Github上的十大机器学习项目

3

机器学习 (http://www.csdn.net/tag/机器学习/news)

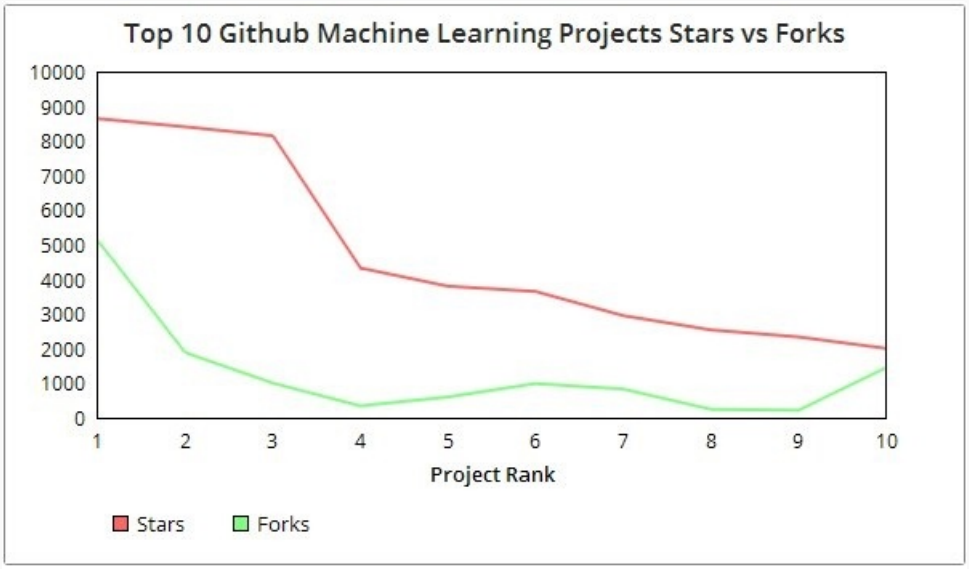
原文作者：Matthew Mayo (https://twitter.com/mattmayo13)
译文地址：Top 10 Machine Learning Projects on Github (http://www.kdnuggets.com/2015/12/top-10-machine-learning-github.html)
文章译者：赵屹华 (http://www.cnblogs.com/naive/)，搜狗计算广告工程师，前生物医学工程师，关注推荐算法、机器学习领域。
文章审校：刘帝伟

Github上的十大机器学习项目涵盖了一系列函数库、框架和教学资源。我们来看看别人使用的工具和学习的资源。

开源软件是数据科学很重要的一部分。

根据最近的KDnuggets数据科学软件投票的结果 (http://www.kdnuggets.com/2015/05/poll-r-rapidminer-python-big-data-spark.html)，73%的数据科学家在过去12个月里使用过免费软件。互联网上有着各式各样的这类工具，而Github事实上则成为了所有开源软件的交流平台，包括数据科学社区里所用的工具。机器学习在数据科学界的重要性和中心地位已经不言而喻。

下图是Github十大机器学习项目 (https://github.com/search?o=desc&q=Machine+Learning&s=stars&type=Repositories&utf8=✓)的概览。



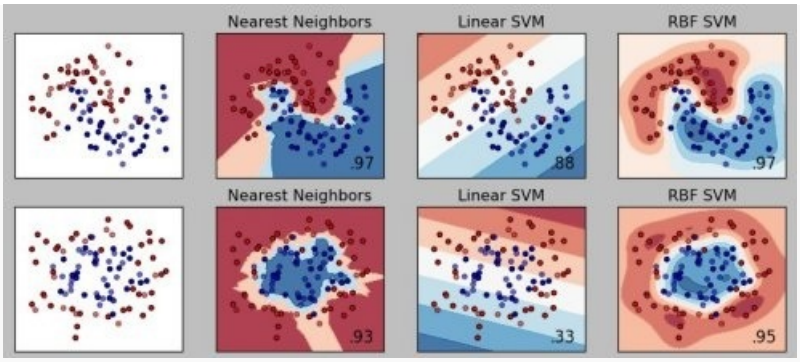
1. Scikit-learn (https://github.com/scikit-learn/scikit-learn)

