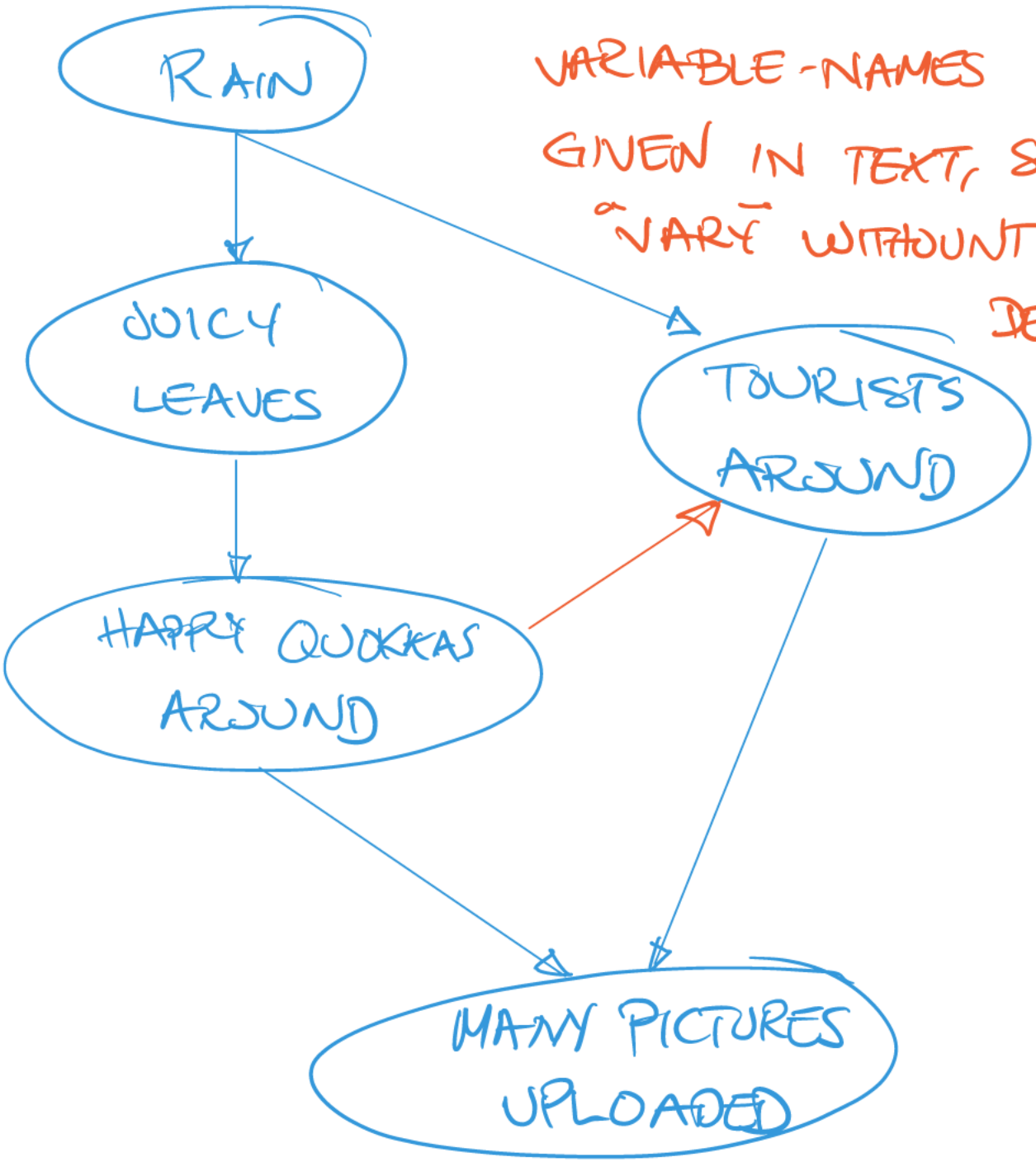
1 Please model the following as a Bayesian network. Your model is to be drawn on paper.

We are interested in the relationship between raining in an Australian national park and the number of photos of happy quokkas uploaded to some social media service. The quokka, see example photo below, likes to eat juicy leaves, so whenever it can find those it is happy. The leaves get juicy whenever it rains. Tourists coming to the national park think the happy quokkas are cute, so whenever a tourist sees a quokka that looks happy, there is a good chance they will take a picture and upload it to their social media account. However, tourists don't like rain, so whenever it rains the number of tourists will decrease. This will then influence the number of uploaded quokka images negatively.

Make a Bayesian network describing this domain. The end-goal for your model is to be able to estimate the status of rain in the national park only by monitoring the number of quokka pictures at a given social media site, but you are nevertheless asked to **make the Bayesian network so that it is as easy to understand as possible**, and not structure the model particularly for the rain-prediction task. If you make additional assumptions about the domain then please write them down explicitly together with your model.

