Bayesian belief networks

Milos Hauskrecht
milos@cs.pitt.edu
5329 Sennott Square

Modeling the uncertainty.

Key challenges:

• How to represent the relations in the presence of uncertainty?
• How to manipulate such knowledge to make inferences?
  – **Humans can reason with uncertainty.**
Modeling uncertainty with probabilities

Probabilistic extension of propositional logic

- **Propositions:**
  - statements about the world
  - Represented by the assignment of values to random variables

- **Random variables:**
  - **Boolean**
    - *Pneumonia* is either *True, False*
  - **Multi-valued**
    - *Pain* is one of \{*Nopain, Mild, Moderate, Severe*\}
  - **Continuous**
    - *HeartRate* is a value in \(<0, 180>\)

---

## Probability distribution

Defines probability for all possible value assignments

**Example 1:**

\[
P(\text{Pneumonia} = \text{True}) = 0.001 \\
P(\text{Pneumonia} = \text{False}) = 0.999
\]

\[
P(\text{Pneumonia} = \text{True}) + P(\text{Pneumonia} = \text{False}) = 1
\]

Probabilities sum to 1 !!!

<table>
<thead>
<tr>
<th>Pneumonia</th>
<th>(P(\text{Pneumonia}))</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td>0.001</td>
</tr>
<tr>
<td>False</td>
<td>0.999</td>
</tr>
</tbody>
</table>

**Example 2:**

\[
P(\text{WBCcount} = \text{high}) = 0.005 \\
P(\text{WBCcount} = \text{normal}) = 0.993 \\
P(\text{WBCcount} = \text{high}) = 0.002
\]

<table>
<thead>
<tr>
<th>WBCcount</th>
<th>(P(\text{WBCcount}))</th>
</tr>
</thead>
<tbody>
<tr>
<td>high</td>
<td>0.005</td>
</tr>
<tr>
<td>normal</td>
<td>0.993</td>
</tr>
<tr>
<td>low</td>
<td>0.002</td>
</tr>
</tbody>
</table>
Joint probability distribution

Joint probability distribution (for a set variables)
• Defines probabilities for all possible assignments of values to variables in the set

Example: Assume variables:

\( Pneumonia \) (2 values)
\( WBC\text{count} \) (3 values)
\( Pain \) (4 values)

\( P(pneumonia, WBC\text{count}, Pain) \) is represented by \( 2 \times 3 \times 4 \) array

Joint probabilities

Marginalization
• reduces the dimension of the joint distribution
• Sums variables out

\( P(pneumonia, WBC\text{count}) \) \( 2 \times 3 \) matrix

\[ \begin{array}{c|ccc}
\text{Pneumonia} & \text{high} & \text{normal} & \text{low} \\
\hline
\text{True} & 0.0008 & 0.0001 & 0.0001 \\
\text{False} & 0.0042 & 0.9929 & 0.0019 \\
\end{array} \]

\( P(\text{Pneumonia}) \)

\( 0.001 \)
\( 0.999 \)

\( P(\text{WBCcount}) \)

\( 0.005 \)
\( 0.993 \)
\( 0.002 \)

Marginalization (here summing of columns or rows)
Marginalization

- reduces the dimension of the joint distribution

\[ P(X_1, X_2, \ldots, X_{n-1}) = \sum_{X_n} P(X_1, X_2, \ldots, X_{n-1}, X_n) \]

- We can continue doing this

\[ P(X_2, \ldots, X_{n-1}) = \sum_{X_1, X_n} P(X_1, X_2, \ldots, X_{n-1}, X_n) \]

What is the maximal joint probability distribution?
- Full joint probability

---

Full joint distribution

The joint distribution for all variables in the problem
- It defines the complete probability model for the problem

Example: pneumonia diagnosis

- Variables: Pneumonia, Fever, Paleness, WBCcount, Cough

Full joint defines the probability for all possible assignments of values to these variables

\[ P(\text{Pneumonia} = T, \text{WBCcount} = \text{High}, \text{Fever} = T, \text{Cough} = T, \text{Paleness} = T) \]
\[ P(\text{Pneumonia} = T, \text{WBCcount} = \text{High}, \text{Fever} = T, \text{Cough} = T, \text{Paleness} = F) \]
\[ P(\text{Pneumonia} = T, \text{WBCcount} = \text{High}, \text{Fever} = T, \text{Cough} = F, \text{Paleness} = T) \]
\[ \ldots \text{ etc} \]
- How many probabilities are there?
**Full joint distribution**

