Learning With Bayesian Networks

Introduction

A bayesian network is a graphical model for probabilistic relationships among a set of variables

What do Bayesian Networks and Bayesian Methods have to offer ?

- **Handling of Incomplete Data Sets**
- **Learning about Causal Networks**
- **Facilitating the combination of domain knowledge and data**
- **Efficient and principled approach for avoiding the over fitting of data**

The Bayesian Approach to Probability and Statistics

Bayesian Probability : the degree of belief in that event

Classical Probability : true or physical probability of an event

Some Criticisms of Bayesian Probability

- **Why degrees of belief satisfy the rules of probability**
- **On what scale should probabilities be measured?**
- **What probabilites are to be assigned to beliefs that are not in extremes?**

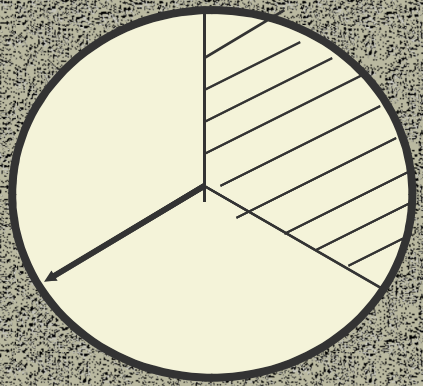
Some Answers ……

- Researchers have suggested different sets of properties that are satisfied by the degrees of belief

Scaling Problem

The probability wheel : a tool for assessing probabilities

What is the probability that the fortune wheel stops in the shaded region?



Probability assessment

An evident problem : SENSITIVITY

How can we say that the probability of an event is 0.601 and not .599 ?

Another problem : ACCURACY

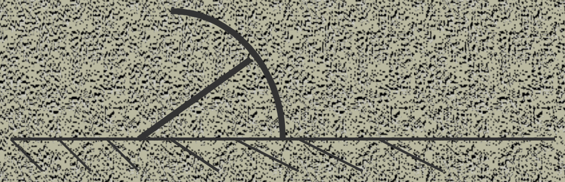
Methods for improving accuracy are available in decision analysis techniques

Learning with Data

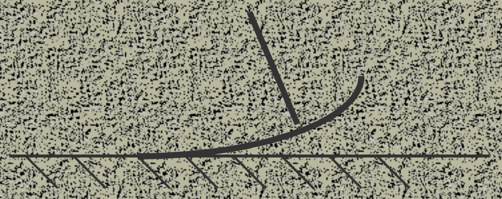
Thumbtack problem

When tossed it can rest on either heads or tails

Heads



Tails



Problem ………

From N observations we want to determine the probability of heads on the N+1 th toss.

Two Approaches

Classical Approach :

- assert some physical probability of heads (unknown)
- Estimate this physical probability from N observations
- Use this estimate as probability for the heads on the N+1 th toss.

The other approach

Bayesian Approach

- Assert some physical probability
- Encode the uncertainty about these physical probability using the Bayesian probabilities
- Use the rules of probability to compute the required probability

Bayes Theorem

- Bayes' theorem is stated mathematically as the following equation:
- $$P(A \mid B) = \frac{P(B \mid A)\,P(A)}{P(B)}$$

Naive Bayes Classifier

A joint model of observations $\mathbf{x}$ and the corresponding class label c using a Bayesian network of the form

$$p(\mathbf{x}, c) = p(c) \prod_{i=1}^{D} p(x_i|c)$$

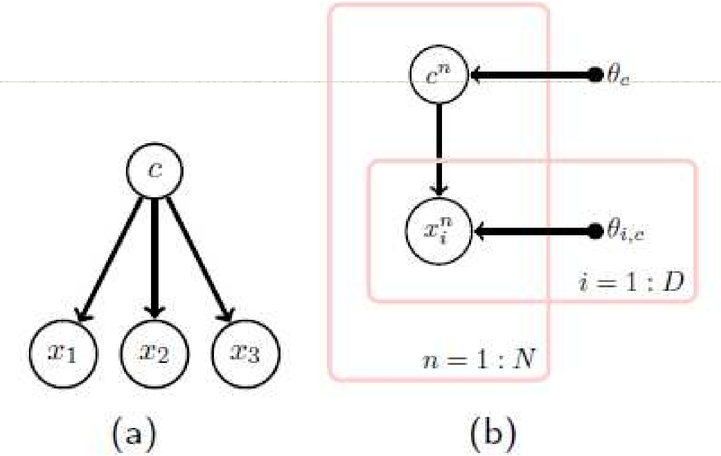


Figure: Naive Bayes classifier. **(a)**: The central assumption is that given the class c, the attributes x_i are independent. **(b)**: Assuming the data is i.i.d., Maximum Likelihood learns the optimal parameters of the distribution $p(c)$ and the class-dependent attribute distributions $p(x_i|c)$.