You should only make the qualitative structure (the directed acyclic graph), and you are **not asked to provide conditional probability tables**. However, make your model so that it would be as easy as possible for a domain expert to provide the required probabilities.

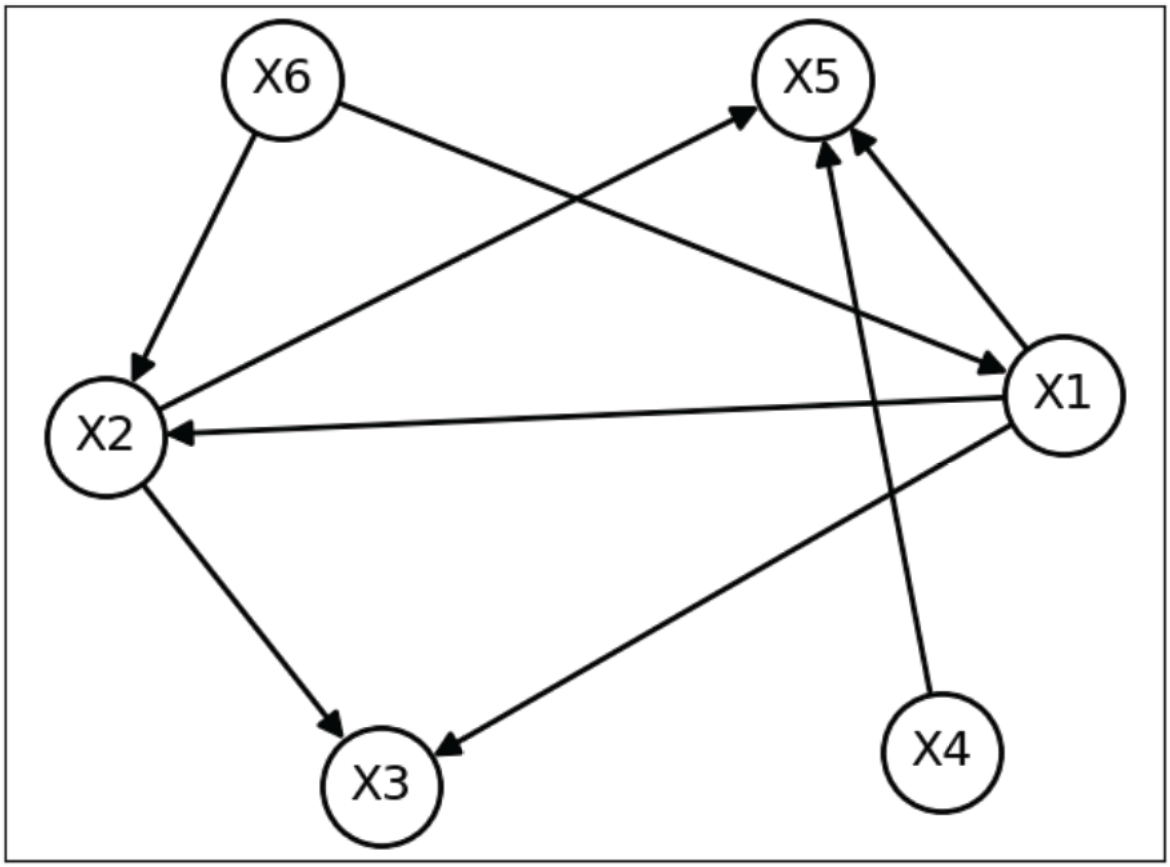
Maximum marks: 12



VARIABLE-NAMES NOT GIVEN IN TEXT, SO MAY "VARY" WITHOUT POINT-DEDUCTION.

WHILE THERE IS NO NEED TO INCLUDE THE LINK FROM QUOKKA TO TOURISTS IT CAN ALSO BE DEFENDED → NO POINT DEDUCTION!

2 Answer each question below. You will get 1 point for each correct answer, -1 point for each wrong answer. No penalty for leaving a question blank. If the total number of points is negative you will get 0 points as your total.



