BN Semantics

Graphical Models – 10708
Carlos Guestrin
Carnegie Mellon University
September 14th, 2005
Announcements

- **Homework 1:**
  - Out later today
  - Due October 3rd – **beginning of class**!
  - It’s hard – start early, ask questions

- **Collaboration policy**
  - OK to discuss in groups
  - Tell us on your paper who you talked with
  - Each person must write their own unique paper
  - No searching the web, papers, etc. for answers, we trust you want to learn

- We are looking into room changes

- We cannot take official auditors for this class 😞
  - Too many people already
Basic concepts for random variables

- Atomic outcome: assignment \( x_1, \ldots, x_n \) to \( X_1, \ldots, X_n \)

- Conditional probability: \( P(X,Y) = P(X)P(Y|X) \)

- Bayes rule: \( P(X|Y) = \)

- Chain rule:
  - \( P(X_1, \ldots, X_n) = P(X_1)P(X_2|X_1) \cdots P(X_k|X_1, \ldots, X_{k-1}) \)
Conditionally independent random variables

- **Sets** of variables $X, Y, Z$
- $X$ is independent of $Y$ given $Z$ if
  - $P \vdash (X \perp Y \mid Z) = z$, $\forall x \in \text{Val}(X), y \in \text{Val}(Y), z \in \text{Val}(Z)$

- **Shorthand:**
  - **Conditional independence:** $P \vdash (X \perp Y \mid Z)$
  - For $P \vdash (X \perp Y \mid \emptyset)$, write $P \vdash (X \perp Y)$

- **Notation:** $I(P)$ – independence properties entailed by $P$

- **Proposition:** $P$ satisfies $(X \perp Y \mid Z)$ if and only if
  - $P(X,Y|Z) = P(X|Z) P(Y|Z)$
Properties of independence

- **Symmetry:**
  - $(X \perp Y \mid Z) \Rightarrow (Y \perp X \mid Z)$

- **Decomposition:**
  - $(X \perp Y,W \mid Z) \Rightarrow (X \perp Y \mid Z)$

- **Weak union:**
  - $(X \perp Y,W \mid Z) \Rightarrow (X \perp Y \mid Z,W)$

- **Contraction:**
  - $(X \perp W \mid Y,Z) \& (X \perp Y \mid Z) \Rightarrow (X \perp Y,W \mid Z)$

- **Intersection:**
  - $(X \perp Y \mid W,Z) \& (X \perp W \mid Y,Z) \Rightarrow (X \perp Y,W \mid Z)$
  - *Only for positive distributions! $(P(\alpha)>0, \forall \alpha, \alpha\neq\emptyset)$*
Bayesian networks

- One of the most exciting advancements in statistical AI in the last 10-15 years
- Compact representation for exponentially-large probability distributions
- Fast marginalization too
- Exploit conditional independencies
Let’s start on BNs…

- Consider $P(X_i)$
  - Assign probability to each $x_i \in \text{Val}(X_i)$
  - Independent parameters

- Consider $P(X_1,\ldots,X_n)$
  - How many independent parameters if $|\text{Val}(X_i)|=k$?
What if variables are independent?

- \((X_i \perp X_j), \, \forall \, i, j\)
- Not enough!!! (See homework 1 😊)
- Must assume that \((X \perp Y), \, \forall \, X, Y \text{ subsets of } \{X_1, \ldots, X_n\}\)

- Can write
  \[P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i)\]

- How many independent parameters now?
Conditional parameterization –
two nodes

- Grade is determined by Intelligence
Conditional parameterization – three nodes

- Grade and SAT score are determined by Intelligence
- \((G \perp S \mid I)\)
The naïve Bayes model –
Your first real Bayes Net

- Class variable: C
- Evidence variables: $X_1, \ldots, X_n$
- Assume that $(X \perp Y \mid C), \forall X, Y$ subsets of \{X_1, \ldots, X_n\}
Causal structure

- Suppose we know the following:
  - The flu causes sinus inflammation
  - Allergies cause sinus inflammation
  - Sinus inflammation causes a runny nose
  - Sinus inflammation causes headaches

- How are these connected?
Possible queries

- Flu
- Allergy
- Headache
- Nose

- Inference
- Most probable explanation
- Active data collection
Car starts BN

- 18 binary attributes

Inference
- $P(\text{BatteryAge}|\text{Starts}=f)$

- $2^{18}$ terms, why so fast?
- Not impressed?
  - HailFinder BN – more than $3^{54} = 58149737003040059690390169$ terms
Factored joint distribution - Preview

Flu → Sinus → Headache
Allergy → Sinus → Nose
Number of parameters

Flu

Allergy

Sinus

Headache

Nose
Key: Independence assumptions

Knowing sinus separates the variables from each other
(Marginal) Independence

- Flu and Allergy are (marginally) independent

More Generally:

<table>
<thead>
<tr>
<th></th>
<th>Flu = t</th>
<th>Flu = f</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allergy = t</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Allergy = f</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Conditional independence

- Flu and Headache are not (marginally) independent

- Flu and Headache are independent given Sinus infection

- More Generally:
The independence assumption

Local Markov Assumption:
A variable $X$ is independent of its non-descendants given its parents
\[(X_i \perp \text{NonDescendants}_{X_i} \mid Pa_{X_i})\]
Explaining away

Local Markov Assumption: A variable $X_i$ is independent of its non-descendants given its parents $(X_i \perp \text{NonDescendants}_{X_i} | \text{Pa}_{X_i})$
What about probabilities?
Conditional probability tables (CPTs)

Flu
Allergy
Sinus
Headache
Nose
Joint distribution

Flu → Sinus → Headache
Allergy → Sinus → Nose

Why can we decompose? Markov Assumption!
A general Bayes net

- Set of random variables
- Directed acyclic graph
- CPTs

Joint distribution:

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

Local Markov Assumption:

- A variable \( X \) is independent of its non-descendants given its parents – \( (X_i \perp \text{NonDescendants}_{X_i} \mid \text{Pa}_{X_i}) \)
Questions???

- What distributions can be represented by a BN?
- What BNs can represent a distribution?
- What are the independence assumptions encoded in a BN?
  - in addition to the local Markov assumption
Today: The Representation Theorem – Joint Distribution to BN

BN: Encodes independence assumptions

If conditional independencies in BN are subset of conditional independencies in $P$

Joint probability distribution:

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid \text{Pa}_{X_i})$$
Today: The Representation Theorem – BN to Joint Distribution

**BN:**

Obtain

Encodes independence assumptions

If joint probability distribution:

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

Obtain

Then conditional independencies in BN are subset of conditional independencies in \( P \)
Let’s start proving it for naïve Bayes – From joint distribution to BN

- Independence assumptions:
  - $X_i$ independent given $C$

- Let’s assume that $P$ satisfies independencies must prove that $P$ factorizes according to BN:
  - $P(C, X_1, \ldots, X_n) = P(C) \prod_i P(X_i|C)$

- Use chain rule!
Let’s start proving it for naïve Bayes – From BN to joint distribution 1

- Let’s assume that $P$ factorizes according to the BN:
  $$P(C, X_1, \ldots, X_n) = P(C) \prod_i P(X_i | C)$$

- Prove the independence assumptions:
  - $X_i$ independent given $C$
  - Actually, $(X \perp Y | C)$, $\forall X, Y$ subsets of $\{X_1, \ldots, X_n\}$
Let’s start proving it for naïve Bayes – From BN to joint distribution 2

- Let’s consider a simpler case
  - Grade and SAT score are determined by Intelligence
  - \( P(I,G,S) = P(I)P(G|I)P(S|I) \)
  - Prove that \( P(G,S|I) = P(G|I) \ P(S|I) \)
Today: The Representation Theorem

If conditional independencies in BN are subset of conditional independencies in $P$

Obtain

Joint probability distribution:

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

If joint probability distribution:

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

Obtain

Then conditional independencies in BN are subset of conditional independencies in $P$
Local Markov assumption & I-maps

- Local independence assumptions in BN structure $G$:

- Independence assertions of $P$:

- BN structure $G$ is an \textit{l-map} (independence map) if:

\[
\text{Local Markov Assumption:} \quad \text{A variable } X \text{ is independent of its non-descendants given its parents} \]
\[
(X_i \perp \text{NonDescendants}_{X_i} \mid Pa_{X_i})
\]
Factorized distributions

- Given
  - Random vars $X_1, \ldots, X_n$
  - $P$ distribution over vars
  - BN structure $G$ over same vars
- $P$ factorizes according to $G$ if

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P \left( X_i \mid \text{Pa}_{X_i} \right)$$
BN Representation Theorem – I-map to factorization

If conditional independencies in BN are subset of conditional independencies in $P$

Obtain

Joint probability distribution:

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P\left(X_i \mid \text{Pa}_{X_i}\right)$$

$G$ is an I-map of $P$

$P$ factorizes according to $G$
BN Representation Theorem – I-map to factorization: Proof

Obtain

$G$ is an I-map of $P$

$P$ factorizes according to $G$

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

ALL YOU NEED:

Local Markov Assumption:
A variable $X$ is independent of its non-descendants given its parents

$$(X_i \perp \text{NonDescendants}_{X_i} \mid Pa_{X_i})$$
Defining a BN

- Given a set of variables and conditional independence assumptions
- Choose an ordering on variables, e.g., $X_1, \ldots, X_n$
- For $i = 1$ to $n$
  - Add $X_i$ to the network
  - Define parents of $X_i$, $\text{Pa}_{X_i}$, in graph as the minimal subset of $\{X_1, \ldots, X_{i-1}\}$ such that local Markov assumption holds – $X_i$ independent of rest of $\{X_1, \ldots, X_{i-1}\}$, given parents $\text{Pa}_{X_i}$
  - Define/learn CPT – $P(X_i| \text{Pa}_{X_i})$
BN Representation Theorem – Factorization to I-map

