Belief Propagation in Bayesian Networks

Céline Comte

Reading Group “Network Theory”
November 5, 2018
Introduction

(First-order) logic
Represent causal relations between variables by a directed acyclic graph

Probabilities
Weight these causal relations by probabilities that implicitly account for non-represented variables
Introduction

“Belief networks are directed acyclic graphs in which the nodes represent propositions (or variables), the arcs signify direct dependencies between the linked propositions, and the strengths of these dependencies are quantified by conditional probabilities” (Pearl, 1986)
Bayesian networks are also ...

- A memory-efficient way of storing a PMF
- Based on simple probability rules
  (more details in a few slides)
- Inspired by human causal reasoning (Pearl, 1986, 1988)
- Used for decision taking if a utility function is provided
- Applied in many fields: medicine diagnoses, turbo-codes, (programming) language detection, ...
- Related to other models: Markov random fields, Markov chains, hidden Markov models, ...
References: Pearl’s articles and book

  → Belief propagation in causal trees

- J. Pearl (1986). “Fusion, propagation, and structuring in belief networks”. In: Artificial Intelligence
  → Belief propagation in causal trees and polytrees

  → A complete reference
  (thanks Achille for providing me with this book)
References: Textbooks

  → A lot of examples in Chapters 2 and 3


  → Definition and belief propagation
    (thanks Nathan for pointing this reference)
Outline

Reminders on probability theory

Bayesian networks

Belief propagation in trees

Belief propagation in polytrees
Outline

Reminders on probability theory

Bayesian networks

Belief propagation in trees

Belief propagation in polytrees
Independence and conditional independence

**Remark:** We work exclusively with discrete random variables
Independence and conditional independence

**Remark:** We work exclusively with discrete random variables

- A and B are **marginally independent** (written $A \perp B$) if one of these three equivalent conditions is satisfied:
  - $P(A, B) = P(A)P(B)$
  - $P(A \mid B) = P(A)$
  - $P(B \mid A) = P(B)$
Independence and conditional independence

**Remark:** We work exclusively with discrete random variables

- A and B are **marginally independent** (written $A \perp B$) if one of these three equivalent conditions is satisfied:
  
  1. $P(A, B) = P(A)P(B)$
  2. $P(A | B) = P(A)$
  3. $P(B | A) = P(B)$

- A and B are **conditionally independent** given C (written $A \perp B | C$) if one of these three equivalent conditions is satisfied:
  
  1. $P(A, B | C) = P(A | C)P(B | C)$
  2. $P(A | B, C) = P(A | C)$
  3. $P(B | A, C) = P(B | C)$
Useful rules

- **The chain rule of probabilities**
  If $A_1, \ldots, A_n$ are random variables, we have

$$P(A_1, \ldots, A_n) = P(A_1) \times P(A_2 | A_1) \times P(A_3 | A_1, A_2) \times \cdots \times P(A_n | A_1, \ldots, A_{n-1})$$
Useful rules

- **The chain rule of probabilities**
  If $A_1, \ldots, A_n$ are random variables, we have

  $$P(A_1, \ldots, A_n) = P(A_1) \times P(A_2 | A_1) \times P(A_3 | A_1, A_2) \times \cdots \times P(A_n | A_1, \ldots, A_{n-1})$$

- **Law of total probability**
  If $A$ and $B$ are two random variables,

  $$P(B) = \sum_A P(B | A) P(A)$$
Useful rules

- **Bayes’ rule**
  If A and B are two random variables,

  \[
  P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A)}
  \]

  We can see P(A) as a normalizing constant: we can first compute \(P(B \mid A) \propto P(A \mid B)P(B)\) for each value of B and then normalize to obtain \(P(B \mid A)\) without computing P(A)
• **Belief** in a random variable (**conviction** in french)
Marginal distribution of this random variable
(given the value of some observed variables)
Glossary

- **Belief** in a random variable (*conviction* in french)
  Marginal distribution of this random variable (given the value of some observed variables)

- **Observe** a random variable
Glossary

- **Belief** in a random variable (*conviction* in french)
  Marginal distribution of this random variable
  (given the value of some observed variables)

- **Observe** a random variable

- **Evidence** (piece of evidence)
  The set of random variables that have been observed
Outline

- Reminders on probability theory
- Bayesian networks
- Belief propagation in trees
- Belief propagation in polytrees
The Student example

Course difficulty

Student intelligence

Grade

Reference letter

Baccalauréat

Borrowed from (Koller, Friedman, and Bach, 2009)
The Student example

Course difficulty

Student intelligence

Grade

Baccalauréat

Reference letter

Borrowed from (Koller, Friedman, and Bach, 2009)
The Student example

- **Local Markov property**
  Each node is conditionally independent of its non-descendants given its parents:

  \[
  D \perp \{I, B\}, \quad I \perp D, \quad G \perp B \mid \{D, I\}, \\
  B \perp \{D, G, L\} \mid I, \quad L \perp \{D, I, B\} \mid G
  \]
The Student example

- **Local Markov property**
  Each node is conditionally independent of its non-descendants given its parents:

  \[ D \perp \{I, B\}, \quad I \perp D, \quad G \perp B \mid \{D, I\}, \]
  \[ B \perp \{D, G, L\} \mid I, \quad L \perp \{D, I, B\} \mid G \]

- **Chain rule of Bayesian networks**
  By the chain rule of probabilities:

  \[
  P(D, I, G, B, L) = P(D)P(I \mid D)P(G \mid D, I)P(B \mid D, I, G)P(L \mid D, I, G, B),
  \]
  \[
  = P(D)P(I)P(G \mid D, I)P(B \mid I)P(L \mid G)
  \]
The Student example