Coupled with a suitable choice for each conditional distribution $p(x_i|c)$, we can then use Bayes' rule to form a classifier for a novel attribute vector $\mathbf{x}^*$:

$$p(c|\mathbf{x}^*) = \frac{p(\mathbf{x}^*|c)p(c)}{p(\mathbf{x}^*)} = \frac{p(\mathbf{x}^*|c)p(c)}{\sum_c p(\mathbf{x}^*|c)p(c)}$$

Naive Bayes example

Consider the following vector of attributes:

(likes shortbread, likes lager, drinks whiskey, eats porridge, watched England play football)

Together with each vector $\mathbf{x}$, there is a label nat describing the nationality of the person, $\mathrm{dom}(nat) = \{\text{scottish}, \text{english}\}$.
We can use Bayes' rule to calculate the probability that $\mathbf{x}$ is Scottish or English:

$$p(\text{scottish}|\mathbf{x}) = \frac{p(\mathbf{x}|\text{scottish})p(\text{scottish})}{p(\mathbf{x})}$$
$$= \frac{p(\mathbf{x}|\text{scottish})p(\text{scottish})}{p(\mathbf{x}|\text{scottish})p(\text{scottish}) + p(\mathbf{x}|\text{english})p(\text{english})}$$

For $p(\mathbf{x}|nat)$ under the Naive Bayes assumption:

$$p(\mathbf{x}|nat) = p(x_1|nat)p(x_2|nat)p(x_3|nat)p(x_4|nat)p(x_5|nat)$$

Y=1

0	1	1	1	0	0
0	0	1	1	1	0
1	1	0	0	0	0
1	1	0	0	0	1
1	0	1	0	1	0

(a) English

1	1	1	1	1	1	1
0	1	1	1	1	0	0
0	0	1	0	0	1	1
1	0	1	1	1	1	0
1	1	0	0	1	0	0

(b) Scottish

Y=0

For $\mathbf{x} = (1, 0, 1, 1, 0)^{\mathsf{T}}$, we get

0	1	1	1	0	0
0	0	1	1	1	0
1	1	0	0	0	0
1	1	0	0	0	1
1	0	1	0	1	0

(a) English

1	1	1	1	1	1	1
0	1	1	1	1	0	0
0	0	1	0	0	1	1
1	0	1	1	1	1	0
1	1	0	0	1	0	0

(b) Scottish

Using Maximum Likelihood we have: $p(\text{scottish}) = 7/13$ and $p(\text{english}) = 6/13$.

$$
\begin{aligned}
p(x_1 = 1|\text{english}) &= 1/2 & \qquad p(x_1 = 1|\text{scottish}) &= 1 \\
p(x_2 = 1|\text{english}) &= 1/2 & \qquad p(x_2 = 1|\text{scottish}) &= 4/7 \\
p(x_3 = 1|\text{english}) &= 1/3 & \qquad p(x_3 = 1|\text{scottish}) &= 3/7 \\
p(x_4 = 1|\text{english}) &= 1/2 & \qquad p(x_4 = 1|\text{scottish}) &= 5/7 \\
p(x_5 = 1|\text{english}) &= 1/2 & \qquad p(x_5 = 1|\text{scottish}) &= 3/7
\end{aligned}
$$

For $\mathbf{x} = (1, 0, 1, 1, 0)^\mathsf{T}$, we get

$$
p(\text{scottish}|\mathbf{x}) = \frac{1 \times \frac{3}{7} \times \frac{3}{7} \times \frac{5}{7} \times \frac{4}{7} \times \frac{7}{13}}{1 \times \frac{3}{7} \times \frac{3}{7} \times \frac{5}{7} \times \frac{4}{7} \times \frac{7}{13} + \frac{1}{2} \times \frac{1}{2} \times \frac{1}{3} \times \frac{1}{2} \times \frac{1}{2} \times \frac{6}{13}} = 0.8076
$$

Since this is greater than 0.5, we would classify this person as being Scottish.

