Bayesian Network: Structure Learning

Machine Learning (BITS F464)

Structure Learning Problem

- Bayesian network structure learning concerns learning the DAG in a Bayesian network from data.
- To accomplish this, we need to learn a DAG that satisfies the Markov condition with the probability distribution P that is generating the data.
- Note that we do not know P; all we know are the data.
- The process of learning such a DAG is called model selection.

Score-based learning

- Each DAG is assigned a score, based on how well it fits the data.
- We will look at two different scoring methods:
  - Bayesian score (with Dirichlet priors)
  - BIC score

Bayesian Score I

The following result is known:

Theorem Suppose we are about to repeatedly toss a coin (or perform any repeatable experiment with two outcomes), we assume exchangeability, and we represent our prior belief concerning the probability of heads using a Dirichlet distribution with parameters a and b, where a and b are positive real numbers and $m = a + b$. Let D be data that consist of s heads and t tails in n trials. Then

$$P(D) = \frac{\Gamma(m)}{\Gamma(m+n)} \frac{\Gamma(a+s)\Gamma(b+t)}{\Gamma(a)\Gamma(b)}$$

Bayesian Score II

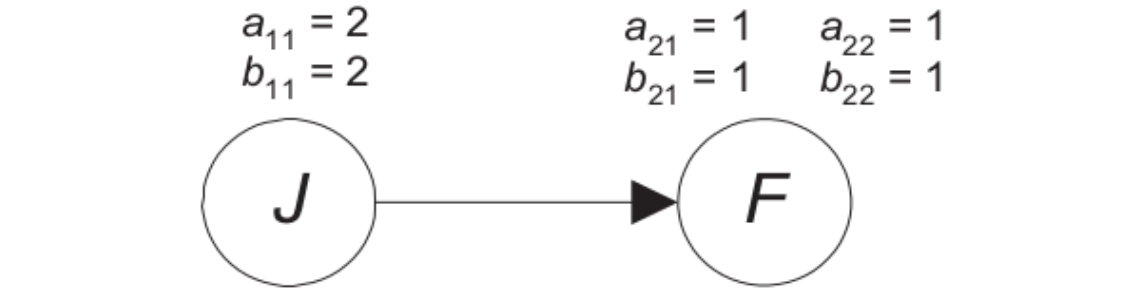
Example Suppose that, before tossing a coin, we assign $a = 3$ and $b = 5$ to model the slight belief that tails is more probable than heads. We then toss the coin 10 times and obtain 4 heads and 6 tails. So,

$$P(D) = \frac{\Gamma(8)}{\Gamma(8+10)} \frac{\Gamma(3+4)\Gamma(5+6)}{\Gamma(3)\Gamma(5)}$$

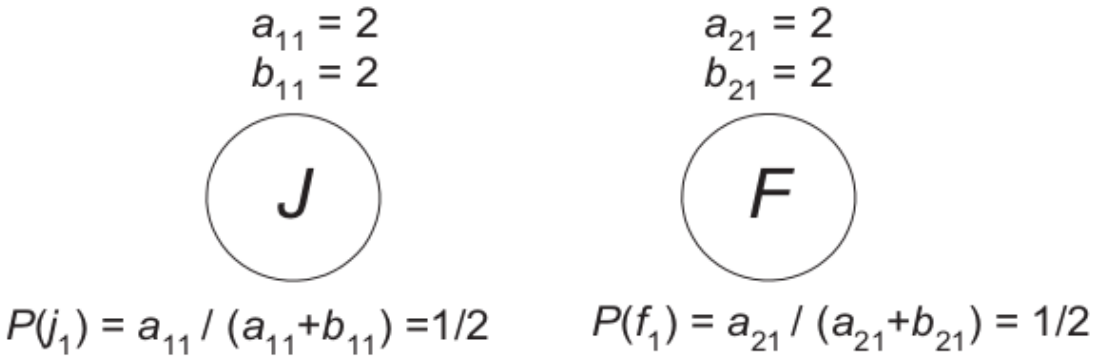
Note: $\Gamma(n) = (n-1)!$

Bayesian Score III

Given the following two DAGs:



(a)



(b)

Bayesian Score IV

Each network is called a 'DAG model'.

