Directed Graphical Models: Bayesian Networks

40-957 Special Topics in Artificial Intelligence: Probabilistic Graphical Models
Sharif University of Technology

Soleymani
Spring 2015
Basics

- Multivariate distributions with large number of variables
- Independencies assumptions are useful
  - Independence and conditional independence relationships simplify representation and alleviate inference complexities
- Bayesian networks enables us to incorporate domain knowledge and structures
  - Modular combination of heterogeneous parts
  - Combining data and knowledge (Bayesian philosophy)
Conditional and marginal independence

- *X* and *Y* are **conditionally independent** given *Z* if:

  \[ X \perp Y | Z \]
  \[ P(X, Y | Z) = P(X | Z)P(Y | Z) \]
  \[ \forall x \in Val(X), y \in Val(Y), z \in Val(Z) \]
  \[ P(X = x, Y = y | Z = z) = P(X = x | Z = z)P(Y = y | Z = z) \]

- *X* and *Y* are **marginal independent** if:

  \[ X \perp Y | \emptyset \]
  \[ P(X, Y) = P(X)P(Y) \]
Bayesian network definition

- **Bayesian Network**
  - Qualitative specification by a **Directed Acyclic Graph (DAG)**
    - Each node denotes a random variable
    - Edges denote dependencies
      - $X \rightarrow Y$ shows a "direct influence" of $X$ on $Y$ ($X$ is a parent of $Y$)
  - Quantitative specification by CPDs
    - CPD for each node $X_i$ defines $P(X_i \mid Pa(X_i))$

- Bayesian Network represents a joint distribution over variables (via DAG and CPDs) compactly in a factorized way:

$$ P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid Pa(X_i)) $$
Burglary example

John do not perceive burglaries directly

John do not perceive minor earthquakes
Burglary example

- Bayesian networks define joint distribution (over the variables) in terms of the graph structure and conditional probability distributions

\[
P(B, E, A, J, M) = P(B)P(E)P(A|B, E)P(J|A)P(M|A)
\]
Burglary example: DAG + CPTs

Burglary: \( P(B) = .001 \)

Earthquake: \( P(E) = .002 \)

Short-hands:

- \( j \): JohnCalls = True
- \( b \): Burglary = False

\[
P(j | A) = \begin{cases} 
  .90 & \text{if } A = t \\
  .05 & \text{if } A = f 
\end{cases}
\]

\[
P(m | A) = \begin{cases} 
  .70 & \text{if } A = t \\
  .01 & \text{if } A = f 
\end{cases}
\]

CPDs as quantitative specification:

| \( B \) | \( E \) | \( P(a | B, E) \) |
|---|---|---|
| t | t | .95 |
| t | f | .94 |
| f | t | .29 |
| f | f | .001 |

…
Burglary example: full joint probability

\[ P(J, M, A, B, E) = P(J \mid A) P(M \mid A) P(A \mid B, E) P(B) P(E) \]

\[ P(j, m, a, b, \bar{e}) = P(j \mid a) P(m \mid a) P(a \mid b, \bar{e}) P(\bar{b}) P(\bar{e}) \]

\[ = 0.9 \times 0.7 \times 0.001 \times 0.999 \times 0.998 = 0.000628 \]
Burglary example: inference

- Partial joint probability distribution: Joint probability distribution for a subset variables
  \[ P(J, M, B) = \sum_A \sum_E P(J, M, A, B, E) \]

- Conditional probability distribution:
  \[ P(B|J = t, M = f) = \frac{P(J=t, M=f, B)}{P(J=t, M=f)} = \frac{\sum_A \sum_E P(J=t, M=f, A, B, E)}{\sum_B \sum_A \sum_E P(J=t, M=f, A, B, E)} \]
Student example

\[
P(I = 0.55)
\]

Intelligence

Difficulty

Rank

Grade

Letter

\[
P(D = 0.65)
\]

| \(I\) | \(P(R = 1|I)\) | \(P(G|I, D)\) |
|------|----------------|----------------|
| \(f\) | 0.1            | \(G = 1\) |
| \(t\) | 0.7            | \(G = 2\) |
| \(t\) | 0.5            | \(G = 3\) |

\[
G | \(P(L = t|G)\)
---|----------------|
1  | 0.9
2  | 0.5
3  | 0.05
\]
Compact representation

