Lecture 7: Probabilistic Graphical Models

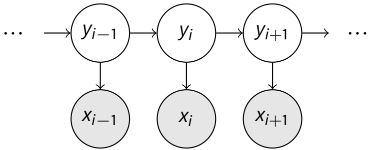
Vlad Niculae & André Martins

Deep Structured Learning Course, Fall 2019

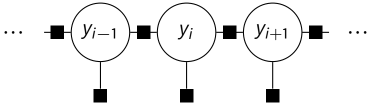
Graphical Models

In this unit, we will formalize & extend these graphical representations encountered in previous lectures.

Directed
(today)



Undirected
(next time)



Outline

Outline

1. **Directed Models**

   Bayes networks

   Conditional independence and D-separation

   Causal graphs & the *do* operator

2. **Undirected Models**

   Markov networks

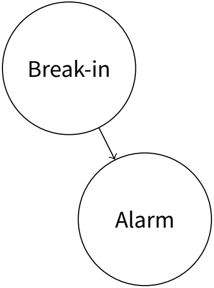
   Factor graphs

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?



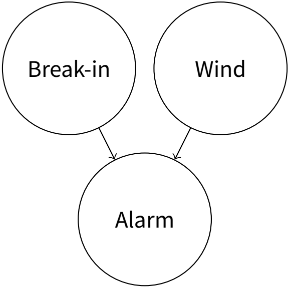
P(B)	B=yes	B=no
	.05	.95

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

- $\mathrm{P}(B \mid A) = ?$

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?



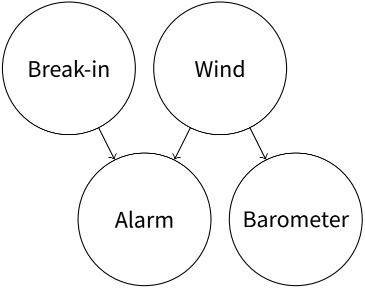
P(B)	B=yes	B=no
	.05	.95

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

- $\mathrm{P}(B \mid A) = ?$
- Can we observe wind? $\mathrm{P}(B \mid A, W) = ?$

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?



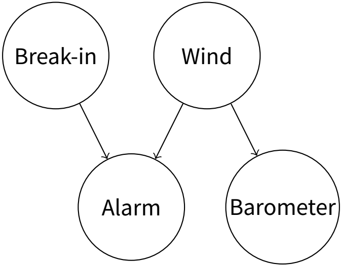
P(B)	B=yes	B=no
	.05	.95

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

- $\mathrm{P}(B \mid A) = ?$
- Can we observe wind? $\mathrm{P}(B \mid A, W) = ?$
- 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.

$$\boxed{\text{In general:} \quad P(X_1, \ldots, X_n) = \prod_i P\left(X_i \mid \text{parents}(X_i)\right)}$$

For example: $\quad$ P(Break-in, Wind, Alarm, Barometer)

$$= P(\text{Break-in})\, P(\text{Wind})\, P(\text{Alarm} \mid \text{Break-in}, \text{Wind})\, P(\text{Barometer} \mid \text{Wind})$$

Without any structure, P(Break-in, Wind, Alarm, Barometer) would have to be stored & estimated like

Brk.	Wind	Alarm	Bar.	P
yes	lo	on	lo	0.0243
yes	lo	on	med	0.0002
yes	lo	on	hi	0.0002
yes	lo	off	lo	0.0002
yes	lo	off	med	2.50e-06
yes	lo	off	hi	2.50e-06
yes	med	on	lo	0.0001
yes	med	on	med	0.0146
yes	med	on	hi	0.0001
yes	med	off	lo	1.50e-06
yes	med	off	med	0.0001
yes	med	off	hi	1.50e-06
yes	hi	on	lo	9.99e-05
yes	hi	on	med	9.99e-05
yes	hi	on	hi	0.0098
yes	hi	off	lo	1.00e-07
yes	hi	off	med	1.00e-07
yes	hi	off	hi	9.80e-06

Brk.	Wind	Alarm	Bar.	P
no	lo	on	lo	0.0047
no	lo	on	med	4.75e-05
no	lo	on	hi	4.75e-05
no	lo	off	lo	0.4608
no	lo	off	med	0.0047
no	lo	off	hi	0.0047
no	med	on	lo	0.0001
no	med	on	med	0.0140
no	med	on	hi	0.0001
no	med	off	lo	0.0027
no	med	off	med	0.2653
no	med	off	hi	0.0027
no	hi	on	lo	0.0005
no	hi	on	med	0.0005
no	hi	on	hi	0.0466
no	hi	off	lo	0.0014
no	hi	off	med	0.0014
no	hi	off	hi	0.1397

Without any structure, P(Break-in, Wind, Alarm, Barometer) would have to be stored & estimated like

