

Recitation 1: Bayesian Networks

Teaching Assistant: Eitan Richardson

Recommended reading:

- PGM Book, chapters 3, 5 [1]

1.1 Reminder – Bayesian Networks

We have a set of (discrete) random variables and we want to represent the joint distribution. An explicit joint probability distribution (table) grows *exponentially* with the number of variables. We want a more compact representation that exploits (and expresses) *conditional independence* between variables.

☞ Any joint distribution can be factorized (in many ways) as a product of conditional distributions per the chain rule: $P(X_1, \dots, X_k) = P(X_1)P(X_2|X_1) \cdots P(X_k|X_1, \dots, X_{k-1})$, but without assuming conditional independencies, this will not reduce the complexity.

1.1.1 Naive Bayes Model

In this simplified model, we assume that all random variables depend on a single one (the class) and are conditionally independent of each other given the class:

$$\forall i \neq j, P \models (X_i \perp X_j \mid C).$$

Therefore:

$$P(C, X_1, \dots, X_k) = P(C) \prod_{i=1}^k P(X_i \mid C)$$

Number of parameters (assuming $|Val(C)| = n_c$ and $\forall i, |Val(X_i)| = n_x$) is just: $n_c(1 + k \cdot n_x)$

Example 1.1 A Naive Bayes model with two features

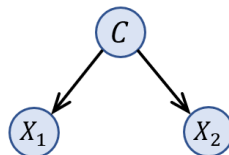


Figure 1.1: Naive Bayes example

¹Original LaTeX template courtesy of UC Berkeley.

Lets verify that $P \models (X_1 \perp X_2 | C)$:

Proof:

$$P(X_1, X_2 | C) = \frac{P(X_1, X_2, C)}{P(C)} = \frac{P(C)P(X_1 | C)P(X_2 | C)}{P(C)} = P(X_1 | C)P(X_2 | C)$$

■

Example for prior and *conditional probability* tables $P(C)$, $P(X_1 | C)$ and $P(X_2 | C)$:

C	$P(C)$	C	$X_1 = 0$	$X_1 = 1$	C	$X_2 = 0$	$X_2 = 1$
0	0.4	0	0.5	0.5	0	0.9	0.1
1	0.6	1	0.2	0.8	1	0.4	0.6

The resulting *joint probability* table $P(C, X_1, X_2) = P(C)P(X_1 | C)P(X_2 | C)$:

X_1	X_2	$C = 0$	$C = 1$
0	0	0.18	0.048
0	1	0.02	0.072
1	0	0.18	0.192
1	1	0.02	0.288

Lets check if in our examples, X_1 and X_2 are *marginally independent* (i.e. $P \models (X_1 \perp X_2)$?)

$$P(X_1, X_2) = \sum_c P(C = c, X_1, X_2)$$

$$P(X_1) = \sum_{x_2} P(X_1, X_2 = x_2), \quad P(X_2) = \sum_{x_1} P(X_1 = x_1, X_2)$$

X_1	X_2	$P(X_1, X_2)$	X_1	$P(X_1)$	X_2	$P(X_2)$	X_1	X_2	$P(X_1)P(X_2)$
0	0	0.228	0	0.32	0	0.6	0	0	0.192
0	1	0.092	1	0.68	1	0.4	0	1	0.128
1	0	0.372					1	0	0.408
1	1	0.308					1	1	0.272

The result shows that, as expected, $P(X_1, X_2) \neq P(X_1)P(X_2) \implies P \not\models (X_1 \perp X_2)$.

1.1.2 A General Bayesian Network

Formally, a *BN* is a pair $\mathcal{B} = (\mathcal{G}, P)$ of a *DAG* \mathcal{G} , whose nodes are *RVs* and whose edges indicate "direct influences", and of a set P of local *conditional probability distributions* – *CPDs* defining the direct influences – the conditional distributions of nodes given their parents.

