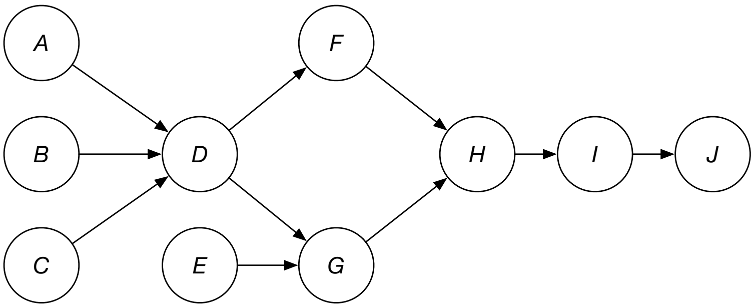
CS5340: Uncertainty Modeling in AI

Tutorial 3: Solutions

Released: Feb. 3, 2022

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?

Solution: No, possibly dependent. Unblocked trail: HIJ

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

Solution: Yes, independent. All trails blocked.

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

Solution: Yes, independent. All trails blocked.

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

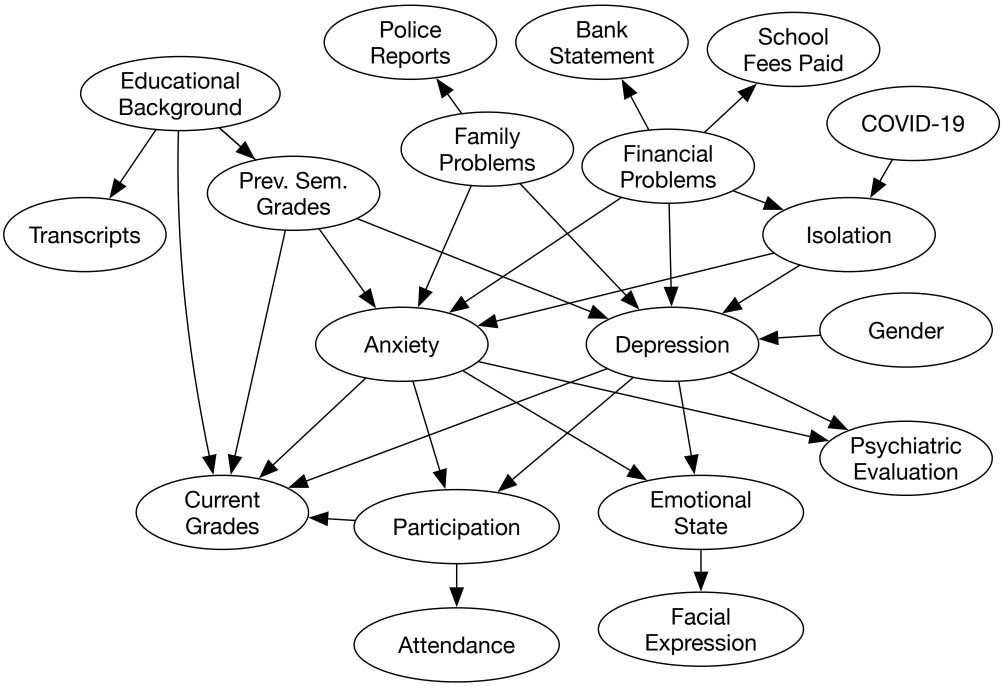
Solution: No, possibly dependent. Unblocked trail: ADGE (since J is observed)

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.

Solution: First, consider the random variables that are involved: the types of mental health problems faced, the causes, and the effects). After coming up with a suitable list, consider how the variables are related. One potential Bayesian Network is the following (it is by no means a complete network or something we recommend you actually implement; we designed it to promote discussion):



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?

Solution: This is a discussion question, which relates to how much you trust that the system would be used appropriately. The system may be used to assist students and the university, by alerting them to possible

[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.

mental health issues. Early support can help students cope with difficulties, which can have long-term benefits.

