Lecture 7: Probabilistic Graphical Models

Vlad Niculae & André Martins

Deep Structured Learning Course, Fall 2019
In this unit, we will formalize & extend these graphical representations encountered in previous lectures.
Directed Models

Bayes networks

Conditional independence and D-separation

Causal graphs & the $do$ operator

Undirected Models

Markov networks

Factor graphs
Directed Models

Bayes networks
  Conditional independence and D-separation
  Causal graphs & the do operator

Undirected Models

Markov networks
  Factor graphs
Bayes (belief) networks

- Common task: Characterize how some related events co-occur. Specifically, in terms of probabilities!
- A car alarm is going off. Was there a break-in?
Bayes (belief) networks

- Common task: Characterize how some related events co-occur. Specifically, in terms of probabilities!

- A car alarm is going off. Was there a break-in?

<table>
<thead>
<tr>
<th>P(B)</th>
<th>B=yes</th>
<th>B=no</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>.05</td>
<td>.95</td>
</tr>
</tbody>
</table>

| P(A | B)    | A=on | A=off |
|--------|------|-------|
| B=yes  | .99  | .01   |
| B=no   | .10  | .90   |

- P(B | A) =?
Bayes (belief) networks

- Common task: Characterize how some related events co-occur. Specifically, in terms of probabilities!
- A car alarm is going off. Was there a break-in?

<table>
<thead>
<tr>
<th>P(B)</th>
<th>B=yes</th>
<th>B=no</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>.05</td>
<td>.95</td>
</tr>
</tbody>
</table>

| P(A | B, W) | A=on | A=off |
|---------|------|-------|
| B=yes W=lo | .99  | .01   |
| B=yes W=med | .99  | .01   |
| B=yes W=hi  | .999 | .001  |
| B=no W=lo   | .01  | .99   |
| B=no W=med  | .05  | .95   |
| B=no W=hi   | .25  | .75   |

- \( P(B | A) =? \)
- Can we observe wind? \( P(B | A, W) =? \)
Bayes (belief) networks

- Common task: Characterize how some related events co-occur. Specifically, in terms of probabilities!
- A car alarm is going off. Was there a break-in?

<table>
<thead>
<tr>
<th>P(B)</th>
<th>B=yes</th>
<th>B=no</th>
</tr>
</thead>
<tbody>
<tr>
<td>.05</td>
<td>.95</td>
<td></td>
</tr>
</tbody>
</table>

| P(A | B, W) | A=on | A=off |
|---------|------|------|
| B=yes   | W=lo | .99  | .01  |
| B=yes   | W=med| .99  | .01  |
| B=yes   | W=hi | .999 | .001 |
| B=no    | W=lo | .01  | .99  |
| B=no    | W=med| .05  | .95  |
| B=no    | W=hi | .25  | .75  |

- \( P(B \mid A) =? \)
- Can we observe wind? \( P(B \mid A, W) =? \)
  - Maybe we’re in the basement, but have a barometer.
Bayes networks

Toolkit for encoding knowledge about interaction structures between random variables.

Directed acyclic graph (DAG). Nodes = variables. Arrows = statistical dependencies.

In general: \[ P(X_1, \ldots, X_n) = \prod_i P(X_i | \text{parents}(X_i)) \]

For example: \[ P(\text{Break-in}, \text{Wind}, \text{Alarm}, \text{Barometer}) \]
\[ = P(\text{Break-in}) P(\text{Wind}) P(\text{Alarm} | \text{Break-in}, \text{Wind}) P(\text{Barometer} | \text{Wind}) \]
Without any structure, \( P(\text{Break-in}, \text{Wind}, \text{Alarm}, \text{Barometer}) \) would have to be stored & estimated like

<table>
<thead>
<tr>
<th>Brk.</th>
<th>Wind</th>
<th>Alarm</th>
<th>Bar.</th>
<th>( P )</th>
</tr>
</thead>
<tbody>
<tr>
<td>yes</td>
<td>lo</td>
<td>on</td>
<td>lo</td>
<td>0.0243</td>
</tr>
<tr>
<td>yes</td>
<td>lo</td>
<td>on</td>
<td>med</td>
<td>0.0002</td>
</tr>
<tr>
<td>yes</td>
<td>lo</td>
<td>on</td>
<td>hi</td>
<td>0.0002</td>
</tr>
<tr>
<td>yes</td>
<td>lo</td>
<td>off</td>
<td>lo</td>
<td>0.0002</td>
</tr>
<tr>
<td>yes</td>
<td>lo</td>
<td>off</td>
<td>med</td>
<td>2.50e-06</td>
</tr>
<tr>
<td>yes</td>
<td>lo</td>
<td>off</td>
<td>hi</td>
<td>2.50e-06</td>
</tr>
<tr>
<td>yes</td>
<td>med</td>
<td>on</td>
<td>lo</td>
<td>0.0001</td>
</tr>
<tr>
<td>yes</td>
<td>med</td>
<td>on</td>
<td>med</td>
<td>0.0146</td>
</tr>
<tr>
<td>yes</td>
<td>med</td>
<td>on</td>
<td>hi</td>
<td>0.0001</td>
</tr>
<tr>
<td>yes</td>
<td>med</td>
<td>off</td>
<td>lo</td>
<td>1.50e-06</td>
</tr>
<tr>
<td>yes</td>
<td>med</td>
<td>off</td>
<td>med</td>
<td>0.0001</td>
</tr>
<tr>
<td>yes</td>
<td>med</td>
<td>off</td>
<td>hi</td>
<td>1.50e-06</td>
</tr>
<tr>
<td>yes</td>
<td>hi</td>
<td>on</td>
<td>lo</td>
<td>9.99e-05</td>
</tr>
<tr>
<td>yes</td>
<td>hi</td>
<td>on</td>
<td>med</td>
<td>9.99e-05</td>
</tr>
<tr>
<td>yes</td>
<td>hi</td>
<td>on</td>
<td>hi</td>
<td>0.0098</td>
</tr>
<tr>
<td>yes</td>
<td>hi</td>
<td>off</td>
<td>lo</td>
<td>1.00e-07</td>
</tr>
<tr>
<td>yes</td>
<td>hi</td>
<td>off</td>
<td>med</td>
<td>1.00e-07</td>
</tr>
<tr>
<td>yes</td>
<td>hi</td>
<td>off</td>
<td>hi</td>
<td>9.80e-06</td>
</tr>
</tbody>
</table>

\[
P(Break-in=yes, \text{Alarm}=on) = \sum_b P(Break-in=b, \text{Alarm}=on)
\]

