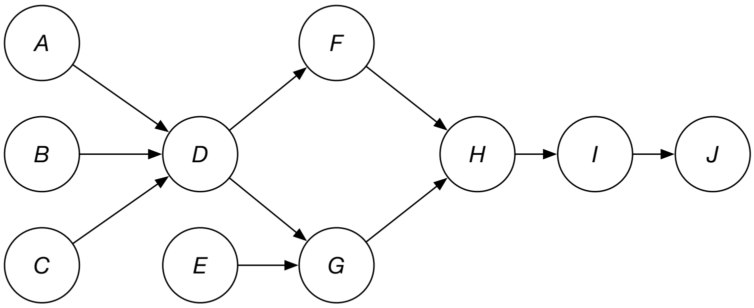
CS5340: Uncertainty Modeling in AI

Tutorial 3

Released: Jan. 31, 2024

Problem 1. (D-separation Test)
Test your d-separation process by answering the following conditional independence questions about the Bayes net shown below.



*Hint: Examine all paths between the two stated nodes and ensure that **every** trail is blocked by examining the 3-node structures along the trail. If every trail is blocked, the two nodes are d-separated and hence, conditionally independent. Otherwise, they are not guaranteed to be conditionally independent.*

Problem 1.a. Are H and J conditionally independent given no observations?

Problem 1.b. Are A and G conditionally independent if we know (given) D?

Problem 1.c. Are A and E conditionally independent given no observations?

Problem 1.d. Are A and E conditionally independent given F and J?

Problem 2. (HealthyStudents)

There has been discussion around campus about mental health issues. A new company HealthyStudents Inc. has approached the university with a plan: HealthyStudents will build an AI system that will monitor students to predict the occurrence of mental health issues[1]. They can then inform the university of such occurrences so that interventions can be taken and support can be given to at-risk students. They plan to use available sensors such as cameras on campus, as well as records of student performance in courses and participation in extra-curricular activities.

Problem 2.a. You are the chief AI specialist at HealthyStudents. Design a Bayesian Network along with the relevant conditional probability distributions. Consider what nodes are observed and which are latent (hidden or unobserved). This is an "open exercise" and there are multiple ways to specify the model.

Problem 2.b. Consider if it is *ethical* to develop this system — can it be misused to cause more harm than good? Do the benefits outweigh the risks? If you were on the university board, would you agree to HealthyStudent's plan?

[1] This is a fictional name and scenario. Any similarity to an actual company is purely coincidental. Read: I made this scenario up simply as a pedagogical exercise.

Problem 3. (Your CS5340 Grade)
In CS5340, we like to model everything, including how well students perform. Suppose that there are four possible final grades (a random variable Z) for the class, i.e., A, B, C, and D. Only two components affect a student's final grade: the student's project (X) and quiz score (Y).

X and Y have two possible outcomes each: Pass (1) or Fail (0), i.e., they are both binary random variables. In our simple model, assume that whether or not a student does well for the project and quiz depends *only* on how hard they work (W). Again, let us assume that this is a binary random variable where someone either works hard (1) or not (0).

Problem 3.a. Given the information above, define the necessary variables and give appropriate distributions. Draw the corresponding Bayesian network, and write down how the joint distribution $p(W, X, Y, Z)$ factorizes. *Hint: What distributions apply to discrete variables?*

Problem 3.b. Given the following observations from last year (assume iid), estimate the distribution parameters for each variable using MLE. *Note: Each row is an iid observation of all four random variables. Pay attention to how the factorization provided by the Bayesian network simplifies maximum likelihood estimation..*

W	X	Y	Z
0	0	0	0
0	0	0	1
0	1	0	1
0	1	1	2
1	0	0	1
1	0	1	2
1	1	1	2
1	1	1	3

Problem 4. (Label Errors)

You work as a data scientist for a YouWork! — the hottest startup in town. You are developing a model for predicting a person is likely to be promoted in the coming year. Given data point $\mathbf{x} \in \mathbb{R}^d$, your model predicts $p(y|\mathbf{x})$ where $y \in \{0, 1\}$ (1 indicates the person with data features $\mathbf{x}$ will be promoted and 0 otherwise). Note: we will abuse notation slightly in this exercise and use lower case letters for random variables.

Problem 4.a. Assume a logistic regression model where $y \sim \text{Bern}[\rho]$ and

$$\rho = \sigma(\mathbf{w}^\top \mathbf{x}) = \frac{1}{1 + \exp(-\mathbf{w}^\top \mathbf{x})}.$$

Construct a Bayesian network for this classification problem. Also make clear any prior and conditional distributions in your model. *Hint:* consider the linear regression example we saw in the lectures.

Problem 4.b. You wish to learn the model parameters $\mathbf{w}$ using maximum likelihood estimation (MLE) on a dataset $\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_N, y_N)\} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$. Assume independent and identically distributed samples. Write down the log-likelihood and show that maximizing the log-likelihood is equivalent to minimizing

$$\mathcal{L} = -\sum_i y_i \log \rho_i + (1 - y_i) \log(1 - \rho_i)$$

where each $\rho_i = \sigma(\mathbf{w}^\top \mathbf{x}_i)$. You may recognize this function as the cross entropy loss and here, we demonstrate how this loss emerges from assuming a Bernoulli likelihood.

Problem 4.c. During inspection of your training data, you find that some of the data points are mislabelled! You could look through the data manually to find the mislabelled data but this seems rather labor intensive. Can you adjust your model to account for the wrong labels? Let us introduce a new random variable z which represents the observed (possibly wrong) label. The actual y is now hidden (or "latent"). You know that the variables are related via the conditional distribution,

z	y	$p(z\|y)$
0	0	0.75
0	1	0.05
1	0	0.25
1	1	0.95

Given this information, design a new Bayesian network (*hint:* extend the basic classification model with z) and derive the log-likelihood (*Hint:* recall the sum rule). How is the new MLE optimization function different from the one you derived in the previous subsection?