ASSIGNMENT 1

Large-Scale Statistical Methods

Theodoros Efthymiadis
Maria Moutti

Contents

1. Premier League Case Study

In this case study we would like to tackle a typical question, and probably suspicion, of the football fans: Do football referees favor specific teams?

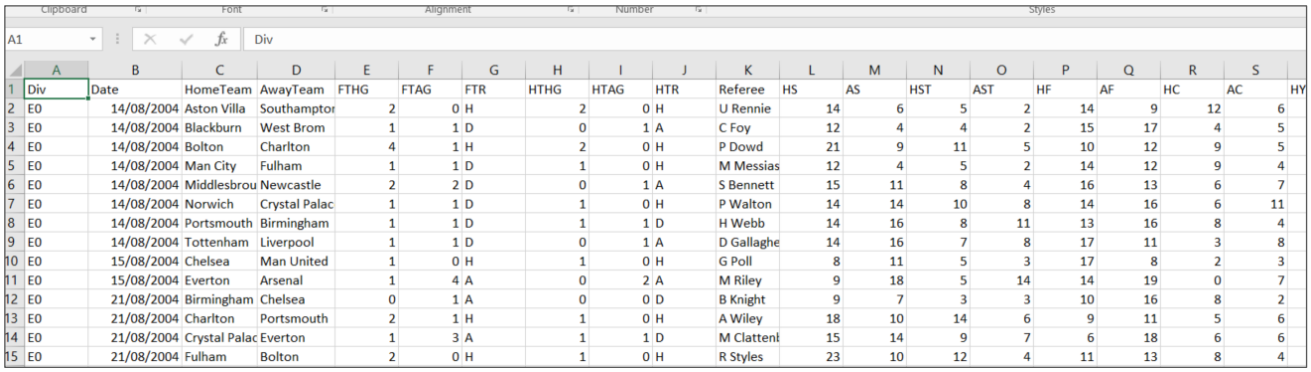
In order to answer this question, one has to consider the average performance of a specific football team and compare it with its performance when playing in presence of each different referee. In order to apply Bayes rule to the specific problem, we would need a large amount of football data, given that each match is handled by a single referee.

1.1 The data set

The most interesting and studied football championship in the world is the Premier League of the United Kingdom. Moreover, it is supported by a large data analytics and gambling industry, which is becoming increasingly more data driven. As a result, finding a clean data set to study our problem was rather easy. The data we used can be found at two different locations:

a) https://datahub.io/sports-data/english-premier-league
b) https://github.com/terrydolan/lfc

These two sources provided data in the exact same format. Description of the various columns can be found in the datahub link. By combining these two data sources, we created a data set of all Premier League data from 2004 to 2019. The most important columns for our analysis were ['Date', 'HomeTeam', 'AwayTeam', 'FTR'(full time result), 'Referee']. A sample screenshot is provided below:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	Div	Date	HomeTeam	AwayTeam	FTHG	FTAG	FTR	HTHG	HTAG	HTR	Referee	HS	AS	HST	AST	HF	AF	HC	AC
2	E0	14/08/2004	Aston Villa	Southamptor	2	0	H	2	0	H	U Rennie	14	6	5	2	14	9	12	6
3	E0	14/08/2004	Blackburn	West Brom	1	1	D	0	1	A	C Foy	12	4	4	2	15	17	4	5
4	E0	14/08/2004	Bolton	Charlton	4	1	H	2	0	H	P Dowd	21	9	11	5	10	12	9	5
5	E0	14/08/2004	Man City	Fulham	1	1	D	1	0	H	M Messias	12	4	5	2	14	12	9	4
6	E0	14/08/2004	Middlesbrou	Newcastle	2	2	D	0	1	A	S Bennett	15	11	8	4	16	13	6	7
7	E0	14/08/2004	Norwich	Crystal Palac	1	1	D	1	0	H	P Walton	14	14	10	8	14	16	6	11
8	E0	14/08/2004	Portsmouth	Birmingham	1	1	D	1	1	D	H Webb	14	16	8	11	13	16	8	4
9	E0	14/08/2004	Tottenham	Liverpool	1	1	D	0	1	A	D Gallaghe	14	16	7	8	17	11	3	8
10	E0	15/08/2004	Chelsea	Man United	1	0	H	1	0	H	G Poll	8	11	5	3	17	8	2	3
11	E0	15/08/2004	Everton	Arsenal	1	4	A	0	2	A	M Riley	9	18	5	14	14	19	0	7
12	E0	21/08/2004	Birmingham	Chelsea	0	1	A	0	0	D	B Knight	9	7	3	3	10	16	8	2
13	E0	21/08/2004	Charlton	Portsmouth	2	1	H	1	0	H	A Wiley	18	10	14	6	9	11	5	6
14	E0	21/08/2004	Crystal Palac	Everton	1	3	A	1	1	D	M Clattent	15	14	9	7	6	18	6	6
15	E0	21/08/2004	Fulham	Bolton	2	0	H	1	0	H	R Styles	23	10	12	4	11	13	8	4