<table>
<thead>
<tr>
<th>Brk.</th>
<th>Wind</th>
<th>Alarm</th>
<th>Bar.</th>
<th>( P )</th>
</tr>
</thead>
<tbody>
<tr>
<td>no</td>
<td>lo</td>
<td>on</td>
<td>lo</td>
<td>0.0047</td>
</tr>
<tr>
<td>no</td>
<td>lo</td>
<td>on</td>
<td>med</td>
<td>4.75e-05</td>
</tr>
<tr>
<td>no</td>
<td>lo</td>
<td>on</td>
<td>hi</td>
<td>4.75e-05</td>
</tr>
<tr>
<td>no</td>
<td>lo</td>
<td>off</td>
<td>lo</td>
<td>0.4608</td>
</tr>
<tr>
<td>no</td>
<td>lo</td>
<td>off</td>
<td>med</td>
<td>0.0047</td>
</tr>
<tr>
<td>no</td>
<td>lo</td>
<td>off</td>
<td>hi</td>
<td>0.0047</td>
</tr>
<tr>
<td>no</td>
<td>med</td>
<td>on</td>
<td>lo</td>
<td>0.0001</td>
</tr>
<tr>
<td>no</td>
<td>med</td>
<td>on</td>
<td>med</td>
<td>0.0140</td>
</tr>
<tr>
<td>no</td>
<td>med</td>
<td>on</td>
<td>hi</td>
<td>0.0001</td>
</tr>
<tr>
<td>no</td>
<td>med</td>
<td>off</td>
<td>lo</td>
<td>0.0027</td>
</tr>
<tr>
<td>no</td>
<td>med</td>
<td>off</td>
<td>med</td>
<td>0.2653</td>
</tr>
<tr>
<td>no</td>
<td>med</td>
<td>off</td>
<td>hi</td>
<td>0.0027</td>
</tr>
<tr>
<td>no</td>
<td>hi</td>
<td>on</td>
<td>lo</td>
<td>0.0005</td>
</tr>
<tr>
<td>no</td>
<td>hi</td>
<td>on</td>
<td>med</td>
<td>0.0005</td>
</tr>
<tr>
<td>no</td>
<td>hi</td>
<td>on</td>
<td>hi</td>
<td>0.0466</td>
</tr>
<tr>
<td>no</td>
<td>hi</td>
<td>off</td>
<td>lo</td>
<td>0.0014</td>
</tr>
<tr>
<td>no</td>
<td>hi</td>
<td>off</td>
<td>med</td>
<td>0.0014</td>
</tr>
<tr>
<td>no</td>
<td>hi</td>
<td>off</td>
<td>hi</td>
<td>0.1397</td>
</tr>
</tbody>
</table>
Without any structure, $P(\text{Break-in, Wind, Alarm, Barometer})$ would have to be stored & estimated like:

$$P(\text{Break-in}=\text{yes}, \text{Alarm}=\text{on}) = 0.0496$$
Without any structure, \( P(\text{Break-in, Wind, Alarm, Barometer}) \) would have to be stored & estimated like

\[
\begin{align*}
\text{Brk.} & & \text{Wind} & & \text{Alarm} & & \text{Bar.} & & P \\
\text{yes} & & \text{lo} & & \text{on} & & \text{lo} & & 0.0243 \\
\text{yes} & & \text{lo} & & \text{on} & & \text{med} & & 0.0002 \\
\text{yes} & & \text{lo} & & \text{on} & & \text{hi} & & 0.0002 \\
\text{yes} & & \text{lo} & & \text{off} & & \text{lo} & & 0.0002 \\
\text{yes} & & \text{lo} & & \text{off} & & \text{med} & & 2.50e-06 \\
\text{yes} & & \text{lo} & & \text{off} & & \text{hi} & & 2.50e-06 \\
\text{yes} & & \text{med} & & \text{on} & & \text{lo} & & 0.0001 \\
\text{yes} & & \text{med} & & \text{on} & & \text{med} & & 0.0146 \\
\text{yes} & & \text{med} & & \text{on} & & \text{hi} & & 0.0001 \\
\text{yes} & & \text{med} & & \text{off} & & \text{lo} & & 1.50e-06 \\
\text{yes} & & \text{med} & & \text{off} & & \text{med} & & 0.0001 \\
\text{yes} & & \text{med} & & \text{off} & & \text{hi} & & 1.50e-06 \\
\text{yes} & & \text{hi} & & \text{on} & & \text{lo} & & 9.99e-05 \\
\text{yes} & & \text{hi} & & \text{on} & & \text{med} & & 9.99e-05 \\
\text{yes} & & \text{hi} & & \text{on} & & \text{hi} & & 0.0098 \\
\text{yes} & & \text{hi} & & \text{off} & & \text{lo} & & 1.00e-07 \\
\text{yes} & & \text{hi} & & \text{off} & & \text{med} & & 1.00e-07 \\
\text{yes} & & \text{hi} & & \text{off} & & \text{hi} & & 9.80e-06 \\
\end{align*}
\]

