Python语言下的机器学习库

2015/03/14 · 工具与框架 · Python, 机器学习

Python是最好的编程语言之一，在科学计算中用途广泛：计算机视觉、人工智能、数学、天文等。它同样适用于机器学习也是意料之中的事。

当然，它也有些缺点；其中一个是工具和库过于分散。如果你是拥有unix思维（unix-minded）的人，你会觉得每个工具只做一件事并且把它做好是非常方便的。但是你也需要知道不同库和工具的优缺点，这样在构建系统时才能做出合理的决策。工具本身不能改善系统或产品，但是使用正确的

这篇文章的目的就是列举并描述Python可用的最有用的机器学习工具和库。这个列表中，我们不要求这些库是用Python写的，只要有Python接口就够了。我们在最后也有一小节关于深度学习（Deep Learning）的内容，因为它最近也吸引了相当多的关注。

我们的目的不是列出Python中**所有**机器学习库（搜索“机器学习”时Python包索引(PyPI)返回了139个结果），而是列出我们所知的有用并且维护良好的那些。另外，尽管有些模块可以用于多种机器学习任务，我们只列出主要焦点在机器学习的库。比如，虽然Scipy包含一些聚类算法，但是它的主焦点不是机器学习而是全面的科学计算工具集。因此我们排除了Scipy（尽管我们也使用它！）。

另一个需要提到的是，我们同样会根据与其他科学计算库的集成效果来评估这些库，因为机器学习（有监督的或者无监督的）也是数据处理系统的一部分。如果你使用的库与数据处理系统其他的库不相配，你就要花大量时间创建不同库之间的中间层。在工具集中有个很棒的库很重要，但这个库能与其他库良好集成也同样重要。

如果你擅长其他语言，但也想使用Python包，我们也简单地描述如何与Python进行集成来使用这篇文章列出的库。

Scikit-Learn

Scikit Learn是我们在CB Insights选用的机器学习工具。我们用它进行分类、特征选择、特征提取和聚集。我们最爱的一点是它拥有易用的一致性API，并提供了**很多**开箱可用的求值、诊断和交叉验证方法（是不是听起来很熟悉？Python也提供了“电池已备(译注：指开箱可用)”的方法）。锦上添花的是它底层使用Scipy数据结构，与Python中其余使用Scipy、Numpy、Pandas和Matplotlib进行科学计算的部分适应地很好。因此，如果你想可视化分类器的性能（比如，使用精确率与反馈率(precision-recall)图表，或者接收者操作特征(Receiver Operating Characteristics，ROC)曲线），Matplotlib可以帮助进行快速可视化。考虑到花在清理和构造数据的时间，使用这个库会非常方便，因为它可以紧密集成到其他科学计算包上。

另外，它还包含有限的自然语言处理特征提取能力，以及词袋（bag of words）、tfidf（Term Frequency Inverse Document Frequency算法）、预处理（停用词/stop-words，自定义预处理，分析器）。此外，如果你想快速对小数据集（toy dataset）进行不同基准测试的话，它自带的数据集模块提供了常见和有用的数据集。你还可以根据这些数据集创建自己的小数据集，这样在将模型应用到真实世界中之前，你可以按照自己的目的来检验模型是否符合期望。对参数最优化和参数调整，它也提供了网格搜索和随机搜索。如果没有强大的社区支持，或者维护得不好，这些特性都不可能实现。我们期盼它的第一个稳定发布版。

Statsmodels

Statsmodels是另一个聚焦在统计模型上的强大的库，主要用于预测性和探索性分析。如果你想拟合线性模型、进行统计分析，或者预测性建模，那么Statsmodels非常适合。它提供的统计测试相当全面，覆盖了大部分情况的验证任务。如果你是R或者S的用户，它也提供了某些统计模型的R语法。它的模型同时也接受Numpy数组和Pandas数据帧，让中间数据结构成为过去！

