2017种子杯初赛 比赛报告

—— by 三只小菜鸡

语言及运行环境

使用 Python 3.6.1 编写。

依赖的第三方库有：

- scikit-learn (0.19.0)
- numpy (1.13.1)
- pandas (0.20.3)

操作系统：

- macOS Sierra (10.12.6)

集成开发环境：

- PyCharm CE (2016.3)

代码运行方法

整个工程只有一个源文件：main.py

使用时，需要把 `matchDataTest.csv` `matchDataTrain.csv` `teamData.csv` 文件与源码放在同一个文件夹下，输出的 `predicPro.csv` 也会添加到该文件夹中。具体文件结构如下：

```
-SeedCup
    |---main.py
    |---matchDataTest.csv
    |---matchDataTrain.csv
    |---teamData.csv
    |---predictPro.csv (输出文件)
```

需要注意：为了找出合适的训练集划分方法，我们使用了无限循环，随机划分训练集与测试集，直至准确率高于某一特定数值：

```
while(True):
  ......
  if score >= 0.73:
    break
  ......
```

由于 0.73 是经过多次实验后找到的“极限值”，所以并不能保证每次运行都能达到该值。如果仅是测试代码能否正常工作，可以适当将 0.73 调小一些。如果不改变该值，在相当长的时间内又无法退出循环，请手动终止程序，并多尝试几次。按经验来讲，如果不能很快达到该值，则在相当长的时间内也无法达到。如果顺利，则很快就能达到（很多时候重新运行几次后，第一次进入循环就能达到 0.73）。

由于随机因素的存在，导致输出的 predictPro.csv 文件每次都不同，AUC 的值也会因此产生波动，不能保证每次都能达到最好的效果（不过差距不会很大）。

各函数的作用和几个重要变量的意义可以之间看代码注释。

数据特征提取

训练使用的特征包括：

- 两队历史的比赛数据，直接从 csv 文件中解析出来，包括客场胜利次数、客场失败次数、主场胜利次数、主场失败次数。
- 队员的数据读入后，以上场时间为标准排序，取上场时间最多的前五名队员，对其各项指标取平均。 之后使用 `ExtraTreesClassifier` 得到各指标的信息量，取权重最大的前 15 个指标。

模型选取

分类器使用了逻辑回归 `LogisticRegression` ，在实验中发现特征的划分对结果影响至关重大，于是采用十折交叉验证，并在外层嵌套了无限循环，直到对测试集验证的准确率高于某一特定的极限值（这里是73%)。为了提高训练的准确率，采用了 `StratifiedKFold` 算法来划分训练集和测试集。一旦达到该极限值，则立刻跳出循环，使用此时的模型预测测试集，并输出相应的概率。

由于模型每次对训练集的划分都是随机的，所以输出的结果每次也不相同，所以 AUC 值也会有波动（而且我们以准确率为标准，而不是 AUC）。甚至，73% 的准确率也不是每次都能达到，一般需要手动终止程序并重新运行几次。根据提交的数据，最好的一次 AUC 达到 0.7531。