CS 343: Artificial Intelligence
Bayesian Networks

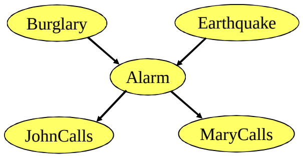
Raymond J. Mooney
University of Texas at Austin

Graphical Models

- If no assumption of independence is made, then an exponential number of parameters must be estimated for sound probabilistic inference.
- No realistic amount of training data is sufficient to estimate so many parameters.
- If a blanket assumption of conditional independence is made, efficient training and inference is possible, but such a strong assumption is rarely warranted.
- **Graphical models** use directed or undirected graphs over a set of random variables to explicitly specify variable dependencies and allow for less restrictive independence assumptions while limiting the number of parameters that must be estimated.
  - **Bayesian Networks**: Directed acyclic graphs that indicate causal structure.
  - **Markov Networks**: Undirected graphs that capture general dependencies.

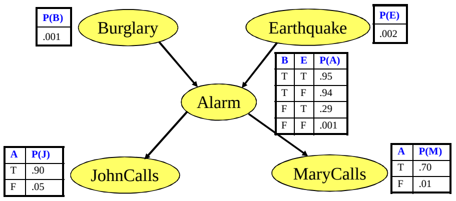
Bayesian Networks

- Directed Acyclic Graph (DAG)
  - Nodes are random variables
  - Edges indicate causal influences



Conditional Probability Tables

- Each node has a **conditional probability table** (**CPT**) that gives the probability of each of its values given every possible combination of values for its parents (conditioning case).
  - Roots (sources) of the DAG that have no parents are given prior probabilities.



P(B)
.001

P(E)
.002

B	E	P(A)
T	T	.95
T	F	.94
F	T	.29
F	F	.001

A	P(J)
T	.90
F	.05

A	P(M)
T	.70
F	.01

CPT Comments

- Probability of false not given since rows must add to 1.
- Example requires 10 parameters rather than $2^5–1 = 31$ for specifying the full joint distribution.
- Number of parameters in the CPT for a node is exponential in the number of parents (fan-in).

Joint Distributions for Bayes Nets

- A Bayesian Network implicitly defines a joint distribution.

$$P(x_1, x_2, \ldots x_n) = \prod_{i=1}^{n} P(x_i \mid \text{Parents}(X_i))$$

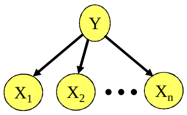
- Example

$$P(J \wedge M \wedge A \wedge \neg B \wedge \neg E)$$
$$= P(J \mid A) P(M \mid A) P(A \mid \neg B \wedge \neg E) P(\neg B) P(\neg E)$$
$$= 0.9 \times 0.7 \times 0.001 \times 0.999 \times 0.998 = 0.00062$$

- Therefore an inefficient approach to inference is:
  - 1) Compute the joint distribution using this equation.
  - 2) Compute any desired conditional probability using the joint distribution.

Naïve Bayes as a Bayes Net

- Naïve Bayes is a simple Bayes Net



- Priors P(Y) and conditionals P($X_i|Y$) for Naïve Bayes provide CPTs for the network.

Independencies in Bayes Nets

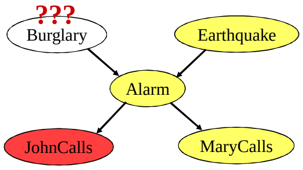
- If removing a subset of nodes S from the network renders nodes X_i and X_j disconnected, then X_i and X_j are independent given S, i.e. P(X_i | X_j, S) = P(X_i | S)
- However, this is too strict a criteria for conditional independence since two nodes will still be considered independent if their simply exists some variable that depends on both.
  - For example, Burglary and Earthquake should be considered independent since they both cause Alarm.

Independencies in Bayes Nets (cont.)

