CSCI 4152/6509 — Natural Language Processing

26-Oct-2009

Lecture 18: Bayesian Belief Networks

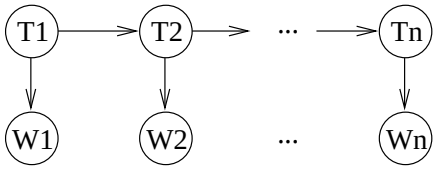
Room: FASS 2176
Time: 11:35 – 12:25

Previous Lecture

- Smoothing:
  - Add-one (Laplace) smoothing,
  - Bell-Witten smoothing;
- Hidden Markov Model,
  - graphical representations,
  - assumption,
  - POS example

11.1 HMM POS Example

When using HMMs for POS tagging, we assume that the hidden internal states of the HMM correspond to correct POS tags of words, while the words correspond to generated observed variables. According to this, a sentence of n words would be associated with the following HMM graph:



The variables $W_1, \ldots, W_n$ are assigned to words in the sentence, while variables $T_1, \ldots, T_n$ are assigned POS tags.
All tables of conditional probabilities for $\mathrm{P}(T_2|T_1)$, $\mathrm{P}(T_3|T_2)$, $\ldots$, $\mathrm{P}(T_n|T_{n-1})$, as well as tables $\mathrm{P}(W_1|T_1)$, $\mathrm{P}(W_2|T_2), \ldots, \mathrm{P}(W_n|T_n)$, are equal.
Having this in mind, suppose that we are given the following training data:

```
swat V flies N like P ants N
time N flies V like P an D arrow N
```

To accommodate for unseen words, we can assign a special symbol $\star$ to unknown words, and assume that it occurred 0.5 "times" with each tag.
The following probability tables are generated using this smoothing technique:

T_1	$\mathrm{P}(T_1)$
N	0.5
V	0.5

,

T_{i-1}	T_i	$\mathrm{P}(T_i\|T_{i-1})$
D	N	1
N	P	0.5
N	V	0.5
P	D	0.5
P	N	0.5
V	N	0.5
V	P	0.5

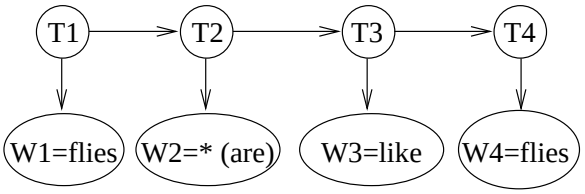
, and

T_i	W_i	$\mathrm{P}(W_i\|T_i)$
D	an	$2/3 \approx 0.666666667$
D	*	$1/3 \approx 0.333333333$
N	ants	$2/9 \approx 0.222222222$
N	arrow	$2/9 \approx 0.222222222$
N	flies	$2/9 \approx 0.222222222$
N	time	$2/9 \approx 0.222222222$
N	*	$1/9 \approx 0.111111111$
P	like	0.8
P	*	0.2
V	flies	0.4
V	swat	0.4
V	*	0.2

.

The above tables may not seem to be complete in the sense that some combinations of variable assignments are not included. For example, $\mathrm{P}(T_i = V|T_{i-1} = D)$ is not included. For the sake of saving space, we have not included the combinations having probability 0, such as this one.

Let us use the Hidden Markov Model to POS tag the sentence "flies are like flies."



The problem of POS tagging in this case is the problem of finding the most probable values of the variables T_i given the values of variables W_i, i.e., it is the completion problem of finding

$$\begin{aligned}\arg\max_T \mathrm{P}(T|W = \text{sentence}) &= \arg\max_T \frac{\mathrm{P}(T, W = \text{sentence})}{\mathrm{P}(W = \text{sentence})} = \arg\max_T \mathrm{P}(T, W = \text{sentence}) \\ &= \arg\max_T \mathrm{P}(T_1) \cdot \mathrm{P}(W_1 = \text{flies}|T_1) \cdot \cdot \mathrm{P}(T_2|T_1) \cdot \mathrm{P}(W_2 = *|T_2) \\ &\quad \cdot \mathrm{P}(T_3|T_2) \cdot \mathrm{P}(W_3 = \text{like}|T_3) \cdot \mathrm{P}(T_4|T_3) \cdot \mathrm{P}(W_4 = \text{flies}|T_4)\end{aligned}$$

