

搜狐 > 科技 > 正文



机器之心

1528
文章

311万
阅读

[查看TA的文章>](#)

深度 | Kaggle创始人Quora问答：深度学习会淘汰其他的机器学习方法吗？

2016-09-30 11:39

选自Quora

机器之心编译

参与：Leonardo Luke、杜夏德

Kaggle 联合创始人兼 CTO Anthony Goldbloom 在 Quora 上公开回答问题，内容涉及了 数据科学、Kaggle 入门方法等。机器之心整理了他回答的所有问题。

1.未来 5 年，数据科学将会发生什么变化？

在这个问题上，我不打算谈太多数据前沿如何发展，而更多关注于数据科学逐渐成为主流变得无处不在。

在预测数据科学未来五年的发展时，回想其过去五年的进展大有裨益。2010 年 Kaggle 成立时，「数据科学」一词还不常见。我们社群的同僚们用其他的词汇来描述自己的工作，比如高级分析、统计学、机器学习、生物信息学、计量经济学或者其他和数据以及统计技术相关的专业之一。很多公司也用职能来称呼其负责数据相关工作的部门：市场分析、风险控制、包销、化学信息学等等。

2011 年 O`Reilly 的 Strata 大会让数据科学一词红了起来。会议聚集了一千五百名「数据科学家」，也让有着不同职称的人们有了根据他们的技能来称呼自己的方式。会议也让高级管理人员明白，不同部门的数据人员实际上有着同样的技能组合。

如果说 O`Reilly 的 Strata 大会是棒球里面的第一局（注：棒球共 9 局），我认为我们现在正在进入第二局。现在很多的公司将他们的数据科学家整合成单个数据科学部门。最有效率的公司架构就是数据科学部门将其数据科学家分配给业务部门（市场、风险等等）。这种架构十分有效，因为数据科学部门会学习如何吸引和招募数据科学团队，同时也允许数据科学家和具体问题语境中的人一起工作。Airbnb 就是有效利用这种架构的公司之一。

随着公司不断挖掘现存数据科学团队的价值，这些团队也会继续发展。最终，我认为中央的数据科学团队会消失，每个业务部门会有自己专门的大型数据科学团队。如果数据科学成为了公司中主要的决策工具，那么它就获得了成功——当需要作出决定时，管理层的第一反应是问「数据科学怎么说？」

从另外一个角度回答这个问题的话，我认为在下一个十年，数据科学领域的规模会比软件工程更大。如果我们把使用 R 或者 Python 数据工具的人定义为数据科学家，目前大约有 150 至 300 万数据科学家（根据 Kaggle 的用户量 65 万和 Jupyter 计划用户数约 300 万估算，目前全球软件工程师约有 2000 万）。同时，目前全球有约 800 万 SAS 用户和 1.2 亿

大家都在搜：苹果或



搜狐号推荐



IT之家
IT之家是业界知名网站。IT之家

搜狐科技社
搜狐科技官
件、大趋势



果壳网
面向都市科
的泛科技兴

大咖说科技
科技大咖说

最黑科技
带你认识全



24小时热文

1 据利
新机

2 没有
不是

3 三星
订单



Python。

)
——初学者如何在 **Kaggle** 上入门？

分享到

aggle 对于喜欢边学边做（而不是通过读书或者看讲座）的人来说是一个非常好的入门方式。

对于喜欢从非常具体问题开始的人，我建议从四个「入门」竞赛（compeition）开始。我们最简单的竞赛是根据性别、舱位等级来预测谁能在泰坦尼克号事故中幸存。

如果你的电脑上还没有 Python 或者 R 的运行环境，我们为你准备了 Kernels 工具，这是一个在线的脚本编辑器，可以让你在不安装 R 或者 Python 的情况下运行代码（并且已经连接好了数据）。

我建议从「fork」（程序员词汇，意为复制）别人的 Kernel 并进行编辑入手，而不是自己从头开始。如果你想从 Python 开始（我推荐），我建议用 Omar El Gabry 的 Kernel，这个 Kernel 是一个优秀的端到端的工作流程，从探索数据开始，到基本的机器学习模型结束。如果你喜欢 R，我推荐 Megan Risdal 的 Kernel。如果你还没准备好从 Python 或者 R 开始，我们准备了一个简单的 Excel 教程 (<https://www.kaggle.com/c/titanic/details/getting-started-with-excel>)。

