Bayesian belief networks

Data:

Objective: try to estimate the underlying true probability distribution over variables $X$, $p(X)$, using examples in $D$

true distribution $p(X)$  \hspace{1cm} n samples $D = \{D_1, D_2, ..., D_n\}$  \hspace{1cm} estimate $\hat{p}(X)$

Standard (iid) assumptions: Samples

- are independent of each other
- come from the same (identical) distribution (fixed $p(X)$)

Density estimation

Data: $D = \{D_1, D_2, ..., D_n\}$
$D_i = x_i$ a vector of attribute values

Objective: try to estimate the underlying true probability distribution over variables $X$, $p(X)$, using examples in $D$
Modeling complex distributions

**Question:** How to model and learn complex multivariate distributions $\hat{p}(X)$ with a large number of variables?

**Example: modeling of disease – symptoms relations**

- **Disease:** pneumonia
- **Patient symptoms (findings, lab tests):**
  - Fever, Cough, Paleness, WBC (white blood cells) count, Chest pain, etc.
- **Model of the full joint distribution:**
  
  $P(\text{Pneumonia}, \text{Fever}, \text{Cough}, \text{Paleness}, \text{WBC}, \text{Chest pain})$

One probability per assignment of values to variables:

$P(\text{Pneumonia} = T, \text{Fever} = T, \text{Cough} = T, \text{WBC} = \text{High}, \text{Chest pain} = T)$

- **How many probabilities are there?**

---

**Marginalization**

**Joint probability distribution (for a set variables)**

- Defines probabilities for all possible assignments to values of variables in the set

$P(\text{pneumonia, WBC count})$  

$\begin{array}{|c|c|c|}
\hline
\text{WBC count} & \text{high} & \text{normal} & \text{low} \\
\hline
\text{Pneumonia} & 0.0008 & 0.001 & 0.001 & 0.001 \\
\text{False} & 0.0042 & 0.9929 & 0.0019 & 0.993 \\
\hline
\end{array}$

- $P(\text{Pneumonia}) = 0.001$  

- $P(\text{WBC count})$

  **Marginalization** (summing of rows, or columns)

- Summing out variables
Full joint distribution

- Any joint probability over a subset of variables can be obtained via marginalization from the full joint

\[ P(\text{Pneumonia}, \text{WBC count}, \text{Fever}) = \sum_{c, p = \{T, F\}} P(\text{Pneumonia}, \text{WBC count}, \text{Fever}, \text{Cough} = c, \text{Paleness} = p) \]

- **Question:** Is it possible to recover the full joint from the joint probabilities over a subset of variables?

---

Joint probabilities

- **Is it possible to recover the full joint from the joint probabilities over a subset of variables?**

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

<table>
<thead>
<tr>
<th>Pneumonia</th>
<th>high</th>
<th>normal</th>
<th>low</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>False</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
</tbody>
</table>

\[ P(\text{WBC count}) \]

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

\[ P(\text{WBC count}) \]

\[ \begin{bmatrix} 0.005 & 0.993 & 0.002 \end{bmatrix} \]
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 !!!

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

2×3 matrix

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

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

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

0.001

0.999

0.005

0.993

0.002

Variable independence

• The two events A, B are said to be independent if:
  \[ P(A, B) = P(A)P(B) \]

• The variables X, Y are said to be independent if their joint can be expressed as a product of marginals:
  \[ P(X, Y) = P(X)P(Y) \]
Conditional probability

Conditional probability:

- **Probability of A given B**
  \[ P(A \mid B) = \frac{P(A, B)}{P(B)} \]

- Conditional probability is defined in terms of joint probabilities
- Joint probabilities can be expressed in terms of conditional probabilities
  \[ P(A, B) = P(A \mid B)P(B) \] **(product rule)**
  \[ P(X_1, X_2, \ldots X_n) = \prod_{i=1}^{n} P(X_i \mid X_1, \ldots X_{i-1}) \] **(chain rule)**

- Conditional probability – is useful for **various probabilistic inferences**
  \[ P(\text{Pneumonia} = \text{True} \mid \text{Fever} = \text{True}, \text{WBC count} = \text{high}, \text{Cough} = \text{True}) \]

---

Conditional probabilities

*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 of values to target variables, given a fixed assignment of other variable values

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

\[
P(Pneumonia \mid WBC\text{count})
\]

3 element vector of 2 elements

\[
Pneumonia
\]

