Bayesian Networks – Representation (cont.)

Inference

Machine Learning – 10701/15781
Carlos Guestrin
Carnegie Mellon University

March 22st, 2006
Announcements

- One page project proposal due now

- We’ll go over midterm in this week’s recitation

- Homework 4 out later today, due April 5th
  - two weeks from today
Handwriting recognition

Character recognition, e.g., kernel SVMs
Handwriting recognition 2

- Context
- Examples not i.i.d.
+ Correlations between labels!!
Car starts BN

- 18 binary attributes
- Inference
  - $P(\text{BatteryAge}|\text{Starts}=f) \propto P(\text{BA}, S=f)$
  - Marginalization
    - $\sum P(\text{BA}, F \uparrow S=f) = P(\text{BA}, S=\epsilon)$
- $2^{18}$ terms, why so fast?
- Not impressed?
  - HailFinder BN – more than $3^{54} = 58149737003040059690390169$ terms
Factored joint distribution - Preview

\[ P(F, A, S, H, N) = P(F) \cdot P(A) \cdot P(S|F, A) \cdot P(H|S) \cdot P(N|S) \]

- **Flu**: \( P(F) = \frac{3}{4} \cdot \frac{1}{3} \)
- **Allergy**: \( P(A) \)
- **Sinus**: \( P(S|F, A) \)
- **Headache**: \( P(H|S) \)
- **Nose**: \( P(N|S) \)

- **Independence**: \( F \perp A \) and \( F \perp N|S \)

**Table**:

<table>
<thead>
<tr>
<th></th>
<th>t</th>
<th>f</th>
</tr>
</thead>
<tbody>
<tr>
<td>s</td>
<td>.8</td>
<td>.3</td>
</tr>
<tr>
<td>f</td>
<td>.2</td>
<td>.7</td>
</tr>
</tbody>
</table>
The independence assumption

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

($F \perp A$)
($N \perp \{F,A,H\} \mid S$)
Explaining away

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

\[
(F \perp A) \quad \text{marginally}
\]

\[
\text{what if } S = t
\]

\[
S = t \quad P(F = t \mid S = t) > P(F = t)
\]

but

\[
S = t \text{ and } A = t:
\]

\[
P(F = t \mid S = t) > P(F = t \mid S = t, A = t) > P(F = t)
\]

\[
F \text{ not independent given } S
\]
Chain rule & Joint distribution

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

\[ P(F, A, S, H, N) = P(F) P(A|F) P(S|FA) P(H|SFA) P(N|FAHS) \]

with local Markov Assumption:
\[ P(F) P(A) P(S|FA) P(H|S) P(N|S) \]
Two (trivial) special cases

Edgeless graph

\[ (X_1 \perp X_4) \]
\[ (X_2 \perp X_3 \mid X_5) \]
\[ \vdots \]

Fully-connected graph

\[ \text{only if all vars indep.} \]

\[ \text{give you some } P \]

\[ \text{always!} \]

\[ \text{no indep. in graph} \]
The Representation Theorem – Joint Distribution to BN

**BN:**

- **Obtain** Joint probability distribution:
  \[ P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid Pa_{X_i}) \]

- If conditional independencies in BN are subset of conditional independencies in \( P \)

- Encodes independence assumptions

\[ \forall P \text{ exists at least one BN} \]
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: $X_1, X_2, X_3, \ldots, X_n$
- Directed acyclic graph
  - Encodes independence assumptions
- CPTs
- Joint distribution:
  \[ P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid \text{Pa}_{X_i}) \]
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
  \[ (X_i \perp \text{NonDescendants}_{X_i} \mid Pa_{X_i}) \]

- 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:
\[ X \rightarrow Z \rightarrow Y \]

Common effect:
\[ X \leftarrow Z \leftarrow Y \]
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 ($X_i \notin O$)
  - $X_{i-1} \leftarrow X_i \leftarrow X_{i+1}$, and $X_i$ is not observed ($X_i \notin O$)
  - $X_{i-1} \leftarrow X_i \rightarrow X_{i+1}$, and $X_i$ is not observed ($X_i \notin O$)
  - $X_{i-1} \rightarrow X_i \leftarrow X_{i+1}$, 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 \( P \)

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

Obtain

Joint probability distribution:

Important because:
Every \( P \) has at least one BN structure \( G \)

If joint probability distribution:

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

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 data</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Missing data</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Data

\[
\begin{align*}
X^{(1)} & \\
\vdots & \\
X^{(m)} & \\
\end{align*}
\]

Structure

CPTs – \( P(X_i | Pa_{X_i}) \)

Parameters
Learning the CPTs

For each discrete variable $X_i$

MLE: $P(X_i = x_i \mid X_j = x_j) = \frac{\text{Count}(X_i = x_i, X_j = x_j)}{\text{Count}(X_j = x_j)}$
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!!! 😊
General probabilistic inference

- Query: \( P(X \mid e) \)

- Using Bayes rule:
  \[
P(X \mid e) = \frac{P(X, e)}{P(e)}
  \]

- Normalization:
  \[
P(X \mid e) \propto P(X, e)
  \]
Marginalization

Flu → Sinus

Allergy=t → Sinus
Inference seems exponential in number of variables!
Inference is NP-hard (Actually $\#P$-complete)

