Machine Learning 10-601

Tom M. Mitchell
Machine Learning Department
Carnegie Mellon University

February 18, 2015

Today:

- Graphical models
- Bayes Nets:
  - Representing distributions
  - Conditional independencies
  - Simple inference
  - Simple learning

Readings:

- Bishop chapter 8, through 8.2

Graphical Models

- Key Idea:
  - Conditional independence assumptions useful
  - but Naïve Bayes is extreme!
  - Graphical models express sets of conditional independence assumptions via graph structure $G = \langle V, E \rangle$
  - Graph structure plus associated parameters define *joint probability distribution over set of variables*

Vertex: (node)
Edge

10-601

- Two types of graphical models:
  - Directed graphs (aka Bayesian Networks)
  - Undirected graphs (aka Markov Random Fields)

Graphical Models – Why Care?

- Among most important ML developments of the decade

- Graphical models allow combining:
  - Prior knowledge in form of dependencies/independencies
  - Prior knowledge in form of priors over parameters
  - Observed training data

- Principled and ~general methods for
  - Probabilistic inference
  - Learning

- Useful in practice
  - Diagnosis, help systems, text analysis, time series models, ...

Conditional Independence

Definition: X is conditionally independent of Y given Z, if the probability distribution governing X is independent of the value of Y, given the value of Z

