Probabilistic Graphical Models

Atif Khan

October 22, 2012

Bayesian Networks

Tree Structure CPDs

Table CPDs are problematic when there are many many parents for a conditional variable. Tree structure CPDs can handle a large set of parents given that there is certain context provided.

We start from the root and traverse the branches. Each branch represents a CPD for a given set of conditions. For example if a student does not apply for a job ($A = a^0$), then the probability of the student getting a job ($J = j^1$) is

- $P(j^1|A, S, L) = 0.2$
- the probability is independent of S, L

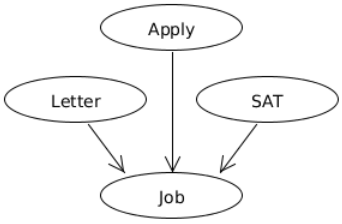


Figure 1: A simple CPD with four parents

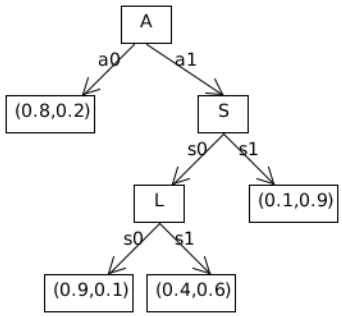


Figure 2: Tree based CPD for Figure 1

Independence of Causal Influence

Tree CPS (§- Tree Structure CPDs) are good for adding context with many parents. But when the number of the parents is quite large and most of the parents contribute (see Figure 3), then the Tree CPD is not a good representation. For example, *Cough* can be caused by an array of diseases. We utilize a noisy OR CPD for this purpose.

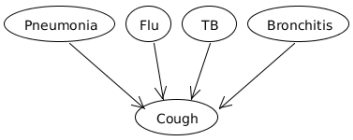


Figure 3: Multiple parents contributing towards a single variable. This does not lead it self into a Tree CPD.

Noisy OR CPD

For each parent variable X, we introduce an intermediate variable Z (filter). Z represents the event of a parent X being *true*, causing Y to be true by itself. Ultimately Y is true if any Z succeeded in making it true. Therefore, Y is a deterministic *OR* based of its parents Z.

$$P(Z_0 = 1) = \lambda_0 \quad \textit{Leak}$$

$$P(Z_i = 1|X_i) = \begin{cases} 0 & X_i = 0 \\ \lambda_i & X_i = 1 \quad \textit{Penetrance} \end{cases}$$

Now consider, what is the probability that $Y = 0$ given all the parents X:

$$P(Y = 0|X_1, X_2 \ldots, X_k) = (1 - \lambda_0) \prod_{i:X_i=1} (1 - \lambda_i)$$

where, $\prod_{i:X_i=1} (1 - \lambda_i)$ represents the parents that are on.

For the probability that $Y = 1$, we have

$$P(Y = 1|X_1, X_2 \ldots, X_k) = 1 - P(Y = 0|X_1, X_2 \ldots, X_k)$$

Figure 4: Noisy OR CPD
Penetrance defines how good is X_i in turning Z_i, where as *Leak* defines Y turning on by itself.

Generalization of the noisy OR CPD: Figure 5 represents the generalization of the noisy OR CPD. The variable Z is a deterministic variable that can represent different functions such as and *AND* operation, *MAX* operation etc.

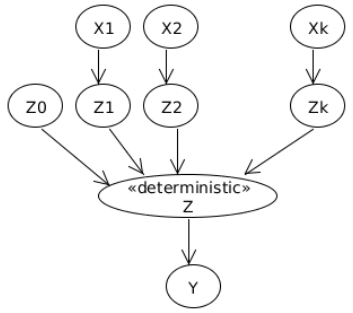


Figure 5: Generalization of the noisy OR CPD

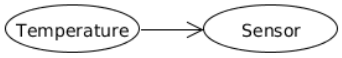
Sigmoid CPD

Given $Z = w_0 + \sum_{i=1}^{k} w_i X_i$, where $Z_i = w_i X_i$., a sigmoid CPD is

$$P(y^1|X_1, X_2, \ldots, X_k) = sigmoid(Z)$$

Where $sigmoid(z) = \frac{e^z}{1+e^z}$, z is a continuous variable. The result of the *sigmoid* function is to reduce the value of z to $[0, 1]$.