Table 7: The Premier League Data set

1.2 The Mathematical Approach

In order to assess the probability of a team to win in its next match against a random opponent, given that the next match is handled by a specific referee, we will apply Bayes rule, as follows:

$$P(Win|Ref) = \frac{P(Ref|Win)P(Win)}{P(Ref|Win)P(Win) + P(Ref|Loss)P(Loss) + P(Ref|Draw)P(Draw)},$$

$P(Win)$ is the prior probability of the team to win

$P(Ref|Win)$ is the likelihood that the team won in the past in presence of the specific referee

$P(Win|Ref)$ the posterior probability for the team to win the next match in presence of the specific referee

Similar expressions can be used to calculate the posterior probability of the team to lose or for the game to result to a draw, given the presence of the specific referee. It's worth noting that in our case the likelihood terms will sum to 1, given that a match that was played and judged by a specific referee resulted to either win, loss, or draw for the team. As a result, the

outcomes of victory, defeat and draw form a partition of the sample space 'Outcome of the game judged by the specific referee' and thus:

$$P(Ref|Win) + P(Ref|Loss) + P(Ref|Draw) = 1$$

In order to estimate both the prior probability and the likelihood we will make use of the historical data. The prior probability will be estimated based on the average performance of the team over a past time period. However, the team consists of a roster that is ever – changing and, thus, the team performance in the past 10 years will be a poor predictor of the team's future performance. In order to account for that, the prior probability was calculated many times, based on a flexible time period. Given that our data can be found in the timeframe between 2004 and 2019, the prior period will be estimated to start at one specific year between 2004 and 2018 and always finish at 2019. Hence:

$$P(Win) = \frac{\sum_{t}^{2019} N_{wins}}{\sum_{t}^{2019} N_{total}}, t \in [2004, 2018]$$

N_{wins} is the number of wins throughout the prior period

N_{total} is the total number of games throughout the prior period

The prior probabilities for defeat and draw are calculated in a similar manner.

In order to calculate the likelihood, the data has to be grouped based on the teams and the referees. In contrast to the prior probability, there is no reason to assume that the likelihood changes over time. If the referee is biased in favor of specific teams, he/she will continue to do so throughout the years. In that regard, the likelihood is calculated by using the whole dataset from 2004 to 2019. It's also worth noting that the data points per year are significantly less in this calculation, given that each team plays 38 matches per year, there are many different referees in the championship and there aren't the same teams in the premier league every year. Consequently, the extension of the likelihood period was not only reasonable, but rather necessary. Hence, the likelihood of a game where the team won, was judged by the specific referee is:

$$P(Ref|Win) = \frac{\sum_{2004}^{2019} N_{wins|Ref}}{\sum_{2004}^{2019} N_{totalgames|Ref}}$$

$N_{wins|Ref}$ is the number of games the team won in presence of the specific referee

$N_{totalgames|Ref}$ is the total number of games the team played in presence of the specific referee

The likelihood for defeat and draw are calculated in a similar manner. In the programming implementation, all probabilities (prior, likelihood and posterior) were calculated for the possible outcomes of victory, defeat and draw simultaneously.

