Homework 2 of CS 165A (Spring 2023)

University of California, Santa Barbara

To be discussed on Apr 26 and May 3, 2023 (Wednesdays)

Notes:

- The homework is optional. You do not need to submit your solutions anywhere and you will not be evaluated by these.
- To maximize your learning, you should try understanding the problems and try solving them as much as you can before the discussion class.
- Feel free to discuss with your peers / form small groups to solve these problems.
- Feel free to discuss any questions with the instructor and the TA in office hours or on Piazza.

---

1 Why should I do this homework?

This homework is given for you to practice what you learned in BayesNet (Problem 1 - 3) and in Search (Problem 4-5). In Problem 1, you will practice modeling with BayesNet.. In Problem 2, you will practice reading conditional independences from the graph. Problem 3 teaches you something about the notorious Hidden Markov Models (HMM). While Problem 3 is a challenge question, part (a) - (c) are short and highly doable.

Problem 4 asks you to write down what we have brainstormed in the lecture on the Missionary and Cannibal example. Problem 5 is a good chance to understand and practice different search algorithms by hands (something that you should perhaps expect one question in the midterm).

2 Homework problems

Problem 1 A patient has a probability to recover from a disease that depends on whether s/he receives the drug, how old s/he is and which gender the patient has. A doctor gives a patient a drug dependent on their age and gender. Additionally it is known that age and gender are independent.

(a) Draw the Bayesian network which describes this situation.

Let R represent the patient recovering, D represent the drug being administered, A represent the patient's age, and G represent the patient's gender. Lastly $X \to Y$ means that there's an arrow from X to Y in the Bayesian network.

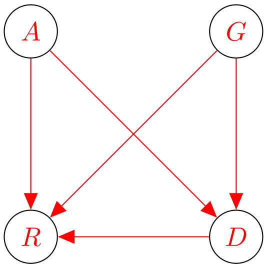
Doctor gives a patient a drug dependent on their age and gender

There should be edges from $A \to D$, and $G \to D$

A patient has a probability to recover from a disease that depends on whether s/he receives the drug, how old s/he is and which gender the patient has

There should be edges from $A \to R$, $G \to R$, $D \to R$

All together, this gives:



Other solutions are valid too. For example, if the doctor does not decide on the drug based on either age or gender, but something else independent to these features, then you have a different Bayesian network.

(b) Factorize the joint probability distributions into CPTs.

By traversing the graph from the parents to the children, we can write

$$P(R, D, A, G) = P(A)P(G)P(D|A, G)P(R|D, A, G)$$

One can also do it algebraically and apply the conditional independences

$$\begin{aligned} P(R, D, A, G) &= P(R|D, A, G)P(D, A, G) \\ &= P(R|D, A, G)P(D|A, G)P(A, G) \\ &= P(R|D, A, G)P(D|A, G)P(A)P(G) \end{aligned}$$

The last step applies $A \perp G$.

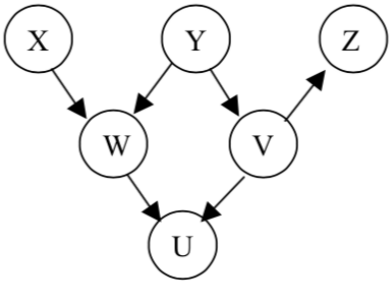
(c) Write down the formula to compute the probability that a patient recovers given that you know if s/he gets the drug. Write down the formula using only probabilities which are part of the factorized probability distribution (i.e., the CPTs).

We apply marginalization to get the CPTs

$$\begin{aligned} P(R|D) &= \frac{P(R,D)}{P(D)} \\ &= \frac{\sum_A \sum_G P(A)P(G)P(D|A,G)P(R|D,A,G)}{\sum_R \sum_A \sum_G P(A)P(G)P(D|A,G)P(R|D,A,G)} \\ &= \frac{\sum_A \sum_G P(A)P(G)P(D|A,G)P(R|D,A,G)}{\sum_A \sum_G P(A)P(G)P(D|A,G)} \end{aligned}$$

The second step simplifies the expression a bit by eliminating variable R. The solution is correct with or without the last step.

Problem 2 Consider the Bayes Net below:



(a) Is it true that $P(X|Y,W) = P(X|W)$? Explain.