- Unless we know something about a common effect of two “independent causes” or a descendent of a common effect, then they can be considered independent.
  - For example, if we know nothing else, Earthquake and Burglary are independent.
- However, if we have information about a common effect (or descendent thereof) then the two “independent” causes become probabilistically linked since evidence for one cause can “explain away” the other.
  - For example, if we know the alarm went off that someone called about the alarm, then it makes earthquake and burglary dependent since evidence for earthquake decreases belief in burglary. and vice versa.

Bayes Net Inference

- Given known values for some **evidence variables**, determine the posterior probability of some **query variables**.
- Example: Given that John calls, what is the probability that there is a Burglary?

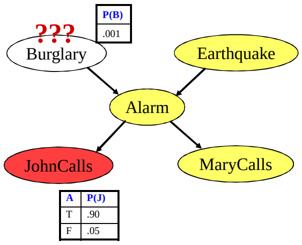


John calls 90% of the time there is an Alarm and the Alarm detects 94% of Burglaries so people generally think it should be fairly high.

However, this ignores the prior probability of John calling.

Bayes Net Inference

- Example: Given that John calls, what is the probability that there is a Burglary?



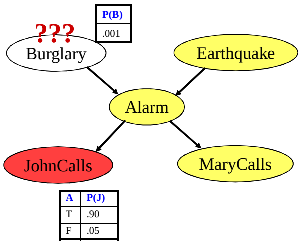
P(B)
.001

A	P(J)
T	.90
F	.05

John also calls 5% of the time when there is no Alarm. So over 1,000 days we expect 1 Burglary and John will probably call. However, he will also call with a false report 50 times on average. So the call is about 50 times more likely a false report: P(Burglary | JohnCalls) ≈ 0.02

Bayes Net Inference

- Example: Given that John calls, what is the probability that there is a Burglary?

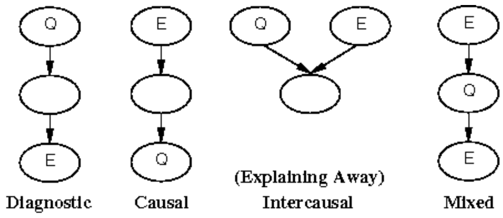


P(B)
.001

A	P(J)
T	.90
F	.05

Actual probability of Burglary is 0.016 since the alarm is not perfect (an Earthquake could have set it off or it could have gone off on its own). On the other side, even if there was not an alarm and John called incorrectly, there could have been an undetected Burglary anyway, but this is unlikely.

Types of Inference



Sample Inferences

- **Diagnostic (evidential, abductive)**: From effect to cause.
  - P(Burglary | JohnCalls) = 0.016
  - P(Burglary | JohnCalls ∧ MaryCalls) = 0.29
  - P(Alarm | JohnCalls ∧ MaryCalls) = 0.76
  - P(Earthquake | JohnCalls ∧ MaryCalls) = 0.18
- **Causal (predictive)**: From cause to effect
  - P(JohnCalls | Burglary) = 0.86
  - P(MaryCalls | Burglary) = 0.67
- **Intercausal (explaining away)**: Between causes of a common effect.
  - P(Burglary | Alarm) = 0.376
  - P(Burglary | Alarm ∧ Earthquake) = 0.003
- **Mixed**: Two or more of the above combined
  - (diagnostic and causal) P(Alarm | JohnCalls ∧ ¬Earthquake) = 0.03
  - (diagnostic and intercausal) P(Burglary | JohnCalls ∧ ¬Earthquake) = 0.017

Probabilistic Inference in Humans

- People are notoriously bad at doing correct probabilistic reasoning in certain cases.
- One problem is they tend to ignore the influence of the prior probability of a situation.

Monty Hall Problem

One Line Demo:

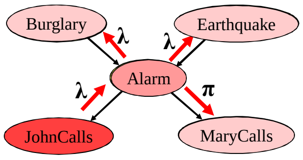
http://math.ucsd.edu/~crypto/Monty/monty.html

Complexity of Bayes Net Inference

- In general, the problem of Bayes Net inference is NP-hard (exponential in the size of the graph).
- For **singly-connected networks** or **polytrees** in which there are no undirected loops, there are linear-time algorithms based on **belief propagation**.
  - Each node sends local evidence messages to their children and parents.
  - Each node updates belief in each of its possible values based on incoming messages from it neighbors and propagates evidence on to its neighbors.
- There are approximations to inference for general networks based on **loopy belief propagation** that iteratively refines probabilities that converge to accurate values in the limit.