\[
\begin{align*}
\text{no} & & \text{lo} & & \text{on} & & \text{lo} & & 0.0047 \\
\text{no} & & \text{lo} & & \text{on} & & \text{med} & & 4.75e-05 \\
\text{no} & & \text{lo} & & \text{on} & & \text{hi} & & 4.75e-05 \\
\text{no} & & \text{lo} & & \text{off} & & \text{lo} & & 0.4608 \\
\text{no} & & \text{lo} & & \text{off} & & \text{med} & & 0.0047 \\
\text{no} & & \text{lo} & & \text{off} & & \text{hi} & & 0.0047 \\
\text{no} & & \text{med} & & \text{on} & & \text{lo} & & 0.0001 \\
\text{no} & & \text{med} & & \text{on} & & \text{med} & & 0.0140 \\
\text{no} & & \text{med} & & \text{on} & & \text{hi} & & 0.0001 \\
\text{no} & & \text{med} & & \text{off} & & \text{lo} & & 0.0027 \\
\text{no} & & \text{med} & & \text{off} & & \text{med} & & 0.2653 \\
\text{no} & & \text{med} & & \text{off} & & \text{hi} & & 0.0027 \\
\text{no} & & \text{hi} & & \text{on} & & \text{lo} & & 0.0005 \\
\text{no} & & \text{hi} & & \text{on} & & \text{med} & & 0.0005 \\
\text{no} & & \text{hi} & & \text{on} & & \text{hi} & & 0.0466 \\
\text{no} & & \text{hi} & & \text{off} & & \text{lo} & & 0.0014 \\
\text{no} & & \text{hi} & & \text{off} & & \text{med} & & 0.0014 \\
\text{no} & & \text{hi} & & \text{off} & & \text{hi} & & 0.1397 \\
\end{align*}
\]

\[
P(\text{Break-in}=\text{yes}, \text{Alarm}=\text{on}) = 0.0496
\]

\[
P(\text{Break-in}=\text{no}, \text{Alarm}=\text{on}) = 0.0665
\]
Without any structure, \( P(\text{Break-in, Wind, Alarm, Barometer}) \) would have to be stored & estimated like

<table>
<thead>
<tr>
<th>Brk.</th>
<th>Wind</th>
<th>Alarm</th>
<th>Bar.</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>yes</td>
<td>lo</td>
<td>on</td>
<td>lo</td>
<td>0.0243</td>
</tr>
<tr>
<td>yes</td>
<td>lo</td>
<td>on</td>
<td>med</td>
<td>0.0002</td>
</tr>
<tr>
<td>yes</td>
<td>lo</td>
<td>on</td>
<td>hi</td>
<td>0.0002</td>
</tr>
<tr>
<td>yes</td>
<td>lo</td>
<td>off</td>
<td>lo</td>
<td>0.0002</td>
</tr>
<tr>
<td>yes</td>
<td>lo</td>
<td>off</td>
<td>med</td>
<td>2.50e-06</td>
</tr>
<tr>
<td>yes</td>
<td>lo</td>
<td>off</td>
<td>hi</td>
<td>2.50e-06</td>
</tr>
<tr>
<td>yes</td>
<td>med</td>
<td>on</td>
<td>lo</td>
<td>0.0001</td>
</tr>
<tr>
<td>yes</td>
<td>med</td>
<td>on</td>
<td>med</td>
<td>0.0146</td>
</tr>
<tr>
<td>yes</td>
<td>med</td>
<td>on</td>
<td>hi</td>
<td>0.0001</td>
</tr>
<tr>
<td>yes</td>
<td>med</td>
<td>off</td>
<td>lo</td>
<td>1.50e-06</td>
</tr>
<tr>
<td>yes</td>
<td>med</td>
<td>off</td>
<td>med</td>
<td>0.0001</td>
</tr>
<tr>
<td>yes</td>
<td>med</td>
<td>off</td>
<td>hi</td>
<td>1.50e-06</td>
</tr>
<tr>
<td>yes</td>
<td>hi</td>
<td>on</td>
<td>lo</td>
<td>9.99e-05</td>
</tr>
<tr>
<td>yes</td>
<td>hi</td>
<td>on</td>
<td>med</td>
<td>9.99e-05</td>
</tr>
<tr>
<td>yes</td>
<td>hi</td>
<td>on</td>
<td>hi</td>
<td>0.0098</td>
</tr>
<tr>
<td>yes</td>
<td>hi</td>
<td>off</td>
<td>lo</td>
<td>1.00e-07</td>
</tr>
<tr>
<td>yes</td>
<td>hi</td>
<td>off</td>
<td>med</td>
<td>1.00e-07</td>
</tr>
<tr>
<td>yes</td>
<td>hi</td>
<td>off</td>
<td>hi</td>
<td>9.80e-06</td>
</tr>
</tbody>
</table>

\[
P(\text{Break-in}=\text{yes}, \text{Alarm}=\text{on}) = 0.0496 \quad P(\text{Break-in}=\text{yes} | \text{Alarm}=\text{on}) = \frac{P(\text{Break-in}=\text{yes}, \text{Alarm}=\text{on})}{\sum_b P(\text{Break-in}=b, \text{Alarm}=\text{on})} = 0.0665
\]

\[
= 0.427
\]
Knowing the model structure (statistical dependencies), complicated models become manageable.

\[ P(Br, W, A, Ba) = P(Br) P(W) P(A | Br, W) P(Ba | W) \]
Knowing the model structure (statistical dependencies), complicated models become manageable.

\[ P(Br, W, A, Ba) = P(Br) P(W) P(A | Br, W) P(Ba | W) \]

- Can estimate parts in isolation e.g. \( P(Wind) \) from weather history.

<table>
<thead>
<tr>
<th>( P(Br) )</th>
<th>yes</th>
<th>no</th>
</tr>
</thead>
<tbody>
<tr>
<td>.05</td>
<td>.95</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( P(W) )</th>
<th>lo</th>
<th>mid</th>
<th>hi</th>
</tr>
</thead>
<tbody>
<tr>
<td>.5</td>
<td>.3</td>
<td>.2</td>
<td></td>
</tr>
</tbody>
</table>

| \( P(A | Br, W) \) | on | off |
|-------------------|----|-----|
| Br=yes W=lo       | .99| .01 |
| Br=yes W=med      | .99| .01 |
| Br=yes W=hi       | .99| .001|
| Br=no W=lo        | .01| .99 |
| Br=no W=med       | .05| .95 |
| Br=no W=hi        | .25| .75 |

| \( P(Ba | W) \) | lo | mid | hi |
|----------------|----|-----|----|
| W=lo           | .98| .01 | .01|
| W=mid          | .01| .98 | .01|
| W=hi           | .01| .01 | .98|
Knowing the model structure (statistical dependencies), complicated models become manageable.

\[
P(\text{Br, W, A, Ba}) = P(\text{Br}) P(\text{W}) P(\text{A} | \text{Br, W}) P(\text{Ba} | \text{W})
\]

- Can estimate parts in isolation e.g. \(P(\text{Wind})\) from weather history.
- Can sample by following the graph from roots to leaves.

