01.112 Machine Learning, Fall 2018
Homework 5

Due Friday 7 Dec 2018, 5pm

Sample Solutions

1 Question 1 (total 30 points)

Consider the Bayesian network below, where we have 11 variables.

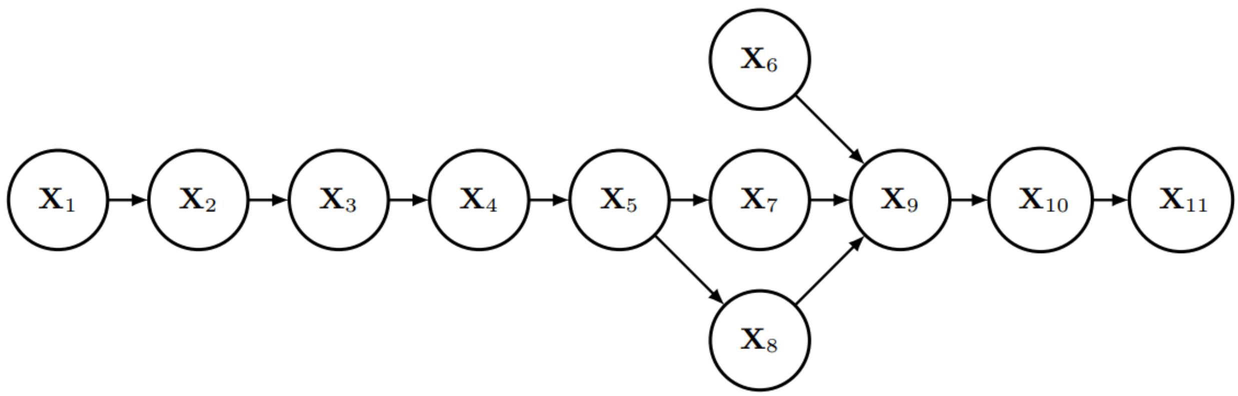

Figure 1: The Bayesian network consists of 11 variables

1.1 (4 points) Assume all variables are taking values from $\{1, 2, 3\}$. What is the number of free parameters? What if we assume all variables are taking values from $\{1, 2, 3, 4\}$?

Answer:

Consider a node i with its parents pa_i, the number of *free parameters*, also called independent parameters, for the node i: $(r_i - 1)\prod_{j\in pa_i} rj$.

In case all variables are taking values from $\{1, 2, 3\}$. According to Fig. 1, we can calculate the number of free parameters as follows:

$$\begin{aligned}\sum_{i=1}^{11}(r_i - 1)\prod_{j\in pa_i} r_j =& 2_{X_1} + 2_{X_2} \times 3_{pa_{X_2}} + 2_{X_3} \times 3_{pa_{X_3}} + 2_{X_4} \times 3_{pa_{X_4}} + \\ & 2_{X_5} \times 3_{pa_{X_5}} + 2_{X_6} + 2_{X_7} \times 3_{pa_{X_7}} + 2_{X_8} \times 3_{pa_{X_8}} + \\ & 2_{X_9} \times 3^3_{pa_{X_9}} + 2_{X_{10}} \times 3_{pa_{X_{10}}} + 2_{X_{11}} \times 3_{pa_{X_{11}}} = 106.\end{aligned} \tag{1}$$

In case all variables are taking values from $\{1, 2, 3, 4\}$. Similarly, we can calculate the number of free parameters as follows:

$$\begin{aligned}\sum_{i=1}^{11}(r_i - 1)\prod_{j\in pa_i} r_j =& 3_{X_1} + 3_{X_2} \times 4_{pa_{X_2}} + 3_{X_3} \times 4_{pa_{X_3}} + 3_{X_4} \times 4_{pa_{X_4}} + \\ & 3_{X_5} \times 4_{pa_{X_5}} + 3_{X_6} + 3_{X_7} \times 4_{pa_{X_7}} + 3_{X_8} \times 4_{pa_{X_8}} + \\ & 3_{X_9} \times 4^3_{pa_{X_9}} + 3_{X_{10}} \times 4_{pa_{X_{10}}} + 3_{X_{11}} \times 4_{pa_{X_{11}}} = 294.\end{aligned} \tag{2}$$

1.2 (4 points) What is the Markov blanket for the variable X_1 in the Bayesian network? What is the Markov blanket for the variable X_7?

