007 | LDA模型的前世今生

2017-10-18 洪亮劼

AI技术内参 进入课程 >

讲述：初明明
时长 11:12 大小 5.13M

在文本挖掘中，有一项重要的工作就是分析和挖掘出文本中隐含的结构信息，而不依赖任何提前标注的信息。今天我要介绍的是一个叫做**LDA**（Latent Dirichlet Allocation）的模型，它在过去十年里开启了一个领域叫**主题模型**。

从 LDA 提出后，不少学者都利用它来分析各式各样的文档数据，从新闻数据到医药文档，从考古文献到政府公文。一段时间内，LDA 成了分析文本信息的标准工具。从最原始的 LDA 发展出来的各类模型变种，则被应用到了多种数据类型上，包括图像、音频、混合信息、推荐系统、文档检索等等，各类主题模型变种层出不穷。下面我来简单剖析一下 LDA 这个模型，聊聊它的模型描述以及训练方法等基础知识。

LDA 的背景介绍

LDA 的论文作者是戴维·布雷（David Blei）、吴恩达和迈克尔·乔丹（Michael Jordan）。这三位都是今天机器学习界炙手可热的人物。论文最早发表在 2002 年的神经信息处理系统大会（Neural Information Processing Systems，简称 NIPS）上，然后长文章（Long Paper）于 2003 年在机器学习顶级期刊《机器学习研究杂志》（Journal of Machine Learning Research）上发表。迄今为止，这篇论文已经有超过 1 万 9 千次的引用数，也成了机器学习史上的重要文献之一。

论文发表的时候，戴维·布雷还在加州大学伯克利分校迈克尔手下攻读博士。吴恩达当时刚刚从迈克尔手下博士毕业来到斯坦福大学任教。戴维 2004 年从伯克利毕业后，先到卡内基梅隆大学跟随统计学权威教授约翰·拉弗蒂（John Lafferty）做了两年的博士后学者，然后又到东部普林斯顿大学任教职，先后担任助理教授和副教授。2014 年转到纽约哥伦比亚大学任统计系和计算机系的正教授。戴维在 2010 年获得斯隆奖（Alfred P. Sloan Fellowship，美国声誉极高的奖励研究人员的奖项，不少诺贝尔奖获得者均在获得诺贝尔奖多年之前获得过此奖），紧接着又在 2011 年获得总统青年科学家和工程师早期成就奖（Presidential Early Career Award for Scientists and Engineers，简称 PECASE）。目前他所有论文的引用数超过了 4 万 9 千次 。

吴恩达在斯坦福晋升到副教授后，2011 年到 2012 年在 Google 工作，开启了谷歌大脑（Google Brain）的项目来训练大规模的深度学习模型，是深度学习的重要人物和推动者之一。2012 年他合作创建了在线学习平台 Coursera，可以说是打开了慕课（Massive Open Online Course，简称 MOOC）运动的大门。之后吴恩达从 2014 年到 2017 年间担任百度首席科学家，并创建和运行了百度在北美的研究机构。目前他所有论文的引用数超过 8 万 3 千次。

文章的第三作者迈克尔·乔丹是机器学习界的泰斗人物。他自 1998 年在加州大学伯克利任教至今，是美国三个科学院院士（American Academy of Arts and Sciences、National Academy of Engineering 以及 National Academy of Sciences），是诸多学术和专业组织的院士（比如 ACM、IEEE、AAAI、SIAM 等）。迈克尔可以说是桃李满天下，而且其徒子徒孙也已经遍布整个机器学习领域，不少都是学术权威。他的所有论文有多达 12 万次以上的引用量。

