Michael I. Jordan《更好的分布式机器学习》笔记

马春杰

Jodan认为，统计学和计算学应该融合发展，并且会在明年发布他们的分布式机器学习框架Ray，集**统计推理、机器学习、大数据处理、计算**为一体，超越Spark。

2015年，艾伦研究所发布了[Semantic Scholar](https://www.semanticscholar.org/)搜索服务，意图超越其他在线学术搜素引擎，目前的目标是**衡量一位科学家或者一所研究机构对之后研究的影响**。

### Jordan 演讲核心：融合计算理论与统计理论

* 在基础性的层面，统计学和计算科学之间有着突出的分歧，在计算科学中，数据点越多，意味着越复杂，需要通过算法或者硬件训练才能达到更好。而对于统计学来说，数据点越多，意味着推理越准确（**因为样本点变多了**）。
* 在形式的层面，两者之间的概念不对称，相互之间缺乏融合。

Jodan教授认为，统计或者是机器学习需要更加深入的计算机科学系统和数据库接触，并不仅仅与具有人工智能的人。

目前的深度学习涉及了监督学习以及大量的数据，这是一个很有吸引力的组合，但是却没有**神经**的感觉，尤其是需要对数据进行标注。

而接近非监督式学习的一种方式就是将好的**特征**或者是**表示**的正式特征记下，然后与现实世界的其他假设联系在一起。Jodan认为，要把这些特征放进优化函数或者贝叶斯先验中优化整合。这个意思是不是**用已知推未知？**

### 演讲内容

* Jordan 从一个职位描述讲起，介绍大数据带来的挑战，引出演讲主题：“大数据”时代同时需要计算思维和推理思维。

**计算思维**：提取、建模、扩展性、鲁棒性等；

**推理思维**：思考数据背后的真实世界现象，考虑抽样模型，开发能从数据回馈到潜在现象的程序。

* 隐私与数据分析，隐私与推理：差分隐私
* 计算与推理。推理质量与经典的计算资源，比如时间和空间，如何平衡？很难！
* 计算与推理的机制和边界
* Jordan 所在研究机构与合作伙伴
* 大数据软件的增长情况
* 介绍了Ray

开发Ray的动力源于构建更好的分布式机器学习框架，目标是在单机上运行相同的代码和簇；对既有的代码进行最小化的修正，让其变得可分布；有效地支持大量的小任务；在任务间有效地分享数据。

应用：商品检测、欺诈检测、认知助手、物联网。

特点：低延迟、个性化和快速变化。

结论：随着数据的增多，计算与推理应该在**基础层面**有更多的**融合**。