Answer:

The Markov blanket for the variable X_1 in the given Bayesian network is $m(X_1) = \{X_2\}$.
The Markov blanket for the variable X_7 in the given Bayesian network is $m(X_7) = \{X_5, X_9, X_6, X_8\}$.

1.3 (6 points) Are X_1 and X_6 independent or dependent of each other if no other variable is given? Why? Are X_1 and X_6 independent or dependent of each other if both X_7 and X_{10} are given? Why?

Answer:

If no other variable is given, X_1 and X_6 are *independent* because there is no path from X_1 to X_6 (according to the Bayes' ball algorithm).

If both X_7 and X_{10} are given, X_1 and X_6 are *dependent*. This is because there exists a path from X_1 to X_6 according to the Bayes' ball algorithm with the boundary conditions. Specifically, the paths are: $X_1 - X_2 - X_3 - X_4 - X_5 - X_8 - X_9 - X_{10} - X_9 - X_6$ and $X_1 - X_2 - X_3 - X_4 - X_5 - X_7 - X_5 - X_8 - X_9 - X_{10} - X_9 - X_6$.

1.4 (8 points) Now, assume the probability tables for all nodes are shown below:

	X_1 = 1	X_1 = 2
	0.5	0.5

X_1	X_2 = 1	X_2 = 2
1	0.2	0.8
2	0.3	0.7

X_2	X_3 = 1	X_3 = 2
1	0.3	0.7
2	0.3	0.7

X_3	X_4 = 1	X_4 = 2
1	0.1	0.9
2	0.5	0.5

X_4	X_5 = 1	X_5 = 2
1	0.5	0.5
2	0.6	0.4

	X_6 = 1	X_6 = 2
	0.6	0.4

X_6	X_7	X_8	X_9 = 1	X_9 = 2
1	1	1	0.8	0.2
1	1	2	0.1	0.9
1	2	1	0.9	0.1
1	2	2	0.7	0.3
2	1	1	0.3	0.7
2	1	2	0.2	0.8
2	2	1	0.2	0.8
2	2	2	0.9	0.1

X_5	X_7 = 1	X_7 = 2
1	0.2	0.8
2	0.3	0.7

X_5	X_8 = 1	X_8 = 2
1	0.8	0.2
2	0.7	0.3

X_9	X_{10} = 1	X_{10} = 2
1	0.8	0.2
2	0.8	0.2

X_{10}	X_{11} = 1	X_{11} = 2
1	0.7	0.3
2	0.8	0.2

Calculate the following conditional probability:

$$P(\mathbf{X}_3 = 2|\mathbf{X}_4 = 1)$$

(Hint: find a short answer.)
Answer:
First, the conditional probability is given as follows:

$$P(\mathbf{X}_3 = 2|\mathbf{X}_4 = 1) = \frac{P(\mathbf{X}_3 = 2, \mathbf{X}_4 = 1)}{P(\mathbf{X}_4 = 1)} \tag{3}$$

The numerator of the Eq. 3 is given as follows:

$$P(\mathbf{X}_3 = 2, \mathbf{X}_4 = 1) = \sum_{\mathbf{X}_1, \mathbf{X}_2} P(\mathbf{X}_1)P(\mathbf{X}2|\mathbf{X}1)P(\mathbf{X}_3|\mathbf{X}_2)P(\mathbf{X}_4|\mathbf{X}_3).$$

And, note that X_3 takes only two values 1 or 2, the denominator of the Eq. 3 is calculated as follows:

$$P(\mathbf{X}_4 = 1) = P(\mathbf{X}_3 = 1, \mathbf{X}_4 = 1) + P(\mathbf{X}_3 = 2, \mathbf{X}_4 = 1).$$

