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

Edge

JDT $\Rightarrow p(Y|X)$

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 $= \int p(x_2|x_1) f(x_2) dx_2$
  - Observed training data

$\mathbb{E}_{p(x_2|x_1)}[f(x_2)]$

- Principled and ~general methods for
  - Probabilistic inference $p(x_1, x_2, x_3)$, $p(x_2|x_1)$, $p(x_2)$
  - Learning → Parameter → MLE/MAP

- 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 <u>marginally independent</u> of Y if

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

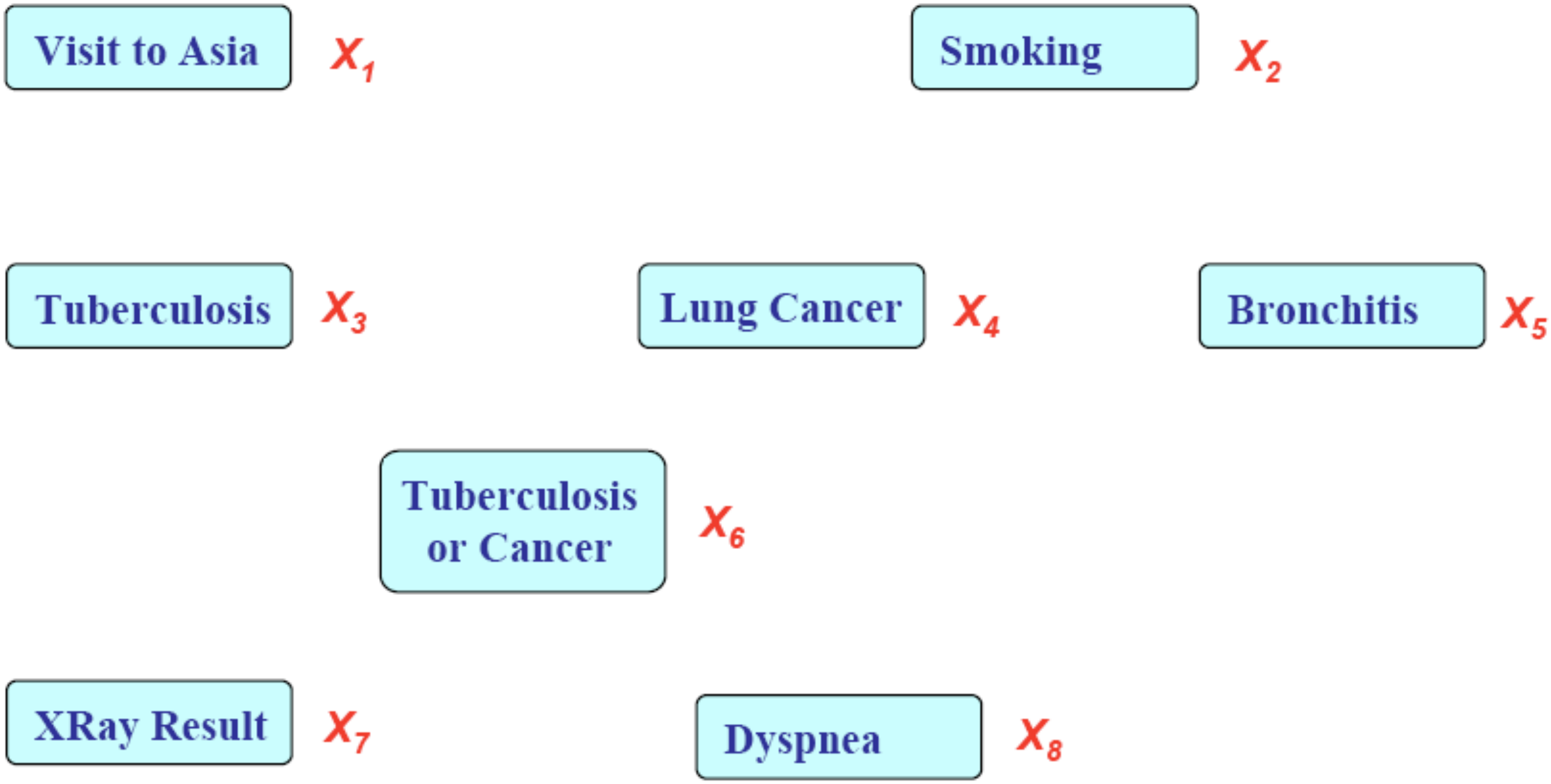
Equivalently, if

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