Bayes Theorem

- Suppose that you are worried that you might have a rare disease. You decide to get tested, and suppose that the testing methods for this disease are correct 99 percent of the time (in other words, if you have the disease, it shows that you do with 99 percent probability, and if you don't have the disease, it shows that you do not with 99 percent probability). Suppose this disease is actually quite rare, occurring randomly in the general population in only one of every 10,000 people.
- If your test results come back positive, what are your chances that you actually have the disease?

- $P(D=1) = 1/10000 = 0.0001$
- $P(T=1 | D=1) = 99\ \% = 0.99$
- $P(T=0 \mid D=0) == 99\% = 0.99$

- $P(D=1 \mid T=1) = ????$

$$= \frac{P(T=1 \mid D=1)\, P(D=1)}{P(T=1)}$$

Test \ Disease	D=1 Yes	D=0 No
Test =1 Yes	0.99	0.01
Test =0 No	0.01	0.99

- $P(D=1 \mid T=1) = ????$

$$= \frac{P(T=1 \mid D=1)\, P(D=1)}{P(T=1)}$$

$$= \frac{P(T=1 \mid D=1)\, P(D=1)}{P(T=1|D=1)\, P(D=1) + P(T=1|D=0)\, P(D=0)}$$

$= 0.0098 = 0.98\ \% = < 1\%$

Definition

A Bayesian Network for a set of variables

X = { X1,…….Xn} contains

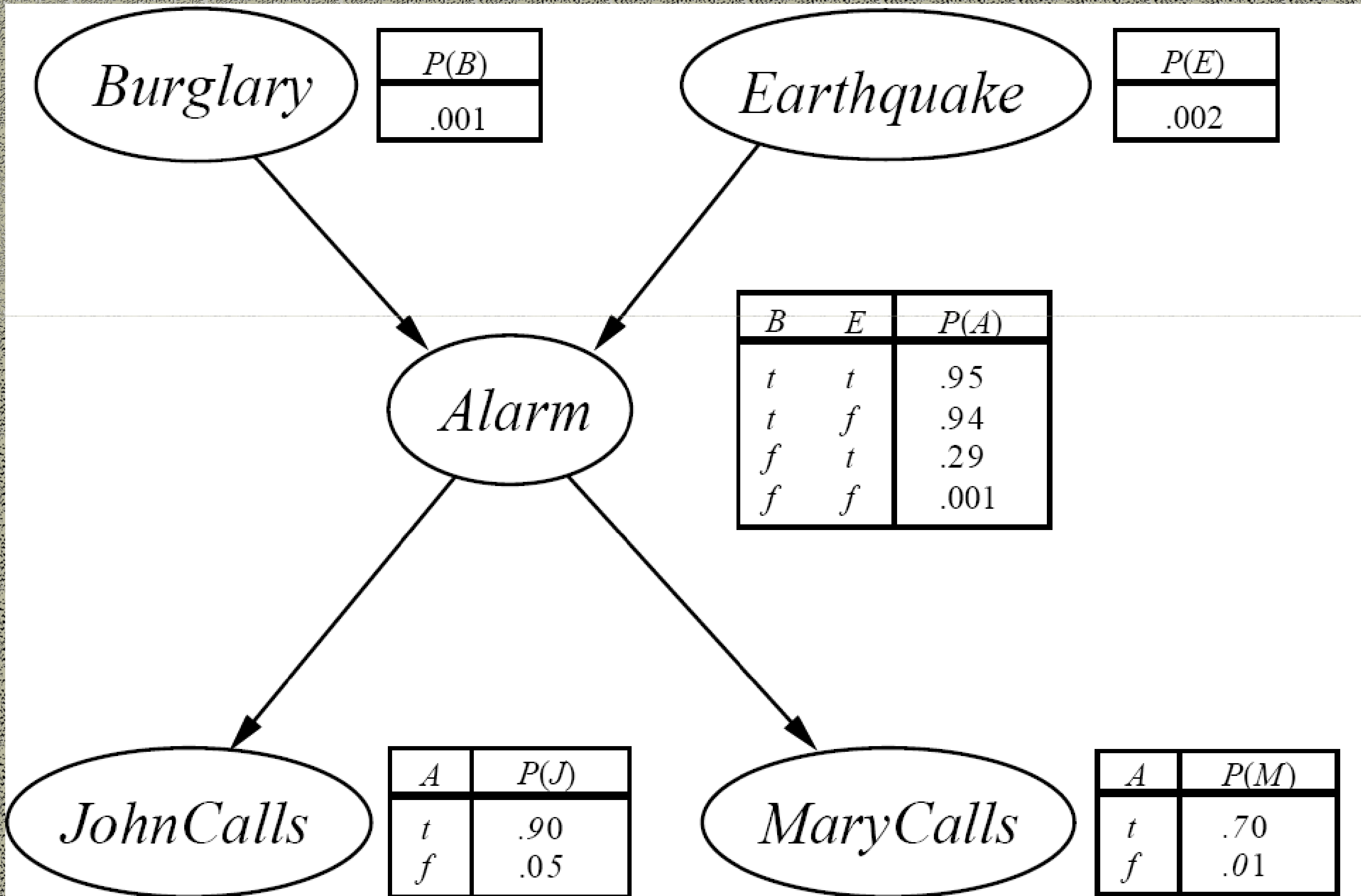
- network structure S encoding conditional independence assertions about X
- a set P of local probability distributions

