Probabilistic knowledge: representation and reasoning

Problem 1.

Suppose we have been traveling the whole day and we have left our laptop switched on at home, plugged in and with the battery fully charged, performing a series of calculations. When we go back home, we could meet the **laptop turned off**. Basically, there are two reasons that could influence this: that a **power failure** had ocurred, or that the **battery had failed**. Batteries are known to fail more often on **hot days**. Also on hot days it is more likely that due to massive consumption of air conditioning, the network gets overloaded and we have a power failure. Also another possible cause of electrical cuts may be a **thunderstorm**. We have been out and we cannot tell if there has been a storm, but a storm occasionally comes accompanied by **rain**, and this in turn **wets the garden** at the entrance to the house.

- Model this situation using a bayesian network, clearly specifying the random variables, their possible values, the order used to construct the network and the conditional independences assumed. Give also the probability tables (with reasonable probabilities).
- Why a bayesian network is a compact representation of the full joint distribution. Explain it using the previous bayesian network. Give two reasons why this representation is better than the full joint distribution.
- According to that bayesian network, are to meet the laptop turned off and the garden wet independent? Are they conditionally independent given that we know if we had a thunderstorm or not? Justify your answers using the d-separation criteria.

Problem 2. The fact that the garden of my house is **flooded** on a weekend, is subject to uncertainty, and several factors may influence that. For example, it could be relevant the fact that the the **sprinklers** were working, or that it **rains** heavily that weekend; it may even influence whether that weekend we stay **at home** or not, since if we're at home we could avoid possible flooding more likely. Assume that rain on a weekend is influenced by the fact that previous Friday has been **cloudy**, and also by the temperature (where temperature may be *high*, *middle* or *low*). Also, the higher is the temperature, the more likely to go outside the weekend (but that is not influenced neither by the rain nor the clouds). Besides, sometimes, if it's cloudy on Friday, we disconnect the sprinklers for the weekend.

Model this situation using a bayesian network, making clear that random variables and their possible values, the order in which the variables are drawn, and the conditional independences assumed. Give also (reasonable) probability tables for the network.

Problem 3. Assume we have five random variables A, B, C, D and E, such that:

- B is independent from A.
- C is independent from A and from B.
- D is conditionally independent from C given A and B.

- E is conditionally independent, given B and C, from the rest of variables.

Draw a bayesian network which expresses the dependency/independency relationships listed above. Assume that we know the following probabilities: $P(a) = 0{,}2$, $P(b) = 0{,}5$, $P(c) = 0{,}8$, $P(d|\neg a, \neg b) = 0{,}9$, $P(d|\neg a, b) = 0{,}6$, $P(d|a, \neg b) = 0{,}5$, $P(d|a, b) = 0{,}1$, $P(e|\neg b, \neg c) = 0{,}2$, $P(e|\neg b, c) = 0{,}4$, $P(e|b, \neg c) = 0{,}8$ and $P(e|b, c) = 0{,}3$.

We need to calculate $P(a, b, c, d, e)$, $P(\neg a|b, c, d, e)$ and $P(e|a, \neg b)$.

Problem 4. Assume that we have a probabilistic model that expresses the way how electrical issues and hardware malfunction affect informatic failures. We have thus three random variables, E (electrical issue), H (hardware malfunction) and I (informatic failure), and we assume that E and H are independent. Besides, we know the following probabilities: $P(e) = 0{,}1$, $P(h) = 0{,}2$, $P(i|\neg e, \neg h) = 0$, $P(i|\neg e, h) = 0{,}5$, $P(i|e, \neg h) = 1$ and $P(i|e, h) = 1$. Draw the corresponding bayesian network and calculate $P(\neg e, h, i)$, $P(h|e)$ and $P(e|i)$.

Problem 5. A study on the influence of smoking related to lung cancer is designed using four variables: a person develops cancer (C), a person is a smoker (S), is a passive smoker (PS) and has smoking parents (SP). Represent, by means of a bayesian network, a causal model describing the influence of these variables among them. After drawing the network, provide a motivated answer (using d-separation criterion) to all questions of the type: is X conditionally independent from Y given Z? (where X, Y and Z are any possible choice of three different variables in the network).

