Machine Learning 10-601, 10-301

Tom M. Mitchell
Machine Learning Department
Carnegie Mellon University

April 12, 2021

Today:

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

Readings:

- Mitchell: Naïve Bayes and Logistic Regression http://www.cs.cmu.edu/~tom/mlbook/NBayesLogReg.pdf
- *Directed Graphical Models (Bayes nets)*. Kevin P. Murphy (2014). Machine Learning: A Probabilistic Perspective. Chapter 10

Naïve Bayes Algorithm – discrete X_i

- Train Naïve Bayes (examples)

  for each* value y_k

  estimate $\pi_k \equiv P(Y = y_k)$

  for each* value x_{ij} of each attribute X_i

  estimate $\theta_{ijk} \equiv P(X_i = x_{ij} | Y = y_k)$

- Classify (X^{new})

$$Y^{new} \leftarrow \arg\max_{y_k} \; P(Y = y_k) \prod_i P(X_i^{new} | Y = y_k)$$

$$Y^{new} \leftarrow \arg\max_{y_k} \; \pi_k \prod_i \theta_{ijk}$$

* probabilities must sum to 1, so need estimate only n-1 of these...

Example: Taking 10-601 or 10-301? P(G|F,B,U)

- G=1 iff you're taking 10-601
- U=1 iff taking undergrad class
- F=1 iff first year at CMU
- B=1 iff Birthday is before July 1

P(G=1) :
P(F=1 | G=1) :
P(F=1 | G=0) :
P(B=1 | G=1) :
P(B=1 | G=0) :
P(U=1 | G=1) :
P(U=1 | G=0) :

P(G=0) :
P(F=0 | G=1) :
P(F=0 | G=0) :
P(B=0 | G=1) :
P(B=0 | G=0) :
P(U=0 | G=1) :
P(U=0 | G=0) :

$$Y^{new} \leftarrow \arg\max_{y_k} \; P(Y = y_k) \prod_i P(X_i^{new} | Y = y_k)$$

$$G^{new} \leftarrow \arg\max_{g_k \in \{0,1\}} P(G = g_k) \; P(F = f^{new} | G = g_k) \; P(B = b^{new} | G = g_k) \; P(U = u^{new} | G = g_k)$$

Example: Taking 10-601 or 10-301? P(G|F,B,U)

- G=1 iff you're taking 10-601
- U=1 iff taking undergrad class
- F=1 iff first year at CMU
- B=1 iff Birthday is before July 1

P(G=1) : .60
P(F=1 | G=1) : .67
P(F=1 | G=0) : .92
P(B=1 | G=1) : .48
P(B=1 | G=0) : .62
P(U=1 | G=1) : .07
P(U=1 | G=0) : 1.0

P(G=0) : .40
P(F=0 | G=1) : .33
P(F=0 | G=0) : .08
P(B=0 | G=1) : .52
P(B=0 | G=0) : .38
P(U=0 | G=1) : .93
P(U=0 | G=0) : 0

$$Y^{new} \leftarrow \arg\max_{y_k} P(Y = y_k) \prod_i P(X_i^{new} | Y = y_k)$$

$$G^{new} \leftarrow \arg\max_{g_k \in \{0,1\}} P(G = g_k)\ P(F = f^{new} | G = g_k)\ P(B = b^{new} | G = g_k)\ P(U = u^{new} | G = g_k)$$

301

F = 0 B = 0 U = 1 ⇒ P(G=0)

G=0 = 0.012

G=1 = .007 = 0.60 0.33 0.52 0.07

0.08 0.38 1.0

$$P(G=0|x) = \frac{0.012}{0.012 + 0.007} \approx 0.63$$

What if we have continuous X_i ?

Eg., we include real-valued Age as an X_i in our 10-301 vs. 10-601 classifier

E.g., image classification: X_i is ith pixel

What if we have continuous X_i ?

image classification: X_i is ith pixel, Y = mental state

Still have:

$$P(Y = y_k | X_1 \ldots X_n) = \frac{P(Y = y_k) \prod_i P(X_i | Y = y_k)}{\sum_j P(Y = y_j) \prod_i P(X_i | Y = y_j)}$$

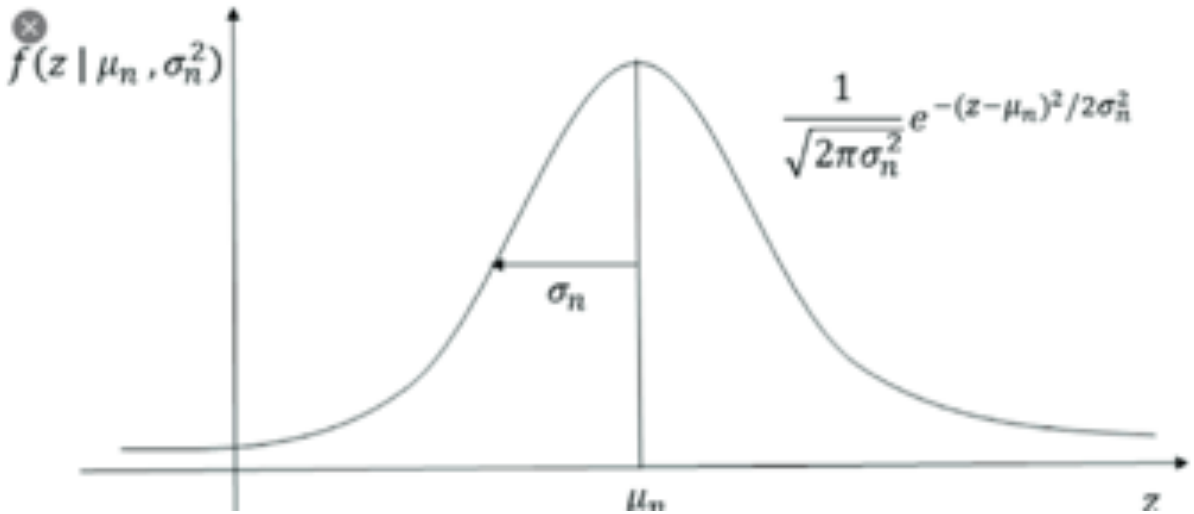
Just need to decide how to represent $P(X_i \mid Y)$

What if we have continuous X_i ?

