Machine Learning 2 - Homework 3

Andrii Skliar, 11636785

deadline: April 23, 2018

During the process of solving the homework problems, I have collaborated with the following colleagues:

Gabriele Bani Davide Belli Linda Petrini Gautier Dagan
Pascal Esser Sindy Loewe Gabriele Cesa

NB: credits for the Latex-format go to Iris Verweij, 2nd year MSc AI Student.

Problem 1:

1. Given the denition of the entropy, conditional entropy

$$H(X) = \mathop{\mathbb{E}}_{p(x)}[-\log\ p(x)]$$
$$H(Y|X) = \mathop{\mathbb{E}}_{p(x,y)}[-\log\ p(y|x)]$$

show that the following equalities hold:

$$\begin{aligned} H(X,Y) &= H(X) + H(Y|X) \\ &= H(Y) + H(X|Y) \end{aligned}$$

Solution: I will show derivation for $H(X,Y) = H(X) + H(Y|X)$. Derivation for $H(X,Y) = H(Y) + H(X|Y)$ can be done in the same way.

$$
\begin{aligned}
H(X,Y) &= \mathop{\mathbb{E}}_{p(x,y)}[-\log\ p(y|x)] = -\int\int p(x,y)\log p(x,y)\ dx\ dy \\
&= -\int\int p(x|y)p(y)\log p(x|y)p(y)\ dx\ dy \\
&= -\int\int p(x|y)p(y)[\log p(x|y) + \log p(y)]\ dx\ dy \\
&= -\int\int p(x|y)p(y)\log p(x|y)\ dx\ dy\ -\int\int p(x|y)p(y)\log p(y)\ dx\ dy \\
&= -\int\int p(x|y)p(y)\log p(x|y)\ dx\ dy\ -\int p(y)\log p(y)\Big(\int p(x|y)\ dx\Big)dy \\
&= -\int\int p(x,y)\log p(x|y)\ dx\ dy\ -\int p(y)\log p(y)dy \\
&= H(X|Y) + H(Y)
\end{aligned}
$$

Similarly, using the fact that $p(x,y) = p(y|x)p(x)$ can be shown that $H(X,Y) = H(Y|X) + H(X)$.

2. Consider the conditional mutual information dened by

$$I(X;Y|Z) = \mathop{\mathbb{E}}_{p(z)}[KL(p(x,y,z)||p(x|z)p(y|z))]$$

and show that

$$
\begin{aligned}
I(X;Y|Z) &= H(X|Z) - H(X|Y,Z) \\
&= H(Y|Z) - H(Y|X,Z)
\end{aligned}
$$

Solution:

$$
\begin{aligned}
I(X;Y|Z) &= \int\int\int p(z)p(x,y|z)\log\frac{p(x,y|z)}{p(x|z)p(y|z)}\ dx\ dy\ dz \\
&= \int\int\int p(z)p(x|y,z)p(y|z)\log\frac{p(x|y,z)}{p(x|z)}\ dx\ dy\ dz \\
&= \int\int\int p(z)p(x|y,z)p(y|z)\log p(x|y,z)\ dx\ dy\ dz \\
&- \int\int\int p(z)p(x|y,z)p(y|z)\log p(x|z)\ dx\ dy\ dz \\
&= \int\int\int p(x|y,z)p(y,z)\log p(x|y,z)\ dx\ dy\ dz\ - \int\int\int p(x|y,z)p(y,z)\log p(x|z)\ dx\ dy\ dz \\
&= \int\int\int p(x,y,z)\log p(x|y,z)\ dx\ dy\ dz\ - \int\int\int p(x,y,z)\log p(x|z)\ dx\ dy\ dz \\
&= -H(X|Y,Z) - \int\int \log p(x|z)\Big(\int p(x,y,z)\ dy\Big)\ dx\ dz \\
&= -H(X|Y,Z) - \int\int \log p(x|z)p(x,z)\ dx\ dz \\
&= -H(X|Y,Z) + H(X|Z) = H(X|Z) - H(X|Y,Z)
\end{aligned}
$$

Similarly, using the fact that $p(x,y|z) = p(y|x,z)p(x|z)$ can be shown that $I(X;Y|Z) = H(Y|Z) - H(Y|X,Z)$.

