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

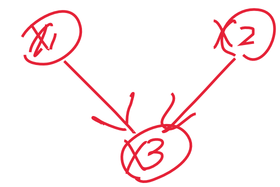
Readings:

- Bishop chapter 8, through 8.2

Graphical Models

$G = \langle V, E \rangle$

RV Dependency



$P(X_1, X_2, X_3)$

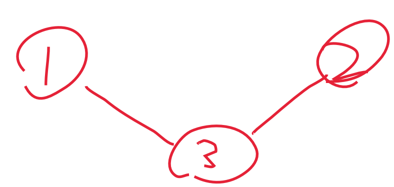
$P(x_1, x_2, x_3)$

$2^3 = 8$

- Key Idea:
  - Conditional independence assumptions useful
  - but Naïve Bayes is extreme!
  - Graphical models express sets of conditional independence assumptions via graph structure
  - Graph structure plus associated parameters define *joint probability distribution over set of variables*

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

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

$P(Y|Z)$

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

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

$P(X|Z) = \sum_{i=1}^{n} P(X_i|Z)$

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

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

Marginal Independence

Definition: X is marginally independent of Y if

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

$$\underline{P(X,Y)} = P(X) \cdot P(Y) \qquad \underline{P(X)} = \prod_{i=1}^{n} P(X_i)$$

$$P(X|Y)\,P(Y) \Rightarrow P(X|Y) = P(X)$$

Equivalently, if

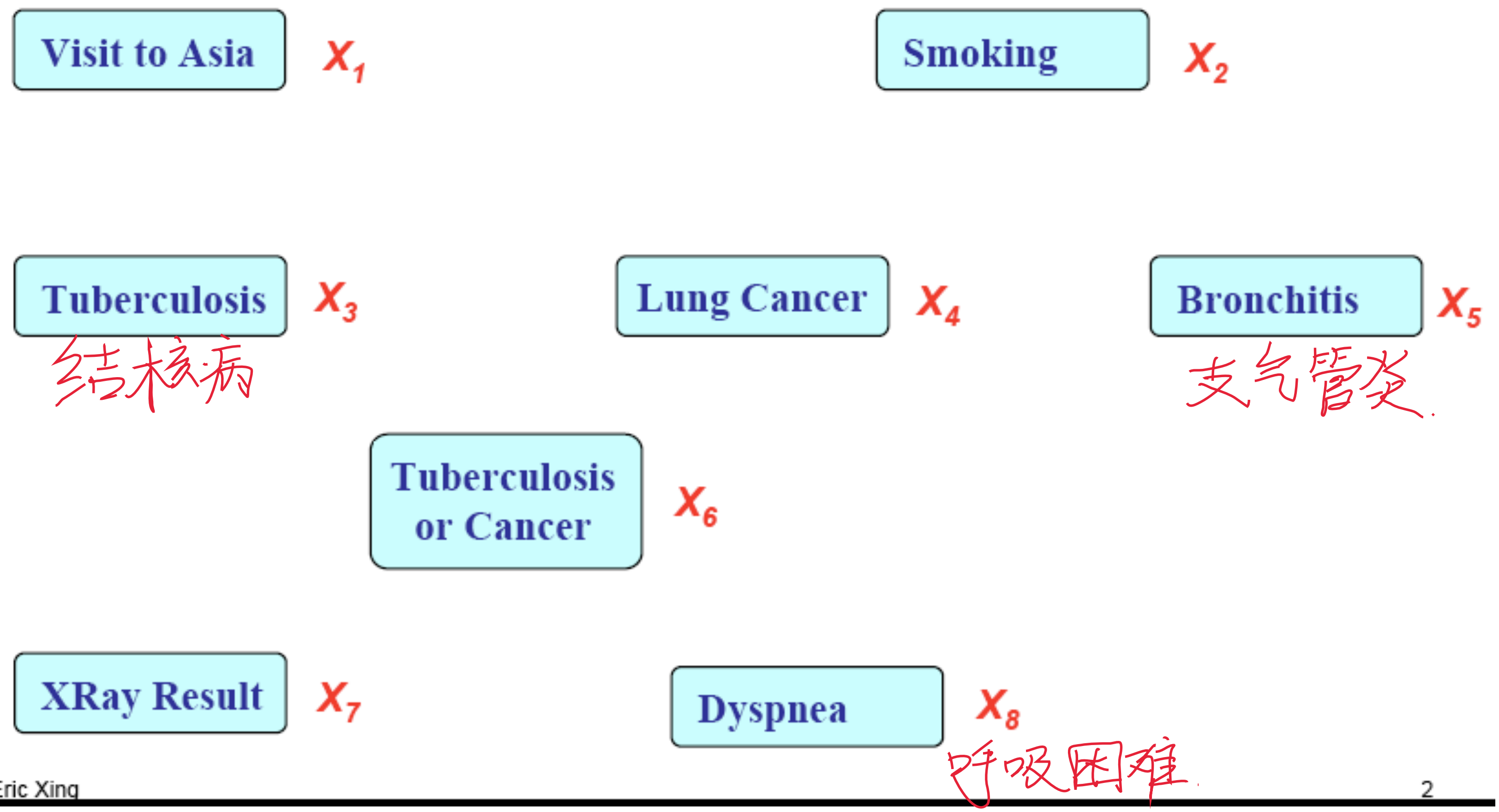
$$P(Y|X)\,P(X) \Rightarrow P(Y|X) = P(Y)$$