Eg., image classification: X_i is ith pixel

Gaussian Naïve Bayes (GNB): assume

$$P(X_i = x \mid Y = y_k) = \frac{1}{\sigma_{ik}\sqrt{2\pi}} \; e^{\frac{-(x-\mu_{ik})^2}{2\sigma_{ik}^2}}$$



Naïve Bayes Algorithm – discrete X_i

- Train Naïve Bayes (examples)

  for each* value y_k

  estimate $\pi_k \equiv P(Y = y_k)$

  for each* value x_{ij} of each attribute X_i

  estimate $\theta_{ijk} \equiv P(X_i = x_{ij} | Y = y_k)$

- Classify (X^{new})

$$Y^{new} \leftarrow \arg\max_{y_k} \; P(Y = y_k) \prod_i P(X_i^{new} | Y = y_k)$$

$$Y^{new} \leftarrow \arg\max_{y_k} \; \pi_k \prod_i \theta_{ijk}$$

* probabilities must sum to 1, so need estimate only n-1 of these...

Gaussian Naïve Bayes Algorithm – continuous X_i (but still discrete Y)

- Train Naïve Bayes (examples)

  for each value y_k

  estimate* $\pi_k \equiv P(Y = y_k)$

  for each attribute X_i estimate $P(X_i = x_{ij}|Y = y_k)$

  class conditional mean μ_{ik}, variance σ_{ik}

- Classify (X^{new})

$$Y^{new} \leftarrow \arg\max_{y_k} \; P(Y = y_k) \prod_i P(X_i^{new}|Y = y_k)$$

$$Y^{new} \leftarrow \arg\max_{y_k} \; \pi_k \prod_i Normal(X_i^{new}, \mu_{ik}, \sigma_{ik})$$

* probabilities must sum to 1, so need estimate only n-1 parameters...

Learning to classify text documents

- Classify which emails are spam?
- Classify which emails promise an attachment?
- Classify which web pages are student home pages?

How shall we represent text documents for Naïve Bayes?

Baseline: Bag of Words Approach

aardvark	0
about	2
all	2
Africa	1
apple	0
anxious	0
...	
gas	1
...	
oil	1
…	
Zaire	0

Learning to classify document: P(Y|X) the "Bag of Words" model

- Y discrete valued. e.g., Spam or not
- $X = \langle X_1, X_2, \ldots X_n \rangle$ = document

- X_i is a random variable describing the word at position i in the document
- possible values for X_i : any word w_k in English

- Document = bag of words: the vector of counts for all w_k's
  - like #heads, #tails, but we have many more than 2 values
  - assume word probabilities are position independent (i.i.d. rolls of a 50,000-sided die)

Naïve Bayes Algorithm – discrete X_i

- Train Naïve Bayes (examples)

  for each value y_k

  estimate $\pi_k \equiv P(Y = y_k)$

  for each value x_j of each attribute X_i

  estimate $\theta_{ijk} \equiv P(X_i = x_j | Y = y_k)$

  prob that word x_j appears in position i, given $Y=y_k$

- Classify (X^{new})

$$Y^{new} \leftarrow \arg\max_{y_k} P(Y = y_k) \prod_i P(X_i^{new} | Y = y_k)$$

$$Y^{new} \leftarrow \arg\max_{y_k} \pi_k \prod_i \theta_{ijk}$$

* Additional assumption: word probabilities are position independent

$$\theta_{ijk} = \theta_{mjk} \quad \text{for all } i, m$$

MAP estimates for bag of words

Map estimate for multinomial

$$\theta_i = \frac{\alpha_i + \beta_i - 1}{\sum_{m=1}^{k} \alpha_m + \sum_{m=1}^{k} (\beta_m - 1)}$$

Observed count for word m

Hallucinated count for word m

What β's should we choose?

Twenty NewsGroups

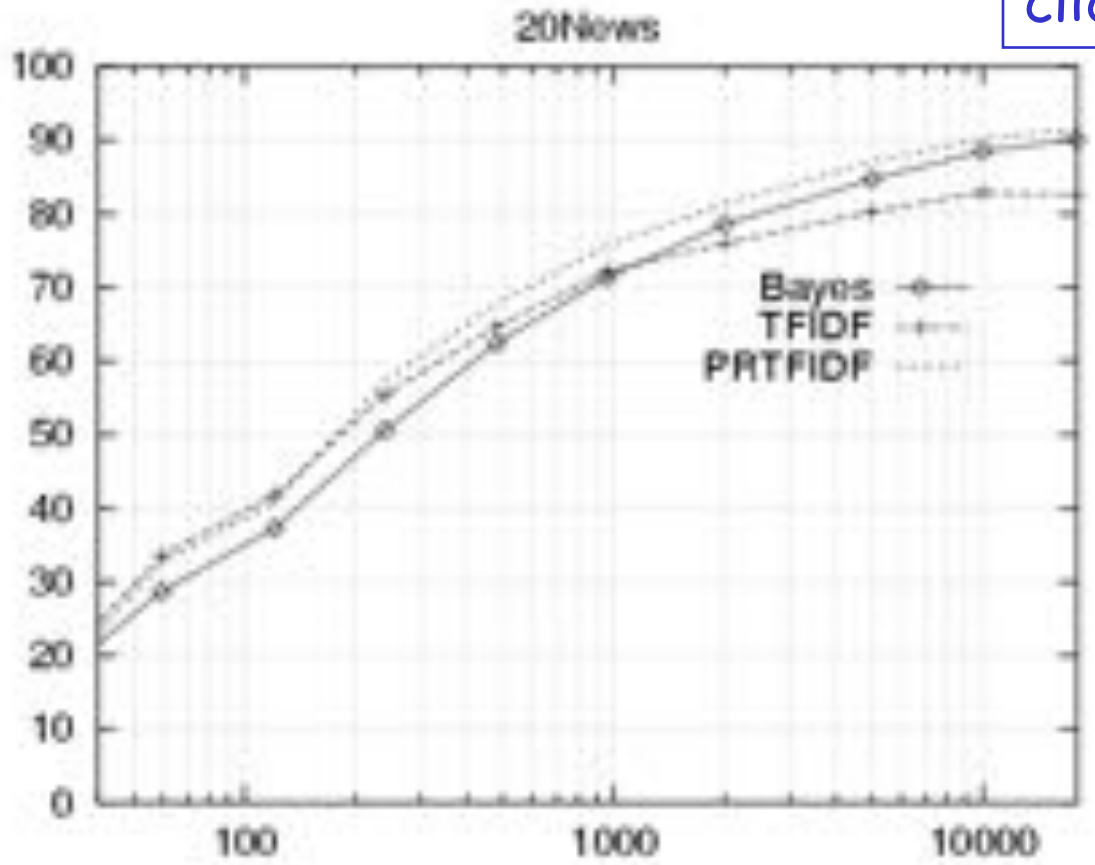
Given 1000 training documents from each group
Learn to classify new documents according to
which newsgroup it came from

comp.graphics	misc.forsale
comp.os.ms-windows.misc	rec.autos
comp.sys.ibm.pc.hardware	rec.motorcycles
comp.sys.mac.hardware	rec.sport.baseball
comp.windows.x	rec.sport.hockey
alt.atheism	sci.space
soc.religion.christian	sci.crypt
talk.religion.misc	sci.electronics
talk.politics.mideast	sci.med
talk.politics.misc	
talk.politics.guns	