The network structure S is a **directed acyclic graph**

And the nodes are in one to one correspondence with the variables X. Lack of an arc denotes a conditional independence.

Some conventions……….

- Variables depicted as nodes
- Arcs represent probabilistic dependence between variables
- Conditional probabilities encode the strength of dependencies


Burglary
P(B)
.001
Earthquake
P(E)
.002
Alarm
B E P(A)
t t .95
t f .94
f t .29
f f .001
JohnCalls
A P(J)
t .90
f .05
MaryCalls
A P(M)
t .70
f .01

Tasks

- Correctly identify the goals of modeling
- Identify many possible observations that may be relevant to a problem
- Determine what subset of those observations is worthwhile to model
- Organize the observations into variables having mutually exclusive and collectively exhaustive states.

Finally we are to build a Directed A cyclic Graph that encodes the assertions of conditional independence

A technique of constructing a Bayesian Network

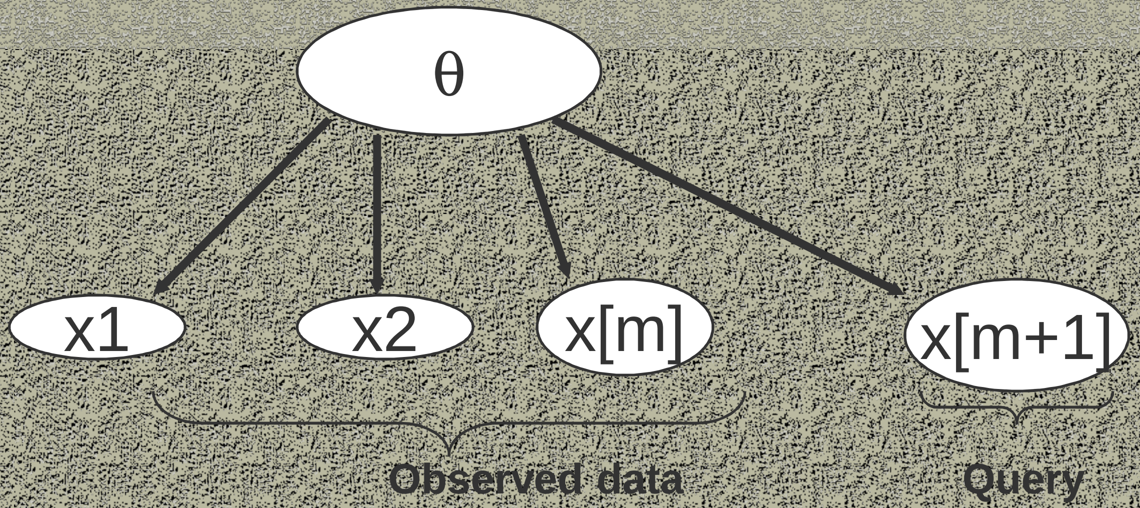
The approach is based on the following observations :

- People can often readily assert causal relationships among the variables
- Casual relations typically correspond to assertions of conditional dependence

To construct a Bayesian Network we simply draw arcs for a given set of variables from the cause variables to their immediate effects.In the final step we determine the local probability distributions.

