CS 486/686 Assignment 1
Spring 2024
(90 marks)

Wenhu Chen

Due Date:

1 Reconstructing Bayesian Network (marks)

Given four boolean random variables A, B, C, D, there are some underlying (conditional) independence between these variables. You are supposed to identify these (conditional) independence between variables, and then reconstruct the **most compact Bayesian Network** (minimum number probabilities to encode the network). The joint probability distribution is in Table 1.

A	B	C	D	Prob
T	T	T	T	0.0080
T	T	T	F	0.0120
T	T	F	T	0.0080
T	T	F	F	0.0120
T	F	T	T	0.0576
T	F	T	F	0.0144
T	F	F	T	0.1344
T	F	F	F	0.0336
F	T	T	T	0.0720
F	T	T	F	0.1080
F	T	F	T	0.0720
F	T	F	F	0.1080
F	F	T	T	0.0864
F	F	T	F	0.0216
F	F	F	T	0.2016
F	F	F	F	0.0504

TABLE 1: Joint probability distribution of $p(A, B, C, D)$.

(a) Is variable A independent from variable D?

Marking Scheme: (marks)

- (marks) Correct calculation

Solutions: No. Show that $P(A, D) \neq P(A) \times P(D)$, there are four cases to show their inequalities.

(b) Given A, are the random variable C and D independent from each other?

Solutions: No. Show that $P(C, D|A) \neq P(C|A) \times P(D|A)$, there are eight cases to show their inequalities.

Marking Scheme: (marks)

- (marks) Correct calculation

(c) Given B, is the random variable A independent from C?

Marking Scheme: (marks)

- (marks) Correct calculation

Solutions: Yes. Show that $P(A, C|B) = P(A|B) \times P(C|B)$, there are eight cases to show their equalities.

(d) Given B, is the random variable A independent from D?

Marking Scheme: (marks)

- (marks) Correct calculation

Solutions: Yes. Show that $P(A, D|B) = P(A|B) \times P(D|B)$, there are eight cases to show their equalities.

(e) Given B, is the random variable C independent from D?

Solutions: Yes. Show that $P(C, D|B) = P(C|B) \times P(D|B)$, there are eight cases to show their equalities.

Marking Scheme: (marks)

- (marks) Correct calculation

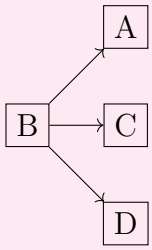
(f) Show the Bayesian Network if we construct it in the order of B, A, C, D. How many probabilities do we need to represent this constructed Bayesian Network?

Marking Scheme: (marks)

- (marks) Correct construction.

- (marks) Correctly compute the number of probabilities needed.

Solutions:

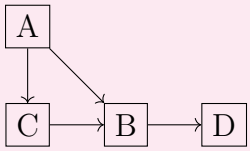


(g) Show the Bayesian Network if we construct it in the order of A, C, B, D. How many probabilities do we need to represent this constructed Bayesian Network?

Marking Scheme: (marks)

- (marks) Correct Construction.
- (marks) Correctly compute the number of probabilities needed.

Solutions:

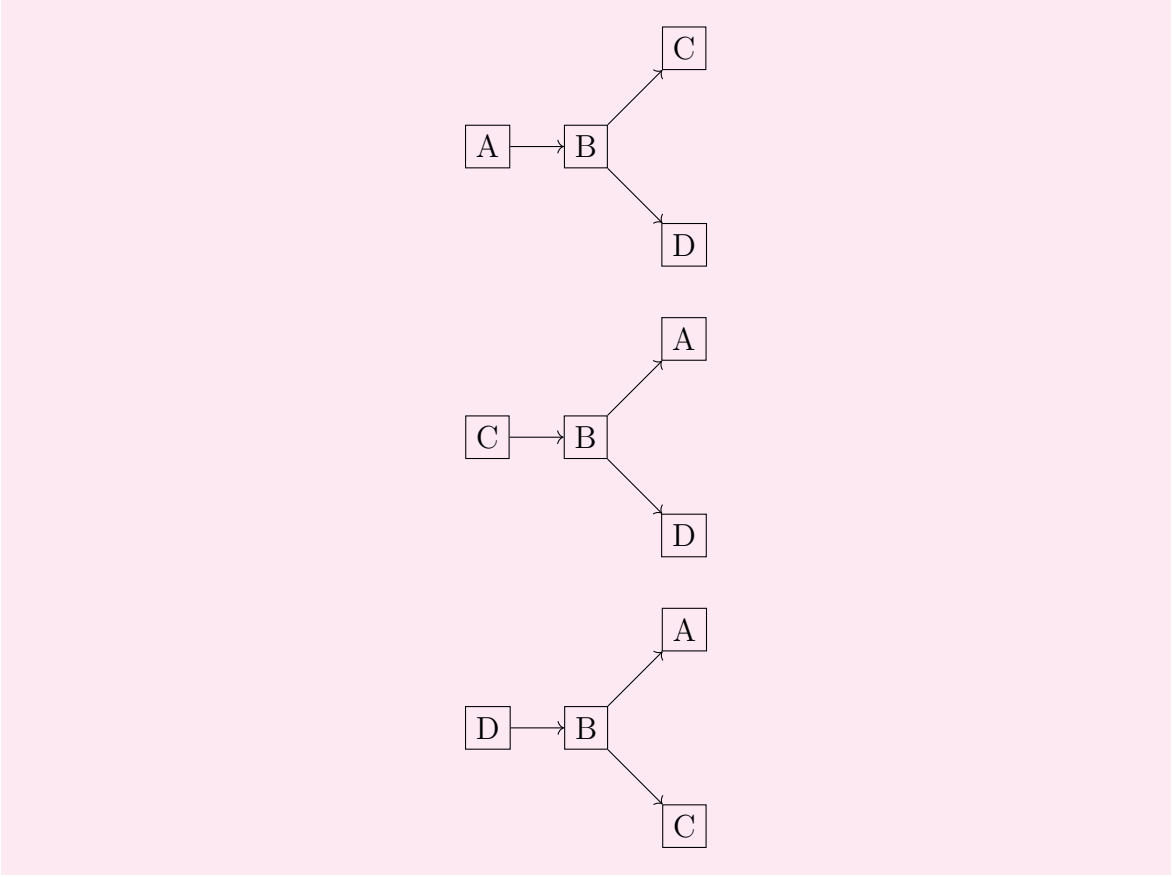


(h) We have already derived two Bayesian Networks so far based on the given order. Are they the most compact Bayesian Networks, i.e., having the smallest amount of probabilities? Is there any other Bayesian Network that is most compact?

Marking Scheme: (marks)

- (marks) Correctly Answer the question

Solutions: Q(f) is compact, Q(g) is not compact.



2 Solutions of Q2

2.1 Q2.1

Considering the following weights and network's structure:

$$W^{(1)} = \begin{bmatrix} 0.2 & -0.3 \\ 0.4 & 0.1 \end{bmatrix}, \quad W^{(2)} = \begin{bmatrix} 0.7 & 0.5 \\ -0.6 & 0.2 \end{bmatrix}$$

we have:
1. $\mathbf{a}^{(1)}$ is:

$$\mathbf{a}^{(1)} = \mathbf{x}W^{(1)} = [0.5, 1.0] \begin{bmatrix} 0.2 & -0.3 \\ 0.4 & 0.1 \end{bmatrix} = [0.5, -0.05]$$

2. $\mathbf{z}^{(1)}$ is:

$$\mathbf{z}^{(1)} = g(\mathbf{a}^{(1)}) + \mathbf{x} = sig([0.5, -0.05]) + [0.5, 1.0] = [0.62, 0.49] + [0.5, 1.0] = [1.12, 1.49]$$

