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

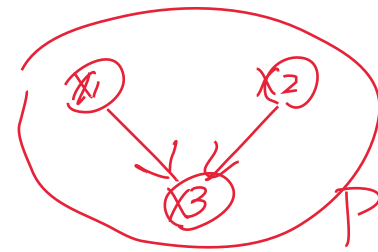
Readings:

- Bishop chapter 8, through 8.2

Graphical Models

$G = \langle V, E \rangle$

RV Dependency



$P(X_3 | X_1, X_2)$

$P(X_1, X_2, X_3)$

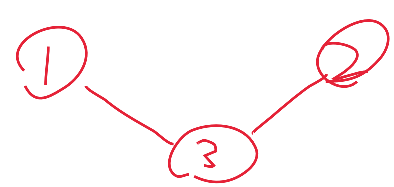
JD $P(x_1, x_2, x_3)$

$2^3 = 8$

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

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

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

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

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

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

$$\| $$

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

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

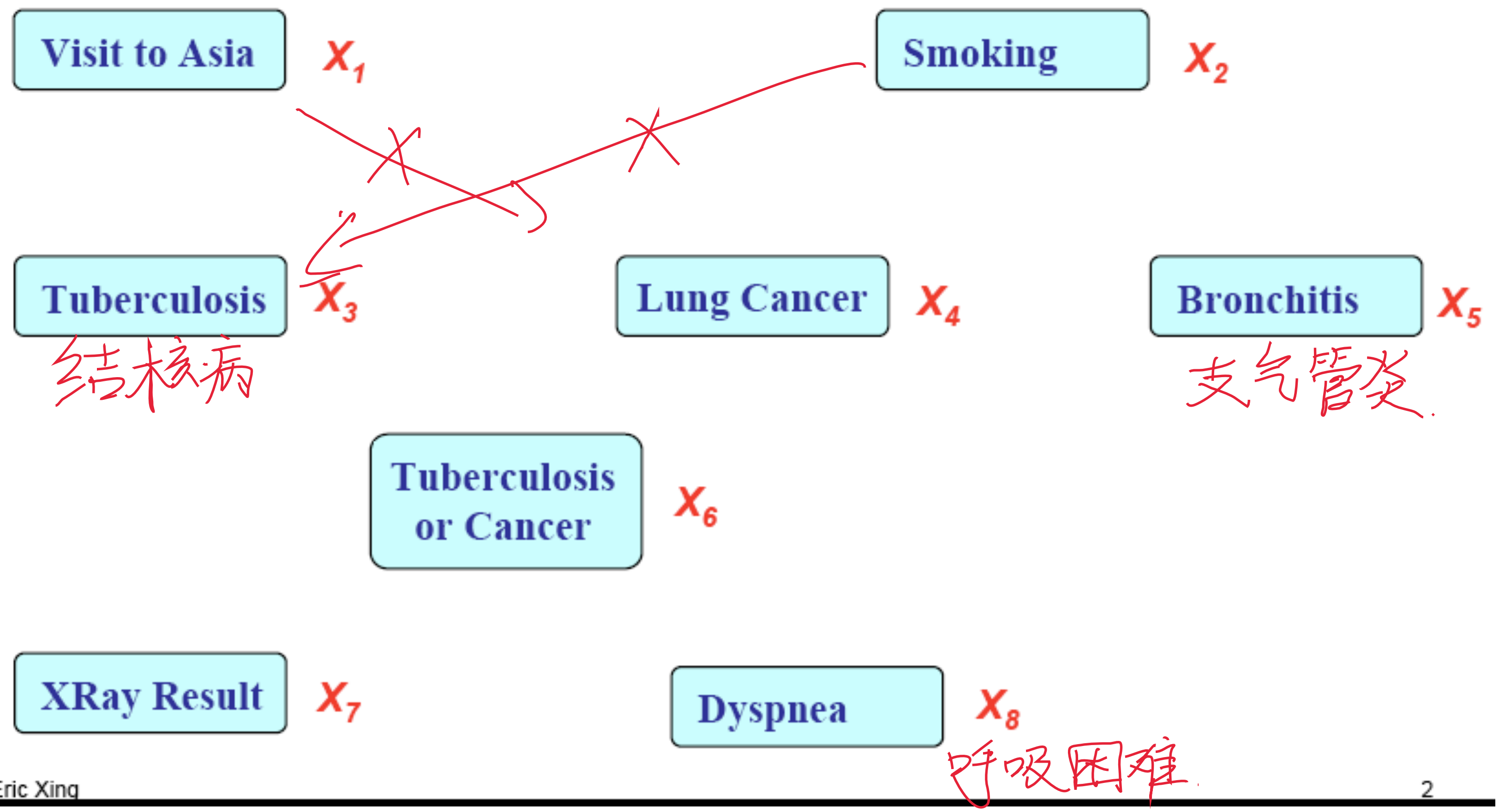
Equivalently, if

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

Equivalently, if

