DD2434/FDD3434 Machine Learning, Advanced Course
Module 2 Exercise Solutions

November 2022

1 Directed Graphical Models (DGM)

1.1 Bayes Ball

Question: List all variables that are independent of A given evidence on the shaded node for each of the DGMs a), b) and c) below.
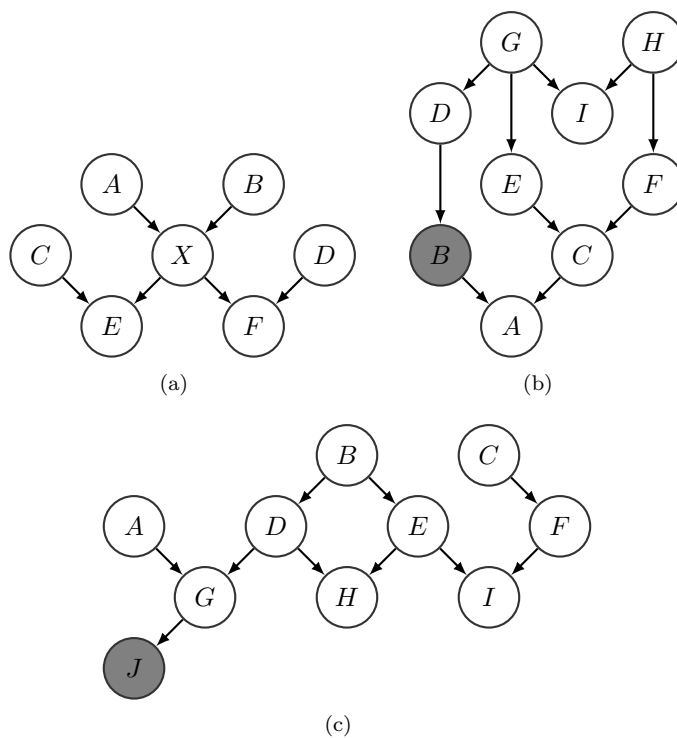

Figure 1: Some DGMs.

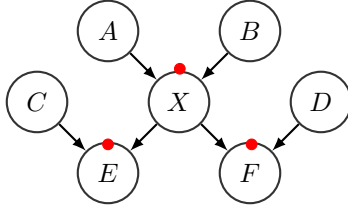**Solution:** a) See Figure 2, b) See Figure 3, c) See Figure 4.

Figure 2: There is no conditioning on any node. Putting the blocks for d-separations, we see that B,C, and D are separated from A or independent of A.
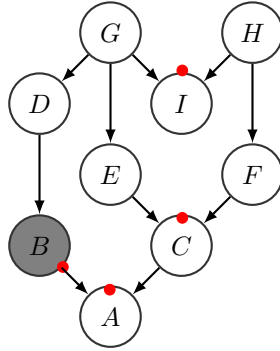

Figure 3: There is a path to A from all variables hence no variable is independent of A.
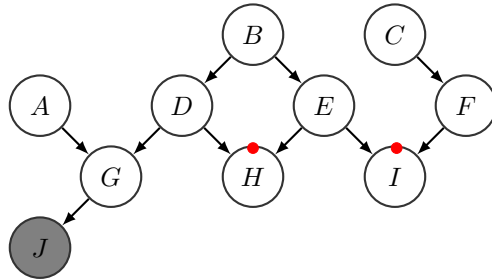

Figure 4: There is a path to A from all variables except F and C. Therefore, only C and F are independent of A.

1.2   PyClone DGM

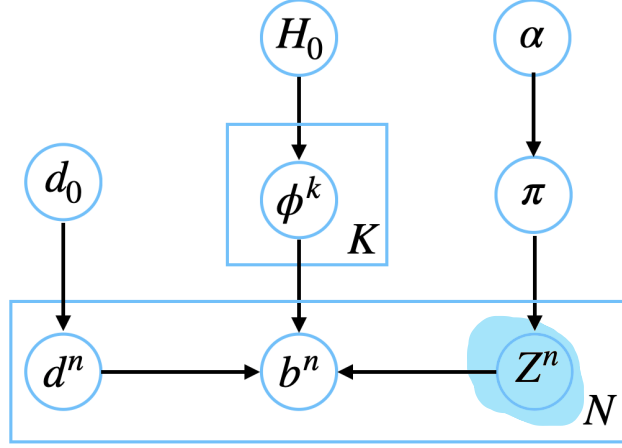Consider the graphical model shown in Figure 5. Answer "yes" or "no" to each question:

Figure 5: Graphical model of PyClone in plate notation

$$\pi \sim Dirichlet(\alpha) \tag{1}$$
$$Z^n \sim Categorical(\pi) \tag{2}$$
$$\phi^k \sim H_0 = Beta(a_0, a_1) \tag{3}$$
$$d^n \sim Poisson(d_0) \tag{4}$$
$$b^n \sim Bin(d^n, \phi^{Z^n}) \tag{5}$$

- $b^n \perp b^{n+1} \mid d^n, d^{n+1}$? no because phi

- $d^n \perp Z^n \mid \alpha, H_0$?   yes bn not given and not descends

- $d^n \perp Z^n \mid b^n$?   no bn given

- $\phi^k \perp d^n \mid d_0, \pi$? yes each all bn not given and not descendants

- $b^{1:N} \perp \pi \mid Z^{1:N}$?

Solution:

- No

- Yes

- No

- Yes

- Yes

3

1.3 Bayes nets for a rainy day (Exercise 10.5 from Murphy [1])

Question: (Source: Nando de Freitas) In this question you must model a problem with 4 binary variables: G = "gray", V = "Vancouver", R = "rain" and S = "sad". Consider the directed graphical model describing the relationship between these variables shown in Figure 6 (and the probability tables shown in Table 1).
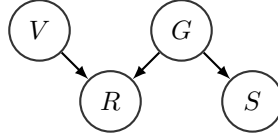

Figure 6: Bayesian net for a rainy day

Table 1: Probability tables of Bayes net for a rainy day

$V = 0$	$V = 1$
δ	$1 - \delta$

$G = 0$	$G = 1$
α	$1 - \alpha$

	$S = 0$	$S = 1$
$G = 0$	γ	$1 - \gamma$
$G = 1$	β	$1 - \beta$

	$R = 0$	$R = 1$
$VG = 00$	0.6	0.4
$VG = 01$	0.3	0.7
$VG = 10$	0.2	0.8
$VG = 11$	0.1	0.9

a. Write down an expression for $P(S = 1|V = 1)$ in terms of $\alpha, \beta, \gamma, \delta$.

b. Write down an expression for $P(S = 1|V = 0)$. Is this the same or different to $P(S = 1|V = 1)$? Explain why.

c. Find maximum likelihood estimates of α, β, γ using the following data set, where each row is a training case. (You may state your answers without proof.)

V	G	R	S
1	1	1	1
1	1	0	1
1	0	0	0

Solution: (The solution is taken from Elin Samuelsson's notes from December 2018 and modified slightly.)

a. Write down an expression for $P(S = 1|V = 1)$ in terms of $\alpha, \beta, \gamma, \delta$.

$$
\begin{aligned}
P(S = 1|V = 1) &= \frac{P(S = 1, V = 1)}{P(V = 1)} \\
&= \frac{1}{P(V = 1)} \sum_{g=0}^{1} \sum_{r=0}^{1} P(S = 1, V = 1, R = r, G = g) \\
&= \frac{1}{P(V = 1)} \sum_{g=0}^{1} \sum_{r=0}^{1} P(S = 1, V = 1, R = r|G = g)P(G = g) \\
&\quad \{G \text{ is tail-to-tail and blocks the path} \to S \perp\!\!\!\perp \{R, V\}|G\} \\
&= \frac{1}{P(V = 1)} \sum_{g=0}^{1} P(S = 1|G = g)P(G = g) \sum_{r=0}^{1} P(V = 1, R = r|G = g) \\
&= \frac{1}{P(V = 1)} \sum_{g=0}^{1} P(S = 1|G = g)P(G = g) \sum_{r=0}^{1} P(R = r|V = 1, G = g)P(V = 1) \\
&= \frac{P(V = 1)}{P(V = 1)} \sum_{g=0}^{1} P(S = 1|G = g)P(G = g) \sum_{r=0}^{1} P(R = r|V = 1, G = g) \\
&= \sum_{g=0}^{1} P(S = 1|G = g)P(G = g) \sum_{r=0}^{1} P(R = r|V = 1, G = g) \\
&\quad \{\sum_{r=0}^{1} P(R = r|V = 1, G = g) = 1, \text{ regardless of the value of } G \text{ and } V\} \\
&= \sum_{g=0}^{1} P(S = 1|G = g)P(G = g) \\
&= P(S = 1|G = 0)P(G = 0) + P(S = 1|G = 1)P(G = 1) \\
&= \alpha(1 - \gamma) + (1 - \alpha)(1 - \beta) \\
&= 1 - \beta + \alpha\beta - \alpha\gamma
\end{aligned}
$$

$$(6)$$

b. Write down an expression for $P(S = 1|V = 0)$. Is this the same or different to $P(S = 1|V = 1)$? Explain why.

$$
\begin{aligned}
P(S = 1|V = 0) &= \sum_{g=0}^{1} P(S = 1|G = g)P(G = g) \sum_{r=0}^{1} P(R = r|V = 0, G = g) \\
&= \sum_{g=0}^{1} P(S = 1|G = g)P(G = g) \\
&= P(S = 1|V = 1)
\end{aligned}
$$

$$(7)$$

since $\sum_{r=0}^{1} P(R = r|V = v, G = g) = 1$, regardless of the value of G and V.

c. Find maximum likelihood estimates of α, β, γ using the following data set, where each row is a training case. (You may state your answers without proof.)

