Introduction to Bayesian Belief Nets

Russ Greiner
Dep’t of Computing Science
Alberta Ingenuity Centre for Machine Learning
University of Alberta

http://www.cs.ualberta.ca/~greiner/bn.html
The future of software may lie in the obscure theories of an 18th century cleric named Thomas Bayes.
Motivation

Gates says [LATimes, 28/Oct/96]:

Microsoft’s competitive advantages is its expertise in “Bayesian networks”

Current Products

- Microsoft Pregnancy and Child Care (MSN)
- Answer Wizard (Office, …)
- Print Troubleshooter
  - Excel Workbook Troubleshooter
  - Office 95 Setup Media Troubleshooter
  - Windows NT 4.0 Video Troubleshooter
  - Word Mail Merge Troubleshooter
Motivation (II)

US Army: **SAIP** (Battalion Detection from SAR, IR… GulfWar)

NASA: **Vista** (DSS for Space Shuttle)

GE: **Gems** (real-time monitor for utility generators)

Intel: (infer possible processing problems from end-of-line tests on semiconductor chips)

**KIC:**

- **Medical:** sleep disorders, pathology, trauma care, hand and wrist evaluations, dermatology, home-based health evaluations
- **DSS for capital equipment:** locomotives, gas-turbine engines, office equipment
Motivation (III)

- Lymph-node pathology diagnosis
- Manufacturing control
- Software diagnosis
- Information retrieval

*Types of tasks*
- Classification/Regression
- Sensor Fusion
- Prediction/Forecasting
Outline

- Existing uses of Belief Nets (BNs)
- What is a BN?
  - Specific Examples of BNs
  - Contrast with Rules, Neural Nets, ...
  - Possible applications of BNs
- Challenges
  - How to reason efficiently
  - How to learn BNs
Symptoms
Chief complaint
History, …

Signs
Physical Exam
Test results, …

Plan
Treatment, …

Diagnosis
Blah blah ouch yak ouch blah ouch blah blah ouch blah
Objectives: Decision Support System

- Determine
  - which *tests* to perform
  - which *repair* to suggest
  based on *costs*, *sensitivity/specificity*, ...

- Use all sources of information
  - *symbolic* (discrete observations, history, …)
  - *signal* (from sensors)

- Handle *partial* information

- *Adapt* to track fault distribution
Underlying Task

- **Situation**: Given observations \( \{O_1=\nu_1, \ldots, O_k=\nu_k\} \) (symptoms, history, test results, …) what is best DIAGNOSIS \( D_{x_i} \) for patient?

  - **Approach1**: Use set of \( \text{obs}_1 \& \ldots \& \text{obs}_m \rightarrow D_{x_i} \) rules

  but… *Need rule for each situation*

  - for each diagnosis \( D_{x_r} \)
  - for each set of possible values \( \nu_j \) for \( O_j \)
  - for each subset of obs. \( \{O_{x1}, O_{x2}, \ldots\} \subset \{O_j\} \)

    Can’t use

    | If Temp>100 & BP = High & Cough = Yes → DiseaseX |
    |-----------------------------------------------|
    | if only know Temp and BP                       |

- *Seldom Completely Certain*
**Underlying Task**

- **Situation:** Given observations \( \{O_1 = v_1, \ldots, O_k = v_k\} \) (symptoms, history, test results, …)
  what is best \( \text{DIAGNOSIS } D_{x_i} \) for patient?