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

Represent Joint Probability Distribution over Variables



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

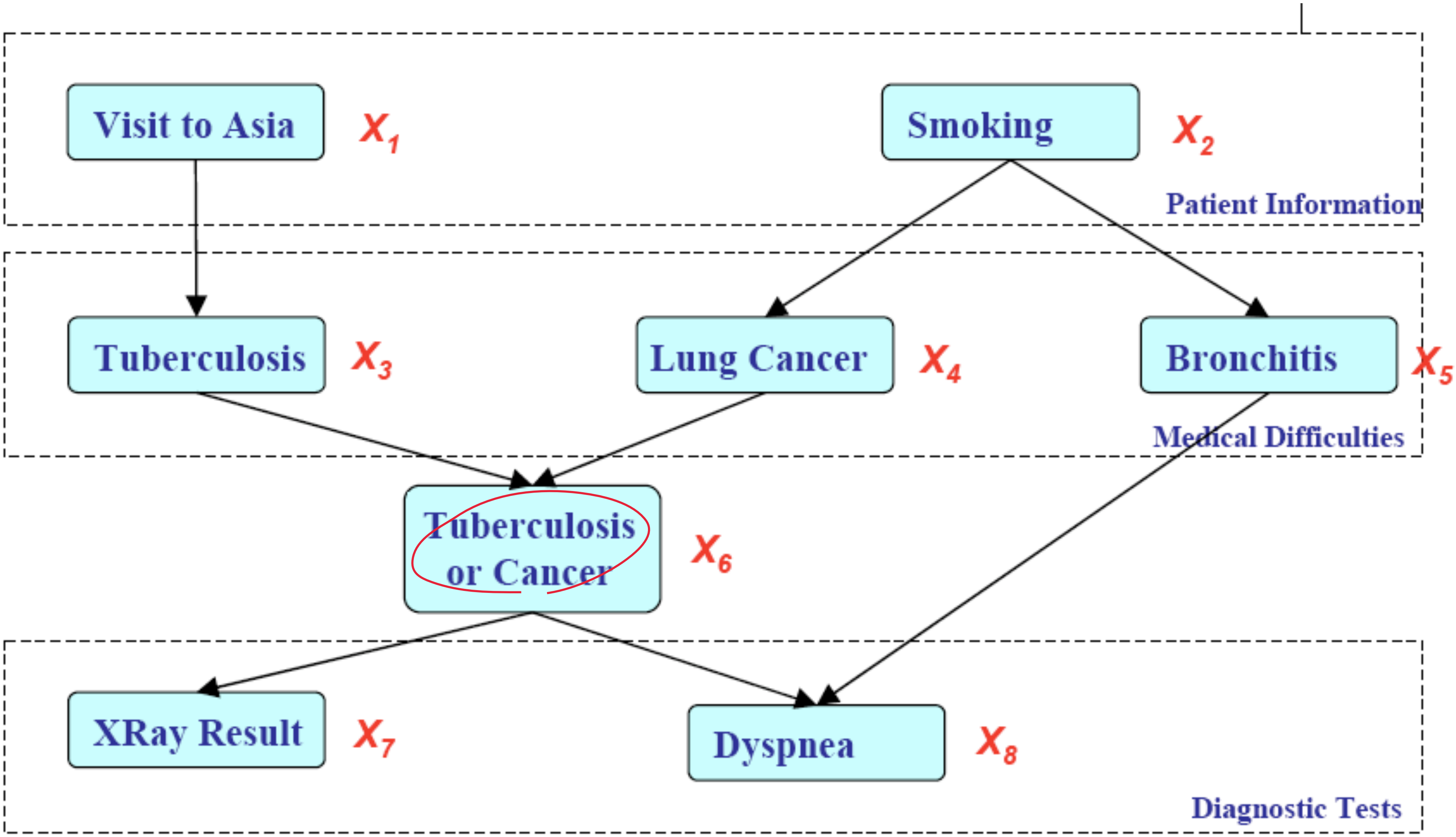
$$P(X_8 \mid X_1, X_2, \ldots, X_7)$$

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

$$p(x_1, x_2, \ldots, x_8) = p(x_1)\, p(x_2|x_1) \cdots p(x_8|x_1, x_2, \ldots, x_7)$$

$(1 \to 2 \to \cdots \to 8)$

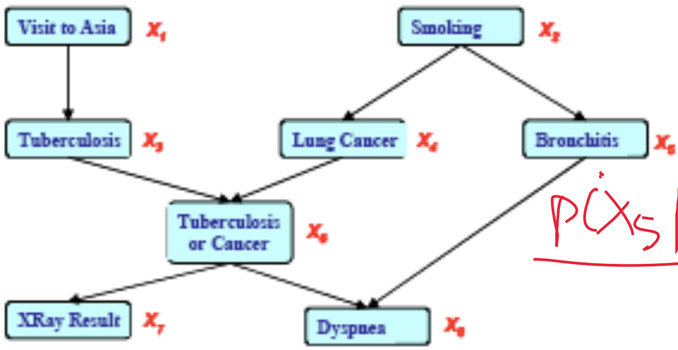
Describe network of dependencies



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

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

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



$p(X_5 | X_2)$

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

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

- $P(X_1, X_2, \ldots, X_8) = P(X_1)\, p(X_2 | X_1)\, p(X_3 | X_1, X_2) \cdots P(X_8 | X_1, \ldots, X_7)$

