

刘建平Pinard

十年码农，对数学统计学，数据挖掘，机器学习，大数据平台，大数据平台应用开发，大数据可视化感兴趣。

博客园 首页 新随笔 联系 订阅 管理

scikit-learn决策树算法类库使用小结

之前对决策树的算法原理做了总结，包括决策树算法原理(上)和决策树算法原理(下)。今天就从实践的角度来介绍决策树算法，主要是讲解使用scikit-learn来跑决策树算法，结果的可视化以及一些参数调参的关键点。

1. scikit-learn决策树算法类库介绍

scikit-learn决策树算法类库内部实现是使用了调优过的CART树算法，既可以做分类，又可以做回归。分类决策树的类对应的是DecisionTreeClassifier，而回归决策树的类对应的是DecisionTreeRegressor。两者的参数定义几乎完全相同，但是意义不全相同。下面就对DecisionTreeClassifier和DecisionTreeRegressor的重要参数做一个总结，重点比较两者参数使用的不同点和调参的注意点。

2. DecisionTreeClassifier和DecisionTreeRegressor重要参数调参注意点

为了便于比较，这里我们用表格的形式对DecisionTreeClassifier和DecisionTreeRegressor重要参数要点做一个比较。

参数	DecisionTreeClassifier	DecisionTreeRegressor
特征选择标准 criterion	可以使用"gini"或者"entropy"，前者代表基尼系数，后者代表信息增益。一般说使用默认的基尼系数"gini"就可以了，即CART算法。除非你更喜欢类似ID3, C4.5的最优特征选择方法。	可以使用"mse"或者"mae"，前者是均方差，后者是和均值之差的绝对值之和。推荐使用默认的"mse"。一般来说"mse"比"mae"更加精确。除非你想比较二个参数的效果的不同之处。

公告

★珠江追梦，饮岭南茶，恋鄂北家★  
昵称：刘建平Pinard  
园龄：1年5个月  
粉丝：1180  
关注：13  
+加关注

2018年4月						
日	一	二	三	四	五	六
25	26	27	28	29	30	31
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	1	2	3	4	5

常用链接

我的随笔  
我的评论  
我的参与

特征划分点选择标准splitter	<p>可以使用"best"或者"random"。前者在特征的所有划分点中找出最优的划分点。后者是随机的在部分划分点中找局部最优的划分点。</p> <p>默认的"best"适合样本量不大的时候，而如果样本数据量非常大，此时决策树构建推荐"random"</p>	最新评论
划分时考虑的最大特征数 max_features	<p>可以使用很多种类型的值，默认是"None",意味着划分时考虑所有的特征数；如果是"log2"意味着划分时最多考虑<math>\log_2 N</math>个特征；如果是"sqrt"或者"auto"意味着划分时最多考虑<math>\sqrt{N}</math>个特征。如果是整数，代表考虑的特征绝对数。如果是浮点数，代表考虑特征百分比，即考虑（百分比xN）取整后的特征数。其中N为样本总特征数。</p> <p>一般来说，如果样本特征数不多，比如小于50，我们用默认的"None"就可以了，如果特征数非常多，我们可以灵活使用刚才描述的其他取值来控制划分时考虑的最大特征数，以控制决策树的生成时间。</p>	我的标签
决策树最大深度 max_depth	<p>决策树的最大深度，默认可以不输入，如果不输入的话，决策树在建立子树的时候不会限制子树的深度。一般来说，数据少或者特征少的时候可以不管这个值。如果模型样本量多，特征也多的情况下，推荐限制这个最大深度，具体的取值取决于数据的分布。常用的可以取值10-100之间。</p>	随笔分类(101)
内部节点再划分所需最小样本数 min_samples_split	<p>这个值限制了子树继续划分的条件，如果某节点的样本数少于min_samples_split，则不会继续再尝试选择最优特征来进行划分。默认是2.如果样本量不大，不需要管这个值。如果样本量数量级非常大，则推荐增大这个值。我之前的一个项目例子，有大概10万样本，建立决策树时，我选择了min_samples_split=10。可以作为参考。</p>	0040. 数学统计学(4) 0081. 机器学习(62) 0082. 深度学习(10) 0083. 自然语言处理(23) 0121. 大数据挖掘(1) 0122. 大数据平台(1) 0123. 大数据可视化
叶子节点最少样本数 min_samples_leaf	<p>这个值限制了叶子节点最少的样本数，如果某叶子节点数目小于样本数，则会和兄弟节点一起被剪枝。默认是1，可以输入最少的样本数的整数，或者最少样本数占样本总数的百分比。如果样本量不大，不需要管这个值。如果样本量数量级非常大，则推荐增大这个值。之前的10万样本项目使用min_samples_leaf的值为5，仅供参考。</p>	随笔档案(101)
叶子节点最小的样本权重和 min_weight	<p>这个值限制了叶子节点所有样本权重和的最小值，如果小于这个值，则会和兄弟节点一起被剪枝。默认是0，就是不考虑权重问题。一般来说，如果我们有较多样本有缺失值，或者分类树样本的分布类别偏差很大，就会引入样本权重，这时我们就要注意这个值了。</p>	2017年8月 (1) 2017年7月 (3) 2017年6月 (8) 2017年5月 (7) 2017年4月 (5) 2017年3月 (10) 2017年2月 (7) 2017年1月 (13) 2016年12月 (17) 2016年11月 (22) 2016年10月 (8)
		常去的机器学习网站
		52 NLP Analytics Vidhya 机器学习库 机器学习路线图