If joint probability distribution:

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

Then conditional independencies in BN are subset of conditional independencies in \( P \)

\( P \) factorizes according to \( G \)

\( G \) is an I-map of \( P \)
BN Representation Theorem – Factorization to I-map: **Proof**

If joint probability distribution:

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

*Proof :* Then conditional independencies in BN are subset of conditional independencies in \( P \)

\( P \) factorizes according to \( G \)

**G is an I-map of \( P \)**

**Homework 1!!!! 😊**
The BN Representation Theorem

If conditional independencies in BN are subset of conditional independencies in $P$

Joint probability distribution:

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

Important because:
Every $P$ has at least one BN structure $G$

If joint probability distribution:

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

Then conditional independencies in BN are subset of conditional independencies in $P$

Important because:
Read independencies of $P$ from BN structure $G$
Independencies encoded in BN

- We said: All you need is the local Markov assumption
  - \( (X_i \perp \text{NonDescendants}_{X_i} | \text{Pa}_{X_i}) \)
- But then we talked about other (in)dependencies
  - e.g., explaining away

- What are the independencies encoded by a BN?
  - Only assumption is local Markov
  - But many others can be derived using the algebra of conditional independencies!!!
Understanding independencies in BNs – BNs with 3 nodes

Local Markov Assumption: A variable X is independent of its non-descendants given its parents

Indirect causal effect:

Indirect evidential effect:

Common cause:

Common effect:
Understanding independencies in BNs – Some examples
Understanding independencies in BNs – Some more examples
An active trail – Example

When are A and H independent?
A path $X_1 - X_2 - \cdots - X_k$ is an **active trail** when variables $O \subseteq \{X_1, \ldots, X_n\}$ are observed if for each consecutive triplet in the trail:

- $X_{i-1} \rightarrow X_i \rightarrow X_{i+1}$, and $X_i$ is **not observed** ($X_i \not\in O$)
- $X_{i-1} \leftarrow X_i \leftarrow X_{i+1}$, and $X_i$ is **not observed** ($X_i \not\in O$)
- $X_{i-1} \leftarrow X_i \rightarrow X_{i+1}$, and $X_i$ is **not observed** ($X_i \not\in O$)
- $X_{i-1} \rightarrow X_i \leftarrow X_{i+1}$, and $X_i$ is **observed** ($X_i \in O$), or one of its descendents
Theorem: Variables $X_i$ and $X_j$ are independent given $Z \subseteq \{X_1, \ldots, X_n\}$ if there is no active trail between $X_i$ and $X_j$ when variables $Z \subseteq \{X_1, \ldots, X_n\}$ are observed.
More generally:

**Soundness of d-separation**

- Given BN structure $G$
- Set of independence assertions obtained by d-separation:
  - $I(G) = \{(X \perp Y | Z) : \text{d-sep}_G(X; Y | Z)\}$

**Theorem: Soundness of d-separation**
- If $P$ factorizes over $G$ then $I(G) \subseteq I(P)$

**Interpretation:** d-separation only captures true independencies

Proof discussed when we talk about undirected models
Existence of dependency when not d-separated

**Theorem:** If X and Y are not d-separated given Z, then X and Y are dependent given Z under some P that factorizes over G

**Proof sketch:**
- Choose an active trail between X and Y given Z
- Make this trail dependent
- Make all else uniform (independent) to avoid “canceling” out influence
More generally:
Completeness of d-separation

Theorem: Completeness of d-separation

- For “almost all” distributions that $P$ factorize over to $G$, we have that $I(G) = I(P)$
- “almost all” distributions: except for a set of measure zero of parameterizations of the CPTs (assuming no finite set of parameterizations has positive measure)

Proof sketch:
Interpretation of completeness

- **Theorem: Completeness of d-separation**
  - For “almost all” distributions that $P$ factorize over to $G$, we have that $I(G) = I(P)$

- BN graph is usually sufficient to capture all independence properties of the distribution!!!!

- But only for complete independence:
  - $P \models (X=x \perp Y=y \mid Z=z), \forall x \in \text{Val}(X), y \in \text{Val}(Y), z \in \text{Val}(Z)$

- Often we have context-specific independence (CSI)
  - $\exists x \in \text{Val}(X), y \in \text{Val}(Y), z \in \text{Val}(Z): P \models (X=x \perp Y=y \mid Z=z)$
  - Many factors may affect your grade
  - But if you are a frequentist, all other factors are irrelevant 😊
What you need to know

- Independence & conditional independence
- Definition of a BN
- The representation theorems
  - Statement
  - Interpretation
- d-separation and independence
  - soundness
  - existence
  - completeness
Acknowledgements

- JavaBayes applet