

3-补充知识

SKLEARN实践

主讲：蔡 波

武汉大学网络安全学院

sklearn库

sklearn是scikit-learn的简称，是一个基于Python的第三方模块。sklearn库集成了一些常用的机器学习方法，在进行机器学习任务时，并不需要实现算法，只需要简单的调用sklearn库中提供的模块就能完成大多数的机器学习任务。

[Home](#)[Installation](#)[Documentation](#)[Examples](#)



scikit-learn

Machine Learning in Python

Machine Learning in Python

- Simple and efficient tools for data mining and data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

Classification

Identifying to which category an object belongs to.

Applications: Spam detection, Image recognition.

Algorithms: SVM, nearest neighbors, random forest, ...

Regression

Predicting a continuous-valued attribute associated with an object.

Applications: Drug response, Stock prices.

Algorithms: SVM, ridge regression, Lasso, ...

Clustering

Automatic grouping of similar objects into sets.

Applications: Customer segmentation, Grouping experiment subtasks.

Algorithms: k-means, spectral clustering, mean-shift, ...

Dimensionality reduction

Reducing the number of random variables to consider.

Applications: Visualization, Increased efficiency.

Algorithms: PCA, feature selection, non-negative matrix factorization, ...

Model selection

Comparing, validating and choosing parameters and models.

Goal: Improved accuracy via parameter tuning.

Modules: grid search, cross validation, metrics, ...

Preprocessing

Feature extraction and normalization.

Application: Transforming input data such as text for use with machine learning algorithms.

Modules: preprocessing, feature extraction, ...

News

Ongoing development: What's new (Changing)

November 2016: scikit-learn 0.16.1 is available for download (Changing).

September 2016: scikit-learn 0.16.0 is available for download (Changing).

November 2015: scikit-learn 0.17.2 is available for download (Changing).

March 2015: scikit-learn 0.16.2 is available for download (Changing).

July 2014: scikit-learn 0.15.2 is available for download (Changing).

July 1st 2014, 2014: international sprint. During this week-long sprint, we gathered 18 of the core contributors in Paris. We want to thank our sponsors: Paris-Saclay Center for Data Science & Engineering and our hosts: La Perleaux, Citius, Inria, and InriaLyon.

August 2013: scikit-learn 0.14 is available for download (Changing).

Community

Meet us: See authors and contributing

More Machine Learning: Find related projects

Questions? See FAQ and stackoverflow

mailing list: scikit-learn@python.org

IRC: #scikit-learn @ freenode

[Join us, download](#) [Cite us!](#)

[Read more about sponsors](#)

Who uses scikit-learn?



Qualtrics

"I think it's the most well-designed ML package I've seen so far."

[More testimonials](#)

Funding provided by:      

[More information on our contributors](#)

Python Scikit-learn

- ❖ 一组简单有效的工具集
- ❖ 依赖Python的NumPy, SciPy和matplotlib库。
- ❖ 开源、可复用

Scikit Learn 开源库中机器学习模型非常丰富，在使用中可以根据要解决的问题类型选择适当的机器学习模型。

Scikit-learn 常用函数

	应用 (Applications)	算法 (Algorithm)
分类 (Classification)	异常检测, 图像识别, 等	KNN, SVM, etc.
聚类 (Clustering)	图像分割, 群体划分, 等	K-Means, 谱聚类, etc.
回归 (Regression)	价格预测, 趋势预测, 等	线性回归, SVR, etc.
降维 (Dimension Reduction)	可视化	PCA, NMF, etc.