<table>
<thead>
<tr>
<th>(P(\text{Br}))</th>
<th>yes</th>
<th>no</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>.05</td>
<td>.95</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(P(\text{W}))</th>
<th>lo</th>
<th>mid</th>
<th>hi</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>.5</td>
<td>.3</td>
<td>.2</td>
</tr>
</tbody>
</table>

| \(P(\text{A} | \text{Br, W})\) | on | off |
|------------------------|----|-----|
| \(\text{Br=yes}\) | \(\text{W=lo}\) | .99 | .01 |
| \(\text{Br=yes}\) | \(\text{W=med}\) | .99 | .01 |
| \(\text{Br=yes}\) | \(\text{W=hi}\) | .999 | .001 |
| \(\text{Br=no}\) | \(\text{W=lo}\) | .01 | .99 |
| \(\text{Br=no}\) | \(\text{W=med}\) | .05 | .95 |
| \(\text{Br=no}\) | \(\text{W=hi}\) | .25 | .75 |

| \(P(\text{Ba} | \text{W})\) | lo | mid | hi |
|----------------|----|-----|----|
| \(\text{W=lo}\) | .98 | .01 | .01 |
| \(\text{W=mid}\) | .01 | .98 | .01 |
| \(\text{W=hi}\) | .01 | .01 | .98 |
Bayes Nets:
reduce number of parameters & aid estimation
let us reason about **independencies** in a model
are a building-block for modeling **causality**
Bayes Nets:

are not neural network diagrams
encode structure, not parametrization
are non-unique for a distribution
encode independence requirements, not necessarily all
BN are not neural net diagrams

Recall the RNN language model:

- In statistical terms, what are we modeling?
Recall the RNN language model:

- In statistical terms, what are we modeling?

\[
P(X_1, \ldots, X_n) = P(X_1) P(X_2 \mid X_1) P(X_3 \mid X_1, X_2) \ldots
\]
BN are not neural net diagrams

Recall the RNN language model:

- In statistical terms, what are we modeling?

\[
P(X_1, \ldots, X_n) = P(X_1) P(X_2 \mid X_1) P(X_3 \mid X_1, X_2) \ldots
\]

- Bayes Net:

- Not useful! Everything conditionally depends on everything. (more later)
Neural net diagrams (and computation graphs) show **how to compute something**

Bayes networks show **how a distribution factorizes** (what is assumed independent)
A BN tells us: **how the distribution decomposes**  
A BN can’t tell us: **what the probabilities are!**

Example: \( X \in \mathcal{X} = \text{all English sentences}, Y \in \{\text{sports, music, ...}\} \).

BN for a generative model: \[ Y \rightarrow X \]

We must posit what are \( P(Y) \) and \( P(X \mid Y) \). **Many possible options!**
A BN tells us: **how the distribution decomposes**

A BN can’t tell us: **what the probabilities are!**

Example: \( X \in \mathcal{X} = \) all English sentences, \( Y \in \{\text{sports, music, } \ldots \} \).

BN for a generative model:

We must posit what are \( P(Y) \) and \( P(X \mid Y) \). Many possible options!

**P(Y):** uniform: \( P(Y = \text{sports}) = P(Y = \text{music}) = \frac{1}{|y|} \),
A BN tells us: **how the distribution decomposes**
A BN can’t tell us: **what the probabilities are!**

Example: $X \in \mathcal{X} = \text{all English sentences}, Y \in \{\text{sports, music, ...}\}$. 

BN for a generative model: 

We must posit what are $P(Y)$ and $P(X \mid Y)$. Many possible options!

$P(Y)$: uniform: $P(Y = \text{sports}) = P(Y = \text{music}) = \frac{1}{|\mathcal{Y}|}$, or estimated from data.
A BN tells us: **how the distribution decomposes**

A BN can’t tell us: **what the probabilities are!**

Example: $X \in \mathcal{X} =$ all English sentences, $Y \in \{\text{sports, music, ...}\}$.

BN for a generative model:

We must posit what are $P(Y)$ and $P(X \mid Y)$. **Many possible options!**

$P(Y)$: uniform: $P(Y = \text{sports}) = P(Y = \text{music}) = \frac{1}{|Y|}$, or estimated from data.

$P(X \mid Y)$ (remember: values of $X$ are sentences)
BN encode structure, not parametrization

A BN tells us: **how the distribution decomposes**
A BN can’t tell us: **what the probabilities are!**

Example: $X \in \mathcal{X} =$ all English sentences, $Y \in \{ \text{sports, music, ...} \}$.  

![BN Diagram]

We must posit what are $P(Y)$ and $P(X \mid Y)$. **Many possible options!**

- **P(Y): uniform**: $P(Y = \text{sports}) = P(Y = \text{music}) = \frac{1}{|Y|}$, or estimated from data.
- **P(X | Y)** (remember: values of $X$ are sentences)
  - Naive Bayes:  
    $$ P(X \mid Y) = \prod_{j=1}^{L} P(X_j \mid Y) $$
  - Per-class Markov language model:  
    $$ P(X \mid Y) = \prod_{j=1}^{L} P(X_j \mid X_{j-1}, Y) $$
  - Per-class recurrent NN language model:  
    $$ P(X \mid Y) = \text{LSTM}(x_1, ..., x_L; w_{vy}) $$

Variables need not be discrete! mixture of Gaussians: $P(X \mid Y = y) \sim \mathcal{N}(\mu_y, \Sigma_y)$. 

