf{w}) = \prod_{n=1}^N p(w_n)$$

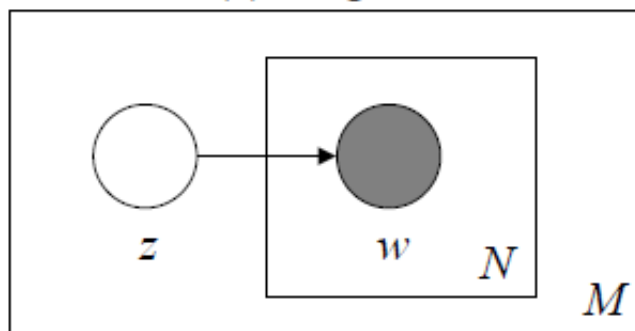
图模型：



(a) unigram

2. Mixture of unigrams (LDA 4.2)

假如我们假设一篇文档是有一个主题的（有且仅有一个主题），可以引入主题变量z，那么就成了mixture of unigrams model。它的图模型如下图：



(b) mixture of unigrams

$$p(w) = \sum_z p(z) \prod_{n=1}^N p(w_n|z)$$

这个模型的generate过程是，首先选择一个topic z for each document，然后根据这个 z 以及 $p(w|z)$ 独立同分布产生 w 。观察这个图， z 是在 N 饼外面的，所以每一个 w 均来自同一个 z ，就是说一个文档 N 个词只有一个topic。这和LDA中 z 在 N 饼里面不一样。

四、LDA

接下来正式说LDA的产生过程，对于一个文档 w ：

1. 选择 $N \sim \text{Poisson}(\xi)$

这一步其实只是选个单词的个数，对整个模型没啥影响

2. 选择一个多项分布参数 $\theta \sim \text{Dir}(\alpha)$

这 α 是狄利克雷分布的参数（ $k+1$ 维）， $\vec{\theta} = (\theta_1, \dots, \theta_k)$ 是产生主题的多项分布的参数，其中每一个 θ_i 代表第 i 个主题被选择的概率。从狄利克雷产生参数 θ 之后，再用 θ 去产生 z

3. 上两步完成后，开始产生文档中的 N 个词

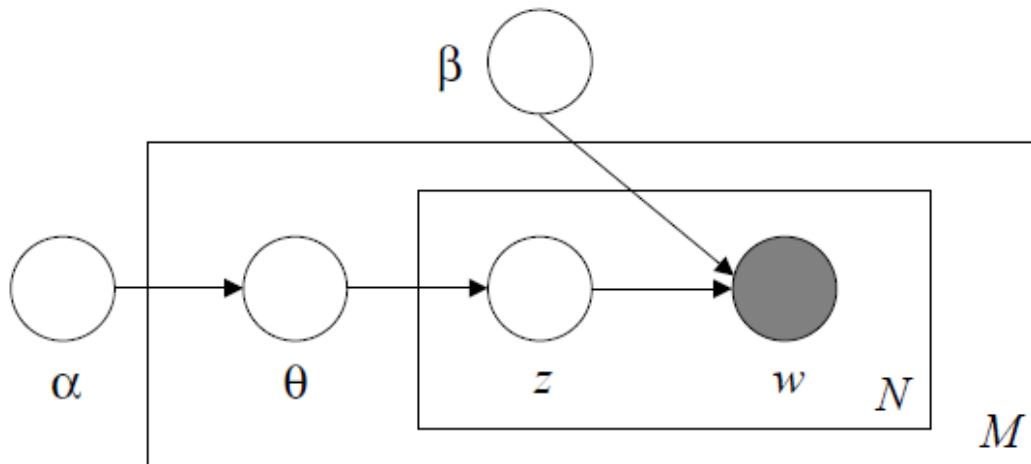
(a) 首先选个一个topic $z \sim \text{Multinomial}(\theta)$

z 是从以 θ 为参数的多项分布中挑选出来的，总共有 k 个topic，根据 θ 的概率参数选择其中一个topic作为 z

(a) 然后选择一个word from $p(w_n|z_n, \beta)$

这个参数 β 也是多项分布，是一个 $k \times V$ 的矩阵，表示从 z_i 到 w_j 的产生概率即 $\beta_{ij} = p(w^j = 1|z^i = 1)$ 。若已选定 z_n ，则矩阵的第 n 行就成了用来选择产生 w 的多项分布，根据这个多项分布产生一个 w