The joint distribution for all variables in the problem
- It defines the complete probability model for the problem

**Example:** pneumonia diagnosis

**Variables:** Pneumonia, Fever, Paleness, WBCcount, Cough

Full joint defines the probability for all possible assignments of values to these variables

\[ P(\text{Pneumonia} = T, \text{WBCcount} = \text{High}, \text{Fever} = T, \text{Cough} = T, \text{Paleness} = T) \]
\[ P(\text{Pneumonia} = T, \text{WBCcount} = \text{High}, \text{Fever} = T, \text{Cough} = T, \text{Paleness} = F) \]
\[ P(\text{Pneumonia} = T, \text{WBCcount} = \text{High}, \text{Fever} = T, \text{Cough} = F, \text{Paleness} = T) \]
\[ \ldots \text{ etc} \]

- **How many probabilities are there?**

---

**Joint probabilities and independence**

- Is it possible to recover the full joint from the joint probabilities over a subset of variables?
- Only if the variables are independent !!!

\[ P(\text{pneumonia, WBCcount}) \] \(2 \times 3\) matrix

\[ P(\text{Pneumonia}) \]

\[ P(\text{WBCcount}) \]

\[ \begin{array}{c|ccc}
\text{Pneumonia} & \text{high} & \text{normal} & \text{low} \\
\hline
\text{True} & ? & ? & ? \\
\text{False} & ? & ? & ? \\
\hline
0.005 & 0.993 & 0.002
\end{array} \]

\[ 0.001 \]

\[ 0.999 \]
Conditional probabilities

• **Conditional probability distribution.**
  \[ P(A \mid B) = \frac{P(A, B)}{P(B)} \text{ s.t. } P(B) \neq 0 \]

• **Product rule.** A join probability can be expressed in terms of conditional probabilities
  \[ P(A, B) = P(A \mid B)P(B) \]

• **Chain rule.** Any joint probability can be expressed as a product of conditionals
  \[
  P(X_1, X_2, \ldots, X_n) = P(X_n \mid X_1, \ldots, X_{n-1})P(X_1, \ldots, X_{n-1}) \\
  = P(X_n \mid X_1, \ldots, X_{n-1})P(X_{n-1} \mid X_1, \ldots, X_{n-2})P(X_1, \ldots, X_{n-2}) \\
  = \prod_{i=1}^{n} P(X_i \mid X_1, \ldots, X_{i-1})
  \]

---

**Conditional probability**

• Is defined in terms of the joint probability:
  \[ P(A \mid B) = \frac{P(A, B)}{P(B)} \text{ s.t. } P(B) \neq 0 \]

• **Example:**
  \[
  P(\text{pneumonia} = \text{true} \mid \text{WBC count} = \text{high}) = \frac{P(\text{pneumonia} = \text{true}, \text{WBC count} = \text{high})}{P(\text{WBC count} = \text{high})} \\
  P(\text{pneumonia} = \text{false} \mid \text{WBC count} = \text{high}) = \frac{P(\text{pneumonia} = \text{false}, \text{WBC count} = \text{high})}{P(\text{WBC count} = \text{high})}
  \]
Conditional probabilities

Conditional probability distribution
- Defines probabilities for all possible assignments, given a fixed assignment to some other variable values

\[ P(\text{Pneumonia} = \text{true} \mid \text{WBCcount} = \text{high}) \]

\[ \mathbf{P}(\text{Pneumonia} \mid \text{WBCcount}) \]

3 element vector of 2 elements

<table>
<thead>
<tr>
<th>WBCcount</th>
<th>high</th>
<th>normal</th>
<th>low</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td>0.08</td>
<td>0.0001</td>
<td>0.0001</td>
</tr>
<tr>
<td>False</td>
<td>0.92</td>
<td>0.9999</td>
<td>0.9999</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
</tbody>
</table>

\[ P(\text{Pneumonia} = \text{true} \mid \text{WBCcount} = \text{high}) \]
\[ + P(\text{Pneumonia} = \text{false} \mid \text{WBCcount} = \text{high}) \]