Reduction – 3-SAT

\[(\bar{X}_1 \lor X_2 \lor X_3) \land (\bar{X}_2 \lor X_3 \lor X_4) \land \ldots\]

Inference unlikely to be efficient in general, but…
Fast probabilistic inference example – Variable elimination

(Potential for) Exponential reduction in computation!
Understanding variable elimination – Exploiting distributivity

Flu → Sinus → Nose=t
Understanding variable elimination – Order can make a HUGE difference
Understanding variable elimination – Intermediate results

Intermediate results are probability distributions
Understanding variable elimination – Another example
Pruning irrelevant variables

Prune all non-ancestors of query variables
Variable elimination algorithm

- Given a BN and a query $P(X|e) \propto P(X,e)$
- Instantiate evidence $e$
- Prune non-ancestors of $\{X,e\}$
- Choose an ordering on variables, e.g., $X_1, \ldots, X_n$
- For $i = 1$ to $n$, if $X_i \notin \{X,e\}$
  - Collect factors $f_1,\ldots,f_k$ that include $X_i$
  - Generate a new factor by eliminating $X_i$ from these factors
  \[
g = \sum_{X_i} \prod_{j=1}^{k} f_j
\]
  - Variable $X_i$ has been eliminated!
- Normalize $P(X,e)$ to obtain $P(X|e)$

IMPORTANT!!!
Complexity of variable elimination – (Poly)-tree graphs

Variable elimination order:
Start from “leaves” up – find topological order, eliminate variables in reverse order

Linear in number of variables!!! (versus exponential)
Complexity of variable elimination –
Graphs with loops

Exponential in number of variables in largest factor generated
Complexity of variable elimination – Tree-width

Moralize graph:
Connect parents into a clique and remove edge directions

Complexity of VE elimination:
(“Only”) exponential in tree-width
Tree-width is maximum node cut +1
Example: Large tree-width with small number of parents

Compact representation ⇛ Easy inference ☹
Choosing an elimination order

- Choosing best order is NP-complete
  - Reduction from MAX-Clique
- Many good heuristics (some with guarantees)
- Ultimately, can’t beat NP-hardness of inference
  - Even optimal order can lead to exponential variable elimination computation
- In practice
  - Variable elimination often very effective
  - Many (many many) approximate inference approaches available when variable elimination too expensive
Most likely explanation (MLE)

- Query: \( \arg \max_{x_1, \ldots, x_n} P(x_1, \ldots, x_n \mid e) \)

- Using Bayes rule:
  \[
  \arg \max_{x_1, \ldots, x_n} P(x_1, \ldots, x_n \mid e) = \arg \max_{x_1, \ldots, x_n} \frac{P(x_1, \ldots, x_n, e)}{P(e)}
  \]

- Normalization irrelevant:
  \[
  \arg \max_{x_1, \ldots, x_n} P(x_1, \ldots, x_n \mid e) = \arg \max_{x_1, \ldots, x_n} P(x_1, \ldots, x_n, e)
  \]
Max-marginalization

Flu

Allergy=t

Sinus
Example of variable elimination for MLE – Forward pass

Flu

Allergy

Sinus

Headache

Nose=t
Example of variable elimination for MLE – Backward pass

- Flu
- Allergy
- Headache
- Sinus
- Nose=t
MLE Variable elimination algorithm – Forward pass

- Given a BN and a MLE query $\max_{x_1, \ldots, x_n} P(x_1, \ldots, x_n, e)$
- Instantiate evidence $e$
- Choose an ordering on variables, e.g., $X_1, \ldots, X_n$
- For $i = 1$ to $n$, If $X_i \notin \{e\}$
  - Collect factors $f_1, \ldots, f_k$ that include $X_i$
  - Generate a new factor by eliminating $X_i$ from these factors

$$g = \max_{x_i} \prod_{j=1}^{k} f_j$$

- Variable $X_i$ has been eliminated!
MLE Variable elimination algorithm – Backward pass

- \( \{x_1^*, \ldots, x_n^*\} \) will store maximizing assignment
- For \( i = n \) to 1, If \( X_i \notin \{e\} \)
  - Take factors \( f_1, \ldots, f_k \) used when \( X_i \) was eliminated
  - Instantiate \( f_1, \ldots, f_k \), with \( \{x_{i+1}^*, \ldots, x_n^*\} \)
    - Now each \( f_j \) depends only on \( X_i \)
  - Generate maximizing assignment for \( X_i \):
    \[
    x_i^* \in \arg\max_{x_i} \prod_{j=1}^{k} f_j
    \]
What you need to know

- Bayesian networks
  - A useful compact representation for large probability distributions
- Inference to compute
  - Probability of X given evidence e
  - Most likely explanation (MLE) given evidence e
  - Inference is NP-hard
- Variable elimination algorithm
  - Efficient algorithm (“only” exponential in tree-width, not number of variables)
  - Elimination order is important!
  - Approximate inference necessary when tree-width too large
    - not covered this semester
  - Only difference between probabilistic inference and MLE is “sum” versus “max”
Acknowledgements

- JavaBayes applet