

# A Tutorial on Learning with Bayesian Networks

---

**David Heckerman**

Presented by: Krishna V Chengavalli  
April 21 2003

## Outline

---

- Introduction
  - Different Approaches
  - Bayesian Networks
  - Learning Probabilities and Structure
  - Bayesian Networks and Supervised and Unsupervised Learning
  - Conclusion
- 

## Probability

---

- Classical
    - The probability that a coin lands head
    - (also called True or Physical Probability)
  - Bayesian
    - Person's belief in the event
    - (also called Personal Probability)
- 

## Example

---

- Telephone Example
    - Consider listening to a friend talking over phone to some person x.
    - We try to guess who is he talking to.
    - Based on what they talk, language they use and tone we assign a certain probability to some persons. (Background Info)
    - We update probability till we assign a maximum likely person.
- 

## Probability Assessment

---

- The process of measuring one's degree of belief
    - A wheel with shaded region analogy
- 

## Illustrating Approaches

---

- Classical Approach
    - A Common Thumbtack Problem.
    - Assert some physical probability
    - Estimate the probability of head/tail from N observations.
    - Estimate for N+1th trial.
-

## Bayesian Approach

- Bayesian Approach to the same problem
  - Assert the physical probability (unknown)
  - Encode the uncertainty about their physical probability using Bayesian Probability
  - Use rules of probability to compute the required probability

## Probability Formulas

Bayes theorem - The posterior probability for  $\theta$  given  $D$  and a background knowledge  $\xi$  :

$$p(\theta/D, \xi) = \frac{p(\theta / \xi) p(D / \theta, \xi)}{P(D / \xi)}$$

Note :  $\theta$  is an uncertain variable whose value corresponds to the possible true values of the physical probability

## Likelihood Function

- How good is a particular value of  $\theta$  ?

How likely it is capable of generating the observed data

$$L(\theta : D) = P(D / \theta)$$

The likelihood of the sequence H, T, H, T, T

$$L(\theta : D) = \theta \cdot (1 - \theta) \cdot \theta \cdot (1 - \theta) \cdot (1 - \theta)$$

- Finally we average over  $\theta$  to determine the probability of  $N+1$ th toss being head  
This is also called as Expectation of  $\Theta$  wrt  $P(\Theta / D, \xi)$

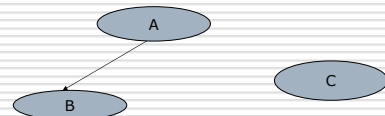
$$\begin{aligned} p(X_{N+1} = \text{heads} | D, \xi) &= \int p(X_{N+1} = \text{heads} | \theta, \xi) p(\theta | D, \xi) d\theta \\ &= \int \theta p(\theta | D, \xi) d\theta \equiv E_{p(\theta | D, \xi)}(\theta) \end{aligned}$$

## Bayesian Network

- A directed acyclic graph
- Encodes set of conditional assertions about variables
- Encodes set of local probability for each variable
- Together they define the joint probability distribution for the structure.

## More about Bayesian Networks

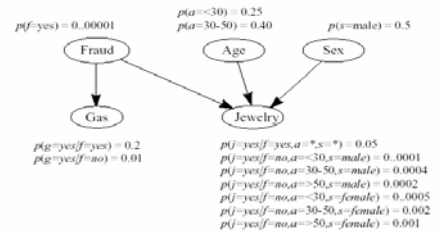
- Presence of Arcs between variables denote conditional dependence



## Why Bayesian Networks?

- Handle Incomplete Data
- Learn Causal Relationships
- Avoids Overfitting
- Combination of knowledge with data

## Bayesian Network



## Conditional (In)Dependence

$$p(a|f) = p(a)$$

$$p(s|f, a) = p(s)$$

$$p(g|f, a, s) = p(g|f)$$

$$p(j|f, a, s, g) = p(j|f, a, s)$$

## Constructing Bayesian Network

### □ Use of Causal Relationships

- Given a set of variables draw arcs between cause variables and their immediate effects.

Use of causal semantics is largely responsible for the success of Bayesian Networks.

## Constructing a Bayesian Network cont....

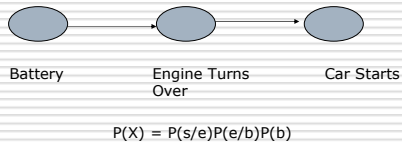
- Identify as many observations as possible
- Determine the subset that is worthwhile to model
- Organize those observations into mutually exclusive and collectively exhaustive states

## Steps in Construction

- model the system of interest with random variables
- arrange nodes that represent the random variables by influence
- number the nodes in order of influence and depth (roots have lowest numbers)
- obtain prior probabilities at root nodes
- construct CPT's at non-root nodes
- compute the joint probabilities
- program Bayes' rules and other equations to be used for the query nodes