Vlad Niculae & André Martins (IST)  
Lecture 7: Probabilistic Graphical Models  
IST, Fall 2019
BN encode structure, not parametrization

A BN tells us: **how the distribution decomposes**

A BN can’t tell us: **what the probabilities are!**

Example: \( X \in \mathcal{X} = \text{all English sentences}, Y \in \{\text{sports, music, ...}\}. \)

BN for a generative model:

\[
\begin{array}{c}
\text{Y} \\
\rightarrow \\
\text{X}
\end{array}
\]

We must posit what are \( P(Y) \) and \( P(X \mid Y) \). Many possible options!

\( P(Y) \): uniform: \( P(Y = \text{sports}) = P(Y = \text{music}) = \frac{1}{|Y|}, \) or estimated from data.

\( P(X \mid Y) \)  (remember: values of \( X \) are sentences)

- Naive Bayes \( P(X \mid Y) = \prod_{j=1}^{L} P(X_j \mid Y) \)
- Per-class Markov language model \( P(X \mid Y) = \prod_{j=1}^{L} P(X_j \mid X_{j-1}, Y) \)
BN encode structure, not parametrization

A BN tells us: **how the distribution decomposes**
A BN can’t tell us: **what the probabilities are!**

Example: $X \in \mathcal{X} =$ all English sentences, $Y \in \{\text{sports, music, ...}\}$.

BN for a generative model:

We must posit what are $P(Y)$ and $P(X \mid Y)$. Many possible options!

$P(Y)$: uniform: $P(Y = \text{sports}) = P(Y = \text{music}) = \frac{1}{|Y|}$, or estimated from data.

$P(X \mid Y)$ (remember: values of $X$ are sentences)

- **Naive Bayes**
  \[ P(X \mid Y) = \prod_{j=1}^{L} P(X_j \mid Y) \]

- **Per-class Markov language model**
  \[ P(X \mid Y) = \prod_{j=1}^{L} P(X_j \mid X_{j-1}, Y) \]

- **Per-class recurrent NN language model**
  \[ P(X \mid Y) = \text{LSTM}(x_1, \ldots, x_L; w_Y) \]
BN encode structure, not parametrization

A BN tells us: **how the distribution decomposes**
A BN can’t tell us: **what the probabilities are!**

Example: $X \in X = \text{all English sentences}$, $Y \in \{\text{sports, music, ...}\}$.

BN for a generative model:

We must posit what are $P(Y)$ and $P(X \mid Y)$. **Many possible options!**

$P(Y)$: uniform: $P(Y = \text{sports}) = P(Y = \text{music}) = \frac{1}{|Y|}$, or estimated from data.

$P(X \mid Y)$ (remember: values of $X$ are sentences)

- **Naive Bayes** $P(X \mid Y) = \prod_{j=1}^{L} P(X_j \mid Y)$
- **Per-class Markov language model** $P(X \mid Y) = \prod_{j=1}^{L} P(X_j \mid X_{j-1}, Y)$
- **Per-class recurrent NN language model** $P(X \mid Y) = \text{LSTM}(x_1, \ldots, x_L; w_y)$

$P(X \mid Y)$ need not be parametrized as a table.
BN encode structure, not parametrization

A BN tells us: how the distribution decomposes
A BN can’t tell us: what the probabilities are!

Example: \( X \in \mathcal{X} = \) all English sentences, \( Y \in \{ \text{sports, music, ...} \} \).

BN for a generative model:

We must posit what are \( P(Y) \) and \( P(X \mid Y) \). Many possible options!

- \( P(Y) \): uniform: \( P(Y = \text{sports}) = P(Y = \text{music}) = \frac{1}{|Y|} \), or estimated from data.
- \( P(X \mid Y) \) (remember: values of \( X \) are sentences)
  - Naive Bayes: \( P(X \mid Y) = \prod_{j=1}^{L} P(X_j \mid Y) \)
  - Per-class Markov language model: \( P(X \mid Y) = \prod_{j=1}^{L} P(X_j \mid X_{j-1}, Y) \)
  - Per-class recurrent NN language model: \( P(X \mid Y) = \text{LSTM}(x_1, \ldots, x_L; w_Y) \)

\( P(X \mid Y) \) need not be parametrized as a table.

Variables need not be discrete! mixture of Gaussians: \( P(X \mid Y = y) \sim \mathcal{N}(\mu_y, \Sigma_y) \).
Equivalent factorizations

There are many possible factorizations! \( P(X, Y) = \)
Equivalent factorizations

There are many possible factorizations! $P(X, Y) =$

$$
\begin{align*}
P(X, Y) & = P(X) P(Y | X) \\
& = P(Y, X) \\
& = P(Y) P(X | Y)
\end{align*}
$$

The first two are valid Bayes nets for any $P(X, Y)$!

In fact, recall generative vs discriminative classifiers!

- **Generative (e.g. naïve Bayes):**
  - To classify, we would compute $P(Y | X)$ via Bayes' rule.

- **Discriminative (e.g. logistic regression):**
  - In LR, we don't model $P(X)$, we assume $X$ is always observed (gray).

Some arrow direction choices are harder to estimate.

Some make more sense (why?): Wind barometer vs. Wind barometer.
There are many possible factorizations! $P(X, Y) =$

\[
\begin{align*}
X & \rightarrow Y \\
& \quad \text{P}(X) \ P(Y \mid X) \\
\end{align*}
\]

\[
\begin{align*}
X & \rightarrow Y \\
& \quad \text{P}(Y) \ P(X \mid Y) \\
\end{align*}
\]
Equivalent factorizations

There are many possible factorizations! $P(X, Y) =$

\[ P(X) P(Y \mid X) \quad P(Y) P(X \mid Y) \quad P(X) P(Y) \]

The first two are valid Bayes nets for any $P(X, Y)$!

In fact, recall generative vs discriminative classifiers!