where T and W denote arrays of variables T_i and W_i. One way to find the values of T that maximize the given probability is to test all variations of their values. This number is exponential in general, and in this case it is $4^4 = 256$. A much more efficient solution can be obtained by applying a dynamic programming approach, known as the Viterbi algorithm. The idea of the Viterbi algorithm is to incrementally calculate maximal values of the following parts of the above product:

$$\mathrm{P}(T_1) \cdot \mathrm{P}(W_1 = \text{flies}|T_1)$$

for all possible values of T_1, then

$$\mathrm{P}(T_1) \cdot \mathrm{P}(W_1 = \text{flies}|T_1) \cdot \cdot \mathrm{P}(T_2|T_1) \cdot \mathrm{P}(W_2 = *|T_2)$$

for all possible values of T_2 and so on. The computation can be summarized in the following table:

	T_1 ($W_1 =$ flies)	T_2 ($W_2 =$ *)	T_3 ($W_3 =$ like)	T_4 ($W_4 =$ flies)
	$\mathrm{P}(T_1)\mathrm{P}(W_1\|T_1)$	$p \cdot \mathrm{P}(T_2\|T_1)\mathrm{P}(W_2\|T_2)$	$p \cdot \mathrm{P}(T_2\|T_1)\mathrm{P}(W_2\|T_2)$	$p \cdot \mathrm{P}(T_2\|T_1)\mathrm{P}(W_2\|T_2)$
D	$0 \times 0 = 0$	DD: $0 \times 0 \times \frac{1}{3} = 0$ ND: $\frac{1}{9} \times 0 \times \frac{1}{3} = 0$ PD: 0 VD: 0 max: 0	DD: $0 \times 0 \times 0 = 0$ ND: $\frac{1}{90} \times 0\times = 0$ PD: $\frac{1}{50} \times \frac{1}{2} \times 0 = 0$ VD: $\frac{1}{90} \times 0 \times 0 = 0$ max: 0	DD: $0 \times 0 \times 0 = 0$ ND: $0 \times 0 \times 0 = 0$ PD: $\frac{1}{225} \times 0.5 \times 0 = 0$ VD: $0 \times 0 \times 0 = 0$ max: 0
N	$0.5 \times \frac{2}{9} = \frac{1}{9}$	DN: $0 \times 1 \ldots = 0$ NN: $\frac{1}{9} \times 0 \ldots = 0$ PN: $0 \times \ldots = 0$ VN: $0.2\times0.5\times\frac{1}{9} = \frac{1}{90}$ max: $\frac{1}{90}$	DN: $0 \times 1 \times 0 = 0$ NN: $\frac{1}{90} \times 0 \ldots = 0$ PN: $\frac{1}{50}\times0.5\times0 = 0$ VN: $\frac{1}{90} \times 0.5 \times 0 = 0$ max: 0	DN: $0 \times 1 \times \frac{2}{9} = 0$ NN: $0 \times 0 \times \frac{2}{9} = 0$ PN: $\frac{1}{225}\times0.5\times\frac{2}{9} = \frac{1}{2025}$ VN: $0 \times 0.5 \times \frac{2}{9} = 0$ max: $\frac{1}{2025}$
P	$0 \times 0 = 0$	DP: $0 \times \ldots = 0$ NP: $\frac{1}{9}\times0.5\times0.2 = \frac{1}{90}$ PP: $0 \times \ldots = 0$ VP: $0.2\times0.5\times0.2 = \frac{1}{50}$ max: $\frac{1}{50}$	DP: $0 \times 0 \times 0.8 = 0$ NP: $\frac{1}{90} \times 0.5 \times 0.8 = \frac{1}{225}$ PP: $\frac{1}{50} \times 0 \times 0.8 = 0$ VP: $\frac{1}{90}\times0.5\times0.8 = \frac{1}{225}$ max: $\frac{1}{225}$	DP: $0 \times 0 \times 0 = 0$ NP: $0 \times 0.5 \times 0 = 0$ PP: $\frac{1}{225} \times 0 \times 0 = 0$ VP: $0 \times 0.5 \times 0 = 0$ max: 0
V	$0.5 \times 0.4 = 0.2$	DV: $0 \times \ldots = 0$ NV: $\frac{1}{9}\times0.5\times0.2 = \frac{1}{90}$ PV: $0 \times \ldots = 0$ VV: $0.2 \times 0 \ldots = 0$ max: $\frac{1}{90}$	DV: $0 \times 0 \times 0 = 0$ NV: $\frac{1}{90} \times 0.5 \times 0 = 0$ PV: $\frac{1}{50} \times 0 \times 0 = 0$ VV: $\frac{1}{90} \times 0 \times 0 = 0$ max: 0	DV: $0 \times 0 \times 0.4 = 0$ NV: $0 \times 0.5 \times 0.4 = 0$ PV: $\frac{1}{225} \times 0 \times 0.4 = 0$ VV: $0 \times 0 \times 0.4 = 0$ max: 0

