On Probabilistic Modeling and Bayesian Networks

Petri Myllymäki, Ph.D., Academy Research Fellow
Complex Systems Computation Group (CoSCo)
Helsinki Institute for Information Technology (HIIT)
Finland

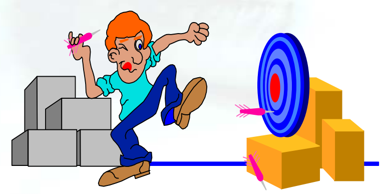
Petri.Myllymaki@hiit.FI, http://www.hiit.FI/Petri.Myllymaki/

Uncertain reasoning and data mining

- Real-world environments are complex
  - pure logic is not a feasible tool for describing the underlying stochastic rules
- It is possible to learn about the underlying uncertain dependencies via observations
  - as shown by the success of some human experts
- Obtaining and communicating this type of deep knowledge is difficult
  - the objective: to develop clever algorithms and methods that help people in these tasks

Different approaches to uncertain reasoning

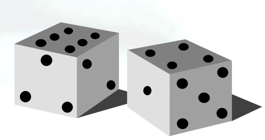
- (Bayesian) probability theory
- neural networks
- fuzzy logic
- possibility measures
- case-based reasoning
- kernel estimators
- support vector machines
- etc....



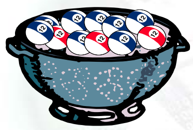
Two perspectives on probability

- The classical frequentist approach (Fisher, Neyman, Cramer, ...)
  - probability of an event is the long-run frequency with which it happens
    - but what then is the probability that the world ends tomorrow?
  - the goal is to find ”the true model”
  - hypothesis testing, classical orthodox statistics
- The modern subjectivist approach (Bernoulli, Bayes, Laplace, Jeffreys, Lindley, Jaynes, …)
  - probability is a degree of belief
  - models are believed to be true with some probability (”All models are false, but some are useful”)

  ⇒ **Bayesian networks**



The Bayes rule



Model M

Data D

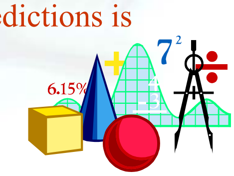
Thomas Bayes (1701-1761)

$$P(M|D) = \frac{P(D|M)P(M)}{P(D)} \propto P(D|M)P(M)$$

- *"The probability of a model M after observing data D is proportional to the likelihood of the data D assuming that M is true, times the prior probability of M."*
- *Bayesianism = subjective probability theory*

Advantages of the Bayesian approach

- A consistent calculus for uncertain reasoning
  - the Cox theorem: constructing a non-Bayesian consistent calculus is difficult
- Decision theory offers a theoretical framework for optimal decision-making
  - **requires** probabilities!
- Transparency
  - A "white box": all the model parameters have a clear semantic interpretation
  - The certainty associated to probabilistic predictions is intuitively understandable
  - cf. "black boxes" like neural networks

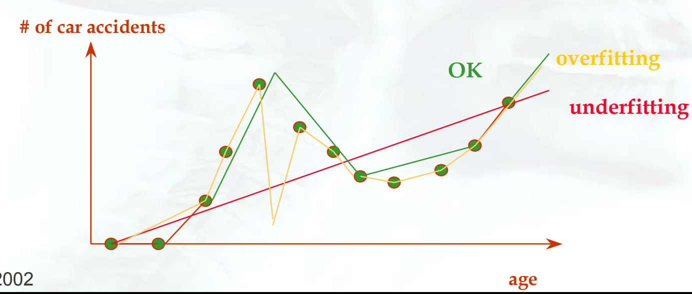


More advantages of Bayesianism

- Versatility
  - Probabilistic inference: compute P(what you want to know | what you already know).
  - cf. single-purpose models like decision trees
- An elegant framework for learning models from data
  - Works with any size data sets
  - Can be combined with prior expert knowledge
  - Incorporates an automatic **Occam's razor principle**, avoids overfitting

The Occam's razor principle

- "If two models of different complexity both fit the data approximately equally well, then the simpler one usually is a better predictive model in the future."
- Overfitting: fitting an overly complex model to the observed data



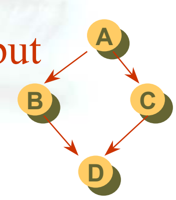
Bayesian metric for learning

$$P(M|D) = \frac{P(D|M)P(M)}{P(D)} \propto P(D|M)P(M)$$

- P(D) is constant with respect to different models, so it can be considered constant.
- Prior P(M) can be determined by experts, or ignored if no prior knowledge is available.
- *The evidence criterion (data marginal likelihood) P(D|M) is an integral over the model parameters, which causes the criterion to automatically penalize too complex models.*