- $1 \to 2 \to 3 \to \ldots \to 8$

~~$2 \to 4 \to \ldots \to 8$~~

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

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

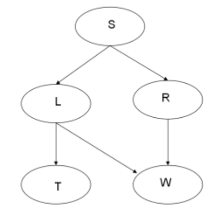
Benefits of Bayes Nets:

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

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

$P(X_8 | X_5, X_6)$

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)

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

$P(X_2|X_1)$

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

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

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

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

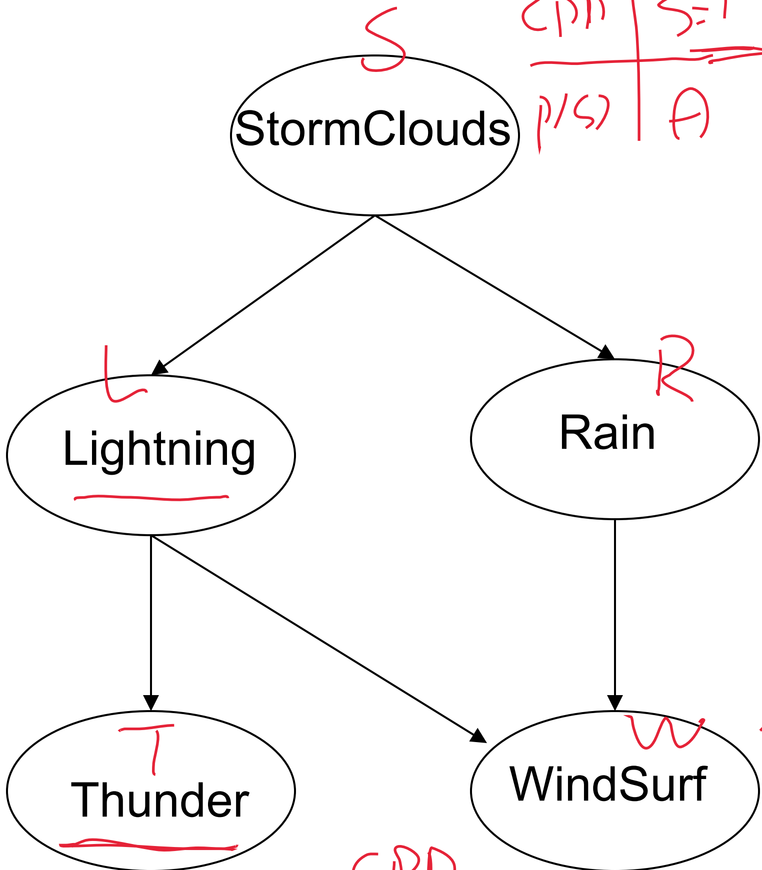
$Pa(X_i)$

Bayesian Network

DAG

CPD

boolean



CPD	S=1	S=0
P(S)	θ	1-θ

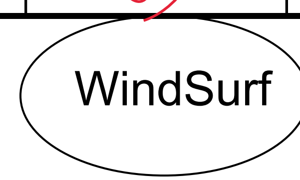
Nodes = random variables

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

P 1-P

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

i-params = 4



2^2

P(W|L,R)

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)

$2^2=4$

The joint distribution over all variables:

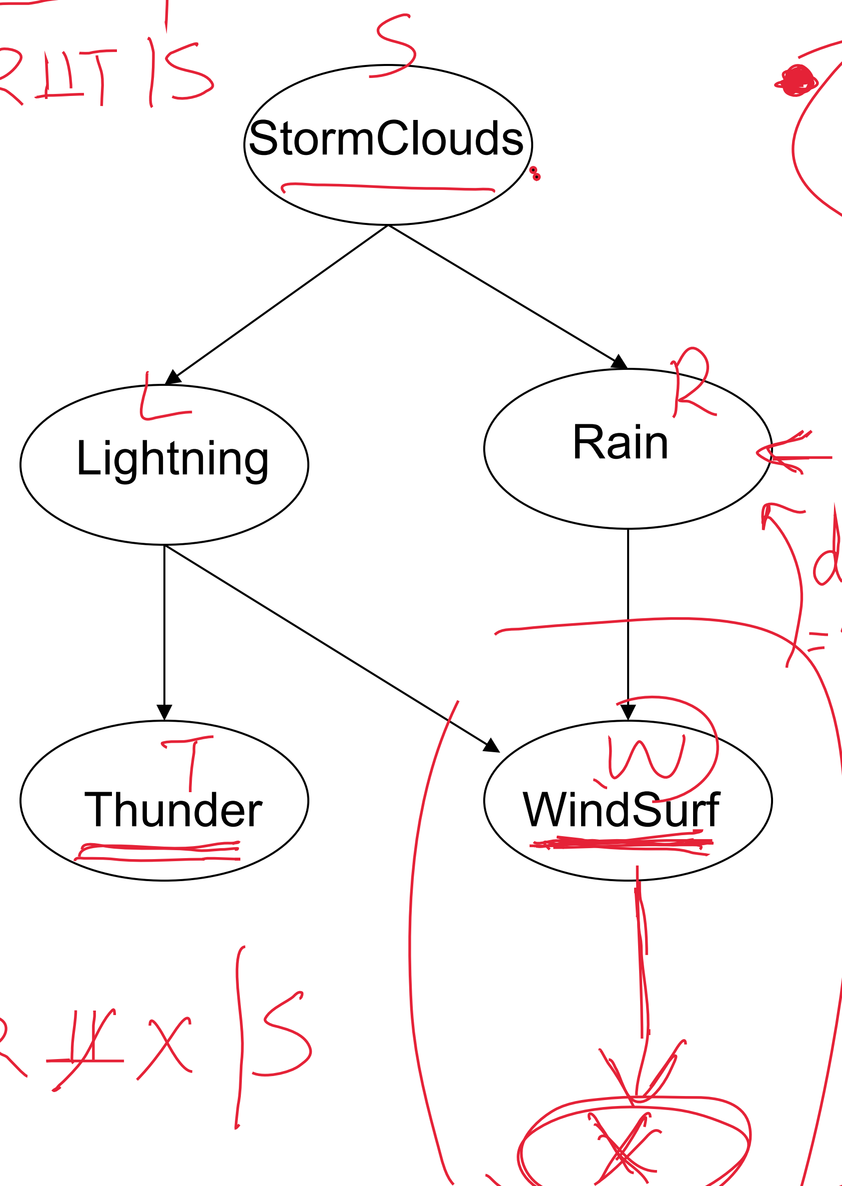
$$\underbrace{P(X_1 \dots X_n) = \prod_i P(X_i|Pa(X_i))}_{\text{CPD}}$$

P(W=1, T=0, R=0, L=1, S=1) =
P(S)·P(R|S) P(L|S) P(W|L,R) P(T|L) =

Bayesian Network

$R \perp L \mid S$

$R \perp T \mid S$



$R \not\perp X \mid S$

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.

$de(R) = \{W, X\}$

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

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

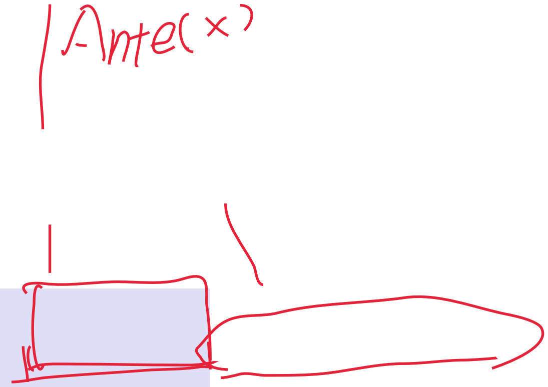
$P(X \mid Y, Z) = P(X \mid Z)$

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

$(P(T \mid L))$

$Pa(x) \subseteq An(x)$

Ante(x)



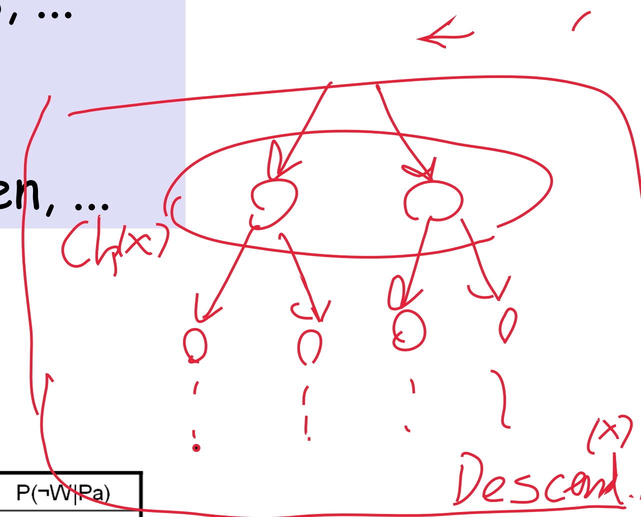
Some helpful terminology

Pa(x). Parents = Pa(X) = immediate parents

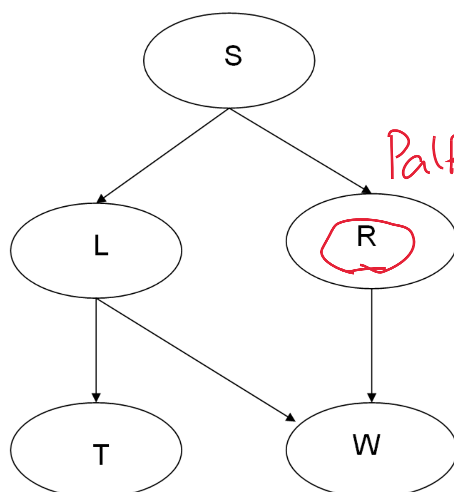
An(x). Antecedents = parents, parents of parents, ...