Problem 2: Consider the multinomial distribution:

$$Mult(\mathbf{x}|\boldsymbol{\pi}) = \frac{M!}{x_1!x_2!\ldots x_K!}\pi_1^{x_1}\pi_2^{x_2}\ldots\pi_K^{x_K}$$

where x_i are non-negative integers such that $\sum_{i=1}^K x_i = M$ and π_i are constraints with $\pi_i > 0$ and $\sum_{i=1}^K \pi_i = 1$.

1. Show that it is a member of an exponential family. Derive the minimal representation, i.e. express it in terms of a minimal number of parameters, sufficient statistics and log partition function.

Solution:

$$
\begin{aligned}
Mult(\mathbf{x}|\boldsymbol{\pi}) &= \frac{M!}{x_1!x_2!\dots x_K!}\pi_1^{x_1}\pi_2^{x_2}\dots\pi_K^{x_K} \\
&= \frac{M!}{\prod_{i=1}^K x_i!}\exp\Big[\log\prod_{i=1}^K \pi_i^{x_i}\Big] \\
&= \frac{M!}{\prod_{i=1}^K x_i!}\exp\Big[\sum_{i=1}^K x_i\log\pi_i\Big] \\
&= \frac{M!}{\prod_{i=1}^K x_i!}\exp\Big[\sum_{i=1}^{K-1} x_i\log\pi_i + (M-\sum_{i=1}^{K-1}x_i)\log(1-\sum_{i=1}^{K-1}\pi_i)\Big] \\
&= \frac{M!}{\prod_{i=1}^K x_i!}\exp\Big[\sum_{i=1}^{K-1} x_i\log\frac{\pi_i}{1-\sum_{i=1}^{K-1}\pi_i} + M\log(1-\sum_{i=1}^{K-1}\pi_i)\Big]
\end{aligned}
$$

$$
\boldsymbol{\eta} = \begin{pmatrix} log\frac{\pi_1}{1-\sum_{i=1}^{K-1}\pi_1} \\ \vdots \\ log\frac{\pi_{K-1}}{1-\sum_{i=1}^{K-1}\pi_i} \end{pmatrix}; \quad \mathbf{T}(\mathbf{x}) = \begin{pmatrix} x_1 \\ \vdots \\ x_{K-1} \end{pmatrix}; \quad h(\mathbf{x}) = \frac{M!}{\prod_{i=1}^K x_i!}
$$

$$
A(\boldsymbol{\eta}) = -M\log(1-\sum_{i=1}^{K-1}\pi_i)
$$

Now we need to rewrite $A(\boldsymbol{\eta})$ to express it in terms of $\boldsymbol{\eta}$.

$$
\begin{aligned}
e^{\eta_i} &= \frac{\pi_i}{1-\sum_{i=1}^{K-1}\pi_i} \\
\sum_{i=1}^{K-1} e^{\eta_i} &= \sum_{i=1}^{K-1}\frac{\pi_i}{1-\sum_{i=1}^{K-1}\pi_i} \\
&= \frac{\sum_{i=1}^{K-1}\pi_i}{1-\sum_{i=1}^{K-1}\pi_i} \\
&= \frac{1-1+\sum_{i=1}^{K-1}\pi_i}{1-\sum_{i=1}^{K-1}\pi_i} \\
&= \frac{1}{1-\sum_{i=1}^{K-1}\pi_i} - 1 \\
\frac{1}{1-\sum_{i=1}^{K-1}\pi_i} &= 1+\sum_{i=1}^{K-1} e^{\eta_i} \\
1-\sum_{i=1}^{K-1}\pi_i &= \frac{1}{1+\sum_{i=1}^{K-1} e^{\eta_i}} \\
A(\boldsymbol{\eta}) &= -M\log\frac{1}{1+\sum_{i=1}^{K-1} e^{\eta_i}} \\
&= M\log(1+\sum_{i=1}^{K-1} e^{\eta_i})
\end{aligned}
$$

