CS 188 Fall 2018 — Introduction to Artificial Intelligence — Written HW 9 Sol.

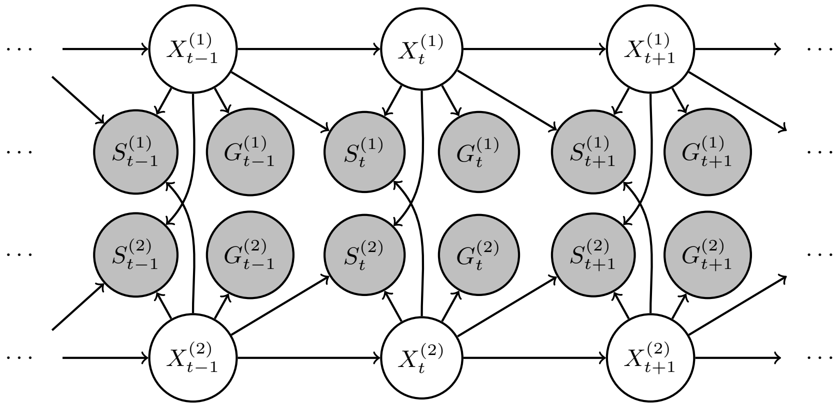
Self-assessment due: Tuesday 11/13/2018 at 11:59pm (submit via Gradescope)

For the self assessment, **fill in the self assessment boxes in your original submission** (you can download a PDF copy of your submission from Gradescope – be sure to delete any extra title pages that Gradescope attaches). For each subpart where your original answer was correct, write "correct." Otherwise, write and explain the correct answer. **Do not leave any boxes empty.** If you did not submit the homework (or skipped some questions) but wish to receive credit for the self-assessment, we ask that you first complete the homework without looking at the solutions, and then perform the self assessment afterwards.

Q1. Particle Filtering: Where are the Two Cars?

As before, we are trying to estimate the location of cars in a city, but now, we model two cars jointly, i.e. car i for $i \in \{1, 2\}$. The modified HMM model is as follows:

- $X^{(i)}$ – the location of car i
- $S^{(i)}$ – the noisy location of the car i from the signal strength at a nearby cell phone tower
- $G^{(i)}$ – the noisy location of car i from GPS



d	$D(d)$	$E_L(d)$	$E_N(d)$	$E_G(d)$
-4	0.05	0	0.02	0
-3	0.10	0	0.04	0.03
-2	0.25	0.05	0.09	0.07
-1	0.10	0.10	0.20	0.15
0	0	0.70	0.30	0.50
1	0.10	0.10	0.20	0.15
2	0.25	0.05	0.09	0.07
3	0.10	0	0.04	0.03
4	0.05	0	0.02	0

The signal strength from one car gets noisier if the other car is at the same location. Thus, the observation $S_t^{(i)}$ also depends on the current state of the other car $X_t^{(j)}$, $j \neq i$.

The transition is modeled using a drift model D, the GPS observation $G_t^{(i)}$ using the error model E_G, and the observation $S_t^{(i)}$ using one of the error models E_L or E_N, depending on the car's speed and the relative location of both cars. These drift and error models are in the table above. **The transition and observation models are:**

$$P(X_t^{(i)}|X_{t-1}^{(i)}) = D(X_t^{(i)} - X_{t-1}^{(i)})$$

$$P(S_t^{(i)}|X_{t-1}^{(i)}, X_t^{(i)}, X_t^{(j)}) = \begin{cases} E_N(X_t^{(i)} - S_t^{(i)}), & \text{if } |X_t^{(i)} - X_{t-1}^{(i)}| \geq 2 \text{ or } X_t^{(i)} = X_t^{(j)} \\ E_L(X_t^{(i)} - S_t^{(i)}), & \text{otherwise} \end{cases}$$

$$P(G_t^{(i)}|X_t^{(i)}) = E_G(X_t^{(i)} - G_t^{(i)}).$$

Throughout this problem you may give answers either as unevaluated numeric expressions (e.g. $0.1 \cdot 0.5$) or as numeric values (e.g. 0.05). The questions are decoupled.

(a) Assume that at $t = 3$, we have the single particle $(X_3^{(1)} = -1, X_3^{(2)} = 2)$.