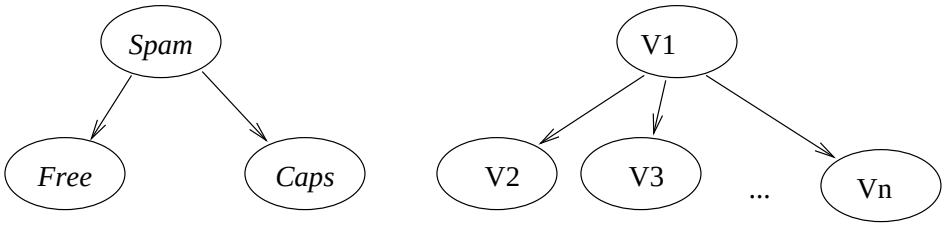
The table is filled column by column. We can see now that the largest value that the expression

$$\mathrm{P}(T_1) \cdot \mathrm{P}(W_1 = \text{flies}|T_1) \cdot \cdot \mathrm{P}(T_2|T_1) \cdot \mathrm{P}(W_2 = *|T_2) \cdot \mathrm{P}(T_3|T_2) \cdot \mathrm{P}(W_3 = \text{like}|T_3) \cdot \mathrm{P}(T_4|T_3) \cdot \mathrm{P}(W_4 = \text{flies}|T_4)$$

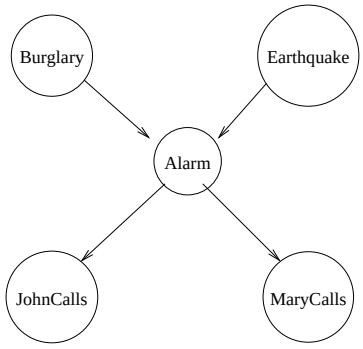
can obtain is $\frac{1}{2025}$, and is achieved with $T_4 =$ N. If we work backwards through the table, we can obtain the optimal values for previous variables as well: $T_3 =$ P, $T_2 =$ V, and $T_1 =$ N. We can also choose $T_2 =$ N, but in this case we have $T_1 =$ V.

12 Bayesian networks

Bayesian Networks (also known as **belief networks, Bayesian belief networks,** or **decision networks**) provide a way to create structured probabilistic models. They make a balance between too strong independence assumptions, as the ones found in Naïve Bayes model or Hidden Markov Model, and the number of parameters of the joint distribution model.
The following graphs shows dependence structure for the Naïve Bayes model of the spam detection, and for the Naïve Bayes model in general:



Examples:
Let us consider the known Burglar-Earthquake example, frequently used in the literature (e.g., the AI textbook by Russell and Norvig):



We can follow a topological sort of the above graph and use the chain rule for the conditional probability, which can be derived from the definition of the conditional probability, to obtain the equation:

$$\mathrm{P}(B, E, A, J, M) = \mathrm{P}(B)\mathrm{P}(E|B)\mathrm{P}(A|B, E)\mathrm{P}(J|A, B, E)\mathrm{P}(M|J, A, B, E)$$