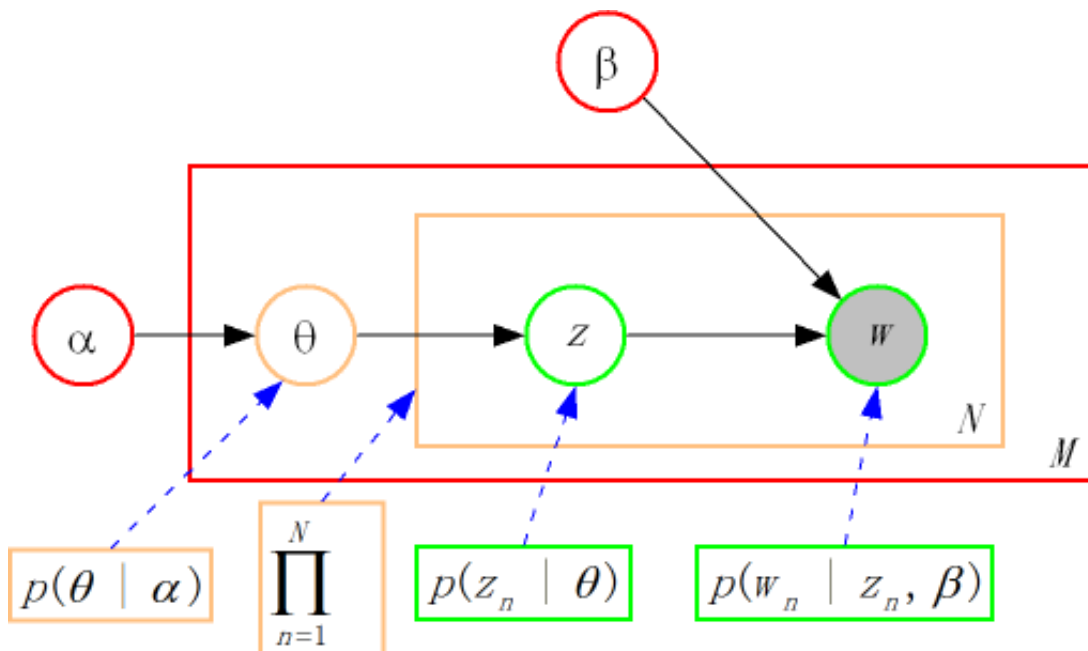
至此，产生过程完成。上概率图模型：



整个图的联合概率为(只算单个文档，不算整个corpus的M个文档)：

$$p(\theta, \mathbf{z}, \mathbf{w} | \alpha, \beta) = p(\theta | \alpha) \prod_{n=1}^N p(z_n | \theta) p(w_n | z_n, \beta)$$

