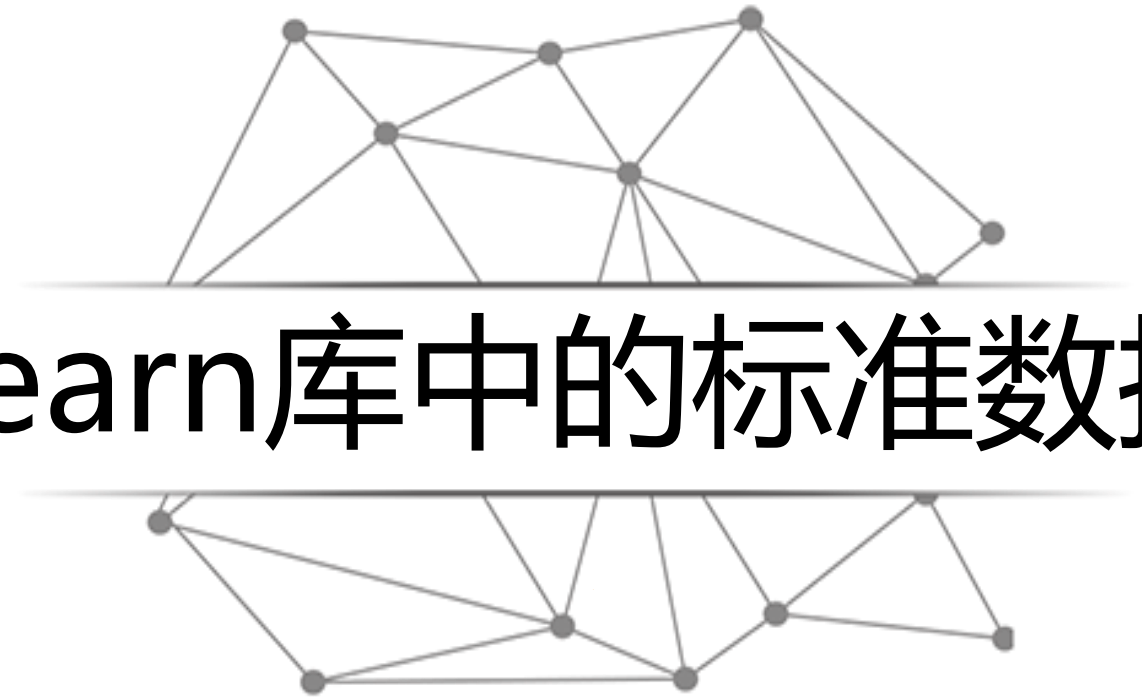
sklearn库中的标准数据集及基本功能

ML03

礼欣
www.python123.org



sklearn库中的标准数据集

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	fetch_olivetti_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

输入

CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
0.00632	18	2.31	0	0.538	6.575	65.2	4.09	1	296	15.3	396.9	4.98	24
0.02731	0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	396.9	9.14	21.6
0.02729	0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7
0.03237	0	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4
0.06905	0	2.18	0	0.458	7.147	54.2	6.0622	3	222	18.7	396.9	5.33	36.2
0.02985	0	2.18	0	0.458	6.43	58.7	6.0622	3	222	18.7	394.12	5.21	28.7
0.08829	12.5	7.87	0	0.524	6.012	66.6	5.5605	5	311	15.2	395.6	12.43	22.9
0.14455	12.5	7.87	0	0.524	6.172	96.1	5.9505	5	311	15.2	396.9	19.15	27.1
0.21124	12.5	7.87	0	0.524	5.631	100	6.0821	5	311	15.2	386.63	29.93	16.5
0.17004	12.5	7.87	0	0.524	6.004	85.9	6.5921	5	311	15.2	386.71	17.1	18.9
0.22489	12.5	7.87	0	0.524	6.377	94.3	6.3467	5	311	15.2	392.52	20.45	15
0.11747	12.5	7.87	0	0.524	6.009	82.9	6.2267	5	311	15.2	396.9	13.27	18.9
0.09378	12.5	7.87	0	0.524	5.889	39	5.4509	5	311	15.2	390.5	15.71	21.7
0.62976	0	8.14	0	0.538	5.949	61.8	4.7075	4	307	21	396.9	8.26	20.4