We can score a DAG model G based on data D by determining how probable the data are given the DAG model. That is, we compute $P(D|G)$.

Bayesian Score V

So, the probability of data D given the DAG model G_1 (Fig. (a)) is given by

$$P(D|G_1) = \frac{\Gamma(m_{11})}{\Gamma(m_{11}+n_{11})}\frac{\Gamma(a_{11}+s_{11})\Gamma(b_{11}+t_{11})}{\Gamma(a_{11})\Gamma(b_{11})} \times \frac{\Gamma(m_{21})}{\Gamma(m_{21}+n_{21})}\frac{\Gamma(a_{21}+s_{21})\Gamma(b_{21}+t_{21})}{\Gamma(a_{21})\Gamma(b_{21})} \times \frac{\Gamma(m_{22})}{\Gamma(m_{22}+n_{22})}\frac{\Gamma(a_{22}+s_{22})\Gamma(b_{22}+t_{22})}{\Gamma(a_{22})\Gamma(b_{22})}$$

Bayesian Score VI

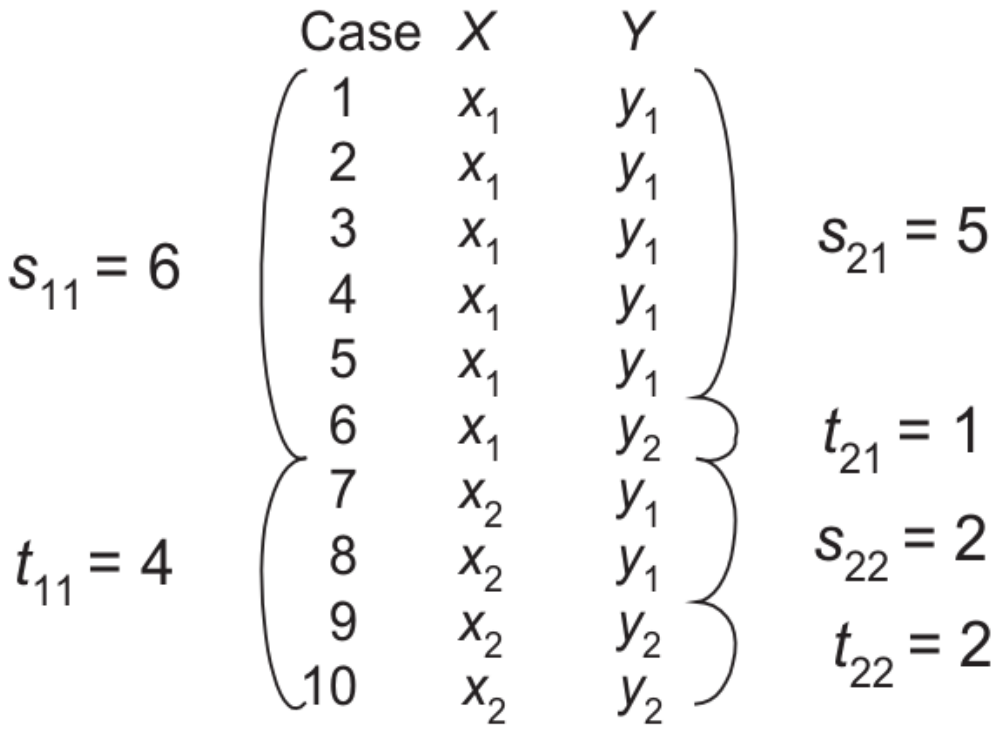
Similarly, the probability of data D given the DAG model G_2 (Fig. (b)) is given by

$$P(D|G_2) = \frac{\Gamma(m_{11})}{\Gamma(m_{11}+n_{11})} \frac{\Gamma(a_{11}+s_{11})\Gamma(b_{11}+t_{11})}{\Gamma(a_{11})\Gamma(b_{11})} \times \frac{\Gamma(m_{21})}{\Gamma(m_{21}+n_{21})} \frac{\Gamma(a_{21}+s_{21})\Gamma(b_{21}+t_{21})}{\Gamma(a_{21})\Gamma(b_{21})}$$

The parameters a, b, s, t are relevant to each DAGs.

