

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/Default.html) 学院 (//edu.csdn.net?ref=toolbar)

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

更多 ▾



0



weixin_3506...

(//write.blog.csdn.net/postedit/activ...
ref=toolbar)source=csdnblor

Python-sklearn机器学习的第一个样例（3）

翻译

2017年05月19日 14:23:00

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)

720

接上一篇

Step 3：数据整理

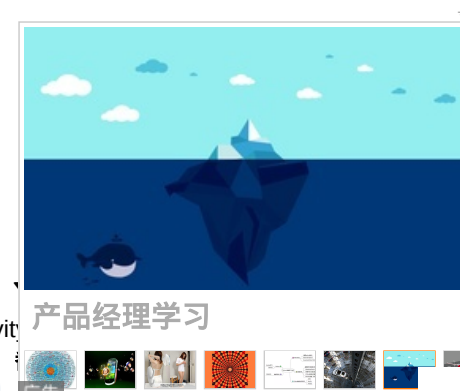
让我们逐个解决以上的问题：出现5个种类的问题，经过数据探索，发现是有部分类名忘记加上Iris-的前缀，另外有部分setossa是拼写错误。下面纠正这些错误：

In [6]:

```
iris_data.loc[iris_data['class'] == 'versicolor', 'class'] = 'Iris-versicolor'
iris_data.loc[iris_data['class'] == 'Iris-setossa', 'class'] = 'Iris-setosa'

iris_data['class'].unique()
```

Out[6]:



立即体

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

码云

未开通

(https://gite...
utm_sourc

原创

4

粉丝

4

喜欢

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/77367504)

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



内容举报



返回顶部

array(['Iris-setosa', 'Iris-versicolor', 'Iris-virginica'], dtype=object)

这样看起来好多了! 现在数据集中只有三种类型了。想象一下，如果我们创建的数据模型中，使用了错误的分类，那多尴尬。

下面解决第二个问题。

解决异常值需要一些技巧。通常我们很难弄清楚产生异常值的原因，到底是测量失误？用了错误的数据单位？确实发生了异常事件？我们通常不知道。因此，我们在处理异常的时候需要保持清醒。如果我们决定要排除任何数据，就应该有强有力的理由，并且在文档中注明。

对于一个异常Iris-setosa条目,经过咨询Iris-setosa领域研究人员知道，不可能有萼片宽度小于2.5厘米。显然这个条目是错误,我们很难知道造成错误的原因，最好的方法是删掉这条记录。

In [7]:

```
# This line drops any 'Iris-setosa' rows with a sepal width less than 2.5 cm
iris_data = iris_data.loc[(iris_data['class'] != 'Iris-setosa') | (iris_data['sepal_width_cm']
>= 2.5)]
iris_data.loc[iris_data['class'] == 'Iris-setosa', 'sepal_width_cm'].hist()
```

Out[7]:<matplotlib.axes._subplots.AxesSubplot at 0x9ca7ab0>

好了，现在所有的 Iris-setosa 花萼宽度都比2.5大。

下一个要解决“一些Iris-versicolor的花萼长度sepal_length_cm接近于零”的问题。让我们看看这些行:

In [8]:

```
iris_data.loc[(iris_data['class'] == 'Iris-versicolor') &
(iris_data['sepal_length_cm'] < 1.0)]
```

Out[8]:

	sepal_length_cm	sepal_width_cm	petal_length_cm	petal_width_cm	class
77	0.067	3.0	5.0	1.7	Iris-versicolor
78	0.060	2.9	4.5	1.5	Iris-versicolor
79	0.057	2.6	3.5	1.0	Iris-versicolor
80	0.055	2.4	3.8	1.1	Iris-versicolor
81	0.055	2.4	3.7	1.0	Iris-versicolor

看起来，这些接近0的数值，好像都比正常值低100倍，也许是小数点点错了？经过咨询研究人员，果然是有些人忘记把长度单位从“米”改为“厘米”。这就好办了：

Python-sklearn机器学习的第一个样例

(6) (http://blog.csdn.net/le/details/72528667)

相关推荐

python sklearn 分类算法简单调用 (http://blog.csdn.net/u010005)

sklearn画ROC曲线 (http://blog.csdn.net/eshao Liu/article/details/72528667)

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

python sklearn 分类算法简单调用 (http://blog.csdn.net/Bryan_/article/details/51288953)



Python机器学习 达内可靠吗

望京soho 移民澳大利亚 公司的网站..

