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, especially 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 the latter children can have relatively easy access to guns can result in enormous tragedies, as seen in the past. This kind of environment is unacceptable and something has to be done in order to prevent further violence. In this project we are focusing on some of the individual's characteristics (and the connections between them) in order to predict whether a teenager is likely to carry a gun in school.

2 Data

In our project, we use data from the 2015 "Youth Risk Behavior Survey (YRBS)", which analyses health-risk behaviours in teenagers in the United States. This data is public and contains the answers given by 15.624 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.

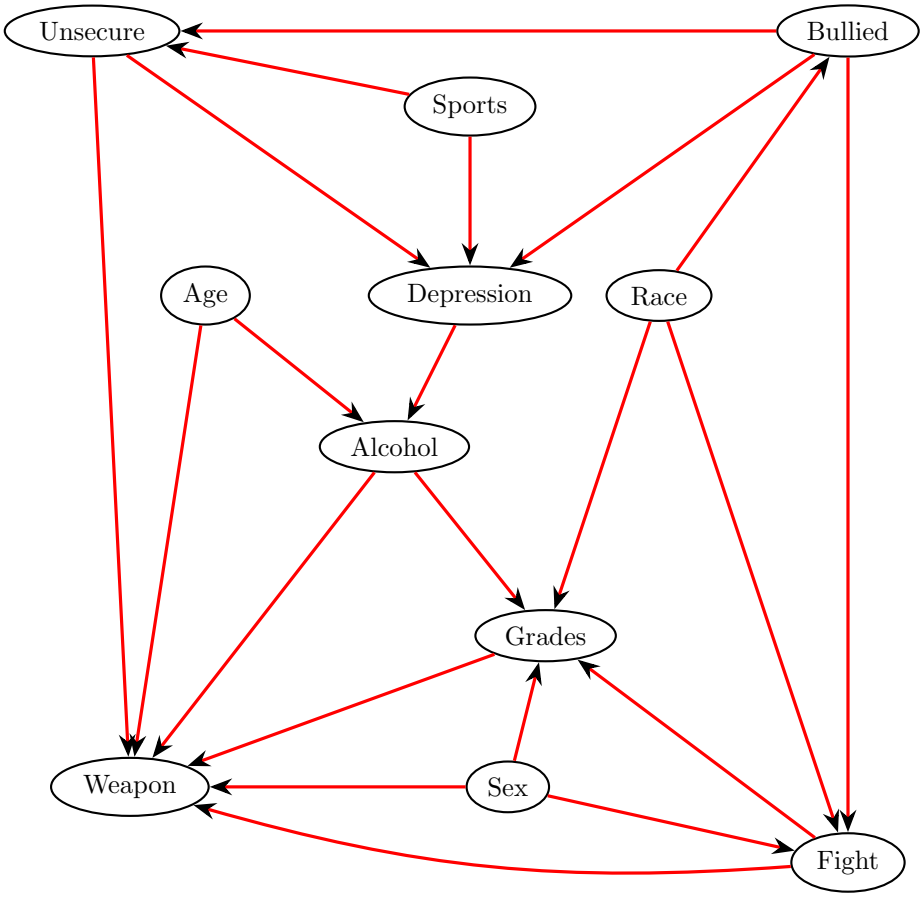
Some of the students are discarded for our study because they didn't answer some questions or their answers were already discarded by the YRBS because they were inconsistent.

We use the following attributes for our Bayesian network. They are all related to some question asked in the mentioned survey.

Name	Type	#Levels	YPBS question
Race	categorical	8	Raceeth (Q4 & Q5)
Age	categorical	7	Q1
Sex	categorical	2	Q2
Carry weapon in school	categorical	2	Q15
Feel unsecure in school	categorical	2	Q16
Participate in fight last year	categorical	4	Q20
Bullied in school	categorical	2	Q24
Suffered depression last year	categorical	2	Q26
Alcohol use last month	categorical	4	Q43
Sports practiced	categorical	4	Q80
Grades	categorical	4	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 Network



[TODO: replace .5 with probabilities extracted by software.]

For this assignment, we did not infer the relations between variables from data, but we assumed the network structure based on our understanding of the domain. Following is a short argumentation for each structural decision we made.

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:

Race → Bullied:

Race → Fight:

Race → Hidden variable:

Sex → Fight:

Sex → Grades:

Sex → Weapon:

Hidden variable → Grades:

Unsecure → Depression:
Unsecure → Weapon:
Depressed → Alcohol:
Alcohol → Grades:
Alcohol → Weapon:
Fight → Grades:
Grades → Weapon:

Bullied → Depression: Students that get bullied are more likely to get depressed. Sports $\not\rightarrow$ Anything: We couldn't find a sensible relation with anything, but we wanted to keep the variable for the second assignment, so as to find out if the model can improve by incorporating this variable.

We implement our Bayesian Network using probability tables and Python's pgmpy library. We implement our Bayesian Network using probability tables and Python's libpgm library.

Report mean squared error. (or use another error measurement) (or say that we didn't test)

4 Inference problem

Our main goal is to determine the importance of the selected attributes in predicting how likely 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.