University of Pittsburgh
CS2750 Machine Learning
Professor Milos Hauskrecht

Handout 7
March 17, 2022

Problem assignment 7

Due: Thursday, March 24, 2022

Problem 1. Conditional independence

Random variables A,B are conditionally independent given C when:

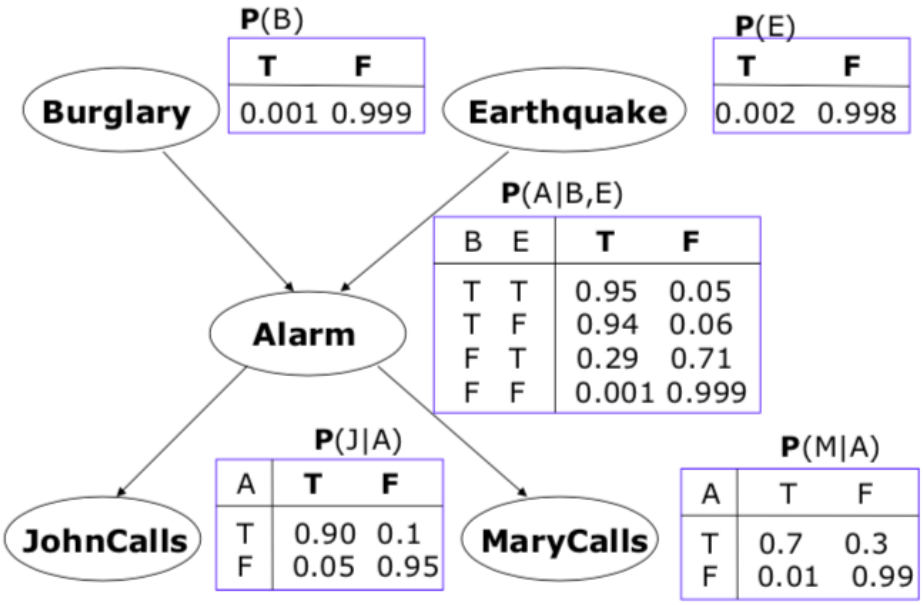
$$P(A, B|C) = P(A|C)P(B|C).$$

Prove that this implies:

$$P(A|B, C) = P(A|C).$$

Problem 2. Bayesian belief networks

Assume the Alarm network from Lecture 16 that is shown the figure bellow.

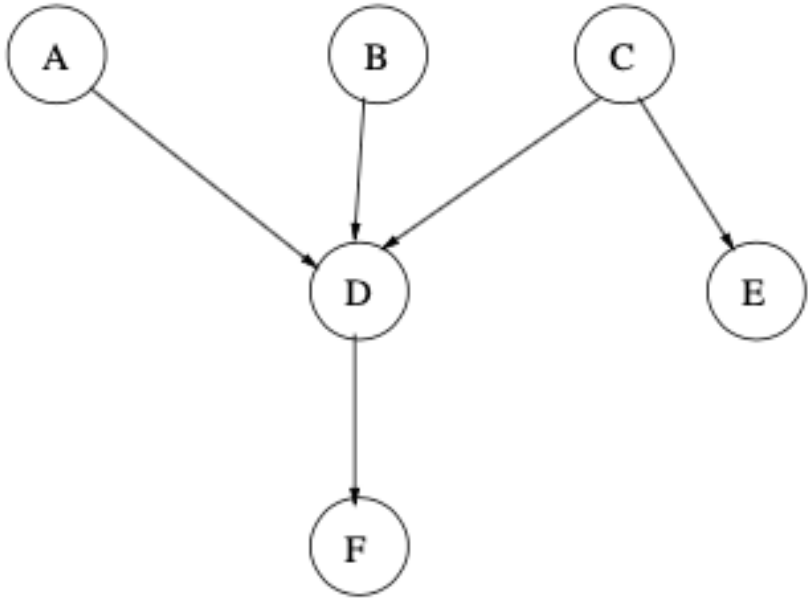


Part a. Express the joint probability $P(E = F, B = F, A = F, J = T, M = F)$ in terms of local conditional probabilities parameterizing the Bayesian belief network.

Part b. Calculate the probability $P(E = F, B = F, A = F, J = T, M = F)$ by substituting the local conditional probabilities to numbers.