$$(\forall i, j, k) P(X = x_i | Y = y_j, Z = z_k) = P(X = x_i | Z = z_k)$$

Which we often write $P(X|Y,Z) = P(X|Z)$

E.g., $P(Thunder|Rain, Lightning) = P(Thunder|Lightning)$

Marginal Independence

Definition: X is marginally independent of Y if

$$(\forall i, j) P(X = x_i, Y = y_j) = P(X = x_i) P(Y = y_j)$$

$= P(Y=y_j | X=x_i) \, P(X=x_i)$

$= P(X=x_i | Y=y_j) \, P(Y=y_j)$

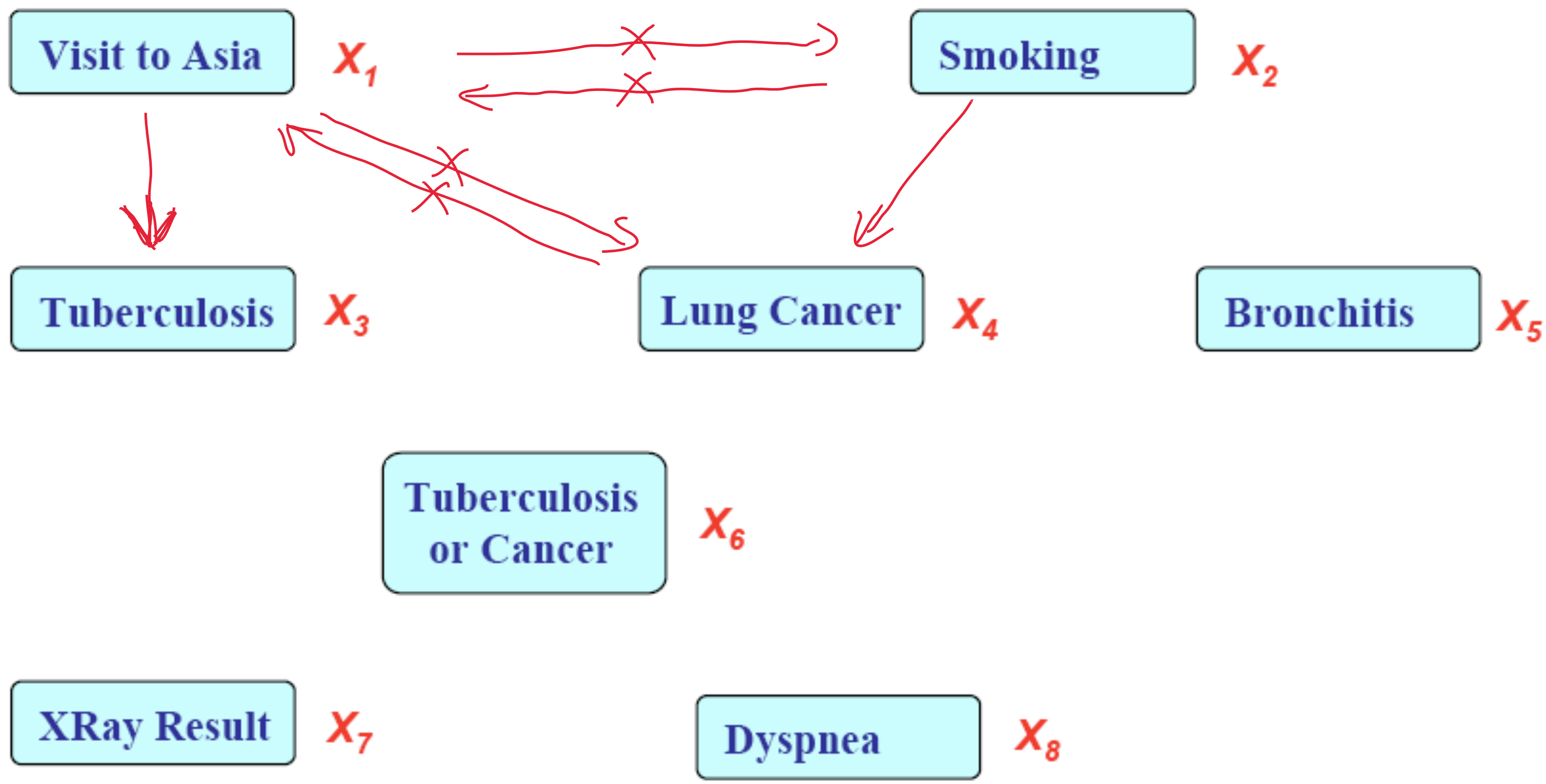
Equivalently, if

$$(\forall i, j) P(X = x_i | Y = y_j) = P(X = x_i)$$

Equivalently, if

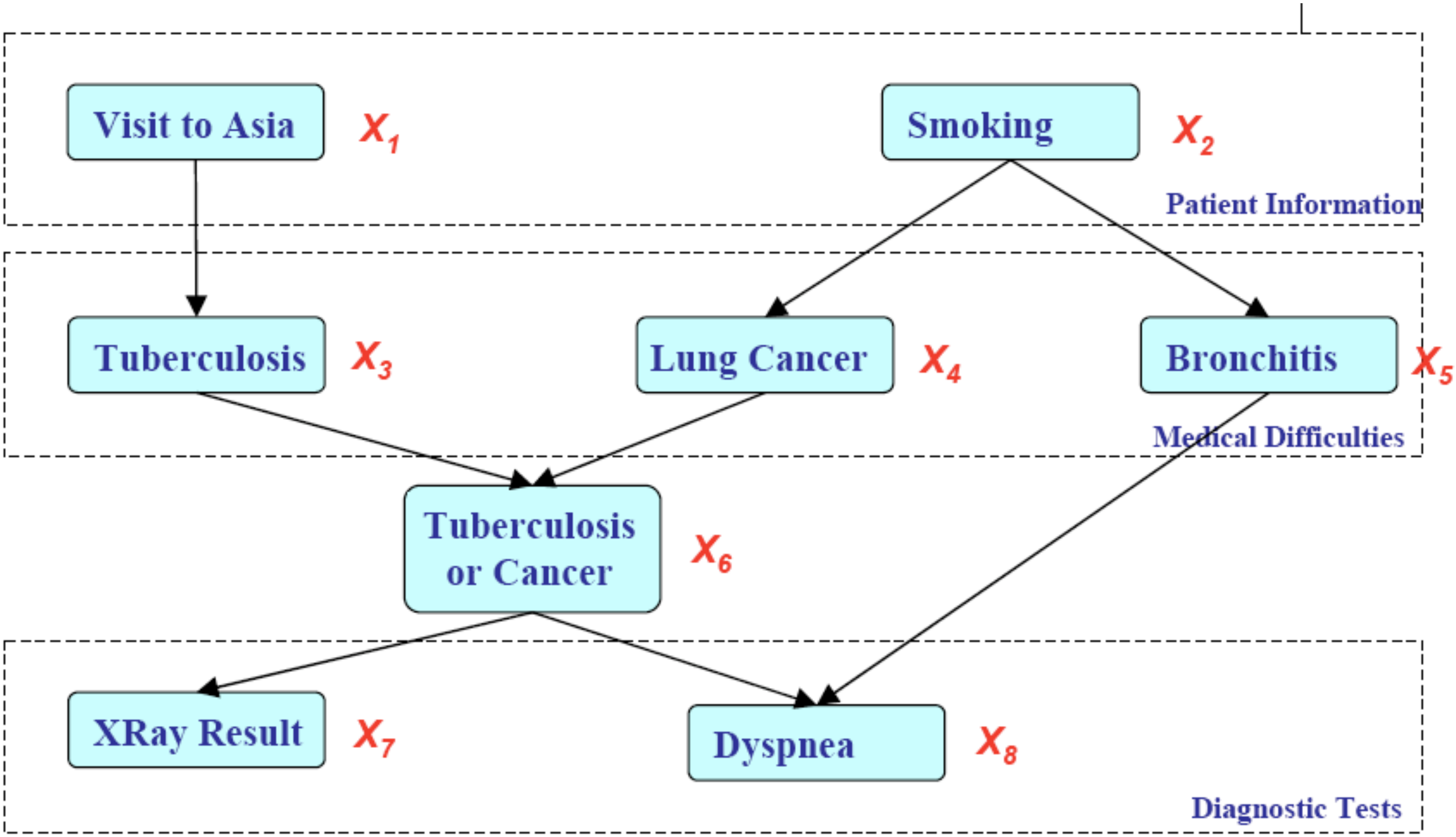
$$(\forall i, j) P(Y = y_i | X = x_j) = P(Y = y_i)$$

Represent Joint Probability Distribution over Variables

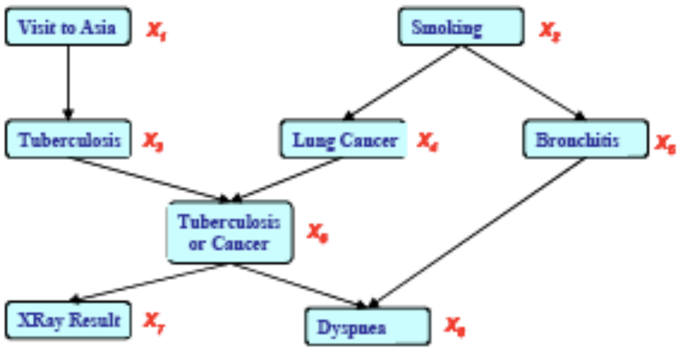


$$P(x_1, x_2, \ldots, x_8) = P(x_1)\underline{P(x_2)}P(x_3|x_1)\underline{P(x_4|x_2)P(x_5|x_2)}P(x_6|x_3, x_4)$$
$$P(x_7|x_6)\ P(x_8|x_5, x_6)$$

Describe network of dependencies



Bayes Nets define Joint Probability Distribution in terms of this graph, plus parameters



$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$$

$$= \overset{1}{P(X_1)}\ \overset{1}{P(X_2)}\ \overset{2}{P(X_3 | X_1)}\ \overset{2}{P(X_4 | X_2)}\ \overset{2}{P(X_5 | X_2)}$$