- **Local Markov property**
  Each node is conditionally independent of its non-descendants given its parents:

  \[
  D \perp \{I, B\}, \quad I \perp D, \quad G \perp B \mid \{D, I\},
  \]

  \[
  B \perp \{D, G, L\} \mid I, \quad L \perp \{D, I, B\} \mid G
  \]

- **Chain rule of Bayesian networks**
  By the chain rule of probabilities:

  \[
  P(D, I, G, B, L) = P(D)P(I \mid D)P(G \mid D, I)P(B \mid D, I, G)P(L \mid D, I, G, B),
  \]

  \[
  = P(D)P(I)P(G \mid D, I)P(B \mid I)P(L \mid G)
  \]

  These two definitions are equivalent
The Student example

- **Course difficulty**
- **Student intelligence**
- **Grade**
- **Baccalauréat**
- **Reference letter**

Probability distributions:
- \( P(D) \)
- \( P(I) \)
- \( P(G | D, I) \)
- \( P(L | G) \)
- \( P(B | I) \)
Bayesian networks in general

Described by

- A **directed acyclic graph**
  - Nodes ~ (discrete) random variables $X_1, \ldots, X_n$
  - Arrows ~ conditional (in)dependencies

- Local **conditional probability tables (CPT)**
  - $P(X_i \mid \text{parents}(X_i))$ for each node $X_i$
Bayesian networks in general

Two equivalent definitions

- **Local Markov property**
  Each node is conditionally independent of its non-descendants given its parents

- **Chain rule of Bayesian networks**

\[
P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid \text{parents}(X_i))
\]

Proof of the equivalence: Corollary 4 p.20 of (Pearl, 1988)
Base case: serial connection

\[ X \perp Z \mid Y \quad P(X, Y, Z) = P(X)P(Y \mid X)P(Z \mid Y) \]
Base case: serial connection

\[ X \perp Z \mid Y \quad P(X, Y, Z) = P(X)P(Y \mid X)P(Z \mid Y) \]

- Interpretation: chain of causality
  X “causes” Y that “causes” Z

\[ X \rightarrow Y \rightarrow Z \]
Base case: serial connection

\[ X \perp Z \mid Y \quad P(X, Y, Z) = P(X)P(Y \mid X)P(Z \mid Y) \]

- Interpretation: chain of causality
  - \( X \) “causes” \( Y \) that “causes” \( Z \)
- Information can flow between \( X \) and \( Z \) through \( Y \) (that is, observing \( X \) changes our belief in \( Z \) and vice versa), unless \( Y \) is observed
Base case: serial connection

\[ X \perp Z \mid Y \quad P(X, Y, Z) = P(X)P(Y \mid X)P(Z \mid Y) \]

- Interpretation: chain of causality
  - \( X \) “causes” \( Y \) that “causes” \( Z \)
- Information can flow between \( X \) and \( Z \) through \( Y \) (that is, observing \( X \) changes our belief in \( Z \) and vice versa), unless \( Y \) is observed
Base case: serial connection

\[ X \perp Z \mid Y \quad P(X,Y,Z) = P(X)P(Y\mid X)P(Z\mid Y) \]

- Interpretation: chain of causality
  X “causes” Y that “causes” Z
- Information can flow between X and Z through Y (that is, observing X changes our belief in Z and vice versa), unless Y is observed
Base case: serial connection

\[ X \perp Z \mid Y \quad P(X, Y, Z) = P(X)P(Y \mid X)P(Z \mid Y) \]

- Interpretation: chain of causality
  \( X \) “causes” \( Y \) that “causes” \( Z \)
- Information can flow between \( X \) and \( Z \) through \( Y \) (that is, observing \( X \) changes our belief in \( Z \) and vice versa), unless \( Y \) is observed
- Example: Markov chains
Base case: diverging connection

\[ X \perp Z \mid Y \]

\[ P(X, Y, Z) = P(X \mid Y)P(Y)P(Z \mid Y) \]
Base case: diverging connection

\[ X \perp Z \mid Y \quad P(X, Y, Z) = P(X \mid Y)P(Y)P(Z \mid Y) \]

- Interpretation: a single root cause \( Y \) with two observable consequences \( X \) and \( Z \)
Base case: diverging connection

\[ X \perp Z \mid Y \quad P(X, Y, Z) = P(X \mid Y)P(Y)P(Z \mid Y) \]

- Interpretation: a single root cause Y with two observable consequences X and Z
- Information can flow between X and Z through Y, unless Y is observed
Base case: diverging connection

\[ X \perp Z \mid Y \quad P(X, Y, Z) = P(X \mid Y)P(Y)P(Z \mid Y) \]

- Interpretation: a single root cause \( Y \) with two observable consequences \( X \) and \( Z \)
- Information can flow between \( X \) and \( Z \) through \( Y \), unless \( Y \) is observed
Base case: diverging connection

$X \perp Z \mid Y \quad P(X,Y,Z) = P(X \mid Y)P(Y)P(Z \mid Y)$

- Interpretation: a single root cause $Y$ with two observable consequences $X$ and $Z$
- Information can flow between $X$ and $Z$ through $Y$, unless $Y$ is observed
Base case: converging connection

\[ X \perp Z \quad \text{P}(X, Y, Z) = \text{P}(X)\text{P}(Y \mid X, Z)\text{P}(Z) \]
Base case: converging connection

\[ X \perp Z \quad P(X, Y, Z) = P(X)P(Y | X, Z)P(Z) \]

- Interpretation: two possible explanations X and Z for an observed consequence Y
Base case: converging connection

\[ X \perp Z \quad P(X, Y, Z) = P(X)P(Y \mid X, Z)P(Z) \]

- Interpretation: two possible explanations \( X \) and \( Z \) for an observed consequence \( Y \)
- “Explaining away” effect: information cannot flow between \( X \) and \( Z \), unless \( Y \) is observed
Base case: converging connection

\[ X \perp Z \quad P(X, Y, Z) = P(X)P(Y \mid X, Z)P(Z) \]

- Interpretation: two possible explanations X and Z for an observed consequence Y
- “Explaining away” effect: information cannot flow between X and Z, unless Y is observed
Base case: converging connection

\[ X \perp Z \quad P(X, Y, Z) = P(X)P(Y \mid X, Z)P(Z) \]

- Interpretation: two possible explanations \( X \) and \( Z \) for an observed consequence \( Y \)
- “Explaining away” effect: information cannot flow between \( X \) and \( Z \), unless \( Y \) is observed
Implied independencies

Similar to the “Strong Markov property” of Markov chains
Implied independencies

Similar to the “Strong Markov property” of Markov chains

Which are correct?
Implied independencies

Similar to the “Strong Markov property” of Markov chains

Which are correct?

