CS 189/289A Introduction to Machine Learning
Fall 2023 Jennifer Listgarten, Jitendra Malik

DIS10

1 Probabilistic Graphical Models

Recall that we can represent joint probability distributions with directed acyclic graphs (DAGs). Let G be a DAG with vertices $X_1, ..., X_k$. If P is a (joint) distribution for $X_1, ..., X_k$ with (joint) probability mass function p, we say that G represents P if

$$p(x_1, \cdots, x_k) = \prod_{i=1}^{k} P(X_i = x_i | \text{pa}(X_i)), \tag{1}$$

where $\text{pa}(X_i)$ denotes the parent nodes of X_i. (Recall that in a DAG, node Z is a parent of node X iff there is a directed edge going out of Z into X.)

Consider the following DAG

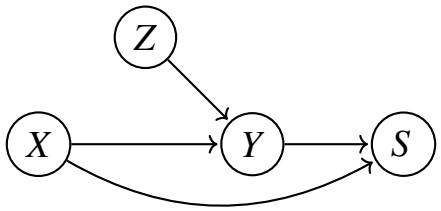

Figure 1: G, a DAG

(a) Write down the joint factorization of $P_{S,X,Y,Z}(s, x, y, z)$ implied by the DAG G shown in Figure 1.

Solution:

$$P_{S,X,Y,Z}(s, x, y, z) = P(X = x)P(Z = z)P(Y = y|X = x, Z = z)P(S = s|X = x, Y = y)\,.$$

(b) Is $S \perp Z \mid Y$?

Solution: No. As a counterexample, consider the case where all nodes represent binary random variables, $P(X = 1) = P(Z = 1) = 0.5$, $Y = X \otimes Z$, and $S = X \otimes Y$, where $\otimes$ is the XOR operator. Then we can see that $S = Z$, whereas knowing Y does not fully determine S or Z.

A version of these solutions from a previous semester erroneously said that this conditional independence did hold. As a result, you may have wrongly heard in section that this statement is true, via faulty algebraic manipulation and/or other algorithms such as the Bayes ball (d-separation). Running these algorithms correctly should show that S and Z are indeed not conditionally independent given Y.

If X is fully removed from G, then we do indeed have $S \perp Z \mid Y$. This is left as an exercise in algebraic manipulation of probability distributions.

(c) Is $S \perp X \mid Y$?

Solution: No. Consider the same example from above with binary random variables. Knowing Y does not determine S, but knowing both X and Y does.

2 PGMs: Sleeping in Class

In this question, you'll be reasoning about a Dynamic Bayesian Network (DBN), a form of a Probabilistic Graphical Model.

Your favorite discussion section TA wants to know if their students are getting enough sleep. Each day, the TA observes the students in their section, noting if they fall asleep in class or have red eyes. The TA makes the following conclusions:

1. The prior probability of getting enough sleep, S, with no observations, is 0.7.
2. The probability of getting enough sleep on night t is 0.8 given that the student got enough sleep the previous night, and 0.3 if not.
3. The probability of having red eyes R is 0.2 if the student got enough sleep, and 0.7 if not.
4. The probability of sleeping in class C is 0.1 if the student got enough sleep, and 0.3 if not.

(a) Formulate this information as a dynamic Bayesian network that the professor could use to filter or predict from a sequence of observations. If you were to reformulate this network as a hidden Markov model instead (that has only a single observation variable), how would you do so? Give a high-level description (probability tables for the HMM formulation are not necessary.)

Solution: Our Bayesian Network has three variables: S_t, whether the student gets enough sleep, R_t, whether they have red eyes in class, and C_t, whether the student sleeps in class. S_t is a parent of S_{t+1}, R_t, and C_t. The network can be provided pictorally, or fully through conditional probability tables (CPTs.) The CPTs for this problem are given by:
$P(s_0) = 0.7$
$P(s_{t+1}|s_t) = 0.8$
$P(s_{t+1}|\neg s_t) = 0.3$
$P(r_t|s_t) = 0.2$
$P(r_t|\neg s_t) = 0.7$
$P(c_t|s_t) = 0.1$
$P(c_t|\neg s_t) = 0.3$

To reformulate this problem as an HMM with a single observation node, we can combine the 2-valued variables r_t and c_t into a single 4-valued variable, multiplying together the emission probabilities.