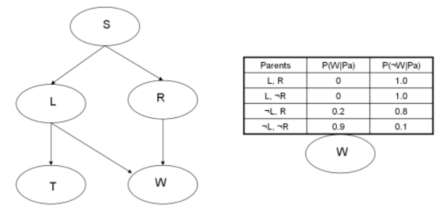
$$\underset{2^2}{P(X_6 | X_3, X_4)}\ \underset{2}{P(X_7 | X_6)}\ \underset{2^2}{P(X_8 | X_5, X_6)}$$

params: 18

Benefits of Bayes Nets:

- Represent the full joint distribution in fewer parameters, using prior knowledge about dependencies
- Algorithms for inference and learning

Bayesian Networks Definition



Parents	P(W|Pa)	P(¬W|Pa)
L, R	0	1.0
L, ¬R	0	1.0
¬L, R	0.2	0.8
¬L, ¬R	0.9	0.1

A Bayes network represents the joint probability distribution over a collection of random variables

(DAG)

A Bayes network is a directed acyclic graph and a set of conditional probability distributions (CPD's)

- Each node denotes a random variable
- Edges denote dependencies
- For each node X_i its CPD defines $P(X_i \mid Pa(X_i))$
- The joint distribution over all variables is defined to be

$$P(X_1 \ldots X_n) = \prod_i P(X_i|Pa(X_i))$$

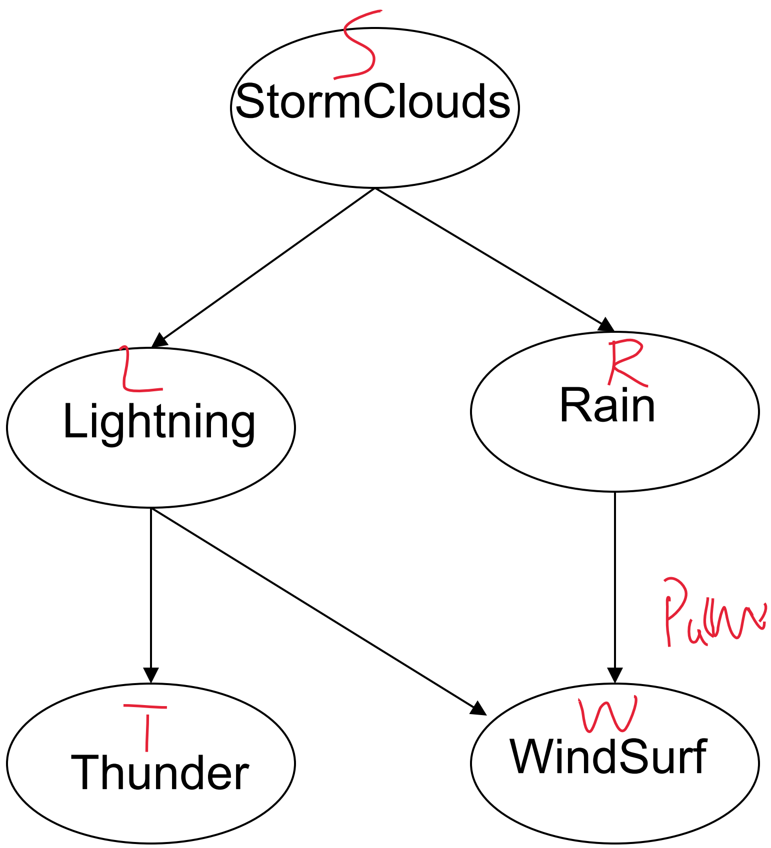
BN

(DAG). Graph ← prior.

CPD ← MLE MAP

Pa(X) = immediate parents of X in the graph

Bayesian Network

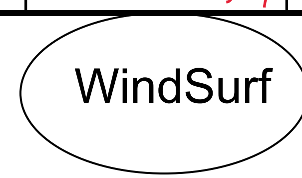


$Pa(W) = \{L, R\}$

Nodes = random variables

A conditional probability distribution (CPD) is associated with each node N, defining P(N | Parents(N))

Parents	P(W|Pa)	P(¬W|Pa)
L, R	0 θ_1	1.0 $1-\theta_1$
L, ¬R	0 θ_2	1.0 $1-\theta_2$
¬L, R	0.2 θ_3	0.8 $1-\theta_3$
¬L, ¬R	0.9 θ_4	0.1 $1-\theta_4$