But there is also a great potential for invasive monitoring using such a system, e.g., to invade student privacy or to shift responsibility to the inference system (e.g., "we banned the student from attending classes because the system said there was a 63% possibility the student would harm themselves or others" or "we didn't help the student because the system said the student did not have anxiety"). Biases and prejudices can creep into the system, either through its design or data it is trained with (e.g., see the ProPublica report on machine bias[2]). There is also a potential for the system to exploited (e.g., via malicious actors who hack the system).

Many systems that result in the greater good for society also have the potential to cause harm. The fact that misuse *could* occur doesn't mean one should not build such systems. Rather, one should consider the trade-offs: the benefits, as well as risk and potential harms. Covering the entirety of ethical decision-making is beyond the scope of CS5340. If you are interested, see this chapter on Ethical Decision Making[3].

[2] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[3] https://us.sagepub.com/sites/default/files/upm-assets/90084_book_item_90084.pdf

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 the final exam (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 final exam 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?*

Solution:

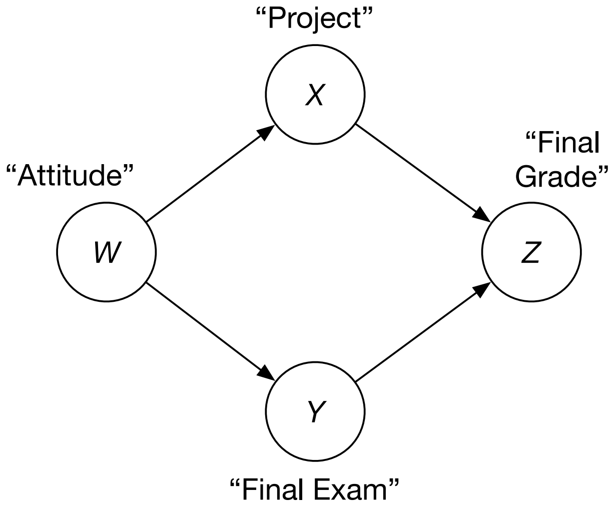


Figure 1: Grade Bayesian Network

We have four Categorical random variables[4]:

- W: Work Attitude. 0 = lazy; 1 = works hard.
- X: Project. 0 = fail; 1= pass.
- Y: Exam. 0 = fail; 1= pass.
- Z: Final grade. 0 = D; 1 = C; 2 = B; 3 = A.

Given the information, we draw the Bayesian network above and the corresponding factorization is: $p(w, x, y, z) = p(z|x, y)p(x|w)p(y|w)p(w)$.

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..*

[4] Equivalent to Bernoulli distributions for W, X, and Y.

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

Solution: Let's derive the general form first and substitute in the observations. The most important thing to realize is that the factorization given by the Bayes net simplifies MLE. From before, we know the joint factorizes as $p(w, x, y, z) = p(z|x, y)p(x|w)p(y|w)p(w)$. Consider a single data point, and we introduce our distribution parameters $\theta = \{\theta_z, \theta_x, \theta_y, \theta_w\}$ and

$$p_\theta(w, x, y, z) = p(z|x, y; \theta_z)p(x|w; \theta_x)p(y|w; \theta_y)p(w; \theta_w) \tag{1}$$

When we perform MLE, we seek to find

$$\text{argmax}_{\theta_w,\theta_x,\theta_y,\theta_z} \log p(z|x, y; \theta_z)p(x|w; \theta_x)p(y|w; \theta_y)p(w; \theta_w) \tag{2}$$

which is equivalent to

$$\text{argmax}_{\theta_w,\theta_x,\theta_y,\theta_z} \log p(z|x, y; \theta_z) + \log p(x|w; \theta_x) + \log p(y|w; \theta_y) + \log p(w; \theta_w). \tag{3}$$

When we are maximizing over a particular variable, we can treat the other terms as a constant and ignore them. For example, when solving for θ_z, we simply have to solve,

$$\text{argmax}_{\theta_z} \log p(z|x, y; \theta_z). \tag{4}$$

and likewise for all the other random variables. As you can see, we no longer have to maximize all the parameters jointly. The remainder of this solution serves to provide details on how to do this for the particular distributions assumed in this problem.

