

刘建平Pinard

十年码农，对数学统计学，机器学习，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 Adaboost类库使用小结

在集成学习之Adaboost算法原理小结中，我们对Adaboost的算法原理做了一个总结。这里我们就从实用的角度对scikit-learn中Adaboost类库的使用做一个小结，重点对调参的注意事项做一个总结。

1. Adaboost类库概述

scikit-learn中Adaboost类库比较直接，就是AdaBoostClassifier和AdaBoostRegressor两个，从名字就可以看出AdaBoostClassifier用于分类，AdaBoostRegressor用于回归。

AdaBoostClassifier使用了两种Adaboost分类算法的实现，SAMME和SAMME.R。而AdaBoostRegressor则使用了我们原理篇里讲到的Adaboost回归算法的实现，即Adaboost.R2。

当我们对Adaboost调参时，主要要对两部分内容进行调参，第一部分是对我们的Adaboost的框架进行调参，第二部分是对我们选择的弱分类器进行调参。两者相辅相成。下面就对Adaboost的两个类：AdaBoostClassifier和AdaBoostRegressor从这两部分做一个介绍。

2. AdaBoostClassifier和AdaBoostRegressor框架参数

我们首先来看看AdaBoostClassifier和AdaBoostRegressor框架参数。两者大部分框架参数相同，下面我们一起讨论这些参数，两个类如果有不同点我们会指出。

1) **base_estimator**: AdaBoostClassifier和AdaBoostRegressor都有，即我们的弱分类学习器或者弱回归学习器。理论上可以选择任何一个分类或者回归学习器，不过需要支持样本权重。我们常用的一般是CART决策树或者神经网络MLP。默认是决策树，即AdaBoostClassifier默认使用CART分类树DecisionTreeClassifier，而AdaBoostRegressor默认使用CART回归树DecisionTreeRegressor。另外有一个要注意的点是，如果我们选择的AdaBoostClassifier算法是SAMME.R，则我们的弱分类学习器还需要支持概率预测，也就是在scikit-learn中弱分类学习器对应的预测方法除了predict还需要有predict_proba。

2) **algorithm**: 这个参数只有AdaBoostClassifier有。主要原因是scikit-learn实现了两种Adaboost分类算法，SAMME和SAMME.R。两者的主要区别是弱学习器权重的度量，SAMME使用了和我们的原理篇里二元分类Adaboost算法的扩展，即用对样本集分类效果作为弱学习器权重，而SAMME.R使用了对于样本集分类的预测概率大小来作为弱学习器权重。由于SAMME.R使用了概率度量的连续值，迭代一般比SAMME快，因此AdaBoostClassifier的默认算法algorithm的值也是SAMME.R。我们一般使用默认的SAMME.R就够了，但是要注意的是使用了SAMME.R，则弱分类学习器参数base_estimator必须限制使用支持概率预测的分类器。SAMME算法则没有这个限制。

3) **loss**: 这个参数只有AdaBoostRegressor有，Adaboost.R2算法需要用到。有线性'linear'，平方'square'和指数'exponential'三种选择，默认是线性，一般使用线性就足够了，除非你怀疑这个参数导致拟合程度不好。这个值的意义在原理篇我们也讲到了，它对应了我们对第k个弱分类器的中第i个样本的误差的处理，即：如果是线性误差，则 $e_{ki} = \frac{|y_i - G_k(x_i)|}{E_k}$ ；如果是平方误差，则 $e_{ki} = \frac{(y_i - G_k(x_i))^2}{E_k^2}$ ，如果是指数误差，则 $e_{ki} = 1 - \exp(-\frac{y_i - G_k(x_i)}{E_k})$ ， E_k 为训练集上的最大误差 $E_k = \max |y_i - G_k(x_i)|$ $i = 1, 2, \dots, m$

4) **n_estimators**: AdaBoostClassifier和AdaBoostRegressor都有，就是我们的弱学习器的最大迭代次数，或者说最大的弱学习器的个数。一般来说n_estimators太小，容易欠拟合，n_estimators太大，又容易过拟合，一般选择一个适中的数值。默认是50。在实际调参的过程中，我们常常将n_estimators和下面介绍的参数learning_rate一起考虑。

5) **learning_rate**: AdaBoostClassifier和AdaBoostRegressor都有，即每个弱学习器的权重缩减系数 ν ，在原理篇的正则化章节我们也讲到了，加上了正则化项，我们的强学习器的迭代公式为 $f_k(x) = f_{k-1}(x) + \nu \alpha_k G_k(x)$ 。 ν 的取值范围为 $0 < \nu \leq 1$ 。对于同样的训练集拟合效果，较小的 ν 意味着我们需要更多的弱学习器的迭代次数。通常我们用步长和迭代最大次数一起来决定算法的拟合效果。所以这两个参数n_estimators和learning_rate要一起调参。一般来说，可以从一个小一点的 ν 开始调参，默认是1。