Ch(x). Children = immediate children

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



$Ch(x) \subseteq De(x)$



Pa(R)=S

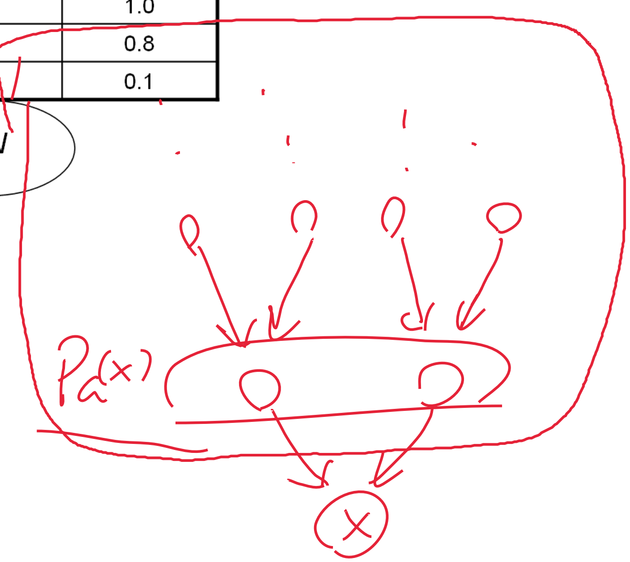
Pa(L)=S

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

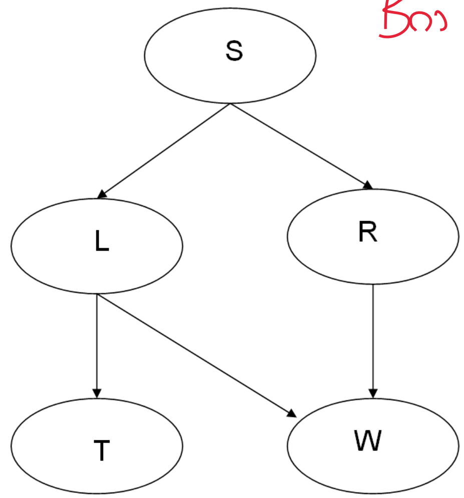
W



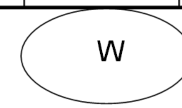
Bayesian Networks

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 2 2 2 $2^2 = 4$

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

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

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

cond. independ.

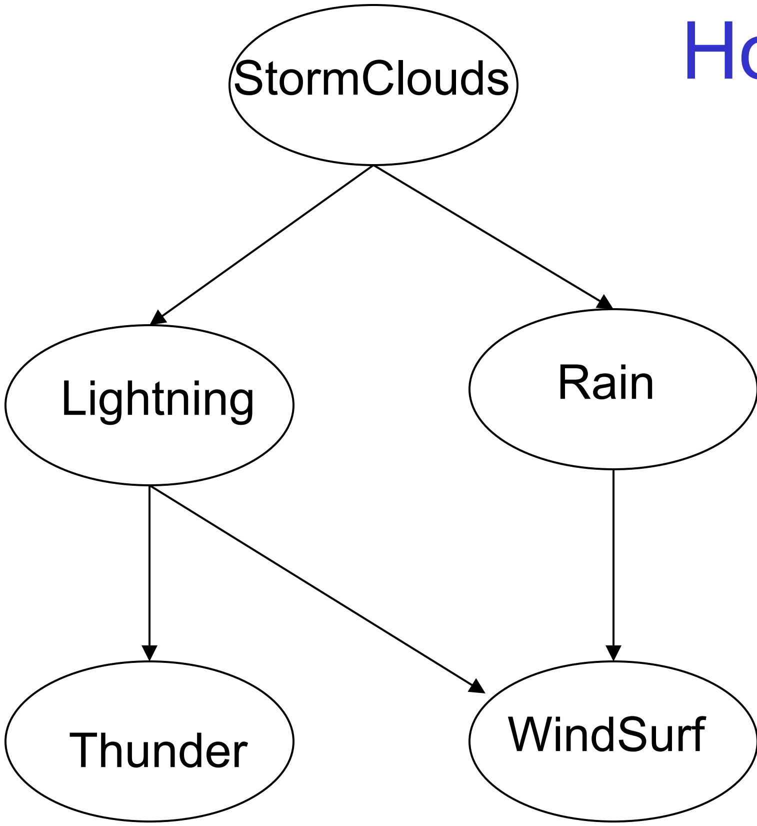
$P(R|S,L) = P(R|S)$

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

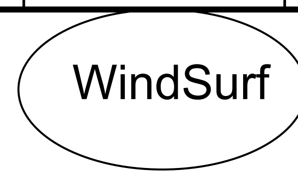
$R \perp\!\!\!\perp L \mid S$

$\# \text{iparams} = 11$

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



n boolean

To define joint distribution in general? 2^n exponetial

To define joint distribution for this Bayes Net?