值得注意的是，对于三位作者来说，LDA 论文都是他们单篇论文引用次数最多的文章。

LDA 模型

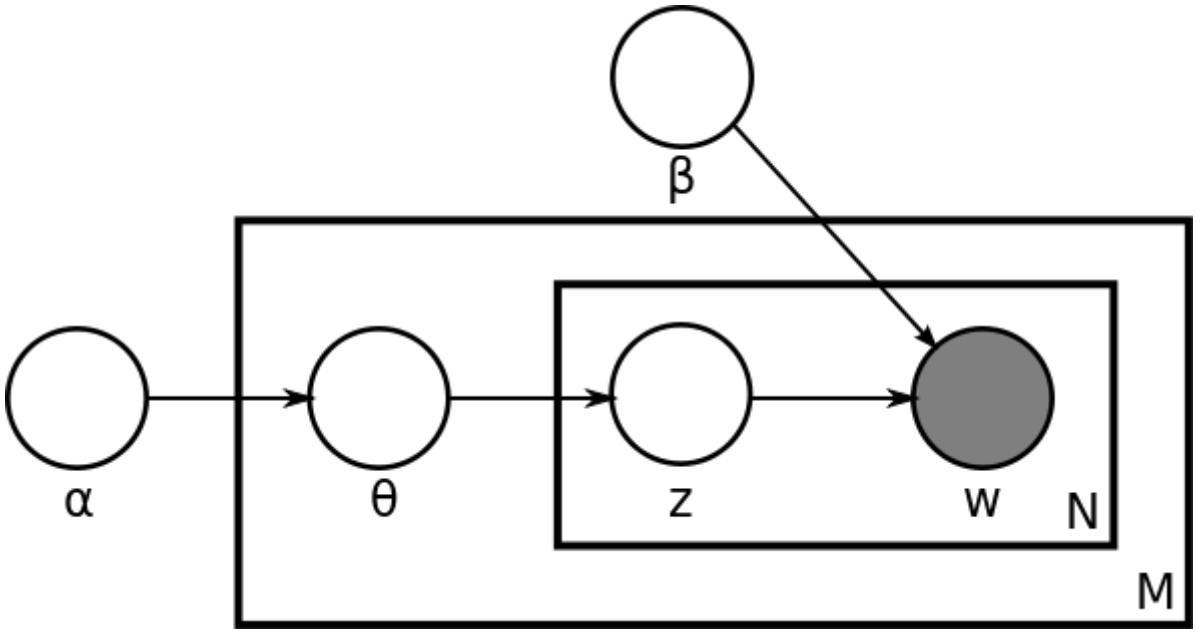
要描述 LDA 模型，就要说一下 LDA 模型所属的**产生式模型**（Generative Model）的背景。产生式模型是相对于**判别式模型**（Discriminative Model）而说的。这里，我们假设需要建模的数据有特征信息，也就是通常说的 X，以及标签信息，也就是通常所说的 Y。

判别式模型常常直接对 Y 的产生过程（Generative Process) 进行描述，而对特征信息本身不予建模。这使得判别式模型天生就成为构建分类器或者回归分析的有利工具。而产生式模型则要同时对 X 和 Y 建模，这使得产生式模型更适合做无标签的数据分析，比如聚类。当然，因为产生式模型要对比较多的信息进行建模，所以一般认为对于同一个数据而言，产生式模型要比判别式模型更难以学习。

一般来说，产生式模型希望通过一个产生过程来帮助读者理解一个模型。注意，这个产生过程本质是描述一个**联合概率分布**（Joint Distribution）的分解过程。也就是说，这个过程是一个虚拟过程，真实的数据往往并不是这样产生的。这样的产生过程是模型的一个假设，一种描述。任何一个产生过程都可以在数学上完全等价一个联合概率分布。

LDA 的产生过程描述了文档以及文档中文字的生成过程。在原始的 LDA 论文中，作者们描述了对于每一个文档而言有这么一种生成过程：

1. 首先，从一个全局的泊松（Poisson）参数为β的分布中生成一个文档的长度 N；
2. 从一个全局的狄利克雷（Dirichlet）参数为α的分布中生成一个当前文档的θ；
3. 然后对于当前文档长度 N 的每一个字执行以下两步，一是从以θ为参数的多项（Multinomial）分布中生成一个主题（Topic）的下标（Index）z_n；二是从以φ和 z 共同为参数的多项分布中产生一个字（Word）w_n。



从这个描述我们可以马上得到这些重要的模型信息。第一，我们有一个维度是 K 乘以 V 的主题矩阵（Topic Matrix）。其中每一行都是一个φ，也就是某一个生成字的多项分布。当然，这个主题矩阵我们在事先并不知道，是需要学习得到的。另外，对每一个文档而言，θ是一个长度为 K 的向量，用于描述当前文档在 K 个主题上的分布。产生过程告诉我们，我们对于文档中的每一个字，都先从这个θ向量中产生一个下标，用于告诉我们现在要从主题矩阵中的哪一行去生成当前的字。

这个产生模型是原论文最初提出的，有两点值得注意。

第一，原始论文为了完整性，提出了使用一个泊松分布来描述文档的长度这一变化信息。然而，从模型的参数和隐变量的角度来说，这个假设并不影响整个模型，最终作者在文章中去除了这个信息的讨论。在主题模型的研究中，也较少有文献专注这个信息。

第二，原始论文并没有在主题矩阵上放置全局的狄利克雷分布作为先验概率分布。这一缺失在后续所有的主题模型文献中得到修正。于是今天标准的 LDA 模型有两类狄利克雷的先验信息，一类是文档主题分布的先验，参数是α，一类是主题矩阵的先验，参数是β。

文章作者们把这个模型和当时的一系列其他模型进行了对比。比如说，LDA 并不是所谓的狄利克雷 - 多项（Dirichlet-Multinomial）聚类模型。这里，LDA 对于每个文档的每一个字都有一个主题下标。也就是说，从文档聚类的角度来看，LDA 是没有一个文档统一的聚类标签，而是每个字有一个聚类标签，在这里就是主题。这也是 LDA 是**Mixed-Membership 模型**的原因——每个字有可能属于不同的类别、每个文档也有可能属于不同的类别。

LDA 很类似在 2000 年初提出的另外一类更简单的主题模型——概率隐形语义索引（Probabilistic Latent Semantic Indexing），简称**PLSI**。其实从本质上来说，LDA 借用了 PLSI 的基本架构，只不过在每个文档的主题分布向量上放置了狄利克雷的先验概率，以及在主题矩阵上放置了另外一个狄利克雷的先验概率。