Brk.	Wind	Alarm	Bar.	P
yes	lo	on	lo	0.0243
yes	lo	on	med	0.0002
yes	lo	on	hi	0.0002
yes	lo	off	lo	0.0002
yes	lo	off	med	2.50e-06
yes	lo	off	hi	2.50e-06
yes	med	on	lo	0.0001
yes	med	on	med	0.0146
yes	med	on	hi	0.0001
yes	med	off	lo	1.50e-06
yes	med	off	med	0.0001
yes	med	off	hi	1.50e-06
yes	hi	on	lo	9.99e-05
yes	hi	on	med	9.99e-05
yes	hi	on	hi	0.0098
yes	hi	off	lo	1.00e-07
yes	hi	off	med	1.00e-07
yes	hi	off	hi	9.80e-06

Brk.	Wind	Alarm	Bar.	P
no	lo	on	lo	0.0047
no	lo	on	med	4.75e-05
no	lo	on	hi	4.75e-05
no	lo	off	lo	0.4608
no	lo	off	med	0.0047
no	lo	off	hi	0.0047
no	med	on	lo	0.0001
no	med	on	med	0.0140
no	med	on	hi	0.0001
no	med	off	lo	0.0027
no	med	off	med	0.2653
no	med	off	hi	0.0027
no	hi	on	lo	0.0005
no	hi	on	med	0.0005
no	hi	on	hi	0.0466
no	hi	off	lo	0.0014
no	hi	off	med	0.0014
no	hi	off	hi	0.1397

P(Break-in=yes, Alarm=on) $=$ 0.0496

Without any structure, P(Break-in, Wind, Alarm, Barometer) would have to be stored & estimated like

Brk.	Wind	Alarm	Bar.	P
yes	lo	on	lo	0.0243
yes	lo	on	med	0.0002
yes	lo	on	hi	0.0002
yes	lo	off	lo	0.0002
yes	lo	off	med	2.50e-06
yes	lo	off	hi	2.50e-06
yes	med	on	lo	0.0001
yes	med	on	med	0.0146
yes	med	on	hi	0.0001
yes	med	off	lo	1.50e-06
yes	med	off	med	0.0001
yes	med	off	hi	1.50e-06
yes	hi	on	lo	9.99e-05
yes	hi	on	med	9.99e-05
yes	hi	on	hi	0.0098
yes	hi	off	lo	1.00e-07
yes	hi	off	med	1.00e-07
yes	hi	off	hi	9.80e-06

Brk.	Wind	Alarm	Bar.	P
no	lo	on	lo	0.0047
no	lo	on	med	4.75e-05
no	lo	on	hi	4.75e-05
no	lo	off	lo	0.4608
no	lo	off	med	0.0047
no	lo	off	hi	0.0047
no	med	on	lo	0.0001
no	med	on	med	0.0140
no	med	on	hi	0.0001
no	med	off	lo	0.0027
no	med	off	med	0.2653
no	med	off	hi	0.0027
no	hi	on	lo	0.0005
no	hi	on	med	0.0005
no	hi	on	hi	0.0466
no	hi	off	lo	0.0014
no	hi	off	med	0.0014
no	hi	off	hi	0.1397

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

$$P(\text{Break-in=no, Alarm=on}) = 0.0665$$

Without any structure, P(Break-in, Wind, Alarm, Barometer) would have to be stored & estimated like

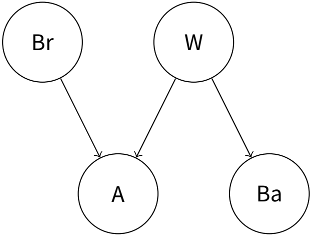
Brk.	Wind	Alarm	Bar.	P
yes	lo	on	lo	0.0243
yes	lo	on	med	0.0002
yes	lo	on	hi	0.0002
yes	lo	off	lo	0.0002
yes	lo	off	med	2.50e-06
yes	lo	off	hi	2.50e-06
yes	med	on	lo	0.0001
yes	med	on	med	0.0146
yes	med	on	hi	0.0001
yes	med	off	lo	1.50e-06
yes	med	off	med	0.0001
yes	med	off	hi	1.50e-06
yes	hi	on	lo	9.99e-05
yes	hi	on	med	9.99e-05
yes	hi	on	hi	0.0098
yes	hi	off	lo	1.00e-07
yes	hi	off	med	1.00e-07
yes	hi	off	hi	9.80e-06

