EQ2415 – Machine Learning and Data Science
HT22

Tutorial 6 A. Honoré, A. Ghosh

Graphical models in a Bayesian framework

1 Bayesian networks

Question 1. (a) Bishop 8.1 [1]

Solution: We know by the directed graph factorization, for a directed acyclic graph (DAG) with K nodes, the joint distribution is given by

$$p(\mathbf{x}) = \prod_{k=1}^{K} p(x_k|\mathrm{pa}_k), \tag{1}$$

where pa_k refers to the set of parent nodes of x_k. Also, by the property of directed acyclic graphs pa_k always contains members from $x_1, x_2, \ldots, x_{k-1}$. We need to show that if $p(\mathbf{x})$ obeys (1) then $p(\mathbf{x})$ is normalized correctly provided each conditional distribution is correctly normalized. This is straightforward to show by using the sum and product rules of probability:

$$\begin{aligned}
\int p(\mathbf{x})\, d\mathbf{x} &= \int \prod_{k=1}^{K} p(x_k|\mathrm{pa}_k)\, d\mathbf{x} \\
&= \int\int \cdots \int \prod_{k=1}^{K} p(x_k|\mathrm{pa}_k)\, dx_1 dx_2 \ldots dx_K \\
&\quad \text{(K times)} \\
&= \int\int \cdots \int p(x_1)\, p(x_2|\mathrm{pa}_2) \ldots p(x_K|\mathrm{pa}_K)\, dx_1 dx_2 \ldots dx_K \\
&\quad \text{(K times)} \\
&= \int p(x_K|\mathrm{pa}_K) \left[\int p(x_{K-1}|\mathrm{pa}_{K-1}) \ldots \left[\int p(x_2|\mathrm{pa}_2) \left[\int p(x_1)\, dx_1 \right] dx_2 \right] \ldots dx_{K-1} \right] dx_K
\end{aligned} \tag{2}$$

If every conditional distribution is normalized i.e. $\int p(x_k|\mathrm{pa}_k)\, dx_k = 1$, then the above result also goes to 1. Hence, $\int p(\mathbf{x})\, d\mathbf{x} = 1$ (since each of the products go to 1). ■

Question 1. (b) Bishop 8.3. Also, draw the graphical model by first showing that $p(a, b, c) = p(a)p(c|a)p(b|c)$.

Solution: Using the table 8.2, first find $p(a, b)$ by using sum-rule. You should be able to get One

a	b	$p(a, b)$
0	0	0.336
0	1	0.264
1	0	0.256
1	1	0.144

Table 1: Probabilty mass function for $p(a, b)$

more step of the sum rule on Table 1, gives $p(a = 0) = 0.6, p(a = 1) = 0.4$ and $p(b = 0) = 0.592, p(b = 1) = 0.408$. Also, we can find $p(c = 0) = 0.48, p(c = 1) = 0.52$. Now try comparing $p(a, b)$ and $p(a)p(b)$ to check independence.

Similarly, if you compute for $p(a, c)$ and (b, c), you can check that For one case say $a = 0, b = 0, c = 0$, you can show that $p(a = 0, b = 0|c = 0) = 0.4 = p(b = 0|c = 0)\, p(a = 0|c = 0)$. Furthermore, using this

a	c	$p(a,b)$
0	0	0.240
0	1	0.360
1	0	0.240
1	1	0.160

Table 2: Probabilty mass function for $p(a,c)$

b	c	$p(b,c)$
0	0	0.384
0	1	0.208
1	0	0.096
1	1	0.312

Table 3: Probabilty mass function for $p(b,c)$

conditional independence $p(a,b|c) = p(a|c)p(b|c)$, you can show that

$$\begin{aligned} p(a,b,c) &= p(a)p(c|a)p(b|a,c) \\ &= p(a)p(c|a)\frac{p(a,b,c)}{p(a,c)} \\ &= p(a)p(c|a)\frac{p(a,b|c)p(c)}{p(a|c)p(c)} \\ &= p(a)p(c|a)\frac{p(a|c)p(b|c)p(c)}{p(a|c)p(c)} \\ &= p(a)p(c|a)p(b|c) \end{aligned} \tag{3}$$

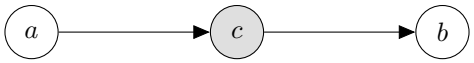

Figure 1: Graphical model for solution 1. (b), where c is observed.

