DAT405/DIT407 – Part 2: Statistical methods in Data Science and AI

DAT405, DIT 407 LP2 2022, Module 4.

CHALMERS
UNIVERSITY OF TECHNOLOGY

Module 4.2: Graphical models

Today's topics:

- Joint probability distributions
- Graphical models
- Chain rule
- Bayesian networks
- Sampling of Bayesian networks
- MCMC and Gibbs sampling
- Naïve Bayes

Join probability distributions

- Consider a t-shirt shop.
- The shirts come in two colors: $\{\text{blue}, \text{white}\}$
- In three sizes: $\{\text{S}, \text{M}, \text{L}\}$
- Which t-shirt is most probably sold next?

color	Size	Sail frequency
BLUE	S	0.1
BLUE	L	0.2
BLUE	M	0.2
WHITE	S	0.1
WHITE	L	0.1
WHITE	M	0.3

Join probability distributions

- Define two random variables
  - $X \in \{\text{blue}, \text{white}\}$
  - $Y \in \{\text{S}, \text{M}, \text{L}\}$
- The joint distribution can be represented in a table
- If $X \in \{x_1, \dots, x_n\}$ and $Y \in \{y_1, \dots, y_m\}$ the size of the table is $n \cdot m$.
- Becomes large fast!

	S	L	M
BLUE	0.1	0.2	0.2
WHITE	0.1	0.1	0.3

Joint probability distributions can be represented more compactly using graphical models

Conditional probability distributions

X	Y	$\mathbb{P}(X, Y)$
TRUE	TRUE	0.16
FALSE	TRUE	0.24
TRUE	FALSE	0.12
FALSE	FALSE	0.48

Joint distribution

X	Y	$P(X\|Y)$
TRUE	TRUE	0.4
FALSE	TRUE	0.6
TRUE	FALSE	0.2
FALSE	FALSE	0.8

Conditional distribution

Y	$\mathbb{P}(Y)$
TRUE	0.4
FALSE	0.6

Marginal distribution Y

Conditional probability distributions

- **Consider a conditional distribution**

$$\mathbb{P}(\boldsymbol{X} = \boldsymbol{x} | \boldsymbol{Y} = \boldsymbol{y})$$

$\boldsymbol{Y}$	$\mathbb{P}(\boldsymbol{X} = \mathbf{TRUE}\,|\boldsymbol{Y})$
TRUE	0.4
FALSE	0.2

$\mathbb{P}(X = \text{TRUE}|\ Y = \text{TRUE})$

$\mathbb{P}(X = \text{TRUE}|\ Y = \text{FALSE})$

Does not have to sum to 1:
$\mathbb{P}(X) = \mathbb{P}(X|Y)\mathbb{P}(Y) + \mathbb{P}(X|Y^c)\mathbb{P}(Y^c)$

Independent events

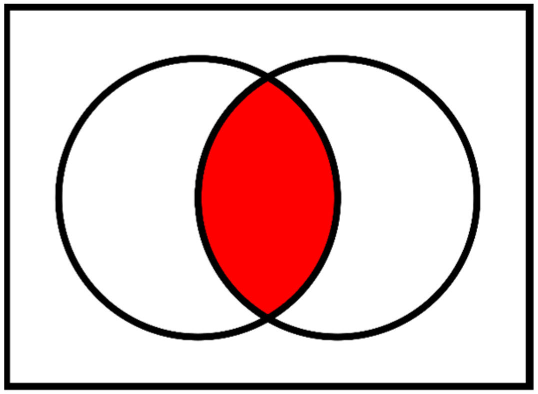
- Two events A and B are ***independent*** if information about one does not affect the probability of the other.
- Definition: A and B are ***independent*** if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$$

- Consequently, if A and B are independent

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A)\mathbb{P}(B)}{\mathbb{P}(B)} = \mathbb{P}(A)$$

- Note: mutually exclusive $\neq$ inependence

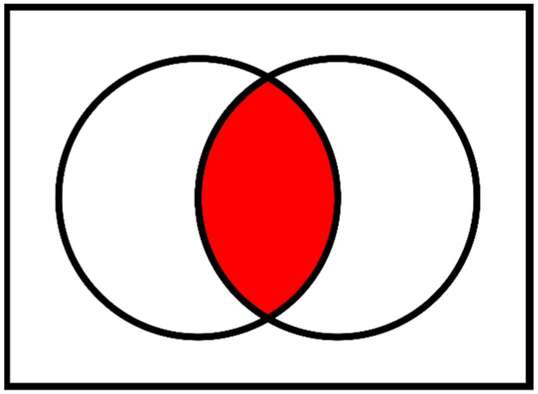


Thus, if $\mathbb{P}(A \cap B) \neq \mathbb{P}(A)\mathbb{P}(B)$, then A and B are dependent.

Independent random variables

- A set of random variables $X_1, X_2, \dots, X_n$ are ***independent*** if

$$\mathbb{P}(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) \\ = \mathbb{P}(X_1 = x_1) \cdots \mathbb{P}(X_n = x_n)$$

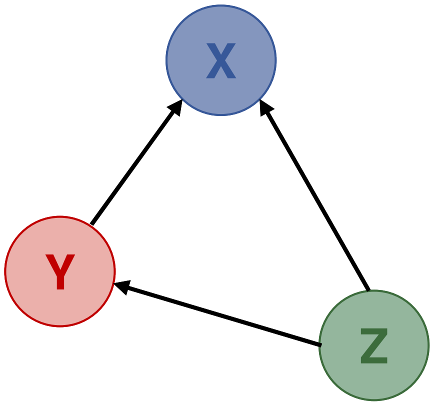


Chain rule

- Random variables: $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}$
- Chain rule

$$\mathbb{P}(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}) = \mathbb{P}(\boldsymbol{X}|\boldsymbol{Y}, \boldsymbol{Z})\mathbb{P}(\boldsymbol{Y}, \boldsymbol{Z}) = \mathbb{P}(\boldsymbol{X}|\boldsymbol{Y}, \boldsymbol{Z})\mathbb{P}(\boldsymbol{Y}|\boldsymbol{Z})\mathbb{P}(\boldsymbol{Z})$$

- The factorization can be represented by a graph



Chain rule

- **In general, for any random variables $\boldsymbol{X_1, X_2, \dots, X_n}$**

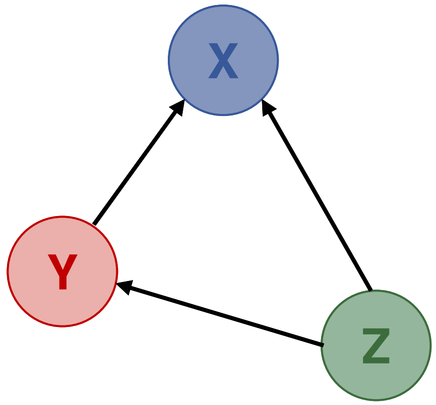
$$\mathbb{P}\,(\boldsymbol{X_1, X_2, \dots, X_n}) =$$
$$= \mathbb{P}(\boldsymbol{X_1})\mathbb{P}(\boldsymbol{X_2}|\boldsymbol{X_1})\mathbb{P}(\boldsymbol{X_3}|\boldsymbol{X_1, X_2}) \cdots \mathbb{P}(\boldsymbol{X_n}|\boldsymbol{X_1, \dots, X_{n-1}})$$

Note: if X_3 only depends on X_2 then $\mathbb{P}(X_3|X_2, X_1) = P(X_3|X_2)$

This factor requires a table withwith 2^{n-1} rows: one for each value combination of $X_1, \dots, X_n$. If $n = 40$ we need a TB-sized memory to store.

Note that we can choose any ordering of the n! possible

Chain rule

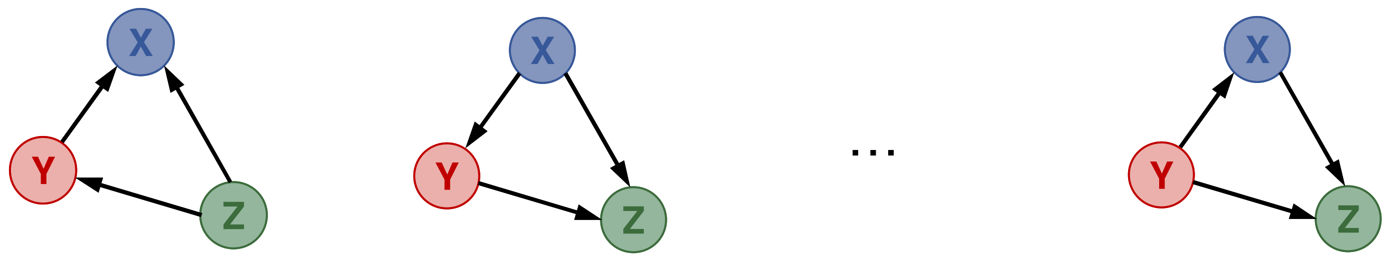


- Note that the factorization is not unique

$$\mathbb{P}(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}) = \mathbb{P}(\boldsymbol{X}|\boldsymbol{Y}, \boldsymbol{Z})\mathbb{P}(\boldsymbol{Y}|\boldsymbol{Z})\mathbb{P}(\boldsymbol{Z}) =$$
$$= \mathbb{P}(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z})\mathbb{P}(\boldsymbol{X}|\boldsymbol{Z})\mathbb{P}(\boldsymbol{Z})$$

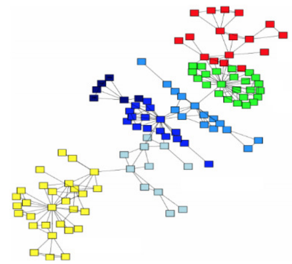
...

In total n! = 6 permutations and different graph representations.

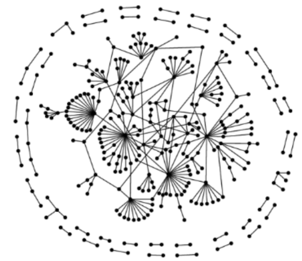


Graphical models

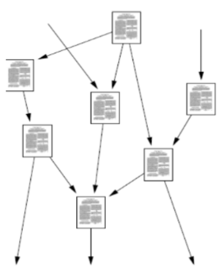
Social networks



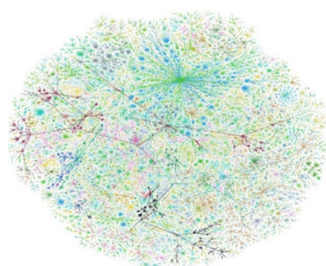
Economic networks



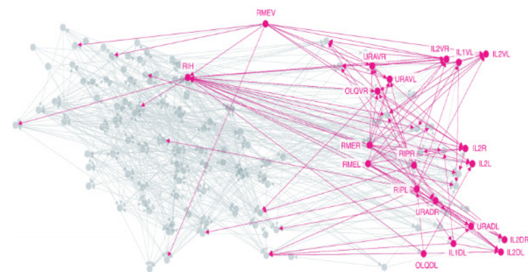
Biomedical networks



Information networks



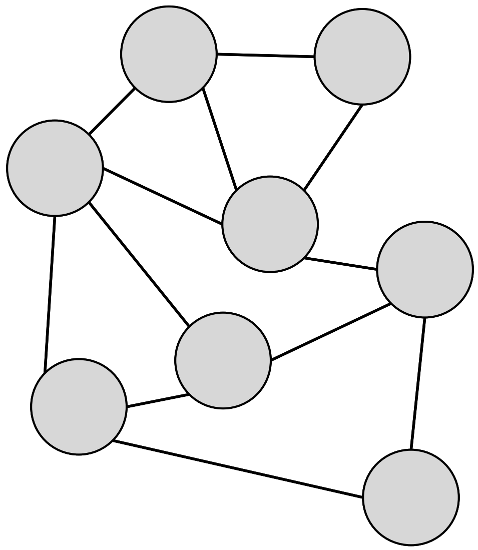
Network of neurons



Internet

Graphical models

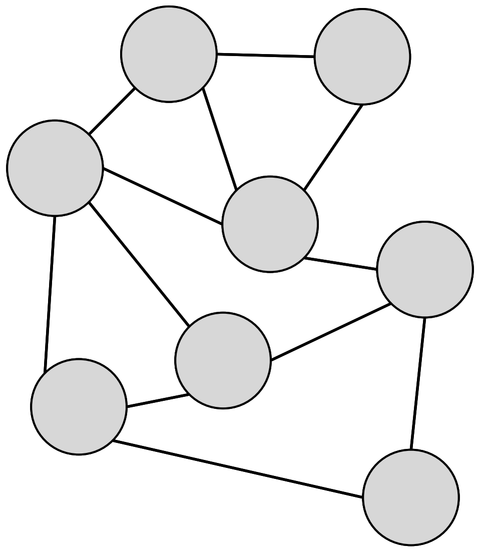
- **Diagrammatic representations of various connections and dependencies**
- **Informative visualization of the structure**
- **Efficient computer algorithms acting directly on the graph model**