如果你喜欢自由探索，或者就是不喜欢竞赛，我建议看一看我们的开源数据集。这些数据集和竞赛没有关系，但通过共享代码和论坛讨论，他们依然对学习有所帮助。你可以从美国婴儿名字开始，这是一个简单、有趣的数据集，记录了过去一百多年以来美国婴儿起名的趋势变化。再说一遍，我建议从 fork 别人的 Kernel 开始，相比从空白屏幕上的一个光标开始，这样学习不那么困难。

3.哪些工具可以让数据科学家更有效率？

我们认为目前数据科学工具差不多和 15 年前软件工程的工具一样。现在做数据科学工作比 5 至 10 年之后要痛苦得多。目前，数据科学工作流程的共享和协作非常痛苦（甚至让其他人的分析在你的机器上运行都需要一些技巧），将机器学习模型用于生产也是一项挑战。

实际上，我们正在开发一个叫做 Kaggle Kernels 的环境，致力于让数据科学工作流的共享、协作更加便捷（最适合的类比是给数据科学家的 Github）。有人用了 Kaggle Kernels 进行分析之后，你就可以 fork 他的分析（复制代码，Docker 容器以及数据连接）。这样你就可以立刻运行他们的分析，并可以迭代。

目前，Kaggle Kernels 只对 Kaggle 竞赛和 Kaggle 上分享的开源数据集开放。明年年中，我们将会商业化的产品，数据科学团队可以使用其进行团队内协作并分享结果。

至于将模型投入生产，是大型的云服务提供商正在专注的领域（比如微软的 AzureML、亚马逊的 Amazon Machine Learning）。对于他们现存的计算和数据存储业务来说，这是顺其自然的扩展。目前还没有大型云服务商成功，但我预测在未来几个月至几年的时间我们会看到他们的进展。

目前，数据科学家需要将专用于工作流上不同细分领域的工具拼凑起来。Git 是控制代码版



一年
仁修



热门图集



普京自费请官员吃冰淇淋



顽童李宗盛：60岁了 还得年少时的梦吗



联系我们

的工具具有 Jupyter Notebooks 和 Shiny。Kaggle 之前使用 Make 米作为编制工具。我也听

一些公司使用 PMML 部署简单的模型，取得了成功。

分享到深度学习会淘汰其他的机器学习方法吗？

我不这么认为。我认为在一些场合中深度神经网络是最佳选择，而在其他的地方则需要其他的技术。

目前，我们发现深度神经网络在一些无结构的数据竞赛中能够获胜（比如图像、视频、脑电图）。对图像和视频来说，卷积神经网络较好；对于脑电图数据和其他包含序列的数据来说，循环神经网络较好。对于结构化的数据问题，我们发现巧妙的特征工程结合梯度提升机器（特别是整合 XGBoost）获胜最多。

对于自然语言处理问题，情况会复杂一些：有些时候需要循环神经网络，有些时候需要结合 XGBoost 的信息检索方法。

深度神经网络的支持者说在数据集「足够大」时，深度神经网络就会超越其他方法。在 Kaggle 的竞赛中我们还没有发现能够佐证这一说法的证据。可能是因为我们的竞赛还没有「足够大」的数据。然而，即便足够大的数据会让深度神经网络显出优势，很多应用还是只有小、中规模的数据。

作为相关方，我们在大部分的竞赛结束之后都会在我们的博客上对获胜者进行采访，问他们采用了什么方法。我们把这看作是记录有监督的机器学习问题中的优秀方法的「活日志」。我们也把所有获胜者的采访作为数据集发布在了 Kaggle 上，你可以使用 Kaggle Kernel 来发现机器学习的发展趋势，看出在哪些场合深度学习占优，哪些场合深度学习处于劣势（比如这个简单的 Kernel 就说明深度学习在在 Kaggle 竞赛中呈现上涨趋势）。

