

CSDN新首页上线啦，邀请你来立即体验！(http://blog.csdn.net/)

立即体

验

CSDN

博客 (//blog.csdn.net/?ref=toolbar) 学院 (//edu.csdn.net/?ref=toolbar)

下载 (//download.csdn.net/?ref=toolbar) GitChat (//gitbook.cn/?ref=csdn)

更多

0



Python-sklearn机器学习的第一个样例（2）



2017年05月19日 14:15:21

标签：Python (http://so.csdn.net/so/search/s.do?q=Python&t=blog) /

机器学习 (http://so.csdn.net/so/search/s.do?q=机器学习&t=blog) /

大数据 (http://so.csdn.net/so/search/s.do?q=大数据&t=blog)

592

接上一篇

Step 2：数据探索

数据探索是为了回答这几个问题：数据集有什么缺陷或错误？是否有异常数据？是否需要数据集进行修补或删除？首先，从把数据从csv中读入，保存在一个pandas DataFrame中：

In [1]:



weixin_3506...

(//my.csdn.net/?ref=toolbar)

(//write.blog.csdn.net/postedit?ref=toolbar)

ref=toolbar)source=csdnblor

番番要吃肉 (http://blog.cs...



+ 关注

(http://blog.csdn.net/xiexf189)

码云

未开通

原创

粉丝

喜欢

(https://gite
utm_sourc

4

4

0

他的最新文章

更多文章 (http://blog.csdn.net/xiexf189)

使用python进行简单的分词与词云 (http://blog.csdn.net/xiexf189/article/details/77477283)

Python数据分析练习：北京、广州PM 2.5空气质量分析（2）(http://blog.csdn.net/xiexf189/article/details/77368583)

Python数据分析练习：北京、广州PM 2.5空气质量分析（1）(http://blog.csdn.net/xiexf189/article/details/77368583)

```
import pandas as pd

iris_data = pd.read_csv('iris-data.csv')
iris_data.head()
```

Out[1]:

	sepal_length_cm	sepal_width_cm	petal_length_cm	petal_width_cm	class
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa

非常幸运，数据是按格式保存的。

接下来首要的工作是找到缺失的数据，感谢研究者，已经把缺失数据用'NA'表示了。

当然，我们也可以使用以下方法，把缺失值用'NA'表示。

In [2]:

```
iris_data = pd.read_csv('iris-data.csv', na_values=['NA'])
```

下一步，我们通常要看看数据集的分布情况，尤其是异常值。

我们先来看看数据集的概要。

In [3]:

```
iris_data.describe()
```

Out[3]:

	sepal_length_cm	sepal_width_cm	petal_length_cm	petal_width_cm
count	150.000000	150.000000	150.000000	145.000000

n.net/xiexf189/article/details/77367504)

Python-sklearn 机器学习的第一个样例 (7) (<http://blog.csdn.net/xiexf189/article/details/72598976>)

Python-sklearn机器学习的第一个样例 (6) (<http://blog.csdn.net/xiexf189/article/details/72598910>)

相关推荐

Python-sklearn机器学习的第一个样例 (3) (<http://blog.csdn.net/xiexf189/article/details/72528755>)

【Iris】【Keras】神经网络分类器和【scikit-learn】逻辑回归分类器的构建 (http://blog.csdn.net/lixiaowang_327/article/details/54094235)

Python-sklearn 机器学习的第一个样例 (1) (<http://blog.csdn.net/xiexf189/article/details/72518860>)

Python-sklearn机器学习的第一个样例 (3) (<http://blog.csdn.net/xiexf189/article/details/72528755>)

	sepal_length_cm	sepal_width_cm	petal_length_cm	petal_width_cm
mean	5.644627	3.054667	3.758667	1.236552
std	1.312781	0.433123	1.764420	0.755058
min	0.055000	2.000000	1.000000	0.100000
25%	5.100000	2.800000	1.600000	0.400000
50%	5.700000	3.000000	4.350000	1.300000
75%	6.400000	3.300000	5.100000	1.800000
max	7.900000	4.400000	6.900000	2.500000