1.3 The Programming Implementation

In order to calculate the probabilities that were described in the previous section, a Python script was developed. Given the possibilities for automation and iteration that Python is offering, the script is able to calculate the prior probability for all teams that participated in

the Premier League **during the prior period** and the likelihood of those teams related to all different referees that were active from 2004 and 2019. The results are stored in a large pandas dataframe with the following structure:

Team	**Prior Probability**			**Likelihood**			
	P(Win)	**P(Loss)**	**P(Draw)**		**P(Ref|Win)**	**P(Ref|Loss)**	**P(Ref|Draw)**
Team 1	P(Win1)	P(Loss1)	P(Draw1)	Referee 1	P(Ref1|Win)	P(Ref1|Loss)	P(Ref1|Draw)
				Referee 2	P(Ref2|Win)	P(Ref2|Loss)	P(Ref2|Draw)
				Referee 3	P(Ref3|Win)	P(Ref3|Loss)	P(Ref3|Draw)
				…			
				Referee N	P(RefN|Win)	P(RefN|Loss)	P(RefN|Draw)
Team 2	P(Win2)	P(Loss2)	P(Draw2)	Referee 1	P(Ref1|Win)	P(Ref1|Loss)	P(Ref1|Draw)
				Referee 2	P(Ref2|Win)	P(Ref2|Loss)	P(Ref2|Draw)
				Referee 3	P(Ref3|Win)	P(Ref3|Loss)	P(Ref3|Draw)
				…			
				Referee N	P(RefN|Win)	P(RefN|Loss)	P(RefN|Draw)
…							
Team M	P(WinM)	P(LossM)	P(DrawM)	Referee 1	P(Ref1|Win)	P(Ref1|Loss)	P(Ref1|Draw)
				Referee 2	P(Ref2|Win)	P(Ref2|Loss)	P(Ref2|Draw)
				Referee 3	P(Ref3|Win)	P(Ref3|Loss)	P(Ref3|Draw)
				…			
				Referee N	P(RefN|Win)	P(RefN|Loss)	P(RefN|Draw)

Table 8: The pandas Dataframe with all Prior and Likelihood for all team – referee combinations

We should keep in mind that the prior distribution is calculated by using the data of the prior period (defined by the user), while the likelihood is calculated by using the entirety of the data set from 2004 to 2019.

We will now go through a brief description of the Premier_League_Posterior.py python script that was used to calculate the data. The following flow chart is quite indicative of the programs operation:

Figure 5: Flowchart of the python script

1.3.1 Libraries & User Parameters

This introductory section (Lines 1 – 20) is devoted to importing the necessary python libraries for the execution of the script. A complete list of the installed libraries is provided in the 'README.txt' file that accompanies this report and the script.

Then, the user is requested to provide the names of the team and the referee that will be reported at the end of the script. This stage requires accurate input, or the algorithm won't be able to process the request. In order to address that, the complete list viable team and referee names is provided in the 'ALL_teams.txt' and 'ALL_Refs.txt' files. Moreover, the user is requested to select the starting year of the prior period. This value can range between 2004 and 2019.

1.3.2 Reading Files

In this section (Lines 21 – 54) the different data files that are contained inside the 'data' folder are loaded as pandas data frames. However, the data are stored in two different data frames, one that will account for the prior period, and one that will account for the likelihood period, namely for the entirety of the data set.

Finally, an if statement is there to ensure that the team that was specified by the user in the previous step did actually participate in the premier league for at least one year during the prior period. Otherwise, an error will be raised and the execution of the script will be interrupted. For example, if the user selects the team to be 'Aston Villa' and the start of the prior period to be '2018', the algorithm will be interrupted because Aston Villa did not participate in Premier League between 2018 and 2019 and, thus, no prior probability can be calculated. The script will reply with the message:

>'Aston Villa did not participate in Premier League in the period 2018 to 2019'

>'Alter the input parameters and try again'

1.3.3 Data Pre – Processing

The section of data pre – processing (Lines 54 – 80) is mainly concerned with the creation of a long list object, which contains all interesting data in a useful format. That list is called teams_refs and it is structured as follows:

teams_refs = [

[Team1, [[Game1], [Game2], …, [GameN]]],

[Team2, [[Game1], [Game2], …, [GameN]]],

…

[TeamM, [[Game1], [Game2], …, [GameN]]]

]

This list allows for the mapping of a team to all games that it was involved in and it was proven very handy in calculating both the prior and the posterior probability.