2. Derive the mean and covariance from the log partition function.

Solution:

$$\mathbb{E}\, x_i = \frac{\partial A(\boldsymbol{\eta})}{\partial \eta_i} = \frac{M e^{\eta_i}}{1 + \sum_{i=1}^{K-1} e^{\eta_i}} = M \frac{\frac{\pi_i}{1-\sum_{i=1}^{K-1}\pi_i}}{\frac{1}{1-\sum_{i=1}^{K-1}\pi_i}} = M\pi_i$$

$$\begin{aligned} Cov(x_i, x_j) = \frac{\partial^2 A(\boldsymbol{\eta})}{\partial \eta_i \eta_j} &= M \frac{\partial(\frac{e^{\eta_i}}{1+\sum_{i=1}^{K-1} e^{\eta_i}})}{\partial \eta_j} = -M \frac{e^{\eta_i} e^{\eta_j}}{(1 + \sum_{i=1}^{K-1} e^{\eta_i})^2} \\ &= -M \frac{e^{\eta_i + \eta_j}}{(1 + \sum_{i=1}^{K-1} e^{\eta_i})^2} = -M \frac{\frac{\pi_i}{1-\sum_{i=1}^{K-1}\pi_i} \frac{\pi_j}{1-\sum_{i=1}^{K-1}\pi_i}}{(\frac{1}{1-\sum_{i=1}^{K-1}\pi_i)^2}} \\ &= -M\pi_i\pi_j \end{aligned}$$

3. Construct the conjugate prior family of this distribution (up to a normalization constant). To which family belongs this conjugate prior?

Solution:

$$\begin{aligned} p(\boldsymbol{\eta}|\boldsymbol{\tau}, \nu) &= \exp\Big[\sum_{i=1}^{K-1} \tau_i log \frac{\pi_i}{1 - \sum_{i=1}^{K-1}\pi_i} - \nu M \log(\frac{1}{1 - \sum_{i=1}^{K-1}\pi_i}) \Big] \\ &= \prod_{i=1}^{K-1} \Big(\frac{\pi_i}{\pi_K}\Big)^{\tau_i} \cdot \pi_K^{\nu M} \\ &= \prod_{i=1}^{K-1} \Big(\frac{\pi_i}{\pi_K}\Big)^{\tau_i} \cdot \pi_K^{\sum_{i=1}^{K-1}\tau_i} \cdot \pi_K^{\nu M - \sum_{i=1}^{K-1}\tau_i} \\ &= \prod_{i=1}^{K-1} \pi_i^{\tau_i} \cdot \pi_K^{\nu M - \sum_{i=1}^{K-1}\tau_i} \end{aligned}$$

Therefore, we can see that conjugate prior of the multinomial distribution is a Dirichlet distribution $Dir(\boldsymbol{\pi}|\boldsymbol{\tau})$, where $\tau_j = \begin{cases} \tau_j, & j < K \\ \nu M - \sum_{i=1}^{K-1}\tau_i, & j = K \end{cases}$

4. For n i.i.d. multinomial observations write down the prior-to-posterior update rule for the hyperparameters.

Solution:
Note: $x_i^{(j)}$ denotes i-th element in vector $\mathbf{x}^{(j)}$. Upper index (j) denotes a value for

j-th observation.

$$\nu \to \nu + n$$

$$\tau_i \to \begin{cases} \tau_i + \sum_{j=1}^{n} x_i^{(j)}, & i < K \\ \tau_i + \sum_{j=1}^{n} (M^j - \sum_{i=1}^{K-1} x_i^{(j)}), & i = K \end{cases}$$

$$\sum_{j=1}^{n} (M^{(j)} - \sum_{i=1}^{K-1} x_i^{(j)}) = \sum_{j=1}^{n} x_K^{(j)}$$

