MAC6916 PROBABILISTIC GRAPHICAL MODELS
LECTURE 2: BAYESIAN NETWORKS

DENIS D. MAUÁ

1. Motivation

We previously saw that every probabilistic model can be represented by a set of variables and their domains and a joint distribution over all the variables. While this structure reduces significantly the space requirements, it is still exponential in the number of variables. For example, it requires $2^n - 1$ numbers to represent a probability function on a language with binary variables $X_1, \dots, X_n$.[1] Clearly, this doesn't scale to domains with more than a handful of variables. So our goal is to find convenient representations that scale *linearly* with the number of variables, allowing models with hundreds, thousands and even millions of variables to be represented efficiently. The key to achieving this goal is the correspondence between (conditional) independence and factorization. For example, consider the toss of three coins represented by binary variables A, B, C. If we *assume* the variables are independent, then the joint distribution factorizes as $p(A, B, C) = p(A)p(B)p(C)$. Hence, instead of the $2^3 - 1 = 7$ numbers we need only $3 \cdot (2^2 - 1) = 3$ numbers to fully specify the probability function. As we will see, the amount of space saved is directly related to number of independence statements among variables. In this example, we assumed that $\{A, B\} \perp\!\!\!\perp C$, $\{A, C\} \perp\!\!\!\perp B$, $\{B, C\} \perp\!\!\!\perp A$, $A \perp\!\!\!\perp B|C$, $A \perp\!\!\!\perp C|B$, $A \perp\!\!\!\perp B|C$, $A \perp\!\!\!\perp B$ and many others (the list is long!). Some of these independences are induced by others (and can be omitted), while some are not (and need to be explicitly represented). Full independence is easy to characterize, but what about more complex phenomena such as intercausal reasoning: A and B are unconditionally independent but conditionally independent given C. This is the purpose of Bayesian networks: to allow for a compact and intuitive language in which to express independence statements and (consequently) factorized probability functions.

2. A few concepts from graph theory

A *graph* is a pair $(\mathcal{V}, \mathcal{E})$ where $\mathcal{V}$ is a set of vertices or nodes and $\mathcal{E}$ is a binary relation on $\mathcal{V}$. The elements of $\mathcal{E}$ are called edges or arcs. The graph is *undirected* if $\mathcal{E}$ is symmetric, that is, if $(X, Y) \in \mathcal{E}$ if and only if $(Y, X) \in \mathcal{E}$. Otherwise, the graph is *directed*.[2] The edges of directed graphs are also called arcs. Let U

graph

undirected graph

directed graph

Date: February 29, 2016.

[1] The -1 is due to the constraint that the probabilities must add to one, so one of the numbers is determined by the others.

children
parents
neighbors
ancestor
descendant

be a node in $\mathcal{V}$. If $(X, Y) \in \mathcal{E}$ then Y is a child of X, X is a parent of X, and X and Y are neighbors. The set of *children*, *parents* and *neighbors* of a node X are denoted, respectively, as $\mathrm{ch}(X), \mathrm{pa}(X), \mathrm{ne}(X)$. The in-degree of a node u is the number of parents and is written as $\mathrm{indeg}(X) = |\mathrm{pa}(X)|$. The out-degree is $\mathrm{outdeg}(X) = |\mathrm{ch}(X)|$. A node X is an *ancestor* of node Y if X is a parent of Y in the transitive closure of $\mathcal{E}$; to put it differently, X is an ancestor of Y if X is a parent or an ancestor of Y. A node X is a *descendant* x if Y is an ancestor of X. The descendants (ancestors) of X are denoted $\mathrm{de}(X)$ $(\mathrm{an}(X))$. Its complement with respect to $\mathcal{V}$ is called the non-descendants and written $\mathrm{nd}(X)$.

acyclic graph
sparse graph

A graph is *acyclic* if no node is an ancestor of itself. The class of *sparse graphs* is a set of graphs $G_n = (\mathcal{V}, \mathcal{E})$ parametrized by $n = |\mathcal{V}|$ and such that $|\mathcal{E}| = O(n)$. Informally, we say that a graph is sparse if $|\mathcal{E}| \ll |\mathcal{V}|^2$.