Continuous Variables



Imagine that the temperature is a *continuous variable* and the sensor provides an approximation of the temperature. That is *sensor S* is a normal distribution defined using *linear Gaussian* as:

$$S \sim \mathcal{N}(T; \sigma_s^2)$$

No imagine that the temperature soon *Temperature*′ depends on current temperature, outside temperature and the conditionally on the door being opened or closed (as shown in Figure 7). We have the following *conditional linear Gaussian* distributions:

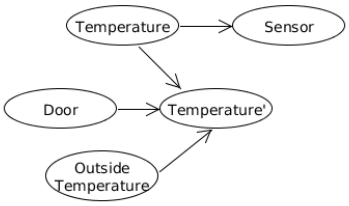


Figure 7: Example of continuous variables with condition

$$T' \sim \mathcal{N}(\alpha_0 T + (1 - \alpha_0)O; \sigma_{0T}^2) \quad \text{when } D^0$$
$$T' \sim \mathcal{N}(\alpha_1 T + (1 - \alpha_1)O; \sigma_{1T}^2) \quad \text{when } D^1$$

Linear Gaussian

For the given graph in Figure 8, we have a variable Y with parents X, then we have a *linear Gaussian* defined as follows:

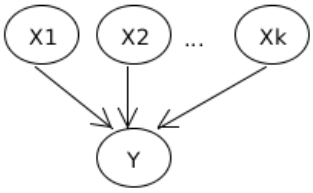


Figure 8: Model for linear Gaussian.

$$Y \sim \mathcal{N}(w_0 + \sum w_i X_i; \sigma^2)$$

where, the mean of the Gaussian distribution ($w_0 + \sum w_i X_i$) is a linear function (of the parents X_i, and the variance σ^2 does not depend on the parents.

Conditional Linear Gaussian

We can now define a *conditional linear Gaussian* (see Figure 9) with a discrete parent variable A as follows:

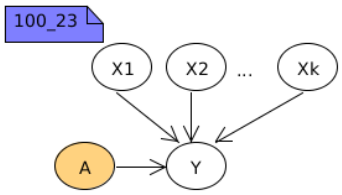


Figure 9: Model for conditional linear Gaussian. Variable A is a discrete parent. There can be more than one discrete parents.

$$Y \sim \mathcal{N}(w_{a0} + \sum w_i X_i; \sigma_a^2)$$

Note that the variance σ_a^2 depends on the discrete parent A but not X.

Undirected Graphical Models

Factors

The characteristics of factors are defined as follows:

- Factors can not be directed in an undirected graphical model.
- Factors subsumes both *joint distribution* and *conditional probability distribution*.
- A joint distribution over D is a factor over D. It specifies a real number for every assignment of values of D.
- A conditional distribution $P(X|U)$ is a factor over $\{X\} \bigcup U$.
- There are no constraints on a parameters in a factor.
- A factor represents *compatibility/affinity* between the joining nodes.
- A factor is only one contribution to the overall joint distribution.

for example, in a joint distribution, the numbers must add up to 1.

Example: Consider a fully connected graph over $\mathcal{X}$. Let all the variables be binary. The we have:

In a fully connected graph, there are no conditional independence.

- Each factor over an edge will have four parameters.
- total number of parameters in the graph would be $4\binom{n}{2}$.

Note The number of parameters required to specify a joint distribution over n binary variables is $2^n - 1$. The number of parameters available in an undirected graph is $4\binom{n}{2}$.

$$4\binom{n}{2} <<< 2^n - 1$$

Therefore, pairwise factors do not have enough parameters to completely cover the joint distribution space. A more general representation can be obtained by allowing factors over an arbitrary subsets of variables.

Factor Product

Let X, Y, and Z be three disjoint variable sets and let $\phi_1(X, Y)$ and $\phi_2(Y, Z)$ be two factors. Then we define *factor product* $\psi = \phi_1 \times \phi_2$, where $\psi : Val(X, Y, Z) \mapsto \mathbb{R}$. $\psi(X, Y, Z) = \phi_1(X, Y) \cdot \phi_2(Y, Z)$. Note that the two factors are multiplied in a way that matches up the common part Y.

Example Let $\phi_1(A, B)$ and $\phi_2(B, C)$ be defined as follows:

$$\begin{bmatrix} a^1 & b^1 & 0.5 \\ a^1 & b^2 & 0.8 \\ a^2 & b^1 & 0.1 \\ a^2 & b^2 & 0 \\ a^3 & b^1 & 0.3 \\ a^3 & b^2 & 0.9 \end{bmatrix} \cdot \begin{bmatrix} b^1 & c^1 & 0.5 \\ b^1 & c^2 & 0.7 \\ b^2 & c^1 & 0.1 \\ b^2 & c^2 & 0.2 \end{bmatrix} = \begin{bmatrix} a^1 & b^1 & c^1 & 0.5 \cdot 0.5 = 0.25 \\ a^1 & b^1 & c^2 & 0.5 \cdot 0.7 = 0.35 \\ a^1 & b^2 & c^1 & 0.8 \cdot 0.1 = 0.08 \\ a^1 & b^2 & c^2 & 0.8 \cdot 0.2 = 0.16 \\ a^2 & b^1 & c^1 & 0.05 \\ a^2 & b^1 & c^2 & 0.07 \\ a^2 & b^2 & c^1 & 0 \\ a^2 & b^2 & c^2 & 0 \\ a^3 & b^1 & c^1 & 0.15 \\ a^3 & b^1 & c^2 & 0.21 \\ a^3 & b^2 & c^1 & 0.09 \\ a^3 & b^2 & c^2 & 0.18 \end{bmatrix}$$

Gibbs Distribution

General notion of *factors product* to define an undirected parametrization of a distribution.