- **Approach 2:** Compute Probabilities of \( D_{x_i} \) given observations \( \{ \text{obs}_j \} \)

  \[
  P( D_x = u \mid O_1 = v_1, \ldots, O_k = v_k )
  \]

- **Challenge:** How to express Probabilities?
How to deal with Probabilities

- **Sufficient: “atomic events”:**

  \[ P(D_x = u, O_1 = v_1, ..., O_k = v_k, ..., O_N = v_N) \]

  for all \( 2^{1+N} \) values \( u \in \{T, F\}, \ v_j \in \{T, F\} \)

  - \( P(D_x = T, O_1 = T, O_2 = T, ..., O_N = T) = 0.03 \)
  - \( P(D_x = T, O_1 = T, O_2 = T, ..., O_N = F) = 0.4 \)
  - \( \Rightarrow \)
  - \( P(D_x = T, O_1 = F, O_2 = F, ..., O_N = T) = 0 \)
  - ... ...
  - \( P(D_x = F, O_1 = F, O_2 = F, ..., O_N = F) = 0.01 \)

- **Then: Marginalize.**

  \[ P(D_x = u, O_1 = v_1, ... O_7 = v_7) = \sum_{v_8, ..., v_N} P(D_x = u, O_1 = v_1, ... O_7 = v_7, ... O_N = v_N) \]

- **Conditionalize:**

  \[ P(D_x = u \mid O_1 = v_1, ... O_7 = v_7) = \frac{P(D_x = u, O_1 = v_1, ... O_7 = v_7)}{P(O_1 = v_1, ... O_7 = v_7)} \]

  - But… even if binary \( D_x \), 20 binary obs.’s. \( \Rightarrow \) >2,097,000 numbers!
Problems with “Atomic Events”

- Representation *is not intuitive*
  - Should make “connections” explicit
  - Use “local information”
  
  \[ P(\text{Jaundice} \mid \text{Hepatitis}), \ P(\text{LightDim} \mid \text{BadBattery}), \ldots \]

- Too many numbers – \( O(2^N) \)
  - Hard to store
  - Hard to use
    - [Must add \( 2^r \) values to marginalize \( r \) variables]
  - Hard to learn
    - [Takes \( O(2^N) \) samples to learn \( 2^N \) parameters]
  
  \[ \Rightarrow \text{Include only necessary “connections”} \]

\[ \Rightarrow \text{Belief Nets} \]
? Hepatitis?

Jaunticed

? Hepatitis, not Jaunticed but +BloodTest

BloodTest
Encoding Causal Links

- **Simple Belief Net:**

  - **Table:**
    
    | h | P(B=1 | H=h) | P(B=0 | H=h) |
    |---|-----------|-----------|
    | 1 | 0.95      | 0.05      |
    | 0 | 0.03      | 0.97      |

- **Node ~ Variable**
- **Link ~ “Causal dependency”**
- **“CPTable” ~ P(child | parents)**

- **Table:**
  
  | h | b | P(J=1|h,b) | P(J=0|h,b) |
  |---|---|------------|------------|
  | 1 | 1 | 0.8        | 0.2        |
  | 1 | 0 | 0.8        | 0.2        |
  | 0 | 1 | 0.3        | 0.7        |
  | 0 | 0 | 0.3        | 0.7        |
### Encoding Causal Links

- $P(J \mid H, B=0) = P(J \mid H, B=1) \ \forall J, H$
  - $\Rightarrow P(J \mid H, B) = P(J \mid H)$
- $J$ is **INDEPENDENT** of $B$, once we know $H$
- Don’t need $B \rightarrow J$ arc!
Encoding Causal Links

- $P(J \mid H, B=0) = P(J \mid H, B=1)$ $\forall J, H$
  $\Rightarrow P(J \mid H, B) = P(J \mid H)$

- $J$ is INDEPENDENT of $B$, once we know $H$

- Don’t need $B \rightarrow J$ arc!
Encoding Causal Links

- \( P(J \mid H, B=0) = P(J \mid H, B=1) \quad \forall J, H \)
  \[ \implies P(J \mid H, B) = P(J \mid H) \]
- \( J \) is **INDEPENDENT** of \( B \), once we know \( H \)
- Don’t need \( B \rightarrow J \) arc!