A topological ordering of the nodes is a total ordering of the nodes $<$ such that $X < Y$ if and only Y is not an ancestor of X. A graph has a topological ordering if and only if it is acyclic.

3. An Algebra of Functions

Let $\mathcal{V}$ be an *ordered* finite set of variables (with their respective domains), and consider a set of real-valued functions over *ordered* subsets of $\mathcal{X} \subseteq \mathcal{V}$. We write $f(\mathcal{X})$ to denote a function on the domain $\prod_{X \in \mathcal{X}} \mathrm{dom}(X)$, where the product denotes the Cartesian product and is performed according to the ordering of the variables. Hence, there is no ambiguity when writing $f(\mathcal{X})$. We write $f(\mathcal{X}, \mathcal{Y})$ to denote $f(\mathcal{X} \cup \mathcal{Y})$.

product of functions

Consider a function $f(X_1, \ldots, X_n)$ and a valuation ν of the corresponding variables. We denote by $f(X_1, \ldots, X_n)\nu$ the image of the function at the point $(\nu(X_1), \ldots, \nu(X_n))$. The *product of functions* $f(\mathcal{X})$ and $g(\mathcal{Y})$ is the function $h(\mathcal{X}, \mathcal{Y}) = f(\mathcal{X})g(\mathcal{Y})$ such that $h(\mathcal{X}, \mathcal{Y})\nu = [f(\mathcal{X})\nu] \cdot [g(\mathcal{Y})\nu]$. Division of function is defined analogously (taking care of division by zero). The product of a scalar (a real-value) c and a function $f(\mathcal{X})$ is the function $c \cdot f(\mathcal{X})$ such that $[c \cdot f(\mathcal{X})]\nu = c \cdot [f(\mathcal{X})\nu]$. We write $f(\mathcal{X}) = g(\mathcal{Y})$ to denote $f(\mathcal{X})\nu = g(\mathcal{Y})\nu$. In particular, we write $f(\mathcal{X}) = c$ for a real-value c to denote that $f(\mathcal{X})$ is everywhere equal to c (i.e., $f(\mathcal{X})\nu = c$).

variable elimination

The *elimination* of a variable $Y \in \mathcal{X}$ from $f(\mathcal{X})$ is the function $g(\mathcal{Z}) = \sum_Y f(\mathcal{X})$ such that $g(\mathcal{Z})\nu = f(\mathcal{X})\nu + \sum_{\nu' : \nu'(Y) \neq \nu(Y)} f(\mathcal{X})\nu'$, where $\mathcal{Z} = \mathcal{X} - \{Y\}$ and the sum is carried over the valuations which differ from ν w.r.t. the assignment of Y. For $Y \notin \mathcal{X}$, we define $\sum_Y f(\mathcal{X}) = f(\mathcal{X})$. Other operations such as maximization can be defined analogously (e.g. $\max_{\mathcal{Y}} f(\mathcal{X})$).

Many arithmetic properties of the reals extend to real-valued functions on product spaces. For example, product of functions and elimination are associative and commutative. Hence, we can unambiguously write a product of several functions as $\prod_i f_i(\mathcal{X}_i)$ and a sequence of eliminations of variables $Y_1, \ldots, Y_n$ as $\sum_{\mathcal{Y}} f(\mathcal{X})$, where $\mathcal{Y} = \{Y_1, \ldots, Y_n\}$. Distributivity also holds on functions: if $\mathcal{X} \cap \mathcal{Y} = \emptyset$ then $\sum_{\mathcal{Y}} f(\mathcal{X})g(\mathcal{Z}) = f(\mathcal{X}) \sum_{\mathcal{Y}} g(\mathcal{Z})$.

[2] Undirected graphs are also commonly defined as a set of nodes and a set of unordered pairs of nodes. Under this definition, there is a single edge connecting two nodes in a undirected graph. Many definitions are made simpler by defining undirected graphs this way.