1 order : $2n$

2 order : $2^2 n$

3 order : $2^3 n$

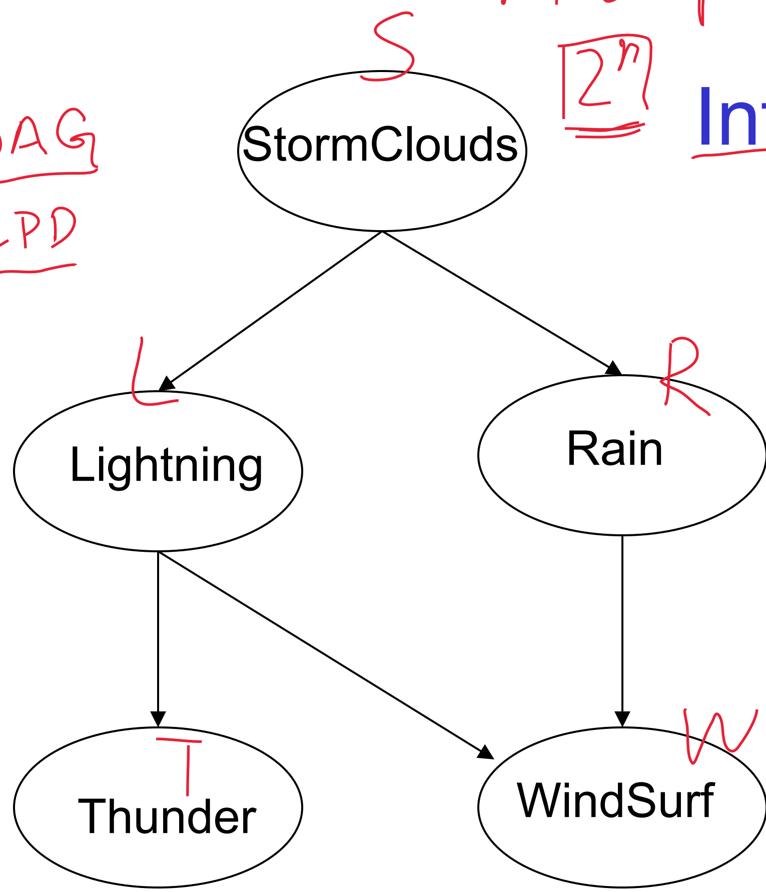
→ linear

NP complete: $P(W=1)$

$\boxed{2^n}$

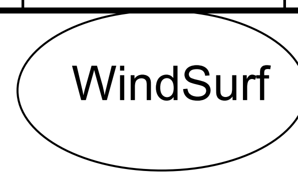
Inference in Bayes Nets

BN < DAG, 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

CPD



$$P(s, l, r, t) = \sum_{w=\{0,1\}} P(W=w, s, l, r, t)$$

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

0.2

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

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

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

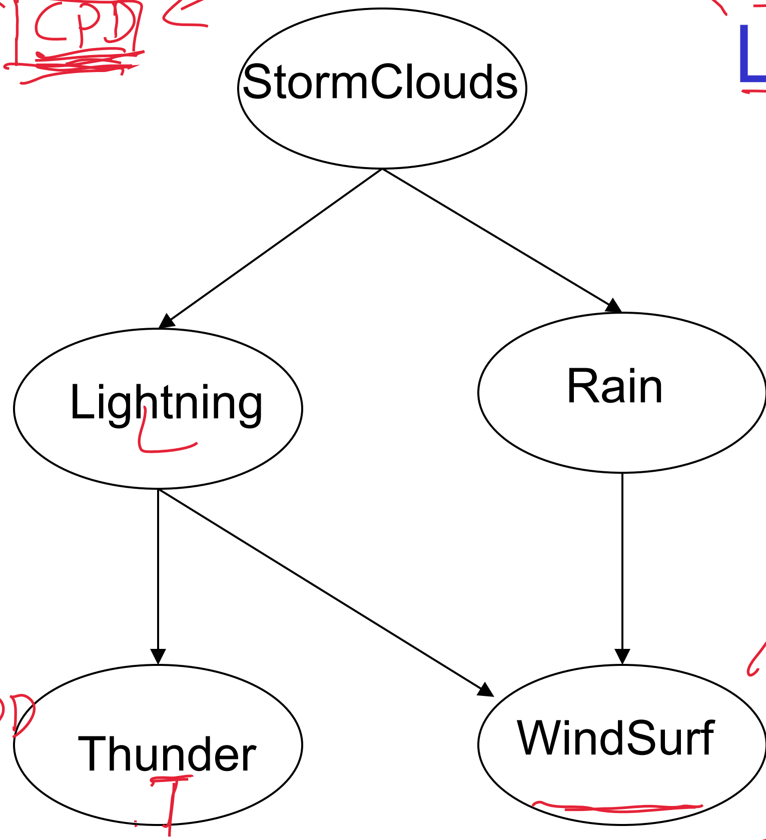
Training Set: $D = \{x_1, x_2, \ldots, x_m\}$ $\quad x_m \in \mathbb{B}^5$