This question asks about whether X and Y are conditional independent given W

So the answer is **No**, X Y might not be conditionally independent given W by the common descendant rule of d-separation.

(b) Write down the expression for computing $P(X|Y)$ using the above Bayes Net.

Since X and Y are marginally independent. $P(X|Y) = P(X)$

(c) Are variables X,W conditionally independent of variables V,Z, given Y? Explain.

Path $W \rightarrow U \leftarrow V$ to node V, this path is blocked by U. This is because U is not observed, according to the Collider rule, the ball can not pass through U.

Path $W \leftarrow Y \rightarrow V$ to node V, this path is blocked by Y. This is because Y is not observed, according to the Fork rule, the ball can not pass through Y.

Since every possible path is blocked, we can say X,W are conditionally independent of variables V, Z, given Y

(d) Are variables X,W conditionally independent of variables V,Z, given U? Explain.

Not necessarily. There is an active path $W \rightarrow U \leftarrow V$, where U is observed.

As long as there is an active path, we can not say X,W are conditionally independent of variables V, Z, given U

(e) Are variables W and Z independent? Explain.

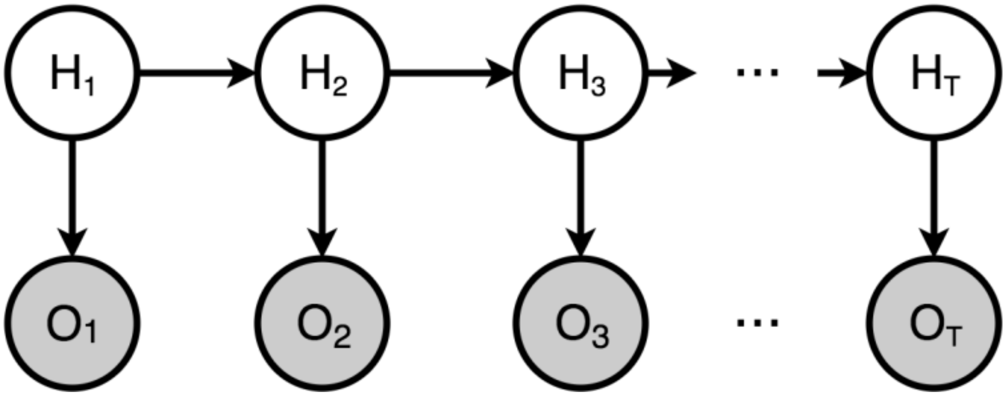
Not necessarily. Because there is an active path $W \leftarrow Y \rightarrow V \rightarrow Z$.

(f) Write down the Markov Blanket of variable W and variable Y.

The Markov blanket of W is $\{X, Y, U, V\}$. The Markov blanket of Y is $\{W, X, V\}$

(g) Assume all the variables are binary, either take value 0 or 1. Write down the expression to compute $P(U = 1, V = 1, W = 1, X = 0, Y = 0, Z = 1)$ using notation like $P(X = 1|W = 0)$.

$P(U = 1|W = 1, V = 1)P(Z = 1|V = 1)P(W = 1|X = 0, Y = 0)P(V = 1|Y = 0)P(X = 0)P(Y = 0)$



Problem 3 Hidden Markov Models (Challenge Problem) Let all variables be discrete. In particular, let O_i be a discrete random variable that could take d possible values, and H_i be a discrete random variable that could take k possible values.

The parameters of the HMM model are simply the CPTs of the graphical model, i.e.,

$$P(H_1) = \theta \in \mathbb{R}^k,$$
$$P(H_{i+1}|H_i) = A \in \mathbb{R}^{k\times k} \text{ for all } i = 1, 2, 3, ..., T-1,$$
$$P(O_i|H_i) = B \in \mathbb{R}^{d\times k} \text{ for all } i = 1, 2, 3, ..., T.$$

Canonically, parameter θ, A, B are called the "initial state distribution", "transition probabilities" and "emission probabilities" in standard HMM jargon.

Convince yourself the dimensionality of these CPTs are correct.

Note that the transition and emission probabilities are *the same* for all $i = 1, ..., T$.

