python 数据挖掘领域工具包

原文：http://qxde01.blog.163.com/blog/static/67335744201368101922991/

Python 在科学计算领域，有两个重要的扩展模块：Numpy 和 Scipy。其中 Numpy 是一个用 python 实现的科学计算包。包括：

- 一个强大的 N 维数组对象 Array；
- 比较成熟的（广播）函数库；
- 用于整合 C/C++和 Fortran 代码的工具包；
- 实用的线性代数、傅里叶变换和随机数生成函数。

SciPy 是一个开源的 Python 算法库和数学工具包，SciPy 包含的模块有最优化、线性代数、积分、插值、特殊函数、快速傅里叶变换、信号处理和图像处理、常微分方程求解和其他科学与工程中常用的计算。其功能与软件 MATLAB、Scilab 和 GNU Octave 类似。

Numpy 和 Scipy 常常结合着使用，Python 大多数机器学习库都依赖于这两个模块，绘图和可视化依赖于 matplotlib 模块，matplotlib 的风格与 matlab 类似。Python 机器学习库非常多，而且大多数开源，主要有：

1. scikit-learn

scikit-learn 是一个基于 SciPy 和 Numpy 的开源机器学习模块，包括分类、回归、聚类系列算法，主要算法有 SVM、逻辑回归、朴素贝叶斯、Kmeans、DBSCAN 等，目前由 INRI 资助，偶尔 Google 也资助一点。

项目主页：

https://pypi.python.org/pypi/scikit-learn/

http://scikit-learn.org/

https://github.com/scikit-learn/scikit-learn

SciKits Index

scikit-aero	Aeronautical engineering calculations in Python.
scikit-bio	Data structures, algorithms and educational resources for bioinformatics.
scikit-commpy	Digital Communication Algorithms with Python
scikit-fmm	An extension module implimenting the fast marching method
scikit-image	Image processing routines for SciPy

scikit-learn	A set of python modules for machine learning and data mining
scikit-monaco	Python modules for Monte Carlo integration
scikit-nano	Python toolkit for generating and analyzing nanostructure data
scikit-rf	Open Source RF Engineering
scikit-tensor	Python module for multilinear algebra and tensor factorizations
scikit-tracker	Object detection and tracking for cell biology
scikit-vi	Scikit providing Virtual Instruments
scikit-video	Video processing routines for SciPy
scikits-image	Image processing routines for SciPy
ann	Approximate Nearest Neighbor library wrapper for Numpy
audiolab	A python module to make noise from numpy arrays
bootstrap	Bootstrap confidence interval estimation routines for SciPy
bvp1lg	Boundary value problem (legacy) solvers for ODEs
bvp_solver	Python package for solving two-point boundary value problems
cuda	Python interface to GPU-powered libraries
datasmooth	Scikits data smoothing package
eartho	Earth Observation routines for SciPy
example	Scikits example package
fitting	Framework for fitting functions to data with SciPy
hydroclimpy	Environmental time series manipulation
learn	A set of python modules for machine learning and data mining
odes	A python module for ordinary differential equation anddifferential algebraic equation solvers
optimization	A python module for numerical optimization
samplerate	A python module for high quality audio resampling
scattpy	Light Scattering methods for Python
sparse	Scikits sparse matrix package
statsmodels	Statistical computations and models for use with SciPy
talkbox	Talkbox, a set of python modules for speech/signal processing
timeseries	Time series manipulation
vectorplot	Vector fields plotting algorithms.

2. NLTK

NLTK(Natural Language Toolkit)是 Python 的自然语言处理模块，包括一系列的字符处理和语言统计模型。NLTK 常用于学术研究和教学，应用的领域有语言学、认知科学、人工智能、信息检索、机器学习等。 NLTK 提供超过 50 个语料库和词典资源，文本处理库包括分类、分词、词干提取、解析、语义推理。可稳定运行在 Windows, Mac OS X 和 Linux 平台上.