3. $\mathbf{a}^{(2)}$ is:

$$\mathbf{a}^{(2)} = \mathbf{z1}W^{(2)} = [1.12, 1.49] \begin{bmatrix} 0.7 & 0.5 \\ -0.6 & 0.2 \end{bmatrix} = [-0.11, 0.86]$$

4. $\mathbf{z}^{(2)}$ is:

$$\mathbf{z}^{(2)} = g(\mathbf{a}^{(2)}) + \mathbf{z}^{(1)} = sig([-0.11, 0.86]) + [1.12, 1.49] = [0.47, 0.70] + [1.12, 1.49] = [1.59, 2.19]$$

2.2 Q2.2

$$\begin{aligned} E &= \frac{1}{2}\|z^{(2)} - y\|_2^2 = \frac{1}{2}\sum_{i=1}^{2}(z_i^{(2)} - y_i)^2 \\ &= \frac{1}{2}\left((1.59 - 1)^2 + (2.19 - 0)^2\right) = 2.58 \end{aligned}$$

2.3 Q2.3

1. Calculating $\frac{\partial E}{\partial W^{(2)}}$

$$\begin{aligned} \frac{\partial E}{\partial \mathbf{w}^{(2)}} &= \frac{\partial E}{\partial \mathbf{z}^{(2)}} \frac{\partial \mathbf{z}^{(2)}}{\partial \mathbf{w}^{(2)}} \\ &= \frac{\partial E}{\partial \mathbf{z}^{(2)}} \frac{\partial (g(\mathbf{a}^{(2)}) + \mathbf{z}^{(1)})}{\partial \mathbf{w}^{(2)}} \\ &= \frac{\partial E}{\partial \mathbf{z}^{(2)}} \frac{\partial g(\mathbf{a}^{(2)})}{\partial \mathbf{w}^{(2)}} \\ &= \frac{\partial E}{\partial \mathbf{z}^{(2)}} \frac{\partial g(\mathbf{a}^{(2)})}{\partial \mathbf{a}^{(2)}} \frac{\partial \mathbf{a}^{(2)}}{\partial \mathbf{w}^{(2)}} \\ &= \mathbf{z}^{(1)T} \left[(\mathbf{z}^{(2)} - \mathbf{y}) \operatorname{diag}(g(\mathbf{a}^{(2)})(1 - g(\mathbf{a}^{(2)})))\right] \end{aligned}$$

$$= \begin{bmatrix} 1.12 \\ 1.49 \end{bmatrix} \begin{bmatrix} 0.59 & 2.19 \end{bmatrix} \begin{bmatrix} 0.25 & 0 \\ 0 & 0.21 \end{bmatrix}$$