OA办公系统 电脑硬件学习 大数据分..

人工智能课程 人脸识别算法

广告

内容举报

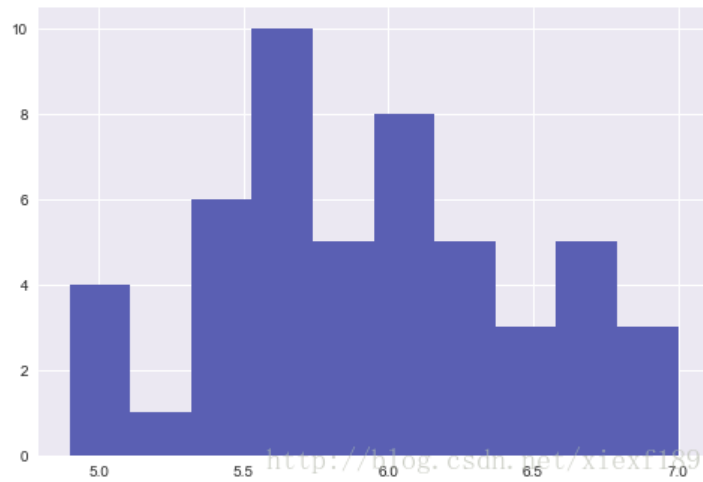
返回顶部

In [8]:

```
iris_data.loc[(iris_data['class'] == 'Iris-versicolor') &
              (iris_data['sepal_length_cm'] < 1.0),
              'sepal_length_cm'] *= 100.0

iris_data.loc[iris_data['class'] == 'Iris-versicolor', 'sepal_length_cm'].hist()
```

Out[8]:<matplotlib.axes._subplots.AxesSubplot at 0x9682450>



我们已经把异常值处理完毕！

下面解决第三个问题：有些行的数据有缺失。

In [10]:

```
iris_data.loc[(iris_data['sepal_length_cm'].isnull()) |
              (iris_data['sepal_width_cm'].isnull()) |
              (iris_data['petal_length_cm'].isnull()) |
              (iris_data['petal_width_cm'].isnull())]
```

Out[10]:

	sepal_length_cm	sepal_width_cm	petal_length_cm	petal_width_cm	class
7	5.0	3.4	1.5	NaN	Iris-setosa
8	4.4	2.9	1.4	NaN	Iris-setosa
9	4.9	3.1	1.5	NaN	Iris-setosa



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>)

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



内容举报



返回顶部

	sepal_length_cm	sepal_width_cm	petal_length_cm	petal_width_cm	class
10	5.4	3.7	1.5	NaN	Iris-setosa
11	4.8	3.4	1.6	NaN	Iris-setosa

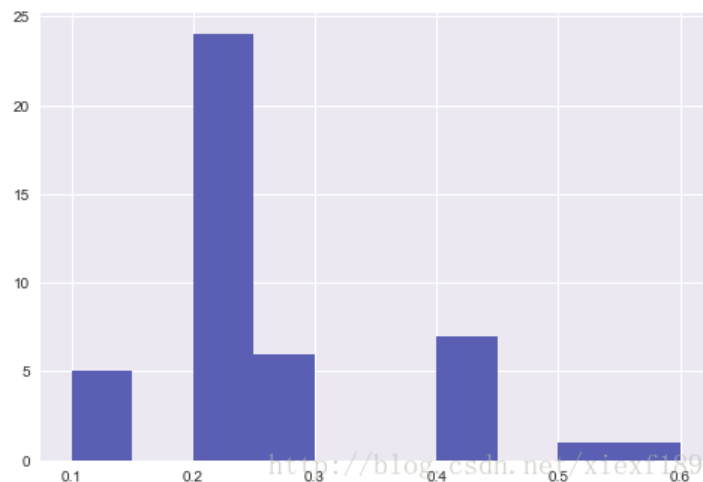
把这些数据全部丢弃，显然不是很理想，特别是它们都属于setosa类。丢弃这部分数据，容易对数据分析结果造成不利影响。

