DM825

Introduction to Machine Learning

Lecture 11
Bayesian Networks

Marco Chiarandini

Department of Mathematics & Computer Science
University of Southern Denmark

Outline

Outline

1. Introduction

2. Theory

Introduction

- Bayesian Networks in Artificial Intelligence from Computer Science
- Probabilistic graphical models in Statistics
- pedigrees and their associated phenotype/genotype information in genetic linkage analysis
  reliability block diagrams in reliability analysis
  hidden Markov models in speech recognition

Bayesian networks provide:

- a systematic and localized method for structuring probabilistic information:
- compact representation of a probability distribution that is usually too large to be handled using tables and equations.
- provide a suite of algorithms that allow one to automatically derive many implications without need for specialization on the application

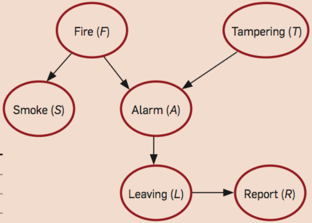
Bayesian Networks

Bayesian network

- has two components: a directed acyclic graph (called a structure),
  - nodes correspond to the variables
  - edges have a formal interpretation in terms of probabilistic independence (direct causal influences)
- a set of conditional probability tables (CPTs) that quantifies the relationship between that variable and its parents in the network via network parameters.

Fire	Smoke	$\theta_{s\|f}$
true	true	.90
false	true	.01

Fire	Tampering	Alarm	$\theta_{a\|f,t}$
true	true	true	.5000
true	false	true	.9900
false	true	true	.8500
false	false	true	.0001



Features

- every variable in the structure is assumed to become independent of its non-descendants once its parents are known (Markovian assumption)
- Guaranteed consistency and completeness:
  unique probability distribution over the 64 instantiations of the variables.
  enough information to attribute a probability to every event that can be expressed using the variables
- existence of efficient algorithms for computing such probabilities without having to explicitly generate the underlying probability distribution

Independence and Causality

The structure represents independence.

Constructed by causality: each variable becomes independent of its non-effects (non-descendant) once its direct causes (parents) are known (humans are good at identifying causes)

(eg: variable L is assumed to become independent of its non-descendants T, F, S once its parent A is known)

The distribution induced by a Bayesian network typically satisfies additional independencies, beyond the Markovian ones. All such independencies can be identified efficiently using a graphical test known as d-separation.

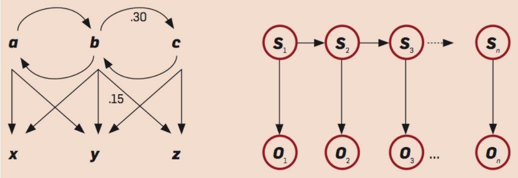
X and Y are guaranteed to be independent given variables Z if every path between X and Y is blocked by Z.

d-separation test (in linear time) can be used to directly derive results that have been proven for specialized probabilistic models used in a variety of fields.

Canonical Models (1)

Some canonical structures have been defined for specific tasks:

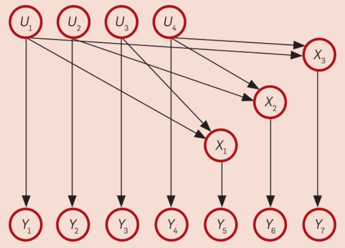
- hidden Markov models (HMMs) or dynamic Bayesian networks, which are used to model dynamic systems whose states are not observable, yet their outputs are.
  Used to make inference about these changing states, given the sequence of outputs they generate (temporal pattern recognition: speech, handwriting, and gesture recognition)



d-separation: once the state of the system at time t is known, its states and outputs at times $> t$ become independent of its states and outputs at times $< t$

Canonical Models (2)

- passing information over a noisy channel
  goal is to compute the most likely message sent over such a channel, given the channel output



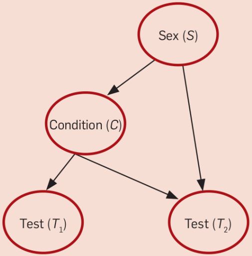
- image restoration and stereo vision
- analysis of documents
- machine translation

Constructing Bayesian Networks

BN can be constructed by:

1. subjective construction
2. synthesis from other specifications
3. learning from data

Subjective construction:



C = yes means that the condition C is present, and T_i = +ve means the result of test T_i is positive.

S	θ_s
male	.55

S	C	$\theta_{c\|s}$
male	yes	.05
female	yes	.01

C	T_1	$\theta_{t_1\|c}$
yes	+ve	.80
no	+ve	.20

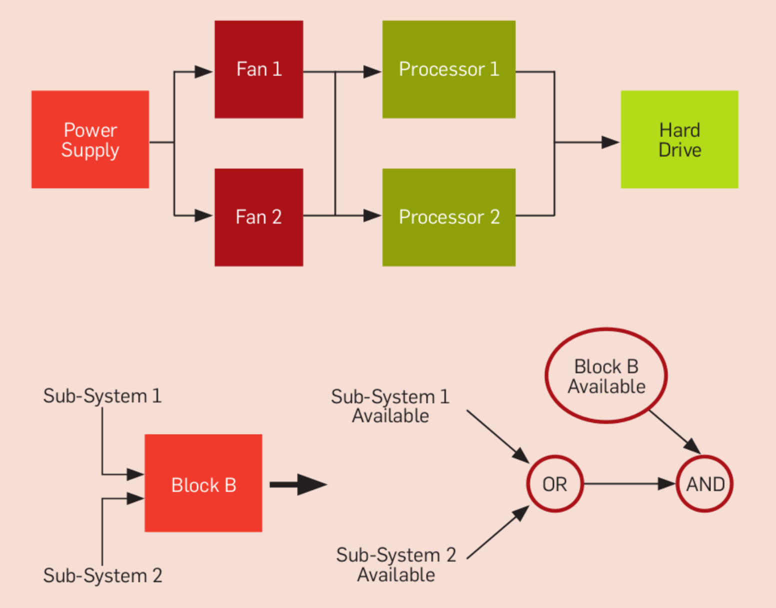
S	C	T_2	$\theta_{t_2\|s,c}$
male	yes	+ve	.80
male	no	+ve	.20
female	yes	+ve	.95
female	no	+ve	.05

