CART分类树：

分类树的建树的一般步骤：

1. 读取数据集
2. 进行训练
3. 进行验证

1. 读取数据集—要有简单的可视化
2. 对特征进行处理
3. 进行训练
4. 进行验证

CART分类树：

将离散型特征变成连续性特征

```
import pandas as pd
pd.get_dummies()
#将离散特征变成多列连续特征
```

4 回归树：

1. mean_squar_error(groups) 最小均方误差#修改
2. split(data,index,value) 划分数据集
3. get_split(data) 获得最优特征与二分标准**##修改**
4. toLeafNode(labelList) 变成叶子节点#需要修改的
5. createTree(data,max_depth,min_size,depth,stop) 递归建树，每个节点记录{index，value，left，right，stop} #需要添加stop
6. predict(tree ,example) 递归解树，进行预测
7. m_s_e(tree, test_data, test_label)计算预测的均方误差

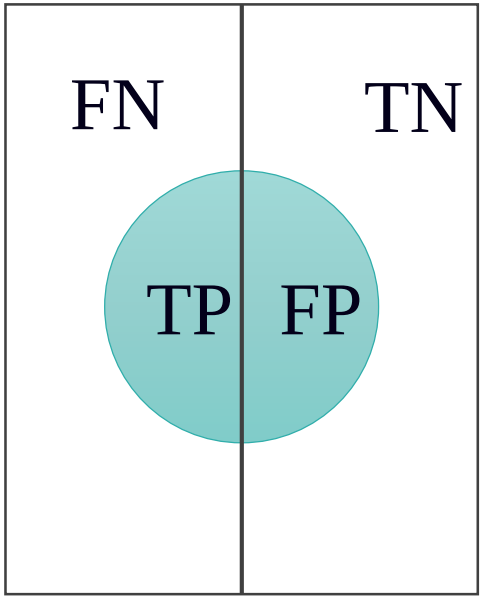
4.1 水雷-岩石数据

将水雷作为数值0，岩石作为数值1进行回归
然后预测值小于0.5是水雷，大于0.5是岩石

预测值小于0.4是水雷，大于等于0.4是岩石
可以吗？
预测值小于0.6是水雷，大于等于0.6是岩石
可以吗？

5 准确率、精确率和召回率

预测结果	真实结果	
	正类	负类
正	TP	FP
负	FN	TN



(总体)准确率 (accuracy) = (TP+TN)/(TP+FN+FP+TN)

(正类的)精确率 (precision) = TP/(TP+FP)

(正类的)召回率 (recall) = TP/(TP+FN)