Therefore :

$$\tau_i \to \tau_i + \sum_{j=1}^{n} x_i^{(j)}$$

Problem 3: Consider a time sequence of T samples from K_s *statistically independent* sound sources: $\{s_{it}\} = (s_{i1}, \ldots s_{iT})$, where i labels the source and t the time it was emitted. We record K_x *noisy* linear mixtures of these sound sources:

$$x_{kt} = \sum_{i=1}^{K_s} A_{ki} + \epsilon_{kt} \quad t = 1, \ldots T, \quad k = 1 \ldots K_x$$

$$s_{it} \sim \mathcal{T}(0, \nu_i), \quad \epsilon_{kt} \sim N(0, \sigma_k^2)$$

where s_{it} is distributed as a zero mean Student's T distribution with ν_i degree of freedom and $_{kt}$ a noise random variable drawn from a zero mean normal (Gaussian) distribution with standard deviation σ_k. We assume that the sources were generated independently and we also assume that there are no statistical dependencies between samples generated at different points in time.

1. Explain why this is an ICA model.

Solution: To define what ICA model is, we use notation used in Bishop , namely: ICA models can be described as such in which the observed variables are related linearly to the latent variables. As you can see, this definition is satisfied by our model as it is consists of several statistically independent noisy linear combinations (observed variables, x_{kt}) of multiple sound sources (latent variables, s_{it}). Also, simplification properties are satisfied, namely, sounds sources are distributed as a non-Gaussian distribution (Student's T distribution) and sound sources are i.i.d.

2. Write a general (Bayesian network) expression for the joint probability distribution

$$p(\{s_{1t}\}, \{s_{2t}\}, \{x_{1t}\}, \{x_{2t}\}, \{x_{3t}\}), t = 1 \ldots T$$

Factorize the distribution into smaller conditional and marginal distributions as much as possible. Use explicit (conditional) distributions such as Normal and Students T distributions instead of a generic form "p" as much as possible.

Solution:

$$
\begin{aligned}
p(\{s_{1t}\},\{s_{2t}\},\{x_{1t}\},\{x_{2t}\},\{x_{3t}\}) &= p(s_{1t}|\nu_1)p(s_{2t}|\nu_2)p(x_{1t}|s_{1t},s_{2t},A_1,\sigma_1) \\
&\quad p(x_{2t}|s_{1t},s_{2t},A_2,\sigma_2)p(x_{3t}|s_{1t},s_{2t},A_3,\sigma_3) \\
&= T(s_{1t}|0,\nu_1)T(s_{2t}|0,\nu_2) \\
&\quad \prod_{i=1}^{3}[N(x_{it}|0,\sigma^2) + A_{i1}T(s_{1t}|0,\nu_1) + A_{i2}T(s_{2t}|0,\nu_2)] \\
&= \prod_{i=1}^{2}[T(s_{it}|0,\nu_i)] \prod_{i=1}^{3}[N(x_{it}|\begin{pmatrix}A_{i1} & A_{i2}\end{pmatrix}\begin{pmatrix}T(s_{1t}|0,\nu_1)\\ T(s_{2t}|0,\nu_2)\end{pmatrix},\sigma^2)]
\end{aligned}
$$

3. Explain what the term explaining away" means and indicate if this explaining away phenomenon is present in the ICA model under discussion.

Solution: "Explaining away" happens when there is a collider case in the Bayesian network. The idea is following: imagine there are two random variables X and Y influencing the third random variables Z and X and Y are independent. However, if we observe Z, X and Y become dependent. This leads to that Y "explains away" X, thus, observing Y makes probability of X lower.

This phenomenon is present in the current ICA model. We can see signals as influencing variables (X and Y) and mixture as variable being influenced (Z). If we observe value of x_{1t} and value of s_{1t}, we can say with high certainty that value of s_{2t} will be close to $\frac{x_{1t}-A_{11}s_{1t}}{A_{12}}$.