ht_fraction_leaf		
最大叶子节点数 max_leaf_nodes	通过限制最大叶子节点数，可以防止过拟合，默认是“None”，即不限制最大的叶子节点数。如果加了限制，算法会建立在最大叶子节点数内最优的决策树。如果特征不多，可以不考虑这个值，但是如果特征分成多的话，可以加以限制，具体的值可以通过交叉验证得到。	
类别权重 class_weight	指定样本各类别的权重，主要是为了防止训练集某些类别的样本过多，导致训练的决策树过于偏向这些类别。这里可以自己指定各个样本的权重，或者用“balanced”，如果使用“balanced”，则算法会自己计算权重，样本量少的类别所对应的样本权重会高。当然，如果你的样本类别分布没有明显的偏倚，则可以不管这个参数，选择默认的“None”	不适用于回归树
节点划分最小不纯度 min_impurity_split	这个值限制了决策树的增长，如果某节点的不纯度(基尼系数，信息增益，均方差，绝对差)小于这个阈值，则该节点不再生成子节点。即为叶子节点。	
数据是否预排序 presort	这个值是布尔值，默认是False不排序。一般来说，如果样本量少或者限制了一个深度很小的决策树，设置为true可以让划分点选择更加快，决策树建立的更加快。如果样本量太大的话，反而没有什么好处。问题是样本量少的时候，我速度本来就不慢。所以这个值一般懒得理它就可以了。	

除了这些参数要注意以外，其他在调参时的注意点有：

- 1) 当样本少数量但是样本特征非常多的时候，决策树很容易过拟合，一般来说，样本数比特征数多一些会比较容易建立健壮模型
- 2) 如果样本数量少但是样本特征非常多，在拟合决策树模型前，推荐先做维度规约，比如主成分分析（PCA），特征选择（Lasso）或者独立成分分析（ICA）。这样特征的维度会大大减小。再来拟合决策树模型效果会好。
- 3) 推荐多用决策树的可视化（下节会讲），同时先限制决策树的深度（比如最多3层），这样可以先观察下生成的决策树里数据的初步拟合情况，然后再决定是否要增加深度。
- 4) 在训练模型先，注意观察样本的类别情况（主要指分类树），如果类别分布非常不均匀，就要考虑用class\_weight来限制模型过于偏向样本多的类别。

深度学习进阶书  
深度学习入门书

积分与排名

积分 - 303959  
排名 - 604

阅读排行榜

1. 梯度下降（Gradient Descent）小结(106306)
2. 梯度提升树(GBDT)原理小结(53037)
3. 线性判别分析LDA原理总结(35187)
4. scikit-learn决策树算法类库使用小结(30037)
5. 谱聚类（spectral clustering）原理总结(23301)

评论排行榜

1. 梯度提升树(GBDT)原理小结(86)
2. 谱聚类（spectral clustering）原理总结(75)
3. 梯度下降（Gradient Descent）小结(65)
4. 卷积神经网络(CNN)反向传播算法(59)
5. 集成学习之Adaboost算法原理小结(50)

推荐排行榜

1. 梯度下降（Gradient Descent）小结(44)
2. 奇异值分解(SVD)原理与在降维中的应用(18)
3. 集成学习原理小结(15)
4. 卷积神经网络(CNN)反向传播算法(14)
5. 集成学习之Adaboost算法原理小结(13)

5) 决策树的数组使用的是numpy的float32类型，如果训练数据不是这样的格式，算法会先做copy再运行。

6) 如果输入的样本矩阵是稀疏的，推荐在拟合前调用csc\_matrix稀疏化，在预测前调用csr\_matrix稀疏化。

## 3. scikit-learn决策树结果的可视化

决策树可视化可以方便我们直观的观察模型，以及发现模型中的问题。这里介绍下scikit-learn中决策树的可视化方法。

### 3.1 决策树可视化环境搭建

scikit-learn中决策树的可视化一般需要安装graphviz。主要包括graphviz的安装和python的graphviz插件的安装。

第一步是安装graphviz。下载地址在：<http://www.graphviz.org/>。如果你是linux，可以用apt-get或者yum的方法安装。如果是windows，就在官网下载msi文件安装。无论是linux还是windows，装完后都要设置环境变量，将graphviz的bin目录加到PATH，比如我是windows，将C:/Program Files (x86)/Graphviz2.38/bin/加入了PATH

第二步是安装python插件graphviz：pip install graphviz

第三步是安装python插件pydotplus。这个没有什么好说的: pip install pydotplus

这样环境就搭好了，有时候python会很笨，仍然找不到graphviz，这时，可以在代码里面加入这一行：

```
os.environ["PATH"] += os.pathsep + 'C:/Program Files (x86)/Graphviz2.38/bin/'
```

注意后面的路径是你自己的graphviz的bin目录。

### 3.2 决策树可视化的三种方法

这里我们有一个例子讲解决策树可视化。

首先载入类库：

```
from sklearn.datasets import load_iris
from sklearn import tree
import sys
import os
os.environ["PATH"] += os.pathsep + 'C:/Program Files (x86)/Graphviz2.38/bin/'
```

