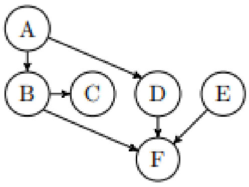
Exercise 5 – Bayesian Networks

Part A: Theoretical Part

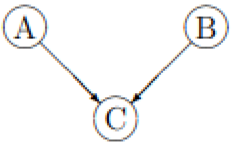
Consider the following Bayesian network:



1. Rewrite the joint probability distribution $P(A, B, C, D, E, F)$ using the conditional independencies expressed by the network
2. Suppose that all the random variables A, B, C, D, E, F in the Bayesian network can only have two possible values: "yes" and "no". What's the minimum number of parameters required to fully define the Bayesian network whose structure is given above?
   - Hint: Remember that e.g. $P(E = yes) = 1 - P(E = no)$
3. How many parameters would be required to define the full joint distribution over A, B, C, D, E, F if we could not assume the conditional independencies expressed by the Bayesian network?
4. Which of the following propositions hold true for the given Bayesian network? Show an active path (for false) or that all of the paths are blocked (for true)
   - $D \perp\!\!\!\perp E \mid \emptyset$
   - $C \perp\!\!\!\perp E \mid F$
   - $C \perp\!\!\!\perp E \mid A$
   - $A \perp\!\!\!\perp F \mid \{B, D\}$

   Note: "$X \perp\!\!\!\perp Y \mid Z$" denotes "$X$ and Y are d-separated by Z"
5. Consider the following joint probability distribution over the three random variables A, B, C

p_{ABC}	$A = a_1$		$A = a_2$	
	$B = b_1$	$B = b_2$	$B = b_1$	$B = b_2$
$C = c_1$	$^3/_{32}$	$^1/_4$	$^1/_{40}$	$^3/_{32}$
$C = c_2$	$^1/_{32}$	$^1/_8$	$^1/_{10}$	$^9/_{32}$



Can the graph depicted next to the table describe the joint distribution in the table? Why?

Instructions/comments for Part A of the exercise:

- Submit all your answers in a single self-contained and well-explained PDF file.

- Part A of the exercise is independent of part B. Instructions for part B are provided at the end of the document.

Part B: Practical part -- The Bayesian Mechanic

On your last road trip your old car started to make weird noises, you noticed smoke rising from the engine, and a high-temperature alert started flashing on the dashboard. Having to wait for a tow truck to take your car to the mechanic ruined your trip, and you decided that you must use your newly acquired knowledge in Bayesian networks to create an automatic diagnosis tool for car problems called The Bayesian Mechanic™ so that you won't be so hopeless next time.

You wrote down a list that sums up your domain knowledge about car problems:

- You know that cars that are made in China are more likely to have faulty dashboards that randomly show high temperature or low water level alerts.
- You know that the low water level dashboard alert pops up when either there's not enough water in the water tank, or when the dashboard is faulty.
- You know that the high-temperature dashboard alert pops up when there's not enough oil in the oil reservoir, when there's not enough water in the water tank, or when the dashboard is faulty.
- A friend of yours once said that even though data scientists are really smart, they don't know the first thing about car maintenance, and so the water and oil levels in their cars are more likely to be low.

You post a survey online to get data to train your network, and get lots of results. The data are provided to you in a csv file called 'bayesian_mechanic_data_2000_results.csv'. Take a look at the file to see the name convention (e.g., NOOIL means no-oil).

It's time to get to work!

1. Use your domain knowledge to create a Bayesian network model for the car temperature problem. (draw the network by using the Pomegranate python package)

2. Use the data you have to learn the parameters of the model. You should do it by building your graph based on the Pomegranate python package

3. To get a better understanding of the model, please use d-separation rules to answer:
   i. Is {NOOIL} d-separated from {NOWATER} given {DATASCI}? Why?
   ii. Is {DATASCI} d-separated from {WATERALERT} given {NOWATER}? Why
   iii. Is {CHINESE} d-separated from {NOOIL} given {DATASCI} ? Why?
   iv. Is {CHINESE} d-separated from {DATASCI, NOOIL} given {TEMPALERT,NOWATER}? Why?

4. Answer the following questions based on the model you have built:
    i. What is P(NOWATER|WATERALERT=2,TEMPALERT=2)
    ii. What is P(NOOIL|TEMPALERT=2)
    iii. What is P(FAULTY|TEMPALERT=2, WATERALERT=2)
    iv. What is P(TEMPALERT|DATASCI=2)?
    v. What is P(TEMPALERT|DATASCI=1)?
    vi. What is P(WATERALERT|CHINESE=2)?
    vii. What is P(NOWATER)?
    viii. What is P(FAULTY)?
    ix. What is P(FAULTY|OR=2)?
    x. What is P(FAULTY|OR=2,NOWATER=2)?

5. A simple way to estimate the model's performance is using the log-probability of the model over all instances it was trained on.
    i. Calculate the log-probability of your model
    ii. Re build your model, but this time use the **Chow-Liu algorithm**. Re calculate the log-probability and compare with the results you get in the previous subsection. Does results make sense? Shortly explain why

Instructions/comments for Part B of the exercise:

- All of the nodes in the model are going to be binary, e.g there is either enough oil or not, a car is either made in China or not, etc.
- The data are provided as a csv file with the node names in the first row.
    - The OR column is a binary or operator between the FAULTY and NOWATER variables. You can use it as a node in your model.
- You should submit your answers and your full code as a single Jupiter notebook. Make sure the notebook is readable and self-explained. The notebook must include:
    - A visualization of your network and an explanation regarding your modeling choices (why you chose to connect nodes, and why in that specific direction). Don't waste a lot of time creating a fancy drawing.
    - Results to all sections of the HW (1 to 5)
- Take some time to explore the Pomegranate package and understand the logic behind the classes/functions you can use for building your Bayesian network.

Good luck and a have a safe journey!