<table>
<thead>
<tr>
<th>WBC\text{count}</th>
<th>True</th>
<th>False</th>
</tr>
</thead>
<tbody>
<tr>
<td>high</td>
<td>0.08</td>
<td>0.92</td>
</tr>
<tr>
<td>normal</td>
<td>0.0001</td>
<td>0.9999</td>
</tr>
<tr>
<td>low</td>
<td>0.0001</td>
<td>0.9999</td>
</tr>
</tbody>
</table>

Variable we condition on

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

**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 a set of variables**, given other variables’ values is obtained through marginalization and definition of conditionals

\[
P(D = d \mid A = a, C = c) = \frac{P(A = a, C = c, D = d)}{P(A = a, C = c)}
\]

\[
= \frac{\sum_i P(A = a, B = b_i, C = c, D = d)}{\sum_i \sum_j P(A = a, B = b_i, C = c, D = d_j)}
\]
Inference

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

\[ P(X_1, X_2, \ldots, X_n) = P(X_n \mid X_{n-1}, \ldots, X_1)P(X_{n-1} \mid X_{n-2}, \ldots, X_1) \]

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

Why this may be important?

• It is often easier to define the distribution in terms of conditional probabilities:
  – E.g. \( P(Fever \mid Pneumonia = T) \)
  – \( P(Fever \mid Pneumonia = F) \)

Probabilistic inference

Various probabilistic inference tasks:

• Diagnostic task. (from effect to cause)
  \( P(Pneumonia \mid Fever = T) \)

• Prediction task. (from cause to effect)
  \( P(Fever \mid Pneumonia = T) \)

• Other probabilistic queries (queries on joint distributions).
  \( P(Fever) \)
  \( P(Fever, ChestPain) \)
Modeling complex distributions

- Defining the **full joint distribution** makes it possible to represent and reason with the probabilities.
- 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.** How to acquire/learn all these probabilities?

Pneumonia 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: $2 \times 2 \times 2 \times 3 \times 2 = 48$
  - We need to define at least 47 probabilities.

- **Time complexity.**
  - Assume we need to compute the marginal of Pneumonia=T from the full joint:

$$P(\text{Pneumonia} = T) = \sum_{i \in \{T,F\}} \sum_{j \in \{T,F\}} \sum_{k = h,n,l} \sum_{u \in \{T,F\}} P(\text{Fever} = i, \text{Cough} = j, \text{WBC count} = k, \text{Pale} = u)$$
  - Sum over: $2 \times 2 \times 3 \times 2 = 24$ combinations
Bayesian belief networks (BBNs)

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

**Key features:**
- 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
- **X and Y are independent** \( P(X, Y) = P(X)P(Y) \)
- **X and Y are conditionally independent given Z**
  \[
  P(X, Y | Z) = P(X | Z)P(Y | Z) \\
  P(X | Y, Z) = P(X | Z)
  \]

---

**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 diagram](image-url)
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 being is influenced by Earthquake,
     The chance of John calling is affected by the Alarm

2. **Local conditional distributions**
   - relating variables and their parents
Bayesian belief network

<table>
<thead>
<tr>
<th></th>
<th>T</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>P(B)</td>
<td>0.001</td>
<td>0.999</td>
</tr>
<tr>
<td>P(E)</td>
<td>0.002</td>
<td>0.998</td>
</tr>
</tbody>
</table>

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

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

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>T</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>P(M</td>
<td>A)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>0.7</td>
<td>0.3</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>0.01</td>
<td>0.99</td>
<td></td>
</tr>
</tbody>
</table>

Full joint distribution in BBNs

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

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

**Example:**

Assume the following assignment of values to random variables:

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

Then its probability is:

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

\[
P(B = T)P(E = T)P(A = T \mid B = T, E = T)P(J = T \mid A = T)P(M = F \mid A = T)
\]
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
  \[
  P(A | C, B) = P(A | C) \\
  P(A, B | C) = P(A | C)P(B | C)
  \]
- The graph structure implies the decomposition !!!