接着载入scikit-learn的自带数据，有决策树拟合，得到模型：

```
iris = load_iris()
clf = tree.DecisionTreeClassifier()
clf = clf.fit(iris.data, iris.target)
```

现在可以将模型存入dot文件iris.dot。

```
with open("iris.dot", 'w') as f:
    f = tree.export_graphviz(clf, out_file=f)
```


这时候我们有3种可视化方法，第一种是用graphviz的dot命令生成决策树的可视化文件，敲完这个命令后当前目录就可以看到决策树的可视化文件iris.pdf.打开可以看到决策树的模型图。

```
#注意，这个命令在命令行执行
dot -Tpdf iris.dot -o iris.pdf
```


第二种方法是用pydotplus生成iris.pdf。这样就不用再命令行去专门生成pdf文件了。

```
import pydotplus
dot_data = tree.export_graphviz(clf, out_file=None)
graph = pydotplus.graph_from_dot_data(dot_data)
graph.write_pdf("iris.pdf")
```

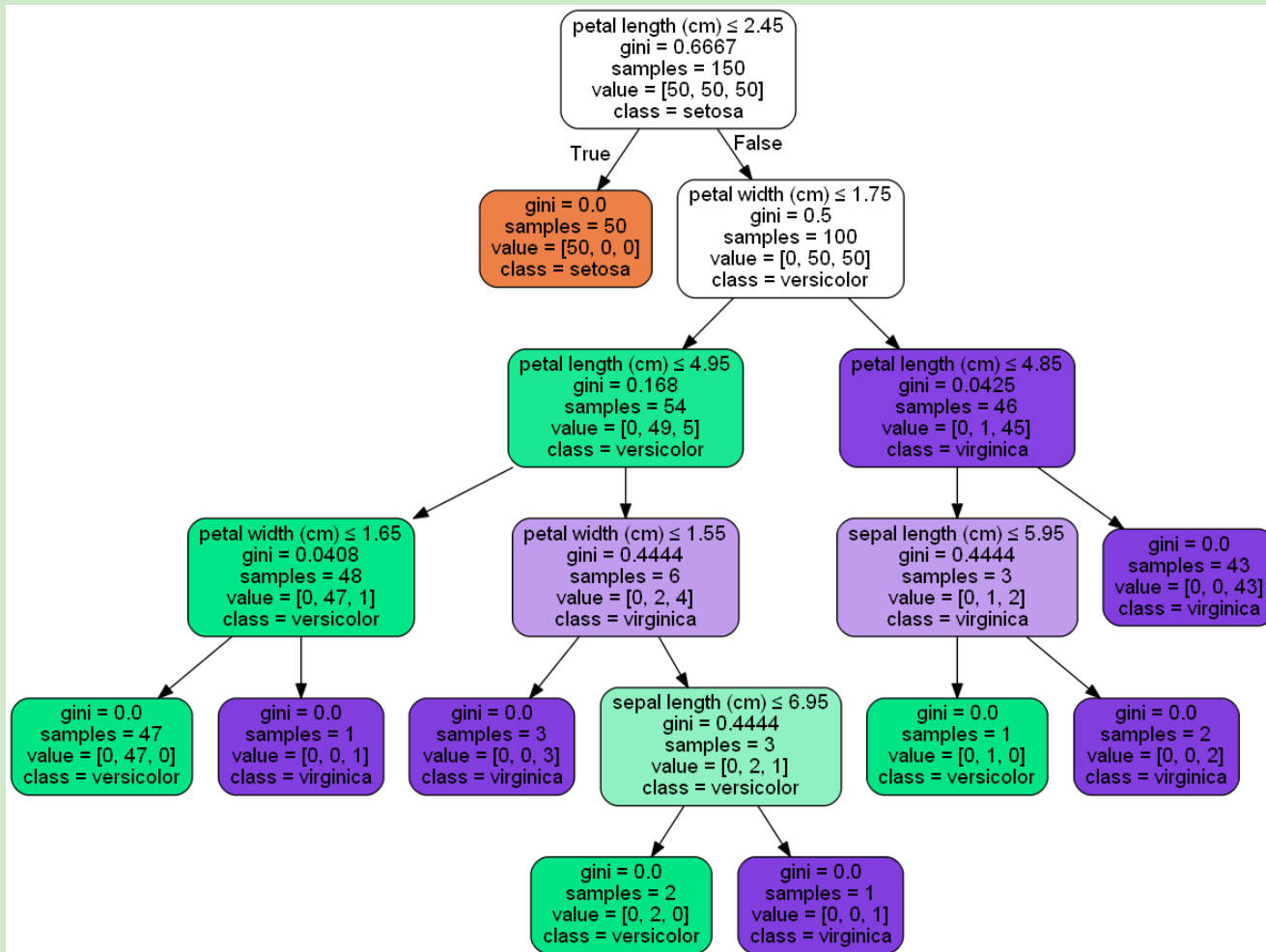
第三种办法是个人比较推荐的做法，因为这样可以直接把图产生在ipython的notebook。代码如下：



```
from IPython.display import Image
dot_data = tree.export_graphviz(clf, out_file=None,
                                feature_names=iris.feature_names,
                                class_names=iris.target_names,
                                filled=True, rounded=True,
                                special_characters=True)
graph = pydotplus.graph_from_dot_data(dot_data)
Image(graph.create_png())
```



