PROBABILISTIC REASONING

Unit # 12

ACKNOWLEDGEMENT

The material in this presentation is taken from Richard Neapolitan's book "Probabilistic Method for Bioinformatics: With an Introduction to Bayesian Networks"

STRUCTURE LEARNING

Structure Learning consists of learning the DAG in a Bayesian network from data.

We need to learn a DAG that satisfies the Markov condition with the probability distribution P that is generating the data.

We do not know P, all we know are the data.

Two popular ways of structure learning:

- Score-based Structure Learning
- Search based Structure Learning

SCORE-BASED STRUCTURE LEARNING

In score-based structure learning, we assign a score to a DAG based on how well the DAG fits the data.

Two popular scores are:

- Bayesian Score
- Bayesian Information Criterion

PROBABILITY OF DATA

Suppose that we are about to repeatedly toss a thumbtack (or perform any repeatable experiment with two outcomes).

Suppose further that we assume exchangeability, and we represent our prior belief concerning the probability of heads using Dirichlet distribution with parameters a and b, where a and b are positive integers and m = a + b.

Let D be data that consists of s heads and t tails in n trials.

Then

$$P(\mathsf{D}) = \frac{(m-1)!}{(m+n-1)!} \times \frac{(a+s-1)!(b+t-1)!}{(a-1)!(b-1)!}.$$

NOTATIONS

Sometimes the following equation

$$P(\mathsf{D}) = \frac{(m-1)!}{(m+n-1)!} \times \frac{(a+s-1)!(b+t-1)!}{(a-1)!(b-1)!}.$$

is written as

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

Γ denotes the gamma function. When n is an integer ≥ 1, we have

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

EXAMPLE 8.2 (SOURCE: NEAPOLITAN)

Suppose that, before tossing a thumbstack, we assign a=3 and b=5 to model the slight belief that tails is more probable than heads.

We then toss the thumbstack ten times and obtain four heads and six tails.

The probability of obtaining these data D is given by

$$P(D) = \frac{(8-1)!}{(8+10-1)!} \times \frac{(3+4-1)!\,(5+6-1)!}{(3-1)!\,(5-1)!}$$

$$P(\mathrm{D}) = \frac{(m-1)!}{(m+n-1)!} \times \frac{(a+s-1)!(b+t-1)!}{(a-1)!(b-1)!}.$$

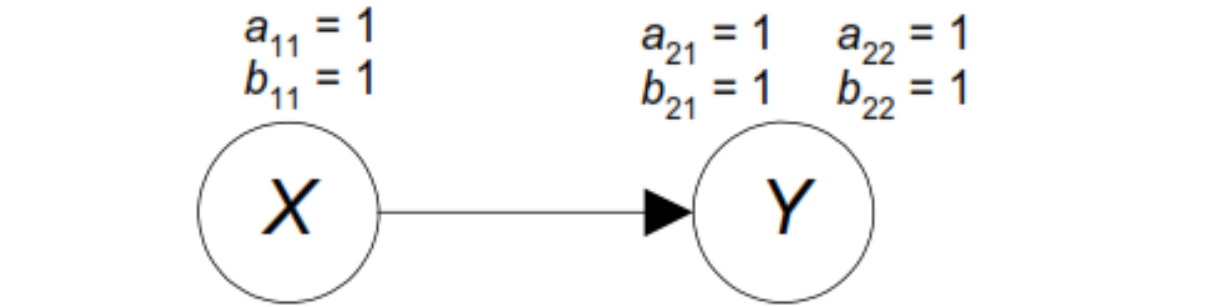
LEARNING DAG MODELS USING BAYESIAN SCORE

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 $(D|G)$.

The formula for this probability is the same as discussed in the previous slides, except there is a term for each probability in the network.

NOTATION

For each probability in the network there is a pair (a_{ij}, b_{ij}). The *i* indexes the variable; the *j* indexes the value of the parent(s) of the variable. For example, the pair (a_{11},b_{11}) is for the first variable (X) and the first value of its parent (in this case there is a default of one parent value since has no parent). The pair (a_{21},b_{21}) is for the second variable (Y) and the first value of its parent, namely x_1 . The pair (namely a_{22},b_{22}) is for the second variable (Y) and the second value of its parent, x_2 . We have attempted to represent prior ignorance as to the value of all probabilities by taking $a_{ij} = b_{ij} = 1$. We compute the prior probabilities using these pairs, just as we did when we were considering a single parameter.