sklearn库的安装

sklearn库

sklearn库是在Numpy、Scipy和matplotlib的基础上开发而成的，因此在介绍sklearn的安装前，需要先安装这些依赖库。

Numpy库

Numpy（ Numerical Python的缩写）是一个开源的Python科学计算库。

Scipy库是sklearn库的基础，它是基于Numpy的一个集成了多种数学算法和函数的Python模块。

matplotlib是基于Numpy的一套Python工具包，它提供了大量的数据绘图工具。

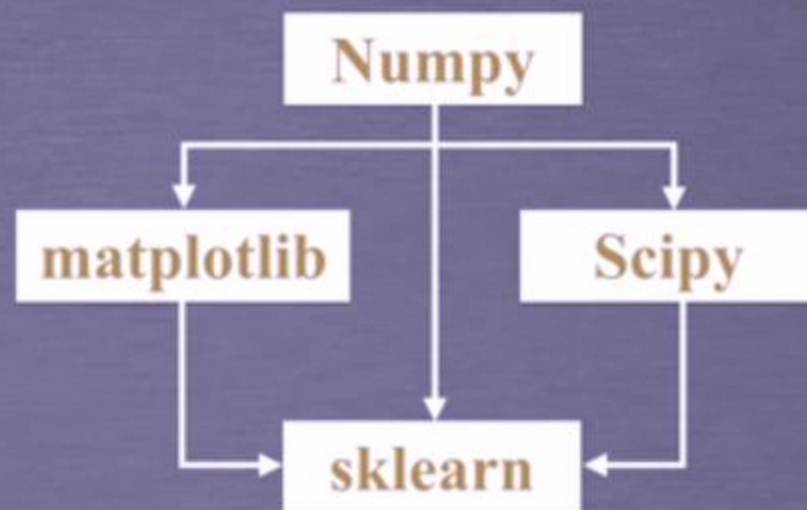
安装包的下载

下载地址: <http://www.lfd.uci.edu/~gohlke/pythonlibs/#>
(官方下载链接)

<https://www.lfd.uci.edu/~gohlke/pythonlibs/>

安装顺序

- ❖ Numpy库
- ❖ Scipy库
- ❖ matplotlib库
- ❖ sklearn库



依赖库之Numpy的安装

- ❖ 访问Numpy的相关下载链接
- ❖ 依据Python的具体版本下载对应的文件。例如：本课程使用的是Python3.5的64位版，请下载win_amd64.whl文件。

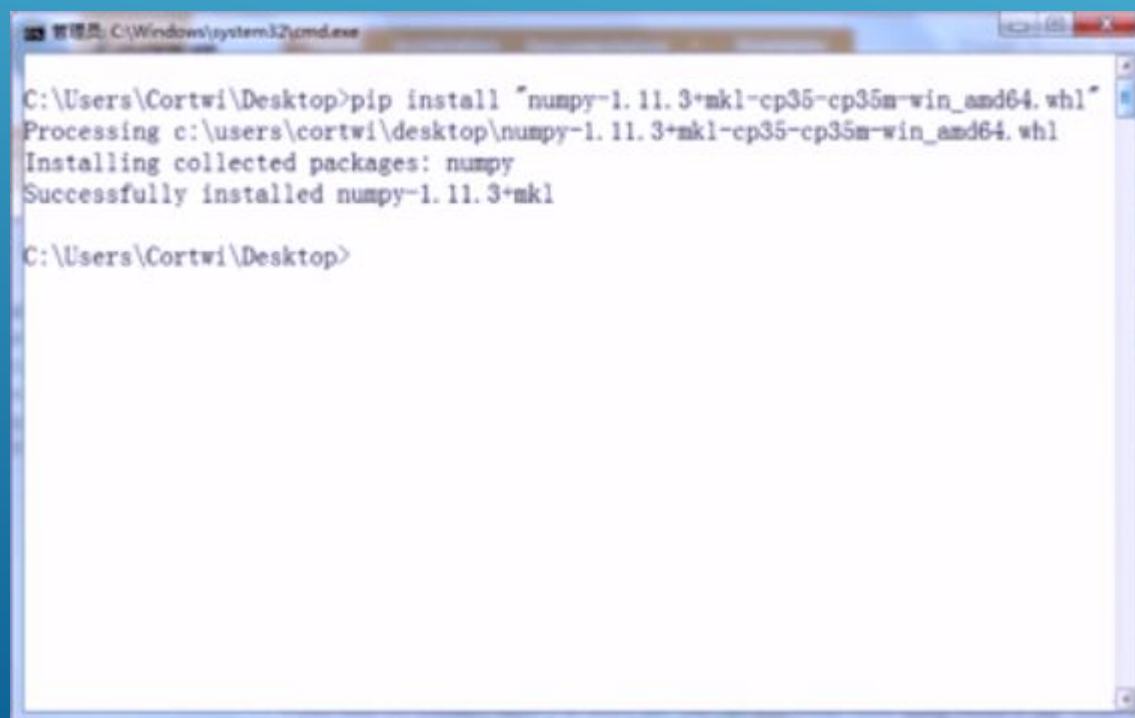
NumPy, a fundamental package needed for scientific computing with Python.
NumPy+MKL is linked to the Intel® Math Kernel Library and includes required

- [numpy-1.11.3+mkl-cp27-cp27m-win32.whl](#)
- [numpy-1.11.3+mkl-cp27-cp27m-win_amd64.whl](#)
- [numpy-1.11.3+mkl-cp34-cp34m-win32.whl](#)
- [numpy-1.11.3+mkl-cp34-cp34m-win_amd64.whl](#)
- [numpy-1.11.3+mkl-cp35-cp35m-win32.whl](#)
- [numpy-1.11.3+mkl-cp35-cp35m-win_amd64.whl](#)
- [numpy-1.11.3+mkl-cp36-cp36m-win32.whl](#)
- [numpy-1.11.3+mkl-cp36-cp36m-win_amd64.whl](#)
- [numpy-1.12.1+mkl-cp27-cp27m-win32.whl](#)
- [numpy-1.12.1+mkl-cp27-cp27m-win_amd64.whl](#)
- [numpy-1.12.1+mkl-cp34-cp34m-win32.whl](#)
- [numpy-1.12.1+mkl-cp34-cp34m-win_amd64.whl](#)
- [numpy-1.12.1+mkl-cp35-cp35m-win32.whl](#)
- [numpy-1.12.1+mkl-cp35-cp35m-win_amd64.whl](#)
- [numpy-1.12.1+mkl-cp36-cp36m-win32.whl](#)
- [numpy-1.12.1+mkl-cp36-cp36m-win_amd64.whl](#)