5.1 真正率和假正率

	真实结果		
预测结果		正类	负类
	正	TP	FP
	负	FN	TN

真正率
True positive rate

$$TPR = \frac{TP}{TP + FN}$$

假正率
False positive rate

$$FPR = \frac{FP}{FP + TN}$$

假负率 False negative rate

$$FNR = \frac{TP}{TP + FN}$$

真负率Ture negative rate

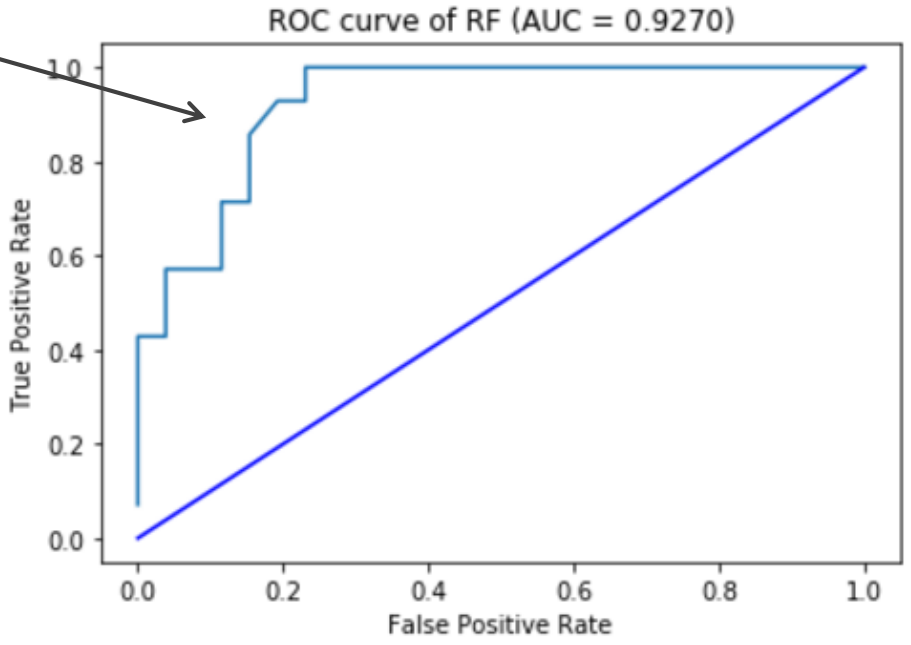
$$TNR = \frac{FP}{FP + TN}$$

5.2 ROC曲线

受试者工作特征曲线(receiver operating characteristic curve，简称ROC曲线)，又称为感受性曲线(sensitivity curve)

ROC曲线
横坐标是FPR
纵坐标是TPR

1. 以连续值(概率)表示分类结果
2. 截断点或阈值



5.2 截断点

ROC曲线不容易理解主要是因为TPR和FPR的关系隐含着”**截断点**”。

机器学习算法对D样本集进行预测后，可以输出各样本对某个类别的**置信度**。比如d1是P类别的概率为0.3，一般我们认为概率低于0.5，d1就属于类别N。这里的0.5，就是”截断点”。

当“截断点”取值不同时，TPR和FPR是随之变化的

真实	1	1	1	0	0	0	1	0	0	0
预测	0.49	0.94	0.56	0.05	0.53	0.27	0.9	0.58	0.32	0.21
断点：0.25	1	1	1	0	1	1	1	1	1	0
断点：0.5	0	1	1	0	1	0	1	1	0	0
断点：0.8	0	1	0	0	0	0	1	0	0	0

5.2 截断点

截断点取不同的值，TPR和FPR的计算结果也不同。将截断点不同取值下对应的TPR和FPR结果画于二维坐标系中得到的曲线，就是ROC曲线。

截断点	TP&FN	FP&TN	TPR&FNR	FPR&TNR
0.25	4	4	**1**	**0.67**
	0	2	0	0.33

TP是预测为**1**，真实也为**1**; FP是预测为**1**，但真实为**0**
FN是预测为**0**，但真实为**1**； TN是预测为**0**，真实也为**0**

0.5	3	2	**0.75**	**0.33**
	1	4	0.25	0.67

5.2 截断点

截断点	TP&FN	FP&TN	TPR&FNR	FPR&TNR
0.25	4	4	**1**	**0.67**
	0	2	0	0.33
0.5	3	2	**0.75**	**0.33**
	1	4	0.25	0.67
0.8	2	0	**0.5**	**0**
	2	6	0.5	1

