3. Learning Bayesian Networks from Data
A. Introduction

Concha Bielza, Pedro Larrañaga

Computational Intelligence Group
Departamento de Inteligencia Artificial
Universidad Politécnica de Madrid

Master Universitario en Inteligencia Artificial

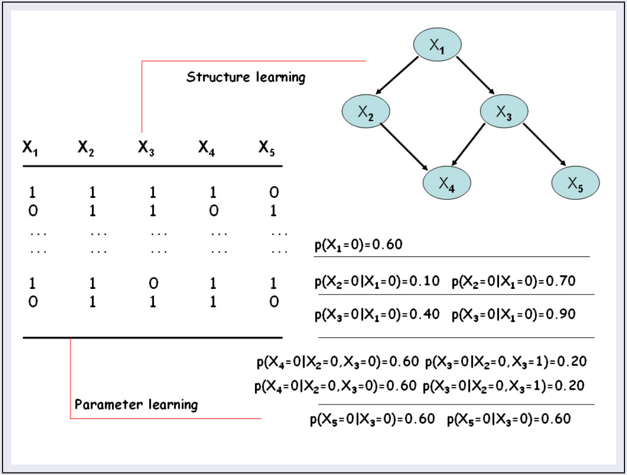
Outline

Outline

1. Introduction
2. Advantages on using Bayesian networks
3. Building Bayesian networks
4. Three different tasks
5. Software

Learning Bayesian networks from data

Structure and parametric learning of Bayesian networks



Outline

Reasons for using Bayesian networks

Benefits

- It is possible to handle explicitly the uncertain knowledge
  - Almost all the human being knowledge presents any type of uncertainty
- Founded in probability theory, provide a clair semantic and a sound theoretical foundation
  - Other paradigms in Artificial Intelligence are based on *ad hoc* theories
- Representation of the knowledge
  - Graphical
  - Intuitive
  - Reasoning close to the one make by human beings

Reasons for using Bayesian networks

Modularity

- The joint probability distribution (global model) is specified through marginal and conditional distributions (local models) taking into account conditional independence relationships between triplet of variables
- This modularity:
  - Provides an easy maintenance
  - Reduces the number of parameters necessary to specify the global model
    - Estimation or elicitation of the parameters is easier
    - Reduction of the storing needs
    - Efficient reasoning (inference)

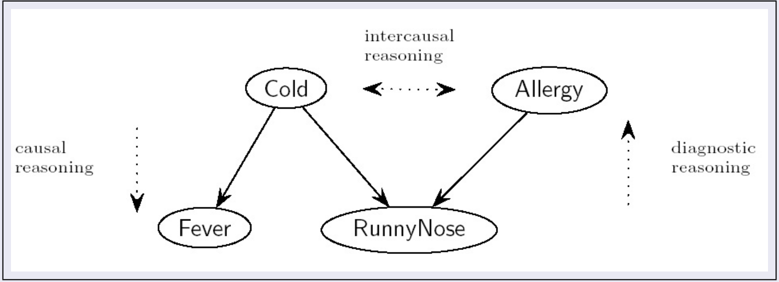
Reasoning

Inference types

- Inference (reasoning) can be done by exact and approximate algorithms for evidence propagation
- Types of inference:
  - Probability of an event given evidence
  - Most probable explanation (abductive inference) as the event that best explains the current evidence
  - Decision making using influence diagrams (a generalization of Bayesian networks that incorporates decision nodes and utilities)
- Bayesian networks provides tools for:
  - Diagnosis as the value of the target variable with highest a posteriori probability
  - Predictive reasoning by means of dynamic Bayesian networks

