Sklearn-随机森林

上一章中学习了决策树，这次我们继续来学习随机森林相关的内容。

讨论随机森林之前，我们需要先了解一下集成算法

1. 集成算法
2. 基本概念

集成算法会考虑多个评估器的建模结果，汇总之后得到一个综合的结果，以此来获取比单个模型更好的回归或分类表现。

多个模型集成成为的模型叫做**集成评估器**，组成集成评估器的每个模型都叫做**基评估器**

1. 集成算法类别

袋装法（Bagging）、提升法（Boosting）和Stacking



**袋装法的代表模型就是随机森林**

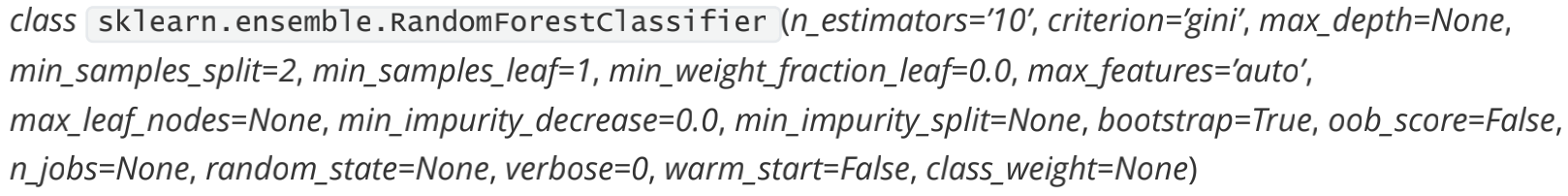
1. sklearn中的集成算法

sklean中的集成算法在ensemble 模块下



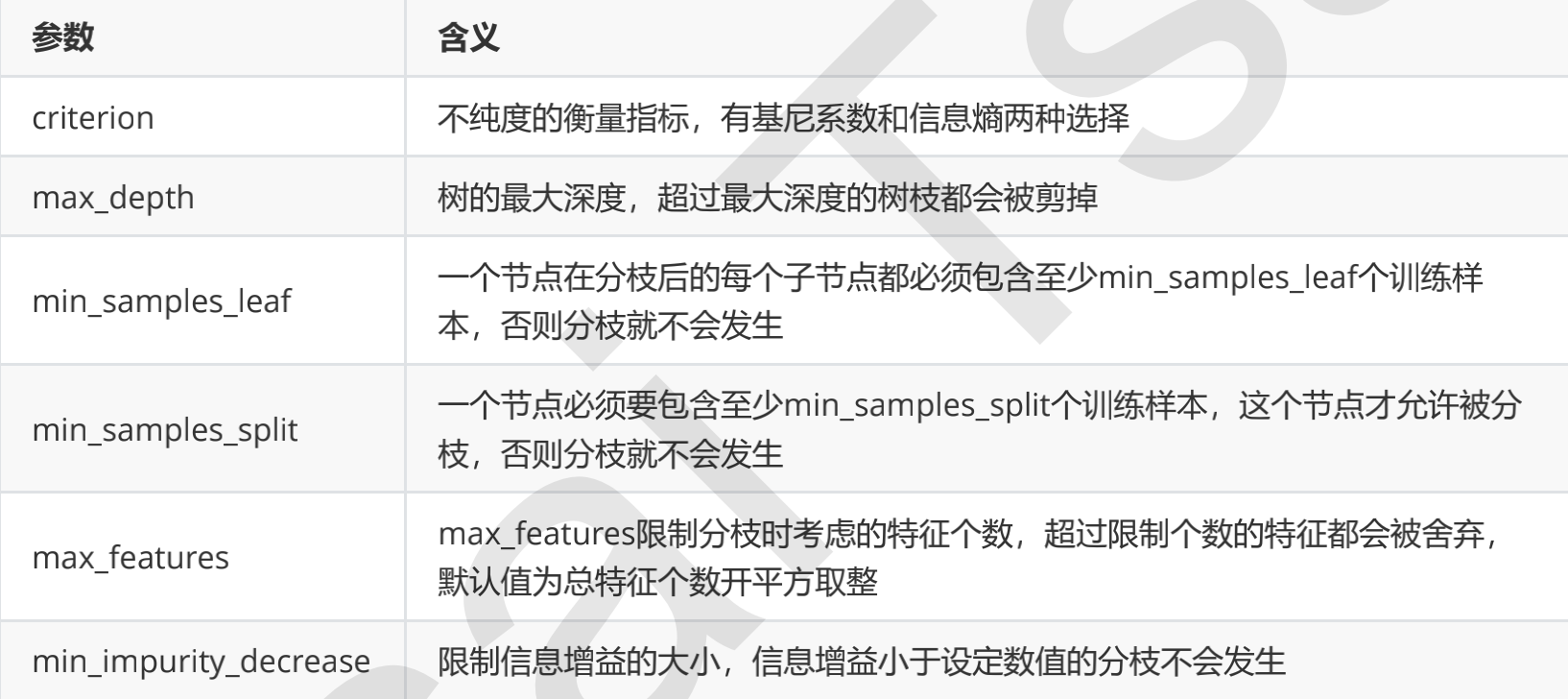
集成算法中，有一半以上都是树的集成模型

1. 随机森林分类RandomForestClassifier
2. sklearn中的RandomForestClassifier



**随机森林是非常具有代表性的Bagging集成算法，它的所有基评估器**都是决策树

1. 重要参数
2. 控制基评估器的参数