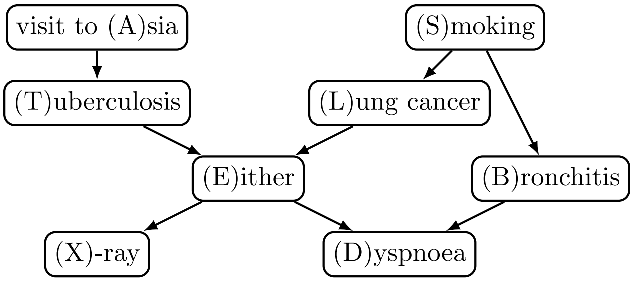

Figure 1. The structure (graph) of the Asia Bayesian network.

4. Bayesian Networks

Consider a probability function $\mathbb{P}$ defined over a language $\mathcal{L}$ with variables in a set $\mathcal{V}$. We write $p(\mathcal{X}|\mathcal{Y})$ to denote the respective conditional distribution of variables $\mathcal{X}$ given $\mathcal{Y}$ induced by $\mathbb{P}$. To avoid ambiguity we assume that the variables in $\mathcal{V}$ are ordered, and that distributions $p(\mathcal{X})$ are defined over the corresponding ordered subsets.

Definition 1. A probabilistic model satisfies (or respects) the *Markov property* represented by (or in) an acyclic directed graph (DAG) $G = (\mathcal{V}, \mathcal{E})$ if each variable X is conditionally independent of its non-descendant non-parents given its parents (i.e., $X \perp\!\!\!\perp \text{nd}(X) - \text{pa}(X)|\text{pa}(X)$).

Markov property

We often notate the Markov property as $p(X|\text{nd}(X)) = p(X|\text{pa}(X))$. Alternatively, the Markov property can be stated as $p(X, \text{nd}(X) - \text{pa}(X)|\text{pa}(X)) = p(X|\text{pa}(X))p(\text{nd}(X)|\text{pa}(X))$.

Definition 2. A *Bayesian network* is a triple $(G, \mathcal{L}, \mathbb{P})$ where

Bayesian network

(BN1) G is a acyclic directed graph with node set $\mathcal{V}$;
(BN2) $\mathcal{L}$ is a generalized propositional language with variables $\mathcal{V}$;
(BN3) The underlying probabilistic model satisfies the Markov property in G.

The graph G is often called the *structure* of the Bayesian network. The distribution $p(\mathcal{V})$ induced by the probability function is called the *joint probability distribution* of the network.

structure

joint probability distribution

Example 1 (Asia network)**.** *Taken from [Cowell et al. 1999]. The chances of developing* dyspnoea *are higher on patients that had* tuberculosis*,* lung cancer *or* bronchitis*. A recent* visit to Asia *increases the chances of tuberculosis, while* smoking *is a risk factor for both* lung cancer *and* bronchitis*. An* x-ray *does not distinguish between* lung cancer *and* tuberculosis*, as neither the presence or absence of* dyspnoea*. The graph in Figure 1 represents the Markov properties implied by the previous statements.*

The following result shows that $\mathbb{P}$ can be efficiently encoded if G is sparse. Define $p(X|\text{pa}(X))$ to be zero at any point where $p(\text{pa}(X))$ is zero. Note that if $p(\mathcal{X})$ is zero at some point that $p(\mathcal{X}, \mathcal{Y})$ is also zero at any consistent point.

Theorem 1 (Factorization Theorem)**.** *If $(G, \mathcal{L}, \mathbb{P})$ is a Bayesian network then*

$$p(\mathcal{V}) = \prod_{X \in \mathcal{V}} p(X|pa(X)) \,,$$

where $p(X|\emptyset) = p(X)$.

Before proving that result, we need the following proposition.

Proposition 1 (Decomposition)**.** *If sets of variables $\mathcal{X}$ and $\mathcal{Y}$ are conditionally independent given set $\mathcal{Z}$, and $\mathcal{S} \subset \mathcal{Y}$ then $\mathcal{X}$ and $\mathcal{S}$ are conditionally independent given $\mathcal{Z}$.*