$P(x_1) = a_{11} / (a_{11}+b_{11}) = 1/2$

$P(y_1 \mid x_1) = a_{21} / (a_{21}+b_{21}) = 1/2$

$P(y_1 \mid x_2) = a_{22} / (a_{22}+b_{22}) = 1/2$

DAGS GIVEN DATA

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

Given this data, which DAGs are possible?

LEARNING DAG USING BAYESIAN SCORE (CONT'D)

Let's consider these two structures

$\mathbb{G}_1$

$\mathbb{G}_2$

$$P(\mathsf{D}|\mathbb{G}_1) = \frac{\Gamma(m_{11})}{\Gamma(m_{11}+n_{11})} \times \frac{\Gamma(a_{11}+s_{11})\Gamma(b_{11}+t_{11})}{\Gamma(a_{11})\Gamma(b_{11})} \times \frac{\Gamma(m_{21})}{\Gamma(m_{21}+n_{21})} \times \frac{\Gamma(a_{21}+s_{21})\Gamma(b_{21}+t_{21})}{\Gamma(a_{21})\Gamma(b_{21})} \times \frac{\Gamma(m_{22})}{\Gamma(m_{22}+n_{22})} \times \frac{\Gamma(a_{22}+s_{22})\Gamma(b_{22}+t_{22})}{\Gamma(a_{22})\Gamma(b_{22})}.$$

$$P(\mathsf{D}|\mathbb{G}_2) = \frac{\Gamma(m_{11})}{\Gamma(m_{11}+n_{11})} \times \frac{\Gamma(a_{11}+s_{11})\Gamma(b_{11}+t_{11})}{\Gamma(a_{11})\Gamma(b_{11})} \times \frac{\Gamma(m_{21})}{\Gamma(m_{21}+n_{21})} \times \frac{\Gamma(a_{21}+s_{21})\Gamma(b_{21}+t_{21})}{\Gamma(a_{21})\Gamma(b_{21})}.$$

$$P(\mathsf{D}|\mathbb{G}_1) = \frac{\Gamma(m_{11})}{\Gamma(m_{11}+n_{11})} \times \frac{\Gamma(a_{11}+s_{11})\Gamma(b_{11}+t_{11})}{\Gamma(a_{11})\Gamma(b_{11})} \times$$
$$\frac{\Gamma(m_{21})}{\Gamma(m_{21}+n_{21})} \times \frac{\Gamma(a_{21}+s_{21})\Gamma(b_{21}+t_{21})}{\Gamma(a_{21})\Gamma(b_{21})} \times$$
$$\frac{\Gamma(m_{22})}{\Gamma(m_{22}+n_{22})} \times \frac{\Gamma(a_{22}+s_{22})\Gamma(b_{22}+t_{22})}{\Gamma(a_{22})\Gamma(b_{22})}.$$

G_1

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

$$\frac{(4-1)!}{(4+8-1)!} \times \frac{(2+5-1)!\,(2+3-1)!}{(2-1)!\,(2-1)!} \times \frac{(2-1)!}{(2+5-1)!} \times \frac{(1+4-1)!\,(1+1-1)!}{(1-1)!\,(1-1)!} \times \frac{(2-1)!}{(2+3-1)!} \times \frac{(1+1-1)!\,(1+2-1)!}{(1-1)!\,(1-1)!}$$

$$\frac{(4-1)!}{(4+8-1)!} \times \frac{(2+5-1)!\,(2+3-1)!}{(2-1)!\,(2-1)!} \times \frac{(4-1)!}{(4+8-1)!} \times \frac{(2+5-1)!\,(2+3-1)!}{(2-1)!\,(2-1)!}$$