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

Equivalently, if

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

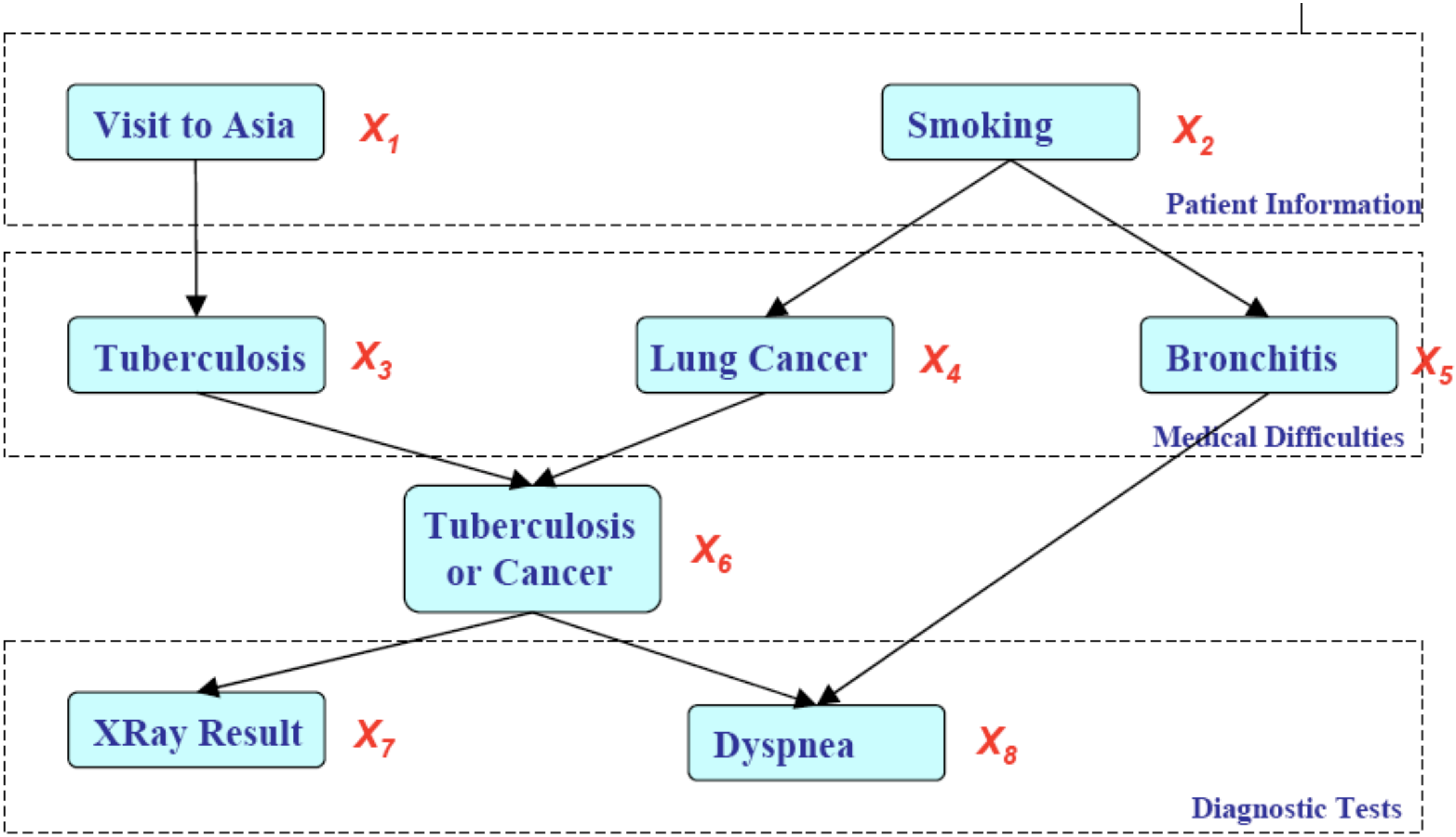
Represent Joint Probability Distribution over Variables