常用的处理数据丢失的方法是：“均值归一”法，就是用所有测量值的平均值来填补缺失值。

In [11]:

```
iris_data.loc[iris_data['class'] == 'Iris-setosa', 'petal_width_cm'].hist()
```

Out[11]:<matplotlib.axes._subplots.AxesSubplot at 0x9d18f30>



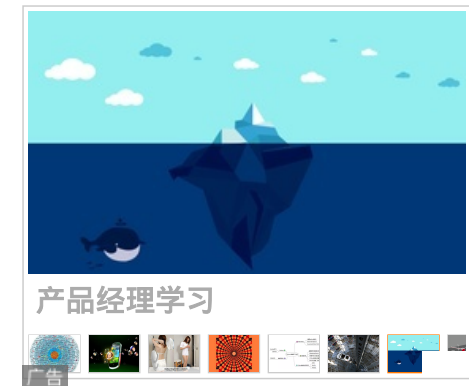
看得出来，大部分setosa类的花瓣宽度都在0.2-0.3cm之间，这就更有理由用平均值来填补缺失值了。

In [12]:

```
average_petal_width = iris_data.loc[iris_data['class'] == 'Iris-setosa', 'petal_width_cm'].mean()

iris_data.loc[(iris_data['class'] == 'Iris-setosa') &
              (iris_data['petal_width_cm'].isnull()),
              'petal_width_cm'] = average_petal_width

iris_data.loc[(iris_data['class'] == 'Iris-setosa') &
              (iris_data['petal_width_cm'] == average_petal_width)]
```



内容举报

返回顶部

Out[12]:

	sepal_length_cm	sepal_width_cm	petal_length_cm	petal_width_cm	class
7	5.0	3.4	1.5	0.25	Iris-setosa
8	4.4	2.9	1.4	0.25	Iris-setosa
9	4.9	3.1	1.5	0.25	Iris-setosa
10	5.4	3.7	1.5	0.25	Iris-setosa
11	4.8	3.4	1.6	0.25	Iris-setosa

In [13]:

```
iris_data.loc[(iris_data['sepal_length_cm'].isnull() |
               (iris_data['sepal_width_cm'].isnull() |
                (iris_data['petal_length_cm'].isnull() |
                 (iris_data['petal_width_cm'].isnull())))]
```

Out[13]:

非常棒，已经没有缺失值了。当然如果你不喜欢对数据进行这样的计算处理，把所有缺失值都去除，也是一种方法。

注意: 如果你对这样修补数据感到不爽，也可以把包含所有缺失数据的行删除，调用 `dropna()` 方法:

```
iris_data.dropna(inplace=True)
```

接下来，把清理后的数据保持到文件里，这样下次就不用再处理一次了。

In [14]:

```
iris_data.to_csv('iris-data-clean.csv', index=False)

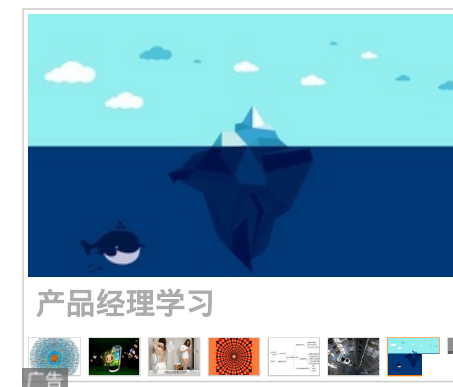
iris_data_clean = pd.read_csv('iris-data-clean.csv')
```