在ipython的notebook生成的图如下：



## 4. DecisionTreeClassifier实例

这里给一个限制决策树层数为4的DecisionTreeClassifier例子。



```
from itertools import product
```

```
import numpy as np
import matplotlib.pyplot as plt

from sklearn import datasets
from sklearn.tree import DecisionTreeClassifier

# 仍然使用自带的iris数据
iris = datasets.load_iris()
X = iris.data[:, [0, 2]]
y = iris.target

# 训练模型，限制树的最大深度4
clf = DecisionTreeClassifier(max_depth=4)
#拟合模型
clf.fit(X, y)

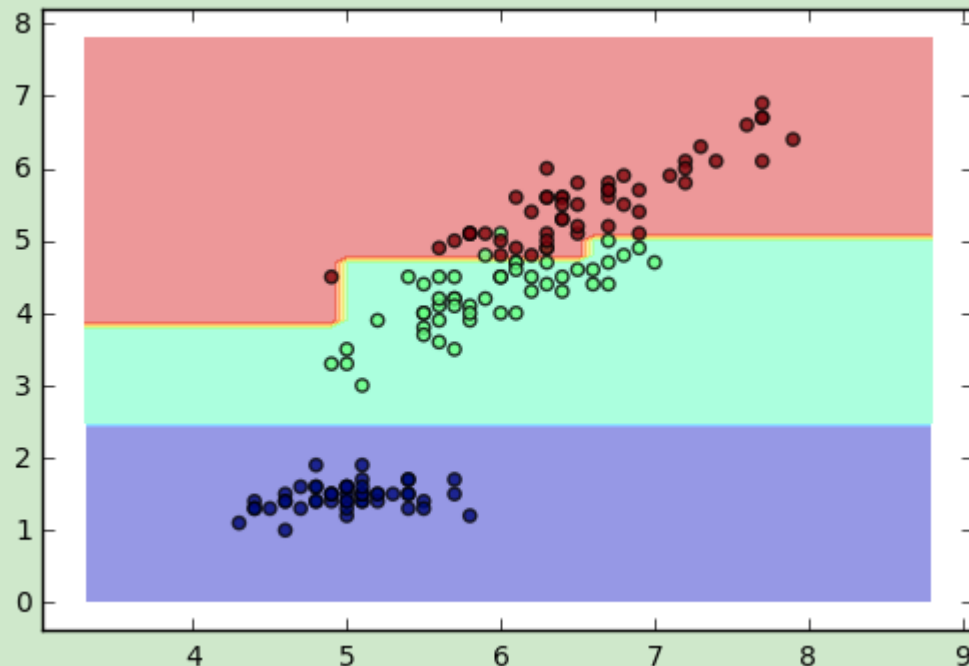
# 画图
x_min, x_max = X[:, 0].min() - 1, X[:, 0].max() + 1
y_min, y_max = X[:, 1].min() - 1, X[:, 1].max() + 1
xx, yy = np.meshgrid(np.arange(x_min, x_max, 0.1),
                     np.arange(y_min, y_max, 0.1))

Z = clf.predict(np.c_[xx.ravel(), yy.ravel()])
Z = Z.reshape(xx.shape)

plt.contourf(xx, yy, Z, alpha=0.4)
plt.scatter(X[:, 0], X[:, 1], c=y, alpha=0.8)
plt.show()
```



得到的图如下：



接着我们可视化我们的决策树，使用了推荐的第三种方法。代码如下：

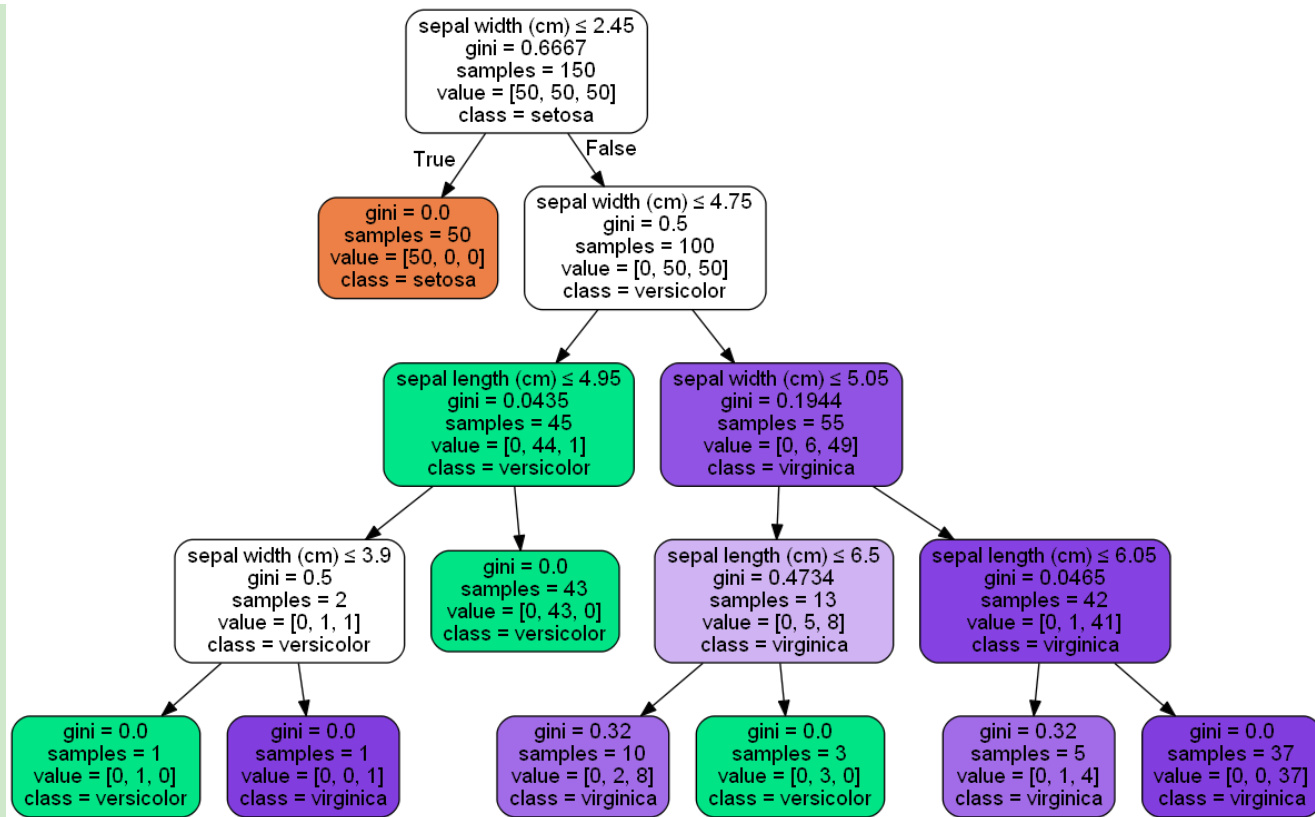


```
from IPython.display import Image
from sklearn import tree
import pydotplus
dot_data = tree.export_graphviz(clf, out_file=None,
                                feature_names=iris.feature_names,
                                class_names=iris.target_names,
                                filled=True, rounded=True,
                                special_characters=True)
graph = pydotplus.graph_from_dot_data(dot_data)
Image(graph.create_png())
```