Brk.	Wind	Alarm	Bar.	P
no	lo	on	lo	0.0047
no	lo	on	med	4.75e-05
no	lo	on	hi	4.75e-05
no	lo	off	lo	0.4608
no	lo	off	med	0.0047
no	lo	off	hi	0.0047
no	med	on	lo	0.0001
no	med	on	med	0.0140
no	med	on	hi	0.0001
no	med	off	lo	0.0027
no	med	off	med	0.2653
no	med	off	hi	0.0027
no	hi	on	lo	0.0005
no	hi	on	med	0.0005
no	hi	on	hi	0.0466
no	hi	off	lo	0.0014
no	hi	off	med	0.0014
no	hi	off	hi	0.1397

$$\mathrm{P}(\text{Break-in=yes}, \text{Alarm=on}) = 0.0496$$
$$\mathrm{P}(\text{Break-in=no}, \text{Alarm=on}) = 0.0665$$

$$\mathrm{P}(\text{Break-in=yes} \mid \text{Alarm=on}) = \frac{\mathrm{P}(\text{Break-in=yes}, \text{Alarm=on})}{\sum_b \mathrm{P}(\text{Break-in=}b, \text{Alarm=on})} = .427$$

Knowing the model structure (statistical dependencies), complicated models become manageable.



$$P(\mathrm{Br}, \mathrm{W}, \mathrm{A}, \mathrm{Ba})$$
$$= P(\mathrm{Br})\, P(\mathrm{W})\, P(\mathrm{A} \mid \mathrm{Br}, \mathrm{W})\, P(\mathrm{Ba} \mid \mathrm{W})$$

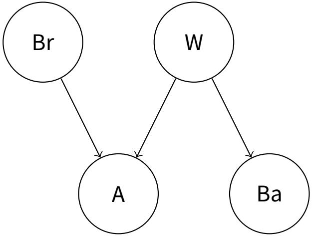
P(Br)	yes	no
	.05	.95

P(W)	lo	mid	hi
	.5	.3	.2

P(A | Br, W)		on	off
Br=yes	W=lo	.99	.01
Br=yes	W=med	.99	.01
Br=yes	W=hi	.999	.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.



$$P(Br, W, A, Ba) = P(Br)\,P(W)\,P(A \mid Br, W)\,P(Ba \mid W)$$

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

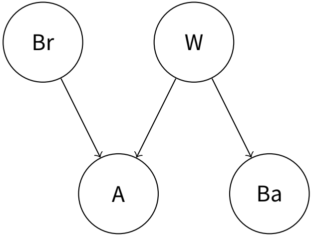
P(Br)	yes	no
	.05	.95

P(W)	lo	mid	hi
	.5	.3	.2

P(A | Br, W)		on	off
Br=yes	W=lo	.99	.01
Br=yes	W=med	.99	.01
Br=yes	W=hi	.999	.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.



$$P(\mathrm{Br}, \mathrm{W}, \mathrm{A}, \mathrm{Ba})$$
$$= P(\mathrm{Br})\, P(\mathrm{W})\, P(\mathrm{A} \mid \mathrm{Br}, \mathrm{W})\, P(\mathrm{Ba} \mid \mathrm{W})$$

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

P(Br)	yes	no
	.05	.95

P(W)	lo	mid	hi
	.5	.3	.2

P(A | Br, W)		on	off
Br=yes	W=lo	.99	.01
Br=yes	W=med	.99	.01
Br=yes	W=hi	.999	.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

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

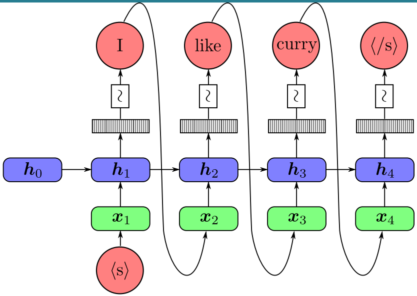
encode structure, not parametrization

are non-unique for a distribution

encode independence **requirements**, not necessarily all

BN are not neural net diagrams

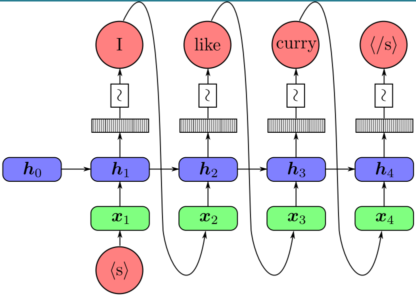
Recall the RNN language model:



- In statistical terms, what are we modeling?

BN are not neural net diagrams

Recall the RNN language model:

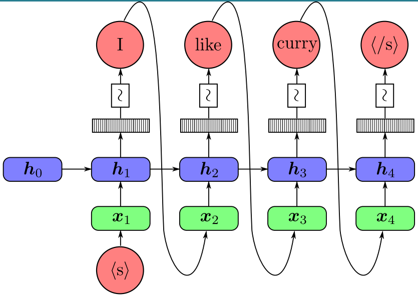


- In statistical terms, what are we modeling?

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

BN are not neural net diagrams

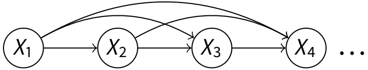
Recall the RNN language model:



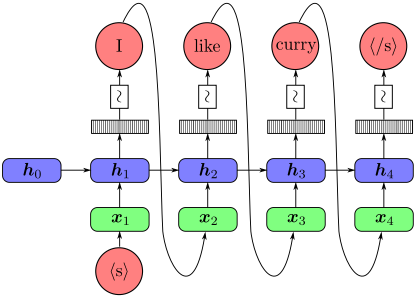
- In statistical terms, what are we modeling?