(a) Write down the joint probability of $P(H_1, ..., H_T, O_1, ..., O_T)$ in factorized form as function of the CPTs θ, A, B.

$$P(H_1, ..., H_T, O_1, ..., O_T; \theta, A, B) = P(H_1)\prod_{t=1}^{T} P(O_t|H_t)\prod_{t=2}^{T} P(H_t|H_{t-1})$$
$$= \theta[H_1]\prod_{t=1}^{T} B[O_t, H_t]\prod_{t=2}^{T} A[H_t|H_{t-1})$$

(b) Write down the probability distribution of the observed variables $P(O_1, ..., O_T)$ as a function of the CPTs θ, A, B.

(hint: This is identical to expressing $P(O_1, ..., O_T)$ using CPTs, but the parameters are shared. The final expression (if you use a matrix form, will be quite clean))

Remark: The above probability distribution $P(O_1, ..., O_T)$ is jointly parametrized by values of $O_1, ..., O_T$, and the values of θ, A, B. When we view it as a function of θ, A, B, while keeping $O_1, ..., O_T$ fixed, Then this function is known as the likelihood function: $L(O_1, ..., O_T; \theta, A, B)$. This measures the likelihood of observing $O_1, ..., O_T$ when the data generating distribution is specified by θ, A, B.

Given a sequence of observation $[O_1, ..., O_T] = [o_1, ..., o_T]$, the parameters A, B, θ that maximizes the likelihood, i.e.

$$[\hat{\theta}, \hat{A}, \hat{B}] = \operatorname*{argmax}_{A,B,\theta} L(O_1 = o_1, ..., O_T = o_t; \theta, A, B)$$

is called the maximum likelihood estimator.

Solving the optimization for this MLE is not easy. It is not a convex optimization problem and we will have to use the EM algorithm to find a local optimal solution. The E-step alone requires using dynamic programming — a Forward-Backward algorithm (closely related to the more famous Viterbi algorithm). The EM solution itself is known as the Baum-Welch algorithm. Rest assured. You are *not* going to derive that in this homework.

We will take an alternative route using only things that we have learned from the class.

$$P(O_1, ..., O_T; \theta, A, B) = \sum_{H_1,...,H_T} P(H_1) \prod_{t=1}^{T} P(O_t|H_t) \prod_{t=2}^{T} P(H_t|H_{t-1})$$

$$= \sum_{H_1,...,H_T} \theta[H_1] \prod_{t=1}^{T} B[O_t, H_t] \prod_{t=2}^{T} A[H_t, H_{t-1}]$$

$$= \sum_{H_T} B[O_T, H_T] \sum_{H_{T-1}} B[O_T, H_T] A[H_T|H_{T-1}] \ldots \sum_{H_2} B[O_2, H_2] A[H_2, H_1]$$

$$\sum_{H_1} B[O_1, H_1] A[H_2, H_1] \theta[H_1].$$

(c) Show (using the rules of d-separation or otherwise) that for $2 \leq i \leq T-1$, O_{i-1}, O_i, O_{i+1} are conditionally independent given H_i.

By d-separation (Chain), $(O_{i-1}, H_{i-1}) \perp O_i | H_i$, and $(O_{i-1}, H_{i-1}) \perp (H_{i+1}, O_{i+1}) | H_i$.

(d) Use the conditional independence in (c) to show that:

$$P(O_1, O_2, O_3) = \sum_{i=1}^{k} P(H_2 = i) P(O_1|H_2 = i) P(O_2|H_2 = i) P(O_3|H_2 = i). \qquad (1)$$

Given the conditional independence that O_1, O_2, and O_3 are conditionally independent given H_2, we can express the joint probability $P(O_1, O_2, O_3|H_2)$ as the product of the conditional probabilities:

$$P(O_1, O_2, O_3|H_2) = P(O_1|H_2)P(O_2|H_2)P(O_3|H_2)$$

Now, to find $P(O_1, O_2, O_3)$, we marginalize out the hidden state H_2:

$$\begin{aligned} P(O_1, O_2, O_3) &= \sum_{i=1}^{k} P(H_2 = I)P(O_1, O_2, O_3|H_2) \\ &= \sum_{i=1}^{k} P(H_2 = i)P(O_1|H_2 = i)P(O_2|H_2 = i)P(O_3|H_2 = i) \end{aligned}$$

(e) Let O_1, O_2, O_3 be discrete random variables with d possible values and H_2 be a discrete random variable with k possible values.