生成的决策树图如下：





以上就是scikit-learn决策树算法使用的一个总结，希望可以帮到大家。

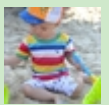
(欢迎转载，转载请注明出处。欢迎沟通交流：pinard.liu@ericsson.com)

分类: [0081. 机器学习](#)

好文要顶

关注我

收藏该文



刘建平Pinard

关注 - 13

粉丝 - 1182

+加关注

10

推荐

0

反对

« 上一篇：[决策树算法原理\(下\)](#)

» 下一篇：[K近邻法\(KNN\)原理小结](#)

posted @ 2016-11-12 14:28 刘建平Pinard 阅读(30037) 评论(47) 编辑 收藏


## 评论列表

#1楼 2017-03-01 10:48 又要起名了 



great job !


支持(0) 反对(0)

#2楼 2017-03-01 11:12 又要起名了 



3.1最后一行“路劲” 改为“路径”

支持(0) 反对(0)


#3楼[楼主 



@ 又要起名了

非常感谢指出错误，已修改。


支持(0) 反对(0)

#4楼 2017-03-16 11:00 robin\_hsu 



awesome!

支持(0) 反对(0)

#5楼 2017-06-18 10:43 niudong 



“第二种方法是用pydotplus生成iris.pdf。这样就不用再命令行去专门生成pdf文件了“

第二部的时候报错了，

```
>>> dot_data = tree.export_graphviz(clf, out_file=None)
```

Traceback (most recent call last):

File "<stdin>", line 1, in <module>

File "C:\Anaconda\lib\site-packages\sklearn\tree\export.py", line 126, in export\_graphviz

out\_file.write("digraph Tree {\n")

AttributeError: 'NoneType' object has no attribute 'write'

这个问题你知道怎么解决么？谢谢了

是我包装错了么？

支持(0) 反对(0)

#6楼[楼主] 2017-06-18 10:52 刘建平Pinard



@ niudong

你好，请先确认你是否完成3.1的所有步骤。然后请问你按第一种方法是否可以生成决策树图。如果第一种不行，第二种肯定失败。

支持(0) 反对(0)

#7楼 2017-06-18 11:09 niudong



使用最后一种方法绘图也没有成功，错误刚开始参数不对应，一个一个删除以后，报错情况和第二种方式是一样的，这种情况有碰到过么，如果有，应该用怎样的方式来进行解决，谢谢。

支持(0) 反对(0)

#8楼[楼主] 2017-06-18 11:19 刘建平Pinard



@ niudong

你好，必须保证第一种方法可以成功，这样才能证明你安装好了graphviz。第一种方法不成功你第二种第三种肯定失败。

支持(0) 反对(0)

#9楼 2017-06-18 11:25 niudong



第一种可以，就是第二种，第三种有些问题，不知道是什么情况

支持(0) 反对(0)

#10楼[楼主] 2017-06-18 11:29 刘建平Pinard



@ niudong

这一句在代码有没有？

```
os.environ["PATH"] += os.pathsep + 'C:/Program Files (x86)/Graphviz2.38/bin/'
```

注意路径是你自己的安装路径，一般第一种可以后面的不行就是环境变量的问题了

支持(0) 反对(0)

#11楼 2017-06-18 11:33 niudong



如果没有，第一种方式就显示不出来了，不是么？

具体显示问题如下：

```
>>> dot_data = tree.export_graphviz(clf, out_file=None)
```

Traceback (most recent call last):

