AI基础：机器学习和深度学习的练习数据

原创 机器学习初学者 机器学习初学者 前天

0.导语

初学者学习机器学习和深度学习的时候，经常会找不到练习的数据，本文提供了获取数据的一些方法。

目前我在编写AI基础系列，目前已经发布：

AI 基础：简易数学入门

AI 基础：Python开发环境设置和小技巧

AI 基础：Python 简易入门

AI 基础：Numpy 简易入门

AI 基础：Pandas 简易入门

AI 基础：Scipy(科学计算库) 简易入门

AI基础：数据可视化简易入门（matplotlib和seaborn）

AI基础：机器学习库Scikit-learn的使用

AI基础：机器学习简易入门

AI基础：机器学习的损失函数

AI基础：特征工程-类别特征

AI基础：特征工程-数字特征处理

AI基础：特征工程-文本特征处理

AI基础：词嵌入基础和Word2Vec

AI基础：图解Transformer

AI基础：一文看懂BERT

AI基础：入门人工智能必看的论文

AI基础：走进深度学习

AI基础：卷积神经网络

AI基础：深度学习论文阅读路线（127篇经典论文下载）

AI基础：数据增强方法综述

后续持续更新

一、scikit-learn自带数据集

Scikit-learn内置了很多可以用于机器学习的数据，可以用两行代码就可以使用这些数据。

一、自带数据集
自带的小的数据集为：sklearn.datasets.load_<name>

load_boston	Boston房屋价格	回归	506*13
fetch_california_housing	加州住房	回归	20640*9
load_diabetes	糖尿病	回归	442*10
load_digits	手写字	分类	1797*64
load_breast_cancer	乳腺癌	分类、聚类	(357+212)*30
load_iris	鸢尾花	分类、聚类	(50*3)*4
load_wine	葡萄酒	分类	(59+71+48)*13
load_linnerud	体能训练	多分类	20

怎么用：

数据集的信息关键字：

- DESCR：
  数据集的描述信息
- data：
  内部数据（即：X）
- feature_names：

数据字段名

- target：
  数据标签（即：y）
- target_names：
  标签字段名(回归数据集无此项）

使用方法（以load_iris为例）
数据介绍：

- 一般用于做分类测试
- 有150个数据集，共分为3类，每类50个样本。每个样本有4个特征。
- 每条记录都有 4 项特征：包含4个特征（Sepal.Length（花萼长度）、Sepal.Width（花萼宽度）、Petal.Length（花瓣长度）、Petal.Width（花瓣宽度）），特征值都为正浮点数，单位为厘米。
- 可以通过这4个特征预测鸢尾花卉属于（iris-setosa（山鸢尾）, iris-versicolour（杂色鸢尾）, iris-virginica（维吉尼亚鸢尾））中的哪一品种。

第一步：
导入数据

```
from  sklearn.datasets  import  load_iris
iris = load_iris()
```

第二步：
定义X和y

```
X, y = iris.data, iris.target
```

此外，可以看下数据的维度：

```
X.shape,y.shape
```

输出为：

```
((150, 4), (150,))
```

查看特征名：

```
iris.feature_names
```

输出为：

```
['sepal length (cm)',
 'sepal width (cm)',
 'petal length (cm)',
 'petal width (cm)']
```

查看标签名：

```
iris.target_names
```

输出为：

```
array(['setosa', 'versicolor', 'virginica'], dtype='<U10')
```

划分训练集和测试集:

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25)
```

这样就把训练集和测试集按照3比1划分了，接下来就可以用机器学习算法进行训练和测试了。

小技巧：将数据转换为Dataframe格式（两种方法都可以）：

```
import pandas as pd
df_X = pd.DataFrame(iris.data, columns=iris.feature_names)
#这个是X
df_y = pd.DataFrame(iris.target, columns=["target"])
#这个是y
df=pd.concat([df_X,df2],axis=1)#横向合并
df.head()
```

或者：

```
import numpy as np
import pandas as pd
col_names = iris['feature_names'] + ['target']
df = pd.DataFrame(data= np.c_[iris['data'], iris['target']], columns=col_names)
df.head()
```

输出结果一致：

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	target
0	5.1	3.5	1.4	0.2	0
1	4.9	3.0	1.4	0.2	0
2	4.7	3.2	1.3	0.2	0
3	4.6	3.1	1.5	0.2	0
4	5.0	3.6	1.4	0.2	0

二、可在线下载的数据集（需要下载）

下载的数据集为：sklearn.datasets.fetch_<name>

这类数据需要在线下载，有点慢

fetch_20newsgroups	用于文本分类、文本挖据和信息检索研究的国际标准数据集之一。数据集收集了大约20,000左右的新闻组文档，均匀分为20个不同主题的新闻组集合。返回一个可以被文本特征提取器
fetch_20newsgroups_vectorized	这是上面这个文本数据的向量化后的数据，返回一个已提取特征的文本序列，即不需要使用特征提取器

fetch_california_housing	加利福尼亚的房价数据，总计20640个样本，每个样本8个属性表示，以及房价作为target，所有属性值均为number，详情可调用fetch_california_housing()['DESCR']了解每个属性的具体含义；
fetch_covtype	森林植被类型，总计581012个样本，每个样本由54个维度表示（12个属性，其中2个分别是onehot4维和onehot40维），以及target表示植被类型1-7，所有属性值均为number，详情可调用fetch_covtype()['DESCR']了解每个属性的具体含义
fetch_kddcup99	KDD竞赛在1999年举行时采用的数据集，KDD99数据集仍然是网络入侵检测领域的事实Benckmark，为基于计算智能的网络入侵检测研究奠定基础，包含41项特征
fetch_lfw_pairs	该任务称为人脸验证：给定一对两张图片，二分类器必须预测这两个图片是否来自同一个人。
fetch_lfw_people	打好标签的人脸数据集
fetch_mldata	从 mldata.org 中下载数据集
fetch_olivetti_faces	Olivetti 脸部图片数据集
fetch_rcv1	路透社新闻语聊数据集
fetch_species_distributions	物种分布数据集

使用方法与自带数据集一致，只是多了下载过程（示例：fetch_20newsgroups）