$$\boxed{P(X_1, X_2, \ldots, X_8)}$$

$$P(X_8 \mid X_1, X_2, \ldots, X_7) = \frac{P(X_8, X_1, X_2, \ldots, X_7)}{P(X_1, X_2, \ldots, X_7)}$$

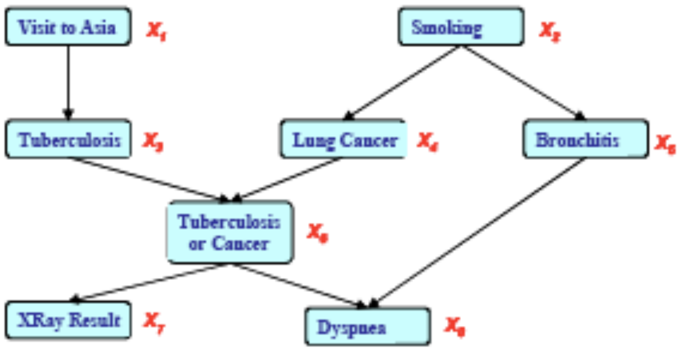
Describe network of dependencies



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

$P(X_2 | X_1) = P(X_2)$

$X_1 \perp\!\!\!\perp X_2$



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

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

$$P(X_1, X_2, \ldots, X_8) = P(X_1)\, \underline{P(X_2 | X_1)}\, \underline{P(X_3 | X_1, X_2)} \cdots P(X_8 | X_1, \ldots, X_7)$$

Benefits of Bayes Nets:

$P(X_3 | X_1) = P(X_3 | X_1, X_2)$

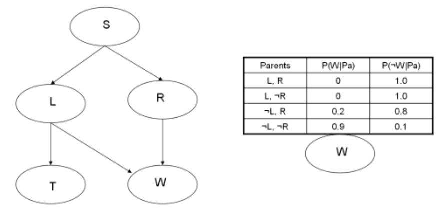
$X_2 \perp\!\!\!\perp X_3 \mid X_1$

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

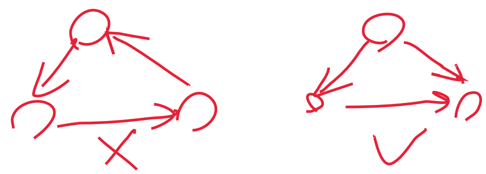
$P(X_8 | X_1, \cancel{X_2}, \ldots, X_7)$

$\underline{P(X_8 | X_5, X_6)}$

Bayesian Networks Definition



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



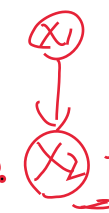
BN (DAG)

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

BN — DAG: $G = \langle V, E \rangle$; CPD (CPT)