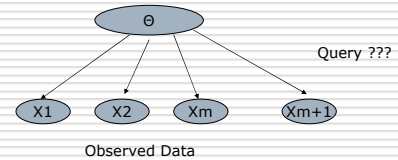
## Chain Rule of Probability

$$p(\mathbf{x}) = \prod_{i=1}^n p(x_i | x_1, \dots, x_{i-1})$$



## Bayesian Inference

- Determining various probabilities of interest from the model



## Bayesian Inference cont..

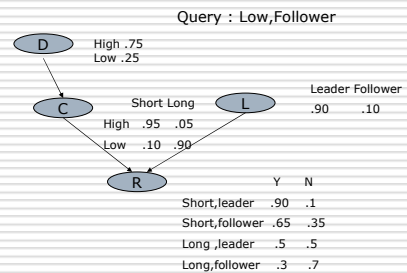
- The probability that the battery is up given the car starts

$$P(b/s) = P(b,s) / P(s)$$

$P(b,s)$  computed by summation over  $t$  for given  $b$  and  $s$

$P(s)$  computed by summation over  $b$  and  $t$  for given  $s$

## Example



## Handling of Incomplete Data

- Gibbs Sampling

- Simulates missing data according to available data
- Time consuming

## Incomplete Data cont...

- Expectation Maximization

- Basic idea is to augment observed data with missing data
  - Expectation Step – taken wrt missing conditioned on observed values
  - Maximization Step
  - Repeat 1 and 2 until estimation converges
- The idea is to converge the parameter being estimated to its maximum likelihood. Faster than Gibbs'.

## Learning Probabilities

- Using Data to update the probabilities in the network
- In thumbtack problem no probability is learned but the posterior distribution of variable that represents physical probability of heads is updated
- We assume no missing data and parameters are independent

## Known Structure

- If structure is known the problem is to learn the parameters for the graph from the database
  - Example - flipping the coin, we have a database of outcomes, learn distribution of the variable
  - More complex situations, more variables, complex formulas

## Unknown Structure

- First select the model.
- Done by generating a model search space.
- Evaluate the models given the values in the dataset

## Model Selection

- Criteria - degree to which the structure fits the prior knowledge and data
  - Use of Relative Posterior Probability
  - Local Criteria

## Relative Posterior Probability

- Relative Posterior Probability

$$\log p(D, S^h) = \log p(S^h) + \log p(D|S^h).$$

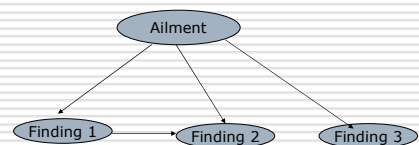
log prior      log marginal  
likelihood

This can be interpreted as

$$\log p(D|S^h) = \sum_{i=1}^N \log p(x_i|x_1, \dots, x_{i-1}, S^h)$$

## Local Criteria

- Medical Example



## Priors

---

- Structure Priors
- Parameter Priors

We use them to compute the relative posterior probability

---

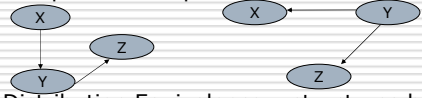
## Priors on Network Parameters

---

### □ Basic Idea

- Independence Equivalence
- Distribution Equivalence

### Independence Equivalence



Distribution Equivalence – structures have same joint distribution

---

## Search Methods

---

- Greedy Search
  - Greedy Search with restarts
  - Best First Search
  - Monte Carlo Methods
- 

## Search Method

---

- Variable Specific Criteria
  - Make successive Arc Changes
  - Make changes according to valid changes  $e$  in  $E$
  - Compute  $C(X_i, Pa_i, D)$  to determine  $\Delta(e)$
- 

## Greedy Search

---

- $\Delta(e)$  is computed for all edges and change  $e$  is done where it is maximum
  - If criterion is separable we need to compute only for that change to determine  $\Delta(e)$
  - Problem – local maximum
  - Way out – Random restarts, Simulated annealing
- 

## Supervised Learning Vs Bayesian

---

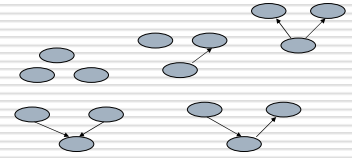
- When input variables cause outcome, data is complete – identical approaches
  - If cause-effect is not found or incomplete data – differences arise
  - Small sample sizes handled well by Bayesian Nets.
-

## Unsupervised Learning

- Assumption of hidden variables
- Generate models with and without hidden variables.
- Compare score.
- Eg AutoClass – performs data clustering.

## Example

Obj	A	B	C
1	T	F	T
2	T	T	T
3	F	T	T
4	F	T	T



$$P(\text{example} | S, A_i) = \prod_i P(\text{object}_i | S_i, A_i) \quad P(\text{object}_i | S_i, A_i) = \dots$$

## Applications

Implementations in real life :

- Used in the Microsoft products(Microsoft Office)
- Medical applications and Biostatistics (BUGS)
- In NASA Autoclass project for data analysis
- Collaborative filtering (Microsoft – MSBN)
- Fraud Detection (ATT)
- Speech recognition (UC , Berkeley )

## Limitations

- Require initial knowledge of many probabilities...quality and extent of prior knowledge play an important role
- Significant computational cost(NP hard task)
- Unanticipated probability of an event is not taken care of.