File "<stdin>", line 1, in <module>

File "C:\Anaconda\lib\site-packages\sklearn\tree\export.py", line 126, in export\_graphviz

out\_file.write("digraph Tree {\n")

AttributeError: 'NoneType' object has no attribute 'write'

我在想是不是我这个sklearn的版本和你的版本不一样？

支持(0) 反对(0)

#12楼[楼主] 2017-06-18 11:37 刘建平Pinard



@ niudong

你好，第一种方法，不依赖于python的graphviz环境变量。因为在python代码中并没有使用graphviz。所以它可以成功说明你的graphviz软件可以正确使用。第二种和第三种方法则需要依赖graphviz的python插件与对应的环境变量。所以你看下代码加入我说的环境变量看看是否有问题。

p.s.我的scikit-learn是0.18

支持(0) 反对(0)

#13楼 2017-06-18 12:24 niudong



嗯嗯，我自己在试试更新版本，感觉很有可能是版本的问题。都忘了，很感谢的的博客，又让我快速的了解决策树，真的写的很棒，谢谢你，您能这么及时的回复也真的很感谢，祝周末愉快。

支持(1) 反对(0)

#14楼 2017-06-30 14:27 AI\_dream



您的博客对我帮助很大，以前我都是从树上学习，进展慢，学得不深，您写的博客让我容易理解，非常感谢

支持(0) 反对(0)

#15楼 2017-08-29 16:34 立顿经典醇



AttributeError: 'NoneType' object has no attribute 'write'

的问题，安装0.18及以上版本的scikit-learn可以解决：

pip install --upgrade scikit-learn

支持(0) 反对(0)

#16楼 2017-09-28 11:21 董浩Razor



请问特征划分点选择标准splitter是指在gini下找到了最佳的特征，然后特征对应着各种值，而分割的时候采用的是二叉分割，那么我们就需要找到最佳分割点，这个时候best表示用了所有的样本，而random使用了部分样本，是这样理解吗？

然后划分时考虑的最大特征数max\_features，选择sqrt或者log的时候，就相当于考虑了部分特征，这些特征是随机选择的吗？是在这部分特征里面用gini找最优吗？这样做是防止过拟合？还是其他的一些功能。。问题有点多~

支持(0) 反对(0)

#17楼[楼主] 2017-09-28 16:35 刘建平Pinard



@ 董浩Razor

对于问题1，是这样的。主要是大样本的时候计算量太大，用random是准确度和计算量之间的一个权衡。

对于问题2，其实主要的目的还是计算量太大的时候的一个折衷。当然，这样做也是做了正则化，但是个人认为不是最主要的目的。

支持(0) 反对(0)

#18楼 2017-09-28 17:02 董浩Razor



@ 刘建平Pinard

因为我看了随机森林里面有这段话，在使用决策树的基础上，RF对决策树的建立做了改进，对于普通的决策树，我们会在节点上所有的n个样本特征中选择一个最优的特征来做决策树的左右子树划分，但是RF通过随机选择节点上的一部分样本特征，这个数字小于n。

所以我对这个问题有点矛盾，既然本来就没有使用全部的特征，那为什么还刻意的表示一下呢，还是说CART里面有个机理选择部分特征是固定的，运行代码每次都一样的特征？

支持(0) 反对(0)

#19楼[楼主] 2017-09-29 10:22 刘建平Pinard



@ 董浩Razor

这个问题其实是算法相互借鉴的过程，最早的经典决策树是没有随机选择部分样本做特征分裂的。

现在普通的CART决策树也可以像RF一样随机选择一部分特征来建立模型，只是RF是多个决策树而已。

至于选择部分特征固定不固定的问题，由于一般的机器学习库比如sklearn都是用的伪随机，所以如果你只要指定random\_state的值的话，那么CART树做随机的时候选择的部分特征就是固定的。不指定则每次运行时选择的特征不一样。

支持(0) 反对(0)

#20楼 2017-10-10 15:37 Yuki727



非常感谢博主，对我的学习有很大的帮助！

还想请教博主一个问题，用这个sklearn的决策树进行计算时，如果数据中既有字符型变量又有数值型变量，该怎么处理？是不是应该将字符型变量转换为哑变量，也就是您上文提到的“输入的样本矩阵是稀疏的”？

支持(0) 反对(0)

#21楼[楼主] 2017-10-10 15:43 刘建平Pinard




@ Yuki727

你好，你说的对，需要对离散型特征比如字符串变量进行编码比如One-Hot Encoding，这样单个离散特征就变成了一个个稀疏的特征向量。而连续特征一般就不需要做特殊处理了。

对于离散特征的编码，这篇文章讲的很好，应该可以直接帮到你。

<http://blog.csdn.net/u012328159/article/details/71617381>

支持(0) 反对(0)


#22楼 2017-10-10 15:48 Yuki727 



@ 刘建平Pinard

好的，谢谢您


支持(0) 反对(0)

#23楼 2017-10-12 15:12 Yuki727 



不好意思再请教一下博主，sklearn的tree包能否将训练样本按叶节点输出，查看因变量的情况；而且我在官网上看到，用tree包进行预测分类情况只能输出一个值，不知道能否输出被分配到哪一个叶节点？

支持(0) 反对(0)

#24楼[楼主 



@ Yuki727


你好，这个sklearn API里面可以用DecisionTreeClassifier的decision\_path方法查看，它会输出每个输入样本是否经过决策树中的每一个节点。

