CS221 Problem Session Solutions

Final Review

1) **Problem 1: Inferencia**

You are the president of the small nation of Inferencia, and you have been charged with choosing which of your country's two rival soccer teams - the Bayesians or the Markovians - should represent Inferencia at the upcoming Olympics. You'd like to send whichever team is more popular, so you decide to model the monthly evolution of the two teams' fanbases during the months leading up to the Olympics using a dynamic Bayesian network.

Let B_t denote the number of fans that the Bayesians have in month t, and let M_t denote the number of fans that the Markovians have in month t. You have no way of observing these quantities directly, but you can observe two other quantities which they influence: let J_t denote the number of jerseys sold by the Bayesians in month t, and let A_t denote the attendance of the monthly exhibition game between the Bayesians and the Markovians in month t.

The fanbases of the two teams evolve according to the following model, where each month a fan is either gained or lost with equal probability:

$$\Pr(M_{t+1}|M_t) = \begin{cases} \frac{1}{2} & \text{if } M_{t+1} = M_t - 1 \\ \frac{1}{2} & \text{if } M_{t+1} = M_t + 1 \\ 0 & \text{otherwise} \end{cases} \qquad \Pr(B_{t+1}|B_t) = \begin{cases} \frac{1}{2} & \text{if } B_{t+1} = B_t - 1 \\ \frac{1}{2} & \text{if } B_{t+1} = B_t + 1 \\ 0 & \text{otherwise} \end{cases}$$

The Bayesian fans are big spenders - almost every fan buys a jersey each month! We model the fanbase size's influence on jersey sales by:

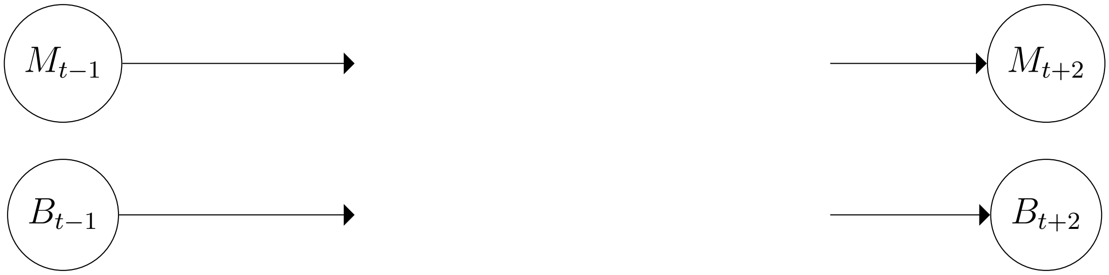
$$\Pr(J_t|B_t) = \begin{cases} 0.3 & \text{if } J_t = B_t \\ 0.25 & \text{if } J_t = B_t - 1 \\ 0.2 & \text{if } J_t = B_t - 2 \\ 0.15 & \text{if } J_t = B_t - 3 \\ 0.1 & \text{if } J_t = B_t - 4 \\ 0 & \text{otherwise} \end{cases}$$

Lastly, because most fans attend each monthly exhibition (although sometimes more, and sometimes fewer), we model the influence of the fanbase sizes on the exhibition attendance by:

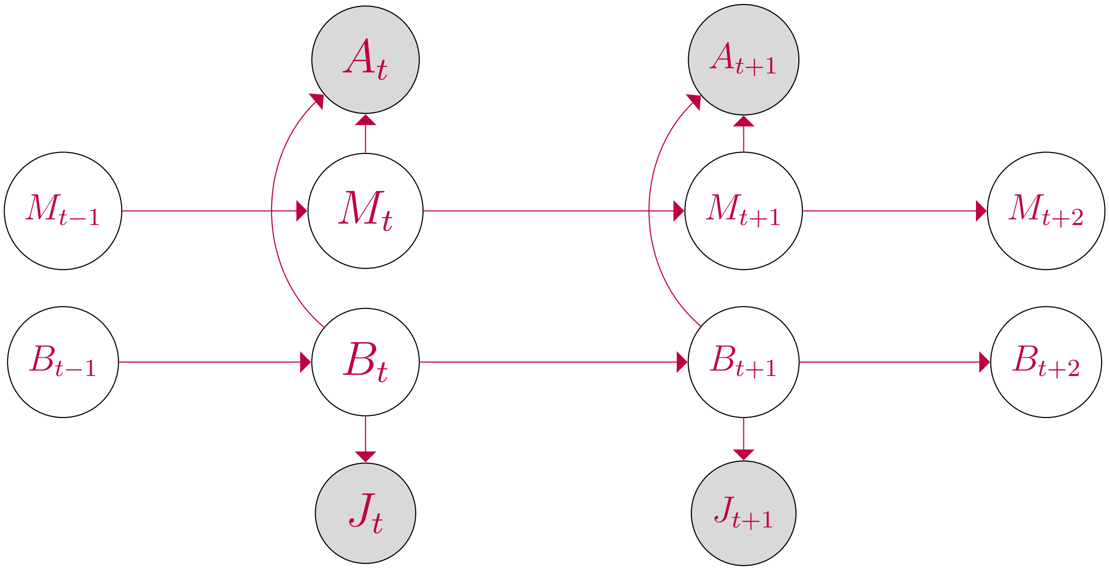
$$\Pr(A_t|B_t, M_t) = \begin{cases} 0.14 & \text{if } A_t = B_t + M_t \\ 0.13 & \text{if } |A_t - (B_t + M_t)| = 1 \\ 0.11 & \text{if } |A_t - (B_t + M_t)| = 2 \\ 0.09 & \text{if } |A_t - (B_t + M_t)| = 3 \\ 0.06 & \text{if } |A_t - (B_t + M_t)| = 4 \\ 0.04 & \text{if } |A_t - (B_t + M_t)| = 5 \\ 0 & \text{otherwise} \end{cases}$$

Note that the assumptions and inferences made in individual parts (i.e. **(a)**, **(b)**, etc.) of this problem do *not* carry over from one to the next; the only assumptions you may make in a given part are those which are explicitly stated in that part's description.

(a) Model the changing fanbases as a Bayesian network. You should create 8 nodes: B_t, B_{t+1}, M_t, M_{t+1}, A_t, A_{t+1}, J_t, and J_{t+1}. Indicate which nodes correspond to latent/hidden fanbase counts and which correspond to the observable emissions.



Solution



The changing fanbases process modeled as a dynamic Bayesian network. The unshaded nodes correspond to the latent/hidden fanbase counts, and the shaded nodes correspond to the observable emissions.

(b) **Domain Consistencies**

As a first step, we will not concern ourselves with which fanbase counts are *probable*, but instead which counts are even *possible*. Suppose that we observe, in our first month of collecting data, that $J_1 = 75$ and $A_1 = 100$. Give the domains for M_1 and B_1 that are consistent with these observations. You need only give the consistent domains (using either set notation or inequality notation).

Solution The only values of B_1 which are consistent (i.e. yield nonzero probability under $\Pr(J_1|B_1)$) with $J_1 = 75$ are $B_1 \in \{75, 76, 77, 78, 79\}$.

Now, we can use this reduced domain for B_1 together with the fact that $A_1 = 100$ to reason about the domain of M_1. To yield nonzero probability under $\Pr(A_1|B_1, M_1)$, we must have:

$$|A_1 - (B_1 + M_1)| = |100 - (B_1 + M_1)| \leq 5$$

And to satisfy this inequality, we must have:

$$95 \leq B_1 + M_1 \leq 105$$

If $B_1 = 75$, we see that $20 \leq M_1 \leq 30$. Because M_1 is largest when B_1 is smallest, this gives an upper bound of 30 on the domain of M_1. Similarly, if $B_1 = 79$, we see that $16 \leq M_1 \leq 26$. Because M_1 is smallest when B_1 is largest, this gives a lower bound of 16 on the domain of M_1. Thus, we conclude that $16 \leq M_1 \leq 30$.

(c) **Inference**

Suppose the Bayesian's manager took a nationwide poll in month t that concluded they had exactly 75 fans. Suppose additionally that in month $t+2$, the Bayesians sell 73 jerseys. What is the probability that in month $t+2$ the Bayesians have 77 fans?

i. What is the probability that in month $t+1$ the Bayesians sell 72 jerseys?

