World Cup Prediction

一、环境介绍

语言：`Python3.6`

requirements.txt:

```
numpy==1.14.3
pandas==0.23.0
sklearn==0.0
seaborn==0.8.1
pydot==1.2.4
pydot-ng==1.0.0
```

开发平台: `Ubuntu16.04 LTS Linux 4.16.0-999`

注：如果要使用 pydot , 还需要安装相应的包： `sudo apt install graphviz`

二、摘要

最近，火热的世界杯拉开序幕，本文使用 Kaggle 提供的世界杯足球数据集，根据各个国家队的历史战绩进行数据处理与分析。文中使用了机器学习中较为简单的算法——逻辑回归模型，推测世界杯赛程概率结果。

三、数据简介

1. FIFA_Rankings.csv

简介：记录了世界上两百多个国家和地区的历史排名

示例如下：

Position	Team	Points
1	Germany	1533.00
2	Brazil	1384.00
3	Belgium	1346.00
4	Portugal	1306.00
5	Argentina	1254.00
6	Switzerland	1179.00
7	France	1166.00
8	Spain	1162.00
9	Chile	1146.00
10	Poland	1118.00

2. result.csv

简介：该数据集主要记录的主要信息如下：

- 日期
- 主队名称
- 客队名称
- 主队进球数
- 客队进球数
- 比赛的类型， 如 Friendly，FIFA World Cup
- 比赛所在城市
- 比赛所在国家

示例如下：

date	home_team	away_team	home_score	away_score	tournament	city	country
1872-11-30	Scotland	England	0	0	Friendly	Glasgow	Scotland
1873-03-08	England	Scotland	4	2	Friendly	London	England
1874-03-07	Scotland	England	2	1	Friendly	Glasgow	Scotland
1875-03-06	England	Scotland	2	2	Friendly	London	England
1876-03-04	Scotland	England	3	0	Friendly	Glasgow	Scotland
1876-03-25	Scotland	Wales	4	0	Friendly	Glasgow	Scotland
1877-03-03	England	Scotland	1	3	Friendly	London	England
1877-03-05	Wales	Scotland	0	2	Friendly	Wrexham	Wales
1878-03-02	Scotland	England	7	2	Friendly	Glasgow	Scotland
1878-03-23	Scotland	Wales	9	0	Friendly	Glasgow	Scotland

3. Schedule

简介：该数据集记录了2018世界杯的赛程

该数据集的主要信息如下：

- Round Number，轮次，如 1 表示第一轮，即所有国家队打的第一场比赛，在决出 32 强的比赛中每个队伍需要打三场比赛
- Date，日期
- Location， 比赛场地
- Home Team， 主队
- Away Team，客队
- Group，组别
- Result，结果

Round Number	Date	Location	Home Team	Away Team	Group	Result
1	14/06/2018 18:00	Luzhniki Stadium, Moscow	Russia	Saudi Arabia	Group A	
1	15/06/2018 15:00	Ekaterinburg Stadium	Egypt	Uruguay	Group A	
1	15/06/2018 18:00	Saint Petersburg Stadium	Morocco	Iran	Group B	
1	15/06/2018 21:00	Fisht Stadium, Sochi	Portugal	Spain	Group B	
1	16/06/2018 13:00	Kazan Arena	France	Australia	Group C	

