


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

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（4）



2017年05月19日 15:23:29

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)

408

彩蛋：数据的完整性测试

使用assert语句，可以快速进行数据测试。如果测试结果是True，notebook不会显示任何信息并继续向下执行，否则终止运行，并显示错误提示。

In [16]:

```
assert 1 == 2
```



weixin_3506...

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

(//my.csdn.

番番要吃肉 (ht

+ 关注

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

码云

未开通

原创
4

粉丝
4

喜欢
0

(https://gite
utm_sourc

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PM2.5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PM2.5空气质量分析（1）(http://blog.csdn.net/xiexf189/article/details/77368583)

立即体验



注意力测试



智能停车场

广告



内容举报



返回顶部

```
-----
AssertionError                                Traceback (most recent call last)
```

```
<ipython-input-16-a810b3a4aded> in <module>()
```

```
----> 1 assert 1 == 2
```

```
AssertionError:
```

```
In [17]:
```

```
# We know that we should only have three classes
assert len(iris_data_clean['class'].unique()) == 3
```

```
In [18]:
```

```
# We know that sepal lengths for 'Iris-versicolor' should never be below 2.5 cm
assert iris_data_clean.loc[iris_data_clean['class'] == 'Iris-versicolor', 'sepal_length_cm'].min() >= 2.5
```

```
In [19]:
```

```
# We know that our data set should have no missing measurements
assert len(iris_data_clean.loc[(iris_data_clean['sepal_length_cm'].isnull() |
                                (iris_data_clean['sepal_width_cm'].isnull() |
                                (iris_data_clean['petal_length_cm'].isnull() |
                                (iris_data_clean['petal_width_cm'].isnull()))]) == 0
```

就像这样的测试，如果不能通过测试，会终止程序并返回例外信息，我们必须回头继续对数据进行整理。

Step 4：数据的探索性分析

探索性分析，是在剔除了异常值和错误之后，对数据集更深入的分析和研究。在这一步里，我们试图回答这样几个问题：

数据是如何分布的？

数据之间是否存在相关性？

有什么混杂的因素，可以解释这种相关性？

n.net/xiexf189/article/details/7736750

4)

Python-sklearn 机器学习的
(7) (<http://blog.csdn.net/cle/details/72598976>)

Python-sklearn机器学习的
(6) (<http://blog.csdn.net/cle/details/72598910>)



相关推荐

android应用程序的签名(Signature) 签名
机制 (<http://blog.csdn.net/niepengpeng33/article/details/7064371>)

对android应用程序的理解 (http://blog.csdn.net/Amo_te_ama_me/article/details/51082561)

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

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



内容举报



返回顶部

在这个阶段，我们会采用各种方法绘数据图，当然不要太考虑美观，因为都是内部使用。先从散点图矩阵开始吧。

In [20]:

```
sb.pairplot(iris_data_clean)
```

Out[20]:

<seaborn.axisgrid.PairGrid at 0xb783f90>



0



他的热门文章

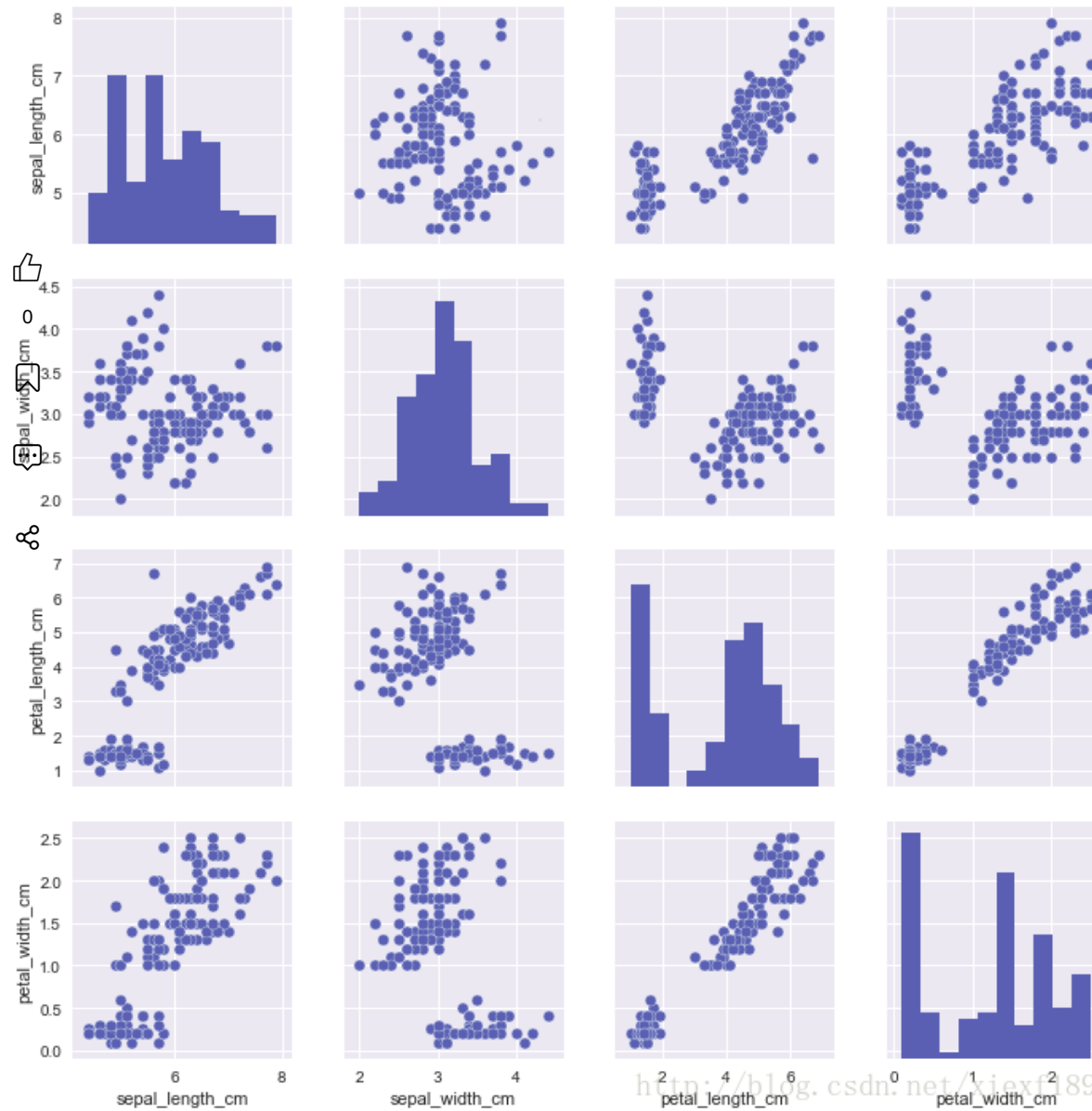
Python数据分析练习：北京、广州PM2.5 内容举报
 空气质量分析（1）(<http://blog.csdn.net/xiexf189/article/details/77367504>)

826



返回顶部

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



737

Python-sklearn机器学习的
(3) (<http://blog.csdn.net/e/details/72528755>)

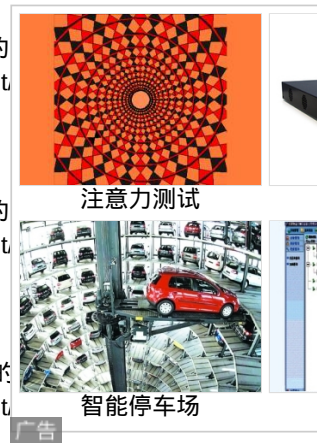
718

Python-sklearn机器学习的
(2) (<http://blog.csdn.net/e/details/72528667>)

589

Python-sklearn 机器学习的
(1) (<http://blog.csdn.net/e/details/72518860>)

