Data Analytics
ECS 784P

Coursework 2

-Semil Halani
(210147012)

Question 1.

This research study analyses the effect of a home game versus an away game for a team in the English Premier League. It illustrates all the effects in the respective statistics for the home team and the away team. The dataset used for this research is of the 2000-2018 seasons for the English Premier League (EPL), describing the respective details for our research. As it is seen that most variables in this dataset are causally related, it is suitable for structured learning algorithms which reflect the correlation between the home team statistics in a game and their winning chances.

As a football fan, one can say that for a home team in EPL, if the home crowd is more and outnumbers the away fans, it leads to more home cheers and boos for the away team. Consequently, this results in appositive impact on home team's confidence. Thus, the home ground advantages strongly influence the games.

	FTHG	FTAG	FTR	HTHG	HTAG	HTR	HS	AS	HST	AST	HF	AF	HC	AC	HY	AY	HR	AR
0	4	0	1	2	0	1	17	8	14	4	13	12	6	6	1	2	0	0
1	4	2	1	1	0	1	17	12	10	5	19	14	7	7	1	2	0	0
2	4	3	-1	1	1	0	6	16	3	9	15	21	8	4	5	3	1	0
3	4	2	0	1	2	-1	6	13	4	6	11	13	5	8	1	1	0	0
4	4	0	1	2	0	1	17	12	8	6	21	20	6	4	1	3	0	0
5	4	0	0	0	0	0	5	5	4	3	12	12	5	4	2	3	0	0
6	4	0	1	0	0	0	16	3	10	2	8	8	6	1	1	1	0	0
7	4	0	1	0	0	0	8	14	2	7	10	21	2	9	3	1	0	1
8	4	1	1	2	1	1	20	15	6	5	14	13	3	4	0	0	0	0
9	4	0	1	1	0	1	19	9	9	6	7	13	7	1	0	1	0	0

Figure 1. A sample of the dataset in use

Question 2.

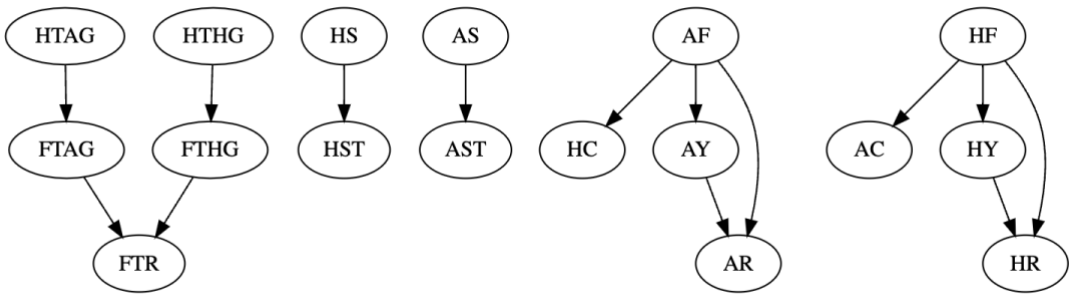

Figure 2. Our knowledge-based DAG

Being a football fan, I believed that my own knowledge is sufficient for the same and so, I did not refer to any literature for gathering the required knowledge for producing this graph. History shows that teams have a higher winning rate at their home. Usually, it is usually seen that the strength of home fans is significantly more than away fans. As the number of fans increase, so does the cheers and chants for the team and players and thus, the spirit and confidence of the players is uplifted to its peak. Thus, the increase in home cheers and better mental state for the home team leads to a better play for the home players and so, more shots are played, and so, more shots on target as the confidence is high and the same for away shots. Also, due to the fouls committed by the away team, the home team wins corners and yellow cards for the away team, and subsequently red card and similarly for the home team fouls. Also, the half team goals for both the teams also affect full-time goals for the respective teams as along with the half time goals being accounted for the full-time goals, it also builds or destroys the players' confidence and thus, leading to a full-time result.

Question 3.

Algorithm	CPDAG scores			Log-Likelihood (LL) score	BIC score	Free parameters	Structure learning elapsed time (seconds)
	BSF	SHD	F1				
HC_CPDAG	0.418	17.000	0.438	-195124.125	-202883.061	1226	20
HC_DAG	0.382	17.500	0.406	-195124.125	-202883.061	1226	20
TABU_CPDAG	0.418	17.000	0.438	-195124.125	-202883.061	1226	8
TABU_DAG	0.382	17.500	0.406	-195124.125	-202883.061	1226	8
SaiyanH	0.454	16.500	0.469	-194855.087	203240.561	1325	12
MAHC	0.522	12.000	0.571	-196809.371	-203777.225	1101	7

Our algorithms tend to form graphs with around 17 nodes and 18 edges and 6460 samples, which is more than 9 nodes, 15 edges and 10^3 samples (Network 1) and less than 27 nodes, 31 edges and 10^4 samples (Network 2). Thus, we compare it with the algorithms for a network with 27 nodes, 31 edges and 10^4 samples.