5.Kaggle 如何盈利？

目前我们有两个主要的现金流来源于主持的竞赛和我们提供的工作招聘版块。

我们有三种比赛：

1. 特色比赛：公司可以利用这种比赛提升自己在数据科学社群的知名度，接触新的方法或者解决疑难杂症
2. 招聘比赛：公司利用这种比赛吸引并招揽之前通过其他方式招募不到的人才
3. 研究比赛：研究人员用把这种比赛作为与数据科学家协作的另一种方式。这对于非常具体且具有挑战性的问题大有裨益。

在未来，Kaggle 计划增加额外的服务。我们打算让 Kaggle Kernels 成为免费增值服务，付费后，公司可以将其用为数据科学团队的协作环境。在更远的未来，还有更多的有意义 d 附加服务（包括数据科学家咨询服务的市场）。

6.招聘者在职位申请上看到 Kaggle 竞赛会怎么想？

他们会对竞赛中表现出色的人进行面试。Google DeepMind 这类的公司会关注我们的竞赛，并主动接洽表现出色的人才。我们了解到有些公司要求在工作申请中放上 Kaggle 资料链接。

分享到

我们听说招聘方喜欢在简历中有 Kaggle 结果的数据科学家，因为这样显得应聘者对数据科学有一定的热情（即并不只是为了工资），也让招聘方了解到应聘者的能力如何。

我们也了解到招聘方希望看到应聘者在竞赛之外的其他信息。就这一点，我们最近上线了开源数据平台，并对 Kaggle 资料进行了改版，以突显个人在优秀 Kernel 以及我们论坛上的贡献（在竞赛表现之外）。

7.Kaggle 目前的重点是什么？

Kaggle 目前的重点是从数据科学家业余项目的站点，变为数据科学家常用工作站点。

为了完成这一点，我们正在着重发展 Kaggle Kernel，让其成为数据科学家共享和协作工作的地方。目前，数据科学家可以使用 Kaggle Kernel 来共享代码、竞赛结果和开源数据集。

到明年年中，我们计划把 Kernel 向小型团队开放，让他们在团队内部使用。小型团队将可以使用 Kaggle Kernel 来 fork 同事的工作，或者从来自社群公开分享的庞大资源中选择并 fork。

8.数据科学家最佳的协作方式是什么？

目前还没有好的解决方案。正如上面说到的一样，现在就连把一个人的分析放到另外一台机器上运行都是一个挑战（需要一样的数据，一样的语言版本，一样的库，有时候库的版本也需要一致）。

这也是我们在 Kaggle Kernel 上积极解决的问题。Kernel 整合了 Git（代码版本控制）、Docker（运行环境版本控制）以及数据连接。Kernel 可以用于竞赛协作和 Kaggle 上的开源数据集。目前 Kernel 还不对小团队开放。

目前，小团队的高效协作方式之一是自己把这些技术整合起来。

9.Kaggle 竞赛和数据科学家的工作有多相似？

Kaggle 竞赛涵盖了很多数据科学家的工作内容。缺少的两块内容是：

1. 把一个业务问题具体化为一个数据科学问题（包括提取数据，进行结构化，以解决业务问题）
2. 将模型投入产品

Kaggle 竞赛间接帮助训练了数据科学家对问题进行结构化。竞赛让我们的社区了解到了大量的、精巧的结构问题。所以即便社区并没有直接参与，他们也会看到我们如何对不同问题进行结构化，并且将这些结构应用到解决实际的业务问题中。

合

新闻

体育

汽车

房产

旅游

教育

时尚

科技

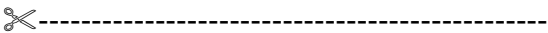
财经

娱乐

更多

) 前我们的社区很少有模型产品化的接触。

分享到 本文由机器之心编译，转载请联系本公众号获得授权。



加入机器之心（全职记者/实习生）：hr@almosthuman.cn

投稿或寻求报道：editor@almosthuman.cn

广告&商务合作：bd@almosthuman.cn

声明：本文由入驻搜狐号的作者撰写，除搜狐官方账号外，观点仅代表作者本人，不代表搜狐立场。

阅读 (58)

不感兴趣

投诉



¥328.00 ¥899



¥6000



¥118.00 ¥299

一体机

小家具

衣柜门

广告机

铁艺广告

我来说两句

0人参与，0条评论

来说两句吧.....