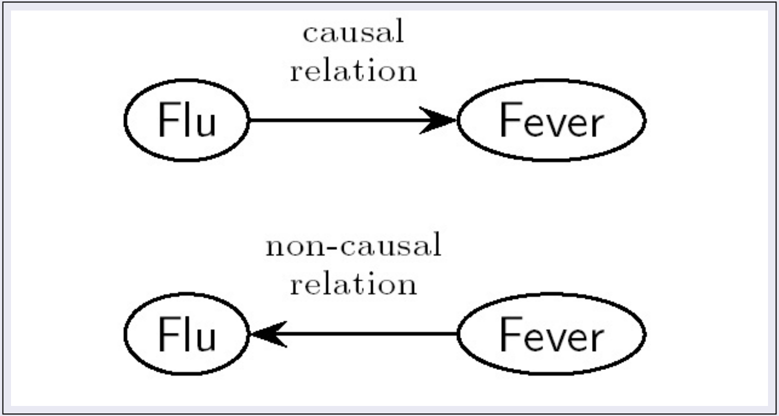
Reasoning

Causal. Intercausal. Diagnosis



Reasoning

Causality



Outline

Expert vs learning from data vs mixed

Three ways for building Bayesian networks

- To take advantage of this paradigm for uncertain knowledge representation, it is necessary to build a model (structure and parameters of the Bayesian network)
- Three main ways for building Bayesian networks:
  - By hand (with the help of an expert in the domain to be modelled)
  - Inducting it form a data base of cases (learning from data)
  - Mixed approach (mixed)

Using the knowledge of an expert

Structure of the Bayesian network

- Step 1: Choose the variables, one per node in the DAG
- Step 2: Determinate the range of values for each variable
- Step 3: Set up the conditional (in)dependencies between triplets of variables
- Step 4: Draw a DAG that express the conditional (in)dependencies obtained in Step 3

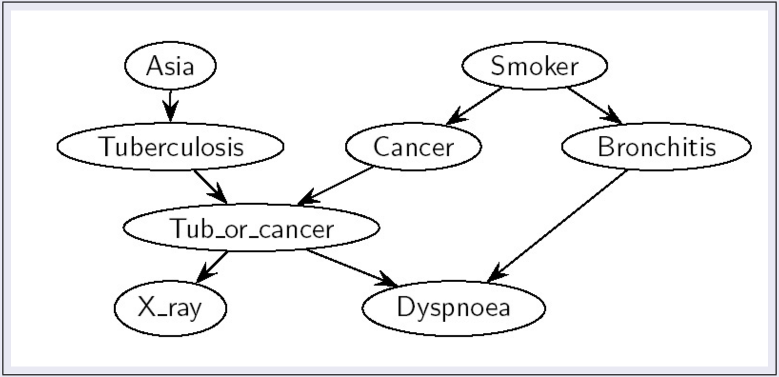
Using the knowledge of an expert

The Asia network (Lauritzen and Spigelhalter, 1988)

- Physician wants to diagnose her patients with respect to three diseases:
  - `Tuberculosis`
  - `Lung cancer`
  - `Bronchitis`
- Symptoms:
  - `Dyspnoea` (shortness-of-breath) may be due to `Tuberculosis`, `Lung cancer`, `Bronchitis`, non of them, or more than one of them
  - Recent visit to `Asia` increases the chances of `Tuberculosis`
  - `Smoking` risk factor for both `Lung cancer` and `Bronchitis`
  - `X-Ray` not discriminate `Lung cancer` and `Tuberculosis`, as neither does the presence or absence of `Dyspnoea`

Using the knowledge of an expert

The Asia network



Using the knowledge of an expert

Parameters of the Bayesian network (in a medical domain)

- From a data base of cases
  - Pros: Fast and cheap
  - Against: Size of the data base, bias
- From subjective estimation
  - Pros: Low cost
  - Against: Psychological bias
- From the medical literature
  - Pros: Reliability, low cost
  - Against: Few quantitative data relates with conditional probabilities
- From epidemiological studies
  - Pros: Direct
  - Against: Time and cost

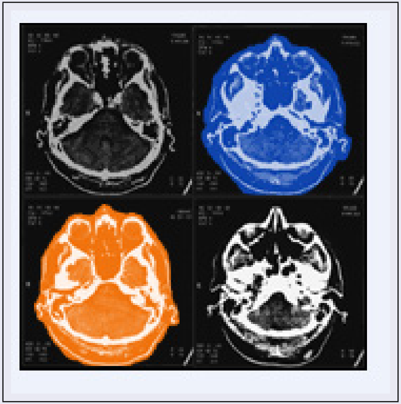
Learning from data

Machine learning methods: from data to knowledge

- Increasing availability of data bases in a great number of domains
- For an expert it is difficult to transform these data into knowledge (slow and unreliable process)
- Apart from this, in some domains it is difficult to contact with experts

