

051 | 精读2017年NIPS最佳研究论文之一：如何解决非凸优化问题？

2018-01-29 洪亮劼

AI技术内参

[进入课程 >](#)



讲述：初明明

时长 07:26 大小 3.41M



机器学习与人工智能领域的顶级会议 NIPS (Conference on Neural Information Processing Systems, 神经信息处理系统大会) 从 1987 年开始举办, 已经有 30 多年的历史。NIPS 2017 大会于 2017 年 12 月 4 日到 9 日在美国加利福尼亚州的长滩 (Long Beach) 举行。

每年大会都会在众多的学术论文中挑选出几篇最有新意和价值的论文作为最佳研究论文。在 NIPS 2017 上, 一共有三篇论文获得了最佳论文的称号。今天, 我就来带你认真剖析一下其中的一篇《具有凸目标的基于方差的正则化》([Variance-based Regularization with Convex Objectives](#))。这篇论文的两作者都是来自斯坦福大学的学者。

这篇文章理论性很强，主要研究的是一种“健壮和优化问题”（Robust Optimization），也就是说我们在优化一个“损失函数”（Loss Function）的时候，不仅要考虑损失函数的“均值”（Mean），还要考虑损失函数的“方差”（Variance）。然而，一个既要考虑均值又要考虑方差的综合的损失函数，往往是一个“非凸”（Non Convex）的问题。对于一般的非凸优化问题来说，我们往往不能找到一个全局的最优解，甚至是找到局部最优解也很困难。这篇文章就是要来解决这么一个问题。

作者群信息介绍

第一作者洪升·南空（Hongseok Namkoong）是斯坦福大学“运筹学”（Operations Research）的一名在读博士研究生。他的导师分别是约翰·达齐（John C. Duchi）和彼得·格林（Peter W. Glynn）。2013年到斯坦福之前，南空在韩国的韩国科学与技术高级研究所（Korea Advanced Institute of Science and Technology），有时候又称为 KAIST，获得工业工程和数学学士学位。最近两三年，南空已经在发表了两篇 NIPS 的文章（包括这篇最佳论文），以及一篇 ICML 的论文。

第二作者约翰·达齐（John C. Duchi）是南空的导师之一。达奇可以说是师出名门，他于 2007 年从斯坦福本科毕业，接着在斯坦福跟随机器学习权威达菲·科勒（Daphne Koller），拿到了计算机科学的硕士学位；然后又到加州大学伯克利分校跟随统计学习权威迈克尔·乔丹（Michael Jordan）拿到了计算机科学的博士学位。在博士阶段的暑假里，达奇还到 Google 研究院中追随约然·辛格（Yoram Singer）积累了非常有价值的实习经验。之后，他来到了斯坦福大学担任统计和电气电子工程系的助理教授。

有了这些良好的基础，达奇的学术成绩也是非常扎实。他于 2010 年获得了 ICML 最佳论文奖。紧接着，2011 年在 Google 实习期间的工作 AdaGrad，成为了现在机器学习优化领域的经典算法，这个工作的论文有超过 2500 次的引用，而且也是深度学习优化算法的一个重要基础。目前，达奇所有论文的引用数超过 6 千次。

论文的主要贡献

我们首先来看一下这篇文章的主要贡献，理解文章主要解决了一个什么场景下的问题。

很多机器学习问题其实都可以最终归结于优化一个目标函数（Objective Function）或者有时候叫做损失函数（Loss Function）的问题。针对训练数据集上损失函数的优化（即最大化或最小化）并且在测试集上表现优异，是可以被证明为最终能够较好“泛化”（Generalization）的一种体现。

那么，通常情况下，这个损失函数都是针对均值的一个描述，比如在整个训练数据集上的平均误差，或者说在整个训练数据集上的平均准确度。然而，我们都知道，在一些很“偏斜”（Skewed）的数据分布上，均值并不是很好的一个数据描述。即便我们的函数能够在“平均”的情况下优化一个损失函数，这个函数也有可能在一些，甚至大部分数据点上表现得不尽如人意。

于是，研究人员就引入了“健壮和优化问题”。也就是我们希望损失函数在更多的点上有优异的表现。那么，**损失函数的健壮性是用损失函数的方差来衡量的**。也就是说，我们希望损失函数在不同数据点上的波动要小。

有了这个概念之后，下一步就显得比较自然了，那就是把损失函数的均值部分，也就是我们通常要做的部分和有方差的部分串联起来，形成一个新的目标函数。**这个目标函数有两个部分，第一部分就是均值部分，第二个部分就是方差的部分，中间有一个自由的参数，把这两个部分衔接起来**。这样，我们就有了一个既考虑均值又考虑方差的新的健壮化的优化问题。