这些参数的详细解释，在上一章中已详细介绍，如果需要了解，可以关注该公众号进行查阅。

1. n\_estimators

* 这个参数决定森林中树木的数量，即**基评估器的数量**
* 通常来说，这个参数越大，模型越精确，当然了，当达到一定阈值之后，模型的精确率不在上升，甚至开始波动。

1. random\_state

* 和决策树类似，这个参数**决定随机森林的生成模式**，是控制生成一片森林，而不是让一个森林中只有一棵树
* 当random\_state固定时，随机森林中生成一组固定的树，但是每棵树依然是不一致的

一般来说，随机性越大，袋装法的效果越好，随机森林中，除了random\_state，还有其他控制随机性的参数，比如bootstrap

1. Bootstrap & oob\_score

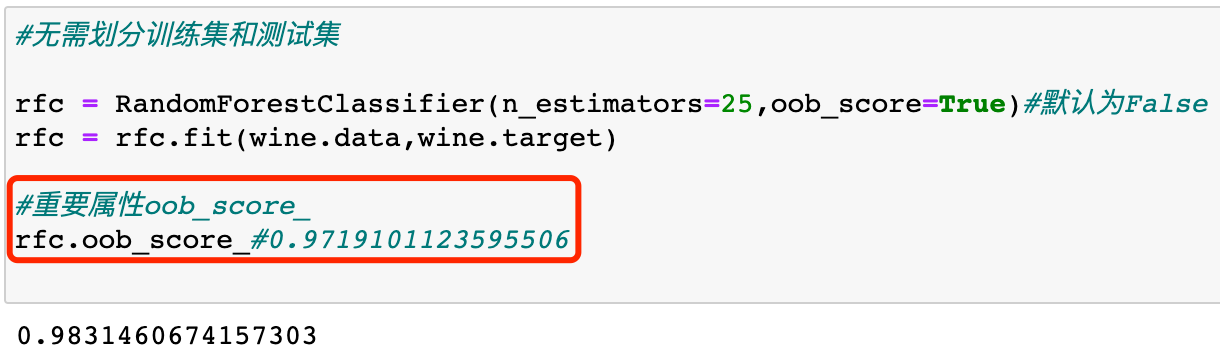
* 要让基分类器尽量不一样，一种很容易理解的方法是使用不同的训练集来进行训练，而袋装法就是通过有放回的随机抽样技术来形成不同的训练数据，**boostrap就是用来控制抽样技术的参数**
* boostrap参数默认为True，代表采用这种有放回的随机抽样技术，通常，我们不会将这个参数设置为false。