Graphical models

Three main objectives:

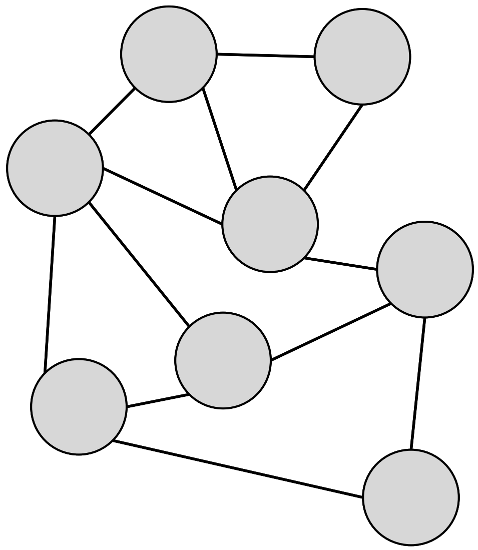
- **Representation**
  - **model structure**
- **Inference**
  - **queries to ask using model**
- **Learning**
  - **fit model to observed data**



Graphical models: some basics

A *simple graph* $G = (V, E)$ consists of

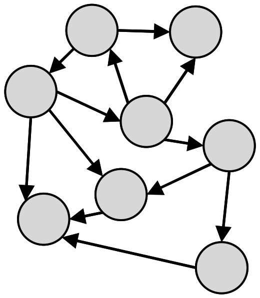
- A set V of *vertices* or *nodes*
- A set E of *edges* or *links*

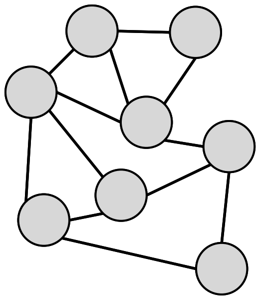


Graphical models: some basics

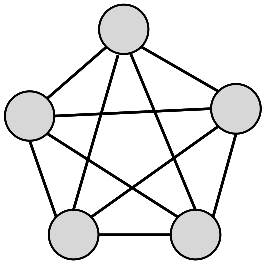
The graph can be

- ***directed*** **or**
- ***undirected***



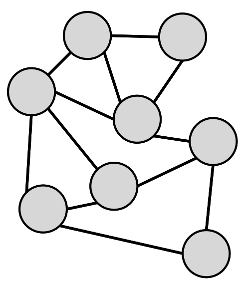


A *complete graph* has a connection between every pair of vertices

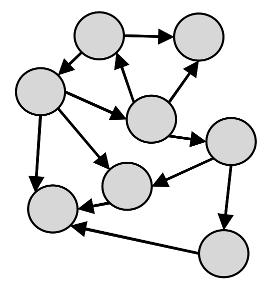


Graphical models: some basics

Undirected
- Joint probability distribution
- Links without arrows
- Indicating relationships (correlation)

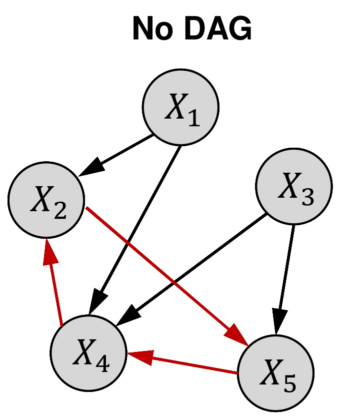


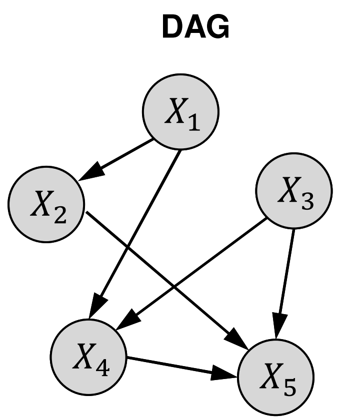
Directed
- Conditional prob distribution
- Directional links (with arrows)
- Indicating conditional dependence



Directed acyclic graphs (DAGs)

- Directed edges
- Contains no cycles/loops.
- Topological ordering of nodes





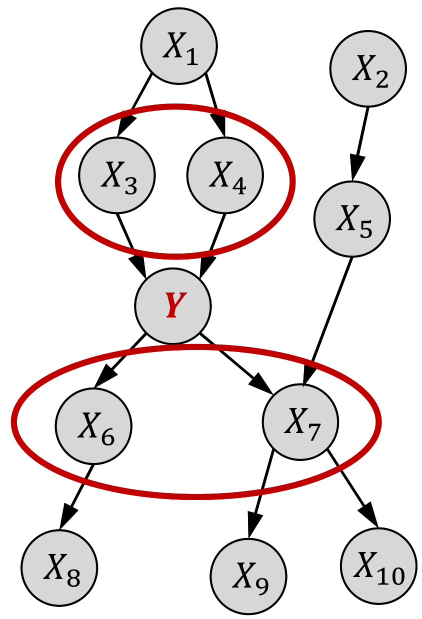
Directed acyclic graphs (DAGs)

- The ***parents*** of a node are the nodes with links into it.

$$\mathrm{pa}(Y) = \{X_3, X_4\}$$

- The ***children*** of a node are the nodes with links to them from that node.

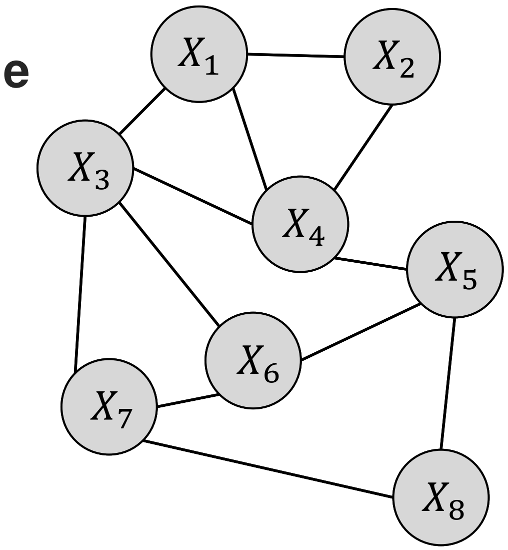
$$\mathrm{ch}(Y) = \{X_6, X_7\}$$



Probabilistic graphical models

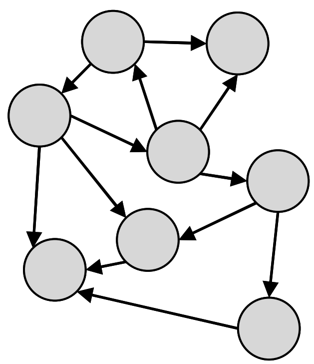
A graph that represents the joint distribution of the random variables

- Vertices: random variables
- Edges: probabilistic relationships

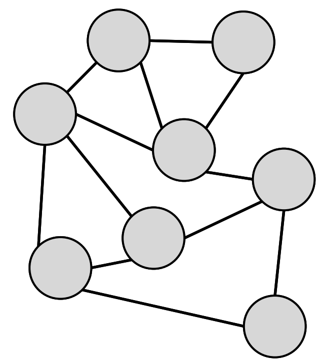


Examples of probabilistic graphical models

Directed
- Naïve Bayes
- Bayesian networks
- Markov chains
- Neural networks



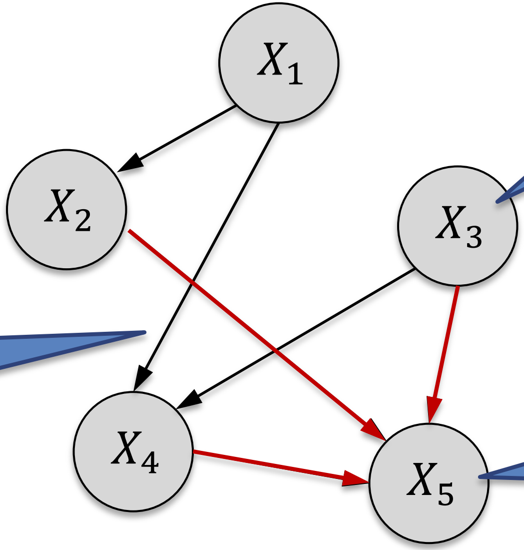
Undirected
- Markov random fields
- Conditional random fields



Intuitive interpretation of the graph

The graph represents the joint probability of all the random variables in it.

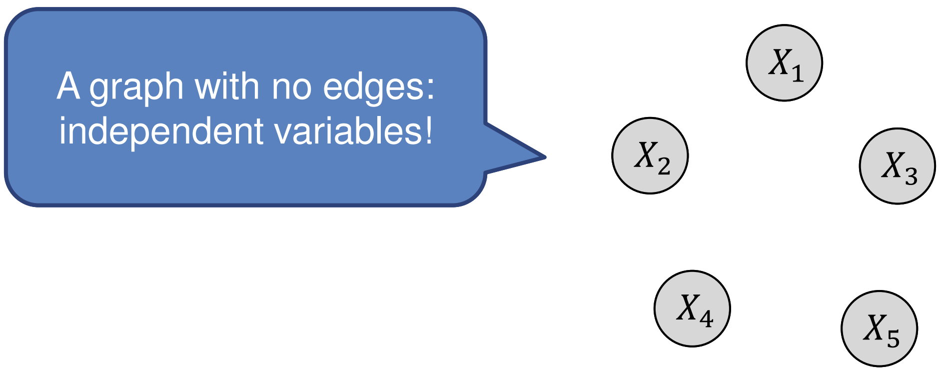
The edges give which variables depend on which.



Each node corresponds to a conditional probability, of that variable given the others.

X_5 depends on X_2, X_3 and X_4 but not on X_1, given X_2, X_3 and X_4

Example 1: chain rule

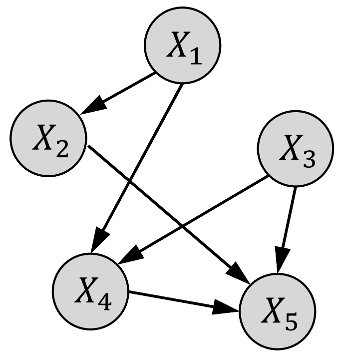


Chain rule:

$$\mathbb{P}\,(\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3, \boldsymbol{X}_4, \boldsymbol{X}_5) = \mathbb{P}(\boldsymbol{X}_1)\mathbb{P}(\boldsymbol{X}_2)\mathbb{P}(\boldsymbol{X}_3)\mathbb{P}(\boldsymbol{X}_4)\mathbb{P}(\boldsymbol{X}_5)$$

CHALMERS
UNIVERSITY OF TECHNOLOGY

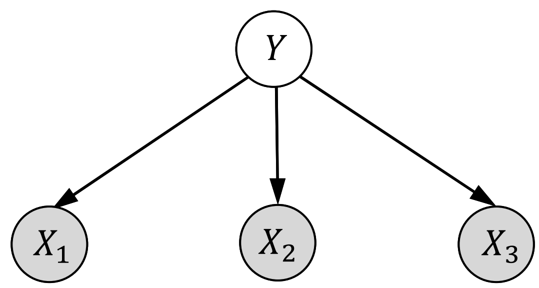
Example 2: chain rule



Chain rule:

$$\mathbb{P}\,(\boldsymbol{X_1}, \boldsymbol{X_2}, \boldsymbol{X_3}, \boldsymbol{X_4}, \boldsymbol{X_5}) = \mathbb{P}(\boldsymbol{X_1})\mathbb{P}(\boldsymbol{X_2}|\boldsymbol{X_1})\mathbb{P}(\boldsymbol{X_3})\mathbb{P}(\boldsymbol{X_4}|\boldsymbol{X_1}, \boldsymbol{X_3})\mathbb{P}(\boldsymbol{X_5}|\boldsymbol{X_2}, \boldsymbol{X_3}, \boldsymbol{X_4})$$

Example 3: chain rule



Chain rule:

$$\mathbb{P}\,(\boldsymbol{Y}, \boldsymbol{X_1}, \boldsymbol{X_2}, \boldsymbol{X_3}) = \mathbb{P}(\boldsymbol{Y})\mathbb{P}(\boldsymbol{X_1}|\boldsymbol{Y})\mathbb{P}(\boldsymbol{X_2}|\boldsymbol{Y})\mathbb{P}(\boldsymbol{X_3}|\boldsymbol{Y})$$