■

2 Conditional independence

Question 2. Show both graphically and analytically if $x_1, x_2, \ldots, x_{t-1} \perp\!\!\!\perp x_{t+1}, x_{t+2}, \ldots, x_T | x_t, y_t$, where y_t is assumed to depend only on x_t, then $x_1, x_2, \ldots, x_{t-1} \perp\!\!\!\perp x_{t+2}, x_{t+3}, \ldots, x_T | x_t, y_t$.

Solution:

Analytical solution: We know from the question that $x_1, x_2, \ldots, x_{t-1} \perp\!\!\!\perp x_{t+1}, x_{t+2}, \ldots, x_T | x_t, y_t$, with y_t being only function of x_t. Writing this out in terms of probability means that

$$p\left(x_1, x_2, \ldots, x_{t-1}, x_{t+1}, x_{t+2}, \ldots, x_T | x_t, y_t\right) = p\left(x_1, x_2, \ldots, x_{t-1} | x_t, y_t\right) p\left(x_{t+1}, x_{t+2}, \ldots, x_T | x_t, y_t\right) \tag{4}$$

Now, if we would like to find $p\left(x_1, x_2, \ldots, x_{t-1}, x_{t+2}, \ldots, x_T | x_t, y_t\right)$, we have:

$$\begin{aligned} p\left(x_1, x_2, \ldots, x_T | x_t, y_t\right) &= \sum_{x_{t+1}} p\left(x_1, x_2, \ldots, x_{t-1} | x_t, y_t\right) p\left(x_{t+1}, x_{t+2}, \ldots, x_T | x_t, y_t\right) \\ &= p\left(x_1, x_2, \ldots, x_{t-1} | x_t, y_t\right) \sum_{x_{t+1}} p\left(x_{t+1}, x_{t+2}, \ldots, x_T | x_t, y_t\right) \\ &= p\left(x_1, x_2, \ldots, x_{t-1} | x_t, y_t\right) p\left(x_{t+2}, \ldots, x_T | x_t, y_t\right) \sum_{x_{t+1}} p\left(x_{t+1} | x_{t+2}, \ldots, x_T, x_t, y_t\right) \\ &= p\left(x_1, x_2, \ldots, x_{t-1} | x_t, y_t\right) p\left(x_{t+2}, \ldots, x_T | x_t, y_t\right) \end{aligned} \tag{5}$$

Graphical solution: We can actually think of x_i being in a Markovian dependency with y_t coming directly from x_t. Then x_t is a head to tail node with the tail part coming from $x_{t+1:T}$ and head part coming from $x_{1:t-1}$. ■

Question 3. A popular architecture recent architecture that can be represented as a graphical model is a dynamical variational autoencoder (DVAE) [2]. It has been shown to perform reasonably well when applied to signal generation and modeling tasks. Here below is the graphical model of a dynamical VAE called deep Kalman filter (DKF): In the inference mode, this becomes

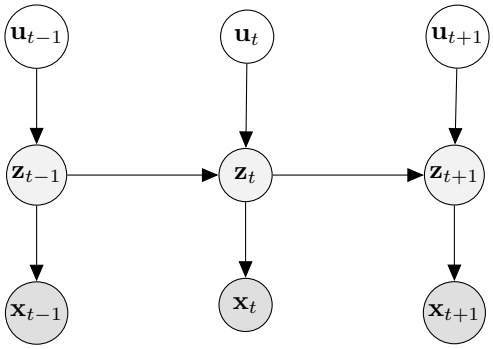

Figure 2: Graphical model of the DKF in generation mode, displayed within 3 states of $\mathbf{z}_{t-1}$. The arrows signify dependence between random variables in the graphical model.

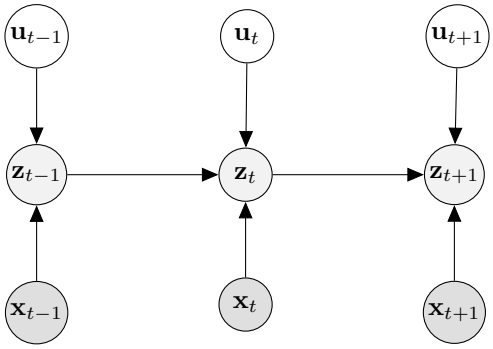

Figure 3: Graphical model of the DKF in inference mode, displayed within 3 states of $\mathbf{z}_{t-1}$. The arrows signify dependence between random variables in the graphical model.