j_1 f_1 $a_{11} = b_{11} = 2$ $\Big|$ $m_{11} = 4$

j_1 f_2 $a_{21} = b_{21} = a_{22} = b_{22} = 1$

j_2 f_1 $m_{21} = 2$

j_2 f_2 $m_{22} = 2$

$$P(\mathsf{D}|\mathbb{G}_1) = \frac{\Gamma(m_{11})}{\Gamma(m_{11}+n_{11})} \times \frac{\Gamma(a_{11}+s_{11})\Gamma(b_{11}+t_{11})}{\Gamma(a_{11})\Gamma(b_{11})} \times \frac{\Gamma(m_{21})}{\Gamma(m_{21}+n_{21})} \times \frac{\Gamma(a_{21}+s_{21})\Gamma(b_{21}+t_{21})}{\Gamma(a_{21})\Gamma(b_{21})} \times \frac{\Gamma(m_{22})}{\Gamma(m_{22}+n_{22})} \times \frac{\Gamma(a_{22}+s_{22})\Gamma(b_{22}+t_{22})}{\Gamma(a_{22})\Gamma(b_{22})}.$$

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_2	f_2
6	j_2	f_2
7	j_2	f_2
8	j_2	f_2

$$\frac{(4-1)!}{(4+8-1)!} \times \frac{(2+4-1)!\,(2+4-1)!}{(2-1)!\,(2-1)!} \times \frac{(2-1)!}{(2+4-1)!} \frac{(1+4-1)!\,(1+0-1)!}{(1-1)!\,(1-1)!} \times \frac{(2-1)!}{(2+4-1)!} \frac{(1+0-1)!\,(1+4-1)!}{(1-1)!\,(1-1)!}$$

$$\frac{(4-1)!}{(4+8-1)!} \frac{(2+4-1)!\,(2+4-1)!}{(2-1)!\,(2-1)!} \times \frac{(4-1)!}{(4+8-1)!} \frac{(2+4-1)!\,(2+4-1)!}{(2-1)!\,(2-1)!}$$

$a_{11} = b_{11} = 1/2$

$a_{21} = b_{21} = a_{22} = b_{22} = 1$

$m_{21} = 2$, $m_{22} = 2$

SCORING DAGS

$$
\begin{aligned}
P(\mathsf{D}|\mathbb{G}_1) &= \frac{\Gamma(m_{11})}{\Gamma(m_{11}+n_{11})} \times \frac{\Gamma(a_{11}+s_{11})\Gamma(b_{11}+t_{11})}{\Gamma(a_{11})\Gamma(b_{11})} \times \\
&\quad \frac{\Gamma(m_{21})}{\Gamma(m_{21}+n_{21})} \times \frac{\Gamma(a_{21}+s_{21})\Gamma(b_{21}+t_{21})}{\Gamma(a_{21})\Gamma(b_{21})} \times \\
&\quad \frac{\Gamma(m_{22})}{\Gamma(m_{22}+n_{22})} \times \frac{\Gamma(a_{22}+s_{22})\Gamma(b_{22}+t_{22})}{\Gamma(a_{22})\Gamma(b_{22})} \\
&= \frac{\Gamma(4)}{\Gamma(4+8)} \times \frac{\Gamma(2+5)\,\Gamma(2+3)}{\Gamma(2)\Gamma(2)} \times \\
&\quad \frac{\Gamma(2)}{\Gamma(2+5)} \times \frac{\Gamma(1+4)\,\Gamma(1+1)}{\Gamma(1)\Gamma(1)} \times \\
&\quad \frac{\Gamma(2)}{\Gamma(2+3)} \times \frac{\Gamma(1+1)\,\Gamma(1+2)}{\Gamma(1)\Gamma(1)} \\
&= 7.2150 \times 10^{-6}.
\end{aligned}
$$

$$
\begin{aligned}
P(\mathsf{D}|\mathbb{G}_2) &= \frac{\Gamma(m_{11})}{\Gamma(m_{11}+n_{11})} \times \frac{\Gamma(a_{11}+s_{11})\Gamma(b_{11}+t_{11})}{\Gamma(a_{11})\Gamma(b_{11})} \times \\
&\quad \frac{\Gamma(m_{21})}{\Gamma(m_{21}+n_{21})} \times \frac{\Gamma(a_{21}+s_{21})\Gamma(b_{21}+t_{21})}{\Gamma(a_{21})\Gamma(b_{21})} \times \\
&= \frac{\Gamma(4)}{\Gamma(4+8)} \times \frac{\Gamma(2+5)\,\Gamma(2+3)}{\Gamma(2)\Gamma(2)} \times \\
&\quad \frac{\Gamma(4)}{\Gamma(4+8)} \times \frac{\Gamma(2+5)\,\Gamma(2+3)}{\Gamma(2)\Gamma(2)} \\
&= 6.7465 \times 10^{-6}.
\end{aligned}
$$