3. AdaBoostClassifier和AdaBoostRegressor弱学习器参数

这里我们再讨论下AdaBoostClassifier和AdaBoostRegressor弱学习器参数，由于使用不同的弱学习器，则对应的弱学习器参数各不相同。这里我们仅仅讨论默认的决策树弱学习器的参数。即CART分类树DecisionTreeClassifier和CART回归树DecisionTreeRegressor。

DecisionTreeClassifier和DecisionTreeRegressor的参数基本类似，在scikit-learn决策树算法类库使用小结这篇文章中我们对这两个类的参数做了详细的解释。这里我们只拿出调参时需要尤其注意的最重要几个的参数再拿出来说一说：

1) 划分时考虑的最大特征数**max_features**: 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 $\log_2 N$ 个特征；如果是"sqr"或者"auto"意味着划分时最多考虑 \sqrt{N} 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。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。

公告

★珠江追梦，饮岭南茶，恋鄂北情
你的支持是我写作的动力：



昵称：刘建平Pinard
园龄：4年
粉丝：6704
关注：16
-取消关注

积分与排名

积分 - 473116
排名 - 707

随笔分类 (135)

0040. 数学统计学(9)
0081. 机器学习(71)
0082. 深度学习(11)
0083. 自然语言处理(23)
0084. 强化学习(19)
0121. 大数据挖掘(1)
0122. 大数据平台(1)

随笔档案 (135)

2019年7月(1)
2019年6月(1)
2019年5月(2)
2019年4月(3)
2019年3月(2)
2019年2月(2)
2019年1月(2)
2018年12月(1)
2018年11月(1)
2018年10月(3)
2018年9月(3)
2018年8月(4)
2018年7月(3)
2018年6月(3)
2018年5月(3)
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)

- 2) 决策树最大深**max_depth**: 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。
- 3) 内部节点再划分所需最小样本数**min_samples_split**: 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。
- 4) 叶子节点最少样本数**min_samples_leaf**: 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。
- 5) 叶子节点最小的样本权重和**min_weight_fraction_leaf**: 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。
- 6) 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。

4. AdaBoostClassifier实战

这里我们用一个具体的例子来讲解AdaBoostClassifier的使用。

完整代码参见我的github: <https://github.com/ljpzzz/machinelearning/blob/master/ensemble-learning/adaboost-classifier.ipynb>

首先我们载入需要的类库:

```
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
from sklearn.ensemble import AdaBoostClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.datasets import make_gaussian_quantiles
```

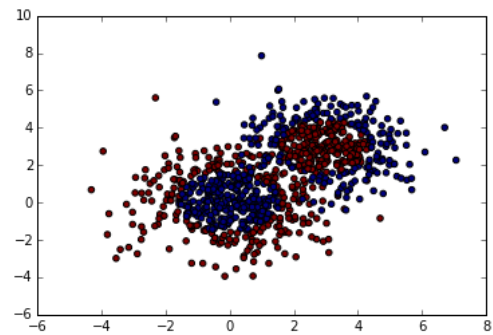
接着我们生成一些随机数据来做二元分类, 如果对如何产生随机数据不熟悉, 在另一篇文章机器学习算法的随机数据生成中有比较详细的介绍。

```
# 生成2维正态分布, 生成的数据按分位数分为两类, 500个样本, 2个样本特征, 协方差系数为2
X1, y1 = make_gaussian_quantiles(cov=2.0, n_samples=500, n_features=2, n_classes=2, random_state=1)
# 生成2维正态分布, 生成的数据按分位数分为两类, 400个样本, 2个样本特征均值都为3, 协方差系数为2
X2, y2 = make_gaussian_quantiles(mean=(3, 3), cov=1.5, n_samples=400, n_features=2, n_classes=2, random_state=1)
#讲两组数据合成一组数据
X = np.concatenate((X1, X2))
y = np.concatenate((y1, - y2 + 1))
```

我们通过可视化看看我们的分类数据, 它有两个特征, 两个输出类别, 用颜色区别。

```
plt.scatter(X[:, 0], X[:, 1], marker='o', c=y)
```

输出为下图:



可以看到数据有些混杂, 我们现在用基于决策树的Adaboost来做分类拟合。

```
bdt = AdaBoostClassifier(DecisionTreeClassifier(max_depth=2, min_samples_split=20, min_samples_leaf=5),
                        algorithm="SAMME",
                        n_estimators=200, learning_rate=0.8)
bdt.fit(X, y)
```

这里我们选择了SAMME算法, 最多200个弱分类器, 步长0.8, 在实际运用中你可能需要通过交叉验证调参而选择最好的参数。拟合完了后, 我们用网格图来看看它拟合的区域。

常去

- 52 |
- Ana
- 深度
- 深度
- 机器
- 机器
- 强化学习入门书

阅读排行榜

1. 梯度下降 (Gradient Desce
2. 梯度提升树(GBDT)原理小结
3. word2vec原理(一) CBOW±
- 基础(198964)
4. 奇异值分解(SVD)原理与在
- 69)
5. 线性判别分析LDA原理总结(

评论排行榜

1. 梯度提升树(GBDT)原理小结
2. 集成学习之Adaboost算法原
3. 决策树算法原理(下)(304)
4. word2vec原理(二) 基于Hie
- 的模型(271)
5. 谱聚类 (spectral clusterin

推荐排行榜

1. 梯度下降 (Gradient Desce
2. 奇异值分解(SVD)原理与在
3. 梯度提升树(GBDT)原理小结
4. 集成学习原理小结(46)
5. 集成学习之Adaboost算法原

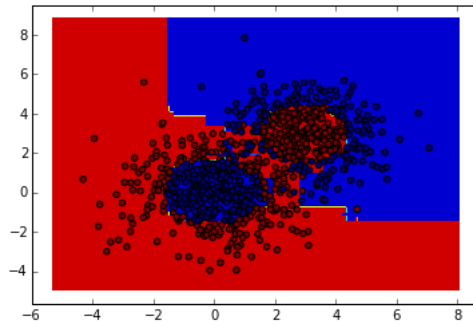


```
x_min, x_max = X[:, 0].min() - 1, X[:, 0].max() + 1
y_min, y_max = X[:, 1].min() - 1, X[:, 1].max() + 1
xx, yy = np.meshgrid(np.arange(x_min, x_max, 0.02),
                     np.arange(y_min, y_max, 0.02))

Z = bdt.predict(np.c_[xx.ravel(), yy.ravel()])
Z = Z.reshape(xx.shape)
cs = plt.contourf(xx, yy, Z, cmap=plt.cm.Paired)
plt.scatter(X[:, 0], X[:, 1], marker='o', c=y)
plt.show()
```



输出的图如下:



从图中可以看出, Adaboost的拟合效果还是不错的, 现在我们看看拟合分数:

```
print "Score:", bdt.score(X, y)
```

输出为:

Score: 0.913333333333

也就是说拟合训练集数据的分数还不错。当然分数高并不一定好, 因为可能过拟合。

现在我们将最大弱分离器个数从200增加到300。再看看拟合分数。

```
bdt = AdaBoostClassifier(DecisionTreeClassifier(max_depth=2, min_samples_split=20, min_samples_leaf=5),
                        algorithm="SAMME",
                        n_estimators=300, learning_rate=0.8)

bdt.fit(X, y)
print "Score:", bdt.score(X, y)
```

此时的输出为:

Score: 0.962222222222

这印证了我们前面讲的, 弱分离器个数越多, 则拟合程度越好, 当然也越容易过拟合。

现在我们降低步长, 将步长从上面的0.8减少到0.5, 再看看拟合分数。

```
bdt = AdaBoostClassifier(DecisionTreeClassifier(max_depth=2, min_samples_split=20, min_samples_leaf=5),
                        algorithm="SAMME",
                        n_estimators=300, learning_rate=0.5)

bdt.fit(X, y)
print "Score:", bdt.score(X, y)
```

此时的输出为:

Score: 0.894444444444

可见在同样的弱分类器的个数情况下, 如果减少步长, 拟合效果会下降。

最后我们看看当弱分类器个数为700, 步长为0.7时候的情况:

```
bdt = AdaBoostClassifier(DecisionTreeClassifier(max_depth=2, min_samples_split=20, min_samples_leaf=5),
                        algorithm="SAMME",
                        n_estimators=600, learning_rate=0.7)

bdt.fit(X, y)
print "Score:", bdt.score(X, y)
```

此时的输出为:

Score: 0.961111111111

此时的拟合分数和我们最初的300弱分类器, 0.8步长的拟合程度相当。也就是说, 在我们这个例子中, 如果步长从0.8降到0.7, 则弱分类器个数要从300增加到700才能达到类似的拟合效果。