项目主页：

http://sourceforge.net/projects/nltk/

https://pypi.python.org/pypi/nltk/

http://nltk.org/

3. Mlpy

Mlpy 是基于 NumPy/SciPy 的 Python 机器学习模块，它是 Cython 的扩展应用。包含的机器学习算法有：

l 回归

least squares, ridge regression, least angle regression, elastic net, kernel ridge regression, support vector machines (SVM), partial least squares (PLS)

l 分类

linear discriminant analysis (LDA), Basic perceptron, Elastic Net, logistic regression, (Kernel) Support Vector Machines (SVM), Diagonal Linear Discriminant Analysis (DLDA), Golub Classifier, Parzen-based, (kernel) Fisher Discriminant Classifier, k-nearest neighbor, Iterative RELIEF, Classification Tree, Maximum Likelihood Classifier

l 聚类

hierarchical clustering, Memory-saving Hierarchical Clustering, k-means

l 维度约减

(Kernel) Fisher discriminant analysis (FDA), Spectral Regression Discriminant Analysis (SRDA), (kernel) Principal component analysis (PCA)

项目主页：

http://sourceforge.net/projects/mlpy

https://mlpy.fbk.eu/

4. **Shogun**

Shogun 是一个开源的大规模机器学习工具箱。目前 Shogun 的机器学习功能分为几个部分：feature 表示，feature 预处理，核函数表示,核函数标准化，距离表示，分类器表示，聚类方法，分布，性能评价方法，回归方法，结构化输出学习器。

SHOGUN 的核心由 C++实现，提供 Matlab、 R、 Octave、 Python 接口。主要应用在 linux 平台上。

项目主页：

http://www.shogun-toolbox.org/

5. MDP

The Modular toolkit for Data Processing (MDP)，用于数据处理的模块化工具包，一个 Python 数据处理框架。

从用户的观点，MDP 是能够被整合到数据处理序列和更复杂的前馈网络结构的一批监督学习和非监督学习算法和其他数据处理单元。计算依照速度和内存需求而高效的执行。从科学开发者的观点，MDP 是一个模块框架，它能够被容易地扩展。新算法的实现是容易且直观的。新实现的单元然后被自动地与程序库的其余部件进行整合。MDP 在神经科学的理论研究背景下被编写，但是它已经被设计为在使用可训练数据处理算法的任何情况中都是有用的。其站在用户一边的简单性，各种不同的随时可用的算法，及应用单元的可重用性，使得它也是一个有用的教学工具。

项目主页：

http://mdp-toolkit.sourceforge.net/

https://pypi.python.org/pypi/MDP/

6. PyBrain

PyBrain(Python-Based Reinforcement Learning, Artificial Intelligence and Neural Network)是 Python 的一个机器学习模块，它的目标是为机器学习任务提供灵活、易应、强大的机器学习算法。（这名字很霸气）

PyBrain 正如其名，包括神经网络、强化学习(及二者结合)、无监督学习、进化算法。因为目前的许多问题需要处理连续态和行为空间，必须使用函数逼近(如神经网络)以应对高维数据。PyBrain 以神经网络为核心，所有的训练方法都以神经网络为一个实例。

项目主页：

http://www.pybrain.org/

https://github.com/pybrain/pybrain/

7. **BigML**

BigML 使得机器学习为数据驱动决策和预测变得容易，BigML 使用容易理解的交互式操作创建优雅的预测模型。BigML 使用 BigML.io,捆绑 Python。

项目主页：

https://bigml.com/

https://pypi.python.org/pypi/bigml

http://bigml.readthedocs.org/

8. **PyML**

PyML 是一个 Python 机器学习工具包， 为各分类和回归方法提供灵活的架构。它主要提供特征选择、模型选择、组合分类器、分类评估等功能。

项目主页：

http://cmgm.stanford.edu/~asab/pyml/tutorial/

http://pyml.sourceforge.net/

9. **Milk(目前只有 2.6,2.7 版本)**

Milk 是 Python 的一个机器学习工具箱,其重点是提供监督分类法与几种有效的分类分析：SVMs(基于 libsvm)，K-NN，随机森林经济和决策树。它还可以进行特征选择。这些分类可以在许多方面相结合，形成不同的分类系统。

对于无监督学习，它提供 K-means 和 affinity propagation 聚类算法。

项目主页：

https://pypi.python.org/pypi/milk/

http://luispedro.org/software/milk

10. **PyMVPA**

PyMVPA(Multivariate Pattern Analysis in Python)是为大数据集提供统计学习分析的 Python 工具包，它提供了一个灵活可扩展的框架。它提供的功能有分类、回归、特征选择、数据导入导出、可视化等

项目主页：

http://www.pymvpa.org/

https://github.com/PyMVPA/PyMVPA

11. **Pattern**

Pattern 是 Python 的 web 挖掘模块，它绑定了 Google、Twitter 、Wikipedia API，提供网络爬虫、HTML 解析功能，文本分析包括浅层规则解析、WordNet 接口、句法与语义分析、TF-IDF、LSA 等，还提供聚类、分类和图网络可视化的功能。