Proof. Let $\mathcal{T} = \mathcal{Y} - \mathcal{S}$. We have by the total rule and the chain rule for variables that

$$p(\mathcal{X}, \mathcal{S}|\mathcal{Z}) = \sum_{\mathcal{T}} p(\mathcal{X}, \mathcal{S}, \mathcal{T}|\mathcal{Z}) = \sum_{\mathcal{T}} p(\mathcal{X}, \mathcal{Y}|\mathcal{Z}) = \sum_{\mathcal{T}} p(\mathcal{X}|\mathcal{Y}, \mathcal{Z}) p(\mathcal{Y}|\mathcal{Z}) .$$

Since $\mathcal{X}$ and $\mathcal{Y}$ are independent given $\mathcal{Z}$, it follows that $p(\mathcal{X}|\mathcal{Y}, \mathcal{Z}) = p(\mathcal{X}|\mathcal{Z})$. Hence,

$$p(\mathcal{X}, \mathcal{S}|\mathcal{Z}) = \sum_{\mathcal{T}} p(\mathcal{X}|\mathcal{Z}) p(\mathcal{Y}|\mathcal{Z}) = p(\mathcal{X}|\mathcal{Z}) \sum_{\mathcal{T}} p(\mathcal{Y}|\mathcal{Z}) = p(\mathcal{X}|\mathcal{Z}) p(\mathcal{S}|\mathcal{Z}) .$$

The second equality used the fact that $p(\mathcal{X}|\mathcal{Z})$ is constant with respect to $\mathcal{T}$. □

We can now prove the theorem.

Proof of Theorem 1. Let $<$ be a total topological ordering of the nodes/variables in G and for each X denote by $\mathcal{A}_X = \{Y : Y < X\}$ the set of variables that are smaller than X. We have by the Chain Rule for variables that

$$p(\mathcal{V}) = \prod_{X \in \mathcal{V}} p(X|\mathcal{A}_X) .$$

For each X, $\mathcal{A}_X$ is a subset of $\mathrm{nd}(X)$ and contains $\mathrm{pa}(X)$. Hence, by applying Proposition 1 with $\mathcal{X} = \{X\}$, $\mathcal{Y} = \mathrm{nd}(X) - \mathrm{pa}(X)$, $\mathcal{Z} = \mathrm{pa}(X)$ and $\mathcal{S} = \mathcal{A}_X - \mathrm{pa}(X)$, it follows that $p(X|\mathcal{A}_X) = p(X|\mathrm{pa}(X))$. □

local distributions

The distributions $p(X|\mathrm{pa}(X))$ are called *local distributions*. Theorem 1 allows us to specify a probabilistic model by

$$\sum_{X \in \mathcal{V}} (|\mathrm{dom}(X)| - 1) \prod_{Y \in \mathrm{pa}(X)} |\mathrm{dom}(Y)|$$

numbers. This is much smaller than the number of numbers required to specify the joint probability when G is sparse. For instance if k is an upper bound on the in-degree of any node and v is an upper bound on the cardinality of (the domain of) any variable, then the number of numbers required to specify a Bayesian network over n variables is $O(n \cdot v^k)$, which is polynomial in the input size.

Example 2 (Asia cont'd)**.** *The joint (variable) distribution of any probabilistic models satisfying the Markov property with respect to the graph in Figure 1 factorizes as*

$$p(A, S, T, L, E, B, X, D) = p(A)p(S)p(T|A)p(L|S)p(E|T, L)p(B|S)p(X|E)p(D|E, B).$$

Considering that all variables are binary, this specification requires $1 + 1 + 1 \cdot 2 + 1 \cdot 2 + 1 \cdot 2 \cdot 2 + 1 \cdot 2 + 1 \cdot 2 + 1 \cdot 2 \cdot 2 = 18$ *values (compare with the* $2^8 - 1 = 255$ *values required by the joint distribution).*

barren node

Definition 3. A node X in an acyclic directed graph G is *barren* w.r.t. to a set of variables/nodes $\mathcal{S}$ if $(\{X\} \cup \mathrm{de}(X)) \cap \mathcal{S} = \emptyset$.

A node is barren w.r.t. $\mathcal{S}$ if it is a leaf node not in $\mathcal{S}$ in G or any graph obtained from G by a sequence of removal of barren nodes.

