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:

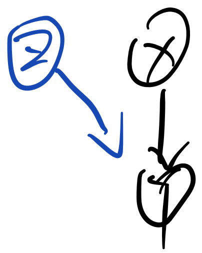
- Bishop chapter 8, through 8.2

Graphical Models

有向无环

概率模型图

- Key Idea:
  - Conditional independence assumptions useful
  - but Naïve Bayes is extreme! 假设所有都独立的假设过强
  - Graphical models express sets of conditional independence assumptions via graph structure
  - Graph structure plus associated parameters define *joint probability distribution over set of variables*



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

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

$P(Z|X) = P(Z)$ → X.Z条件独立.

X.Z不是独立的(Y) 若没观测Y 则X.Z独立.但它们条件独立.

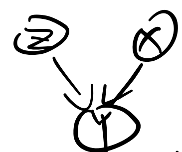
10-601

- 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



	$Y=0$	$Y=1$
$X=1, Z=1$	θ_{11}	
$X=1, Z=0$	θ_{10}	
$X=0, Z=1$	θ_{01}	
$X=0, Z=0$	θ_{00}	

$\vdash \theta_{ij}$

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

$$P(Y) = \sum_{x,z} P(X=x, Y=y, Z=z)$$

$$= \sum_{x,z} P(X=x) P(Y=y \mid X=x, Z=z) P(Z=z)$$

但随 r.v. ↑ complexity ↑ 简单时可如上去计算

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

r.v. 多 => 采样. $E_{P(X|Y)}[F(x)] = \int P(x|Y) F(x) dx$

Monte Carol

$= \frac{1}{K} \sum_{k=1}^{K} F(x_k), \quad x_k \sim P(X|Y)$

或 变分 (用 Gaussian distribution 逼近)

$\min_{\phi} KL(q_{\phi}(X) \| P(X|Y))$ → 高斯拟合的结果

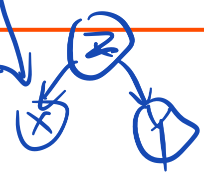
变成优化问题.

$X \perp\!\!\!\perp Y \mid Z$：在给定Z的情况下 X,Y独立.　　$X \perp\!\!\!\perp Y$ 或 $X \perp\!\!\!\perp Y \mid \phi$：边缘独立.

Conditional Independence

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

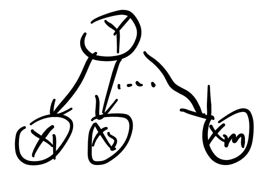
$$(\forall i, j, k) P(X = x_i | Y = y_j, Z = z_k) = P(X = x_i | Z = z_k)$$



：$P(X,Y|Z) = \frac{P(X,Y,Z)}{P(Z)} = \frac{P(X|Z)P(Y|Z)P(Z)}{P(Z)} = P(X|Z)P(Y|Z)$

是独立的 => 得证.

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



Naive Bayes' Assumption：$P_m = P(X_1, X_2, \dots, X_m|Y) = \prod_{i=1}^{m} P(X_i|Y) \Rightarrow$