A *BN* can be viewed as:

- A factorization structure of the joint distribution:

$$P_{\mathcal{B}}(X_1, \dots, X_n) = \prod_{i=1}^n P_i(X_i | Pa_{X_i}^{\mathcal{G}})$$

- A map of the *conditional independence* assumptions about the joint distribution:

$$\mathcal{I}_{LM}(\mathcal{G}) = \forall i : (X_i \perp \text{NonDescendants}_{X_i} | Pa_{X_i}^{\mathcal{G}})$$

In addition to the local independencies $\mathcal{I}_{LM}(\mathcal{G})$, the BN graph \mathcal{G} defines other independencies, which will be discussed later on.

Example 1.2 The Stopped Car

We have 5 binary random variables describing different events that may occur on the way to the university (Mechanical problems, Forgot to fill the tank, need a New car, car Stopped and Late for Class).

A general joint probability distribution will have $2^5 - 1 = 31$ free parameters.

If we have some prior knowledge or assumptions about *conditional independence* between variables, we can use a more compact and modular representation like the *Bayesian Network* in Fig. 1.2.

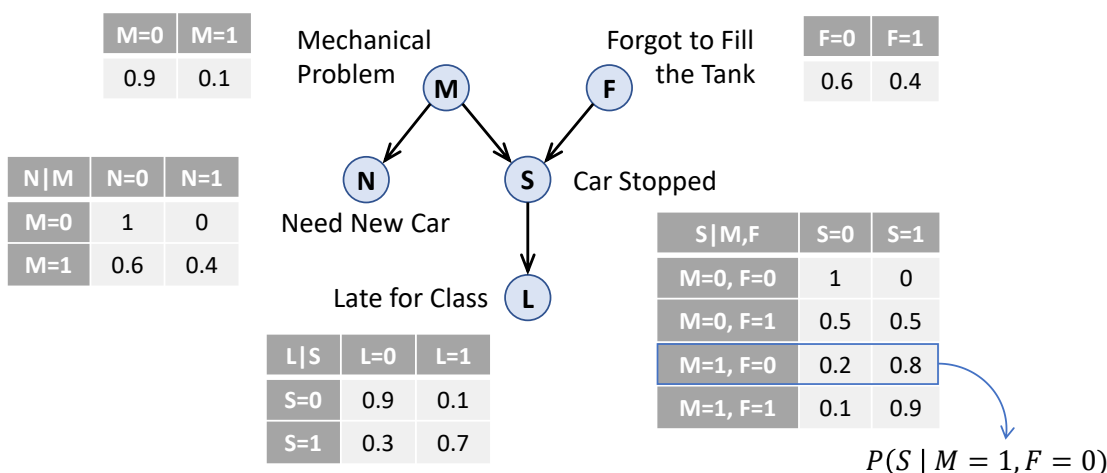


Figure 1.2: Bayesian Network example – The Stopped Car

The *local conditional independence* assumptions expressed in our *BN* are:

$$(M \perp F), \quad (N \perp S, F, L \mid M), \quad (S \perp N \mid M, F) \quad \text{and} \quad (L \perp N, M, F \mid S)$$

The number of free parameters in the our *BN* is only 10 (one for each CPD row, in the case of binary variables), as opposed to 31 in the general joint distribution.

The resulting joint probability distribution is:

$$P(M, F, N, S, L) = P(M)P(F)P(N \mid M)P(S \mid M, F)P(L \mid S) =$$

M	F	N	S	L = 0	L = 1
0	0	0	0	0.486	0.054
0	0	0	1	0.0	0.0
0	0	1	0	0.0	0.0
0	0	1	1	0.0	0.0
0	1	0	0	0.162	0.018
0	1	0	1	0.054	0.126
0	1	1	0	0.0	0.0
0	1	1	1	0.0	0.0
1	0	0	0	0.00648	0.00072
1	0	0	1	0.00864	0.02016
1	0	1	0	0.00432	0.00048
1	0	1	1	0.00576	0.01344
1	1	0	0	0.00216	0.00024
1	1	0	1	0.00648	0.01512
1	1	1	0	0.00144	0.00016
1	1	1	1	0.00432	0.01008

☞ Typically (when we have many random variables) we wouldn't want to compute or hold the full joint-distribution table.