1. $G \perp B$?
Implied independencies

Similar to the “Strong Markov property” of Markov chains

Which are correct?

1. \( G \perp B? \) No
Implied independencies

Similar to the “Strong Markov property” of Markov chains

Which are correct?

1. $G \perp B$? No
2. $B \perp L$?
Implied independencies

Similar to the “Strong Markov property” of Markov chains

Which are correct?

1. $G \perp B$? No
2. $B \perp L$? No
Implied independencies

Similar to the “Strong Markov property” of Markov chains

Which are correct?

1. G ⊥ B? No
2. B ⊥ L? No
3. D ⊥ L?
Implied independencies

Similar to the “Strong Markov property” of Markov chains

Which are correct?

1. $G \perp B$? No
2. $B \perp L$? No
3. $D \perp L$? No

4. $D \perp B$? Yes (by the local Markov property applied to $D$)
5. $D \perp B_j G$? No (“explaining away” effect)
6. $D \perp B_j \{I, G\}$? Yes
Implied independencies

Similar to the “Strong Markov property” of Markov chains

Which are correct?

1. $G \perp B$? No
2. $B \perp L$? No
3. $D \perp L$? No
4. $D \perp B$?
Implied independencies

Similar to the “Strong Markov property” of Markov chains

Which are correct?

1. G ⊥ B? No
2. B ⊥ L? No
3. D ⊥ L? No
4. D ⊥ B? Yes (by the local Markov property applied to D)
Implied independencies

Similar to the “Strong Markov property” of Markov chains

Which are correct?

1. $G \perp B$? No
2. $B \perp L$? No
3. $D \perp L$? No
4. $D \perp B$? Yes (by the local Markov property applied to D)
5. $D \perp B \mid G$?
Implied independencies

Similar to the “Strong Markov property” of Markov chains

Which are correct?

1. $G \perp B$? No
2. $B \perp L$? No
3. $D \perp L$? No
4. $D \perp B$? Yes (by the local Markov property applied to $D$)
5. $D \perp B \mid G$? No (“explaining away” effect)
Implied independencies

Similar to the “Strong Markov property” of Markov chains

Which are correct?

1. \( G \perp B? \) No
2. \( B \perp L? \) No
3. \( D \perp L? \) No
4. \( D \perp B? \) Yes (by the local Markov property applied to \( D \))
5. \( D \perp B | G? \) No (“explaining away” effect)
6. \( D \perp B | \{I, G\}? \)
Implied independencies

Similar to the “Strong Markov property” of Markov chains

Which are correct?

1. $G \perp B$? No
2. $B \perp L$? No
3. $D \perp L$? No
4. $D \perp B$? Yes (by the local Markov property applied to $D$)
5. $D \perp B \mid G$? No (“explaining away” effect)
6. $D \perp B \mid \{I, G\}$? Yes
Implied independencies

Proof of $\mathbb{D} \independent B \mid \{I, G\}$

$$P(D, B \mid I, G)$$
Implied independencies

Proof of $\mathbf{D \perp B | \{I, G\}}$

\[
P(D, B | I, G) = P(D | G, I)P(B | I, G)
\]
Implied independencies

Proof of $D \perp B \mid \{I, G\}$

$$P(D, B \mid I, G) = \frac{P(G \mid D, I, B)P(D, B \mid I)}{P(G \mid I)} \quad \text{(Bayes' rule)}$$

$$= P(D \mid G, I)P(B \mid I, G)$$
Implied independencies

Proof of $\text{D} \perp \text{B} \mid \{\text{I}, \text{G}\}$

$$P(\text{D}, \text{B} \mid \text{I}, \text{G}) = \frac{P(\text{G} \mid \text{D}, \text{I}, \text{B})P(\text{D}, \text{B} \mid \text{I})}{P(\text{G} \mid \text{I})} \quad \left(\text{Bayes' rule}\right)$$

$$= \frac{P(\text{G} \mid \text{D}, \text{I})P(\text{D}, \text{B} \mid \text{I})}{P(\text{G} \mid \text{I})} \quad \left(\text{local Markov property applied to G}\right)$$

$$= P(\text{D} \mid \text{G}, \text{I})P(\text{B} \mid \text{I}, \text{G})$$
Implied independencies