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

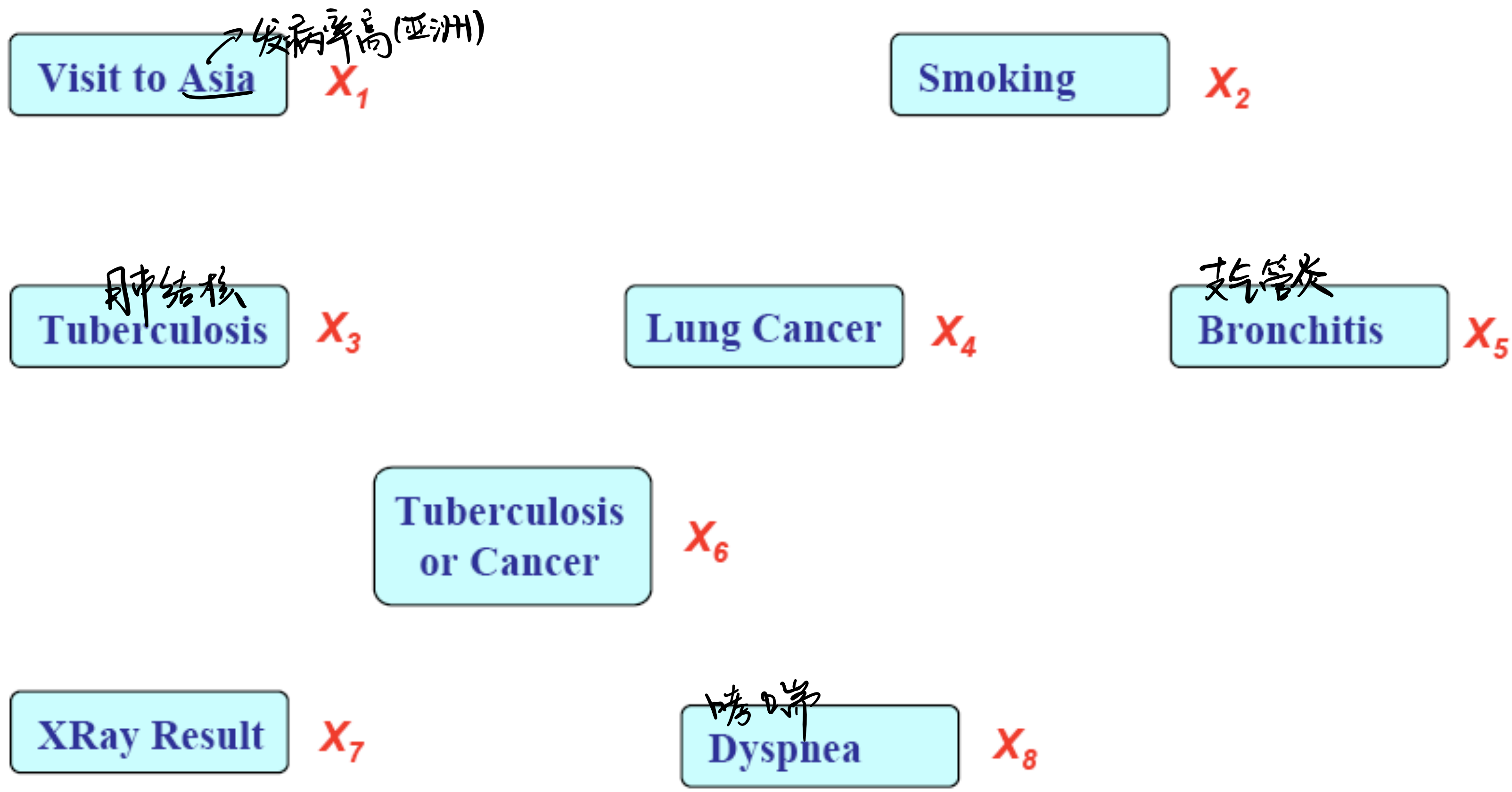
$$(\forall i, j) P(X = x_i | Y = y_j) = P(X = x_i)$$

Equivalently, if

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

Represent Joint Probability Distribution over Variables

肺结核相关因素



Describe network of dependencies

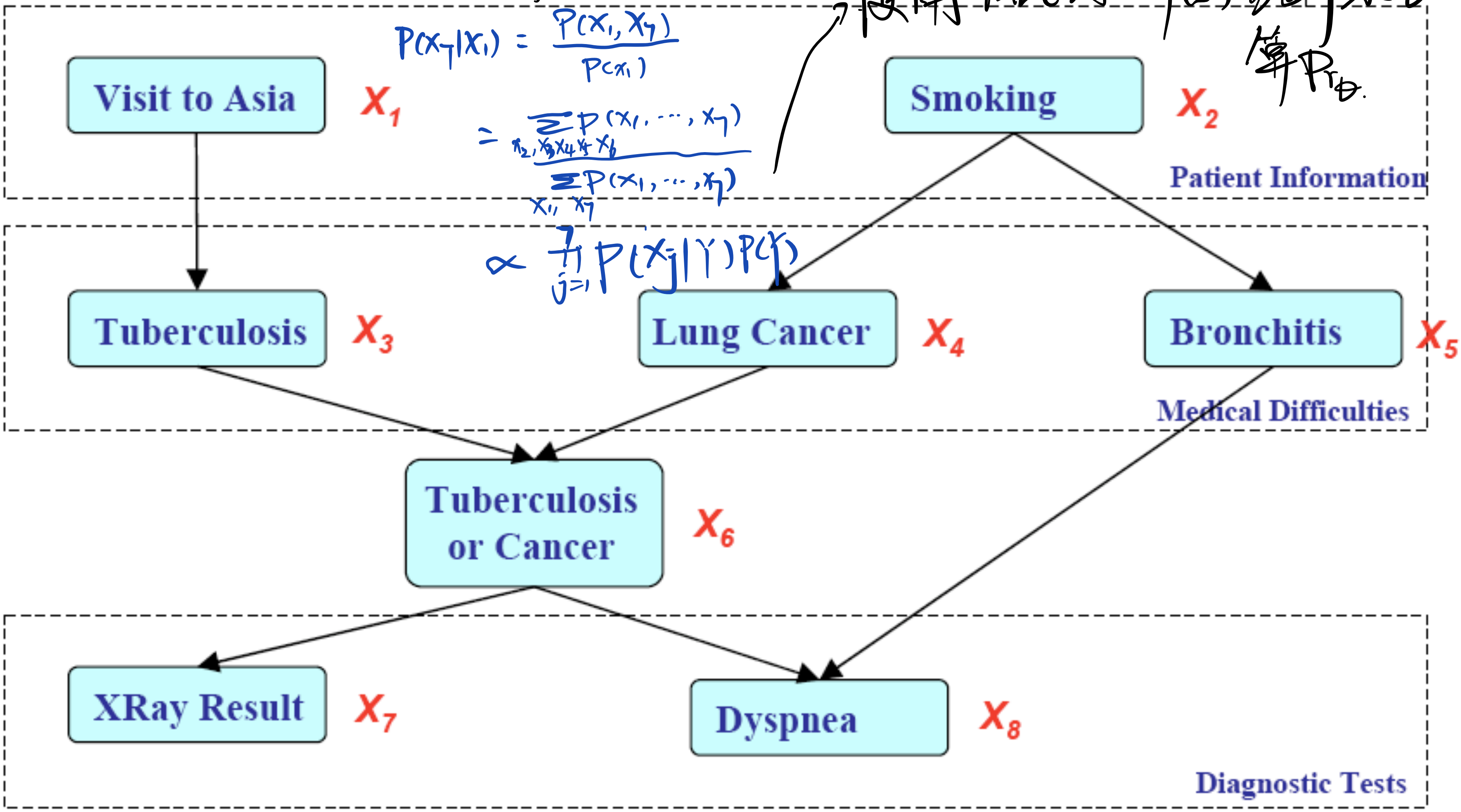
$G<V,E>$. $V: X_1, X_2, \cdots, X_8$ E：箭头.