Python的机器学习库

★8641，

5125

十大之首，毫无悬念地就是工业界和学术界Python开发者首选的机器学习函数库。Scikit利用了Python的科学计算工具，它基于Numpy，Scipy和matplotlib。Scikit-learn拥有一般工具包的常规功能，包括分类、回归和聚类算法，也包括数据预处理和模型评价模块。



2. Awesome Machine Learning (https://github.com/josephmisiti/awesome-machine-learning)

> 一系列绝妙的机器学习框架、函数库和软件。

★ 8404，

1885

这是一系列绝妙的机器学习框架、函数库和软件。这个列表先按照语言来分类，然后按照机器学习的类别（通用型，计算机视觉，自然语言处理，等等）。它还包括数据可视化工具，从某种意义上来说它比数据科学的通用列表更丰富，这是一件好事。

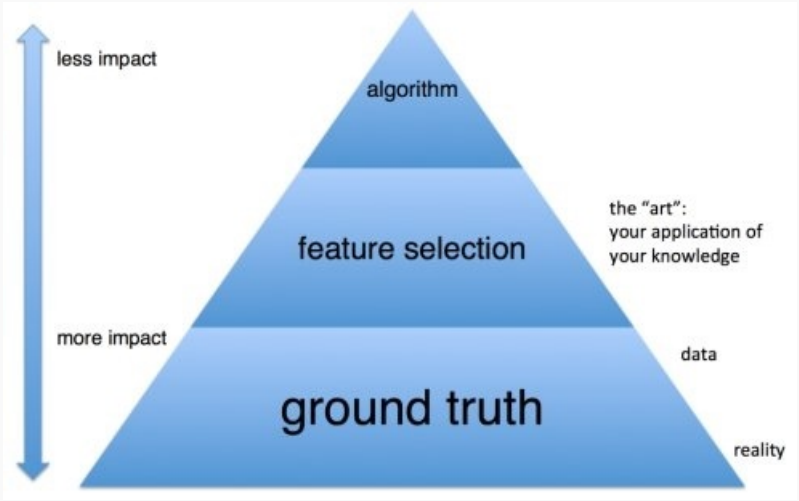
3. PredictionIO (https://github.com/PredictionIO/PredictionIO)

> PredictionIO是开发者和ML工程师的一个机器学习服务器。它基于Apache Spark、HBase和Spray。

★ 8145，

1002

PredictionIO （https://github.com/PredictionIO/PredictionIO)是一个通用型框架。它包括一些处理常规问题的模板引擎，比如分类和推荐，也可以用户自定义修改，通过REST APIs或者SDKs与现有的应用连接。由于它是建立在Spark基础上并且利用了Spark的生态系统，因此PredictionIO主要用Scala开发也就不足为奇了。



4. Dive into Machine Learning (https://github.com/hangtwenty/dive-into-machine-learning)

> 使用Python Jupyter和Scikit-learn深入研究机器学习。

★ 4326，

342

这是scikit-learn的一个教程集合，有大量IPython　notebook脚本，和许多Python相关的或者通用型的机器学习话题的链接，以及更多的数据科学信息。作者并不很贪婪，如果某一个教程不足以吸引你，它们会很快发掘更多的其它类似内容。代码库里并没有软件，但如果你不熟悉Python环境下的机器学习，则值得读一下。

5. Pattern (https://github.com/clips/pattern)

> Python的网页挖掘模块，包括爬虫、自然语言处理、机器学习、网络分析和可视化等工具。

★ 3799，

598

Pattern (http://www.clips.ua.ac.be/pages/pattern)是基于Python的网页挖掘工具包，由Antwerp大学的计算语言学和心理语言学研究中心（Computational Linguistics & Psycholinguistics, CLiPS (http://www.clips.ua.ac.be/pages/pattern)）开发完成。它可以用来完成爬虫、机器学习、自然语言处理、网络分析和可视化等任务。Pattern还可以从一些熟知的网络服务中挖掘数据。这个项目有完整的文档，并且包含了大量的例子和单元测试。