以上就是scikit-learn Adaboost类库使用的一个总结，希望可以帮到朋友们。

(欢迎转载，转载请注明出处。欢迎沟通交流：liujianping-ok@163.com)

分类: [0081. 机器学习](#)

好文要顶

已关注

收藏该文



刘建平Pinard

关注 - 16

粉丝 - 6704

我在关注他 取消关注

26

0

« 上一篇: [集成学习之Adaboost算法原理小结](#)
» 下一篇: [梯度提升树\(GBDT\)原理小结](#)

posted @ 2016-12-06 19:41 刘建平Pinard 阅读(61288) 评论(95) 编辑 收藏

评论列表

- #51楼 2019-05-22 10:12 CassielLee

大神你好，请问用BP神经网络做弱分类器吗？我看到有一些文章是这样用的，但是具体实现方式却没有给出

支持(0) 反对(0)
- #52楼 [楼主] 2019-05-22 10:28 刘建平Pinard

@ CassielLee
你好，肯定可以做。但是目前深度学习一个大的神经网络已经可以满足需求了，所以现在神经网络做集成并不主流。

之前是因为神经网络不能做的特别深，所以就出现了很多集成学习+神经网络的做法。目前单个大神经网络可以满足需求，不建议走这一块。

支持(0) 反对(0)
- #53楼 2019-05-22 10:30 CassielLee

@ 刘建平Pinard 但是比较深的网络训练起来比较麻烦，我想的是构建一个最简单的三层BP，然后用集成学习，可能会容易一些。

支持(0) 反对(0)
- #54楼 [楼主] 2019-05-23 09:55 刘建平Pinard

@ CassielLee
你好，如果只是3层你可以尝试使用sklearn的API来做，弱学习器使用sklearn.neural_network.MLPClassifier

支持(0) 反对(0)
- #55楼 2019-06-17 20:11 叶落锦

老师您好，我想问一个问题
每一个弱分类器的输入是原始数据的特征直接乘上更新后的权值吗？
比如，对于初始阶段，输入的每一个样本乘以 1/m，m是样本个数，然后使用第一个分类器分类？

支持(0) 反对(0)
- #56楼 [楼主] 2019-06-18 10:30 刘建平Pinard

@ 叶落锦
你好，权值并不需要和特征相乘，而只是针对分类结果输出是否正确对应的损失大小而言。可以参看原理篇里面的讲解。

至于每个弱分类器的样本权重，如果你是用的算法库，这块对你来说是透明的。当然对于初始阶段，输入的每一个样本的权重的确是1/m

支持(0) 反对(0)
- #57楼 2019-06-19 17:18 叶落锦

@ 刘建平Pinard
对的，权重和对应的损失相关，但是如果对每一个弱分类器输入的训练样本都是一样的（权重只是在计算分类误差时用到了），那么会不会每次得到的弱分类器也是一样的呢，因为训练数据都是一样的，尤其对于CART这种没有确定性的分类器？

支持(0) 反对(0)
- #58楼 2019-06-19 17:21 叶落锦

@ 刘建平Pinard
讲错啦，CART这种确定性的，没有随机因素的分类器

支持(0) 反对(0)
- #59楼 [楼主] 2019-06-20 09:55 刘建平Pinard

@ 叶落锦
你好！
弱学习器不会都一样的。因为权重影响损失，影响损失函数的大小，而我们的弱学习器的建立的目标就是最小化当前的损失函数，才得到当前弱学习器的模型参数的。

对于CART用于Adaboost，在做节点分裂的时候也要考虑权重，比如是分类即在选择各个特征最优划分点的基尼系数的时候，基尼系数公式里，每个样本
- https://www.cnblogs.com/pinard/p/6136914.html
- 4/9

的权重也是要乘上来才能比较的。

否则就会出现你说的，在决策树弱学习器建立的时候，权重不起作用，那么决策树每次建立的就几乎一样了，没有意义。

支持(0) 反对(0)

#60楼 2019-07-31 15:26 最后的战役aag

回复 引用

@ 刘建平Pinard

您好，请问样本权重如何处理？是不是将样本权重乘以样本之后再训练弱学习器？

支持(0) 反对(0)

#61楼 [楼主] 2019-08-01 10:16 刘建平Pinard

回复 引用

@ 最后的战役aag

你好，对的。样本权重会对弱学习器的误分类计算由影响。如果误分类的样本权重重大，则损失会更大。也就是在优化损失函数的时候考虑样本权重，一般是乘以权重后再进行训练。

比如Adaboost如果你使用CART回归树做弱学习器，那么在分裂子树的时候，计算均方误差时，所有的 样本特征都要乘以自己权重后再参与计算。

支持(1) 反对(0)