依赖库之Numpy的安装

找到下载的文件的路径，打开windows的DOS命令行窗口，执行如下命令：

```
pip install "numpy-1.11.3+mk1-cp35-cp35m-win_amd64.whl"
```



```
C:\Users\Cortwi\Desktop>pip install "numpy-1.11.3+mk1-cp35-cp35m-win_amd64.whl"
Processing c:\users\cortwi\desktop\numpy-1.11.3+mk1-cp35-cp35m-win_amd64.whl
Installing collected packages: numpy
Successfully installed numpy-1.11.3+mk1
C:\Users\Cortwi\Desktop>
```

依赖库之Scipy的安装

- ❖ 访问Scipy的相关下载链接
- ❖ 依据Python的具体版本下载对应的文件。同样这里需要下载右侧红框中*win_amd64.whl文件。

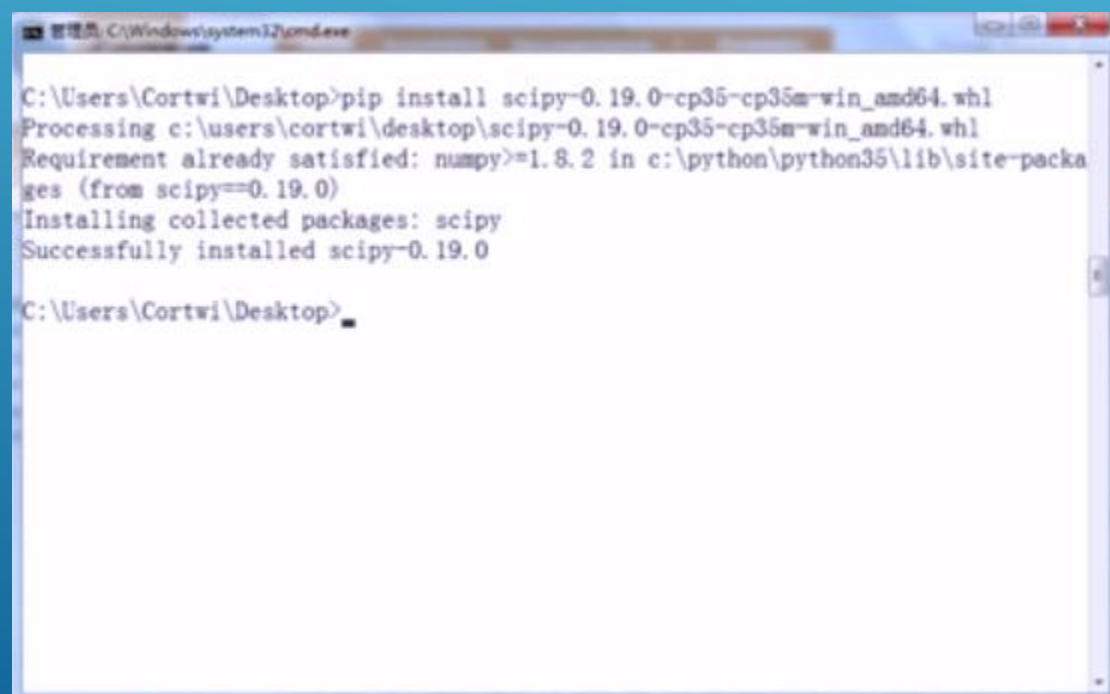
SciPy is software for mathematics, science, and engineering.
Install numpy+mkl before installing scipy.