   $\Pr(J_{t+1} = 72|B_t = 75) =$

   Solution Marginalizing out B_{t+1} gives:

   $$\Pr(J_{t+1} = 72|B_t = 75) = \sum_x \Pr(J_{t+1} = 72|B_{t+1} = x)\Pr(B_{t+1} = x|B_t = 75)$$

   Given that $B_t = 75$, we know that B_{t+1} equals either 74 or 76 with equal probability. Plugging this in to the above expression gives:

   $$\Pr(J_{t+1} = 72|B_t = 75) = 0.5 \cdot \Pr(J_{t+1} = 72|B_{t+1} = 74) + 0.5 \cdot \Pr(J_{t+1} = 72|B_{t+1} = 76)$$

   And lastly, appealing to our model for jersey counts yields the expression:

   $$\Pr(J_{t+1} = 72|B_t = 75) = 0.5 \cdot 0.2 + 0.5 \cdot 0.1 = 0.5 \cdot 0.3 = 0.15$$

ii. What is the probability that in month $t+2$ that the Bayesians have 77 fans given that that they had 75 in month t and sold 73 jerseys in month $t+2$?

   $\Pr(B_{t+2} = 77|B_t = 75, J_{t+2} = 73) =$

   Solution By Bayes rule, we have:

   $$\Pr(B_{t+2} = 77|B_t = 75, J_{t+2} = 73) = \frac{\Pr(J_{t+2} = 73|B_t = 75, B_{t+2} = 77)\Pr(B_{t+2} = 77|B_t = 75)}{\Pr(J_{t+2} = 73|B_t = 75)}$$

   We'll begin with the first term in the numerator; because J_{t+2} is conditionally independent of B_t given B_{t+2}, we have $\Pr(J_{t+2} = 73|B_t = 75, B_{t+2} = 77) = \Pr(J_{t+2} = 73|B_{t+2} = 77)$. This is simply given by our jersey sales model; the probability that the Bayesians sell four fewer jerseys than they have fans is 0.1.

   We turn next to the second term in the numerator; if there are 75 fans in month t, then with equal probability there are either 74 or 76 fans in month $t+1$. If there were 74 in month $t+1$, then there would be either 73 or 75 in month $t+2$ with equal probability, and if there were 76 in month $t+1$, then there would be either 75 or 77 in month $t+2$ with equal probability. Thus, we have that $\Pr(B_{t+2} = 73|B_t = 75) = \Pr(B_{t+2} = 77|B_t = 75) = 0.25$, and $\Pr(B_{t+2} = 75|B_t = 75) = 0.5$.

Now, to compute the denominator, we simply sum the expression in the numerator across all possible values for B_{t+2}:

$$\Pr(J_{t+2} = 73 | B_t = 75) = \sum_x \Pr(J_{t+2} = 73 | B_t = 75, B_{t+2} = x) \Pr(B_{t+2} = x | B_t = 75)$$

Following the same reasoning as we used for the numerator, this evaluates to:

$$\begin{aligned}\Pr(J_{t+2} = 73 | B_t = 75) =& 0.25 \cdot \Pr(J_{t+2} = 73 | B_{t+2} = 73) \\ &+ 0.5 \cdot \Pr(J_{t+2} = 73 | B_{t+2} = 75) \\ &+ 0.25 \cdot \Pr(J_{t+2} = 73 | B_{t+2} = 77) \\ =& 0.25 \cdot 0.3 + 0.5 \cdot 0.2 + 0.25 \cdot 0.1 = 0.075 + 0.1 + 0.025 \\ =& 0.2\end{aligned}$$

So altogether, we have:

$$\Pr(B_{t+2} = 77 | B_t = 75, J_{t+2} = 73) = \frac{0.1 \cdot 0.25}{0.2} = \frac{1}{2} \cdot \frac{1}{4} = \frac{1}{8}$$

(d) **Gibbs Sampling**

Inference is exhausting; you decide that you'd be satisfied with simply being able to draw samples from distributions rather than specifying them exactly. In particular, you want to sample joint assignments to the variables $\{B_t, M_t, A_t, J_t\}_{t=1}^T$ for some time horizon T. You decide to implement Gibbs sampling for this purpose, but something's not right! What additional information, beyond what we've given you, would allow you to perform Gibbs sampling? Briefly explain.

Solution (The following argument applies identically to M_t as well as B_t): In order to sample B_t, we need to have first assigned a value to B_{t-1}; but in order to have sampled a value for B_{t-1}, we need to have first assigned a value to B_{t-2}, and so on. Continuing in this way, we realize that we must have a way of assigning a value to B_1 in order to perform Gibbs sampling. But to do this, we would either need to specify a fixed value for B_1, or specify a prior distribution $\Pr(B_1)$ from which to sample.

(e) **Exact Filtering**

You now want to begin making inferences as to the sizes of the teams' fanbases given only observations of attendances and jersey sales. Recall that exact inference of this kind in dynamic Bayesian networks can be achieved using a dynamic programming approach - for example, in the context of Hidden Markov Models, we used the forward-backward algorithm to do filtering and smoothing.