CPD	$X_2=1$	$X_2=0$
$X_1=1$	θ_1	$1-\theta_1$
$X_1=0$	θ_0	$1-\theta_0$

- 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



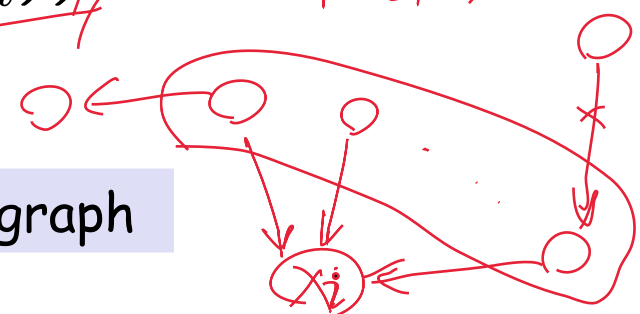
$P(X_2|X_1)$

$$\text{BN} \to P(X_1 \ldots X_n) = \prod_i P(X_i | Pa(X_i))$$

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

$= \prod_{i=1}^{n} P(X_i | Pa(X_i))$

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

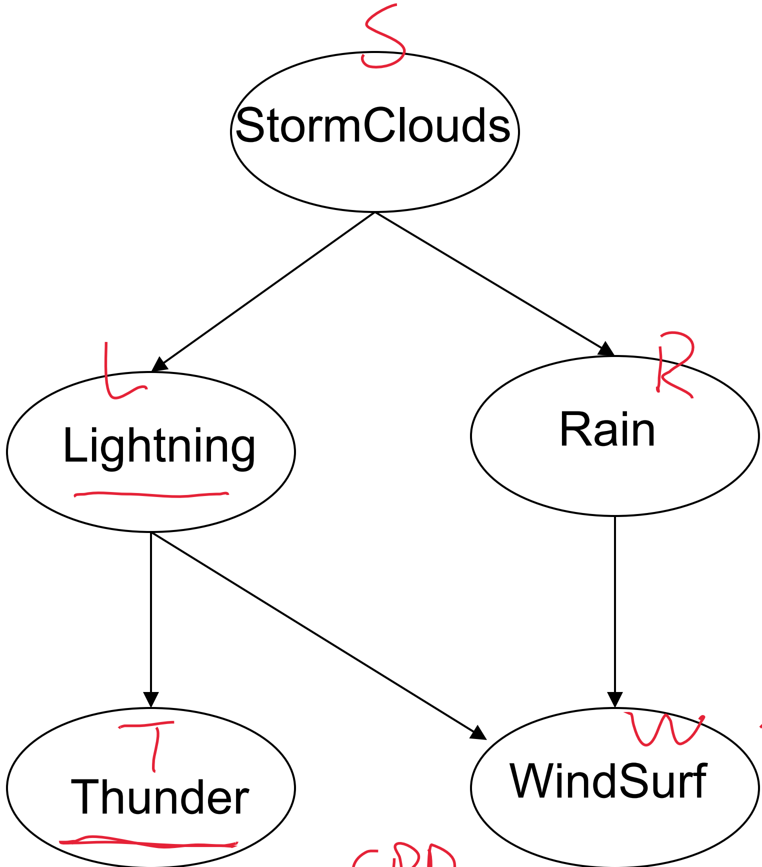


$Pa(X_i)$

Bayesian Network

DAG
CPD

boolean



CPD:
P(T|L)

