LDA 浅易入门指南

摘要：本文主要从一个较宏观的角度介绍了潜在狄利克雷分配(Latent Dirichlet Allocation)的动机，过程以及作为一个极为重要的机器学习算法，它的具体用途。

注：由于本人才疏学浅，文章中肯定会出现许多不严谨、错误的地方，恭请指证。

一、简介

请想象一个概率为王的世界：你的行为受你的大脑支配，你的大脑受概率支配。你所做的一切都是符合一些概率分布的——你有一定概率今天早上喝一杯咖啡；有一定概率晚上吃鸡扒饭；也有一定概率走路去教室。当然，你可以说我有“自由意志”，我做的事都是我选择做的而不是遵从什么狗屁概率。但你无法否认一点：你的认知是不可抑制地受概率支配的，或者准确来说，受后验概率支配。什么是认知呢？这里的认知是你的一些“判断”，例如听到一些声音，你能判断出它的语言、意义，看到一些现象，你能判断出它的原因。

这是一个多么矛盾但又无比符合“理性”的假设：一方面，人为拥有“自由意志”而欢呼；另一方面，这一“自由意志”又能被理性所模型化。这个模型最直观的理解就是：“你的知识与观察决定着你的判断”。在这个概率世界里，“知识与观察”就是先验，“判断”就是后验，而且往往是概率最高的后验。医生们根据诊断报告与他们的知识判断病人“最可能”得了哪种病，你根据听到音节与音调来判断“最可能”的单词与语义——这就是模型解释下，人脑面对一个“分类问题”时的过程。

直观地理解了“后验”与“先验”，那么这些“先验”又是怎么得到的呢？这就是人所做的工作了，聪明的人类引入了概率分布来描述随机变量的取值规律 (Wikipedia)，例如泊松分布可以用于描述单位时间内随机事件发生次数的概率分布 (Wikipedia)，各种各样的心理学测试分数和物理现象比如光子计数则都被发现近似地服从正态分布 (Wikipedia)。

在机器学习的领域中，一个重要的有监督的学习方法就是生成方法 (Generative approach)，这个方法学习到的则是一个模型。什么模型呢？自然是生成模型 (Generative Model)。生成模型是一个描述一个“结果”是如何得到（生成）的，构成它的基本元素就是前文所提到的各种各样的概率分布。举个例子，某次考试，考试的分数分布经过观察，符合正态分布的钟型曲线，从生成方法的观点来看，则可以看成这些分数是由某些参数确定下的正态分布所“生成的”。

Latent Dirichlet Allocation（潜在狄利克雷分配，LDA）就是一个最初被用来解决文本分类问题的算法，本质上它是一个生成方法。比之前谈到的简单生成模型复杂，它假设了一篇文本、一个文本库的生成模型，然后从这个定义好的模型出发，通过机器学习的方法，学习这个模型的参数。最终，再根据这个学习到的模型来对每个词所属主题(topic)进行分类。

二、LDA 结构

初学 LDA 的人往往因繁杂的公式而陷入细节之中，包括我在学习的时候也是如此，经常在某一个公式或定理上绞尽脑汁，因为它的开山论文实在是太难读啦，前段介绍模型，后段讲推导，模型就十分复杂，更不要说推导部分的各种变分法、参数估计了。实际上，在查阅了许多资料后我发现，相比于细节上的公式定理，更重要的是略去枝叶，看清主干。

怎样确定主干呢？这又需要回到原始的定义：生成方法。只需要紧紧抓住这一点，就可以大致理解 LDA 的思路：

1) 定义一个生成模型（LDA 模型）

2) 学习这么一个模型的参数（概率分布的参数）

这里的生成模型就是 LDA 模型——糅合了各种概率分布、参数的复杂模型，要理解他它，首先得理解贝叶斯定理。

1、贝叶斯定理

什么是贝叶斯定理？从数学角度上来说，它是这样的：

$$\mathrm{P(A|B)} = \frac{\mathrm{P(A,B)}}{\mathrm{P(B)}}$$

或者说，是这样的：$\mathrm{P(A|B)} = \mathrm{P(B|A)} * \frac{\mathrm{P(A)}}{P(B)}$

从数学语言的角度上来说它是这样的：在 B 出现的前提下, A 出现的概率等于 A 出现的前提下 B 出现的概率乘以 A 出现的概率再除以 B 出现的概率 (Wikipedia)。