Give recursive expressions for the following filtering queries. Leave your expressions in terms of known probabilities.

i. Let's start by making inferences based only on observed jersey sales. Denote $F_t(b_t) = \Pr(B_t = b_t | J_1 = j_1, \ldots, J_t = j_t)$. Give a recursive expression for $F_t(b_t)$ assuming that you've already computed $F_{t-1}(b_{t-1})$ for all b_{t-1}.

Solution This is exactly the "forward" computation in an HMM. We can compute the unnormalized quantity, which we'll denote $\tilde{F}_t(b_t)$, using the standard forward update:

$$\tilde{F}_t(b_t) = \sum_{b_{t-1}} F_{t-1}(b_{t-1}) \cdot \Pr(B_t = b_t | B_{t-1} = b_{t-1}) \cdot \Pr(J_t = j_t | B_t = b_t)$$

and can subsequently produce the required probability by normalizing:

$$F_t(b_t) = \frac{\tilde{F}_t(b_t)}{\sum_{b'_t} \tilde{F}_t(b'_t)}$$

ii. Let's bring in the observed attendances as well! Now, denote $F_t(b_t, m_t) = \Pr(B_t = b_t, M_t = m_t | J_1 = j_1, \ldots, J_t = j_t, A_1 = a_1, \ldots, A_t = a_t)$. Give a recursive expression for $F_t(b_t, m_t)$ assuming that you've already computed $F_{t-1}(b_{t-1}, m_{t-1})$ for all b_{t-1} and all m_{t-1}.

Solution This closely mirrors the "forward" computation in an HMM, but now we must account for the dynamics of both hidden states, as well as the probabilities of both observed emissions. We can compute the unnormalized quantity, which we'll denote $\tilde{F}_t(b_t, m_t)$, using the following forward update:

$$\begin{aligned}\tilde{F}_t(b_t, m_t) = \sum_{b_{t-1}, m_{t-1}} & F_{t-1}(b_{t-1}, m_{t-1}) \cdot \Pr(B_t = b_t | B_{t-1} = b_{t-1}) \\ & \cdot \Pr(M_t = m_t | M_{t-1} = m_{t-1}) \cdot \Pr(J_t = j_t | B_t = b_t) \\ & \cdot \Pr(A_t = a_t | B_t = b_t, M_t = m_t)\end{aligned}$$

and can subsequently produce the required probability by normalizing:

$$F_t(b_t, m_t) = \frac{\tilde{F}_t(b_t, m_t)}{\sum_{b'_t, m'_t} \tilde{F}_t(b'_t, m'_t)}$$

(f) **Particle Filtering**

Throughout this problem, you are free to leave quantities in terms of unevaluated expressions (i.e. you may write $0.75 \cdot 0.5$ instead of 0.375).

Computing all of those terms exactly seems tedious, so you instead decide to employ particle filtering to quickly and painlessly provide you with approximate solutions. You're fine with a (very) crude approximation, so you only use two particles.

i. Suppose you begin with the two particles $(B_1 = 80, M_1 = 75)$ and $(B_1 = 82, M_1 = 74)$. You then observe that $J_1 = 79$ and $A_1 = 154$. Compute the weights that you should assign to the two particles based on this evidence.

Solution For the first particle, we have $\Pr(A_1 = 154|B_1 = 80, M_1 = 75) = 0.13$ and $\Pr(J_1 = 79|B_1 = 80) = 0.25$. Thus, the first particle should get a weight of $0.13 * 0.25 = 0.0325$.
Similarly, for the second particle, we have $\Pr(A_1 = 154|B_1 = 82, M_1 = 74) = 0.11$ and $\Pr(J_1 = 79|B_1 = 82) = 0.15$. Thus, the second particle should get a weight of $0.11 * 0.15 = 0.0165$.

ii. Using these weights, we now resample two new particles. Provide this sampling distribution.
Probability of sampling a new particle to be $(B_1 = 80, M_1 = 75) =$

Solution $\frac{0.0325}{0.0325+0.0165}$
Probability of sampling a new particle to be $(B_1 = 82, M_1 = 74) =$

Solution $\frac{0.0165}{0.0325+0.0165}$

iii. Suppose both of our new particles are sampled to be $(B_1 = 80, M_1 = 75)$. We now extend these particles using our dynamics models. What is the probability that a particular one of these two particles is extended to:
$(B_1 = 80, M_1 = 75, B_2 = 78, M_2 = 76)$?

