

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 (7)

翻译

2017年05月21日 16:14:26

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)

338

我们终于有了自己的一个分类器。下面我们用图形看看它的表现。

In [37]:

```
dt_scores = cross_val_score(decision_tree_classifier, all_inputs, all_classes, cv=10)

sb.boxplot(dt_scores)
sb.stripplot(dt_scores, jitter=True, color='white')
```

Out[37]:

<matplotlib.axes._subplots.AxesSubplot at 0x113cd4b38>



weixin_3506...

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

source=csdnblog



番番要吃肉 (ht

+ 关注

(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



望京soho



未来三年房价

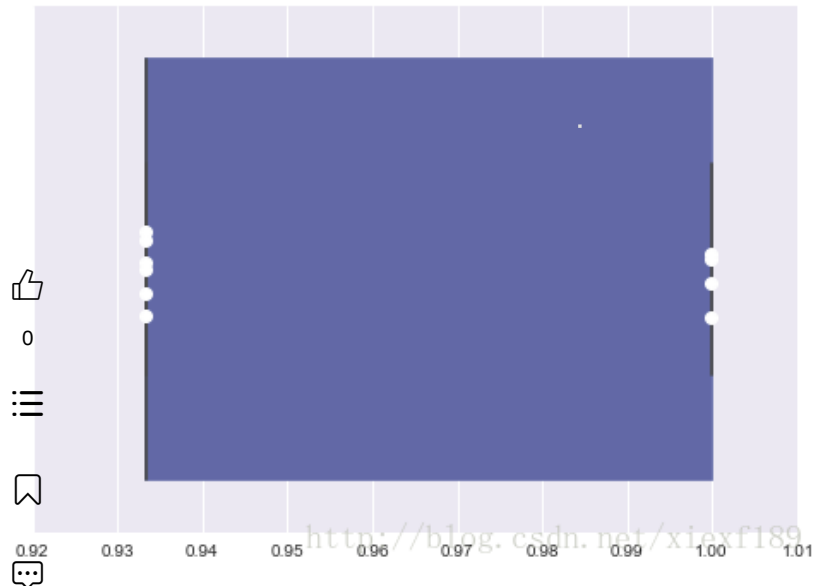
广告



内容举报



返回顶部



别急，还没结束。我们还应该用其他的分类算法（分类器）进行对比，看看这个决策树的表现如何。

下面我们使用“随机森林”分类器做一个对比。

我们已经知道，随机森林分类器通常比独立决策树的表现更好。决策树的通病就是过度拟合，它对训练集数据可以获得接近完美的分类，但它对于测试集，或者说对它没有见过的数据则可能表现不佳。

随机森林分类器的原理：创建一串决策树，每棵树的训练集是从总训练集里随机有放回抽取，而特征值是从所有特征里按比例无放回抽取。通过这一串决策树的共同工作，达到更高的分类精确度。

让我们来看看随机森林分类器是不是表现更好。

scikit-learn的妙处就在于：训练、测试、参数调优等过程对所有建模来说都是一样的，因此我们只需要选择新的分类器即可。

In [40]:

n.net/xiexf189/article/details/7736750

4)

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

Python-sklearn机器学习的
(5) (<http://blog.csdn.net/cle/details/72560725>)



相关推荐

PythonMachineLearning-Chap4.CodeExample (<http://blog.csdn.net/bojackhosreman/article/details/65633966>)

【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/article/details/62889012)

Python-sklearn机器学习的第一个样例
(5) (<http://blog.csdn.net/xiexf189/article/details/72560725>)

sklearn.model_selection.KFold (<http://blog.csdn.net/kancy110/article/details/74910185>)



内容举报



返回顶部

```

from sklearn.ensemble import RandomForestClassifier

random_forest_classifier = RandomForestClassifier()

parameter_grid = {'n_estimators': [5, 10, 25, 50],
                  'criterion': ['gini', 'entropy'],
                  'max_features': [1, 2, 3, 4],
                  'warm_start': [True, False]}

cross_validation = StratifiedKFold(all_classes, n_folds=10)

grid_search = GridSearchCV(random_forest_classifier,
                           param_grid=parameter_grid,
                           cv=cross_validation)

grid_search.fit(all_inputs, all_classes)
print('Best score: {}'.format(grid_search.best_score_))
print('Best parameters: {}'.format(grid_search.best_params_))

grid_search.best_estimator_