We will introduce some notation to help us keep track of the variables[5]. We use N to denote the number of observations and U to denote the number of variables, i.e. the number of variables in the Bayesian Network. D denotes the entire observation set. The log likelihood of the observations is:

$$\log p(D|\theta) = \sum_{n=1}^{N}\sum_{u=1}^{U} \log p(R_{u,n} = r_{u,n}|R_{\pi_u,n}; \theta_u)$$

where R_u denotes the u^{th} random variable (node), R_{π_u} denotes the *set* of parents variables of R_u and θ_u denotes the parameters related to the u_{th} variable. The additional subscript n indicates the particular datum. When solving for θ_u, we can ignore all terms that do not involves θ_u.

$$\text{argmax}_{\theta_u} \log p(D|\theta) = \sum_{n=1}^{N} \log p(R_{u,n}|R_{\pi_u,n}; \theta_u)$$

Assume that variables follow categorical distribution, we denote parameters related to u_{th} variable, θ_u, as $\lambda_u = \{\lambda_{u11}, ..., \lambda_{uck}, ...\lambda_{uCK}\}$, where C is the total number of states that R_{π_u} takes and K is the total number of states that R_u takes.

[5] We overload/abuse some notation. If anything is unclear, please ask on Piazza.

Define indicator variables: $I_{uck} = 1$ when $R_u = k, R_{\pi_u} = c$ and 0 otherwise.

$$\begin{aligned} p(R_u|R_{\pi_u}, \lambda_u) &= \text{Cat}_{R_u|R_{\pi_u}}[\lambda_u] \\ &= \Pi_{c=1}^{C} \Pi_{k=1}^{K} \lambda_{uck}^{I_{uck}} \\ &s.t. \sum_k \lambda_{uck} = 1 \end{aligned}$$

To perform MLE, we solve the following maximization problem:

$$\underset{\lambda_{uc1} \dots \lambda_{ucK}}{\text{argmax}} \sum_n \sum_c \sum_k \log \lambda_{uck}^{x_{uck,n}}, s.t. \sum_k \lambda_{uck} = 1$$

Drop $\sum_c$ because we need to optimize under each possible c.

$$\underset{\lambda_{uc1} \dots \lambda_{ucK}}{\text{argmax}} \sum_n \sum_k \log \lambda_{uck}^{x_{uck,n}}, s.t. \sum_k \lambda_{uck} = 1$$

$$\underset{\lambda_{uc1} \dots \lambda_{ucK}}{\text{argmax}} \sum_k \log \lambda_{uck}^{x_{N_{uck}}}, s.t. \sum_k \lambda_{uck} = 1$$

where N_{uck} denotes the number of times we observe $(R_u = k, R_{\pi_u} = c)$. Then apply Lagrange multiplier v on the constraint, we get the auxiliary function:

$$L = \sum_k N_{uck} \log \lambda_{uck} + v \left(\sum_k \lambda_{uck} - 1 \right)$$

Take derivatives of L w.r.t λ_{uck}, v and set them to zero.

$$\frac{N_{uck}}{\lambda_{uck}} + v = 0 \qquad \forall k$$

$$\sum_k \lambda_{uck} - 1 = 0$$

$$\hat{\lambda}_{uck} = \frac{N_{uck}}{\sum_{m=1}^{K} N_{ucm}}$$

Now we substitute in the observations and get:

W	
0	1
1/2	1/2

	X	
W	0	1
0	1/2	1/2
1	1/2	1/2

	Y	
W	0	1
0	3/4	1/4
1	1/4	3/4

		Z			
X	Y	0	1	2	3
0	0	1/3	2/3	0	0
1	0	0	1/1	0	0
0	1	0	0	1/1	0
1	1	0	0	2/3	1/3

Problem 4. (Label Errors)