让我们再用散点图矩阵来看看经过清洗后的数据分布。

In [15]:

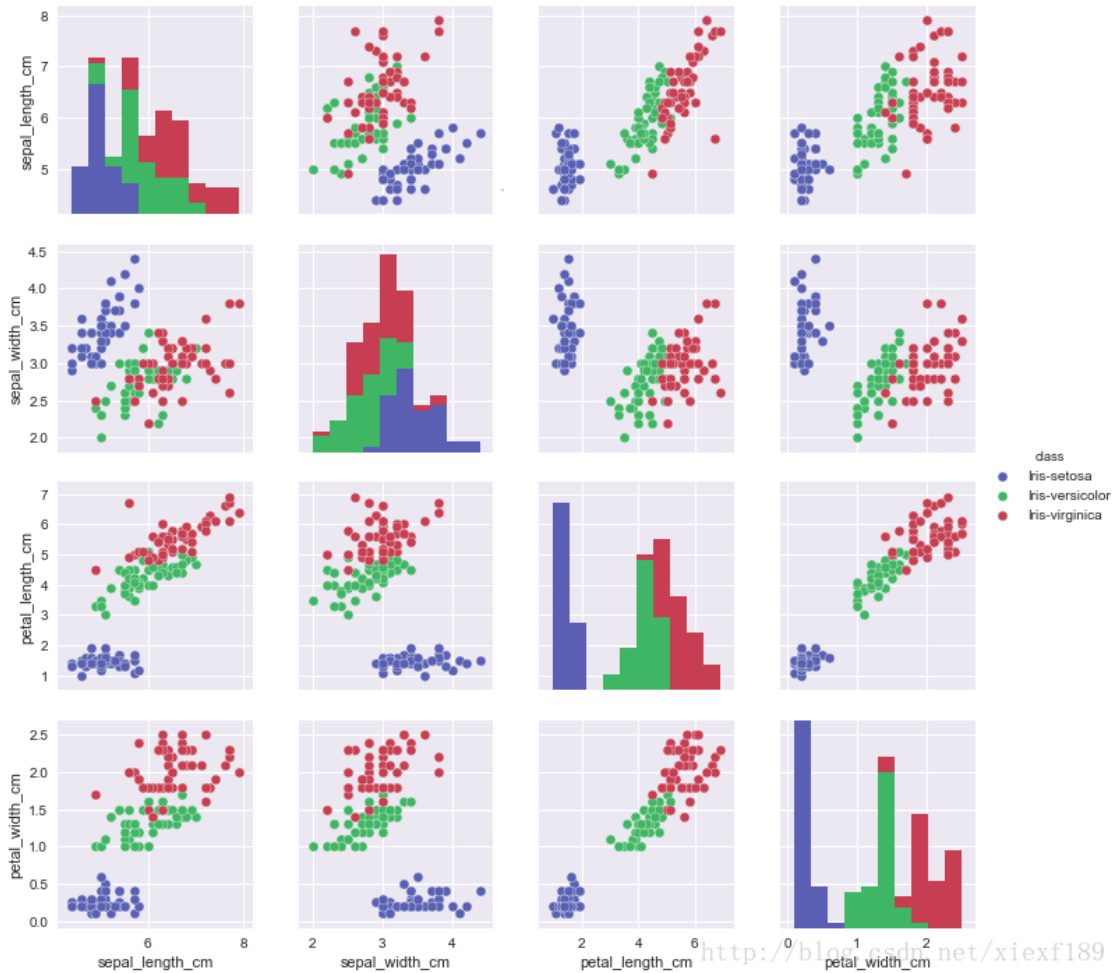
```
sb.pairplot(iris_data_clean, hue='class')
```

