Student Performance Bayesian Networks

Gio Formichella
Master's Degree in Artificial Intelligence, University of Bologna
gio.formichella@studio.unibo.it

April 18, 2025

Abstract

This project compares three different Bayesian networks: one constructed via PC algorithm, one via Tree Search and one expert-defined, for modeling student attributes and math performance from UCI's Student Performance dataset. The networks were queried about the impact of absences, study time and internet access on final grades and about whether long travel times from home to school negatively affect health. The query execution time varied depending on network structure but all three models agreed on the results: low absence rate is more impactful on performance than longer study time, internet access helps students' education and long travel times do not affect health.

Introduction

Domain

The Bayesian networks were modelled on UCI's Student Perfomance dataset (Cortez 2008) containing information about secondary school students' performance in math and Portuguese language courses. The 33 data attributes include student grades, demographic, social and school related features and have discrete domains. The project focuses on the math course data.

Aim

The goal of the project is to compare both memory usage and query efficiency of Bayesian networks constructed with different strategies in modeling students' math performance. While two networks use a structure learning approach, one constrained-based and one score-based, the third is defined through domain knowledge and prior understanding of factors influencing student grades.

Method

The pgmpy library (Ankan and Textor 2024) was used to implement the three networks and run exact inference queries via variable elimination. The query execution time was measured 5 times, via the Python timeit library, and then averaged to reduce measurement error. The average execution time, the query results and the total network CPT size were compared.

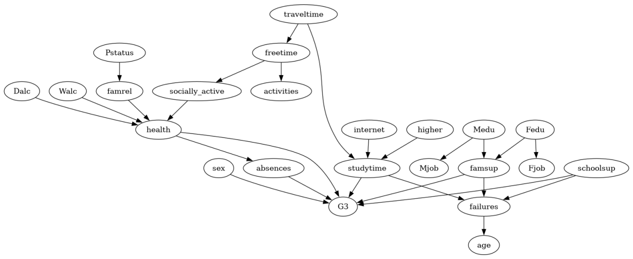

Figure 1: Expert-defined network

Results

Surprisingly, the PC network, the sparsest of the three, had the largest memory usage due to the interconnection of high cardinality nodes, resulting in a huge CPT.

Depending on the structure, the networks had equivalent query answers but very different average execution times, with no clear winner.

Model

The PC network, constructed via the constrained-based structure learning PC algorithm (Spirtes, Glymour, and Scheines 2000), and the Tree network, constructed via the score-based structure learning Tree Search algorithm (Chow and Liu 1968), use the full set of features in the dataset with no data manipulation or re-binning. The expert-defined network (Figure 1), on the other hand, was constructed from preprocessed data. The G3 variable, corresponding to the final grade, was re-binned from [0, 20] to [0,3], corresponding to ”Insufficient”, ”Sufficient”, ”Good” and ”Excellent”, while the absences variable was re-binned to 0, 1 and 2, corresponding respectively to ”Low”, ”Medium” and ”High”. The mother and father education and job features were turned binary and the romantic and goout attributes were merged together into the binary ”socially_active” feature. Finally, the characteristics not deemed highly relevant for the grade prediction were dropped. These simplifying assumptions allowed for a smaller network but at the cost of contained information loss.

The probabilities in the CPTs of the three models were estimated from the data.

Analysis

Experimental setup

In the structure comparison of the three networks, the number of edges were counted along with the total number of entries of all the CPTs.

The following queries were carried out and their average execution time measured:

1. In order to have a better chance of achieving a higher final grade, is it better to have fewer absences at the cost of less study time or is it better to have more absences but more study time ? In mathematical terms:

   $$\mathbb{P}(G3 \geq good|absences = low, studytime = medium) \gtrless$$
   $$\mathbb{P}(G3 \geq good|absences = medium, studytime = high)$$

   For the non-re-binned networks, the left side formula (right side is analogous) corresponds to:

   $$\frac{\sum_{j=0}^{9}\left(\mathbb{P}(absences{=}j|studytime{=}3)\sum_{i=15}^{20}\mathbb{P}(G3{=}i|absences{=}j,studytime{=}3)\right)}{\sum_{j=0}^{9}\mathbb{P}(absences{=}j|studytime{=}3)}$$

   While for the re-binned network:

   $$\sum_{i=2,3}\mathbb{P}(G3 = i|absences = 0, studytime = 3)$$

2. Does having a long travel time from home to school affect students' health ? That is to say, does the probability of having good health change with evidence of long traveltime ?
3. How does internet access at home affect student performance ? Does it support them in their studies or does it allow for more distractions ? The distributions $\mathbf{P}(G3|internet = yes)$ and $\mathbf{P}(G3|internet = no)$ were compared.