用平实的语言来讲它是这样的：给定事件 B 出现（Evidence，先验），那么事件 A 出现的概率受 A、B 均出现的概率、B 出现概率的影响，计算公式为 A、B 均出现的概率除以 B 出现的概率。

用一个生动的例子来讲是这样的：已知有蛀牙且牙痛的概率为$P1$，而出现牙痛的概率$P2$，那么档你的医生听到了你说牙齿痛，判断出有蛀牙的概率就为$P3 = \frac{P1}{P2}$。

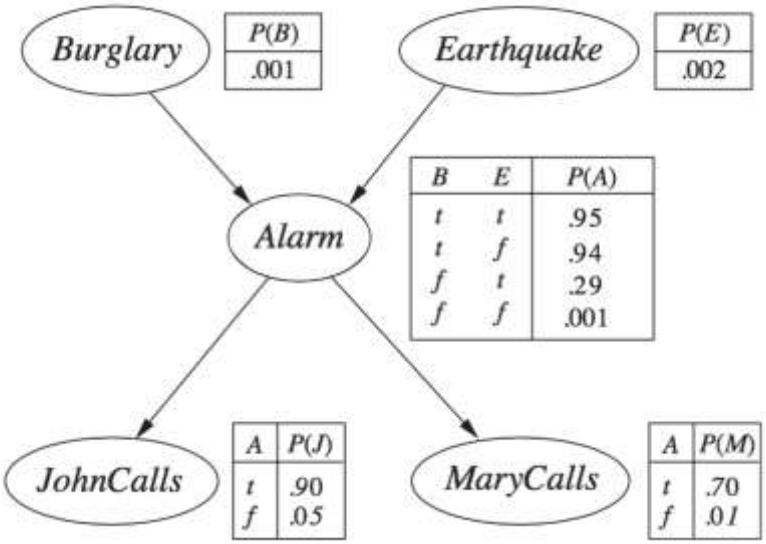
很容易看出来，贝叶斯定理玩弄的是先验与后验，它的魔力是由观察到的已知推测出未知。那么事物之间的影响仅仅有这么简单么？事件就只能有 A 和 B 了么？

当然不是，当影响因素多了，它们就构成了一个网络：

2、贝叶斯网络

贝叶斯网络是一个可视化的表示事件之间相互影响的概率网络。父节点有一根指向子节点的有向线段，代表父节点对子节点有影响，是子节点的先验。

一个简单的贝叶斯网络如下图所示 (Russell & Norvig, 2010)：



它讲述的是这么一个故事：家中入贼（Burglary）与发生地震（Earthquake）均有一定几率触发警报（Alarm），而你的两个邻居均因为警报而有一定几率打电话给你（JohnCalls, MaryCalls）。根据这一网络，我们即可根据需求计算某事件发生的概率，例如：警报响了，而又不是窃贼触发，也不是地震触发，而且你的两个邻居都打电话过来了。这个事件的概率等价于：

$$\begin{aligned} P(j, m, a, \neg b, \neg e) &= P(j \mid a) P(m \mid a) P(a \mid \neg b \wedge \neg e) P(\neg b) P(\neg e) \\ &= 0.90 \times 0.70 \times 0.001 \times 0.999 \times 0.998 = 0.000628 \end{aligned}$$

通常的，对于一个贝叶斯网络，有全概率公式：

$$\begin{aligned} P(x_1, \ldots, x_n) &= P(x_n \mid x_{n-1}, \ldots, x_1) P(x_{n-1} \mid x_{n-2}, \ldots, x_1) \cdots P(x_2 \mid x_1) P(x_1) \\ &= \prod_{i=1}^{n} P(x_i \mid x_{i-1}, \ldots, x_1) \,. \end{aligned}$$

