# CSE 575: Homework #3

Due: November 4, 2022

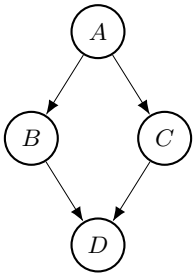


Figure 1: Bayesian network

## Problem 1

Consider the Bayesian network structure given in Figure 1.

a) (6pt) Use the Markovian assumption on each variable to derive independence statements described by this Bayesian network.

**Solution:** Using the Markovian assumption on each variable:

- A: *none*
- B: $B \perp C \mid A$
- C: $C \perp B \mid A$
- D: $D \perp A \mid B, C$

Thus, this Bayesian network implies these two conditional independencies:

$$B \perp C \mid A, \quad D \perp A \mid B, C$$

b) (6pt) For each variable, **(i)** identify its Markov blanket and **(ii)** state any independence that can be derived using the Markov blanket.

| $A$ | $B$ | $C$ | $D$ | $P(A,B,C,D)$ |
|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 1/16 |
| 0 | 0 | 0 | 1 | 1/16 |
| 0 | 0 | 1 | 0 | 1/16 |
| 0 | 0 | 1 | 1 | 1/16 |
| 0 | 1 | 0 | 0 | 1/32 |
| 0 | 1 | 0 | 1 | 1/32 |
| 0 | 1 | 1 | 0 | 3/32 |
| 0 | 1 | 1 | 1 | 3/32 |
| 1 | 0 | 0 | 0 | 5/64 |
| 1 | 0 | 0 | 1 | 7/64 |
| 1 | 0 | 1 | 0 | 5/64 |
| 1 | 0 | 1 | 1 | 7/64 |
| 1 | 1 | 0 | 0 | 1/32 |
| 1 | 1 | 0 | 1 | 1/32 |
| 1 | 1 | 1 | 0 | 1/32 |
| 1 | 1 | 1 | 1 | 1/32 |

Table 1: Joint distribution table

**Solution:** Markov blanket for each variable:

- A: $\{B, C\}$
- B: $\{A, C, D\}$
- C: $\{A, B, D\}$
- D: $\{B, C\}$

Independence that can be derived: $A \perp D \mid B, C$

c) (6pt) Write down the factorization of the joint probability distribution over $A, B, C, D$ that corresponds to this structure.

**Solution:** Order children before parents and apply chain rule of probability.

$$P(A,B,C,D) = P(D \mid C,B,A)P(C \mid B,A)P(B \mid A)P(A) = P(D \mid B,C)P(C \mid A)P(B \mid A)P(A)$$

d) (6pt) Consider the joint probability table in Table 1. Is this a valid distribution for the Bayesian network in Figure 1? Explain your answer.

**Solution:** No, the joint probability distribution does not satisfy the independence $B \perp C \mid A$ implied by the Bayesian network structure. For example,

$$P(C = 1 \mid A = 0) = \frac{1/16 + 1/16 + 3/32 + 3/32}{1/2} = \frac{5}{8}$$

$$P(C = 1 \mid A = 0, B = 1) = \frac{3/32 + 3/32}{1/4} = \frac{3}{4}$$

In fact, it does not satisfy $D \perp A \mid B, C$ either.

| $A$ | $B$ | $C$ | $D$ |
|---|---|---|---|
| 0 | 1 | 1 | 0 |
| 1 | 1 | 1 | 1 |
| 1 | 0 | 0 | 0 |
| 0 | 1 | 1 | 1 |
| 1 | 1 | 0 | 1 |
| 0 | 0 | 0 | 0 |
| 0 | 1 | 0 | 1 |
| 0 | 0 | 1 | 1 |

Table 2: Training data

# Problem 2

Again consider the Bayesian network structure in Figure 1 and the associated conditional probability tables (CPTs) below. Assume all variables are binary.

| $P(A=1)$ |
|---|
| $\theta_A$ |

| $A$ | $P(B=1\|A)$ |
|---|---|
| 0 | $\theta_{B\|0}$ |
| 1 | $\theta_{B\|1}$ |

| $A$ | $P(C=1\|A)$ |
|---|---|
| 0 | $\theta_{C\|0}$ |
| 1 | $\theta_{C\|1}$ |

| $B$ | $C$ | $P(D=1\|B,C)$ |
|---|---|---|
| 0 | 0 | $\theta_{D\|0,0}$ |
| 0 | 1 | $\theta_{D\|0,1}$ |
| 1 | 0 | $\theta_{D\|1,0}$ |
| 1 | 1 | $\theta_{D\|1,1}$ |

a) (6pt) Write down the expression for the probability $P(A = 1 \mid B = 1, C = 0)$ in terms of the parameters shown in the above CPTs.