According to the value of the Table, we can get:

$$\begin{aligned}
P(\mathbf{X}_3 = 1, \mathbf{X}_4 = 1) =& P(\mathbf{X}_1 = 1)P(\mathbf{X}_2 = 1|\mathbf{X}_1 = 1)P(\mathbf{X}_3 = 1|\mathbf{X}_2 = 1)P(\mathbf{X}_4 = 1|\mathbf{X}_3 = 1) \\
&+ P(\mathbf{X}_1 = 1)P(\mathbf{X}_2 = 2|\mathbf{X}_1 = 1)P(\mathbf{X}_3 = 1|\mathbf{X}_2 = 2)P(\mathbf{X}_4 = 1|\mathbf{X}_3 = 1) \\
&+ P(\mathbf{X}_1 = 2)P(\mathbf{X}_2 = 1|\mathbf{X}_1 = 2)P(\mathbf{X}_3 = 1|\mathbf{X}_2 = 1)P(\mathbf{X}_4 = 1|\mathbf{X}_3 = 1) \\
&+ P(\mathbf{X}_1 = 1)P(\mathbf{X}_2 = 2|\mathbf{X}_1 = 2)P(\mathbf{X}_3 = 1|\mathbf{X}_2 = 2)P(\mathbf{X}_4 = 1|\mathbf{X}_3 = 1) \\
&= 0.5 \times 0.2 \times 0.3 \times 0.1 \\
&+ 0.5 \times 0.8 \times 0.3 \times 0.1 \\
&+ 0.5 \times 0.3 \times 0.3 \times 0.1 \\
&+ 0.5 \times 0.7 \times 0.3 \times 0.1 \\
&= 0.03.
\end{aligned}$$

$$\begin{aligned}
P(\mathbf{X}_3 = 2, \mathbf{X}_4 = 1) =& P(\mathbf{X}_1 = 1)P(\mathbf{X}_2 = 1|\mathbf{X}_1 = 1)P(\mathbf{X}_3 = 2|\mathbf{X}_2 = 1)P(\mathbf{X}_4 = 1|\mathbf{X}_3 = 2) \\
&+ P(\mathbf{X}_1 = 1)P(\mathbf{X}_2 = 2|\mathbf{X}_1 = 1)P(\mathbf{X}_3 = 2|\mathbf{X}_2 = 2)P(\mathbf{X}_4 = 1|\mathbf{X}_3 = 2) \\
&+ P(\mathbf{X}_1 = 2)P(\mathbf{X}_2 = 1|\mathbf{X}_1 = 2)P(\mathbf{X}_3 = 2|\mathbf{X}_2 = 1)P(\mathbf{X}_4 = 1|\mathbf{X}_3 = 2) \\
&+ P(\mathbf{X}_1 = 1)P(\mathbf{X}_2 = 2|\mathbf{X}_1 = 2)P(\mathbf{X}_3 = 2|\mathbf{X}_2 = 2)P(\mathbf{X}_4 = 1|\mathbf{X}_3 = 2) \\
&= 0.5 \times 0.2 \times 0.7 \times 0.5 \\
&+ 0.5 \times 0.8 \times 0.7 \times 0.5 \\
&+ 0.5 \times 0.3 \times 0.7 \times 0.5 \\
&+ 0.5 \times 0.7 \times 0.7 \times 0.5 \\
&= 0.35.
\end{aligned}$$

So that, we can get the result as follows:

$$\begin{aligned} P(\mathbf{X}_3 = 2|\mathbf{X}_4 = 1) &= \frac{P(\mathbf{X}_3 = 2, \mathbf{X}_4 = 1)}{P(\mathbf{X}_4 = 1)} \\ &= \frac{0.35}{0.03 + 0.35} \approx 0.92105 \end{aligned}$$

1.5 (8 points) Calculate the following conditional probability based on the above probability tables.

$$P(\mathbf{X}_5 = 2|\mathbf{X}_3 = 1, \mathbf{X}_{11} = 2, \mathbf{X}_1 = 1)$$

(Hint: find a short answer. The values in some of the probability tables may reveal some useful information.)

Answer:

First, we have following observations based on the probability tables:

- Probability of X_3 does not change regarding X_2. In other words, X_3 and X_2 are independent.
- Probability of X_{10} does not change regarding X_9. In other words, X_{10} and X_9 are independent.