4. Since samples across time t are independent, we will ignore the index t in the following two questions (you may imagine t = 1). For all of the (conditional) independence expressions below, state if they are true or (typically) false:

Solution:

a) False

b) True

c) False

d) True

e) False

f) False

g) False

h) False

5. What is the Markov blanket of s_1 ? What is the Markov blanket of x_1?

Solution: Due to the fact that Markov blankets haven't been extensively defined and discussed during the lecture and definitions on the internet differ, I assumed that parameters, which are not random variables are not part of Markov blanket. Taking this fact into the account, we get following results:

$$MB(s_1) = \{x_1, x_2, x_3, s_2\}$$
$$MB(x_1) = \{s_1, s_2\}$$

6. Write an explicit expression in terms of W and the sources students T distributions $\mathcal{T}(s_i|0, \nu_i)$ of the probability:

$$p(\{x_{kt}|\mathbf{W}, \{\nu_i\}),\ t = 1 \dots T, k = 1 \dots K$$

Solution: First we need to calculate $Jac(s \to x)$.

$$Jac(s \to x) = Jac(\sum_{k=1}^{K} W_{ik} x_{kt}) p(\{x_{kt}|\mathbf{W}, \{\nu_i\}) = \frac{\partial \sum_{k=1}^{K} W_{ik} x_{kt}}{\partial (x_1, \dots x_k)} = \mathbf{W}$$

$$\begin{aligned}
\prod_{t=1}^{T} p_X(x) &= \{\text{using formulas provided}\} \\
&= \prod_{t=1}^{T} p_S(s(x)) |det Jac(s \to x)| \\
&= \prod_{t=1}^{T} \prod_{i=1}^{K} (p(s_{it}))| \det \mathbf{W}| \\
&= \prod_{t=1}^{T} \prod_{i=1}^{K} (\tau(0, \nu_i))| \det \mathbf{W}| \\
&= \prod_{t=1}^{T} | \det \mathbf{W}| \prod_{i=1}^{K} \tau(0, \nu_i)
\end{aligned}$$

7. Write down the log-likelihood of the complete deterministic ICA model above.

Solution:

$$\log p(\{x_{kt}|\mathbf{W}, \{\nu_i\}) = \log \Big[\prod_{t=1}^{T} |\det \mathbf{W}| \prod_{i=1}^{K} \tau(0, \nu_i) \Big]$$
$$= \sum_{t=1}^{T} \log |\det \mathbf{W}| + \sum_{i=1}^{K} \log \tau(0, \nu_i)$$

8. Explain in detail the stochastic gradient ascent optimization algorithm to maximize the loglikelihood of the previous question. Note: you do not have to derive or provide the expression of the gradient; instead you can provide a general description of the algorithm.

Solution: Stochastic Gradient Descent for the ICA looks as following:

Algorithm 1 ICA Stochastic Gradient Descent

Initialize $\mathbf{W}^{(0)} = RandomMatrix(K, K)$
Set time: $\tau = 0$
Choose a learning rate η
while $||\mathbf{W}^{\tau+1} - \mathbf{W}^{\tau}|| > \epsilon$ **do**
    for each datapoint $\mathbf{x}(t)$ **do**
        Put $\mathbf{x}$ through a linear mapping: $\mathbf{a}(t) = \mathbf{W}\mathbf{x}(t)$
        $\mathbf{z} = \mathbf{0}$
        for $j := 1, \dots K$ **do**
            Put $\mathbf{a}$ through a nonlinear map: $z_j(t) = \phi_j(a_j(t))$
        end for
        Put $\mathbf{a}$ back through $\mathbf{W}$: $\mathbf{x}' = \mathbf{W}^T\mathbf{a}$
        Update $\mathbf{W}$: $\mathbf{W}^{(\tau+1)} = \mathbf{W}^{(\tau)} + \eta(\mathbf{W}^{(\tau)} + \mathbf{z}\mathbf{x}'^T)$
    end for
    Update time: $\tau = \tau + 1$
end while
return $\mathbf{W}^{(\tau)}$

The idea is that we would like to maximize log-likelihood and to do that, we find the direction of greatest ascend by taking a gradient. We update matrix $\mathbf{W}$ in the direction of the greatest ascend. This is done in each iteration until the convergence. In the end, it results in finding $\mathbf{W}$, which maximizes the log-likelihood.