Results

As depicted in Figure 2, the PC network is the sparsest of the three, with just 13 edges, but has the largest amount of CPT entries. On its own, the CPT of the G3 variable, connected to absences and G2, accounts for 10404 probabilities. Meanwhile, the expert-defined network, thanks to its simplifying assumptions, was the smallest memory-wise.

The query answers were equivalent, the absolute probabilities differed but the comparative results stayed the same; the execution times are presented in Table 1. All three networks confirmed that having fewer absences is more important that having more time to study outside class. The fastest answer came from the expert-defined net, due to its simpler formula.

For the health query, the probability of having good health was almost exactly the same as the probability conditioned by long travel times. It was thus concluded that long travel time does not affect health. For this inquiry, the fastest result came from the PC network, due to the independence of the health variable from all other features.

For the final query, the distributions are shown in figure 3. The PC network, due to G3 being independent from the internet variable, has matching plots, signaling no effect of internet access on performance. The Tree network, however, shows that with internet access the probabilities of higher grades are higher while the probabilities of lower grades are lower, thus confirming its usefulness. The expert-defined network was not able to represent the distribution over the full grade scale because of the the re-binning of the G3 variable. However, it shows that with internet access, the probabilities of ”Good” and ”Excellent” grades are higher, and the probabilities of ”Insufficient” and ”Low” grades are lower than without internet access. This time, the fastest network was the Tree one thanks to smaller factors during variable elimination.

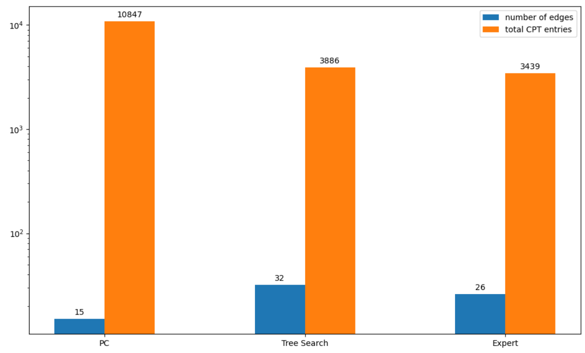

Figure 2: Structure comparison

	Average time (ms)		
Query	PC	Tree	Expert
1	7.1	6.7	3.1
2	0.9	1.3	2. 4
3	2.2	1.6	4.6

Table 1: Query execution time

Conclusion

The use of domain knowledge and simplifying assumptions was pivotal in reducing the size of the expert-defined network but it came at the cost of some information loss. The structure of networks affected the query outcomes and the inference execution time, with no clear best model. The choice of the best approach thus relies on the application queries.

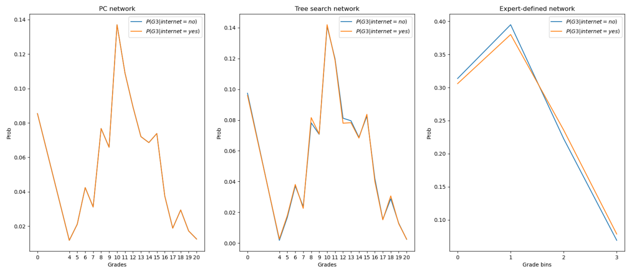

Figure 3: Internet access at home

References

Ankan, A., and Textor, J. 2024. pgmpy: A python toolkit for bayesian networks. *Journal of Machine Learning Research* 25(265):1–8.

Chow, C. K., and Liu, C. N. 1968. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory* 14(3):462–467.

Cortez, P. 2008. Student Performance. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5TG7T.

Spirtes, P.; Glymour, C.; and Scheines, R. 2000. *Causation, Prediction, and Search*. MIT Press, 2nd edition.