Proof of \( \mathcal{D} \perp B \mid \{I, G\} \)

\[
P(D, B \mid I, G) = \frac{P(G \mid D, I, B)P(D, B \mid I)}{P(G \mid I)} \quad \text{(Bayes' rule)}
\]

\[
= \frac{P(G \mid D, I)P(D, B \mid I)}{P(G \mid I)} \quad \text{(local Markov property applied to G)}
\]

\[
= \frac{P(G \mid D, I)P(D \mid I)P(B \mid D, I)}{P(G \mid I)} \quad \text{(definition of conditional probabilities)}
\]

\[
= P(D \mid G, I)P(B \mid I, G)
\]
Implied independencies

Proof of $\mathsf{D} \perp \mathsf{B} \mid \{\mathsf{I}, \mathsf{G}\}$

\[
P(\mathsf{D}, \mathsf{B} \mid \mathsf{I}, \mathsf{G}) = \frac{P(\mathsf{G} \mid \mathsf{D}, \mathsf{I}, \mathsf{B})P(\mathsf{D}, \mathsf{B} \mid \mathsf{I})}{P(\mathsf{G} \mid \mathsf{I})} \quad \text{(Bayes' rule)}
\]

\[
= \frac{P(\mathsf{G} \mid \mathsf{D}, \mathsf{I})P(\mathsf{D} \mid \mathsf{I})P(\mathsf{B} \mid \mathsf{D}, \mathsf{I})}{P(\mathsf{G} \mid \mathsf{I})} \quad \text{(local Markov property applied to $\mathsf{G}$)}
\]

\[
= \frac{P(\mathsf{G} \mid \mathsf{D}, \mathsf{I})P(\mathsf{D} \mid \mathsf{I})P(\mathsf{B} \mid \mathsf{D}, \mathsf{I})}{P(\mathsf{G} \mid \mathsf{I})} \cdot \frac{P(\mathsf{G} \mid \mathsf{I})}{P(\mathsf{G} \mid \mathsf{I})} \quad \text{(definition of conditional probabilities)}
\]

\[
= P(\mathsf{D} \mid \mathsf{G}, \mathsf{I})P(\mathsf{B} \mid \mathsf{I}, \mathsf{G})
\]
Implied independencies

Proof of \( D \perp B \mid \{I, G\} \)

\[
P(D, B \mid I, G) = \frac{P(G \mid D, I, B)P(D, B \mid I)}{P(G \mid I)} \quad \text{(Bayes' rule)}
\]

\[
= \frac{P(G \mid D, I)P(D, B \mid I)}{P(G \mid I)} \quad \text{(local Markov property applied to \( G \))}
\]

\[
= \frac{P(G \mid D, I)P(D \mid I)P(B \mid D, I)}{P(G \mid I)} \quad \text{(definition of conditional probabilities)}
\]

\[
= \frac{P(G \mid D, I)P(D \mid I)}{P(G \mid I)}P(B \mid D, I)
\]

\[
= P(D \mid G, I)P(B \mid D, I) \quad \text{(Bayes' rule)}
\]

\[
= P(D \mid G, I)P(B \mid I, G)
\]
Implied independencies

Proof of $\mathcal{D} \perp \mathcal{B} \mid \{\mathcal{I}, \mathcal{G}\}$

$$P(\mathcal{D}, \mathcal{B} \mid \mathcal{I}, \mathcal{G}) = \frac{P(\mathcal{G} \mid \mathcal{D}, \mathcal{I}, \mathcal{B})P(\mathcal{D}, \mathcal{B} \mid \mathcal{I})}{P(\mathcal{G} \mid \mathcal{I})} \quad \left(\text{Bayes' rule}\right)$$

$$= \frac{P(\mathcal{G} \mid \mathcal{D}, \mathcal{I})P(\mathcal{D} \mid \mathcal{I})P(\mathcal{B} \mid \mathcal{D}, \mathcal{I})}{P(\mathcal{G} \mid \mathcal{I})} \quad \left(\text{local Markov property applied to } \mathcal{G}\right)$$

$$= \frac{P(\mathcal{G} \mid \mathcal{D}, \mathcal{I})P(\mathcal{D} \mid \mathcal{I})P(\mathcal{B} \mid \mathcal{D}, \mathcal{I})}{P(\mathcal{G} \mid \mathcal{I})} \quad \left(\text{definition of conditional probabilities}\right)$$

$$= \frac{P(\mathcal{G} \mid \mathcal{D}, \mathcal{I})P(\mathcal{D} \mid \mathcal{I})}{P(\mathcal{G} \mid \mathcal{I})}P(\mathcal{B} \mid \mathcal{D}, \mathcal{I})$$

$$= P(\mathcal{D} \mid \mathcal{G}, \mathcal{I})P(\mathcal{B} \mid \mathcal{D}, \mathcal{I}) \quad \left(\text{Bayes' rule}\right)$$

$$= P(\mathcal{D} \mid \mathcal{G}, \mathcal{I})P(\mathcal{B} \mid \mathcal{I}, \mathcal{G}) \quad \left(\text{local Markov property applied to } \mathcal{B}\right)$$
Memory and time complexity

Parameters
Memory and time complexity

Parameters

- **n** number of random variables
  (typically, \( n \sim \) hundreds or thousands)

- **r** number of values each variable can take
- **d** maximum number of parents of a node

Memory complexity

- If we store the probability distribution:
  \[ O(rn) \] entries

- If we store the node parents and the conditional probability tables:
  \[ O(n(r \times rd)) \approx O(nrd) \] entries

What about the time complexity?
Memory and time complexity

Parameters

- $n$ number of random variables (typically, $n \sim$ hundreds or thousands)
- $r$ number of values each variable can take

Memory complexity

- If we store the probability distribution: $O(r^n)$ entries
- If we store the node parents and the conditional probability tables: $O(n \cdot (r \cdot r^d)) \approx O(nr^d)$ entries

What about the time complexity?
Memory and time complexity

Parameters

- \( n \) number of random variables
  (typically, \( n \sim \) hundreds or thousands)
- \( r \) number of values each variable can take
- \( d \) maximum number of parents of a node
Memory and time complexity

Parameters

- \( n \) number of random variables
  (typically, \( n \sim \) hundreds or thousands)
- \( r \) number of values each variable can take
- \( d \uparrow \) maximum number of parents of a node

Memory complexity
Memory and time complexity

Parameters

- $n$ number of random variables
  (typically, $n \sim$ hundreds or thousands)
- $r$ number of values each variable can take
- $d$ maximum number of parents of a node