The graph denotes some dependence assumptions, which include: B and E are independent variables, hence $\mathrm{P}(E|B) = \mathrm{P}(E)$, J depends only on A, hence $\mathrm{P}(J|A, B, E) = \mathrm{P}(J|A)$, and M depends only on A, hence $\mathrm{P}(M|J, A, B, E) = \mathrm{P}(M|A)$. If we make these substitutions in the above equation, we obtain the following,
Bayesian network assumption:

$$\mathrm{P}(B, E, A, J, M) = \mathrm{P}(B)\mathrm{P}(E)\mathrm{P}(A|B, E)\mathrm{P}(J|A)\mathrm{P}(M|A)$$

This assumption implies that to evaluate any probability of a complete configuration of the model we need to keep only the parameters for the following conditional probabilities: $\mathrm{P}(B)$, $\mathrm{P}(E)$, $\mathrm{P}(A|B, E)$, $\mathrm{P}(J|A)$, and $\mathrm{P}(M|A)$. These parameters are given in corresponding **conditional probability tables** (CPTs).
Let us assume that the conditional tables for our example are:

B	$\mathrm{P}(B)$
T	0.001
F	0.999

E	$\mathrm{P}(E)$
T	0.002
F	0.998

B	E	A	$\mathrm{P}(A\|B,E)$
T	T	T	0.95
T	T	F	0.05
T	F	T	0.94
T	F	F	0.06
F	T	T	0.29
F	T	F	0.71
F	F	T	0.001
F	F	F	0.999

A	J	$\mathrm{P}(J\|A)$
T	T	0.90
T	F	0.10
F	T	0.05
F	F	0.95

A	M	$\mathrm{P}(M\|A)$
T	T	0.70
T	F	0.30
F	T	0.01
F	F	0.99

– Inference by brute force

By using the tables we can easily compute the probability of any complete configuration. With appropriate summations and by using the definitions of marginal and conditional probability, we can also solve other inference problems. For example, if we want to calculate $\mathrm{P}(B = T|J = T)$, i.e., the probability that a burglar is in the house if John told us over the phone that the alarm is on, we first use the definition of the conditional probability:

$$\mathrm{P}(B = T|J = T) = \frac{\mathrm{P}(B = T, J = T)}{\mathrm{P}(J = T)}$$

The marginal probability $\mathrm{P}(B = T, J = T)$ can be calculated using the formula:

$$\begin{aligned}\mathrm{P}(B = T, J = T) &= \sum_{E,A,M} \mathrm{P}(B = T, E, A, J = T, M) \\ &= \sum_{E,A,M} \mathrm{P}(B = T)\mathrm{P}(E)\mathrm{P}(A|B = T, E)\mathrm{P}(J = T|A)\mathrm{P}(M|A)\end{aligned}$$