使用MLE对8个点,分别求θ.
算Pr_θ.

$$P(x_7|x_1) = \frac{P(x_1,x_7)}{P(x_1)}$$

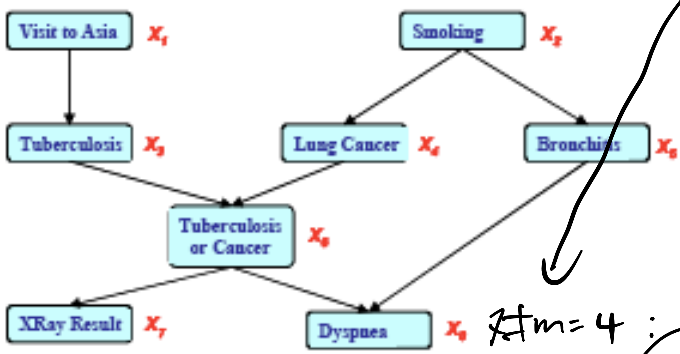
$$= \frac{\sum_{x_2,x_3,x_4,x_5,x_6} P(x_1,\cdots,x_7)}{\sum_{x_1, x_7} P(x_1,\cdots,x_7)}$$

$$\propto \prod_{j=1}^{7} P(x_j|Y)P(Y)$$



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

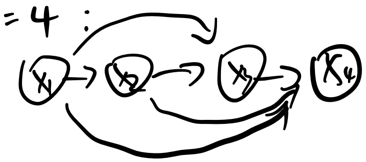
链式法则：$P(x_1,\cdots,x_m)=P(x_1)P(x_2|x_1)P(x_3|x_1,x_2)\cdots P(x_m|x_1,x_2\cdots x_m)$



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

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

对m=4：



故对点n：有2^n-1个参数去估计（改节点不同取值）

↳ 注意：这个可以换顺序 故要找最优

自己估的 故顺序影响大.

但有m!种可能.

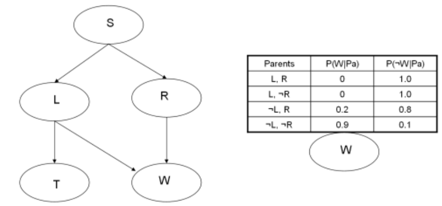
Benefits of Bayes Nets:

- Represent the full joint distribution in fewer parameters, using prior knowledge about dependencies

如果是全连通图则估参过多. 故通过上述先验知识减化计算.

- Algorithms for inference and learning

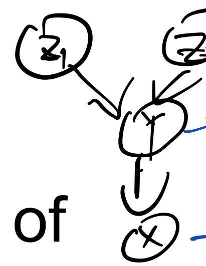
Bayesian Networks Definition



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



DAG：有向无环图

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

条件概率分布.