9. In which limit do you expect overfitting: $K >> T$ or $T >> K$? Explain your answer.

Solution: We can expect overfitting in the case of $K >> T$, because in that case we would have much fewer datapoints than dimensions. This would mean that we have to estimate matrix W, which has K^2 parameters while we only have $K \cdot T$ values in our

dataset. This means, that we have an under-determined system and this would result in our model overfitting to the provided data.

Problem 4: Given a graphical model, show that:

1. $p(\mathbf{x}_1, \dots \mathbf{x}_{n-1}|\mathbf{x}_n, \mathbf{z}_n) = p(\mathbf{x}_1, \dots \mathbf{x}_{n-1}|\mathbf{z}_n)$

Solution: To prove this equality, we can use d-separation. First, let's declare A, B and C.
$A = \{\mathbf{x}_1, \dots \mathbf{x}_{n-1}\}; B = \{\mathbf{x}_n\}; C = \{\mathbf{z}_n\}$.
For arbitrary $n > 1$, every path π from A to B goes through $\mathbf{z}_n$, which is not an end node and a non-collider, which satisfies $\mathbf{z}_n \in C$, therefore, every π from A to B is blocked by C.

This means, that $A \perp^d B|C \implies A \perp\!\!\!\perp_p B|C$.

As far as this holds for any subset of A, B and C, we can see that:
$\{\mathbf{x}_1, \dots \mathbf{x}_{n-1}\} \perp\!\!\!\perp \mathbf{x}_n|\mathbf{z}_n \implies p(\mathbf{x}_1, \dots \mathbf{x}_{n-1}|\mathbf{x}_n, \mathbf{z}_n) = p(\mathbf{x}_1, \dots \mathbf{x}_{n-1}|\mathbf{z}_n)$.

2. $p(\mathbf{x}_1, \dots \mathbf{x}_{n-1}|\mathbf{z}_{n-1}, \mathbf{z}_n) = p(\mathbf{x}_1, \dots \mathbf{x}_{n-1}|\mathbf{z}_{n-1})$

Solution: To prove this equality, we can use d-separation. First, let's declare A, B and C.
$A = \{\mathbf{x}_1, \dots \mathbf{x}_{n-1}\}; B = \{\mathbf{z}_n\}; C = \{\mathbf{z}_{n-1}\}$.
For arbitrary $n > 1$, every path π from A to B goes through $\mathbf{z}_{n-1}$, which is not an end node and a non-collider, which satisfies $\mathbf{z}_{n-1} \in C$, therefore, every π from A to B is blocked by C.

This means, that $A \perp^d B|C \implies A \perp\!\!\!\perp_p B|C$.

As far as this holds for any subset of A, B and C, we can see that:
$\{\mathbf{x}_1, \dots \mathbf{x}_{n-1}\} \perp\!\!\!\perp \mathbf{z}_n|\mathbf{z}_{n-1} \implies p(\mathbf{x}_1, \dots \mathbf{x}_{n-1}|\mathbf{z}_{n-1}, \mathbf{z}_n) = p(\mathbf{x}_1, \dots \mathbf{x}_{n-1}|\mathbf{z}_{n-1})$.