	T=1	T=0
L=1	θ_1	$1-\theta_1$
L=0	θ_0	$1-\theta_2$

#params=2

CPD
↓ P(W|L,R)
$2^2=4$

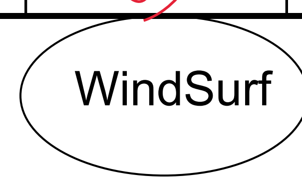
Nodes = random variables

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

p 1−p

#params =4

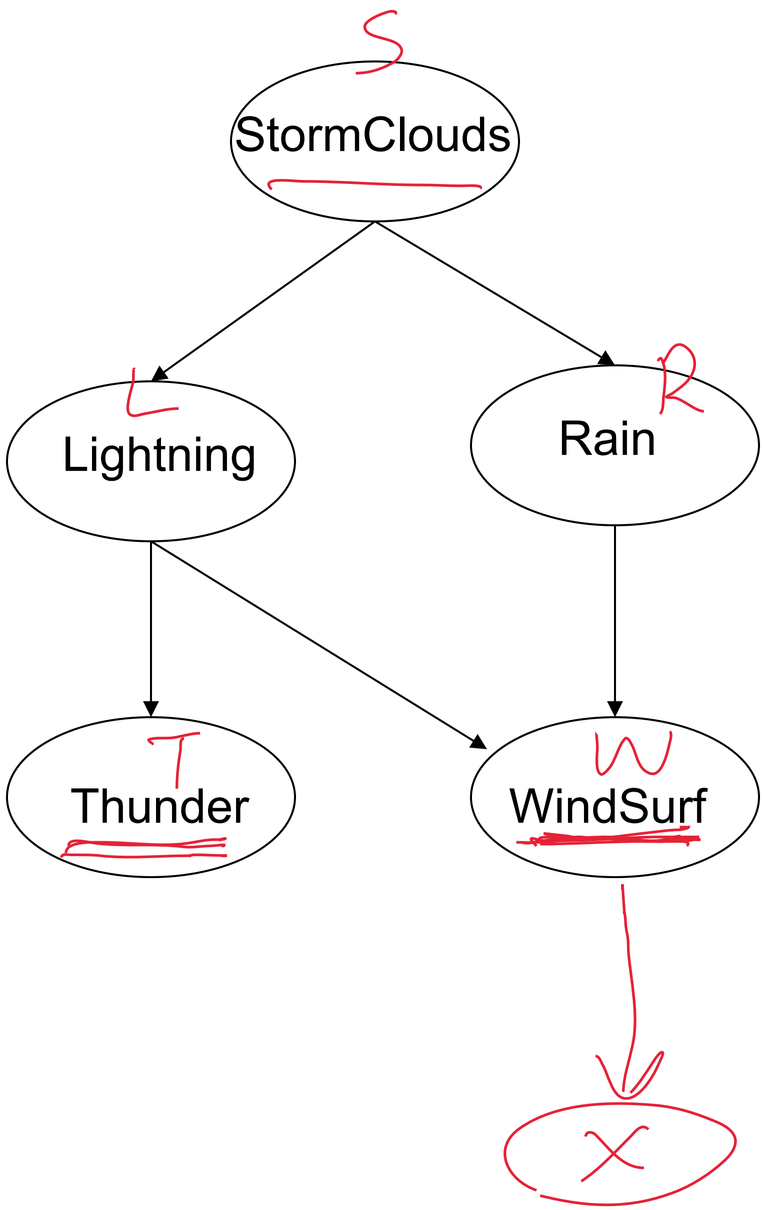
Parents	P(W|Pa)	P(¬W|Pa)
L, R	0	1.0
L, ¬R	0	1.0
¬L, R	0.2	0.8
¬L, ¬R	0.9	0.1



The joint distribution over all variables:

$$P(X_1 \ldots X_n) = \prod_i \underbrace{P(X_i|Pa(X_i))}_{\text{CPD}}$$

Bayesian Network



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.

