084 | LDA变种模型知多少

2018-04-16 洪亮劼

AI技术内参 进入课程 >

讲述：初明明
时长 07:48 大小 3.57M

我们在之前的分享中曾经介绍过文本挖掘（Text Mining）中的重要工具 LDA（Latent Diriclet Allocation）的基本原理。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（Labeled）的信息。也就是说，我们希望能够利用文本挖掘技术来对无标签的数据进行挖掘，这是典型的无监督学习。

LDA 就是一个出色的无监督学习的文本挖掘模型。这个模型在过去的十年里开启了主题模型（Topic Model）这个领域。不少学者都利用 LDA 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。在一段时间内，LDA 成为了分析文本信息的标准工具。而从最原始的 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，可以说各类主题模型变种层出不穷。

今天我们就结合几篇经典论文，来看一看**LDA 的各种扩展模型**。当然，在介绍今天的内容之前，我们首先来回顾一下 LDA 模型的一些基本信息。

LDA 模型的回顾

LDA 模型是一个典型的**产生式模型**（Generative Model）。产生式模型的一大特点就是通过一组概率语言，对数据的产生过程进行描述，从而对现实数据建立一个模型。注意，这个产生过程的本质是描述的一个**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的过程，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产生过程。在原始的 LDA 论文中，作者们描述了对于每一个文档而言的产生过程。

LDA 模型的前世今生

相比于传统的文本聚类方法，LDA 对于每个文档的每一个字都有一个主题下标，也就是说，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。

LDA 模型的训练一直是一个难点。传统上，LDA 的学习属于**贝叶斯推断**（Bayesian Inference），而在 2000 年初期，只有**MCMC 算法**（Markov chain Monte Carlo，马尔科夫链蒙特卡洛）以及 **VI**（Variational Inference，变分推断）作为工具可以解决。在最初的 LDA 论文里，作者们采用了 VI；后续大多数 LDA 相关的论文都选择了 MCMC 为主的吉布斯采样（Gibbs Sampling）来作为学习算法。

LDA 的扩展

当 LDA 被提出以后，不少学者看到了这个模型的潜力，于是开始思考怎么把更多的信息融入到 LDA 里面去。通过我们上面的讲解，你可以看到，LDA 只是对文档的文字信息本身进行建模。但是绝大多数的文档数据集还有很多额外的信息，如何利用这些额外信息，就成为了日后对 LDA 扩展的最重要的工作。

第一个很容易想到的需要扩展的信息就是作者信息。特别是 LDA 最早期的应用，对于一般的文档来说，比如科学文档或者新闻文档，都有作者信息。很多时候我们希望借用作者在写

文档时的遣词造句风格来分析作者的一些写作信息。那么，如何让 LDA 能够分析作者的信息呢?

这里我们分享一篇论文《用于作者和文档信息的作者主题模型》（The author-topic model for authors and documents）[1]，这是最早利用额外信息到 LDA 模型中的扩展模型。文章提出的模型叫作“**作者 LDA**”（Author LDA）。这个模型的主要思想是，每篇文档都会有一些作者信息，我们可以把这些作者编码成为一组下标（Index）。对于每一个文档来说，我们首先从这组作者数组中，选出一个当前的作者，然后假定这个作者有一组相对应的主题。这样，文档的主题就不是每个文档随机产生了，而是每个作者有一套主题。这个时候，我们从作者相对应的主题分布中取出当前的主题，然后再到相应的语言模型中，采样到当前的单词。

可以看到，作者 LDA 和普通的 LDA 相比，最大的不同就是**主题分布不是每个文档有一个，而是每个作者有一个。**这个主题分布决定着当前的单词是从哪一个语言模型中采样的单词。作者 LDA 也采用吉布斯采样的方法学习，并且通过模型的学习之后，能够看得出不同作者对于文档的影响。

从作者 LDA 之后，大家看出了一种扩展 LDA 的思路，那就是**依靠额外的信息去影响主题分布，进而影响文档字句的选择。**这种扩展的方法叫作“**上游扩展法**”（Upstream）。什么意思呢？就是说把希望对模型有影响的信息，放到主题分布的上游，去主动影响主题分布的变化。**这其实是概率图模型的一种基本的思路，那就是把变量放到这个产生式模型的上游，使得下游的变量受到影响。**

那你可能要问，有没有把需要依赖的变量放到下游的情况呢？答案是肯定的。我们再来看一篇论文《同时进行图像分类和注释》（Simultaneous image classification and annotation）[2]，这篇文章就发明了一种方法。具体来说，文章希望**利用 LDA 到多模数据领域**（Multiple Modal）。也就是数据中可能有文字，也可能有图像，还可能有其他信息。在这样的多模数据的情况下，如何让 LDA 能够对多种不同的数据进行建模呢?