4. WorldCup2018Dataset.csv

Team	Group	Previous appearances	Previous titles	Previous finals	Previous semifinals	Current FIFA rank	First match against	Match index	history with first opponent W-L	history with first opponent goals	Second match against	Match index	history with second opponent W-L	history with second opponent goals	Third match against	Match index	history with third opponent W-L
Russia	A	10	0	0	1	65	Saudi Arabia	1	-1	-2	Egypt	17	N/A	N/A	Uruguay	33	0
Saudi Arabia	A	4	0	0	0	63	Russia	1	1	2	Uruguay	18	1	1	Egypt	34	-5
Egypt	A	2	0	0	0	31	Uruguay	2	-1	-2	Russia	17	N/A	N/A	Saudi Arabia	34	5
Uruguay	A	12	2	2	5	21	Egypt	2	1	2	Saudi Arabia	18	-1	-1	Russia	33	0
Porugal	B	6	0	0	2	3	Spain	3	-12	-31	Morocco	19	-1	-2	Iran	35	2
Spain	B	14	1	1	2	6	Portugal	3	12	31	Iran	20	N/A	N/A	Morocco	36	5
Morocco	B	4	0	0	0	40	Iran	4	-2	-2	Portugal	19	1	2	Spain	36	-5
IRAN	B	4	0	0	0	32	Morocco	4	2	2	Spain	20	N/A	N/A	Portugal	35	-2
France	C	14	1	2	5	9	Australia	5	1	6	Peru	21	-1	-1	Denmark	37	4
Australia	C	4	0	0	0	39	France	5	-1	-6	Denmark	22	-1	-3	Peru	38	N/A

四、数据预处理

1. 获取 result 数据，同时增加获胜队伍字段
2. 根据2018年世界杯名单，从以往的比赛数据中挑出32强的数据

示例如下：

	date	home_team	away_team	home_score	away_score	tournament	city	country	winning_team	goal_difference	match_year	
1230	1930-01-01	Spain	Czechoslovakia	1	0	Friendly	Barcelona	Spain	Spain	1	1930	
1231	1930-01-12	Portugal	Czechoslovakia	1	0	Friendly	Lisbon	Portugal	Portugal	1	1930	
1237	1930-02-23	Portugal	France	2	0	Friendly	Porto	Portugal	Portugal	2	1930	
1238	1930-03-02	Germany	Italy	0	2	Friendly	Frankfurt am Main	Germany	Italy	2	1930	
1240	1930-03-23	France	Switzerland	3	3	Friendly	Colombes	France	Draw	0	1930	

3. 丢弃不必要字段，并挑出 自从1930年开始的比赛 FIFA 比赛数据

   FIFA 从1930 年开始

   示例如下：winning_team中2代表主队获胜，1代表平局，0代表主队失败

	home_team	away_team	winning_team
0	England	Scotland	2
1	England	Scotland	1
2	England	Scotland	0
3	England	Wales	2
4	England	Scotland	2

4. 将处理的数据横铺，便于后期进行逻辑回归

五、数据分析

1. 选取 X 和 Y

这里我们将最后处理完的数据中的 "winning_team" 字段作为 Y，由于 "winning_team" 的取值只有2，1，0 三个值，很适合作为测试结果的 Y。其他数据作为 X 训练数据。

注：“winning_team" 的 2 表示主队胜利，1表示平局，0表示主队失败

2. 划分数据集和测试集

这里我们按照 1:9 的比例划分数据集和测试集，同时设定随机 random_state 为 100

```
X_train, X_test, Y_train, Y_test = train_test_split(
    X, Y, test_size=0.10, random_state=100)
```

3. 开始预测

主体使用训练数据进行逻辑回归预测，选择逻辑回归的原因则是，数据集比较简单规则，同时逻辑回归可以预测出概率，而且也比较简单，适合足球预测。

- 预测 32 强对抗结果

  部分示例如下：

home_team	away_team	winning_team	Probability of home_team	Probability of away_team	Probability of Draw
Russia	Saudi Arabia	Russia	0.683	0.095	0.222
Uruguay	Egypt	Uruguay	0.650	0.057	0.294
Iran	Morocco	Draw	0.314	0.329	0.358
Portugal	Spain	Spain	0.303	0.328	0.369
France	Australia	France	0.648	0.120	0.232

