

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



weixin\_3506...

(//my.csdn.net/)

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

ref=toolbar)source=csdnblog

番番要吃肉 (http://my.csdn.net/)



+ 关注

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

码云

未开通  
(https://gitee.com/xiexf189)

原创  
4

粉丝  
4

喜欢  
0

未开通  
(https://gitee.com/xiexf189)

## Python-sklearn机器学习的第一个样例（6）



2017年05月21日 16:06:27

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)

743

## 交叉检验（Cross-validation）

正是由于这个问题，大部分的数据科学家都会对数据模型进行“K层交叉检验（K-fold cross-validation）”：把原始的数据集划分为K个子集，使用其中一个子集作为测试集，其他子集都用作训练集。这个过程重复K次，这样每个子集都会成为一次测试集。

10层交叉验证是最常用的。



### 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)



内容举报



返回顶部

In [65]:

```
import numpy as np
from sklearn.cross_validation import StratifiedKFold
```

```
def plot_cv(cv, n_samples):
```

```
    masks = []
```

```
    for train, test in cv:
```

```
        mask = np.zeros(n_samples, dtype=bool)
```

```
        mask[test] = 1
```

```
        masks.append(mask)
```

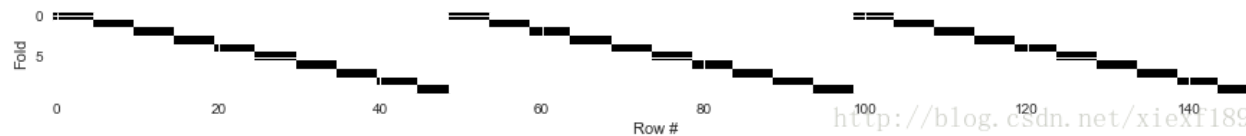
```
    plt.figure(figsize=(15, 15))
```

```
    plt.imshow(masks, interpolation='none')
```

```
    plt.ylabel('Fold')
```

```
    plt.xlabel('Row #')
```

```
plot_cv(StratifiedKFold(all_classes, n_folds=10), len(all_classes))
```



你已经注意到，在以上的代码中我们使用了分层的K次交叉验证。这种分层的K次交叉验证，可以保证在每一次验证中的，每一类中的数据量比例一致，这样才能保证数据子集的代表性。毕竟我们不可能在每一个子集中，包含某个类别的所有记录。

In [66]:

```
from sklearn.cross_validation import cross_val_score
```

```
decision_tree_classifier = DecisionTreeClassifier()
```

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  
(5) (<http://blog.csdn.net/cle/details/72560725>)



## 相关推荐

Python实现HMM（隐马尔可夫模型）([http://blog.csdn.net/sinat\\_36005594/article/details/69568538](http://blog.csdn.net/sinat_36005594/article/details/69568538))

Faster-RCNN训练问题解决：GPU内存 ([http://blog.csdn.net/forest\\_world/article/details/78151803](http://blog.csdn.net/forest_world/article/details/78151803))

时间序列分析 (<http://blog.csdn.net/pipisorry/article/details/62053938>)

python实现的四种抽样方法 (<http://blog.csdn.net/wang1127248268/article/details/53576325>)



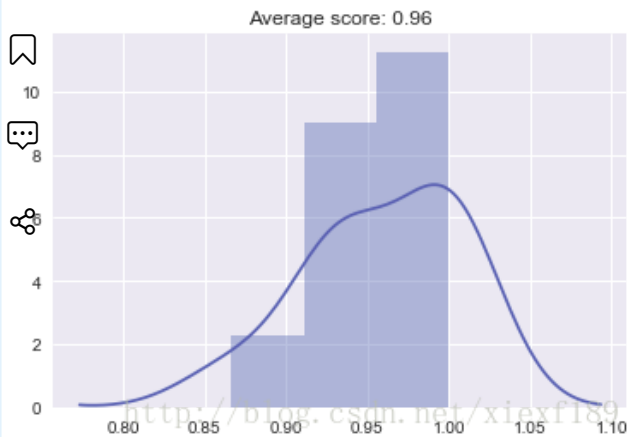
内容举报



返回顶部

```
# cross_val_score returns a list of the scores, which we can visualize
# to get a reasonable estimate of our classifier's performance
cv_scores = cross_val_score(decision_tree_classifier, all_inputs, all_classes, cv=10)
sb.distplot(cv_scores)

plt.title('Average score: {}'.format(np.mean(cv_scores)))
f:\Anaconda3\lib\site-packages\statsmodels\nonparametric\kdetools.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[66]:
matplotlib.text.Text at 0x217ef54860>
```



现在这个分类器要好多了，相对来说有了更一致的分类准确性。

## 参数调优

每个机器学习的模型都伴随着大量的参数调优，这些参数对模型的表现至关重要。例如，我们是否严格限制决策树的深度。

In [67]:



### 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

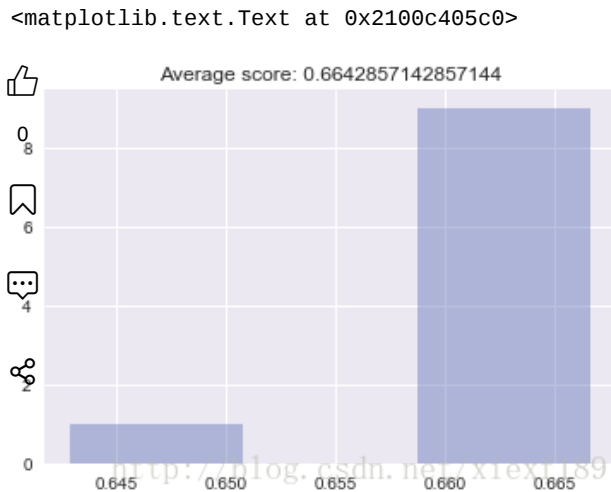
内容举报  
TOP  
返回顶部

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

```
decision_tree_classifier = DecisionTreeClassifier(max_depth=1)

cv_scores = cross_val_score(decision_tree_classifier, all_inputs, all_classes, cv=10)
sb.distplot(cv_scores, kde=False)
```

Out[67]:



把最大深度限制为1，分类器的精确度当然非常差。

因此，我们应该找到一个系统性的方法，探寻模型和数据集的最佳参数。

最通常的模型参数调优方法是：网格搜索（Grid Search）。原理其实很简单：探测整个范围内的参数，寻找表现最佳的参数组合。

下面开始对我们的决策树分类器进行调优。这里主要聚焦两个参数，实际应用中，可能需要面对多个参数的调优。

In [68]:

```
from sklearn.grid_search import GridSearchCV

decision_tree_classifier = DecisionTreeClassifier()
```

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

```

parameter_grid = {'max_depth': [1, 2, 3, 4, 5],
                  'max_features': [1, 2, 3, 4]}

cross_validation = StratifiedKFold(all_classes, n_folds=10)

grid_search = GridSearchCV(decision_tree_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_))
Best score: 0.959731543624161
Best parameters: {'max_depth': 3, 'max_features': 3}

```

现在，让我们用图形的方式，来看看网格搜索的参数关系。

In[32]:

```

grid_visualization = []

for grid_pair in grid_search.grid_scores_:
    grid_visualization.append(grid_pair.mean_validation_score)

grid_visualization = np.array(grid_visualization)
grid_visualization.shape = (5, 4)
sb.heatmap(grid_visualization, cmap='Blues')
plt.xticks(np.arange(4) + 0.5, grid_search.param_grid['max_features'])
plt.yticks(np.arange(5) + 0.5, grid_search.param_grid['max_depth'][:-1])
plt.xlabel('max_features')

```

Out[32]:

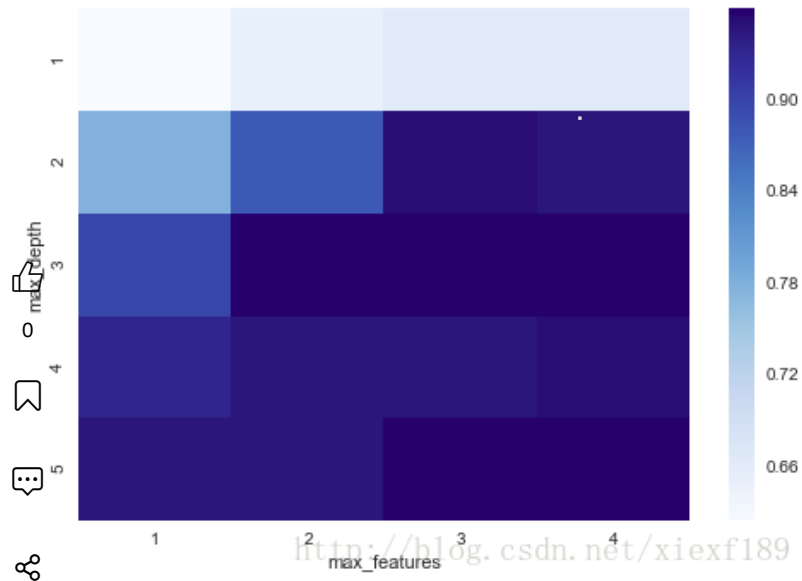
<matplotlib.text.Text at 0x217ae4f978>



内容举报



返回顶部



现在我们对这个模型的参数有了更好的感觉：决策树的最大深度max\_depth至少是2，而不是做一次性的决定。

max\_features 参数对模型的影响似乎不大，只要有2个就够了。考虑到我们的数据集只有4个参数，而且相对来说比较容易分类。

让我们继续使用一个更宽泛的网格搜索，寻找一个最佳的参数组合。

In [33]:

```
decision_tree_classifier = DecisionTreeClassifier()

parameter_grid = {'criterion': ['gini', 'entropy'],
                  'splitter': ['best', 'random'],
                  'max_depth': [1, 2, 3, 4, 5],
                  'max_features': [1, 2, 3, 4]}

cross_validation = StratifiedKFold(all_classes, n_folds=10)
```