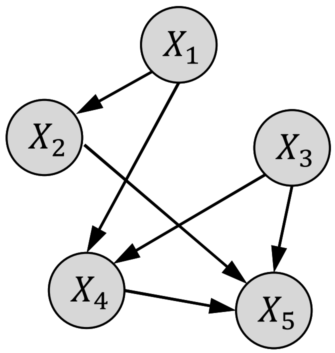
Chain rule for DAGs

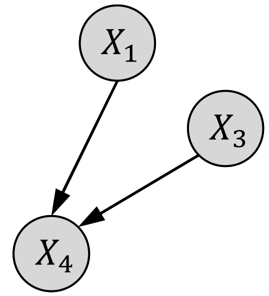
- Can deduce probabilistic model *from* graph

  $$\mathbb{P}(\boldsymbol{X_1}, \boldsymbol{X_2}, \dots, \boldsymbol{X_5}) = \mathbb{P}(\boldsymbol{X_1})\mathbb{P}(\boldsymbol{X_3})\mathbb{P}(\boldsymbol{X_2}|\boldsymbol{X_1})\mathbb{P}(\boldsymbol{X_4}|\boldsymbol{X_1}, \boldsymbol{X_3})\mathbb{P}(\boldsymbol{X_5}|\boldsymbol{X_2}, \boldsymbol{X_3}, \boldsymbol{X_4})$$

- A link going from $X_1 \rightarrow X_2$ means that X_1 is a ***parent node*** of X_2.
- The probability of each node X_i is conditioned only on its parents $\mathrm{pa}(X_i)$

$$\mathbb{P}\,(\boldsymbol{X_1}, \dots, \boldsymbol{X_n}) = \prod_{i=1}^{n} \mathbb{P}(\boldsymbol{X_i}|\mathrm{pa}(\boldsymbol{X_i}))$$

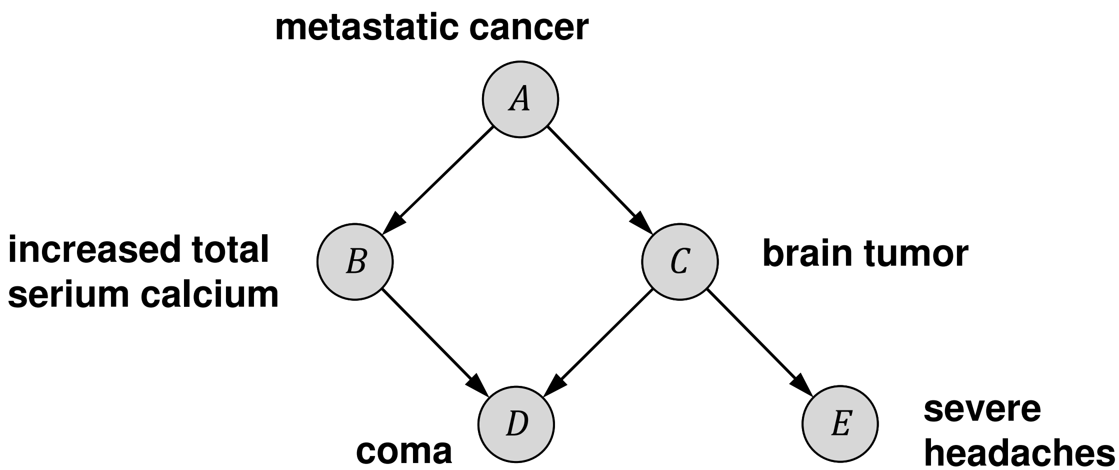




$\mathrm{pa}(\boldsymbol{X_4}) = \{\boldsymbol{X_1}, \boldsymbol{X_3}\}$

Example of a Bayesian network

– Diagnosis of cancer

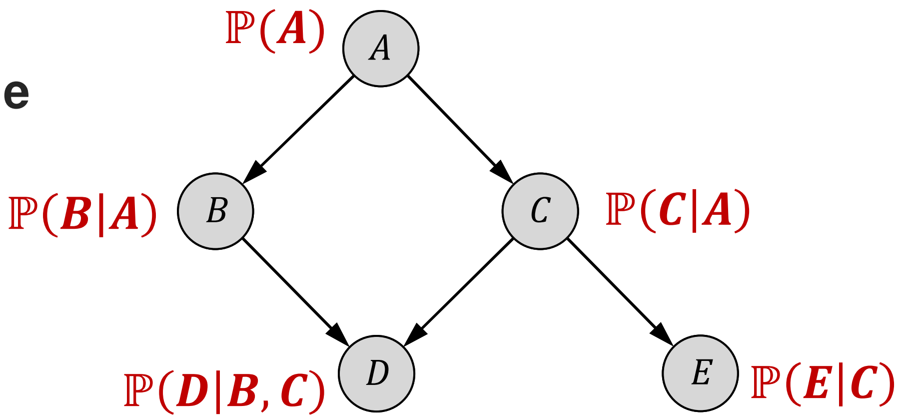


Example of a Bayesian network

The entire network represents the joint probability

$$\mathbb{P}(A, B, C, D, E)$$

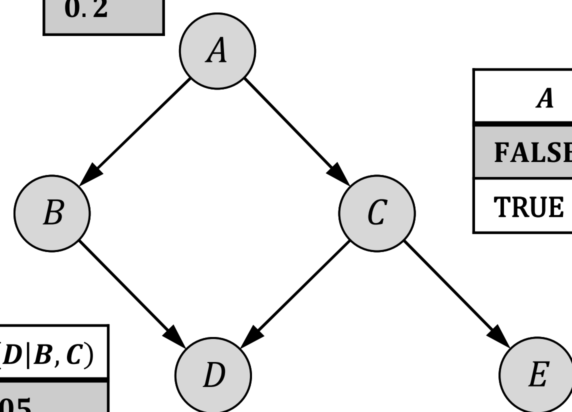
This can be factorized according to the dependencies given by the edges



$$\mathbb{P}(A, B, C, D, E) = \mathbb{P}(\boldsymbol{A})\,\mathbb{P}(\boldsymbol{B}|\boldsymbol{A})\mathbb{P}(\boldsymbol{C}|\boldsymbol{A})\mathbb{P}(\boldsymbol{D}|\boldsymbol{B},\boldsymbol{C})\mathbb{P}(\boldsymbol{E}|\boldsymbol{C})$$

Example of a Bayesian network

$\mathbb{P}(A)$
0.2



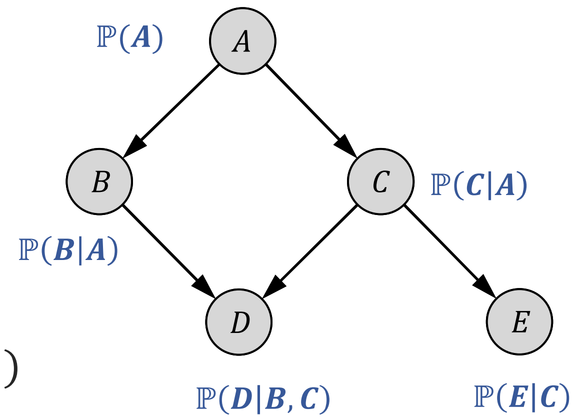
A	$\mathbb{P}(B\|A)$
FALSE	0.2
TRUE	0.8

A	$\mathbb{P}(C\|A)$
FALSE	0.05
TRUE	0.2

B	C	$\mathbb{P}(D\|B,C)$
FALSE	FALSE	0.05
FALSE	TRUE	0.8
TRUE	FALSE	0.8
TRUE	TRUE	0.8

C	$\mathbb{P}(C\|E)$
FALSE	0.6
TRUE	0.8

Example of a Bayesian network

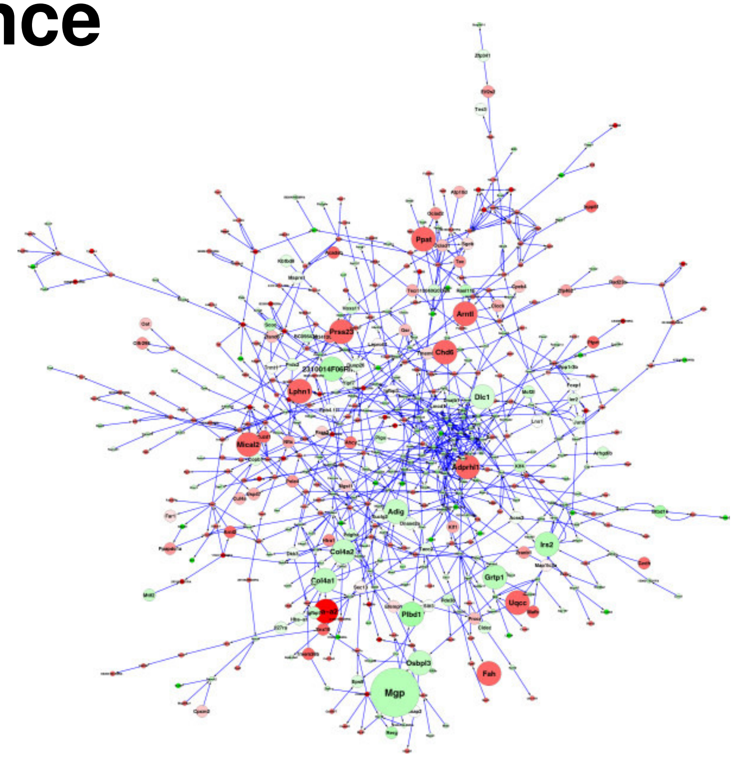


Now we can compute the joint probability for any combination of interest

$$
\begin{aligned}
&\mathbb{P}(\boldsymbol{A}^+,\boldsymbol{B}^-,\boldsymbol{C}^+,\boldsymbol{D}^-,\boldsymbol{E}^+) = \\
&= \mathbb{P}(\boldsymbol{A}^+)\mathbb{P}(\boldsymbol{B}^-|\boldsymbol{A}^+)\mathbb{P}(\boldsymbol{C}^+|\boldsymbol{A}^-)\mathbb{P}(\boldsymbol{D}^-|\boldsymbol{B}^-,\boldsymbol{C}^+)\mathbb{P}(\boldsymbol{E}^+|\boldsymbol{C}^+) \\
&= \mathbb{P}(\boldsymbol{A}^+)\big(\mathbf{1} - \mathbb{P}(\boldsymbol{B}^+|\boldsymbol{A}^+)\big)\mathbb{P}(\boldsymbol{C}^+|\boldsymbol{A}^-)\big(\mathbf{1} - \mathbb{P}(\boldsymbol{D}^+|\boldsymbol{B}^-,\boldsymbol{C}^+)\big)\mathbb{P}(\boldsymbol{E}^+|\boldsymbol{C}^+) \\
&= \cdots = \mathbf{0.00128}
\end{aligned}
$$

However: this needs to be put in relation to all other value combinations ($2^5 = 32$ joint probabilities)…

Bayesian network inference



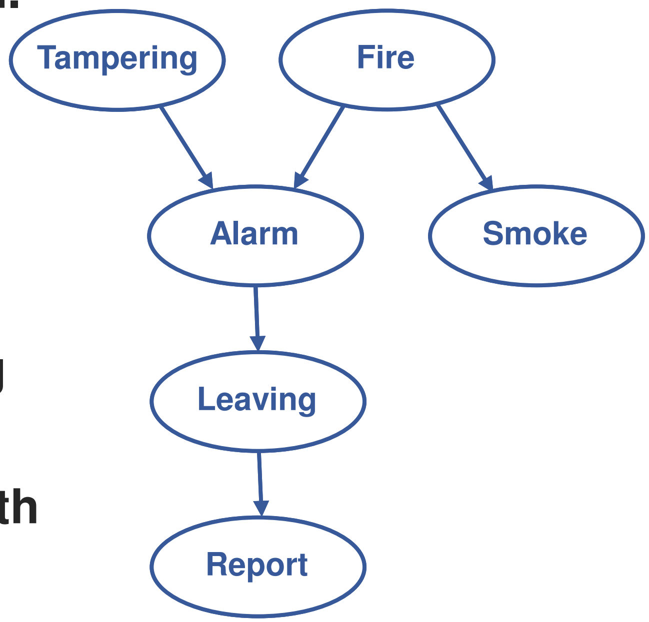
Bayesian network inference

We want to diagnose the true cause of a fire alarm.