Problem 6. An automatic diagnosis system can help to diagnose the influenza virus (IV) or the smoking habit (SH) as causes for bronchitis (B). In order to do so, we take into account several symptoms of a patient: those directly related to bronchitis, like cough (C) and breathing difficulties (D); and also those related to the influenza virus, fever (F) and throat ache (T). Represent, by means of a bayesian network, a causal model describing the influence of these variables among them.

Problem 7. The risk of myocardial infarction (I) is known to be associated, among other factors, with both fluid retention (L) and hypertension (H). A possible cause of both fluid retention and hypertension is excessive salt intake (S). Assuming as random variables the four mentioned above, model the situation described by means of a bayesian network, choosing a suitable order for drawing the variables and specifying the relationships of conditional independence that have been assumed.

Problem 8. The IEEE is designing a prototype of a robot for domestic assistance, and we are in charge of preparing public demos for potential buyers. We have received a load of 100 robots. We know that only one of them comes from a damaged series, but all of them look exactly the same.

The tech guys from the factory tell us that te robots from this damaged series have a 50 % probability of experimenting some small failures during the demo, while for the rest of robots the failure probability is only 2 %.

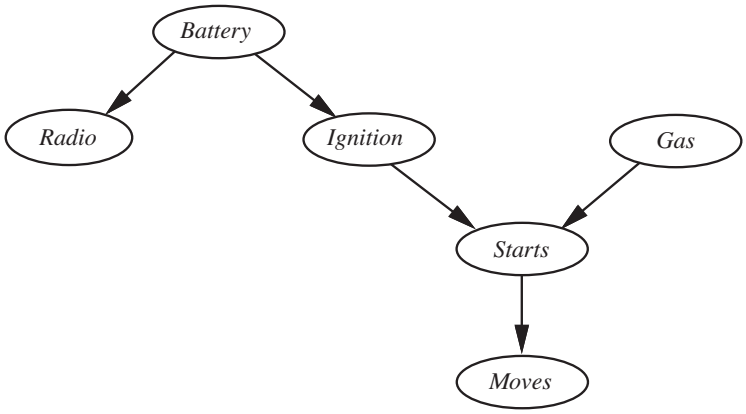
On the other hand, the marketing team has observed that 80 % of the clients that see a failure-free demo will actually buy the robot, while only 0,5 % of the clients who see a demo where a failure happens decide to buy it anyway.

You should:

- Design a bayesian network (structure and associated tables) capturing all the above information.

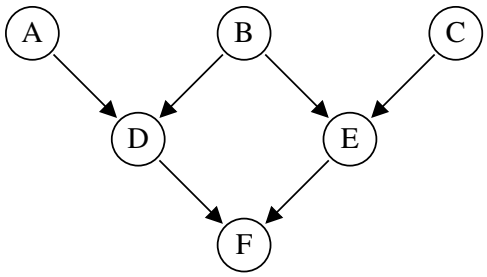
- Calculate the probability that the robot used in the demo was the damaged one, knowing that the client bought the robot.

Problem 9. (**Russell & Norvig, ej. 14.8**) Consider the following network for car malfunction diagnosis (all variables are Boolean):



- Extend the network by adding two variables $IcyWeather$ and $StarterMotor$
- Give reasonable conditional probability tables for all the nodes.
- How many independent values are contained in the joint probability distribution for eight Boolean nodes, assuming that no conditional independence relations are known to hold among them?
- How many independent probability values do your network tables contain?
- Using these probability values, calculate the probability that a car that does not start has a problem with the battery.