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

BN < DAG, CPD

Learning a Bayes Net

#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 $1-\theta_{00}$

WindSurf

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

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

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	$1-\theta_0$

$$P(T \mid L) = \theta_1^{LT} (1-\theta_1)^{L(1-T)} \theta_0^{(1-L)T} (1-\theta_0)^{(1-L)(1-T)}$$

$\{L=1\}T \quad \{L=1\}(1-T) \quad \{L=0\}T \quad \{L=0\}(1-T)$

iid

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

$$\frac{\partial L(\theta)}{\partial \theta_1} = \theta_1^{\sum_{i=1}^{m} \{L=1\}T} (1-\theta_1)^{\sum_{i=1}^{m} \{L=1\}(1-T)} \cdot \theta_0^{\alpha_{01}} (1-\theta_0)^{\alpha_{00}} = 0$$

$\alpha_{11} \quad \alpha_{10}$

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

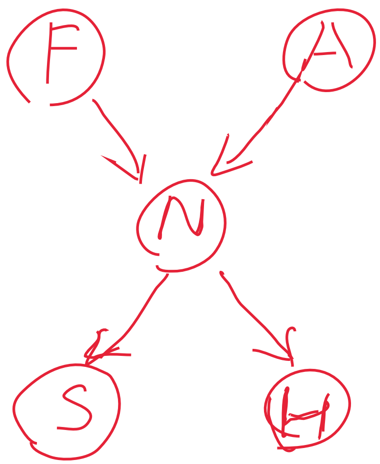
Boolean

Prior:

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

F A N S H

BN: DAG: CPD:



$A \perp F$

$A \not\perp F \mid N$

$D = \{(f, a, n, s, h)\}$

MLE:

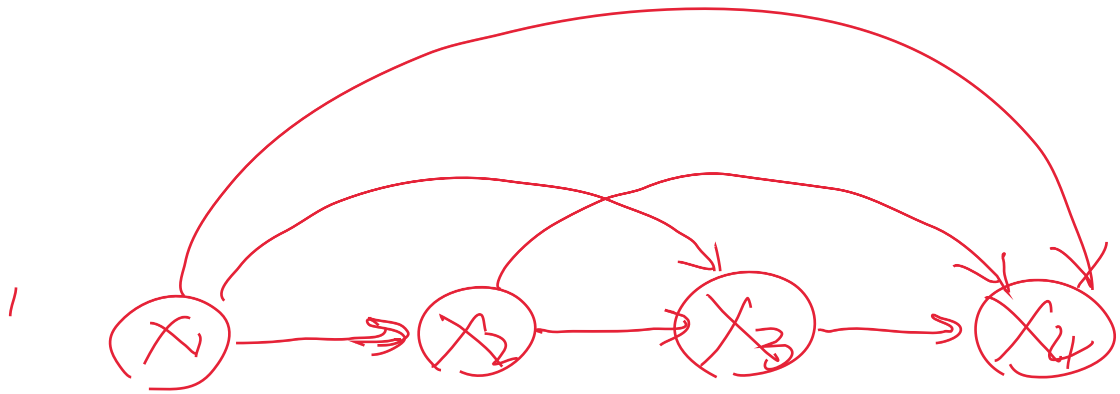
MAP

1. PDF
2. Likelihood
3. Derivative

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

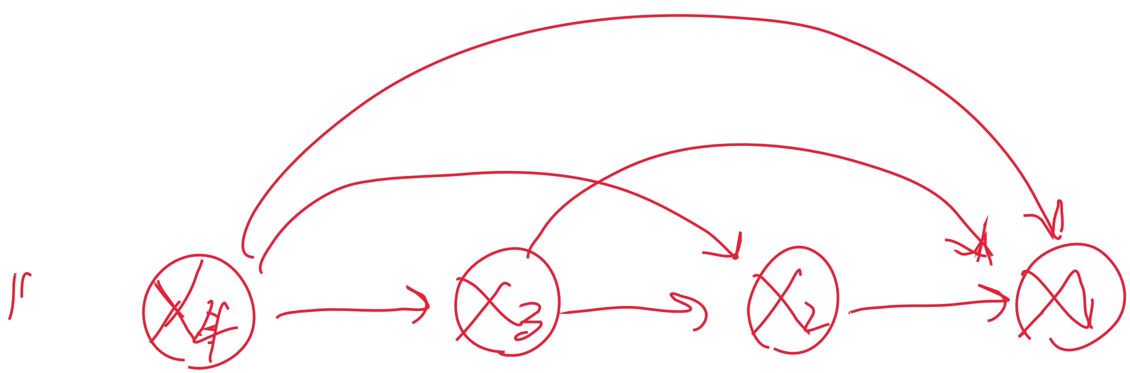
What is the Bayes Network for X1,…X4 with NO assumed conditional independencies?