Variabels (all true/false):

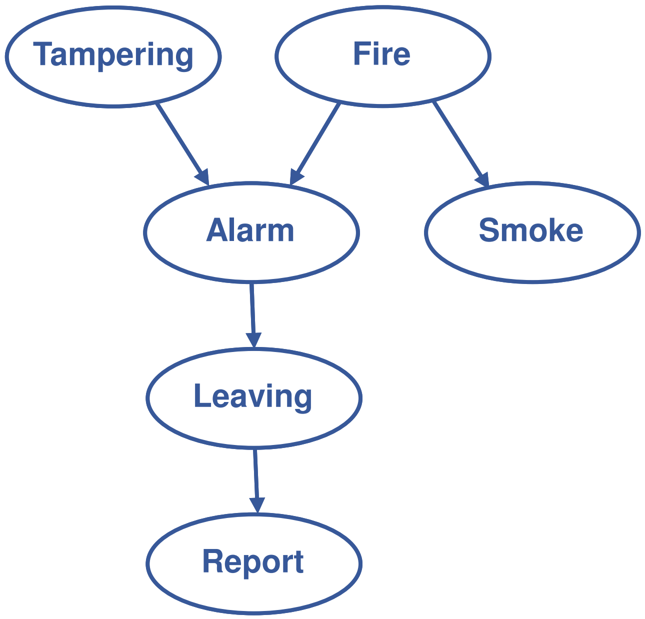
- *Fire*: true when there is a fire
- *Alarm*: true when the alarm sounds
- *Smoke*: true when there is smoke
- *Leaving*: true if many people leave the building
- *Report*: true if reports of people leaving
- *Tampering*: true when alarm were tampered with

Conditional dependencies are given by the DAG.



Bayesian network inference

- Exact inference: variable elimination
- Approximate inference: sampling

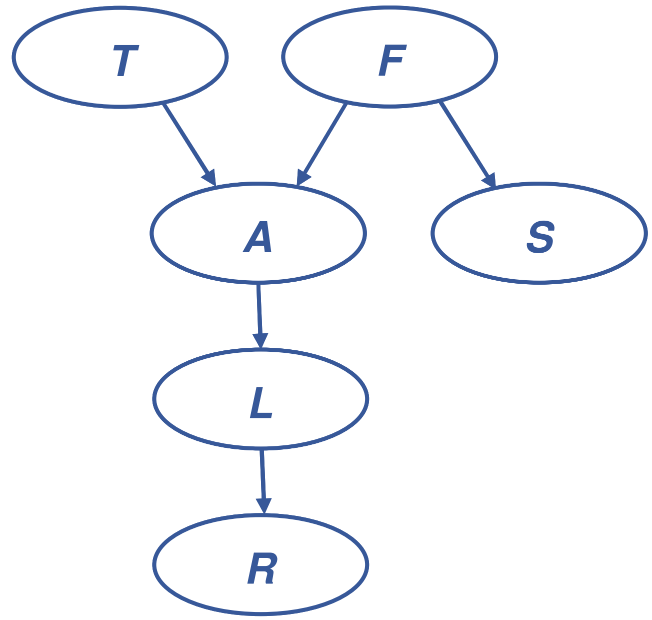


Sampling in Bayesian networks: example

Factorization

$\mathbb{P}(\boldsymbol{T},\boldsymbol{F},\boldsymbol{A},\boldsymbol{S},\boldsymbol{L},\boldsymbol{R}) =$

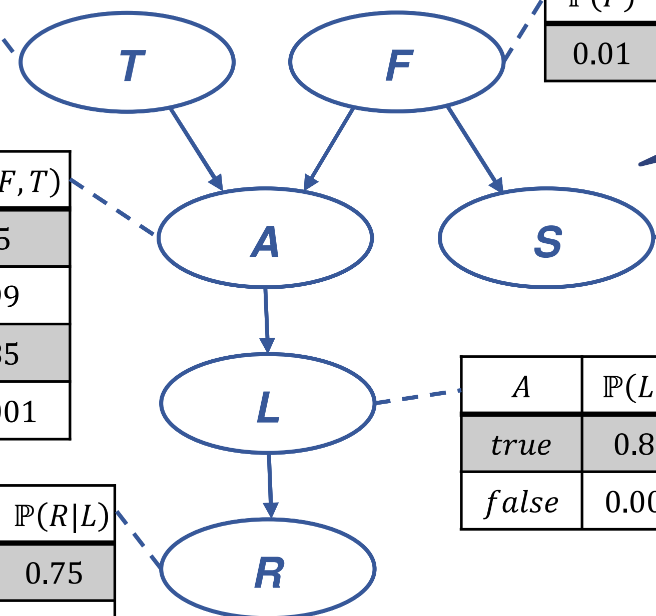
$= \mathbb{P}(\boldsymbol{T})\mathbb{P}(\boldsymbol{F})\mathbb{P}(\boldsymbol{A}|\boldsymbol{T},\boldsymbol{F})\mathbb{P}(\boldsymbol{S},\boldsymbol{F})\mathbb{P}(\boldsymbol{L}|\boldsymbol{A})\mathbb{P}(\boldsymbol{R}|\boldsymbol{L})$



Exact inference

Suppose we want to compute $P(R^+, S^-)$.

We could compute all parts of the factorization and combine as before.



$\mathbb{P}(T)$
0.02

$\mathbb{P}(F)$
0.01

F	T	$\mathbb{P}(A \mid F, T)$
$true$	$true$	0.5
$true$	$false$	0.99
$false$	$true$	0.85
$false$	$false$	0.0001

F	$\mathbb{P}(S \mid F)$
$true$	0.9
$false$	0.01

A	$\mathbb{P}(L \mid A)$
$true$	0.88
$false$	0.001

L	$\mathbb{P}(R \mid L)$
$true$	0.75
$false$	0.01

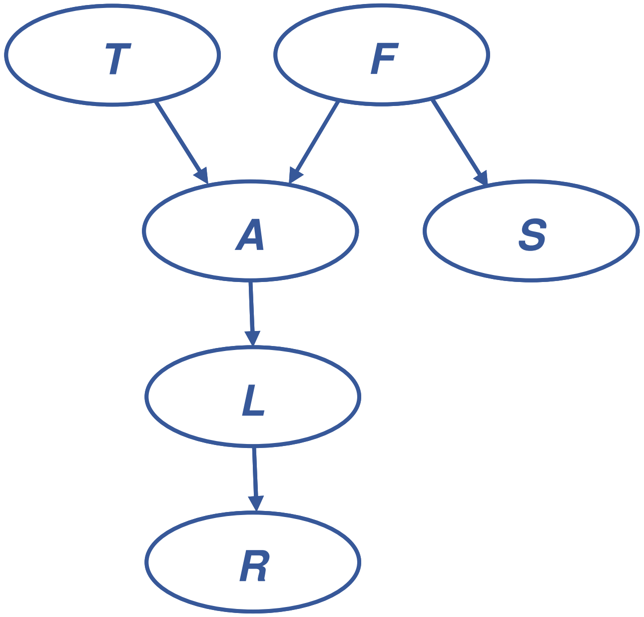
Exact inference

$$\mathbb{P}(R^+, S^-) =$$

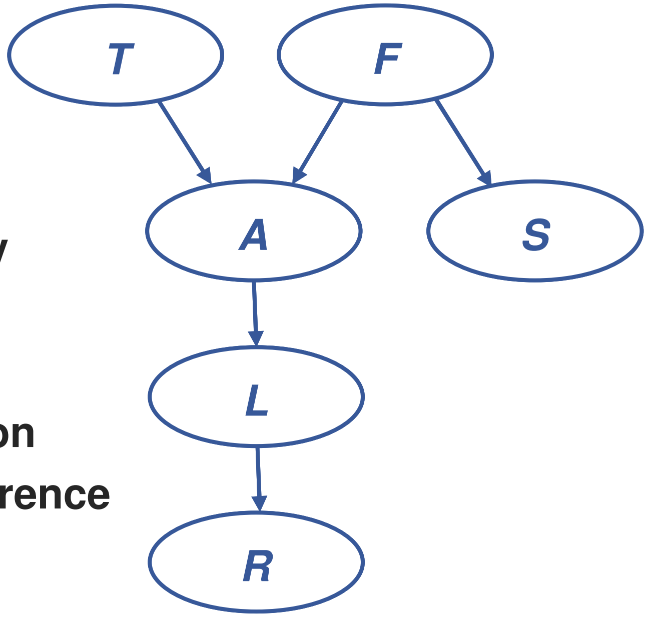
$$= \sum_L \sum_A \sum_T \sum_F \mathbb{P}(R^+, S^- | L, A, T, F)\, \mathbb{P}(L, A, T, F)$$

$$= \sum_{L,A,T,F} \mathbb{P}(R^+|L)\mathbb{P}(L|A)\mathbb{P}(A|T,F)\mathbb{P}(T)\mathbb{P}(F)\mathbb{P}(S^-|F)$$

Very time consuming. Approximation by sampling scales better.



Sampling in Bayesian networks



Basic idea

- Draw *N* samples from a sampling distribution
- Estimate the posterior probability
- Show that this converges to the desired probability

Why sampling?

- Learning: get samples from an unknown distribution
- Inference: faster and more scalable than exact inference

Sampling – basic idea

Sampling from a given distribution:

Step 1: Split the interval $[0, 1)$ into subintervals proportional to the desired sampling distribution

Step 2: Sample from the uniform distribution $U[0, 1)$
(i.e. a random number $0 \leq u < 1$)

Step 3: Associate u with the corresponding subinterval

Repeat *N* times.

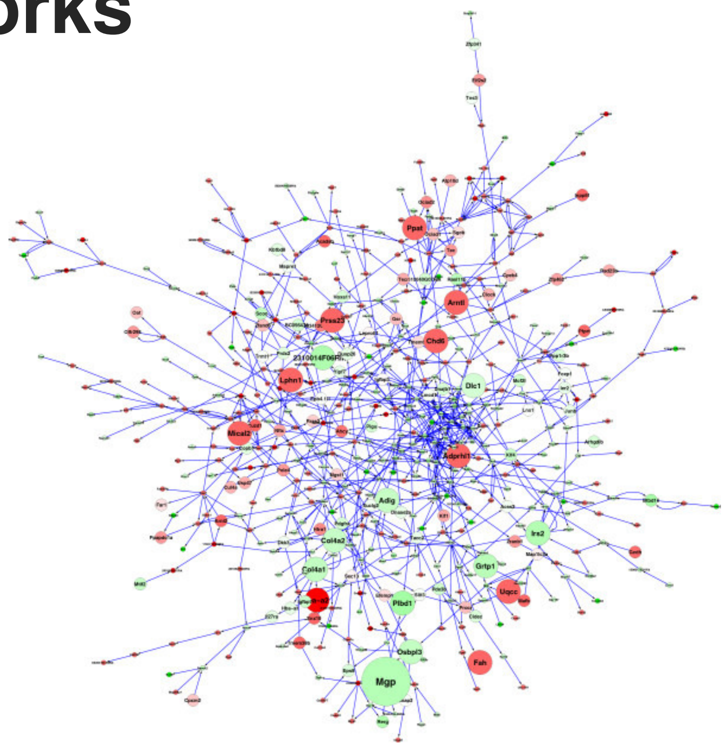
Example:

	Class C	$\mathbb{P}(C)$
$0 \leq u < 0.6$	red	0.6
$0.6 \leq u < 0.7$	green	0.1
$0.7 \leq u < 1$	blue	0.3

- If $random()$ returns $u = 0.83$, then $C = \text{blue}$.