- A CPT for a Boolean variable with $k$ Boolean parents requires:
  - $2^k$ rows: different combinations of parent values
    - $k = 0$: one row showing the prior probability

- If each variable has no more than $k$ parents
  - **Full joint distribution** requires $2^n - 1$ numbers
  - **Bayesian network** requires atmost $n \times 2^k - 1$ numbers (linear with $n$)
  - $\Rightarrow$ Exponential reduction in number of parameters
Continuous variables example

- Linear Gaussian

\[ X \sim N(0, 1) \]
\[ Y | X \sim N(b + X, \sigma) \]

\[ b = 0.5 \]
\[ \sigma = 0.1 \]
Bayesian network semantics

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

\[
X_i \perp \text{Non\_Descendants}(X_i) | Pa(X_i)
\]

- Are local independencies all of the conditional independencies implied by a BN?
Factorization & independence

- Let $G$ be a graph over $X_1, \ldots, X_n$, distribution $P$ factorizes over $G$ if:

  $$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | Pa(X_i))$$

- Factorization $\Rightarrow$ Independence
  - If $P$ factorizes over $G$, then any variable in $P$ is independent of its non-descendants given its parents (in $G$)
    - Factorization according to $G$ implies the associated conditional independencies.

- Independence $\Rightarrow$ Factorization
  - If any variable in the distribution $P$ is independent of its non-descendants given its parents (in the graph $G$) then $P$ factorizes over $G$
    - Conditional independencies imply factorization of the joint distribution (into a product of simpler terms)
Independence $\Rightarrow$ factorization

- Consider the chain rule:

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|X_1, \ldots, X_{i-1})$$

- We can simplify it through conditional independencies assumptions
  - $P(X_i|X_1, X_2, \ldots, X_{i-1}) = P(X_i|\text{Parents}(X_i))$ can be showed using $X_i \perp \text{Non.Descendants}(X_i)|\text{Pa}(X_i)$
Equivalence Theorem

For a graph $G$:
- Let $D_1$ denote the family of all distributions that satisfy conditional independencies of $G$
- Let $D_2$ denote the family of all distributions that factor according to $G$
- $\Rightarrow D_1 \equiv D_2$. 
Other independencies

- Are there other independences that hold for every distribution $P$ that factorizes over $G$?

- According to the graphical criterion called D-separation, we can find independencies from the graph:
  - If $P$ factorizes over $G$, can we read these independencies from the structure of $G$?
Basic structures

- $X \perp Y | Z$
- $X \perp Y | Z$
- $X \perp Y$

Explaining away
Let $A, B, C$ denote three disjoint sets of nodes, $A$ is **d-separated** from $B$ by $C$ iff $A \perp B | C$

$A$ is **d-separated** from $B$ by $C$ if all undirected paths between $A$ and $B$ are **blocked** by $C$.
Undirected path blocking

- Head-to-tail at a node $Z \in C$

  ![Diagram of head-to-tail at a node $Z \in C$]

  $X \in A$  $Z \in C$  $Y \in B$

- Tail-to-tail at a node $Z \in C$

  ![Diagram of tail-to-tail at a node $Z \in C$]

  $X \in A$  $Z \in C$  $Y \in B$

- Head-to-head (i.e., v-structure) at a node $Z$ ($Z \notin C$ & none of its descendants are in $C$)

  ![Diagram of head-to-head at a node $Z$ ($Z \notin C$ & none of its descendants are in $C$)]

  $X \in A$  $Y \in B$
Undirected path blocking

In all trails (undirected paths) between A and B:

- A node in the path is in C where the path at the node do not meet head-to-head.
- Or a head-to-head node in the path, and neither the node, nor any of its descendants, is in C.
D-separation: active trail view

- **Definition:** \(X\) and \(Y\) are d-separated in \(G\) given \(Z\) if there is no active trail in \(G\) between \(X\) and \(Y\) given \(Z\).