```

Best score: 0.9731543624161074

Best parameters: {'n_estimators': 5, 'max_features': 3, 'warm_start': True, 'criterion': 'gini'}

Out[40]:

```

RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
                       max_depth=None, max_features=3, max_leaf_nodes=None,
                       min_samples_leaf=1, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, n_estimators=5, n_jobs=1,
                       oob_score=False, random_state=None, verbose=0, warm_start=True)

```

下面可以对比两个分类器的表现：

In [42]:



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>)

```

random_forest_classifier = grid_search.best_estimator_

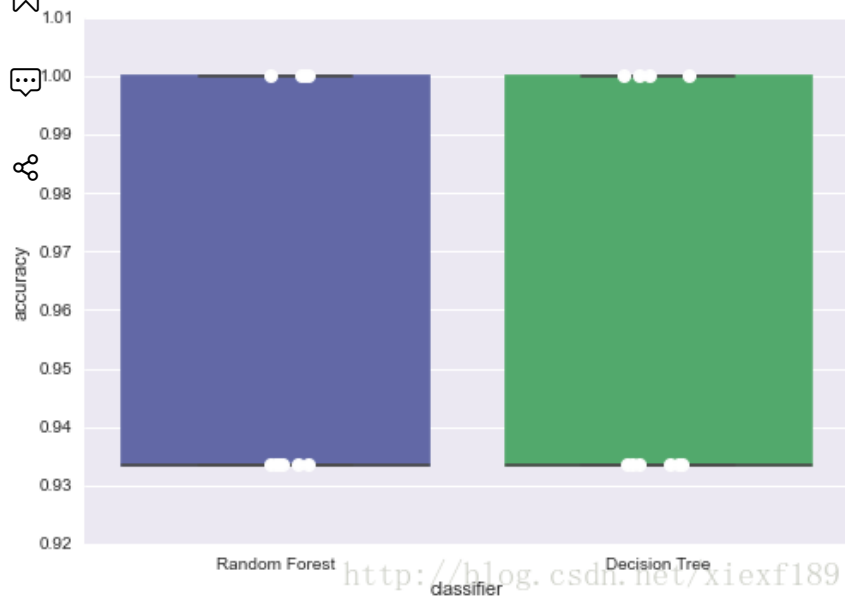
rf_df = pd.DataFrame({'accuracy': cross_val_score(random_forest_classifier, all_inputs, all_classes, cv=10),
                    'classifier': ['Random Forest'] * 10})
dt_df = pd.DataFrame({'accuracy': cross_val_score(decision_tree_classifier, all_inputs, all_classes, cv=10),
                    'classifier': ['Decision Tree'] * 10})
both_df = rf_df.append(dt_df)

sb.boxplot(x='classifier', y='accuracy', data=both_df)
sb.stripplot(x='classifier', y='accuracy', data=both_df, jitter=True, color='white')

```

Out[42]:

<matplotlib.axes._subplots.AxesSubplot at 0x1141bff28>



怎么样？看起来两者的表现差不多。这可能是我们的数据集只有4个特征值用于分类，而随机森林分类器在几百个可能特征值的情况下才能表现出优越性。换句话说，这个数据集没有太大的改进空间。

737

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

718

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

589

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

497



内容举报

返回顶部

Step 6 : 可重复性

确保我们的工作是可重复的，是任何分析的最后一步，也许是最重要的步骤。我们不能把太大的赌注压在一个我们不能重现的发现上。如果我们的分析不能重现，我们也许就根本不应该做这件事。

这个笔记完整记录了我们所做的每一个步骤，而且解释了为什么这么做。

In [43]:

```
%install_ext https://raw.githubusercontent.com/rasbt/watermark/master/watermark.py
```

Installed watermark.py. To use it, type:

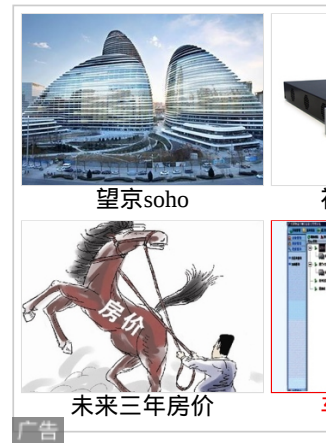
```
%load_ext watermark
```

In [44]:

```
%load_ext watermark
```

In [45]:

```
%watermark -a 'Randal S. Olson' -nmv --packages numpy,pandas,scikit-learn,matplotlib,Seaborn
```