Problem 10. Consider the following bayesian network for the random variables A, B, C, D, E and F:



with the following probability distribution tables:

$P(a)$
0,4

$P(b)$
0,3

$P(c)$
0,7

A	B	$P(d\|A,B)$
a	b	0,8
a	$\neg b$	0,2
$\neg a$	b	0,7
$\neg a$	$\neg b$	0,3

B	C	$P(e\|B,C)$
b	c	0,2
b	$\neg c$	0,3
$\neg b$	c	0,5
$\neg b$	$\neg c$	0,8

D	E	$P(f\|D,E)$
d	e	0,1
d	$\neg e$	0,8
$\neg d$	e	0,2
$\neg d$	$\neg e$	0,7

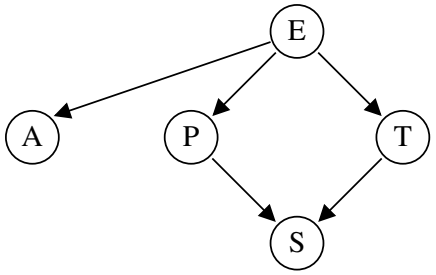
You should:

- Calculate the distribution of the random variable B assuming that we have observed that F is True, by using the **algorithm of inference by enumeration**.
- Calculate the distribution of the random variable B assuming that we have observed that F is False, by using the **algorithm of variable elimination**.

Problem 11. Consider the following random variables describing several circumstances related to the exam to get a driving license:

- E: time devoted to study was not enough.
- T: the theoretical exam was not good.
- P: the practical exam was not good.
- A: all traffic pannels have been studied.
- S: the license was not obtained.

Assume that all the (uncertain) knowledge about this domain is expressed by means of the following bayesian network:



with the following probability distribution tables:

$P(e)$
0,7

E	$P(a\|E)$
e	0,2
$\neg e$	0,8

E	$P(p\|E)$
e	0,3
$\neg e$	0,8

E	$P(t\|E)$
e	0,8
$\neg e$	0,4

P	T	$P(s\|P,T)$
p	t	0,9
p	$\neg t$	0,6
$\neg p$	t	0,5
$\neg p$	$\neg t$	0,1

Give a motivated answer to the following questions.

- Is it possible to calculate any value of the joint distribution, given the network and its associated tables?

- Are P and A independent? Is it true that our degree of belief about S knowing the value of T can be different if we also know the value of E? (that is, the question is wether $\mathbf{P}(S|T,E) = \mathbf{P}(S|T)$ holds) Is S conditionally independent from A given E? Are P and T conditionally independent given S? In the motivation of the answers, use the d-separation criteria.

- Calculate the probability of the event: the time devoted to study was not enough, and both exams (theoretical and practical) were not good, traffic pannels have not been studied, and the license is obtained.

- Assume that the license was not obtained. Which is the probability that the time devoted to study was not enough?

Problem 12.

In *St Peter's High School* there are four types of **Students (S)** which can be classified (in decreasing order) as *a, b, c* and *d*, according to their working capabilities and studying skills. Thus, students of type *a* are the best prepared ones for the examination test, and type *d* are the worst. On the other hand, students think that there are three types of **Teachers (T)**: 70 % of them are *fair (f)*; 10 % are *highly demanding (h)*; and 20 % are *moderate (m)*. This classification is made according to the number of students that **Pass (P)** their tests, as follows:

- With *fair* teachers, usually 100 % of students *a* and *b pass the test*, while for types *c* and *d* it's only 50 % and 10 %, respectively.

- With *highly demanding* teachers, usually 90 % of students *a pass the test*, together with 70 % of students *b*, 30 % of students *c*, and 0 % of students *d*.

- With *moderate* teachers, usually 100 % of students *a* and *b pass the test*, while for types *c* and *d* it's 60 % and 30 %, respectively.

There is a final test today, and the ratios of each type of students are 10 %, 40 %, 30 % and 20 % for types *a*, *b*, *c* and *d*, respectively. The teacher in charge of giving marks for this test will be randomly chosen. Among type *a* students who pass, 20 % of them will be given a **Job offer (J)**, together with 10 % of type *b* students who pass.