项目主页：

http://www.clips.ua.ac.be/pages/pattern

https://pypi.python.org/pypi/Pattern

12. **pyrallel**

Pyrallel(Parallel Data Analytics in Python)基于分布式计算模式的机器学习和半交互式的试验项目，可在小型集群上运行，适用范围：

l focus on small to medium dataset that fits in memory on a small (10+ nodes) to medium cluster (100+ nodes).

l focus on small to medium data (with data locality when possible).

l focus on CPU bound tasks (e.g. training Random Forests) while trying to limit disk / network access to a minimum.

l do not focus on HA / Fault Tolerance (yet).

l do not try to invent new set of high level programming abstractions (yet): use a low level programming model (IPython.parallel) to finely control the cluster elements and messages transfered and help identify what are the practical underlying constraints in distributed machine learning setting.

项目主页：

https://pypi.python.org/pypi/pyrallel

http://github.com/pydata/pyrallel

13. **Monte**

Monte (machine learning in pure Python)是一个纯 Python 机器学习库。它可以迅速构建神经网络、条件随机场、逻辑回归等模型，使用 inline-C 优化，极易使用和扩展。

项目主页：

https://pypi.python.org/pypi/Monte

http://montepython.sourceforge.net

14. **Orange**

Orange 是一个基于组件的数据挖掘和机器学习软件套装，它的功能即友好，又很强大，快速而又多功能的可视化编程前端，以便浏览数据分析和可视化，基绑定了 Python 以进行脚本开发。它包含了完整的一系列的组件以进行数据预处理，并提供了数据帐目，过渡，建模，模式评估和勘探的功能。其由 C++ 和 Python 开发，它的图形库是由跨平台的 Qt 框架开发。

项目主页：

https://pypi.python.org/pypi/Orange/

http://orange.biolab.si/

15. **Theano**

Theano 是一个 Python 库，用来定义、优化和模拟数学表达式计算，用于高效的解决多维数组的计算问题。Theano 的特点：

l 紧密集成 Numpy

l 高效的数据密集型 GPU 计算

l 高效的符号微分运算

l 高速和稳定的优化

l 动态生成 c 代码

l 广泛的单元测试和自我验证

自 2007 年以来，Theano 已被广泛应用于科学运算。theano 使得构建深度学习模型更加容易，可以快速实现下列模型：

l Logistic Regression

l Multilayer perceptron

l Deep Convolutional Network

l Auto Encoders, Denoising Autoencoders

l Stacked Denoising Auto-Encoders

l Restricted Boltzmann Machines

l Deep Belief Networks

l HMC Sampling

l Contractive auto-encoders

Theano，一位希腊美女，Croton 最有权势的 Milo 的女儿，后来成为了毕达哥拉斯的老婆。

项目主页：

http://deeplearning.net/tutorial/

https://pypi.python.org/pypi/Theano

16. **Pylearn2**

Pylearn2 建立在 theano 上，部分依赖 scikit-learn 上，目前 Pylearn2 正处于开发中，将可以处理向量、图像、视频等数据，提供 MLP、RBM、SDA 等深度学习模型。Pylearn 2 的目标是：

- Researchers add features as they need them. We avoid getting bogg ed down by too much top-down planning in advance.
- A machine learning toolbox for easy scientific experimentation.

- All models/algorithms published by the LISA lab should have reference implementations in Pylearn2.
- Pylearn2 may wrap other libraries such as scikits.learn when this is practical
- Pylearn2 differs from scikits.learn in that Pylearn2 aims to provide great flexibility and make it possible for a researcher to do almost anything, while scikits.learn aims to work as a "black box" that can produce good results even if the user does not understand the implementation
- Dataset interface for vector, images, video, ...
- Small framework for all what is needed for one normal MLP/RBM/SDA/Convolution experiments.
- Easy reuse of sub-component of Pylearn2.
- Using one sub-component of the library does not force you to use / learn to use all of the other sub-components if you choose not to.
- Support cross-platform serialization of learned models.
- Remain approachable enough to be used in the classroom (IFT6266 at the University of Montreal).

项目主页：

http://deeplearning.net/software/pylearn2/

https://github.com/lisa-lab/pylearn2

还有其他的一些 Python 的机器学习库，如：

pmll(https://github.com/pavlov99/pmll)

pymining(https://github.com/bartdag/pymining)

ease (https://github.com/edx/ease)

textmining(http://www.christianpeccei.com/textmining/)

更多的机器学习库可通过 https://pypi.python.org/pypi 查找。
分类: Data Mining, Programming(Python,Java)