Proposition 2. *Let p_G be a distribution that satisfies the Markov properties in a graph $G = (\mathcal{V}, \mathcal{E})$, $\mathcal{B}$ be a set of barren nodes in G with respect to a node set $\mathcal{S}$, and $H = (\mathcal{V} - \mathcal{B}, \mathcal{E}')$ be the graph obtained by removing the nodes (and respective arcs) in $\mathcal{B}$ from G. Then $p_G(\mathcal{S}) = p_H(\mathcal{S})$, where $p_H(\mathcal{V} - \mathcal{B}) = \prod_{X \in \mathcal{V}-\mathcal{B}} p_G(X|pa(X))$.*

Proof. First note that

$$\begin{aligned} p_G(\mathcal{S}) &= \sum_{\mathcal{V}-\mathcal{S}} \prod_{X \in \mathcal{V}} p_G(X|\mathrm{pa}(X)) \\ &= \sum_{\mathcal{V}-\mathcal{S}-\mathcal{B}} \prod_{X \in \mathcal{V}-\mathcal{B}} p_G(X|\mathrm{pa}(X)) \sum_{\mathcal{B}} \prod_{X \in \mathcal{B}} p_G(X|\mathrm{pa}(X)) . \end{aligned}$$

Now for any X it follows that $\sum_X p(X|\mathrm{pa}(X)) = 1$. Thus, by performing the eliminations $\sum_X$ for each $X \in \mathcal{B}$ in reverse topological order, we have (by an inductive argument) that $\sum_{\mathcal{B}} \prod_{X \in \mathcal{B}} p_G(X|\mathrm{pa}(X)) = 1$, which proves the result. □

A simple corollary of the previous result is that if $\mathcal{B}$ is a set of barren nodes w.r.t. $\mathcal{R} \cup \mathcal{S}$, then $p_G(\mathcal{R}|\mathcal{S}) = p_H(\mathcal{R}|\mathcal{S})$, where H is the graph obtained by removing $\mathcal{B}$. This holds since Proposition 2 can be applied to both the numerator and the denominator of the definition of conditional distribution. Another simple corollary is the following.

Corollary 1. *If $(\mathcal{R}, \mathcal{S})$ is a partition of the nodes of a Bayesian network structure G such that $\mathcal{S}$ is barren (w.r.t. $\mathcal{R}$ in G) then $p(\mathcal{R}) = \prod_{X \in \mathcal{R}} p(X|pa(X))$.*

Proof. The graph obtained by removing $\mathcal{S}$ is a Bayesian network over $\mathcal{R}$ and hence encodes a distribution $p(\mathcal{R})$ which factorizes as the product of local conditional distributions. □

We can now prove the following result.

Theorem 2. *If the joint distribution of a probabilistic model factorizes according to an acyclic directed graph G, then the probabilistic model satisfies the Markov properties represented by G.*

Proof. The set $\mathrm{de}(X)$ is barren w.r.t. $\{X\} \cup \mathrm{nd}(X)$, thus by Corollary 1,

$$p(X, \mathrm{nd}(X)) = p(X|\mathrm{pa}(X)) \prod_{Y \in \mathrm{nd}(X)} p(Y|\mathrm{pa}(Y)) .$$

Similarly, $\{X\} \cup \mathrm{de}(X)$ is barren w.r.t. $\mathrm{nd}(X)$, hence

$$p(\mathrm{nd}(X)) = \prod_{Y \in \mathrm{nd}(X)} p(Y|\mathrm{pa}(Y)) .$$

Combining these results, we have that

$$\begin{aligned} p(X, \mathrm{nd}(X)|\mathrm{pa}(X)) &= p(X|\mathrm{pa}(X)) \frac{\prod_{Y \in \mathrm{nd}(X)} p(Y|\mathrm{pa}(Y))}{p(\mathrm{pa}(X))} \\ &= p(X|\mathrm{pa}(X)) p(\mathrm{nd}(X)|\mathrm{pa}(X)) . \end{aligned}$$

which is true iff the Markov property is satisfied. □

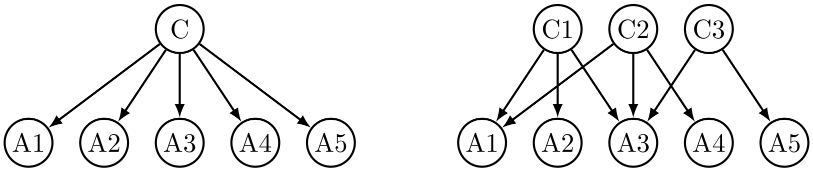

Figure 2. A naive Bayes model (left) and a bipartite Bayesian network (right).

