12 Structured probability models

12.1 Bayesian networks

Bayesian networks are an important method for representing restricted forms of joint distributions that have certain conditional independence structures. To define a Bayesian network we will exploit the general fact that for *any* joint distribution we have the following *chain rule of probability*

$$\begin{aligned}
&\mathrm{P}(X_1\!=\!x_1, ..., X_n\!=\!x_n) \\
&\quad = \mathrm{P}(X_1\!=\!x_1)\,\mathrm{P}(X_2\!=\!x_2|X_1\!=\!x_1)\,\mathrm{P}(X_3\!=\!x_3|X_2\!=\!x_2, X_1\!=\!x_1)\,\cdots \\
&\qquad \cdots \mathrm{P}(X_n\!=\!x_n|X_{n-1}\!=\!x_{n-1}, ..., X_2\!=\!x_2, X_1\!=\!x_1)
\end{aligned}$$

Definition A *Bayesian network* is defined by a directed acyclic graph (DAG) and a collection of conditional probability tables

- *Nodes* in the graph represent random variables
- *Directed edges* in the graph represent direct dependencies between variables (which indirectly specifies conditional independence assumptions)

Order the variables so that X_j's parents appear before X_j in the graph.
Let $\pi(j)$ denote the indices of the parents of X_j in the graph.

Then the conditional independence assumptions encoded by the graph are: Any random variable X_k is independent of any ancestor variable X_j, $j < k$, given X_k's parents, $\mathbf{X}_{\pi(k)}$. That is,

$$\mathrm{P}(X_k\!=\!x_k|\mathbf{X}_{\pi(k)}\!=\!\mathbf{x}_{\pi(k)}, X_j\!=\!x_j) \quad = \quad \mathrm{P}(X_k\!=\!x_k|\mathbf{X}_{\pi(k)}\!=\!\mathbf{x}_{\pi(k)})$$

for any X_j such that $j < k$.

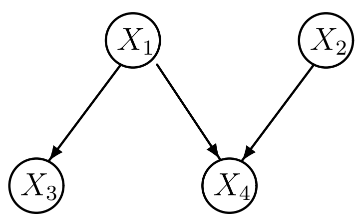
To represent a Bayesian network we first need to store the graph, and then store a lookup table for each variable X_j which represents the conditional probability of X_j given each possible configuration of its parents.

Note that for a random variable X_j, we can represent $\mathrm{P}(X_j\!=\!x_j|\mathbf{X}_{\pi(j)}\!=\!\mathbf{x}_{\pi(j)})$ by a lookup table with $V \times |\pi(j)|$ positive numbers, minus one constraint for each configuration of the parents $\mathbf{X}_{\pi(j)}$. That is, let $\theta_{j,x,\mathbf{v}} = \mathrm{P}(X_j\!=\!x|\mathbf{X}_{\pi(j)}\!=\!\mathbf{v})$. These numbers are positive and satisfy the constraint $\sum_{x=1}^{V} \theta_{j,x,\mathbf{v}} = 1$ for each j and $\mathbf{v}$. Thus, the joint distribution over $X_1, ..., X_n$

can be represented by $\sum_{j=1}^{n} V \times V^{|\pi(j)|}$ positive numbers minus $\sum_{j=1}^{n} V^{|\pi(j)|}$ constraints.

If the maximum number of parents in the graph is bounded by k then this can be a severe restriction on the structure of the joint distribution, since the number of free parameters defining the distribution is reduced from $V^n - 1$ to $n(V-1)V^k$.

12.2 Example



$$\begin{aligned} &\mathrm{P}(X_1{=}x_1, X_2{=}x_2, X_3{=}x_3, X_4{=}x_4) \\ &\quad= \ \mathrm{P}(X_1{=}x_1)\ \mathrm{P}(X_2{=}x_2)\ \mathrm{P}(X_3{=}x_3|X_1{=}x_1)\ \mathrm{P}(X_4{=}x_4|X_1{=}x_1, X_2{=}x_2) \end{aligned}$$

How many parameters to represent?

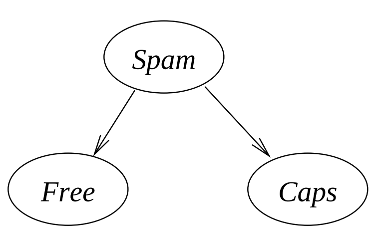
$$\begin{array}{ll} V + V + V^2 + V^3 & \text{parameters} \\ -1 - 1 - V - V^2 & \text{constraints} \end{array}$$

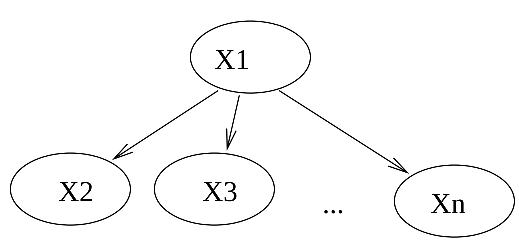
For each variable store a conditional probability table of size

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

12.3 Example: Naive Bayes model

In the Naive Bayes model one assumes that there is a single parent variable and a collection of child variables whose values are conditionally independent from one another given the parent. The following two graphs show the Naive Bayes model applied to the spam detection example, and in general