(b) Consider the following evidence values at timesteps 1, 2, and 3:

(a) e_1 = not red eyes, not sleeping in class
(b) e_2 = red eyes, not sleeping in class
(c) e_3 = red eyes, sleeping in class

Compute state estimates for timesteps t at 1, 2, and 3; that is, calculate $P(S_t|e_{1:t})$. Assume a prior on $P(s_0)$ that is consistent with the prior in the previous part; that is, $P(s_0) = 0.7$.
Solution: We can apply the filtering (forward computation) method. We walk through the computation step-by-step:

$$\begin{aligned}
P(S_0) &= \langle 0.7, 0.3\rangle \\
P(S_1) &= \sum_{s_0} P(S_1|s_0)P(s_0) \\
&= \langle 0.8, 0.2\rangle 0.7 + \langle 0.3, 0.7\rangle 0.3 \\
&= \langle 0.65, 0.35\rangle \\
P(S_1|e_1) &= \alpha P(e_1|S_1)P(S_1) \\
&= \alpha\langle 0.8 * 0.9, 0.3 * 0.7\rangle\langle 0.65, 0.35\rangle
\end{aligned}$$

After normalizing, we get the following (the rest of the solution(s) will normalize implicitly.)

$$\begin{aligned}
&= \langle 0.8643, 0.1357\rangle \\
P(S_2|e_1) &= \sum_{s_1} P(S_2|s_1)P(s_1|e_1) \\
&= \langle 0.7321, 0.2679\rangle \\
P(S_2|e_1 : e_2) &= \alpha P(e_2|S_2)P(S_2|e_1) \\
&= \langle 0.5010, 0.4990\rangle \\
P(S_3|e_1 : e_2) &= \sum_{s_2} P(S_3|s_2)P(s_1|e_1 : e_2) \\
&= \langle 0.5505, 0.4495\rangle \\
P(S_3|e_1 : e_3) &= \alpha P(e_3|S_3)P(S_3|e_1 : e_2) \\
&= \langle 0.1045, 0.8955\rangle
\end{aligned}$$

(c) Compute smoothing estimates $P(S_t|e_{1:3})$ for each timestep, using the same evidence as the previous part.

Solution:

First, we do the backwards computations:

$$\begin{aligned}
P(e_3|S_3) &= \langle 0.2 * 0.1, 0.7 * 0.3\rangle \\
&= \langle 0.02, 0.21\rangle \\
P(e_3|S_2) &= \sum_{s_3} P(e_3|s_3)P(s_3|S_2) \\
&= \langle 0.02 * 0.8 + 0.21 * 0.2, 0.02 * 0.3 + 0.21 * 0.7\rangle \\
&= \langle 0.0588, 0.153\rangle
\end{aligned}$$

$$P(e_2 : e_3|S\,1) = \sum_{s_2} P(e_2|s_2)P(e_3|s_2)P(s_2|S_1)$$
$$= \langle 0.0233, 0.0556 \rangle$$

Now, we can combine them with the forwards computation and normalize.

$$P(S_1|e_1 : e_3) = \alpha P(S_1|e_1)P(e_2 : e_3|S_1)$$
$$= \langle 0.7277, 0.2723 \rangle$$
$$P(S_2|e_1 : e_3) = \alpha P(S_2|e_1 : e_2)P(e_3|S_2)$$
$$= \langle 0.2757, 0.7243 \rangle$$
$$P(S_3|e_1 : e_3) = \langle 0.1045, 0.8955 \rangle$$

(d) Compare, in plain English, the filtered estimates you computed for timesteps 1 and 2 with the smoothed estimates. How do the two analyses differ?

Solution:

The smoothed analysis shows that the time the student started sleeping poorly is one timestep earlier than filtering only computation by incorporating future observations that indicated lack of sleep at the last step.