X_3 is independent of X_5

Select one alternative:

- False ✔
- True

X_2 is conditionally independent of X_4 given X_3

Select an alternative

- False
- True ✔

X_1 is conditionally independent of X_4 given $\{X_3, X_6\}$

Select an alternative

- False
- True ✔

X_2 is conditionally independent of X_4 given $\{X_1, X_3, X_5\}$

Select an alternative

- False ✔
- True

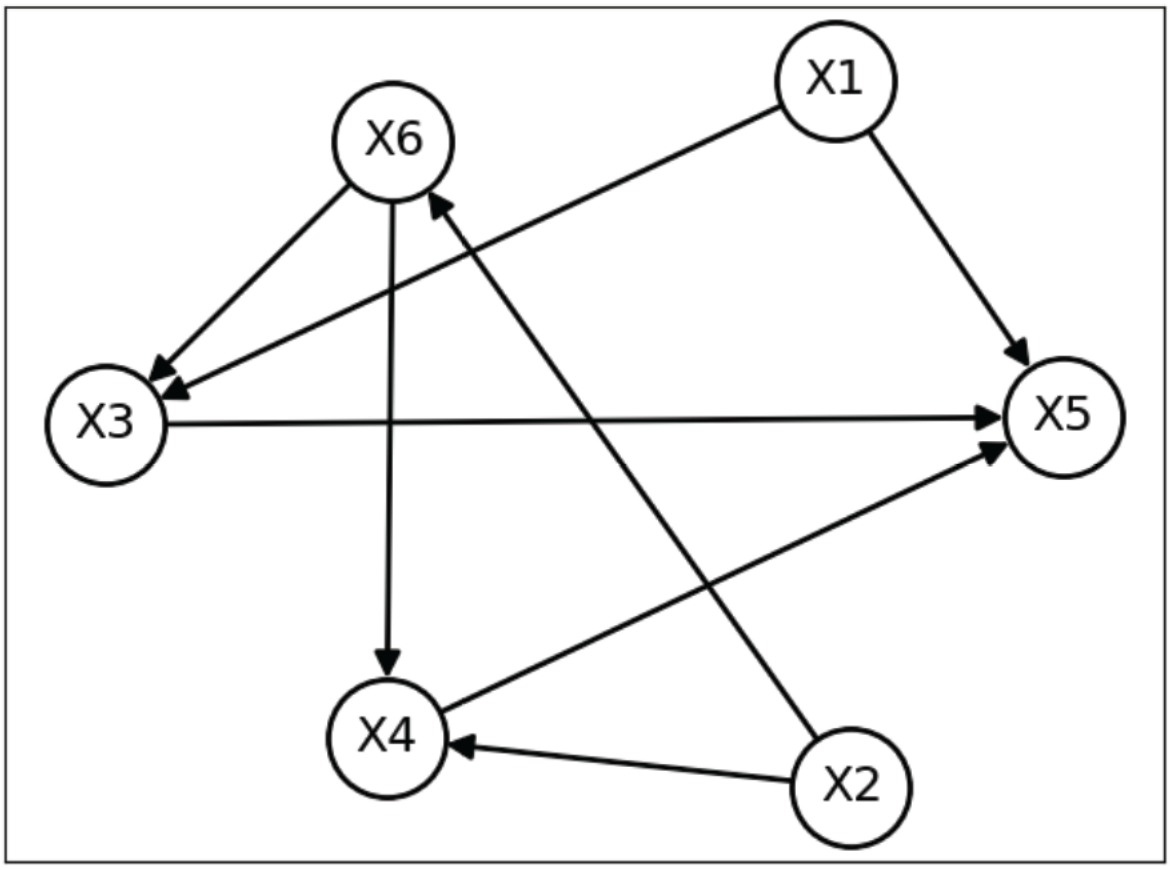
X_3 is conditionally independent of X_6 given $\{X_1, X_2, X_4, X_5\}$

Select an alternative

- True ✔
- False

Maximum marks: 5

3 Answer each question below. You will get 1 point for each correct answer, -1 point for each wrong answer. No penalty for leaving a question blank. If the total number of points is negative you will get 0 points as your total.



Is X_1 independent of X_4?
Select one alternative:

- False
- True ✔

Is X_1 independent of X_4 given X_2?
Select an alternative

- True ✔
- False

Is X_1 independent of X_4 given $\{X_2, X_3\}$?

Select an alternative

- ○ True
- ○ False ✔

Is X_1 independent of X_4 given $\{X_2, X_3, X_6\}$?
Select an alternative

- ○ False
- ○ True ✔

Is X_1 independent of X_4 given $\{X_2, X_3, X_5, X_6\}$?
Select an alternative

- ○ True
- ○ False ✔

Maximum marks: 5

4 Rational behavior can be described using mathematical terms
Select one alternative:

- ○ True ✔
- ○ False

Maximum marks: 1

5 Consider a Markov decision process that we want to solve using *value iteration*. Assume that all rewards are non-negative, and only depend on the current state.

Value iteration defines that in iteration *t+1* expressed by
$\hat{U}_{t+1}(s) \leftarrow R(s) + \gamma \max_a \sum_{s'} P(s'|s,a) \cdot \hat{U}_t(s')$

We initialize our solution with zero for all possible states, $\hat{U}_0(s)$.