- A trail between \(X\) and \(Y\) is **active**: 
  - for any v-structure node \(U\) in the trail \(X \rightarrow U \leftarrow \cdots Y\), neither \(U\) nor any of its descendants are in \(Z\).
  - other nodes in this trail are not in \(Z\).
D-separation: example

- Intelligence
- Difficulty
- Rank
- Grade
- Letter

\[
\begin{align*}
R \perp G | I \\
R \perp D | I \\
R \perp D | G & \times \\
R \perp D | L & \times \\
R \perp L | G \\
D \perp L | G
\end{align*}
\]
Markov Blanket in Bayesian Network

- A variable is conditionally independent of all other variables given its Markov blanket

- Markov blanket of a node:
  - All parents
  - Children
  - Co-parents of children
D-Separation: soundness & completeness

- **Soundness**: Any conditional independence properties we can derive from \( G \) should hold for the probability distribution that factorize over \( G \)
  - **Theorem**: If \( P \) factorizes over \( G \), and \( d\text{-sep}_G(X,Y|Z) \) then \( P \) satisfies \( X \perp Y|Z \)

- **Weak completeness**:
  - For almost all distributions \( P \) that factorize over \( G \), if \( X \perp Y|Z \) is in \( P \) then \( X \) and \( Y \) are d-separated given \( Z \) in the graph \( G \)
    - There can be independencies in \( P \) that are not in conditional independence properties of \( G \)
I-map

- \( I(G) = \{(X \perp Y|Z) : \text{d-sep}_G(X, Y|Z)\} \)

- **Definition:** If \( P \) satisfies \( I(G) \), we say that \( G \) is an I-map (independencies map) of \( P \)
  - \( I(G) \subseteq I(P) \) where \( I(P) = \{(X \perp Y|Z) : P \models (X, Y|Z)\} \)

\( I(P) \): All conditional independence relations satisfied in \( P \)
Minimal I-map

- When more independence relations exist in the graph
  - $\Rightarrow$ sparser representation (fewer parameters)
  - $\Rightarrow$ more informative or intuitive representation

- We want a graph that captures as much of the structure (conditional independence relations) in $P$ as possible

- $G$ is a **minimal I-map** for $P$ if it is an I-map for $P$, and also the removal of each edge from $G$ renders it not an I-map.
  - Minimal I-map may still not capture $I(P)$
Perfect map

- $G$ is a **D-map** for a distribution $P$ if every conditional independence relation satisfied by the distribution is also in $I(G)$.

- $G$ is a **perfect map** (**P-map**) for a distribution $P$ is both an $I$-map and a $D$-map for that distribution.

Theorem: Not every distribution has a perfect map as a DAG.

- A distribution $P$ with the independencies
  \[ I(P) = \{A \perp C \mid \{B, D\}, \ B \perp D \mid \{A, C\}\} \]
  cannot be represented by any Bayesian network.
Perfect map

- A perfect map of a distribution is great, but may not exist for many distributions.

- A distribution \( P \) can have many P-map graphs (all of them are I-equivalent).

- A minimal I-map graph \( G \) for distribution \( P \) may be far from a guarantee that \( G \) contains all independencies in \( P \).
I-equivalence

- **Definition:** Two graphs $G_1$ and $G_2$ over a set of variables are I-equivalent if $I(G_1) = I(G_2)$

- Most graphs have many I-equivalent variants
I-equivalence

- If $G_1$ and $G_2$ have the same skeleton and the same set of immoralities (v-structures without direct edge between parents) then they are I-equivalent.
Bayesian networks: summary

- **Bayesian network** is a pair \((G, \text{CPDs})\) where \(G\) is a DAG and CPDs can be used to find a joint distribution \(P\) that factorizes over \(G\)
  - Each CPD is the conditional distribution \(P(X_i|Pa(X_i))\) associated to the graph node \(X_i\).

- We can show “causality”, “generative schemes”, “asymmetric influences”, etc., between variables via a Bayesian network

- We can find local and global independence relations from the graph structure via \(d\)-separation criteria