Out[15]:<seaborn.axisgrid.PairGrid at 0xa30f370>



内容举报

返回顶部



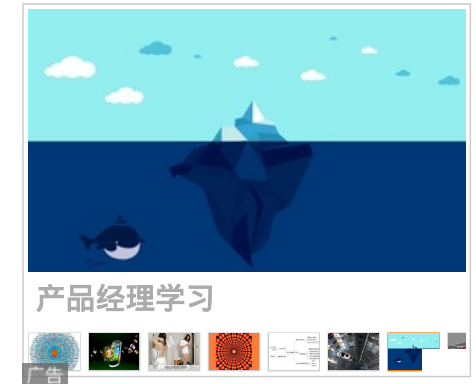
关于数据清洗的几个要点：

确保你的数据采用正确的方式编码

确保你的数值落在合理的区间，尽可能采用专业领域的知识，判断合理的数值区间范围

处理缺失数据的两个方法：替换或删除

不要手工整理数据，因为这很难重现



内容举报



返回顶部

使用程序代码，这样可以比较好地记录数据处理的过程

尽可能画图展现所有数据，可视化地方式确认数据正确性

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
```

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
```

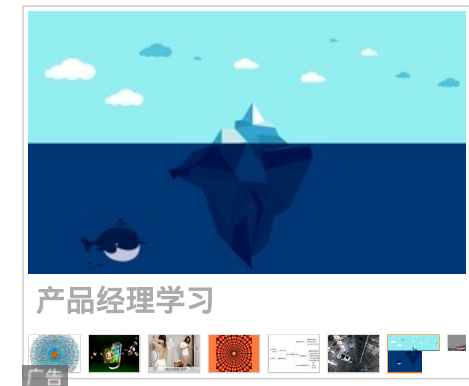
In [18]:

```
# 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。



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



0

**python sklearn画ROC曲线 (http://blog.csdn.net/u010454729/article/details/45098305)**

preface : 最近《生物信息学》多次谈到AUC, ROC这两个指标, 正在做的project, 要求画ROC曲线, sklearn里面有相应的函数, 故学习学习。 AUC: ROC: 具体使用参考sklear...



u010454729 (http://blog.csdn.net/u010454729) 2015年04月17日 16:11 17780

sklearn画R O C 曲线 (http://blog.csdn.net/eshaoliu/article/details/51187156)

```
#coding:utf-8 print(__doc__) import numpy as np from scipy import interp import matplotlib.pyplot...
```



eshaoliu (http://blog.csdn.net/eshaoliu) 2016年04月19日 06:52 2544

**太任性！学AI的应届学弟怒拒20K Offer，他想要多少钱？**

AI改变命运呀！！前段时间在我司联合举办的校招招聘会上，一名刚刚毕业的学弟陆续拒绝2份Offer，企业给出18K、23K高薪，学弟拒绝后直接来了一句...

(http://www.baidu.com/cb.php?c=lgF_pyfqHmknjnvPjn0IZ0qnfK9ujYzP1f4PjDs0Aw-

5Hc3rHnYnHb0TAq15HfLPWRznjb0T1YYnvD3mWu-uAc4uhu-

mWT40AwY5HDdnHfzrHDdnHD0lgF_5y9YIZ0IQzq-

uZR8mLPbUB48ugfEIAqspynEmybz5LNYUNq1ULNzmvRqmhkEu1Ds0ZFb5HD0mhYqn0KsTWYs0ZNGujYkPHTYn1mk0AqGujYknWb3rjDY0APGujYLnWm4n1c0ULI85H0QTZbqnWl

Python-sklearn机器学习的第一个样例(2) (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 14:15 592




内容举报



返回顶部


python sklearn 分类算法简单调用 (http://blog.csdn.net/Bryan___/article/details/51288953)

scikit-learn已经包含在Anaconda中。也可以在官方下载源码包进行安装。本文代码里封装了如下机器学习算法，我们修改数据加载函数，即可一键测试：[python] view p...

 Bryan___ (http://blog.csdn.net/Bryan___) 2016年05月01日 00:58 13120

sklearn学习代码 (<http://blog.csdn.net/kunlong0909/article/details/47956125>)


