

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 (<http://blog.csdn.net/>) 学院 (<http://edu.csdn.net/?ref=toolbar>)

下载 (<http://download.csdn.net/?ref=toolbar>) GitChat (<http://gitbook.cn/?ref=csdn>)

更多 ▾



0



weixin_3506...



(<http://write.blog.csdn.net/postedit?ref=toolbar>)

(<http://my.csdn.net/?ref=toolbar>)

番番要吃肉 (<http://blog.cs...>)



+ 关注

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

码云

未开通

(<https://github.com/xiexf189>)

原创

粉丝

喜欢

4

4

0

Python-sklearn机器学习的第一个样例（5）

翻译

2017年05月19日 20:27:50

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>)

381

Step 5：分类

虽然数据清理令人厌烦，但它却是数据分析的关键步骤。如果我们跳过这个阶段直接进入建模，会导致错误的数据模型。

记住：错误的数据导致错误的模型。永远要从检查数据开始。

现在我们已经尽可能地把数据清洗了，并且对数据集的分布和关系有了初步的认识。接下来的重要步骤就是把数据集分为：训练集和测试集。

训练集是数据集的一个随机子集，用于训练模型。

测试集也是数据集的一个随机子集（与训练集互斥分开），用于验证模型的准确性。

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>)

尤其对于我们这样一个比较稀疏的数据集来说，容易造成“过度拟合”，就是说对训练集的拟合度过高，以至于不能处理哪些没有见过的数据。这就是为什么我们要把数据集分开，用训练集建模，而用测试集评价模型。

需要注意的是，一旦我们把数据集划分为训练集和测试集，那么我们在建模的过程中，就不能再使用测试集的任何数据，否则就是作弊哦。

In [61]:

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

# We're using all four measurements as inputs
# 注意到 scikit-learn 要求所有的记录都要用“列表”list形式表示, e.g.,
# [ [val1, val2, val3],
#   [val1, val2, val3],
#   ... ]
# 所以, 我们要把输入数据集转化为一个列表的列表 (a list of lists)

# We can extract the data in this format from pandas like this:
all_inputs = iris_data_clean[['sepal_length_cm', 'sepal_width_cm',
                              'petal_length_cm', 'petal_width_cm']].values

# Similarly, we can extract the classes
all_classes = iris_data_clean['class'].values

# Make sure that you don't mix up the order of the entries
# all_inputs[5] inputs should correspond to the class in all_classes[5]

# Here's what a subset of our inputs looks like:
all_inputs[:5]
```

Out[61]:

```
array([[ 5.1,  3.5,  1.4,  0.2],
       [ 4.9,  3. ,  1.4,  0.2],
       [ 4.7,  3.2,  1.3,  0.2],
       [ 4.6,  3.1,  1.5,  0.2],
       [ 5. ,  3.6,  1.4,  0.2]])
```

下面要划分数据集

In [62]:

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

非参数估计：核密度估计KDE (<http://blog.csdn.net/pipisorry/article/details/53635895>)

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>)

Python-sklearn机器学习的第一个样例
(4) (<http://blog.csdn.net/xiexf189/article/details/72530805>)



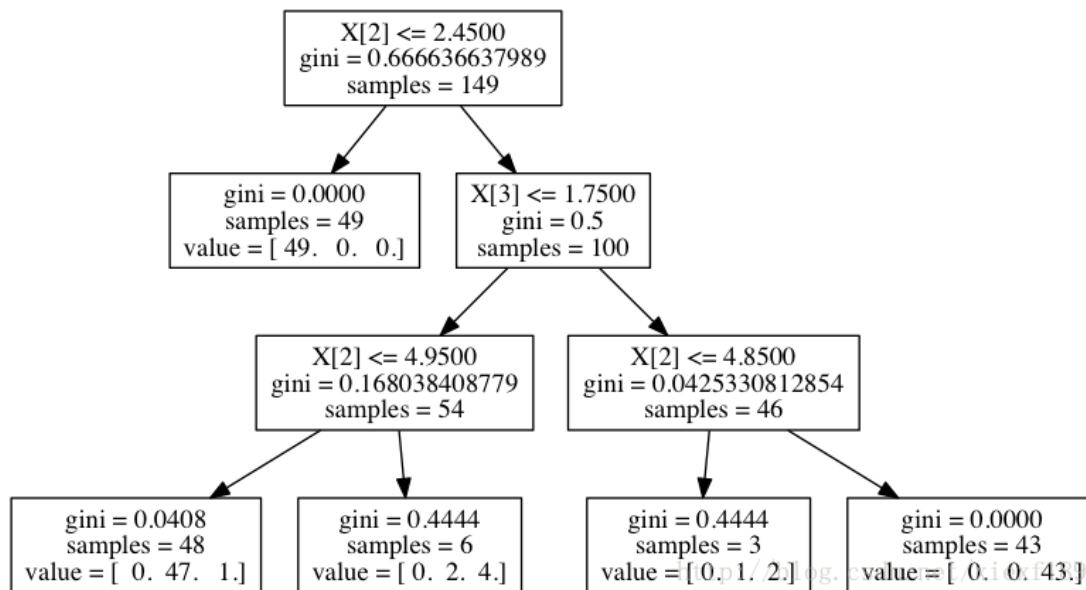
```
from sklearn.cross_validation import train_test_split