Belief Propagation Example

- λ messages are sent from children to parents representing abductive evidence for a node.
- π messages are sent from parents to children representing causal evidence for a node.



Belief Propagation Details

- Each node *B* acts as a simple processor which maintains a vector λ(*B*) for the total evidential support for each value of its corresponding variable and an analogous vector π(*B*) for the total causal support.
- The belief vector *BEL*(*B*) for a node, which maintains the probability for each value, is calculated as the normalized product:

$$BEL(B) = \alpha\, \lambda(B)\, \pi(B)$$

- Computation at each node involve λ and π **message vectors** sent between nodes and consists of simple matrix calculations using the CPT to update belief (the λ and π node vectors) for each node based on new evidence.

Belief Propagation Details (cont.)

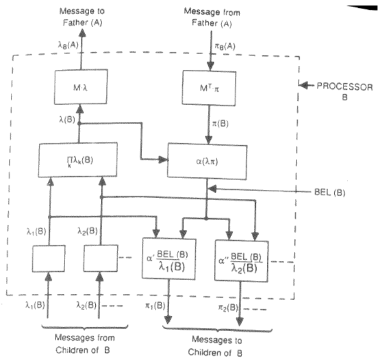
- Assumes the CPT for each node is a matrix (*M*) with a column for each value of the node's variable and a row for each conditioning case (all rows must sum to 1).

Matrix *M* for the Alarm node

$$\begin{array}{r|c|c|}
 & \multicolumn{2}{c}{\text{Value of Alarm}} \\
 & \text{T} & \text{F} \\
\text{Values of Burglary and Earthquake} \quad \text{TT} & 0.95 & 0.05 \\
\text{TF} & 0.94 & 0.06 \\
\text{FT} & 0.29 & 0.71 \\
\text{FF} & 0.001 & 0.999
\end{array}$$

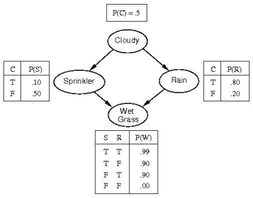
- Propagation algorithm is simplest for trees in which each node has only one parent (i.e. one cause).
- To initialize, λ(*B*) for all leaf nodes is set to all 1's and π(*B*) of all root nodes is set to the priors given in the CPT. Belief based on the root priors is then propagated down the tree to all leaves to establish priors for all nodes.
- Evidence is then added incrementally and the effects propagated to other nodes.

Processor for Tree Networks



Multiply Connected Networks

- Networks with undirected loops, more than one directed path between some pair of nodes.



C	P(S)
T	.10
F	.50

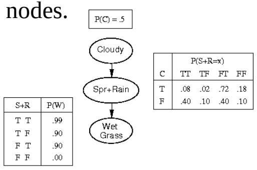
C	P(R)
T	.80
F	.20

S	R	P(W)
T	T	.99
T	F	.90
F	T	.90
F	F	.00

- In general, inference in such networks is NP-hard.
- Some methods construct a polytree(s) from given network and perform inference on transformed graph.

Node Clustering

- Eliminate all loops by merging nodes to create **meganodes** that have the cross-product of values of the merged nodes.



C	P(S+R=x) TT	TF	FT	FF
T	.08	.02	.72	.18
F	.40	.10	.40	.10

S+R		P(W)
T	T	.99
T	F	.90
F	T	.90
F	F	.00

- Number of values for merged node is exponential in the number of nodes merged.
- Still reasonably tractable for many network topologies requiring relatively little merging to eliminate loops.

Bayes Nets Applications

- Medical diagnosis
  - Pathfinder system outperforms leading experts in diagnosis of lymph-node disease.
- Microsoft applications
  - Problem diagnosis: printer problems
  - Recognizing user intents for HCI
- Text categorization and spam filtering
- Student modeling for intelligent tutoring systems.