It now holds that $\hat{U}_t(s)$ is non-decreasing in *t* for all possible states *s*. This is true for every choice of non-negative rewards R(s), any discount factor $\gamma \in (0,1)$ and every possible transition model $P(s'|s,a)$.

Select one alternative:

- False
- True ✔

Maximum marks: 1

6 Case Based Reasoning uses a distance measure to compare a new problem (*query*) with the examples stored in the case-base

Select one alternative:

- True ✔
- False

Maximum marks: 1

7 Reinforcement learning is a kind of deep learning

Select one alternative:

- False ✔
- True

Maximum marks: 1

8 The reason for making the assumption of stationarity when building Hidden Markov models is that stationarity induces conditional independences in the model that makes the calculations more efficient.

Select one alternative:

- False ✔
- True

Maximum marks: 1

9 When a probability *P(x)* is *subjective* then two intelligent agents can hold different opinions about how the probability should be quantified even though both agents are rational.

Select one alternative:

- True ✔
- False

Maximum marks: 1

10 The computational cost of doing smoothing in Hidden Markov Models grows linearly in the number of time-steps for which you have observed data.

Select one alternative:

- False
- True ✔

Maximum marks: 1

11 If two agents are both rational, then they must have the same underlying utility function.
Select one alternative:

- True
- False ✔

Maximum marks: 1

12 If you want to make the Markov assumption you fist have to employ Bayes rule.
Select one alternative:

- False ✔
- True

Maximum marks: 1

13 When using Hidden Markov Models we have to make both the Markov-assumption and the sensor Markov assumption.
Select one alternative:

- False
- True ✔

Maximum marks: 1

14 Let **A** be the set of all actions and **S** be the set of states for some Markov decision process. We assume that the number of states is much larger than the number of actions, $|\mathbf{S}| \gg |\mathbf{A}|$.
For this model, one iteration of *value iteration* is generally faster than one iteration of *policy iteration*.
Select one alternative:

- True ✔
- False

Maximum marks: 1

15 When one is to make a deep learning model to classify images, it is common to use convolutions among other things to ensure that the same mathematical operations are done on all parts of the input-image.
Select one alternative:

- True ✔
- False

Maximum marks: 1

16 For any Markov decision process (MDP), changing the discount factor (typically called γ) does not affect the optimal policy for the MDP.
Select one alternative:

- True
- False ✔

Maximum marks: 1

17 The reason deep learning typically works better than a single perceptron in terms of training loss is that the gradient descent algorithm scales well with respect to the number of observations in the training data, and therefore avoids overfitting.

Select one alternative:

- False ✔
- True

Maximum marks: 1

18 There always exists a utility-function that can be used to explain the behavior of a human being precisely.

Select one alternative:

- True
- False ✔

Maximum marks: 1

19 Weight-regularization as used in deep learning (when a term that increases with the length of the weight-vector is added to the loss) is a technique to avoid overfitting.

Select one alternative:

- True ✔
- False

Maximum marks: 1

20 Consider the following dataset that describes the relationship between two binary input features (denoted *x1* and *x2*) and the class *c*. Note that the data is given so that each *column* is an observation, this is done to save space on your screen.

x1	0	1	1	1	0	0	1	1	0
x2	0	1	0	0	1	1	1	0	0
c	1	1	1	1	1	0	1	1	0

As you can see, the dataset has some noise. For instance, the first observation shows the example (x1=0, x2=0, c=1) while the last example classifies the observation (x1=0, x2=0) to class *c*=0 instead.

Mark exactly those of the following machine learning techniques that are well-behaved (e.g., to be understood as converging to a well-defined optimum for the iterative learning approaches below), also when training on datasets with this kind of noise.

You get **1point per correct answer, -1 per wrong answer**, max 5 points available, minimum total score from the question is 0 points.

Select one or more alternatives:

- ☐ Percetpton learned using the delta-rule
- ☐ Decision trees learned using ID3 (the entropy-based method discussed in class) ✔
- ☐ Deep learning using gradient descent ✔
- ☐ Perceptron learned using gradient descent ✔
- ☐ Case-based reasoning with Euclidean distance meassure ✔

Maximum marks: 5

21 In the following, we shall consider natural language processing using both classical probabilistic techniques and deep learning.

Each correct answer gives **1 point, each wrong or unanswered question gives -1 point**. The minimum total number of points from these questions is nevertheless 0 points.

Question (a):
Word embeddings translate words to vector-representations.
Select one alternative:

- ○ True ✔
- ○ False

Question (b):
The main benefit from word embeddings is that we change the input representation from a single binary (existence of a word) to a high-level vector, and deep learning models prefer high-dimensional data.
Select an alternative

- ○ True
- ○ False ✔

Question (c):
Probabilistic models for natural language are not needed after the major breakthroughs of the modern large language models.