Naive Bayes: 89% classification accuracy

Learning Curve for 20 Newsgroups

For code and data, see
www.cs.cmu.edu/~tom/mlbook.html
click on "Software and Data"



Accuracy vs. Training set size (1/3 withheld for test)

What you should know:

- Training and using classifiers based on Bayes rule
- Conditional independence
  - What it is
  - Why it's important
- Naïve Bayes
  - What it is
  - Why we use it so much
  - Training using MLE, MAP estimates
  - Discrete variables and continuous (Gaussian)

Questions:

- How can we extend Naïve Bayes if just 2 of the X_i's are dependent?
- What does the decision surface of a Naïve Bayes classifier look like?
- What error will the classifier achieve if Naïve Bayes assumption is satisfied and we have infinite training data?
- Can you use Naïve Bayes for a combination of discrete and real-valued X_i?

Graphical Models

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*

- Two types of graphical models:
  - Directed graphs (aka Bayesian Networks)
  - Undirected graphs (aka Markov Random Fields)

our focus

Graphical Models – Why Care?

- Unify statistics, probability, machine learning
- Graphical models allow combining:
  - Prior knowledge about 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, ...
- Increasingly, deep nets are also probabilistic 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 $P(X|Y,Z) = P(X|Z)$

or $X \perp Y \mid Z$

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

Marginal Independence

Definition: X is marginally independent of Y if

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

Which we may write as $X \perp Y$

Equivalently, if

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

Equivalently, if

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

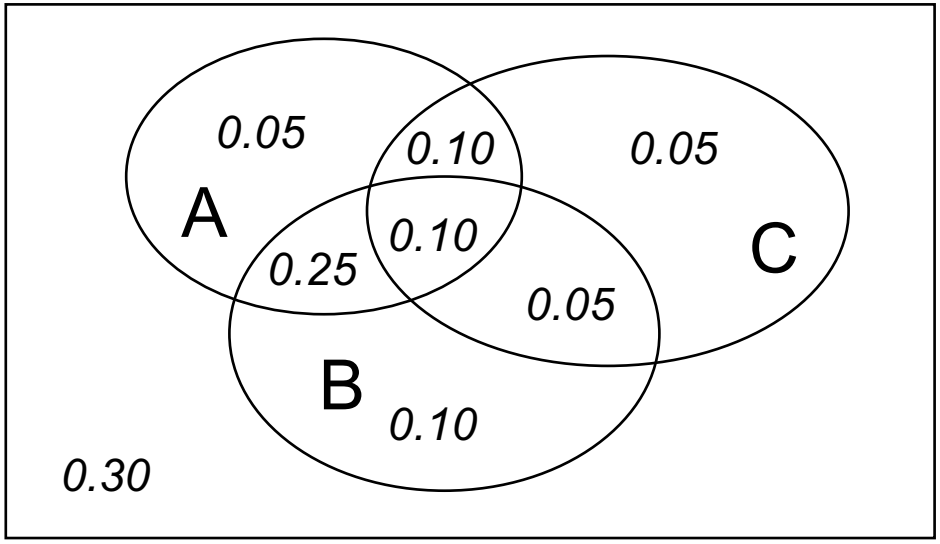
1. Representing Joint Probability Distributions using Bayesian Networks

The Joint Distribution

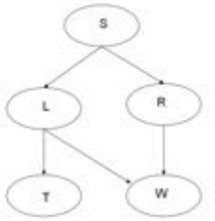
Joint distribution assigns a probability to each possible joint assignment of variables

P(A, B, C)

A	B	C	Prob
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10



Bayesian Networks Definition



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

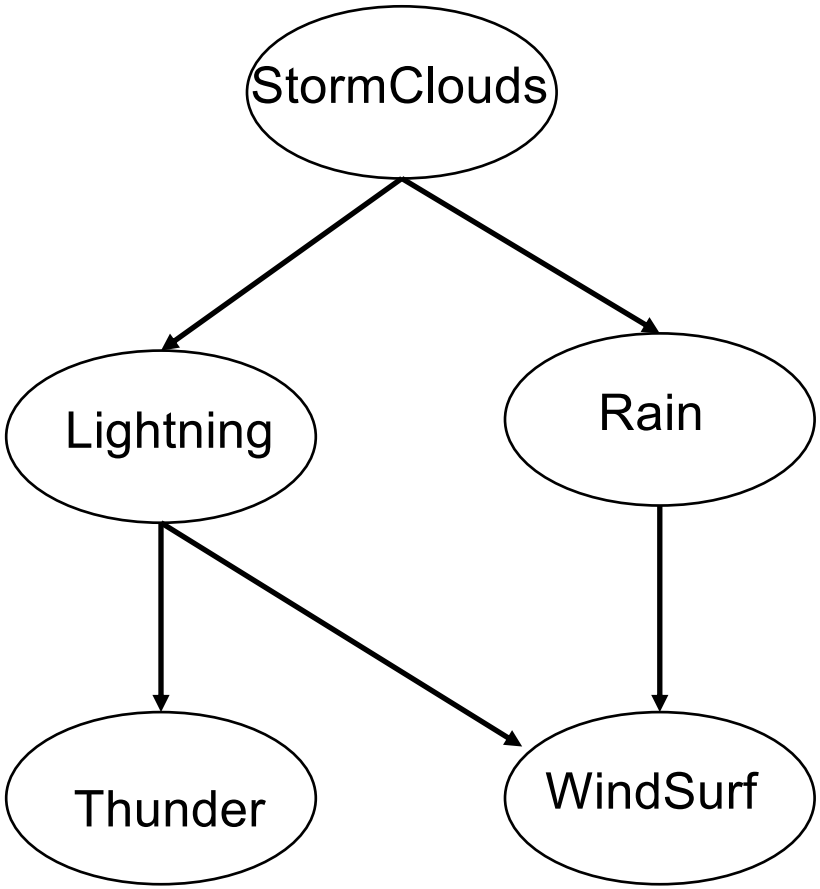
A Bayes network is a directed acyclic graph and a set of Conditional Probability Distributions (CPD's)