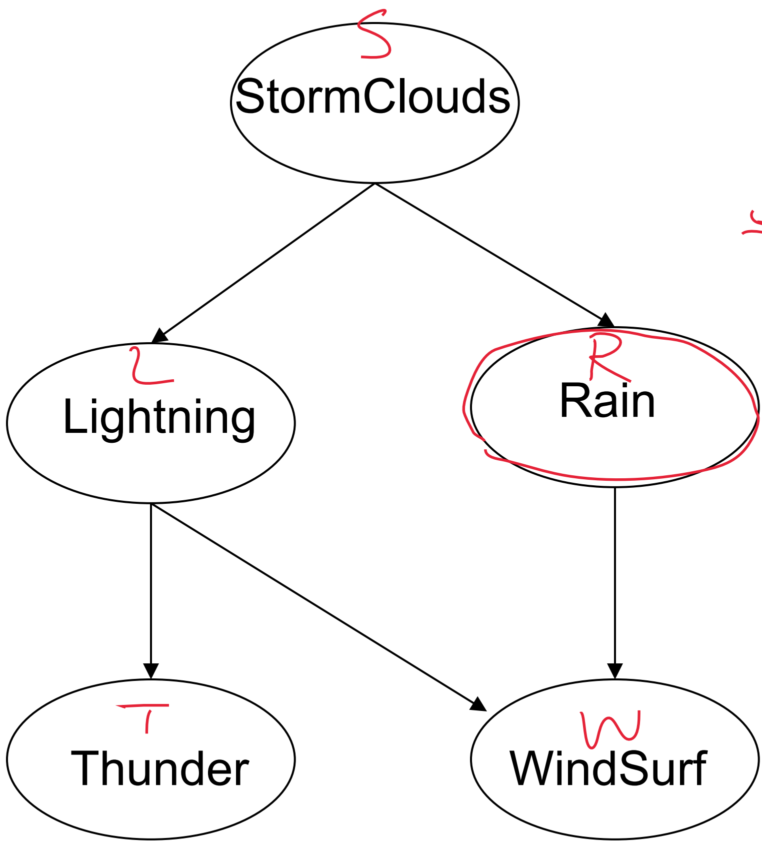
$P(W=1, T=0, R=0, L=1) \cdot = \cdot$

The joint distribution over all variables:

$$P(X_1 \ldots X_n) = \prod_i P(X_i | Pa(X_i))$$

Bayesian Network

$W \perp\!\!\!\perp T \mid \{L, R\}$



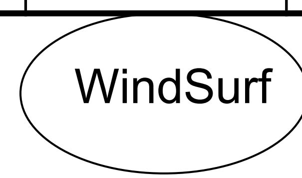
What can we say about conditional independencies in a Bayes Net?

One thing is this:

Each node is conditionally independent of its non-descendents, given only its immediate parents.

$R \perp\!\!\!\perp \{L, T\} \mid S$

Parents	P(W|Pa)	P(¬W|Pa)
L, R	0	1.0
L, ¬R	0	1.0
¬L, R	0.2	0.8
¬L, ¬R	0.9	0.1



$P(S, L, R, T, W) = P(S)\,P(L \mid S)\,P(R \mid S)\,P(T \mid L)\,P(W \mid L, R)$

$\Rightarrow P(S, L, R) = P(S)\,P(L \mid S)\,P(R \mid S)$

$= P(L, R \mid S)\,P(S)$

$L \perp\!\!\!\perp R \mid S$

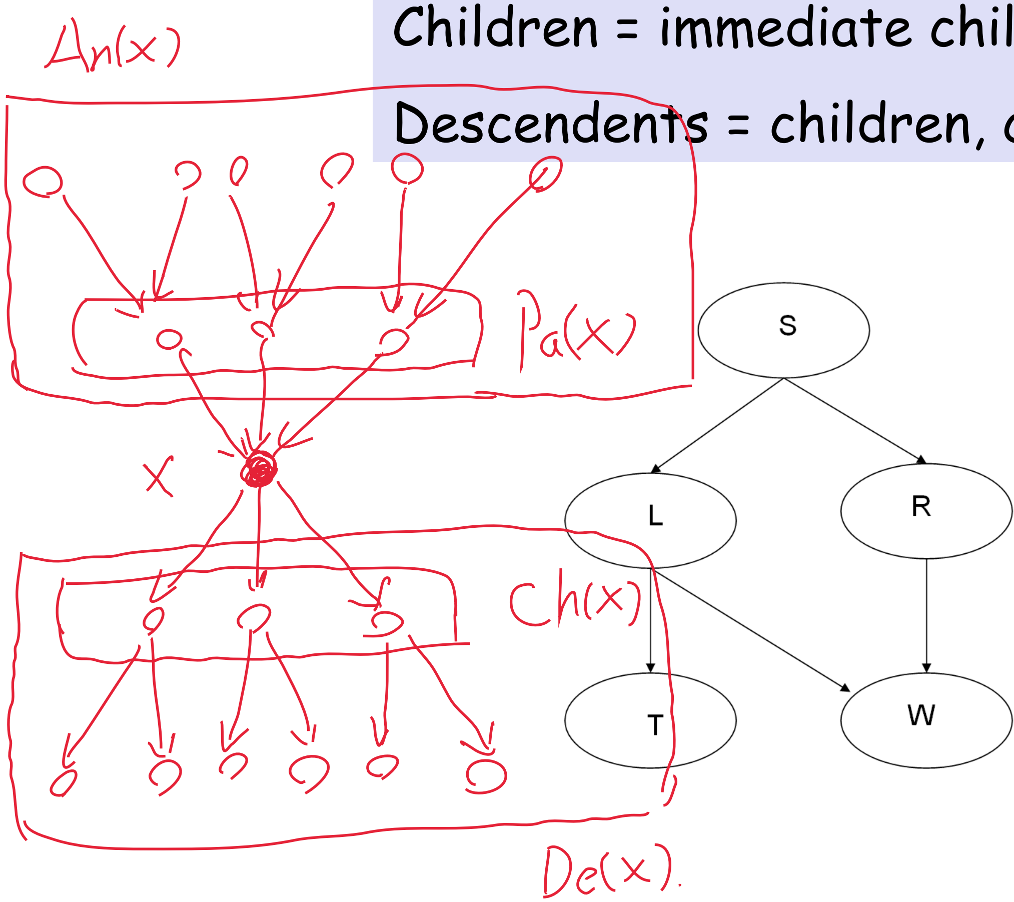
Some helpful terminology

Parents = Pa(X) = immediate parents

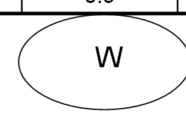
An(x) Antecedents = parents, parents of parents, ...

