Baysian Networks - First Assignment

Alejandro González Rogel
Athena Iakovidi
Juraj Šušnjara

12 December 2016

1 Problem domain

Violence in schools constitutes a serious problem in many countries. This problem could get even worst in those places where weapons, such as guns, are involved. Although school is meant to be a secure and friendly place, many children are afraid to go there because they are often threatened and bullied by other children. Moreover, the fact that any of those children (both the bully and the bullied) can have relatively easy access to guns can result in enormous tragedies, as seen in the past. One of the researches of importance and statistics of students carrying weapons can be seen in [1].

In order to deal with this problem, we propose a solution based on Bayesian Networks that could help us to predict if a student is likely to carry any short of weapon to school. For that, we will be focusing on some of the children characteristics and the connections between them.

2 Data

In our project, we use data from the 2015 from [2], which analyses health-risk behaviours in teenagers in the United States. This data is public and contains the answers given by more than 15.000 students to 99 different questions (plus a few extra parameters about the students and their schools). There is also a "Data User's Guide" that contains additional information about the survey. All the information (together with the data) can be found at [2].

Some of the students have been discarded for our study because they didn't answer some questions or their answers were already discarded by the YRBS because they were inconsistent. Thus, the total data used for this project contained 12.669 entries.

We use the attributes shown in Table 1. They are all related to some question asked in the mentioned survey. However, it is important to know that the number of possible outcomes of most of the variables has changed with respect to the original dataset. This is due to the high number of parameters the Bayesian network would have in the case we wanted to use the original distributions. This will be further discuss in subsection 5.2, where we talk about binning variables.

Name	Type	#Levels	Description	YPBS question
Race	categorical	8	Race	Raceeth (Q4 & Q5)
Age	categorical	5	Age	Q1
Sex	categorical	2	Sex (Gender)	Q2
Weapon	categorical	2	Carry weapon in school	Q15
Unsecure	categorical	2	Feel unsecure in school	Q16
Fight	categorical	2	Participate in fight last year	Q20
Bulled	categorical	2	Bullied in school	Q24
Depression	categorical	2	Suffered depression last year	Q26
Alcohol	categorical	2	Alcohol use last month	Q43
Sports	categorical	4	Sports practiced	Q80
Grades	categorical	4	Grades	Q89

Table 1: The variables of our network. The table includes the name, type and number of levels of each variable (categorical variables with only 2 levels can be considered as binary).

3 Our Network

To begin with, we created a prototype by assuming the network structure based on our understanding of the domain (see the red connections in Figure 1). Following is a short argumentation for each structural decision we made.

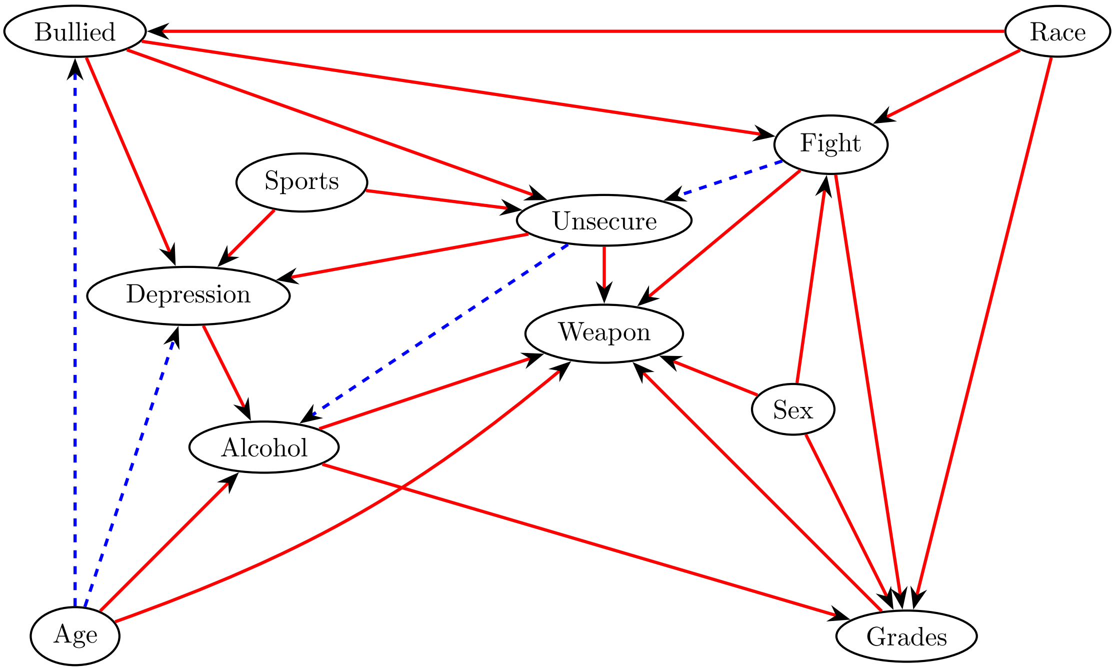

Figure 1: Final Bayesian network proposed. The connections between nodes were suggested based on our own knowledge of the domain. Four extra connections were added (blue, dashed arrows) after a conditional independency test was performed.

