Bayesian AI

Introduction

for
IEEE Computational Intelligence Society
IEEE Computer Society

Kevin Korb
Monash University&
Bayesian Intelligence Pty Ltd

kbkorb@gmail.com
www.bayesian-intelligence.com

Contents

Probabilistic Graphical Models

- First systematic use of qualitative graphs for reasoning

  Wigmore (1913) charts for legal reasoning

- First systematic use of quantitative causal graphs

  Sewall Wright (1921) linear path models for analysing population genetics

- Idiot Bayes models for diagnosis (1960s)
- Judea Pearl's (1988) *Probabilistic Reasoning for Intelligent Systems*

Bayes' Theorem

Forward Inference (likelihoods) $P(e|h)$ – e.g., what is the probability of heads given a fair coin?

Bayes' Inverse Inference Rule
Discovered by Rev Thomas Bayes; published posthumously in 1763

$$P(h|e) = \frac{P(e|h)P(h)}{P(e)}$$

$$\text{posterior} = (\text{likelihood} \times \text{prior})\alpha$$

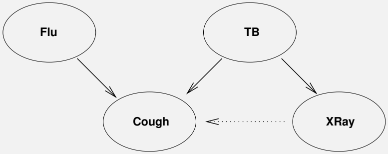
- Forward inference tells us likelihoods
- Finding priors is the main problem in applying Bayes' Rule

Idiot Bayes

First Computer Implementation

If you can't fully implement Bayes, then simplify. Assumptions:

- Diseases marginally independent
  - E.g., Flu and TB independent
- Symptoms independent given disease
  - E.g., Sneezing & Cough given Flu (!?)
- Diseases *remain* independent given symptoms
  - E.g., $P(Flu|Cough, \neg TB) = P(Flu|Cough)$
  - This is obviously wrong!
  - Indeed, if $P(TB \vee Flu) = 1$, $P(Flu|Cough, \neg TB) = 1$



Pearl's Alarm Example

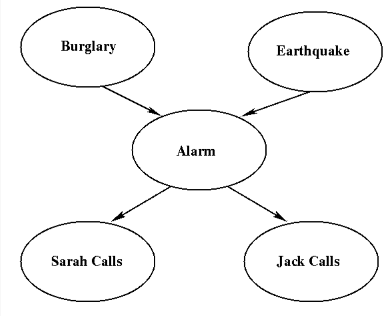


Figure: Pearl's Alarm Example

Probability and Causality

For chains and common causes, the ends are:

- Marginally dependent
- Conditionally independent

E.g., Burglary $\rightarrow$ *Alarm* $\rightarrow$ *Sarah Calls*
Sarah Calls $\leftarrow$ *Alarm* $\rightarrow$ *Jack Calls*

For common effects, the opposite; the ends are

- Marginally independent
- Conditionally dependent ("explaining away")

E.g., Burglary $\rightarrow$ *Alarm* $\leftarrow$ *Earthquake*

The Markov condition

A precondition for proper modeling:

Definition (Markov Condition)

There are no direct dependencies in the system being modeled which are not explicitly shown via arcs.

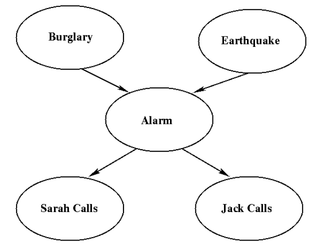
Most problems of interest then admit sparse networks