On comparing the results shown below, to that of the respective row from table 2.1 of the manual, it is seen that the according to our expectations, BSF scores for all algorithms lie between both the range for both the cases (Network 1 & 2) except the SaiyanH algorithm whose BSF score is less than the respective ones. On the other hand, the SHD scores and the runtime is significantly higher for all algorithms in our experiment scenarios. Moreover, similar to the BSF score trend, the F1 scores for all algorithms in our case are less than the respective scores

as stated in the Table 2.1 of the manual except that for the MAHC algorithm which lies between both the scores from the table.

Algorithm	BSF		SHD		F1		Runtime (secs)	
	Network 1	Network 2	Network 1	Network 2	Network 1	Network 2	Network 1	Network 2
HC	0.305	0.697	11	13	0.5	0.721	0	0
TABU	0.305	0.697	11	13	0.5	0.721	0	0
SaiyanH	0.6	0.755	6	8.5	0.75	0.797	0	5
MAHC	0.333	0.72	10	10.5	0.5	0.776	0	0

Question 4.

The three causal classes in the CPDAG are highlighted below:

1. Causal chain (X → Y → Z):

   HTHG -> FTHG -> FTR

   The causal chain illustrate above indicates that the number of goals for the home team at half time influences the goals for the home team at full time, which further influences the result for the game at full time.

2. Common cause (X ← Y → Z):

   HTHG <- HTR -> HTAG

   The common cause described above states that the result for the game at half time is related to the number of goals for the home team and also the away team at half time.

3. Common effect (X → Y ← Z):

   FTHG -> FTR <- FTHG

   As the number of goals for the home team at full time and for the away team collectively decides the result of the game at full time, whether it is a win or a lose for the home team, or maybe a draw for both the teams. Thus, the common effect stated above holds true.

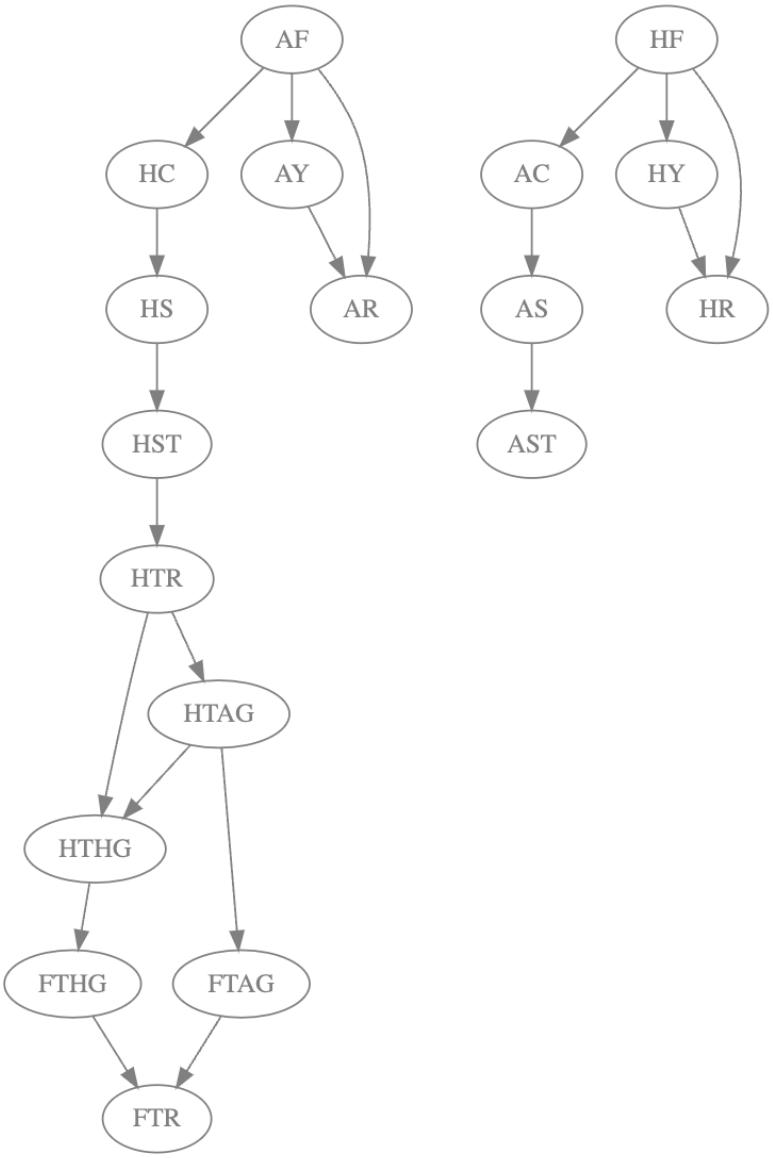

Figure 3. The graph visualised for the CPDAGlearned.csv generated by HC_CPDAG

Question 5.