- What is the total number of independent numbers to describe $P(H_2)$, $P(O_2|H_2)$, $P(O_1|H_2), P(O_3|H_2)$ in terms of k and d?

  To describe $P(H_2)$, we need $(k - 1)$ independent numbers since the probabilities of all hidden states must sum to 1.

  To describe $P(O_2|H_2)$, we need $k * (d - 1)$ independent numbers, as for each of the k hidden states, we have $(d - 1)$ independent probabilities for the observed states (the last one is determined by the fact that probabilities must sum to 1).

  Similarly, for $P(O_1|H_2)$ and $P(O_3|H_2)$, we also need $k * (d - 1)$ independent numbers for each.

  So the total number of independent numbers needed to describe $P(H_2)$, $P(O_2|H_2)$, $P(O_1|H_2)$, and $P(O_3|H_2)$ is: $k - 1 + 3k(d - 1)$

- Let us enumerate all combinations of O_1, O_2, O_3 in (1), how many equations do we get in total?

  Solution: kd^3

- Note that the LHS of (1) can be estimated from the data directly and the RHS are all unknown parameters. By solving the system of (nonlinear) equations, we can potentially identify the unknowns: $P(H_2), P(O_2|H_2), P(O_1|H_2), P(O_3|H_2)$. What is a condition on k, d such that we have enough equations to identify all unknowns variables? (Assume that we need one equation for one unknown.)

(Hint: the number of unknown variables are the same as the number of independent parameters)

Solution:
For the system of equations to have enough equations to identify all unknown variables, we need the total number of equations to be greater than or equal to the total number of independent parameters.
$kd^3 \geq k - 1 + 3k(d-1)$

(f) If we can solve the nonlinear equations about, we can then identify

$$P(H_2), P(O_2|H_2), P(O_1|H_2), P(O_3|H_2).$$

But these are not the CPTs. If the CPTs are ultimately what we want to learn, then we need an set of equations to convert these quantities back to CPTs.

Write $P(O_2|H_2)$, $P(O_3|H_2)$ and $P(O_1|H_2)$ in terms of the model parameters (the CPTs): θ, A, B.

The first expression is trivially:

$$P(O_2|H_2) = B[O_2, H_2]$$

By the sum rule:

$$P(O_3|H_2) = \sum_{H_3} B[O_3, H_3]A[H_3, H_2]$$

Finally, by the Bayes rule:

$$\begin{aligned} P(O_1|H_2) &= \sum_{H_1} P(H_1|H_2)P(O_1|H_1) = \sum_{H_1} \frac{P(H_2|H_1)P(H_1)}{P(H_2)} P(O_1|H_1) \\ &= \sum_{H_1} \frac{A[H_2, H_1]\theta(H_1)}{P(H_2)} B[O_1, H_1] \end{aligned}$$

Note that $P(H_2) = A[H_2, :]\theta$. Everything above has a matrix representation too.

In matrix form:

$$\begin{aligned} P(H_2) &= A\theta \\ P(O_2|H_2) &= B \\ P(O_3|H_2) &= BA, \\ P(O_1, H_2) &= (B\text{diag}(\theta)A^T) \end{aligned}$$

and $P(O_1|H_2)$ is constructed by element-wise dividing every column of $(B\text{diag}(\theta)A^T)$ by $A\theta$.

These equations tell us how we can make use of what we can solve for using the nonlinear system of equations — $P(O_2|H_2)$, $P(O_3|H_2)$ and $P(O_1|H_2)$ — to recover the parameters of the HMM model.

Problem 4 The missionaries and cannibals problem is usually stated as follows. Three missionaries and three cannibals are on one side of a river, along with a boat that can hold one or two people. Find a way to get everyone to the other side without ever leaving a group of missionaries in one place outnumbered by the cannibals in that place.

(a) Define a state representation.

There are many possibilities, one way is to use numbers (as we did in the lecture), that use $3, 3, 1$ to denote 3 missionaries and 3 cannibals on the left-hand side and the boat is on the left-hand side.

Another example is to represent the missionaries by M and the cannibals by C. Let the boat be B.

Each state can be represented by the items on each side, e.g., Side1 $\{M, M, C, C\}$, Side2 $\{M, C, B\}$

(b) Give the initial and goal states in this representation.