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

Cond. independ. $P(X,Y|Z) = P(X|Z)\,P(Y|Z)$

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

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

$(P(T|L)$

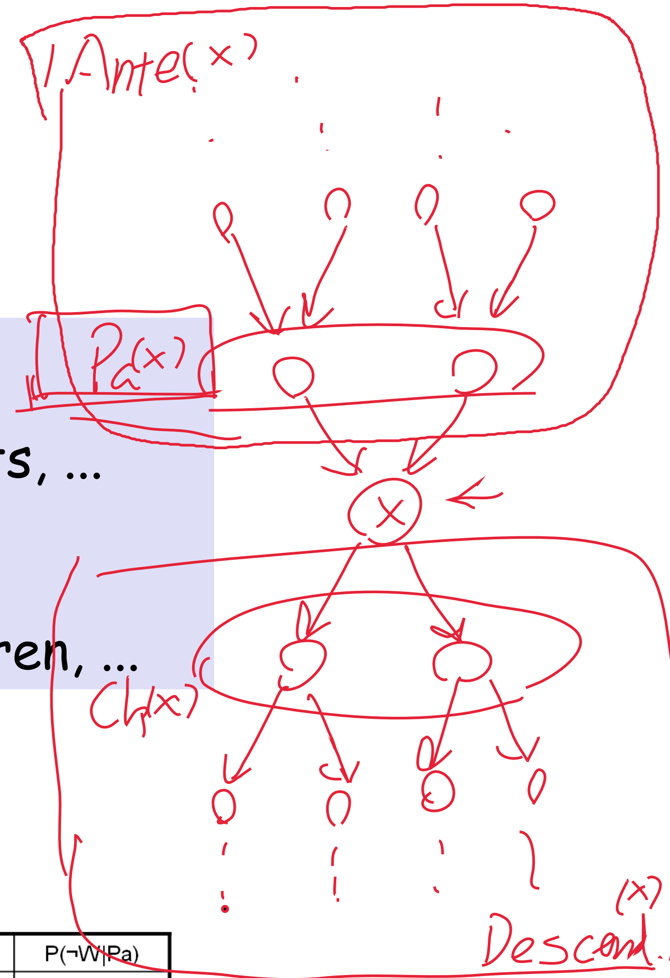
Some helpful terminology

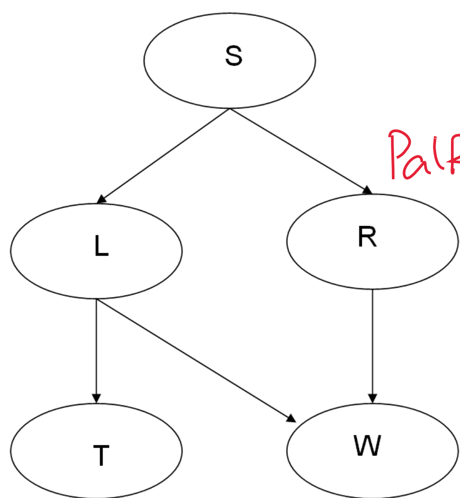
Parents = Pa(X) = immediate parents

Antecedents = parents, parents of parents, ...

Children = immediate children

Descendents = children, children of children, ...





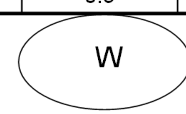
Pa(L) = S

Pa(R) = S

Pa(T) = L

Pa(W) = { L, R }

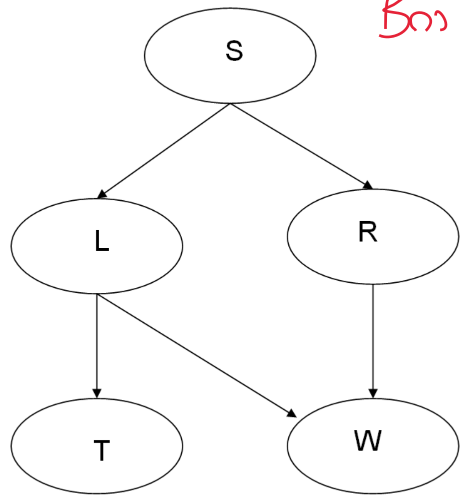
Parents	P(W|Pa)	P(¬W|Pa)
L, R	0	1.0
L, ¬R	0	1.0
¬L, R	0.2	0.8
¬L, ¬R	0.9	0.1



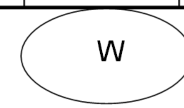
Bayesian Networks

Boolean

- 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



$\#\text{iparams} = 2^5 - 1 = 31$

Chain rule of probability says that in general:

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

no cond. independ.

$1 \quad 2 \quad 2 \quad 2 \quad 2^2 = 4$

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

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

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

cond. independ.