Equivalently, if

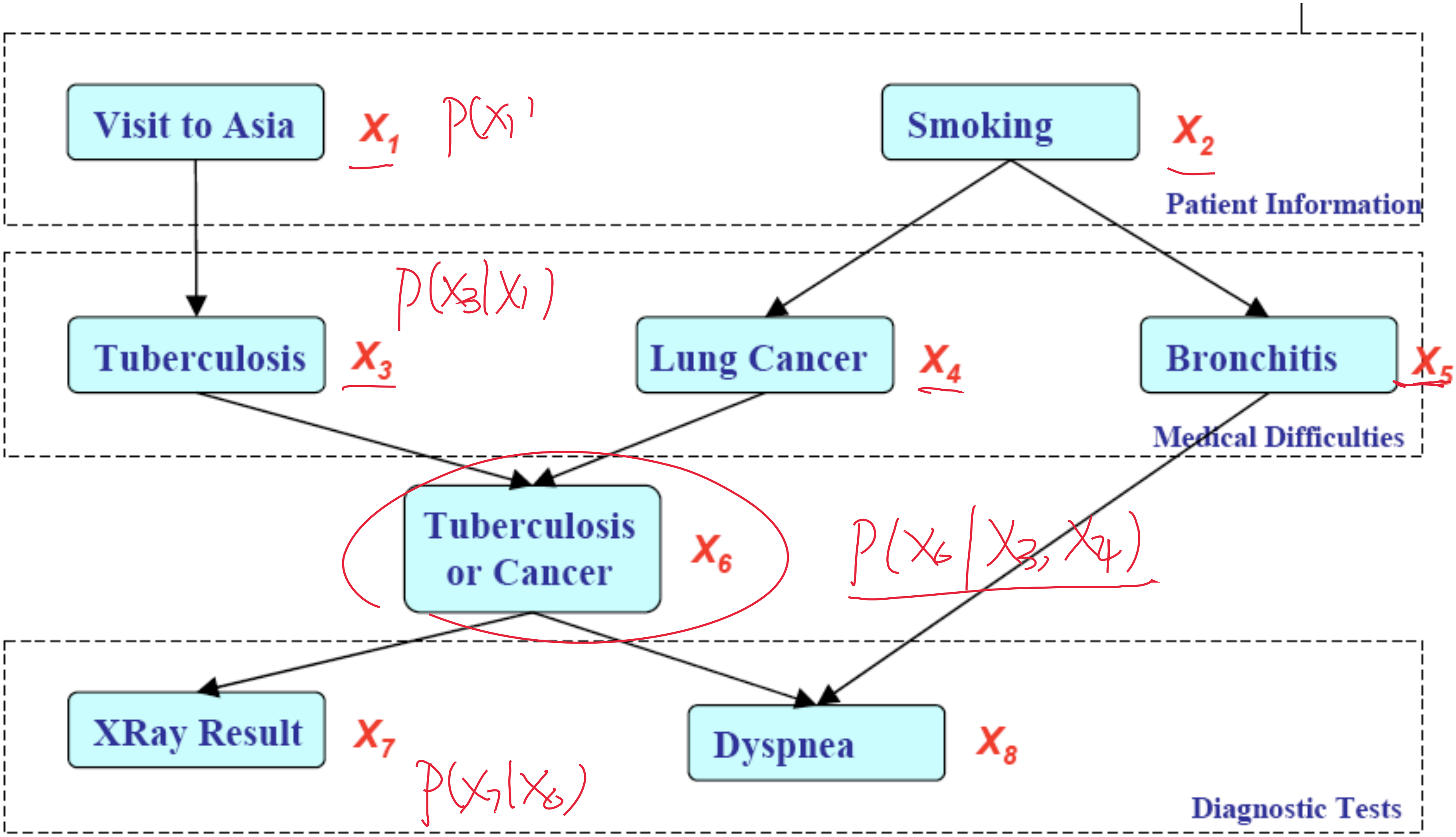
$$(\forall i, j) P(Y = y_i | X = x_j) = P(Y = y_i)$$

Represent Joint Probability Distribution over Variables



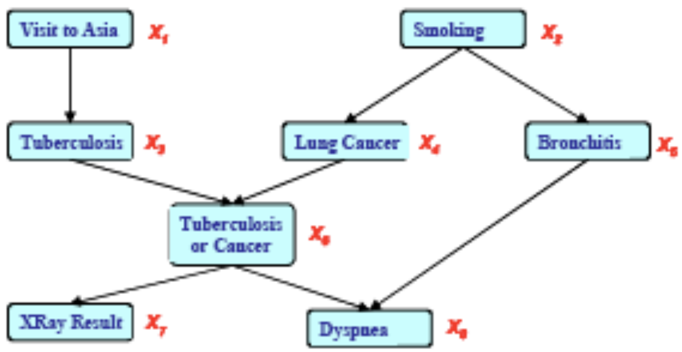
$G = \langle V, E \rangle$

Describe network of dependencies



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

Chain rule



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

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

2^2

$P(X_1)\, P(X_2|X_1)\, P(X_3|X_1,X_2) \ldots$

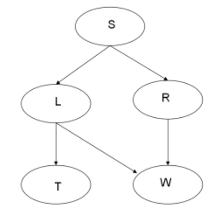
$P(X_8|X_1,X_2,\ldots,X_7)$

2^7

Benefits of Bayes Nets:

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

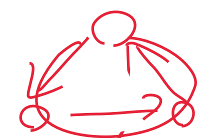
Bayesian Networks Definition



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

W

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




BN (DAG)

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

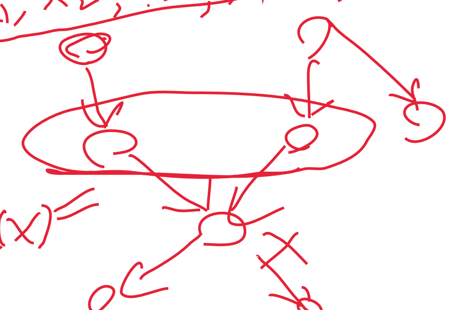
- Each node denotes a random variable
- Edges denote dependencies
- For each node X_i its CPD defines $P(X_i \,/\, Pa(X_i))$
- The joint distribution over all variables is defined to be

BN < DAG, CPD

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

$= p(X_1)\, p(X_2 | X_1) \ldots p(X_n | X_1, X_2, \ldots, X_{n-1})$

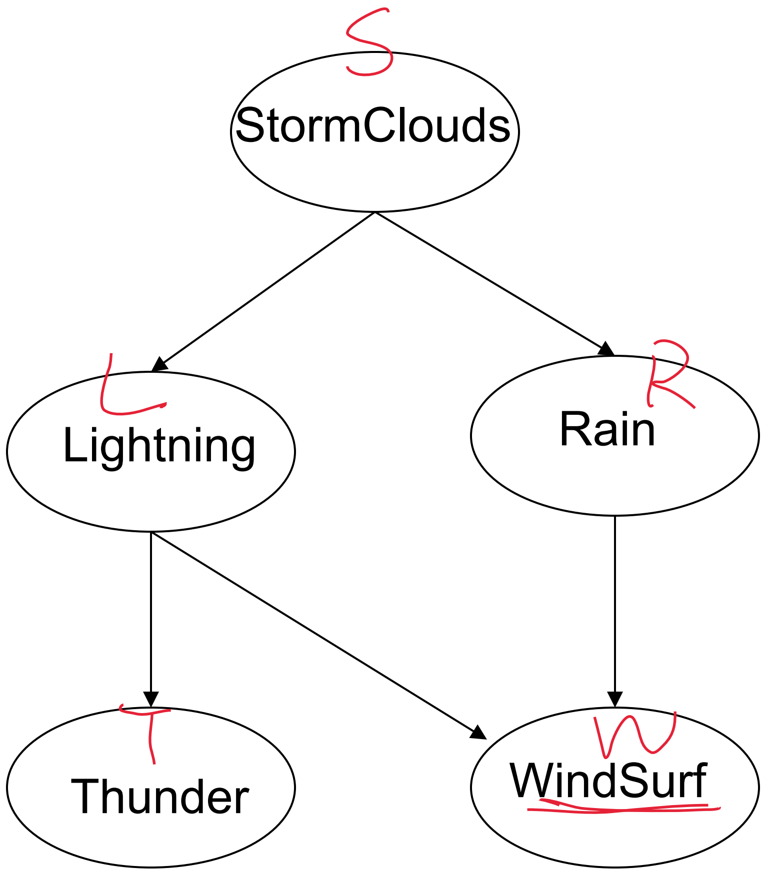
Pa(X) = immediate parents of X in the graph



Pa(x)=

Bayesian Network

#para (CPD) $= 2^{|Pa(x)|}$



P(X | Pa(X))

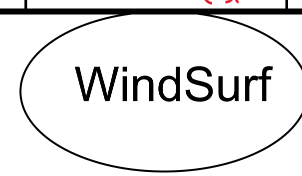
P(W | L,R)

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_{11}	1.0
L, ¬R	0 θ_{10}	1.0
¬L, R	0.2 θ_{01}	0.8
¬L, ¬R	0.9 θ_{00}	0.1