Select an alternative

- ○ True
- ○ False ✔

Question (d):
Monogram / Bag-of-Word representations are chosen over Trigram models because they are better at capturing the structure of the sentences we learn the models from.

Select an alternative

- ○ True
- ○ False ✔

Question (e):
Major challenges for language models are that natural language is ambiguous, subjective, and used inconsistently.

Select an alternative

- ○ False
- ○ True ✔

Maximum marks: 5

22 Use the grid below to define what is happening in each of the steps of the CBR cycle. The codes for the rows are as follows:

(1) Find the case or cases closest to the query using a similarity (or distance) measure.
(2) Adapt the solution(s) from the case(s) selected in **(1)** to work in the context of the query.
(3) After trying out the solution generated in **(2)**, make further changes to the provided solution if required.
(4) Store the final case from **(3)** in the case-base, if needed.

Note! You get **1 point** per correct answer, no penalty for wrong answers.

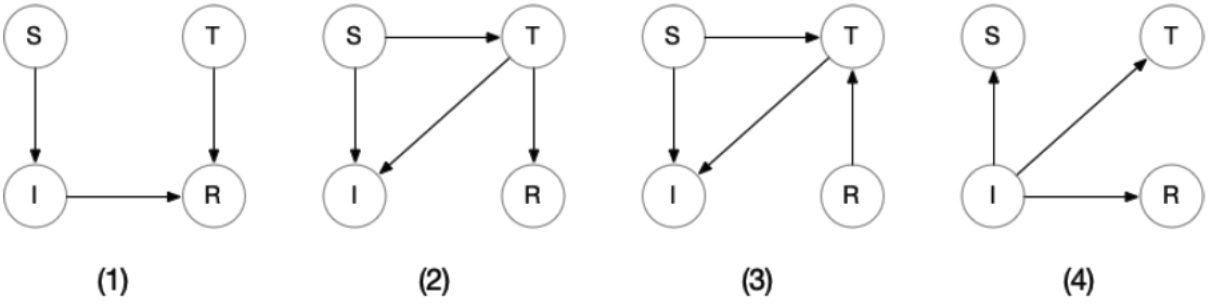
Please match the values:

	Reuse	Revise	Retain	Retrieve
(1)	○	○	○	○ ✔
(3)	○	○ ✔	○	○
(4)	○	○	○ ✔	○
(2)	○ ✔	○	○	○

Maximum marks: 4

23 Ice-cream sales (I) increases if it is sunny (S) and with increasing temperature (T). Temperature is measured by the reading (R) of a thermostat. We assume that when it is sunny, the temperature also increases.

Consider the models below:



Note!
The max points obtainable for this task is 10 points. You get that if you answer all sub-questions correctly. If you make at least one mistake or fail to answer at least one sub-question you get 0 points.

Sub-question (a):
Which model(s) - if any - incorporate exactly the causal relations mentioned in the text above?

Select one or more alternatives:

- [ ] Model (1)
- [ ] Model (2) ✔
- [ ] Model (3)
- [ ] Model (4)

For simplicity we assume all variables are measured on a 3-point scale, with states "Low", "Medium" and "High".

Sub-question (b):
Which model is the most compact (needing the fewest number of parameters in the conditional probability tables)?

Select one alternative:

- Model (1)
- Model (2)
- Model (3)
- Model (4) ✔

Sub-question (c):

Using Model (1), what is the correct and most compact expression for the probability P(R=High|S=Low), when we are only allowed to use probabilities represented in Model (1)?

Select one alternative:

Maximum marks: 10

Part (c) is unsolvable, as none of the alternatives are correct. Also, typos (extra "|"-s several places).

Everyone got "full score" for part (c), meaning (a) and (b) correct would give 10 points, independently of (c).

24 Consider the variables X_1, X_2, ..., X_t, ... and assume that we model these as a first order Markov chain (hence, when referring to "the Markov assumption" below, it is the *first order* Markov assumption we talk about).

Four of the following statements are true. Which are they? You will get **1 point** for each correct statement that you have marked, and you are **only allowed to select four options**.

Select one or more alternatives:

- [ ] The Markov assumption asserts that the future is conditionally independent of the p t given the present. ✔
- [ ] The Markov assumption ensures that $P(X_t|x_1, x_2, ..., x_{t-1}) = P(X_t|x_{t-1})$ ✔
- [ ] The Markov assumption can only be made for models where all variables are discrete.
- [ ] The Markov assumption ensures that $P(x_t) = P(x_{t+1})$
- [ ] The Markov assumption ensures that $P(X_t|x_1, x_2, ..., x_{t-1}) = P(X_t|x_1, x_{t-1})$ ✔
- [ ] The Markov assumption ensures that $P(X_t|x_1, x_2, ..., x_{t-1}) = P(X_{t+1}|x_1, x_2, ..., x_{t-1}, x_t)$
- [ ] The Markov assumption asserts that $P(X_t|x_{t-1}) = P(X_{t+1}|x_t)$
- [ ] The Markov assumption asserts that the past is conditionally independent of the fut given the present. ✔

Maximum marks: 4

25 Consider a perceptron model that takes a single input *x*. The weight of the model is *w*, and the offset *b*. As nonlinearity we use the ReLU, which for a real-valued *z* is defined as ReLU(*z*) = max(0, *z*). In total, the perceptron will thus produce the output $\mathbf{max}(0, w \cdot x + b)$ when given input *x*.