Bayes rule

Bayes rule:

\[ P(A \mid B) = \frac{P(A, B)}{P(B)} \]

\[ P(A, B) = P(B \mid A)P(A) \]

When is it useful?
- When we are interested in computing the diagnostic query from the causal probability
  \[ P(\text{cause} \mid \text{effect}) = \frac{P(\text{effect} \mid \text{cause})P(\text{cause})}{P(\text{effect})} \]
- **Reason:** It is often easier to assess causal probability
  - E.g. Probability of pneumonia causing fever vs. probability of pneumonia given fever
Probabilistic inference

Various inference tasks:

- **Diagnostic task.** (from effect to cause)
  \[ P(\text{Pneumonia} \mid \text{Fever} = T) \]

- **Prediction task.** (from cause to effect)
  \[ P(\text{Fever} \mid \text{Pneumonia} = T) \]

- **Other probabilistic queries** (queries on joint distributions).
  \[ P(\text{Fever}) \]
  \[ P(\text{Fever}, \text{ChestPain}) \]

---

Inference

Any query can be computed from the full joint distribution !!!

- **Joint over a subset of variables** is obtained through marginalization
  \[ P(A = a, C = c) = \sum_i \sum_j P(A = a, B = b_j, C = c, D = d_j) \]

- **Conditional probability over set of variables**, given other variables’ values is obtained through marginalization and definition of conditionals
  \[ P(D = d \mid A = a, C = c) = \frac{P(A = a, C = c, D = d)}{P(A = a, C = c)} \]
  \[ = \frac{\sum_i P(A = a, B = b_i, C = c, D = d)}{\sum_i \sum_j P(A = a, B = b_j, C = c, D = d_j)} \]
Modeling uncertainty with probabilities

• Defining the **full joint distribution** makes it possible to represent and reason with uncertainty in a uniform way
• We are able to handle an arbitrary probabilistic inference problem

**Problems:**

– **Space complexity.** To store a full joint distribution we need to remember $O(d^n)$ numbers.
  
  $n$ – number of random variables, $d$ – number of values

– **Inference (time) complexity.** To compute some queries requires $O(d^n)$ steps.

– **Acquisition problem.** Who is going to define all of the probability entries?

---

Modeling uncertainty with probabilities

• **Knowledge based system era (70s – early 80’s)**
  – **Extensional non-probabilistic models**
  – Solve the space, time and acquisition bottlenecks in probability-based models
  – Froze the development and advancement of KB systems and contributed to the slow-down of AI in 80s in general

• Breakthrough (late 80s, beginning of 90s)
  – **Bayesian belief networks**
    • Give solutions to the space, acquisition bottlenecks
    • Partial solutions for time complexities
  • Bayesian belief network
Bayesian belief networks (BBNs)

Bayesian belief networks.
- Represent the full joint distribution over the variables more compactly with a **smaller number of parameters**.
- Take advantage of **conditional and marginal independences** among random variables

- **A and B are independent**
  \[ P(A, B) = P(A)P(B) \]

- **A and B are conditionally independent given C**
  \[ P(A, B | C) = P(A | C)P(B | C) \]
  \[ P(A | C, B) = P(A | C) \]

---

Alarm system example

- Assume your house has an **alarm system** against **burglary**. You live in the seismically active area and the alarm system can get occasionally set off by an **earthquake**. You have two neighbors, **Mary** and **John**, who do not know each other. If they hear the alarm they call you, but this is not guaranteed.
- We want to represent the probability distribution of events:
  - Burglary, Earthquake, Alarm, Mary calls and John calls

Causal relations

![Diagram of causal relations showing Burglary, Earthquake, Alarm, JohnCalls, and MaryCalls]
Bayesian belief network

1. Directed acyclic graph
   - **Nodes** = random variables
     Burglary, Earthquake, Alarm, Mary calls and John calls
   - **Links** = direct (causal) dependencies between variables.
     The chance of Alarm is influenced by Earthquake, The chance of John calling is affected by the Alarm

2. Local conditional distributions
   - relate variables and their parents
Bayesian belief networks (general)

Two components: $B = (S, \Theta_S)$

- **Directed acyclic graph**
  - Nodes correspond to random variables
  - (Missing) links encode independences