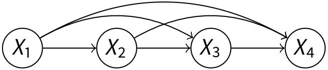
$$P(X_1, \ldots, X_n) = P(X_1)\, P(X_2 \mid X_1)\, P(X_3 \mid X_1, X_2) \ldots$$

- 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)

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, $Y \in \{\texttt{sports}, \texttt{music}, \ldots\}$.

BN for a generative model: $\text{Y} \longrightarrow \text{X}$

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

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, $Y \in \{\texttt{sports}, \texttt{music}, \ldots\}$.

BN for a generative model: $Y \rightarrow X$

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

$\mathrm{P}(Y)$: uniform: $\mathrm{P}(Y = \texttt{sports}) = \mathrm{P}(Y = \texttt{music}) = \frac{1}{|\mathcal{Y}|}$,

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, $Y \in \{\texttt{sports}, \texttt{music}, \ldots\}$.

BN for a generative model: $Y \longrightarrow X$

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

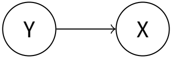
$\mathrm{P}(Y)$: uniform: $\mathrm{P}(Y = \texttt{sports}) = \mathrm{P}(Y = \texttt{music}) = \frac{1}{|\mathcal{Y}|}$, or estimated from data.

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, $Y \in \{\texttt{sports}, \texttt{music}, \ldots\}$.

BN for a generative model:



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

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

$\mathrm{P}(X \mid Y)$ (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, $Y \in \{\texttt{sports}, \texttt{music}, \ldots\}$.

BN for a generative model: $Y \rightarrow X$

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

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

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

Naive Bayes $\mathrm{P}(X \mid Y) = \prod_{j=1}^{L} \mathrm{P}(X_j \mid Y)$

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, $Y \in \{\texttt{sports}, \texttt{music}, \dots\}$.

BN for a generative model: $Y \rightarrow X$

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

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

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

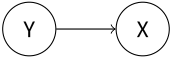
Naive Bayes $\mathrm{P}(X \mid Y) = \prod_{j=1}^{L} \mathrm{P}(X_j \mid Y)$

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

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, $Y \in \{\texttt{sports}, \texttt{music}, \ldots\}$.

BN for a generative model: 

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

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

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

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

BN encode structure, not parametrization

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

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

BN for a generative model: $Y \rightarrow X$

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

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

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

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

$\mathrm{P}(X \mid Y)$ 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, $Y \in \{\texttt{sports}, \texttt{music}, \ldots\}$.

BN for a generative model: $Y \rightarrow X$

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

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

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

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

$\mathrm{P}(X \mid Y)$ need not be parametrized as a table.

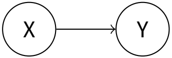
Variables need not be discrete! mixture of Gaussians: $\mathrm{P}(X \mid Y = y) \sim \mathcal{N}(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y)$.

Equivalent factorizations

There are many possible factorizations! $\mathrm{P}(X, Y) =$

Equivalent factorizations

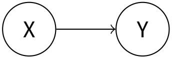
There are many possible factorizations! $\mathrm{P}(X, Y) =$



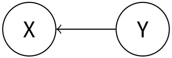
$\mathrm{P}(X)\,\mathrm{P}(Y \mid X)$

Equivalent factorizations

There are many possible factorizations! $\mathrm{P}(X, Y) =$



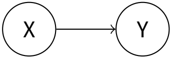
$\mathrm{P}(X)\,\mathrm{P}(Y \mid X)$



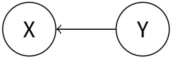
$\mathrm{P}(Y)\,\mathrm{P}(X \mid Y)$

Equivalent factorizations

There are many possible factorizations! $\mathrm{P}(X, Y) =$



$\mathrm{P}(X)\,\mathrm{P}(Y \mid X)$

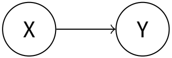


$\mathrm{P}(Y)\,\mathrm{P}(X \mid Y)$

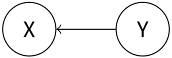
$\mathrm{P}(X)\,\mathrm{P}(Y)$

Equivalent factorizations

There are many possible factorizations! $\mathrm{P}(X, Y) =$



$\mathrm{P}(X)\,\mathrm{P}(Y \mid X)$



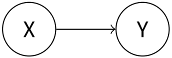
$\mathrm{P}(Y)\,\mathrm{P}(X \mid Y)$

$\mathrm{P}(X)\,\mathrm{P}(Y)$

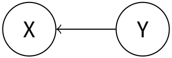
The first two are valid Bayes nets for **any** $\mathrm{P}(X, Y)$!