1.1.3 Reasoning using the Bayesian Network

Below are some examples for different types of *reasoning* that can be performed using the *BN*:

Causal (downstream) reasoning:

$$\begin{aligned} P(L = 1 | F = 1) &= \frac{P(L = 1, F = 1)}{P(F = 1)} \\ &= \frac{\sum_m \sum_s P(L = 1, M = m, S = s, F = 1)}{P(F = 1)} \\ &= \sum_m \sum_s [P(M = m)P(S = s | M = m, F = 1)P(L = 1 | S = s)] \\ &= (0.9 \cdot 0.5 \cdot 0.1) + (0.9 \cdot 0.5 \cdot 0.7) + (0.1 \cdot 0.1 \cdot 0.1) + (0.1 \cdot 0.9 \cdot 0.7) = 0.424 \end{aligned}$$

Evidential (upstream) reasoning:

$$P(F = 1 | L = 1) = \frac{P(L = 1, F = 1)}{P(L = 1)} = \frac{P(L = 1, F = 1)}{P(L = 1, F = 0) + P(L = 1, F = 1)} \approx 0.6563$$

Another example:

$$P(M | S = 1) = \frac{P(M, S = 1)}{P(S = 1)} = \frac{P(M, S = 1)}{\sum_m P(M = m, S = 1)} \quad (\text{notice how the numerator is reused})$$

$$\text{where } P(M, S = 1) = \sum_f P(M, F = f, S = 1) = \sum_f P(M)P(F = f)P(S = 1 | M, F = f)$$

Doing the calculation gives $P(M = 1 | S = 1) \approx 0.318$

Inter-causal reasoning (explaining away):

$$P(M | S = 1, F = 1) = \frac{P(M, S = 1, F = 1)}{P(S = 1, F = 1)} = \frac{P(M, S = 1, F = 1)}{\sum_m P(M = m, S = 1, F = 1)}$$

Doing the calculation gives $P(M = 1 | S = 1, F = 1) \approx 0.166$

☞ Looking at the last two results, we see that knowing we forgot to fill the fuel tank reduced the probability that we have a mechanical problem given that our car stopped. This is called explaining away.

1.2 Local and Global Independence in Bayesian Networks

1.2.1 I-Maps

As discussed earlier, the *BN* is both a factorization of the joint distribution and a "map" defining conditional independencies. Here we provide more formal definitions.

Definition 1.3 $\mathcal{I}(P)$

We define $\mathcal{I}(P)$ to be the set of all independencies of the form $(\mathbf{X} \perp \mathbf{Y} | \mathbf{Z})$ that hold in P .

Definition 1.4 *I-Map*

A BN graph \mathcal{G} is an *I-Map* for a distribution P if the set of local independencies it defines hold in P i.e. $\mathcal{I}_{LM}(\mathcal{G}) \subseteq \mathcal{I}(P)$.

Theorem 1.5 \mathcal{G} is an *I-Map* for $P \iff P$ factorizes according to \mathcal{G} .

We saw the proof for theorem 1.5 in class.

1.2.2 D-separation

So far we discussed the *local independencies* in a BN: $\mathcal{I}_{LM}(\mathcal{G})$ – a node is independent of its non-descendants given its parents. Here we define additional independencies encoded in the graph.

Random variables (nodes) influence each other via trails in the graph. We can define the following cases in which X influences Y via Z :

- Causal trail $X \rightarrow Z \rightarrow Y$
- Evidential trail $X \leftarrow Z \leftarrow Y$
- Common cause trail $X \leftarrow Z \rightarrow Y$
- Common effect trail $X \rightarrow Z \leftarrow Y$ a.k.a v-structure

Returning to example 1.2, the trail $L \leftarrow S \leftarrow F$ enables L to influence F (if someone sees us come late to class, it increases the probability that we forgot to fill the fuel tank), that is: $P(F = 1 | L = 1) > P(F = 1)$. This means that $P_{\mathcal{B}} \not\models (F \perp L)$.

What happens if S is observed? Intuitively, once we know that our car stopped, the fact that we're late (or not) to class, does no longer affect the probability that we forgot to fill the tank, i.e. $P_{\mathcal{B}} \models (F \perp L | S)$. The fact that S is observed *blocked* the trail from L to F and it is no longer *active*.