#62楼 2019-08-01 10:20 最后的战役aag

回复 引用

@ 刘建平Pinard

明白了，谢谢

支持(0) 反对(0)

#63楼 2019-08-01 16:06 机器学习蔡鸟

回复 引用

博主，请问为什么用两个高斯分布，是每个高斯分布对应一个类吗，分为两类。还有怎么看最后的函数表达式，就像李航例题的最后的f(x)

支持(0) 反对(0)

#64楼 [楼主] 2019-08-02 10:54 刘建平Pinard

回复 引用

@ 机器学习蔡鸟

你好，对的，是期望人工生成2类数据，用于做算法的数据集。

至于函数表达式，由于这里用的是决策树，不太好用多项式来表达。如果你使用的线性分类器，那么就可以用函数表达式写出来。

支持(0) 反对(0)

#65楼 2019-08-07 16:00 freedom_hu

回复 引用

您好，我想请问一下，集成学习都是提高模型的准确度，并要求在50%以上。但对于样本不平衡的二分类问题，想提高捕捉少数类的能力，即提高模型的precision和recall.如果采用集成学习的话，提高了准确度，但结果使得少数类基本全部分类错误，这种情况下，该如何使用集成学习呢？对于样本不平衡的问题，是不是不在适合集成学习了？还是说要对样本先进行重采样？

支持(0) 反对(0)

#66楼 [楼主] 2019-08-08 10:52 刘建平Pinard

回复 引用

@ freedom_hu

你好，样本不平衡的时候，使用任何分类算法模型都会有很大的问题，模型准确度不高。

此时不是要考虑是否使用集成学习，而是首先要考虑解决样本失衡问题，你可以看看我写的这篇<特征工程之特征预处理>的第三节：
<https://www.cnblogs.com/pinard/p/9093890.html>

支持(0) 反对(0)

#67楼 2019-11-25 11:19 windelvin

回复 引用

请问刘老师，SAMME和SAMME.R算法都可以用于多分类任务吗？具体算法是什么呢？与matlab里的AdaBoost.M2算法有何不同？

支持(0) 反对(0)

#68楼 [楼主] 2019-11-26 09:02 刘建平Pinard

回复 引用

@ windelvin

SAMME和SAMME.R算法都可以用于多分类任务，和AdaBoost.M2是类似的Adaboost的优化版算法，原理基本相同。

如果你对 SAMME和SAMME.R算法详情感兴趣，可以看看这篇论文
<https://web.stanford.edu/~hastie/Papers/samme.pdf>

支持(0) 反对(0)

#69楼 2020-03-02 01:11 niunai96

回复 引用

老师您好：

adaboost算法本身没有限制使用神经网络作为基学习器。但是sklearn的adaboost算法对基学习器有要求，必须支持样本权重和类属性 classes_ 以及 n_classes_

我想问下classes_代表的是什么？ n_classes_是协方差系数是吗？谢谢老师！

支持(0) 反对(0)

#70楼 [楼主] 2020-03-02 09:28 刘建平Pinard

回复 引用

@niunai96

你好，其实sklearn的文档里都有的，classes_是所有类别的标签列表，比如[0,1,2]这样的.n_classes就是类别数量，比如3，就是一共三个类别。

支持(0) 反对(0)