You should:

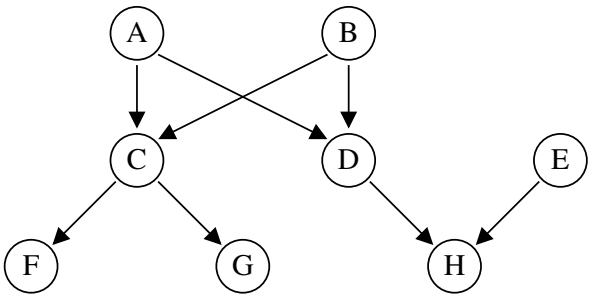
- Design a bayesian network capturing this information, identifying precisely the variables and their domains, nodes, dependencies, arcs and probability tables.

  Use the studied techniques of probabilistic inference to answer the following questions (explaining the intermediate steps):

- What is the probability for a type *b* student to pass the test?

- Consider the case of a student who did not get the job offer, what is the probability that the teacher giving the marks was a highly demanding one? In order to compute this probability, you should use the **variable elimination** algorithm.

Problem 13. Consider the following bayesian network, with random variables A, B, C, D, E, F, G y H:



with the following probability tables:

$P(a)$
0,3

$P(b)$
0,4

A	B	$P(c\|A,B)$
a	b	0,4
a	$\neg b$	0,25
$\neg a$	b	0,4
$\neg a$	$\neg b$	0,3

A	B	$P(d\|A,B)$
a	b	0,8
a	$\neg b$	0,75
$\neg a$	b	0,1
$\neg a$	$\neg b$	0,25

$P(e)$
0,9

C	$P(f\|C)$
c	0,8
$\neg c$	0,5

C	$P(g\|C)$
c	0,2
$\neg c$	0,1

D	E	$P(h\|D,E)$
d	e	0,5
d	$\neg e$	0,05
$\neg d$	e	0,3
$\neg d$	$\neg e$	0,7

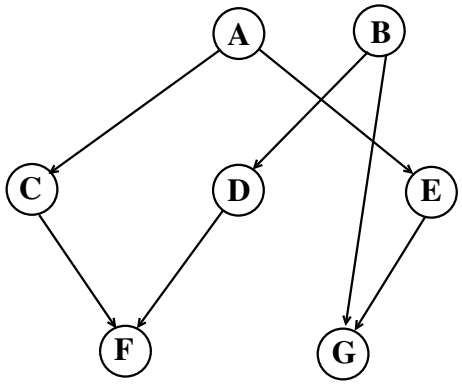
You should:

a) Compute $P(a, b, \neg c, d, \neg e, f, g, \neg h)$

b) Suppose we want to compute $P(\neg f|a, b, \neg d)$ Which variables can be ignored for this particular query? Apply, detailing every step, the variable elimination algorithm to compute that probability.

Problem 14.

Consider the following bayesian network, with random variables A, B, C, D, E, F y G:



with the following probability tables:

$P(a)$
0,3

$P(b)$
0,6

A	$P(c\|A)$
a	0,7
$\neg a$	0,1

B	$P(d\|B)$
a	0,1
$\neg a$	0,9

A	$P(e\|A)$
a	0,3
$\neg a$	0,8

C	D	$P(f\|C,D)$
c	d	0,9
c	$\neg d$	0,7
$\neg c$	d	0,5
$\neg c$	$\neg d$	0,2

B	E	$P(g\|B,E)$
b	e	0,2
b	$\neg e$	0,8
$\neg b$	e	0,2
$\neg b$	$\neg e$	0,9

You should:

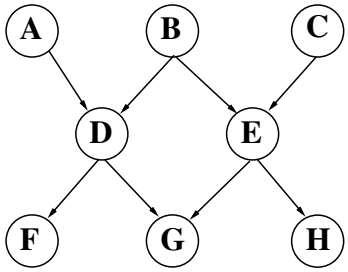
1. Suppose that the network has been constructed following the algorithm given in class, and drawing the variables in alphabetical order. Which conditional independences have been assumed?