Statistical Revolution

- Across AI there has been a movement from logic-based approaches to approaches based on probability and statistics.
  - Statistical natural language processing
  - Statistical computer vision
  - Statistical robot navigation
  - Statistical learning
- Most approaches are feature-based and “propositional” and do not handle complex relational descriptions with multiple entities like those typically requiring predicate logic.

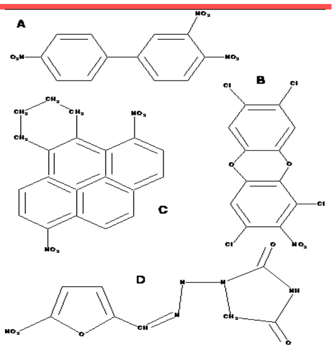
Structured (Multi-Relational) Data

- In many domains, data consists of an unbounded number of entities with an arbitrary number of properties and relations between them.
  - Social networks
  - Biochemical compounds
  - Web sites

Biochemical Data

Predicting mutagenicity

[Srinivasan et. al, 1995]



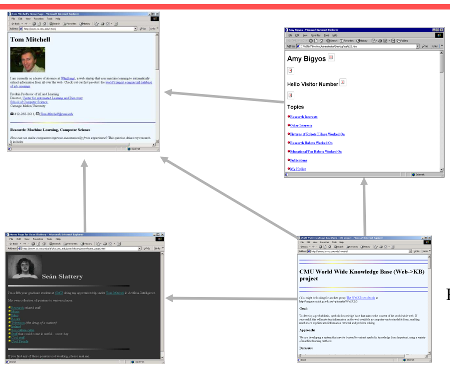
Web-KB Dataset [Slattery & Craven, 1998]

Faculty

Other

Grad Student

Research Project



Tom Mitchell

I am currently on a leave of absence at WhizBang!, a web startup that uses machine learning to automatically extract information from all over the web. Check out our first product: the world's largest commercial database of job openings.

Fredkin Professor of AI and Learning
Director, Center for Automated Learning and Discovery
School of Computer Science.
Carnegie Mellon University

☎ 412-268-2611, Tom.Mitchell@cmu.edu

Research: Machine Learning, Computer Science

How can we make computers improve automatically from experience? This question drives my research. It includes

Amy Bigyos

Hello Visitor Number

Topics

- Research Interests
- Other Interests
- Pictures of Robots I Have Worked On
- Research Robots Worked On
- Educational/Fun Robots Worked On
- Publications
- My Hotlist

Seán Slattery

I'm a fifth year graduate student at CMU doing my apprenticeship under Tom Mitchell in Artificial Intelligence.

My own collection of pointers to various places:

- Research related stuff
- Music
- Films
- Books
- Television *(the drug of a nation)*
- Ireland
- Soc.culture.celtic
- Stuff that could come in useful... some day.
- Cool stuff
- Cool People

If you find any of these pointers not working, please mail me.

CMU World Wide Knowledge Base (Web->KB) project

(You might be looking for another group: The WebKB set of tools at http://meganesia.int.gu.edu.au/~phmartin/WebKB/)

Goal:

To develop a probabilistic, symbolic knowledge base that mirrors the content of the world wide web. If successful, this will make text information on the web available in computer-understandable form, enabling much more sophisticated information retrieval and problem solving.

Approach:

We are developing a system that can *be trained* to extract symbolic knowledge from hypertext, using a variety of machine learning methods.