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

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

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)). For example:

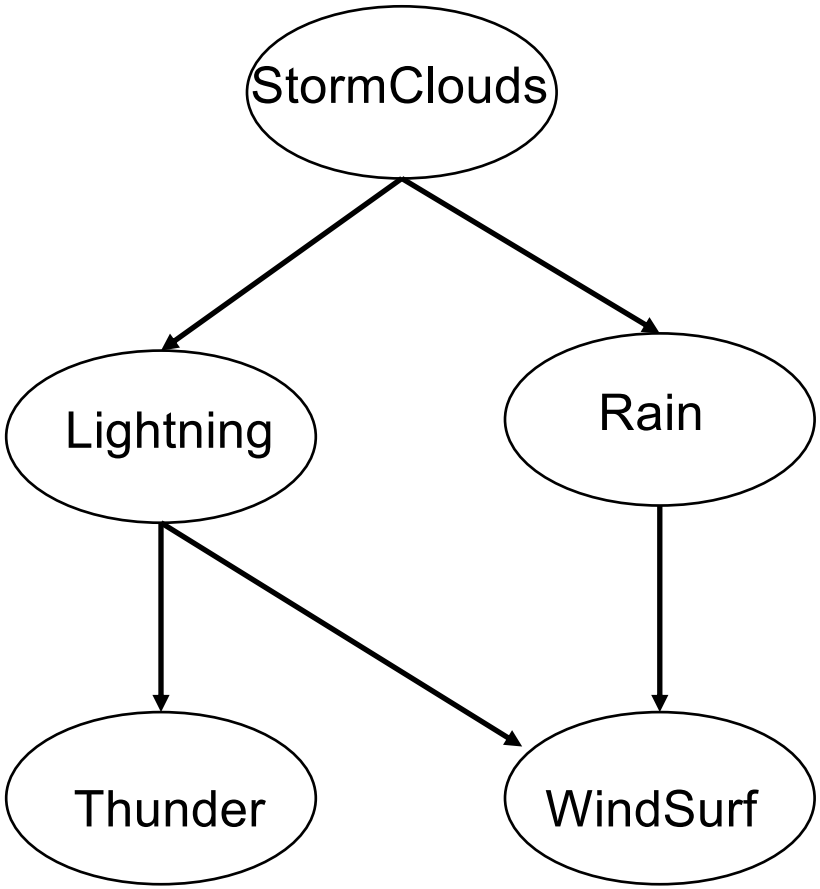
Parents	P(W|Pa)	P(¬W|Pa)
L, R	0	1.0
L, ¬R	0	1.0
¬L, R	0.2	0.8
¬L, ¬R	0.9	0.1

WindSurf

The joint distribution over all variables:

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

Bayesian Network



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

One thing we can say:

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

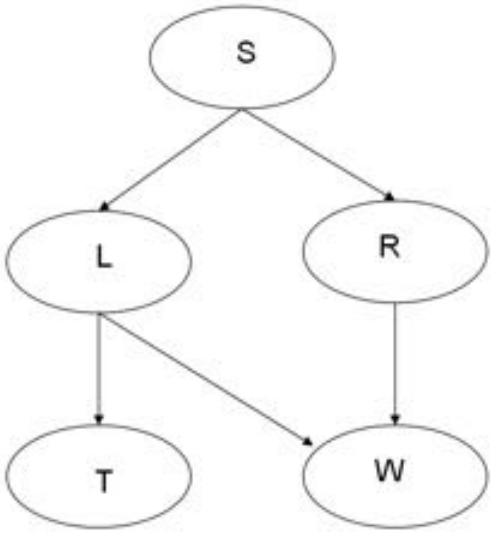
Some helpful terminology

Parents = Pa(X) = immediate parents of X

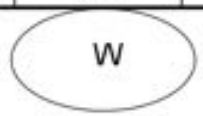
Antecedents = parents, parents of parents, ...

Children = immediate children

Descendents = children, children of children, ...

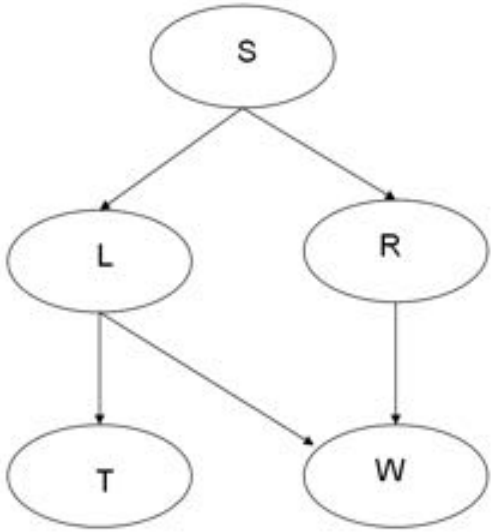


Parents	P(W|Pa)	P(¬W|Pa)
L, R	0	1.0
L, ¬R	0	1.0
¬L, R	0.2	0.8
¬L, ¬R	0.9	0.1



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

W

Chain rule of probability says that in general:

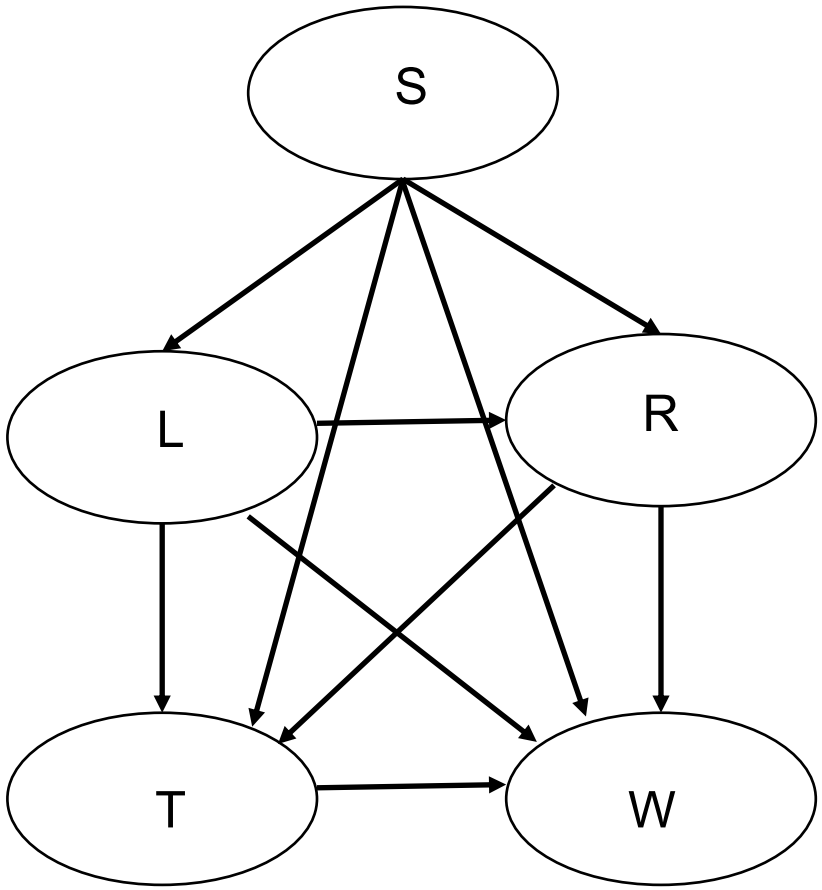
$$P(S, L, R, T, W) = P(S)P(L|S)P(R|S, L)P(T|S, L, R)P(W|S, L, R, T)$$

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

What is the Bayes Net assuming **no** conditional independencies?

Chain rule of probability says that in general:

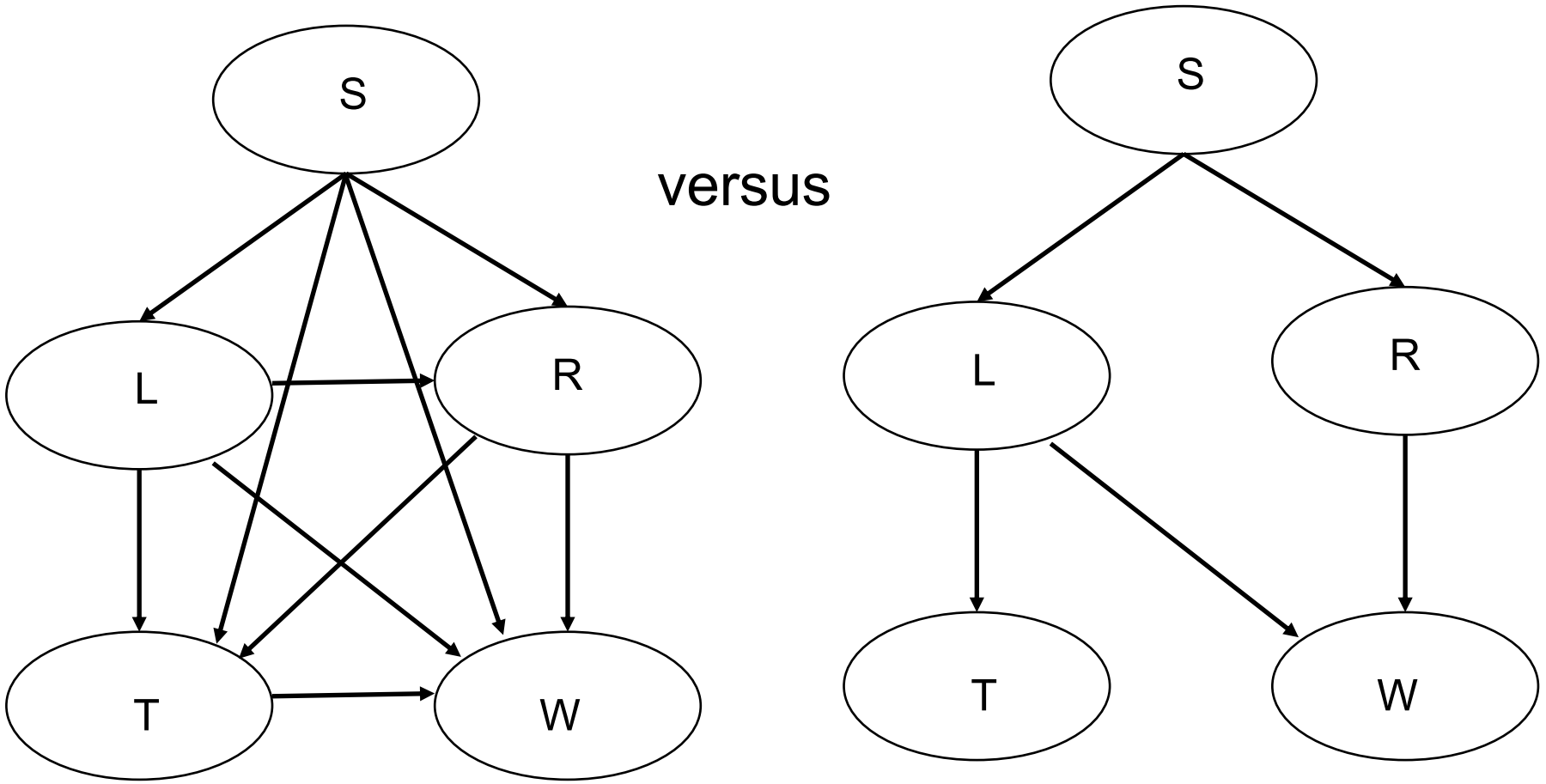
$$P(S, L, R, T, W) = P(S)P(L|S)P(R|S, L)P(T|S, L, R)P(W|S, L, R, T)$$