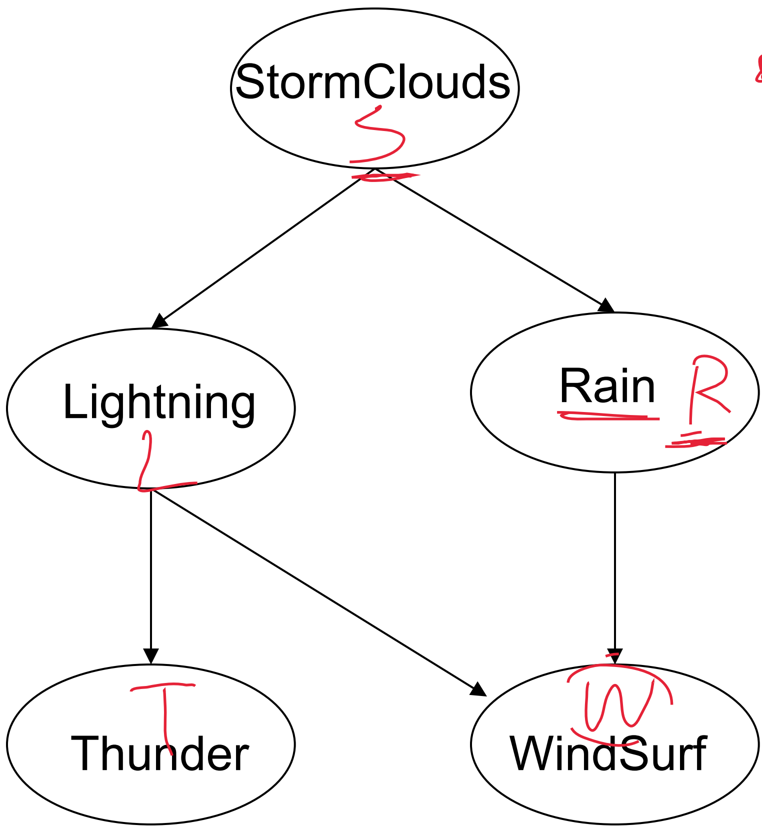
$2^2 = 4$



The joint distribution over all variables:

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

Bayesian Network



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

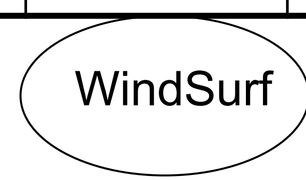
One thing is this:

X

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

Pa(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



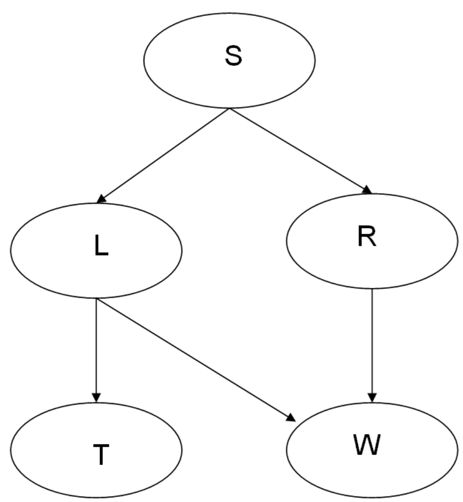
Some helpful terminology

Parents = Pa(X) = immediate parents

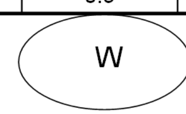
An(x) Antecedents = parents, parents of parents, ...

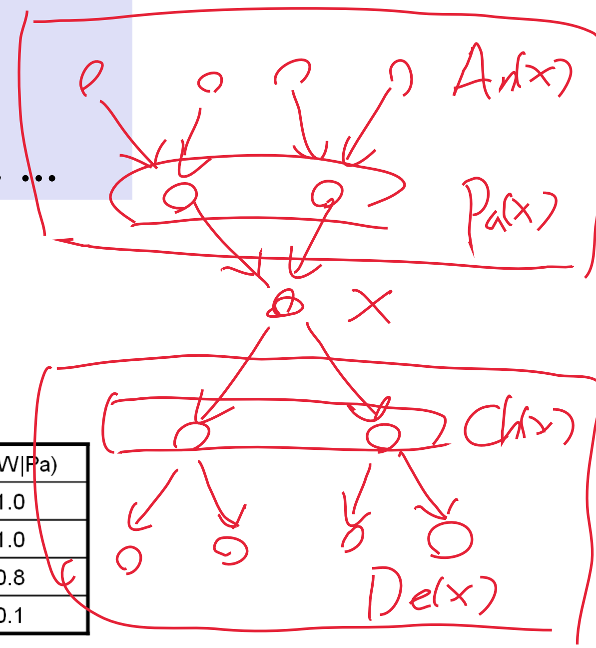
Ch(x) Children = immediate children

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



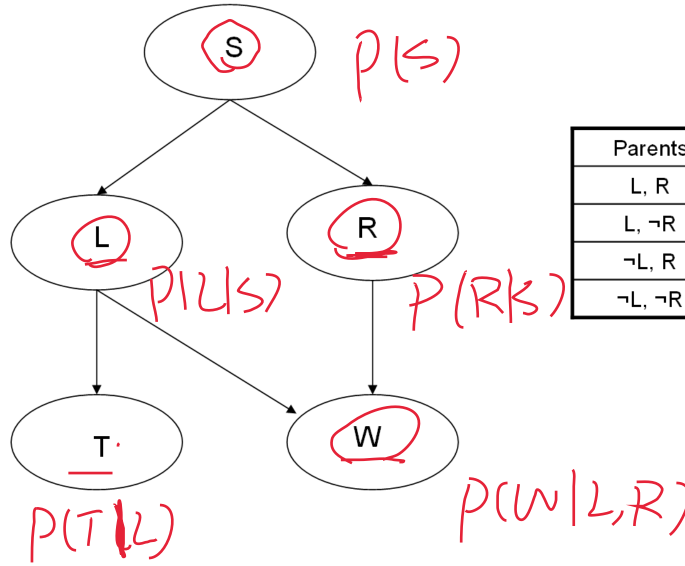
Parents	P(W|Pa)	P(¬W|Pa)
L, R	0	1.0
L, ¬R	0	1.0
¬L, R	0.2	0.8
¬L, ¬R	0.9	0.1