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



fully-connected BN

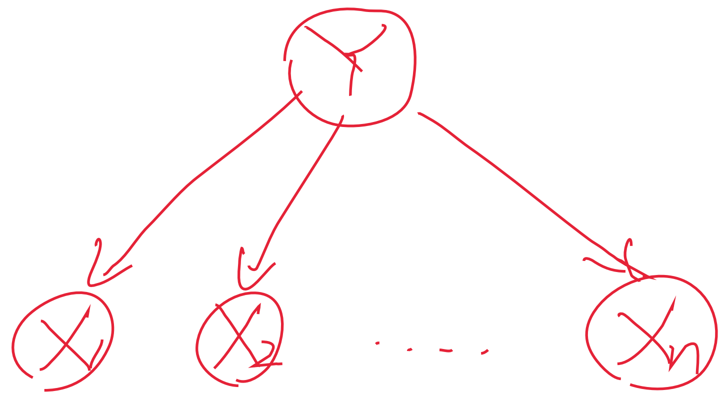
$4! = 4 \times 3 \times 2 \times 1$



$P(X_1, X_2, X_3, X_4) = \text{-} \quad \text{——} \quad \text{-}$

What is the Bayes Network for Naïve Bayes?

NB:



CPD

$X_i \perp\!\!\!\perp X_j | Y \quad \forall i \neq j$

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

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

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

GNB

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

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

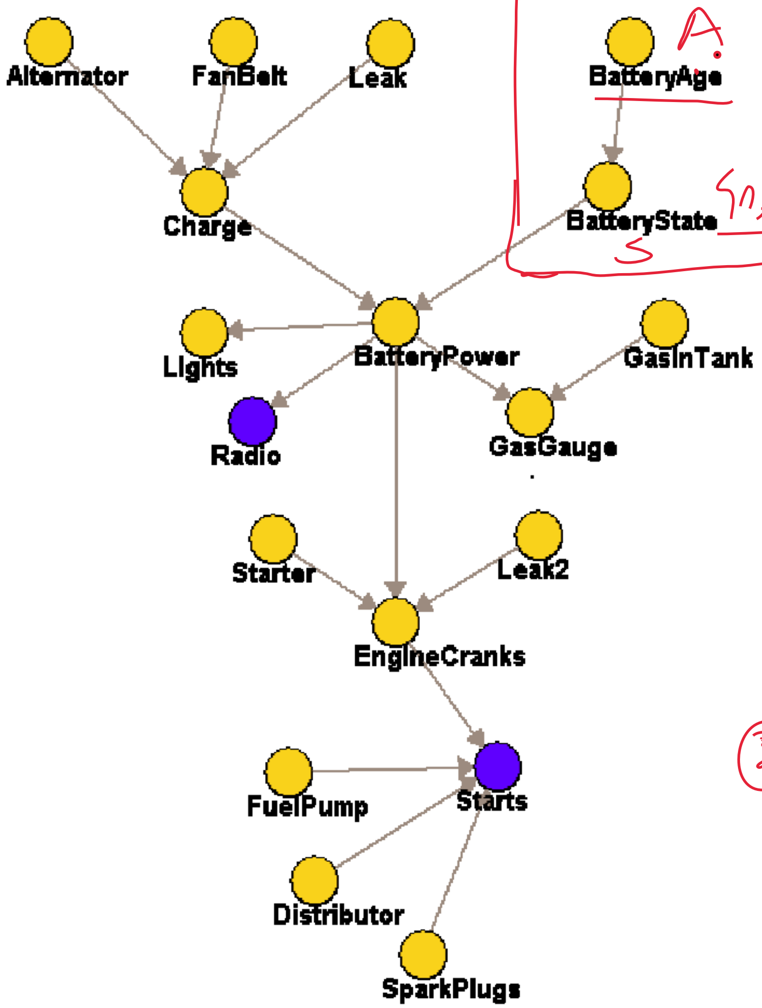
$P(Y=1 \mid X_1=1, X_2=0, \dots, X_n=1)$

$= \dfrac{P(y=1, x_1, \dots, x_n)}{P(x_1, x_2, \dots, x_n)}$

$= \sum_{y\in\{0,1\}} P(Y=y, x_1, \dots, x_n).$

Q

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



[0, 50]

1) [0, 10] [10, 20] — — [40, 50]

A ← continuous

{0,1} S ← discrete

$P(S|A)$

$P(S\|A)$	$S=1$	$S=0$
$A=0$		
$A=1$		
$A=3$		
⋮		
$A=5$		

infinite

② parameterized model

$\sigma(\beta A+\beta_0)$

Logistic / sigmoid

$$p(x) = \sigma(x) = \frac{1}{1+e^{-x}}$$

$\beta^T x+\beta_0$

$p(x)$, $\frac{1}{2}$, 0, x

$$p(Y|X) \leftarrow \sigma(\beta^T x+\beta_0)$$