然而，一个既要考虑均值又要考虑方差的综合的损失函数，往往是一个“非凸”（Non Convex）的问题。什么叫做非凸函数？**一个“凸”（Convex）问题可以简单理解为函数只有唯一的最小值，并且我们具备有效算法来找到这个最小值**。而对于非凸问题来说，我们往往不能找到一个全局的最优解，或者找到局部最优解也很困难。

健壮优化问题已经在之前的研究中提了出来，那么这篇文章的主要贡献在于，为健壮优化问题找到了一个“凸”问题的逼近表达，并基于此提出了一个优化算法，解决了这个新提出的凸问题的近似解。

这里，值得注意的一点是，**对于非凸问题提出凸问题的近似表达，是解决非凸问题的一个重要思路**。有很多经典的非凸问题，都是通过凸问题的近似来得到解决或者部分解决的。从这个思路来说，这篇文章是延续了解决这种问题的一贯的策略。

论文的核心方法

这篇论文的核心方法以及证明都有很强的理论性，需要有一定的数学功底和类似研究背景，才能更好地理解。如果对文章内容有兴趣，建议不仅要阅读原本的 NIPS 论文，还需要去阅读其附加的文档，一共有 50 多页，才能比较全面地理解这篇文章的细节。我们在这里仅仅从概念上做一个高度浓缩的概括。

作者们在文章中**提出了一种叫“健壮化的正则风险”（Robustly Regularized Risk）的目标函数**。这个新的目标函数是建立在一个叫“经验分布”（Empirical Distribution）上的“散度”（Divergence）。而这个新的健壮化正则风险是一个凸问题。

直白一点说，这个健壮化的正则风险可以被认为是一个包含两项的式子，这两项是在数据集上的损失函数的期望加上一个损失函数的方差。在这个新的两项的式子中，期望和方差都是定义在数据的经验分布上的。于是这样就把这个新提出的风险式子和我们实际需要解决的问题挂上了钩。当然后面大段的论文就是要证明这两个式子之间的差距到底有多少，是不是新的式子提供了一个比较“紧”的“界限”（Bound）。

紧接着，这篇文章其实讨论了这个健壮化的正则风险可以写成一个更加简单的优化问题，然后文章在附录中提供了这个简化后的优化问题的求解。

方法的实验效果

虽然这篇文章的核心内容是一个理论结果，或者是算法革新。但是这篇文章依然是在两个数据集中做了实验，一个是在 UCI ML 的数据集上，展示了提出的新的健壮化的目标函数达到了比一般的目标函数更好的效果；另外一个则是在 RCV1 文本分类的问题上比一般的优化目标函数有更好的效果。

小结

今天我为你讲了 NIPS 2017 年的最佳研究论文之一，文章非常理论化。文章的一个核心观点是希望能够通过对损失函数的均值和方差同时建模从而达到让目标函数健壮化的目的。

一起来回顾下要点：第一，我们简要介绍了这篇文章的作者群信息。第二，我们详细介绍了这篇文章要解决的问题以及贡献。第三，我们简要地介绍的文章提出方法的核心内容。

最后，给你留一个思考题，要想控制目标函数的预测结果的方差，除了本文提出的把均值和方差都设计到目标函数里，还有没有别的方法？

欢迎你给我留言，和我一起讨论。


AI 技术内参

你的360度人工智能信息助理

洪亮劼

Etsy 数据科学主管
前雅虎研究院资深科学家



新版升级：点击「 请朋友读」，10位好友免费读，邀请订阅更有**现金**奖励。

© 版权归极客邦科技所有，未经许可不得传播售卖。页面已增加防盗追踪，如有侵权极客邦将依法追究其法律责任。

上一篇 内参特刊 | 和你聊聊每个人都关心的人工智能热点话题

下一篇 052 | 精读2017年NIPS最佳研究论文之二：KSD测试如何检验两个分布的异同？

精选留言 (2)

 写留言



林彦

2018-01-29

 4

常规的降低方差，也就是减少过拟合的方法有

- 1) 找到一个合适的复杂度，比如目标函数多项式的次数，降低次数可以降低方差；
- 2) 引入合适的正则化参数lambda，lambda越大过拟合程度越低；
- 3) 集成学习中的bagging（传统视角）；
- 4) 减少特征数量；...

展开 



yy

2018-06-29



非常感谢！分享的真好！

展开 ✓