实验 1：基于决策树的英雄联盟游戏胜负预测实验报告

陈泓宇 2022/6/22

实验任务

本次实验的任务是：根据已有的对局前 10 分钟特征信息，预测最后获胜方是蓝色方还是红色方，了解执行一个**机器学习任务的大致流程**。

数据分析

在本次实验给出的数据中，有共 9879 组数据，每个数据对应有一个标签（blueWins）和 38 个特征。其中红队与蓝队特征各 19 项。

特征处理：

在所有特征中，队伍间的金钱差距、经验差距、前十分钟的每秒击杀、这些数值本身并不非常重要，而它们在红蓝两队间的差值比较重要。因此去掉这些特征，并加入它们的差值作为新的特征。

离散化特征时用到了 pd.qcut 函数。qcut 可以基于给定数组，将样本离散化为数据量相等的 bin。这里简单把除了特征中，除了击杀野怪/推塔数量的特征除外的其他数据警醒一个二分。

数据集划分

数据集划分使用 sklearn.model_selection（此处 import 为 train_test_split）函数，将原始数据集以 8：2 划分为训练集和测试集。

模型设计

决策树的基本核心循环为 5 步：

1. 从当前根节点出发，找到下一步的**决策 Feature A**
2. 将 A 作为当前节点的决策属性
3. 对属性 A（vj）的每个值，创建与其对应的新的子节点。
4. 根据属性值将训练样本分配到各个节点
5. 判断是否需要**停止分裂**。如果还可以分裂则向下分裂新的叶节点

找到最佳决策：信息增益最大的 feature。其中信息增益的计算方法为：

$$\sum_{v \in Values(A)} \frac{S_v}{S} Entropy(S_v)$$

（此公式的实现在 gain 函数中）

其中 Entropy 即熵，与 Gini 相等。因此可以使用 Gini 公式计算熵的值，即：

$$1-\sum_{j} p^2(w_j)$$

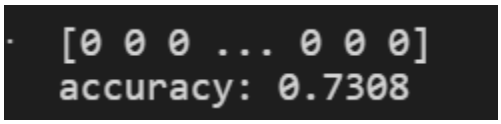
（此公式的实现在 impurity 函数中）

判断停止分裂有 3 种情况：

1. 当前节点中的样例都有相同输出类别
2. 当前节点中的样例都有相同的输入特征值
3. 当前节点的深度超过设定的最大深度

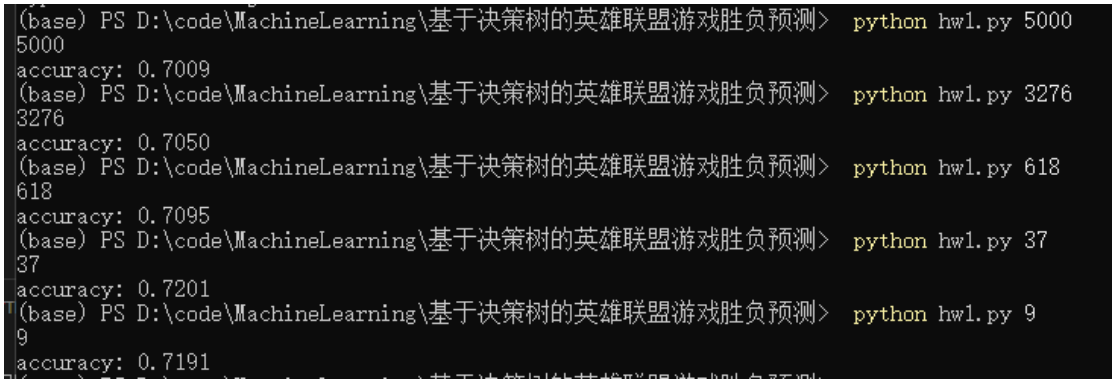
（此公式的实现在 expand_node 函数中）

训练效果测试



修改随机数种子会得到不同的结果

多次重复实验得到结果：



准确率大致在 0.70~0.72 之间

（统计有效性检验的计算没太听明白 QAQ，这里不会算，希望老师或者助教有空还能再讲讲）