Sklearn-决策树

最近在学习Sklearn的具体使用，所以计划输出模型相关的类以及针对每个模型的重要参数做简单介绍，同时也会给出代码示例。

1. **sklearn中的决策树**

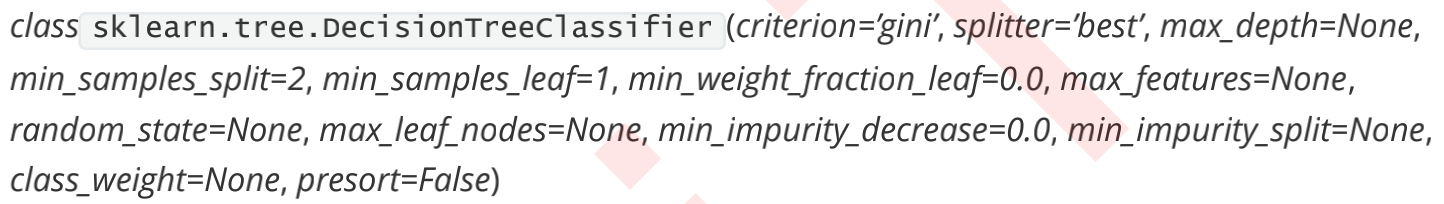
Sklearn中的决策树类都在“tree”这个模块下，这个模块主要包含五个类：



这篇文章中，我们主要介绍的是分类和回归树。

1. **分类树-DecisionTreeClassifier**

2.1 sklearn中的DecisionTreeClassifier



2.2 重要参数

1. Criterion

* 背景知识

在训练模型的过程中，决策树需要找出**最佳节点和最佳的分支**方法，对于分类树来说，这个“最佳”的衡量指标叫做不纯度。那么这个不纯度，如何计算呢？通常来说有两种计算方式：信息熵和基尼系数，而Criterion这个参数就是来指定不纯度的计算方法。

