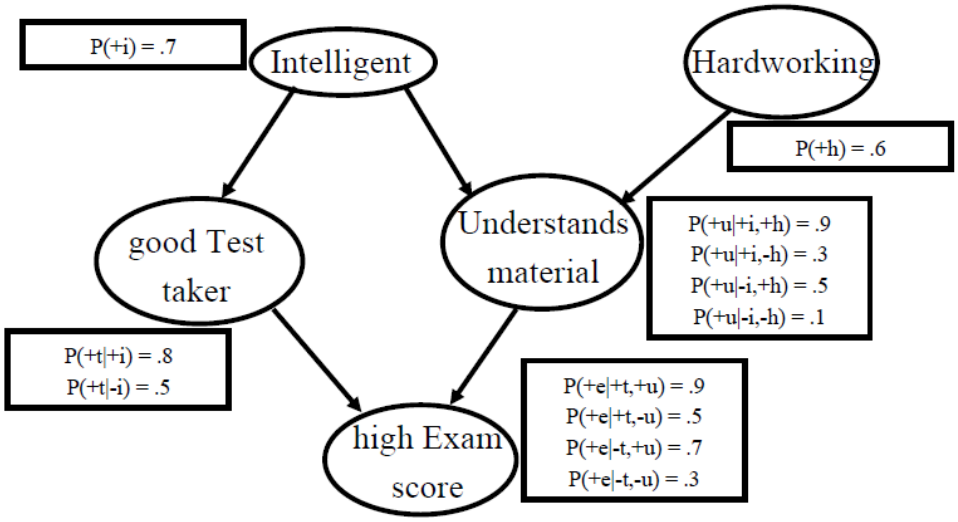
Homework 1

Q1

We are going to take the perspective of an instructor who wants to determine whether a student has understood the material, based on the exam score. Figure 1 gives a Bayes net for this. As you can see, whether the student scores high on the exam is influenced both by whether she is a good test taker, and whether she understood the material. Both of those, in turn, are influenced by whether she is intelligent; whether she understood the material is also influenced by whether she is a hard worker.



1) Using any algorithm (by hand!), compute the probability that a student who did well on the test actually understood the material, that is, compute P(+u| + e).

2) For the above Bayesian network, label the following statements about conditional independence as true or false.
For this question, you should consider only the structure of the Bayesian network, not the specific probabilities.

Explain each of your answers.
1. T and U are independent.
2. T and U are conditionally independent given I, E, and H.
3. T and U are conditionally independent given I and H.
4. E and H are conditionally independent given U.
5. E and H are conditionally independent given U, I, and T.
6. I and H are conditionally independent given E.
7. I and H are conditionally independent given T.
8. T and H are independent.
9. T and H are conditionally independent given E.
10. T and H are conditionally independent given E and U.

Q2

In this part you will be analyzing risk factors for certain health problems (heart disease, stroke, heart attack, diabetes). The data is from the 2011 Behavioral Risk Factor Surveillance System (BRFSS) survey, which is run by the Centers for Disease Control (CDC). The distilled data is in the spreadsheet **RiskFactorData.csv**.
The variables and their meanings are as follows:

• income - Annual personal income level. 1 (< $10,000) 2 ($10,000 - $15,000) 3 ($15,000, - $20,000) 4 ($20,000 - $25,000) 5 ($25,000 - $35,000) 6 ($35,000 - $50,000) 7 ($50,000 - $75,000) 8 (> $75,000)

• exercise - Exercised in past 30 days. 1 (yes) 2 (no)

• smoke - Smoked 100 or more cigarettes in lifetime. 1 (yes) 2 (no)

• bmi - Body mass index (category). 1 (underweight) 2 (normal) 3 (overweight) 4 (obese)

• bp - Has high blood pressure. 1 (yes) 2 (only when pregnant) 3 (no) 4 (pre-hypertensive)

• cholesterol - Has high cholesterol. 1 (yes) 2 (no)

• angina - Had heart disease (angina). 1 (yes) 2 (no)

• stroke - Had a stroke. 1 (yes) 2 (no) • attack - Had a heart attack. 1 (yes) 2 (no)

• diabetes - Had diabetes. 1 (yes) 2 (only during pregnancy) 3 (no) 4 (pre-diabetic)

1) Create the following Bayesian network to analyze the survey results. **(Write a piece of code to read the CSV file and calculate the probabilities of CPTs)**

What is the size (in terms of the number of probabilities needed) of this network? Alternatively, what is the total number of probabilities needed to store the full joint distribution?
2. For each of the four health outcomes (diabetes, stroke, heart attack, and angina), answer the following by querying your network (**Write your own code or use GeNIe** https://dslpitt.org/genie/):

(a) What is the probability of the outcome if I have bad habits (smoke and don't exercise)? How about if I have good habits (don't smoke and do exercise)?

(b) What is the probability of the outcome if I have poor health (high blood pressure, high cholesterol, and overweight)? What if I have good health (low blood pressure, low cholesterol, and normal weight)?

3. Evaluate the effect a person's income has on their probability of having one of the four health outcomes (diabetes, stroke, heart attack, and angina). For each of these four outcomes, plot their probability given income status (your horizontal axis should be i = 1; 2; : : : ; 8; and your vertical axis should be P(y = 1/ income = i), where y is the outcome). What can you conclude?