Now lets look at the trail $M \rightarrow S \leftarrow F$. Does the fact we forgot to fill the tank changes the probability we'll have a mechanical problem? Intuitively it doesn't.

Proof:

$$P(M | F) = \frac{P(M, F)}{P(F)} = \frac{\sum_s P(M, F, S = s)}{P(F)} = \frac{\sum_s [P(M)P(F)P(S = s | M, F)]}{P(F)} = P(M)$$

■

What happens if S is observed?

$$P(M | F, S) = \frac{P(M, F, S)}{\sum_m P(m, F, S)} = \frac{P(M)P(S | M, F)}{\sum_m P(m)P(S | m, F)} \neq P(M | S)$$

We can see that in this trail, observing the middle node Z has an opposite effect – it made the trail *active* and the influence possible.

Definition 1.6 *Active Trail*

A trail $X_1 - \dots - X_n$ in a BN is active given a set of observed RVs \mathbf{Z} if, whenever there is a v-structure

along the trail: $X_{i-1} \rightarrow X_i \leftarrow X_{i+1}$, then X_i or one of its descendants is in \mathbf{Z} , and all other nodes along the trail are not in \mathbf{Z} .

Definition 1.7 *D-separation*

The sets \mathbf{X} and \mathbf{Y} are d-separated given \mathbf{Z} , denoted $d\text{-sep}_{\mathcal{G}}(\mathbf{X}; \mathbf{Y} | \mathbf{Z})$ if there are no active trails between any node $X \in \mathbf{X}$ and $Y \in \mathbf{Y}$ given \mathbf{Z} .

Definition 1.8 *Global Markov Independencies*

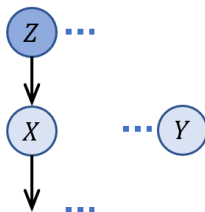
The set of all independencies that correspond to d-separation:

$$\mathcal{I}(\mathcal{G}) = \{(\mathbf{X} \perp \mathbf{Y} | \mathbf{Z}) : d\text{-sep}_{\mathcal{G}}(\mathbf{X}; \mathbf{Y} | \mathbf{Z})\}$$

What is the relationship between $\mathcal{I}_{LM}(\mathcal{G})$ and $\mathcal{I}(\mathcal{G})$?

Claim 1.9 $\mathcal{I}_{LM}(\mathcal{G}) \subseteq \mathcal{I}(\mathcal{G})$

Proof: We need to prove that a node is d-separated from its non-descendants given its parents i.e. $d\text{-sep}_{\mathcal{G}}(X; Y | Z)$ and we know that Y is not a descendant of X .



If there is no trail between X and Y than X and Y are d-separated (no active trail).

If a trail exists in \mathcal{G} , it can pass either via one of X 's parents or one of its children.

Considering the first option, the trail is: $X \leftarrow Z - \dots - Y$. Z is observed (being a parent) and is not a v-structure, hence the trail is blocked.

Considering the second option, the trail is: $X \rightarrow \dots - Y$. We can start at X and continue along the trail until we reach a left (opposite direction) edge. If we reached Y before encountering such an edge, Y is a descendant of X , which is not the case. This means that we have some v-structure along the trail $X \rightarrow \dots \rightarrow W \leftarrow \dots - Y$. For the trail to be active, W or one of its descendants must be observed i.e. must be a parent of X , but since W is a descendant of X , this would create a cycle in \mathcal{G} , hence the trail is blocked in this option as well.


■

The next three theorems define the relation between the conditional independencies defined by d-separation in \mathcal{G} to those that hold in the probability distributions that factorize according to \mathcal{G} .