Bayesian Score VII

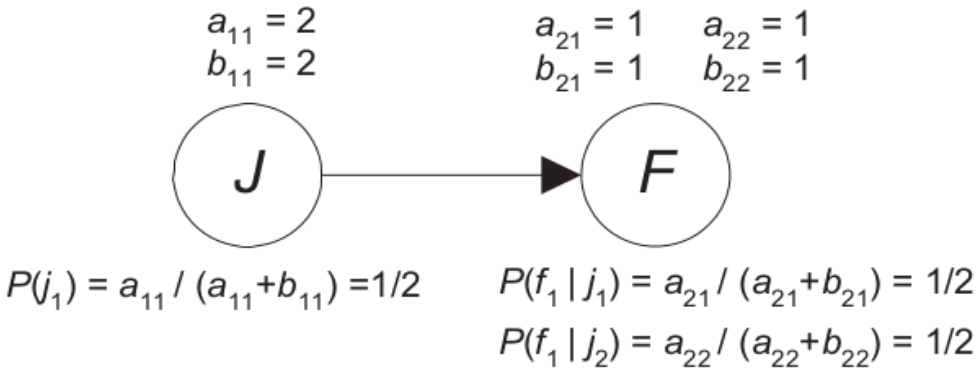
Example of the notations:



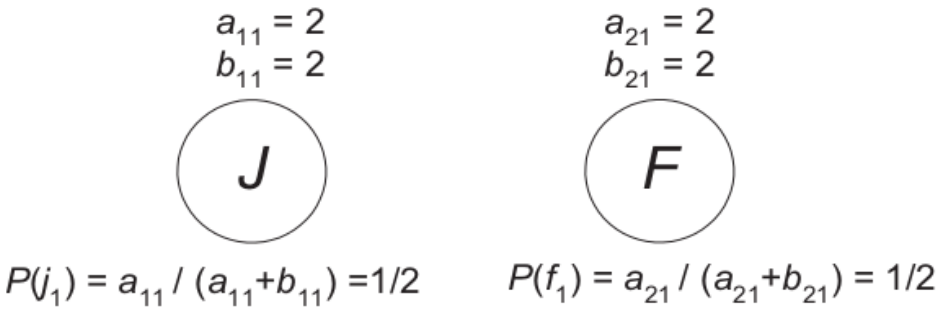
s always represents count of heads, t always represents count of tails. The subscripts represent the r.v.

Bayesian Score VIII

Problem Given two r.v. J and F, each with two outcomes j_1, j_2 and f_1, f_2. Let's assume that we are given two DAG structures: Fig.(a) and Fig.(b) (re-showing here:)



(a)



(b)

Bayesian Score IX

Now, suppose we obtain the following data, D:

Case	J	F
1	j_1	f_1
2	j_1	f_1
3	j_1	f_1
4	j_1	f_1
5	j_1	f_2
6	j_2	f_1
7	j_2	f_2
8	j_2	f_2

Goal: Obtain the DAG (G_1 or G_2) that best-fits D.