X_5 and X_1 are independent, and X_5 and X_{11} are independent as well (i.e., no path). Based on the given graph of the Bayesian network, we can get the following formula:

$$\begin{aligned} &P(\mathbf{X}_5 = 2|\mathbf{X}_3 = 1, \mathbf{X}_{11} = 2, \mathbf{X}_1 = 1) \\ =&P(\mathbf{X}_5 = 2|\mathbf{X}_3 = 1) \\ =&\frac{P(\mathbf{X}_5 = 2, \mathbf{X}_3 = 1)}{P(\mathbf{X}_3 = 1)} \\ =&\frac{\sum_{x_4} P(\mathbf{X}_3 = 1)P(\mathbf{X}_4 = x_4|\mathbf{X}_3 = 1)P(\mathbf{X}_5 = 2|\mathbf{X}_4 = x_4)}{P(\mathbf{X}_3 = 1)} \\ =&\sum_{x_4} P(\mathbf{X}_4 = x_4|\mathbf{X}_3 = 1)P(\mathbf{X}_5 = 2|\mathbf{X}_4 = x_4). \end{aligned}$$

Replacing $x_4 \in \{1, 2\}$ and replace the probabilities based on the values from the tables.

$$\begin{aligned} &P(\mathbf{X}_5 = 2|\mathbf{X}_3 = 1, \mathbf{X}_{11} = 2, \mathbf{X}_1 = 1) \\ =&P(\mathbf{X}_4 = 1|\mathbf{X}_3 = 1)P(\mathbf{X}_5 = 2|\mathbf{X}_4 = 1) + P(\mathbf{X}_4 = 2|\mathbf{X}_3 = 1)P(\mathbf{X}_5 = 2|\mathbf{X}_4 = 2) \\ =&0.1 \times 0.5 + 0.9 \times 0.4 \\ =&0.41 \end{aligned}$$

2 Question 2 (total 10 points)

Now consider the following two Bayesian network structures, where all variables are binary. In other words, they are taking values from $\{1, 2\}$.

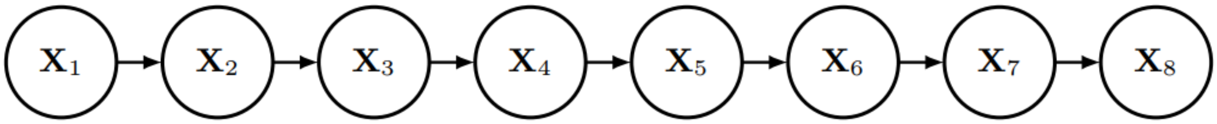

Figure 2: G_1

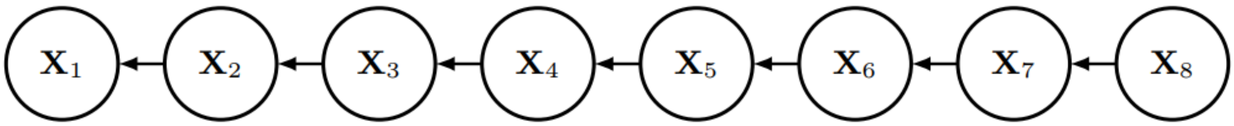

Figure 3: G_2

Now, you would like to use BIC as the criterion for selecting a better structure of the Bayesian network between G_1 and G_2 based on a large collection of samples. Construct a case (i.e., provide a collection of samples, where each sample is of the form "$X_1 = 1, X_2 = 2, \cdots, X_8 = 2$", for example) where the final BIC of the first structure G_1 would be strictly higher than G_2. If you believe no such case exists, clearly explain why.

Answer:

The BIC scores for the graphs G_1 and G_2 are defined as follows:

$$BIC(D; \theta; G_1) = l(D; \theta; G_1) - \frac{dim(G_1)}{2} \log(m), \tag{4}$$

$$BIC(D; \theta; G_2) = l(D; \theta; G_2) - \frac{dim(G_2)}{2} \log(m), \tag{5}$$

where $l(D; \theta; G_1)$ and $l(D; \theta; G_2)$ are the log-likelihood of the two structures, $dim(G_1)$ and $dim(G_2)$ are the number of free parameters of the two structures, and m is the number of data points in the collection of samples.

