Machine Learning 10-701

Tom M. Mitchell
Machine Learning Department
Carnegie Mellon University

February 8, 2011

Today:

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

Readings:

Required:
- 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
  - Graph structure plus associated parameters define *joint probability distribution over set of variables/ nodes*

- Two types of graphical models:
  - Directed graphs (aka Bayesian Networks) today
  - 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
  - Observed data to estimate parameters
- 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

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

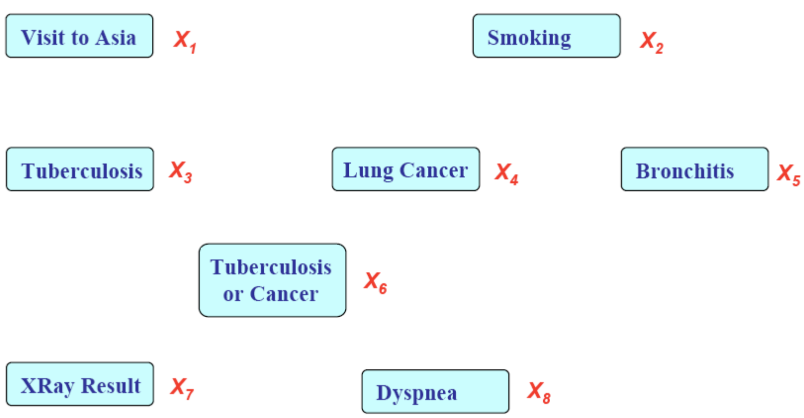
Equivalently, if

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

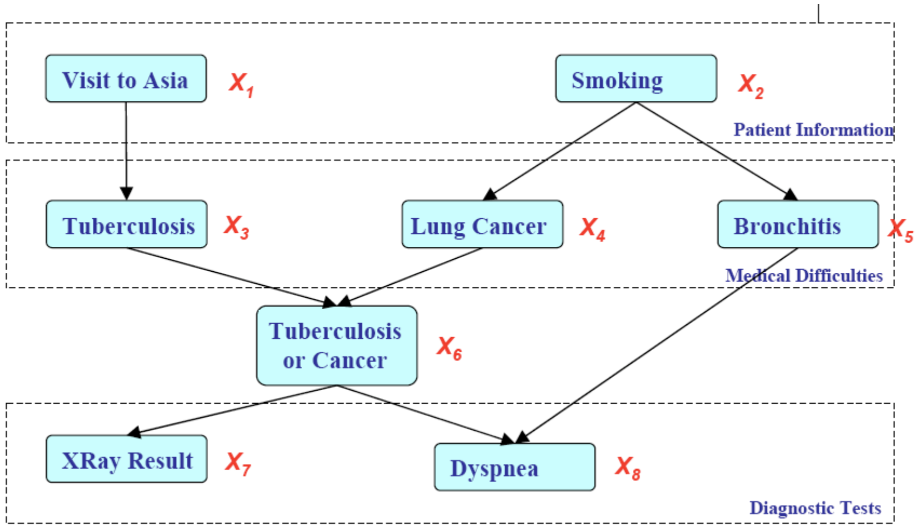
Equivalently, if

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

Represent Joint Probability Distribution over Variables

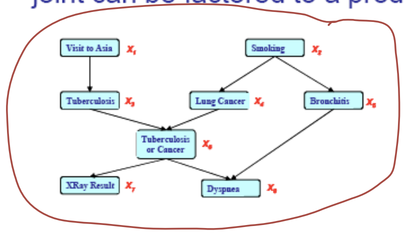


Describe network of dependencies



Bayesian Networks define Joint Distribution in terms of this graph, plus parameters

- If X_i's are **conditionally independent** (as described by a **PGM**), the joint can be factored to a product of simpler terms, e.g.,



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

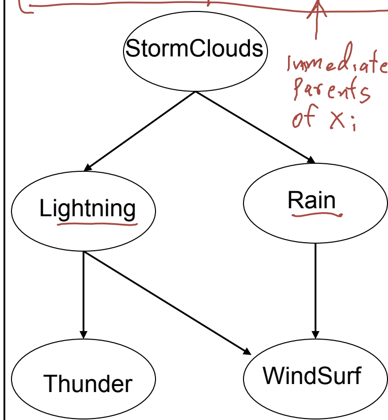
- Why we may favor a PGM?
  - Representation cost: how many probability statements are needed?