2. According to the conditional independences that can be *deduced* from the network, indicate if the following statements are tru or false (justifying your answers using the d-separation criteria)

   - If we know the value taken by A, our degree of belief in the value that C may take is not updated if in addition we knew the value taken by G.
   - F and G are conditionally independent given A
   - $\boldsymbol{P}(F|A,B) = \boldsymbol{P}(F|A,B,G)$

3. Apply the variable elimination algorithm to compute the probability of A being false, given that F is observed to be false and G is observed to be true

Problem 15. Consider the following bayesian network:



$P(a)$
0,2

$P(b)$
0,9

$P(c)$
0,1

A	B	$P(d\|A,B)$
a	b	0,1
a	$\neg b$	0,3
$\neg a$	b	0,7
$\neg a$	$\neg b$	0,9

B	C	$P(e\|B,C)$
b	c	0,9
b	$\neg c$	0,8
$\neg b$	c	0,2
$\neg b$	$\neg c$	0,1

D	$P(f\|D)$
d	0,8
$\neg d$	0,2

D	E	$P(g\|D,E)$
d	e	0,9
d	$\neg e$	0,3
$\neg d$	e	0,1
$\neg d$	$\neg e$	0,8

E	$P(h\|E)$
e	0,2
$\neg e$	0,9

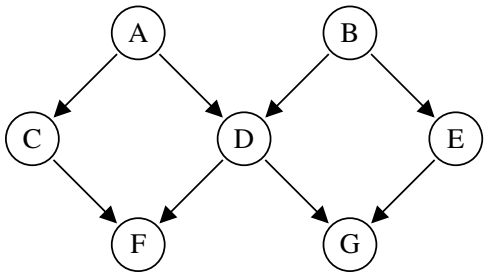
You should:

- Answer the following questions, using the d-separation criteria:

1. Are B and C independent?
2. Are B and C conditionally independent given E?
3. Are F and H independent?
4. Are F and H independent given B?

- Apply the variable elimination algorithm to compute $P(A|b, \neg f)$?

Problem 16. Consider the following bayesian network for the random variables A, B, C, D, E, F and G:



with the following probability distribution tables:

$P(a)$
0,4

$P(b)$
0,3

A	$P(c\|A)$
a	0,7
$\neg a$	0,2

A	B	$P(d\|A,B)$
a	b	0,8
a	$\neg b$	0,2
$\neg a$	b	0,7
$\neg a$	$\neg b$	0,3

B	$P(e\|B)$
b	0,2
$\neg b$	0,5

C	D	$P(f\|C,D)$
c	d	0,1
c	$\neg d$	0,5
$\neg c$	d	0,7
$\neg c$	$\neg d$	0,9

D	E	$P(g\|D,E)$
d	e	0,9
d	$\neg e$	0,7
$\neg d$	e	0,6
$\neg d$	$\neg e$	0,1

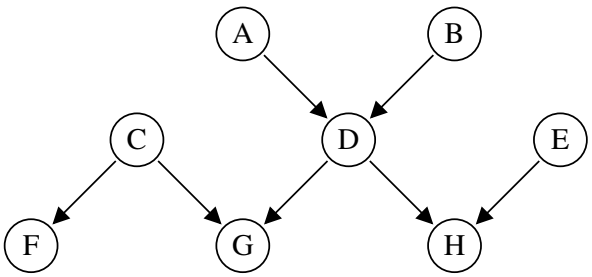
You should:

- Calculate the distribution of the random variable B assuming that we have observed that F is True, by using the **algorithm of variable elimination**.
- Calculate the distribution of the random variable B assuming that we have observed that D is True and F is False, by using the **algorithm of likelihood weighting** with 5 samples, indicating the generated samples and their corresponding weights. Is it necessary to generate random values for all variables in the samples, or can some of them be eliminated? give a motivated answer.

  Consider the following sequence of random numbers in the process of sample generation: 0.13, 0.07, 0.57, 0.94, 0.13, 0.78, 0.48, 0.38, 0.75, 0.93, 0.55, 0.16, 0.91, 0.06, 0.74, 0.02, 0.71, 0.48, 0.10, 0.04, 0.86, 0.70, 0.49, 0.40, 0.77

Problem 17.