We want to use the model for a regression problem, so we predict real-valued outputs. Our dataset consists of *N* examples of the type (x_i, y_i), so one observation can for instance be (x=3.14159, y=2.71828). We use the sum of squared prediction errors as our learning loss: $\mathcal{L}(w, b) = \sum_i (y_i - \mathbf{max}(0, w \cdot x_i + b))^2$.

Before learning, our estimates for the two parameters are $\hat{w} = 1.0, \hat{b} = 0.0$

Which of the following statements (if any) are correct? **You will get 1 point for each correctly marked statement, -1 for each error.** The total from this question cannot be negative, and is thus from 0 to 4 points.

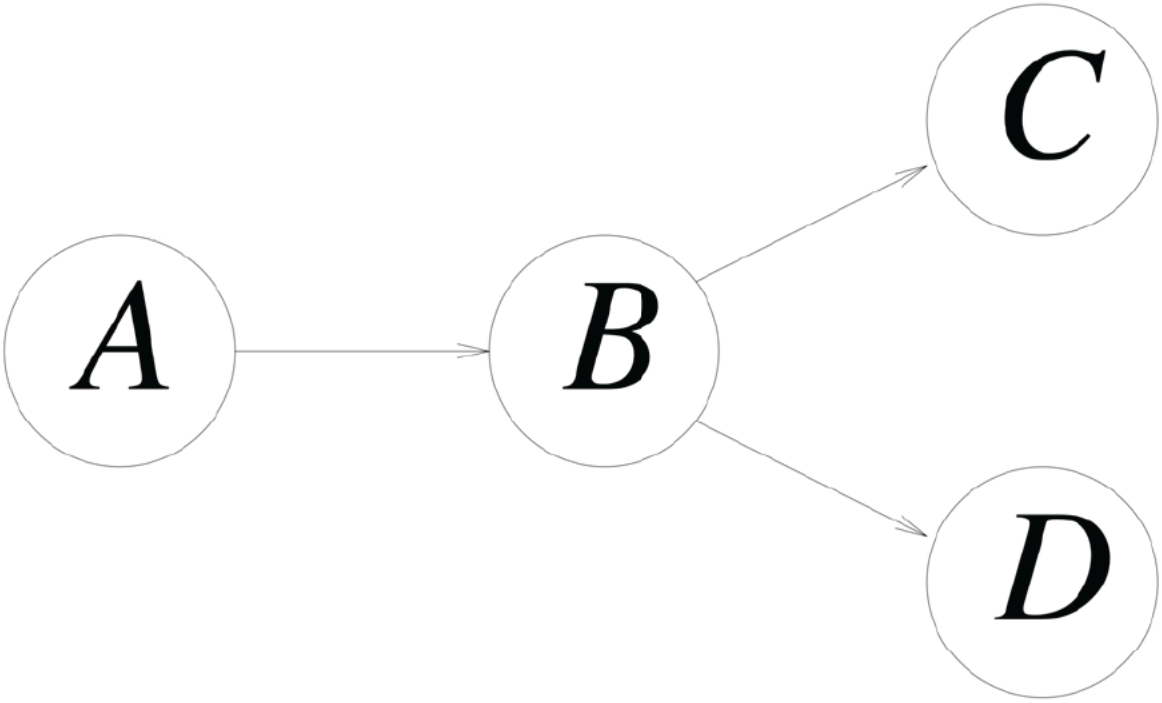
Select one or more alternatives:

- ☐ Since we are using the fairly modern ReLU function as transfer-function, this perceptron can represent any target function.
- ☐ If we make a prediction for the observation x=-3, the model with parameters given ✔ ve will produce the output-value 0.0.
- ☐ If we try to learn from a new example (x=-3, y=10) using gradient decent, the updated model is identical to the one we have defined above (the gradient method does not ✔ nge the parameters).
- ☐ If we try to learn from a new example (x=+3, y=10) using gradient decent, the learn✔ will result in increased values for both $\hat{w}$ and $\hat{b}$.

Maximum marks: 4

26 For this question, each correct answer gives 2 points. There is no penalty for wrong answers.

Consider this Bayesian network structure.