<table>
<thead>
<tr>
<th>( h )</th>
<th>( P(B=1 \mid H=h) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.95</td>
</tr>
<tr>
<td>0</td>
<td>0.03</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( h )</th>
<th>( P(J=1 \mid h) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.8</td>
</tr>
<tr>
<td>0</td>
<td>0.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( P(H=1) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
</tr>
</tbody>
</table>
Sufficient Belief Net

\[ P(B=1 \mid H=h) \]

<table>
<thead>
<tr>
<th>h</th>
<th>P(B=1 \mid H=h)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.95</td>
</tr>
<tr>
<td>0</td>
<td>0.03</td>
</tr>
</tbody>
</table>

\[ P(H=1) \]

<table>
<thead>
<tr>
<th>h</th>
<th>P(H=1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.05</td>
</tr>
</tbody>
</table>

\[ P(J=1 \mid h) \]

<table>
<thead>
<tr>
<th>h</th>
<th>P(J=1 \mid h)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.8</td>
</tr>
<tr>
<td>0</td>
<td>0.3</td>
</tr>
</tbody>
</table>

- **Requires:**
  - \( P(H=1) \) known
  - \( P(J=1 \mid H=1) \) known
  - \( P(B=1 \mid H=1) \) known

(Only 5 parameters, not 7)

\[ P(H=1 \mid B=1, J=0) = \frac{1}{\alpha} \quad P(H=1) \quad P(B=1 \mid H=1) \quad P(J=0 \mid B=1, H=1) \]

Hence:

\[ P(H=1 \mid B=1, J=0) = \frac{1}{\alpha} \quad P(H=1) \quad P(B=1 \mid H=1) \quad P(J=0 \mid B=1, H=1) \]
“Factoring”

- **B does depend on J:**
  
  *If J=1, then likely that H=1  \( \Rightarrow \)  B = 1*

- **but... ONLY THROUGH H:**
  
  - If know H=1, then likely that B=1
  - ... doesn’t matter whether J=1 or J=0!

  \( \Rightarrow \)  \( P(J=0 \mid B=1, H=1) = P(J=0 \mid H=1) \)

N.b., **B and J ARE correlated a priori**  \( P(J \mid B ) \neq P(J) \)

**GIVEN H, they become uncorrelated**  \( P(J \mid B, H) = P(J \mid H) \)
Factored Distribution

- **Symptoms independent, given Disease**
  
<table>
<thead>
<tr>
<th>Symptom</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>Hepatitis</td>
</tr>
<tr>
<td>J</td>
<td>Jaundice</td>
</tr>
<tr>
<td>B</td>
<td>(positive) Blood test</td>
</tr>
</tbody>
</table>

  \[
  P(B | J) \neq P(B) \quad \text{but} \quad P(B | J, H) = P(B | H)
  \]

- **ReadingAbility and ShoeSize are dependent,**

  \[
  P(\text{ReadAbility} | \text{ShoeSize}) \neq P(\text{ReadAbility})
  \]

  but become independent, given Age

  \[
  P(\text{ReadAbility} | \text{ShoeSize}, \text{Age}) = P(\text{ReadAbility} | \text{Age})
  \]
"Naïve Bayes"

**Classification Task:**
Given \( \{ O_1 = v_1, \ldots, O_n = v_n \} \)
Find \( h_i \) that maximizes \( (H = h_i \mid O_1 = v_1, \ldots, O_n = v_n) \)

\[
P(H = h_i) \cdot \prod_{j} P(O_j = v_j \mid H = h_i)
\]

**Given**

- \( P(O_j = v_j \mid H = h_i) \)
- \( \text{Independent: } P(O_j \mid H, O_k, \ldots) = P(O_j \mid H) \)

\[
P(H = h_i \mid O_1 = v_1, \ldots, O_n = v_n) = \frac{1}{\alpha} P(H = h_i) \prod_{j} P(O_j = v_j \mid H = h_i)
\]

**Find** \( \text{argmax} \ \{ h_i \} \)
Naïve Bayes (con’t)

\[ P(H = h_i \mid O_1 = v_1, \ldots, O_n = v_n) = \frac{1}{\alpha} P(H = h_i) \prod_j P(O_j = v_j \mid H = h_i) \]

