# 基于朴素贝叶斯的世界杯胜负预测

想法起源：来自于上课所讲述的朴素贝叶斯算法，通过计算概率进行分类。

既然如此，我想到可以把胜负平看作类别（编码为1,0，-1），

把两只参赛队伍编码后作为条件（一个国家赋予一个编码）

就可以进行使用朴素贝叶斯计算进行分类与预报了

# PART1数据的收集与整理

数据来源于kaggle上的世界杯数据集（文件夹中的最早的原始数据.csv）

可以看到这个数据集之中包含了大量的有关于世界杯的数据（1930-2014）

| Year | Datetime | Stage | Stadium | City | Home Tea | Home Tea | Away Tea | Away Tea | Win condi | Attendanc | Half-time | Half-time | Referee | Assistant | Assistant | RoundID | MatchID | Home Tea | Away Tear |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1930 | 13 Jul 193 | Group 1 | Pocitos | Montevide | France | 4 | 1 | Mexico | | 4444 | 3 | 0 | LOMBARI | CRISTOPH | REGO Gilt | 201 | 1096 | FRA | MEX |
| 1930 | 13 Jul 193 | Group 4 | Parque Ce | Montevide | USA | 3 | 0 | Belgium | | 18346 | 2 | 0 | MACIAS J | MATEUCC | WARNKEN | 201 | 1090 | USA | BEL |
| 1930 | 14 Jul 193 | Group 2 | Parque Ce | Montevide | Yugoslavi | 2 | 1 | Brazil | | 24059 | 2 | 0 | TEJADA A | VALLARIN | BALWAY T | 201 | 1093 | YUG | BRA |
| 1930 | 14 Jul 193 | Group 3 | Pocitos | Montevide | Romania | 3 | 1 | Peru | | 2549 | 1 | 0 | WARNKEN | LANGENU | MATEUCC | 201 | 1098 | ROU | PER |
| 1930 | 15 Jul 193 | Group 1 | Parque Ce | Montevide | Argentina | 1 | 0 | France | | 23409 | 0 | 0 | REGO Gilt | SAUCEDO | RADULES( | 201 | 1085 | ARG | FRA |
| 1930 | 16 Jul 193 | Group 1 | Parque Ce | Montevide | Chile | 3 | 0 | Mexico | | 9249 | 1 | 0 | CRISTOPH | APHESTEC | LANGENU | 201 | 1095 | CHI | MEX |
| 1930 | 17 Jul 193 | Group 2 | Parque Ce | Montevide | Yugoslavi | 4 | 0 | Bolivia | | 18306 | 0 | 0 | MATEUCC | LOMBARI | WARNKEN | 201 | 1092 | YUG | BOL |
| 1930 | 17 Jul 193 | Group 4 | Parque Ce | Montevide | USA | 3 | 0 | Paraguay | | 18306 | 2 | 0 | MACIAS J | APHESTEC | TEJADA A | 201 | 1097 | USA | PAR |
| 1930 | 18 Jul 193 | Group 3 | Estadio Ce | Montevide | Uruguay | 1 | 0 | Peru | | 57735 | 0 | 0 | LANGENU | BALWAY T | CRISTOPH | 201 | 1099 | URU | PER |
| 1930 | 19 Jul 193 | Group 1 | Estadio Ce | Montevide | Chile | 1 | 0 | France | | 2000 | 0 | 0 | TEJADA A | LOMBARI | REGO Gilt | 201 | 1094 | CHI | FRA |
| 1930 | 19 Jul 193 | Group 1 | Estadio Ce | Montevide | Argentina | 6 | 3 | Mexico | | 42100 | 3 | 1 | SAUCEDO | ALONSO ( | RADULES( | 201 | 1086 | ARG | MEX |
| 1930 | 20 Jul 193 | Group 2 | Estadio Ce | Montevide | Brazil | 4 | 0 | Bolivia | | 25466 | 1 | 0 | BALWAY T | MATEUCC | VALLEJO ( | 201 | 1091 | BRA | BOL |
| 1930 | 20 Jul 193 | Group 4 | Estadio Ce | Montevide | Paraguay | 1 | 0 | Belgium | | 12000 | 1 | 0 | VALLARIN | MACIAS J | LOMBARI | 201 | 1089 | PAR | BEL |
| 1930 | 21 Jul 193 | Group 3 | Estadio Ce | Montevide | Uruguay | 4 | 0 | Romania | | 70022 | 4 | 0 | REGO Gilt | WARNKEN | SAUCEDO | 201 | 1100 | URU | ROU |
| 1930 | 22 Jul 193 | Group 1 | Estadio Ce | Montevide | Argentina | 3 | 1 | Chile | | 41459 | 2 | 1 | LANGENU | CRISTOPH | SAUCEDO | 201 | 1084 | ARG | CHI |
| 1930 | 26 Jul 193 | Semi-final | Estadio Ce | Montevide | Argentina | 6 | 1 | USA | | 72886 | 1 | 0 | LANGENU | VALLEJO ( | WARNKEN | 202 | 1088 | ARG | USA |
| 1930 | 27 Jul 193 | Semi-final | Estadio Ce | Montevide | Uruguay | 6 | 1 | Yugoslavi | | 79867 | 3 | 1 | REGO Gilt | SAUCEDO | BALWAY T | 202 | 1101 | URU | YUG |
| 1930 | 30 Jul 193 | Final | Estadio Ce | Montevide | Uruguay | 4 | 2 | Argentina | | 68346 | 1 | 2 | LANGENU | SAUCEDO | CRISTOPH | 405 | 1087 | URU | ARG |
| 1934 | 27 May 19 | Preliminar | Stadio Ber | Turin | Austria | 3 | 2 | France | Austria wi | 16000 | 0 | 0 | VAN MOC | CAIRONI ( | BAERT Lou | 204 | 1104 | AUT | FRA |
| 1934 | 27 May 19 | Preliminar | Giorgio As | Naples | Hungary | 4 | 2 | Egypt | | 9000 | 2 | 2 | BARLASSI | DATTILO ( | SASSI Ote | 204 | 1119 | HUN | EGY |
| 1934 | 27 May 19 | Preliminar | San Siro | Milan | Switzerlan | 3 | 2 | Netherlan | | 33000 | 2 | 1 | EKLIND Iv | BERANEK | BONIVEN | 204 | 1133 | SUI | NED |
| 1934 | 27 May 19 | Preliminar | Littorale | Bologna | Sweden | 3 | 2 | Argentina | | 14000 | 1 | 1 | BRAUN Eu | CARRARO | TURBIANI | 204 | 1102 | SWE | ARG |
| 1934 | 27 May 19 | Preliminar | Giovanni I | Florence | Germany | 5 | 2 | Belgium | | 8000 | 1 | 2 | MATTEA F | MELANDR | BAERT Jac | 204 | 1108 | GER | BEL |
| 1934 | 27 May 19 | Preliminar | Luigi Ferra | Genoa | Spain | 3 | 1 | Brazil | | 21000 | 3 | 0 | BIRLEM Al | CARMINA | IVANCSIC | 204 | 1111 | ESP | BRA |
| 1934 | 27 May 19 | Preliminar | Nazionale | Rome | Italy | 7 | 1 | USA | | 25000 | 3 | 0 | MERCET F | ESCARTIN | ZENISEK E | 204 | 1135 | ITA | USA |
| 1934 | 27 May 19 | Preliminar | Littorio | Trieste | Czechoslo | 2 | 1 | Romania | | 9000 | 0 | 1 | LANGENU | SCARPI Gi | SCORZON | 204 | 1141 | TCH | ROU |
| 1934 | 31 May 19 | Quarter-fi | Stadio Ber | Turin | Czechoslo | 3 | 2 | Switzerlan | | 12000 | 1 | 1 | BERANEK | MOHAME | BAERT Jac | 418 | 1143 | TCH | SUI |