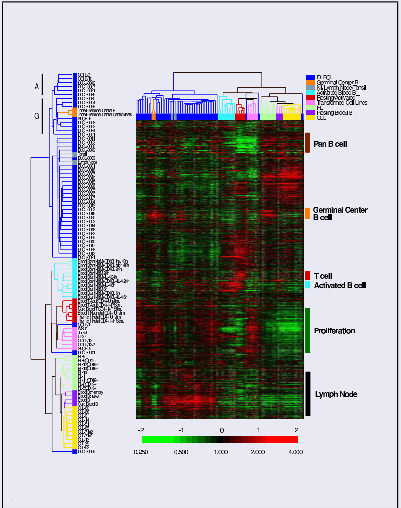
Learning from data

A difficult domain: Neurology



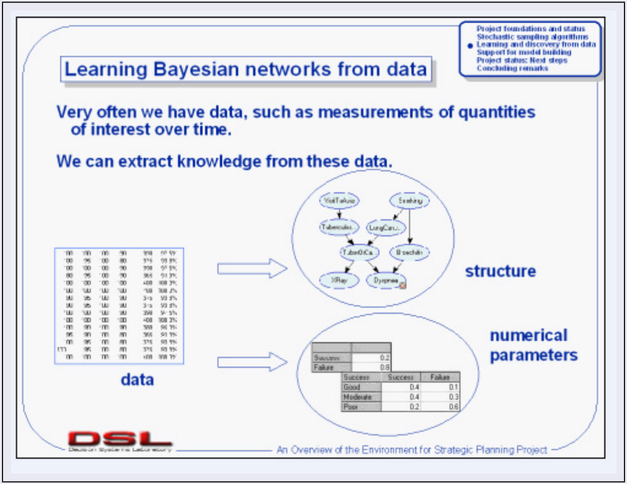
Learning from data

Another difficult domain: Microarray of DNA



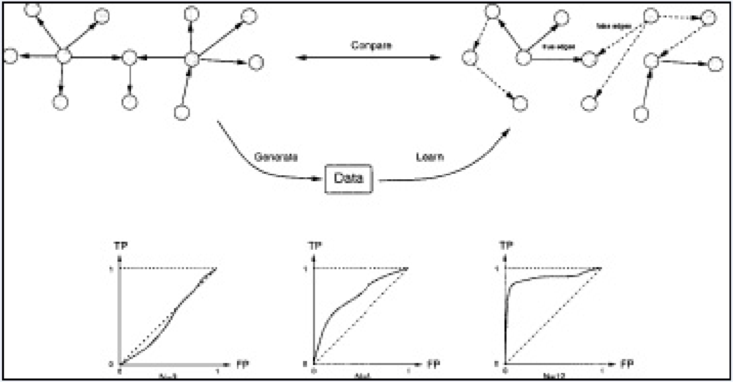
Learning from data

Structural and parametric learning



Learning from data

Generator model. Data. Learnt model



Learning from data

R. Neapolitan (2004). Prentice Hall

R. Neapolitan (2004)

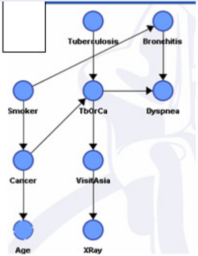
Learning Bayesian networks

Prentice Hall Series in Artificial Intelligence

Outline

Association discovery

Smoker	Cancer	Age	Tuber...	TbOrCa	VisitAsia	XRay	Bronchitis	Dyspnea
False	False	89.058...	False	False	False		False	False
False	False	54.474...	False	False	False	Normal	False	True
True	False		False	False	False	Normal	False	False
False	False	93.482...	False	False		Abnormal	False	False
	False	42.232...	False	False	False	Normal	False	False
True	False	45.994...	False	False	False	Normal	False	False
False	False	80.948...	False	False	False	Normal	False	False
False	False	30.676...		False	False	Normal	False	False
False	False	65.221...	False	False	False	Normal	True	True
False	False	68.297...	False	False	False	Normal	False	False
False	False	47.897...	False	False	False	Normal	False	False
False	False	49.409...	False	False	False	Normal	False	False
True	False	95.256...	False	False	False	Normal	True	False