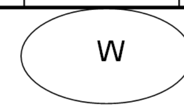


Bayesian Networks

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



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:

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

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

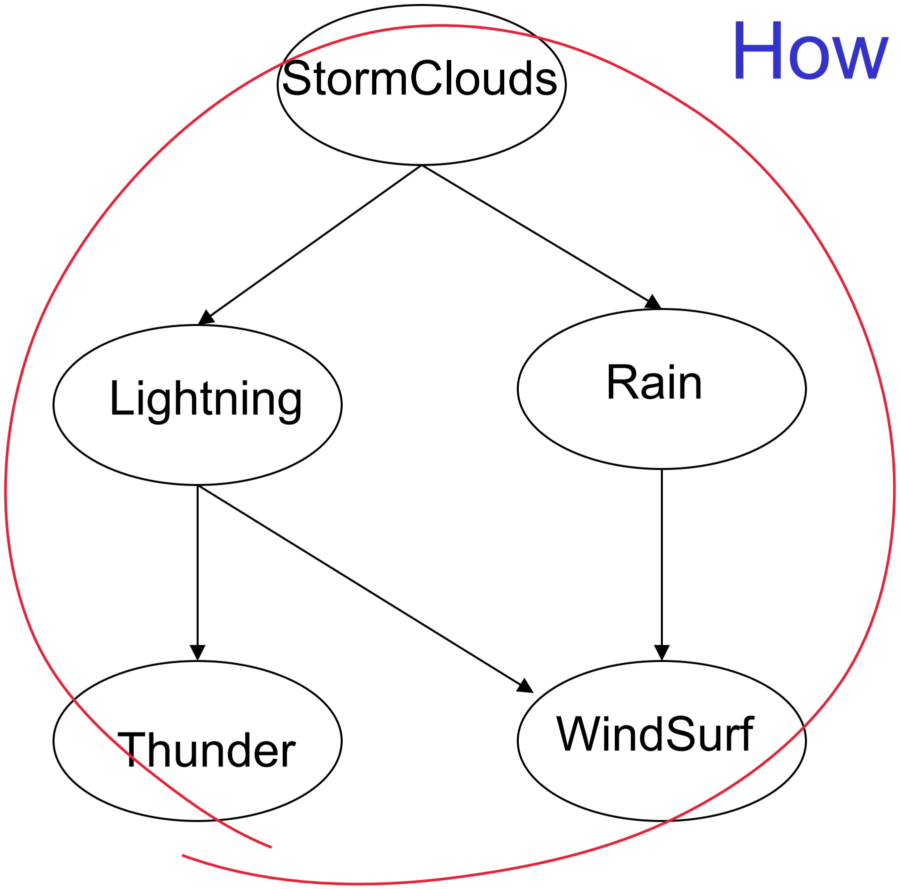
But in a Bayes net: $P(X_1 \dots X_n) = \prod_i P(X_i|Pa(X_i))$