#71楼 2020-03-03 01:06 niunai96

回复 引用

@刘建平Pinard 谢谢老师，是我偷懒了，下回我先自己看一下sklearn文档，谢谢老师的回复！！		支持(0)	反对(0)
# 72楼	2020-03-18 17:20 xiaozhu1024	回复	引用
老师，我一直有一个疑问，就是像SVM和adaboost在做分类时，算法设置的y取值为-1和+1，而在实践中，都是设置为0和1，python内置的模块是怎么实现的呢		支持(0)	反对(0)
# 73楼	[楼主] 2020-03-19 09:06 刘建平Pinard	回复	引用
@xiaozhu1024 你好，设置[-1,1]还是[0,1]与你使用的算法的损失函数有关。并不是实践中都是设置为0和1。如果你的损失函数是类似逻辑回归那样的损失，那么一般都是[0,1],如果是类似Adaboost,SVM这样的，那么一般都是[-1,1]		支持(0)	反对(0)
# 74楼	2020-03-19 09:37 xiaozhu1024	回复	引用
@刘建平Pinard 老师，感谢你的回答，但是我还有点疑问，为什么你上面的实例代码中生成的数据，y的取值是0和1呢？那用adaboost分类是怎么实现的呢？是python内置的模块中进行了转换吗		支持(0)	反对(0)
# 75楼	[楼主] 2020-03-19 09:48 刘建平Pinard	回复	引用
@xiaozhu1024 你好，这个是sklearn自己Adaboost算法库做了类别值的转化，不是python自带的内置模块。		支持(0)	反对(0)
# 76楼	2020-03-19 09:54 xiaozhu1024	回复	引用
@刘建平Pinard 现在完全清楚了，非常感谢老师的指点		支持(0)	反对(0)
# 77楼	2020-04-07 20:09 supermaer	回复	引用
@刘建平Pinard 老师您好，CART结合Adaboost时，对特征加权，是指特征内的每一个值都要先和权值相乘再去计算GINI值么，这样的话，之后生成的每一棵树的特征值都发生了改变（阈值也发生了改变），这样生成的树还能作用于检验样本吗？ 书上说的“误差率e”（我理解的作用和GINI是一样的，用于确定分支点），应该是权值去乘的y值（y取值1，-1），但这样的话，权值就影响不了GINI了（因为是分类），所以我十分困惑。 另外，Adaboost tree是不是只能是树桩，能不能产生多层的分支呢？		支持(0)	反对(0)
# 78楼	[楼主] 2020-04-08 09:22 刘建平Pinard	回复	引用
@supermaer 你好！ 1，指特征内的每一个值都要先和权值相乘再去计算GINI值么===》对的。 2，这样生成的树还能作用于检验样本吗===》单颗CART弱学习器不行，但是整体可以 3，书上说的“误差率e”应该是指的弱学习器的加权误差率，即权重乘以误分类的样本计数1，你说的y会让人以为是分类的结果输出，而不是误分类的样本。 4，Adaboost tree可以产生多层的分支。		支持(0)	反对(0)
# 79楼	2020-04-08 11:40 supermaer	回复	引用
@刘建平Pinard，谢谢老师！ 关于3，加权误差率在CART树中是不是就没有作用了，因为CART树（尤其是分类树）是按GINI来确定分支的。似乎和误差率没有关系？		支持(0)	反对(0)
# 80楼	[楼主] 2020-04-09 09:13 刘建平Pinard	回复	引用
@supermaer 你好，对的，加权误差率在CART构建是没有用的，只是影响对应弱学习器的系数。		支持(0)	反对(0)
# 81楼	2020-05-18 23:30 阿尔法小杰	回复	引用
刘老师好： 有几个问题还需要再咨询下： 1 样本的权重是在弱学习器中有所体现，您在61楼说权重乘以样本后在进行训练是什么意思？这个能不能再具体讲解下，比如决策树作为弱学习器，在求每个特征的基尼指数的时候是如何操作的？ 2 adaboost的弱学习器不能是使用MLP，是因为MLP不支持样本权重，那么你在文中提到的神经网络是哪一种？ 3 类别权重在adaboost中是怎么考虑的？ 4 在类别不平衡的数据集中，普通学习器一般对少类别样本的学习识别能力偏差，那么在adaboost的学习中，正好是加大了误分类样本的权重来增加对它的学习，按理说，在学习类别不平衡样本的时候，adaboost应该表现出很好的效果，怎么实际做出来的效果和普通学习器的效果差不多呢？		支持(0)	反对(0)
# 82楼	[楼主] 2020-05-19 09:29 刘建平Pinard	回复	引用
@阿尔法小杰 你好！			

1, 这个其实就是CART决策树里样本权重的用法, 你在计算特征GINI系数 (分类) 或者MSE (回归) , 做决策树分裂的时候, 样本参与的计算都需要乘以一个样本权重。

以你说的分类的基尼系数为例, 我们知道计算方式是这样的: $Gini(p) = \sum_{k=1}^K p_k(1 - p_k) = 1 - \sum_{k=1}^K p_k^2$