(i) What is the probability that this particle becomes $(X_4^{(1)} = -3, X_4^{(2)} = 3)$ after passing it through the dynamics model?

$$\begin{aligned} P(X_4^{(1)} = -3, X_4^{(2)} = 3|X_3^{(1)} = -1, X_3^{(2)} = 2) &= P(X_4^{(1)} = -3|X_3^{(1)} = -1) \cdot P(X_4^{(2)} = 3|X_3^{(2)} = 2) \\ &= D(-3 - (-1)) \cdot D(3 - 2) \\ &= 0.25 \cdot 0.10 \\ &= 0.025 \end{aligned}$$

Answer: 0.025

(ii) Assume that there are no sensor readings at $t = 4$. What is the joint probability that the *original* single particle (from $t = 3$) becomes $(X_4^{(1)} = -3, X_4^{(2)} = 3)$ and then becomes $(X_5^{(1)} = -4, X_5^{(2)} = 4)$?

$$\begin{aligned}
&P(X_4^{(1)} = -3, X_5^{(1)} = -4, X_4^{(2)} = 3, X_5^{(2)} = 4 | X_3^{(1)} = -1, X_3^{(2)} = 2) \\
&= P(X_4^{(1)} = -3, X_5^{(1)} = -4 | X_3^{(1)} = -1) \cdot P(X_4^{(2)} = 3, X_5^{(2)} = 4 | X_3^{(2)} = 2) \\
&= P(X_5^{(1)} = -4 | X_4^{(1)} = -3) \cdot P(X_4^{(1)} = -3 | X_3^{(1)} = -1) \cdot P(X_5^{(2)} = 4 | X_4^{(2)} = 3) \cdot P(X_4^{(2)} = 3 | X_3^{(2)} = 2) \\
&= D(-4 - (-3)) \cdot D(-3 - (-1)) \cdot D(4 - 3) \cdot D(3 - 2) \\
&= 0.10 \cdot 0.25 \cdot 0.10 \cdot 0.10 \\
&= 0.00025
\end{aligned}$$

Answer: 0.00025

For the remaining of this problem, we will be using 2 particles at each time step.

(b) At $t = 6$, we have particles $[(X_6^{(1)} = 3, X_6^{(2)} = 0), (X_6^{(1)} = 3, X_6^{(2)} = 5)]$. Suppose that after weighting, resampling, and transitioning from $t = 6$ to $t = 7$, the particles become $[(X_7^{(1)} = 2, X_7^{(2)} = 2), (X_7^{(1)} = 4, X_7^{(2)} = 1)]$.

(i) At $t = 7$, you get the observations $S_7^{(1)} = 2, G_7^{(1)} = 2, S_7^{(2)} = 2, G_7^{(2)} = 2$. What is the weight of each particle?

Particle	Weight
$(X_7^{(1)} = 2, X_7^{(2)} = 2)$	$P(S_7^{(1)} = 2 \| X_6^{(1)} = 3, X_7^{(1)} = 2, X_7^{(2)} = 2) \cdot P(G_7^{(1)} = 2 \| X_7^{(1)} = 2) \cdot$ $P(S_7^{(2)} = 2 \| X_6^{(2)} = 0, X_7^{(2)} = 2, X_7^{(1)} = 2) \cdot P(G_7^{(2)} = 2 \| X_7^{(2)} = 2)$ $= E_N(2-2) \cdot E_G(2-2) \cdot E_N(2-2) \cdot E_G(2-2)$ $= 0.30 \cdot 0.50 \cdot 0.30 \cdot 0.50$ $= 0.0225$
$(X_7^{(1)} = 4, X_7^{(2)} = 1)$	$P(S_7^{(1)} = 2 \| X_6^{(1)} = 3, X_7^{(1)} = 4, X_7^{(2)} = 1) \cdot P(G_7^{(1)} = 2 \| X_7^{(1)} = 4) \cdot$ $P(S_7^{(2)} = 2 \| X_6^{(2)} = 5, X_7^{(2)} = 1, X_7^{(1)} = 4) \cdot P(G_7^{(2)} = 2 \| X_7^{(2)} = 1)$ $= E_L(4-2) \cdot E_G(4-2) \cdot E_N(1-2) \cdot E_G(1-2)$ $= 0.05 \cdot 0.07 \cdot 0.20 \cdot 0.15$ $= 0.000105$

(ii) Suppose both cars' cell phones died so you only get the observations $G_7^{(1)} = 2, G_7^{(2)} = 2$. What is the weight of each particle?

Particle	Weight
$(X_7^{(1)} = 2, X_7^{(2)} = 2)$	$P(G_7^{(1)} = 2 \| X_7^{(1)} = 2) \cdot P(G_7^{(2)} = 2 \| X_7^{(2)} = 2)$ $= E_G(2-2) \cdot E_G(2-2)$ $= 0.50 \cdot 0.50$ $= 0.25$
$(X_7^{(1)} = 4, X_7^{(2)} = 1)$	$P(G_7^{(1)} = 2 \| X_7^{(1)} = 4) \cdot P(G_7^{(2)} = 2 \| X_7^{(2)} = 1)$ $= E_G(4-2) \cdot E_G(1-2)$ $= 0.07 \cdot 0.15$ $= 0.0105$

(c) To decouple this question, assume that you got the following weights for the two particles.

Particle	Weight
$(X_7^{(1)} = 2, X_7^{(2)} = 2)$	0.09
$(X_7^{(1)} = 4, X_7^{(2)} = 1)$	0.01

What is the belief for the location of car 1 and car 2 at $t = 7$?