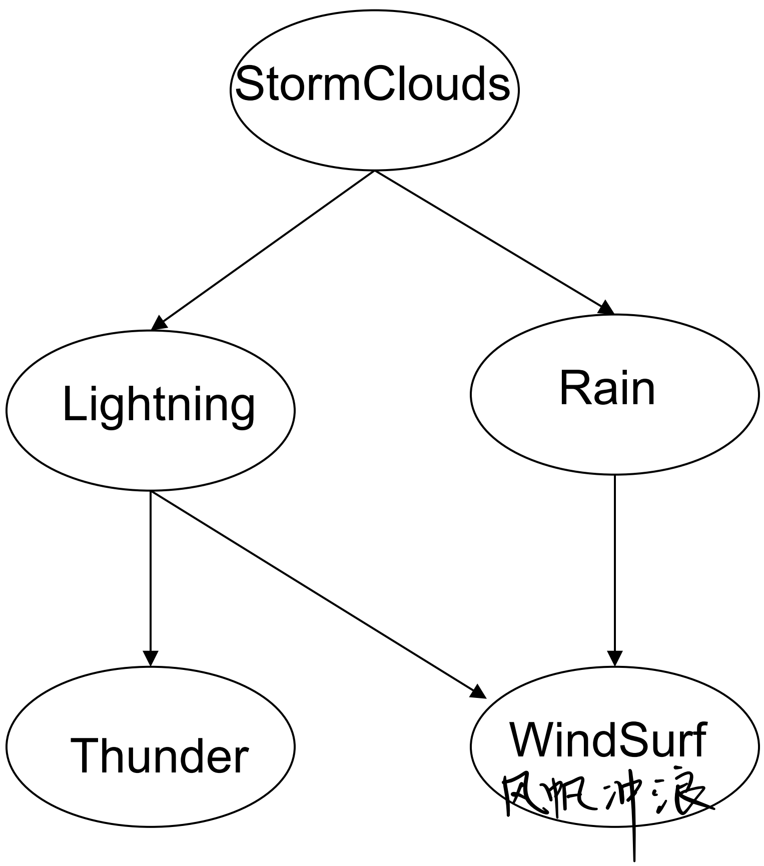
- Each node denotes a random variable
- Edges denote dependencies
- For each node X_i its CPD defines $P(X_i \mid Pa(X_i))$
- The joint distribution over all variables is defined to be

直接父节点（直接相连）

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

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

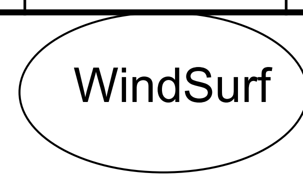
Bayesian Network



Nodes = random variables

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

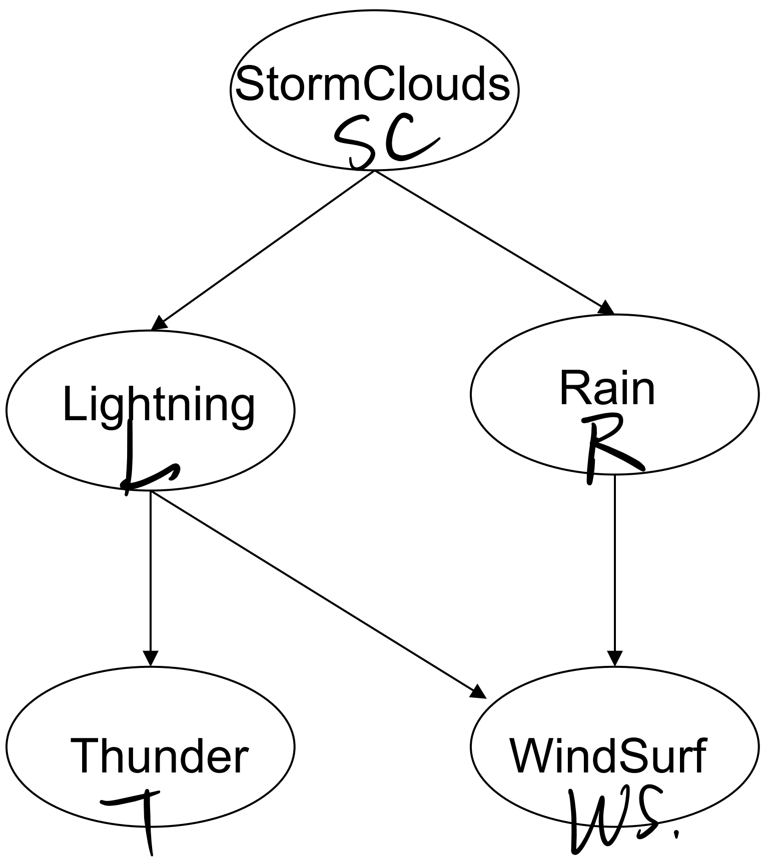
Parents	P(W|Pa)	P(¬W|Pa)
L, R	0	1.0
L, ¬R	0	1.0
¬L, R	0.2	0.8
¬L, ¬R	0.9	0.1



The joint distribution over all variables:

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

Bayesian Network



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

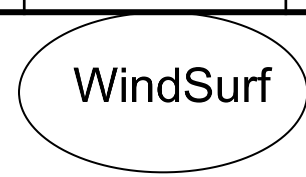
One thing is this:

Each node (X_i) is conditionally independent of its non-descendents, given only its immediate parents. $Pa(X_i)$

↗注意还是有例外的.