$$= \begin{bmatrix} 0.16 & 0.51 \\ 0.22 & 0.68 \end{bmatrix}$$

2. Calculating $\frac{\partial E}{\partial W^{(1)}}$

$$\begin{aligned}
\frac{\partial E}{\partial \mathbf{w}^{(1)}} &= \frac{\partial E}{\partial \mathbf{z}^{(2)}} \frac{\partial \mathbf{z}^{(2)}}{\partial \mathbf{w}^{(1)}} \\
&= \frac{\partial E}{\partial \mathbf{z}^{(2)}} \frac{\partial (g(\mathbf{a}^{(2)}) + \mathbf{z}^{(1)})}{\partial \mathbf{w}^{(1)}} \\
&= \frac{\partial E}{\partial \mathbf{z}^{(2)}} \left(\frac{\partial g(\mathbf{a}^{(2)})}{\partial \mathbf{w}^{(1)}} + \frac{\partial \mathbf{z}^{(1)}}{\partial \mathbf{w}^{(1)}} \right) \\
&= \frac{\partial E}{\partial \mathbf{z}^{(2)}} \left(\frac{\partial g(\mathbf{a}^{(2)})}{\partial \mathbf{a}^{(2)}} \frac{\partial \mathbf{a}^{(2)}}{\partial \mathbf{w}^{(1)}} + \frac{\partial \mathbf{z}^{(1)}}{\partial \mathbf{w}^{(1)}} \right) \\
&= \frac{\partial E}{\partial \mathbf{z}^{(2)}} \left(\frac{\partial g(\mathbf{a}^{(2)})}{\partial \mathbf{a}^{(2)}} \frac{\partial \mathbf{a}^{(2)}}{\partial \mathbf{z}^{(1)}} \frac{\partial \mathbf{z}^{(1)}}{\partial \mathbf{w}^{(1)}} + \frac{\partial \mathbf{z}^{(1)}}{\partial \mathbf{w}^{(1)}} \right) \\
&= \frac{\partial E}{\partial \mathbf{z}^{(2)}} \left(\text{diag}(g(\mathbf{a}^{(2)}) \odot (1 - g(\mathbf{a}^{(2)}))) \mathbf{w}^{(2)T} + I \right) \frac{\partial \mathbf{z}^{(1)}}{\partial \mathbf{w}^{(1)}} \\
&= \frac{\partial E}{\partial \mathbf{z}^{(2)}} \left(\text{diag}(g(\mathbf{a}^{(2)}) \odot (1 - g(\mathbf{a}^{(2)}))) \mathbf{w}^{(2)T} + I \right) \frac{\partial (g(\mathbf{a}^{(1)}) + \mathbf{x})}{\partial \mathbf{w}^{(1)}} \\
&= \frac{\partial E}{\partial \mathbf{z}^{(2)}} \left(\text{diag}(g(\mathbf{a}^{(2)}) \odot (1 - g(\mathbf{a}^{(2)}))) \mathbf{w}^{(2)T} + I \right) \frac{\partial g(\mathbf{a}^{(1)})}{\partial \mathbf{a}^{(1)}} \frac{\partial \mathbf{a}^{(1)}}{\partial \mathbf{w}^{(1)}} \\
&= \mathbf{x}^T (\mathbf{z}^{(2)} - \mathbf{y}) (\text{diag}(g(\mathbf{a}^{(2)}) \odot (1 - g(\mathbf{a}^{(2)}))) \mathbf{w}^{(2)T} + I) \\
&\quad \cdot \text{diag}(g(\mathbf{a}^{(1)}) \odot (1 - g(\mathbf{a}^{(1)})))
\end{aligned}$$

$$\begin{bmatrix} 0.5 \\ 1 \end{bmatrix} \begin{bmatrix} 0.59 & 2.19 \end{bmatrix} \begin{bmatrix} 0.25 & 0 \\ 0 & 0.21 \end{bmatrix} \left(\begin{bmatrix} 0.7 & -0.6 \\ 0.5 & 0.2 \end{bmatrix} + I \right) \begin{bmatrix} 0.23 & 0 \\ 0 & 0.25 \end{bmatrix}$$

$$= \begin{bmatrix} 0.11 & 0.27 \\ 0.22 & 0.55 \end{bmatrix}$$

2.4 Q2.4

$$
\begin{aligned}
\mathbf{W}^{(i),\text{new}} &= \mathbf{W}^{(i),\text{old}} - \eta \frac{\partial E}{\partial \mathbf{w}^{(i)}} \\
\Rightarrow \mathbf{W}^{(1)} &= \begin{bmatrix} 0.2 & -0.3 \\ 0.4 & 0.1 \end{bmatrix} - 0.1 \begin{bmatrix} 0.11 & 0.27 \\ 0.22 & 0.55 \end{bmatrix} \\
&= \begin{bmatrix} 0.19 & -0.33 \\ 0.38 & 0.05 \end{bmatrix} \\
\mathbf{W}^{(2)} &= \begin{bmatrix} 0.7 & 0.5 \\ -0.6 & 0.2 \end{bmatrix} - 0.1 \begin{bmatrix} 0.16 & 0.51 \\ 0.22 & 0.68 \end{bmatrix} \\
&= \begin{bmatrix} 0.68 & 0.45 \\ -0.62 & 0.13 \end{bmatrix}
\end{aligned}
$$