3. $p(\mathbf{x}_{n+1}, \dots \mathbf{x}_N|\mathbf{z}_n, \mathbf{z}_{n+1}) = p(\mathbf{x}_{n+1}, \dots \mathbf{x}_N|\mathbf{z}_{n+1})$

Solution: Using factorization properties of the Bayesian Networks, we can rewrite full joint probability in a following way:

$$p(\mathbf{x}_1, \dots \mathbf{x}_N, \mathbf{z}_1, \dots \mathbf{z}_N) = p(\mathbf{z}_1)p(\mathbf{x}_1|\mathbf{z}_1)\prod_{i=2}^{N} p(\mathbf{z}_i|\mathbf{z}_{i-1})p(\mathbf{x}_i|\mathbf{z}_i)$$

Therefore, it can be seen from that equation that:

$$\begin{aligned}
p(\mathbf{x}_{n+1}, \dots \mathbf{x}_N | \mathbf{z}_n, \mathbf{z}_{n+1}) &= \frac{p(\mathbf{z}_n, \mathbf{z}_{n+1} | \mathbf{x}_{n+1}, \dots \mathbf{x}_N) p(\mathbf{x}_{n+1}, \dots \mathbf{x}_N)}{p(\mathbf{z}_n, \mathbf{z}_{n+1})} \\
&= \frac{\cancel{p(\mathbf{z}_n | \mathbf{z}_{n+1})} p(\mathbf{z}_{n+1} | \mathbf{x}_{n+1}, \dots \mathbf{x}_N) p(\mathbf{x}_{n+1}, \dots \mathbf{x}_N)}{\cancel{p(\mathbf{z}_n | \mathbf{z}_{n+1})} p(\mathbf{z}_{n+1})} \\
&= \frac{p(\mathbf{z}_{n+1} | \mathbf{x}_{n+1}, \dots \mathbf{x}_N) p(\mathbf{x}_{n+1}, \dots \mathbf{x}_N)}{p(\mathbf{z}_{n+1})} \\
&= \{\text{Using Bayes rule}\} \\
&= p(\mathbf{x}_{n+1}, \dots \mathbf{x}_N | \mathbf{z}_{n+1})
\end{aligned}$$

4. $p(\mathbf{z}_{N+1} | \mathbf{z}_N, \mathbf{X}) = p(\mathbf{z}_{N+1} | \mathbf{z}_N)$

Solution: Here we assume, that z_{N+1} is a new node. Therefore, using factorization properties of the Bayesian Networks, we can rewrite full joint probability in a following way:

$$p(\mathbf{x}_1, \dots \mathbf{x}_N, \mathbf{z}_1, \dots \mathbf{z}_{N+1}) = p(\mathbf{z}_1) p(\mathbf{x}_1 | \mathbf{z}_1) \prod_{i=2}^{N} p(\mathbf{z}_i | \mathbf{z}_{i-1}) p(\mathbf{x}_i | \mathbf{z}_i) \cdot p(\mathbf{z}_{N+1} | \mathbf{z}_N)$$

Therefore, it can be seen from that equation that:

$$\begin{aligned}
p(\mathbf{z}_{N+1} | \mathbf{z}_N, \mathbf{X}) &= \frac{p(\mathbf{z}_{N+1}, \mathbf{X} | \mathbf{z}_N) p(\mathbf{z}_{N+1})}{p(\mathbf{z}_N, \mathbf{X})} \\
&= \frac{p(\mathbf{X} | \mathbf{z}_N, \mathbf{z}_{N+1}) p(\mathbf{z}_N | \mathbf{z}_{N+1}) p(\mathbf{z}_{N+1})}{p(\mathbf{X}, \mathbf{z}_N)} \\
&= \{\text{Using task 4.2}\} \\
&= \frac{\cancel{p(\mathbf{X} | \mathbf{z}_N)} p(\mathbf{z}_N | \mathbf{z}_{N+1}) p(\mathbf{z}_{N+1})}{\cancel{p(\mathbf{X} | \mathbf{z}_N)} p(\mathbf{z}_N)} \\
&= \{\text{Using Bayes rule}\} \\
&= p(\mathbf{z}_{N+1} | \mathbf{z}_N)
\end{aligned}$$