Bayesian inference

On construction of a Bayesian network we need to determine the various probabilities of interest from the model



Computation of a probability of interest given a model is probabilistic inference

Learning Probabilities in a Bayesian Network

Problem **: Using data to update the probabilities of a given network structure**

Thumbtack problem **: We do not learn the probability of the heads , we update the posterior distribution for the variable that represents the physical probability of the heads**

The problem restated **:Given a random sample D compute the posterior probability .**

Assumptions to compute the posterior probability

- There is no missing data in the random sample D.
- Parameters are independent .

But……

Data may be missing and then how do we proceed   ?????????

Obvious concerns….

Why was the data missing?

- Missing values
- Hidden variables

Is the absence of an observation dependent on the actual states of the variables?

We deal with the missing data that are independent of the state

Incomplete data (contd)

Observations reveal that for any interesting set of local likelihoods and priors the exact computation of the posterior distribution will be intractable.

We require approximation for incomplete data

The various methods of approximations for Incomplete Data

- Monte Carlo Sampling methods
- Gaussian Approximation
- MAP and ML Approximations and EM algorithm

Gibb's Sampling

The steps involved :

Start :

- Choose an initial state for each of the variables in X at random

Iterate :

- Unassign the current state of X1.
- Compute the probability of this state given that of n-1 variables.
- Repeat this procedure for all X creating a new sample of X

- After " burn in " phase the possible configuration of X will be sampled with probability p(x).

Problem in Monte Carlo method

Intractable when the sample size is large

Gaussian Approximation

Idea : Large amounts of data can be approximated to a multivariate Gaussian Distribution.

Criteria for Model Selection

Some criterion must be used to determine the degree to which a network structure fits the prior knowledge and data

Some such criteria include

- Relative posterior probability
- Local criteria

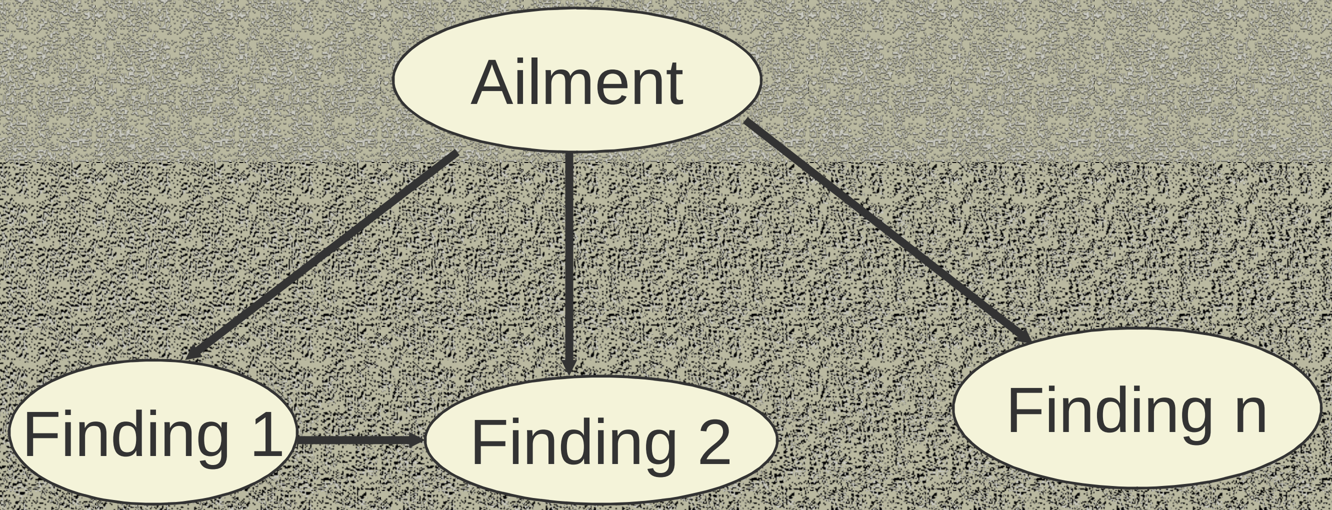
Relative posterior probability

A criteria for model selection is the logarithm of the relative posterior probability given as follows :

$$\text{Log } p(D/Sh) = \underbrace{\log p(Sh)}_{\text{log prior}} + \underbrace{\log p(D/Sh)}_{\text{log marginal likelihood}}$$

Local Criteria

An Example :