可以看出有5条数据缺失了“花瓣宽度”数据。

除了使用表格，我们还经常用画图的方式探索数据，特别是表格太大的时候。

In [4]:

```
# This line tells the notebook to show plots inside of the notebook
%matplotlib inline

import matplotlib.pyplot as plt
import seaborn as sb
```

接下来创建一个散点图矩阵。散点图矩阵中，每一列的分布情况在对角线中表示，其他部分标识了每两列之间的关系。这是用来查找数据异常的有效工具。

我们可以使用plot包，对不同的种类着色，观察各个种类的趋势。

In [5]:

```
# We have to temporarily drop the rows with 'NA' values
# because the Seaborn plotting function does not know
# what to do with them
sb.pairplot(iris_data.dropna(), hue='class')
```

Out[5]:<seaborn.axisgrid.PairGrid at 0x615fb30>



联合办公

it培训机构排名

超强注意力 橡树湾 未来三年房价

嵌入式工程师待遇 人脸识别

舆情监测系统 OA办公系统 Python机..

视频会议终端 达内的 csv

他的热门文章

Python数据分析练习：北京、广州PM2.5
空气质量分析（1）(<http://blog.csdn.net/xiexf189/article/details/77367504>)

826

Python-sklearn机器学习的第一个样例
(6) (<http://blog.csdn.net/xiexf189/article/details/72598910>)

通过散点图矩阵，我们可以发现以下几个问题：

1. 原本应该只有三个种类的鸢尾花，但却出现了五个种类。意味着有编码错误。
2. 有一些明显的异常值的测量可能是错误的:一个Iris-setosa种的sepal_width_cm字段，游离在其正常范围之外；几个Iris-versicolor种类的sepal_length_cm字段不知为何接近于零。
3. 我们不得不放弃那些具有缺失值的行。因此，我们需要考虑如何处理这些错误的数值，进入下一步。



0



发表你的评论

(http://my.csdn.net/weixin_35068028)

737

Python-sklearn机器学习的第一个样例
(3) (<http://blog.csdn.net/xiexf189/article/details/72528755>)

718

Python-sklearn机器学习的第一个样例
(2) (<http://blog.csdn.net/xiexf189/article/details/72528667>)

589

Python-sklearn 机器学习的第一个样例
(1) (<http://blog.csdn.net/xiexf189/article/details/72518860>)

497

相关文章推荐

Python-sklearn机器学习的第一个样例(3) (<http://blog.csdn.net/xiexf189/article/details/72...>)

本文翻译自Randal S. Olson的文章《An example machine learning notebook》，原文：点击打开链接 这篇文章可以作为机器学习的第一个学习案例，通过这个案例，...



xiexf189 (<http://blog.csdn.net/xiexf189>) 2017年05月19日 14:23 719

【Iris】【Keras】神经网络分类器和【scikit-learn】逻辑回归分类器的构建 (<http://blog.csdn...>)

针对鸢尾花(Iris)数据集，基于scikit-learn训练logistic Regression分类器，基于Keras构建并训练三层前馈神经网络分类器，对比两者的正确率差异。Keras深度学习库...



lixiaowang_327 (http://blog.csdn.net/lixiaowang_327) 2017年01月05日 16:07 2492



【前端逆袭记】我是怎么从月薪4k到40k的！

谨以此篇文章献给我奋斗过的程序人生！我第一次编码是在我大一的时候....