> *And substantial computational savings over full joint updating!*

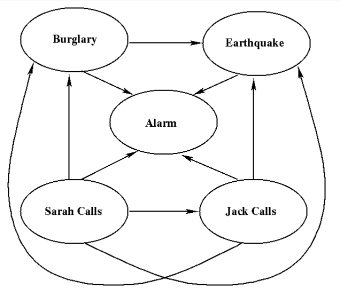
Basic Modeling Rules

Causal Order

Using $\langle B, E, A, S, J \rangle$



Using $\langle S, J, B, E, A \rangle$



Basic Modeling Rules

Causal Order

Symptoms are often confused with causes. E.g.,

$$\begin{array}{lcl} \text{Obesity} & \searrow & \\ \text{Alcohol} & \rightarrow & \text{Steatosis} \\ \text{Triglicerides} & \nearrow & \end{array}$$

versus

$$\begin{array}{ccc} \text{Obesity} & & \text{Alcohol} \\ & \searrow \;\; \swarrow & \\ & \text{Steatosis} & \\ & \downarrow & \\ & \text{Triglicerides} & \end{array}$$

These give rise to **opposite** *prob signatures!*

Basic Modeling Rules

Variable Identification

Variables must be merged when they are

- Semantically dependent
- Functionally dependent

E.g., Var 1 = mean of a binomial, Var 2 = variance of the binomial

Basic Modeling Rules

Minimize Parameters

The more parents, the more parameters. Reduce promiscuity and divorce some parents!

- If parents can be meaningfully grouped, that is

Example

Pesticides in Water, *Flow*, *Habitat Diversity*, *Habitat Range* are all parents of *Population Health*.

Then: channel them through *Water Quality* and *Habitat*

Basic Modeling Rules

Start Small

The easiest mistake to make is to build the full solution first, rather than last!

Bayesian net modeling is like other software engineering tasks — *hard and easily underestimated!*

- We use the iterative prototyping approach of Boehm and Brooks.

Extensions to Bayesian Networks

Decision networks (Influence Diagrams):
For decision making under uncertainty.

Dynamic Bayesian networks:
For reasoning about changes over time: planning under uncertainty; time series modeling.

Rational Decision Making

Bayesian networks can be extended to support rational decision making.

- **Decision theory** = Utility theory + Probability theory.

Definition (Expected Utility)

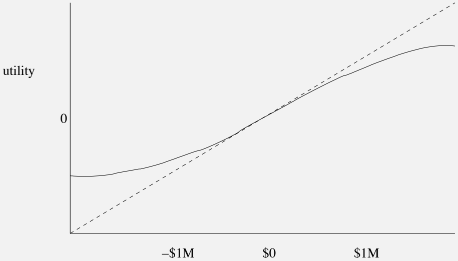
$$EU(A|E) = \sum_i P(O_i|E, A) \times U(O_i|A)$$

- E = available evidence,
- A = an action
- O_i = possible outcome state
- U = utility

Utility

Utility functions map from *Outcomes* $\times$ *Actions* $\rightarrow$ *R*

- Often utility is simplistically equated with money



Formal utility elicitation began with Frank Ramsey (1931).

Type of Nodes

Chance nodes (○) :
Represent random variables, as in plain Bayesian networks. Parents can be decision nodes or other chance nodes.

Decision nodes (□) :
Represent choices between actions.

Utility nodes (♢) :
Represent the utility function. Parents are any variables that directly affect utility, whether chance or decision nodes. Has an associated table representing multi-attribute utility function.

Decision Networks

There are now two kinds of updating possible:

(1) Add observational evidence and update to get posterior probabilities.

(2) Select an alternative decision value. Update to get the decision's consequences.

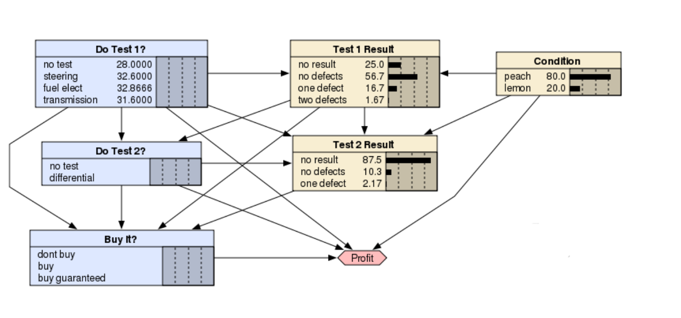
(Of course, these can be combined.)