Problem 3. Bayesian belief networks

Assume the Bayesian belief network in the figure below. Assume that every variable in the network is binary representing T,F values, except variable D that can take on three possible values T,F,X (X stands for undecided).



Assume you want to compute $P(B = T, E = T)$.

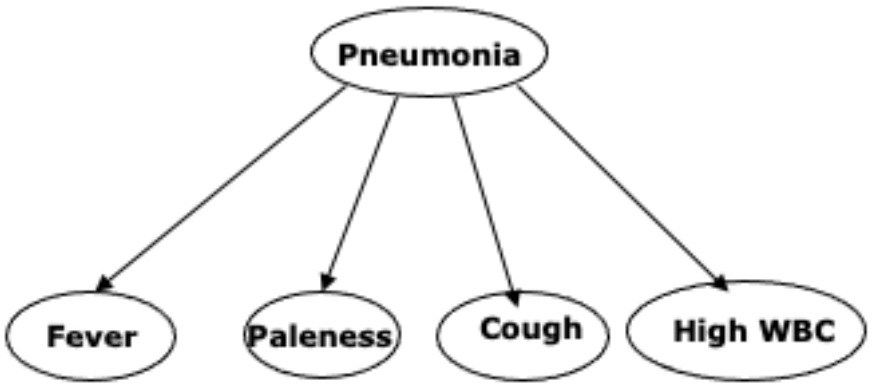
Part a. Assess the computational cost of the (blind) solution that (1) calculates $P(B = T, E = T)$ from the full joint, and (2) uses the BBN representation to express the full joint, in terms of the number of additions and multiplications and report the results in the report.

Part b. Show how would you compute the expression more efficiently by interleaving products and sums. Provide a new expression for calculating $P(B = T, E = T)$ and show its computational cost in terms of the number of additions and multiplications. Compare the two solutions and discuss the results in your report.

Problem 4. Pneumonia diagnosis.

Assume a Bayesian belief network with the Naive-Bayes structure for a simplified version of the pneumonia problem (see figure below) The variable Pneumonia is at the root of the Naive Bayes network and features (attributes) Fever, Paleness, Cough and HighWBC are conditionally independent given Pneumonia. Assume that random variables in our model are discrete with the following set of values:

- **Pneumonia:** True, False
- **Fever:** True, False
- **Paleness:** True, False
- **Cough:** True, False
- **HighWBC:** True, False



Part a. Write and submit a program PD_learning.py that takes the data in the file 'pneumonia.csv', learns the ML estimates of parameters of the BBN network, reports the BBN parameters and finally saves the learned network. You can use different formats to save the BBN model using either json format or pickle libraries. Please make sure you know how to recover (load) the model back from the file.

The pneumonia file given to you includes 500 examples (in rows). The features are in columns, in the following order: Fever, Paleness, Cough, HighWBC. Last column represents Pneumonia variable. The data are represented such that True corresponds to 1 and False to 0. Report the parameters of the network in your report.

You can use different formats to save the BBN model: json format, pickle or others. Please make sure you know how to recover (load) the model back from the file (see below).

Part b. Assume that the patient comes with the following set symptoms: Fever and Cough are true; Paleness and HighWBC are false. What is the probability $P(Pneumonia = T|Fever = T, Paleness = F, Cough = T, HighWBC = F)$, that is, the probability that the patient suffers from Pneumonia, given the symptoms? Simplify the expression as much as possible before plugging in the values.

Part c. Assume that the patient reports Cough and a Fever (they are true); and values of Paleness and HighWBC are not known. Compute the probability $P(Pneumonia = T|Fever = T, Cough = T)$ of the patient suffering from the pneumonia, given the symptoms? Simplify the expression as much as possible before plugging in the numbers.

Part d. Write and **submit** a Python program $PD\_inference.py$ that reads the previously saved BBN model, a combination of patient symptoms (the values of Fever, Paleness, Cough, HighWBC) from the file Example.csv and computes the probability of $P(Pneumonia = True|CurrentSymptoms)$. The values of symptoms for the current patient case are given in the following order:

$$Fever, Paleness, Cough, HighWBC.$$

The values are encoded as follows:

- 0 for False,
- 1 for True, and
- -1 for not given (not known).

For example, the symptoms of the patient case in part b are encoded as a vector of 4 values:
(1 0 1 0).
and symptoms in part c as:
(1 −1 1 −1).

Note: Please note that your program should work on an arbitrary combination of input values in the example file.