497



内容举报

返回顶部

看起来我们的数据大致是正常分布。对于数据建模来说，数据分布正常是一个很棒的消息。

不过，我们看到有些花瓣的尺寸数据有一点奇怪，是不是由于花的不同种类导致的呢？我们可以再一次通过有色的散点图来观察一次。

In [21]:

```
sb.pairplot(iris_data_clean, hue='class')
```

Out[21]:

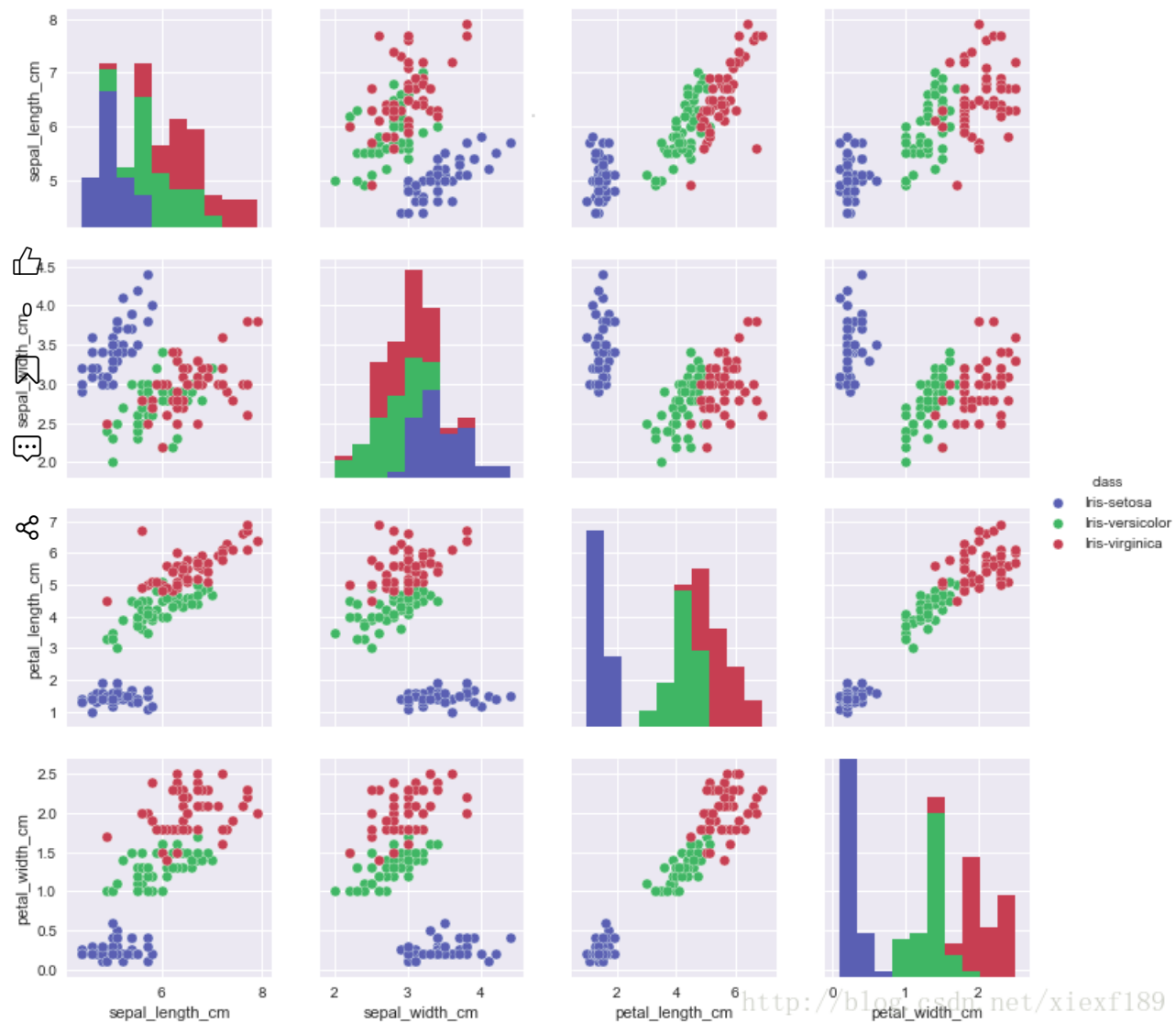
```
<seaborn.axisgrid.PairGrid at 0xc04a6b0>
```



