

Bayesian Networks and Markov Random Fields

1 Bayesian Networks

We will use capital letters for random variables and lower case letters for values of those variables. A Bayesian network is a triple $\langle V, G, \mathcal{P} \rangle$ where V is a set of random variables X_1, \dots, X_n , G is a directed acyclic graph (DAG) whose nodes are the variables in V , and \mathcal{P} is a set of conditional probability tables as described below. The conditional probability tables determine a probability distribution over the values of the variables. If there is a directed edge in G from X_j to X_i then we will say that X_j is a parent of X_i and X_i is a child of X_j . To determine values for the variables one first selects values for variables that have no parents and then repeatedly picks a value for any node all of whose parents already have values. When we pick a value for a variable we look only at the values of the parents of that variable. We will write $P(x_i \mid \text{parents of } x_i)$ to abbreviate $P(x \mid x_{i_1}, \dots, x_{i_k})$ where x_{i_1}, \dots, x_{i_k} are the parents of x . For example, if x_7 has parents x_2 and x_4 then $P(x_7 \mid \text{parents of } x_7)$ abbreviates $P(x_7 \mid x_2, x_4)$. Formally, the probability distribution on the variables of a Bayesian network is determined by the following equation.

$$P(x_1, \dots, x_n) = \prod_{i=1}^n P(x_i \mid \text{parents of } x_i) \quad (1)$$

The conditional probabilities of the form $P(x_i \mid \text{parents of } x_i)$ are called conditional probability tables (CPTs). Suppose that each of the variables x_i has d possible values (this is not required in general). In this case if a variable x has k parents then $P(x \mid \text{parents of } x)$ has d^{k+1} values (with $(d-1)d^k$ degrees of freedom). These d^{k+1} values can be stored in a table with $k+1$ indices. Hence the term “table”. Note that the number of indices of the CPTs (conditional probability tables) is different for the different variables.

Note that an HMM is a Bayesian network with a variable for each hidden state and each observable token.

Bayesian networks are often used in medical diagnosis where variables represent the presence or absence of certain diseases or the presence or absence of certain measurable quantities such as blood sugar or presence of a certain protein in the blood.

In a Bayesian network the edges of the directed graph are often interpreted as “causation” with the parent causally influencing the child and parents getting assigned values temporally before children.

We are interested in “Bayesian inference” which means, intuitively, inferring

causes by observing their effects using Bayes' rule. In an HMM for example, we want to infer the hidden states from the observations that they cause.

In general we can formulate the inference problem as the problem of determining the probability of an unknown variable (a hidden cause) from observed values of other variables. In general we can consider the variables in any order.

$$P(x_5, x_7 \mid x_3, x_2) = \frac{P(x_5, x_7, x_2, x_3)}{P(x_2, x_3)}$$

So in general, for inference it suffices to be able to compute probabilities of the form $P(x_{i_1}, \dots, x_{i_k})$. We give an algorithm for doing this in section 3.

2 Inference in Bayesian Network is #P hard

A Boolean variable (also called Bernoulli variable) is a variable that has only the two possible values of 0 or 1. A disjunctive clause is a disjunction of literals where each literal is either a Boolean variable or the negation of a Boolean variable. For example we have that $(X_5 \vee \neg X_2 \vee X_3)$ is a clause with three literals. A 3SAT problem is a set of clauses with three literals in each clause. It is hard (in fact #P hard) to take a 3SAT problem and determine the number of complete assignments of values to variables that satisfy the clauses, i.e., that make every clause true.

Take X_1, \dots, X_n be independent Boolean (Bernoulli) variables with $P(X_i = 1) = 1/2$. Let Σ be a set of clauses over these variables where each clause has only three literals. Let C_j be a random variable which is 1 if the j th clause is satisfied. Since we can compute c_j from x_1, \dots, x_n using a (deterministic) conditional probability table having only three parents. Let A_j be a Boolean variable that is true if all of C_1, \dots, C_j are true. a_1 can be computed with a CPT from c_1 and for $j > 1$ we have that a_j can be computed using a CPT from a_{j-1} and c_j .

Now we have that $P(A_k = 1)$ is proportional to the number of truth assignments that satisfying all the clauses. This implies that computing the probability of a partial assignment in a Bayesian network is #P hard. It is widely believed that there are no polynomial time algorithms for #P hard problems.

3 Recursive Conditioning

Although inference in Bayesian networks is hard in general, there exist algorithms that work well when the Bayesian network has special structure.