	Our rankings			Rankings according to the Bayesys manual		
Rank	**BSF [single score]**	**SHD [single score]**	**F1 [single score]**	**BSF [average score]**	**SHD [Normalised score]**	**F1 [average score]**
1	MAHC [0.522]	HC_DAG [17.500]	MAHC [0.571]	TABU_CPDAG [0.533]	MAHC [0.481]	SaiyanH [0.576]
2	SaiyanH [0.454]	TABU_DAG [17.500]	SaiyanH [0.469]	SaiyanH [0.515]	TABU_CPDAG [0.44]	TABU_CPDAG [0.564]
3	HC_CPDAG [0.418]	HC_CPDAG [17.000]	HC_CPDAG [0.438]	HC_CPDAG [0.506]	SaiyanH [0.438]	MAHC [0.562]
4	TABU_CPDAG [0.418]	TABU_CPDAG [17.000]	TABU_CPDAG [0.438]	MAHC [0.499]	HC_CPDAG [0.402]	HC_CPDAG [0.537]
5	HC_DAG [0.382]	SaiyanH [16.500]	HC_DAG [0.406]	TABU_DAG [0.484]	TABU_DAG [0.397]	TABU_DAG [0.53]
6	TABU_DAG [0.382]	MAHC [12.000]	TABU_DAG [0.406]	HC_DAG [0.438]	HC_DAG [0.314]	HC_DAG [0.479]

Upon comparison of our rankings to that of the ranking from the Bayesys manual, it is seen that only the values for our MAHC algorithm are close to those mentioned in the manual. Apart from that, rest all values are different and considerably lower in both BSF and F1 scores. However, the order for the SHD scores is almost the opposite when compared to the respective order mentioned in the manual. I believe that the main reason for this difference can be the missing constraints that need to be used that are not used here, the knowledge approaches.

Question 6.

As mentioned in the answer for Question 3, the runtime for our algorithms is considerably higher than the runtime mentioned in the manual for the respective algorithms for similar datasets. It seems that the problem is too complex, and the algorithm has to work through completely resulting in more time being taken for the same. Also, the lack of usage of the knowledge approaches also seems to be play an important role for the same.

Question 7.

	Your Task 4 results				Your Task 5 results		
Algorithm	**BIC score**	**Log-Likelihood**	**Free Parameters**	**Algorithm**	**BIC score**	**Log-Likelihood**	**Free Parameters**
My knowledge-based graph	-207645.499	-196374.157	1781	**HC_CPDAG**	-202883.061	-195124.125	1226
				HC_DAG	-202883.061	-195124.125	1226
				TABU_CPDAG	-202883.061	-195124.125	1226
				TABU_DAG	-202883.061	-195124.125	1226
				SaiyanH	203240.561	-194855.087	1325
				MAHC	-203777.225	-196809.371	1101

The BIC score, Log-likelihood and the number of free parameters for the algorithm based on our knowledge-based graph as performed in the Task 4, is mentioned in the table above. Similarly, the respective values for the different algorithms used in the Task 5 in mentioned alongside it, in the same table. Upon comparison between both, it is seen that the calculation using the knowledge-based graph produced a better BIC score and more free parameters than all the other algorithms of Task 5. Furthermore, the Log-likelihood is also better for the knowledge-based graph as compared to the other algorithms of Task 5, except the MAHC algorithm whose Log-likelihood value is slightly better than the one obtained through the knowledge-based graph. This seems as a result of better clarity for the dataset with the help of our knowledge-based graph and thus, a better output.

Question 8.

From among the many knowledge approaches available like Directed, Undirected, Forbidden, Temporal, Initial graph, Target nodes etc., the following 2 approaches have been used here:

- Directed
- Temporal

ID	Parent	Child
1	FTHG	FTR
2	FTAG	FTR
3	HTHG	FTHG
4	HTAG	FTAG
5	HS	HST
6	AS	AST
7	AF	HC
8	HF	AC
9	HF	HY
10	AF	AY
11	HF	HR
12	AF	AR
13	HY	HR
14	AY	AR

Figure 4. constraintsDirected.csv

ID	Tier 1	Tier 2	END
1	HS	FTR	
2	AS		
3	HF		
4	AF		

Figure 5. constraintsTemporal.csv

The above illustrated constraints have been used for the respective knowledge approaches for this solution.

Knowledge Approach	CPDAG scores			LL	BIC	Free Parameters	Number of edges	Runtime
	BSF	SHD	F1					
Without knowledge	0.418	17.000	0.438	-195124.125	-202883.061	1226	18	20
With Knowledge (Directed constraints)	0.951	6.000	0.824	-196374.157	-207645.499	1781	20	14
With Knowledge (Temporal constraints)	0.498	15.000	0.516	-194895.39	-202920.130	1268	17	19

After careful calculations and observations, it can be observed that the solution is improved with the use of the knowledge approaches. Also, as the constraintsDirected.csv for the Directed knowledge approach is much more specific than the constraintsTemporal.csv for the Temporal knowledge approach, it is seen that the Directed approach works far better than the other method. Thus, it can be concluded that the solutions can be improved with more detailed knowledge-based approaches for the respective problems.