Sampling in Bayesian networks

- Prior sampling
- Rejection sampling
- Likelihood Weighting
- Gibbs sampling (MCMC)



Prior sampling

Sample	T	F	A	S	L	R
s_1	*false*	*false*	*true*	*false*	*false*	*true*
s_2	*false*	*true*	*false*	*true*	*false*	*false*
s_3	*false*	*true*	*true*	*false*	*true*	*true*
s_4	*false*	*true*	*false*	*true*	*false*	*false*
s_5	*false*	*true*	*true*	*true*	*true*	*true*
s_6	*false*	*false*	*false*	*true*	*false*	*false*
s_7	*true*	*false*	*false*	*false*	*true*	*false*
s_8	*true*	*true*	*true*	*true*	*true*	*true*
…						
s_{1000}	*true*	*false*	*true*	*false*	*true*	*false*

Example: estimate $P(T^+)$

Forward sampling from BN

$$P(T^+) = \frac{\#\{T^+\}}{\#\{samples\}}$$

Rejection sampling

Reject the samples that conflict with our evidence

Sample	T	F	A	S	L	R	
s_1	*false*	*false*	*true*	*false*	*false*	*true*	
s_2	*false*	*true*	*false*	$\boxed{\textit{true}}$	*false*	$\boxed{\textit{false}}$	✓
s_3	*false*	*true*	*true*	*false*	*true*	*true*	
s_4	*false*	*true*	*false*	$\boxed{\textit{true}}$	*false*	$\boxed{\textit{false}}$	✓
s_5	*false*	*true*	*true*	*true*	*true*	*true*	
s_6	*false*	*false*	*false*	$\boxed{\textit{true}}$	*false*	$\boxed{\textit{false}}$	✓
s_7	*true*	*false*	*false*	*false*	*true*	$\boxed{\textit{false}}$	
s_8	*true*	*true*	*true*	$\boxed{\textit{true}}$	*true*	*true*	
…							
s_{1000}	*true*	*false*	*true*	*false*	*true*	$\boxed{\textit{false}}$	

<u>Example</u>: $\mathbb{P}(T^+|S^+, R^-)$

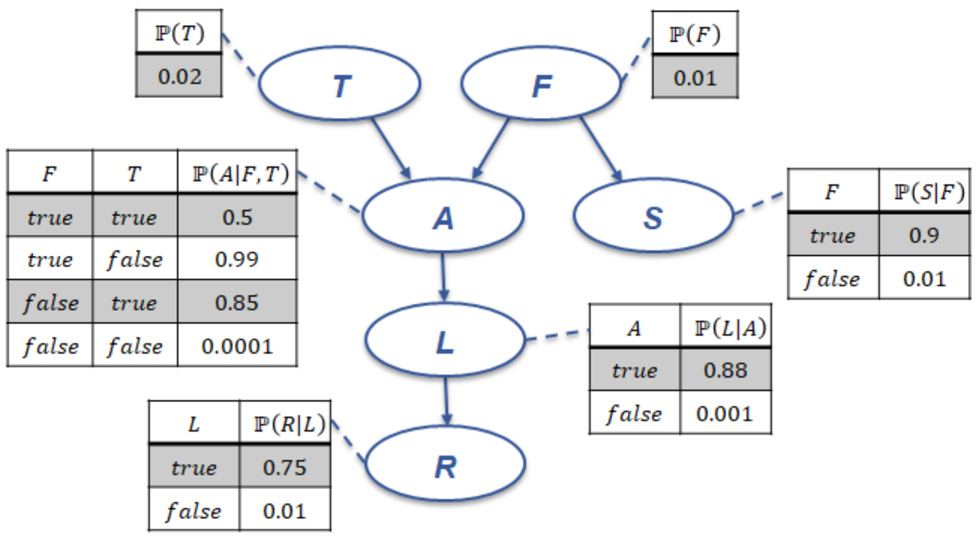
Reject all samples that conflict with our evidence (S^+, R^-)

Estimate $\mathbb{P}(T^+|S^+, R^-)$ from the accepted:

$$\mathbb{P}(T^+|S^+, R^-) = \frac{\#\{\text{accepted}, T^+\}}{\#\{\text{accepted}\}}$$

Downside: requires MANY samples

Likelihood weighting



$\mathbb{P}(T)$
0.02

$\mathbb{P}(F)$
0.01

F	T	$\mathbb{P}(A \mid F,T)$
$true$	$true$	0.5
$true$	$false$	0.99
$false$	$true$	0.85
$false$	$false$	0.0001

F	$\mathbb{P}(S \mid F)$
$true$	0.9
$false$	0.01

A	$\mathbb{P}(L \mid A)$
$true$	0.88
$false$	0.001

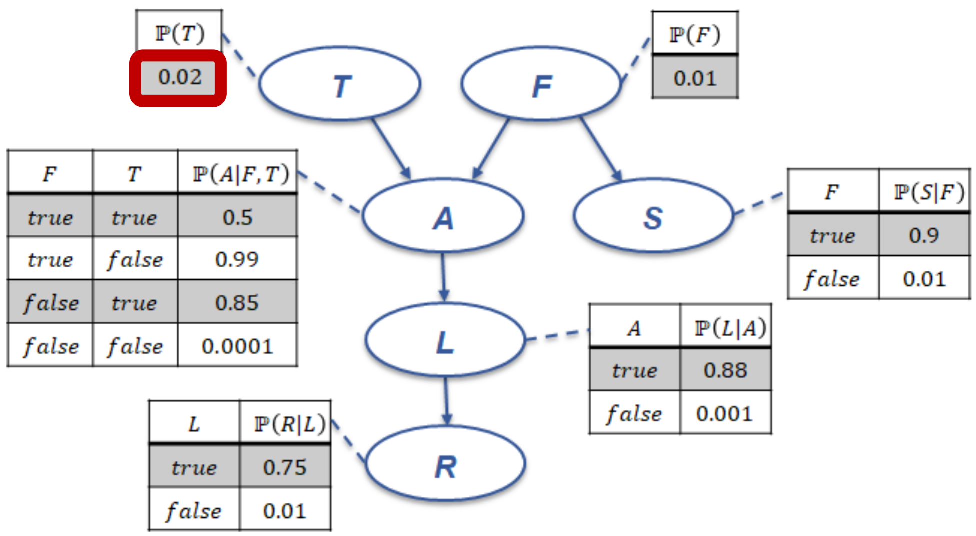
L	$\mathbb{P}(R \mid L)$
$true$	0.75
$false$	0.01

Example: estimate $\mathbb{P}(R^+, S^-)$

Variables $S = true$ and $R = false$ are fixed *evidence*

Generate sample 1:

Likelihood weighting



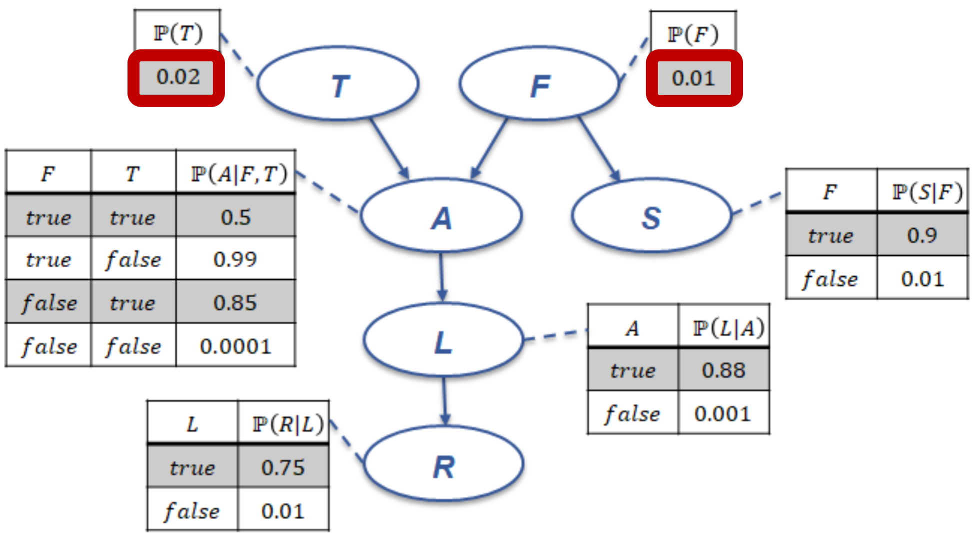
Example: estimate $\mathbb{P}(R^+, S^-)$

Variables $S = true$ **and** $R = false$ **are fixed** ***evidence***

Generate sample 1:

1. **Sample** T**: e.g** $T = false$

Likelihood weighting



$\mathbb{P}(T)$
0.02

$\mathbb{P}(F)$
0.01

F	T	$\mathbb{P}(A\|F,T)$
$true$	$true$	0.5
$true$	$false$	0.99
$false$	$true$	0.85
$false$	$false$	0.0001

F	$\mathbb{P}(S\|F)$
$true$	0.9
$false$	0.01

A	$\mathbb{P}(L\|A)$
$true$	0.88
$false$	0.001

L	$\mathbb{P}(R\|L)$
$true$	0.75
$false$	0.01

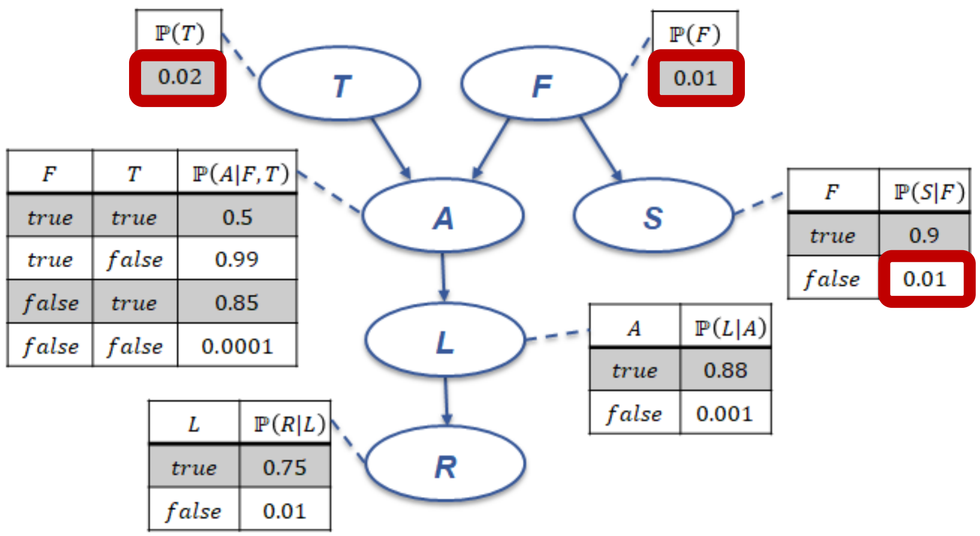
Example: estimate $\mathbb{P}(R^+, S^-)$

Variables $S = true$ **and** $R = false$ **are fixed** ***evidence***

Generate sample 1:

1. **Sample** T**: e.g** $T = false$
2. **Sample** F**: e.g.** $F = false$

Likelihood weighting



$\mathbb{P}(T)$
0.02

$\mathbb{P}(F)$
0.01

F	T	$\mathbb{P}(A\|F,T)$
$true$	$true$	0.5
$true$	$false$	0.99
$false$	$true$	0.85
$false$	$false$	0.0001

F	$\mathbb{P}(S\|F)$
$true$	0.9
$false$	0.01

A	$\mathbb{P}(L\|A)$
$true$	0.88
$false$	0.001

L	$\mathbb{P}(R\|L)$
$true$	0.75
$false$	0.01

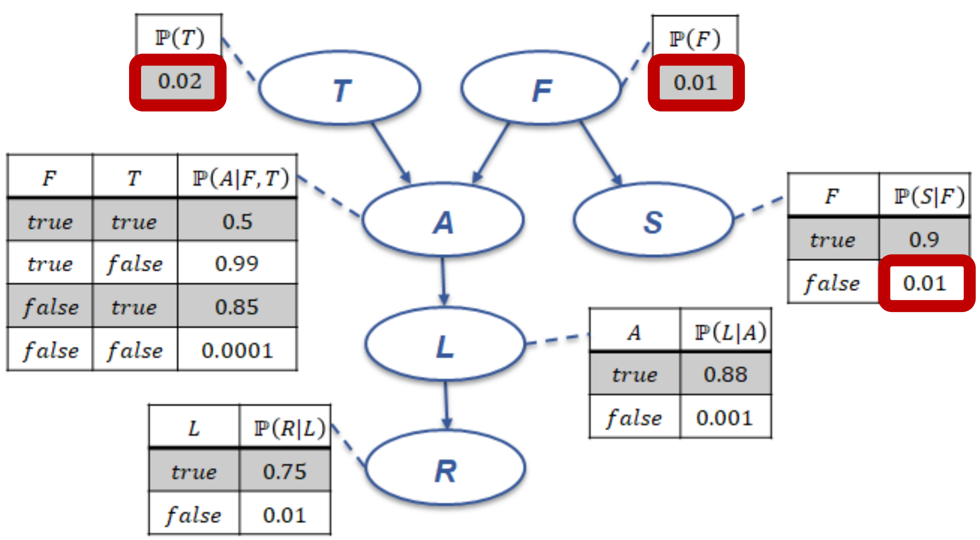
Example: estimate $\mathbb{P}(R^+, S^-)$

Variables $S = true$ **and** $R = false$ **are fixed** ***evidence***

Generate sample 1:

1. **Sample** T**: e.g** $T = false$
2. **Sample** F**: e.g.** $F = false$
3. **Sample** $S|F$**: e.g.** $S = true$

Likelihood weighting



$\mathbb{P}(T)$
0.02

$\mathbb{P}(F)$
0.01

F	T	$\mathbb{P}(A\|F,T)$
$true$	$true$	0.5
$true$	$false$	0.99
$false$	$true$	0.85
$false$	$false$	0.0001