SCORING DAGS

If our prior belief is that neither model is more probable than the other, we assign

$$P(\mathbb{G}_1) = P(\mathbb{G}_2) = .5.$$

Then, owing to Bayes' Theorem,

$$\begin{aligned} P(\mathbb{G}_1|\mathsf{D}) &= \frac{P(\mathsf{D}|\mathbb{G}_1)P(\mathbb{G}_1)}{P(\mathsf{D}|\mathbb{G}_1)P(\mathbb{G}_1) + P(\mathsf{D}|\mathbb{G}_2)P(\mathbb{G}_2)} \\ &= \frac{7.2150 \times 10^{-6} \times .5}{7.2150 \times 10^{-6} \times .5 + 6.7465 \times 10^{-6} \times .5} \\ &= .517 \end{aligned}$$

and

$$\begin{aligned} P(\mathbb{G}_2|\mathsf{D}) &= \frac{P(\mathsf{D}|\mathbb{G}_2)P(\mathbb{G}_2)}{P(\mathsf{D})} \\ &= \frac{6.7465 \times 10^{-6}(.5)}{7.2150 \times 10^{-6} \times .5 + 6.7465 \times 10^{-6} \times .5} \\ &= .483. \end{aligned}$$

EXERCISE

What is the score of two DAGS under consideration given the data is:

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_2	f_2
6	j_2	f_2
7	j_2	f_2
8	j_2	f_2

$$P(\mathsf{D}) = \frac{(m-1)!}{(m+n-1)!} \times \frac{(a+s-1)!(b+t-1)!}{(a-1)!(b-1)!}.$$

EXAMPLES

Compute $P(\mathbb{G}_1 \mid D)$ and $P(\mathbb{G}_2 \mid D)$ for the following cases. Assume $P(\mathbb{G}_1) = P(\mathbb{G}_2) = 0.5$ and assume a prior equivalent sample size of 4.

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

(0.517, 0.483)

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_2	f_2
6	j_2	f_2
7	j_2	f_2
8	j_2	f_2

(0.959, 0.041)

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

(0.331,0.661)

BAYESIAN INFORMATION CRITERIA

BAYESIAN INFORMATION CRITERION (BIC)

The Bayesian information criterion (BIC) score is as follows:

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

where m is the number of data items and d is the number of parameters in the DAG model.

ADVANTAGES OF BIC

The BIC score is intuitively appealing because it contains

1. a term that shows how well the model predicts the data when the parameter set is equal to its ML value, and
2. 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, which means there is no need to assess one.

EXAMPLES REVISITED (USING BIC SCORE)

Compute BIC (G1: D) and BIC(G2: 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

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_2	f_2
6	j_2	f_2
7	j_2	f_2
8	j_2	f_2

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

COMPUTING BIC

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

$$
\begin{aligned}
& P(\mathsf{D}|\hat{\mathsf{P}}, \mathbb{G}_1) \\
&= \left[\hat{P}(f_1|j_1)\hat{P}(j_1)\right]^4 \left[\hat{P}(f_2|j_1)\hat{P}(j_1)\right] \left[\hat{P}(f_1|j_2)\hat{P}(j_2)\right] \left[\hat{P}(f_2|j_2)\hat{P}(j_2)\right]^2 \\
&= \left(\frac{4}{5}\frac{5}{8}\right)^4 \left(\frac{1}{5}\frac{5}{8}\right) \left(\frac{1}{3}\frac{3}{8}\right) \left(\frac{2}{3}\frac{3}{8}\right)^2 \\
&= 6.1035 \times 10^{-5},
\end{aligned}
$$

and therefore

$$
\begin{aligned}
BIC\,(\mathbb{G}_1 : \mathsf{D}) &= \ln\left(P(\mathsf{D}|\hat{\mathsf{P}}, \mathbb{G}_1)\right) - \frac{d}{2}\ln m \\
&= \ln\left(6.1035 \times 10^{-5}\right) - \frac{3}{2}\ln 8 \\
&= -12.823.
\end{aligned}
$$