这里的 p_k 如果没有样本权重, 那么就是第k类样本的占比, 但是有了样本权重, 那么 p_k 就不能简单是第k类样本的占比, 而是第k类样本的加权占比。

2, adaboost的弱学习器的确不能使用MLP, 但是稍微定制一下即可用了, 我一般是这么用的。

```
from sklearn.neural_network import MLPClassifier
from sklearn.datasets import load_iris
from sklearn.ensemble import AdaBoostClassifier
class customMLPClassifier(MLPClassifier):
def resample_with_replacement(self, X_train, y_train, sample_weight):
```

```
# normalize sample_weights if not already
sample_weight = sample_weight / sample_weight.sum(dtype=np.float64)

X_train_resampled = np.zeros((len(X_train), len(X_train[0])), dtype=np.float32)
y_train_resampled = np.zeros((len(y_train)), dtype=np.int)
for i in range(len(X_train)):
    # draw a number from 0 to len(X_train)-1
    draw = np.random.choice(np.arange(len(X_train)), p=sample_weight)

    # place the X and y at the drawn number into the resampled X and y
    X_train_resampled[i] = X_train[draw]
    y_train_resampled[i] = y_train[draw]

return X_train_resampled, y_train_resampled

def fit(self, X, y, sample_weight=None):
    if sample_weight is not None:
        X, y = self.resample_with_replacement(X, y, sample_weight)

    return self._fit(X, y, incremental=(self.warm_start and
                                        hasattr(self, "classes_")))
```

```
X,y = load_iris(return_X_y=True)
adabooster = AdaBoostClassifier(base_estimator=customMLPClassifier())
adabooster.fit(X,y)
```

参见: <https://stackoverflow.com/questions/55632010/using-scikit-learns-mlpclassifier-in-adaboostclassifier>

3, 类别权重的影响仅仅是样本权重变为: 类别权重x样本权重。但是 Adaboost不支持 class weight

4, 类别不平衡的时候, 是特征工程要解决的问题, 不能拖到让算法自己来解决。就算是Adaboost, 也一样解决不了这个问题。

回到Adaboost, Adaboost的增加误分类样本的权重, 仅仅是针对他的算法迭代流程中误分类的样本, 而不是你说的少数类别的样本。你的少数类别的样本误分类后权重再大, 仍然需要和其他多数类别样本综合考虑损失。

支持(0)

反对(0)

#83楼 2020-06-18 20:46 pigpoo

回复

引用

@刘建平Pinard

谢谢老师, 对第一个问题还是我还是有点钻牛角尖了。我们使用adaboost是给了每一个样本权重, 计算基尼系数的时候, 用的是类别权重。举个例子, 样本(1,1,1,0,0), 权重(0.3,0.2,0.2,0.1,0.2), 实在想不到怎么把权重考虑进去。

支持(0)

反对(0)

#84楼 [楼主] 2020-06-19 09:03 刘建平Pinard

回复

引用

@pigpoo

你好, 总样本数为5, 如果没设置权重, 那么每个样本出现的概率是1/5, 所以类别1的概率是3/5,类别0的概率是2/5。设置了权重后, 类别1的概率为:

$$\frac{\frac{1}{5} * 0.3 + \frac{1}{5} * 0.3 + \frac{1}{5} * 0.2}{\frac{1}{5} * 0.3 + \frac{1}{5} * 0.3 + \frac{1}{5} * 0.2 + \frac{1}{5} * 0.1 + \frac{1}{5} * 0.2}$$

支持(0)

反对(0)

#85楼 2020-06-23 14:31 HULU-葫芦

回复

引用

刘老师您好, 想请教下 ADABOOST 子分类器如果是决策树的话, 应该怎么调参。因为是子分类器, 不可能一个个分别去调参, 如果是所有分类器用一套参数的话, 不知道怎么用GRID 或者贝叶斯优化了, 是只能通过经验去调参了么。谢谢

支持(0)

反对(0)

#86楼 [楼主] 2020-06-24 08:53 刘建平Pinard

回复

引用

@HULU-葫芦

你好, 一般是所有的决策树弱学习器使用一套参数即可, 我们通过调整决策树的数量, 以及所有决策树公用的正则化参数来调参。不然假如你有100颗决策树弱学习器, 那不是要调100颗决策树的参数, 那个组合就太多了。

支持(0)

反对(0)

#87楼 2020-10-16 11:15 刘大佬的小跟班

回复

引用

老师, 我用了你定制的那个感知机作为基学习器的adaboost算法, 我在导出模型转成pmml文件是出现了问题, 我想问下, 这个算法能不能转成Pmml文件, 我是做的二分类, 错误粘贴如下, 望老师回复。Standard output is empty
java.lang.IllegalArgumentException: The transformer object (Python class sklearn.ensemble._weight_boosting.AdaBoostClassifier) is not a supported Transformer
RuntimeError: The JPMML-SkLearn conversion application has failed. The Java executable should have printed more information about the failure into its standard output and/or standard error streams
进程已结束, 退出代码 1





支持(0)

反对(0)

