**机器学习决策树文档**

1. 策略：

（1）、数据的离散化流程：首先数据归一化，然后以0.5为界限离散化数据。当value >= 0.5时，令value = 1，当value < 0.5 时，value = 0。

（2）、训练集与验证集比例：9：1

（3）、决策树最后结果通过文件保存于“txtOfmyTree.txt”

（4）、程序运行方法：编译运行后，在shell窗口输入rungo()，回车可分别执行程序。

（5）、各叶子节点代表标签，类别标签：1：极具魅力的人，2：魅力一般的人，3：不喜欢的人。

决策树结果：

{'每年飞行里程': {0.0: {'玩游戏时间': {0.0: {'冰激凌消耗量': {0.0: 2.0, 1.0: 2.0}}, 1.0: {'冰激凌消耗量': {0.0: 3.0, 1.0: 3.0}}}}, 1.0: {'冰激凌消耗量': {0.0: {'玩游戏时间': {0.0: 1.0, 1.0: 1.0}}, 1.0: {'玩游戏时间': {0.0: 1.0, 1.0: 1.0}}}}}}

决策树：

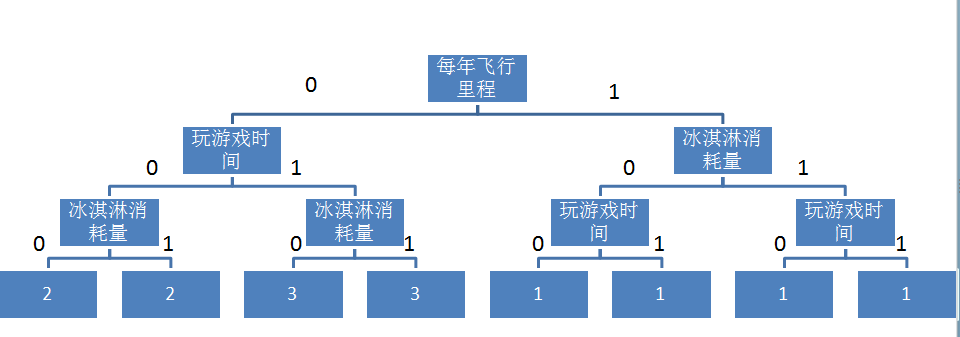


图1.决策树

验证集中的判断错误个数：num\_error: 19.0，精度为:。

1. 预剪枝决策树

预剪枝决策树决策树结果：

{'每年飞行里程': {0.0: {'玩游戏时间': {0.0: 2.0, 1.0: 3.0}}, 1.0: 1.0}}

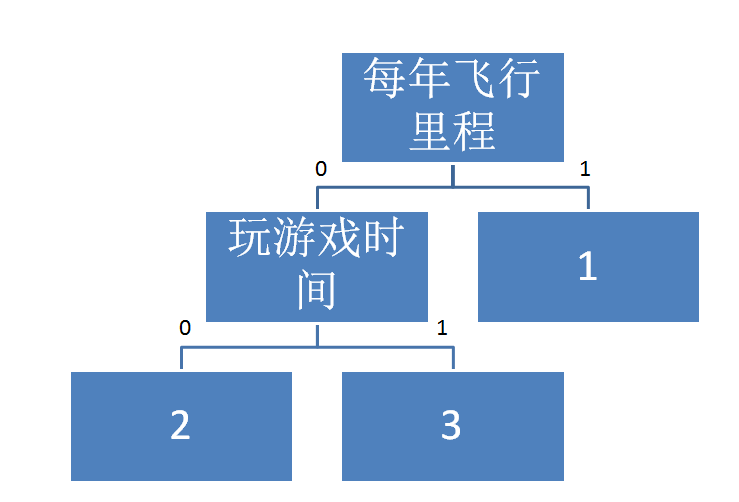


图2.预剪枝决策树

验证集中的判断错误个数：num\_error: 19.0，精度为:。

1. 后剪枝决策树

后剪枝决策树决策树结果：

{'每年飞行里程': {0.0: {'玩游戏时间': {0.0: {'冰激凌消耗量': {0.0: 2.0, 1.0: 2.0}}, 1.0: {'冰激凌消耗量': {0.0: 3.0, 1.0: 3.0}}}}, 1.0: {'冰激凌消耗量': {0.0: {'玩游戏时间': {0.0: 1.0, 1.0: 1.0}}, 1.0: {'玩游戏时间': {0.0: 1.0, 1.0: 1.0}}}}}}

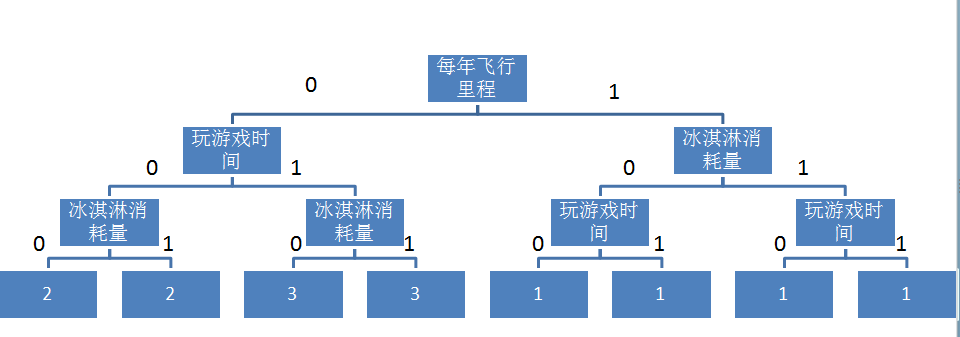


图3.后剪枝决策树

验证集中的判断错误个数：num\_error: 19.0，精度为:。

1. 结果分析

（1）、可以很明显看出最初构建的决策树与后剪枝的决策树是一样的，这是由于在后剪枝过程中没有出现减少验证集中错误率的子节点。

（2）上述三个决策树都没有提高精度！

（3）可以看出，飞行里程对个人魅力的影响最大，飞行里程越多的人往往魅力越高。