- We can now find the expected utilities.

If we iterate through all alternative decisions, we can maximize expected utility.

Example

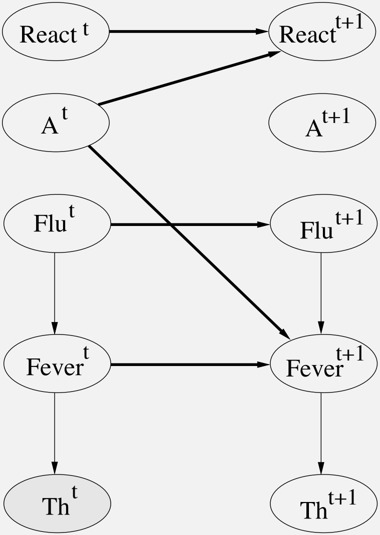
Neapolitan's Car Buyer



Dynamic Belief Networks

- One node for each variable for each time step.
- **Intra-slice** arcs $Flu^T \longrightarrow Fever^T$
- **Inter-slice (temporal)** arcs
  1. $Flu^T \longrightarrow Flu^{T+1}$
  2. $Aspirin^T \longrightarrow Fever^{T+1}$

Fever DBN



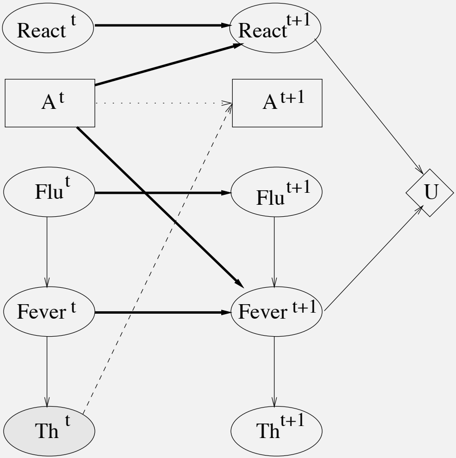
DBN reasoning

- Can calculate distributions for S_{t+1} and further: **probabilistic projection**.
- Reasoning can be done using standard BN updating algorithms
- This type of DBN gets very large, very quickly.
- Usually only keep two time slices of the network.

Dynamic Decision Network

Planning Under Uncertainty

Decision Networks can be expanded dynamically…



BN Tools

GeNie: Bayesian and decision networks; programmable interface; Windows GUI; hierarchical Bayesian networks. Free.
`genie.sis.pitt.edu`

Hugin: Bayesian and decision networks; programmable; GUI; PC learning algorithm; handles large networks. Expensive.
`www.hugin.com`

Netica: Bayesian and decision networks; programmable; easy-to-use Windows GUI. Medium cost; free for small networks.
`www.norsys.com`

Causal Discovery

- Parameterization
  - Linear models: see path modeling
- Structure Learning
  - Constraint-based learning = CI learning
  - Metric learning: Bayesian (or non-Bayesian) scoring function plus search

Learning Conditional Probability Tables

Spiegelhalter & Lauritzen (1990):

- assume parameter independence
- each CPT cell has α_i = a parameter in the Dirichlet distribution:

$$D[\alpha_1, \ldots, \alpha_i, \ldots, \alpha_K]$$

  for K values of the child
- prob of outcome i is $\alpha_i / \Sigma_{k=1}^{K} \alpha_k$
- observing outcome i update D to

$$D[\alpha_1, \ldots, \alpha_i + 1, \ldots, \alpha_K]$$

Available in Netica and Hugin.

(Different techniques are available with Weka and CaMML.)

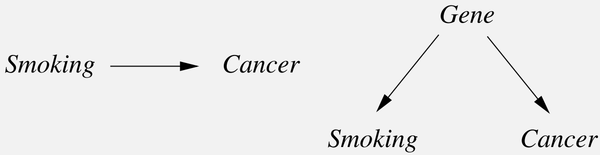
Causal Discovery: Possible?