Equivalent factorizations

There are many possible factorizations! $\mathrm{P}(X, Y) =$



$\mathrm{P}(X)\,\mathrm{P}(Y \mid X)$



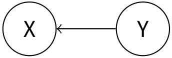
$\mathrm{P}(Y)\,\mathrm{P}(X \mid Y)$

$\mathrm{P}(X)\,\mathrm{P}(Y)$

The first two are valid Bayes nets for **any** $\mathrm{P}(X, Y)$!

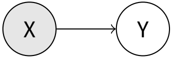
In fact, recall generative vs discriminative classifiers!

- Generative (e.g. naïve Bayes):



  To classify, we would compute $\mathrm{P}(Y \mid X)$ via Bayes' rule.

- Discriminative (e.g. logistic regression)



  in LR, we don't model $\mathrm{P}(X)$, we assume X is always observed (gray).

Equivalent factorizations

There are many possible factorizations! $\mathrm{P}(X, Y) =$

$X \rightarrow Y$	$X \leftarrow Y$	$X \quad Y$
$\mathrm{P}(X)\,\mathrm{P}(Y \mid X)$	$\mathrm{P}(Y)\,\mathrm{P}(X \mid Y)$	$\mathrm{P}(X)\,\mathrm{P}(Y)$

The first two are valid Bayes nets for **any** $\mathrm{P}(X, Y)$!

In fact, recall generative vs discriminative classifiers!

- Generative (e.g. naïve Bayes): $X \leftarrow Y$

  To classify, we would compute $\mathrm{P}(Y \mid X)$ via Bayes' rule.

- Discriminative (e.g. logistic regression) $X \rightarrow Y$

  in LR, we don't model $\mathrm{P}(X)$, we assume X is always observed (gray).

Some arrow direction choices are harder to estimate.

Some make more sense (why?): Barmtr. $\leftarrow$ Wind vs. Barmtr. $\rightarrow$ Wind

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): 

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:

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

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

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:

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**.

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): $X \qquad Y$

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

Example parametrization:

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

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

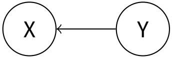
Does it mean we *must* have $X \not\perp\!\!\!\perp Y$?

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): 

Bayes net (2): 

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 $X \not\perp\!\!\!\perp Y$? **NO!**

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

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

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): X Y

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

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 $X \not\perp\!\!\!\perp Y$? **NO!**

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

$P(X \mid Y)$	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**.

Outline

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)

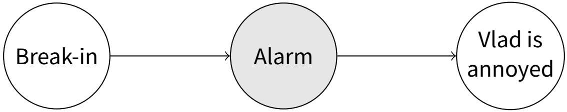
A is independent of B given a set of variables $\mathcal{C} = \{C_1, \ldots, C_n\}$, denoted as

$$A \perp\!\!\!\perp B \mid \mathcal{C},$$

if and only if

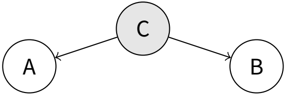
$$\mathrm{P}(A, B \mid C_1, \ldots, C_n) = \mathrm{P}(A \mid C_1, \ldots, C_n)\,\mathrm{P}(B \mid C_1, \ldots, C_n).$$

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



Three fundamental relationships in BN

The Fork

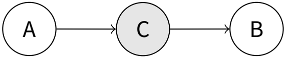


$A \perp\!\!\!\perp B \mid C$

Given C, A and B are independent.

Example: Alarm $\leftarrow$ Wind $\rightarrow$ Barometer

The Chain



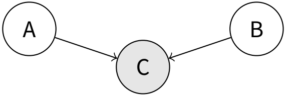
$A \perp\!\!\!\perp B \mid C$

After observing C,

further observing A would not tell us about B.

Example: Burglary $\rightarrow$ Alarm $\rightarrow$ Vlad distracted

The Collider



Surprisingly, $A \perp\!\!\!\perp B$

but **not** $A \perp\!\!\!\perp B \mid C$!

Example: Burglary $\rightarrow$ Alarm $\leftarrow$ 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 $\mathcal{C}$, implying:

$$A \perp\!\!\!\perp B \mid \mathcal{C}.$$

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:

$$X \leftarrow C \rightarrow Y \qquad (\text{with } C \in \mathcal{C})$$

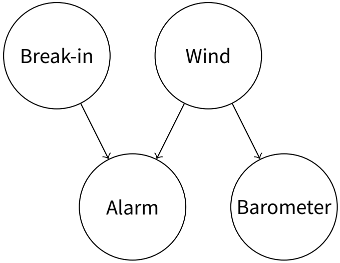
2. P includes a chain with observed middle:

$$X \rightarrow C \rightarrow Y \quad \text{or} \quad X \leftarrow C \leftarrow Y \qquad (\text{with } C \in \mathcal{C})$$

3. P includes a collider

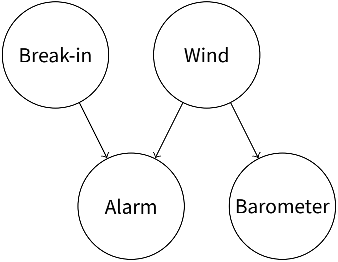
$$X \rightarrow U \leftarrow Y \qquad (\text{with } U \notin \mathcal{C})$$

Examples



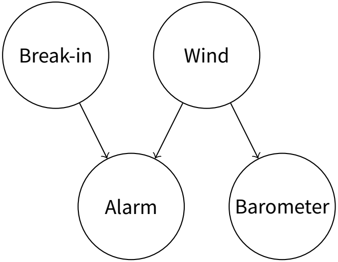
Wind $\perp\!\!\!\perp$ Barometer?