(training_inputs,
testing_inputs,
training_classes,
testing_classes) = train_test_split(all_inputs, all_classes, train_size=0.75, random_state=1)
```

下面用决策树的方法进行分类建模。

决策树的理论其实并不复杂，简单来说，决策树分类就是要求回答一系列的“Yes/No”，这样逐步划分出所有字段的分类。

下面是一个决策树分类器的例子：



决策树模型的一个有趣的特征是“尺度不变性（scale-invariant）”，就是说特征值的尺度规模，不会影响模型的表现。换句话说，一个特征值的范围是从0到1，还是从0到1000，决策树分类的处理方式是一样的。

对于决策树，我们可以使用许多参数进行调优，但目前我们用最基本的决策树模型。

In [63]:



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

```

from sklearn.tree import DecisionTreeClassifier

# Create the classifier
decision_tree_classifier = DecisionTreeClassifier()

# Train the classifier on the training set
decision_tree_classifier.fit(training_inputs, training_classes)

# Validate the classifier on the testing set using classification accuracy
decision_tree_classifier.score(testing_inputs, testing_classes)

```

Out[63]:

0.97368421052631582

哇，我们的分类模型准确率达到了97%，而且这么轻易就实现了哦。

可是，别高兴太早，这也许是瞎猫碰上死耗子。因为模型的准确率，是依赖训练集和测试集的取样不同，在大约80%到100%之间变化。

In [64]:

```

model_accuracies = []

for repetition in range(1000):
    (training_inputs,
     testing_inputs,
     training_classes,
     testing_classes) = train_test_split(all_inputs, all_classes, train_size=0.75)

    decision_tree_classifier = DecisionTreeClassifier()
    decision_tree_classifier.fit(training_inputs, training_classes)
    classifier_accuracy = decision_tree_classifier.score(testing_inputs, testing_classes)
    model_accuracies.append(classifier_accuracy)

sb.distplot(model_accuracies)

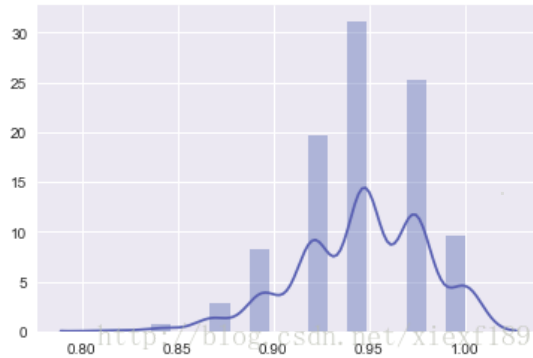
```

f:\Anaconda3\lib\site-packages\statsmodels\nonparametric\kdtools.py:20: VisibleDeprecationWarning: using a non-integer number instead of an integer will result in an error in the future

```
y = X[:m/2+1] + np.r_[0, X[m/2+1:], 0]*1j
```

Out[64]:

<matplotlib.axes._subplots.AxesSubplot at 0x2100be8278>



问题很明显，模型的表现与训练集的选择有很大关系。这种现象被称为“过度拟合”，模型针对训练集的分类表现太好，以至于对那些没有见过的数据则表现很差。



发表你的评论

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

相关文章推荐

非参数估计：核密度估计KDE (<http://blog.csdn.net/pipisorry/article/details/53635895>)

<http://blog.csdn.net/pipisorry/article/details/53635895>核密度估计Kernel Density Estimation(KDE)概述密度估计的问题由...



pipisorry (<http://blog.csdn.net/pipisorry>) 2016年12月14日 11:38 11417

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 16:14 338



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_pyfqHmknjnvPjn0lZ0qnfK9ujYzP1f4PjDs0Aw-5Hc3rHnYnHb0TAq15HfLPWRznjb0T1Y4PjD3PjNbPvm4uyfvrAmk0AwY5HDdnHfzrHDLnWR0lgF_5y9YIZ0lQzq-uZR8mLPbUB48ugfEIAqspynEmybz5LNYUNq1ULNzmvrQmhkEu1Ds0ZFb5HD0mhYqn0KsTWYs0ZNGujYkPHTYn1mk0AqGujYknWb3rjDY0APGujYLnWm4n1c0ULI85H00TZbqnW)

uZR8mLPbUB48ugfEIAqspynEmybz5LNYUNq1ULNzmvrQmhkEu1Ds0ZFb5HD0mhYqn0KsTWYs0ZNGujYkPHTYn1mk0AqGujYknWb3rjDY0APGujYLnWm4n1c0ULI85H00TZbqnW

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 16:06 743

Python-sklearn机器学习的第一个样例（4）(<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 15:23 408

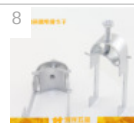
Python-sklearn机器学习的第一个样例（2）(<http://blog.csdn.net/xiexf189/article/details/72...>)

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 14:15 593



180.00/件
订做精图 网络箱 ONU
箱 综合箱 网络配线箱





2.00/只
桥架线槽电缆卡子，
K09角钢电缆卡，K-09



30.00/件
12芯 24芯 ST SC FC
机架式光纤盒，光缆终



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（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



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



【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  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 3734



0




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/cymy...>)

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/c...>)

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/78000000>)

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/76000000>)

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/hlike/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/bruce1993/article/details/76000000>)

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://blog.csdn.net/a593651986/article/details/7560725>)

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：e2ku...



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