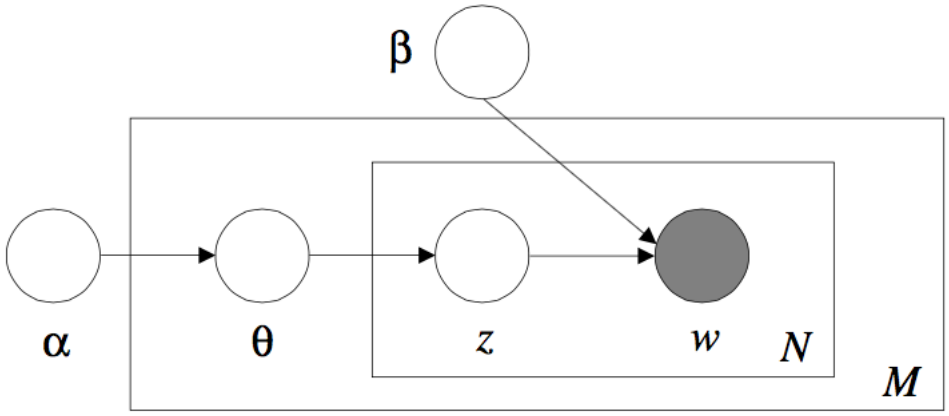
与

$$\mathbf{P}(X_i \mid X_{i-1}, \ldots, X_1) = \mathbf{P}(X_i \mid Parents(X_i))$$

话说回来，这些又与 LDA 有什么关系呢？

3、LDA 的图模型

实际上，LDA 的模型是贝叶斯网络的延伸——它的节点变成了概率分布，甚至控制概率分布的分布。它的图模型如下所示 (Blei, Ng, & Jordan, 2003)：



其中，w 表示单词（word），z 表示话题（topic），α，β，θ 均为控制概率分布的参数。

那么，根据之前贝叶斯网络的语义，给定参数α，β，我们可以得到概率公式：

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

再对照上面的语义公式，我们就可以理解上面这一个等式了（注意 z，w 均为向量，因为一篇文章有许多单词 w，也就对应了多个话题 z）：根据图模型，α 为θ 的父节点，θ 为 z 的父节点，z 与β 共为 w 的父节点。由于 w 在文章中假设相互独立（Bag of words 模型），所以可以直接用连乘将概率直接相乘。

实际上，LDA 的确也是这样设计的，它假设了这样一个文档（同一文本库内）的生成过程：

1. Choose $N \sim \text{Poisson}(\xi)$.
2. Choose $\theta \sim \text{Dir}(\alpha)$.
3. For each of the N words w_n:
   (a) Choose a topic $z_n \sim \text{Multinomial}(\theta)$.
   (b) Choose a word w_n from $p(w_n \mid z_n, \beta)$, a multinomial probability conditioned on the topic z_n.

用人话来讲，是这样的：

1. 由泊松分布选一个 N，这个 N 是这篇文档的字数。
2. 根据狄利克雷分布选一个θ，这个θ是多项分布的参数。
3. 对于这篇文档的 N 个单词，每个单词 w 都是这样得到的：首先根据之前的以 θ 为参数的多项分布选一个话题 z，再由以 z 与β为父节点的在单词 w 上的条件概率分布，选择一个 w。

4、生成模型的使用

理解了 LDA 的生成模型，它又是怎么使用的呢？回忆之前讲到的贝叶斯定理，它神奇地利用已知推出未知，还能将先验转化为后验。

还是牙痛的例子，一般来说，我们正着使用是这样的：给定一个人牙痛，可以根据贝叶斯定理得到他有蛀牙的概率；如果我们反着用，就是这样的：给定一个人有蛀牙，根据贝叶斯定理，我们还能知道他发生牙痛的概率。