- Normalizing term
  \[ \alpha = P(O_1 = v_1, \ldots, O_n = v_n) = \sum_i P(H = h_i) \prod_j P(O_j = v_j \mid H = h_i) \]
  (No need to compute, as same for all \( h_i \))

- Easy to use for Classification
- Can use even if some \( v_j \)'s not specified

- If \( k \) \( Dx \)'s and \( n \) \( O_i \)'s,
  requires only \( k \) priors, \( n \times k \) pairwise-conditionals
  (Not \( 2^{n+k} \ldots \) relatively easy to learn)

<table>
<thead>
<tr>
<th>( n )</th>
<th>1+2n</th>
<th>( 2^{n+1} - 1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>21</td>
<td>2,047</td>
</tr>
<tr>
<td>30</td>
<td>61</td>
<td>2,147,438,647</td>
</tr>
</tbody>
</table>
Bigger Networks

Intuition: Show \textit{CAUSAL} connections:

\textbf{GeneticPH CAUSES Hepatitis}; \quad \textbf{Hepatitis CAUSES Jaundice}

If \textbf{GeneticPH}, then expect \textbf{Jaundice}:

GeneticPH $\Rightarrow$ Hepatitis $\Rightarrow$ Jaundice

But only via \textbf{Hepatitis}:

GeneticPH and not Hepatitis $\Rightarrow$ Jaundice

\[
P(J \mid G) \neq P(J) \quad \text{but} \quad P(J \mid G,H) = P(J \mid H)
\]
Belief Nets

- **DAG structure**
  - Each node $\equiv$ Variable $\nu$
  - $\nu$ depends (only) on its parents

+ conditional prob: $P(\nu_i \mid \text{parent}_i = \langle 0, 1, \ldots \rangle)$

- $\nu$ is INDEPENDENT of non-descendants, given assignments to its parents

Given $H = 1$,
- D has no influence on J
- J has no influence on B
- etc.
Less Trivial Situations

- *N.b.*, obs₁ is *not* always independent of obs₂ given H

- *Eg.*, FamilyHistoryDepression *‘causes’* MotherSuicide and Depression
  
  MotherSuicide *causes* Depression *(w/ or w/o F.H.Depression)*

  $P(D \mid MS, FHD) \neq P(D \mid FHD)$!

  Can be done using Belief Network,
  but need to specify:

  $P(FHD)$

  $P(MS \mid FHD)$

  $P(D \mid MS, FHD)$

  | f | P(MS=1 | FHD=f) |
  |---|----------------|
  | 1 | 0.10           |
  | 0 | 0.03           |

  | f | m | P(D=1 | FHD=f, MS=m) |
  |---|---|----------------|
  | 1 | 1 | 0.97           |
  | 1 | 0 | 0.90           |
  | 0 | 1 | 0.08           |
  | 0 | 0 | 0.04           |
A Logical Alarm Reduction Mechanism
• 8 diagnoses, 16 findings, …
Troup Detection

Diagram illustrating the relationship between various factors in troup detection, including terrain, unit type, vehicle classification, formation, sub-units, cluster, and vehicle detections.
ARCO1: Forecasting Oil Prices

Diagram showing the relationship between US Tax Policy, Price, OPEC Politics, Demand, Supply, World Growth, and Historical Values.
ARCO1: Forecasting Oil Prices
Forecasting Potato Production
Warning System
Extensions

- Find best values (posterior distr.) for SEVERAL (> 1) “output” variables
- Partial specification of “input” values
  - only subset of variables
  - only “distribution” of each input variable
- General Variables
  - Discrete, but domain > 2
  - Continuous (Gaussian: \( x = \sum_i b_i y_i \) for parents \( \{Y\} \))
- Decision Theory \( \Rightarrow \) Decision Nets (Influence Diagrams)
  Making Decisions, not just assigning prob’s
- Storing \( P(v | p_1, p_2, ..., p_k) \)
  General “CP Tables” 0(2^k)
  Noisy-Or, Noisy-And, Noisy-Max
  “Decision Trees”
Outline