内容举报



返回顶部

Randal S. Olson Fri Aug 21 2015

CPython 3.4.3

IPython 3.2.1

numpy 1.9.2

pandas 0.16.2

scikit-learn 0.16.1

matplotlib 1.4.3

Seaborn 0.6.0

⋮

compiler : GCC 4.2.1 (Apple Inc. build 5577)

system : Darwin

release : 14.5.0

machine : x86_64

processor : i386

CPU cores : 8

interpreter: 64bit

最后，我们把步骤1-5的核心部分，转化为一个独立的程序段：

In [46]:



内容举报



返回顶部

```

%matplotlib inline
import pandas as pd
import seaborn as sb
from sklearn.ensemble import RandomForestClassifier
from sklearn.cross_validation import train_test_split
from sklearn.cross_validation import cross_val_score

# We can jump directly to working with the clean data because we saved our cleaned data set
iris_data_clean = pd.read_csv('iris-data-clean.csv')

# Testing our data: Our analysis will stop here if any of these assertions are wrong
# We know that we should only have three classes
assert len(iris_data_clean['class'].unique()) == 3

# 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

# 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

all_inputs = iris_data_clean[['sepal_length_cm', 'sepal_width_cm',
                                'petal_length_cm', 'petal_width_cm']].values

all_classes = iris_data_clean['class'].values

# This is the classifier that came out of Grid Search
random_forest_classifier = RandomForestClassifier(bootstrap=True, class_weight=None,
criterion='gini',
                                                    max_depth=None, max_features=3, max_leaf_nodes=None,
                                                    min_samples_leaf=1, min_samples_split=2,
                                                    min_weight_fraction_leaf=0.0, n_estimators=5, n_jobs=1,
                                                    oob_score=False, random_state=None, verbose=0,
warm_start=True)

```



内容举报



返回顶部


```

# All that's left to do now is plot the cross-validation scores
rf_classifier_scores = cross_val_score(random_forest_classifier, all_inputs, all_classes, cv=10)
sb.boxplot(rf_classifier_scores)
sb.stripplot(rf_classifier_scores, jitter=True, color='white')

# ...and show some of the predictions from the classifier
(training_inputs,
testing_inputs,
training_classes,
testing_classes) = train_test_split(all_inputs, all_classes, train_size=0.75)

random_forest_classifier.fit(training_inputs, training_classes)

for input_features, prediction, actual in zip(testing_inputs[:10],
random_forest_classifier.predict(testing_inputs[:10]),
testing_classes[:10]):
    print('{}\t-->\t{}\t(Actual: {})'.format(input_features, prediction, actual))

```

```

[ 4.6  3.6  1.  0.2] --> Iris-setosa      (Actual: Iris-setosa)
[ 5.2  2.7  3.9  1.4] --> Iris-versicolor (Actual: Iris-versicolor)
[ 7.1  3.  5.9  2.1] --> Iris-virginica  (Actual: Iris-virginica)
[ 6.3  3.3  4.7  1.6] --> Iris-versicolor (Actual: Iris-versicolor)
[ 6.7  3.3  5.7  2.5] --> Iris-virginica  (Actual: Iris-virginica)
[ 6.9  3.1  5.4  2.1] --> Iris-virginica  (Actual: Iris-virginica)
[ 5.1  3.3  1.7  0.5] --> Iris-setosa      (Actual: Iris-setosa)
[ 6.3  2.8  5.1  1.5] --> Iris-versicolor (Actual: Iris-virginica)
[ 5.2  3.4  1.4  0.2] --> Iris-setosa      (Actual: Iris-setosa)
[ 6.1  2.6  5.6  1.4] --> Iris-virginica  (Actual: Iris-virginica)

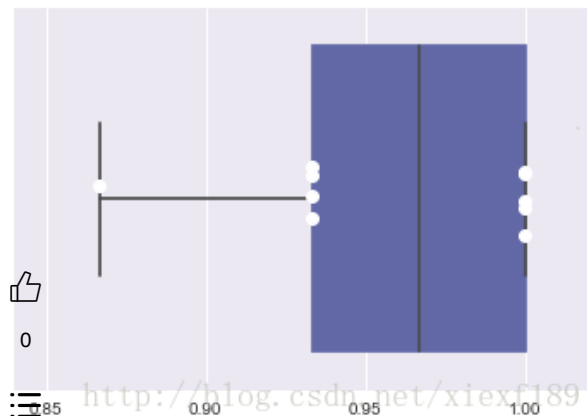
```