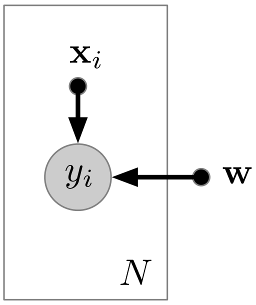
You work as a data scientist for a YouWork! — the hottest startup in town. You are developing a model for predicting a person is likely to be promoted in the coming year (hopefully, your model is more accurate that the Horoscopes on the internet!). Given data point $\mathbf{x} \in \mathbb{R}^d$, your 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.

Solution:



We have conditional distributions for each y_i, i.e.,

$$p(y_i|\mathbf{x}_i, \mathbf{w}) = \text{Bern}[\rho_i] = \rho_i^{y_i}(1 - \rho_i)^{1-y_i}$$

where $\rho_i = \sigma(\mathbf{w}^\top \mathbf{x}_i)$.

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.

Solution:
We wish to maximize the log likelihood

$$\arg\max_{\mathbf{w}} \log \prod_i^N p(y_i|\mathbf{x}_i, \mathbf{w}) \tag{5}$$

So, which is equivalent to minimizing the negative log-likelihood,

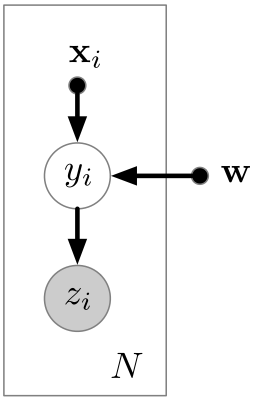
$$\begin{aligned}\mathcal{L}(\mathbf{w}) &= -\log \prod_i^N p(y_i|\mathbf{x}_i, \mathbf{w}) \\ &= -\sum_i^N \log p(y_i|\mathbf{x}_i, \mathbf{w}) \\ &= -\sum_i^N \log \rho_i^{y_i}(1-\rho_i)^{1-y_i} \\ &= -\sum_i^N y_i \log \rho_i + (1-y_i)\log(1-\rho_i)\end{aligned}$$

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?

Solution:



Again, we wish to maximize the log likelihood of the data, so

$$\arg\max_{\mathbf{w}} \log \prod_i^N p(z_i|\mathbf{x}_i, \mathbf{w}) \tag{6}$$

Again, we minimize the negative log-likelihood:

$$\arg\min_{\mathbf{w}} \mathcal{L} \tag{7}$$

where $\mathcal{L} = -\log \prod_i^N p(z_i|\mathbf{x}_i, \mathbf{w})$. To compute $p(z_i|\mathbf{x}_i, \mathbf{w})$, we need to marginalize out the unseen y_i's.

$$p(z_i|\mathbf{x}_i, \mathbf{w}) = \sum_{y_i} p(z_i|y_i)p(y_i|\mathbf{w}, \mathbf{x}_i) \tag{8}$$

Substituting (8) for $p(z_i|\mathbf{x}_i, \mathbf{w})$,

$$\begin{aligned} \mathcal{L}(\mathbf{w}) = -\sum_i^N z_i \log[p(z_i = 1|y_i = 1)\rho_i + p(z_i = 1|y_i = 0)(1 - \rho_i)] + \\ (1 - z_i) \log[p(z_i = 0|y_i = 1)\rho_i + p(z_i = 0|y_i = 0)(1 - \rho_i)] \end{aligned} \tag{9}$$

For this special case, we can substitute in the precise values for the conditional $p(z|y)$,

$$\mathcal{L}(\mathbf{w}) = -\sum_i^N z_i \log[0.95\rho_i + 0.25(1 - \rho_i)] + (1 - z_i) \log[0.05\rho_i + 0.75(1 - \rho_i)]$$