```
from sklearn.ensemble import RandomForestClassifier import pandas as pd from numpy import * import t...
```

 kunlong0909 (<http://blog.csdn.net/kunlong0909>) 2015年08月24日 19:54 2542

 <p>10 桥架线槽 2.00/只 桥架线槽电缆卡子，K09角钢电缆卡，K-09</p>	 <p>11 铝合金线槽 4.80/米 铝合金线槽方型外开型线槽2020。厂家直销</p>	 <p>12 机房布线桥架 135.00/米 【供应】JY-J101 400*125机房布线桥架</p>
--	--	---


将sklearn生成的决策树进行图形化展示 (<http://blog.csdn.net/u010736419/article/details/735...>)

1，工具和平台：python2.7 windows 2，决策树的可视化展示据我所知有两种途径：一是将生成的结果导出为pmmi文件，工具包为sklearn2pmmi等，具体可见<https://gi...>

 u010736419 (<http://blog.csdn.net/u010736419>) 2017年06月22日 11:42 2518

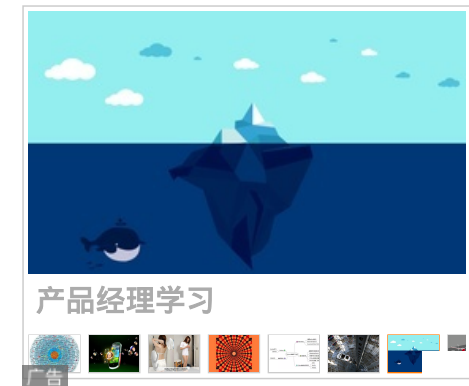
sklearn decision-tree实验 (<http://blog.csdn.net/u012420309/article/details/50389712>)

《机器学习技法》作业需要用到决策树。一直以为要用matlab自己实现一个，但今早起床的时候突然想起，不是有个叫sklearn的东西，是否可以直接拿来用呢？下午做了一些实验，作业里提供的数据，参考sk...

 u012420309 (<http://blog.csdn.net/u012420309>) 2015年12月23日 21:00 1907


机器学习逻辑回归模型总结——从原理到sklearn实践 (<http://blog.csdn.net/u011721501/artic...>)

0x00 基本原理逻辑回归算法，从名字上看似乎是个回归问题，但实际上逻辑回归是个典型的分类算法。对于分类问题，一般都是些离散变量，且y的取值如下： $y \in \{0, 1, 2, 3, \dots, n\}$ y \in \{0, 1, 2, 3, \dots, n\}





内容举报


返回顶部

 u011721501 (<http://blog.csdn.net/u011721501>) 2015年11月05日 13:31 7667

CART决策树的sklearn实现及其GraphViz可视化 (http://blog.csdn.net/chai_zheng/article/de...)

这一部分,我使用了sklearn来调用决策树模型对葡萄酒数据进行分类。除此之外,使用Python调用AT&T实验室开源的画图工具GraphViz软件以实现决策树的可视化。from sklearn.da...

 chai_zheng (http://blog.csdn.net/chai_zheng) 2017年10月13日 15:25 672


使用sklearn优雅地进行数据挖掘 (<http://blog.csdn.net/GoodShot/article/details/60588009>)

作者: jasonfreak 1 使用sklearn进行数据挖掘 1.1 数据挖掘的步骤 数据挖掘通常包括数据采集,数据分析,特征工程,训练模型,模型评估等步骤。使用sklearn工具...

 GoodShot (<http://blog.csdn.net/GoodShot>) 2017年03月06日 20:46 498