$\forall X_i$, X_i 与非$Pa(X_i)$和子代节点以外的node都条件独立.

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



$Pa(WS) = L, R. \Rightarrow WS \perp\!\!\!\perp L, R \mid \{L, R\}$, conditional independent

$Pa(T) = L \Rightarrow T \perp\!\!\!\perp WS, R, SC \mid L.$

注意, L: $Pa(L) = SC \Rightarrow L \perp\!\!\!\perp R, WS \mid SC$ 没有T: 子代节点.

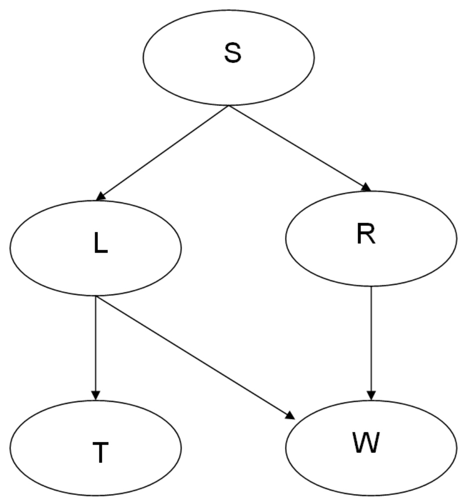
Some helpful terminology

Parents = Pa(X) = immediate parents

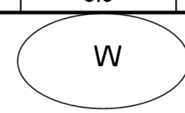
Antecedents = parents, parents of parents, ...

Children = immediate children Ch(X)

Descendents = children, children of children, ...

