085 | 针对大规模数据，如何优化LDA算法?

2018-04-18 洪亮劼

AI技术内参

进入课程 >

讲述：初明明

时长 07:21　大小 3.36M

周一，我们分享了 LDA（Latent Diriclet Allocation）的各种扩展模型，介绍了基于上游的和下游的两种把额外信息融入到 LDA 模型中的方法。同时，我们也讨论了在时间尺度上如何把 LDA 模型扩展到可以“感知”不同的时间段对于模型的影响。以 LDA 为代表的主题模型在过去的十年间发展出了一整套的扩展，为各式各样的应用场景提供了有力的工具。

尽管 LDA 在模型的表达力上给研究者们提供了把各种场景和模型结合的可能性，但是 LDA 的训练过程比较复杂，而且速度也比较慢。因此，如何能够把 LDA 真正应用到工业级的场景中，对于很多人来说，都是一件煞费苦心的事情。今天我们就来聊聊**LDA 的算法优化问题**。

LDA 模型训练

我们首先来回顾一下 LDA 模型的训练过程，从高维度上为你分析一下为什么这个过程很困难。

LDA 模型中最重要的未知变量就是每个单词对应的**主题下标**（Index）或者说是**主题“赋值”**（Assignment）。这个主题下标是从每个文档对应的主题分布中“采样”得来的。每个文档的主题分布本身也是一个未知的多项式分布，用来表达当前这个文档的所属主题，比如有多少百分比属于运动、有多少百分比属于金融等等。这个分布是从一个全局的狄利克雷（Diriclet）分布中产生的。狄利克雷分布在这里起到了**超参数**的作用，其参数的取值往往也是未知的。但是我们可以根据一些**经验值**对其进行设置。除了每个文档的主题分布和主题赋值以外，我们还需要对全局的主题语言模型进行估计。这些语言模型直接决定了，各类词语出现的概率是多少。

流行的 LDA 训练方法有两个，一个是基于**吉布斯采样**（Gibbs Sampling）的随机方法，一个是基于**变分推断**（Variational Inference）的确定性方法（Deterministic）。这两种方法的初始形态都无法应对大型数据。这里我们来简要介绍一下这两种方法。

吉布斯采样主要是针对主题赋值进行采样，最开始是完全随机的结果，但是慢慢会收敛到参数的后验概率的真值。这里面比较慢的一个原因，是这个收敛过程可能需要几百到几千个不等的迭代。同时，吉布斯采样只能一个文档一个文档进行，所有的数据结构都需要在采样的过程中进行更改。这个过程比较慢的另外一个原因，是吉布斯采样的核心是如何对一个离散分布进行采样。而离散分布采样本身，如果在分布的参数变化的情况下，最好能够达到 O（KlogK），这里 K 是主题的数目。因此，从原理上来说，这也是阻碍吉布斯采样能够扩展到大型数据的一个原因。

变分推断的思路则和吉布斯采样很不一样。它是把对隐含参数的估计问题变成一个确定性的优化问题，这样我们就可以**利用种种优化算法来解决贝叶斯推断的问题**。不过和吉布斯采样相比，变分推断存在一个问题，因为这种方法并不是解决原来的优化问题，因此新的优化问题可能并不能带来原来问题的解。同时，变分推断也需要一个文档一个文档单独处理，因此推广到大规模数据上有其局限性。

LDA 的大规模优化算法

顺着我们刚才提到的问题，为了把吉布斯采样和变分推断扩大到大规模数据上，学者们有针对性地做了很多探索。我们下面就分别对这两种思路展开简要的介绍。

首先，我们来看吉布斯采样。吉布斯采样慢的一个核心就是我们刚才说的，需要从一个离散分布中采样出一个样本，在我们这个例子中也就是每个单词的主题赋值。那么，有没有什么方法让这个步骤加速呢？答案是，有的。

在 KDD 2009 上发表了一篇论文《应用于流文档集合的主题模型推断的高效方法》（Efficient methods for topic model inference on streaming document collections）[1]，算是在这方面取得突出成绩的一个重要参考文献。这篇论文的主要贡献就是，对原有的采样公式进行了一个比较仔细的分析。