**Solution:**

$$\frac{\theta_A \theta_{B|1}(1-\theta_{C|1})}{\theta_A \theta_{B|1}(1-\theta_{C|1}) + (1-\theta_A)\theta_{B|0}(1-\theta_{C|0})}$$

b) (10pt) Suppose you are given the training data in Table 2. What is the log-likelihood function of above Bayesian network given this data? Note: it should be a function of the parameters shown in the above CPTs.

**Solution:** Log-likelihood $\mathrm{LL}(\theta)$:

$$\log \prod_{n=1}^{8} P(A_n, B_n, C_n, D_n) = \sum_{n=1}^{8} \log(P(A_n)P(B_n \mid A_n)P(C_n \mid A_n)P(D_n \mid B_n, C_n))$$
$$= 3\log\theta_A + 5\log(1-\theta_A) + 3\log\theta_{B|0} + 2\log(1-\theta_{B|0}) + 2\log\theta_{B|1} + 1\log(1-\theta_{B|1})$$
$$+ 3\log\theta_{C|0} + 2\log(1-\theta_{C|0}) + 1\log\theta_{C|1} + 2\log(1-\theta_{C|1})$$
$$+ 2\log(1-\theta_{D|0,0}) + 1\log\theta_{D|0,1} + 2\log\theta_{D|1,0} + 2\log\theta_{D|1,1} + 1\log(1-\theta_{D|1,1})$$

c) (10pt) Optimize the log-likelihood function from part (b) to obtain the maximum-likelihood parameters for this Bayesian network. Please show your work.

Problem 2 continued on next page...

**Solution:**

$$\frac{\partial \text{LL}}{\partial \theta_A} = \frac{3}{\theta_A} - \frac{5}{1-\theta_A} = 0 \quad \Rightarrow \theta_A = \frac{3}{8}$$
$$\frac{\partial \text{LL}}{\partial \theta_{B|0}} = \frac{3}{\theta_{B|0}} - \frac{2}{1-\theta_{B|0}} = 0 \quad \Rightarrow \theta_{B|0} = \frac{3}{5}$$
$$\frac{\partial \text{LL}}{\partial \theta_{B|1}} = \frac{2}{\theta_{B|1}} - \frac{1}{1-\theta_{B|1}} = 0 \quad \Rightarrow \theta_{B|1} = \frac{2}{3}$$
$$\frac{\partial \text{LL}}{\partial \theta_{C|0}} = \frac{3}{\theta_{C|0}} - \frac{2}{1-\theta_{C|0}} = 0 \quad \Rightarrow \theta_{C|0} = \frac{3}{5}$$
$$\frac{\partial \text{LL}}{\partial \theta_{C|1}} = \frac{1}{\theta_{C|1}} - \frac{2}{1-\theta_{C|1}} = 0 \quad \Rightarrow \theta_{C|1} = \frac{1}{3}$$
$$\frac{\partial \text{LL}}{\partial \theta_{D|1,1}} = \frac{2}{\theta_{D|1,1}} - \frac{1}{1-\theta_{D|1,1}} = 0 \quad \Rightarrow \theta_{D|1,1} = \frac{2}{3}$$

Be careful with the following cases:

$$\frac{\partial \text{LL}}{\partial \theta_{D|0,0}} = -\frac{2}{1-\theta_{D|0,0}} = 0$$
$$\frac{\partial \text{LL}}{\partial \theta_{D|0,1}} = \frac{1}{\theta_{D|0,1}} = 0$$
$$\frac{\partial \text{LL}}{\partial \theta_{D|1,0}} = \frac{2}{\theta_{D|1,0}} = 0$$

Note that there are no values of $\theta_{D|0,0}, \theta_{D|0,1}, \theta_{D|1,0}$ to set above partial derivatives to zero. However, we can use the fact that the parameters must be between 0 and 1 and that the log-likelihood is monotonically decreasing with respect to $\theta_{D|0,0}$ and increasing with respect to $\theta_{D|0,1}, \theta_{D|1,0}$. Thus, the maximum-likelihood estimates are: $\theta_{D|0,0} = 0, \theta_{D|0,1} = 1, \theta_{D|1,0} = 1$.

d) (6pt) Consider an alternative Bayesian network structure shown in Figure 2. Would the likelihood achieved by this BN be at least that of the BN from part (c), or would it be smaller? Justify your answer.

