2017 种子杯实验报告

队名：JS003
队长：蒋苡杭 队员：王卫星 韩旭

使用语言及运行环境

本次比赛我们队使用的是 python3.6 用于建立数据集，建立模型和训练数据并做出预测，运行环境是 spyder3.1.4，同时用 excel 辅助处理数据。

代码接口以及变量含义

代码在 python3.x 的运行环境下都可以运行。在此说明与文件相关的变量类型，以免裁判组在更改文件存储路径之后无法进行验证。

team_stats:代表读取 Teaminfo.csv 文件后的 DataFrame 类型，即由赛事组委会发布的内部包括各个团队的信息文件 teamData.csv 处理得到。

testdata:代表读取 testdata.csv 文件后的 DataFrame 类型，即赛事组委会发布的测试数据处理后所得的数据集。

result_data: 代表读取 matchresult.csv 文件后的 DataFrame 类型，即由赛事组委会的 matchtrain.csv 文件经过处理产生的，训练集的比赛结果。

数据特征提取思路

1.利用 excel 提取数据部分

由于最后的结果反映的是团队表现，因此将原始文件 teamData.csv 文件中的队员数据整合为团队数据 Teaminfo.csv。整合过程中注意筛选相关数据并且保留必要数据。例如，团队命中率是一个重要的参考量，结合总投篮出手数便可以得到总投篮命中数，因此总投篮命中数便可以不保留。此外一个球队的犯规数并不能反映一只球队的进攻和防守效率，因为可能是因为防守积极造成的犯规数增多，也有可能是一支球队防守不力，造成只能用犯规来阻止对手，因此该数据项被我们删除。

第二部分的处理是将 matchDataTrain.csv 文件中的比赛比分结果简化为胜负结果和两队的主客场关系，得到文件 matchresult.csv。一方面因为主客场关系对于比赛的结果的影响是不可忽视的，另一方面是因为一个球队的进攻和防守能力已经通过 Teaminfo.csv 文件中的球队信息表现出来，因此该部分可以忽略。

2.利用 python 提取数据

借用 python 的 pandas 函数库中遍历的 index 索引可以提取 excel 处理完的数据中的每一列数据并且作为一个整体来参与运算。主要是提取了 Teaminfo 文件中的各种队伍的数据以及根据'Wteam'和'Lteam'提取了 matchresult 文件中各队的胜负关系，两者结合就可以作为一个数据集来使用。

利用 python 还提取得到了预测过程中很重要的一个数据特征：elo 等级分。借用国际比赛中 elo rating system 作为各个球队实力判别的一个依据。ELO 等级分制度是基于统计

学的一个评估国际象棋棋手水平的方法，在这里被我用来估计各支球队的排名。该计分系统使用的是 Logistic distribution。具体数学计算方法如下：

假设棋手A和B的当前等级分分别为R_A和R_B，则按Logistic distribution A对B的胜率期望值当为：

$$E_A = \frac{1}{1+10x^{(R_B-R_A)/400}}$$

类似B对A的胜率为：

$$E_B = \frac{1}{1+10x^{(R_A-R_B)/400}}$$

假如一位棋手在比赛中的真实得分（胜=1分，和=0.5分，负=0分）和他的胜率期望值不同，则他的等级分要作相应的调整。具体的数学公式为：

$$R'_A = R_A + K(S_A - E_A)$$

公式中 R_A 和 R'_A 分别为棋手调整前后的等级分。在大师级比赛中K通常为16。

例如，棋手A等级分为1613，与等级分为1573的棋手B战平。若K取32，则A的胜率期望值为 $\frac{1}{1+10^{(1573-1613)/400}} \approx 0.5573$ ，因而A的新等级分为 $1613+32\times(0.5-0.5573)=1611.166$ 。

预测模型选取

选择 python 自带的机器学习包中的 linear_model.LogisticRegression() 方法作为最终模型，主要是通过 fit（x,y）的方法来训练模型，其中 x 为数据的属性，y 为所属类型。利用线性回归，即通过拟合线性模型的回归系数 W =（w_1，…，w_p）来减少数据中观察到的结果和实际结果之间的残差平方和，并通过线性逼近预测。下面是逻辑回归的数学原理：

逻辑回归的模型 是一个非线性模型，sigmoid 函数，又称逻辑回归函数。但是它本质上又是一个线性回归模型，因为除去 sigmoid 映射函数关系，其他的步骤，算法都是线性回归的。可以说，逻辑回归，都是以线性回归为理论支持的。

只不过，线性模型，无法做到 sigmoid 的非线性形式，sigmoid 可以轻松处理 0/1 分类问题。

下面是逻辑回归函数：

$$f(t) = \frac{e^t}{e^t+1} = \frac{1}{1+e^{-t}},$$

该函数表现了 0,1 分类的形式。

模型参数的选取与优化思路

逻辑回归中 fit(x,y)中的两个参数有不同的含义，其中 x 为提取的特征数据所构成的数组，内部包括每一个球队的总投篮出手数、总三分出手数、总罚球出手数、总得分、总篮板、总助攻、总抢断、总盖帽、总失误、投篮命中率、三分命中率、罚球命中率。y 为利用 random 函数随机分配的只包含 0，1 值的数组。提取参数内特征数据的思路报告前面的部分已经有了详细的描述，在此不再赘述。由于第一次正式提交就上到了排行榜第二的位置，因此没有进一步优化，本来想要再计算一下各支球队总的失分数来体现球队的防守能力，后来觉得反正进复赛了，争个初赛第一也没意思，就随他去了。

src 文件目录中的 csv 文件具体说明

matchresult.csv 是将组委会发布的 matchDataTrain.csv 文件中的比赛比分结果简化为胜负结果和两队的主客场关系所得到的训练数据。

testdata.csv 是将组委会发布的 matchDataTest.csv 文件中的各队对阵情况经过处理后得到的测试数据，包含了对阵双方球队编号和主客场信息。

Teaminfo.csv 是将组委会发布的 teamData 中的队员数据整合为团队数据得来，作为程序中待建立数据集的一部分准备读入。

最后运行完 python 程序后 src 文件夹中会出现 FinalRsult.csv 文件，这个文件就是根据 testdata 得到的最终结果，把空行去除就可以作为提交的文件了。因为用到了随机函数，所以最后的结果每次都不会相同，但是得分每次都能维持在 0.765 以上。