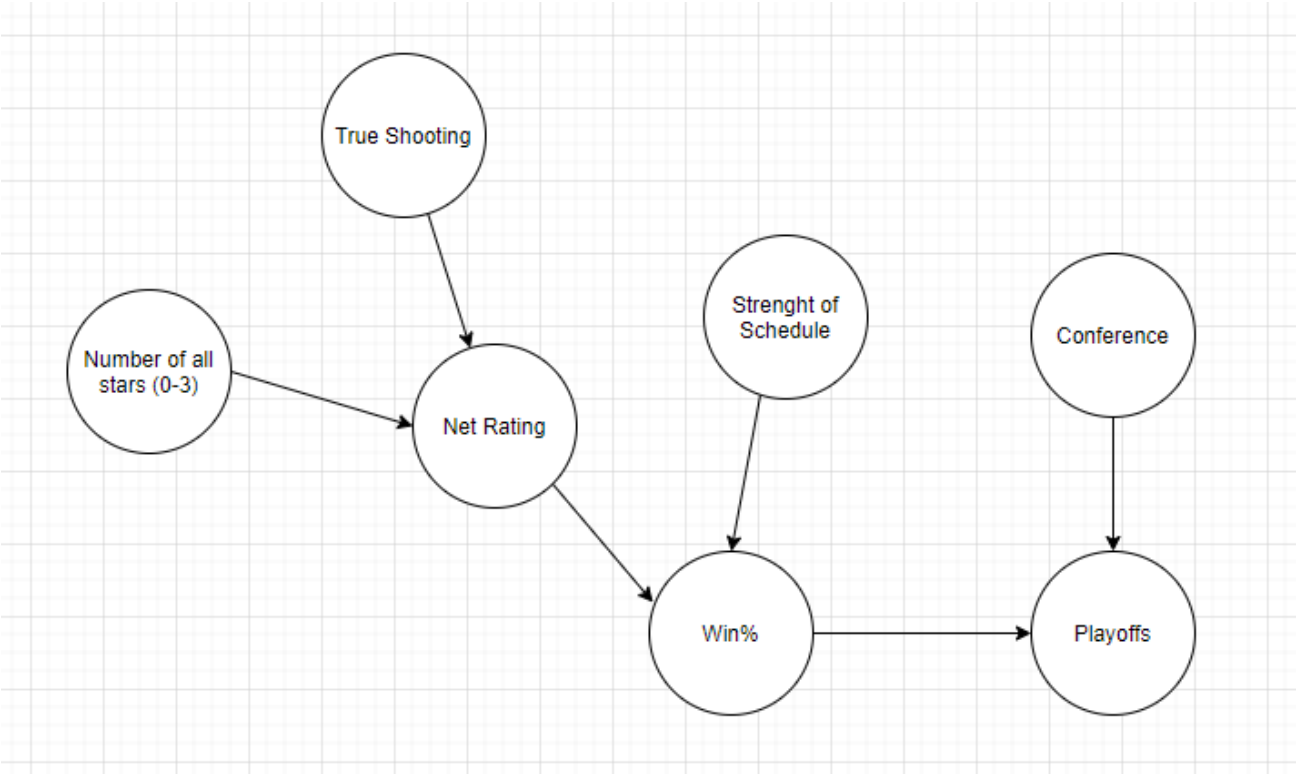
Fundamentals of AI and KR - Module 3

Project Report - Lorenzo Niccolai - 0001004845

1. The working domain

As I'm a fan of basketball, in particular of the NBA (the american top level league), the idea of applying Probabilistic Reasoning and Bayesian Networks to that domain was intriguing. In fact analytics are commonly used in basketball to improve performances, due to the natural translation of game events in statistics. To create the network and the tables I used both my intuition and took inspiration from a paper I found on this same argument (https://core.ac.uk/download/pdf/212850307.pdf).