Which of the following claims is the correct one?
Select one alternative:

- A is conditionally independent of C given D
- P(A|B,C) = P(A|C)
- P(A|B, C) = P(A)
- P(A|B,C) = P(A|B) ✔
- None of the alternatives above are correct in general

Let a and b be two states for variables A and B, respectively. What is the probability P(A=a|B=b)?

Select one alternative

- $P(A=a|B=b) = P(A=a)$
- $P(A=a|B=b) = P(B=b)$
- $P(A=a|B=b) = \frac{P(A=a)}{P(B=b)}$
- $P(A=a|B=b) = \frac{P(B=b|A=a)P(A=a)}{\sum_{a'} P(B=b|A=a')\cdot P(A=a')}$ ✔
- None of the alternatives above are correct in general.

Finally, assume we have observed C = c, D = d (c and d are legal states for the two variables C and D, respectively), but have no observation for B. What is the probability P(A=a|C=c, D=d) if we are constrained to only using numbers that are directly encoded in the Bayesian network with structure given above?

Select one alternative

- $P(A=a|C=c,D=d) = \sum_b P(B=b)\cdot P(A=a|B=b,C=c,D=d)$
- $P(A=a|C=c,D=d) = \sum_b P(B=b|A=a)\cdot P(A=a|B=b,C=c,D=d)$
- $P(A=a|C=c,D=d) = P(A=a)\sum_b P(B=b|A=a)\cdot P(C=c|B=b)\cdot P(D=d|B=b)$
- None of the alternatives above are correct in general. ✔

Maximum marks: 6

27 **Note!** For this question there are 2 points per correctly answered sub-question, and no penalty for wrong answers.

Peter is going to build a deep learning model. He has input-data of dimension *d* (meaning that an observation **x** consists of *d* real numbers). The data **x** gives the value (the closing-price) of a particular stock over the last *d* days, and the goal of the system is to learn "something" from these data to help Peter understand the value of the stocks and thereby get rich.

Initially, Peter has data covering the last ten years, meaning *d* is around 3000, and he has decided to focus on the *N=2* stocks *IBM* and *Apple*.

Question (a):

Among the high-level deep learning architectures given below, which is **least likely** to give good results?

Select one alternative:

- Feed forward neural networks ✔
- Convolutional neural networks
- Recurrent neural network

Question (b):

When building the model, Peter considers to either regard the problem as a *regression*-task or a *classification*-task. If he chooses regression, he has decided that the task will be to train the model to predict the value of the next observation (stock-price the next day). If he chooses classification, he wants to have these three classes:

1. Next observation is all time high for this stock over the ten years of data
2. Next observation is all time low for this stock over the ten years of data
3. Next observation is neither all time high nor all time low

Which model is **least likely** to provide information that Peter can use to get rich?

Select one alternative

- The classification model ✔
- The regression model

Question (c):

Peter observes that his model is **over-fitting**.

Which of the following strategies is **least likely** to help with overfitting issues?

Select one alternative

- Add weight regularization
- Reduce the learning-rate ✔
- Add drop-out layer(s)

Question (d):

Peter gives up on the stock market, and rather focuses on a pet-project: He wants to find out if images of animals contain at least one cat. To do this he collects a large number of images of different animals, uses a convolutional neural network (CNN), and builds a binary classifier that is trained to classify images as belonging to either of the two classes

1. Cat image
2. Non-cat image

Unfortunately the model is not working well, as he experiences a very high **training error**.

Which of the following strategies is **least likely** to help with poor performance on the training data?

Select one alternative

- Remove some of the intermediate layers from the model ✔
- Add more intermediate layers to the model
- Manually check the training data for problems like poor image quality and mislabelling
- Reduce the learning rate

Maximum marks: 8

28 You are playing the game Hearthstone. (Note that if you don't know the game, it doesn't really matter, as all the details you need are given in the text.) You are up against the famous player Alice.