登录并发表

搜狐“我来说两句” 用户公约

还没有评论，快来抢沙发吧！

推荐阅读

百度网盘被指泄露用户隐私 官方：将打击第三方搜索网站



SOHU.com



SOHU.com




SOHU.com



SOHU.com


上游新闻 · 今天 10:58

3




SOHU.com

南京玄武警方推出“滴滴警务”平台，让警察“网约抢单”

 澎湃新闻 · 今天 14:07

12



SOHU.com

苹果公布新专利：用户能用指纹悄悄报警

合

新闻

体育

汽车

房产

旅游

教育

时尚


科技

财经


娱乐

更多


分享到




微信推“争议”功能 公众号不能继续任性了



 华尔街见闻 · 今天 13:01

 1



全新卡罗拉，驱动每个幸福的微笑，让幸福盛放！

广告 · 今天 17:14


你能活多久？让人工智能算一算！


中国经济网 · 今天 11:14




无人车如果出事故，微软百度要负责吗？

人民日报中央厨房 · 今天 14:17


 1



财报连续21季度营收下滑，云服务与AI数据低迷，IBM的“转型”能持续吗？



 36氪 · 今天 14:25



嘲笑鹿晗娘炮的小米，找了吴亦凡当代言人









PingWest品玩 · 今天 08:52

 13

还在守着死工资？成功的人已经加微信学炒股了



广告 · 今天 17:14

供应商讨债新技能，乐视大厦楼底“搭帐篷” | 图说













 36氪 · 今天 12:09


 1



搜狗地图发布智能副驾 主打语音交互


搜狗科技视界 · 今天 11:23

 1



富士康为首的苹果4大代工制造商起诉高通涉嫌垄断

PingWest品玩 · 今天 16:21



新闻 体育 汽车 房产 旅游 教育 时尚 科技 财经 娱乐 更多

SOHU.com

钛媒体 · 今天 16:21

分享到



卡罗拉双擎，驾享精致生活。

广告 · 今天 17:14

努比亚Z17mini，让你坐地铁快人一步

搜狐 SOHU.com

中国派 · 今天 16:30

搜狐 SOHU.com

搜狐 SOHU.com

搜狐 SOHU.com

孙宏斌：只要贾跃亭退出 好多人愿意接盘

PingWest品玩 · 今天 16:25

新零售时代的超市成了「四不像」，开始狂奔的盒马鲜生究竟想做些什么？

搜狐 SOHU.com

极客公园 · 今天 16:27

搜狐 SOHU.com

搜狐 SOHU.com

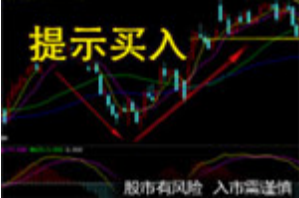
搜狐 SOHU.com

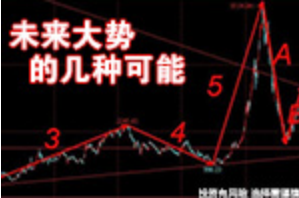
搜狐 SOHU.com

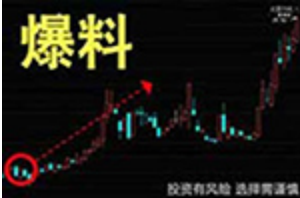
苹果支付在华推最大促销成功？薅完羊毛再用支付宝

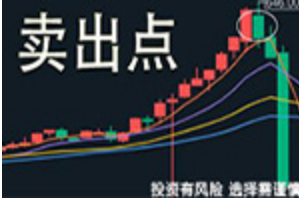
易北辰 · 今天 16:05

通知：A股深跌之后有机会，抢反弹跟进这些股！









广告 · 今天 17:14

亮三点03期 | ofo小黄车：人多，钱傻，速来



刘兴亮 · 今天 15:23

搜狐 SOHU.com

魅族Pro7配Flow耳机售价曝光：6999元

 IT之家 · 今天 14:30

搜狐 SOHU.com

微信的胜利？“苹果税”或取消不再强推应用内购买

搜狐科技视界 · 今天 11:12

7

两年后Google Glass又回来了，这次要主攻企业市场



36氪 · 今天 15:36

合

新闻

体育

汽车

房产

旅游

教育

时尚

科技

财经

娱乐

更多



广告 · 今天 17:14

分享到

加载更多