Bayesian belief networks

Probabilistic inference

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

- **Joint over a subset of variables** is obtained through marginalization

\[ P(A = a, C = c) = \sum_i \sum_j P(A = a, B = b_i, C = c, D = d_j) \]

- **Conditional probability over set of variables**, given other variables’ values is obtained through marginalization and definition of conditionals

\[
\begin{align*}
P(D = d \mid A = a, C = c) &= \frac{P(A = a, C = c, D = d)}{\sum_i P(A = a, B = b_i, C = c, D = d_i)} \\
&= \sum_i \sum_j P(A = a, B = b_i, C = c, D = d_j)
\end{align*}
\]
### Probabilistic inference

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

- Any joint probability can be expressed as a product of conditionals via the **chain rule**.

\[
P(X_1, X_2, \ldots, X_n) = P(X_n | X_1, \ldots, X_{n-1})P(X_1, \ldots, X_{n-1})
= P(X_n | X_1, \ldots, X_{n-1})P(X_{n-1} | X_1, \ldots, X_{n-2})P(X_1, \ldots, X_{n-2})
= \prod_{i=1}^{n} P(X_i | X_1, \ldots, X_{i-1})
\]

### 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 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?
Medical diagnosis example

- **Space complexity.**
  - Pneumonia (2 values: T,F), Fever (2: T,F), Cough (2: T,F), WBC count (3: high, normal, low), paleness (2: T,F)
  - Number of assignments: 2*2*2*3*2=48
  - We need to define at least 47 probabilities.

- **Time complexity.**
  - Assume we need to compute the marginal of Pneumonia=T from the full joint
  
  \[
  P(\text{Pneumonia}=T) = \sum_{i \in \{T,F\}} \sum_{j \in \{T,F\}} \sum_{k \in \{h,n,l\}} \sum_{u \in \{T,F\}} P(\text{Fever}=i, \text{Cough}=j, \text{WBC count}=k, \text{Pale}=u)
  \]
  - Sum over: 2*2*3*2=24 combinations

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

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

![Causal relations diagram](attachment:image.png)
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):

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

**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 parameterizations?

**Answer:**
- **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

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

1. Burglary

2. Alarm

3. JohnCalls

2. MaryCalls is independent of JohnCalls given Alarm

\[ P(J | A, M) = P(J | A) \]

\[ P(J, M | A) = P(J | A)P(M | A) \]

3. Burglary is independent of Earthquake (not knowing Alarm)

Burglary and Earthquake become dependent given Alarm !!

\[ P(B, E) = P(B)P(E) \]
Indepences 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**
Undirected path blocking

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

1. Path blocking with a linear substructure

- X in A
- Z in C
- Y in B
Undirected path blocking

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

1. Path blocking with the wedge substructure

2.  Path blocking with the wedge substructure

\[
\begin{align*}
X \quad Z \quad Y \\
X \text{ in } A & \quad Z \text{ in } C & \quad Y \text{ in } B
\end{align*}
\]

3. Path blocking with the V-structure (explain away)

\[
\begin{align*}
X \quad Z \quad Y \\
X \quad Z \text{ or any of its descendants not in } C & \quad Y
\end{align*}
\]
Independences in BBNs

• Earthquake and Burglary are independent given MaryCalls

• Burglary and MaryCalls are independent (not knowing Alarm)

• 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 \( \text{F} \)
- Burglary and MaryCalls are independent (not knowing Alarm) \( \text{F} \)
- Burglary and RadioReport are independent given Earthquake \( ? \)
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: \( T \)
- Burglary and RadioReport are independent given MaryCalls: \( F \)

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:

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

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

\[ = P(J|A=T)P(B=T, E=T, A=T, M=F) \]
Rewrite the full joint probability using the product rule:

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

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

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

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

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

\[ = P(A = T \mid B = T, E = T)P(B = T, E = T) \]
Full joint distribution in BBNs

Rewrite the full joint probability using the product rule:

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

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

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

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

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

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

\[
P(B=T)P(E=T)
\]
Parameter complexity problem

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

\[ 2^5 - 1 = 31 \]
Bayesian belief network.

- In the BBN the **full joint distribution** is expressed using a set of local conditional distributions

<table>
<thead>
<tr>
<th>B</th>
<th>E</th>
<th>P(B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>T</td>
<td>0.95</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>0.94</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>0.29</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>0.001</td>
</tr>
</tbody>
</table>

| A | P(A|B,E) |
|---|---------|
| T | 0.90    |
| F | 0.05    |

| J | P(J|A) |
|---|------|
| T | 0.9  |
| F | 0.05 |

| M | P(M|A) |
|---|-------|
| T | 0.7   |
| F | 0.01  |
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: 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:
  \[ ? \]
Model acquisition problem

The structure of the BBN
- typically reflects causal relations
  (BBNs are also sometime referred to as causal networks)
- Causal structure is intuitive in many applications domain and it is relatively easy to define to the domain expert

Probability parameters of BBN
- are conditional distributions relating random variables and their parents
- Complexity is much smaller than the full joint
- It is much easier to obtain such probabilities from the expert or learn them automatically from data