Solution Zero. Under the given model for $\Pr(B_{t+1}|B_t)$, the only possible values for B_2 are 79 and 81.

$(B_1 = 80, M_1 = 76, B_2 = 79, M_2 = 75)$?

Solution Zero. The value assigned to M_1 cannot change upon extending the particle.

$(B_1 = 80, M_1 = 75, B_2 = 79, M_2 = 76)$?

Solution $\Pr(B_2 = 79|B_1 = 80) \cdot \Pr(M_2 = 76|M_1 = 75) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$

iv. Suppose now that you have access to a large number of particles which are approximating the distribution over $(B_1, \ldots, B_n, M_1, \ldots, M_n)$. The Olympics are happening in 6 months, but you have to decide now which team to send so that they can start preparing! You decide to make predictions of B_{n+6} and M_{n+6} in order to send whichever team you predict to be more popular during the month in which the Olympics will be held. Explain in a few sentences how you would use your particles for making this decision.

Solution Propagate each particle through the two dynamics models six times in order to sample values of $B_{n+1}, \ldots, B_{n+6}$ and $M_{n+1} \ldots, M_{n+6}$ for each particle. Compute the average values of B_{n+6} and M_{n+6} across all of the particles, and send whichever team has the larger average value.

2) PS9 Problem 4: Knowledge Base

Imagine we are building a knowledge base of propositions in first order logic and want to make inferences based on what we know. We will deal with a simple setting, where we only have three objects in the world: Alice, Carol, and Bob. Our predicates are as follows:

- Employee(x): x is an employee.
- Boss(x): x is a boss.
- Works(x): x works.
- Paid(x): x gets paid.

The knowledge base we have constructed consists of the following propositions:

(a) Boss(Carol)

(b) Employee(Bob)

(c) Paid(Carol) $\wedge$ Works(Carol)

(d) Paid(Alice)

(e) $\forall x$ (Employee(x) $\leftrightarrow \neg$ Boss(x))

(f) $\forall x$ (Employee(x) $\rightarrow$ Works(x))

(g) $\forall x$ ((Paid(x) $\wedge \neg$ Works(x)) $\rightarrow$ Boss(x))

(a) We know from class that one technique we can use to perform inference with our knowledge base is to propositionalize the statements of first-order logic into statements of propositional logic. Practice this by propositionalizing statement (6) from our knowledge base.

Solution (EmployeeAlice $\rightarrow$ WorksAlice) $\wedge$ (EmployeeBob $\rightarrow$ WorksBob) $\wedge$ (EmployeeCarol $\rightarrow$ WorksCarol)

(b) If we translated the statement "Anyone who is not a boss either works or does not get paid" into first-order logic and added it to our knowledge base, how would the size of the set of valid models representing our knowledge base change, and why?

Solution The set of valid would stay the same as the statement is entailed by our current knowledge base.

(c) Using only our original knowledge base (not including the statement from part (b)), we want to answer the question "Does everyone work?" We first translate the sentence "everyone works" into first order logic as statement f. Determine the answer to our query by considering the following questions of satisfiability:

① Is KB $\cup \neg f$ satisfiable? Answer yes/no. If yes, fill in the following table with T for true and F for false to show that there is a satisfying model.

x	Employee(x)	Boss(x)	Works(x)	Paid(x)
Alice				
Bob				
Carol				

Solution Yes

x	Employee(x)	Boss(x)	Works(x)	Paid(x)
Alice	F	T	F	T
Bob	T	F	T	T or F
Carol	F	T	T	T

② Is KB ∪ f satisfiable? Answer yes/no. If yes, fill in the following table with T for true and F for false to show that there is a satisfying model.

x	Employee(x)	Boss(x)	Works(x)	Paid(x)
Alice				
Bob				
Carol				

Solution Yes

x	Employee(x)	Boss(x)	Works(x)	Paid(x)
Alice	T or F	Opposite	T	T
Bob	T	F	T	T or F
Carol	F	T	T	T

③ Based on your answers to the previous two parts, does our knowledge base entail f, contradict f, or is f contingent? And what should the answer to our original question ”Does everyone work?” be?