1.3.4 Prior & Likelihood

This is the main section (Lines 97 – 138) of the program, where the calculation of the prior probability and the likelihood takes place. Both probabilities are calculated for all three outcomes of victory, defeat and draw and the likelihood is calculated for all possible combinations of teams and referees. This involves an iterative process that uses a number of intermediate indices, lists and data frames.

The final results are stored in the teams_total dataframe that was demonstrated earlier.

1.3.5 Posterior & Output

In the final part (Lines 138 – 164) of the script the calculation of the posterior probability for the specific team and referee that were specified by the user is computed. The necessary prior and likelihood values are queried from the teams_total data frame and the posterior probability is calculated by applying Bayes Rule. Finally, the results are printed in the command prompt. A sample output is displayed in the next figure for team = 'Arsenal' and referee = 'C Foy':

```
Arsenal Prior Probability
     P(Win)    P(Loss)    P(Draw)
0  0.526316  0.315789  0.157895
--------------------------------------------------
C Foy / Arsenal Likelihood
P(Ref|Win)      0.791667
P(Ref|Loss)     0.083333
P(Ref|Draw)     0.125000
Name: C Foy, dtype: float64
--------------------------------------------------
Posterior Probability: Arsenal / C Foy
P(Win|Ref) =  0.9004739336492892
P(Loss|Ref) =  0.05687203791469195
P(Draw|Ref) =  0.04265402843601896
```

Figure 6: Sample output of the Python script

The result contains the prior probability of the team, the likelihood of the referee/team combination and the final posterior probability that is calculated. As mentioned earlier, the likelihood terms add up to 1. Given these results, could one argue that C Foy favors Arsenal?

1.3.6 Impact of Prior and Likelihood on Posterior

In order to get a better understanding of the results, we will have a closer look at their graphical representation:

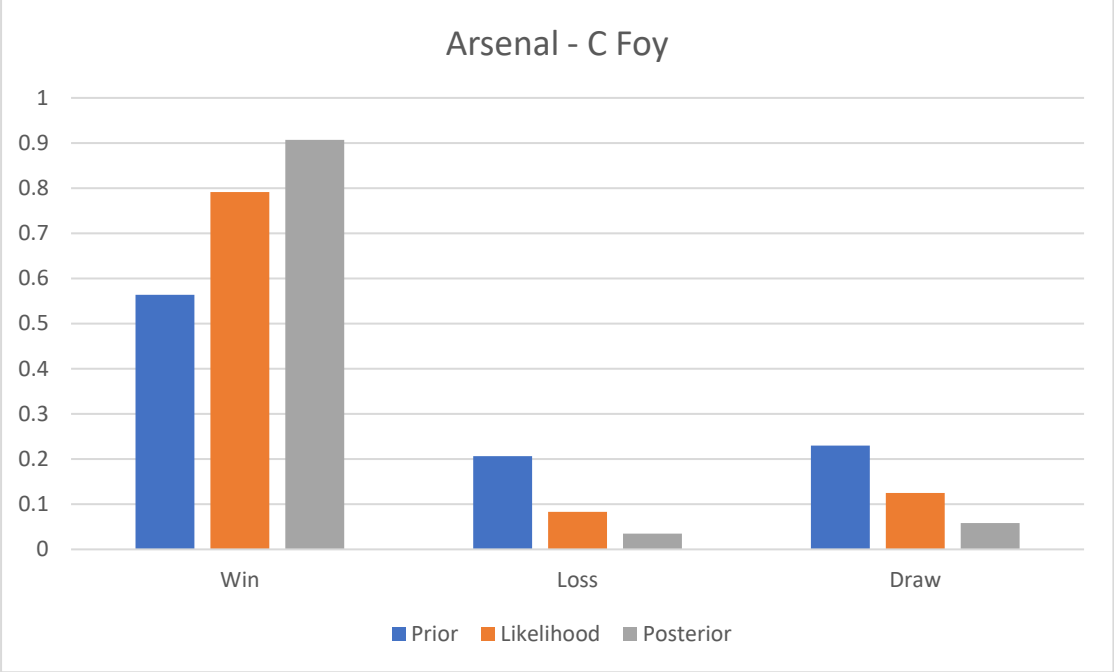


Figure 7: All three probability distributions for Arsenal and the referee 'C Foy'

This calculation was executed for Prior period = '2018,2019' and Likelihood period = '2004:2019'. The posterior is mainly Likelihood driven, given that the large differences between victory, defeat and draw observed in the Likelihood are further magnified in the Posterior. The smoother distribution of the Prior probabilities seems to have little effect on the Posterior.