Bayesian Score X

G_1: counts are

- $s_{11} = 5,\ t_{11} = 3\ (n_{11} = 8)$
- $s_{21} = 4,\ t_{21} = 1\ (n_{21} = 5)$
- $s_{22} = 1,\ t_{22} = 2\ (n_{22} = 3)$

Bayesian Score XI

So,

$$
\begin{aligned}
P(D|G_1) &= \frac{\Gamma(m_{11})}{\Gamma(m_{11}+n_{11})}\frac{\Gamma(a_{11}+s_{11})\Gamma(b_{11}+t_{11})}{\Gamma(a_{11})\Gamma(b_{11})}\times \\
&\quad \frac{\Gamma(m_{21})}{\Gamma(m_{21}+n_{21})}\frac{\Gamma(a_{21}+s_{21})\Gamma(b_{21}+t_{21})}{\Gamma(a_{21})\Gamma(b_{21})}\times \\
&\quad \frac{\Gamma(m_{22})}{\Gamma(m_{22}+n_{22})}\frac{\Gamma(a_{22}+s_{22})\Gamma(b_{22}+t_{22})}{\Gamma(a_{22})\Gamma(b_{22})} \\
&= \frac{\Gamma(4)}{\Gamma(4+8)}\frac{\Gamma(2+5)\Gamma(2+3)}{\Gamma(2)\Gamma(2)}\times \\
&\quad \frac{\Gamma(2)}{\Gamma(2+5)}\frac{\Gamma(1+4)\Gamma(1+1)}{\Gamma(1)\Gamma(1)}\times \\
&\quad \frac{\Gamma(2)}{\Gamma(2+3)}\frac{\Gamma(1+1)\Gamma(1+2)}{\Gamma(1)\Gamma(1)} \\
&= 7.2150 \times 10^{-6}
\end{aligned}
$$

Bayesian Score XII

Similarly, G_2: counts are:

- $s_{11} = 5,\ t_{11} = 3\ (n_{11} = 8)$
- $s_{21} = 5,\ t_{21} = 3\ (n_{21} = 8)$

(Here, the notation is little confusing. The term s_{21} and t_{21} **do not** have anything to do with first r.v.)

Bayesian Score XIII

$$
\begin{aligned}
P(D|G_2) &= \frac{\Gamma(m_{11})}{\Gamma(m_{11}+n_{11})}\frac{\Gamma(a_{11}+s_{11})\Gamma(b_{11}+t_{11})}{\Gamma(a_{11})\Gamma(b_{11})}\times \\
&\quad \frac{\Gamma(m_{21})}{\Gamma(m_{21}+n_{21})}\frac{\Gamma(a_{21}+s_{21})\Gamma(b_{21}+t_{21})}{\Gamma(a_{21})\Gamma(b_{21})} \\
&= \frac{\Gamma(4)}{\Gamma(4+8)}\frac{\Gamma(2+5)\Gamma(2+3)}{\Gamma(2)\Gamma(2)}\times \\
&\quad \frac{\Gamma(4)}{\Gamma(4+8)}\frac{\Gamma(2+5)\Gamma(2+3)}{\Gamma(2)\Gamma(2)} \\
&= 6.7465 \times 10^{-6}
\end{aligned}
$$

Bayesian Score XIV

If we knew the priors $P(G_1)$ and $P(G_2)$. Then,

$$P(D) = \sum_i P(D|G_i)P(G_i)$$

and

$$P(G_i|D) = \frac{P(D|G_i)P(G_i)}{P(D)}$$

Bayesian Score XV

For our problem, let somebody tells us that:
$P(G_1) = P(G_2) = 0.5$. Then,

$$P(G_1|D) = \frac{7.2150 \times 10^{-6} \times 0.5}{7.2150 \times 10^{-6} \times 0.5 + 6.7465 \times 10^{-6} \times 0.5} = 0.517$$

and

$$P(G_2|D) = \frac{6.7465 \times 10^{-6} \times 0.5}{7.2150 \times 10^{-6} \times 0.5 + 6.7465 \times 10^{-6} \times 0.5} = 0.483$$

Since $P(G_1|D) > P(G_2|D)$, G_1 is learned.

BIC Score I

The Bayesian Information Criterion (BIC) score is given as:

$$BIC(G : D) = \ln\left(P(D|\hat{P}, G)\right) - \frac{d}{2}\ln m$$

where

- m is the number of data items
- d is the dimension of the DAG model
- $\hat{P}$ is the set of maximum likelihood values of the parameters.
- The dimension is the number of parameters in the model.