Children = immediate children Ch(x)

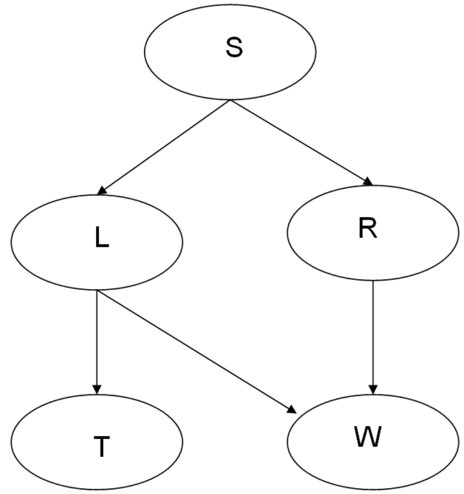
Descendents = children, children of children, ... De(x)



Parents	P(W|Pa)	P(¬W|Pa)
L, R	0	1.0
L, ¬R	0	1.0
¬L, R	0.2	0.8
¬L, ¬R	0.9	0.1



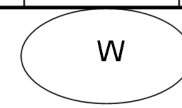
Bayesian Networks



- CPD for each node X_i describes $P(X_i \mid Pa(X_i))$

$P(X \mid Y, Z) = P(X \mid Z)$: $X \perp\!\!\!\perp Y \mid Z$

Parents	P(W|Pa)	P(¬W|Pa)
L, R	0	1.0
L, ¬R	0	1.0
¬L, R	0.2	0.8
¬L, ¬R	0.9	0.1



Chain rule of probability says that in general:

$W \perp\!\!\!\perp \{S, T\} \mid \{L, R\}$

$$P(S,L,R,T,W) = P(S)P(L|S)P(R|S,L)P(T|S,L,R)P(W|S,L,R,T)$$

No Conditional Independence (S → L → R → T → W)

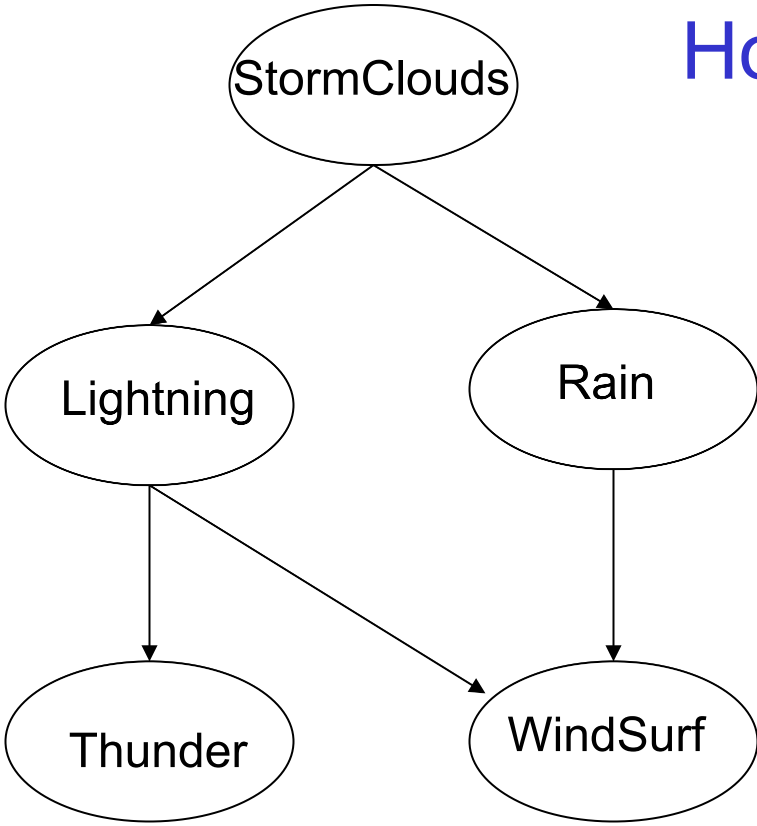
$P(R \mid S, L) = P(R \mid S) \Rightarrow R \perp\!\!\!\perp L \mid S$

But in a Bayes net: $$P(X_1 \ldots X_n) = \prod_i P(X_i | Pa(X_i))$$

Conditional Independ.

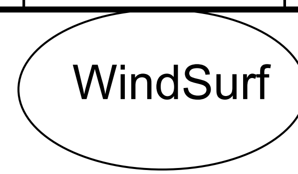
$P(S, L, R, T, W) = P(S)\, P(L \mid S)\, P(R \mid S)\, P(T \mid L)\, P(W \mid L, R)$

$P(T \mid S, L, R) = P(T \mid L) \Rightarrow T \perp\!\!\!\perp \{S, R\} \mid L$.



How Many Parameters?

Parents	P(W|Pa)	P(¬W|Pa)
L, R	0	1.0
L, ¬R	0	1.0
¬L, R	0.2	0.8
¬L, ¬R	0.9	0.1



nodes = n , (boolean)

To define joint distribution in general? # paras: $2^n - 1$ (expenontial)

To define joint distribution for this Bayes Net?

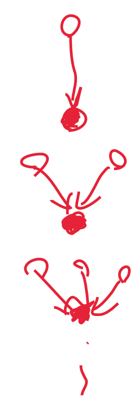
(Linear)