F	$\mathbb{P}(S\|F)$
$true$	0.9
$false$	0.01

A	$\mathbb{P}(L\|A)$
$true$	0.88
$false$	0.001

L	$\mathbb{P}(R\|L)$
$true$	0.75
$false$	0.01

Example: estimate $\mathbb{P}(R^+, S^-)$

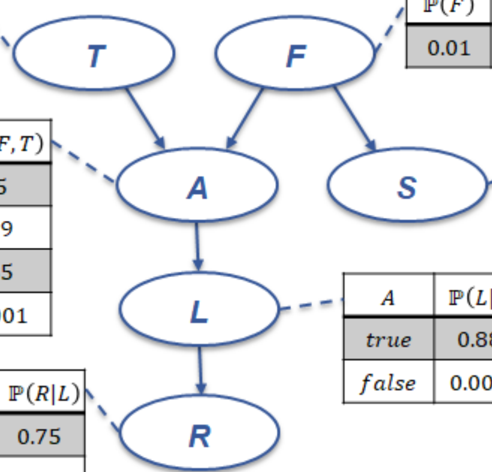
Variables $S = true$ and $R = false$ are fixed *evidence*

Generate sample 1:

1. **Sample T: e.g $T = false$**
2. **Sample F: e.g. $F = false$**
3. **Sample $S|F$: e.g. $S = true$**

1 sample: Hits: 0, Misses: 1

Likelihood weighting



$\mathbb{P}(T)$
0.02

$\mathbb{P}(F)$
0.01

F	T	$\mathbb{P}(A\|F,T)$
$true$	$true$	0.5
$true$	$false$	0.99
$false$	$true$	0.85
$false$	$false$	0.0001

F	$\mathbb{P}(S\|F)$
$true$	0.9
$false$	0.01

A	$\mathbb{P}(L\|A)$
$true$	0.88
$false$	0.001

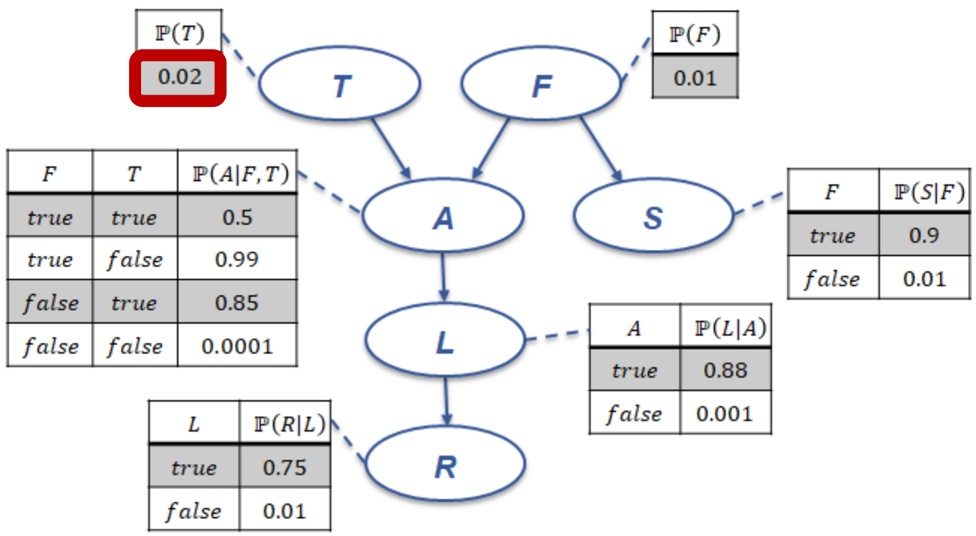
L	$\mathbb{P}(R\|L)$
$true$	0.75
$false$	0.01

Example: estimate $\mathbb{P}(R^+, S^-)$

Variables $S = true$ and $R = false$ are fixed *evidence*

Generate sample 2:

Likelihood weighting



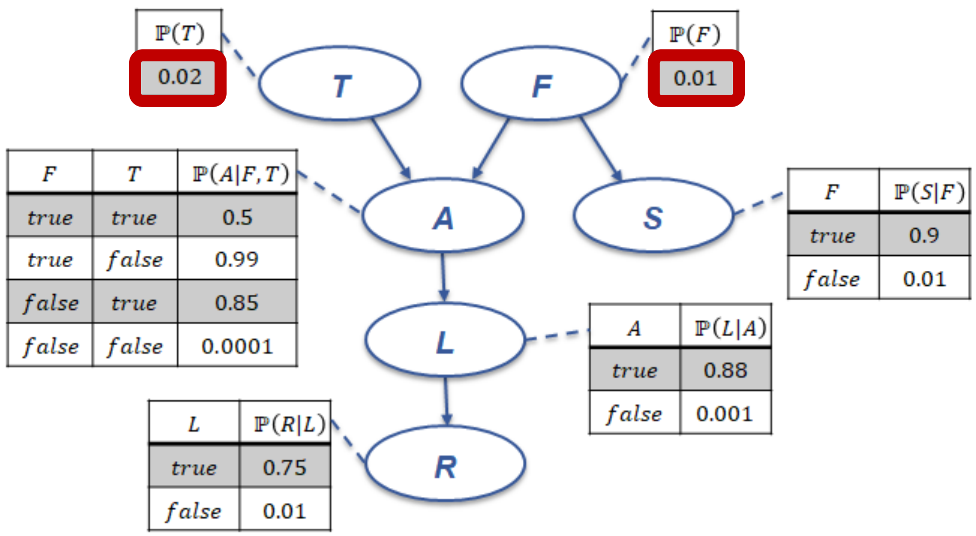
Example: estimate $\mathbb{P}(R^+, S^-)$

Variables $S = true$ **and** $R = false$ **are fixed** ***evidence***

Generate sample 2:

1. **Sample** T**: e.g** $T = false$

Likelihood weighting



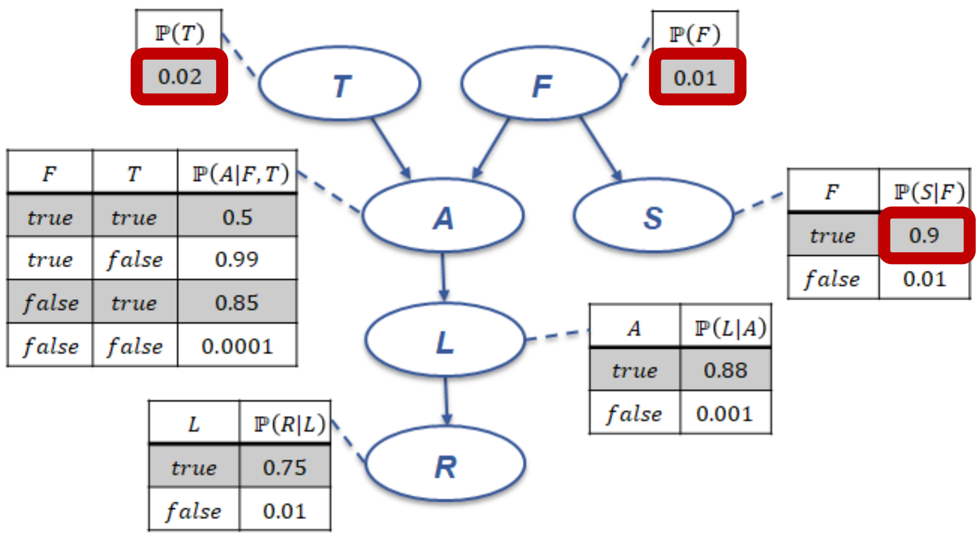
Example: estimate $\mathbb{P}(R^+, S^-)$

Variables $S = true$ and $R = false$ are fixed *evidence*

Generate sample 2:

1. **Sample T: e.g $T = false$**
2. **Sample F: e.g. $F = true$**

Likelihood weighting



$\mathbb{P}(T)$
0.02

$\mathbb{P}(F)$
0.01

F	T	$\mathbb{P}(A\|F,T)$
$true$	$true$	0.5
$true$	$false$	0.99
$false$	$true$	0.85
$false$	$false$	0.0001

F	$\mathbb{P}(S\|F)$
$true$	0.9
$false$	0.01

A	$\mathbb{P}(L\|A)$
$true$	0.88
$false$	0.001

L	$\mathbb{P}(R\|L)$
$true$	0.75
$false$	0.01

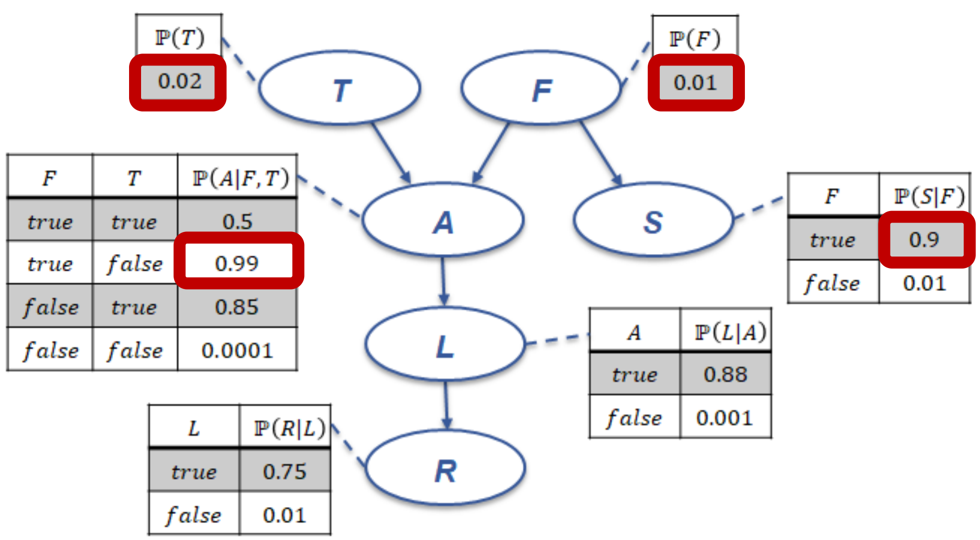
Example: estimate $\mathbb{P}(R^+, S^-)$

Variables $S = true$ and $R = false$ are fixed *evidence*

Generate sample 2:

1. **Sample T: e.g $T = false$**
2. **Sample F: e.g. $F = true$**
3. **Sample $S|F^+$: e.g. $S = false$**

Likelihood weighting



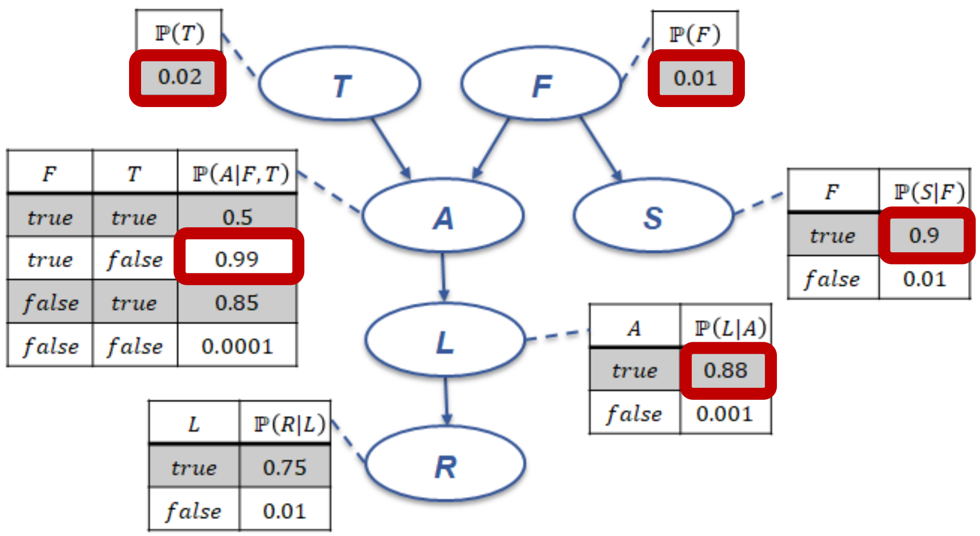
Example: estimate $\mathbb{P}(R^+, S^-)$

Variables $S = true$ **and** $R = false$ **are fixed *evidence***

Generate sample 2:

1. **Sample** T**: e.g** $T = false$
2. **Sample** F**: e.g.** $F = true$
3. **Sample** $S|F^+$**: e.g.** $S = false$
4. **Sample** $A|F^+, T^-$**: e.g.** $A = true$

Likelihood weighting



$\mathbb{P}(T)$
0.02

$\mathbb{P}(F)$
0.01

F	T	$\mathbb{P}(A\|F,T)$
$true$	$true$	0.5
$true$	$false$	0.99
$false$	$true$	0.85
$false$	$false$	0.0001