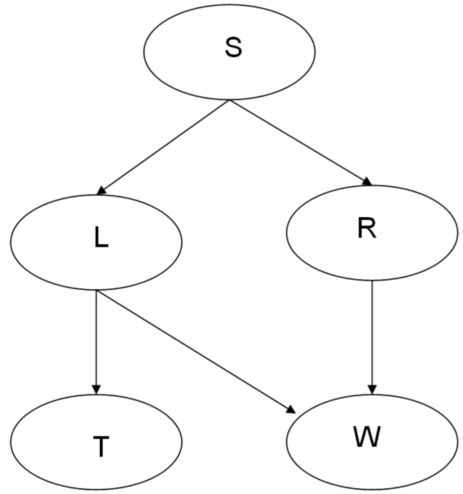


Parents	P(W|Pa)	P(¬W|Pa)
L, R	0	1.0
L, ¬R	0	1.0
¬L, R	0.2	0.8
¬L, ¬R	0.9	0.1

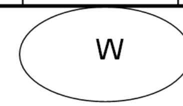


Bayesian Networks

- 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

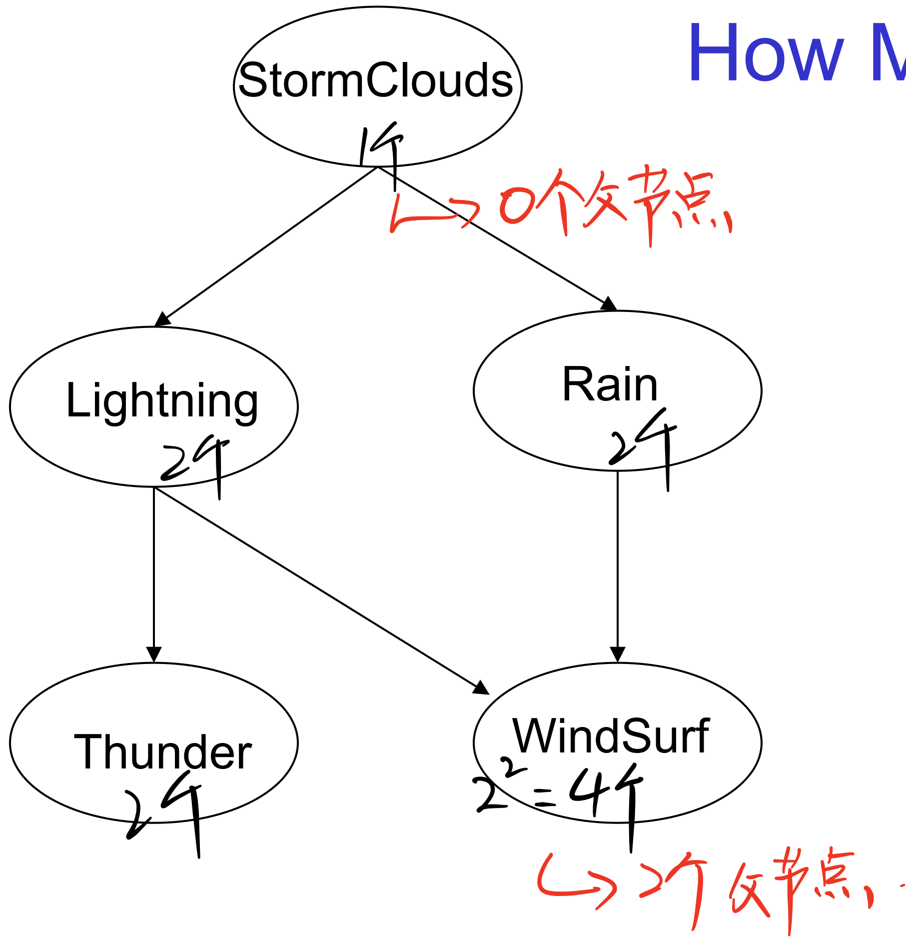


Chain rule of probability says that in general:

$$P(S,L,R,T,W) = P(S)P(L|S)P(R|S,\cancel{L})P(T|\cancel{S},L,\cancel{R})P(W|\cancel{S},L,R,\cancel{T})$$

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

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

WindSurf → Bayes Net

fully-connected BN: 链式法则

全连接: 每个节点与前面的点都相连

To define joint distribution in general?

2^n-1 : $2^5-1=31$个参数. → 指数

To define joint distribution for this Bayes Net?

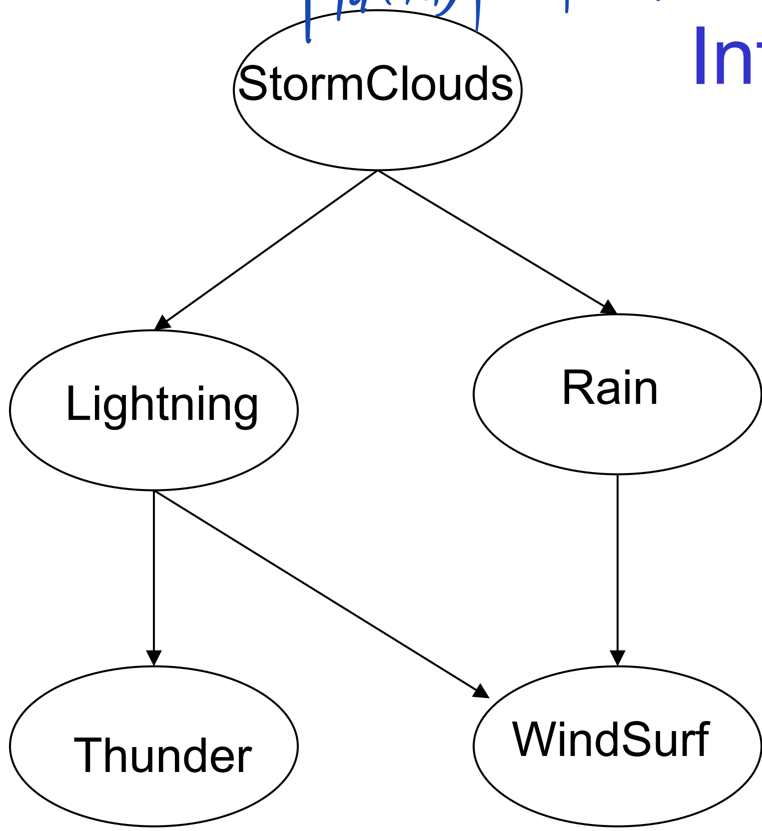
11个参数. → 线性 (参数数量少).

参数: $|Pa(X_i)| \le 1 \Rightarrow 2n$个

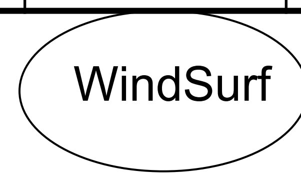
$|Pa(X_i)| \le 2 \Rightarrow 2^2 n = 4n$