```
from  sklearn.datasets  import  fetch_20newsgroups
news = fetch_20newsgroups(subset='all')  #本次使用的数据需要到互联网上下载
from  sklearn.model_selection  import  train_test_split
#对数据训练集和测试件进行划分
X_train, X_test, y_train, y_test = train_test_split(
        news.data, news.target, test_size=0.25, random_state=33)
```

三、生成数据集

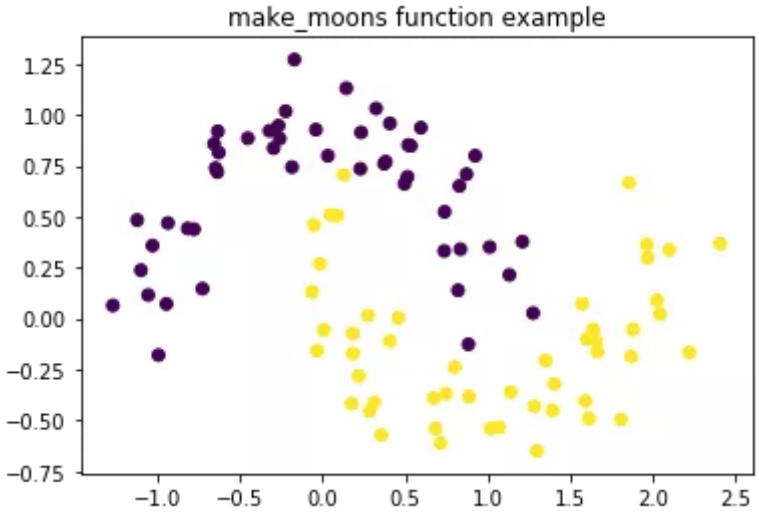
可以用来分类任务，可以用来回归任务，可以用来聚类任务，用于流形学习的，用于因子分解任务的，用于分类任务和聚类任务的：这些函数产生样本特征向量矩阵以及对应的类别标签集合

- make_blobs：多类单标签数据集，为每个类分配一个或多个正态分布的点集
- make_classification：多类单标签数据集，为每个类分配一个或多个正态分布的点集，提供了为数据添加噪声的方式，包括维度相关性，无效特征以及冗余特征等
- make_gaussian-quantiles：将一个单高斯分布的点集划分为两个数量均等的点集，作为两类

- make_hastie-10-2：产生一个相似的二元分类数据集，有10个维度
- make_circle和make_moons：产生二维二元分类数据集来测试某些算法的性能，可以为数据集添加噪声，可以为二元分类器产生一些球形判决界面的数据

举例：

```
import matplotlib.pyplot as plt
from sklearn.datasets import make_moons
X, y = make_moons(n_samples=100, noise=0.15, random_state=42)
plt.title('make_moons function example')
plt.scatter(X[:,0],X[:,1],marker='o',c=y)
plt.show()
```



深度学习数据集

MS-COCO

COCO是一个可用于object detection, segmentation and caption的大型数据集。

http://cocodataset.org/#home

ImageNet

图像总数约1,500,000; 每个都有多个边界框和相应的类标签。

大小：约150GB

http://www.image-net.org

Yelp Reviews

它由数百万用户评论、商业类型和来自多个大型城市的超过20万张照片组成。这在全球都是一个非常常用的NLP挑战级数据集。

大小：2.66 GB JSON，2.9 GB SQL and 7.5 GB Photos（全部已压缩）

数量：5,200,000条评论，174,000条商业类型，20万张图片和11个大型城市

https://www.yelp.com/dataset

……待补充

其它数据集

kaggle：

https://www.kaggle.com

天池：

https://tianchi.aliyun.com/dataset

搜狗实验室：

http://www.sogou.com/labs/resource/list_pingce.php

DC竞赛：

https://www.pkbigdata.com/common/cmptIndex.html

DF竞赛：

https://www.datafountain.cn/datasets

Google数据集

[需要科学上网]

https://toolbox.google.com/datasetsearch

科赛网

https://www.kesci.com/home/dataset

微软数据集

https://msropendata.com/

……待补充

总结

本文为机器学习初学者提供了使用scikit-learn内置数据的方法，用两行代码就可以使用这些数据，可以进行大部分的机器学习实验了。

参考

https://scikit-learn.org/stable/datasets/index.html

https://blog.csdn.net/fendouaini/article/details/79871922

备注：公众号菜单包含了整理了一本AI小抄，非常适合在通勤路上用学习。

往期精彩回顾

- 2019年公众号文章精选
- 适合初学者入门人工智能的路线及资料下载
- 机器学习在线手册
- 深度学习在线手册
- AI基础下载（第一部分）

备注：加入本站微信群或者qq群，请回复“加群”

加入知识星球（4500+用户，ID：92416895），请回复“知识星球”

喜欢文章，点个在看