Finally, we will study the change of the Prior distribution over time. Specifically, we will stick with the example of 'Arsenal' and 'C Foy' and we will compute 15 different Prior distributions, while keeping the same Likelihood and examine the effect on the Posterior. The different prior distributions will vary in the start of the time period that will be used, which will fall in range between 2004 and 2018, while the end of the prior period will always be 2019. In that regard, the first computation will account for a Prior period from 2004 to 2019, the second from 2005 to 2019, while the last one will only account for 2018 to 2019. However, only the probabilities for the outcome of 'Win' will be plotted to avoid clutter. The results are illustrated in the figure below:

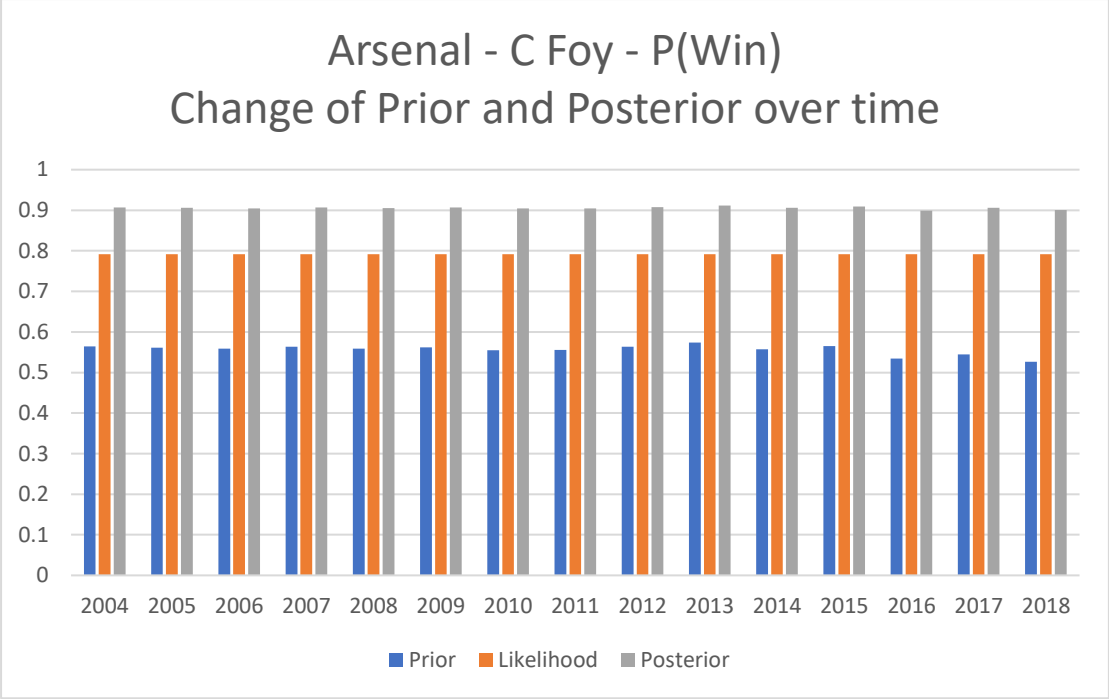


Figure 8: The calculation of the Prior probability for Arsenal to win, over different time periods

By observing the different Prior probabilities that are calculated over time, we realize that the performance of Arsenal has been very consistent throughout the years. Moreover, we assume that the preference of the referee (if it exists) remains more or less consistent as well. This figure also supports what was argued earlier, namely that the Posterior is primarily Likelihood driven.

2. References

1. www.car.gr was used to obtain the car accident data set [accessed on 10/11/2020]
2. https://datahub.io/sports-data/english-premier-league was used to obtain the premier league data set [accessed on 10/11/2020]
3. https://github.com/terrydolan/lfc was used to obtain the premier league data set [accessed on 10/11/2020]