- [scipy-0.19.0-cp27-cp27m-win32.whl](#)
- [scipy-0.19.0-cp27-cp27m-win_amd64.whl](#)
- [scipy-0.19.0-cp34-cp34m-win32.whl](#)
- [scipy-0.19.0-cp34-cp34m-win_amd64.whl](#)
- [scipy-0.19.0-cp35-cp35m-win32.whl](#)
- [scipy-0.19.0-cp35-cp35m-win_amd64.whl](#)
- [scipy-0.19.0-cp36-cp36m-win32.whl](#)
- [scipy-0.19.0-cp36-cp36m-win_amd64.whl](#)

依赖库之Scipy的安装

找到下载的文件的路径，打开windows的DOS命令行窗口，执行如下命令：

```
pip install scipy-0.19.0-cp35-cp35m-win_amd64.whl
```



```
管理员: C:\Windows\system32\cmd.exe
C:\Users\Cortwi\Desktop>pip install scipy-0.19.0-cp35-cp35m-win_amd64.whl
Processing c:\users\cortwi\desktop\scipy-0.19.0-cp35-cp35m-win_amd64.whl
Requirement already satisfied: numpy>=1.8.2 in c:\python\python35\lib\site-packages (from scipy==0.19.0)
Installing collected packages: scipy
Successfully installed scipy-0.19.0
C:\Users\Cortwi\Desktop>_
```

依赖库之matplotlib的安装

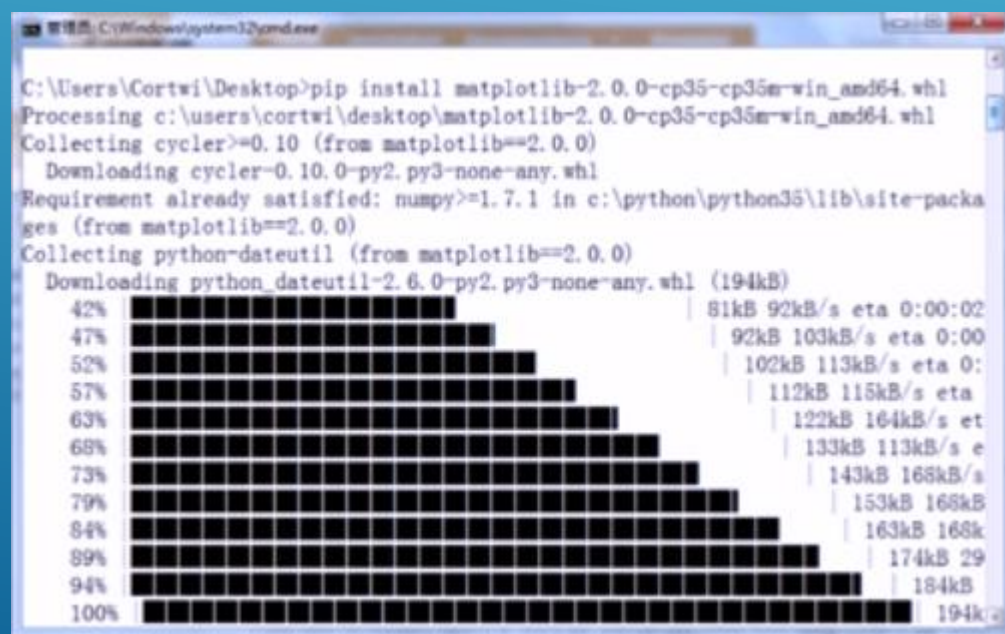
- ❖ 访问Numpy的相关下载链接
- ❖ 依据Python的具体版本下载对应的文件。下载红框中对应的*win_amd64.whl文件。

Matplotlib, a 2D plotting library.
Requires numpy, dateutil, pytz, pyparsing, cycler, setuptools, imageio.
matplotlib-1.5.3-cp27-cp27m-win32.whl
matplotlib-1.5.3-cp27-cp27m-win_amd64.whl
matplotlib-1.5.3-cp34-cp34m-win32.whl
matplotlib-1.5.3-cp34-cp34m-win_amd64.whl
matplotlib-1.5.3-cp35-cp35m-win32.whl
matplotlib-1.5.3-cp35-cp35m-win_amd64.whl
matplotlib-1.5.3-cp36-cp36m-win32.whl
matplotlib-1.5.3-cp36-cp36m-win_amd64.whl
matplotlib-1.5.3.chm
matplotlib-2.0.0-cp27-cp27m-win32.whl
matplotlib-2.0.0-cp27-cp27m-win_amd64.whl
matplotlib-2.0.0-cp34-cp34m-win32.whl
matplotlib-2.0.0-cp34-cp34m-win_amd64.whl
matplotlib-2.0.0-cp35-cp35m-win32.whl
matplotlib-2.0.0-cp35-cp35m-win_amd64.whl
matplotlib-2.0.0-cp36-cp36m-win32.whl
matplotlib-2.0.0-cp36-cp36m-win_amd64.whl
matplotlib-2.0.0.chm
matplotlib-2.x-windows-link-libraries.zip
matplotlib_tests-1.5.3-py2.py3-none-any.whl
matplotlib_tests-2.0.0-py2.py3-none-any.whl