Theorem 1.10 *If P factorizes according to \mathcal{G} than $\mathcal{I}(\mathcal{G}) \subseteq \mathcal{I}(P)$*

☞ *The proof for the above theorem requires some tools we'll learn when discussing undirected representation.*


Theorem 1.11 *If X and Y are not d -separated given Z (i.e. an active trail exists between X to Y given Z), then X and Y are dependent given Z in some distribution P that factorizes over \mathcal{G} .*

 *The above theorem is proved by constructing such distribution by a set of CPDs along the trail.*

Theorem 1.12 *For almost all distributions P that factorizes over \mathcal{G} : $\mathcal{I}(\mathcal{G}) = \mathcal{I}(P)$*

 *Proof sketch in the book [1]*

Returning to our stopped car example, lets check if the fact that we forgot to fill the tank affects the probability that we will need a new car, given that our car has stopped. Examining the trail $N \leftarrow M \rightarrow S \leftarrow F$, we can see that it is *active*, since the only observed variable (S) is part of a v-structure. This means that in almost all sets of CPSs over our graph, N and F will be dependent. In our specific distribution, knowing that we forgot to fill the tank reduces the probability that we'll need a new car (given that our car has stopped).

 *Notice that we sometime refer to BN as the pair $\mathcal{B} = (\mathcal{G}, P)$, which defines a specific distribution $P_{\mathcal{B}}$. In other cases, we only talk about the BN graph structure \mathcal{G} that defines $\mathcal{I}(\mathcal{G})$ and can be thought of as a filter that passes only distributions that factorize according to \mathcal{G} and in which $\mathcal{I}(\mathcal{G}) \subseteq \mathcal{I}(P)$.*

1.3 Local Conditional Probability Distributions

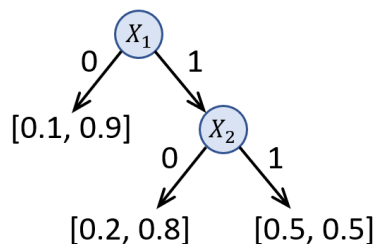
So far we assumed the local CPDs are explicit tables. If a node has many parents, the CPD table can become large (the number of parameters in a regular table CPD grows exponentially with the number of parent random variables). Here we discuss alternative representations of the local CPDs.

1.3.1 Tree CPD

One method to reduce the number of parameters is representing the CPD as a tree, where every leaf represents a distribution over the dependent (child) RV and the path from the root to the leaf represents the combinations of parent values that lead to this distribution.

This representation is useful when different combinations of parent values share the same conditional distribution. For example, the CPD below can be represented by the tree-CPD in Fig. 1.3.

X_1	X_2	$P(Y = 0 X_1, X_2)$	$P(Y = 1 X_1, X_2)$
0	0	0.1	0.9
0	1	0.1	0.9
1	0	0.2	0.8
1	1	0.5	0.5

Figure 1.3: *Tree-CPD* example

1.3.2 Noisy Or

In example 1.2 we defined a CPD for the conditional dependency $P(S | M, F)$. In many cases, we don't need to model a complex interaction between the different causes (combinations of parent values) – we might want the CPD to represent some sort of a probabilistic *OR* model that scales linearly with additional variables. The *Noisy Or* model provides this kind of *independence of causal influence*:

Definition 1.13 *Noisy Or*

A binary RV Y depends on k binary parents X_1, \dots, X_k in a noisy-or model if:

$$P(Y = 0 | x_1, \dots, x_k) = (1 - \lambda_0) \prod_{i=1}^k (1 - \lambda_i)^{x_i}$$

λ_0 is the *leak* parameter, allowing a positive probability for $Y = 1$ even if all X_i are 0.

λ_i are the *noise parameters* defining the amount by which $X_i = 1$ reduces the probability that $Y = 0$. When $\lambda_0 = 0$ and all noise parameters λ_i equal 1, the model behaves like a deterministic OR.

The parameters that match the CPD we defined in example 1.2 are $\lambda_0 = 0$, $\lambda_F = 0.5$, $\lambda_M = 0.8$.

1.3.3 Logistic CPD

Another model that expresses *independence of causal influence* is the logistic CPD defined as:

$$P(Y = 1 | x_1, \dots, x_k) = \sigma\left(\omega_0 + \sum_{i=1}^k \omega_i x_i\right)$$

In this model, the dependent variable Y is influenced by a linear combination of the parent variables, followed by a sigmoid function.

References

- [1] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.