1- order: $2^1 n$

2- order $2^2 \cdot n$

3- order $2^3 n$

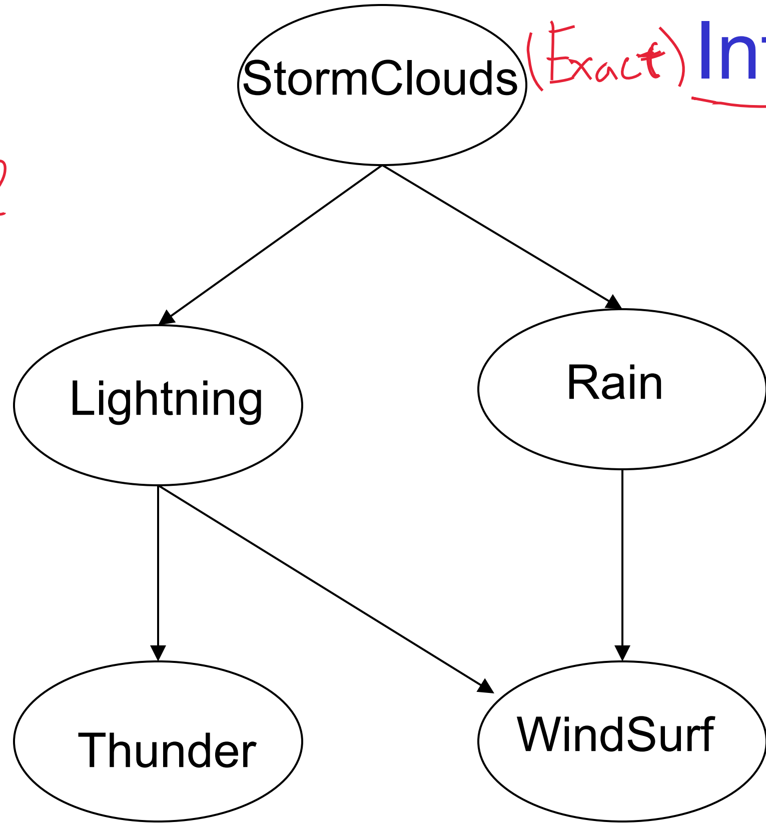
⋮



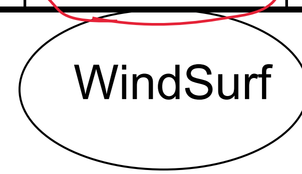
NP-Hard

(Exact) Inference in Bayes Nets

$P(T=0)$



Parents	P(W|Pa)	P(¬W|Pa)
L, R	0	1.0
L, ¬R	0	1.0
¬L, R	0.2	0.8
¬L, ¬R	0.9	0.1



0.2

$P(W=1 \mid L=0, R=1)$

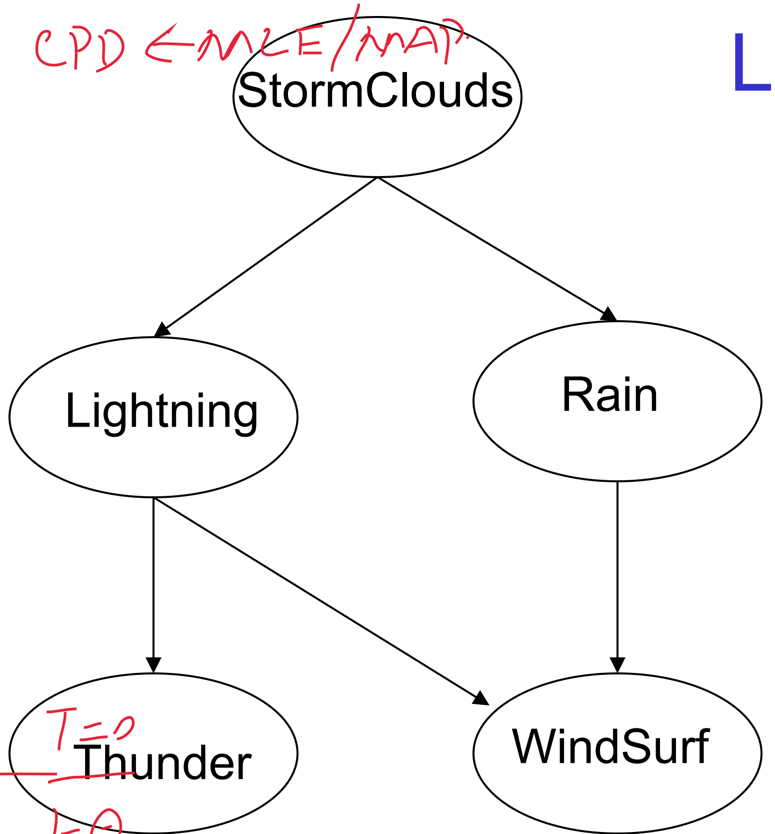
$$P(S=1, L=0, R=1, T=0, W=1) = P(S=1)\,P(L=0 \mid S=1)\,P(R=1 \mid S=1)\,P(T=0 \mid L=0)$$

$$P(S=1, L=0, R=1) = \sum_{T,W \in \{0,1\}} P(S=1, L=0, R=1, W, T)$$

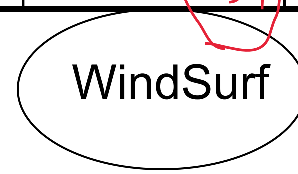
$$P(S=1 \mid L=0) = \frac{P(S=1, L=0)}{P(L=0)} = \frac{\sum_{R,T,W} P(S=1, L=0, R, T, W)}{\sum_{S,R,T,W} P(L=0, R, S, T, W)}$$

BN ← Graph (DAG) ← prior
CPD ← MLE/MAP

Learning a Bayes Net



Parents	P(W|Pa)	P(¬W|Pa)
L, R	0 θ_1	1.0 $1-\theta_1$
L, ¬R	0 θ_2	1.0 $1-\theta_2$
¬L, R	0.2 θ_3	0.8 $1-\theta_3$
¬L, ¬R	0.9 θ_4	0.1 $1-\theta_4$



P	T=1	T=0
L=1	θ_1	$1-\theta_1$
L=0	θ_0	$1-\theta_0$

$\Theta = \{\theta_1, \theta_0\}$

Training data:

$$D = \{\underbrace{(s_j, l_j, r_j, t_j, w_j)}_{x_j}\}_{j=1}^{m} \quad \text{(i.i.d.)}$$

Consider learning when graph structure is given, and data = { <s,l,r,t,w> }

What is the MLE solution? MAP?