- Generative (e.g. naïve Bayes): To classify, we would compute $P(Y \mid X)$ via Bayes' rule.
- Discriminative (e.g. logistic regression): in LR, we don’t model $P(X)$, we assume $X$ is always observed (gray).

Some arrow direction choices are harder to estimate.

Some make more sense (why?): Wind → Baromtr. vs. Wind ← Baromtr.
Equivalent factorizations

There are many possible factorizations! \( P(X, Y) = \)

\[
\begin{align*}
X & \rightarrow Y \\
& \quad P(X) P(Y \mid X) \\
X & \leftarrow Y \\
& \quad P(Y) P(X \mid Y) \\
X & \rightarrow Y \\
& \quad P(X) P(Y)
\end{align*}
\]

The first two are valid Bayes nets for any \( P(X, Y) \)!
Equivalent factorizations

There are many possible factorizations! $P(X, Y) =$

$$P(X) P(Y | X)$$

$$P(Y) P(X | Y)$$

$$P(X) P(Y)$$

The first two are valid Bayes nets for any $P(X, Y)$!

In fact, recall generative vs discriminative classifiers!

- Generative (e.g. naïve Bayes):

  $X \rightarrow Y$

  To classify, we would compute $P(Y | X)$ via Bayes’ rule.

- Discriminative (e.g. logistic regression)

  $X \rightarrow Y$

  in LR, we don’t model $P(X)$, we assume $X$ is always observed (gray).
Equivalent factorizations

There are many possible factorizations: $P(X, Y) =$

\[
\begin{align*}
X & \rightarrow Y \\
P(X) & P(Y | X) \\
Y & \rightarrow X \\
P(Y) & P(X | Y) \\
X & \rightarrow Y \\
P(X) & P(Y)
\end{align*}
\]

The first two are valid Bayes nets for any $P(X, Y)$!

In fact, recall generative vs discriminative classifiers!

- Generative (e.g. naïve Bayes):
  
  To classify, we would compute $P(Y | X)$ via Bayes’ rule.

- Discriminative (e.g. logistic regression)
  
  In LR, we don’t model $P(X)$, we assume $X$ is always observed (gray).

Some arrow direction choices are harder to estimate.

Some make more sense (why?):

- Barmtr. $\rightarrow$ Wind vs. Barmtr. $\rightarrow$ Wind
Recall, we say $X \perp \perp Y$ iff. $P(X, Y) = P(X)P(Y)$
Let $X = \text{grade in DSL}, Y = \text{month you were born}$.

Bayes net (1): $\begin{array}{c}
\text{X} \\
\text{Y} \\
\end{array}$
Recall, we say \( X \perp \!\!\!\!\!\!\!\!\!\!\perp Y \) iff. \( P(X, Y) = P(X)P(Y) \)

Let \( X = \) grade in DSL, \( Y = \) month you were born.

Bayes net (1):

Example parametrization:

\[
\begin{array}{cccc}
P(X) & A^+ & A & B & \ldots \\
.01 & .02 & .04 & \\
\hline
P(Y) & \text{Jan} & \text{Feb} & \text{Mar} & \ldots \\
.10 & .12 & .09 & \\
\end{array}
\]
Minimal independence assumptions

Recall, we say $X \perp \perp Y$ iff. $P(X, Y) = P(X)P(Y)$
Let $X =$ grade in DSL, $Y =$ month you were born.

Bayes net (1): $\text{\circ X \quad \circ Y}$

Example parametrization:

<table>
<thead>
<tr>
<th>P(X)</th>
<th>A+</th>
<th>A</th>
<th>B</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>.01</td>
<td>.02</td>
<td>.04</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>P(Y)</th>
<th>Jan</th>
<th>Feb</th>
<th>Mar</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>.10</td>
<td>.12</td>
<td>.09</td>
<td></td>
</tr>
</tbody>
</table>

BN (1) imposes $X \perp \perp Y$
in any parametrization.
Minimal independence assumptions

Recall, we say $X \perp \perp Y$ iff. $P(X, Y) = P(X)P(Y)$
Let $X =$ grade in DSL, $Y =$ month you were born.

Bayes net (1): $\begin{array}{c} X \\ Y \end{array}$

Bayes net (2): $\begin{array}{c} X \\ Y \end{array}$

Example parametrization:

| P(X) | A+ | A | B | ...
|------|----|---|---|---
|      | .01| .02| .04|  

| P(Y) | Jan | Feb | Mar | ...
|------|-----|-----|-----|---
|      | .10 | .12 | .09 |  

BN (1) imposes $X \perp \perp Y$ in any parametrization.

Does it mean we must have $X \not\perp \perp Y$?
Minimal independence assumptions

Recall, we say $X \perp \perp Y$ iff. $P(X, Y) = P(X)P(Y)$
Let $X = \text{grade in DSL}$, $Y = \text{month you were born}$.

Bayes net (1): $X \rightarrow Y$

Example parametrization:

<table>
<thead>
<tr>
<th>$P(X)$</th>
<th>A+</th>
<th>A</th>
<th>B</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>.01</td>
<td>.02</td>
<td>.04</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$P(Y)$</th>
<th>Jan</th>
<th>Feb</th>
<th>Mar</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>.10</td>
<td>.12</td>
<td>.09</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

BN (1) imposes $X \perp \perp Y$ in any parametrization.

Bayes net (2): $X \rightarrow Y$

Does it mean we must have $X \nparallel Y$? **NO!**

Example parametrization:

<table>
<thead>
<tr>
<th>$P(Y)$</th>
<th>Jan</th>
<th>Feb</th>
<th>Mar</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>.10</td>
<td>.12</td>
<td>.09</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$P(X \mid Y)$</th>
<th>A+</th>
<th>A</th>
<th>B</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y=\text{Jan}$</td>
<td>.01</td>
<td>.02</td>
<td>.04</td>
<td></td>
</tr>
<tr>
<td>$Y=\text{Feb}$</td>
<td>.01</td>
<td>.02</td>
<td>.04</td>
<td></td>
</tr>
<tr>
<td>$Y=\text{Mar}$</td>
<td>.01</td>
<td>.02</td>
<td>.04</td>
<td></td>
</tr>
</tbody>
</table>

...
Minimal independence assumptions

Recall, we say $X \perp \perp Y$ iff. $P(X, Y) = P(X)P(Y)$
Let $X =$ grade in DSL, $Y =$ month you were born.

