# 一、实验目的

1、理解分类的一般过程和基本原理；

2、巩固分类算法的算法思想，能够进行分类操作；

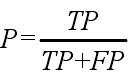
3、学会分类预测问题中的性能评估方法。

# 二、实验原理

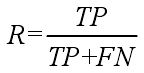
1、常用的预测模型：决策树、朴素贝叶斯分类器和支持向量机（SVM）等。

2、评价预测结果，常用的性能评价指标：

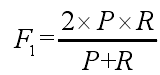
（1）**精确率**P(Precision)是指被正确预测为正例数与所有被预测为正例数的比率，反映了预测模型的准确程度，也称为查准率。



（2）**召回率** R(Recall)是指被正确预测为正例数与实际正例数的比率，反映了一个有缺陷模块被正确预测出的概率，也称为查全率。



（3）**F-Measure** 是信息检索领域的一个评价指标，常用的是F1 度量，即为精确率与召回率的调和平均数。

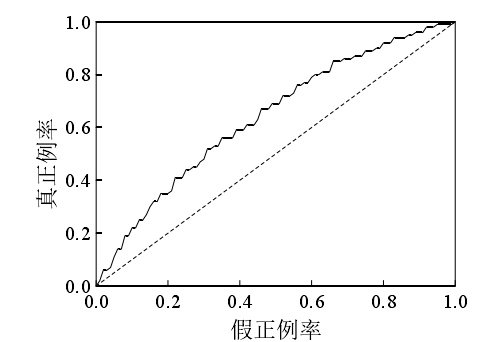


（4）**AUC**

真正例率 TPR(True Positive Rate)与召回率相同，也是指被正确预测为正例数与实际正例数的比率。

假正例率 FPR(False Positive Rate)是指被错误预测为正例数与实际负例数的比率。

接受者操作特征(Receiver Operating Characteristic，ROC)曲线是描述分类模型真正例率 TPR 和假正例率 FPR 之间关系的一种图形化方法，如下图所示。



横坐标表示假正例率，纵坐标表示真正例率。对于一个特定的预测模型和训练数据集，其预测结果对应于 ROC 曲线上的一个点，通过调整该模型的阈值即可得到一条经过(0, 0)和(1, 1)的曲线，曲线下方的面积即为AUC(Area Under the Curve)的值。其中，AUC的取值范围为0~1，当 AUC 为 0.5 时，表示随机猜测模型的性能，如上图中的虚线所示。AUC 值越大，说明该模型的性能越好。因此，好的预测模型应尽可能地靠近坐标系的左上角。

# 三、实验内容：

选取12个 NASA 数据集作为实验对象，采用不同的分类方法分别对这12个数据集进行软件缺陷预测分析。（数据见data文件夹）

## 1. 预处理：

**选出与类别相关度较高的特征子集**：采用上机1 中数据预处理方法，获得特征排序列表。在特征排序列表的基础上，按照顺序依次进行特征子集选择。例如，得到的特征排序列表为，d表示特征个数，排序越靠前，说明该特征与类别的相关程度越高。由此，根据特征排序顺序可以得到d个特征子集，分别为，最后评价所有特征子集的分类性能，并据此选取与类别相关度较高的特征子集，并进一步构建分类预测模型。

## 2. 预测模型方法

分类器构建预测模型：（1）决策树 （2）Naive Bayes (NB) （3）SVM

可在以上分类器中**选择2个**或者**3个**分类器构建预测模型。

## 3. 性能评价

选取**F-Measure**或**AUC**作为性能评价指标；

采用 10 次 10 折交叉验证，保证实验结果的准确性和可靠性。