由于决策树中叶子节点是固定的那几个，且每个样本最终只会落到一个叶子节点，所以你很容易查出每个样本落到的叶子节点位置。

你可以研究下这个API的输出就明白了

[http://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html#sklearn.tree.DecisionTreeClassifier.decision\\_path](http://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html#sklearn.tree.DecisionTreeClassifier.decision_path)

支持(0) 反对(0)

#25楼 2017-10-16 14:25 Yuki727 



@ 刘建平Pinard

好的，谢谢楼主！！

支持(0) 反对(0)

#26楼 2017-12-05 09:56 羽翔77 



谢谢 刘哥 的分享，我有两个问题，想问下：

第一个问题：

我在进行第三种方法进行数据呈现的时候，有报错

```
427 "does not match number of features, %d"
428 % (len(feature_names),
--> 429 decision_tree.n_features_))
430
431 # The depth of each node for plotting with 'leaf' option
ValueError: Length of feature_names, 4 does not match number of features, 2
```

我查找了下原因，应该是之前将

`X = iris.data[:, [0, 2]]` #那么只有两个feature了

如果下面

```
dot_data = tree.export_graphviz(clf, out_file=None,
feature_names=iris.feature_names,#这边的iris.feature_names就有4个feature了，所以报错
```

我改成了

```
feature_names=[iris.feature_names[0],iris.feature_names[2]],
```

之后就没有报错了

不过叶节点的名字就变化了，不知道是否理解正确？

第二个问题：

如果用decisiontreeclassification进行了fit建模，用什么方法查看模型是用什么feature叶节点一层层的分类下去的？

我查了下官网的decisiontreeclassification，没有找到这个方法。

支持(0) 反对(0)

---

#27楼[楼主] 2017-12-05 11:08 刘建平Pinard



@ 羽翔77

你好，第一个问题你做的没问题。

第二个问题，其实DecisionTreeClassifier将决策的过程都放在了内部对象tree\_里面，这个树对象没有暴露方法方便我们直接去查看。

所以最简单的方法还是用graphviz做可视化，这样在每个树节点里面的就可以很清楚的看到决策使用的特征和特征值了。

支持(0) 反对(0)

---

#28楼 2018-01-04 17:06 彭猛



@ niudong

你好，我遇到了和你同样的问题，第一种方式没问题，第二种第三种方式报错AttributeError: 'NoneType' object has no att

ribute 'write'

请问你解决了吗

支持(0) 反对(0)

#29楼[楼主] 2018-01-05 11:12 刘建平Pinard



@ 彭猛

第二种和第三种方法需要依赖graphiz的python插件与对应的环境变量。所以你看下代码加入我说的环境变量看看是否有问题。

支持(0) 反对(0)

#30楼 2018-01-09 17:10 小北潜行



博主，我安装了pydotplus

windows的系统，在cmd界面直接输入python，此处输入Import pydotplus没有问题

但是在notesbooks下面输入Import pydotplus

会报错no module named pydotplus

支持(0) 反对(0)

#31楼[楼主] 2018-01-10 11:33 刘建平Pinard



@ 小北潜行

你好，如果是这样，觉得是你的notebook的python 环境变量有问题，你看看pip安装一个普通的新python包，然后在notebook看能不能import。如果也不能import，那就可以确定是你jupyter notebook环境的问题。

支持(0) 反对(0)

#32楼 2018-01-15 11:57 小北潜行



博主

我有一个3W条的数据集

输入的特征值大部分是字符串型或者中文的

我使用了one-hot encoding

利用pd.get\_dummies()进行转换，转换出了30948个特征值

使用tree.DecisionTreeClassifier进行决策树分类

jupyter报错MemoryError

请问是因为特征值过多吗

支持(0) 反对(0)

#33楼[楼主] 2018-01-15 14:46 刘建平Pinard



“

@ 小北潜行

你好，是的，你的特征很多，要跑的话单机可能比较困难。

既然特征有3万多，而且估计大部分也是稀疏的，你可以尝试先用PCA降维到一个合适的维度（比如300），再去跑分类算法。

当然可以多选择几个降维到的维度，然后比较分类效果，选择一个合适的降维维度。

支持(0) 反对(0)

#34楼 2018-01-15 17:05 小北潜行

“

@ 刘建平Pinard

这里我有个疑问

因为数据集本身的特征值其实并不多

只是因为都是离散的，多为字符串和中文，而且数据量大，类别多

所以使用了one hot encoding后会出现大量的特征值

如果这个时候使用PCA进行降维，必然会消除部分的类别

这样合理吗？

还有就是我的数据集中存在中文

使用one-hot encoding过程中读取出现乱码

对于中文的特征，请问有什么好的处理方法吗？

支持(0) 反对(0)

#35楼[楼主] 2018-01-16 10:37 刘建平Pinard

“

@ 小北潜行

你好！

1. PCA的确会消除一些类别信息，但是好处就是可以简化模型和模型特征处理的计算时间，这是一个权衡。当然如果你觉得所有的信息都很重要，都不能丢弃的话，那就不降维，但是需要的计算量就需要一个分布式的系统来处理了。

2. 中文编码和英文的不一样，我之前在这篇文章有讲中文编码处理，你看看。

<http://www.cnblogs.com/pinard/p/6744056.html>

支持(0) 反对(0)