COMPUTING BIC

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

$$
\begin{aligned}
& P(\mathsf{D}|\hat{\mathsf{P}}, \mathbb{G}_2) \\
&= \left[\hat{P}(f_1)\hat{P}(j_1)\right]^4 \left[\hat{P}(f_2|)\hat{P}(j_1)\right] \left[\hat{P}(f_1)\hat{P}(j_2)\right] \left[\hat{P}(f_2)\hat{P}(j_2)\right]^2 \\
&= \left(\frac{5}{8}\frac{5}{8}\right)^4 \left(\frac{3}{8}\frac{5}{8}\right) \left(\frac{5}{8}\frac{3}{8}\right) \left(\frac{3}{8}\frac{3}{8}\right)^2 \\
&= 2.529\,2 \times 10^{-5},
\end{aligned}
$$

and therefore

$$
\begin{aligned}
BIC\,(\mathbb{G}_2 : \mathsf{D}) &= \ln\left(P(\mathsf{D}|\hat{\mathsf{P}}, \mathbb{G}_2)\right) - \frac{d}{2}\ln m \\
&= \ln\left(2.529\,2 \times 10^{-5}\right) - \frac{2}{2}\ln 8 \\
&= -12.644.
\end{aligned}
$$

CONSISTENT SCORING CRITERION

A consistent scoring criterion for DAG models has the following two properties:

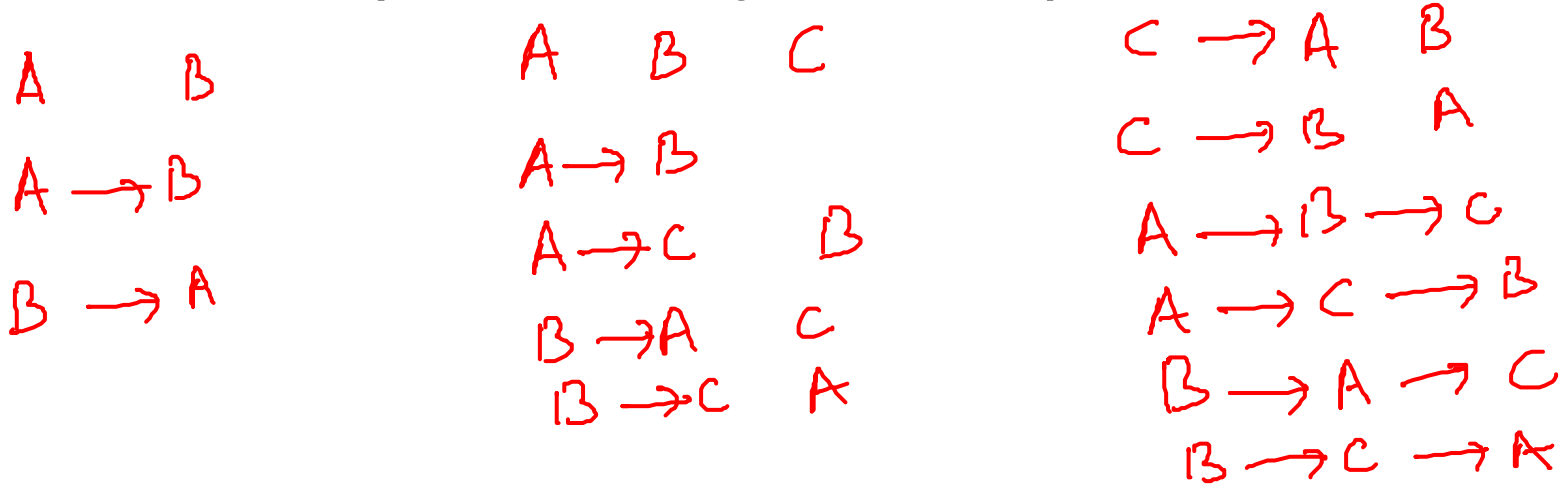
- As the size of the data set approaches infinity, the probability approaches one that a DAG that includes P will score higher than a DAG that does not include P.
- As the size of the data set approaches infinity, the probability approaches one that a smaller DAG that includes P will score higher than a larger DAG that includes P.