6. NuPIC (Numenta Platform for Intelligent Computing) (https://github.com/numenta/nupic)

> 一个大脑启发式的机器智能平台和基于大脑皮层学习算法的生物神经网络。

★ 3647，

987

NuPIC (http://numenta.org/)实现了Hierarchical Temporal Memory (HTM) (http://numenta.com/learn/hierarchical-temporal-memory-white-paper.html)机器学习算法。HTM算法试图以大脑皮层的计算方式来建模，专注于存储和召回空间和时间的模式。NuPIC非常适合模式相关的异常检测。

7. Vowpal Wabbit (https://github.com/JohnLangford/vowpal_wabbit)

> Vowpal Wabbit是一种机器学习系统，它在online, hashing, allreduce, reductions, learning2search, active 和 interactive learning等技术上一直处于领先优势。

★ 2949，

827

Vowpal　Wabbit　(http://hunch.net/~vw/)的目标是对大数据快速建模，并支持并行学习。这个项目由雅虎发起，现在由微软研究院开发。Vowpal　Wabbit采用了外部存储学习算法（out-of-core　(https://en.wikipedia.org/wiki/Out-of-core_algorithm)），它已经实现了用1000个计算节点在一小时内训练TB级的特征数据集 (http://arxiv.org/abs/1110.4198)。

8. aerosolve (https://github.com/airbnb/aerosolve)

> 一个交互友好的机器学习工具包

★ 2538，

245

aerosol与其它的函数包不同，它主要是提供交互友好的调试工具，训练模型的Scala代码，一个用于图像排序的图像内容分析引擎，和一种特征转换语言，用户可以灵活地控制特征。aerosolve采用基于thrift的特征表征，特征按照逻辑被分组后进行变换，或者一次性对所有特征组完成变换。

9. GoLearn (https://github.com/sjwhitworth/golearn)

> 一种Go语言的机器学习工具。

★ 2334，

215

GoLearn是使用Go语言开发的机器学习库，开发的活跃度很高。它的目的是为Go语言开发者提供一套完善的、易用的、可自定义的工具包。GoLearn实现了Scikit-learn中常用的fit/predict接口，简化了预测器的生成方法，并实现了交叉验证、训练集/测试集切分等常用函数。

10. Code for Machine Learning for Hackers (https://github.com/johnmyleswhite/ML_for_Hackers)

> Machine Learning for Hackers一书中的代码

★ 2003，

1446

这个代码库中的代码都来自O’Reilly出版的Machine Learning for Hackers (http://shop.oreilly.com/product/0636920018483.do)一书。代码用R语言实现，其依赖了大量R工具包，它的内容包括常见的分类任务、排序和回归，以及主成分分析和多维标度法等统计方法。

注：上榜依据是在Github上搜索“Machine Learning”关键词所返回的结果，按照星星的数量排序，数据搜集时间是2015年12月10日下午1点。

相关阅读：

- Top 20 Python Machine Learning Open Source Projects (http://www.kdnuggets.com/2015/06/top-20-python-machine-learning-open-source-projects.html)
- Topological Data Analysis – Open Source Implementations (http://www.kdnuggets.com/2015/11/topological-data-analysis-open-source-implementations.html)
- 7 Steps to Mastering Machine Learning With Python (http://www.kdnuggets.com/2015/11/seven-steps-machine-learning-python.html)

（责编/周建丁）

(http://geek.csdn.net/user/publishlist/zhong930)
仲浩 (http://geek.csdn.net/user/publishlist/zhong930)
发布于 人工智能 (http://geek.csdn.net/forum/43) 19小时前

评论

已有0条评论 最新

还没有评论，赶快来抢沙发吧。