PyMC是做**贝叶斯曲线**的工具。它包含贝叶斯模型、统计分布和模型收敛的诊断工具，也包含一些层次模型。如果想进行贝叶斯分析，你应该看看。

Shogun

Shogun是个聚焦在支持向量机（Support Vector Machines, SVM）上的机器学习工具箱，用C++编写。它正处于积极开发和维护中，提供了Python接口，也是文档化最好的接口。但是，相对于Scikit-learn，我们发现它的API比较难用。而且，也没提供很多开箱可用的诊断和求值算法。但是，速度是个很大的优势。

Gensim

Gensim被定义为“人们的主题建模工具（topic modeling for humans）”。它的主页上描述，其焦点是狄利克雷划分（Latent Dirichlet Allocation，LDA）及变体。不同于其他包，它支持自然语言处理，能将NLP和其他机器学习算法更容易组合在一起。如果你的领域在NLP，并想进行聚集和基本的分类，你可以看看。目前，它们引入了Google的基于递归神经网络（Recurrent Neural Network）的文本表示法word2vec。这个库只使用Python编写。

Orange

Orange是这篇文章列举的所有库中唯一带有图形用户界面（Graphical User Interface，GUI）的。对分类、聚集和特征选择方法而言，它是相当全面的，还有些交叉验证的方法。在某些方面比Scikit-learn还要好（分类方法、一些预处理能力），但与其他科学计算系统（Numpy, Scipy, Matplotlib, Pandas）的适配上比不上Scikit-learn。

但是，包含GUI是个很重要的优势。你可以可视化交叉验证的结果、模型和特征选择方法（某些功能需要安装Graphviz）。对大多数算法，Orange都有自己的数据结构，所以你需要将数据包装成Orange兼容的数据结构，这使得其学习曲线更陡。

PyMVPA

PyMVPA是另一个统计学习库，API上与Scikit-learn很像。包含交叉验证和诊断工具，但是没有Scikit-learn全面。

深度学习

Theano

Theano是最成熟的深度学习库。它提供了不错的数据结构（张量，tensor）来表示神经网络的层，对线性代数来说很高效，与Numpy的数组类似。需要注意的是，它的API可能不是很直观，用户的学习曲线会很高。有很多基于Theano的库都在利用其数据结构。它同时支持开箱可用的GPU编程。

PyLearn2

还有另外一个基于Theano的库，PyLearn2，它给Theano引入了模块化和可配置性，你可以通过不同的配置文件来创建神经网络，这样尝试不同的参数会更容易。可以说，如果分离神经网络的参数和属性到配置文件，它的模块化能力更强大。

Decaf

Decaf是最近由UC Berkeley发布的深度学习库，在Imagenet分类挑战中测试发现，其神经网络实现是很先进的（state of art）。

Nolearn

如果你想在深度学习中也能使用优秀的Scikit-learn库API，封装了Decaf的Nolearn会让你能够更轻松地使用它。它是对Decaf的包装，与Scikit-learn兼容（大部分），使得Decaf更不可思议。

OverFeat

OverFeat是最近猫vs.狗（kaggle挑战）的胜利者，它使用C++编写，也包含一个Python包装器（还有Matlab和Lua）。通过Torch库使用GPU，所以速度很快。也赢得了ImageNet分类的检测和本地化挑战。如果你的领域是计算机视觉，你可能需要看看。

Hebel

Hebel是另一个带有GPU支持的神经网络库，开箱可用。你可以通过YAML文件（与Pylearn2类似）决定神经网络的属性，提供了将神级网络和代码友好分离的方式，可以快速地运行模型。由于开发不久，就深度和广度上说，文档很匮乏。就神经网络模型来说，也是有局限的，因为只支持一种神经网络模型（正向反馈，feed-forward）。但是，它是用纯Python编写，将会是很友好的库，因为包含很多实用函数，比如调度器和监视器，其他库中我们并没有发现这些功能。