Definition A distribution P_ϕ is a Gibbs distribution, parametrized by a set of factors $\phi = \{\phi_1(D_1), \ldots, \phi_k(D_k)\}$, if it is defined as follows:

$$P_\phi(X_1, \ldots, X_n) = \frac{1}{Z}\tilde{P}(X_1, \ldots, X_n)$$

where

$$\tilde{P}(X_1, \ldots, X_n) = \phi_1(D_1) \times \phi_2(D_2) \times \ldots \times \phi_k(D_k)$$

and

$$Z = \sum_{X_1, \ldots, X_k} \tilde{P}(X_1, \ldots, X_n) = \text{partition function}$$

To map Gibbs distribution to a graph we inspect the scope of the factors contained in the parametrization. For example if the scope contain both X and Y, then we will introduce an edge between X and Y nodes.

Markov network factorization We say that a distribution P_ϕ with $\phi = \{\phi_1(D_1), \ldots, \phi_k(D_K)\}$ *factorizes* over a *markov network* $\mathcal{H}$ if each $D_k (k = 1, \ldots, K)$ is a complete sub-graph of $\mathcal{H}$.

Clique potentials The factors that parametrize a Markov network are often called *clique potentials*. Note that

- Every complete sub-graph is a subset of some (maximal) clique. Therefore, we can reduce the number of factors in our parametrization by allowing factors only for maximal cliques. Let $C_1, \ldots, C_k$ be

the cliques in $\mathcal{H}$, then we can parametrize P using a set of factors $\phi_1(C_1), \ldots, \phi_k(C_l)$.

- Any factorization (in terms of complete sub-graph) can be converted into this form simply by assigning each factor to a clique that encompasses its scope.
- *Clique potential* is then calculated by multiplying all the factors assigned to each clique.

Pairwise Markov Networks

A pairwise Markov network over a graph $\mathcal{H}$ is associated with a set of node potentials $\{\phi(X_i) : i = 1, \ldots, n\}$ and a set of edge potentials $\{\phi(X_i, X_j) : (X_i, X_j) \in \mathcal{H}\}$. The overall distribution is the normalized product of all of the potentials (both node and edge).

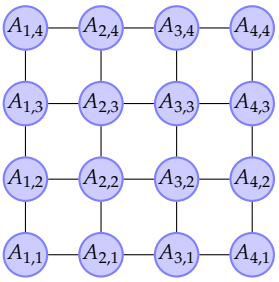


Figure 10: A pairwise Markov network (MRF) structured as a grid.

A distribution where all of the factors are over single variables or pair of variables.

Reduced Markov Networks

Conditioning a distribution on some assignment u to some subset of variables$U$.

Let $\phi(Y)$ be a factor, and $U = u$ an assignment for $U \subseteq Y$. We define the reduction of the factor ϕ to the context $U = u$, denoted by $\phi[U = u] = \phi[u]$, to be a factor over scope $Y' = Y - U$ such that:

$$\phi[u](y', u) = \phi(y', u)$$

Conditioning a distribution corresponds to eliminating a ll entries in the joint distribution that are inconsistent with the event $U = u$, and renormalizing the remaining entries to sum to one.

Exact Inference: Variable Elimination

Conditional Probability Query

The conditional probability query is of the type $P(Y|E = e)$. Such queries allows for many reasoning patterns, including explanation, prediction, intercausal reasoning etc.). From the definition of conditional probability we know that

$$P(Y|E = e) = \frac{P(Y, e)}{P(e)} \tag{1}$$

Where each instantiation of the numerator is a probability express $P(y|e)$, which can be computed by summing out all entries in the joint distribution that correspond to assignments consistent with y, e. Let $W = \mathcal{X} - Y - E$; a set of variables that are neither evidence not query. Then we have

$$P(y, e) = \sum_{w} P(y, e, w) \tag{2}$$

The probability can be computed as

$$P(e) = \sum_{y} P(y, e) \tag{3}$$

Since Y, E, W are all of the network's variables, each term $P(y, e, w)$ in the summation is simply an entry in the joint distribution.

Once we have $P(e)$ and $P(y, e)$ we can then calculate $P(y|e)$ using equation 1[1].

[1] Note that this process corresponds to traking the vector of marginal probabilities $P(y^1, e), \ldots, P(y^k, e)$ where $k = |Val(Y)|$ and renormalizing the entries to sum to 1.

References

D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.