对不同截断点情况总结

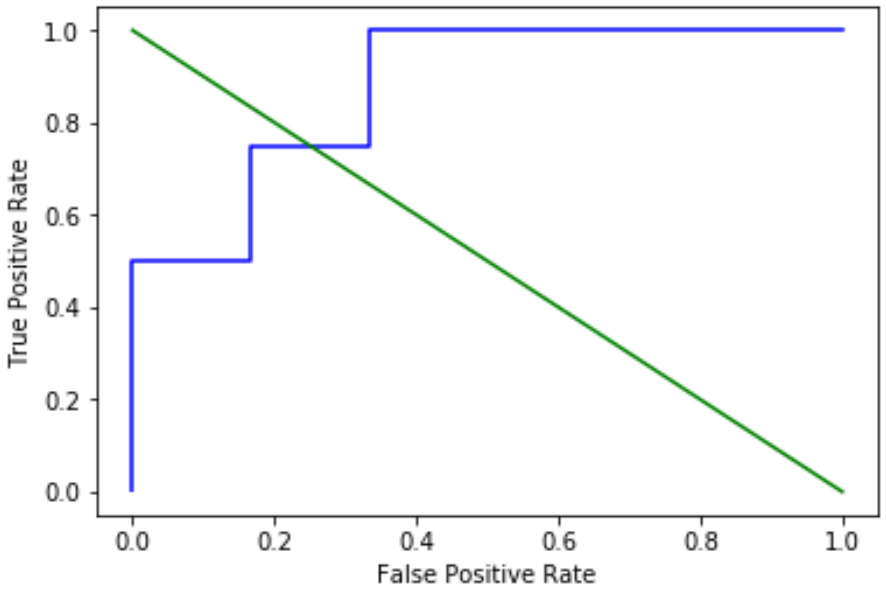
对每一种情况：TP+FN都是相等的，都等于真实的正类数

对每一种情况：TPR+FNR=1

随着截断点的增大，TPR和FPR只减不增

2.2 ROC曲线

截断点	FPR	TPR
0	**1**	**1**
0.25	**0.67**	**1**
0.5	**0.33**	**0.75**
0.8	**0**	**0.5**
1	**0**	**0**