First, we look at the log-likelihood of the two structures. Instead of the log-likelihood, we calculate the likelihood $P(X_1 = x_1, X_2 = x_2, \cdots, X_8 = x_8)$ for the first structure G1 as follows.

$$
\begin{aligned}
&P(X_1 = x_1, X_2 = x_2, \cdots, X_8 = x_8) && (6)\\
=&P(X_1 = x_1)\prod_{i=2}^{8} P(X_i = x_i | X_{i-1} = x_{i-1}) && (7)\\
=&\frac{\text{Count}(X_1 = x_1)}{\#\text{All-samples}} \prod_{i=2}^{8} \frac{\text{Count}(X_i = x_i; X_{i-1} = x_{i-1})}{\text{Count}(X_{i-1} = x_{i-1})} && (8)\\
=&\frac{\text{Count}(X_1 = x_1)}{\#\text{All-samples}} \frac{\prod_{i=2}^{8}\text{Count}(X_i = x_i; X_{i-1} = x_{i-1})}{\prod_{i=1}^{7}\text{Count}(X_i = x_i)} && (9)\\
=&\frac{1}{\#\text{All-samples}} \frac{\prod_{i=2}^{8}\text{Count}(X_i = x_i; X_{i-1} = x_{i-1})}{\prod_{i=2}^{7}\text{Count}(X_i = x_i)}. && (10)
\end{aligned}
$$

Similarly, we can calculate the likelihood for the second structure G_2 as follows.

$$
\begin{aligned}
&P(X_1 = x_1, X_2 = x_2, \cdots, X_8 = x_8) && (11)\\
=&P(X_8 = x_8)\prod_{i=2}^{8} P(X_{i-1} = x_{i-1} | X_i = x_i) && (12)\\
=&\frac{\text{Count}(X_8 = x_8)}{\#\text{All-samples}} \prod_{i=2}^{8} \frac{\text{Count}(X_{i-1} = x_{i-1}; X_i = x_i)}{\text{Count}(X_i = x_i)} && (13)\\
=&\frac{\text{Count}(X_8 = x_8)}{\#\text{All-samples}} \frac{\prod_{i=2}^{8}\text{Count}(X_{i-1} = x_{i-1}; X_i = x_i)}{\prod_{i=2}^{8}\text{Count}(X_i = x_i)} && (14)\\
=&\frac{1}{\#\text{All-samples}} \frac{\prod_{i=2}^{8}\text{Count}(X_i = x_i; X_{i-1} = x_{i-1})}{\prod_{i=2}^{7}\text{Count}(X_i = x_i)}. && (15)
\end{aligned}
$$

As we can see, the last two Eq. 10, and Eq. 15 are the same. Therefore, the likelihood and the log-likelihood of the two structures are the same for any collection of samples, i.e., $l(D;\theta;G_1) = l(D;\theta;G_2)$.

Second, we look at the number of free parameters of the two structures, i.e., $dim(G_1)$ and $dim(G_2)$. Suppose that each variable X_i can take r_i values, i.e., $X_i \in \{1, 2, \cdots, r_i\}$. The free parameters of the two structures are given as follows:

$$dim(G_1) = (r_1 - 1) + \sum_{i=2}^{8} (r_i - 1) r_{i-1} \tag{16}$$

$$= (r_1 - 1) + \sum_{i=2}^{8} (r_i r_{i-1} - r_{i-1}) \tag{17}$$

$$= r_1 - 1 + \sum_{i=2}^{8} r_i r_{i-1} - \sum_{i=2}^{8} r_{i-1} \tag{18}$$

$$= r_1 - 1 + \sum_{i=1}^{7} r_{i+1} r_i - \sum_{i=1}^{7} r_i \tag{19}$$

$$= \sum_{i=1}^{7} r_i r_{i+1} - \sum_{i=2}^{7} r_i - 1. \tag{20}$$

Similarly,

$$dim(G_2) = (r_8 - 1) + \sum_{i=2}^{8} (r_{i-1} - 1) r_i \tag{21}$$

$$= (r_8 - 1) + \sum_{i=2}^{8} (r_{i-1} r_i - r_i) \tag{22}$$

$$= r_8 - 1 + \sum_{i=2}^{8} r_i r_{i-1} - \sum_{i=2}^{8} r_i \tag{23}$$

$$= -1 + \sum_{i=1}^{7} r_{i+1} r_i - \sum_{i=2}^{7} r_i \tag{24}$$

$$= \sum_{i=1}^{7} r_i r_{i+1} - \sum_{i=2}^{7} r_i - 1. \tag{25}$$

As we can see that the last two Eq. 20, and Eq. 25 are the same. Hence, $dim(G_1) = dim(G_2)$.

With these results, from the two Eq. 4, and Eq. 5, we can conclude that the BIC scores of the two structures G_1 and G_2 for any given collection of samples are the same.