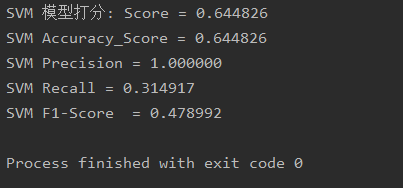
1 svm效果

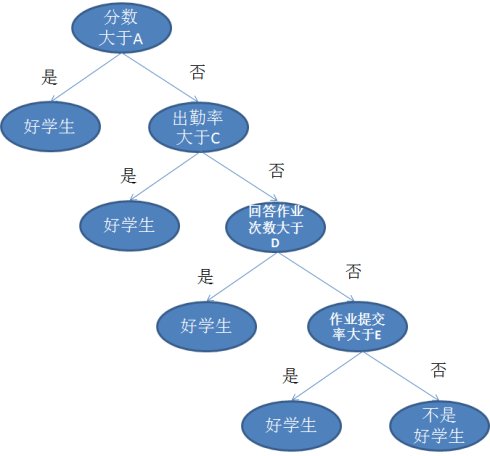
·训练速度慢

·准确率低



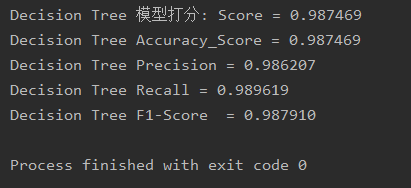
2决策树

以if else的形式不断判断，从而将数据分为多种的分类方法



优点：

(1)速度快: 计算量相对较小, 且容易转化成分类规则. 只要沿着树根向下一直走到叶, 沿途的分裂条件就能够唯一确定一条分类的谓词.  
(2)准确性高: 挖掘出来的分类规则准确性高, 便于理解, 决策树可以清晰的显示哪些字段比较重要, 即可以生成可以理解的规则.  
(3)可以处理连续和种类字段  
(4)不需要任何领域知识和参数假设  
(5)适合高维数据

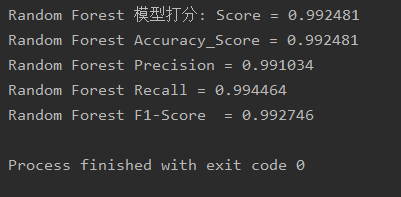


3 决策树集成

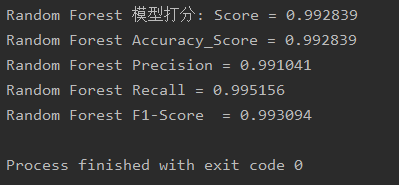
集成（ensemble）是合并多个机器学习模型来构建更强大模型的方法。在以决策树为基础上，有两种集成模型对大量分类的数据集都是有效的，二者分别是随机森林和梯度提升决策树

3.1随机森林

决策树的主要缺点在于经常对训练数据过拟合。随机森林算法的本质是多个决策树的集合，但其中每棵树都和其他树略有不同。如果每棵树的预测结果都相对较好，但又以不同的方式拟合，那么我们可以对这些树的结果取平均值来降低过拟合。



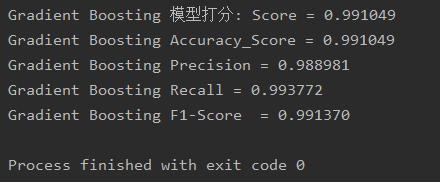
图为100棵树的训练模型



图为10000棵树的训练模型

3.2 梯度提升回归树

梯度提升回归树的主要思想是通过合并多个决策树来构建一个更为强大的模型。与随机森立不同的是，梯度提升采用的是连续的方式，每一棵树都试图纠正前一棵树的错误。梯度提升回归树通常使用深度为1-5之间的树，模型占用的内存更少，预测速度也更快。



图为100棵树，最大深度为3，学习率为0.1的训练模型

学习率为控制每一棵树纠正前一棵树的错误的强度

根据准确率，采用随机森林模型较为合适