(http://www.baidu.com/cb.php?c=lgF_pyfqHmknj0dP1f0lZ0qnfK9ujYzP1ndPWb10Aw-5Hc3rHnYnHb0TAq15HfLPWRznjb0T1Yvm1FhmHnsPHDYP1ckmHRv0AwY5HDdnHfzrHDsnH00lgF_5y9YIZ0lQzq-uZR8nLpUB48ugfEIAqspynElvNBnHqdlAdxTvqdThP-5yF_UvTkn0KzujYk0AFV5H00TZcqn0KdpyfqHRLPjnvnfKEpyfqHc4rj6kP0KWpyfqP1cvrHnz0AqLUWYs0ZK45HcsP6KWThnqPjnYnWm)



Python-sklearn 机器学习的第一个样例(1) (<http://blog.csdn.net/xiexf189/article/details/7...>)



这篇文章可以作为机器学习的第一个学习案例，通过这个案例，基本上可以把机器学习的整个过程接触一遍，对机器学习有了初步的了解。整个过程包括：业务问题、数据探索、数据整理和清洗、建模、模型调优、评估等步骤。...



xiexf189 (<http://blog.csdn.net/xiexf189>) 2017年05月19日 10:16 498

Python-sklearn机器学习的第一个样例(3) (<http://blog.csdn.net/xiexf189/article/details/72...>)


本文翻译自Randal S. Olson的文章《An example machine learning notebook》，原文：点击打开链接 这篇文章可以作为机器学习的第一个学习案例，通过这个案例，...



xiexf189 (<http://blog.csdn.net/xiexf189>) 2017年05月19日 14:23 719

Python-sklearn机器学习的第一个样例(6) (<http://blog.csdn.net/xiexf189/article/details/72...>)


本文翻译自Randal S. Olson的文章《An example machine learning notebook》，原文：点击打开链接 这篇文章可以作为机器学习的第一个学习案例，通过这个案例，...

 xiexf189 (<http://blog.csdn.net/xiexf189>) 2017年05月21日 16:06 742

 <p>1 议价 供应特价销售 制造销售Fu"-50u"镀金 8P8C</p>	 <p>2 0.25/个 批量改价:USB电线扣.汽车线卡.汽车电线固</p>	 <p>3 25.00/包 【批发】塑料定位片 25*25不干胶自粘式扎</p>
--	--	---


Python-sklearn 机器学习的第一个样例(7) (<http://blog.csdn.net/xiexf189/article/details/7...>)

本文翻译自Randal S. Olson的文章《An example machine learning notebook》，原文：点击打开链接 这篇文章可以作为机器学习的第一个学习案例，通过这个案例，...

 xiexf189 (<http://blog.csdn.net/xiexf189>) 2017年05月21日 16:14 336


【机器学习】Python sklearn包的使用示例以及参数调优示例 ([http://blog.csdn.net/wy_0928/...](http://blog.csdn.net/wy_0928/))

coding=utf-8 # !/usr/bin/env python """ 【说明】 1.当前sklearn版本0.18 2.sklearn自带的鸢尾花数据集样例：（1）样本特征矩阵（类型：...

 wy_0928 (http://blog.csdn.net/wy_0928) 2017年03月17日 15:30 4740



用Python开始机器学习(5: 文本特征抽取与向量化) sklearn (http://blog.csdn.net/sherri_d...)

<http://blog.csdn.net/lsidd/article/details/41520953> 假设我们刚看完诺兰的大片《星际穿越》，设想如何让机器来自动分析各位观众对电影的评价到底是“...


 sherri_du (http://blog.csdn.net/sherri_du) 2016年08月03日 19:26 1293



Python机器学习库SKLearn：数据集转换之特征提取 (<http://blog.csdn.net/cheng9981/articl...>)

特征提取：sklearn.feature_extraction模块可以用于从由诸如文本和图像的格式组成的数据集中提取机器学习算法支持的格式的特征。注意：特征提取与特征选择非常不同：前者包括将任意...


 cheng9981 (<http://blog.csdn.net/cheng9981>) 2017年03月13日 20:35  4333

python机器学习sklearn数据集iris介绍 (<http://blog.csdn.net/suibianshen2012/article/detail...>)