Independences in BBNs

3 basic independence structures:

1. \( \text{Burglary} \rightarrow \text{Alarm} \rightarrow \text{JohnCalls} \)
2. \( \text{Burglary} \rightarrow \text{Alarm} \rightarrow \text{Earthquake} \)
3. \( \text{Alarm} \rightarrow \text{JohnCalls} \rightarrow \text{MaryCalls} \)
1. JohnCalls is independent of Burglary given Alarm

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

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

\[ P(B, E) = P(B)P(E) \]
Independences in BBNs

1. JohnCalls
2. JohnCalls
3. MaryCalls is independent of JohnCalls given Alarm

$$P(J \mid A, M) = P(J \mid A)$$
$$P(J, M \mid A) = P(J \mid A)P(M \mid A)$$

Independence in BBN

- BBN distribution models many conditional independence relations relating distant variables and sets
- These are defined in terms of the graphical criterion called d-separation
- **D-separation in the graph**
  - 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 showing undirected path blocking](image-url)
Undirected path blocking

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

• 1. Path blocking with a linear substructure

\begin{center}
\begin{tikzpicture}
  \node (A) at (0,0) [shape=circle,fill=blue!20] {A};
  \node (C) at (2,0) [shape=circle,fill=blue!20] {C};
  \node (B) at (4,0) [shape=circle,fill=blue!20] {B};
  \node (X) at (1,-1) [shape=circle,fill=blue!20] {X};
  \node (Y) at (3,-1) [shape=circle,fill=blue!20] {Y};
  \node (Z) at (2,-1) [shape=circle,fill=red!50] {Z};
  \draw (A) -- (C);
  \draw (C) -- (B);
  \draw (X) -- (Z) -- (Y);
\end{tikzpicture}
\end{center}

X in A \quad Z in C \quad Y in B

• 2. Path blocking with the wedge substructure

\begin{center}
\begin{tikzpicture}
  \node (X) at (0,0) [shape=circle,fill=blue!20] {X};
  \node (Y) at (2,0) [shape=circle,fill=blue!20] {Y};
  \node (Z) at (1,-1) [shape=circle,fill=red!50] {Z};
  \node (ZinC) at (1,-2) [shape=circle,fill=blue!20] {Z in C};
  \draw (X) -- (Z);
  \draw (Z) -- (ZinC);
  \draw (ZinC) -- (Y);
\end{tikzpicture}
\end{center}

X in A \quad Y in B
Undirected path blocking

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

- **3. Path blocking with the vee substructure**

  X in A \hspace{1cm} Y in B
  \[
  \begin{array}{c}
  \circ \quad \circ \quad \circ \quad \circ \quad \circ \\
  \end{array}
  \]

  Z or any of its descendants **not** in C

---

Independences in BBNs

- Earthquake and Burglary are independent given MaryCalls

---
Independences in BBNs

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

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 \( F \)
Full joint distribution in BBNs

Rewrite the full joint probability using the product rule:

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

Full joint distribution in BBNs

Rewrite the full joint probability using the product rule:

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

\[ \text{Product rule} \]

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

Rewrite the full joint probability using the product rule:

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

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

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

Full joint distribution in BBNs

Rewrite the full joint probability using the product rule:

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

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

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

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

Rewrite the full joint probability using the product rule:

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

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

Full joint distribution in BBNs

Rewrite the full joint probability using the product rule:

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

\[ = P(J = T \mid B = T, E = T, A = T, M = F)P(B = T, E = T, A = T, M = F) \]
\[ = P(J = T \mid A = T)P(B = T, E = T, A = T, M = F) \]
\[ \quad \frac{P(M = F \mid B = T, E = T, A = T)}{P(M = F \mid A = T)}P(B = T, E = T, A = T) \]
\[ \quad \frac{P(A = T \mid B = T, E = T)}{P(A = T \mid E = T)}P(B = T, E = T) \]
Full joint distribution in BBNs

Rewrite the full joint probability using the product rule:

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

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

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

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

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

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

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

• In the BBN the **full joint distribution** is defined as:

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

• What did we save?

Alarm example: binary (True, False) variables

# of parameters of the full joint:

? 

---

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:

? 

---
Bayesian belief network: parameters count

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

  \[
  \begin{align*}
  \text{# of parameters of the full joint:} \\
  2^5 &= 32 \\
  \text{One parameter is for free:} \\
  2^5 - 1 &= 31 \\
  \text{# of parameters of the BBN:} \\
  2^3 + 2(2^2) + 2(2) &= 20 \\
  \text{One parameter in every conditional is for free:} \\
  &? 
  \end{align*}
  \]
Bayesian belief network: free parameters

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:
  \[ 2^2 + 2(2) + 2(1) = 10 \]