Memory complexity

- If we store the probability distribution: $O(r^n)$ entries
Memory and time complexity

Parameters

- \( n \) number of random variables
  (typically, \( n \sim \) hundreds or thousands)
- \( r \) number of values each variable can take
- \( d^\dagger \) maximum number of parents of a node

Memory complexity

- If we store the probability distribution: \( O(r^n) \) entries
- If we store the node parents and the conditional probability tables: \( O(n(r + r^{d^\dagger})) = O(nr^{d^\dagger}) \) entries
Memory and time complexity

Parameters

- **n**: number of random variables
  (typically, \( n \sim \) hundreds or thousands)
- **r**: number of values each variable can take
- **d\( \dagger \)**: maximum number of parents of a node

Memory complexity

- If we store the probability distribution: \( O(r^n) \) entries
- If we store the node parents and the conditional probability tables: \( O(n(r + r^{d\dagger})) = O(nr^{d\dagger}) \) entries

What about the time complexity?
Inference

“A guess that you make or an opinion that you form based on the information that you have” (Cambridge dictionary)
Inference

“A guess that you make or an opinion that you form based on the information that you have” (Cambridge dictionary)

→ Bayesian networks: compute or update the belief in each variable given some evidence
Inference

“A guess that you make or an opinion that you form based on the information that you have” (Cambridge dictionary)

→ Bayesian networks: compute or update the belief in each variable given some evidence

**Belief propagation**, a.k.a. **sum-product message passing**: Propagate the information through the network, starting from the evidence node(s)
Inference

“A guess that you make or an opinion that you form based on the information that you have” (Cambridge dictionary)

→ Bayesian networks: compute or update the belief in each variable given some evidence

**Belief propagation**, a.k.a. **sum-product message passing**: Propagate the information through the network, starting from the evidence node(s)

- Each variable is a “separate processor” (a neuron?) that knows its own CPT and the messages received from its direct neighbors (Pearl, 1982)
Inference

“A guess that you make or an opinion that you form based on the information that you have” (Cambridge dictionary)

→ Bayesian networks: compute or update the belief in each variable given some evidence

**Belief propagation, a.k.a. sum-product message passing:**
Propagate the information through the network, starting from the evidence node(s)

- Each variable is a “separate processor” (a neuron?) that knows its own CPT and the messages received from its direct neighbors (Pearl, 1982)
- Dynamic programming
Outline

Reminders on probability theory

Bayesian networks

Belief propagation in trees

Belief propagation in polytrees
Tree Bayesian network

Each node (except the root) has at most one parent
Tree Bayesian network

Each node (except the root) has at most one parent.

Remark: We will explain the propagation algorithm on this toy example borrowed from (Pearl, 1988).
Tree Bayesian network

Each node (except the root) has at most one parent

A

B

C

D

E

F

Remark: We will explain the propagation algorithm on this toy example borrowed from (Pearl, 1988)
Tree Bayesian network

Each node (except the root) has at most one parent.

Each node *separates* the tree: its non-descendants and the subtrees rooted at each of its children are conditionally independent given this node.
Tree Bayesian network

Each node (except the root) has at most one parent.

Each node *separates* the tree: its non-descendants and the subtrees rooted at each of its children are conditionally independent given this node.

**Remark:** We will explain the propagation algorithm on this toy example borrowed from (Pearl, 1988).
No evidence

\[
\begin{align*}
P(A) & = \sum_B P(B) \sum_A P(A|B) P(A) \\
P(B) & = \sum_A P(A) \sum_B P(B|A) P(B) \\
P(C) & = \sum_B P(B) \sum_D P(D|B) P(C|B) P(D) \\
P(D) & = \sum_B P(B) \sum_E P(E|B) P(C|B) P(D) \\
P(E) & = \sum_F P(F|D) P(E|D) P(D) \\
P(F) & = \sum_D P(D|E) P(F|E) P(E) \\
\end{align*}
\]
No evidence

- P(A): parameter
No evidence

- $P(A)$: parameter
- $P(B) = \sum_{A} P(B \mid A)P(A)$
No evidence

- \( P(A) \): parameter

- \( P(B) = \sum_A P(B | A) P(A) \)
No evidence

- $P(A)$: parameter
- $P(B) = \sum_A P(B \mid A)P(A)$
- $P(C) = \sum_B P(C \mid B)P(B)$
No evidence

- $P(A)$: parameter

- $P(B) = \sum_A P(B | A)P(A)$

- $P(C) = \sum_B P(C | B)P(B)$
No evidence

- $P(A)$: parameter
- $P(B) = \sum_A P(B \mid A)P(A)$
- $P(C) = \sum_B P(C \mid B)P(B)$
- $P(D) = \sum_B P(D \mid B)P(B)$
No evidence

- \( P(A) \): parameter

- \( P(B) = \sum_A P(B|A)P(A) \)

- \( P(C) = \sum_B P(C|B)P(B) \)

- \( P(D) = \sum_B P(D|B)P(B) \)
No evidence

- $P(A)$: parameter

- $P(B) = \sum_A P(B \mid A)P(A)$

- $P(C) = \sum_B P(C \mid B)P(B)$

- $P(D) = \sum_B P(D \mid B)P(B)$
No evidence

- $P(A)$: parameter

- $P(B) = \sum_A P(B \mid A)P(A)$

- $P(C) = \sum_B P(C \mid B)P(B)$

- $P(D) = \sum_B P(D \mid B)P(B)$

**Top-down propagation**

Complexity $O(nr^2)$
No evidence

- $P(A)$: parameter
- $P(B) = \sum_A P(B \mid A)P(A)$
- $P(C) = \sum_B P(C \mid B)P(B)$
- $P(D) = \sum_B P(D \mid B)P(B)$