但是我所关心的只有参赛队伍与胜负情况，而且考虑到足球运动员的运动周期，只参考了1998年之后的所有世界杯

删去多余的数据，仅保留参赛队伍与胜负情况

使用excel的查找替换，将前面的队伍对后面的队伍赢的比赛记录为1，平的比赛记录为0，输的比赛记录为-1

然后将所有国家进行编码（对照表见文件夹中的国家序号对照表.txt）

| Year | Home Tea | Away Team | Name |
|---|---|---|---|
| 1998 | 1 | 2 | 1 |
| 1998 | 3 | 4 | 0 |
| 1998 | 5 | 6 | 0 |
| 1998 | 7 | 8 | 0 |
| 1998 | 9 | 10 | 0 |
| 1998 | 11 | 12 | -1 |
| 1998 | 13 | 14 | 1 |
| 1998 | 15 | 16 | -1 |
| 1998 | 17 | 18 | -1 |
| 1998 | 19 | 20 | 0 |
| 1998 | 21 | 22 | 1 |
| 1998 | 25 | 26 | 1 |
| 1998 | 27 | 28 | -1 |
| 1998 | 24 | 30 | 1 |
| 1998 | 31 | 32 | 1 |
| 1998 | 23 | 29 | 1 |
| 1998 | 2 | 4 | 0 |
| 1998 | 1 | 3 | 1 |
| 1998 | 6 | 8 | 0 |
| 1998 | 5 | 7 | 1 |
| 1998 | 14 | 12 | 0 |
| 1998 | 13 | 11 | 1 |
| 1998 | 16 | 10 | 1 |
| 1998 | 15 | 9 | 0 |
| 1998 | 22 | 28 | -1 |
| 1998 | 20 | 18 | 0 |