Initial state: Side1 $\{M, M, M, C, C, C, B\}$, Side2 $\{\}$

Goal state: Side1 $\{\}$, Side2 $\{M, M, M, C, C, C, B\}$

(c) Define the successor function (output available states that are safe) in this representation.

A set of missionaries and/or cannibals (call them Move) can be moved from $Side_a$ to $Side_b$ if:

- The boat is on $Side_a$.
- The set Move consists of 1 or 2 people that are on $Side_a$.
- The number of missionaries in the set formed by subtracting Move from $Side_a$ is 0 or it is greater than or equal to the number of cannibals.
- The number of missionaries in the set formed by adding Move to $Side_b$ is 0 or it is greater than or equal to the number of cannibals.

(d) What is the cost function in your successor function?

Each move has unit cost.

(e) What is the total number of safe states? Give an example of a state that is safe but unreachable?

The total number of safe states is 16:

Side1$\{M, M, M, C, C, C, B\}$, Side2$\{\}$

Side1$\{\}$, Side2$\{M, M, M, C, C, C, B\}$

Side1$\{M, M, M, C, C, B\}$, Side2$\{C\}$

Side1$\{M, M, M, C, C\}$, Side2$\{C, B\}$

Side1$\{M, M, M, C, B\}$, Side2$\{C, C\}$

Side1$\{M, M, M, C\}$, Side2$\{C, C, B\}$

Side1$\{M, M, C, C, B\}$, Side2$\{M, C\}$

Side1$\{M, M, C, C\}$, Side2$\{M, C, B\}$

Side1$\{M, C, B\}$, Side2$\{M, M, C, C\}$

Side1$\{M, C\}$, Side2$\{M, M, C, C, B\}$

Side1$\{C, C, C, B\}$, Side2$\{M, M, M\}$

Side1$\{C\}$, Side2$\{M, M, M, C, C, B\}$

Side1$\{C, C, B\}$, Side2$\{M, M, M, C\}$

Side1$\{C, C, B\}$, Side2$\{M, M, M, C\}$

Side1$\{M, M, M\}$, Side2$\{C, C, C, B\}$

Side1$\{C, B\}$, Side2$\{M, M, M, C, C\}$

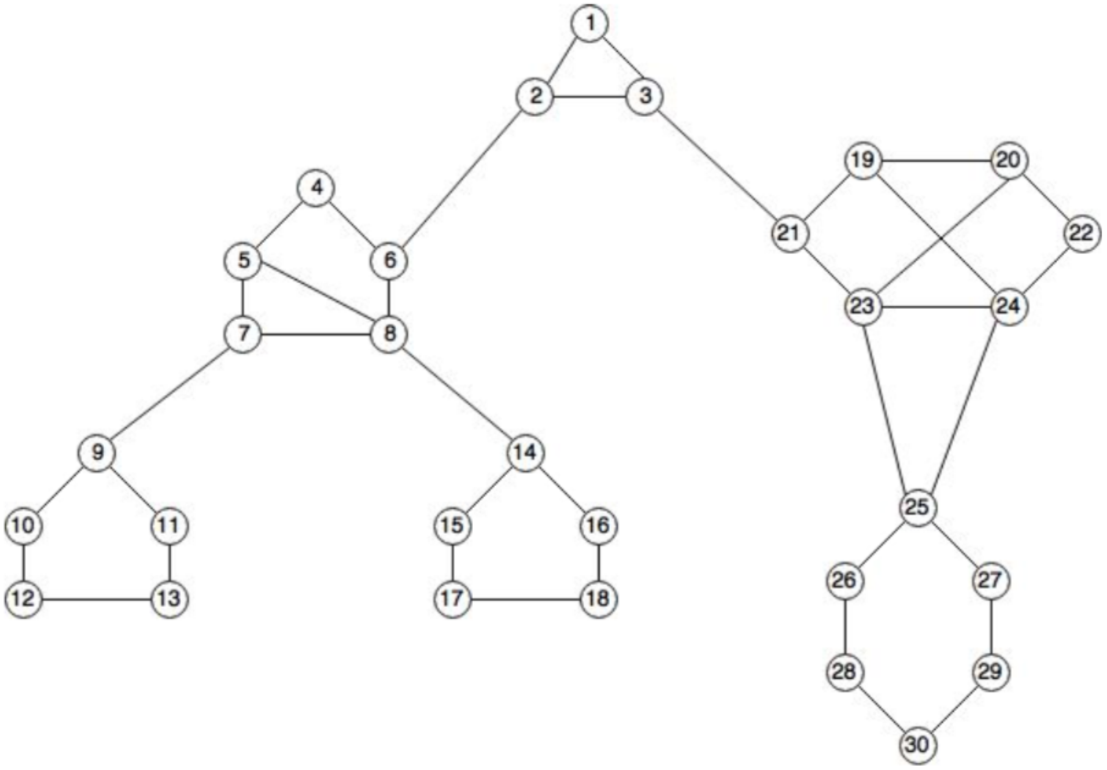
The last one is only reachable through the goal state