依赖库之matplotlib的安装

找到下载的文件的路径，打开windows的DOS命令行窗口，使用如下命令：

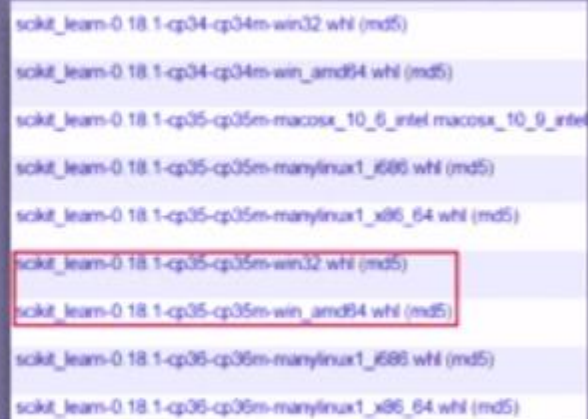
```
pip install matplotlib-2.0.0-cp35-cp35m-win_amd64.whl
```



```
管理員: C:\Windows\system32\cmd.exe
C:\Users\Cortwi\Desktop>pip install matplotlib-2.0.0-cp35-cp35m-win_amd64.whl
Processing c:\users\cortwi\desktop\matplotlib-2.0.0-cp35-cp35m-win_amd64.whl
Collecting cyclor>=0.10 (from matplotlib==2.0.0)
  Downloading cyclor-0.10.0-py2.py3-none-any.whl
Requirement already satisfied: numpy>=1.7.1 in c:\python\python35\lib\site-packa
ges (from matplotlib==2.0.0)
Collecting python-dateutil (from matplotlib==2.0.0)
  Downloading python_dateutil-2.6.0-py2.py3-none-any.whl (194kB)
  42% |#####| 81kB 92kB/s eta 0:00:02
  47% |#####| 92kB 103kB/s eta 0:00
  52% |#####| 102kB 113kB/s eta 0:
  57% |#####| 112kB 115kB/s eta
  63% |#####| 122kB 164kB/s et
  68% |#####| 133kB 113kB/s e
  73% |#####| 143kB 168kB/s
  79% |#####| 153kB 168kB
  84% |#####| 163kB 168k
  89% |#####| 174kB 29
  94% |#####| 184kB
  100% |#####| 194kB
```

sklearn库的安装

- ❖ 访问sklearn的相关下载链接
- ❖ 找到对应的安装文件
- ❖ 下载右侧红框中对应的*win_amd64.whl文件。



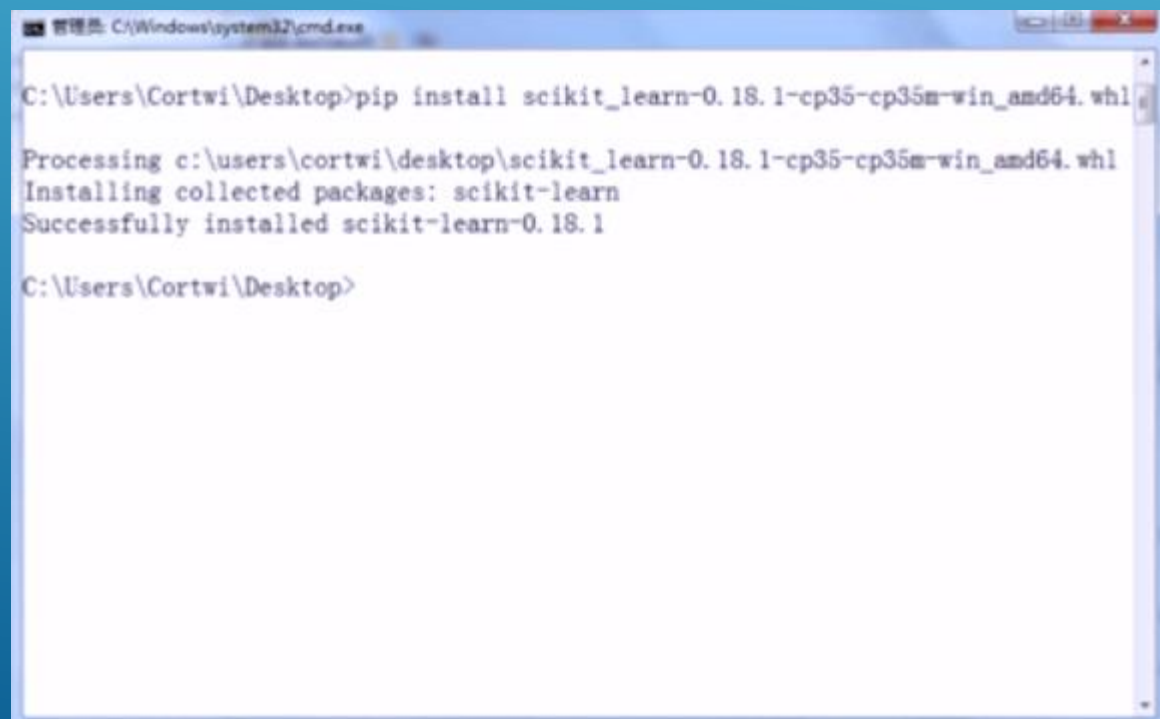
scikit_learn-0.18.1-cp34-cp34m-win32.whl (md5)
scikit_learn-0.18.1-cp34-cp34m-win_amd64.whl (md5)
scikit_learn-0.18.1-cp35-cp35m-macosx_10_6_intel macosx_10_9_intel
scikit_learn-0.18.1-cp35-cp35m-manylinux1_i686.whl (md5)
scikit_learn-0.18.1-cp35-cp35m-manylinux1_x86_64.whl (md5)
scikit_learn-0.18.1-cp35-cp35m-win32.whl (md5)
scikit_learn-0.18.1-cp35-cp35m-win_amd64.whl (md5)
scikit_learn-0.18.1-cp36-cp36m-manylinux1_i686.whl (md5)
scikit_learn-0.18.1-cp36-cp36m-manylinux1_x86_64.whl (md5)