Causal Discovery = **Structure** *Learning*

Is causal discovery even possible?

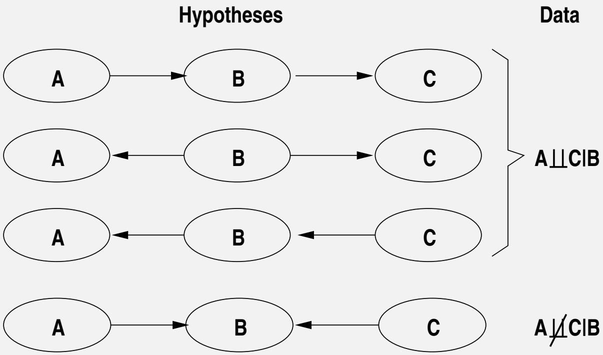
> *RA Fisher: Learning a probabilistic dependency will not advance our causal understanding even one step.*

Fisher's most notorious example:



Causal Discovery: Possible

There are four types of "threesomes":



In Popperian language, we can "falsify" the one causal pattern or the other.

> *When scaled up on sparse networks, this kind of discovery works reasonably well!*

Causal Discovery Programs

PC

IC was first made practical in the PC algorithm, in TETRAD II (now IV) from Spirtes, Glymour and Scheines (1993).

- Replacing the Oracle with statistical significance tests (for correlation with linear data, χ^2 for discrete data).
- Heuristics used to speed up search.
- Ideal result: discovered pattern.
- PC is being implemented by numerous BN tools, including Weka and Hugin
- A better performing constraint-learner is GES (Meek, 1996)

Metric Causal Discovery

A very different approach is *metric* learning of causality:

- Develop a score function which evaluates any Bayesian network *as a whole* relative to the evidence.
- Originally this was done in a brute force Bayesian computation of

  $$P(dag|data)$$

  by Cooper & Herskovits (1991)
- CD then means: search the space of dags looking for that dag which maximizes the score.

Metric Discovery Programs

K2 (Cooper & Herskovits)
Greedy search. Mediocre performance.

MDL (Lam & Bacchus, 1993; Friedman, 1997)
An information-theoretic scoring function with various kinds of search, such as beam search. Friedman allows for hybrid local structure.

BDe/BGe (Heckerman & Geiger, 1995)
A Bayesian score; edit-distance priors supported; returns a pattern. Good performance.

CaMML (Korb & Nicholson, 2004; ch 8)
A Bayesian information-theoretic scoring function with MCMC (sampling search); returns dags and patterns. Performance similar to BDe/BGe. Supports priors and hybrid local structure.

Recent Extensions to CaMML

Two significant enhancements have been added in the last few years.

Expert priors (O'Donnell et al., 2006b)

- Being Bayesian, it is relatively easy to incorporate non-default priors into CaMML. We've done this in various ways, specifying strengths for:
  - A prior dag, computing a prior distribution via edit distance
  - Arc densities
  - Topological orders, total or partial

Hybrid model learning (O'Donnell et al., 2006a)

- Allowing varying representations of local structure (CPTs, d-trees, logit model) throughout the network

Selected Monash Applications

- Bureau of Meterology: Fog forecasting: Nicholson, Korb, Boneh PhD project, 2004-2009
- Intelligent tutoring for decimal understanding: Nicholson, Boneh, University of Melbourne (1999-2003)
- NAG (Nice Argument Generator): Zukerman, Korb, 1997-2000
- Bayesian Poker: Korb, Nicholson, Honours projects 1993-2009
- SARBayes: Twardy, Korb, Victorian Search and Rescue, 2001 Honours project
- User modelling (plan recognition in a MUD, web page pre-fetching): Zukerman, Albrecht, Nicholson (1997-2001)

Selected Monash Applications (cont.)