内容举报

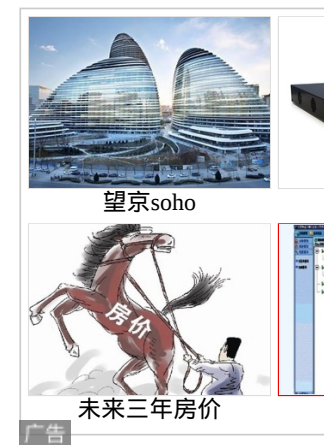


返回顶部



结束语：

针对本文开头的数据集，我们获得了一个完整的、可重复的机器学习演示程序段。我们已经达到了预定的标准：精确度>90%，并且，我们的程序有足够的适应性，可以处理任何新的输入数据。看起来不错吧！



发表你的评论

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

相关文章推荐

PythonMachineLearning-Chap4.CodeExample (<http://blog.csdn.net/bojackhosreman/article...>)


Python Machine Learning Essentials - Code ExamplesChapter 4



内容举报





返回顶部

 bojackhosreman (<http://blog.csdn.net/bojackhosreman>) 2017年03月24日 17:25  373

【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  4741



程序员想转管理有捷径吗？一位老前辈给我指了这条路！靠谱吗？



做程序员5年了收获蛮多，但是最近【中兴跳楼事件】发生后，我在想如果我到了40岁，会被辞退吗...



(http://www.baidu.com/cb.php?c=lgF_pyfqHmknjT3P160lZ0qnfK9ujYzP1nsrjDz0Aw-5Hc348nYnHb0TAq15HfLPWRznjb0T1YdrywBPACzmH04mWDSrjcz0AwY5HDdnHfzrHDLnjc0lgF_5y9YIZ0IQzqBTLn8mLPbUB48ugfEUiqYULKGmzq-uZNxug99UHqdlAdxTvqdThP-5yF_UvTkn0KzujYk0AFV5H00TZcqN0KdpyfqHRLPjnvnfKEpyfqHc4rj6kP0KWpyfqP1cvrHnz0AqLUWYs0ZK45HcsP6KWThnqrj0kPs)

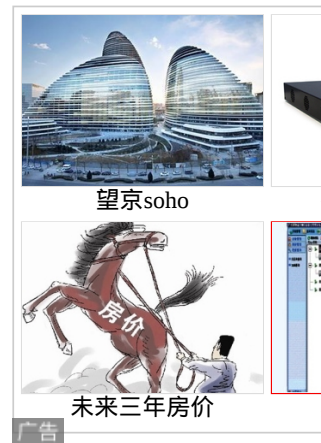
Python-sklearn机器学习的第一个样例 (5) (<http://blog.csdn.net/xiexf189/article/details/72...>)

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 20:27  380

sklearn.model_selection.KFold (<http://blog.csdn.net/kancy110/article/details/74910185>)


K折交叉验证：sklearn.model_selection.KFold(n_splits=3, shuffle=False, random_state=None) 思路：将训练/测试数据集划分n_s...



内容举报




返回顶部

 kancy110 (<http://blog.csdn.net/kancy110>) 2017年07月10日 10:57 2963

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



2.60/米
广电云村村通 电信级
室外铠装光缆 24芯 层




18.00/片
DD001 蓝牙LED灯带
模块 蓝牙4.0模块



1.00/米
光缆4芯6芯8芯12芯24
芯36芯48芯72芯96芯


【模式识别与机器学习】模式识别中的一些基本概念 (http://blog.csdn.net/Harry_lyc/article/d...)

1 特征(feature)：如果有一个区分鱼的类别的系统，可以分类的依据为长度、光泽、宽度、鳍的数目和形状、嘴的位置。这些可以利用的要素称为模式分类的特征。 2 模型(model)：如果鱼的不同类别...

 Harry_lyc (http://blog.csdn.net/Harry_lyc) 2012年07月02日 11:27 3548

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

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




内容举报





返回顶部

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


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：e2kU...

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

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


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/cymy00...>)

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




内容举报


返回顶部