F	$\mathbb{P}(S\|F)$
$true$	0.9
$false$	0.01

A	$\mathbb{P}(L\|A)$
$true$	0.88
$false$	0.001

L	$\mathbb{P}(R\|L)$
$true$	0.75
$false$	0.01

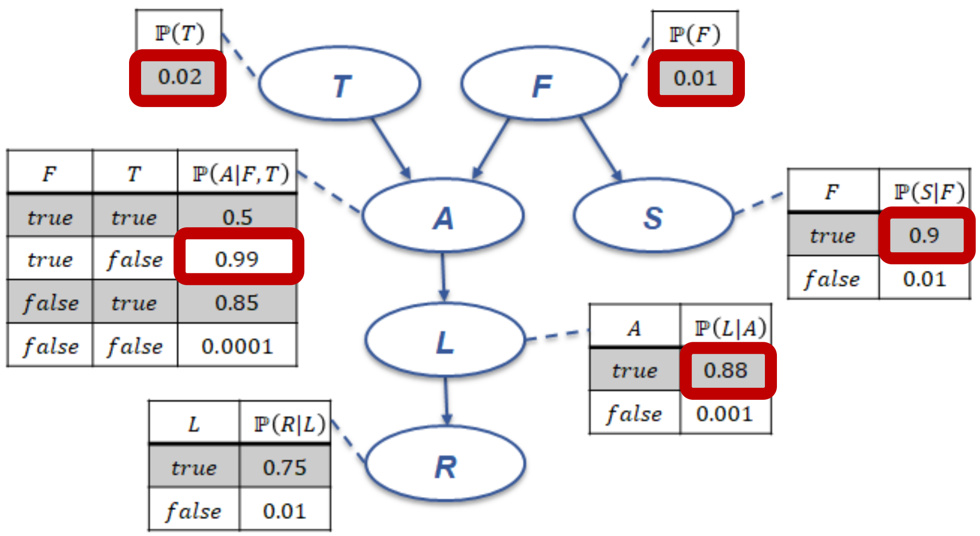
Example: estimate $\mathbb{P}(R^+, S^-)$

Variables $S = true$ **and** $R = false$ **are fixed *evidence***

Generate sample 2:

1. **Sample** T**: e.g** $T = false$
2. **Sample** F**: e.g.** $F = true$
3. **Sample** $S|F^+$**: e.g.** $S = false$
4. **Sample** $A|F^+, T^-$**: e.g.** $A = true$
5. **Sample** $L|A^+$**: e.g.** $L = true$

Likelihood weighting



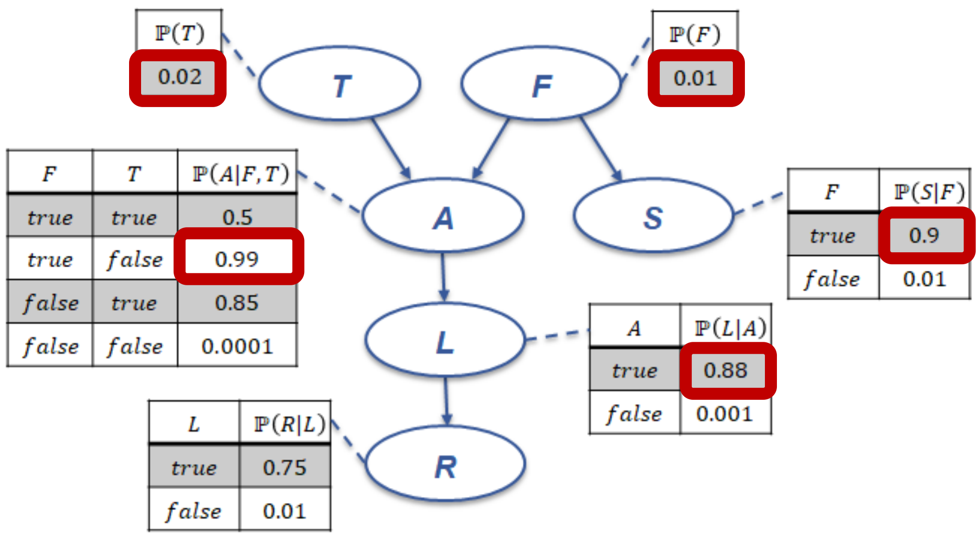
Example: estimate $\mathbb{P}(R^+, S^-)$

Variables $S = true$ **and** $R = false$ **are fixed** ***evidence***

Generate sample 2:

1. **Sample** T**: e.g** $T = false$
2. **Sample** F**: e.g.** $F = true$
3. **Sample** $S|F^+$**: e.g.** $S = false$
4. **Sample** $A|F^+, T^-$**: e.g.** $A = true$
5. **Sample** $L|A^+$**: e.g.** $L = true$
6. **Sample** $R|L^+$**: e.g.** $R = true$

Likelihood weighting



$\mathbb{P}(T)$
0.02

$\mathbb{P}(F)$
0.01

F	T	$\mathbb{P}(A\|F,T)$
$true$	$true$	0.5
$true$	$false$	0.99
$false$	$true$	0.85
$false$	$false$	0.0001

F	$\mathbb{P}(S\|F)$
$true$	0.9
$false$	0.01

A	$\mathbb{P}(L\|A)$
$true$	0.88
$false$	0.001

L	$\mathbb{P}(R\|L)$
$true$	0.75
$false$	0.01

Example: estimate $\mathbb{P}(R^+, S^-)$

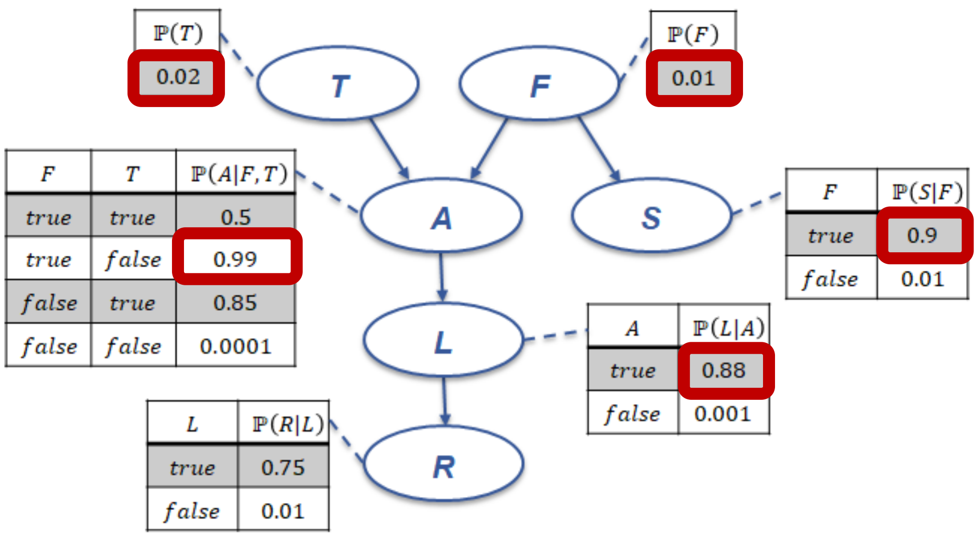
Variables $S = true$ **and** $R = false$ **are fixed** ***evidence***

Generate sample 2:

1. **Sample** T**: e.g** $T = false$
2. **Sample** F**: e.g.** $F = true$
3. **Sample** $S|F^+$**: e.g.** $S = false$
4. **Sample** $A|F^+, T^-$**: e.g.** $A = true$
5. **Sample** $L|A^+$**: e.g.** $L = true$
6. **Sample** $R|L^+$**: e.g.** $R = true$

2 samples: Hits: 1, Misses: 1

Likelihood weighting



$\mathbb{P}(T)$
0.02

$\mathbb{P}(F)$
0.01

F	T	$\mathbb{P}(A\|F,T)$
$true$	$true$	0.5
$true$	$false$	0.99
$false$	$true$	0.85
$false$	$false$	0.0001

F	$\mathbb{P}(S\|F)$
$true$	0.9
$false$	0.01

A	$\mathbb{P}(L\|A)$
$true$	0.88
$false$	0.001

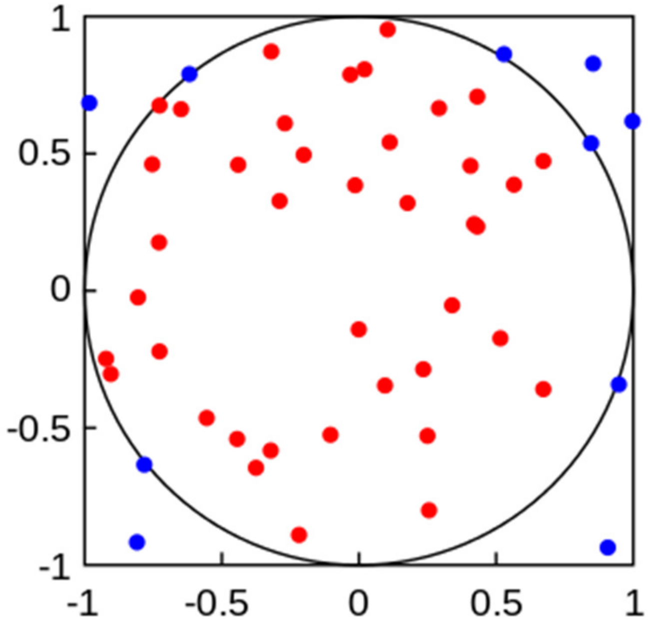
L	$\mathbb{P}(R\|L)$
$true$	0.75
$false$	0.01

Draw many such samples and then estimate $\mathbb{P}(R^+, S^-)$ by

$$\mathbb{P}(R^+, S^-) \approx \frac{Hits}{Hits + Misses}$$

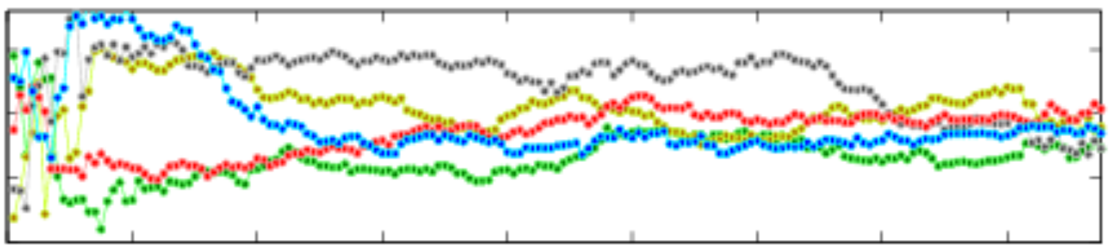
Sampling in Monte Carlo integration

- Let us approximate π
- Generate 100 random points inside the square (”throw darts”)
  - Hits (inside circle): 80
  - Misses (outside circle): 20
- Circle area $\approx \mathbf{0.8} \cdot \text{square area}$
  $= \mathbf{0.8 \cdot 4 = 3.2}$
- Also: circle area $= \boldsymbol{\pi} \cdot \mathbf{1} \approx \mathbf{3.2}$



Markov chain Monte Carlo (MCMC)

- MCMC combines Monte Carlo simulation and Markov chains.
- Idea: Instead of generating every sample from scratch, we create samples that are similar to the previous one



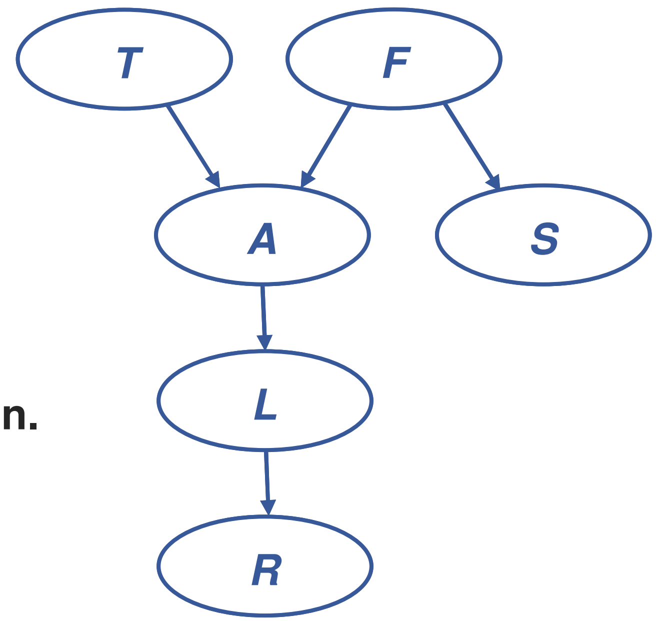
Gibbs sampling in a Bayesian network

Example: estimation of $\mathbb{P}(T^+, A^- | S^+, R^+)$

1. **Keep the evidence S^+, R^+ fixed, sample all other variables from there conditional distributions**
2. **Repeat (as many times as wanted)**
3. **Alternatively march through the variables in some predefined order.**

   E.g. sample $\mathbb{P}(T)$ and $\mathbb{P}(F)$, then $\mathbb{P}(A|T, F)$ and so on.