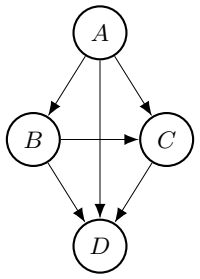


Figure 2: Bayesian network

**Solution:** It will achieve likelihood at least that of the BN from part (c). This is a fully-connected Bayesian network structure, which does not impose any independence. Thus, it can model any joint probability distribution over variables $A, B, C, D$, including the distribution given by maximum-likelihood parameters in part (c).

# Problem 3

Recall that a Gaussian mixture model (GMM) defines a probability density function as follows:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x} \mid \mu_k, \mathbf{\Sigma}_k),$$

where $\pi_k$ are the mixing coefficients satisfying $\pi_k \geq 0, \sum_{k=1}^{K} \pi_k = 1$, and $\mathcal{N}(\cdot \mid \mu_k, \mathbf{\Sigma}_k)$ denotes a multivariate Gaussian with mean $\mu_k$ and covariance $\mathbf{\Sigma}_k$.

Consider a simplified GMM whose covariance matrices are diagonal. In other words, for $k \in \{1, \ldots, K\}$, the covariance matrix $\mathbf{\Sigma}_k$ can be expressed as:

$$\mathbf{\Sigma}_k = \begin{bmatrix} \sigma_{k1}^2 & & & \\ & \sigma_{k2}^2 & & \\ & & \ddots & \\ & & & \sigma_{kD}^2 \end{bmatrix}$$

where $D$ is the dimension of $\mathbf{x}$. Moreover, denote each entry of $\mu_k$ as $\mu_{ki}$ for $i \in \{1, \ldots, D\}$. That is, $\mu_k = (\mu_{k1}, \ldots, \mu_{kD})^T$.

a) (10pt) Show that the density function of such simplified GMM can be expressed in the following form:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \prod_{i=1}^{D} f_{ki}(\mathbf{x}).$$

Derive an expression for $f_{ki}(\mathbf{x})$ in terms of $\mathbf{x}, \mu_{ki}, \sigma_{ki}^2$.

**Solution:**

$$\begin{aligned} p(x) &= \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x} \mid \mu_k, \mathbf{\Sigma}_k) \\ &= \sum_{k=1}^{K} \pi_k (2\pi)^{-\frac{D}{2}} |\mathbf{\Sigma}_k|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} (\mathbf{x} - \mu_k)^T \mathbf{\Sigma}^{-1} (\mathbf{x} - \mu_k) \right\} \\ &= \sum_{k=1}^{K} \pi_k (2\pi)^{-\frac{D}{2}} \prod_{i=1}^{D} \sigma_{ki}^{-1} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{D} \sigma_{ki}^{-2} (\mathbf{x}_i - \mu_{ki})^2 \right\} \\ &= \sum_{k=1}^{K} \pi_k \prod_{i=1}^{D} (2\pi)^{-\frac{1}{2}} \sigma_{ki}^{-1} \exp\left\{ -\frac{1}{2} \sigma_{ki}^{-2} (\mathbf{x}_i - \mu_{ki})^2 \right\} \end{aligned}$$

Therefore,

$$f_{ki}(\mathbf{x}) = (2\pi)^{-\frac{1}{2}} \sigma_{ki}^{-1} \exp\left\{ -\frac{1}{2} \sigma_{ki}^{-2} (\mathbf{x}_i - \mu_{ki})^2 \right\}$$

b) (10pt) Recall the latent variable interpretation of GMMs, by introducing a discrete latent variable $z$ that can take values in $1, \ldots, K$. **(i)** What would the Bayesian network structure of this simplified GMM be? The DAG should contain nodes corresponding to $z, x_1, \ldots, x_D$. **(ii)** Describe the conditional probability table/densities of this Bayesian network using $\pi_k, \mu_{ki}, \sigma_{ki}^2$.

**Solution:** (i) Because the off-diagonal entries in each covariance matrix are all 0s, the variables $x_1$, ..., $x_D$ are independent of one another in each Gaussian mixture component, which is the conditional distribution given the latent variable. In other words, $x_1$, ..., $x_D$ are conditionally independent given $z$. The DAG is as shown in Figure 3.