把上式对应到图上，可以大致解释成这个样子：



在上面这个新图中，LDA的三个表示层被用三种颜色表示了出来：

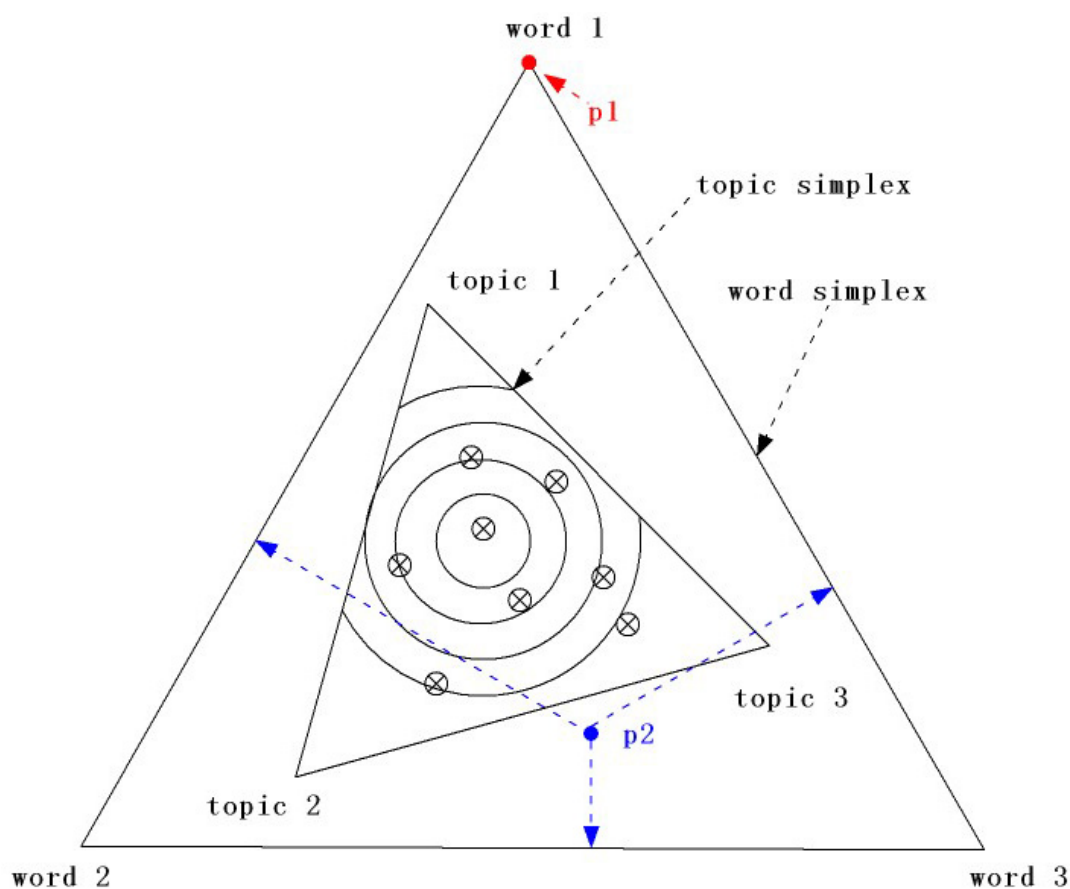
1. **corpus-level (红色)**: α 和 β 是语料级别的参数，也就是说对于每个文档都是一样的，因此在generate过程中只需要sample一次。

2. **document-level (橙色)**: θ 是文档级别的参数，意即每个文档的 θ 参数是不一样的，也就是说每个文档产生topic z 的概率是不同的，所以对于每个文档都要sample一次 θ 。

3. **word-level (绿色)**: 最后 z 和 w 都是文档级别的变量， z 由参数 θ 产生，之后再由 z 和 β 共同产生 w ，一个 w 对应一个 z 。

五、几何学解释

来看下面这个图：



这个图的意思是这样的，外面大三角形的三个顶点代表三个word，这三个word组成一个simplex，那么这个simplex中的一个点，代表什么意思呢？它代表的意思就是一个点就是一个产生这三个word的多项分布的概率密度（对于这个图多项分布的它是一个三维向量）。具体点来说，例如红色的点 $p1$ ，

它就在word1上。这个意思就是说， p_1 是一个多项分布，其参数为(1.0, 0, 0)，也就是它产生word1的概率为1，产生其它两个word的概率为0。再来看蓝色的点 p_2 ，它产生word1的概率正比于它到word1对边的距离（注意可不是到word1那个点的距离哈）。因为正三角形内部任意一点到三边的垂线之和等于高，也就是可以视为等于1。那么正好这个性质满足概率之和等于1。所以 p_2 到三边的垂线非别代表 p_2 产生垂线对面那个顶点的概率。因此， p_2 产生word1的概率看起来像是0.1，word2的概率像是0.4，word3像是0.5。

了解了上面这层意思之后，我们再来看这个topic simplex。它是包含在word simplex里面的(sub-simplex)，所以topic simplex上的一点同时也是word simplex上的一个点。这样topic simplex上的一个点，就有了两层含义，一层含义是它是一个产生word的多项分布概率密度，另一层含义就是它是产生topic的多项分布概率密度。在这个图上，还可以发现topic的点相对于word simplex是已经固定的，其实这topic simplex上的三个顶点到word simplex上的三个顶点对边垂直线总共9个距离，也就是9个概率值，正好是一个 3×3 的矩阵，这个矩阵就是LDA中的 β 参数。