$2^0 = 1$ $\quad 2^1 \quad 2^2$

$1 + 3 \times 2 + 4 = 11$

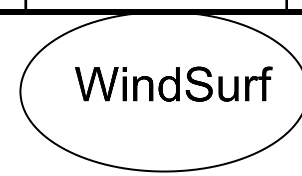
$P(R|S,L) = P(R|S)$

- $P(R,L|S) = P(R|S)\,P(L|S)$



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



2^n

$2^5 - 1 = 31$

To define joint distribution in general? $P(S, L, R, T, W)$

To define joint distribution for this Bayes Net? 11

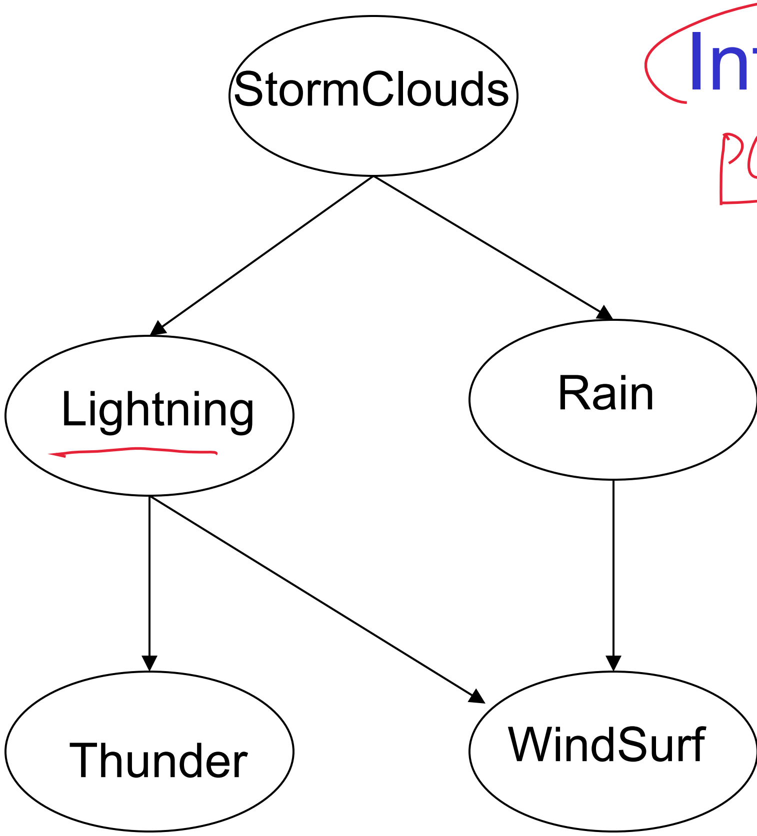
$2^1 n$ 1-st

$2^2 \cdot n$ 2-nd

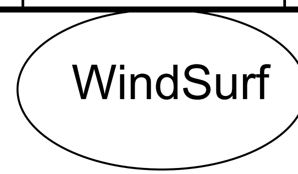
$2^3 \cdot n$ 3-rd

Inference in Bayes Nets

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

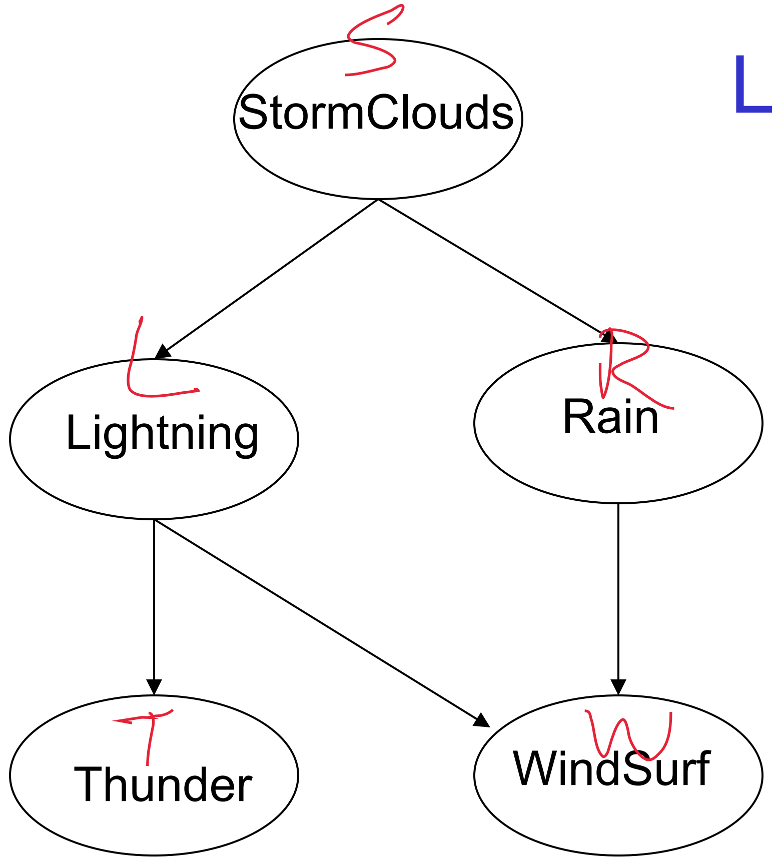


Parents	P(W|Pa)	P(¬W|Pa)
L, R	0	1.0
L, ¬R	0	1.0
¬L, R	0.2	0.8
¬L, ¬R	0.9	0.1



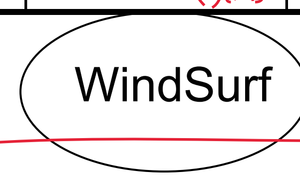
- $P(S=1, L=0, R=1, T=0, W=1)$ = $P(S=1) \cdot P(L=0|S=1)\, P(R=1|S=1)\, P(T=0|L=0)\, P(W=1|L=0,R=1)$
- $P(S=1|L=0,R=1,T=0,W=1) = \dfrac{P(S=1,L=0,R=1,T=0,W=1)}{P(L=0,R=1,T=0,W=1)}$
- $P(S=1) = \sum_{l,r,t,w} P(S=1, R=r, T=t, W=w)$ $= \sum_{s=0,1} P(S=s, L=0, R=1, T=0, W=1)$

Learning a Bayes Net



CPD

Parents	P(W|Pa)	P(¬W|Pa)
L, R	0 θ_{11}	1.0
L, ¬R	0 θ_{10}	1.0
¬L, R	0.2 θ_{01}	0.8
¬L, ¬R	0.9 θ_{00}	0.1



$$P(W=w|L=1,R=1) = \theta_{11}^{w}\,(1-\theta_{11})^{1-w}$$

Consider learning when graph structure is given, and data = $\{ <s,l,r,t,w> \}_{i=1}^{n}$ D

What is the MLE solution? MAP?