- Existing **uses** of Belief Nets (BNs)
- What is a BN?
- Specific **Examples** of BNs
- **Contrast** with Rules, Neural Nets, …
- Possible **applications** of BNs
- **Challenges**
  - How to reason efficiently
  - How to *learn* BNs
Belief Nets vs Rules

- Both have “Locality”
  Specific clusters (rules / connected nodes)

- Often same nodes (rep’ning Propositions) but

  **BN:** Cause $\Rightarrow$ Effect
  “Hep $\Rightarrow$ Jaundice” $P(J | H)$

  **Rule:** Effect $\Rightarrow$ Cause
  “Jaundice $\Rightarrow$ Hep”

  **WHY?:** Easier for people to reason CAUSALLY
  even if use is DIAGNOSTIC

- BN provide OPTIMAL way to deal with
  + Uncertainty
  + Vagueness  (var not given, or only dist)
  + Error

  ...Signals meeting Symbols ...

- BN permits different “direction”’s of inference
Belief Nets vs Neural Nets

- Both have "graph structure" but

<table>
<thead>
<tr>
<th>BN:</th>
<th>Nodes have SEMANTICs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Combination Rules: Sound Probability</td>
</tr>
<tr>
<td>NN:</td>
<td>Nodes: arbitrary</td>
</tr>
<tr>
<td></td>
<td>Combination Rules: Arbitrary</td>
</tr>
</tbody>
</table>

- So harder to
  - Initialize NN
  - Explain NN
    (But perhaps easier to learn NN from examples only?)

- BNs can deal with
  - Partial Information
  - Different "direction"s of inference
Belief Nets vs Markov Nets

- Each uses “graph structure”
  to FACTOR a distribution
  …explicitly specify dependencies, implicitly independencies...

- but subtle differences...
  - BNs capture “causality”, “hierarchies”
  - MNs capture “temporality”

Technical: BNs use DIRECTED arcs
  ⇒ allow “induced dependencies”

  \[ I(A, \{\}, B) \] “A independent of B, given {}”
  \[ \neg I(A, C, B) \] “A dependent on B, given C”

MNs use UNDIRECTED arcs
  ⇒ allow other independencies

  \[ I(A, BC, D) \] A independent of D, given B, C
  \[ I(B, AD, C) \] B independent of C, given A, D
Uses of Belief Nets #1

Medical Diagnosis: “Assist/Critique” MD
- identify diseases not ruled-out
- specify additional tests to perform
- suggest treatments appropriate/cost-effective
- react to MD’s proposed treatment

Decision Support: Find/repair faults in complex machines
[Device, or Manufacturing Plant, or …]
… based on sensors, recorded info, history,…

Preventative Maintenance:
Anticipate problems in complex machines
[Device, or Manufacturing Plant, or …]
…based on sensors, statistics, recorded info, device history,…
Uses (con’t)

- **Logistics Support**: Stock warehouses appropriately ... based on (estimated) freq. of needs, costs,
- **Diagnose Software**: Find most probable bugs, given program behavior, core dump, source code, ...
- **Part Inspection/Classification**: ... based on multiple sensors, background, model of production,...
- **Information Retrieval**: Combine information from various sources, based on info from various “agents”,...

**General: Partial Info, Sensor fusion**
- Classification
- Prediction
- Interpretation
- ...

40
For given BN:

Given \( O_1 = v_1, \ldots, O_n = v_n \)

Compute \( P(H \mid O_1 = v_1, \ldots, O_n = v_n) \)

+ If BN is “poly tree”, \( \exists \) efficient alg.

- If BN is gen’l DAG (>1 path from \( X \) to \( Y \))
  - NP-hard in theory
  - slow in practice

**Tricks:** Get *approximate* answer (quickly)

+ Use abstraction of BN
+ Use “abstraction” of query (range)
**Why Reasoning is Hard**

- BN reasoning may look easy: Just “propagate” information from node to node

- Challenge: What is $P(C=t)$?