HOW MANY DAGS TO SCORE?

NUMBER OF POSSIBLE DAGS

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.



FORMULA FOR NUMBER OF POSSIBLE DAGS

$$
\begin{aligned}
f(n) &= \sum_{i=1}^{n} (-1)^{i+1} \binom{n}{i} 2^{i(n-i)} f(n-i) \qquad n > 2 \\
f(0) &= 1 \\
f(1) &= 1
\end{aligned}
$$

f(2) = 3, f(3) = 25, f(5) = 29,000 and f(10) = 4.2 x 10^{18}

SEARCH ALGORITHMS

K2 ALGORITHM

K2 ALGORITHM

An ordering of the nodes is assumed such that if X_i precedes X_j in the order, an arc from X_j to X_i is not allowed.

Let Pred(X_i) be the set of nodes that precede X_i in the ordering.

Nodes will be visited in sequence according to the ordering.

Initially the set PA_i of parents of X_i is assumed to be empty.

The K2 algorithm incrementally adds parents to each node, as long as it increases the global score. When adding parents to any node does increase the score, the search stops.

Also, given a causal ordering it guarantees that there are not cycles in the graph.

K2 ALGORITHM

Algorithm 8.1 The K2 Algorithm

Require: Set of variables X with a causal ordering, scoring function S, and maximum parents u
Ensure: Set of parents for each variable, $Pa(X_i)$

```
for i = 1 to n do
    oldScore = S(i, Pa(X_i))
    incrementScore = true
    Pa(X_i) = ∅
    while incrementScore and |Pa(X_i)| < u do
        let Z be the node in Predecessors(X_i) − Pa(X_i) that maximizes S
        newScore = S(i, Pa(X_i) ∪ Z)
        if newScore > oldScore then
            oldScore = newScore
            Pa(X_i) = Pa(X_i) ∪ Z
        else
            incrementScore = false
        end if
    end while
end for
return Pa(X_1), Pa(X_2) . . . Pa(X_n)
```

FINDING NODE ORDERING

You might wonder where we could obtain the ordering required by K2.

Such an ordering could possibly be obtained from domain knowledge such as a time ordering of the variables.

For example, we might know that in patients, smoking precedes bronchitis and lung cancer and that each of these conditions precedes fatigue and a positive chest X-ray.

ALGORITHM WITHOUT A PRIOR ORDERING

The following greedy search algorithm that does not require a time ordering.

The search space is again the set of all DAGs containing the n variables.

Following operations are allowed.

1. If two nodes are not adjacent, add an edge between them in either direction.
2. If two nodes are adjacent, remove the edge between them.
3. If two nodes are adjacent, reverse the edge between them.

ALGORITHM WITHOUT A PRIOR ORDERING (CONT'D)

Problem: Find a DAG that approximates maximizing $score(\mathbb{G} : \mathsf{D})$.

Inputs: A set V of n random variables; data D.

Outputs: A set of edges E in a DAG that approximates maximizing $score(\mathbb{G} : \mathsf{D})$.

```
void DAG_search (set_of_variables V, data D,
                 set_of_edges& E)
{
  E = ∅; G = (V, E);
  do
    if (any DAG in the neighborhood of our current DAG
    increases score(G : D))
      modify E according to the one that increases score(G : D) the most;
  while (some operation increases score(G : D));
}
```

HILL CLIMBING APPROACH (STRUCTURE LEARNING)

1. Generate an initial structure-tree.
2. Calculate the fitness measure of the initial structure.
3. Add/invert an arc from the current structure.
4. Calculate the fitness measure of the new structure.
5. If the fitness improves, keep the change; if not, return to the previous structure.
6. Repeat 3–5 until no further improvements exist.

THANKS