Node selection : Dyspnea

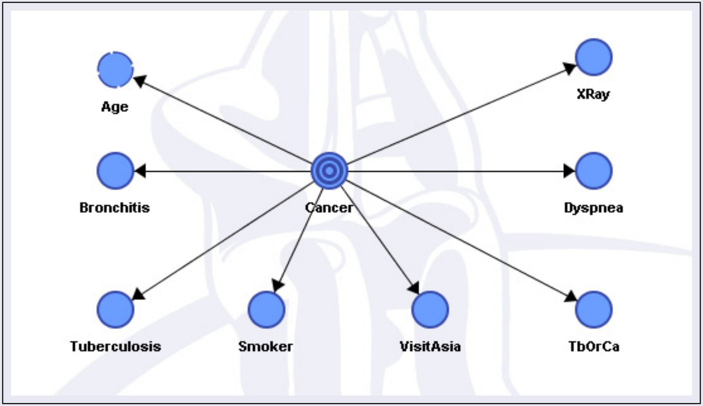
bution | Properties | Classes | Values | Comment

View mode

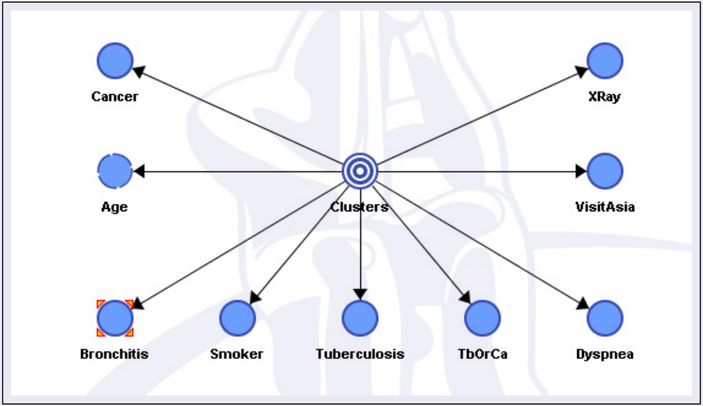
Determinist | Equation | Occurrences

TbOrCa	Bronchitis	False	True
False	False	89.980	10.020
	True	19.882	80.118
True	False	30.216	69.784
	True	10.648	89.352

Supervised classification



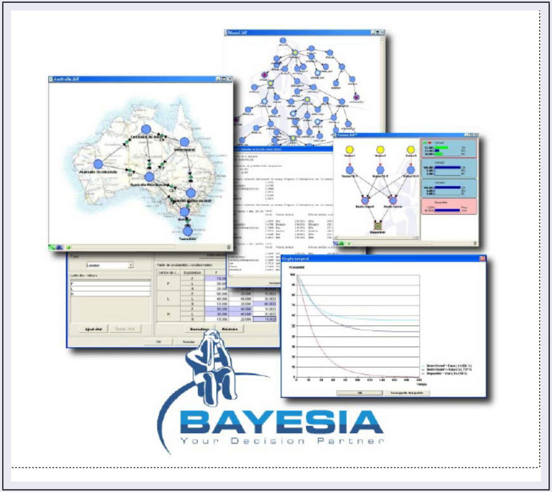
Clustering


Cancer
XRay
Age
Clusters
VisitAsia
Bronchitis
Smoker
Tuberculosis
TbOrCa
Dyspnea

Outline

BayesiaLab 4.1 DE

www.bayesia.com



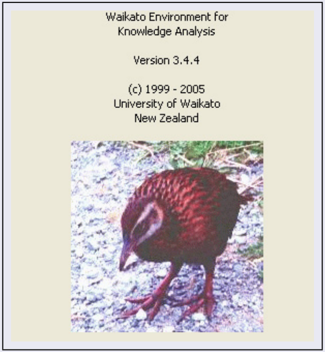
Hugin

www.hugin.com

HUGIN
EXPERT

WEKA

www.cs.waikato.ac.nz/ml/weka



3. Learning Bayesian Networks from Data
A. Introduction

Concha Bielza, Pedro Larrañaga

Computational Intelligence Group
Departamento de Inteligencia Artificial
Universidad Politécnica de Madrid

Master Universitario en Inteligencia Artificial