- $\mathcal{D}(X)$ is the set of values that variable X can have.
- ρ ranges over partial assignments of values to variables — ρ assigns values to *some* of the variables in V .
- $\text{dom}(\rho)$ is the set of variables that are assigned values by ρ . For $X \in \text{dom}(\rho)$ we have that $\rho(X) \in \mathcal{D}(X)$.
- For $X \notin \text{dom}(\rho)$, and $x \in \mathcal{D}(X)$ we let $\rho[X := x]$ be the extension of ρ that assigns X the value x .
- σ ranges over total assignments of values to the variables in V . For $X \in V$ we have $\sigma(X) \in \mathcal{D}(X)$.
- $\sigma \sqsubseteq \rho$ means that the complete assignment σ is compatible with the partial assignment ρ . In other words, for $X \in \text{dom}(\rho)$ we have $\sigma(X) = \rho(X)$.

We can now write equation (1) as follows where T_i is the conditional probability table for the i th variable and $T_i(\sigma)$ is $P(x_i \mid \text{parents of } x_i)$ where the variable values are determined by σ .

$$P(\sigma) = \prod_i T_i(\sigma) \tag{2}$$

For a partial assignment ρ we have the following.

$$P(\rho) = \sum_{\rho \sqsubseteq \sigma} \prod_i T_i(\sigma) \tag{3}$$

Note that $T_i(\sigma)$ depends on only some of the variables assigned by σ . For example suppose we have seven variables X_1, \dots, X_7 arranged in an “inverted tree” where variable X_1 has parents X_2 and X_3 ; variable X_2 has parents X_4 and X_5 ; and variable X_3 has parents X_6 and X_7 . Now suppose that $\text{dom}(\rho) = \{X_1, X_2\}$, i.e., ρ assigns a value only to X_1 and X_2 . Then we can write equation (3) as follows.

$$\begin{aligned}
P(x_1, x_2) &= \sum_{x_3, x_4, x_5, x_6, x_7} \frac{[T_2(x_2, x_4, x_5)T_4(x_4)T_5(x_5)]}{[T_1(x_1, x_2, x_3), T_3(x_3, x_6, x_7)T_6(x_6)T_7(x_7)]} \\
&= \frac{\left[\sum_{x_4, x_5} T_2(x_2, x_4, x_5)T_4(x_4)T_5(x_5) \right]}{\left[\sum_{x_3, x_6, x_7} T_1(x_1, x_2, x_3)T_3(x_3, x_6, x_7)T_6(x_6)T_7(x_7) \right]} \\
&= \tilde{P}(\langle x_2 \rangle, \langle T_2, T_3, T_4 \rangle) \tilde{P}(\langle x_1, x_2 \rangle, \langle T_1, T_3, T_6, T_7 \rangle) \\
&\text{where} \\
\tilde{P}(\rho, \mathcal{T}) &= \sum_{\sigma \sqsubseteq \rho} \prod_{T \in \mathcal{T}} T(\sigma) \\
&\text{where } \sigma \text{ only assigns to variables in } \mathcal{T}
\end{aligned}$$

So we can summarize this as follows.

$$\begin{aligned}
P(x_1, x_2) &= \tilde{P}(\langle x_1, x_2 \rangle, \langle T_1, T_2, T_3, T_4, T_5, T_6, T_7 \rangle) \\
&= \tilde{P}(\langle x_2 \rangle, \langle T_2, T_3, T_4 \rangle) \tilde{P}(\langle x_1, x_2 \rangle, \langle T_1, T_3, T_6, T_7 \rangle)
\end{aligned}$$

To compute probabilities of the form $P(\rho)$ we can compute the quantities $\tilde{P}(\rho, \mathcal{T})$ where these quantities often factor. Recursive conditioning is defined by the following equation where $Y \notin \text{dom}(\rho)$.

$$\tilde{P}(\rho, \mathcal{T}) = \sum_{y \in \mathcal{D}(Y)} \tilde{P}(\rho_1[Y := y], \mathcal{T}_1) \cdots \tilde{P}(\rho_k[Y := y], \mathcal{T}_k)$$

For $i \neq j$ we must have that no variable $Z \notin \text{dom}(\rho) \cup \{Y\}$ appears in both \mathcal{T}_i and \mathcal{T}_j . Also, we require that $\rho_i[Y = y]$ is the restriction of $\rho[Y = y]$ to the variables occurring \mathcal{T}_i . See the above example. To make this algorithm efficient the computations of values for expressions of the form $\tilde{P}(\rho, \mathcal{T})$ must be memoized, i.e., stored in a table so that values can be reused if they are needed again. The choice of the variable $Y \notin \text{dom}(\rho)$ is important for the efficiency of the algorithm.