#36楼 2018-01-23 10:21 小俊俊俊

“

您好，我按您有篇文章的步骤进行了Python2.7.12的安装。（win7,32位系统）


但是在打开IDLE，新建file时，print语句无错误，可在shell中进行显示。

但用您的和官网的例程总是弹出“There's an error in your program:invalid syntax”，例如会在“%matplotlib lib inline”的%处标注红色，在“dot-Tpdf iris.dot-o iris.pdf”的“iris”处标红。且在shell界面无反应，未说明错误原因。

已经进行检查：


```
pip install numpy已安装
pip install scipy已安装
pip install upgrade --scikit-learn已更新
path中也加入了 : C:\python27;C:\python27\scripts
只是numpy和scipy文件很您文章中的文件有出入，在连接中没有找到您指出的版本，找到了：
numpy-1.13.3+mkl-cp27-cp27m-win32.whl
scipy-1.0.0-cp27-cp27m-win32.whl
```

支持(0) 反对(0)

#37楼 2018-01-23 10:23 小俊俊俊 

望您多多指教，非常感谢！

支持(0) 反对(0)

#38楼[楼主 


@ 小俊俊俊

你好，我的文章是16年写的，所以numpy和scipy有更新，没关系的。

会在"%matplotlib lib inline"的%处标注红色：这个例子是在jupyter notebook里面写的，主要是画图用。你如果用的是idle，不需要这几句。

在"dot-Tpdf iris.dot-o iris.pdf"的"iris"处标红：这个命令不是在python里执行的，是在系统命令行执行的，windows是cmd，linux 是shell里面。

支持(0) 反对(0)

#39楼 2018-01-23 11:20 小俊俊俊 

@ 刘建平Pinard

在idle中，画图的话，会在哪里显示呢？

将前边的程序在idle中输入，然后在windows 命令行 输入dot-Tpdf iris.dot-o iris.pdf，cmd显示'dot'不是内部或外部命令，也不是可运行程序，或批处理文件。是哪一个工具包没有安装吗？

如果在windows中命令行用cmd而不是shell，那shell有什么用吗？每次打开idle先出来shell界面，难道是装错了？

我是控制学科的，原来没有接触过这方面知识，现在研一，刚定课题：基于数据驱动的工业过程监测，刚开始学，问题比较幼稚。谢谢博主的耐心！能在起步时遇到您真的很幸运！

支持(0) 反对(0)

#40楼[楼主👤] 2018-01-24 11:50 刘建平Pinard 📧



@ 小俊俊俊

你好，按我上面说的3.1节第一步，你需要安装windows版的graphviz

idle出来的shell是python的解析器，不是操作系统的shell解释器。

支持(0) 反对(0)

#41楼 2018-01-24 14:43 小俊俊俊 📧



@ 刘建平Pinard

明白了，谢谢！

支持(0) 反对(0)

#42楼 2018-01-24 15:20 小俊俊俊 📧



安装windows版的graphviz，并加入了环境变量。

代码中加入了os.environ["PATH"] += os.pathsep + ' C:/Python27/graphviz-2.38/bin/'（我的路径）

在windows命令行，输入命令显示：dot can't open iris.dot

又是哪的问题呢？

支持(0) 反对(0)

#43楼 2018-01-24 15:21 小俊俊俊 📧



安装windows版的graphviz，并加入了环境变量。

代码中加入了os.environ["PATH"] += os.pathsep + ' C:/Python27/graphviz-2.38/bin/'（我的路径）

在windows命令行，输入命令显示：dot can't open iris.dot

又是哪的问题呢？

支持(0) 反对(0)

#44楼[楼主👤] 2018-01-25 11:25 刘建平Pinard 📧



@ 小俊俊俊

你好，你的iris.dot文件在哪个位置呢？



支持(0) 反对(0)

#45楼 2018-01-25 14:59 小俊俊俊 📧



和建的py文件在一个文件夹，在F盘，dot文件可以改变位置吗？

支持(0) 反对(0)

#46楼[楼主 ] 2018-01-25 15:09 刘建平Pinard 




@ 小俊俊俊

不用改变位置，但是建议你跑命令的时候加上iris.dot的绝对路径，比如：

```
F:/Temp/iris.dot
```

```
dot -Tpdf F:/Temp/iris.dot -o F:/Temp/iris.pdf
```

支持(0) 反对(0)

#47楼 2018-01-25 15:23 小俊俊俊 



@ 刘建平Pinard

加上绝对路径，可以了！谢谢！

支持(0) 反对(0)

[刷新评论](#) [刷新页面](#) [返回顶部](#)



注册用户登录后才能发表评论，请 [登录](#) 或 [注册](#)，[访问网站首页](#)。

#### 最新IT新闻:

- [陆奇时代的百度硬件梦](#)
- [阿里背后的女人彭蕾：我拿什么帮马云提高战斗力？](#)
- [Java案虽已尘埃落定，但软件界的连锁反应才刚刚开始](#)
- [彭蕾卸任：8年来她如何带着支付宝逆袭的？](#)
- [抛弃同龄人？还是一分没得到？再看摩拜收购案](#)
- » [更多新闻...](#)

#### 最新知识库文章:

- [写给自学者的入门指南](#)
- [和程序员谈恋爱](#)
- [学会学习](#)
- [优秀技术人的管理陷阱](#)
- [作为一个程序员，数学对你到底有多重要](#)
- » [更多知识库文章...](#)