$|Pa(X_i)| \le k \Rightarrow 2^k n$

Inference in Bayes Nets



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) = 用这个算法 ← 联合概率密度分布

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

积分积回去

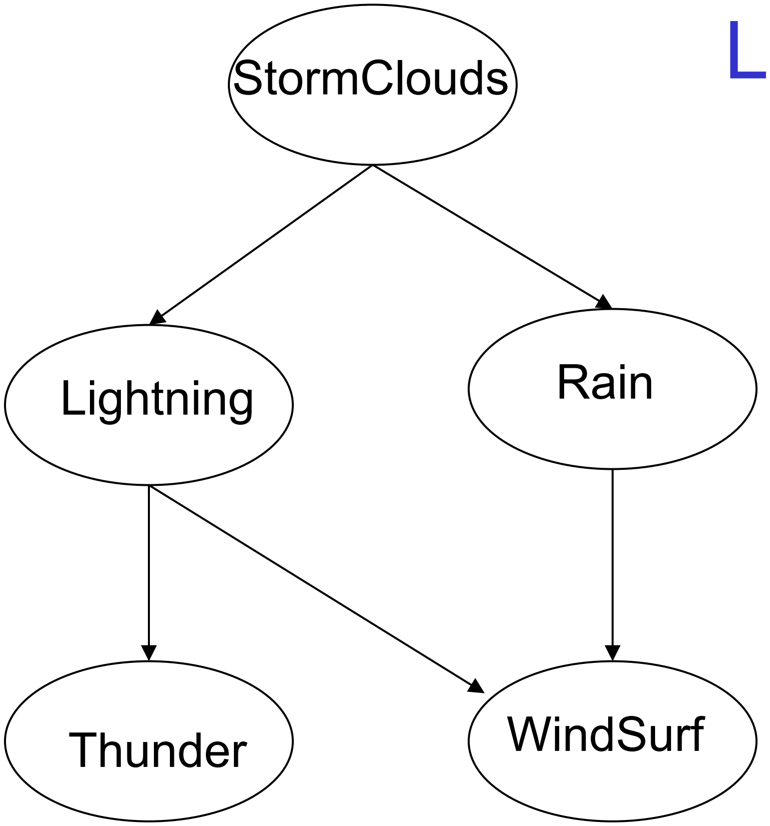
$$P(S=1|L=0,T=1) = \frac{P(S=1,L=0,T=1)}{P(T=1,L=0)} = \frac{\sum_{w\in\{0,1\}}\sum_{r\in\{0,1\}} P(S=1,T=1,L=0,W=w,R=r)}{\sum_{w\in\{0,1\}}\sum_{r\in\{0,1\}}\sum_{s\in\{0,1\}} P(S=s,T=1,L=0,W=w,R=r)}$$

$P(S=1) = \sum\sum\sum\sum P(S=1,T=t,L=l,W=w,R=r)$ 共算 $2^4=16$ 次.

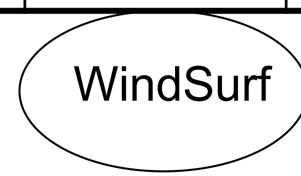
↳算n个变量则有2^{n-1}次计算求和，计算中做n-1次求和 => $(n-1)2^{n-1}$次计算

估计(采样) ⇐ 太大，无法算

Learning a Bayes Net



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



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

What is the MLE solution? MAP?

以T为例建立似然函数.

CPD(T): 条件概率分布

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

$$P(T|L) = \theta_1^{TL}(1-\theta_1)^{(1-T)L} + \theta_0^{T(1-L)}(1-\theta_0)^{(1-T)(1-L)}$$

$$\Rightarrow l(\theta_1,\theta_0) = \sum_{i=1}^{n} \ln P(T=t_i|L=l_i)$$

$$D = \{(s_i, l_i, t_i, w_i, r_i)\}_{i=1}^{n}$$

求 $\max\limits_{\theta_1,\theta_0} l(\theta_1,\theta_0)$：用求导.

假设 $x_1,\cdots,x_{i-1}$ 已建立某种网络.

则考虑 $P(x_i|Pa(x_i))$ 与 $P(x_i|\overbrace{x_1,\cdots,x_{i-1}}^{\text{全概率}})$

判断哪些是独立的.

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

先画 DAG (先验, 根据经验)
再算 (估计) CPD.

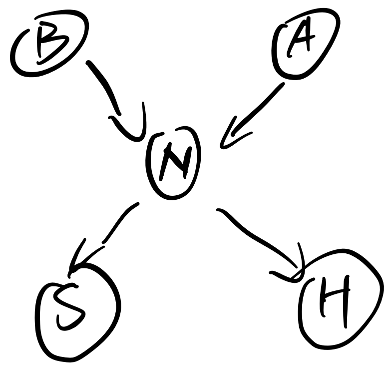
B A N

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

S H

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

画图:



$\Rightarrow$ 给定 $N \Rightarrow S, H$ 独立.

B, A 独立 (边缘情况, 即未给定任何条件).

	$H=1$	$H=0$
$N=1$	θ_1	$1-\theta_1$
$N=0$	θ_0	$1-\theta_0$

$\Rightarrow$ 计算 CPD.