These two are not reachable because the preceding state must have had more cannibals than missionaries on one side of the river:

Side1$\{C, C, C\}$, Side2$\{M, M, M, B\}$

Side1$\{M, M, M, B\}$, Side2$\{C, C, C\}$

Problem 5



Consider the state space diagram shown above. Assume state 12 is the start state and state 30 is the goal state.

1. Assuming a uniform cost of 1 on each edge, simulate the execution of BFS, DFS, IDS (assuming that the depth increases by 1 beginning from 3 to 5) and show the order of states visited. Assume that lower number children are visited first.

   BFS:

   12-10-13-9-11-7-5-8-4-6-14-2-15-16-1-3-17-18-21-19-23-20-24-25-22-26-27-28-29-30

   DFS:

   12-10-9-7-5-4-6-2-1-3-21-19-20-22-24-23-25-26-28-30

   IDS:

   IDS calls DFS for different depths starting from an initial value. In every call, DFS is restricted from going beyond given depth. So basically we do DFS in a BFS fashion.

depth=3: 12-10-9-7-11-13

depth=4: 12-10-9-7-5-8-11-13

depth=5: 12-10-9-7-5-4-8-6-14-11-13

2. Now, simulate the execution of bidirectional search (assuming uniform cost of 1 on each edge and BFS as the basic search from each end). At which state do the two searches meet? (3')

   start:12-10-13-9-11-7-5-8-4-6-14-2-15-16-1-3

   end :30-28-29-26-27-25-23-24-19-20-21-22-3-1-2

   They will intersect when both expand to state 1.

3. Now, we consider non-uniform weights on edges. Assume that edges between even-even and odd-odd numbered edges have a cost of 1 and those between even-odd numbered edges have a cost of 2. Repeat the goal search using uniform-cost search.

   uniform-cost search:

   12-10-13-9-11-7-5-8-4-6-14-2-16-15-1-3-18-21-17-19-23-20-24-25-26-27-22-29-28-30

4. Now, we add a heuristic h to the search. Denote states 1-3 as cluster A, 4-8 as cluster B, 9-13 as cluster C, 14-18 as cluster D, 19-24 as cluster E, and 25-30 as cluster F. Heuristic h estimates costs to the goal state 30 as follows:

   (a) h(30) = 0

   (b) h(all nodes in cluster F except 30) = 1

   (c) h(all nodes in cluster E)=2

   (d) h(all nodes in cluster A)=3

   (e) h(all nodes in cluster B) = 4

   (f) h(all nodes in cluster C)=5

   (g) h(all nodes in cluster D) = 5

   (a) Is this heuristic admissible? Prove or disprove.

   To prove if a heuristic is admissible, we must show that it never overestimates the cost to reach the goal from the current state. In other words, for each node n, h(n) must be less than or equal to the true cost to reach the goal. But, to disprove this statement, you only need to find a counter-example

   In our case, the $h(n)$ is always less or equal to the true cost from all the states, so this heuristic is admissible.

(b) Is it consistent? Prove or disprove.

A heuristic is consistent if it satisfies the triangle inequality, meaning $h(n) \leq c(n \leftarrow n') + h(n')$, where $c(n, n')$ is the cost of transitioning from node n to node n'.

For any two adjacent states, their h(n) difference is either 0 or 1. The path cost between only two choices, 1 or 2. So $h(n) \leq h(n') + c((n \leftarrow n')$.

(c) If the heuristic is consistent, repeat the search for the goal state using A* (GRAPH-SEARCH).

Order of expansion: 12-10-13-9-7-11-5-9-8-4-6-2-1-3-21-19-23-25-27-26-29-30