内容举报



返回顶部



内容举报

返回顶部

果然，花瓣的数值分布，确实跟花的种类有关。这对我们分类的任务来说是个极好的消息，它意味着通过花瓣的尺寸，就可以容易地把iris-sentosa这个品种区分出来。

然而，区分iris-virginica和iris-versicolor这两个品种会困难一下，因为看起来它们的尺寸有些重叠。

花瓣的长度和宽度看起来有一些相关，花萼也有类似现象。我们通过咨生物学家，这是自然现象：较长的花瓣，宽度也会较大，花萼也类似。

我们还可以画出提琴图（violin plots），对比不同种类的数值分布。提琴图和箱型图所包含的信息是类似的，但它还能体现数据的密度。

In [23]:

```
plt.figure(figsize=(10, 10))

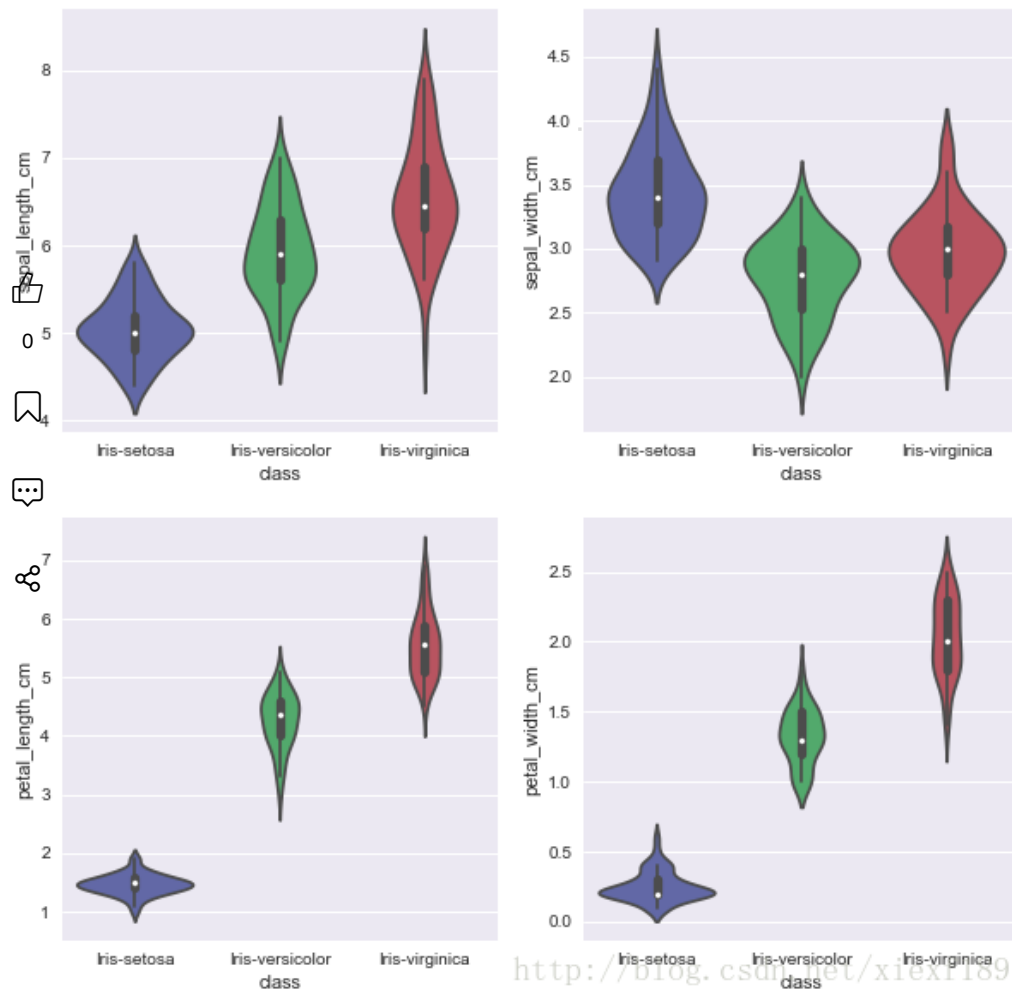
for column_index, column in enumerate(iris_data_clean.columns):
    if column == 'class':
        continue
    plt.subplot(2, 2, column_index + 1)
    sb.violinplot(x='class', y=column, data=iris_data_clean)
```