内容举报



返回顶部

```
grid_search = GridSearchCV(decision_tree_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_))  
Best score: 0.9664429530201343  
Best parameters: {'criterion': 'gini', 'max_depth': 3, 'max_features': 3, 'splitter': 'best'}
```

现在我们可以说通过网格搜索，找到了一个最佳的分类器：

In[35]:

```
decision_tree_classifier = grid_search.best_estimator_  
decision_tree_classifier
```

Out[35]:

```
DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=3,  
                       max_features=3, max_leaf_nodes=None, min_impurity_split=1e-07,  
                       min_samples_leaf=1, min_samples_split=2,  
                       min_weight_fraction_leaf=0.0, presort=False, random_state=None,  
                       splitter='best')
```



发表你的评论

([http://my.csdn.net/weixin\\_35068028](http://my.csdn.net/weixin_35068028))

## 相关文章推荐



内容举报





返回顶部



## Python实现HMM（隐马尔可夫模型）([http://blog.csdn.net/sinat\\_36005594/article/details/6...](http://blog.csdn.net/sinat_36005594/article/details/6...))



前几天用MATLAB实现了HMM的代码，这次用python写了一遍，依据仍然是李航博士的《统计学习方法》由于第一次用python，所以代码可能会有许多缺陷，但是所有代码都用书中的例题进行了测试，结果...

 sinat\_36005594 ([http://blog.csdn.net/sinat\\_36005594](http://blog.csdn.net/sinat_36005594)) 2017年04月07日 16:13  2551



## Faster-RCNN训练问题解决：GPU内存 ([http://blog.csdn.net/forest\\_world/article/details/78...](http://blog.csdn.net/forest_world/article/details/78...))

l1002 16:29:32.222652 27395 layer\_factory.hpp:77] Creating layer bbox\_pred l1002 16:29:32.222658 273...

 forest\_world ([http://blog.csdn.net/forest\\_world](http://blog.csdn.net/forest_world)) 2017年10月02日 17:40  362



### 【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-5Hc3rHnYnHb0TAq15HfLPWRznjb0T1YsPWRsuWTLuAFhPjNhnWK-0AwY5HDdnHfzrHDvP1f0lgF\\_5y9YIZ0lQzq-uZR8mLPbUB48ugfElAqspynElvNBnHqdlAdxTvqdThP-5yF\\_UvTkn0KzujYk0AFV5H00TZcq0KdpYfqHRLPjnvnfKEpyfqHc4rj6kP0KWpyfqP1cvrHnz0AqLUWYs0ZK45HcsP6KWThnqPH01Pj6](http://www.baidu.com/cb.php?c=lgF_pyfqHmknj0dP1f0lZ0qnfK9ujYzP1ndPWb10Aw-5Hc3rHnYnHb0TAq15HfLPWRznjb0T1YsPWRsuWTLuAFhPjNhnWK-0AwY5HDdnHfzrHDvP1f0lgF_5y9YIZ0lQzq-uZR8mLPbUB48ugfElAqspynElvNBnHqdlAdxTvqdThP-5yF_UvTkn0KzujYk0AFV5H00TZcq0KdpYfqHRLPjnvnfKEpyfqHc4rj6kP0KWpyfqP1cvrHnz0AqLUWYs0ZK45HcsP6KWThnqPH01Pj6))

## 时间序列分析 (<http://blog.csdn.net/pipisorry/article/details/62053938>)

<http://blog.csdn.net/pipisorry/article/details/62053938> 时间序列简介时间序列是时间间隔不变的情况下收集的时间点集合。这些集合被分析用来了解长期发展...

 pipisorry (<http://blog.csdn.net/pipisorry>) 2017年03月22日 17:04  3653



内容举报




返回顶部




## python实现的四种抽样方法 (<http://blog.csdn.net/wang1127248268/article/details/53576325>)

一、单纯随机抽样（simple random sampling）将调查总体全部观察单位编号，再用抽签法或随机数字表随机抽取部分观察单位组成样本。优点：操作简单，均数、率及相应的标准误计算简单。...

 wang1127248268 (<http://blog.csdn.net/wang1127248268>) 2016年12月11日 22:36 7204

## python-Pandas学习 如何对数据集随机抽样？ ([http://blog.csdn.net/qq\\_22238533/article/det...](http://blog.csdn.net/qq_22238533/article/det...))

摘要：有时候我们只需要数据集中的一部分，并不需要全部的数据。这个时候我们就要对数据集进行随机的抽样。pandas中自带有的抽样的方法。应用场景：我有10W行数据，每一行都11列的属性。现在，我...

 qq\_22238533 ([http://blog.csdn.net/qq\\_22238533](http://blog.csdn.net/qq_22238533)) 2017年05月02日 14:25 10008



it培训机构排名



超强注意力



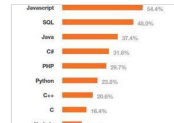
舆情监测系统



人脸识别



未来三年房价




嵌入式工程师



Python机器学习


## 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...>)

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




内容举报



返回顶部

### 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

0


### 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


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




内容举报



返回顶部


## 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

## 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...



内容举报




返回顶部

 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/77992416>)

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/72598910>)

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



内容举报



返回顶部