- **Parameters**
  - Local conditional probability distributions
    for every variable-parent configuration
    $$P(X_i \mid pa(X_i))$$

Where:

$pa(X_i)$ - stand for parents of $X_i$
Full joint distribution in BBNs

**Full joint distribution** is defined in terms of local conditional distributions (obtained via the chain rule):

\[
P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | \text{pa}(X_i))
\]

**Example:**
Assume the following assignment of values to random variables
\[B = T, E = T, A = T, J = T, M = F\]

Then its probability is:
\[
\]

Bayesian belief networks (BBNs)

**Bayesian belief networks**
- Represent the full joint distribution over the variables more compactly using the product of local conditionals.
- **But how did we get to local parameterization?**

**Answer:**
- **Chain rule** +
- **Graphical structure** encodes conditional and marginal independences among random variables:
  - **A and B are independent** \( P(A, B) = P(A)P(B) \)
  - **A and B are conditionally independent given C**
    \[
P(A | C, B) = P(A | C) \quad P(A, B | C) = P(A | C)P(B | C)
    \]
  - **The graph structure implies the decomposition !!!**
Independences in BBNs

3 basic independence structures:

1. JohnCalls is independent of Burglary given Alarm

\[
P(J \mid A, B) = P(J \mid A) \\
P(J, B \mid A) = P(J \mid A)P(B \mid A)
\]
Independences in BBNs

1. Burglary

2. Burglary is independent of Earthquake (not knowing Alarm)
Burglary and Earthquake become dependent given Alarm !!

\[ P(B, E) = P(B)P(E) \]

3. MaryCalls is independent of JohnCalls given Alarm

\[ P(J \mid A, M) = P(J \mid A) \]
\[ P(J, M \mid A) = P(J \mid A)P(M \mid A) \]
Independences in BBN

- BBN distribution models many conditional independence relations among distant variables and sets of variables
- These are defined in terms of the graphical criterion called d-separation
- **D-separation and independence**
  - Let X, Y and Z be three sets of nodes
  - If X and Y are d-separated by Z, then X and Y are conditionally independent given Z
- **D-separation**:
  - A is d-separated from B given C if every undirected path between them is blocked with C
- **Path blocking**
  - 3 cases that expand on three basic independence structures

Undirected path blocking

A is d-separated from B given C if every undirected path between them is **blocked**

![Diagram](image-url)
Undirected path blocking

A is d-separated from B given C if every undirected path between them is \textit{blocked}

\begin{itemize}
  \item 1. Path blocking with a linear substructure
  \end{itemize}
Undirected path blocking

A is d-separated from B given C if every undirected path between them is \textit{blocked}

- \textbf{2. Path blocking with the wedge substructure}

\begin{center}
\begin{tikzpicture}
  \node (X) at (0,0) [circle,draw] {X in A};
  \node (Y) at (3,0) [circle,draw] {Y in B};
  \node (Z) at (1.5,1.5) [circle,fill,red] {Z};
  \draw (X) -- (Z);
  \draw (Z) -- (Y);
\end{tikzpicture}
\end{center}

- \textbf{3. Path blocking with the vee substructure}

\begin{center}
\begin{tikzpicture}
  \node (X) at (0,0) [circle,draw] {X in A};
  \node (Y) at (3,0) [circle,draw] {Y in B};
  \node (Z) at (1.5,1.5) [circle,fill,red] {Z or any of its descendants \textit{not} in C};
  \draw (X) -- (Z);
  \draw (Z) -- (Y);
\end{tikzpicture}
\end{center}
• Earthquake and Burglary are independent given MaryCalls

• Burglary and MaryCalls are independent (not knowing Alarm)
Independences in BBNs

- Earthquake and Burglary are independent given MaryCalls \( F \)
- Burglary and MaryCalls are independent (not knowing Alarm) \( F \)
- Burglary and RadioReport are independent given Earthquake \( ? \)
Independences in BBNs

- Earthquake and Burglary are independent given MaryCalls
- Burglary and MaryCalls are independent (not knowing Alarm)
- Burglary and RadioReport are independent given Earthquake
- Burglary and RadioReport are independent given MaryCalls

Bayesian belief networks (BBNs)