| 1998 | |
|---|---|
| Brazil | 1 |
| Scotland | 2 |
| Morocco | 3 |
| Norway | 4 |
| Italy | 5 |
| Chile | 6 |
| Cameroon | 7 |
| Austria | 8 |
| Paraguay | 9 |
| Bulgaria | 10 |
| Saudi Arabia | 11 |
| Denmark | 12 |
| France | 13 |
| South Africa | 14 |
| Spain | 15 |
| Nigeria | 16 |
| Korea Republic | 17 |
| Mexico | 18 |
| Netherlands | 19 |
| Belgium | 20 |
| Argentina | 21 |
| Japan | 22 |
| Germany | 23 |
| England | 24 |
| Yugoslavia | 25 |
| Iran | 26 |

但这样产生了一个问题，胜负可能由国家前后的顺序改变，因此需要进一步处理

采用的解决办法是将两个国家顺序，因此胜负的编码也就相反，将这段处理后的数据拼接到前一段数据的后面

避免因为前后位置的摆放影响胜负结果，得到最终训练使用的数据（1.TXT）

前面两行为参赛队伍的编码，后面一行为前对后的胜负关系（胜1，平0，负-1）

| | | |
|---|---|---|
| 1 | 2 | 1 |
| 3 | 4 | 0 |
| 5 | 6 | 0 |
| 7 | 8 | 0 |
| 9 | 10 | 0 |
| 11 | 12 | -1 |
| 13 | 14 | 1 |
| 15 | 16 | -1 |
| 17 | 18 | -1 |
| 19 | 20 | 0 |
| 21 | 22 | 1 |
| 25 | 26 | 1 |
| 27 | 28 | -1 |
| 24 | 30 | 1 |
| 31 | 32 | 1 |
| 23 | 29 | 1 |
| 2 | 4 | 0 |
| 1 | 3 | 1 |
| 6 | 8 | 0 |
| 5 | 7 | 1 |
| 14 | 12 | 0 |
| 13 | 11 | 1 |
| 16 | 10 | 1 |
| 15 | 9 | 0 |
| 22 | 28 | -1 |
| 20 | 18 | 0 |
| 19 | 17 | 1 |
| 23 | 25 | 0 |
| 21 | 27 | 1 |
| 29 | 26 | -1 |
| 32 | 30 | 1 |

# 到这里数据就处理完毕!

# PART2 建立模型

In [33]:

```
from sklearn.naive_bayes import GaussianNB
import numpy as np
data=np.loadtxt(r"1.txt")
x=data[:,0:-1]#读入前两列
y=data[:,-1]#读入胜负结果
from sklearn.model_selection import train_test_split
data_train, data_test, target_train, target_test = train_test_split(x, y)
bayes = GaussianNB()
model=bayes.fit(x, y)
pred = model.predict(data_test)
err=pred-target_test
count=[]
for i in err:
    if i==0:
        count.append(i)
```

```
rate=len(count)/len(err)*100
print("以自身作为预报的准确率{:.2f}".format(rate),"%")
```

以自身作为预报的准确率46.43 %

看起来模型的准确率低，可能的原因是小组赛里有的队伍就出现过一次，有时候发生爆冷，数据量不足，对于经常打入世界杯的传统强队来说，可能更加准确一些

# PART 3 下面以2018年俄罗斯世界杯的结果进行预测与验证（训练数据只到2014）

## 建模

In [19]:

```
from sklearn.naive_bayes import GaussianNB
import numpy as np
data=np.loadtxt(r"C:\Users\cm\Desktop\1.txt")
x=data[:,0:-1]
y=data[:,-1]
bayes = GaussianNB()
model=bayes.fit(x, y)
```