$P(\theta) \sim \text{Beta}$

1. PDF
2. Likelihood: $L(\theta) = P(D|\theta) = \prod_{i=1}^{n} P(x_i|\theta)$
3. $\frac{\partial L(\theta)}{\partial \theta} = 0$

$$P(X_1,X_2,\ldots,X_n) = P(X_1)P(X_2|X_1)\cdots P(X_n|X_1,\ldots,X_{n-1}) \leftarrow (n!)$$

DAG

$$= P(X_n)P(X_{n-1}|X_n)\cdots P(X_1|X_n,X_{n-1},\ldots,X_2)$$

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

$(X_1, X_2, \ldots, X_{i-1})$

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

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

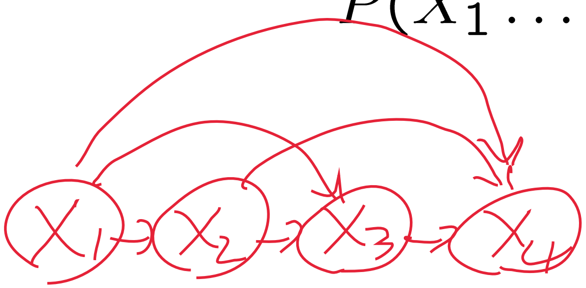
$A \perp B | C$

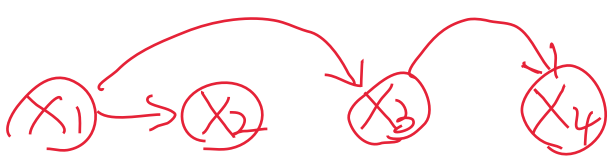
$P(A,B|C) = P(A|C)\,P(B|C)$

Notice this choice of parents assures

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

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



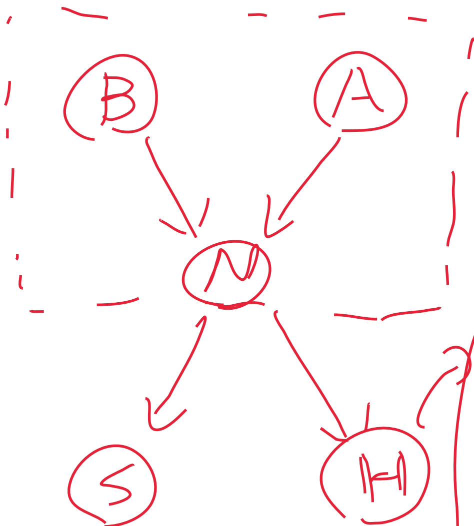


Example

B A N S H

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

BN: DAG, CPD



inference. $P(N=1)$

$$= \sum_{a,b,s,h} P(N=1, B=b, A=a, S=s, H=h)$$

$P(H \mid N)$	$H=1$	$H=0$
$N=1$	θ_1	$1-\theta_1$
$N=0$	θ_0	$1-\theta_0$

$$D = \{(b_i, a_i, n_i, s_i, h_i)\}_{i=1}^{n}$$

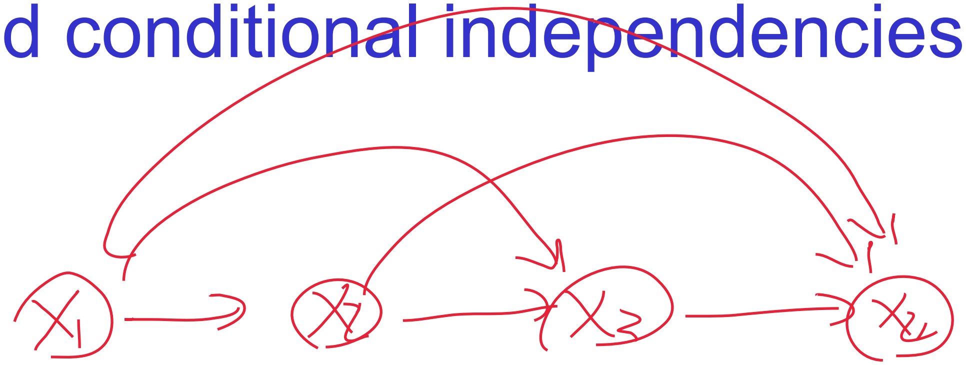
① PDF
② Likelihood $L(\theta)$
③ $\frac{\partial L(\theta)}{\partial \theta} = 0$

Explaining away

$B \perp A$

$B \not\perp A \mid N$

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



Fully-connected BN

4!