下载地址: <https://pypi.python.org/pypi/scikit-learn/0.18.1>

sklearn库的安装

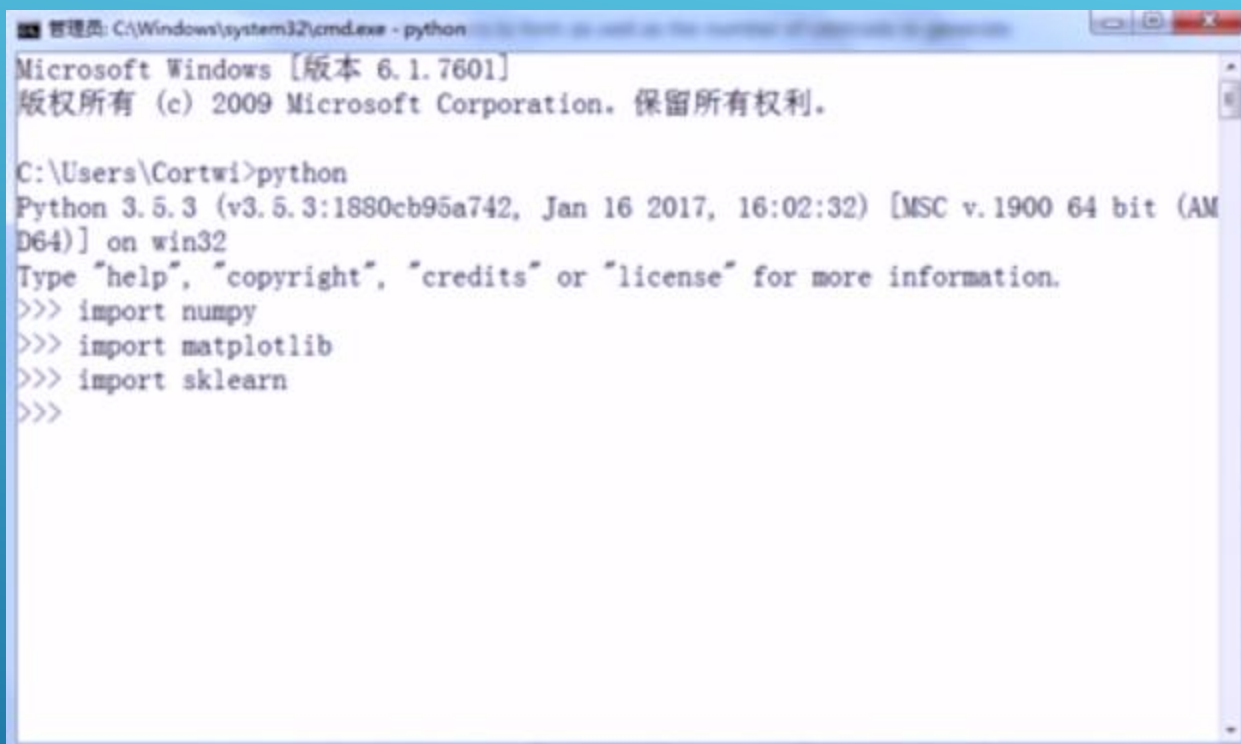
找到下载的文件的路径，打开windows的DOS命令行窗口，使用如下命令：

```
pip install scikit_learn-0.18.1-cp35-cp35m-win_amd64.whl
```



```
管理员: C:\Windows\system32\cmd.exe  
C:\Users\Cortwi\Desktop>pip install scikit_learn-0.18.1-cp35-cp35m-win_amd64.whl  
Processing c:\users\cortwi\desktop\scikit_learn-0.18.1-cp35-cp35m-win_amd64.whl  
Installing collected packages: scikit-learn  
Successfully installed scikit-learn-0.18.1  
C:\Users\Cortwi\Desktop>
```