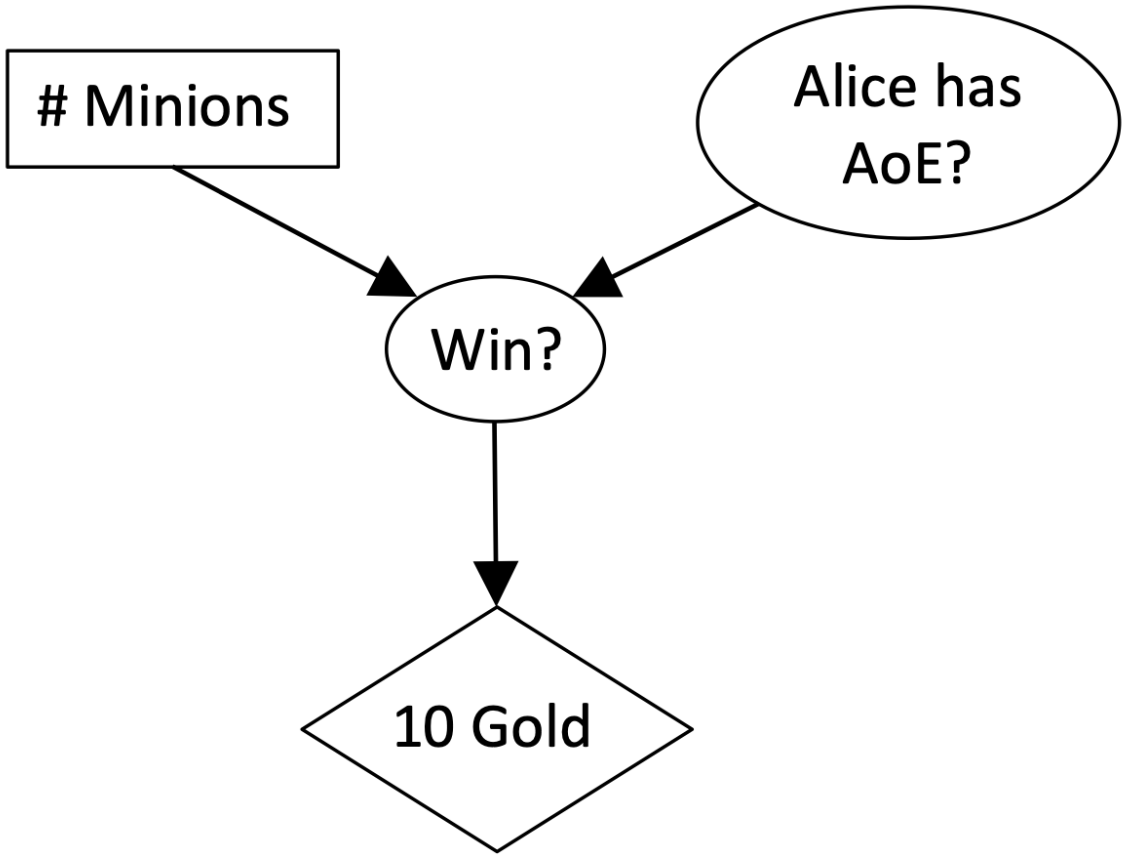
On your turn, you can choose between playing 0, 1, or 2 of your minions. You realize Alice might be holding up an Area of Effect (AoE) card, which is more devastating the more minions you play.

- If Alice has the AoE, then your chances of winning are:
  - 60% if you play 0 minions
  - 50% if you play 1 minion
  - 20% if you play 2 minions

- If Alice does **not** have the AoE, then your chances of winning are:
  - 20% if you play 0 minions
  - 60% if you play 1 minion
  - 90% if you play 2 minions

You know that there is a 50% chance that Alice has an AoE.
Winning this game is worth 10 gold and losing is worth 0.

So, you have to make a decision: How many minions should you play? Having learned about decision graphs, you immediately sketch this following model:



Use this graphical structure and the quantitative information given above to answer the following. For each question, give your answer with one digit after the decimal point, e.g., -2.3 or 4.1. **Each correct answer gives 1 point, each wrong answer 0 points.**

Question (a):

How much gold would you expect to win choosing 0 minions? (4) .

Question (b):

How much gold would you expect to win choosing 1 minion? (5.5)

Question (c):

How much gold would you expect to win by choosing 2 minions? (5.5)

Question (d):

How much gold would you expect to win if you **know** the AoE is in Alice's hand and you play optimally given that information? (6)

Question (e):

How much gold would you expect to win if you *know* the AoE is **not** in Alice's hand and you play optimally given that information? (9)

Question (f):

Assume that your utility of gold is the same as the amount of gold, so for instance is the utility of getting 0 gold 0, the utility of getting 10 gold is +10.

How much gold would you be willing to pay to get to know whether or not the AoE is in Alice's hand before making your decision about the number of minions to play? (2.0)

Maximum marks: 6

29 Discuss the following statement. Your presentation will be regarded as better if you use terminology and arguments from the curriculum wherever appropriate. You can choose to answer in English or Norwegian. Your text is limited to 500 words.

"Large language models (LLMs) are showing sparks of intelligence. LLMs have the possibility to develop self-awareness and free will. With this in mind, work on AI systems world-wide, in particular on LLMs, should be paused for a period of 6 months. During this period, focus must be put on international legislation that must aim at preventing the development of systems with AGI (artificial general intelligence)."

Fill in your answer here

Format | B I U x_2 x^2 | I_x | | | | | | | Ω | | | Σ |

Words: 0/500

Maximum marks: 10