Examples



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

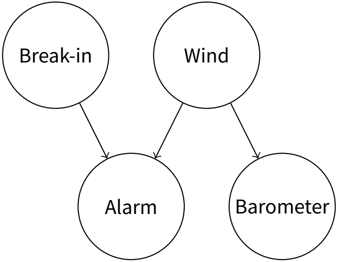
Examples



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

Break-in $\perp\!\!\!\perp$ Wind?

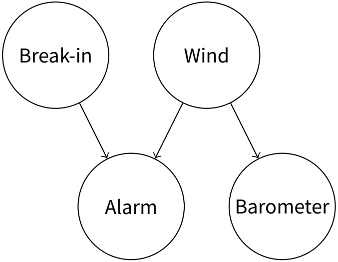
Examples



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

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

Examples

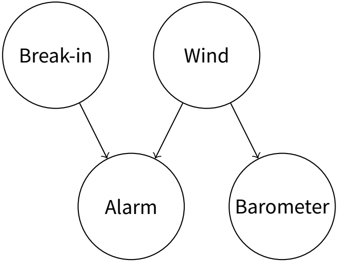


Wind $\perp\!\!\!\perp$ Barometer? **No**

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

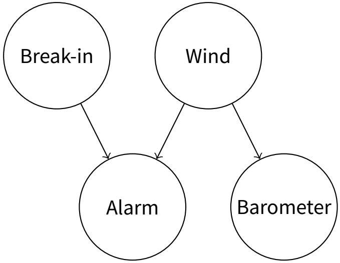
Break-in $\perp\!\!\!\perp$ Barometer?

Examples



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

Examples



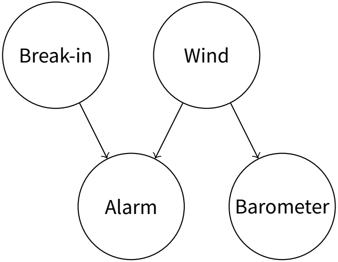
Wind $\perp\!\!\!\perp$ Barometer? **No**

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

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

Break-in $\perp\!\!\!\perp$ Barometer $\mid$ Alarm?

Examples



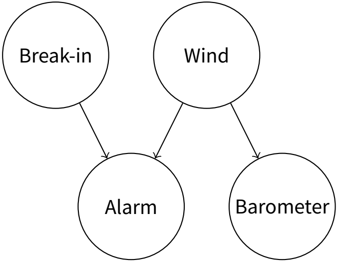
Wind $\perp\!\!\!\perp$ Barometer? **No**

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

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

Break-in $\perp\!\!\!\perp$ Barometer $\mid$ Alarm? **No**

Examples



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

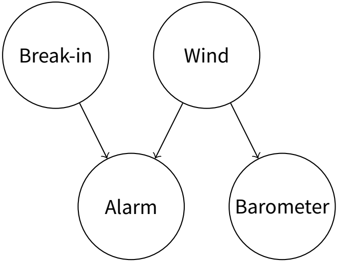
Break-in $\perp\!\!\!\perp$ Wind? **Yes**

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

Break-in $\perp\!\!\!\perp$ Barometer $\mid$ Alarm? **No**

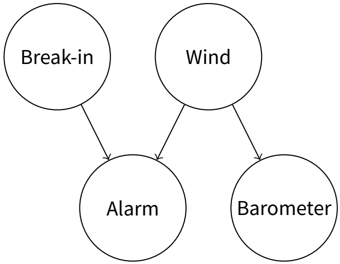
Break-in $\perp\!\!\!\perp$ Barometer $\mid$ Alarm, Wind?