# 八强战，具体的国家在注释上（俄罗斯东道主）

In [35]:

```
p1=[36,45]#乌拉圭，葡萄牙 1
p2=[13,21]#法国，阿根廷 1
p3=[1,18]#巴西，墨西哥 1
p4=[22,20]#日本，比利时
p5=[34,15]#俄罗斯，西班牙
p6=[28,12]#克罗地亚，丹麦
p7=[38,46]#瑞典，瑞士
p8=[32,24]#哥伦比亚，英格兰
px=[p1,p2,p3,p4,p5,p6,p7,p8]
result=[1,1,1,-1,1,1,1,-1]
preresult=[]
for i in px:
    xNew=np.array([i])
    pred = model.predict(xNew)
    pred=list(pred)
    preresult.append(pred)
preresult=list(np.ravel(preresult))
print(preresult)
print(result)
error=[preresult[i]-result[i] for i in range(len(result))]
count=[]
for i in error:
    if i==0:
        count.append(i)
rate=len(count)/len(error)*100
print(error)  # 误判率怎么求？不为0的就是误判
print("俄罗斯世界杯八强赛正确率{:.2f}".format(rate),"%")
```

[1.0, 1.0, -1.0, -1.0, -1.0, -1.0, 1.0, -1.0]

```
[1, 1, 1, -1, 1, 1, 1, -1]
[0.0, 0.0, -2.0, 0.0, -2.0, -2.0, 0.0, 0.0]
俄罗斯世界杯八强赛正确率62.50 %
```

# 四强战 正确率100%

In [25]:

```
p1=[36,13]#乌拉圭，法国
p2=[1,20]#巴西,比利时
p3=[34,28]#俄罗斯 克罗地亚
p4=[38,24]#瑞典 英格兰
px=[p1,p2,p3,p4]
result=[-1,1,-1,-1]
preresult=[]
for i in px:
    xNew=np.array([i])
    pred = model.predict(xNew)
    pred=list(pred)
    preresult.append(pred)
preresult=list(np.ravel(preresult))
print(preresult)
print(result)
error=[preresult[i]-result[i] for i in range(len(result))]

print(error)  # 误判率怎么求？不为0的就是误判
count=[]
for i in error:
    if i==0:
        count.append(i)
rate=len(count)/len(error)*100
print(error)  # 误判率怎么求？不为0的就是误判
print("俄罗斯世界杯四强赛正确率{:.2f}".format(rate),"%")
```

```
[-1.0, 1.0, -1.0, -1.0]
[-1, 1, -1, -1]
[0.0, 0.0, 0.0, 0.0]
[0.0, 0.0, 0.0, 0.0]
俄罗斯世界杯四强赛正确率100.00 %
```

# 半决赛（正确率50%，克罗地亚大黑马┓(´∀`)┏）

In [34]:

```
p1=[13,20]#法国，比利时
p2=[28,24]#克罗地亚，英格兰
px=[p1,p2]
result=[1,1]
preresult=[]
for i in px:
    xNew=np.array([i])
    pred = model.predict(xNew)
    pred=list(pred)
    preresult.append(pred)
preresult=list(np.ravel(preresult))
print(preresult)
print(result)
error=[preresult[i]-result[i] for i in range(len(result))]

print(error)  # 误判率怎么求？不为0的就是误判
count=[]
for i in error:
    if i==0:
```

```
        count.append(i)
rate=len(count)/len(error)*100
print(error)  # 误判率怎么求？不为0的就是误判
print("俄罗斯世界杯半决赛正确率{:.2f}".format(rate),"%")
```

```
[1.0, -1.0]
[1, 1]
[0.0, -2.0]
[0.0, -2.0]
俄罗斯世界杯半决赛正确率50.00 %
```

# 决赛，准确预报

In [28]:

```
pl=[13,28]#法国，克罗地亚
px=[pl]
result=[1]
preresult=[]
for i in px:
    xNew=np.array([i])
    pred = model.predict(xNew)
    pred=list(pred)
    preresult.append(pred)
preresult=list(np.ravel(preresult))
print(preresult)
print(result)
error=[preresult[i]-result[i] for i in range(len(result))]


print(error)  # 误判率怎么求？不为0的就是误判
count=[]
for i in error:
    if i==0:
        count.append(i)
rate=len(count)/len(error)*100
print("俄罗斯世界杯决赛正确率{:.2f}".format(rate),"%")
```

```
[1.0]
[1]
[0.0]
俄罗斯世界杯决赛正确率100.00 %
```

# 误差产生原因分析

毕竟是足球比赛，充满着许多不确定性，本模型只是基于历年数据进行统计预测。

黑马球队的存在将对于模型的准确性造成极大的影响

# 期待明年世界杯再次进行验证