Solution f is contingent. Answer should be ”maybe” or ”it depends”

3) **Problem 3: CA Assignment (Winter 21, Problem 1)**

Every quarter, the Stanford computer science department assigns graduate students as course assistants (CAs). Students who wish to serve as CAs fill out an application in which they can list the classes they'd like to CA for. After the application due date, the department matches applicants to courses, taking into account student preferences as well as how many course assistants each class needs. Here's the formal CA-assignment problem setup:

- There are n students $S_1, \ldots, S_n$ who apply for CAships.
- There are m courses $C_1, \ldots, C_m$ that have CA openings.
- Each student S_i specifies arbitrary non-negative preferences $P_1^{(i)}, \ldots, P_m^{(i)} \geq 0$ for each of the m classes. A large preference value $P_j^{(i)}$ means student S_i really wants to CA for class C_j, and a preference value of 0 for $P_j^{(i)}$ means student S_i does not want to CA for class C_j.

The CA-matching process must adhere to the following requirements:

- Each course C_i can have a maximum of M_i course assistants.
- Every student must be matched to exactly one class for which they have specified a positive preference (assume each student has at least one such preference).

Model the CA-matching process with a CSP with n variables, one for each student $S_1, \ldots, S_n$. Our CSP should find the *maximum weight assignment*, where the weights are determined by student preferences.

(a) What is the domain of each variable and what is the cardinality?

Solution The domain for each student is of cardinality m with values $\{C_1, \ldots, C_m\}$.

(b) What are the factors? State the arity of each.

Solution We have 2 sets of factors. The first group encodes the maximum CA openings for each course. This can be written as n-ary factors $f_1, \ldots, f_m$, where

$$f_j\left(S_1, \ldots, S_n\right) = \left[\sum_{i=1}^{n} \left[S_i = C_j\right] \leq M_j\right]$$

The second set encodes individual student preferences for which class they'd like to CA. These are unary factors $g_1, \ldots, g_n$, where

$$g_i\left(S_i\right) = P_{S_i}^{(i)}$$

which is the preference of the i th student to CA for class S_i.

(c) We imagine a small setting of this problem for 3 students S_1, S_2, S_3 and 3 courses C_1, C_2, C_3. The student preferences are given by the following table:

	C_1	C_2	C_3
S_1	3	0	0
S_2	2	1	3
S_3	5	3	0

Additionally, classes C_1 and C_2 can have a maximum of 1 CA each, and class C_3 can have at most 2 CAs.

Apply the CSP you designed to this small setting and enforce arc-consistency amongst its variables. In particular, write out each variable and its domain after arc-consistency has been enforced. For example, if you have a variable X_i with a domain $\{a, b, c\}$ after enforcing arc-consistency, you should write

$$X_i : \{a, b, c\}$$

Solution

$$S_1 : \{C_1\}$$
$$S_2 : \{C_3\}$$
$$S_3 : \{C_2\}$$

(d) True or False, with justification.

i. The least constrained value (LCV) heuristic would be a useful optimization for our CA-assignment CSP.

Solution False. LCV is a useful optimization when all of our factors are constraints. Since student preferences are arbitrary non-negative values, we need to try all of the consistent values anyway.

ii. The most constrained variable (MCV) heuristic would be a useful optimzation for our CA-assignment CSP.

Solution True. MCV is a useful optimization when some of our factors are constraints. Since we have constraints for the maximum number of CAs for a particular class, MCV can help.

iii. If we use the ICM algorithm to solve our CA-assignment CSP, everytime we modify a single variable assignment our factor recomputation will be on the order of n (recall that n is the number of students applying for a CA assignment).

Solution True. While in general for ICM we only need to compute factors involving the single variable whose assignment we changed, the constraints on the maximum number of classes (which depends on all n variables) will need to be recomputed.

iv. If we use beam search with different beam sizes k to solve our CA-assignment CSP, our solution's assignment weight will always increase as we increase the beam size k.

Solution False. Consider going from $k = 1$ (greedy) to $k = 2$. The greedy solution might be the globally optimal assignment, but when $k = 2$ we may find more partially optimal solutions as we expand more paths that cause us to drop the greedy solution from our beam. We are only guaranteed a global optimum with an unbounded beam size.

(e) Explain how you would modify your CSP from part a. to allow for the possibility that some students aren't matched to a course. You should encode the (realistic) assumption that not receiving a CAship is the least-preferable assignment for the student. (Note that students can still give a preference of 0 for a class if they do not want to be a CA for that class, and your modification should not prohibit this.)

Solution Solution A simple way to do this is to extend the domain for each student to include a non-assignment value, which we can denote $\varnothing$. We can then augment the unary factors that encode student preferences to output a positive value for $\varnothing$ that is smaller then all of the student's positive preferences.