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?

2018-01-31 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

本周我们来分析和探讨 NIPS 2017 上的三篇最佳论文。周一我们分享的文章主要研究的是一种“健壮的优化问题”，也就是说我们在优化一个“损失函数”的时候，不仅要考虑损失函数的“均值”，还要考虑损失函数的“方差”。

今天，我们来看另外一篇最佳论文《线性时间内核拟合优度测试》（A Linear-Time Kernel Goodness-of-Fit Test），讲的是如何来衡量一组数据是否来自于某一个分布。

今天的这篇文章理论性也很强，这里我尝试从更高的维度为你做一个归纳，如果对文章内容感兴趣，建议你一定要去阅读原文。

作者群信息介绍

本文一共有五位作者，我们在这里进行一个简要介绍。

第一作者叫维特瓦特·吉特克鲁特（Wittawat Jitkrittum），刚从伦敦大学学院（University College London）的“加斯比计算人脑科学所”（Gatsby Computational Neuroscience Unit）博士毕业。他在博士期间的主要研究是“统计测试”（Statistical Tests），特别是如何利用“核方法”（Kernel Method）来对“分布特征”（Distributional Features）进行测试。吉特克鲁特在泰国完成本科学习，于日本京的东京科技学院（Tokyo Institute Of Technology）获得硕士学位。最近几年，吉特克鲁特已经在 NIPS、ICML、UAI 等会议连续发表了多篇高质量论文，可以说是统计测试界的学者新秀。

第二作者许文凯（Wenkai Xu）是加斯比计算人脑科学所的一名博士生。

第三作者佐尔坦·萨博（Zoltán Szabó）来自法国一所著名的理工大学“巴黎综合理工学院”（École Polytechnique）。萨博之前也曾在加斯比计算人脑科学所工作过，目前在巴黎综合理工学院任职研究副教授（类似于研究员），长期从事核方法、信息论（Information Theory）、统计机器学习等方面的研究。

第四作者福水健次（Kenji Fukumizu）是“统计数学学院”（The Institute of Statistical Mathematics）的教授，长期从事核方法的研究，可以说是这方面的专家。

最后一个作者阿瑟·格里顿（Arthur Gretton）是加斯比计算人脑科学所的机器学习教授，长期从事机器学习，特别是核方法的研究。他的论文有 9 千多次的引用数。

论文的主要贡献和核心方法

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

在一般的建模场景里，我们常常会对一组数据提出一个模型，来描述产生这些数据背后的过程。这个过程我们通常是看不见的，是一个隐含的过程。那么，当我们提出了模型之后，如何知道用这个模型描述现实就是准确的呢？这时候我们就需要用到一些**统计检验**（Statistical Testing）的方法。

一种比较普遍的方法，那就是假设我们的模型是 P，而数据的产生分布是 Q。说得直白一些，就需要去验证 P 是不是等于 Q，也就是需要验证两个分布是否相等。一个基本的做法就是，从 P 里“产生”（Generate）一组样本，或者叫一组数据，然后我们已经有了一组

从 Q 里产生的数据，于是用“**两个样本假设检验**”（Two Sample Tests）来看这两组数据背后的分布是否相等。

这个想法看似无懈可击，但是在实际操作中往往充满困难。**最大的操作难点就是从 P 中产生样本**。比如 P 是一个深度神经网络模型，那从中产生样本就不是一个简单且计算效率高的流程，这就为基于“两个样本假设检验”带来了难度。

另一方面，我们在做这样的统计检验的时候，最好能够针对每一个数据点，得到一个数值，来描述当前数据点和模型之间的关系，从而能够给我们带来更加直观的认识，看模型是否符合数据。

这里，有一种叫作“**最大均值差别**”（Maximum Mean Discrepancy），或者简称为 **MMD** 的检验方法能够达到这样的效果。MMD 的提出者就是这篇论文的最后一位作者阿瑟·格里顿，MMD 是在 NIPS 2016 提出的一个检验两个样本是否来自同一个分布的一种方法。当 MMD 值大的时候，就说明这两个样本更有可能来自不同的分布。