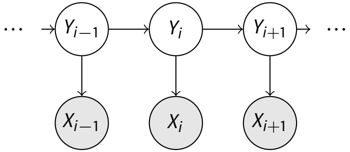
Examples



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

Examples





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

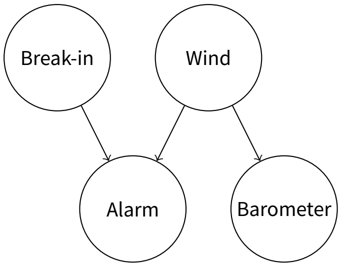
Break-in $\perp\!\!\!\perp$ Wind? **Yes**

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

Break-in $\perp\!\!\!\perp$ Barometer $\mid$ Alarm? **No**

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

Examples



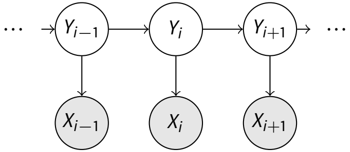
Wind $\perp\!\!\!\perp$ Barometer? **No**

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

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

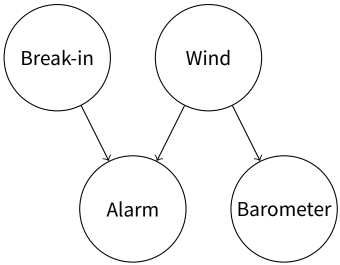
Break-in $\perp\!\!\!\perp$ Barometer $\mid$ Alarm? **No**

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



$Y_{i+1} \perp\!\!\!\perp Y_{i-1}$?

Examples



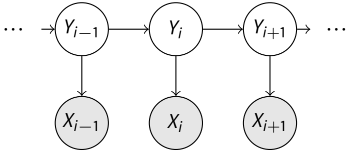
Wind $\perp\!\!\!\perp$ Barometer? **No**

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

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

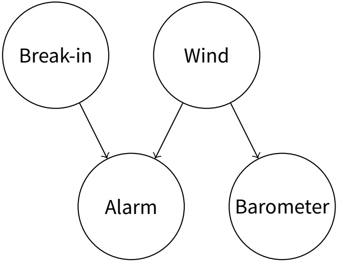
Break-in $\perp\!\!\!\perp$ Barometer $\mid$ Alarm? **No**

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

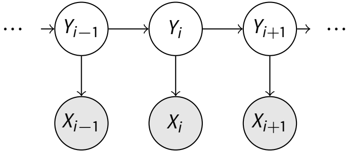


$Y_{i+1} \perp\!\!\!\perp Y_{i-1}$? **No**

Examples

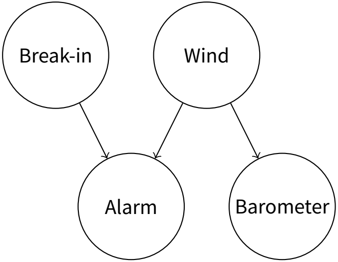


Wind $\perp\!\!\!\perp$ Barometer? **No**
Break-in $\perp\!\!\!\perp$ Wind? **Yes**
Break-in $\perp\!\!\!\perp$ Barometer? **Yes**
Break-in $\perp\!\!\!\perp$ Barometer $\mid$ Alarm? **No**
Break-in $\perp\!\!\!\perp$ Barometer $\mid$ Alarm, Wind? **Yes**



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

Examples



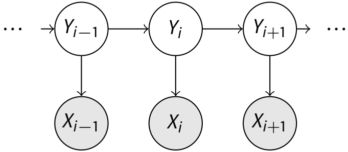
Wind $\perp\!\!\!\perp$ Barometer? **No**

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

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

Break-in $\perp\!\!\!\perp$ Barometer $\mid$ Alarm? **No**

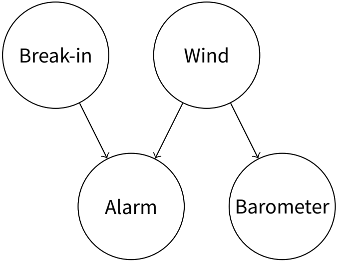
Break-in $\perp\!\!\!\perp$ Barometer $\mid$ Alarm, Wind? **Yes**



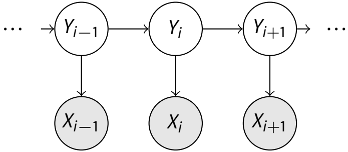
$Y_{i+1} \perp\!\!\!\perp Y_{i-1}$? **No**

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

Examples

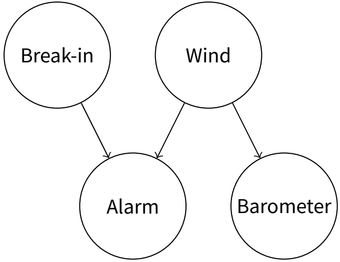


Wind $\perp\!\!\!\perp$ Barometer? **No**
Break-in $\perp\!\!\!\perp$ Wind? **Yes**
Break-in $\perp\!\!\!\perp$ Barometer? **Yes**
Break-in $\perp\!\!\!\perp$ Barometer $\mid$ Alarm? **No**
Break-in $\perp\!\!\!\perp$ Barometer $\mid$ Alarm, Wind? **Yes**

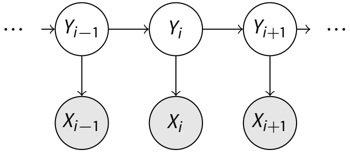


$Y_{i+1} \perp\!\!\!\perp Y_{i-1}$? **No**
$Y_{i+1} \perp\!\!\!\perp Y_{i-1} \mid Y_i$? **Yes**
$Y_{i+1} \perp\!\!\!\perp X_i$?