2.5 Q2.5

Here, the activation function is ReLU. So we have:

$$
g(x) = \begin{cases} x & \text{if } x > 0 \\ 0 & \text{if } x \leq 0 \end{cases} \quad \Rightarrow \quad g'(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x \leq 0 \end{cases}
$$

1. Forward pass:

$$
\mathbf{a}^{(1)} = \mathbf{x}\mathbf{W}^{(1)} = [0.5, 1] \begin{bmatrix} 0.2 & -0.3 \\ 0.4 & 0.1 \end{bmatrix} = [0.5, -0.05]
$$

$$
\mathbf{z}^{(1)} = g(\mathbf{a}^{(1)}) + \mathbf{x} = [0.5, 0] + [0.5, 1] = [1, 1]
$$

$$
\mathbf{a}^{(2)} = \mathbf{z}^{(1)}\mathbf{W}^{(2)} = [1, 1] \begin{bmatrix} 0.7 & 0.5 \\ -0.6 & 0.2 \end{bmatrix} = [0.1, 0.7]
$$

$$
\mathbf{z}^{(2)} = g(\mathbf{a}^{(2)}) + \mathbf{x} = [0.1, 0.7] + [0.5, 1] = [1.1, 1.7]
$$

2. Calculate Error:

$$E = \frac{1}{2}\sum_{i=1}^{2}(z_i^{(2)} - y_i)^2 = \frac{1}{2}\left[(1.1-1)^2 + (1.7-0)^2\right] = 1.45$$

3. Calculating $\frac{\partial E}{\partial W^{(2)}}$

$$\begin{aligned}
\frac{\partial E}{\partial \mathbf{w}^{(2)}} &= \frac{\partial E}{\partial \mathbf{z}^{(2)}}\frac{\partial \mathbf{z}^{(2)}}{\partial \mathbf{w}^{(2)}} \\
&= \frac{\partial E}{\partial \mathbf{z}^{(2)}}\frac{\partial (g(\mathbf{a}^{(2)}) + \mathbf{z}^{(1)})}{\partial \mathbf{w}^{(2)}} \\
&= \frac{\partial E}{\partial \mathbf{z}^{(2)}}\frac{\partial g(\mathbf{a}^{(2)})}{\partial \mathbf{w}^{(2)}} \\
&= \frac{\partial E}{\partial \mathbf{z}^{(2)}}\frac{\partial g(\mathbf{a}^{(2)})}{\partial \mathbf{a}^{(2)}}\frac{\partial \mathbf{a}^{(2)}}{\partial \mathbf{w}^{(2)}} \\
&= \mathbf{z}^{(1)T}\left[(\mathbf{z}^{(2)} - \mathbf{y})\,\mathrm{diag}(g'(\mathbf{a}^{(2)}))\right]
\end{aligned}$$

$$= \begin{bmatrix} 1 \\ 1 \end{bmatrix} \begin{bmatrix} 0.1 & 1.7 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 0.1 & 1.7 \\ 0.1 & 1.7 \end{bmatrix}$$