V	G	R	S
1	1	1	1
1	1	0	1
1	0	0	0

Notation:

$$\mathbf{1}(a = b) = \begin{cases} 1, \text{ if } a = b \\ 0, \text{ if } a \neq b \end{cases} \tag{8}$$

Maximum likelihood estimates:

$$\hat{\alpha} = P(G = 0) = \frac{\sum_{n=1}^{N} \mathbf{1}(G_n = 0)}{N} = \frac{1}{3} \tag{9}$$

$$\hat{\beta} = P(S = 0 | G = 1) = \frac{P(S = 0, G = 1)}{P(G = 1)} = \frac{\sum_{n=1}^{N} \mathbf{1}(S_n = 0, G_n = 1)}{\sum_{n=1}^{N} \mathbf{1}(G_n = 1)} = \frac{0}{2} \tag{10}$$

$$\hat{\gamma} = P(S = 0 | G = 0) = \frac{P(S = 0, G = 0)}{P(G = 0)} = \frac{\sum_{n=1}^{N} \mathbf{1}(S_n = 0, G_n = 0)}{\sum_{n=1}^{N} \mathbf{1}(G_n = 0)} = \frac{1}{1} \tag{11}$$

An Alternative Solution

The more general way to find the MLE is i) write the likelihood (or log-likelihood) ii) take derivative w.r.t the parameter of interest and set it to zero iii) check whether the value actually maximizes the likelihood (by looking at the second derivative [2] is negative or not).

The likelihood of the data is:

$$\begin{aligned} \mathcal{L} &= P(D|\Theta) \\ &= P(D_1, \ldots, D_N | \Theta) \\ &= \prod_{n=1}^{N} P(D_n | \Theta) \\ &= \prod_{n=1}^{N} P(V_n | \delta) P(G_n | \alpha) P(S_n | G_n, \beta, \gamma) P(R_n | V_n, G_n) \\ &= P(V_n = 1 | \delta)^3 P(G_n = 0 | \alpha) P(G_n = 1 | \alpha)^2 P(S_n = 0 | G_n = 0, \beta, \gamma) P(S_n = 1 | G_n = 1, \beta, \gamma)^2 \\ &\quad P(R_n = 0 | V_n = 1, G_n = 0) P(R_n = 0 | V_n = 1, G_n = 1) P(R_n = 1 | V_n = 1, G_n = 1) \\ &= (1 - \delta)^3 \alpha (1 - \alpha)^2 \gamma (1 - \beta)^2 \times 0.2 \times 0.1 \times 0.9 \\ &\propto (1 - \delta)^3 \alpha (1 - \alpha)^2 \gamma (1 - \beta)^2 \end{aligned} \tag{12}$$

Let's look at the likelihood (see Figure 7). For the first subplot, I fixed $\alpha, \gamma, \beta \in (0, 1)$ and ranged $\delta \in [0, 1]$. The subplot shows how the likelihood changes w.r.t δ. I repeated the same method for the rest of the parameters. From the figure, we can clearly see which values of the parameters maximize the likelihood ($\hat{\delta} = 0, \hat{\alpha} = 0.\overline{33}, \hat{\gamma} = 1, \hat{\beta} = 0$).
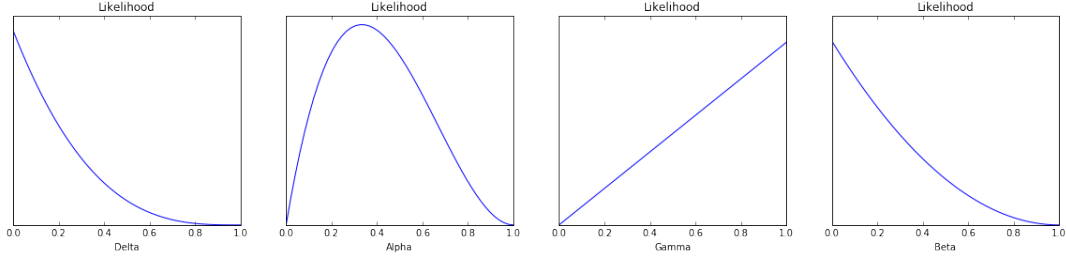
Figure 7: Likelihood of rainy day example with varying parameters