Scikit Learn 测试



```
管理员: C:\Windows\system32\cmd.exe - python
Microsoft Windows [版本 6.1.7601]
版权所有 (c) 2009 Microsoft Corporation. 保留所有权利。

C:\Users\Cortwi>python
Python 3.5.3 (v3.5.3:1880cb95a742, Jan 16 2017, 16:02:32) [MSC v.1900 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>> import numpy
>>> import matplotlib
>>> import sklearn
>>>
```



sklearn库中的标准数据集及基本功能

数据集总览

	数据集名称	调用方式	适用算法	数据规模
小数据集	波士顿房价数据集	load_boston()	回归	506*13
	鸢尾花数据集	load_iris()	分类	150*4
	糖尿病数据集	load_diabetes()	回归	442*10
	手写数字数据集	load_digits()	分类	5620*64
大数据集	Olivetti脸部图像数据集	fetcholivetti_faces()	降维	400*64*64
	新闻分类数据集	fetch_20newsgroups()	分类	-
	带标签的人脸数据集	fetch_lfw_people()	分类; 降维	-
	路透社新闻语料数据集	fetch_rcv1()	分类	804414*47236

注：小数据集可以直接使用，大数据集要在调用时程序自动下载（一次即可）。

波士顿房价数据集

波士顿房价数据集包含506组数据，每条数据包含房屋以及房屋周围的详细信息。其中包括城镇犯罪率、一氧化氮浓度、住宅平均房间数、到中心区域的加权距离以及自住房平均房价等。因此，波士顿房价数据集能够应用到回归问题上。

波士顿房价数据集

使用`sklearn.datasets.load_boston`即可加载相关数据集

重要参数:

❖ `return_X_y`: 表示是否返回`target`（即价格），默认为`False`，只返回`data`（即属性）。

波士顿房价数据集-加载示例

示例1

```
>>> from sklearn.datasets import load_boston
>>> boston = load_boston()
>>> print(boston.data.shape)
(506, 13)
```

示例2

```
>>> from sklearn.datasets import load_boston
>>> data, target = load_boston(return_X_y=True)
>>> print(data.shape)
(506, 13)
>>> print(target.shape)    (506)
```

鸢尾花数据集

鸢尾花数据集采集的是鸢尾花的测量数据以及其所属的类别。

测量数据包括：萼片长度、萼片宽度、花瓣长度、花瓣宽度。

类别共分为三类：Iris Setosa, Iris Versicolour, Iris Virginica。该数据集可用于多分类问题。

萼片长度	萼片宽度	花瓣长度	花瓣宽度	类别
5.1	3.5	1.4	0.2	Iris-setosa
4.9	3	1.4	0.2	Iris-setosa
4.7	3.2	1.3	0.2	Iris-setosa
4.6	3.1	1.5	0.2	Iris-setosa
5	3.6	1.4	0.2	Iris-setosa
5.4	3.9	1.7	0.4	Iris-setosa
4.6	3.4	1.4	0.3	Iris-setosa
5	3.4	1.5	0.2	Iris-setosa
4.4	2.9	1.4	0.2	Iris-setosa
4.9	3.1	1.5	0.1	Iris-setosa
5.4	3.7	1.5	0.2	Iris-setosa
4.8	3.4	1.6	0.2	Iris-setosa
4.8	3	1.4	0.1	Iris-setosa
4.3	3	1.1	0.1	Iris-setosa
5.8	4	1.2	0.2	Iris-setosa

鸢尾花数据集数据示例

鸢尾花数据集

使用`sklearn.datasets.load_iris`即可加载相关数据集

参数：

- ❖ `return_X_y`: 若为True，则以（`data`, `target`）形式返回数据；默认为False，表示以字典形式返回数据全部信息（包括`data`和`target`）。

鸢尾花数据集-加载示例

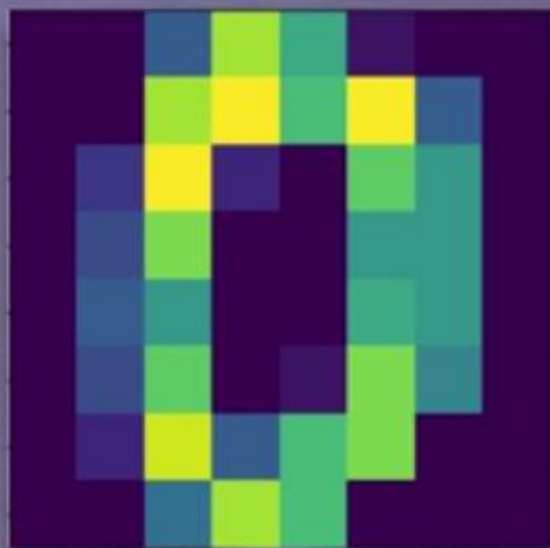
```
>>> from sklearn.datasets import load_iris
>>> iris = load_iris()
>>> print(iris.data.shape)
(150, 4)
>>> print(iris.target.shape)
(150, )
>>> list(iris.target_names)
['setosa', 'versicolor', 'virginica']
```


