Bayesian Networks – Representation

Machine Learning – 10701/15781
Carlos Guestrin
Carnegie Mellon University
March 20th, 2006
Announcements

- Welcome back!
- One page project proposal due Wednesday
- We’ll go over midterm in this week’s recitation
Handwriting recognition

Character recognition, e.g., kernel SVMs
Webpage classification

Company home page vs Personal home page vs Univeristy home page vs ...

Our energy exploration, production, and distribution operations span the globe, with activities in more than 100 countries.

At TOTAL, we draw our greatest strength from our fast-growing oil and gas reserves. Our strategic emphasis on natural gas provides a strong position in a rapidly expanding market.

Our expanding refining and marketing operations in Asia and the Mediterranean Rim complement already solid positions in Europe, Africa, and the U.S.

Our growing specialty chemicals sector adds balance and profit to the core energy business.
Handwriting recognition 2
Webpage classification 2
Today – Bayesian networks

- One of the most exciting advancements in statistical AI in the last 10-15 years
- Generalizes naïve Bayes and logistic regression classifiers
- Compact representation for exponentially-large probability distributions
- Exploit conditional independencies
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
- Sinus
- Headache
- Nose

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

- 18 binary attributes
- Inference
  - $P(BatteryAge|Starts=f)$
- $2^{18}$ terms, why so fast?
- Not impressed?
  - HailFinder BN – more than $3^{54} = 58149737003040059690390169$ terms
Factored joint distribution - Preview

Flu -> Sinus
Allergy -> Sinus
Headache -> Sinus
Nose -> Sinus
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>

<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>
Marginal independence of random variables

- **Sets** of variables $X, Y$
- $X$ is independent of $Y$ if
  - $P \models (X=x | Y=y), \forall x \in \text{Val}(X), y \in \text{Val}(Y)$

- **Shorthand:**
  - **Marginal independence:** $P \models (X \perp Y)$

- **Proposition:** $P$ satisfies $(X \perp Y)$ if and only if
  - $P(X,Y) = P(X) \cdot P(Y)$
Conditional independence

- Flu and Headache are not (marginally) independent

- Flu and Headache are independent given Sinus infection

- More Generally:
**Conditionally independent random variables**

- **Sets** of variables $X$, $Y$, $Z$

  - $X$ is independent of $Y$ given $Z$ if
    - $P \models (X=x, Y=y | Z=z), \forall x \in \text{Val}(X), y \in \text{Val}(Y), z \in \text{Val}(Z)$

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

- **Proposition:** $P$ satisfies $(X \perp Y | Z)$ 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!
  - $P(\alpha)>0, \ \forall \alpha, \ \alpha \neq \emptyset$
The independence assumption

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

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

Local Markov Assumption:
A variable $X$ is independent of its non-descendants given its parents
What about probabilities?
Conditional probability tables (CPTs)

Flu
Allergy
Sinus
Headache
Nose
Joint distribution

Why can we decompose? Markov Assumption!
The chain rule of probabilities

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

More generally:
- \( P(X_1,\ldots,X_n) = P(X_1) \cdot P(X_2|X_1) \cdot \ldots \cdot P(X_n|X_1,\ldots,X_{n-1}) \)
Chain rule & Joint distribution

**Local Markov Assumption:**
A variable $X$ is independent of its non-descendants given its parents.
Two (trivial) special cases

- Edgeless graph
- Fully-connected graph
The Representation Theorem – Joint Distribution to BN

Joint probability distribution:

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

Obtain

Joint probability distribution:

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

- Diagnosis of lymph node disease
- Speech recognition
- Microsoft office and Windows
- Study Human genome
- Robot mapping
- Robots to identify meteorites to study
- Modeling fMRI data
- Anomaly detection
- Fault diagnosis
- Modeling sensor network data
A general Bayes net

- Set of random variables
- Directed acyclic graph
  - Encodes independence assumptions
- CPTs
- Joint distribution:

\[ P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid \text{Pa}_{X_i}) \]
How many parameters in a BN?

- Discrete variables $X_1, \ldots, X_n$
- Graph
  - Defines parents of $X_i$, $\text{Pa}_{X_i}$
- CPTs – $P(X_i | \text{Pa}_{X_i})$
Another example

- **Variables:**
  - B – Burglar
  - E – Earthquake
  - A – Burglar alarm
  - N – Neighbor calls
  - R – Radio report

- Both burglars and earthquakes can set off the alarm
- If the alarm sounds, a neighbor may call
- An earthquake may be announced on the radio
Another example – Building the BN

- B – Burglar
- E – Earthquake
- A – Burglar alarm
- N – Neighbor calls
- R – Radio report
Independencies encoded in BN

- We said: All you need is the local Markov assumption
  \[ (X_i \perp \text{NonDescendants}_{X_i} \mid \text{Pa}_{X_i}) \]
- 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:
\[ X \rightarrow Z \rightarrow Y \]

Indirect evidential effect:
\[ X \leftarrow Z \leftarrow Y \]

Common cause:
\[ X \leftarrow Z \rightarrow Y \]

Common effect:
\[ X \rightarrow Z \rightarrow Y \]
Understanding independencies in BNs – Some examples
An active trail – Example

When are A and H independent?
Active trails formalized

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
Active trails and independence?

**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.
The BN Representation Theorem

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

**Obtain**

**Joint probability distribution:**

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

**Important because:**

Every $P$ has at least one BN structure $G$

---

**If joint probability distribution:**

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

**Obtain**

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

**Important because:**

Read independencies of $P$ from BN structure $G$
“Simpler” BNs

- A distribution can be represented by many BNs:

- Simpler BN, requires fewer parameters
Learning Bayes nets

<table>
<thead>
<tr>
<th></th>
<th>Known structure</th>
<th>Unknown structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fully observable</td>
<td></td>
<td></td>
</tr>
<tr>
<td>data</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Missing data</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Data: $x^{(1)}, \ldots, x^{(m)}$

Structure:

CPTs – $P(X_i | Pa_{X_i})$

Parameters:
Learning the CPTs

For each discrete variable $X_i$

$$\text{MLE: } P(X_i = x_i \mid X_j = x_j) = \frac{\text{Count}(X_i = x_i, X_j = x_j)}{\text{Count}(X_j = x_j)}$$
Queries in Bayes nets

- Given BN, find:
  - Probability of X given some evidence, $P(X|e)$
  - Most probable explanation, $\max_{x_1,\ldots,x_n} P(x_1,\ldots,x_n | e)$
  - Most informative query

- Learn more about these next class
What you need to know

- Bayesian networks
  - A compact **representation** for large probability distributions
  - Not an algorithm
- Semantics of a BN
  - Conditional independence assumptions
- Representation
  - Variables
  - Graph
  - CPTs
- Why BNs are useful
- Learning CPTs from fully observable data
- Play with applet!!! 😊
Acknowledgements

- JavaBayes applet