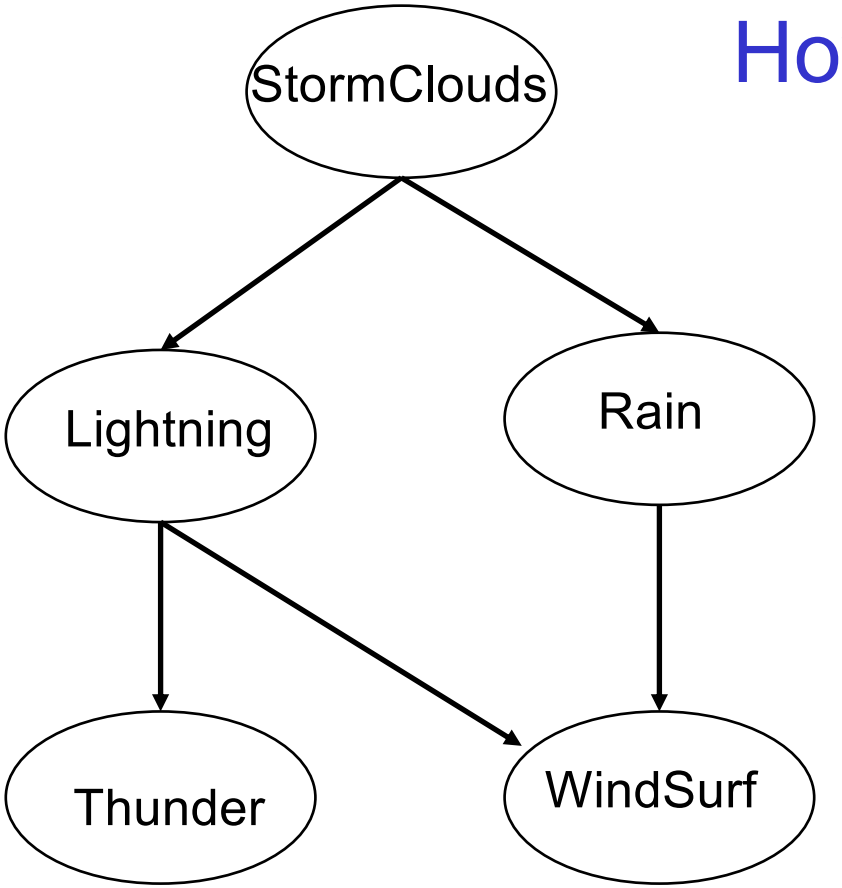
What is the Bayes Net assuming **no** conditional independencies?

Chain rule of probability says that in general:

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



How Many Parameters?



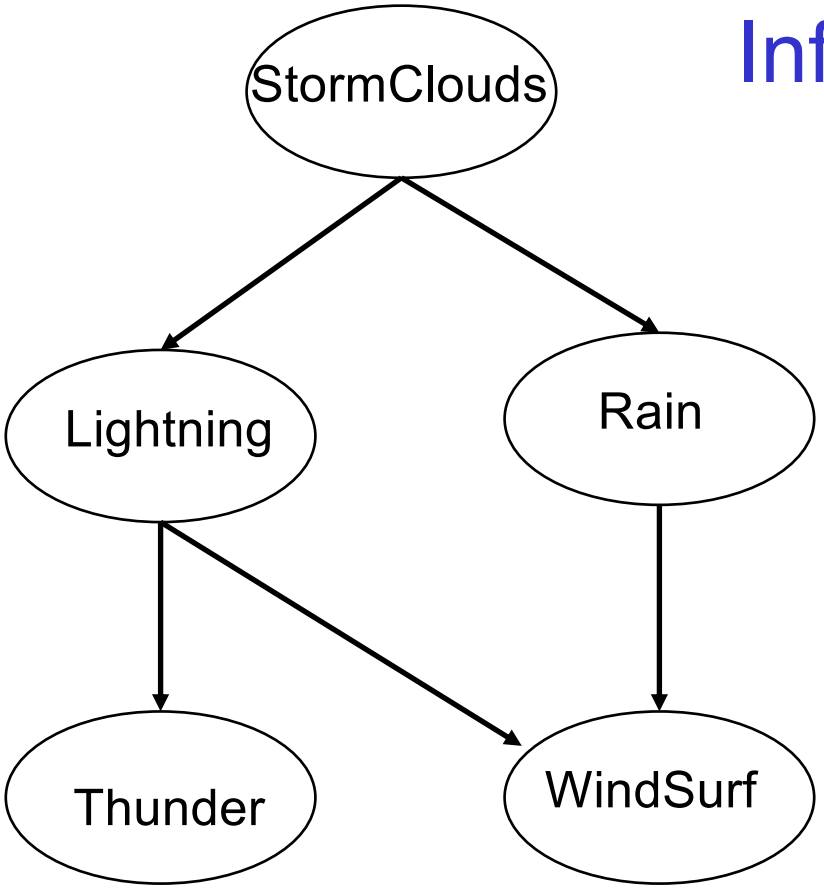
Parents	P(W|Pa)	P(¬W|Pa)
L, R	0	1.0
L, ¬R	0	1.0
¬L, R	0.2	0.8
¬L, ¬R	0.9	0.1

WindSurf

To define joint distribution in general?

To define joint distribution for this Bayes Net?