Bayes net (1):

Example parametrization:

<table>
<thead>
<tr>
<th>$P(X)$</th>
<th>A+</th>
<th>A</th>
<th>B</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>.01</td>
<td>.02</td>
<td>.04</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$P(Y)$</th>
<th>Jan</th>
<th>Feb</th>
<th>Mar</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>.10</td>
<td>.12</td>
<td>.09</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

BN (1) imposes $X \perp \perp Y$ in any parametrization.

Bayes net (2):

Does it mean we must have $X \not\perp \not\perp Y$? NO!

<table>
<thead>
<tr>
<th>$P(Y)$</th>
<th>Jan</th>
<th>Feb</th>
<th>Mar</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>.10</td>
<td>.12</td>
<td>.09</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| $P(X | Y)$ | A+ | A | B | ... |
|-----------|----|---|---|-----|
| Y=Jan     | .01 | .02 | .04|
| Y=Feb     | .01 | .02 | .04|
| Y=Mar     | .01 | .02 | .04|
| ...       |    |    |   |

A BN constraints what independences must be in the model as a minimum.
1 Directed Models

Bayes networks

Conditional independence and D-separation

Causal graphs & the $do$ operator

2 Undirected Models

Markov networks

Factor graphs
Conditional independence in Bayes nets

Identifying independences in a distribution is generally hard. Bayes nets let us reason about it via graph algorithms!

**Definition (conditional independence)**

A is independent of B given a set of variables $C = \{ C_1, \ldots, C_n \}$, denoted as

$$A \perp \!\!\!\perp B \mid C,$$

if and only if

$$P(A, B \mid C_1, \ldots, C_n) = P(A \mid C_1, \ldots, C_n) P(B \mid C_1, \ldots, C_n).$$

**Note.** Equivalently, $P(A \mid B, C_1, \ldots, C_n) = P(A \mid C_1, \ldots, C_n)$. Intuitively: if we observe $C$, does observing $B$ too bring us more info about $A$?

---

Break-in $\rightarrow$ Alarm $\rightarrow$ Vlad is annoyed
Three fundamental relationships in BN

The Fork

\[ A \perp \perp B \mid C \]

Given \( C \), \( A \) and \( B \) are independent.
Example: Alarm ← Wind → Barometer

The Chain

\[ A \perp \perp B \mid C \]

After observing \( C \), further observing \( A \) would not tell us about \( B \).
Example: Burglary → Alarm → Vlad distracted

The Collider

\[ A \perp \perp B \]

Surprisingly, but not \( A \perp \perp B \mid C \)!
Example: Burglary → Alarm ← Wind
Burglaries occur regardless how windy it is.
If alarm rings, hearing wind makes burglary less likely!
Burglary is “explained away” by wind.
Detecting independence: d-separation

Algorithm for deciding if $A$ and $B$ are **d-separated** given set $C$, implying:

$$ A \perp \! \! \! \! \perp B \mid C. $$

For all paths $P$ from $A$ to $B$ in the **skeleton** of the BN, at least one holds:

1. $P$ includes a fork with observed parent:
   $$ X \leftarrow C \rightarrow Y \quad (\text{with } C \in C) $$

2. $P$ includes a chain with observed middle:
   $$ X \rightarrow C \rightarrow Y \quad \text{or} \quad X \leftarrow C \leftarrow Y \quad (\text{with } C \in C) $$

3. $P$ includes a collider
   $$ X \rightarrow U \leftarrow Y \quad (\text{with } U \notin C) $$
Wind $\perp\!\!\!\!\!\!\perp$ Barometer?
Examples

Wind \perp \text{Barometer? No}
Wind ⊥ Barometer? No
Break-in ⊥ Wind?
Examples

Break-in  Wind

Alarm  Barometer

Wind ⊥ Barometer? No
Break-in ⊥ Wind? Yes
Examples

Wind ⊥ Barometer? No
Break-in ⊥ Wind? Yes
Break-in ⊥ Barometer?
Wind $\perp\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
Examples

Wind ⊥ Barometer? No
Break-in ⊥ Wind? Yes
Break-in ⊥ Barometer? Yes
Break-in ⊥ Barometer | Alarm?
Wind $\perp \!\!\!\perp$ Barometer? No
Break-in $\perp \!\!\!\perp$ Wind? Yes
Break-in $\perp \!\!\!\perp$ Barometer? Yes
Break-in $\perp \!\!\!\perp$ Barometer $\mid$ Alarm? No
Wind ⊥ Barometer? No
Break-in ⊥ Wind? Yes
Break-in ⊥ Barometer? Yes
Break-in ⊥ Barometer | Alarm? No
Break-in ⊥ Barometer | Alarm, Wind?
Wind ⊥ Barometer? No
Break-in ⊥ Wind? Yes
Break-in ⊥ Barometer? Yes
Break-in ⊥ Barometer | Alarm? No
Break-in ⊥ Barometer | Alarm, Wind? Yes
Wind \perp \text{ Barometer}\,? \textbf{No}

Break-in \perp \text{ Wind}\,? \textbf{Yes}

Break-in \perp \text{ Barometer}\,? \textbf{Yes}

Break-in \perp \text{ Barometer} \mid \text{ Alarm}\,? \textbf{No}

Break-in \perp \text{ Barometer} \mid \text{ Alarm, Wind}\,? \textbf{Yes}
Examples