知道了这些之后，我们就可以来看mixture of unigrams在图上应该怎么表示了。还记得mixture of unigrams是要先选择一个文档的topic z 的，然后根据这个topic产生word。所以它在这个图上的产生过程就是，先随机挑选topic simplex(注意是topic simplex)三个顶点中的一个，然后根据这个顶点到word simplex顶点对边线的距离，也就是这个顶点在word simplex上的多项分布产生每一个word。

再来看pLSI，图中间每一个带叉的圈圈就是一个pLSI中的文档，每一个文档(在pLSI中文档被视为观察变量，即每个文档都被视为word那样是有编号的)都有一个独立的产生topic的多项分布，文档点的位置就表示了它产生三个topic的概率值。

对于LDA，汗，不是很理解，LDA places a smooth distribution on the topic simplex denoted by the contour lines。只好先放着了。

2012@3@28，关于上面这个LDA的图形为啥是曲线的问题，我专门请教了北大赵鑫大牛，他的回答很给力而且一针见血。要理解LDA为啥是曲线，要先从pLSI为啥是点说起。因为pLSI中，由文档 w 产生topic z 的概率是一个参数，对于每个单独文档这个参数要被估计一次，参数可不是随机变量，而是

固定的值。因此pLSI中每个文档在图中表示为一个确定的点。而LDA呢，文档w产生topic z的概率在论文里后面inference部分已经给出了，它是

$$p(z|w) = p(\theta, z|w, \alpha, \beta) = \frac{p(\theta, z, w|\alpha, \beta)}{p(w|\alpha, \beta)}, \text{ 也就是隐含变量 } z \text{ 的后验分布, 它是一个概率分布, 这也是整个LDA inference部分最需要估计的东东。}$$

因此图中用曲线来表示LDA，也就是说LDA places a smooth distribution on the topic simplex ...

2012@4@18 今天看到《Parameter estimation for text analysis》(PETA)里的内容，可以更深入地解释“LDA places a smooth distribution on the topic simplex denoted by the contour lines”这句话。首先给出PETA里面的原话：LDA with a uniform prior Dir(1) is a full Bayesian estimator for the same model for which PLSA provides an ML or MAP estimator。这句话说明了pLSA用的是最大似然推断或最大后验推断，在最大后验推断中，p(z|w)是一个给定的置信值（这一点PETA中也有说明：最大后验推断中的置信值不等于于概率），这个置信值是一个常量。LDA用的是贝叶斯推断，所以LDA中的p(z|w)是一个概率分布。

分类: **Academics** 标签: **LDA, Paper Comments** 作者: 管理员. 4,251 views

7 THOUGHTS ON “【JMLR'03】LATENT DIRICHLET ALLOCATION (LDA) - DAVID M.BLEI”



haidao

on 2012 年 4 月 18 日 at 下午 1:23 said:

大神啊！！！！ 终于对lad 里的一些数学公式有了点理解了，谢谢啊



恒

on 2012 年 4 月 18 日 at 下午 2:22 said:

这么客气的



chasefornone

on 2012 年 7 月 4 日 at 下午 1:45 said:

到处找资料，终于有点头绪，感谢博主，期待博主更多的美文！



王泽

on 2012 年 8 月 12 日 at 上午 9:57 said:

drichlet 的参数 α 为什么你说是 $k+1$ 维呢，不是 k 维，代表的应该是 k 个topic出现的次数吧？？



王树辰

on 2012 年 9 月 12 日 at 下午 9:40 said:

对于LDA的训练模型的样本，也就是输入文档集有什么要求没？是不是必须和推断新样本不同。



恒

on 2012 年 12 月 3 日 at 上午 9:51 said:

因为 α_0 是一个所有 α 的和，这个在狄利克雷分布前面归一化的那部分中 useful。 α 是topic的先验次数



张吉赓

on 2012 年 11 月 27 日 at 下午 9:42 said:

对我有不小帮助 这个simplex我一直看得很糊涂 今天看了终于明白了