(a) Write down the joint distribution of $p\left(\mathbf{x}_{1:T}, \mathbf{z}_{1:T} | \mathbf{u}_{1:T}\right)$ for the generation mode in simplified form using D-separation.

Solution: We can write the joint distribution as

$$\begin{aligned} p\left(\mathbf{x}_{1:T}, \mathbf{z}_{1:T} | \mathbf{u}_{1:T}\right) &= \prod_{t=1}^{T} p\left(\mathbf{x}_t | \mathbf{x}_{1:t-1}, \mathbf{z}_{1:t}, \mathbf{u}_{1:t}\right) p\left(\mathbf{z}_t | \mathbf{z}_{1:t-1}, \mathbf{x}_{1:t-1}, \mathbf{u}_{1:t}\right) \\ &= \prod_{t=1}^{T} p\left(\mathbf{x}_t | \mathbf{z}_t\right) p\left(\mathbf{z}_t | \mathbf{z}_{t-1}, \mathbf{u}_t\right) \end{aligned} \tag{6}$$

This simplifcation is due to D-separation. This is because $\mathbf{z}_t$ is a head-to-tail node, so that $\mathbf{x_t} \perp\!\!\!\perp \mathbf{z}_{1:t-2}, \mathbf{u}_{1:t}, \mathbf{x}_{1:t-1} | \mathbf{z}_t$. Similar arguments for $\mathbf{z}_{t-1}$ give us $\mathbf{z_t} \perp\!\!\!\perp \mathbf{z}_{1:t-2}, \mathbf{u}_{1:t}, \mathbf{x}_{1:t-1} | \mathbf{z}_{t-1}$ leading to simplification of the second term. ■

(b) Write down the joint distribution $q\left(\mathbf{z}_{1:T} | \mathbf{x}_{1:T}, \mathbf{u}_{1:T}\right)$ for the inference mode in simplified form using D-separation.

Solution: We can write the joint distribution as

$$\begin{aligned} q\left(\mathbf{z}_{1:T}|\mathbf{x}_{1:T},\mathbf{u}_{1:T}\right) &= \prod_{t=1}^{T} q\left(\mathbf{z}_t|\mathbf{z}_{1:t-1},\mathbf{x}_{1:T},\mathbf{u}_{1:T}\right) \\ &= \prod_{t=1}^{T} q\left(\mathbf{z}_t|\mathbf{z}_{t-1},\mathbf{x}_{t:T},\mathbf{u}_{t:T}\right) \end{aligned} \tag{7}$$

This simplifcation is again due to D-separation. This is because $\mathbf{z}_{t-1}$ is a head-to-tail node, so that it 'blocks' / accumulates the information coming from $\mathbf{x}_{1:t-1}, \mathbf{z}_{1:t-2}, \mathbf{u}_{1:t-1}$. ■

3 Markov random fields

Question 5. (a) Bishop 8.12

Solution: Considering we have M distinct random variables, the i^{th} variable can be connected to $M-1$ other variables. So, in total we can have $M(M-1)$ number of connections. But since in an undirected graphical model the connection between nodes i and j is same as the connection between j and i, we actually have $M(M-1)/2$ distinct connections. Finally to obtain the number of such graphs, we note for each connection, we have two choices - either the connection is present in the graph or not, so in total we have $2^{M(M-1)/2}$ such distinct undirected graphs from M distinct random variables. ■

Question 5. (b) Bishop 8.13

Solution: The difference between the energy functions for different states of x_j is

$$\begin{aligned} E\left(\mathbf{x}_j = x_0, \mathbf{y}\right) - E\left(\mathbf{x}_j = x_1, \mathbf{y}\right) &= hx_0 - \eta\sum_i x_0 y_i - \beta\sum_i x_0 x_i - hx_1 - \eta\sum_i x_1 y_i - \beta\sum_i x_1 x_i \\ &= h(x_0 - x_1) - \eta\sum_i (x_0 - x_1) y_i - \beta\sum_i (x_0 - x_1) x_i \end{aligned} \tag{8}$$

So, we can see that the difference in energy functions depend on the neighbouring points of x_j (that actually form the cliques), i.e. local to x_j in the graph (assuming other variables fixed). ■

References

[1] Bishop, Christopher M., and Nasser M. Nasrabadi. Pattern recognition and machine learning. Vol. 4. No. 4. New York: springer, 2006.

[2] Girin, Laurent, et al. "Dynamical Variational Autoencoders: A Comprehensive Review." Foundations and Trends in Machine Learning 15.1-2 (2021): 1-175.