* 参数类别

当设置为“entropy”时，表示使用**信息熵**。

当设置为“gini”，表示使用**基尼系数**。

* 参数选择

**默认使用基尼系数**

**当数据维度很大，噪音很大的时候，使用基尼系数**（因为比起基尼系数，信息熵对不纯度更加敏感，所以信息熵作为指标时，决策树的生长会更加精细，从而很容易导致过拟合，尤其在高维数据或噪音很多的时候）

**维度低、数据比较清晰的时候，两者没区别**

**当决策树的拟合程度不够时，可以使用信息熵**

* 多说几句

不纯度越低，决策树对训练集的拟合越好，现在使用的决策树算法在分支方法上的核心大多数是围绕某个不纯度相关指标的最优化上。

1. random\_state & splitter

* random\_state

该参数用来设置分枝中的随机模式，默认为None

**输入任意整数，会一直长出同一颗树，让模型稳定下来**

在高维度时，随机性会表现更明显，低维度数据时，随机性几乎不会出现

* splitter

该参数也是用来控制决策树的随机性，有两种输入值

**输入“best”**，决策树在分支时，虽然随机但还是会选择更加重要的特征进行分支。

**输入“random”**，决策树在分支时更加随机

* 多说几句

当预测到模型会过拟合的时候，可以用这两个参数来降低模型过拟合的可能性

1. 剪枝参数

当模型训练好之后，可以通过相应的剪枝策略来防止过拟合，正确的剪枝策略，是优化决策树算法的核心，Sklearn为我们提供了下面几种剪枝策略。

* max\_depth

这个参数**限制树的最大深度**，当超过设定深度的树枝全部剪掉

这是用的最广泛的参数，实际使用时，**建议从3开始尝试**，观察拟合效果从而决定是否增加设定的深度

* min\_samples\_leaf

这个参数限定，一个节点在分支后的每个子节点都必须包含至少

min\_samples\_leaf个训练样本，否则分支就不会发生，或者分支会朝着满足每个子节 点都包含min\_samples\_leaf个样本的方向去发生。

这个参数设置的太小，会引起过拟合，设置的太大又会阻止模型的学习，**一般建议从5开始使用**

一般搭配max\_depth使用

* min\_samples\_split

这个参数限定，一个节点必须要包含至少min\_samples\_split个训练样本，这个节点才允许被分枝，否则分枝就不会发生

* min\_features

这个参数限制分枝时考虑的特征个数，超过限制个数的特征都会被丢弃

这个参数是强行使决策树停下的参数，在不清楚各个特征重要性的情况下，强行停止可能会导致模型欠拟合。

* min\_impurity\_decrease

这个参数限制信息增益的大小，信息增益小于设定数值的分支不会发生

1. 目标权重参数

* class\_weight

有时候会存在样本不均衡的问题，所以会使用class\_weight参数对样本进行一定的均衡，给少量样本更多的权重，让模型更偏向于少数类。

参数模型为None，表示自动给予数据集中所有类别相同的权重。

* min\_weight\_fraction\_leaf

输入的样本有了权重之后，对应的剪枝，也需要使用这个基于权重的剪枝参数。指的是在剪枝过程中更少的偏向于主导类。

2.3 重要属性

属性是在模型训练之后，能够调用查看模型的各种性质，对决策树来说，最重要的是feature\_importances\_, 能够查看各个特征对模型的重要性。

2.4 重要方法

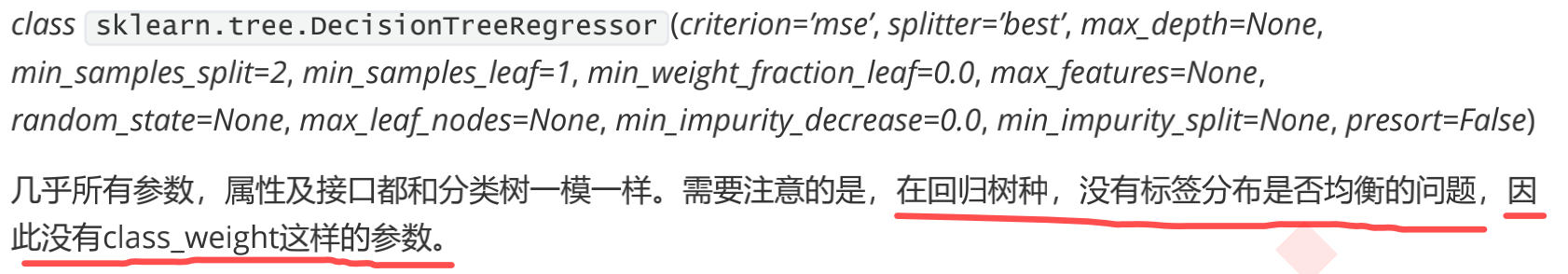
sklearn中许多算法的方法都是相似的，比如fit、socre。决策树中，经常还会用到apply和predict两个方法。其中，