#88楼 [楼主] 2020-10-19 09:25 刘建平Pinard	回复 引用
<div>@刘大佬的小跟班</div> <div>你好，你用的是这个库吗？ https://github.com/alex-pirozhenko/sklearn-pmml 如果是这个的话，adaboost是不支持的。 Supported models DecisionTreeClassifier DecisionTreeRegressor GradientBoostingClassifier RandomForestClassifier</div>	支持(0) 反对(0)
#89楼 2020-10-19 09:27 刘大佬的小跟班	回复 引用
<div>from sklearn2pmml import sklearn2pmml, PMMLPipeline，用的这个</div>	支持(0) 反对(0)
#90楼 2020-10-19 09:28 刘大佬的小跟班	回复 引用
<div>那我应该怎么办？需要在android平台上使用啊</div>	支持(0) 反对(0)
#91楼 [楼主] 2020-10-19 09:43 刘建平Pinard	回复 引用
<div>@刘大佬的小跟班</div> <div>你好，你如果用的是JPMMML-SkLearn里面sklearn2pmml，那么就是支持adaboost的 https://github.com/jpmml/jpmml-sklearn#features 你确认下你的python版本和这个库的版本，看看github的官方文档，确认下具体报错是什么？</div>	支持(0) 反对(0)
#92楼 2020-10-19 15:18 刘大佬的小跟班	回复 引用
<div>所有错误显示链接太多了，发不了，我不知道啥原因，我自己写了一个转换标签的函数，放在apply里面，不知道会不会影响，你有没有解决方法，我邮箱2927322796@qq.com，老师能不能联系下我，</div>	支持(0) 反对(0)
#93楼 [楼主] 2020-10-20 09:57 刘建平Pinard	回复 引用
<div>@刘大佬的小跟班</div> <div>你好，我看了你用的是adaboost的分类器，而sklearn2pmml只支持adaboost的回归器，所以你的确是用不了，可以换个模型再搞。 https://github.com/jpmml/jpmml-sklearn#features 这个库里面的集成学习算法只支持这些： Ensemble Methods: ensemble.AdaBoostRegressor ensemble.BaggingClassifier ensemble.BaggingRegressor ensemble.ExtraTreesClassifier ensemble.ExtraTreesRegressor ensemble.GradientBoostingClassifier ensemble.GradientBoostingRegressor ensemble.HistGradientBoostingClassifier ensemble.HistGradientBoostingRegressor ensemble.IsolationForest ensemble.RandomForestClassifier ensemble.RandomForestRegressor ensemble.StackingClassifier ensemble.StackingRegressor ensemble.VotingClassifier ensemble.VotingRegressor</div>	支持(0) 反对(0)
#94楼 2020-10-20 10:05 刘大佬的小跟班	回复 引用
<div>就是，昨天在Github得到了同样的回复，感觉自己做了无用功，，， 我想请教下老师，对于只有一个属性的特征（但该属性下的变量很多，一千多个）实现2分类，采用啥子算法效果较好呢？还得部署在在Android端</div>	支持(0) 反对(0)
#95楼 [楼主] 2020-10-20 10:51 刘建平Pinard	回复 引用
<div>@刘大佬的小跟班</div> <div>你好，只有一个特征的确有点囧。个人建议你再多搞几个特征再去建模，否则你的模型估计预测效果会很差的。 你可以尝试下朴素贝叶斯看看效果。另外也可以直接去基于统计做二分类，说不定比你建立机器学习模型的效果更好。</div>	支持(0) 反对(0)

发表评论

编辑 预览

B    

支持 Markdown

提交评论

退出

订阅评论

[Ctrl+Enter快捷键提交]

[首页](#) [新闻](#) [博问](#) [专区](#) [闪存](#) [班级](#)

【推荐】超50万行C++/C#：大型组态工控、电力仿真CAD与GIS源码库

【推荐】赋能开发者，葡萄城，全球领先的软件开发技术提供商

【推荐】未知数的距离，毫秒间的传递，声网与你实时互动

相关博文：

- [scikit-learn中的KMeans](#)
 - [scikit-learn与数据预处理](#)
 - [机器学习实战_基于Scikit-Learn和Tensorflow读书笔记](#)
 - [adaboost](#)
 - [Codecademy-LearnHTML](#)
- » [更多推荐...](#)

最新 IT 新闻：

- [正式“退休”的Flash，未来我们会怀念它吗？](#)
 - [王力接任陌陌CEO一职 唐岩为何作出这个选择？](#)
 - [唯品会，模仿京东，却无法成为京东](#)
 - [新东方冲刺港交所：抢先好未来 成最早回归港股教育企业](#)
 - [老人不会用智能手机，就活该被淘汰吗？](#)
- » [更多新闻...](#)