  $A = Z = \neg B$ \quad $P(A=t) = P(B=f) = \frac{1}{2}$

  So...? $P(C=t) = P(A=t, B=t)$

  $$= P(A=t) \times P(B=t) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

- Wrong: $P(C=t) = 0$!

Need to maintain dependencies! $P(A=t, B=t) = P(A=t) \times P(B=t|A=t)$

---

**Table: Conditional Probabilities**

| $A$ | $B$ | $P(C=t|a,b)$ |
|-----|-----|---------------|
| $t$ | $t$ | $1.0$         |
| $t$ | $f$ | $0.0$         |
| $f$ | $t$ | $0.0$         |
| $f$ | $f$ | $0.0$         |

**Diagram:**

- Nodes: $A$, $B$, $C$, $Z$
- Edges: $A \rightarrow C$, $B \rightarrow C$, $Z \rightarrow A$, $Z \rightarrow B$
- Probabilities:
  - $P(Z=t) = 0.5$
  - $P(A=t|Z=t) = 1.0$, $P(A=t|Z=f) = 0.0$
  - $P(B=t|Z=t) = 0.0$, $P(B=t|Z=f) = 1.0$
  - $P(C=t|A=t, B=t) = \frac{1}{4}$
# 2a: Obtaining Accurate BN

BN encodes distribution over \( n \) variables

\[
\text{Not } O(2^n) \text{ values, but “only” } \sum_i 2^{k_i} \\
(\text{Node } n_i \text{ binary, with } k_i \text{ parents})
\]

Still lots of values! …structure ..

⇒ **Qualitative Information**

Structure: “What depends on what?”

- Easy for people (background knowledge)
- But NP-hard to learn from samples...

Knowledge acquisition: from human experts

⇒ **Quantitative Information**

Actual CP-tables

- Easy to learn, given lots of examples.
- But people have hard time…

Simple learning algorithm
Notes on Learning

- **Mixed Sources**: Person provides structure; Algorithm fills-in numbers.

- **Just Learning Algorithm**: $\exists$ algorithms that learn $\left\{ \text{structure values} \right\}$ from sample

- **Just Human Expert**: People produce CP-table, as well as structure
  Relatively few values really required
  Esp. if NoisyOr, NoisyAnd, NaiveBayes, ...

  Actual values not *that* important
  …Sensitivity studies
The world changes. Information in $BN^*$ may be
- perfect at time $t$
- sub-optimal at time $t + 20$
- worthless at time $t + 200$

Need to *MAINTAIN* a BN over time using *on-going* human consultant

Adaptive BN
- Dirichlet distribution (variables)
- Priors over BNs
My Results Related to Belief Nets

- Quantifying Uncertainty in BN Response
  - $\Pr_{\Theta}(C=\text{true} \mid D=\text{false}) = 0.3\pm 0.05$
  - Uses: Good Decision, Bad Outcome Bias$^2+$Variance; Mixture using Variance

- Learning Structure – Generatively
  - BDe, 2-foldCV work well (not MDL)

- Learning Structure – Discriminatively
  - Bias$^2+$Variance works well (not MDL)

- Learning Parameters – Discriminately
  - NaïveBayes : Logistic Regression :: Belief Nets : ELR
Conclusions

- **Belief Nets are PROVEN TECHNOLOGY**
  - Medical Diagnosis
  - DSS for complex machines
  - Forecasting, Modeling, InfoRetrieval…

- **Provide effective way to**
  - Represent complicated, inter-related events
  - Reason about such situations
    - Diagnosis, Explanation, ValueOfInfo
    - Explain conclusions
    - Mix Symbolic and Numeric observations

- **Challenges**
  - Efficient ways to use BNs
  - How to create accurate/effective BNs
  - How to maintain BNs
  - Reason about time…