$$p(t \mid l, \theta_1, \theta_0) = \theta_1^{LT}(1-\theta_1)^{L(1-T)} \times \theta_0^{(1-L)T}(1-\theta_0)^{(1-L)(1-T)}$$

$$\max_{\Theta} L(\Theta) = p(D \mid \Theta) \Rightarrow \prod_{j=1}^{m} p(t_j \mid l_j, \theta_1, \theta_0)$$

log-Likelihood: $\ell(\Theta) \Rightarrow \sum_{j=1}^{m} \ln p(t_j \mid l_j, \theta_1, \theta_0)$

$$\frac{\partial \ell(\Theta)}{\partial \theta_1} = 0 \Rightarrow \hat{\theta}_1 = \frac{|D_{T=1,L=1}|}{|D_{T=1}|}$$

$$\frac{\partial \ell(\Theta)}{\partial \theta_0} = 0 \Rightarrow \hat{\theta}_0 = \frac{|D_{T=1,L=0}|}{|D_{T=1}|}$$

Algorithm for Constructing Bayes Network

- Choose an ordering over variables, e.g., $X_1, X_2, \ldots X_n$
- For i=1 to n
  - Add X_i to the network
  - Select parents $Pa(X_i)$ as minimal subset of $X_1 \ldots X_{i-1}$ such that

$$P(X_i|Pa(X_i)) = P(X_i|\underbrace{X_1, \ldots, X_{i-1}}_{Pa(X_i),\ \overline{Pa}(X_i)})$$

$(X_i \perp \overline{Pa}(X_i) \mid Pa(X_i))$

$P(X_i \mid Pa(X_i), \overline{Pa}(X_i))$

$= P(X_i \mid Pa(X_i))$

Notice this choice of parents assures

$$P(X_1 \ldots X_n) = \prod_i P(X_i|X_1 \ldots X_{i-1}) \quad \text{(by chain rule)}$$

$$= \prod_i P(X_i|Pa(X_i)) \quad \text{(by construction)}$$

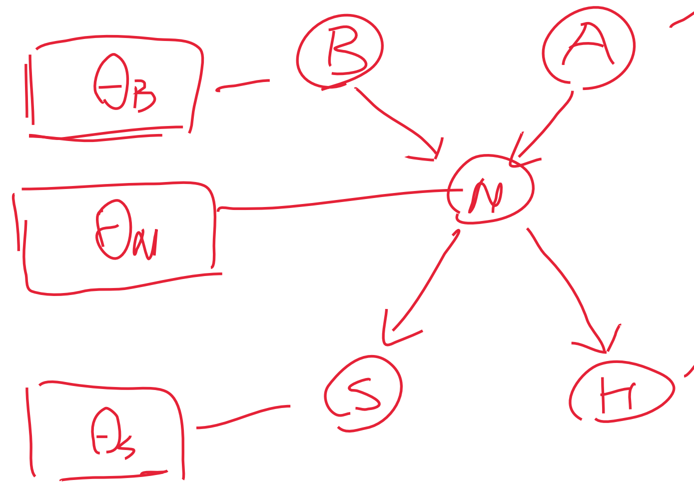
Example

- Bird flu and Allegies both cause Nasal problems
- Nasal problems cause Sneezes and Headaches

B A N

S H

BN: Graph, CPD



P(A)	A=1	A=0
	θ_a	$1-\theta_a$

P(H|N)	H=1	H=0
N=1	θ_{h_1}	$1-\theta_{h_1}$
N=0	θ_{h_0}	$1-\theta_{h_0}$

Training data

$$D=\{\underbrace{(b_j, a_j, n_j, s_j, h_j)}_{x_j}\}_{j=1}^{m}$$

$$\ell(\theta) = \ln P(D \mid \theta)$$

$$= \sum_{j=1}^{m} \ln P(x_j \mid \theta_a, \theta_b, \theta_n, \theta_s, \theta_H)$$

$$= \sum_{j=1}^{m} \ln P(x_j \mid \theta_a) + \sum_{j=1}^{m} \ln P(x_j \mid \theta_b) + \sum_{j=1}^{m} \ln P(x_j \mid \theta_n) + \sum_{j=1}^{m} \ln P(x_j \mid \theta_s) + \sum_{j=1}^{m} \ln P(x_j \mid \theta_h)$$

$$\frac{\partial \ell(\theta)}{\partial \theta} = 0$$

What is the Bayes Network for X1,…X4 with NO assumed conditional independencies?

$$p(X_1, X_2, X_3, X_4) = p(X_1)\,p(X_2|X_1)\,p(X_3|X_1,X_2)\,p(X_4|X_1,X_2,X_3)$$

$1 \to 2 \to 3 \to 4$

$1 \to 2 \to 4 \to 3 \Rightarrow p(X_1)\,p(X_2|X_1)\,p(X_4|X_1,X_2)\,p(X_3|X_1,X_2,X_4)$

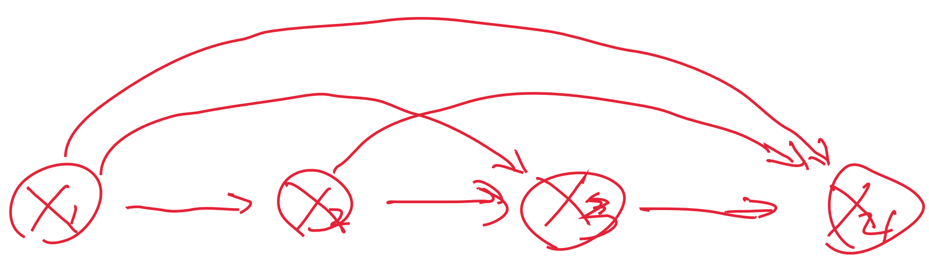
4!