Top-down propagation
Complexity $O(nr^2)$
Three pieces of evidence

Evidence: We observe that \( C = c \), \( E = e \), and \( F = f \)

Objective: Compute the belief \( \text{BEL}(X) \) of each node \( X \)

Principle: Propagate the information through the network, starting from the evidence nodes.
Three pieces of evidence

- **Evidence**: We observe that $C = c$, $E = e$, and $F = f$

- **Objective**: Compute the belief $\text{BEL}(X) = P(X | c, e, f)$ of each node $X$
Three pieces of evidence

- **Evidence:** We observe that $C = c$, $E = e$, and $F = f$

- **Objective:** Compute the belief $\text{BEL}(X) = P(X | c, e, f)$ of each node $X$

- **Principle:** Propagate the information through the network, starting from the evidence nodes
Causal and diagnostic support

By Bayes’ rule:

\[
\frac{P(A|c, e, f)}{P(c, e, f|A)} \times \frac{P(A)}{P(c, e, f)}
\]

\[
\frac{P(B|c, e, f)}{P(c, e, f|B)} \times \frac{P(B)}{P(c, e, f)}
\]

\[
\frac{P(D|c, e, f)}{P(e, f|D, c)} \times \frac{P(D)}{P(e, f|D)}
\]
Causal and diagnostic support

By Bayes’ rule:

\[
P(D \mid c, e, f) = \frac{P(e, f \mid D, c)P(D \mid c)}{P(e, f \mid c)}
\]
Causal and diagnostic support

By Bayes’ rule:

\[ P(D \mid c, e, f) = \frac{P(e, f \mid D, c)P(D \mid c)}{P(e, f \mid c)P(e, f \mid D)P(D \mid c)} \]
Causal and diagnostic support

By Bayes’ rule:

\[ P(D \mid c, e, f) = \frac{P(e, f \mid D, c)P(D \mid c)}{P(e, f \mid c)} \]

\[ = \frac{P(e, f \mid D)P(D \mid c)}{P(e, f \mid c)} \]

\[ \propto P(e, f \mid D)P(D \mid c) \]
Causal and diagnostic support

By Bayes’ rule:

\[ P(B | c, e, f) = \frac{P(c, e, f | B)P(B)}{P(c, e, f)} \]

\[ P(D | c, e, f) = \frac{P(e, f | D, c)P(D | c)}{P(e, f | c)} \]

\[ \propto P(e, f | D)P(D | c) \]
Causal and diagnostic support

By Bayes’ rule:

- \( P(B \mid c, e, f) = \frac{P(c, e, f \mid B)P(B)}{P(c, e, f)} \propto P(c, e, f \mid B)P(B) \)

- \( P(D \mid c, e, f) = \frac{P(e, f \mid D, c)P(D \mid c)}{P(e, f \mid c)} = \frac{P(e, f \mid D)P(D \mid c)}{P(e, f \mid c)} \propto P(e, f \mid D)P(D \mid c) \)
Causal and diagnostic support

By Bayes’ rule:

- \( P(A \mid c, e, f) = \frac{P(c, e, f \mid A)P(A)}{P(c, e, f)} \)

- \( P(B \mid c, e, f) = \frac{P(c, e, f \mid B)P(B)}{P(c, e, f)} \approx P(c, e, f \mid B)P(B) \)

- \( P(D \mid c, e, f) = \frac{P(e, f \mid D, c)P(D \mid c)}{P(e, f \mid c)} \approx P(e, f \mid D)P(D \mid c) \)
Causal and diagnostic support

By Bayes’ rule:

- \( P(A \mid c, e, f) = \frac{P(c, e, f \mid A)P(A)}{P(c, e, f)} \propto P(c, e, f \mid A)P(A) \)
- \( P(B \mid c, e, f) = \frac{P(c, e, f \mid B)P(B)}{P(c, e, f)} \propto P(c, e, f \mid B)P(B) \)
- \( P(D \mid c, e, f) = \frac{P(e, f \mid D, c)P(D \mid c)}{P(e, f \mid c)} = \frac{P(e, f \mid D)P(D \mid c)}{P(e, f \mid c)} \propto P(e, f \mid D)P(D \mid c) \)
Causal and diagnostic support

For each X, we compute

- Diagnostic support $P(\text{evidence below } X | X)$
  Bottom-up propagation

- Causal support $P(X | \text{evidence above } X)$
  Top-down propagation

$BEL(X) \propto P(\text{evidence below } X | X) \times P(X | \text{evidence above } X)$
Diagnostic support $P\left(\text{evidence below } X \mid X\right)$

Bottom-up propagation
Diagnostic support \( P\left(\text{evidence below } X \mid X\right) \)

Bottom-up propagation

- \( P(e,f \mid D) = P(e \mid D)P(f \mid D) \)
Diagnosis support $P\left(\text{evidence below } X \mid X\right)$

Bottom-up propagation

- $P(e, f \mid D) = P(e \mid D)P(f \mid D)$
Diagnostic support $P(\text{evidence below } X \mid X)$

Bottom-up propagation

- $P(e,f \mid D) = P(e \mid D)P(f \mid D)$
- $P(c,e,f \mid B) = P(c \mid B)P(e,f \mid B)$
Diagnostic support \( P\left(\text{evidence below } X \mid X\right) \)

Bottom-up propagation

- \( P(e, f \mid D) = P(e \mid D)P(f \mid D) \)
- \( P(c, e, f \mid B) = P(c \mid B)P(e, f \mid B) \)
Diagnostic support $P(\text{evidence below } X \mid X)$

Bottom-up propagation

- $P(e, f \mid D) = P(e \mid D)P(f \mid D)$
- $P(c, e, f \mid B) = P(c \mid B)P(e, f \mid B)$

Compute $P(e, f \mid B)$:

$P(e, f \mid B) = \sum_{D} P(e, f \mid B, D)P(D \mid B)$

$= \sum_{D} P(e, f \mid D)P(D \mid B)$

31/40 © 2018 Nokia
Diagnostic support $P(\text{evidence below } X \mid X)$

Bottom-up propagation

- $P(e, f \mid D) = P(e \mid D)P(f \mid D)$
- $P(c, e, f \mid B) = P(c \mid B)P(e, f \mid B)$

Compute $P(e, f \mid B)$:

$$P(e, f \mid B) = \sum_D P(e, f \mid B, D)P(D \mid B)$$

$$= \sum_D P(e, f \mid D)P(D \mid B)$$
Diagnostic support $P\left(\text{evidence below } X \mid X\right)$

- $P(c, e, f \mid A) = P(c, e, f \mid A)$

Compute $P(c, e, f \mid A)$:

$$P(c, e, f \mid A) = \sum_B P(c, e, f \mid A, B)P(B \mid A)$$

$$= \sum_B P(c, e, f \mid B)P(B \mid A)$$
Diagnostic support \( P_evidence_below X \mid X \)

- \( P(c, e, f \mid A) = P(c, e, f \mid A) \)

Compute \( P(c, e, f \mid A) \):

\[
P(c, e, f \mid A) = \sum_B P(c, e, f \mid A, B) P(B \mid A)
\]

\[
= \sum_B P(c, e, f \mid B) P(B \mid A)
\]
Diagnostic support $P(evidence \text{ below } X \mid X)$

- $P(c, e, f \mid A) = P(c, e, f \mid A)$

Compute $P(c, e, f \mid A)$:

$$P(c, e, f \mid A) = \sum_{B} P(c, e, f \mid A, B)P(B \mid A)$$

$$= \sum_{B} P(c, e, f \mid B)P(B \mid A)$$
Causal support $P(X \mid \text{evidence above } X)$
Causal support $P(X \mid \text{evidence above } X)$

- $P(A)$: parameter

Bottom-up $P(\text{evidence below } Y \mid X)$
Causal support $P(X \mid \text{evidence above } X)$

- $P(A)$: parameter
  \[ \rightarrow \text{BEL}(A) \propto P(c, e, f \mid A)P(A) \]

Bottom-up $P(\text{evidence below } Y \mid X)$
Causal support $P(X \mid \text{evidence above } X)$

- $P(A)$: parameter
  - $\rightarrow \text{BEL}(A) \propto P(c, e, f \mid A)P(A)$
- $P(B) = \sum_{A} P(B \mid A)P(A)$

Bottom-up

$P(\text{evidence below } Y \mid X)$
Causal support $P(X \mid \text{evidence above } X)$

- $P(A)$: parameter
  $$\rightarrow \text{BEL}(A) \propto P(c, e, f \mid A)P(A)$$

- $P(B) = \sum_A P(B \mid A)P(A)$

Bottom-up $P(\text{evidence below } Y \mid X)$
Causal support $P(X \mid \text{evidence above } X)$

- $P(A)$: parameter
  - $\rightarrow \text{BEL}(A) \propto P(c, e, f \mid A)P(A)$

- $P(B) = \sum_A P(B \mid A)P(A)$
  - $\rightarrow \text{BEL}(B) \propto P(c, e, f \mid B)P(B)$

Bottom-up $P(\text{evidence below } Y \mid X)$
Causal support $P(X \mid \text{evidence above } X)$

A

P(A)

B

C

D

e

f

X

Y

Bottom-up

$P(\text{evidence below } Y \mid X)$
Causal support $P(X \mid \text{evidence above } X)$

- $P(D \mid c) = \sum_B P(D \mid B, c)P(B \mid c)$

Bottom-up $P(\text{evidence below } Y \mid X)$
Causal support $P(X \mid \text{evidence above } X)$

- $P(D \mid c) = \sum_B P(D \mid B, c)P(B \mid c)$
  
  $= \sum_B P(D \mid B)P(B \mid c)$
Causal support $P(X \mid \text{evidence above } X)$

- $P(D \mid c) = \sum_B P(D \mid B, c)P(B \mid c)$
- $= \sum_B P(D \mid B)P(B \mid c)$
Causal support $P(X \mid \text{evidence above } X)$

- $P(D \mid c) = \sum_B P(D \mid B, c)P(B \mid c) = \sum_B P(D \mid B)P(B \mid c)$

Compute $P(B \mid c)$:

$P(B \mid c, e, f) = \frac{P(e, f \mid B, c)P(B \mid c)}{P(e, f \mid c)}$  

i.e. $P(B \mid c) \propto \frac{\text{BEL}(B)}{P(e, f \mid B)}$
Causal support $P(X \mid \text{evidence above } X)$

- $P(D \mid c) = \sum_B P(D \mid B, c)P(B \mid c) = \sum_B P(D \mid B)P(B \mid c)$

Compute $P(B \mid c)$:

$$P(B \mid c, e, f) = \frac{P(e, f \mid B, c)P(B \mid c)}{P(e, f \mid c)}$$

i.e. $P(B \mid c) \propto \frac{\text{BEL}(B)}{P(e, f \mid B)}$

$$\rightarrow \text{BEL}(D) \propto P(e, f \mid D)P(D \mid c)$$
Causal support $P(X \mid \text{evidence above } X)$

- $P(D \mid c) = \sum_B P(D \mid B, c)P(B \mid c)$
  
  $= \sum_B P(D \mid B)P(B \mid c)$

Compute $P(B \mid c)$:

$P(B \mid c, e, f) = \frac{P(e, f \mid B, c)P(B \mid c)}{P(e, f \mid c)}$

i.e. $P(B \mid c) \propto \frac{\text{BEL}(B)}{P(e, f \mid B)}$

$\rightarrow \text{BEL}(D) \propto P(e, f \mid D)P(D \mid c)$
Summary

Algorithm

Bottom-up
\[ P(\text{evidence below } Y \mid X) \]

