2017”贝贝网 • 种子杯”

初赛试题

2017/9/23

*本次最终解释权由大赛组委会所有
*更多详情请访问大赛官网 http://dian.hust.edu.cn/seedpk/

1 赛题描述

Dian 团队小点同学特别喜欢观看篮球比赛，他收集了很多场学校篮球比赛的数据。他想利用这些赛前数据来预测某场比赛两队的胜负结果，请大家一起来帮小点出出主意，使用任何技术（规则、回归、分类）来做一次预测吧。

2 数据描述

本赛题包含 5 个数据集，所有文件均为 UTF-8 编码存储的逗号分隔值文件(如遇到 excel 打开乱码请自行百度解决方案)，文件内容如下：

teamsData.csv：各个球队队员的前赛季数据。（发布时给出）

matchDataTrain.csv: 球队之间的比赛训练数据。（发布时给出）

matchDataTest.csv: 球队之间的比赛测试数据。(测试集在 26 号给出)

predictPro.csv: 各个参赛组预测的需要上交测试的结果。（自己生成需要提交的内容）

predictPro_template.csv: 输出文件的样例版，全 1 输出。(提交的样本文件在 26 号给出)

teamData.csv

该文件包含上个赛季各个队伍队员的各项数据指标。部分数据如下图所示：

队名	队员号	出场次数	首发次数	上场时间	投篮命中	投篮命中	投篮出手	三分命中	三分命中	三分出手	罚球命中	罚球命中	罚球出手	篮板总数	前场篮板	后场篮板	助攻
0	0	81	81	37.4	44.50%	8.8	19.7	33.20%	1.6	4.8	85.80%	5.9	6.9	4.1	1	3.1	7.7
0	1	59	59	31.9	51.00%	7.3	14.3		0	0	70.10%	2.9	4.1	9.6	2.2	7.4	2.5
0	2	82	82	39.1	46.00%	6.5	14.1	34.50%	1.4	4.1	75.30%	3.1	4.1	5.8	1.4	4.4	2.8
0	3	48	48	32.8	52.50%	4.4	8.4	0.00%	0	0	73.90%	2.9	3.9	10.4	3.8	6.6	2.2
0	4	82	0	20.1	43.40%	3	6.8	41.50%	1.5	3.5	88.50%	0.9	1.1	1.8	0.1	1.7	1.5
0	5	80	19	21.8	46.60%	2.9	6.3	12.50%	0	0.1	67.60%	1.3	1.9	5.7	2	3.7	0.7
0	6	81	1	22	48.00%	2.5	5.3	22.20%	0.1	0.3	65.40%	1.1	1.6	3.2	0.7	2.6	1.7
0	7	82	1	13.3	37.10%	1.8	4.8	39.30%	0.5	1.4	74.20%	0.8	1.1	1.1	0.2	1	2.3
0	8	82	82	17.8	40.40%	1.5	3.7	38.00%	1.1	2.9	65.60%	0.3	0.4	1.8	0.2	1.6	1.2
0	9	52	37	22.7	51.10%	1.8	3.6	100.00%	0	0	62.50%	0.4	0.6	5.8	1.4	4.3	1.2
0	10	13	0	9.5	41.50%	1.3	3.2	22.20%	0.2	0.7	46.20%	0.5	1	1.8	0.5	1.4	1.1
0	11	82	0	12.1	55.30%	1	1.7		0	0	50.30%	0.9	1.8	3.7	1.4	2.4	0.4
0	12	6	0	4.3	54.50%	1	1.8	57.10%	0.7	1.2		0	0	0.2	0	0.2	0
0	13	18	0	4.9	52.60%	0.6	1.1	0.00%	0	0.3		0	0	0.4	0.1	0.4	0.3
0	14	2	0	5	33.30%	0.5	1.5	0.00%	0	0.5	0.00%	0	1	0	0	0	0.5
1	0	39	39	35.3	43.50%	7.7	17.8	31.20%	1.4	4.4	81.20%	5	6.1	3.4	0.7	2.7	7.9
1	1	54	54	39.4	41.20%	5.8	14	36.70%	1.5	4	77.00%	2.4	3.1	6.5	1.4	5.1	2.9
1	2	66	66	29.5	53.20%	6.8	12.8	0.00%	0	0	69.30%	1.4	2.1	8.5	1.7	6.8	1.9
1	3	28	28	24.9	45.20%	5	11.1	37.00%	0.6	1.6	78.40%	1	1.3	2.4	0.7	1.6	3
1	4	64	64	30.4	50.80%	3.9	7.7	0.00%	0	0	74.80%	2.4	3.2	9.8	3.8	6	2.5

特别注意以下数据:

投篮命中率:所有投篮的命中率（包括罚篮）

投篮命中个数:所有投球次数（包括罚篮）

投篮出手次数:所有出手次数（包括罚篮）

matchDataTrain.csv

数据包括比赛对阵情况，有主客场队名，两队比分以及本场比赛之前的主客场战绩。部分数据如下：

客场队号	主场队号	客场本场前战绩	主场本场前战绩	比分
180	138	0胜1负	0胜0负	128:132
145	138	0胜1负	1胜0负	91:109
138	152	2胜0负	2胜0负	83:107
187	138	2胜2负	2胜1负	109:115
131	138	2胜2负	3胜1负	78:85
138	14	4胜1负	1胜5负	97:102
138	61	4胜2负	1胜5负	109:102
138	47	5胜2负	3胜4负	122:117
138	0	6胜2负	3胜3负	90:120
138	28	6胜3负	4胜5负	72:79
14	138	4胜6负	6胜4负	97:101

matchDataTest.csv

测试数据给出主客场队号，以及本场比赛之前战绩，需要预测主场球队赢得比赛的概率（置信度）。部分数据如下:

客场队名	主场队名	客场本场前战绩	主场本场前战绩
138	152	2胜0负	2胜0负
138	0	6胜2负	3胜3负
138	28	6胜3负	4胜5负
14	138	4胜6负	6胜4负
138	68	12胜18负	20胜13负
138	194	13胜21负	21胜14负
138	145	15胜21负	11胜24负

predictPro.csv

测试数据提交结果，0-1 的数值表示主场赢得比赛的置信度。部分数据如下:

主场赢得比赛的置信度
0.1
0.4
0.5

3 初赛任务评价指标

初赛题以 AUC 指标来评估预测模型的优劣，计算公式如下:

$$\frac{\sum_{\text{所有正样本}} \text{rank} - M(M+1)/2}{M * N}$$

AUC 物理含义：

假设分类器的输出是样本属于正类的 score（置信度），则 AUC 的物理意义为，任取一对（正、负）样本，正样本的 score 大于负样本的 score 的概率。

符号含义:

M 为真实结果中正样本数，N 为负样本数，rank 为将预测概率升序排序后正样本的排序位置。

AUC 计算示例：

y_true = [0,0,1,1,0]
y_pred = [0.1,0,2,0.3,0.5,0.4]

排序后：
y_true = [0,0,1,0,1]
y_pred = [0.1,0.2,0.3,0.4,0.5
]
M = 2 , N =3

Auc = $\frac{(3+5)-(2)*(2+1)/2}{3*2} = 0.8333$ (两个正样本的概率分别 rank1 = 3,rank2 = 5)

使用问题：
建议自己查阅相关资料，按照上述公式写出测试函数，或者可以借用 scikit-learn 中 AUC 的计算函数（此处方法不同，结果略有差异），裁判组最终测试以上述公式为准。

4 作品提交相关

4.1预测结果提交说明

• 提交一份格式与 predictPro_template.csv（提交样本）格式一样的文件，要求行数一致，顺序要求和 matchDataTest.csv 的预测结果保持一致。
• 提交文件要求采用 utf-8 的编码格式，使用逗号分隔值文件保存。
• 测试集于 2017 年 9 月 25 日 12:00 发放到交流群以及官网，请大家注意下载。
• 测试结果提交到测评网站（测评网站之后发布到官网和交流群），每队每日限制 3 次提交次数，请大家珍惜测试机会，格式不正确的不会浪费一次机会。

4.2最终结果提交说明

• 最终提交时间：2017 年 10 月 1 日（第四周周日）晚 22:00，请大家按时提交，我们不接收晚于提交时间提交的作品。
• 提交内容:最高分数测试样例 csv 文件、代码以及比赛报告，格式如下：
  - seedCup2017 初赛-xxx 队.zip
    - predictPro.csv(要求测试最优的结果)
    - src (源文件目录)
    - xxx.pdf
• 比赛报告内容
    • 使用语言以及运行环境。
    • 提供代码相应的接口并指明运行需要用到的变量含义，以便裁判组进行测试。
    • 数据特征提取思路。
    • 预测模型选取（包括对于规则的描述和最终模型的选择）。
    • 对于模型参数的选择与优化思路。
    • 报告内容不限于以上所述内容。
• 提交邮箱：seedcup@dian.org.cn

4.3最终评判标准

- 以提交 predictPro.csv 的 AUC 计算为主要标准，占比重 90%。
- 代码规范以及比赛报告占比重 10%。
- 为照顾新生，全队均为大一的队伍会适当加分。

5 相关事项

- 本次初赛语言不限。
- 如果发现队伍之间互相抄袭，或者任何作弊行为，直接取消比赛资格。
- 根据提交测试结果和报告，我们会优先筛选 25 支的队伍进入复赛。