尽管看上去这是一个非常小的改动，但是这样做的结果则是 LDA 的参数个数并不随着文档数目的增加而增加。那么，相对于 PLSI 来说，LDA 并不容易对训练数据**过度拟合**（Overfitting）。

值得注意的，原始文章说过度拟合主要是指，对于 PLSI 而言，文档的主题分布向量是必须需要学习的，而这个向量对于 LDA 是可以被忽略或者说是并不需要保存的中间变量。然而

在实际的应用中，我们其实常常也需要这个向量的信息，因此这部分对于过度拟合的讨论在后来的应用中并没有特别体现。

LDA 模型的训练和结果

LDA 虽然从某种意义上来说仅仅是在 PLSI 上增加了先验信息。然而，这一个改动为整个模型的训练学习带来了非常大的挑战。应该说，整个 LDA 的学习直到模型提出后近 10 年，才随着**随机变分推理**（Stochastic Variational Inference）的提出以及基于**别名方法**（Alias Method）的抽样算法（Sampling Method）而得以真正的大规模化。一直以来，LDA 的训练学习都是一件很困难的事情。

不像 PLSI 可以依靠**最大期望（EM）算法**得以比较完美的解决，传统上，LDA 的学习属于**贝叶斯推理**（Bayesian Inference），而在 2000 年代初期，只有**马尔科夫蒙特卡洛**（Markov chain Monte Carlo），简称 MCMC，以及迈克尔·乔丹等人推崇的**变分推理**（Variational Inference），简称 VI，作为工具可以解决。这篇文章因为出自迈克尔的实验室，当仁不让地选择了 VI。比较有意思的是，后续大多数 LDA 相关的论文都选择了 MCMC 为主的**吉布斯**（Gibbs）采样来作为学习算法。

VI 的完整讲解无法在本文涵盖。从最高的层次上来理解，VI 是选取一整组简单的、可以优化的所谓变分分布（Variational Distribution）来逼近整个模型的后验概率分布。当然，由于这组分布的选取，有可能会为模型带来不小的误差。不过好处则是这样就把贝叶斯推理的问题转化成了优化问题。

从 LDA 的角度来讲，就是要为θ以及 z 选取一组等价的分布，只不过更加简单，更不依赖其他的信息。在 VI 进行更新θ以及 z 的时候，算法可以根据当前的θ以及 z 的最新值，更新α的值（这里的讨论依照原始的 LDA 论文，忽略了β的信息）。整个流程俗称**变分最大期望（Variational EM）算法**。

文章在 TREC AP 的文档数据中做了实验。首先，作者们使用了一个叫**困惑度**（Perplexity）的评估值来衡量文档的建模有效程度，这个值越低越好。LDA 在好几个数据集中都明显好于 PLSI 以及其他更加简单的模型。从这篇文章之后，主题模型的发展和对比都离不开困惑度的比较，也算是开启了一个新时代。

然后，作者们展示了利用 LDA 来做文档分类，也就是利用文档主题向量来作为文档的特征，从而放入分类器中加以分类。作者们展示了 LDA 作为文档分类特征的有力证据，在数

据比较少的情况下优于文本本身的特征。不过总体说来，在原始的 LDA 论文中，作者们并没有特别多地展现出 LDA 的所有可能性。

小结

今天我为你梳理了 LDA 提出的背景以及这篇论文所引领的整个领域的情况。你需要掌握的核心要点：第一，论文作者们目前的状态；第二，LDA 模型本身和它的一些特点；第三，LDA 的训练流程概况以及在原始文章中的实验结果。

最后，我为你留一个思考题：LDA 的产生过程决定了对于一个文本而言，每个字都可能来自不同的主题，那么如果你希望，对于某一个段落，所有的文字都来自同一个主题，你需要对 LDA 这个模型进行怎么样的修改呢？

欢迎你给我留言，和我一起讨论。

上一篇　006 | Google的点击率系统模型

下一篇　008 | 曾经辉煌的雅虎研究院

精选留言 (6)

写留言

陈俊
2017-10-24

3

对于非统计学或cs专业的同学，上一期和这一期似乎有点深奥。但是还是十分感谢老师的辛苦付出！

沉
2017-10-19

2

如果能在文中给出一些可以详细参考的资料就更好了 ~

展开

兔子ORZ
2018-04-14

1

有监督的LDA比较好的形式就是LLDA，在plate notation上的theta上再加上一个观察变量表示主题标签，而这个观察变量也是基于狄利克雷超参的

unicornmm
2018-01-01

1

希望能够给出每篇文章的链接，非常感谢老师的讲解

展开

潜行
2018-08-28

希望老师给文章链接，另外提到很多统计学的知识，还看不懂。。得好好消化

展开

惜心（伟祺...
2018-03-27

文章后的问题一段属于同一主题 是不是可以在生成这段字时候公用生成主题的theta参数值就相当于同一段落同一主题了