这里面的基本思路就是认为**所有的这些数据都是通过主题分布产生的。**也就是说，一个数据点，我们一旦知道了这个数据点内涵的主题（比如到底是关于体育的，还是关于金融的），那么我们就可以产生出和这个数据点相关的所有信息，包括文字、图像、影音等。

具体到这篇文章提出的思路，那就是这组数据的图像标签以及图像所属的类别都是主题产生的。我们可以看到，和之前的作者 LDA 的区别，那就是其他信息都是放在主题变量的下游

的，希望通过主题变量来施加影响。

这两种模型代表了一系列丰富的关于 LDA 的扩展思路，那就是**如何把扩展的变量设置在上游或者是下游，从而能够对主题信息产生影响或者是受到主题信息的影响。**

除此以外，LDA 的另外一大扩展就是**把文档放到时间的尺度上，希望去分析和了解文档在时间轴上的变化。**这就要看经典的论文《动态主题模型》（Dynamic topic models）[3]。这篇论文最后获得了 ICML 2010 年的最佳贡献奖。那么，我们怎么修改 LDA 使其能够理解时间的变化呢？很明显，还是需要从主题分布入手，因为主题分布控制了究竟什么文字会被产生出来。因此，我们可以认为主题分布会随着时间的变化而变化。

在之前的模型中，我们已经介绍了，每个文档的主题分布其实来自一个全局的狄利克雷（Diriclet ）先验分布。那么，我们可以认为不同时间的先验分布是不一样的，而这些先验分布会随着时间变化而变化。怎么能够表达这个思想呢？作者们用到了一个叫“**状态空间**”（State-Space）的模型。简而言之，状态空间模型就是把不同时间点的狄利克雷分布的参数给串起来，使得这些分布的参数会随着时间的变化而变化。**把一堆静态的参数用状态空间模型串接起来**，可以说是这篇文章开创的一个新的思维。

总结

今天我为你梳理了 LDA 的扩展模型。LDA 的扩展当然还有很多，我们今天讨论了几个非常经典的扩展思路，分别是基于上游、下游和时间序列的 LDA 扩展模型。

一起来回顾下要点：第一，我们回顾了 LDA 这个模型的核心思想；第二，我们聊了如何把文档的其他信息融入到 LDA 模型中去，以及如何对时间信息进行建模。

最后，给你留一个思考题，如果我们希望利用 LDA 来对“用户对商品的喜好”进行建模，应该怎么对模型进行更改呢？

欢迎你给我留言，和我一起讨论。

参考文献

1. Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. The author-topic model for authors and documents. Proceedings of the 20th

conference on Uncertainty in artificial intelligence (UAI '04). AUAI Press, Arlington, Virginia, United States, 487-494, 2004.

2. C. Wang, D. Blei., and L. Fei-Fei. Simultaneous image classification and annotation. Computer Vision and Pattern Recognition, 2009.

3. D.Blei and J.Lafferty. Dynamic topic models. Proceedings of the 23rd International Conference on Machine Learning, 2006.

上一篇　复盘 2 | 推荐系统核心技术模块

下一篇　085 | 针对大规模数据，如何优化LDA算法?

精选留言 (2)

写留言

林彦
2018-04-18

1

通过LDA生成用户的兴趣主题(商品的语义标签是一种数据来源)，这个过程有些类似于生成文档。然后根据这些用户兴趣主题来寻找匹配的商品，比如计算和商品主题的相似度。

展开

Jack_Saini...

2018-04-16

1

每个用户看做一篇文档，用户选择的商品视作文档中的每个词。

展开