然后可计算所有概率.

$$P(H=1|A=1) = \frac{P(H=1,A=1)}{P(A=1)}$$

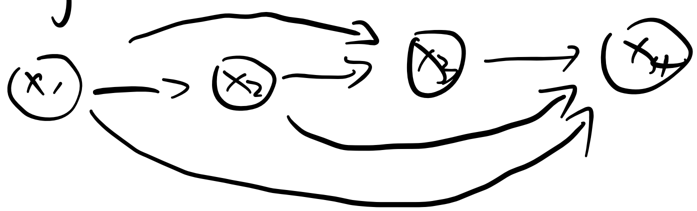
$$= \frac{\sum_{bsn} P(H=1, A=1, B=b, S=s, N=n)}{\sum_{b,s,n,h} P(A=1, B=b, S=s, N=n, H=h)}$$

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

假定顺序：$X_1 \to X_2 \to X_3 \to X_4$

不同顺序有不同链式法则展开，共有 n! 种顺序.

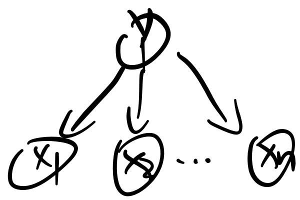
图：



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

What is the Bayes Network for Naïve Bayes?

Native Bayes :



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

	$X_i=1$	$X_i=0$
$Y=1$	θ_1	$1-\theta_1$
$Y=0$	θ_0	$1-\theta_0$

写似然函数，求导等于0

$$P(Y=y \mid X=x) = \frac{P(X=x \mid Y=y)\,P(Y=y)}{P(X=x)}$$

[0,1) [1,2) [n-1,n)

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

→是连续变量，不易估计

{ 离散化：BA = 1, 2, ⋯ n

↳ 一般是3-5份

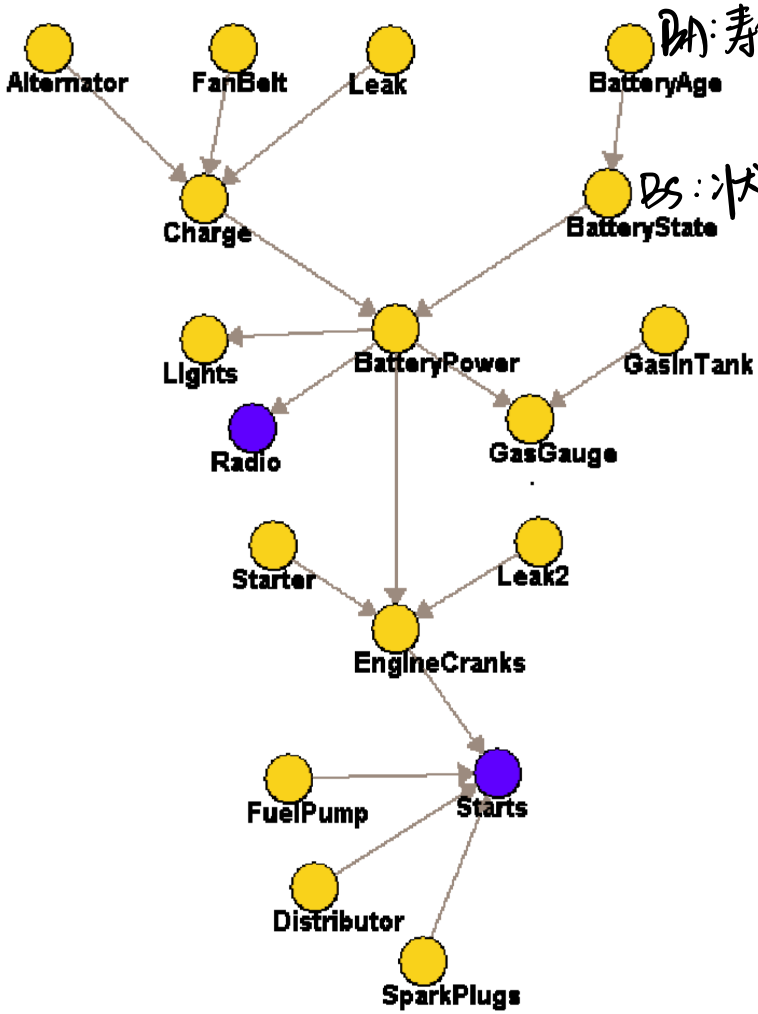
对参数建模.

Sigmoid函数

考虑. $\sigma(x) = \frac{1}{1+e^{x}}$

$\Rightarrow P(BS=s \mid BA=a) = \frac{1}{1+e^{-\beta_1 a+\beta_0}}$

然后对β_1, β_0求解.



BA：寿命

BS：状态