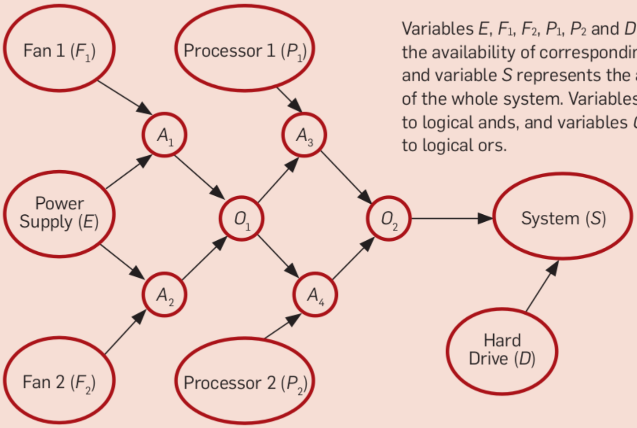
Sex S	Condition C	Test T_1	Test T_2
male	no	?	−ve
male	?	−ve	+ve
female	yes	+ve	?
⋮	⋮	⋮	⋮

Learning from data:

- learn the network parameters given its structure,
- learn both the structure and its parameters.

easier if data set is complete.


Power Supply
Fan 1
Fan 2
Processor 1
Processor 2
Hard Drive
Sub-System 1
Sub-System 2
Block B
Sub-System 1 Available
Sub-System 2 Available
Block B Available
OR
AND



Variables E, F_1, F_2, P_1, P_2 and D represent the availability of corresponding components, and variable S represents the availability of the whole system. Variables A_i correspond to logical ands, and variables O_i correspond to logical ors.

Learning and Using Parameters

Given the structure (for example a canonical naive Bayes) the method to learn parameters are:

1. maximum likelihood
2. Bayesian approach
   prior distribution + posterior distribution: take the mean or average over all possible parameter values according to the distribution

Parameters are used for tasks:

- diagnostic tasks, by inferring the most likely disease given a set of observed symptoms.
- prediction tasks, where we infer the most likely symptom given some diseases

Ingredients:

- subjective nature of the information used in constructing the networks;
- the reliance on Bayes conditioning when reasoning with Bayesian networks;
- ability to perform causal as well as evidential reasoning

Carrying out the Tasks

$\vec{e}$ assignment of values to some variables $\mathbf{E}$ (instantiation, evidence)

- Probability of Evidence $\Pr(\vec{e})$
  Example: probability that an individual will come out positive on both tests $\Pr(T1 = +\text{ve}, T2 = +\text{ve})$
  overall reliability of the system $\Pr(S = \text{avail})$
  related: node marginals query: probability $\Pr(x \mid e)$ for each X and for each of $x \in X$.

- Most Probable Explanation (MPE) $\arg\max_{\vec{q} \in \mathbf{Q}} \Pr(\vec{q} \mid \vec{e}), \mathbf{Q} = \overline{\mathbf{E}}$
  Example: find the most likely group, dissected by sex and condition, that will yield negative results for both tests ($\vec{e} = \{T_1 = \text{-ve}; T2 = \text{-ve}\}$ and $Q = \{S, C\}$)

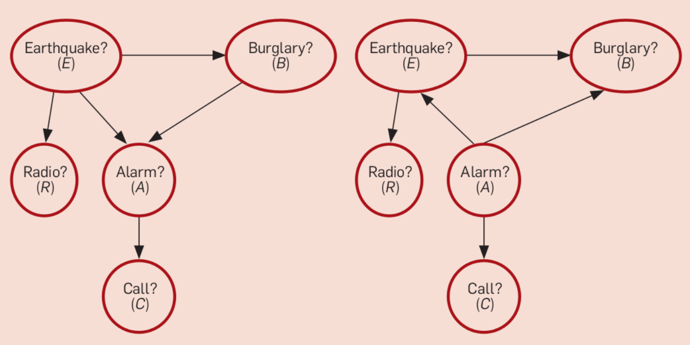
- Maximum a Posteriori Hypothesis (MAP) $\arg\max_{\vec{q} \in \mathbf{Q}} \Pr(\vec{q} \mid \vec{e}), \mathbf{Q} \subseteq \overline{\mathbf{E}}$
  Example: find most likely configuration of the two fans given that the system is unavailable ($\vec{e} = \{S = \text{unavail}\}, Q = \{F1, F2\}$).

Inference Algorithms

Treewidth: graph-theoretic parameter that measures the resemblance of a graph to a tree structure. (trees have $tr \leq 1$)

Elimination and conditioning inference algorithms are exponential in the tree width. $\rightsquigarrow$ Efficient on trees
More generally graphs have tree width $= n$ and approximation algorithms are used (stochastic sampling, variational algs)

A Note on Causality



equivalent representations, they induce the same set of probability distributions. Yet one is consistent with perceptions and one not.

Causal structures allow to integrate consistently interventions

Beyond BN

- Markov Random Field
- Influence diagrams (include utilities associated with decisions. Three node types: chance, utilities and decisions

Outline

Basic Examples

To come

Conditional Independence