NeuroLab是另一个API友好（与Matlabapi类似）的神经网络库。与其他库不同，它包含递归神经网络（Recurrent Neural Network，RNN）实现的不同变体。如果你想使用RNN，这个库是同类API中最好的选择之一。

与其他语言集成

你不了解Python但是很擅长其他语言？不要绝望！Python（还有其他）的一个强项就是它是一个完美的胶水语言，你可以使用自己常用的编程语言，通过Python来访问这些库。以下适合各种编程语言的包可以用于将其他语言与Python组合到一起：

- R -> RPython
- Matlab -> matpython
- Java -> Jython
- Lua -> Lunatic Python
- Julia -> PyCall.jl

不活跃的库

这些库超过一年没有发布任何更新，我们列出是因为你有可能会有用，但是这些库不太可能会进行BUG修复，特别是未来进行增强。

- MDP
- MlPy
- FFnet
- PyBrain

如果我们遗漏了你最爱的Python机器学习包，通过评论让我们知道。我们很乐意将其添加到文章中。

关于作者：douxingxiang

简介还没来得及写：）

个人主页 · 我的文章 · 13

相关文章

- 遗传算法中适值函数的标定与大变异算法
- 为什么你应该学 Python ？
- 遗传算法中几种不同选择算子及Python实现
- 用 Scikit-Learn 和 Pandas 学习线性回归
- 构建多层感知器神经网络对数字图片进行文本识别

可能感兴趣的话题

- 失业码农的再就业——第 0 篇 · 8
- PHP底层原理
- 如何在linux下SSH远程登录另一台LINUX，并且用sftp自动上传文件夹？ · 8
- vue中的参数传递除了父子组件之间传递还有那些方式方法
- 一个关于MVC URL 的相关问题。 · 1
- 现在在学JavaEE Web阶段的知识，觉得自己对Servlet，Filter和Listen... · 5

登录后评论 新用户注册

直接登录

Python小组话题 我有新话题

关于生成器函数递归
加瓦 发起 • 3 回复

做自己喜欢的事情成本有多高？
北冥有沙丁鱼 发起 • 32 回复

Numpy模块安装问题
桃李不言 发起 • 6 回复

有没有非互联网行业的小伙伴自学编程...
叫我小K咯 发起 • 256 回复

Python自学，基础已经学完，现在学...
alexhan 发起 • 36 回复

flask中如何触发request请求的?
加瓦 发起

- 本周热门Python文章
- 本月热门

Python工具资源 **更多资源 »**

Tryton：一个通用商务框架
杂项

NLTK：一个先进的用来处理自然语言数据的Python程序。
自然语言处理 · 2

PyMC：马尔科夫链蒙特卡洛采样工具
科学计算与分析

statsmodels：统计建模和计量经济学
科学计算与分析

[Pylearn2：一个基于Theano的机器学习库](#)
[机器学习](#) · [1](#)

关于 Python 频道

Python频道分享 Python 开发技术、相关的行业动态。

快速链接
网站使用指南 »
加入我们 »
问题反馈与求助 »
网站积分规则 »
网站声望规则 »

关注我们

新浪微博：@Python开发者
RSS：订阅地址
推荐微信号

QQ： 2302462408（加好友请注明来意）

更多频道

小组 – 好的话题、有启发的回复、值得信赖的圈子
头条 – 分享和发现有价值的内容与观点
相亲 – 为IT单身男女服务的征婚传播平台
资源 – 优秀的工具资源导航
翻译 – 翻译传播优秀的外文文章
文章 – 国内外的精选文章
设计 – UI,网页，交互和用户体验
iOS – 专注iOS技术分享
安卓 – 专注Android技术分享
前端 – JavaScript, HTML5, CSS
Java – 专注Java技术分享
Python – 专注Python技术分享

 文章 小组 相亲 加入我们 反馈 沪ICP备14046347号-1