    2+2+4+4+4+8+4+8=36, an 8-fold reduction from 2^8!
  - Algorithms for systematic and efficient inference/learning computation
    - Exploring the graph structure and probabilistic (e.g., Bayesian, Markovian) semantics
  - Incorporation of domain knowledge and causal (logical) structures

Bayesian Network

$$\boxed{P(x_1, \dots x_n) = \prod_i P(x_i | Pa(x_i))}$$

immediate parents of x_i



Bayes network: a directed acyclic graph defining a joint probability distribution over a set of variables

Each node denotes a random variable

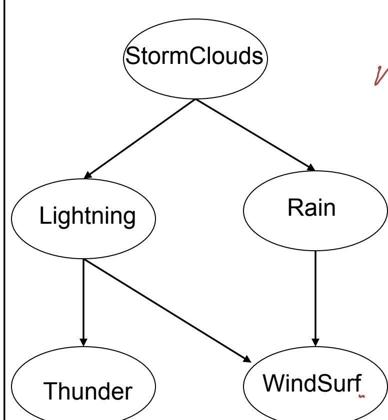
A conditional probability distribution (CPD) is associated with each node N, defining P(N | Parents(N))

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

WindSurf

The joint distribution over all variables in the network is defined in terms of these CPD's, plus the graph

Bayesian Network



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

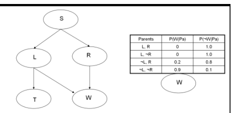
One thing is this:

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

WindSurf

Bayesian Networks Definition



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

A Bayes network is a directed acyclic graph and a set of CPD's

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

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

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

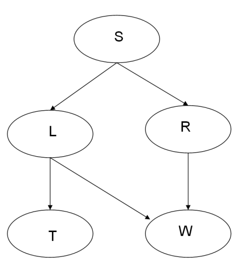
Some helpful terminology

Parents = Pa(X) = immediate parents

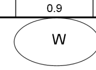
Antecedents = parents, parents of parents, ...

Children = immediate children

Descendents = children, children of children, ...

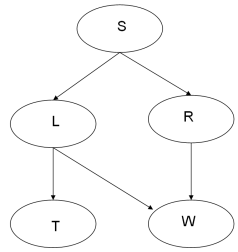


Parents	P(W|Pa)	P(¬W|Pa)
L, R	0	1.0
L, ¬R	0	1.0
¬L, R	0.2	0.8
¬L, ¬R	0.9	0.1

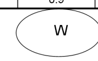


Bayesian Networks

- 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:

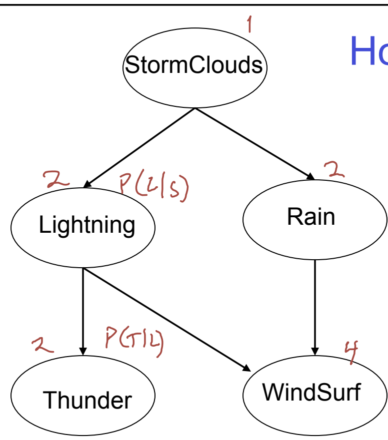
$$P(S,L,R,T,W) = P(S)P(L|S)P(R|S,L)P(T|S,L,R)P(W|S,L,R,T)$$

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

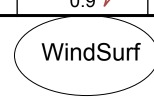
$$P(S\,L\,R\,T\,W) = P(S)\,P(L|S)\,P(R|S)\,P(T|L)\,P(W|L,R)$$

$(\forall s\,l\,r\,t\,w)\ P(S=s, L=l \cdots) = P(S=s)\,P(L=l|S=s) \dots$ "

How Many Parameters?



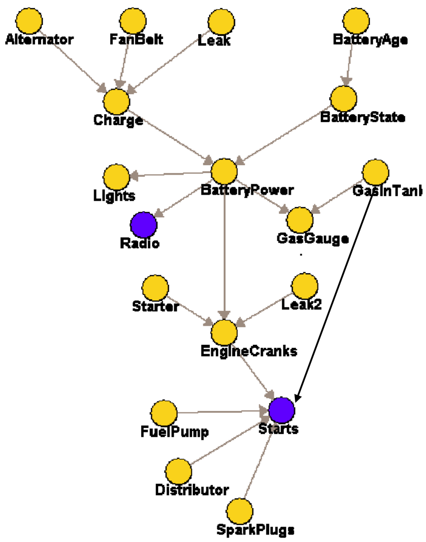
Parents	P(W|Pa)	P(¬W|Pa)
L, R	0	1.0
L, ¬R	0	1.0
¬L, R	0.2	0.8
¬L, ¬R	0.9	0.1



In full joint distribution? $2^5 - 1 = 31$

Given this Bayes Net? $= 11$

Bayes Net



<u>Inference</u>:

P(BattPower=t | Radio=t, Starts=f)

<u>Most probable explanation</u>:

What is most likely value of Leak, BatteryPower given Starts=f?

<u>Active data collection</u>:

What is most useful variable to observe next, to improve our knowledge of node X?