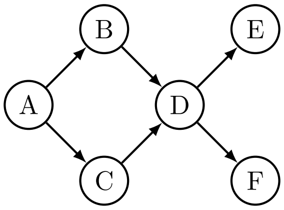

Figure 3. A polytree-shaped Bayesian network.

5. Some common structures

A tree is a graph where each node has at most one parent (note that this definition allows disconnected graphs). A Bayesian network is tree-shaped if its structure is a tree. As we will see later, trees are particularly important as inference (querying) can be performed in linear time on them. A naive Bayes model is a specific type of tree-shaped Bayesian network with a single root variable which has the remaining variables as children, as in the graph in the left in Figure 2. Naive Bayes is a common model for building classifiers in many domains, for instance spam detection in emails. The root variable represents the message type (spam or ham) and the leaves indicate textual features such as word frequencies and amount of capitalization. Recall that a conditional distribution over a set of variables determines a probabilistic model.

Bipartite Bayesian networks are Bayesian networks where the nodes can be partitioned in two sets such that all arcs originate in one set and point to a variable in the other set. The network in the right in Figure 2 is bipartite. Note that a naive Bayes is the simplest case of a bipartite Bayesian network (in terms of structure).

path

singly directed graphs

polytree-shaped Bayesian network

A *path* is a sequence of nodes $X_1, \dots, X_n$ such that X_i, X_{i+1} is an edge of the graph for each $i = 1, \dots, n-1$. An acyclic directed graph is *singly directed* if there is at most one directed path from any two nodes. A *polytree-shaped Bayesian network* has a singly-directed graph. Note that bipartite networks are also polytree-shaped. Figure 3 depicts a polytree-shaped Bayesian network which is not bipartite. The Asia Bayesian network in Figure 1 is not polytree-shaped as e.g. the node D is reachable from node S through two different paths.

6. Exercises

Exercise 1. *Compute the number of probability values necessary to specify a tree-shaped Bayesian network with n binary variables.*

Exercise 2. *Compute the number of probability values necessary to specify a bipartite Bayesian network with m root variables and n leaves. The root variables are binary and the leaf variables are ternary.*

Exercise 3. *Prove that the leaves of a naive Bayes model are pairwise independence conditional on the root node.*

7. Reading

- Chapter 4.2–4.3 and 5.2–5.3 of Darwiche's book
- Chapter 2 of Korb and Nicholson's book (skip section 2.4.5)

8. Assignment

- Write code that allows specification of Bayesian networks over categorical variables and computation of conjunctive queries (by enumeration). Your code should represent distributions explicitly and implement product, division and variable elimination of functions. [Optional: You can test and avoid including distributions associated to barren nodes to speed up computations.]
- Test your code on the Asia network; in particular verify that the network satisfies all the Markov properties and the independences induced by Proposition 1.
- [Optional: Write code that reads a Bayesian network in some of the standard file formats (BIF, NET, DSC).]

Coding tips: If $\text{dom}(X_i) = \{0, 1, \dots, k_i - 1\}$ for every variable X_i, then a function $f(\{X_1, \dots, X_m\})$ can be represented by a list (or set) of variables $(X_1, \dots, X_m)$, a vector of reals $p_1, \dots, p_N$ with $N = \prod_{i=1}^{m} k_i$, and the mapping $f(\{X_1, \dots, X_m\})\nu \mapsto p_i$ such that $i = \nu(X_1) + \sum_{i=2}^{n} \nu(X_i) \prod_{j=1}^{j-1} k_i$. The ordering of the variables can be local (different from function to function) or global (equal for all functions). Functional operations can be performed element-wise by iterating over (relevant) valuations.

If you want a off-the-shelf Bayesian network software to compare your results, I suggest downloading SamIam at `http://reasoning.cs.ucla.edu/samiam/`.