## 代码运行

[点击查看原图](a.png)

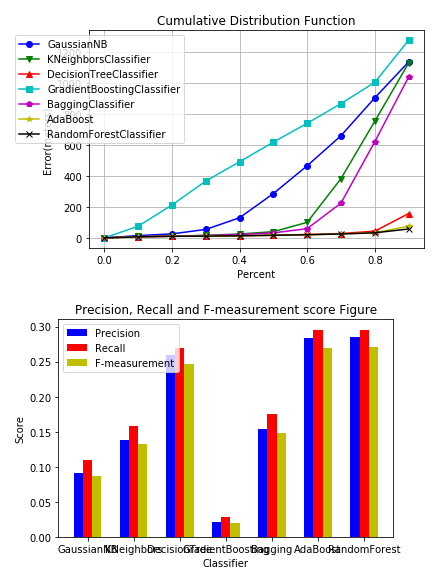
（截图过大此处无法载入，请直接浏览原图）

### 讨论分析

1. 本题使用GaussianNB、KNeighborsClassifier、DecisionTreeClassifier、AdaBoostClassifier、BaggingClassifier、RandomForestClassifier和GradientBoostingClassifier共七种分类器对GPS及信号强度数据进行处理，最终分别获得七种分类模型并进行比较。
2. 其中数据集被随机分为0.8的训练集和0.2的测试集，两者数据不存在重叠，这一点在之后的所有小题中都相同。
3. 首先对七种分类器的特点进行比较：
   1. GaussianNB：借助贝叶斯概率公式直接计算“条件 - 结果”概率
   2. KNeighborsClassifier：使用KNN算法直接计算样本距离，获得一组数字上合理的近邻分类，和前者贝叶斯一起是暴力计算的方法，虽然缺少一定的逻辑支持但是具有不错的效果和性能。
   3. DecisionTreeClassifier：针对多个条件构造基于Y/N的树形分类，通过多重条件判断来确定最后的分类，相较于前两者具有更好的解释性
   4. AdaBoostClassifier：较为直接的集成方法，通过用同一组数据训练多个不同的分类器并进行集合来获得更好的预测结果
   5. BaggingClassifier：在AdaBoost的基础上增加了预测序列和投票机制，构造一个预测函数的系列，并按照投票的结果进行组合，以期获得最好的结果。比Adaboost相比是有选择的集成
   6. RandomForestClassifier：分裂多棵决策子树来复杂化决策过程，以在Y/N决策的基础上获得更为复杂的组合来预测更多的情况
   7. GradientBoostingClassifier：通过多次迭代并从中选优的过程，来从多个弱分类器获得一个较强的分类器。

基于以上特点，七种分类器的性能和效果比较实际就已经能看出一部分了。

1. 具体结果可见代码运行截图最下方，此处补贴一份：



* 1. 分析可见，预测效果为RandomForestClassifier > Adaboost > DecisionTreeClassifier > BaggingClassifier > KNeighborsClassifier > GaussianNB > GradientBoostingClassifier，基本符合之前我们对于七种分类器的分析。
  2. 但是此处存在一个异类，GradientBoostingClassifier的误差明显高于其他六种分类器，并且以一种近乎直线的形态直逼随机猜测下会产生的直线结果，可认为GradientBoostingClassifier在此处并未起到集成学习的效果并且甚至存在反效果。并且这一分类器的训练在此处花费了极长的时间，约在4.5h/训练十次的水平，这一数据将在下面的图表中被除去。
  3. 查找相关资料，GradientBoostingClassifier在过多标签的分类过程中表现极差。而在此处我们的栅格约有5k多个标签，符合这种情况，得到解释。
  4. 第二张图表中，七种分类器的三种评分均未达到0.30以上，此处原因和c点相同，由于大量标签的存在，分类结果极小概率恰好落在对应栅格上，而大概率落在四周的范围中，这一问题并不影响实际的距离误差，将在乘上栅格宽度后大幅度缩小。
  5. 此处RandomForestClassifier，Adaboost，DecisionTreeClassifier三种分类器表现出了极好的效果，中位误差极小。

### 性能比较

前三种分类器时间极短，可打开图标数据查看具体数据，均在1s以下。