Now, let's show these with the derivatives. First, consider α. Take the first derivative of the likelihood w.r.t α.

$$
\begin{aligned}
\frac{\partial}{\partial \alpha}\mathcal{L} &= \frac{\partial}{\partial \alpha}\alpha(1-\alpha)^2 C \\
&= (1-\alpha)^2 C - 2\alpha(1-\alpha)C \\
&= (1-\alpha)(1-\alpha-2\alpha)C \\
&= (1-\alpha)(1-3\alpha)C \\
&= 0 \\
\alpha &= 1 \text{ or } \alpha = \frac{1}{3}
\end{aligned}
\tag{13}
$$

where C is a temporary variable that I used to represent all the other terms in the likelihood beside the parameter of interest (notice that C is non-negative). There are two α values which either maximize or minimize the likelihood. We need to check the second derivative of the likelihood w.r.t α:

$$
\begin{aligned}
\frac{\partial^2}{\partial \alpha^2}\mathcal{L} &= \frac{\partial}{\partial \alpha}\alpha(1-\alpha)^2 C \\
&= -(1-3\alpha)C - 3(1-\alpha)C \\
&= (-4+6\alpha)C
\end{aligned}
\tag{14}
$$

When $\alpha = 1$, the second derivative becomes $2C$, which is non-negative, which means $\alpha = 1$ is not a maximizer of the likelihood. When $\alpha = \frac{1}{3}$, the second derivative becomes $-2C$, which is negative, which means $\hat{\alpha} = \frac{1}{3}$ is the maximum likelihood estimator. We can confirm this result with Figure 7.

Now, we move on to β. We re-write the likelihood as $\mathcal{L} = (1-\beta)^2 C$. It is clear that the $\hat{\beta}$ which maximizes the likelihood must be $\hat{\beta} = 0$ (since C is non-negative and $\mathcal{L}$ gets the highest value, which is $1C$, when $\beta = 0$). Let's look at take the derivatives.

$$
\begin{aligned}
\frac{\partial}{\partial \beta}\mathcal{L} &= \frac{\partial}{\partial \alpha}(1-\beta)^2 C \\
&= -2(1-\beta)C \\
&= 0 \\
\beta &= 1
\end{aligned}
\tag{15}
$$

Now, check the second derivative:

$$
\begin{aligned}
\frac{\partial^2}{\partial \beta^2} \mathcal{L} &= \frac{\partial}{\partial \alpha}(1-\beta)^2 C \\
&= 2C \\
&> 0
\end{aligned}
\tag{16}
$$

Since the second derivative is always non-negative, $\beta = 1$ is the minimizer of the likelihood. Notice that we were unable to find the β which maximizes the likelihood with this approach. Why? In our data D, we don't have any samples where $S = 0$ and $G = 1$.

Finally, let's re-write the likelihood in terms of γ; $\mathcal{L} = \gamma C$. It is clear that the $\hat{\gamma}$ which maximizes the likelihood must be $\hat{\gamma} = 1$ (because the likelihood is a linearly increasing function of γ).

2 Hidden Markov Models (HMM)

2.1 Forward-Backward Algorithm for Posteriors

Derive forward-backward algorithm for:

- **(a)** the marginal posterior distribution of one hidden variable, i.e, $p(z_n|x_{1:N})$

- **(b)** the joint posterior distribution of two hidden variables, i.e, $p(z_{n-1}, z_n|x_{1:N})$

- **(c)** the posterior predictive distribution, i.e, $p(x_{N+1}|x_{1:N})$

Solutions:
(a) For learning and inference in HMMs (e.g. EM), we need to calculate marginal distribution of the hidden variables, such as $p(z_n|x_{1:N}, \theta)$. So we assume to know the model parameters and we have observed the data. For readability, we write $p(z_n|x_{1:N})$. We could calculate this by performing marginalization over all the hidden variables, but, that is costly (we would have to calculate N nested sum with each iterating over J possible values of the hidden variable i.e. $O(J^N)$). Instead we use an approach which takes $O(NJ^2)$ number of calculations. We write the probability as $\frac{p(x_{1:N}|z_n)p(z_n)}{p(x_{1:N})}$ due to Bayes, and then, because of conditional independence, we can write $p(x_{1:N}|z_n)$ as $p(x_{1:n}|z_n)p(x_{n+1:N}|z_n)$. We have now split the problem into two problems which are solved by recursion (this whole process is referred to as dynamic programming). We can rewrite the problem, finally, as: $\frac{p(x_{1:n}, z_n)p(x_{n+1:N}|z_n)}{p(x_{1:N})}$ or $\frac{\alpha(z_n)\beta(z_n)}{\sum_{z_N} \alpha(z_N)}$. We calculate the probabilities $\alpha(z_n)$ and $\beta(z_n)$, forward and backward, as follows:

Forward pass

$$
\begin{aligned}
\alpha(z_n) = p(x_{1:n}, z_n) &= p(x_{1:n}|z_n)p(z_n) = p(x_n|z_n)p(x_{1:n-1}|z_n)p(z_n) \\
&= p(x_n|z_n)p(x_{1:n-1}, z_n) \\
&= p(x_n|z_n) \sum_{z_{n-1}} p(x_{1:n-1}, z_n, z_{n-1}) \\
&= p(x_n|z_n) \sum_{z_{n-1}} p(x_{1:n-1}, z_n|z_{n-1})p(z_{n-1}) \\
&= p(x_n|z_n) \sum_{z_{n-1}} p(x_{1:n-1}|z_{n-1})p(z_{n-1})p(z_n|z_{n-1}) \\
&= p(x_n|z_n) \sum_{z_{n-1}} p(x_{1:n-1}, z_{n-1})p(z_n|z_{n-1}) = p(x_n|z_n) \sum_{z_{n-1}} \alpha(z_{n-1})p(z_n|z_{n-1}) \\
&, where \quad \alpha(z_1) = p(x_1|z_1)p(z_1)
\end{aligned}
$$

Note that we have $\alpha(z_1)$ from the initializations (we initialize the initial and emission probabilities, thus we have $\alpha(z_1)$).

Backward pass

$$\beta(z_n) = p(x_{n+1:N}|z_n)$$
$$= \sum_{z_{n+1}} p(x_{n+1:N}, z_{n+1}|z_n)$$
$$= \sum_{z_{n+1}} p(x_{n+1:N}|z_{n+1}, z_n)p(z_{n+1}|z_n)$$
$$= \sum_{z_{n+1}} p(x_{n+1:N}|z_{n+1})p(z_{n+1}|z_n)$$
$$= \sum_{z_{n+1}} p(x_{n+1}|z_{n+1})p(x_{n+2:N}|z_{n+1})p(z_{n+1}|z_n)$$
$$= \sum_{z_{n+1}} p(x_{n+1}|z_{n+1})\beta(z_{n+1})p(z_{n+1}|z_n)$$
$$, where \quad \beta(z_N) = 1$$

(b)

$$p(z_{n-1}, z_n|x_{1:N}) = \frac{p(x_{1:N}|z_{n-1}, z_n)p(z_{n-1}, z_n)}{p(x_{1:N})}$$
$$= \frac{p(x_{1:n-1}|z_{n-1})p(x_n|z_n)p(x_{n+1:N}|z_n)p(z_n|z_{n-1})p(z_{n-1}))}{p(x_{1:N})}$$
$$= \frac{\alpha(z_{n-1})p(x_n|z_n)p(z_n|z_{n-1})\beta(z_n)}{\sum_{z_n} \alpha(z_n)\beta(z_n)}$$

(c)

$$p(x_{N+1}|x_{1:N}) = \sum_{z_{N+1}} p(x_{N+1}, z_{N+1}|x_{1:N})$$
$$= \sum_{z_{N+1}} p(x_{N+1}|z_{N+1})P(z_{N+1}|x_{1:N})$$
$$= \sum_{z_{N+1}} p(x_{N+1}|z_{N+1}) \sum_{z_N} P(z_{N+1}, z_N|x_{1:N})$$
$$= \sum_{z_{N+1}} p(x_{N+1}|z_{N+1}) \sum_{z_N} P(z_{N+1}|z_N)p(z_N|x_{1:N})$$
$$= \sum_{z_{N+1}} p(x_{N+1}|z_{N+1}) \sum_{z_N} P(z_{N+1}|z_N)\frac{p(z_N, x1:N)}{p(x_{1:N})}$$
$$= \frac{1}{p(x_{1:N})} \sum_{z_{N+1}} p(x_{N+1}|z_{N+1}) \sum_{z_N} P(z_{N+1}|z_N)\alpha(z_N)$$

References

[1]   Kevin P Murphy. *Machine learning: a probabilistic perspective.* MIT press, 2012.

[2]   E Weisstein. *Second derivative test. From MathWorld–A Wolfram Web Resource.* 2004.