手写数字数据集

手写数字数据集包括1797个0-9的手写数字数据，每个数字由 8×8 大小的矩阵构成，矩阵中值的范围是0-16，代表颜色的深度。

手写数字数据集

0	0	5	13	9	1	0	0
0	0	13	15	10	15	5	0
0	3	15	2	0	11	8	0
0	4	12	0	0	8	8	0
0	5	8	0	0	9	8	0
0	4	11	0	1	12	7	0
0	2	14	5	10	12	0	0
0	0	6	13	10	0	0	0



数字0的样本

手写数字数据集

使用`sklearn.datasets.load_digits`即可加载相关数据集

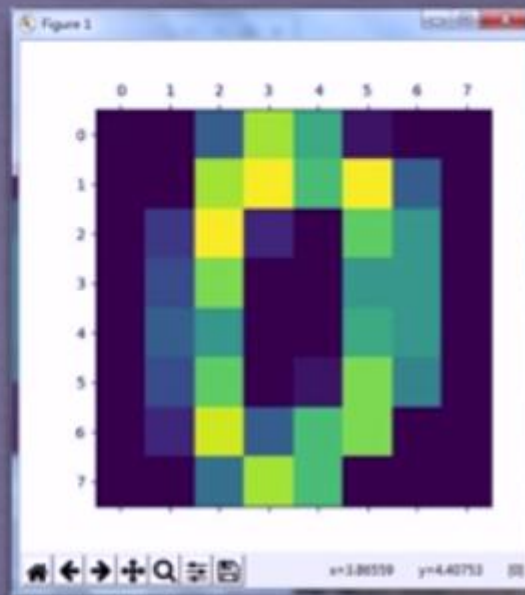
参数：

- ❖ `return_X_y`：若为True，则以（`data, target`）形式返回数据；默认为False，表示以字典形式返回数据全部信息（包括`data`和`target`）。
- ❖ `n_class`：表示返回数据的类别数，如：`n_class=5`，则返回0到4的数据样本。

手写数字数据集

示例

```
>>> from sklearn.datasets import load_digits
>>> digits = load_digits()
>>> print(digits.data.shape)
(1797, 64)
>>> print(digits.target.shape)
(1797, )
>>> print(digits.images.shape)
(1797, 8, 8)
>>> import matplotlib.pyplot as plt
>>> plt.matshow(digits.images[0])
>>> plt.show()
```



sklearn库的基本功能

sklearn库的基本功能

sklearn库的共分为6大部分，分别用于完成分类任务、回归任务、聚类任务、降维任务、模型选择以及数据的预处理。

分类任务

分类模型	加载模块
最近邻算法	<code>neighbors.NearestNeighbors</code>
支持向量机	<code>svm.SVC</code>
朴素贝叶斯	<code>naive_bayes.GaussianNB</code>
决策树	<code>tree.DecisionTreeClassifier</code>
集成方法	<code>ensemble.BaggingClassifier</code>
神经网络	<code>neural_network.MLPClassifier</code>

回归任务

回归模型	加载模块
岭回归	<code>linear_model.Ridge</code>
Lasso回归	<code>linear_model.Lasso</code>
弹性网络	<code>linear_model.ElasticNet</code>
最小角回归	<code>linear_model.Lars</code>
贝叶斯回归	<code>linear_model.BayesianRidge</code>
逻辑回归	<code>linear_model.LogisticRegression</code>
多项式回归	<code>preprocessing. PolynomialFeatures</code>

聚类任务

聚类方法	加载模块
K-means	cluster.KMeans
AP聚类	cluster.AffinityPropagation
均值漂移	cluster.MeanShift
层次聚类	cluster.AgglomerativeClustering
DBSCAN	cluster.DBSCAN
BIRCH	cluster.Birch
谱聚类	cluster.SpectralClustering

降维任务

降维方法	加载模块
主成分分析	decomposition.PCA
截断SVD和LSA	decomposition.TruncatedSVD
字典学习	decomposition.SparseCoder
因子分析	decomposition.FactorAnalysis
独立成分分析	decomposition.FastICA
非负矩阵分解	decomposition.NMF
LDA	decomposition.LatentDirichletAllocation