Age → Alcohol: Children do not have as easy access to alcohol as teenagers (parents are more careful to hide alcohol and shop owners are less likely to give alcohol to so young children).

Age → Weapon: Weapons are less available to younger children and young children might be more terrified of using them. Thus, age should influence probability of carrying a weapon.

Race → Bullied: Based on research [3] in 16% of bullying cases the reason is race.

Race → Fight: Since there is always signs of racism in many schools, race might influence frequency of getting into fights.

Race → Grades: Our first assumption was that race can influence grades. We don't assume that race is the cause of this dependency, but relevant social constructs make this connection be true.

Sex → Fight: Males engage in more physical conflicts than females.

Sex → Grades: Females generally tend to have higher grades than males.

Sex → Weapon: Males are more likely to carry a weapon.

Bullied → Unsecure: Children that are bullied tend to feel less secure in school.

Bullied → Depression: Bullied children are more likely to feel unhappy for a big part of the day, which can lead to depression.

Bullied → Fight: Some bullied children may fight back. Moreover, they may initiate fights with weaker children so as to feel more in control.

Unsecure → Depression: Children that feel unsecure in school are more likely to show signs of depression.

Unsecure → Weapon: Children that feel unsecure in schools might decide to carry a weapon in order to protect themselves.

Depression → Alcohol: Depressed children may seek solution of their problems in alcohol.

Alcohol → Grades: Alcohol consumption might reduce children's motivation and concentration when it comes to studying.

Alcohol → Weapon: Children who drink can be more aggressive and thus more likely to carry a weapon.

Fight → Grades: Children involved in more fights should have lower grades.

Fight → Weapon: Children often involved in fights are more likely to carry a weapon to protect themselves or scare other children.

Grades → Weapon: Children with lower grades might feel less important or neglected so that can lead in carrying a weapon in order to "prove" themselves in some way.

Bullied → Depression: Students that get bullied are more likely to get depressed.

Sports → Unsecure: People who practice sports often believe more on themselves and their capabilities.

Sports → Depression: Sports are known to help counteract or prevent depression.

We tested the network's independencies (see section 5.5.1) and saw that many implied conditional independencies were challenged by the data. We resolved some of the most important of these independencies so as to improve our results. We first made a list of all the inconsistencies and then focused on those conections with a high estimate value. The connections we added can be viewed at Figure 1 (blue, dashed arrows). When doing so we also took into consideration that the relation between the nodes had sense.

Throughout the creation of the before mentioned network, we created some other prototypes by adding even more connections to the structure of the network. These prototypes performed worst than the proposed one (mainly due to the lack of data in our dataset) and will not be present neither in this report nor in the code handed with it.

4 Implementation of the network

4.1 Programming language and libraries

We used R as the programming language of our project. We could not find a single library that contained all the functions required for this assignment, and, thus, we considered two:

- `dagitty` [4]: This library allowed us to easily test the conditional dependencies within our network, so we could update our initial prototype (see section 3).
- `bnlearn` [5]: Using this library we were able to infer the probability tables from the data and test the network results using cross validation.

4.2 Structure of the program

Two scripts have been uploaded together with the report and the dataset:

- A file `networkTesting.R` that contains many of the functions used to test the network. This file was included so it could help to understand what has been done for the assignment.
- A file `BN_As1_Weapons.R` that is the final product. It is a simple script that creates the final network and infers the probability tables from the data. It also contains some examples on how to make queries.

5 Testing

5.1 Structure of the network

Throughout this section, we discuss the methods used to test the network or to modify the nodes (trying to get rid of scalability issues).

5.2 Binning

Most of the variables in our dataset had around 7 or 8 possible outcomes. In the context of our network, this could be a problem, specially considering that some of our nodes (*Grades* or *Weapon*) are connected to a big number of other nodes. For instance, if we wanted to obtain the probability table of the node *Weapon* we would need more parameters than the number of entries in our current dataset.

Due to this problem, we decided to reduce the number of levels of many of our variables:

- Binary variables (*Sex*, *Bullied*, *race* and *Depression*) were not modified.
- Variables *Fight*, *Alcohol*, *Weapon* and *Unsecure* were converted to binary (They had 5 or 7 different values in the original dataset). This was made because we do not care about the grading of these variables. For example, we are interested in knowing if a student carried a weapon to school, we are not interested in knowing if he does that everyday or twice a week.
- Some other variables with either 7 or 8 levels (*Sports* and *Grades*) were reduced to 4 levels. That was done in a way that 1 represents eg. not practicing sports at all or lowest possible grade, 4 represents the two highest grades or a person that trains 6 or 7 times a week. Number 2 and 3 are values in between.
- The outputs of the variable *Age* was reduced from 7 to 5. We merged the 3 first values into one because the two first ones didn't have enough people (less than 0.3% if we sum them).