输出

图. 部分房价数据

波士顿房价数据集-属性描述

CRIM：城镇人均犯罪率。

ZN：住宅用地超过 25000 sq.ft. 的比例。

INDUS：城镇非零售商用土地的比例。

CHAS：查理斯河空变量（如果边界是河流，则为1；否则为0）

NOX：一氧化氮浓度。

RM：住宅平均房间数。

AGE：1940 年之前建成的自用房屋比例。

DIS：到波士顿五个中心区域的加权距离。

RAD：辐射性公路的接近指数。

TAX：每 10000 美元的全值财产税率。

PTRATIO：城镇师生比例。

B：1000（Bk-0.63）^ 2，其中 Bk 指代城镇中黑人的比例。

LSTAT：人口中地位低下者的比例。

MEDV：自住房的平均房价，以千美元计。

波士顿房价数据集

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

其重要参数为：

- **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

图. 鸢尾花部数据集分数据示例

鸢尾花数据集

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

其参数有：

- **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

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

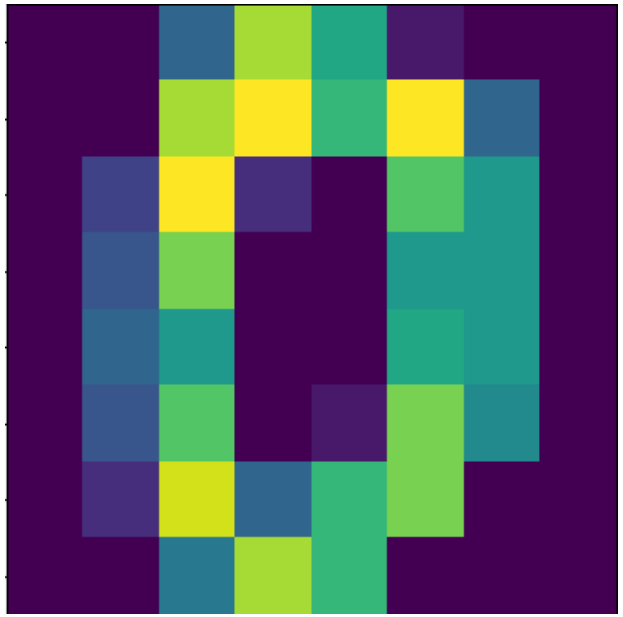


图. 数字0的样本

手写数字数据集

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

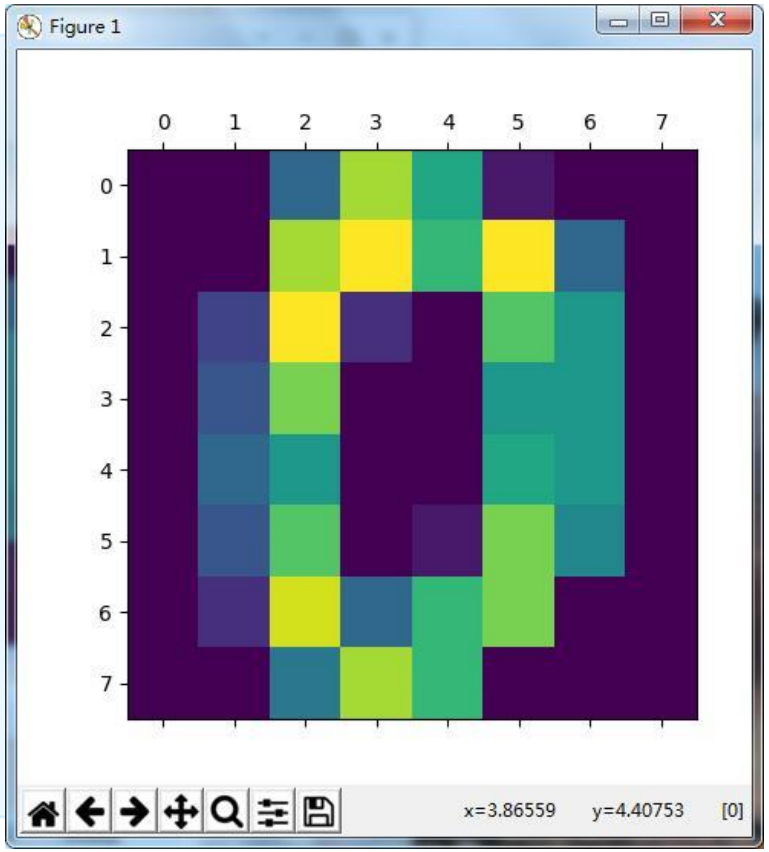
其参数包括：

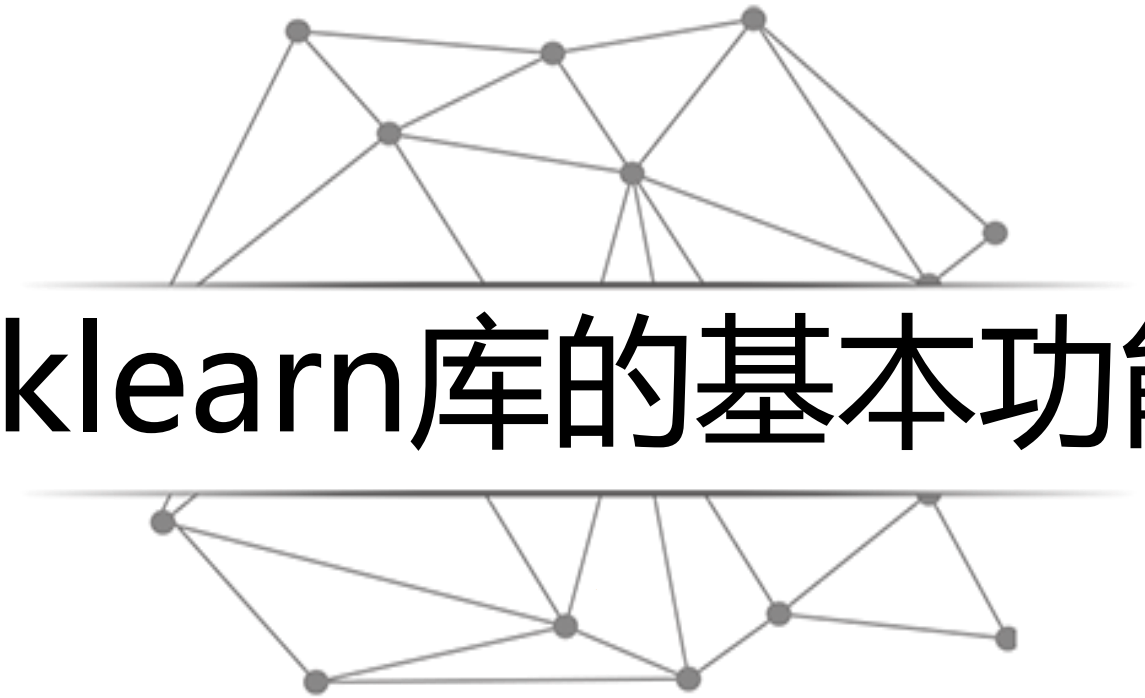
- **return_X_y**:若为True，则以（data, target）形式返回数据；默认为False，表示以字典形式返回数据全部信息（包括data和target）；
- **n_class**：表示返回数据的类别数，如：n_class=5,则返回0到4的数据样本。

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	neighbors.NearestNeighbors
支持向量机	svm.SVC
朴素贝叶斯	naive_bayes.GaussianNB
决策树	tree.DecisionTreeClassifier
集成方法	ensemble.BaggingClassifier
神经网络	neural_network.MLPClassifier

回归任务

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

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