Inference in Bayes Nets

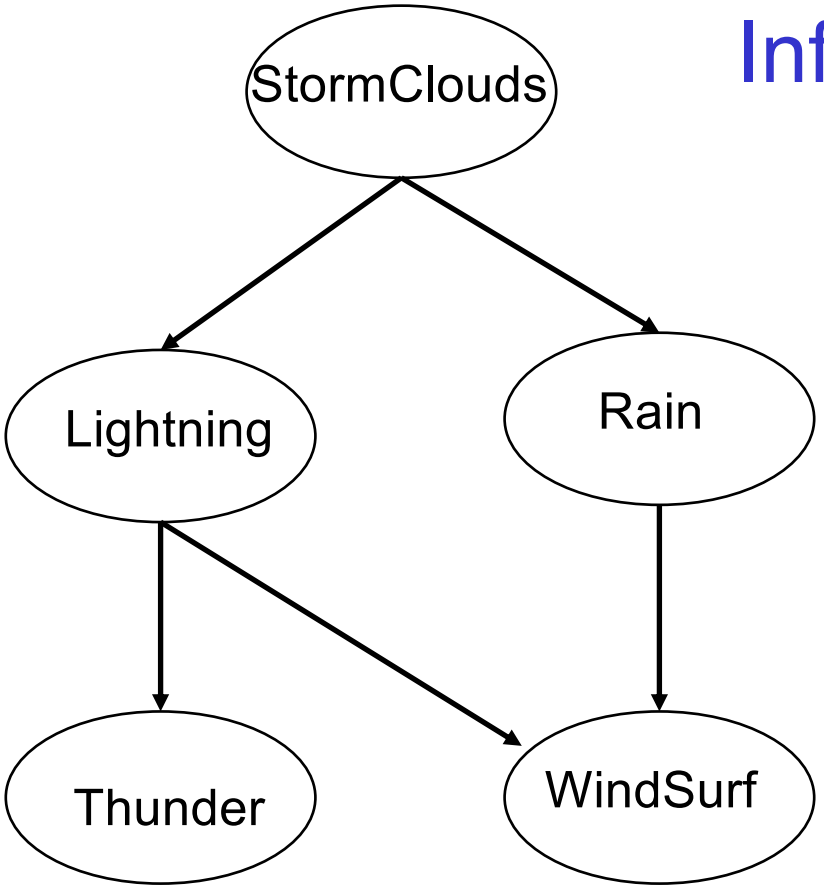


Parents	P(W|Pa)	P(¬W|Pa)
L, R	0	1.0
L, ¬R	0	1.0
¬L, R	0.2	0.8
¬L, ¬R	0.9	0.1

WindSurf

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

Inference in Bayes Nets



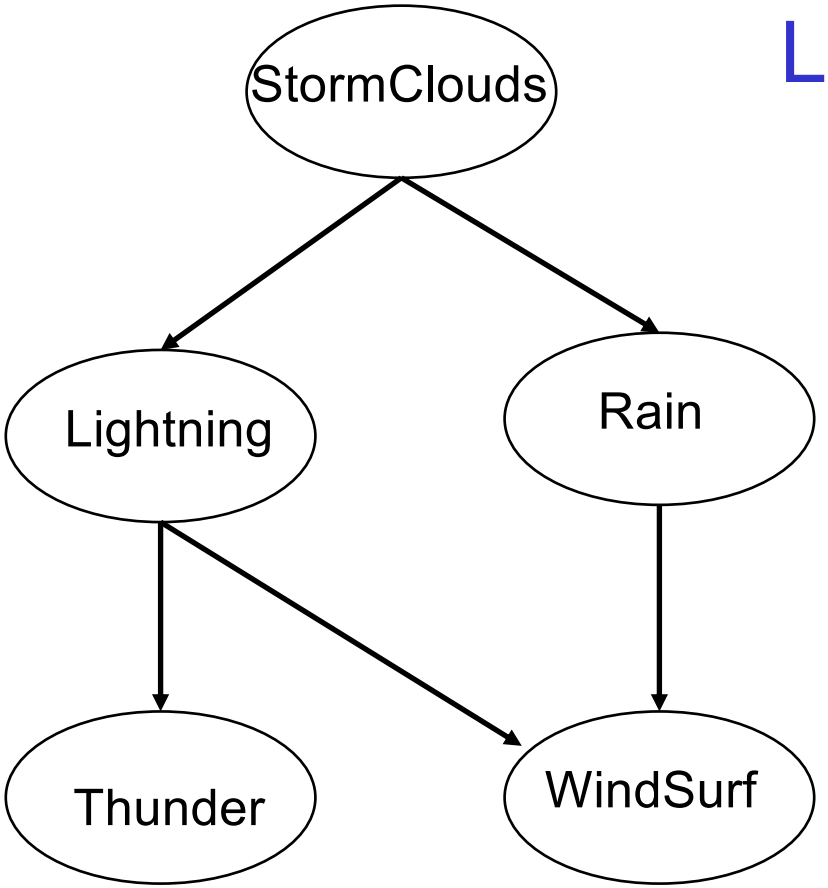
Parents	P(W|Pa)	P(¬W|Pa)
L, R	0	1.0
L, ¬R	0	1.0
¬L, R	0.2	0.8
¬L, ¬R	0.9	0.1

WindSurf

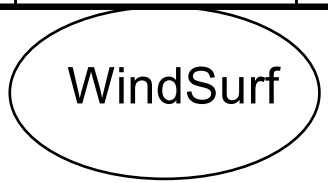
$$P(S=1, L=0, R=1, T=0, W=1) =$$

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

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?

Constructing a 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

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

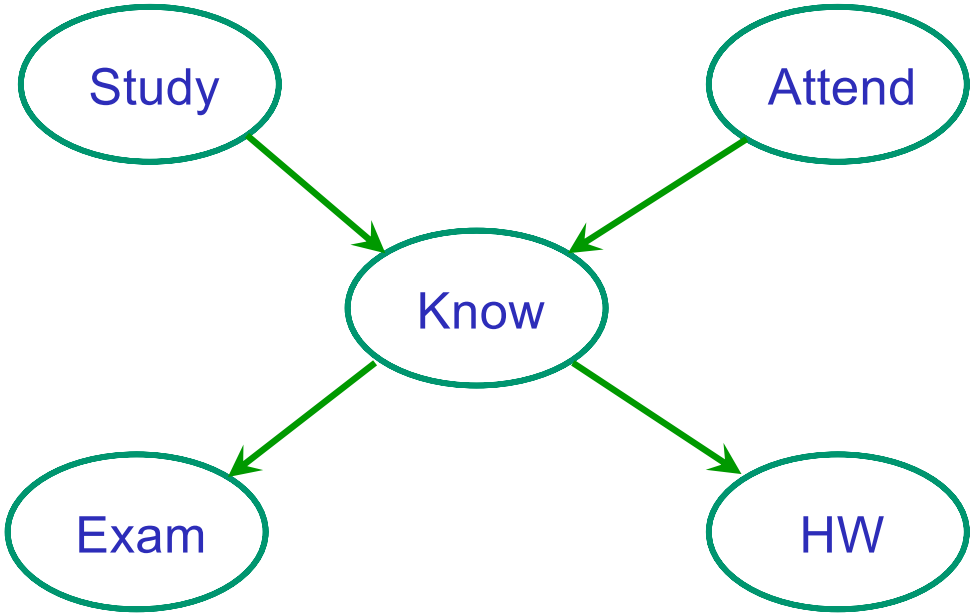
$$= \prod_i P(X_i|Pa(X_i)) \quad \text{(by construction)}$$

Example

- Attending class and Studying both cause you to Know the course material
- Knowing the course material determines whether you pass the Exam, and ace the HW