Top-down
\[ P(X \mid \text{evidence above } Y) \]
Summary

Algorithm

- Diagnostic support $P(\text{evidence below } X \mid X)$
  - Bottom-up propagation

- Causal support $P(X \mid \text{evidence above } X)$
  - Top-down propagation

In general
- Use a topological ordering
- Complexity: $O(\text{rd} \# \text{Å} r^2 \text{Å} r)$
Summary

Algorithm

- Diagnostic support $P(\text{evidence below } X \mid X)$
  Bottom-up propagation
- Causal support $P(X \mid \text{evidence above } X)$
  Top-down propagation

Bottom-up propagation

Top-down propagation
Summary

Algorithm

- Diagnostic support $P(\text{evidence below } X \mid X)$
  Bottom-up propagation
- Causal support $P(X \mid \text{evidence above } X)$
  Top-down propagation

In general

Bottom-up
$P(\text{evidence below } Y \mid X)$

Top-down
$P(X \mid \text{evidence above } Y)$
Summary

Algorithm

- Diagnostic support $P(evidence \, below \, X \, | \, X)$
  Bottom-up propagation
- Causal support $P(X \, | \, evidence \, above \, X)$
  Top-down propagation

In general

- Use a topological ordering
Summary

Algorithm

- Diagnostic support \( P(\text{evidence below } X \mid X) \)
  Bottom-up propagation
- Causal support \( P(X \mid \text{evidence above } X) \)
  Top-down propagation

In general

- Use a topological ordering
- Complexity: \( O(rd^1 + r^2 + r) \)
Additional remarks

- If the evidence node is not a leaf: add a phantom node

```
if the evidence node is not a leaf: add a phantom node
```

The calculations can be written as matrix products.

Belief, causal and diagnostic supports, messages → Vectors

CPT → Matrix

Additional remarks

- If the evidence node is not a leaf: add a phantom node
- The calculations can be written as matrix products
  - Belief, causal and diagnostic supports, messages $\sim$ Vectors
  - CPT $\sim$ Matrix

Bottom-up $P(\text{evidence below } Y \mid X)$

Top-down $P(X \mid \text{evidence above } Y)$
Additional remarks

- If the evidence node is not a leaf: add a phantom node
- The calculations can be written as matrix products
  - Belief, causal and diagnostic supports, messages ~ Vectors
  - CPT ~ Matrix
Outline

Reminders on probability theory

Bayesian networks

Belief propagation in trees

Belief propagation in polytrees
Polytree (or singly-connected) Bayesian network

The underlying undirected graph is a tree
Polytree (or singly-connected) Bayesian network

The underlying undirected graph is a tree

Separation properties
Polytree (or singly-connected) Bayesian network

The underlying undirected graph is a tree

**Separation properties**

- Given a node, the non-descendants and the subtrees rooted at each child are independent
Polytree (or singly-connected) Bayesian network

The underlying undirected graph is a tree

Separation properties

- Given a node, the non-descendants and the subtrees rooted at each child are independent
- If we don’t condition on a node nor any of its descendants, the inverted subtrees rooted at its ancestors are independent
Belief propagation

$$P(X_j | \text{evidence}) / P(\text{evidence} \bar{X}) \leq P(X \bar{evidence} \bar{X})$$

Diagnostic support

Bottom-up propagation

#XC and #XD are independent given X

Causal support

Top-down propagation

"AX and "BX are independent
Belief propagation

\[ P(X \mid \text{evidence}) \propto P(\text{evidence below } X \mid X) \times P(X \mid \text{evidence above } X) \]
Belief propagation

\[ P(X \mid \text{evidence}) \propto P(\text{evidence below } X \mid X) \times P(X \mid \text{evidence above } X) \]
Belief propagation

\[ P(X \mid \text{evidence}) \propto P\left(\text{evidence below } X \mid X\right) \times P\left(X \mid \text{evidence above } X\right) \]
Belief propagation

\[ P(X \mid \text{evidence}) \propto P\left(\text{evidence below } X \mid X\right) \times P\left(X \mid \text{evidence above } X\right) \]
Belief propagation

$$P(X \mid \text{evidence}) \propto P\left(\text{evidence below } X \mid X\right) \times P\left(X \mid \text{evidence above } X\right)$$
Belief propagation

\[ P(X \mid \text{evidence}) \propto P(\text{evidence below } X \mid X) \times P(X \mid \text{evidence above } X) \]

- Diagnostic support \( P(\text{evidence below } X \mid X) \)
  Bottom-up propagation
  \( \downarrow XC \) and \( \downarrow XD \) are independent given \( X \)
Belief propagation

\[ P(X \mid \text{evidence}) \propto P\left(\text{evidence below } X \mid X\right) \times P\left(X \mid \text{evidence above } X\right) \]

- Diagnostic support \( P\left(\text{evidence below } X \mid X\right) \)
  - Bottom-up propagation
  - \( \downarrow XC \) and \( \downarrow XD \) are independent given given \( X \)

- Causal support \( P\left(X \mid \text{evidence above } X\right) \)
  - Top-down propagation
  - \( \uparrow AX \) and \( \uparrow BX \) are independent
Conclusion
Conclusion

- **Bayesian networks**
  A memory-efficient way of storing a PMF by leveraging conditional independencies between variables
Conclusion

- **Bayesian networks**
  A memory-efficient way of storing a PMF by leveraging conditional independencies between variables

- **Belief propagation**
  A time-efficient algorithm for computing the belief
  - Asynchronous, parallelizable
  - Exact in (poly)trees
  - In general, extended to the junction tree algorithm and to other (approximate) algorithms