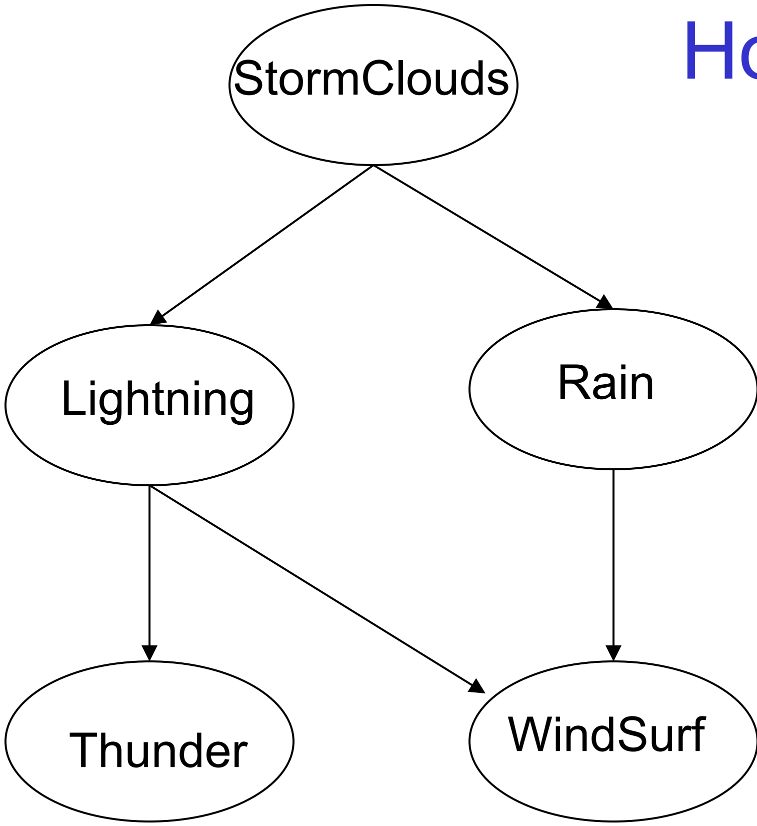
$P(R|S,L) = P(R|S)$

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

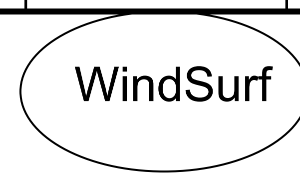
$R \perp\!\!\!\perp L \mid S$

$\#\text{iparams} = 11$

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

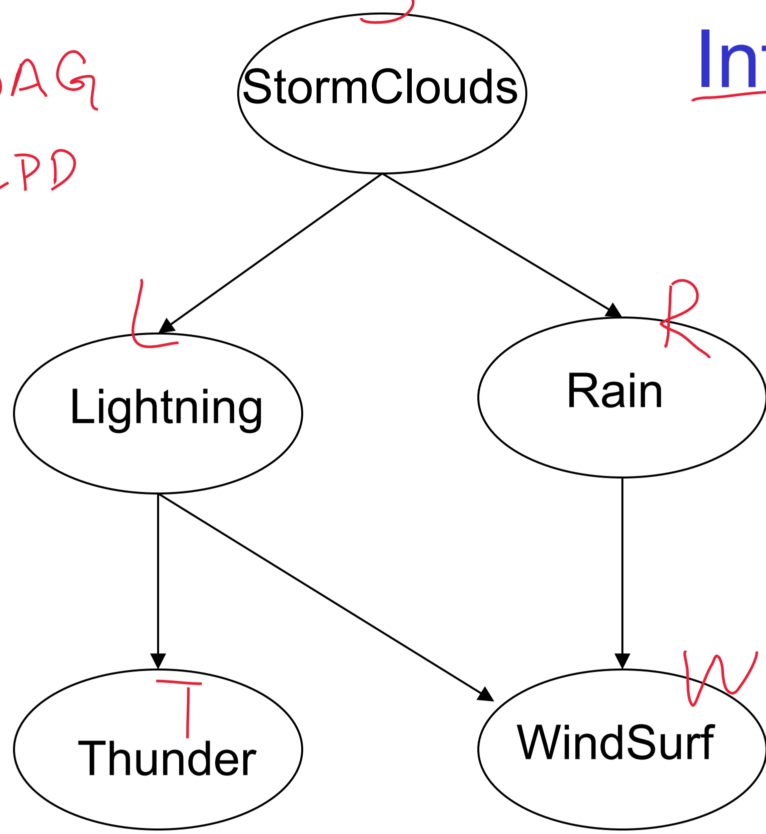


To define joint distribution in general?

To define joint distribution for this Bayes Net?