Example

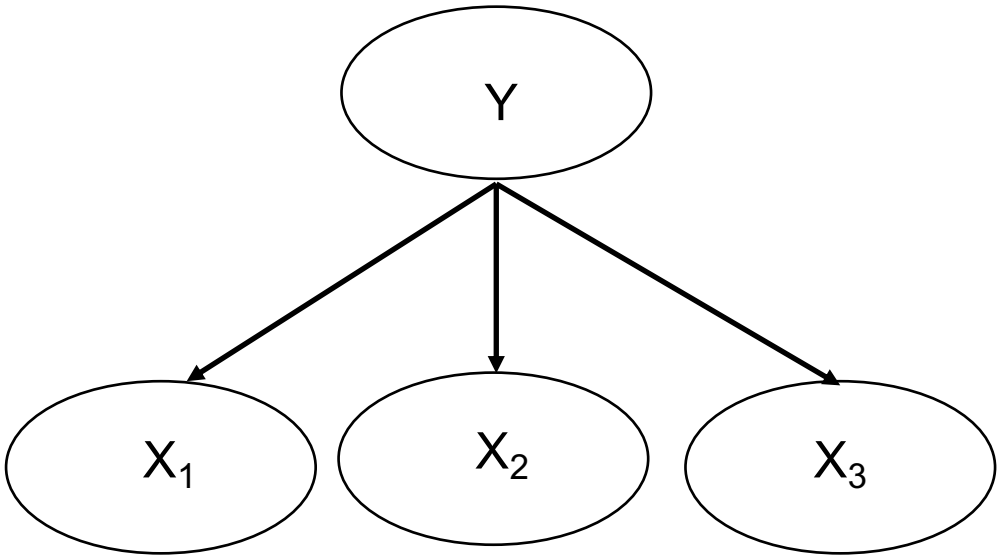
- Attending class and Studying both cause you to Know the course material
- Knowing the course material determines whether you pass the Exam, and ace the HW



What is the Bayes Network for Naïve Bayes?

What is the Bayes Net representing Naïve Bayes?

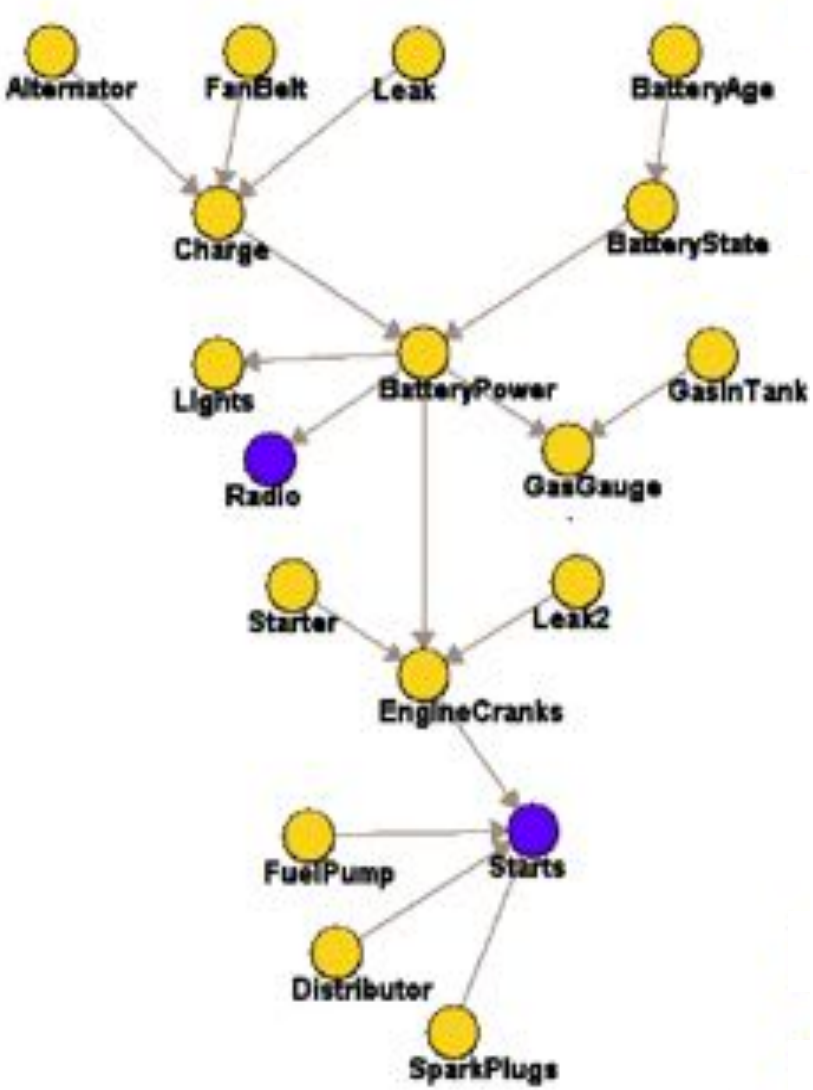
Naïve Bayes assumes all pairs of inputs X_i and X_k conditionally independent, given the label Y



Recall:

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

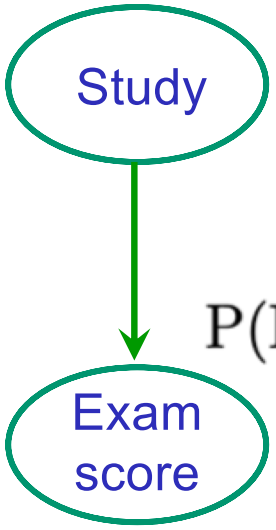
What do we do if variables are mix of discrete and real valued?



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

No problem! Just define CPD as appropriate!

$$\text{Suppose Study} \in \{0, 1\}, \text{ExamScore} \in \Re$$



$$\text{P(ExamScore} \mid \text{Study}) = \begin{array}{l} \text{if Study=0 then } \mathcal{N}(60, 20) \\ \text{if Study=1 then } \mathcal{N}(85, 10) \end{array}$$

where $\mathcal{N}(\mu, \sigma)$ is the Normal distribution

What You Should Know

- Bayes nets are convenient representation for encoding dependencies / conditional independence
- BN = Graph plus parameters of CPD's
  - Defines joint distribution over variables
  - Can calculate everything else from that
  - Though inference may be intractable
- Reading conditional independence relations from the graph
  - Each node is cond indep of non-descendents, given only its parents
  - (Optional) D-Separation is a rule that gives a *complete* accounting of all conditional independencies. (see optional 10 minute video http://www.cs.cmu.edu/~tom/10701-S20/VSV_GrMod_Representation.mp4)