(equivalent but not unique)

⋮

$4 \to 3 \to 2 \to 1 \Rightarrow p(X_4)\,p(X_3|X_4)\,p(X_2|X_3,X_4)\,p(X_1|X_2,X_3,X_4)$

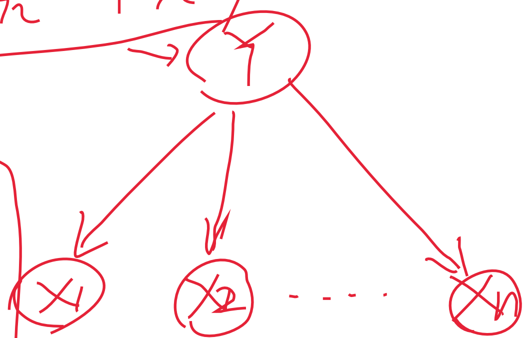
$1 \to 2 \to 3 \to 4$:



What is the Bayes Network for Naïve Bayes?

$(X_i \perp\!\!\!\perp X_j \mid Y, \ \forall i \neq j)$

	$Y=1$	$Y=0$
$P(Y)$	π	$1-\pi$



$P(X_1 \mid Y)$	$X_1=1$	$X_1=0$
$Y=0$	θ_0	$1-\theta_0$
$Y=1$	θ_1	$1-\theta_1$

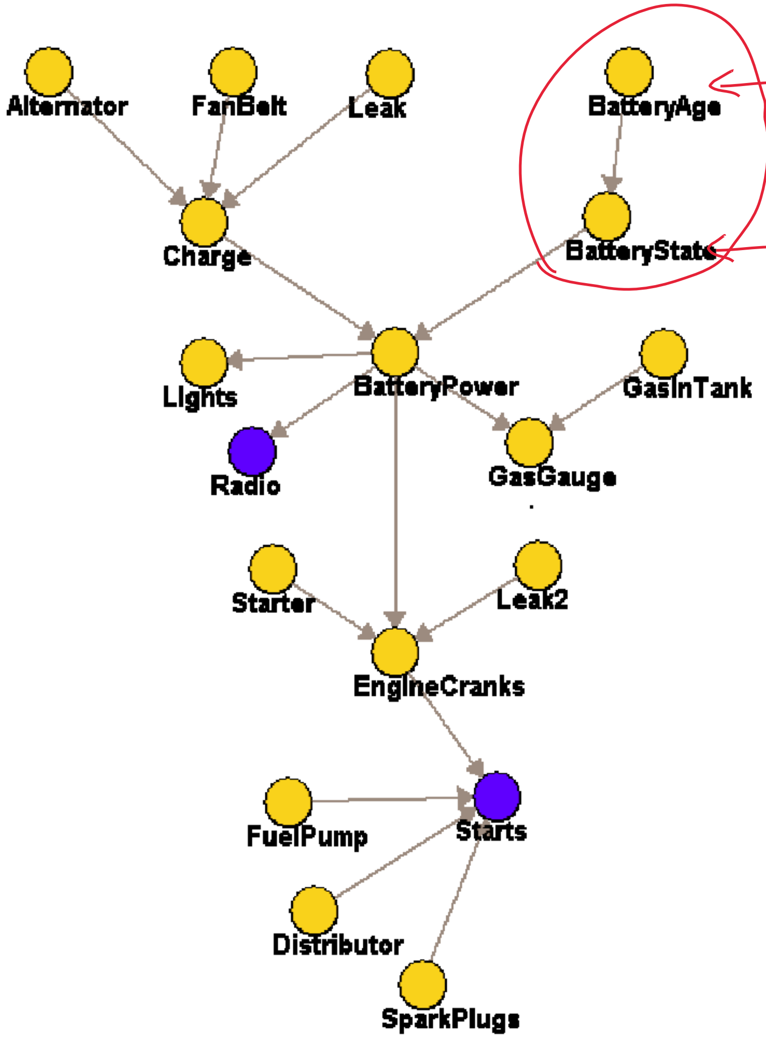
$$P(X_1, X_2, \ldots, X_n, Y)$$
$$= P(Y)\,P(X_1 \mid Y)\,P(X_2 \mid Y) \cdots P(X_n \mid Y)$$
$$= P(Y) \prod_{i=1}^{n} P(X_n \mid Y)$$

$$D = \{(x_j, y_j)\}_{j=1}^{m}$$

$$\ell(\theta; \pi) = \ell(\theta) + \ell(\pi)$$

$$\text{MLE} \begin{cases} \dfrac{\partial \ell(\theta)}{\partial \theta} = 0 \Rightarrow \hat{\theta}_0 = \dfrac{|D_{Y=0, X_1=1}|}{|D_{Y=0}|}, \quad \hat{\theta}_1 = \dfrac{|D_{Y=1, X_1=1}|}{|D_{Y=1}|} \\[2ex] \dfrac{\partial \ell(\pi)}{\partial \pi} = 0 \Rightarrow \hat{\pi} = \dfrac{|D_{Y=1}|}{|D|} \end{cases}$$

What do we do if variables are mix of discrete and real valued?



BatteryAge ← continuos

BatteryState ← discrete

P(BS|BA)	BS=0	BS=1
BA=0.1	θ_1	
BA=0.15	θ_2	
BA=0.2	θ_3	
⋮	⋮	

(infinite rows → intractable)

Solutions

① Discretization

$BA \in \{1, 2, 3, 4, 5\}$

BA	1	2	3	4	5
range	0–10	10–20	20–30	30–40	40–50

② Parameterized distribution

$$P(BS \mid BA) = \sigma(\beta \cdot BA + \beta_0) = \frac{1}{1 + e^{-(\beta BA + \beta_0)}}$$

- CPD of BS can be calculated by β and β_0