1. 重要属性
2. estimators\_

查看基分类器信息

1. oob\_score\_

随机森林的这种有放回抽样，会产生袋外数据，oob\_score就是用来查看模型在袋外数据上的测试结果，具体使用如下：



1. feature\_importances\_

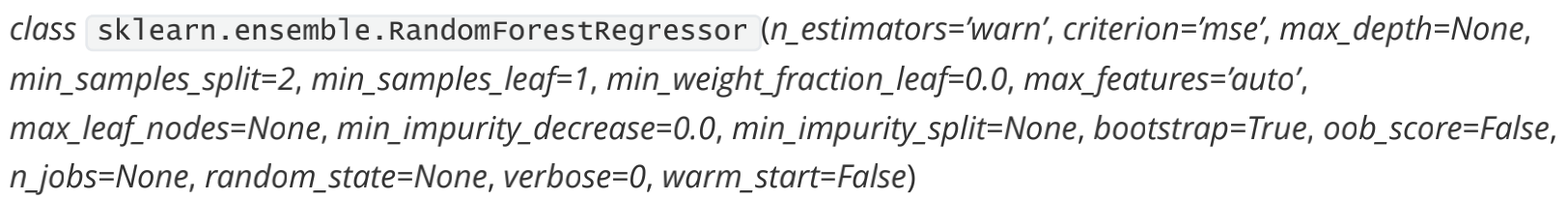
和决策树一样，用于输出重要特征

1. 重要方法

和决策树一样，常用的依然是**apply、fit、predict 和 score** 这四个方法。

不过随机森林中，也需要注意**predict\_proba**这个方法，它返回每个测试样本对应的被分到每一类标签的概率。

1. 随机森林回归RandomForestRegressor
2. sklearn中的RandomForestRegressor



1. 重要参数
2. Criterion

这个参数的具体内容和决策树中一致，这里不再赘述，如需了解，请关注本公众号。

1. 重要属性

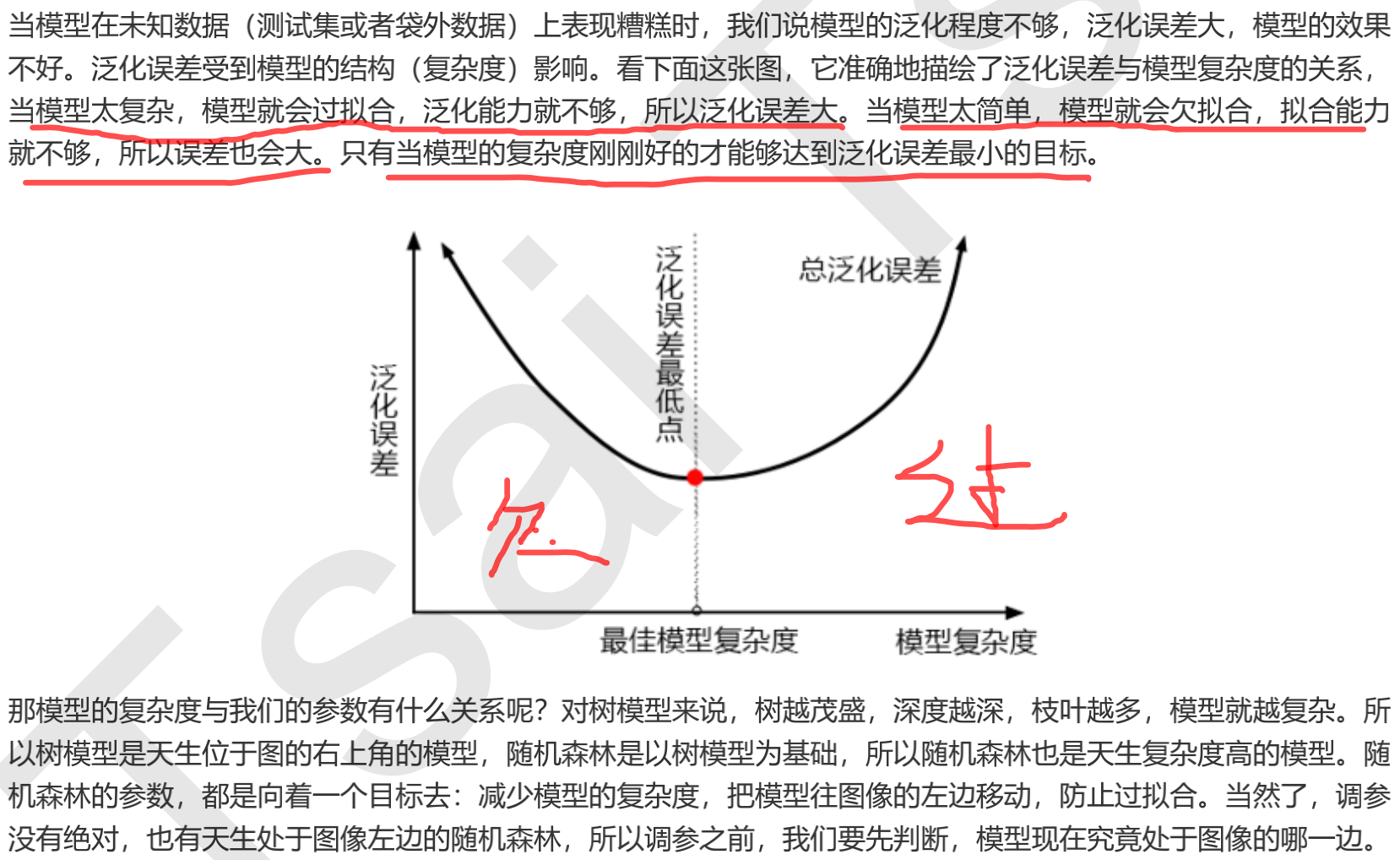
和随机森林分类器保持一致

1. 重要方法

和随机森林分类器保持一致，不过回归树中，没有predict\_proba这个方法

1. 机器学习中调参的基本思想

在机器学习中，我们用来衡量模型在未知数据上的准确率的指标，叫做**泛化误差**



所以在调参时，通过确定模型到底是在最佳模型的左边还是右边，从而决定是将模型复杂化还是简单化，从而针对性的调整对应的参数。

1. 代码示例

请移步github：<https://github.com/codingSally/sklean-learn1>

好了，随机森林就分享到这里，希望对大家有帮助，后续我也会定时更新其他内容，如果觉得对您有帮助，请关注我的公众号，我们一起学习、交流、进步。