内容举报



返回顶部



数据摆弄得差不多了，我们进入建模部分吧。



内容举报



返回顶部



发表你的评论

http://my.csdn.net/weixin_35068028

相关文章推荐

android应用程序的签名(Signature) 签名机制 (<http://blog.csdn.net/niepengpeng333/article/...>)



摘自：<http://hi.baidu.com/%CE%D294%CB%FD/blog/item/95bfd49a5e6704186f068c72.html> 1. 为什么要签名 ...



niepengpeng333 (<http://blog.csdn.net/niepengpeng333>) 2011年12月12日 17:30 528



对android应用程序的理解 (http://blog.csdn.net/Amo_te_ama_me/article/details/51082561)

在判断一个应用程序是系统程序还是用户程序时，经常用到下面一端代码：int flags = packInfo.applicationInfo.flags;//应用程序信息的标记 ...



Amo_te_ama_me (http://blog.csdn.net/Amo_te_ama_me) 2016年04月07日 08:53 742



惊呆了！微博和阿里背后的数据库有多厉害？

想不到！数据库作为最关键的基础设施，渗透技术领域的方方面面，我阿里和微博的师哥们是这么分享的...

(http://www.baidu.com/cb.php?c=lgF_pyfqHmknjTzrjb0IZ0qnfK9ujYzP1nsrjD10Aw-5Hc3rHnYnHb0TAq15HfLPWRznjb0T1Y1uAwbm16vn1NbuhDsPhmk0AwY5HDdnHfzrHDvnjb0lgF_5y9YIZ0IQzq-uZR8mLPbUB48ugfEpZNGXy-jULNzTvRETvNzpyN1gvw-IA7GUatLPjqlAdxTvqdThP-5yF_UvTkn0KzujYk0AFV5H00TZcqN0KdpyfqHRLPjnvnfKEpyfqHc4rj6kP0KWpyfqP1cvrHnz0AqLUWYs0ZK45HcsP6KWThnqnWm3rjf)

Python-sklearn机器学习的第一个样例(2) (<http://blog.csdn.net/xiexf189/article/details/72...>)





内容举报






返回顶部

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 14:15  593



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  500

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

 7 4.60/米 铝合金线槽方型3020 外开型工厂直销可以定	 8 Hello! 厂家直销对讲机民用 10W大功率无线自驾	 9 4.80/米 铝合金线槽方型外开型 线槽2020。厂家直销
--	--	--

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  721





内容举报



返回顶部



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  337

用Python开始机器学习（4：KNN分类算法）sklearn做KNN算法 python (<http://blog.csdn.n...>)

1、KNN分类算法 KNN分类算法（K-Nearest-Neighbors Classification），又叫K近邻算法，是一个概念极其简单，而分类效果又很优秀的分类算法。他的核心思想就是，要确定...

 sherri_du (http://blog.csdn.net/sherri_du) 2016年08月03日 18:19  1676



机器学习（4）岭回归sklearn.linear_model.Ridge (<http://blog.csdn.net/voidfaceless/article...>)

sklearn.linear_model.Ridge class sklearn.linear_model.Ridge(alpha=1.0, fit_intercept=True, normalize=...

 voidfaceless (<http://blog.csdn.net/voidfaceless>) 2017年03月10日 15:31  1739

【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/...)

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  4741


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“...





内容举报


返回顶部

 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/article...>)

特征提取：sklearn.feature_extraction模块可以用于从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 4334

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

...
 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




内容举报


返回顶部