Consider the following random variables with random variables A, B, C, D, E, F, G y H:



with the following probability tables:

$P(a)$
0,3

$P(b)$
0,4

$P(c)$
0,8

A	B	$P(d\|A,B)$
a	b	0,8
a	$\neg b$	0,7
$\neg a$	b	0,4
$\neg a$	$\neg b$	0,3

$P(e)$
0,1

C	$P(f\|C)$
c	0,8
$\neg c$	0,5

C	D	$P(g\|C,D)$
c	d	0,1
c	$\neg d$	0,7
$\neg c$	d	0,5
$\neg c$	$\neg d$	0,4

D	E	$P(h\|D,E)$
d	e	0,3
d	$\neg e$	0,8
$\neg d$	e	0,4
$\neg d$	$\neg e$	0,3

Se pide:

1. Suppose we want to compute $P(d|a, \neg g)$. Which variables can be ignored for this particular query? Apply the variable elimination algorithm to compute this probability.

2. Consider the following atomic events:

   (a) $(a, b, c, \neg d, \neg e, \neg f, g, h)$

   (b) $(\neg a, \neg b, \neg c, \neg d, \neg e, \neg f, \neg g, \neg h)$

   (c) $(a, b, c, d, \neg e, \neg f, \neg g, \neg h)$

   Which of them might be generated by likelihood weighting, if we were applying that algorithm to approximate $P(a|\neg g, \neg h)$? Why? In that case, compute the associated weighting

Problem 18. Assume that we want to learn to classify emails as "SPAM" or "not SPAM", from three boolean features X_1, X_2 y X_3. For that, we have a set of 1000 emails already classified: 750 classified as "not SPAM" and 250 as "SPAM". Half of the "not SPAM" emails have feature X_1, a quarter have feature X_2, and 225 have feature X_3. A quarter of the "SPAM" emails have feature X_1, half of them have feature X_2, and 100 have feature X_3.

- Assume a **Naive Bayes** model for thisproblem, showing the corresponding network.
- Estime the probability tables for that network.

- Given a new email, having feature X_1 but without X_2 and X_3, how will it be classiffied, according to **Naive Bayes**?

Problem 19.

A company is promoting three services (**s1**, **s2** and textbfs3) and wants to create an automatic tool that decides for each customer, which product best fits to his or her profile (it's supposed that each client will contract only one service). To this end, the marketing specialists have developed a survey with four yes/no questions (**A**, **B**, **C** and —textbfD) and they think that from the answers to those questions, it is possible to predict which product best suits the customer needs.

Based on previous sales, we have the following data with the answers of 5000 customers and the product they finally purchased:

- 1500 contracted service **s1**, and these are their answers:

A=yes	B=yes	C=yes	D=yes
750	1000	500	1350

- 1000 purchased service **s2**, and these are their answers:

A=yes	B=yes	C=yes	D=yes
500	900	750	100

- 2500 contracted service **s3**, and their answers were::

A=yes	B=yes	C=yes	D=yes
50	2000	1500	500

You should:

1. Show the bayesian network that corresponds with a **Naive Bayes** approach to this classification problem. What conditional indepependence relationships are assumed?
2. If we have a customer answering "no" to all the questions, which service the model will suggest for her? (use Laplace smoothing for the probability estimates)

Problem 20. A team of biologists who were exploring the Amazon river just discovered a new species of insects, which they decided to call *lepistos*. Unfortunately, they are all missing, and the only information we have about such insects consists on the following set of examples found in a notebook. They classify a series of observed insects, according to attributes such as their COLOUR, whether they have WINGS, their SIZE, and their SPEED:

Example	COLOUR	WINGS	SIZE	SPEED	LEPISTO
E_1	black	yes	small	high	yes
E_2	yellow	no	big	average	no
E_3	yellow	no	big	low	no
E_4	white	yes	medium	high	yes
E_5	black	no	medium	high	no
E_6	red	yes	small	high	yes
E_7	red	yes	small	low	no
E_8	black	no	medium	average	no
E_9	black	yes	small	average	no
E_{10}	yellow	yes	big	average	no

Show a bayesian network as a **Naive Bayes** model for this problem. Estimate the probabiliy tables, using the above training set and Laplace smoothing. Using that model, predict if an insect yellow, small, with no wings and high speed, is a lepisto.