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

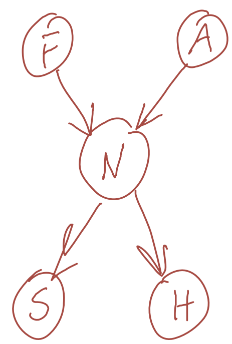
$$P(X_1 \ldots X_n) = \prod_i P(X_i|X_1 \ldots X_{i-1}) \quad \text{(by chain rule)}$$

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

Example

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

F



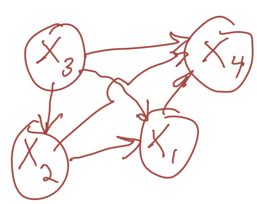
$H \perp S \mid N$

$H \perp A \mid N$

$H \perp F \mid N$

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

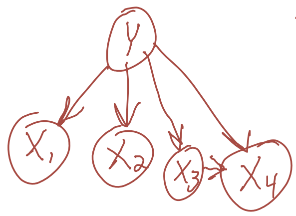
$$P(X_1 X_2 X_3 X_4) \overset{\text{Chain Rule}}{=} P(X_1)\, P(X_2|X_1)\, P(X_3|X_1 X_2)\, P(X_4|X_1 X_2 X_3)$$



What is the Bayes Network for Naïve Bayes?

$P(Y \mid X_1 \cdots X_4)$

$P(Y X_1 X_2 \cdots X_4) = P(Y)\, P(X_1 \mid Y)\, P(X_2 \mid Y) \cdots P(X_4 \mid Y)$

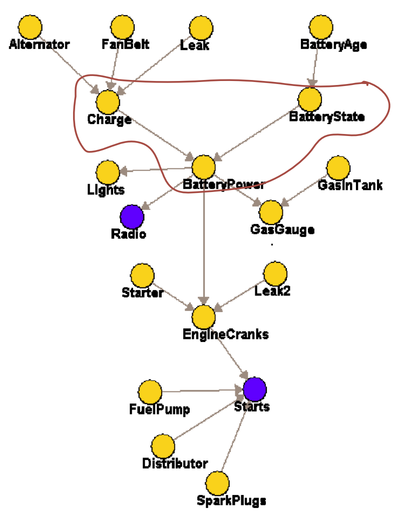


$P(X_3 X_4 \mid Y) = P(X_3 \mid Y X_4)\, P(X_4 \mid Y)$

$= P(X_4 \mid Y X_3)\, P(X_3 \mid Y)$

$X_1 \perp X_2 \mid Y$

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



eg.

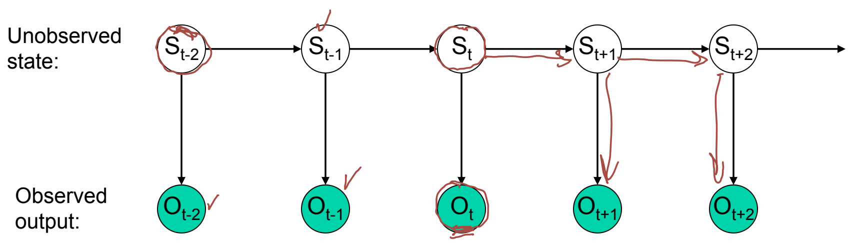
$P(BP \mid BS, C) = N(\mu_{BS,C}, \sigma_{BS,C})$

real valued

Bayes Network for a Hidden Markov Model

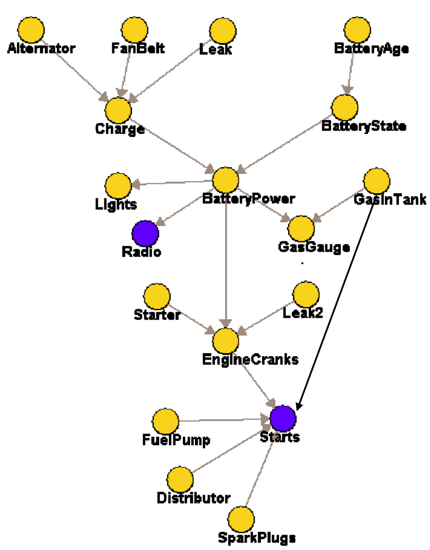
Assume the future is conditionally independent of the past, given the present

$S_t \perp S_{t-2}, O_{t-1}, O_{t-2} \mid S_{t-1}$



$$P(S_{t-2}, O_{t-2}, S_{t-1}, \ldots, O_{t+2}) = P(S_{t-2})\, P(O_{t-2} \mid S_{t-2})\, P(S_{t-1} \mid S_{t-2})\, P(O_{t-1} \mid S_{t-1})\, P(S_t \mid S_{t-1}) \ldots$$

How Can We Train a Bayes Net



1. when graph is given, and each training example gives value of every RV?

Easy: use data to obtain MLE or MAP estimates of θ for each CPD

P(Xi | Pa(Xi); θ)

e.g. like training the CPD's of a naïve Bayes classifier

2. when graph unknown or some RV's unobserved?

this is more difficult… later…