4. Calculating $\frac{\partial E}{\partial W^{(1)}}$

$$
\begin{aligned}
\frac{\partial E}{\partial \mathbf{w}^{(1)}} &= \frac{\partial E}{\partial \mathbf{z}^{(2)}} \frac{\partial \mathbf{z}^{(2)}}{\partial \mathbf{w}^{(1)}} \\
&= \frac{\partial E}{\partial \mathbf{z}^{(2)}} \frac{\partial (g(\mathbf{a}^{(2)}) + \mathbf{z}^{(1)})}{\partial \mathbf{w}^{(1)}} \\
&= \frac{\partial E}{\partial \mathbf{z}^{(2)}} \left(\frac{\partial g(\mathbf{a}^{(2)})}{\partial \mathbf{w}^{(1)}} + \frac{\partial \mathbf{z}^{(1)}}{\partial \mathbf{w}^{(1)}} \right) \\
&= \frac{\partial E}{\partial \mathbf{z}^{(2)}} \left(\frac{\partial g(\mathbf{a}^{(2)})}{\partial \mathbf{a}^{(2)}} \frac{\partial \mathbf{a}^{(2)}}{\partial \mathbf{w}^{(1)}} + \frac{\partial \mathbf{z}^{(1)}}{\partial \mathbf{w}^{(1)}} \right) \\
&= \frac{\partial E}{\partial \mathbf{z}^{(2)}} \left(\frac{\partial g(\mathbf{a}^{(2)})}{\partial \mathbf{a}^{(2)}} \frac{\partial \mathbf{a}^{(2)}}{\partial \mathbf{z}^{(1)}} \frac{\partial \mathbf{z}^{(1)}}{\partial \mathbf{w}^{(1)}} + \frac{\partial \mathbf{z}^{(1)}}{\partial \mathbf{w}^{(1)}} \right) \\
&= \frac{\partial E}{\partial \mathbf{z}^{(2)}} \left(\text{diag}(g'(\mathbf{a}^{(2)})) \mathbf{w}^{(2)T} + I \right) \frac{\partial \mathbf{z}^{(1)}}{\partial \mathbf{w}^{(1)}} \\
&= \frac{\partial E}{\partial \mathbf{z}^{(2)}} \left(\text{diag}(g'(\mathbf{a}^{(2)})) \mathbf{w}^{(2)T} + I \right) \frac{\partial (g(\mathbf{a}^{(1)}) + \mathbf{x})}{\partial \mathbf{w}^{(1)}} \\
&= \frac{\partial E}{\partial \mathbf{z}^{(2)}} \left(\text{diag}(g'(\mathbf{a}^{(2)})) \mathbf{w}^{(2)T} + I \right) \frac{\partial g(\mathbf{a}^{(1)})}{\partial \mathbf{a}^{(1)}} \frac{\partial \mathbf{a}^{(1)}}{\partial \mathbf{w}^{(1)}} \\
&= \mathbf{x}^T (\mathbf{z}^{(2)} - \mathbf{y}) (\text{diag}(g'(\mathbf{a}^{(2)})) \mathbf{w}^{(2)T} + I) \cdot \text{diag}(g'(\mathbf{a}^{(1)}))
\end{aligned}
$$

$$
\begin{bmatrix} 0.5 \\ 1 \end{bmatrix} \begin{bmatrix} 0.1 & 1.7 \end{bmatrix} \left(\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0.7 & -0.6 \\ 0.5 & 0.2 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right) \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}
$$

$$
= \begin{bmatrix} 0.51 & 0 \\ 1.02 & 0 \end{bmatrix}
$$

2.6 Q2.6

(a)

The learning rate controls the step size in the gradient descent algorithm.

- A small learning rate can cause:
  - Slow convergence
  - Getting stuck in local minima
- A large learning rate can cause:

- Overshooting the minimum
- Oscillating around the solution

(b)

In neural networks, the vanishing gradient problem occurs when the gradient of the loss function with respect to the weights becomes very small during backpropagation, especially in earlier layers. This issue often arises due to certain activation functions, like the sigmoid, which cause the gradients to decay after each layer. As a result, learning effectively stops in these earlier layers. Skip connections, also known as residual connections, help mitigate this problem by providing alternative shortcut paths for the gradients. These paths allow gradients to flow more easily, ensuring that earlier layers receive sufficient updates during training. Additionally, skip connections help preserve useful features across layers, improving overall model performance and stability.