A.k.a. forward sampling

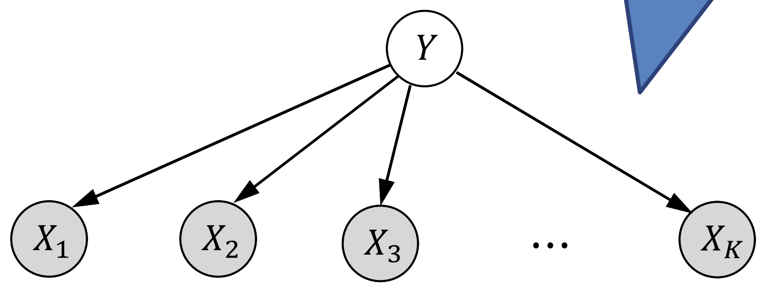


The naïve Bayes classifier

What is naïve here is to assume that each X_i only depends on Y, and not on each other.

- **Assume the graph looks like this**

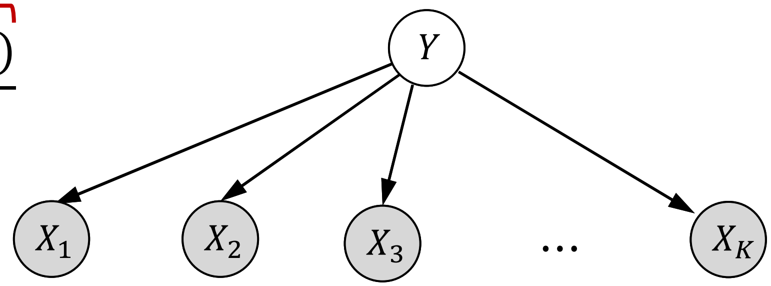
Still, very helpful simplification that give ”good enough” results in many applications.



Naïve Bayes: general description

- **The Naïve Bayes assumption**
  - $\mathbb{P}(\boldsymbol{Y}, \boldsymbol{X}_1, \boldsymbol{X}_2, \ldots \boldsymbol{X}_K) = \mathbb{P}(\boldsymbol{Y}) \prod_{k=1}^{K} \mathbb{P}(\boldsymbol{X}_k | \boldsymbol{Y})$
  - $\mathbb{P}(\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots \boldsymbol{X}_K) = \mathbb{P}(\boldsymbol{X}_1)\mathbb{P}(\boldsymbol{X}_2) \cdots \mathbb{P}(\boldsymbol{X}_K) = \prod_{k=1}^{K} \mathbb{P}(\boldsymbol{X}_k)$
- **Posterior**

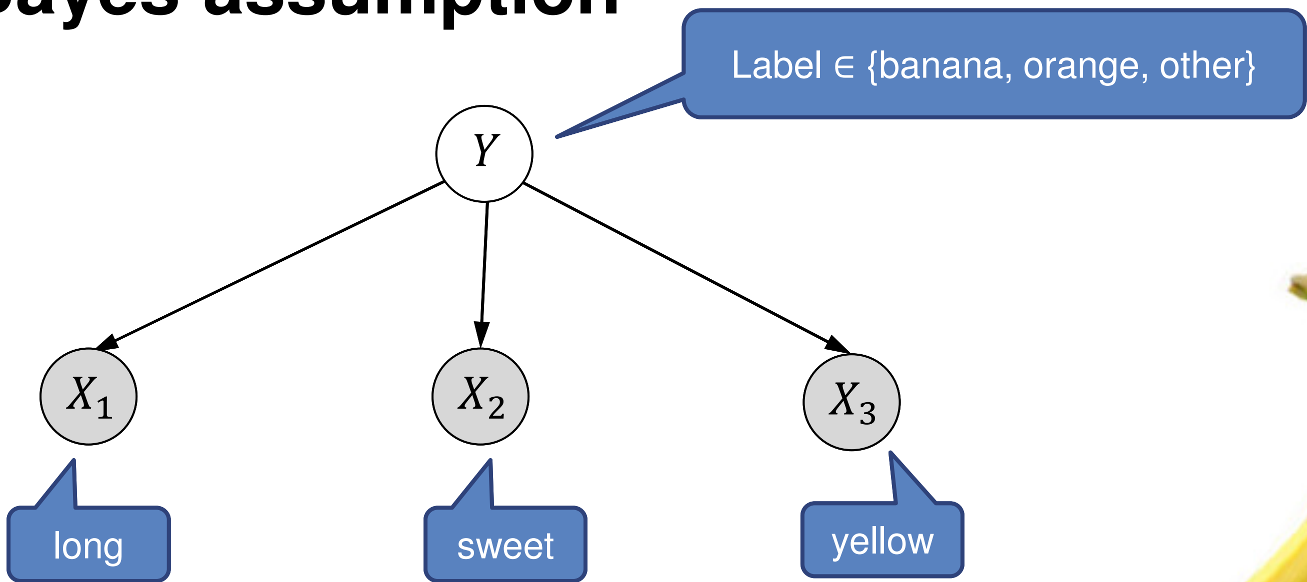
$$\mathbb{P}(\boldsymbol{Y} | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_K) = \frac{\overbrace{\mathbb{P}(\boldsymbol{Y})}^{\text{prior}} \cdot \overbrace{\prod_{k=1}^{K} \mathbb{P}(\boldsymbol{X}_k | \boldsymbol{Y})}^{\text{likelihood}}}{\underbrace{\prod_{k=1}^{K} \mathbb{P}(\boldsymbol{X}_k)}_{\text{normalizer}}}$$



Naïve Bayes: a motivating example

- We have $N = 1000$ fruits, with labels
  - Banana
  - Orange
  - Other
- Three features of each fruit
  - Long $\in \{\mathbf{true}, \mathbf{false}\}$
  - Sweet $\in \{\mathbf{true}, \mathbf{false}\}$
  - Yellow $\in \{\mathbf{true}, \mathbf{false}\}$
- Objective: determine label for a new fruit given the three features

Naïve Bayes assumption



Features are independent of each other: may not be entirely correct, but if true for *most*, then it is ok.

Naïve Bayes: a motivating example

Label	Long	Not long	Sweet	Not sweet	Yellow	Not yellow	Total
Banana	400	100	350	150	450	50	500
Orange	0	300	150	150	300	0	300
Other	100	200	150	50	50	150	200
Total	500	500	650	350	800	200	1000

- **Potential queries**
  - **What is the probability of it being a banana given the features long, sweet and yellow?**

Naïve Bayes: a motivating example

Step 1: Compute the prior probabilities $P(Y)$ for each fruit label

- **from prior information**
- **or from training data**

$$\mathbb{P}(\boldsymbol{Y} = \text{banana}) = \mathbf{500/1000 = 0.5}$$

$$\mathbb{P}(\boldsymbol{Y} = \text{orange}) = \mathbf{300/1000 = 0.3}$$

$$\mathbb{P}(\boldsymbol{Y} = \text{other}) = \mathbf{200/1000 = 0.2}$$

Label	Total
Banana	500
Orange	300
Other	200
Total	1000

Naïve Bayes: a motivating example

Step 2: Compute the denominator

$$\prod_{k=1}^{K} P(X_k)$$

$$\mathbb{P}(X_1 = \text{long}) = \mathbf{500/1000 = 0.5}$$

$$\mathbb{P}(X_2 = \text{sweet}) = \mathbf{650/1000 = 0.65}$$

$$\mathbb{P}(X_3 = \text{yellow}) = \mathbf{800/1000 = 0.8}$$

Label	Long	Sweet	Yellow	Total
Banana	400	350	450	500
Orange	0	150	300	300
Other	100	150	50	200
Total	500	650	800	1000

Naïve Bayes: a motivating example

Step 3: Compute the likelihood

$$\prod_{k=1}^{K} \mathbb{P}(\boldsymbol{X}_k|\boldsymbol{Y}) = \prod_{k=1}^{K} \frac{\#\{\text{fruits with label } \boldsymbol{Y} \text{ and feature } \boldsymbol{X}_k\}}{\#\{\text{fruits with label } \boldsymbol{Y}\}}$$

$$\mathbb{P}(\boldsymbol{X}_1 = \text{long}|\text{banana}) = \mathbf{400/500 = 0.8}$$

$$\mathbb{P}(\boldsymbol{X}_2 = \text{sweet}|\text{banana}) = \mathbf{350/500 = 0.7}$$

$$\mathbb{P}(\boldsymbol{X}_3 = \text{yellow}|\text{banana}) = \mathbf{450/500 = 0.9}$$

Label	Long	Sweet	Yellow	Total
Banana	400	350	450	500

Naïve Bayes: a motivating example

Given that the fruit is long, sweet, and yellow, what is the probability it is a banana?

$$\mathbb{P}(\text{banana}|\text{long}, \text{sweet}, \text{yellow}) =$$

$$= \frac{\mathbb{P}(\text{banana})\mathbb{P}(\text{long}|\text{banana})\mathbb{P}(\text{sweet}|\text{banana})\mathbb{P}(\text{yellow}|\text{banana})}{\mathbb{P}(\text{long})\mathbb{P}(\text{sweet})\mathbb{P}(\text{yellow})}$$

$$= \frac{\mathbf{0.5 \cdot 0.8 \cdot 0.7 \cdot 0.9}}{\mathbf{0.5 \cdot 0.65 \cdot 0.8}} = \mathbf{0.969}$$

Naïve Bayes: a motivating example

Given some features, which is the most likely label?

Which is biggest?

- $\mathbb{P}(\text{banana}|\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K)$**?**
- $\mathbb{P}(\text{orange}|\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K)$**?**
- $\mathbb{P}(\text{other}|\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K)$**?**

All labels have the same denominator

To find out, it is enough to compute the nominator

$$\mathbb{P}(\boldsymbol{Y}|\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K) = \frac{\mathbb{P}(\boldsymbol{Y}) \cdot \prod_{k=1}^{K} \mathbb{P}(\boldsymbol{X}_k|\boldsymbol{Y})}{\prod_{k=1}^{K} \mathbb{P}(\boldsymbol{X}_k)}$$

Naïve Bayes: a motivating example

Step 4: Given that the fruit is long, sweet, and yellow, what is the *most likely label*?

$$\mathbb{P}(\text{banana}|\text{long, sweet, yellow})$$

$$\propto \mathbb{P}(\text{banana})\mathbb{P}(\text{long }|\text{banana})\mathbb{P}(\text{sweet }|\text{banana})\mathbb{P}(\text{yellow }|\text{banana})$$

$$= \mathbf{0.5 \cdot 0.8 \cdot 0.7 \cdot 0.9 = 0.252}$$

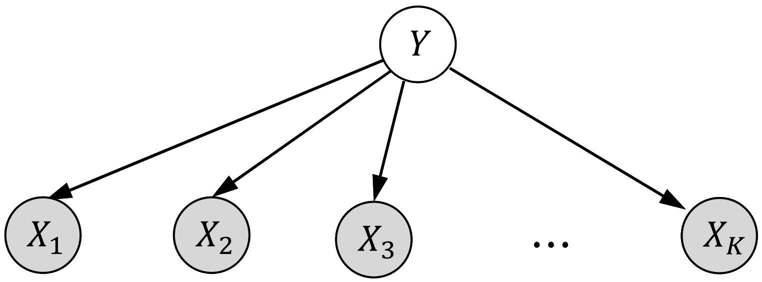
$$\mathbb{P}(\text{orange }|\text{long, sweet, yellow}) \propto \mathbf{0} \;\; \text{because } \mathbb{P}(\text{long}|\text{orange}) = \mathbf{0}$$

$$\mathbb{P}(\text{other }|\text{long, sweet, yellow}) \propto \mathbf{0.01875}$$

The fruit is most likely a banana!

Applications of naïve Bayes

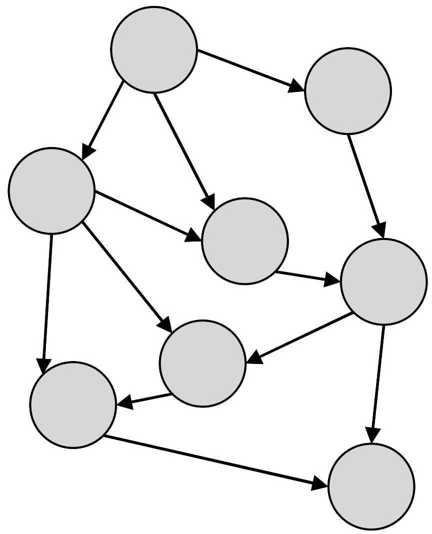
- Real-time prediction (fast, scalable)
- Multi-class prediction
- Text classification/spam filtering/sentiment analysis
- Recommendation system



Inference algorithms in graphical models

Exact inference

- Variable elimination
- Message passing/belief propagation
- Junction trees



Approximative inference

- Stochastic simulation
- Markov chain Monte Carlo (MCMC)
- Variational algorithms