Hence,

$$\begin{aligned}\mathrm{P}(B = T, J = T) &= \\ &\quad \mathrm{P}(B = T)\mathrm{P}(E = T)\mathrm{P}(A = T|B = T, E = T)\mathrm{P}(J = T|A = T)\mathrm{P}(M = T|A = T) \\ &+ \mathrm{P}(B = T)\mathrm{P}(E = T)\mathrm{P}(A = T|B = T, E = T)\mathrm{P}(J = T|A = T)\mathrm{P}(M = F|A = T) \\ &+ \mathrm{P}(B = T)\mathrm{P}(E = T)\mathrm{P}(A = F|B = T, E = T)\mathrm{P}(J = T|A = F)\mathrm{P}(M = T|A = F) \\ &+ \mathrm{P}(B = T)\mathrm{P}(E = T)\mathrm{P}(A = F|B = T, E = T)\mathrm{P}(J = T|A = F)\mathrm{P}(M = F|A = F) \\ &+ \mathrm{P}(B = T)\mathrm{P}(E = F)\mathrm{P}(A = T|B = T, E = F)\mathrm{P}(J = T|A = T)\mathrm{P}(M = T|A = T) \\ &+ \mathrm{P}(B = T)\mathrm{P}(E = F)\mathrm{P}(A = T|B = T, E = F)\mathrm{P}(J = T|A = T)\mathrm{P}(M = F|A = T) \\ &+ \mathrm{P}(B = T)\mathrm{P}(E = F)\mathrm{P}(A = F|B = T, E = F)\mathrm{P}(J = T|A = F)\mathrm{P}(M = T|A = F) \\ &+ \mathrm{P}(B = T)\mathrm{P}(E = F)\mathrm{P}(A = F|B = T, E = F)\mathrm{P}(J = T|A = F)\mathrm{P}(M = F|A = F) \\ &= 0.001 \cdot 0.002 \cdot 0.95 \cdot 0.9 \cdot 0.7 \\ &+ 0.001 \cdot 0.002 \cdot 0.95 \cdot 0.9 \cdot 0.3 \\ &+ 0.001 \cdot 0.002 \cdot 0.05 \cdot 0.05 \cdot 0.01 \\ &+ 0.001 \cdot 0.002 \cdot 0.05 \cdot 0.05 \cdot 0.99 \\ &+ 0.001 \cdot 0.998 \cdot 0.94 \cdot 0.9 \cdot 0.7 \\ &+ 0.001 \cdot 0.998 \cdot 0.94 \cdot 0.9 \cdot 0.3 \\ &+ 0.001 \cdot 0.998 \cdot 0.06 \cdot 0.05 \cdot 0.01 \\ &+ 0.001 \cdot 0.998 \cdot 0.06 \cdot 0.05 \cdot 0.99 \\ &= 8.49017 \cdot 10^{-4}\end{aligned}$$

To calculate $\mathrm{P}(J=T)$, we can represent it as $\mathrm{P}(J=T) = \mathrm{P}(B=T, J=T) + \mathrm{P}(B=F, J=T)$ and first calculate $\mathrm{P}(B=F, J=T)$:

$$\begin{aligned}
\mathrm{P}(B=F, J=T) &= \sum_{E,A,M} \mathrm{P}(B=F, E, A, J=T, M) \\
&= \sum_{E,A,M} \mathrm{P}(B=F)\mathrm{P}(E)\mathrm{P}(A|B=F,E)\mathrm{P}(J=T|A)\mathrm{P}(M|A) \\
&= \mathrm{P}(B=F)\mathrm{P}(E=T)\mathrm{P}(A=T|B=F,E=T)\mathrm{P}(J=T|A=T)\mathrm{P}(M=T|A=T) \\
&+ \mathrm{P}(B=F)\mathrm{P}(E=T)\mathrm{P}(A=T|B=F,E=T)\mathrm{P}(J=T|A=T)\mathrm{P}(M=F|A=T) \\
&+ \mathrm{P}(B=F)\mathrm{P}(E=T)\mathrm{P}(A=F|B=F,E=T)\mathrm{P}(J=T|A=F)\mathrm{P}(M=T|A=F) \\
&+ \mathrm{P}(B=F)\mathrm{P}(E=T)\mathrm{P}(A=F|B=F,E=T)\mathrm{P}(J=T|A=F)\mathrm{P}(M=F|A=F) \\
&+ \mathrm{P}(B=F)\mathrm{P}(E=F)\mathrm{P}(A=T|B=F,E=F)\mathrm{P}(J=T|A=T)\mathrm{P}(M=T|A=T) \\
&+ \mathrm{P}(B=F)\mathrm{P}(E=F)\mathrm{P}(A=T|B=F,E=F)\mathrm{P}(J=T|A=T)\mathrm{P}(M=F|A=T) \\
&+ \mathrm{P}(B=F)\mathrm{P}(E=F)\mathrm{P}(A=F|B=F,E=F)\mathrm{P}(J=T|A=F)\mathrm{P}(M=T|A=F) \\
&+ \mathrm{P}(B=F)\mathrm{P}(E=F)\mathrm{P}(A=F|B=F,E=F)\mathrm{P}(J=T|A=F)\mathrm{P}(M=F|A=F) \\
&= 0.999 \cdot 0.002 \cdot 0.29 \cdot 0.9 \cdot 0.7 \\
&+ 0.999 \cdot 0.002 \cdot 0.29 \cdot 0.9 \cdot 0.3 \\
&+ 0.999 \cdot 0.002 \cdot 0.71 \cdot 0.05 \cdot 0.01 \\
&+ 0.999 \cdot 0.002 \cdot 0.71 \cdot 0.05 \cdot 0.99 \\
&+ 0.999 \cdot 0.998 \cdot 0.001 \cdot 0.9 \cdot 0.7 \\
&+ 0.999 \cdot 0.998 \cdot 0.001 \cdot 0.9 \cdot 0.3 \\
&+ 0.999 \cdot 0.998 \cdot 0.999 \cdot 0.05 \cdot 0.01 \\
&+ 0.999 \cdot 0.998 \cdot 0.999 \cdot 0.05 \cdot 0.99 \\
&= 5.12899587 \cdot 10^{-2}
\end{aligned}$$