Examples

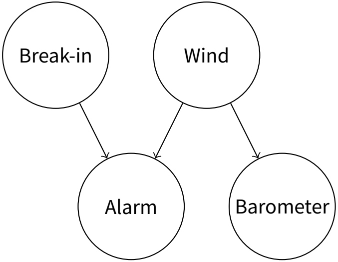


$$\text{Wind} \perp\!\!\!\perp \text{Barometer? } \textbf{No}$$
$$\text{Break-in} \perp\!\!\!\perp \text{Wind? } \textbf{Yes}$$
$$\text{Break-in} \perp\!\!\!\perp \text{Barometer? } \textbf{Yes}$$
$$\text{Break-in} \perp\!\!\!\perp \text{Barometer} \mid \text{Alarm? } \textbf{No}$$
$$\text{Break-in}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

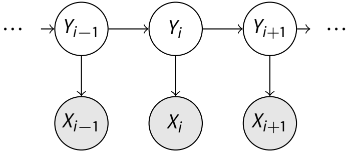


$$Y_{i+1} \perp\!\!\!\perp Y_{i-1}? \ \textbf{No}$$
$$Y_{i+1} \perp\!\!\!\perp Y_{i-1} \mid Y_i? \ \textbf{Yes}$$
$$Y_{i+1} \perp\!\!\!\perp X_i? \ \textbf{No}$$

Examples

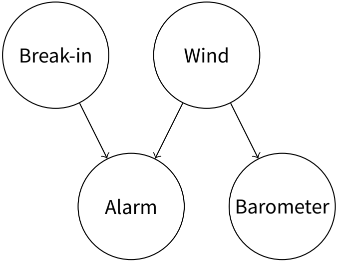


Wind $\perp\!\!\!\perp$ Barometer? **No**
Break-in $\perp\!\!\!\perp$ Wind? **Yes**
Break-in $\perp\!\!\!\perp$ Barometer? **Yes**
Break-in $\perp\!\!\!\perp$ Barometer $\mid$ Alarm? **No**
Break-in $\perp\!\!\!\perp$ Barometer $\mid$ Alarm, Wind? **Yes**

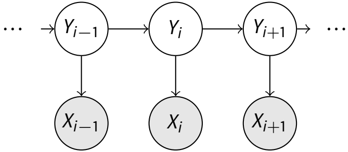


$Y_{i+1} \perp\!\!\!\perp Y_{i-1}$? **No**
$Y_{i+1} \perp\!\!\!\perp Y_{i-1} \mid Y_i$? **Yes**
$Y_{i+1} \perp\!\!\!\perp X_i$? **No**
$Y_{i+1} \perp\!\!\!\perp X_i \mid Y_i$?

Examples



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



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

Generative stories and plate notation

In papers, you'll see statistical models defined through *generative stories*:

$$\begin{aligned} \mu &\sim \text{Uniform}([-1,1]) \\ \sigma &\sim \text{Uniform}([1,2]) \\ X \mid \mu, \sigma &\sim \text{Normal}(\mu, \sigma) \end{aligned}$$

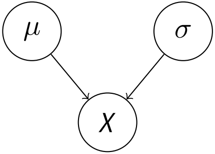
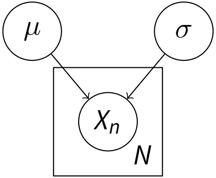


Plate notation is a way to denote **repetition of templates**:

$$\begin{aligned} \mu &\sim \text{Uniform}([-1,1]) \\ \sigma &\sim \text{Uniform}([1,2]) \\ X_n \mid \mu, \sigma &\sim \text{Normal}(\mu, \sigma) \quad i = 1, \ldots, N \end{aligned}$$



Outline

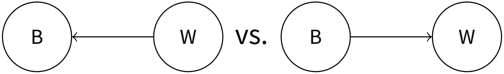
1. **Directed Models**
   - Bayes networks
   - Conditional independence and D-separation
   - Causal graphs & the *do* operator
2. **Undirected Models**
   - Markov networks
   - Factor graphs

Correlation does not imply causation;
but then, *what does?*

Seeing versus doing

Bayes nets only model independence assumptions.

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



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

$P(W \mid B)$	lo	mid	hi
$B = \text{lo}$	.98	.01	.01
$B = \text{mid}$	.01	.98	.01
$B = \text{hi}$	.01	.01	.98

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

We write: $P\left(W \mid \text{do}(B = \text{hi})\right) = ?$

Outline

Outline

Outline

1. **Directed Models**
   - Bayes networks
   - Conditional independence and D-separation
   - Causal graphs & the *do* operator
2. **Undirected Models**
   - Markov networks
   - Factor graphs