Wind $\perp \perp$ Barometer? No
Break-in $\perp \perp$ Wind? Yes
Break-in $\perp \perp$ Barometer? Yes
Break-in $\perp \perp$ Barometer $|$ Alarm? No
Break-in $\perp \perp$ Barometer $|$ Alarm, Wind? Yes

\[ Y_{i+1} \perp \perp Y_{i-1}? \]
Examples

Wind \perp \text{ Barometer? \textcolor{red}{No}}

\text{Break-in} \perp \text{ Wind? \textcolor{red}{Yes}}

\text{Break-in} \perp \text{ Barometer? \textcolor{red}{Yes}}

\text{Break-in} \perp \text{ Barometer | Alarm? \textcolor{red}{No}}

\text{Break-in} \perp \text{ Barometer | Alarm, Wind? \textcolor{red}{Yes}}

Y_{i+1} \perp Y_{i-1}? \textcolor{red}{No}
Examples

Wind $\perp \!\!\!\!\!\perp$ Barometer? No
Break-in $\perp \!\!\!\!\!\perp$ Wind? Yes
Break-in $\perp \!\!\!\!\!\perp$ Barometer? Yes
Break-in $\perp \!\!\!\!\!\perp$ Barometer | Alarm? No
Break-in $\perp \!\!\!\!\!\perp$ Barometer | Alarm, Wind? Yes

\[ Y_{i+1} \perp \!\!\!\!\!\perp Y_{i-1} \, ? \, No \]
\[ Y_{i+1} \perp \!\!\!\!\!\perp Y_{i-1} \mid Y_i ? \]
Examples

Wind ⊥ Barometer? No
Break-in ⊥ Wind? Yes
Break-in ⊥ Barometer? Yes
Break-in ⊥ Barometer | Alarm? No
Break-in ⊥ Barometer | Alarm, Wind? Yes

\[ Y_{i+1} \perp Y_{i-1} ? \text{ No} \]
\[ Y_{i+1} \perp Y_{i-1} | Y_i ? \text{ Yes} \]
Wind $\perp \! \! \! \! \perp$ Barometer? \textbf{No}

Break-in $\perp \! \! \! \! \perp$ Wind? \textbf{Yes}

Break-in $\perp \! \! \! \! \perp$ Barometer? \textbf{Yes}

Break-in $\perp \! \! \! \! \perp$ Barometer $|$ Alarm? \textbf{No}

Break-in $\perp \! \! \! \! \perp$ Barometer $|$ Alarm, Wind? \textbf{Yes}

\[ Y_{i+1} \perp \! \! \! \! \perp Y_{i-1} ? \textbf{ No} \]

\[ Y_{i+1} \perp \! \! \! \! \perp Y_{i-1} \mid Y_i \textbf{ Yes} \]

\[ Y_{i+1} \perp \! \! \! \! \perp X_i ? \]
Examples

Wind ⊥ Barometer? No
Break-in ⊥ Wind? Yes
Break-in ⊥ Barometer? Yes
Break-in ⊥ Barometer | Alarm? No
Break-in ⊥ Barometer | Alarm, Wind? Yes

\[
Y_{i+1} \perp Y_{i-1} ? \text{ No}
\]
\[
Y_{i+1} \perp Y_{i-1} | Y_i ? \text{ Yes}
\]
\[
Y_{i+1} \perp X_i ? \text{ No}
\]
Examples

Wind $\perp$ Barometer? No
Break-in $\perp$ Wind? Yes
Break-in $\perp$ Barometer? Yes
Break-in $\perp$ Barometer $|$ Alarm? No
Break-in $\perp$ Barometer $|$ Alarm, Wind? Yes

$Y_{i+1} \perp Y_{i-1}?$ No
$Y_{i+1} \perp Y_{i-1} \mid Y_i?$ Yes
$Y_{i+1} \perp X_i?$ No
$Y_{i+1} \perp X_i \mid Y_i?$
**Examples**

Wind $\perp$ Barometer? **No**

Break-in $\perp$ Wind? **Yes**

Break-in $\perp$ Barometer? **Yes**

Break-in $\perp$ Barometer $|$ Alarm? **No**

Break-in $\perp$ Barometer $|$ Alarm, Wind? **Yes**

$Y_{i+1} \perp Y_{i-1}$? **No**

$Y_{i+1} \perp Y_{i-1} \mid Y_i$? **Yes**

$Y_{i+1} \perp X_i$? **No**

$Y_{i+1} \perp X_i \mid Y_i$? **Yes**
In papers, you’ll see statistical models defined through *generative stories*:

\[
\mu \sim \text{Uniform}([-1, 1]) \\
\sigma \sim \text{Uniform}([1, 2]) \\
X | \mu, \sigma \sim \text{Normal}(\mu, \sigma)
\]

Plate notation is a way to denote *repetition of templates*:

\[
\mu \sim \text{Uniform}([-1, 1]) \\
\sigma \sim \text{Uniform}([1, 2]) \\
X_n | \mu, \sigma \sim \text{Normal}(\mu, \sigma) \quad i = 1, \ldots, N
\]
1 Directed Models

Bayes networks

Conditional independence and D-separation

Causal graphs & the do operator

2 Undirected Models

Markov networks

Factor graphs
Correlation does not imply causation; but then, *what does?*
Bayes nets only model independence assumptions.

The correlation between the barometer reading $B$ and wind strength $W$ can be represented either way:

- **Seeing** that the barometer reading is high, we can forecast wind.

  \[
  \begin{array}{ccc}
  P(W | B) & \text{lo} & \text{mid} & \text{hi} \\
  B = \text{lo} & .98 & .01 & .01 \\
  B = \text{mid} & .01 & .98 & .01 \\
  B = \text{hi} & .01 & .01 & .98
  \end{array}
  \]

  But **setting** the barometer needle to high manually won’t cause wind!

  We write: $P(W | \text{do}(B = \text{hi})) =$?
Directed Models

Bayes networks
Conditional independence and D-separation
Causal graphs & the do operator

Undirected Models

Markov networks
Factor graphs
Directed Models

- Bayes networks
- Conditional independence and D-separation
- Causal graphs & the \textit{do} operator

Undirected Models

- Markov networks
- Factor graphs
Outline

1 Directed Models

Bayes networks
Conditional independence and D-separation
Causal graphs & the do operator

2 Undirected Models

Markov networks
Factor graphs