Now, we calculate

$$\mathrm{P}(J=T) = \mathrm{P}(B=T, J=T) + \mathrm{P}(B=F, J=T) = 8.49017 \cdot 10^{-4} + 5.12899587 \cdot 10^{-2} = 0.0521389757,$$

and finally

$$\mathrm{P}(B=T|J=T) = \frac{\mathrm{P}(B=T, J=T)}{\mathrm{P}(J=T)} = \frac{8.49017 \cdot 10^{-4}}{0.0521389757} = 0.0162837299467699.$$

Even this small example illustrates inefficiency of this approach.

After this informal description of Bayesian Network, here is a more formal definition:

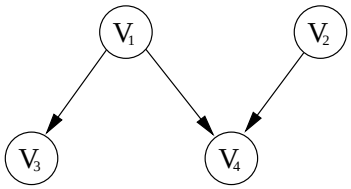
Definition 12.1 (Bayesian Network) *A* Bayesian Network *is defined by a directed acyclic graph (DAG) and a collection of conditional probability tables, where nodes in the graph represent random variables and directed edges in the graph represent conditional independence assumptions. The edges are interpreted in the following way: If V_j ($1 \leq j \leq n$) is a random variable, and $\mathbf{V}_{\pi(j)}$ are parent variables of V_j, i.e., all source nodes for edges whose destination node is V_jm then the probability of V_j given variables $\mathbf{V}_{\pi(j)}$ is independent of any other variable; i.e.,*

$$\mathrm{P}(V_j = x_j | \textit{ all other variables } \} = \mathrm{P}(V_j = x_j | \mathbf{V}_{\pi(j)} = \mathbf{x}_{\pi(j)})$$

Hence, a Bayesian network consists of two components: a directed graph and a set of lookup tables for conditional probabilities for each variable V_i. If $\pi(i)$ are parent nodes of V_i, then for all possible values of variables $\mathbf{V}_{\pi(i)}$ and V_i, the conditional probability table (CPT) specifies the probability of $\mathrm{P}(V_i|\mathbf{V}_{\pi(i)})$. If the number of parent nodes is

k, and if we assume that each variable may have m different values, than the conditional probability table has m^{k+1} rows. For each of m^k combinations of values of parent notes, there is one constraint on the probability distribution $\sum_x \mathrm{P}(V_i = x|\mathbf{V}_{\pi(i)}) = 1$, which will result in m^k constraints; i.e., there are $m^{k+1} - m^k = (m-1)m^k$ free parameters. If the maximal number of parents for a whole network is k, then the total number of free parameters is not grater than $n(m-1)m^k$.

