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

Posted on 2012 年 3 月 22 日 by 管理员

【注：本文为原创】

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.byr.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。

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：

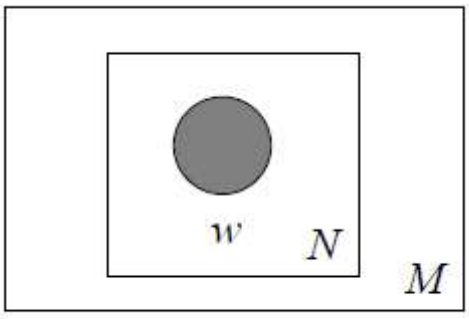
先从两个基础模型说起：

1. Unitgram 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：

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

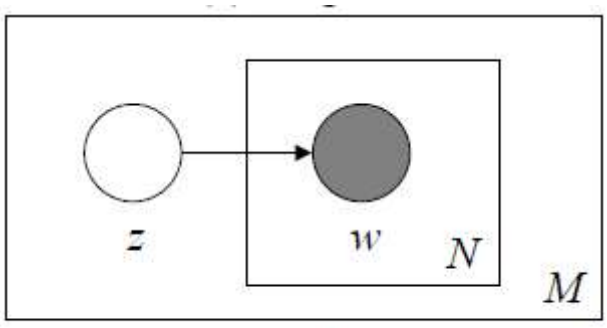
图模型：



(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

这个模型的generate过程是，首先选择一个topic z for each docoment，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 $N \sim Possion(\xi)$

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 $\theta \sim Dir(\alpha)$

这α是狄利克雷分布的参数（k+1维），$\vec{\theta} = (\theta_1, \ldots, \theta_k)$是产生主题的多项分布的参数，其中每一个$\theta_i$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

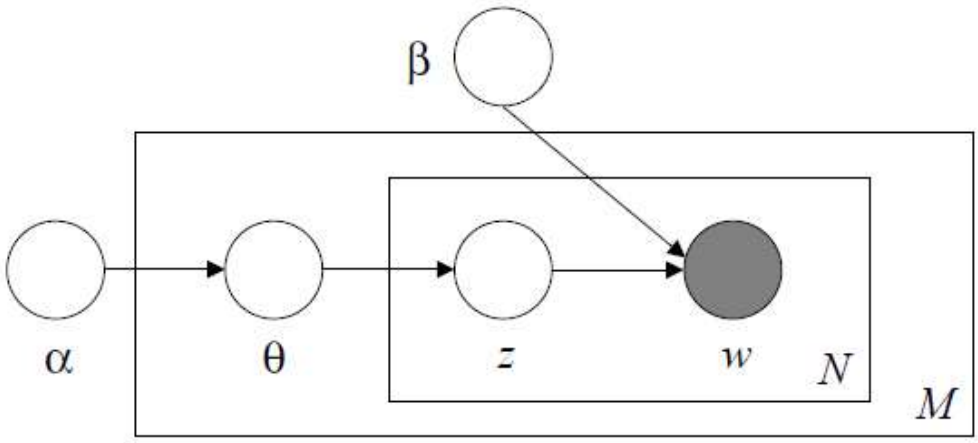
(a) 首先选个一个topic $z \sim Multinomial(\theta)$

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zi到wj的产生概率即$\beta_{ij} = p(w^j = 1|z^i = 1)$。若已选定zn，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