These graph structures correspond to the assumption

$$\begin{aligned} &\mathrm{P}(X_1 = x_1, \ldots, X_n = x_n) \\ &= \ \mathrm{P}(X_1 = x_1)\ \mathrm{P}(X_2 = x_2 | X_1 = x_1)\ \cdots\ \mathrm{P}(X_n = x_n | X_1 = x_1) \end{aligned}$$

Parameters?

$$\begin{array}{ll} V + V^2 + \cdots + V^2 & \text{parameters} \\ -1 - V - \cdots - V & \text{constraints} \end{array}$$

In the spam detection example, one way to apply the Naive Bayes assumption is to assume $\mathrm{P}(\mathit{Free}, \mathit{Caps}, \mathit{Spam}) = \mathrm{P}(\mathit{Spam})\ \mathrm{P}(\mathit{Free}|\mathit{Spam})\ \mathrm{P}(\mathit{Caps}|\mathit{Spam})$. Assume we have the same sample data as before

Free	*Caps*	*Spam*	# messages
Y	Y	Y	20
Y	Y	N	1
Y	N	Y	5
Y	N	N	0
N	Y	Y	20
N	Y	N	3
N	N	Y	2
N	N	N	49
		Total:	100

Then using direct estimates of the probabilities from this data we obtain

Spam	P(*Spam*)
Y	$\frac{20+5+20+2}{100} = 0.47$
N	$\frac{1+0+3+49}{100} = 0.53$

Caps	*Spam*	P(*Caps*|*Spam*)
Y	Y	$\frac{20+20}{20+5+20+2} \approx 0.8511$
Y	N	$\frac{1+3}{1+0+3+49} \approx 0.0755$
N	Y	$\frac{5+2}{20+5+20+2} \approx 0.1489$
N	N	$\frac{0+49}{1+0+3+49} \approx 0.9245$

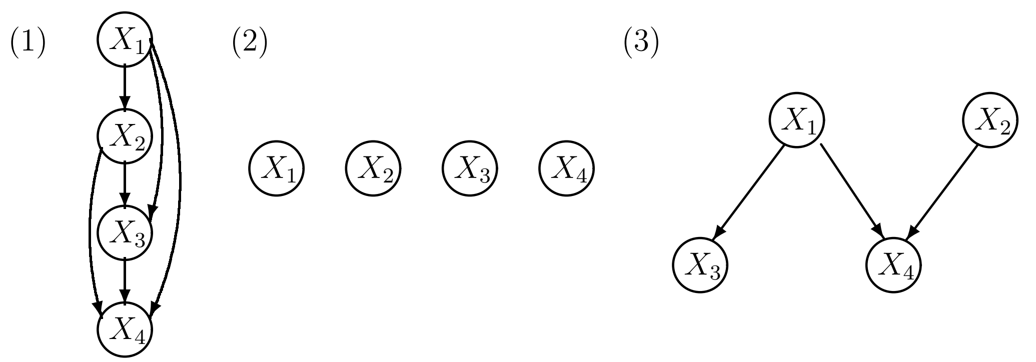
Free	*Spam*	P(*Free*|*Spam*)
Y	Y	$\frac{20+5}{20+5+20+2} \approx 0.5319$
Y	N	$\frac{1+0}{1+0+3+49} \approx 0.0189$
N	Y	$\frac{20+2}{20+5+20+2} \approx 0.4681$
N	N	$\frac{3+49}{1+0+3+49} \approx 0.9811$

The probability of a particular configuration can now be calculated in this model as follows

$$
\begin{aligned}
&\mathrm{P}(\mathit{Free} = Y, \mathit{Caps} = N, \mathit{Spam} = N) \\
&\quad = \mathrm{P}(\mathit{Spam} = N)\ \mathrm{P}(\mathit{Caps} = N|\mathit{Spam} = N)\ \mathrm{P}(\mathit{Free} = Y|\mathit{Spam} = N) \\
&\quad \approx 0.53 \times 0.9245 \times 0.0189 \\
&\quad \approx 0.0093
\end{aligned}
$$

12.4 Representational power

Using a Bayesian network representation one can represent: (1) arbitrary joint distribution, (2) fully independent distribution, and (3) distributions intermediate between these.



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

Bayesian networks cannot represent all possible conditional independence structures, but they are still very useful.

12.5 Elementary tasks

Simulation

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

Evaluation

To compute the probability of a complete configuration, just multiply the local probabilities

$$\mathrm{P}(X_1 = x_1, ..., X_n = x_n) \quad = \quad \prod_{j=1}^{n} \mathrm{P}(X_j = x_j | \mathbf{X}_{\pi(j)} = \mathbf{x}_{\pi(j)})$$

12.6 Inference

For some Bayesian networks inference must be hard (for example, inference with an arbitrary joint model that has an explicit lookup table representation) because the size of the representation is exponentially large in the number of variables n (i.e. a size V^n lookup table). On the other hand, inference in trivial Bayesian networks is easy (such as the complete independent model).

In general, inference (marginalization, conditioning, completion) is NP-hard for Bayesian networks, even if we restrict the graph to at most 2 parents per variable which forces a polynomial size representation. If, however, graph is a *tree*, then efficient (polynomial time) inference algorithms can be derived. This will be the topic of the next lecture.

Readings

Russell and Norvig 2nd Ed: Chapters 14-15
Dean, Allen, Aloimonos: Sect 8.3