Example. For example, let us calculate the number of free parameters of the following Bayesian Network:



The Bayesian assumption for the network above is:

$$\begin{aligned}&\mathrm{P}(V_1{=}x_1, V_2{=}x_2, V_3{=}x_3, V_4{=}x_4)\\ &\quad = \quad \mathrm{P}(V_1{=}x_1)\ \mathrm{P}(V_2{=}x_2)\ \mathrm{P}(V_3{=}x_3|V_1{=}x_1)\ \mathrm{P}(V_4{=}x_4|V_1{=}x_1, V_2{=}x_2)\end{aligned}$$

How many parameters are needed to represent the network?
For each variable store a conditional probability table of size

$$m \cdot m^{\#\text{parents}} - m^{\#\text{parents}} \quad \text{(constraints)}$$

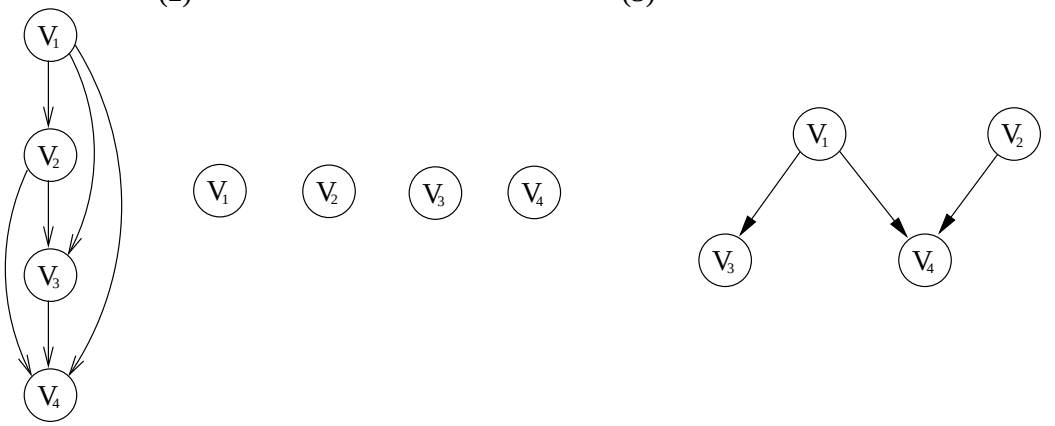
For the above network, the number of free parameters is:

$$\begin{array}{ll} m + m + m^2 + m^3 & \text{parameters} \\ -1 - 1 - m - m^2 & \text{constraints} \\ = m^3 + m - 2 & \end{array}$$

Representational power

Using the Bayesian network model we can represent many other models: the full joint distribution model, fully independent model, Naïve Bayes model, Hidden Markov Model, and other structured models.
Some examples of the Bayesian Networks are:

(1) (2) (3)



Number of parameters in each model:
(1) $(m-1) + (m^2 - m) + (m^3 - m^2) + (m^4 - m^3) = m^4 - 1$
(2) $(m-1) + (m-1) + (m-1) + (m-1) = 4m - 4$
(3) solved above: $m^3 + m - 2$

Example (Burglar-Earthquake cont'd)

The tables for the Burglar-Earthquake example:

B	$\mathrm{P}(B)$
T	0.001
F	0.999

E	$\mathrm{P}(E)$
T	0.002
F	0.998

B	E	A	$\mathrm{P}(A\|B,E)$
T	T	T	0.95
T	T	F	0.05
T	F	T	0.94
T	F	F	0.06
F	T	T	0.29
F	T	F	0.71
F	F	T	0.001
F	F	F	0.999

A	J	$\mathrm{P}(J\|A)$
T	T	0.90
T	F	0.10
F	T	0.05
F	F	0.95

A	M	$\mathrm{P}(M\|A)$
T	T	0.70
T	F	0.30
F	T	0.01
F	F	0.99

12.1 Computational Tasks

Evaluation

The calculate probability of a complete configuration, we multiply corresponding conditional probabilities:

$$\mathrm{P}(V_1{=}x_1, ..., V_n{=}x_n) \quad = \quad \prod_{j=1}^{n} \mathrm{P}(V_j{=}x_j|\mathbf{V}_{\pi(j)}{=}\mathbf{x}_{\pi(j)})$$

Simulation

For $i = 1, ..., n$, draw x_j according to $\mathrm{P}(V_j = x_j|\mathbf{V}_{\pi(j)} = \mathbf{x}_{\pi(j)})$. Conjoin $(x_1, ..., x_n)$ to form a complete configuration.

Learning

Learning with predetermined network graph, and from complete observations can be done by direct MLE; i.e., counting.