## References

- A tutorial to Learning Bayesian Networks
- <http://www.stat.duke.edu/~chris/research/EM.pdf>
- <http://www.cs.huji.ac.il/~nir/Nips01-Tutorial/>

## A Tutorial on Learning With Bayesian Networks

- David Heckerman

Mohammad Saif  
Lecturer: Dr. R. Maclin  
Feature Presentation: Krishna C.  
Date: April 21, 2003

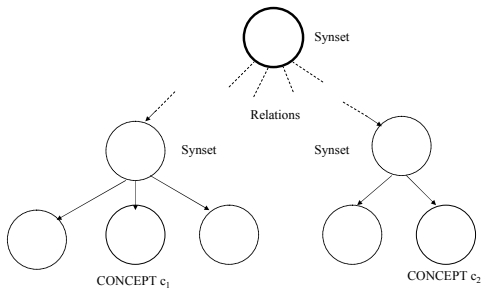
1

## Integration of Analytical and Empirical Components

- Bayesian Networks suitable to combine prior knowledge and evidence from data
- Bayesian network derived from WordNet
- Empirical component composed of suitable probabilistic model
  - Tagged training data
- Janyce Wiebe, Tom O'Hara and Rebecca Bruce

2

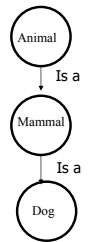
## WordNet



3

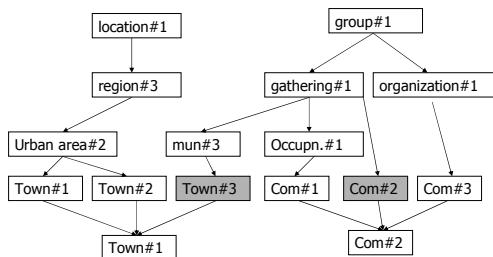
## Creation of the Bayesian N/W

- Let "community" and "town" appear in the same sentence
- Task
  - Assign correct senses to them
- A new Bayesian n/w created for each sentence
- Includes all synsets
  - of the target word
  - Reachable from them through the hypernymy (is a) hierarchy



4

## Bayesian Network using WordNet



5

## Probabilities

- Each child node assigned probability
  - Inversely proportional to number of children hypernym has
  - $P(\text{town} / \text{urban area}) = \frac{1}{2} = .5$
- Empirical classifier developed for each ambiguous word
  - Determines likelihood of each sense based on context
  - Nodes added to each ambiguous word
    - Represent support for each sense derived from the corpus

6



# A Tutorial on Learning with Bayesian Networks

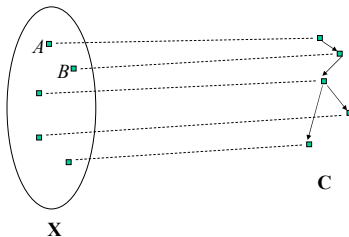
paper by  
David Heckerman  
presented by  
Krishna V Chengavalli  
commentary by  
Siddharth Patwardhan

## Learning Causal Relationships

Causal Markov Condition

*Directed Acyclic Graph C is a causal graph for variables X if the nodes in C have a one-to-one correspondence with the variables in X, and there is an arc from node A to node B in C if and only if A is a direct cause of B.*

## Causal Markov Condition



## Learning Causal Relations

- Learn Bayesian Network Structure from Data.
- Infer Causal Relationships from network structure.

## An Illustration

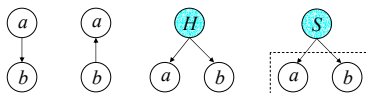
- Given data containing 2 variables  $a$  and  $b$ .
- $a$  is independent of  $b$ , if

$$p(b|a) = p(b|\bar{a}) = p(b)$$

- If we observe

$$p(b|a) \neq p(b|\bar{a})$$

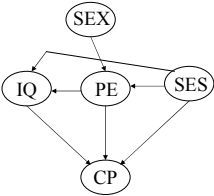
we can conclude (by Causal Markov Condition)



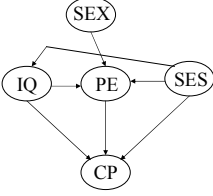
## Learning Network Structure

- To learn a structure on a set of variables  $X$ .
- Consider a random variable  $z$ , whose states correspond to the various network configurations.
- Using Bayesian methods, determine the posterior probability distribution of  $z$ .
- Structures with highest probability are the learnt structures.

# College Plans!! An Example



$$P(\text{Struct}_1 | \text{Data}) = 1.0$$



$$P(\text{Struct}_2 | \text{Data}) = 1.2 \times 10^{-10}$$