A Bayesian network structure for medical diagnosis

Priors

To compute the relative posterior probability

We assess the

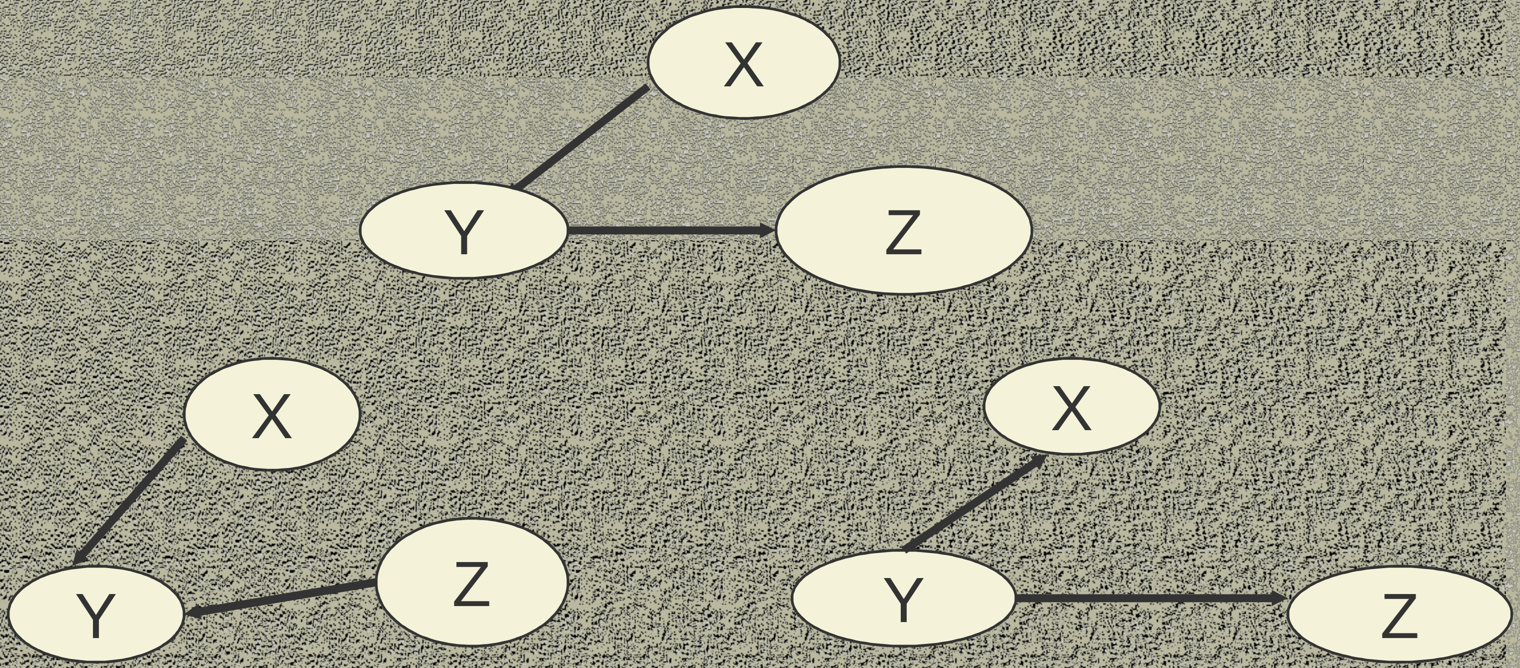
- Structure priors p(Sh)
- Parameter priors p(θs /Sh)

Priors on network parameters

Key concepts :

- Independence Equivalence
- Distribution Equivalence

Illustration of independent equivalence



Independence assertion : X and Z are conditionally independent given Y

Priors on structures

Various methods….

- Assumption that every hypothesis is equally likely (usually for convenience)
- Variables can be ordered and presence or absence of arcs are mutually independent
- Use of prior networks
- Imaginary data from domain experts

Benefits of learning structures

- Efficient learning --- more accurate models with less data
- Compare P(A) and P(B) versus P(A,B) former requires less data
- Discover structural properties of the domain
- Helps to order events that occur sequentially and in sensitivity analysis and inference
- Predict effect of the actions

Search Methods

Problem : We are to find the best network from the set of all networks in which each node has no more than k parents

Search techniques :

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

Bayesian Networks for Supervised and Unsupervised learning

Supervised learning : A natural representation in which to encode prior knowledge

Unsupervised learning :

- **Apply the learning technique to select a model with no hidden variables**
- **Look for sets of mutually dependent variables in the model**
- **Create a new model with a hidden variable**
- **Score new models possibly finding one better than the original.**

What is all this good for anyway????????