apply返回每个测试样本所在叶子节点的索引

predict返回每个测试样本的分类/回归结果

1. **回归树-DecisionTreeRegressor**

3.1 sklearn中的DecisionTreeRegressor



3.2 重要参数

1. Criterion

回归树衡量分枝质量的指标，有如下三种



**MSE的本质，是样本真实数据和回归结果的差异**

**在回归树中，MSE不仅是衡量分枝质量的指标，也是最常用的衡量回归树回归质量的指标**

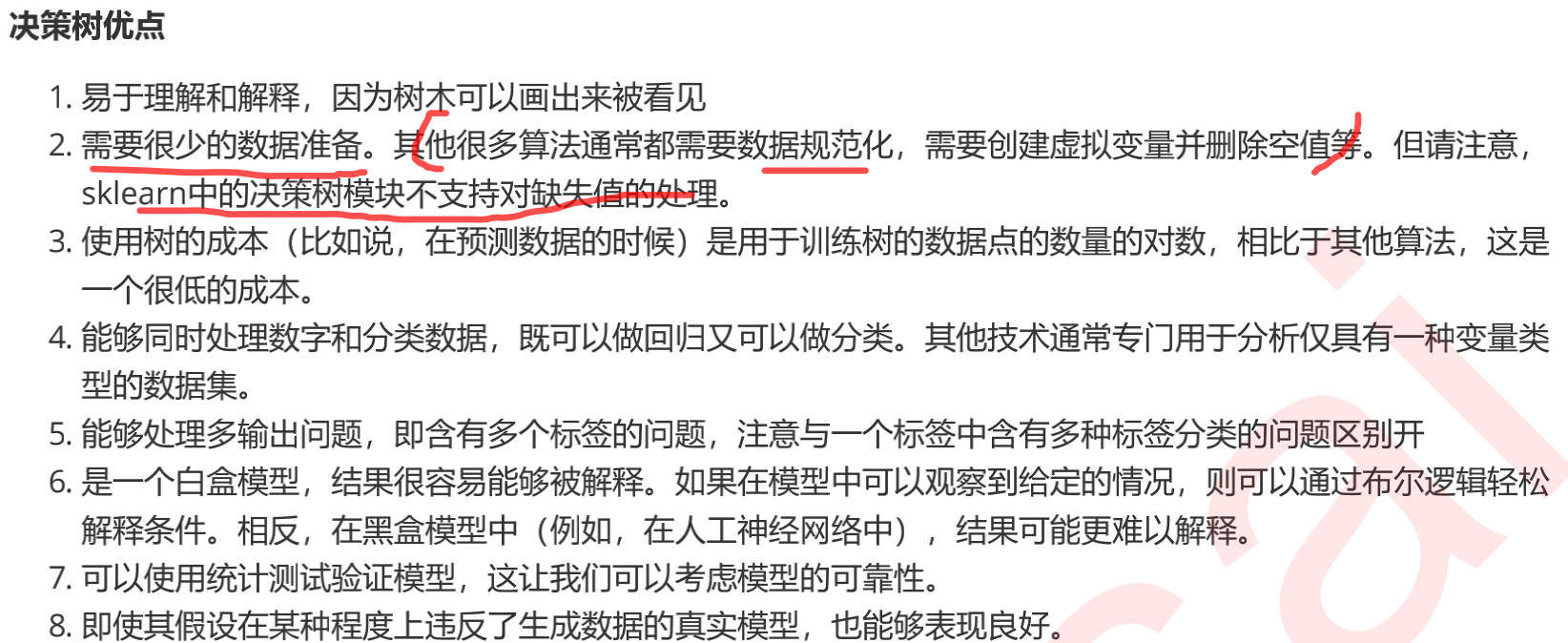
3.3 重要属性

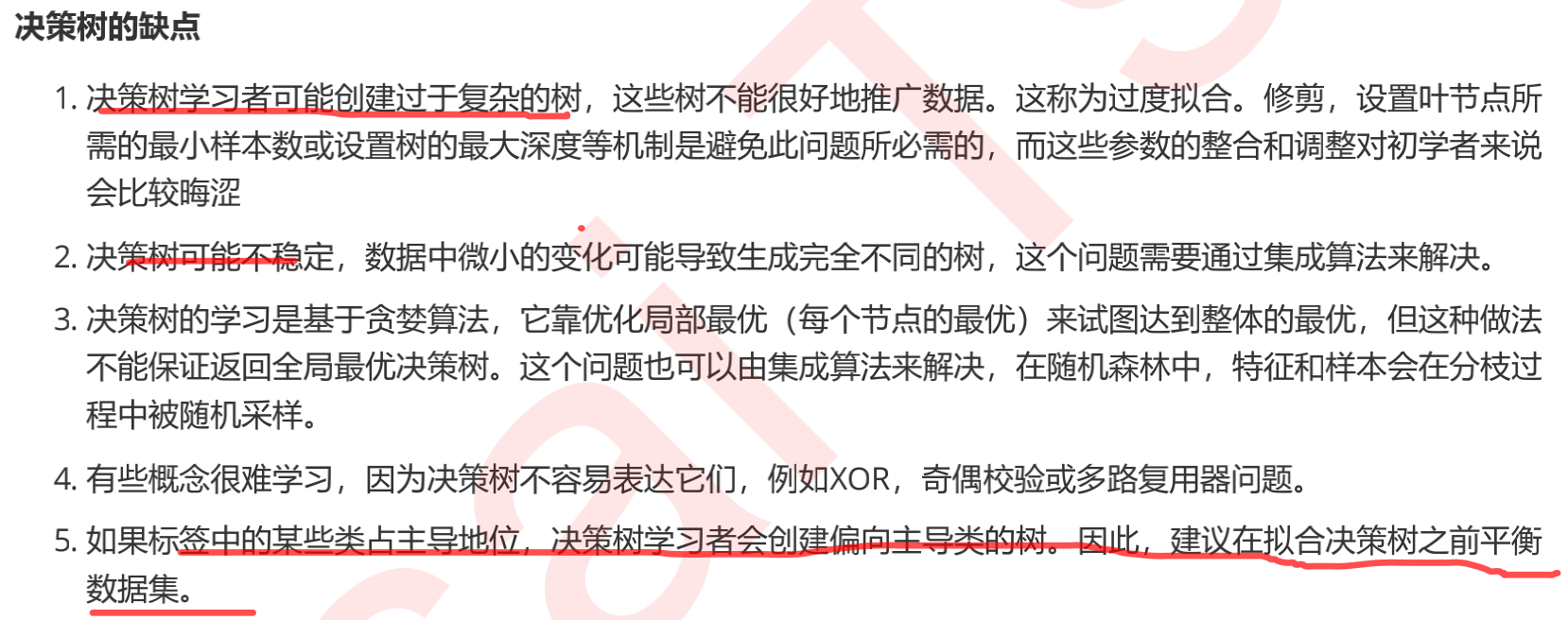
重要属性依然是feature\_importances\_

3.4 重要方法