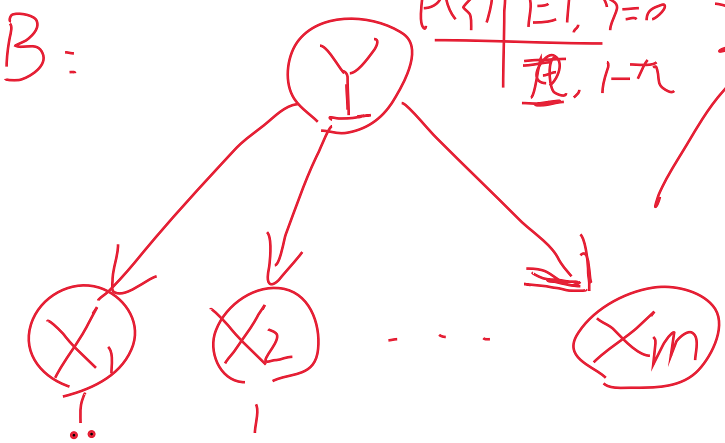
What is the Bayes Network for Naïve Bayes?

$$P(X, Y) = \underline{P(X|Y)}\, P(Y)$$

$$= \prod_{j=1}^{m} \underline{P(X_j|Y)}\, \underline{P(Y)}$$

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

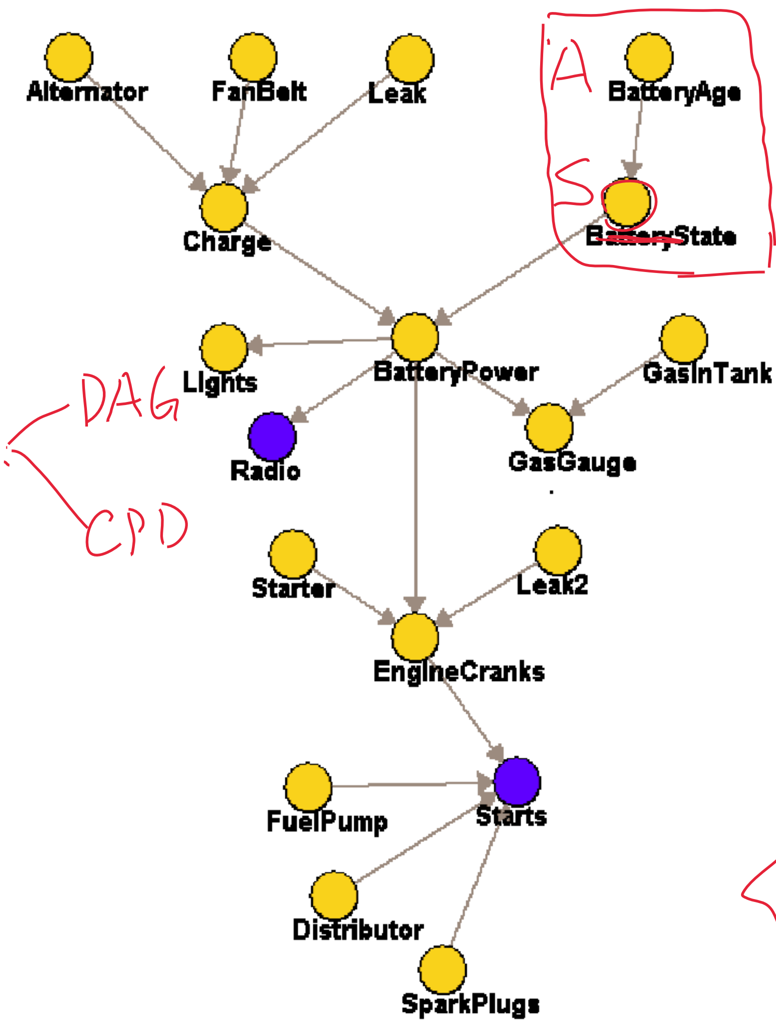
NB:



P(Y)	Y=1	Y=0
	π	$1-\pi$

P(X|Y)	X=1	X=0
Y=1	θ_1	$1-\theta_1$
Y=0	θ_0	$1-\theta_0$

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



BN: DAG, CPD

$A \in [0, 10]$

$S \in \{0, 1\}$

$P(S \mid A)$	$S=1$	$S=0$
$A=0$	θ_1	$1-\theta_1$
$A=0.1$	θ_2	$1-\theta_2$
$A=0.2$		
⋮	⋮	⋮
$A=10$	θ_{inf}	$1-\theta_{inf}$

① A: discretization.

$A \in [0, 3]$: 1

$\in [3, 6]$: 2

$\in [6, 10]$: 3

Logistic

$$P(Y=1 \mid X=x) = \sigma(\beta^T x)$$

Parametric model

② $$P(S=1 \mid A=a) = \frac{1}{1+e^{-\beta^T a}}$$

sigmoid $$\sigma(x) = \frac{1}{1+e^{-\beta^T x}}$$