 ##### #说明：# 撰写本文的原因是，笔者在研究博文“<http://python.jobbole.com/83563/>”中发现 #
... 0

 suibianshen2012 (<http://blog.csdn.net/suibianshen2012>) 2016年07月11日 14:54  3733



Python机器学习库sklearn网格搜索与交叉验证 (<http://blog.csdn.net/cymy001/article/details...>)

 网格搜索一般是针对参数进行寻优，交叉验证是为了验证训练模型拟合程度。sklearn中的相关内容如下：（1）首先，要进行交叉验证，就要对数据集进行切分，构造训练集和测试集，不同的交叉验证方法会对...

 cymy001 (<http://blog.csdn.net/cymy001>) 2017年11月20日 02:57  172

python3机器学习——sklearn0.19.1版本——数据处理（一）（数据标准化、tfidf、独热编码）..

一、数据标准化 1、StandardScaler

 loveliuzz (<http://blog.csdn.net/loveliuzz>) 2017年11月21日 16:04  170


Python机器学习库sklearn自动特征选择（训练集） (<http://blog.csdn.net/cymy001/article/de...>)

1.单变量分析from sklearn.feature_selection import SelectPercentilefrom sklearn.datasets import load_breas...

 cymy001 (<http://blog.csdn.net/cymy001>) 2017年11月19日 19:37  172

Python机器学习库sklearn里利用决策树模型进行回归分析的原理 (<http://blog.csdn.net/cymy001/article/details/78027083>)

决策树的相关理论参考<http://blog.csdn.net/cymy001/article/details/78027083> #原数据网址变了, 新换的数据地址需要处理http://lib.stat....

 cymy001 (<http://blog.csdn.net/cymy001>) 2017年11月17日 04:51 57


Python机器学习库sklearn里利用感知机进行三分类(多分类)的原理 (<http://blog.csdn.net/cymy001/article/details/77992416>)

感知机的理论参考<http://blog.csdn.net/cymy001/article/details/77992416> from IPython.display import Im...

 cymy001 (<http://blog.csdn.net/cymy001>) 2017年11月14日 19:35 184


Python机器学习库sklearn数据预处理, 数据集构建, 特征选择 (<http://blog.csdn.net/cymy001/article/details/78027083>)

from IPython.display import Image %matplotlib inline # Added version check for recent scikit-learn 0...

 cymy001 (<http://blog.csdn.net/cymy001>) 2017年11月15日 23:11 160

Python下机器学习库安装经验——numpy、sklearn (<http://blog.csdn.net/lisasue/article/details/78027083>)

一、查看python安装情况以及pip对应版本pip2 --version pip3 --version 二、下载对应安装包、依赖包<http://www.lfd.uci.edu/~go/hlke/pytho...>

 lisasue (<http://blog.csdn.net/lisasue>) 2017年06月22日 14:57 362

机器学习sklearn库的使用--部署环境 (python2.7 windows7 64bit) (<http://blog.csdn.net/br...>)

最近在学习机器学习的内容, 难免地, 要用到Scikit-learn (sklearn, 下同) 这一机器学习包。为了使用sklearn库, 我们需要安装python2.7, pip install 工具, numpy...



bruce1993 (<http://blog.csdn.net/bruce1993>) 2017年07月03日 18:14 515

Python2.7+pycharm Win7 64bit安装教程 附:机器学习numpy+scipy+sklearn安装组 (<http://b...>)

博主 Win7 64bit机, 实装成功, 资源分享 一键打包相关软件合集下载, 链接: <http://pan.baidu.com/s/1nuPHsdr> 密码: e2k

U...



a593651986 (<http://blog.csdn.net/a593651986>) 2017年05月11日 17:26 998



Python机器学习库SKLearn的特征选择 (<http://blog.csdn.net/cheng9981/article/details/7102...>)



参考地址: http://scikit-learn.org/stable/modules/feature_selection.html#feature-selection sklearn.featur...



cheng9981 (<http://blog.csdn.net/cheng9981>) 2017年04月30日 17:10 1665