2. The network



Nodes description

- Conference
  - Values: 0,1
  - Description: The conference of the Nba team, encoded by a binary variable. 0 for East and 1 for West
- Strength of Schedule
  - Values: 0,1
  - Description: The difficulty of a team's seasonal schedule, encoded by a binary variable. 0 for above average and 1 for below or equal to average

    - Tip: In the Nba a team has more games with teams of the same conference and division, which is a more precise division. Some divisions are stronger than others, so the difficulty of schedule changes.
- Number of All-Stars
    - Values: 0,1,2,3
    - Description: The number of All-Star players of a team, encoded by a 4-ary variable. The numbers are trivially referring to the number of All-Star players in a team.
    - Tip: In principle more All-Stars means a better team, especially in last years. Anyway the all-stars are decided based on people voting and may be misleading in limit cases, but all the top teams have at least 2 all star level players.
- True Shooting
    - Values: 0,1
    - Description: The True Shooting percentage of a team, encoded by a binary variable, 0 for below league average (56) and 1 for equal or above.
    - Tip: True Shooting is a composite statistic (weights more 3s and free throws), it accounts for the efficiency of a player, for the team it averages on the players. It is closely related to the team performance through the season.
- Net Rating
    - Values: 0,1.
    - Description: The net rating of a team, encoded by a binary variable, 0 for negative and 1 for neutral or positive.
    - Tip: This statistic alone encodes very well the level of a team, it takes into account the offensive and defensive points produced but normalized in 100 possessions. In simple terms, the more positive it is the better the team is, the more negative the worse.
- Win
    - Values: 0,1.
    - Description: The win percentage of a team encoded by a binary variable, 0 for below 50% and 1 for equal or above.
    - Tip: This is a variable which is important to predict from the previous ones, since predicting the winning percentage of a team is useful to both the team itself and betting societies for example. It is also conditioned by the schedule.
- Playoffs
    - Values: 0,1,2.
    - Description: The final outcome of the season, encoded by a 3-ary variable, 1 for entering the play-ins (6-10), 2 for entering the playoffs (5-1) and 0 if a team positions 11 or more in his conference.
    - Tip: This is a direct consequence of the Win variable. In addition to the wins it is useful to predict if a team will enter the playoffs, which is an important goal.

3. Queries

With the variable elimination queries I tried to tackle the relation of some variables and ‘Playoff’. With a simple query (q_1) it is possible to see that the a priori distribution of such variables is quite standard, but when we assume things like ‘Conference’ we immediately obtain the confirmation that reaching playoffs in the West is more difficult than in the East (q_2, q_bb). And even if the query is further refined assuming that the team has a positive record (‘Win’:1) the probability of a good result (‘Playoff’:1,2) increases as it should (q_3, q_4). An interesting connection is the one between the number of All Stars players of a team and the final result, this is not taken into account by any statistical database but it is widely

believed to be a strong causation of success in NBA. From the queries it is visible that more All Stars imply a better season record, of course this varies with the conference: in the West having 2 All Stars makes a bit of difference from 0 or 1 (q_5, q_6, q_7), while in the East it nearly guarantees you to clinch the Playoffs (q_9, q_10, q_11). Also the Strength of Schedule is an underestimated factor, given a below average Net Rating team (very bad and unlikely to clinch the playoffs), having an easier schedule increases quite a lot the season good outcomes (q_15, q_16, q_17). The queries of True Shooting given Playoff variable (season outcome) gives a good explanation of the actual level of the team (q_12, q_13, q_14).

Methods

I applied 2 different methods for approximate inference: Rejection sampling and Likelihood Weighting. They both needed roughly 10k samples to reach an acceptable error, an interesting behavior resulted from the test on the P('Conference':1) where the Likelihood Weighting absolute error after a certain number of samples seems to plateau (at least it decreases very slowly), while the Rejection sampling absolute error further approaches the zero.

4. Results

This domain has been already explored and with this small experiment I tried to show how Bayesian Network can nicely represent the performance of an NBA team over a regular season. Even if the model is somewhat simple it encodes a big part of the meaningful stats for an Nba team, another improvement could be to create broader bins for the variables. This network could be augmented with more complex and powerful methods, predicting probability of the final seed is useful both to a team itself and to a betting company for example.