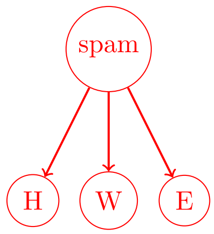
Location	$P(X_7^{(1)})$	$P(X_7^{(2)})$
$X_7^{(i)} = 1$	$\frac{0}{0.09+0.01} = 0$	$\frac{0.01}{0.09+0.01} = 0.1$
$X_7^{(i)} = 2$	$\frac{0.09}{0.09+0.01} = 0.9$	$\frac{0.09}{0.09+0.01} = 0.9$
$X_7^{(i)} = 4$	$\frac{0.01}{0.09+0.01} = 0.1$	$\frac{0}{0.09+0.01} = 0$

Q2. Naive Bayes

Your friend claims that he can write an effective Naive Bayes spam detector with only three features: the hour of the day that the email was received ($H \in \{1, 2, \ldots, 24\}$), whether it contains the word 'viagra' ($W \in \{\text{yes}, \text{no}\}$), and whether the email address of the sender is Known in his address book, Seen before in his inbox, or Unseen before ($E \in \{\text{K}, \text{S}, \text{U}\}$).

(a) Flesh out the following information about this Bayes net:

Graph structure:



Parameters:

$\theta_{spam}, \theta_{H,i,c}, \theta_{W,c}, \theta_{E,j,c}, i \in \{1, \ldots, 23\}, j \in \{K, S\}, c \in \{spam, ham\}$ is a correct minimal parameterization. Note that the sum-to-one constraint on distributions results in one fewer parameter than the number of settings of a variable. For instance, θ_{spam} suffices because $\theta_{ham} = 1 - \theta_{spam}$. *Aside:* a non-minimal but correct parameterization was also accepted since the question did not ask for minimal parameters.

Size of the set of parameters:

$1 + 23 \cdot 2 + 2 + 2 \cdot 2$. The size of the set is the sum of parameter sizes. Every parameter has size = number of values $\times$ number of settings of its parents. For instance, $\theta_{H,i,c}$ has 23 values of hour H and its parent c, the class, has 2.

Suppose now that you labeled three of the emails in your mailbox to test this idea:

spam or ham?	H	W	E
spam	3	yes	S
ham	14	no	K
ham	15	no	K

(b) Use the three instances to estimate the maximum likelihood parameters. The maximum likelihood estimates are the sample proportions.
$\theta_{spam} = 1/3, \theta_{H,3,spam} = 1, \theta_{H,14,ham} = 1/2, \theta_{H,15,ham} = 1/2, \theta_{W,spam} = 1.0, \theta_{E,S,spam} = 1, \theta_{E,K,ham} = 1$

(c) Using the maximum likelihood parameters, find the predicted class of a new datapoint with $H = 3$, $W = \text{no}$, $E = U$.
No prediction can be made. Since $E = U$ is never observed, it has zero likelihood under both classes.

(d) Now use the three to estimate the parameters using Laplace smoothing and $k = 2$. Do not forget to smooth both the class prior parameters and the feature values parameters. The Laplace smoothed estimate for a categorical variable X with parameters $\theta_{1,\cdots,d}$ for the $\{1,\cdots,d\}$ values of X is $\theta_i = \frac{x_i+k}{N+kd}$ where x_i is the number of times value i is observed, N is the total number of observations, and d is the number of values of X.

$\theta_{spam} = 3/7, \theta_{H,3,spam} = 3/49, \theta_{H,other,spam} = 2/49, \theta_{H,14,ham} = 3/50, \theta_{H,15,ham} = 3/50, \theta_{H,other,ham} = 2/50, \theta_{W,spam} = 3/5, \theta_{W,ham} = 2/6, \theta_{E,S,spam} = 3/7, \theta_{E,other,spam} = 2/7, \theta_{E,K,ham} = 4/8, \theta_{E,other,ham} = 2/8.$

(e) Using the parameters obtained with Laplace smoothing, find the predicted class of a new datapoint with $H = 3$, $W = \text{no}$, $E = U$. Ham. The probability under the model for each class is computed as the product of the class prior and the feature conditonals: $p(ham) \propto (1-\theta_{spam})(\theta_{H,other,ham})(1-\theta_{W,ham})(1-\theta_{E,K,ham}-\theta_{E,other,ham})$ and $p(spam) \propto (\theta_{spam})(\theta_{H,3,spam})(1-\theta_{W,spam})(1-\theta_{E,S,spam}-\theta_{E,other,spam})$ where both are proportional because the distribution has not been normalized.

(f) You observe that you tend to receive spam emails in batches. In particular, if you receive one spam message, the next message is more likely to be a spam message as well. Explain a new graphical model which most naturally captures this phenomena.

Graph structure:

The structure is the same as an HMM except each hidden state node has three observation child nodes.

Parameters:

Add 2 parameters: transition to spam from spam and from ham.

Size of the set of parameters:

Add 2 to the expression in the first question.