sklearn 的优雅数据挖掘流程 (<http://blog.csdn.net/ma416539432/article/details/53510277>)

我是大纲!!! 1) 下载数据集,通过统计方法理解数据集,并可视化。2) 构建6个机器学习模型。从中选择最好的。在下载并且安装好了所需的python包之后,我们来看一下各个包的版本。# Check ...

 ma416539432 (<http://blog.csdn.net/ma416539432>) 2016年12月07日 20:44 965

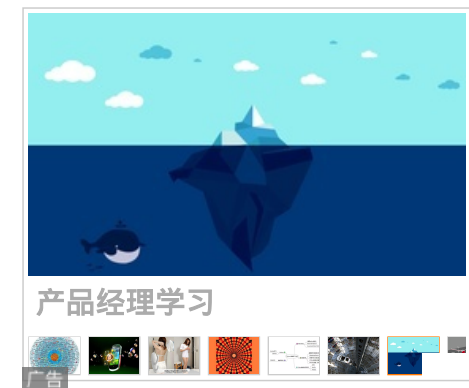
【机器学习】Python sklearn包的使用示例以及参数调优示例 (http://blog.csdn.net/wy_0928/...

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

python sklearn 机器学习库使用 (<http://blog.csdn.net/MakeRoomFor1/article/details/54315...>)

此代码包括了机器学习常用的算法的Python使用方法,随机选取训练集和测试集,需要安装numpy、sklearn库。包括: 'SVM', 'GaussianNB', 'BernoulliNB', 'L...




内容举报


返回顶部

MakeRoomFor1 (<http://blog.csdn.net/MakeRoomFor1>) 2017年01月10日 17:08 976

Python-sklearn机器学习的第一个样例（2）(<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 14:15 592

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

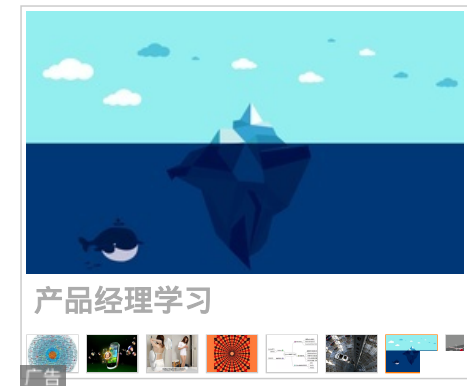
Python-sklearn 机器学习的第一个样例（7）(<http://blog.csdn.net/xiexf189/article/details/7...>)

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

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



内容举报



返回顶部

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

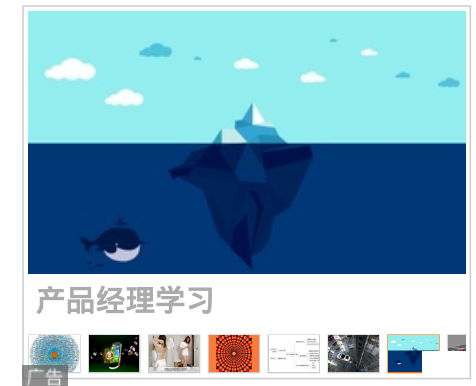
python3机器学习——sklearn0.19.1版本——数据处理（二）（多项式、pipeline、分类模型评...

一、数据变换——多项式 sklearn.preprocessing.PolynomialFeatures类实现多项式的数据转换。用于产生多项式，并且多项式包含的是相互影响的特征集。比如：一个输...

loveliuzz (<http://blog.csdn.net/loveliuzz>) 2017年11月21日 19:34 133

building machine learning system with Python 学习笔记--从零开始机器学习（3）第一个应...

这个小应用是根据已有的网站访问量来预测什么时候到达现有设施的极限，估计是每小时100000个请求。这种问题在初高中求函数极值时经常遇到，只是现在函数形式是未知的，只有一定量的离散数据。机器学习就派上...

qq_25203493 (http://blog.csdn.net/qq_25203493) 2017年05月07日 20:10 134

内容举报



返回顶部