Datasets:

Collective Classification

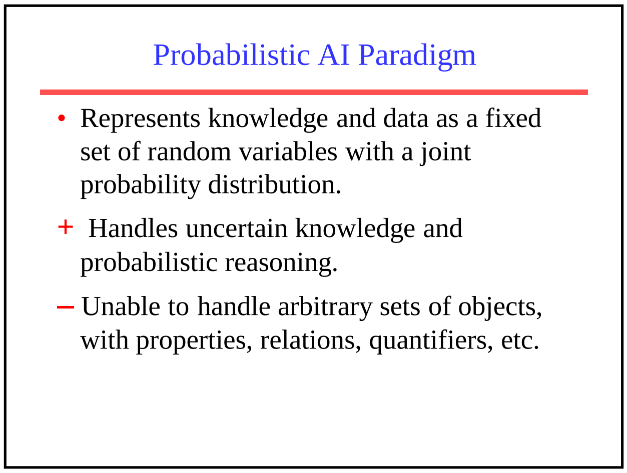
- Traditional learning methods assume that objects to be classified are independent (the first “i” in the i.i.d. assumption)
- In structured data, the class of an entity can be influenced by the classes of related entities.
- Need to assign classes to all objects simultaneously to produce the most probable globally-consistent interpretation.

Logical AI Paradigm

- Represents knowledge and data in a binary symbolic logic such as FOPC.

+ Rich representation that handles arbitrary sets of objects, with properties, relations, quantifiers, etc.

– Unable to handle uncertain knowledge and probabilistic reasoning.

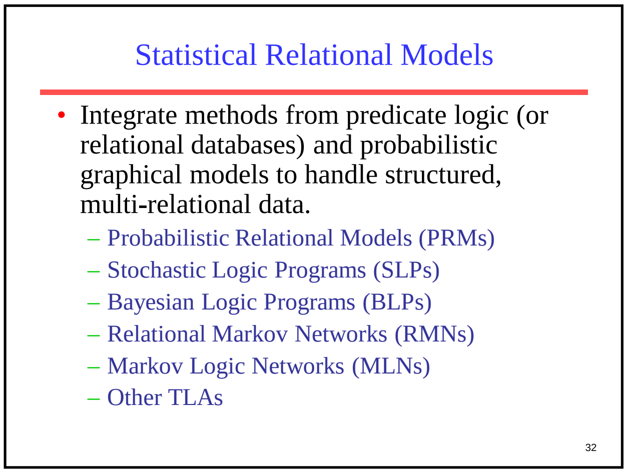


Probabilistic AI Paradigm

- Represents knowledge and data as a fixed set of random variables with a joint probability distribution.

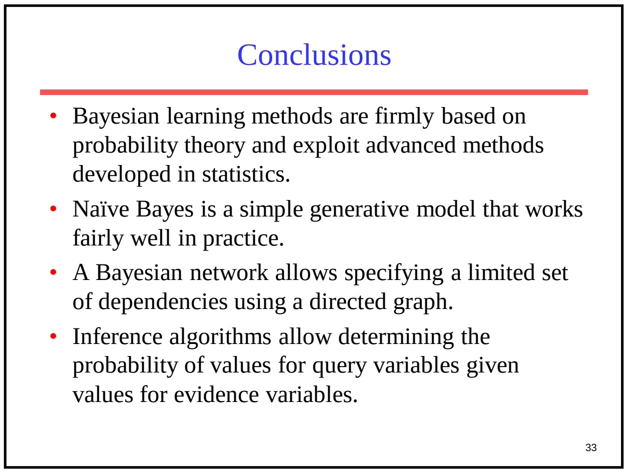
+ Handles uncertain knowledge and probabilistic reasoning.

– Unable to handle arbitrary sets of objects, with properties, relations, quantifiers, etc.



Statistical Relational Models

- Integrate methods from predicate logic (or relational databases) and probabilistic graphical models to handle structured, multi-relational data.
  - Probabilistic Relational Models (PRMs)
  - Stochastic Logic Programs (SLPs)
  - Bayesian Logic Programs (BLPs)
  - Relational Markov Networks (RMNs)
  - Markov Logic Networks (MLNs)
  - Other TLAs



Conclusions

- Bayesian learning methods are firmly based on probability theory and exploit advanced methods developed in statistics.
- Naïve Bayes is a simple generative model that works fairly well in practice.
- A Bayesian network allows specifying a limited set of dependencies using a directed graph.
- Inference algorithms allow determining the probability of values for query variables given values for evidence variables.