和一般的衡量两个分布距离的方法相比，MMD 的不同之处是把两个分布都通过核方法转换到了另外一个空间，也就是通常所说的“**再生核希尔伯特空间**”（Reproducing Kernel Hilbert Space），或者简称为 **RKHS**。在这个空间里，测量会变得更加容易。然而遗憾的是，MMD 依然需要得到两个分布的样本，也就是说我们依然需要从 P 里得到样本。

那么，**这篇文章的最大贡献，就是使用了一系列的技巧让 P 和 Q 的比较不依赖于从 P 中得到样本，从而让数据对于模型的验证，仅仅依赖于 P 的一个所谓的“打分函数”**（Score Function）。

其实在 MMD 里，这个打分函数就是存在的，那就是针对我们从 P 或者是 Q 里抽取出来的样本，我们先经过一个函数 F 的变换，然后再经过一个叫“核函数”T 的操作，最后两个样本转换的结果相减。

在这篇文章里，作者们提出了一个叫“核斯特恩差异”（Kernel Stein Discrepancy），或者叫**KSD 测试**的概念，本质上就是希望能够让这两个式子中关于 P 的项等于零。

什么意思呢？刚才我们说了 MMD 的一个问题是依然要依赖于 P，依赖于 P 的样本。假设我们能够让依赖 P 的样本这一项成为零，那么我们这个测试就不需要 P 的样本了，那也就是绕过了刚才所说的难点。

KSD 的本质就是让 MMD 的第二项在任何时候都成为零。注意，我们这里所说的是“任何时候”，也就是说，KSD 构造了一个特殊的 T，这个 T 叫作“斯特恩运算符”（Stein Operator），使得第二项关于 P 的样本的计算，在任何函数 F 的情况下都是零，这一点在文章中提供了详细证明。于是，整个 KSD 就不依赖于 P 的样本了。

这篇文章不仅阐述了 KSD 的思想，而且在 KSD 的思想上更进了一步，**试图把 KSD 的计算复杂度，也就是在平方级别的计算复杂度变为线性复杂度。**什么意思呢？也就是说，希望能够让 KSD 的计算复杂度随着数据点的增加而线性增加，从而能够应用到大数据上。这个内容我们就不在这里复述了。

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，文章还是在“受限波兹曼机”（Restricted Boltzmann Machine），简称 RBM 上做了实验。本质上就是在 RBM 的某一个链接上进行了简单的改变而整个模型都保持原样。

如果我们有从这两个 RBM 中得到的样本，其实是很难知道他们之间的区别的。在实验中，传统的 MMD 基本上没法看出这两个样本的差别。然而不管是 KSD，还是线性的 KSD 都能够得出正确的结论，而最终的线性 KSD 基本上是随着数据点的增多而性能增加，达到了线性的效果。

最后，作者们用了芝加哥犯罪记录来作为说明，使用“打分函数”来形象地找到哪些点不符合模型。应该说，理论性这么强的论文有如此直观的结果，实在难能可贵。

小结

今天我为你讲了 NIPS 2017 年的另外一篇最佳研究论文，文章的一个核心观点是希望能够通过构建一个特殊的运算符，使得传统的通过样本来检验两个分布的异同的方法，比如 MMD 方法，可以不依赖于目标分布的样本，并且还能达到线性计算速度。

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，我们简要地介绍了文章的实验结果 。

最后，给你留一个思考题，这种衡量分布之间距离的想法，除了在假设检验中使用以外，在机器学习的哪个环节也经常碰到？

欢迎你给我留言，和我一起讨论。

上一篇 051 | 精读2017年NIPS最佳研究论文之一：如何解决非凸优化问题?

下一篇 053 | 精读2017年NIPS最佳研究论文之三：如何解决非完美信息博弈问题?

精选留言 (2)

写留言

林彦 3
2018-01-31

神经网络里的常用损失函数，交叉熵依据的K-L散度是衡量2种概率分布之间的差异。但是不符合对称性，因此不能算一种距离的度量

林彦
2018-01-31

聚类里面也会衡量分布的距离来评估聚类的效果。不知道问题理解对不对。期望看到更多人的答案和得到老师的提示。谢谢