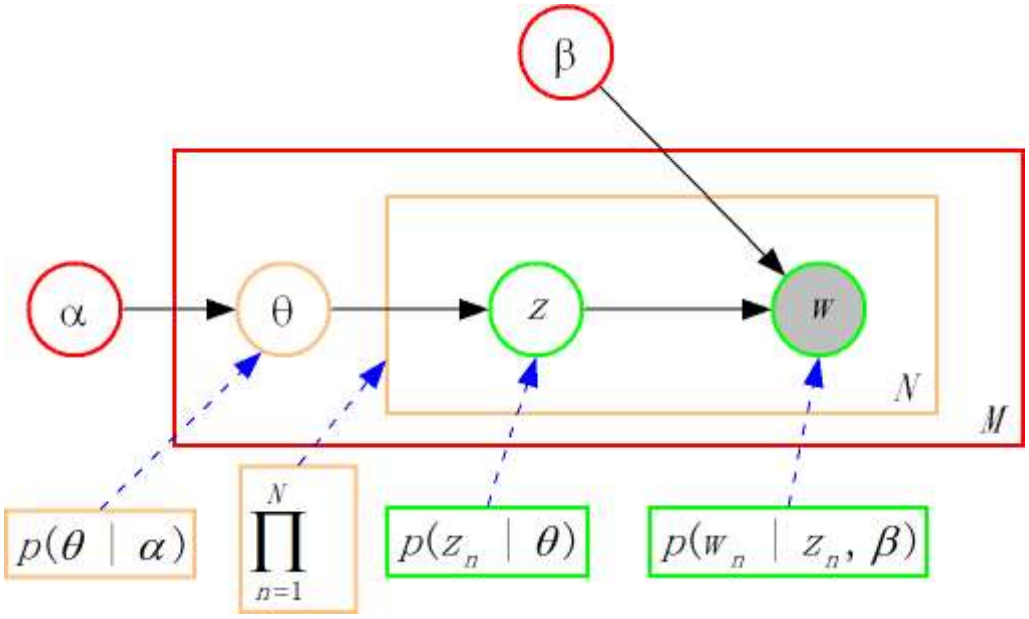
至此，产生过程完成。上概率图模型：



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：



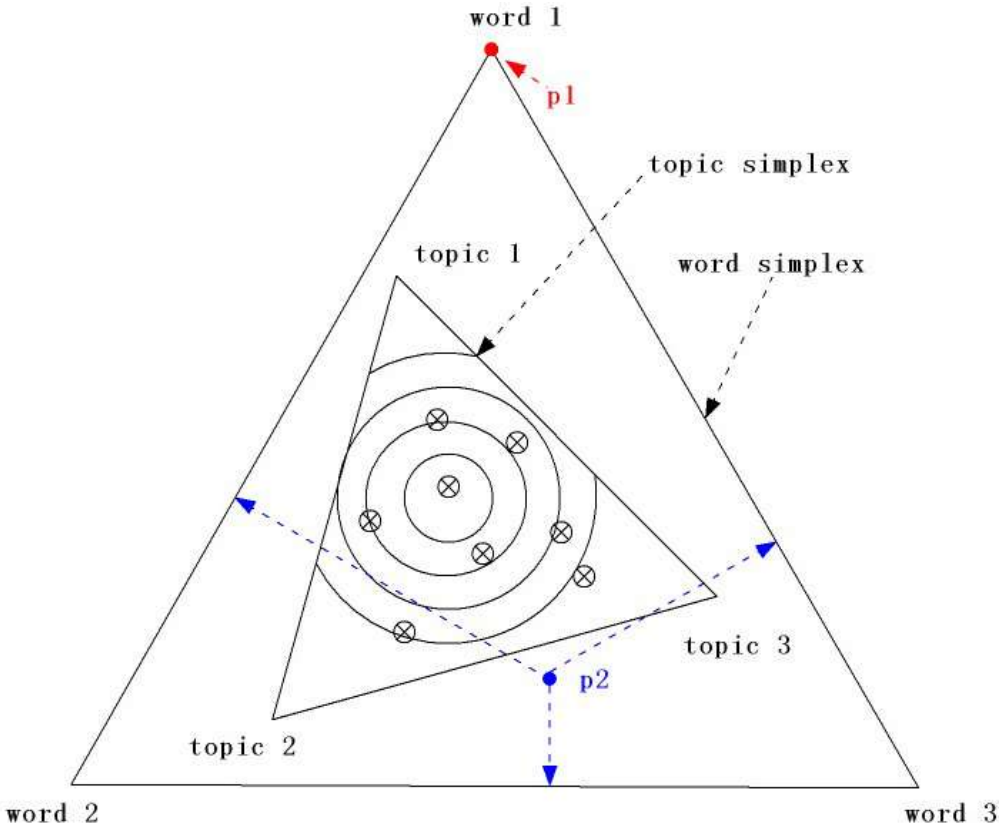
在上面这个新图中，LDA的三个表示层被用三种颜色表示了出来：

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



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，p1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p2，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p2到三边的垂线非别代表p2产生垂线对面那个顶点的概率。因此，p2产生word 1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

word。所以它在这个图上的产生过程就是，先随机挑选topic simplx(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$p(z|w) = p(\theta, z|w, \alpha, \beta) = \frac{p(\theta,z,w|\alpha,\beta)}{p(w|\alpha,\beta)}$，也就是隐含变量z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是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同于概率），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。

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α0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有用。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

恒
on 2013 年 7 月 12 日 at 上午 10:57 said:
你指哪一句呢？是论文原文里的？能贴英文原文么？

yxy
on 2013 年 11 月 19 日 at 下午 3:18 said:
我看倒这儿也产生了和他一样的问题，所以能理解他提出的这个问题，我看看能不能描述清楚。在生成模型的最后一步，对于一个document，得到了topic的分布（k维），每个topic有一个word的二项分布（V维）生成该文档的word（N维）的概率，应该是mul_foreach(n=1 to N) { sum_foreach(j=1 to K){ p(topic_j)* p(word_n | topic_j } }. 但从博文中介绍的mul_foreach(n=1 to N){ p(topic_n)*p(word_n | topic_n) }，看起来，好像每个word必须只能由一个topic生成。不知道是不是我们理解错了。（公司部分有省略，并且形式不好看，见谅）

yxy
on 2013 年 11 月 19 日 at 下午 11:07 said:
我已经理解到其中的意思了，其实是针对每个word根据theta的分布选一个topic，同时在这个topic下有一个对应的phi中的值，选topic类似于抛硬币，按多项式分布选定一个topic后就确定了这个word由该topic生成。解释得有点不清楚，如果博主有时间，帮忙梳理一下。谢谢。

恒
on 2013 年 11 月 22 日 at 下午 4:52 said:
我觉得解释的很对啊！！

Kern
on 2013 年 11 月 27 日 at 下午 7:44 said:

您的一系列博文，令我受益匪浅，非常感谢~~

恒

on **2013** 年 **12** 月 **7** 日 **at** 下午 **11:39** said:

应该有每一个词对应每一个topic的概率啊，可以挑概率最大的几个吧

Pingback: 通俗理解LDA主题模型 - Jeek

Pingback: 十一城-elevencitys.com » NLP自然语言处理系列——LDA主题词模型探析

Pingback: 通俗理解LDA主题模型 | DreamCore