(ii) The conditional probability table for $z$ would contain $K-1$ parameters: $P(z=k)=\pi_k$ for $k=1,\ldots,K-1$ without loss of generality. The conditional probability density for each $x_i$ for $i \in \{1,2,...,D\}$ would take the following form:

$$P(x_i|z=k) = \frac{1}{\sqrt{2\pi}\sigma_{ki}} \exp\left\{-\frac{(x_i-\mu_{ki})^2}{2\sigma_{ki}^2}\right\}$$

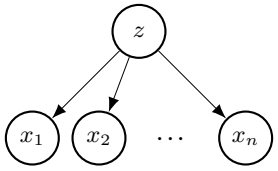


Figure 3: Bayesian network

c) (12pt) You will now derive the Expectation Maximization (EM) algorithm for this simplified GMM given a set of data examples $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$. **(i)** Derive the expression for $\gamma(z_n=k)$ (i.e., $p(z_n=k \mid \mathbf{x}_n)$) for the E step. **(ii)** Derive the update rule for $\sigma_{ki}^2$ in the M step. Your answer may rely on the value of $\mu_{ki}$ in the current M step.

**Solution:** (i)

$$\begin{aligned}
\gamma(z_n=k) &= p(z_n=k|\mathbf{x}_n) \\
&= \frac{p(\mathbf{x}_n|z_n=k)p(z_n=k)}{\sum_{k=1}^K p(\mathbf{x}_n|z_n=k)p(z_n=k)} \\
&= \frac{\pi_k \mathcal{N}(\mathbf{x}|\mu_k,\Sigma_k)}{\sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}|\mu_k,\Sigma_k)} \\
&= \frac{\pi_k \prod_{i=1}^D \frac{1}{\sqrt{2\pi}\sigma_{ki}} \exp\{-\frac{(\mathbf{x}_i-\mu_{ki})^2}{2\sigma_{ki}^2}\}}{\sum_{k=1}^K \pi_k \prod_{i=1}^D \frac{1}{\sqrt{2\pi}\sigma_{ki}} \exp\{-\frac{(\mathbf{x}_i-\mu_{ki})^2}{2\sigma_{ki}^2}\}} \\
&= \frac{\pi_k \prod_{i=1}^D \frac{1}{\sigma_{ki}} \exp\{-\frac{(\mathbf{x}_i-\mu_{ki})^2}{2\sigma_{ki}^2}\}}{\sum_{k=1}^K \pi_k \prod_{i=1}^D \frac{1}{\sigma_{ki}} \exp\{-\frac{(\mathbf{x}_i-\mu_{ki})^2}{2\sigma_{ki}^2}\}}
\end{aligned}$$

(ii)

$$\sigma_{ki}^{2\,(\text{new})} = \frac{\sum_{n=1}^N \gamma(z_n=k)(\mathbf{x}_{ni}-\mu_{ki})^2}{\sum_{n=1}^N \gamma(z_n=k)}$$

Where $\mu_{ki}$ is the value in current M step and $\gamma(z_n=k)$ is the same as in (i).

| $x_1$ | $x_2$ |
|---|---|
| 1.1 | 0.4 |
| -2 | 1 |
| 0 | 0.5 |
| 1 | -1.5 |
| 2.2 | 0 |

Table 3: Training data

d) (12pt) Consider the training data shown in Table 3. Assume that $K = 2$ and $D = 2$. Suppose the parameters are initialized as follows:

$$\pi_1 = 0.2$$
$$\mu_{ki} = \begin{cases} i & \text{if } k = 1 \\ -i & \text{if } k = 2 \end{cases}$$
$$\sigma^2_{ki} = 0.4 \times k \times i$$

Show the parameter values after the one iteration of EM.

**Solution:** The initial parameters are: $\pi_1 = 0.2$, $\pi_2 = 0.8$, $\mu_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$, $\mu_2 = \begin{bmatrix} -1 \\ -2 \end{bmatrix}$, $\Sigma_1 = \begin{bmatrix} 0.4 & 0 \\ 0 & 0.8 \end{bmatrix}$, and $\Sigma_2 = \begin{bmatrix} 0.8 & 0 \\ 0 & 1.6 \end{bmatrix}$

After E step, $z_1 = 0.9047$, $z_2 = 0.0001$, $z_3 = 0.3162$, $z_4 = 0.0003$, $z_5 = 0.9344$.

After M step, $\pi_1 = 0.4317$, $\pi_2 = 0.5682$, $\mu_1 = \begin{bmatrix} 1.4148 \\ 0.2388 \end{bmatrix}$, $\mu_2 = \begin{bmatrix} -0.2653 \\ -0.0406 \end{bmatrix}$, $\Sigma_1 = \begin{bmatrix} 0.6025 & 0 \\ 0 & 0.0500 \end{bmatrix}$, and $\Sigma_2 = \begin{bmatrix} 1.8403 & 0 \\ 0 & 1.2052 \end{bmatrix}$