- Causal discovery via MML (CaMML): Wallace, Korb, Neil, O'Donnell, Dai, Nyberg et al., 1996-2009
- Ecological risk assessment:
  - Nicholson, Korb, Pollino (Monash Centre for Water Studies), 2003-2005 Native Fish abundance in Goulburn Water
  - Predicting recreational water quality: Twardy, Nicholson, NSW EPA, 2003 Honours project
  - Tropical seagrass in great barrier reef: Nicholson, Thomas (Monash Centre for Water Studies), 2004-2006

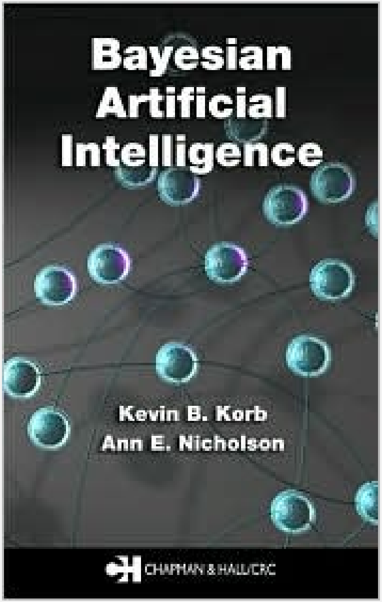
Selected Monash Applications (cont.)

- Predicting cardiovascular risk from epidemiological data: Korb, Nicholson, Twardy, John McNeil (Department of Epidemiology and Preventive Medicine, Monash University), 2004-2006
- Change impact analysis in software architecture design: Nicholson, Tang, Jin, Han (Swinburne)
- Victorian DSE (Bayesian Intelligence)
  - BN Tool for modeling and reporting on threats to fresh water ecologies around the state.

Summary

- Graphical models
  - simplify computation
  - hide parameter details
  - enhance intelligibility for elicitation, validation, explanation
- BNs support combination of knowledge and data
- BNs support flexible GUI/DBMS integration
- Bayesian Intelligence can deliver these benefits
  - `www.bayesian-intelligence.com`

For More Information



Next Computational Intelligence Events

- Computational Intelligence Summer School
  Sponsored by IEEE
- AI'09
  `www.infotech.monash.edu.au/about/news/conferences/ai09/`
- Australian Conf on Artificial Life
  `www.infotech.monash.edu.au/about/news/conferences/acal09/`

References I

G.F. Cooper and E. Herskovits (1991) "A Bayesian Method for Constructing Bayesian Belief Networks from Databases," in D'Ambrosio, Smets and Bonissone (eds.) *UAI 1991,* 86-94.

D. Heckerman and D. Geiger (1995) "Learning Bayesian networks: A unification for discrete and Gaussian domains," in Besnard and Hanks (eds.) *UAI 1995,* 274-284.

K.B. Korb and A.E. Nicholson (2004) *Bayesian Artificial Intelligence.* CRC/Chapman Hall.

Meek, C. (1996). *Graphical models: Selectiong causal and statistical models.* PhD disseration, Philosophy, Carnegie Mellon University.

R. Neapolitan (1990) *Probabilistic Reasoning in Expert Systems*. Wiley.

J. Pearl (1988) *Probabilistic Reasoning in Intelligent Systems*, Morgan Kaufmann.

F.P. Ramsey (1931). *The Foundations of Mathematics,* edited by R.B. Braithwaite. Routledge.

References II

D. Spiegelhalter & S. Lauritzen (1990) "Sequential Updating of Conditional Probabilities on Directed Graphical Structures," *Networks, 20,* 579-605.

P. Spirtes, C. Glymour and R. Scheines (1993) *Causation, Prediction and Search: Lecture Notes in Statistics 81.* Springer Verlag.

J. H. Wigmore (1913). The problem of proof. *Illinois Law Journal 8,* 77-103.

S. Wright (1921). Correlation and causation. *Journal of Agricultural Research, 20*, 557-585.