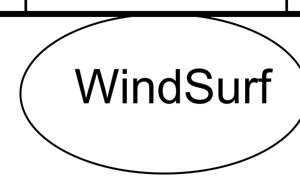
Inference in Bayes Nets

BN $<$ DAG, CPD



CPD

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



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

0.2

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

$$\overset{BN}{=} P(W=1 \mid L=1, R=0) = 0$$

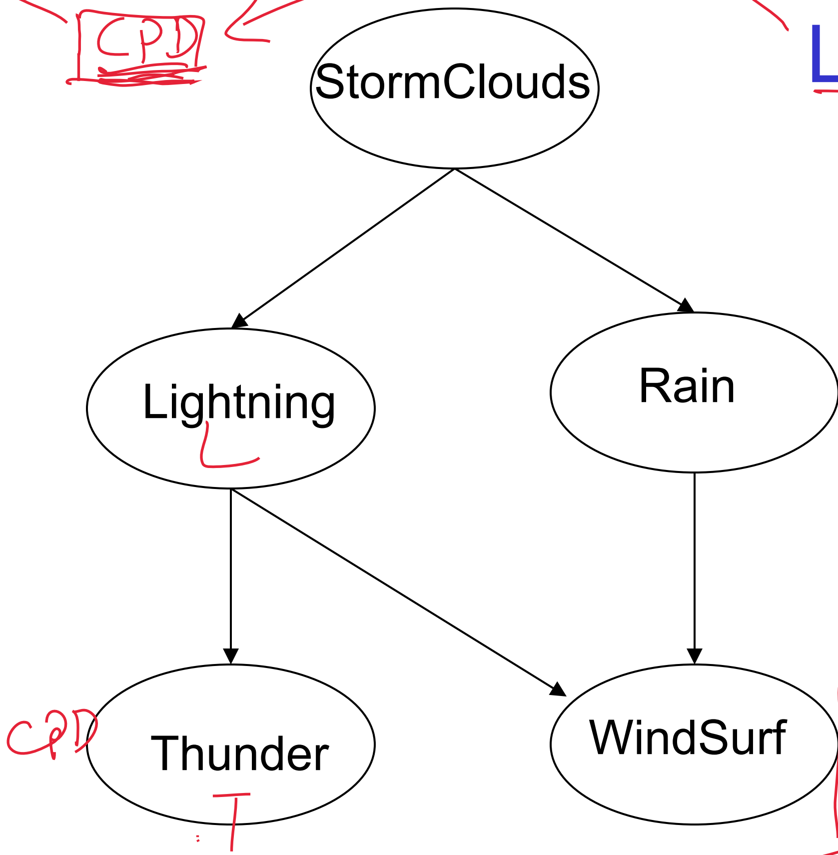
$$\boxed{W \perp \{S, T\} \mid \{L, R\}}$$

BN < DAG, CPD

Training Set: $D = \{x_1, x_2, \ldots, x_m\}$ $\quad x_m \in \mathbb{B}^5$

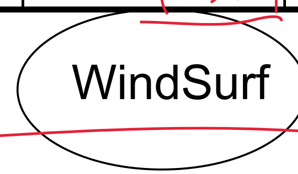
Learning a Bayes Net

$x^T = (S, L, R, T, W)$



params = 4

Parents	P(W|Pa)	P(¬W|Pa)
L, R	0 θ_{11}	1.0 $1-\theta_{11}$
L, ¬R	0 θ_{10}	1.0 $1-\theta_{10}$
¬L, R	0.2 θ_{01}	0.8 $1-\theta_{01}$
¬L, ¬R	0.9 θ_{00}	0.1 $1-\theta_{00}$



$$\hat{\theta}_1 = \frac{\alpha_{11} + \beta_{11}}{(\alpha_{11} + \alpha_{10}) + \beta_1} \qquad \hat{\theta}_0 = \frac{\alpha_{01} + \beta_{01}}{(\alpha_{01} + \alpha_{00}) + \beta_0}$$

$\theta = \langle \theta_1, \theta_0 \rangle$

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

What is the MLE solution? MAP?

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

$$P(T|L) = \theta_1^{[L=1]T} (1-\theta_1)^{[L=1](1-T)} \theta_0^{[L=0]T} (1-\theta_0)^{[L=0](1-T)}$$

$$L(\theta) = P(D|\theta)$$

$$\frac{\partial L(\theta)_n}{\partial \theta_1} = \theta_1^{\sum_{i=1}^{m} [L=1]T}_{\alpha_{11}} (1-\theta_1)^{\sum_{i=1}^{m} [L=1](1-T)}_{\alpha_{10}} \cdot \theta_0^{\alpha_{01}} \cdot (1-\theta_0)^{\alpha_{00}}$$

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

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

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

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

What is the Bayes Network for Naïve Bayes?

Q

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