5.3 Divorcing

We won't consider divorcing as an option because it could lead to a great lost of information. In our opinion, there exist no connections in this network that would be easily divorced.

5.4 Markov Blanket

It is easy to prove that the Markov blanket of node *Weapon* does not include all the other nodes in any of the proposed structures. However, we have not deleted any node that is not included in that blanket. The reason for this is that this network is thought to be used by people that might not know the values of all the nodes and thus, they could use information that they know (even though it is not part of the Markov blanket) to obtain more precise queries.

For example, we can think on a head teacher in a school. When making a query about one of his students, that person probably knows values such as *Age*, *Race* or *Sex*. However, it is unlikely that he would know values such as *Unsecure* or *Alcohol* that are also part of the Markov blanket.

5.5 Checking the network

We used two different methods to test our networks: One that would test the structure and one that would test their ability to predict the value of *Weapon*.

5.5.1 Checking structure

In order to check our network structure we extrapolated all conditional independencies that can be found in our first prototype. Our goal was finding out if these independencies can hold against the data. After extrapolation, we performed regression so we could see estimates, p-values and standard errors for each independency. Those connections with p-value less than 0.05 are considered wrongly assumed. Furthermore, big estimate means that small change in input results in big change in output so we focused on dealing with independencies with big estimates first because they are more relevant.

5.5.2 Checking predictions

We used the function `bn.cv` from the library `bnlearn`. This function allows us to run a 10-fold cross-validation on the network to test its predictions when asked by the node *Weapon*. The predictions done using only information from the parent nodes, since the function does not select random nodes to run the test.

The mean classification error obtained by this method was, in both cases, really similar and do not differ much from trial to trial. However, these results might be misleading. Our dataset contains a considerable number of entries, but it is really unbalanced (most of the teenagers do not carry any kind of weapon to school). Because of this, even though there are small differences between the classification error between both configurations, the algorithm being applied for calculating the classification error might not be the best option, since just predicting that all teenagers will not carry any weapon to school results in quite a high accuracy.

6 Inference problem

Our main goal is to determine the importance of the selected attributes in predicting how likely it is for a given individual to carry some kind of weapon in school.

Users of our network can infer whether specific students are likely to carry a weapon in school, given other attributes such as gender, age, alcohol use, etc. We suppose that the application will be used by people who does not know the values of all the nodes.

7 Use of the network

We already suppose that the users know how to call and use functions in R and, thus, we will not explain in detail how to do that.

In order to try the network the user should open an R command-line interpreter and run our script `BN_As1_Weapons.R`. Note that the dataset used must be in the same folder as the script and that the name must be `"ProjectData.csv"` (Unless the name has been previously changed). A small script will run creating the final Bayesian network and fitting its parameters. The name of such a network will be *fittebn* Now, the user can query any question they want to test. The script contains some examples on how to do this, but it will not automatically run any.

8 Conclusions

We proposed two Bayesian networks (the first prototype and the final one) that could predict the chances of a teenager carrying a weapon in school. We conclude that both configurations perform similarly but we chose the second one because, even though the classification error is almost equal (most probably due to our unbalanced dataset), it has a better structure according to the data.

The huge number of parameters that some of the nodes had (*Grades* and *Weapon*) was an important problem that we had to solve. We tried to do so by binning most of our nodes but, even then, we still have a problem with the node *Weapon*. Several connections end to this node and, if we want to infer the probability table of the node using the data, we would need many entries in our database that contain information about people carrying weapons to school. However, our initial dataset only contains about 500 of these entries, an insufficient number given the total number of parameters that the node must take. Against this situation, we could either keep binning our variables or generate fake data. Because we wanted to keep working with the original dataset and we though that binning the data even more would render the network incapable of predicting any coherent value, we decided to leave the network as is. It is important to know this limitation when querying the network, because it means that some queries could result in a probability of 0 or 1 if we use all the nodes of the Markov Blanket of *Weapon*. The user should keep in mind that this happens because of the lack of data.

References

[1] Child Trends - High School Students Carrying Weapons
`http://www.childtrends.org/indicators/high-school-students-carrying-weapons`

[2] Youth Risk Behavior Survey (YRBS) (1995-2015)
`http://www.cdc.gov/healthyyouth/data/yrbs/data.htm`

[3] Charisse Nixon, Stan Davis - The Youth Voice Project
`http://www.youthvoiceproject.com`

[4] dagitty - Graphical Analysis of Structural Causal Models
`https://cran.r-project.org/web/packages/dagitty/index.html`

[5] bnlearn - Bayesian Network Structure Learning, Parameter Learning and Inference
`https://cran.r-project.org/web/packages/bnlearn/index.html`