4 Markov Random Fields

Throughout this section (and throughout these notes) any variable X is assumed to have an associated set $\mathcal{D}(\cdot)X$ of possible values. A *configuration* of set e of variables is defined to be an assignment of a value in $\mathcal{D}(\cdot)X$ to each variable X in e . For example, consider two variables X and Y that range over integers, i.e., $\mathcal{D}(\cdot)X = \mathcal{D}(\cdot)Y = \mathcal{I}$ the set of all integers. A configuration of the set $\{X, Y\}$ is an assignment of an integer to X and an integer to Y . So the configurations of the set $\{X, Y\}$ correspond to pairs $\langle i, j \rangle$ where i is the value for X and j is the value for Y .

A hypergraph is a set of nodes and a set of hyperedges where each hyperedge is a subset of the nodes. A normal (undirected) graph is a special case of a hypergraph in which each hyperedge contains exactly two nodes. In a general hypergraph, a hyperedge may involve only one node, or may involve three or more nodes.

A Markov random field is hypergraph whose nodes are variables and where each hyperedge e is associated with an energy function mapping each possible configuration of the variables in e to a real number often referred to as energy. If ρ is a configuration of the variables in hyperedge e then we write $E(e, \rho)$ for the energy assigned to hyperedge e under this configuration of its variables. A configuration of a Markov random field is a configuration of its variables, i.e., an assignment of a value to each of the variables in the field. If e is a hyperedge of the field and *sigma* is a configuration of the (whole) field, then $E(e, \sigma)$ is defined to equal $E(e, \sigma|e)$ where $\sigma|e$ is the restriction of the (global) configuration σ to the (local) variables in e . A Markov random field, together with a temperature parameter β , determines a probability distribution over its configurations defined by the following equation.

$$P(\sigma) = \frac{1}{Z} e^{-\beta E(\sigma)} \quad (4)$$

$$Z = \sum_{\sigma} e^{-\beta E(\sigma)} \quad (5)$$

$$E(\sigma) = \sum_e E(e, \sigma) \quad (6)$$

A Bayesian network is a special case of a Markov random field in which there is one hyperedge for each conditional probability table. To convert a Bayesian network to a Markov random field with the same conditional probability table

we define a hyperedge e for each CPT T and define the energy function for e as follows.

$$E(e, \rho) = \frac{\ln \frac{1}{T(\rho)}}{\beta} \quad (7)$$

For Markov random fields satisfying 7 we have $Z = 1$ (problem: show this).

As with Bayesian networks, we let ρ range over configurations of subsets of the variables in G and let σ range over configurations of all nodes in G , i.e., configurations that assign a value to every variable. We first make the following observation about conditional probabilities in Markov random fields.

$$\begin{aligned} P(x_2, x_5 \mid x_6, x_7) &= \frac{P(x_2, x_5, x_6, x_7)}{P(x_6, x_7)} \\ &= \frac{\frac{1}{Z} \sum_{\langle x_2, x_5, x_6, x_7 \rangle \subseteq \sigma} e^{-\beta E(\sigma)}}{\frac{1}{Z} \sum_{\langle x_6, x_7 \rangle \subseteq \sigma} e^{-\beta E(\sigma)}} \\ &= \frac{Z(\langle x_2, x_5, x_6, x_7 \rangle)}{Z(\langle x_6, x_7 \rangle)} \\ Z(\rho) &= \sum_{\rho \subseteq \sigma} e^{-\beta E(\sigma)} \end{aligned} \quad (8)$$

The probability of evidence problem for Markov random fields is the problem of computing $Z(\rho)$ rather than $P(\rho)$ (we have that $P(\rho) = Z(\rho)/Z = Z(\rho)/Z(\emptyset)$). We can again use recursive conditioning. For recursive conditioning the recursive subproblems are on smaller fields. Therefore, it is important to explicitly give the field under consideration. For a Markov random field \mathcal{F} and partial assignment ρ to the variables of \mathcal{F} We define $Z(\rho, \mathcal{F})$ as follows.

$$Z(\rho, \mathcal{F}) = \sum_{\rho \subseteq \sigma} e^{-\beta E(\sigma|\mathcal{F})}$$

where σ only assigns to variables in \mathcal{F}

Recursive conditioning for Markov random fields is defined by the following equation where $Y \notin \text{dom}(\rho)$.

$$Z(\rho, \mathcal{F}) = \sum_{y \in \mathcal{D}(Y)} Z(\rho_1[Y := y], \mathcal{F}_1) \cdots Z(\rho_k[Y := y], \mathcal{F}_k)$$

As with Bayesian networks, for $i \neq j$ we must have that no variable $Z \notin \text{dom}(\rho) \cup \{Y\}$ appears in both \mathcal{F}_i and \mathcal{F}_j . Also, we require that $\rho_i[Y = y]$ is the

restriction of $\rho[Y = y]$ to the variables occurring \mathcal{G}_i . This algorithm is identical to that used for Bayesian networks and the same comments about memoization and the choice of $Y \in \text{dom}(\rho)$ apply.