重要方法依然是fit、socre、apply 和 predict比较常用

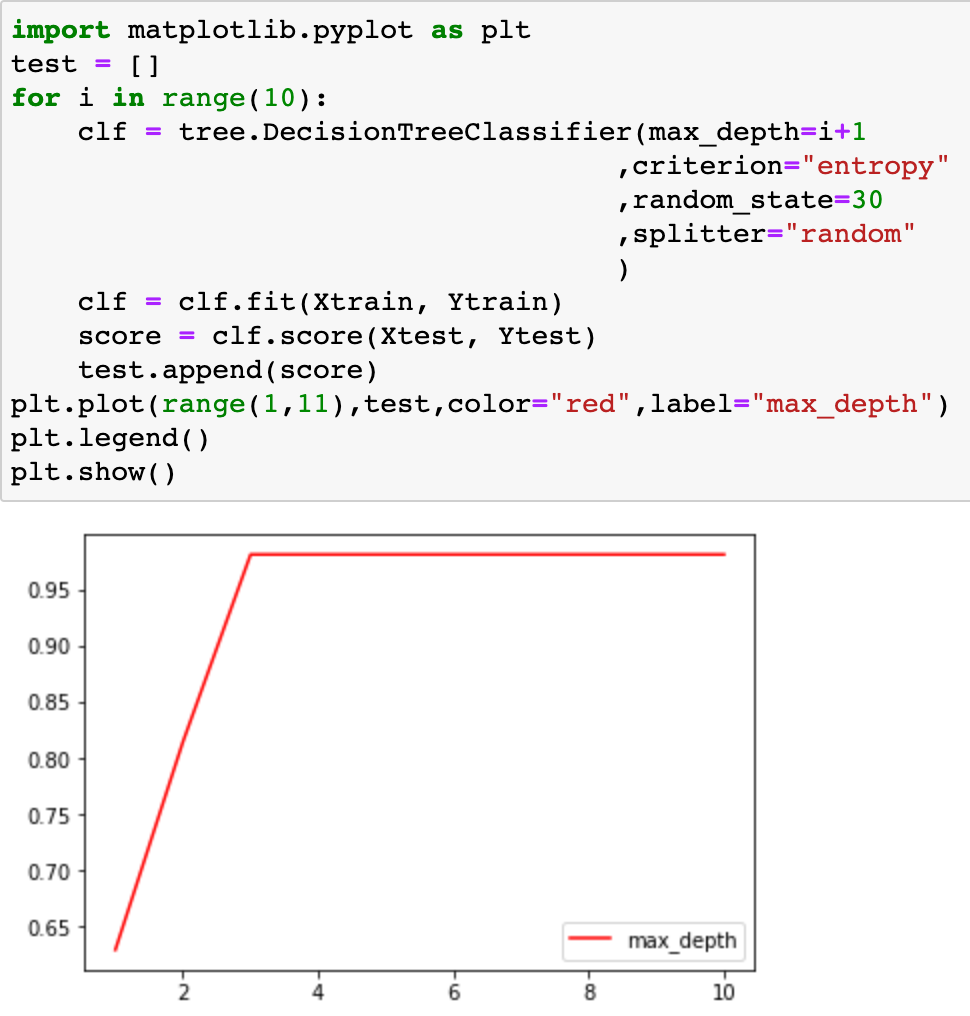
1. **决策树的优缺点**





1. **学习曲线**

上面介绍了各种可以调节的参数，那么针对每个参数，我们如何确定选择哪个好呢？这个时候，就可以通过绘制学习曲线来达到参数选择的目的，比如，我们针对上面的max\_depth来绘制学习曲线



**可以发现，当max\_depth设置到3或4的时候，模型已经达到最优状态**

1. **代码示例**

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

好了，决策树就分享到这里，后续我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。