- 预测 16 强对抗结果

  根据32强对抗结果选出16强，并选出新的16强数据进行预测

  部分示例如下：

home_team	away_team	winning_team	Probability of home_team	Probability of away_team	Probability of Draw
Portugal	Uruguay	Portugal	0.439	0.291	0.270
France	Croatia	France	0.463	0.306	0.231
Brazil	Mexico	Brazil	0.719	0.087	0.194
England	Colombia	England	0.535	0.127	0.338
Spain	Russia	Spain	0.513	0.206	0.281

- 预测 8 强对抗结果

  部分示例如下：

home_team	away_team	winning_team	Probability of home_team	Probability of away_team	Probability of Draw
Portugal	France	Portugal	0.423	0.286	0.291
Argentina	Spain	Argentina	0.516	0.220	0.264
Brazil	England	Brazil	0.526	0.242	0.232
Germany	Belgium	Germany	0.604	0.166	0.232

- 预测 4 强对抗结果

home_team	away_team	winning_team	Probability of home_team	Probability of away_team	Probability of Draw
Brazil	Portugal	Brazil	0.717	0.123	0.159
Germany	Argentina	Germany	0.435	0.290	0.275

- 预测 决赛对抗结果

home_team	away_team	winning_team	Probability of home_team	Probability of away_team	Probability of Draw
Germany	Brazil	Brazil	0.358	0.220	0.422

六、总结

本次实验的处理方法较为简单，只是选取2018年世界杯的数据使用逻辑回归的方法进行预测，预测的依据为参赛队伍的历史战绩，并在最后给出各个队伍获胜的概率，失败的概率，平局的概率，并最后预测巴西队获取本届世界杯冠军。本结果仅供学习观看，无法保证准确率，世界杯的比赛结果无法使用历史战绩来衡量。

七、参考资料

sklearn Linear Regression

pandas

机器学习-逻辑回归

八、附录

1. 文件简介

- src

  代码及数据集

  - datasets

    数据集

  - predict.py

    用于预测世界杯的代码

  - pre_process.py

    数据预处理

  - test.py

    用于做一些小测试

- README.md

  报告

- requirements.txt

  依赖包

2. 相关包简介

- numpy

  NumPy是Python语言的一个扩展包。支持多维数组与矩阵运算，此外也针对数组运算提供大量的数学函数库。NumPy提供了与Matlab相似的功能与操作方式，因为两者皆为直译语言。

NumPy通常与SciPy(Scientific Python)和Matplotlib(绘图库)一起使用，这种组合广泛用于替代Matlab，是一个流行的技术平台。

NumPy中定义的最重要的对象是称为ndarray的N维数组类型。它描述相同类型的元素集合，可以使用基于零的索引访问集合中元素。基本的ndarray是使用NumPy中的数组函数创建的: numpy.array。

NumPy支持比Python更多种类的数值类型。NumPy数值是dtype(数据类型)对象的实例，每个对象具有唯一的特征。

- pandas

  Pandas 是python的一个数据分析包，最初由AQR Capital Management于2008年4月开发，并于2009年底开源出来，目前由专注于Python数据包开发的PyData开发team继续开发和维护，属于PyData项目的一部分。Pandas最初被作为金融数据分析工具而开发出来，因此，pandas为时间序列分析提供了很好的支持。Pandas的名称来自于面板数据（panel data）和python数据分析（data analysis）。panel data是经济学中关于多维数据集的一个术语，在Pandas中也提供了panel的数据类型。

- sklearn

  Scikit-learn（以下简称sklearn）是开源的Python机器学习库，它基于Numpy和Scipy，提供了大量用于数据挖掘和分析的工具，包括数据预处理、交叉验证、算法与可视化算法等一系列接口。sklearn的官方网站是 http://scikit-learn.org/stable/，在上面可以找到相关的Scikit-Learn的资源，模块下载，文档，例程等。sklearn的基本功能主要被分为六个部分，分类，回归，聚类，数据降维，模型选择，数据预处理，具体可以参考官方网站上的文档。