BIC Score II

Looking closely at the equation:

- The first term of the equation shows how well the model predicts the data when the parameter set is equal to its ML value
- a term that punishes for model complexity

Another nice feature of the BIC is that it does not depend on the prior distribution of the parameters.

BIC Score III

Problem Let's solve the same problem that we solved for Bayesian score mething. D is given below

Case	J	F
1	j_1	f_1
2	j_1	f_1
3	j_1	f_1
4	j_1	f_1
5	j_1	f_2
6	j_2	f_1
7	j_2	f_2
8	j_2	f_2

BIC Score IV

The ML parameters are:

- $\hat{P}(j_1) = \frac{5}{8}$
- $\hat{P}(f_1|j_1) = \frac{4}{5}$, $\hat{P}(f_1|j_2) = \frac{1}{3}$

Because there are 3 parameters in the model, the dimension $d = 3$.

BIC Score V

Now,

$$
\begin{aligned}
P(D|\hat{P}, G_1) &= \left[\hat{P}(f_1|j_1)\hat{P}(j_1)\right]^4 \left[\hat{P}(f_2|j_1)\hat{P}(j_1)\right] \\
&\qquad \left[\hat{P}(f_1|j_2)\hat{P}(j_2)\right] \left[\hat{P}(f_2|j_2)\hat{P}(j_2)\right]^2 \\
&= \left[\frac{4}{5}\frac{5}{8}\right]^4 \left[\frac{1}{5}\frac{5}{8}\right] \left[\frac{1}{3}\frac{3}{8}\right] \left[\frac{2}{3}\frac{3}{8}\right] \\
&= 6.1035 \times 10^{-5}
\end{aligned}
$$

So,

$$
BIC(G_1 : D) = \ln\left(6.1035 \times 10^{-5}\right) - \frac{3}{2}\ln 8 = -12.823
$$

BIC Score VI

For the second DAG, we have

- $\hat{P}(j_1) = \frac{5}{8}$
- $\hat{P}(f_1) = \frac{5}{8}$

There are two parameters; $d = 2$.

BIC Score VII

$$
\begin{aligned}
P(D|\hat{P}, G_2) &= \left[\hat{P}(f_1)\hat{P}(j_1)\right]^4 \left[\hat{P}(f_2)\hat{P}(j_1)\right] \left[\hat{P}(f_1)\hat{P}(j_2)\right] \left[\hat{P}(f_2)\hat{P}(j_2)\right]^2 \\
&= \left(\frac{5}{8}\frac{5}{8}\right)^4 \left(\frac{3}{8}\frac{5}{8}\right) \left(\frac{5}{8}\frac{3}{8}\right) \left(\frac{3}{8}\frac{3}{8}\right)^2 \\
&= 2.5292 \times 10^{-5}
\end{aligned}
$$

So,

$$
BIC(G_2 : D) = \ln\left(2.5292 \times 10^{-5}\right) - \frac{2}{2}\ln 8 = -12.644
$$

Note that although the data were more probable given G_1 , G_2 is learned because it is less complex.

BIC Score VIII

- We see that the Bayesian score and the BIC can choose different DAG models.
- The reason is that the dataset is small.
- In the limit they will both choose the same DAG model because the BIC is asymptotically correct*.

*A scoring criterion for DAG models is called asymptotically correct if for a sufficiently large dataset it chooses the DAG that maximizes $P(D|G)$.

How many DAGs can be scored?

- When there are not many variables, we can exhaustively score all possible DAGs. We then select the DAG(s) with the highest score.
- However, when the number of variables is not small, it is computationally unfeasible to find the maximizing DAGs by exhaustively considering all DAG patterns*.

* Number of DAGs with n nodes is $2^{\mathcal{O}(n^2)}$. (Verify this from the literature.)