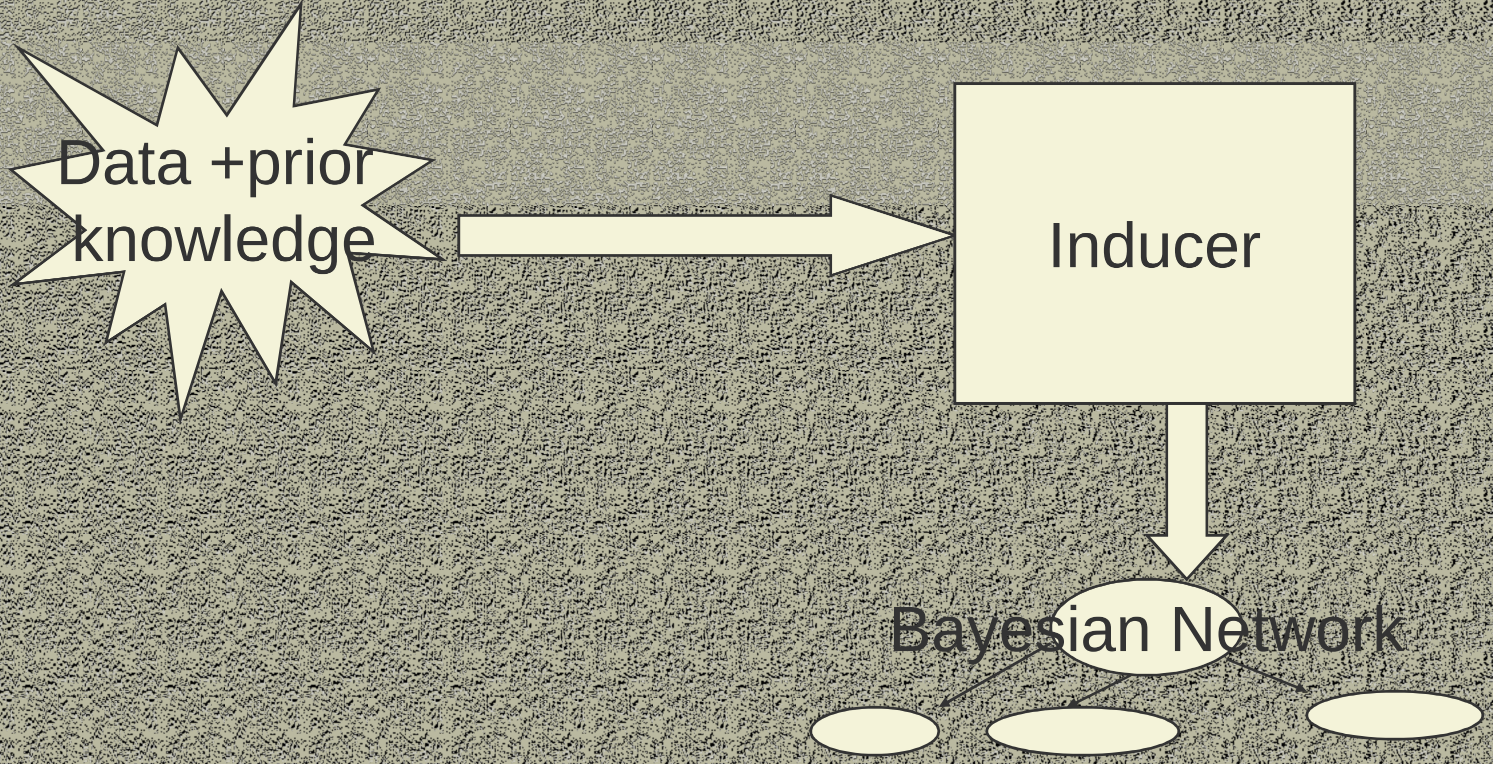
Implementations in real life :

- **It is used in the Microsoft products(Microsoft Office)**
- **Medical applications and Biostatistics (BUGS)**
- **In NASA Autoclass projectfor data analysis**
- **Collaborative filtering (Microsoft – MSBN)**
- **Fraud Detection (ATT)**
- **Speech recognition (UC , Berkeley)**

Limitations Of Bayesian Networks

- Typically 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.

Conclusion



Some Comments

- Cross fertilization with other techniques?

For e.g with decision trees, R trees and neural networks

- Improvements in search techniques using the classical search methods ?
- Application in some other areas as estimation of population death rate and birth rate, financial applications ?