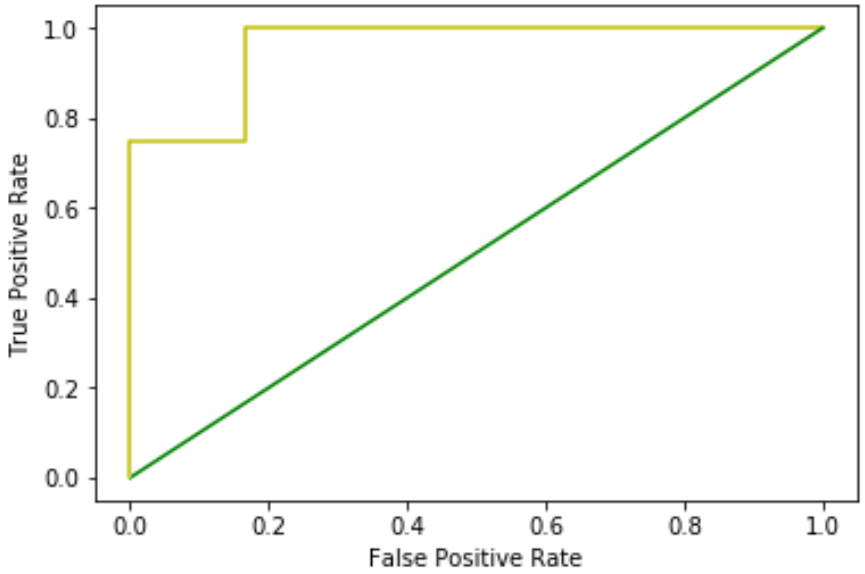


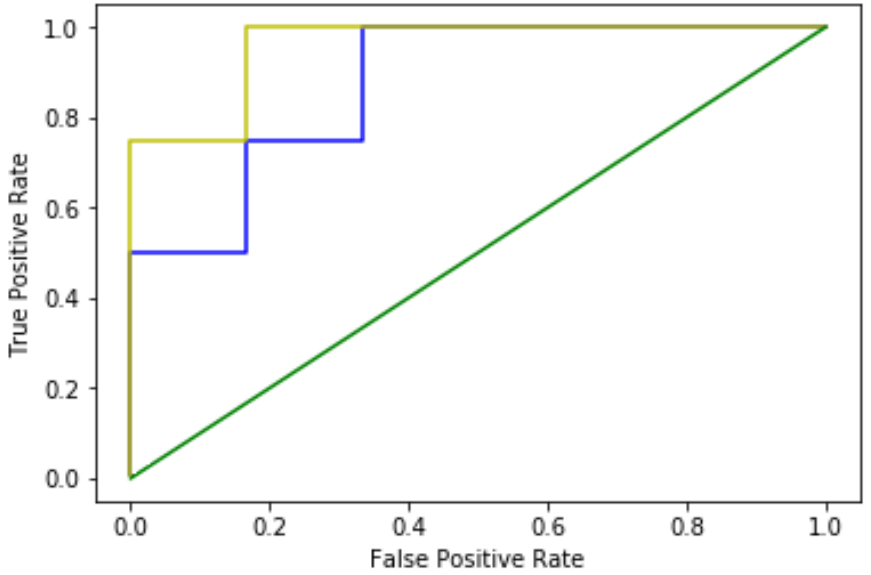
1. 截断点选哪个值，是最优的
一般情况正负样本的均衡，要求假正率(FPR)等于假负率(FNR)是合适的截断点选取位置。
2. ROC曲线越光滑，用模型对新的测试样本的分类效果越稳定

2.3 AUC值

如何通过ROC曲线判断不同模型的预测结果的优劣

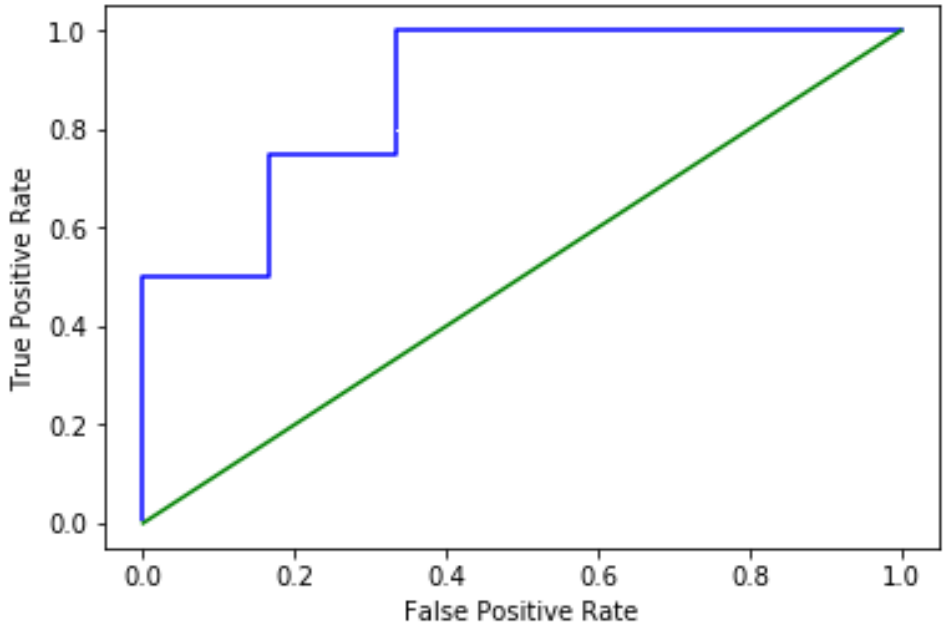
真实	1	1	1	0	0	0	1	0	0	0
预测1	0.49	0.94	0.56	0.05	0.53	0.27	0.9	0.58	0.32	0.21
预测2	0.75	0.52	0.83	0.30	0.13	0.48	0.40	0.14	0.38	0.30





2.3 AUC值

使用AUC值判断两个二分类模型的效果
AUC是Area Under Curve的首字母缩写，就是ROC曲线下区域的面积
假设分类器的输出是样本属于正类的socre(置信度)，则AUC的意义为，任取一对（正、负）样本，正样本的score大于负样本的score的概率



AUC值越大，模型越好
AUC = 1，是完美分类器，
0.5 < AUC < 1，
AUC = 0.5，跟随机猜测一样（例：丢铜板），模型没有预测价值。
一般模型的AUC值处于