作者们发现，原来的吉布斯采样公式可以被分解为几个部分：和全局的语言模型有关、和文档有关以及和当前需要采样的单词有关。这是一个非常有价值的观察，之后很多加速吉布斯采样的工作基本上都采用了类似的思路，也就是**试图把原始的吉布斯采样公式拆分成好几个组成部分，并且每一个部分所代表数据的变化率是不一样的。**

以这篇文章提出的方法来说，全局语言模型在每个文档的采样过程中是不变的，于是这部分的计算不需要每个单词都重算。同理，只与文档相关的部分，也可以每个单词的采样过程中，只算一次，而不需要每个主题算一次。在这样一个简化了的流程里，采样速度得到了极大的提升。

在这篇文章之后，通过吉布斯采样这个方法，LDA 的采样速度还是没有得到明确的提升，直到《降低主题模型的采样复杂度》（Reducing the sampling complexity of topic models）[2] 这篇论文的出现。这篇论文获得了 KDD 2014 年的最佳论文奖。文章的思想还是针对吉布斯采样的公式，不过这一次，拆分的方法略不一样。作者们把采样的公式拆分成了与当前文档有关系的一部分，以及和当前文档没关系的全局语言模型的部分。

同时，作者们提出了一个“**Alias 方法**”（Alias Method），简称**A 算法**，来加速采样。这个 A 算法其实并不是作者们为了 LDA 发明的，而是一个普遍的可以对离散分布采样的一个算法。A 算法的核心思想是，如果我们要针对一个分布进行反复采样，那么就可以建立一种数据结构，使得这种采样只有在第一遍的时候有一定的计算成本，而后都会以 O(1) 的成本进行采样。这个方法极大地加速了 LDA 通过吉布斯采样的效率。值得一提的是，在这篇论文之后，很多研究者发布了一系列的后续工作。

那么在变分推断的道路上，有没有什么方法能够加速呢？答案依然是肯定的。

这方面的代表作无疑就是论文《LDA 的在线学习》（Online learning for Latent Dirichlet Allocation）[3]。

我们回到变分推断的场景中，把一个贝叶斯推断的问题变成了优化的问题。那么，在优化的场景里，是怎么针对大规模数据的呢？

在优化的场景里，特别是基于梯度（Gradient）的优化方法中，大数据的应用往往需要 SGD（Stochastic Gradient Descent，随机梯度下降）的方法。通俗地讲，就是在计算梯度的时候，我们不需要处理完所有的数据之后才计算一次梯度，而是针对每一个文档，都可以计算一次梯度的估计值。

作者们其实就是把这个思想给搬到了变分推断里。总的来说，新发明出来的变分推断其实就是希望能够**推演出一种类似 SGD 的变分方法**，这种方法在后来的很多论文中都有所应用。

总结

今天我为你梳理了 LDA 优化算法的相关知识。

一起来回顾下要点：第一，我们聊了聊 LDA 这个模型的优化算法为什么会有难度，特别是针对吉布斯采样和变分推断这两种思路来说难点在哪里；第二，我们分享了当前加速 LDA 算法的两种思路，主要讨论了两种思路的一些核心思想，希望能够起到抛砖引玉的作用。

最后，给你留一个思考题，除了在算法层面希望能够加速 LDA 以外，我们能否利用并行化对 LDA 进行加速呢？

欢迎你给我留言，和我一起讨论。

参考文献

1. Limin Yao, David Mimno, and Andrew McCallum. Efficient methods for topic model inference on streaming document collections. Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '09). ACM, New York, NY, USA, 937-946, 2009.

2. Aaron Q. Li, Amr Ahmed, Sujith Ravi, and Alexander J. Smola. Reducing the sampling complexity of topic models. Proceedings of the 20th ACM SIGKDD

international conference on Knowledge discovery and data mining (KDD '14). ACM, New York, NY, USA, 891-900, 2014.

3. Matthew D. Hoffman, David M. Blei, and Francis Bach. Online learning for Latent Dirichlet Allocation. Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 1 (NIPS'10), 2010.

上一篇　084 | LDA变种模型知多少

下一篇　086 | 基础文本分析模型之一：隐语义分析

精选留言 (2)

写留言

帅帅
2018-10-23

我在spark mllib中看到了LDA的实现，应该就是并行化的实现

展开

vick_zh

2018-06-22

请问：LDA是否又被深度学习方法替代的趋势？所以在应用价值显得没那么“重要”？