将这样的思想应用到 LDA 的模型上来，可以看到，当正着使用时，给定一个话题，我们能得到出现某个单词的概率，反着使用时（这也是我们的目的），给定一个单词，我们可以根据贝叶斯定理得到出现某话题的概率。具体地，是这样一个概率：

$$p(\theta, \mathbf{z} \mid \mathbf{w}, \alpha, \beta) = \frac{p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta)}{p(\mathbf{w} \mid \alpha, \beta)}$$

由于这个概率公式是难解的（intractable），所以作者使用了变分推断（Variational Inference）、EM 算法等等来进行近似求解，这些求解步骤有着巨量的数学、统计学细节，所以才是迷惑初学者们的大坑，再加上本人确实也才疏学浅，这里也就不深挖了。

三、LDA 的应用

最直接的，是将 LDA 直接用于分词，例如论文中的例子：

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

对方框中的一段文字使用 LDA，就可以将词语所属的话题（Topic）区分出来，从而得到了方框上部的结果。

以话题模型为延伸，LDA 自然也能应用到推荐系统中。比如使用 LDA 将用户查询的词汇对应到话题的维度，再推荐与该话题相关的内容。或者将文章映射到话题的维度，获得相似话题的文章，直接在这些文章的集合中进行推荐 (Not_GOD, 2014)。

与以卷积神经网络为基础的 Deep learning 类似，LDA 所做的也是通过一些方法将细粒度的特征组合到一个新的空间，例如话题空间上去 (丕子, 2013)。它提出的实际上是一种以概率图模型为基础提取特征的方法。一项研究是否有价值，不仅要看它在当时是否能解决一个现实问题，还要看它是否能以自己为沃土激发更多的研究，如今这种方法越来越受到重视，SIGKDD2014 推荐系统部分就有多篇文章从这里发散开来（ (Yuan,

Cong, & Lin, 2015), (Charlin, Zemel, & Larochelle, 2015), (Diao, Qiu, Wu, Smola, Jiang, & Wang, 2015))，充分说明了 LDA 的价值与潜力。

Reference

BleiM.David, NgY.Andrew, & JordanI.Michael. (2003 年 03 月 1 日). Latent Dirichlet Allocation. Journal of Machine Learning Research.

CharlinLaurent, ZemelS.Richard, & LarochelleHugo. (2015). Leveraging User Libraries to Bootstrap Collaborative Filtering. SIGKDD.

DiaoQiming, QiuMinghui, WuChao-Yuan, SmolaJ.Alexander, JiangJing, & WangChong. (2015). Jointly Modeling Aspects, Ratings and Sentiments for Movie Recommendation (JMARS). SIGKDD.

Not_GOD. (2014 年 11 月 24 日). LDA 话题模型与推荐系统 . 检索来源: 简书: http://www.jianshu.com/p/50295398d802

RussellStuart, & NorvigPeter. (2010). Artificial Intelligence: A Modern Approach (Third Edition 版本). PearsonEducation,Inc.,.

Wikipedia. (无日期). Wikipedia/贝叶斯定理. 检索来源: Wikipedia: https://zh.wikipedia.org/wiki/%E8%B4%9D%E5%8F%B6%E6%96%AF%E5%AE%9A%E7%90%86

Wikipedia. (无日期). Wikipedia/泊松分布. 检索来源: Wikipedia: https://zh.wikipedia.org/wiki/%E6%B3%8A%E6%9D%BE%E5%88%86%E4%BD%88

Wikipedia. (无日期). Wikipedia/概率分布. 检索来源: Wikipedia: https://zh.wikipedia.org/wiki/%E6%A6%82%E7%8E%87%E5%88%86%E5%B8%83

Wikipedia. (无日期). Wikipedida/正态分布. 检索来源: Wikipedia/正态分布: https://zh.wikipedia.org/wiki/%E6%AD%A3%E6%80%81%E5%88%86%E5%B8%83

YuanQuan, CongGao, & LinChin-Yew. (2015). COM: a Generative Model for Group Recommendation. SIGKDD.

丕子. (2013 年 03 月 24 日). LDA（latent dirichlet allocation）的应用 . 检索来源: 丕子: http://www.zhizhihu.com/html/y2013/4219.html