Bayesian belief networks
- Represents the full joint distribution over the variables more compactly using the product of local conditionals.
- So how did we get to local parameterizations?

\[ P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid pa(X_i)) \]

- The decomposition is implied by the set of independences encoded in the belief network.
Full joint distribution in BBNs

Rewrite the full joint probability using the product rule:

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

(Insert diagram here)

M. Hauskrecht
Full joint distribution in BBNs

Rewrite the full joint probability using the product rule:

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

Product rule

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

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

M. Hauskrecht
Full joint distribution in BBNs

Rewrite the full joint probability using the product rule:

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

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

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

\[ \quad \frac{P(M = F \mid B = T, E = T, A = T)}{P(M = F \mid A = T)}P(B = T, E = T, A = T) \]

\[ \quad \frac{P(\text{rest of the joint distribution})}{P(A = T \mid B = T, E = T)}P(B = T, E = T) \]

M. Hauskrecht
Full joint distribution in BBNs

Rewrite the full joint probability using the product rule:

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

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

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

\[ \quad P(M = F \mid B = T, E = T, A = T)P(B = T, E = T, A = T) \]

\[ \quad P(M = F \mid A = T)P(B = T, E = T, A = T) \]

\[ \quad P(A = T \mid B = T, E = T)P(B = T, E = T) \]

\[ P(B = T)P(E = T) \]

M. Hauskrecht
Parameter complexity problem

- In the BBN the **full joint distribution** is defined as:
  \[ P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid pa(X_i)) \]
- What did we save?

Alarm example: binary (True, False) variables

# of parameters of the full joint:

\[ ? \]

M. Hauskrecht

---

Parameter complexity problem

- In the BBN the **full joint distribution** is defined as:
  \[ P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid pa(X_i)) \]
- What did we save?

Alarm example: binary (True, False) variables

# of parameters of the full joint:

\[ 2^5 = 32 \]

One parameter is for free:

\[ 2^5 - 1 = 31 \]

# of parameters of the BBN:

\[ ? \]

M. Hauskrecht
### Bayesian belief network: parameters count

#### Parameters of the full joint:

- **Burglary (B)**: 2 parameters
- **Earthquake (E)**: 2 parameters
- **Alarm (A)**: 8 parameters
- **JohnCalls (J)**: 4 parameters
- **MaryCalls (M)**: 4 parameters

Total: **20**

#### Parameter complexity problem

- In the BBN the **full joint distribution** is defined as:
  \[ P(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} P(X_i \mid pa(X_i)) \]

- What did we save?

  **Alarm example:** 5 binary (True, False) variables

  # of parameters of the full joint:
  \[ 2^5 = 32 \]
  
  One parameter is for free:
  \[ 2^5 - 1 = 31 \]

  # of parameters of the BBN:
  \[ 2^3 + 2(2^2) + 2(2) = 20 \]

  One parameter in every conditional is for free:
Bayesian belief network: free parameters

<table>
<thead>
<tr>
<th></th>
<th>B</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>0.95</td>
<td>0.05</td>
</tr>
<tr>
<td>F</td>
<td>0.94</td>
<td>0.06</td>
</tr>
<tr>
<td>T</td>
<td>0.29</td>
<td>0.71</td>
</tr>
<tr>
<td>F</td>
<td>0.001</td>
<td>0.999</td>
</tr>
</tbody>
</table>

\[ P(B) = 1 \]
\[ P(E) = 1 \]

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>0.90</td>
<td>0.1</td>
</tr>
<tr>
<td>F</td>
<td>0.05</td>
<td>0.95</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>0.07</td>
<td>0.3</td>
</tr>
<tr>
<td>F</td>
<td>0.01</td>
<td>0.99</td>
</tr>
</tbody>
</table>

Parameter complexity problem

- In the BBN the full joint distribution is defined as:
  \[ P(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} P(X_i | pa(X_i)) \]
- What did we save?
  Alarm example: 5 binary (True, False) variables

  # of parameters of the full joint:
  \[ 2^5 = 32 \]
  One parameter is for free:
  \[ 2^5 - 1 = 31 \]

  # of parameters of the BBN:
  \[ 2^3 + 2(2^2) + 2(2) = 20 \]

  One parameter in every conditional is for free:
  \[ 2^2 + 2(2) + 2(1) = 10 \]