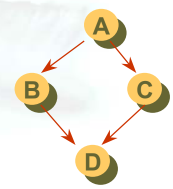
Probability theory in practice

- Bayesian networks: a family of probabilistic models and algorithms enabling computationally efficient
  1. Probabilistic inference
  2. Automated learning of models from sample data
- Based on novel discoveries made in the last two decades by people like Pearl, Lauritzen, Spiegelhalter and many others
- Commercial exploitation growing fast, but still in its infant state



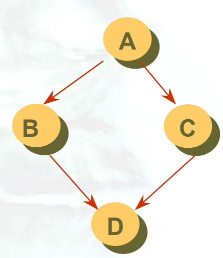
Bayesian networks

- A Bayesian network is a model of probabilistic dependencies between the domain variables.
- The model can be described as a list of dependencies, but is is usually more convenient to express them in a graphical form as a directed acyclic network.
- The nodes in the network correspond to the domain variables, and the arcs reveal the underlying dependencies, i.e., the hidden structure of the domain of your data.
- The strengths of the dependencies are modeled as conditional probability distributions (not shown in the graph).



Dependencies and Bayesian networks

- The Bayesian network on the right represents the following list of dependencies:
  - A and B are dependent on each other no matter what we know and what we don't know about C or D (or both).
  - A and C are dependent on each other no matter what we know and what we don't know about B or D (or both).
  - B and D are dependent on each other no matter what we know and what we don't know about A or C (or both).
  - C and D are dependent on each other no matter what we know and what we don't know about A or B (or both).
  - A and D are dependent on each other if we do not know both B and C.
  - B and C are dependent on each other if we know D or if we do not know D and also do not know A.

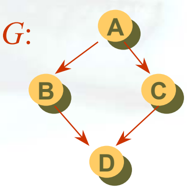


Bayesian networks: the textbook definition

- A Bayesian (belief) network representation for a probability distribution P on a domain $(X_1,\ldots,X_n)$ is a pair (G,θ), where G is a directed acyclic graph whose nodes correspond to the variables $X_1,\ldots,X_n$, and whose topology satisfies the following: each variable X is conditionally independent of all of its non-descendants in G, given its set of parents F_X, and no proper subset of F_X satisfies this condition. The second component θ is a set consisting of all the conditional probabilities of the form $P(X|F_X)$.

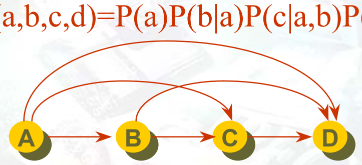
θ = {P(+a), P(+b|+a), P(+b|-a), P(+c|+a), P(+c|-a), P(+d|+b,+c), P(+d|-b,+c), P(+d|+b,-c), P(+d|-b,-c)}

G:

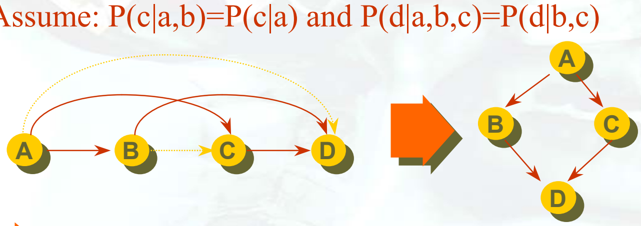


A more intuitive description

- From the axioms of probability theory, it follows that P(a,b,c,d)=P(a)P(b|a)P(c|a,b)P(d|a,b,c)



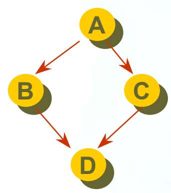
- Assume: P(c|a,b)=P(c|a) and P(d|a,b,c)=P(d|b,c)



$$P(x_1,\ldots,x_n)=\prod_{i=1}^{n} P(x_i\,|\,F_{X_i})$$

Why does it work?

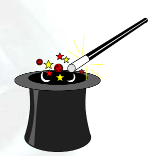
- simple conditional probabilities are easier to determine than the full joint probabilities
- in many domains, the underlying structure corresponds to relatively sparse networks, so only a small number of conditional probabilities is needed



P(+a,+b,+c,+d)=P(+a)P(+b|+a)P(+c|+a)P(+d|+b,+c)
P(–a,+b,+c,+d)=P(–a)P(+b|–a)P(+c|–a)P(+d|+b,+c)
P(–a,–b,+c,+d)=P(–a)P(–b|–a)P(+c|–a)P(+d|–b,+c)
P(–a,–b,–c,+d)=P(–a)P(–b|–a)P(–c|–a)P(+d|–b,–c)
P(–a,–b,–c,–d)=P(–a)P(–b|–a)P(–c|–a)P(–d|–b,–c)
P(+a,–b,–c,–d)=P(+a)P(–b|+a)P(–c|+a)P(–d|–b,–c)
. . .

Computing the evidence

- Under certain natural technical assumptions, the *evidence* criterion P(D|M) for a given BN structure M and database D can be computed exactly in feasible time:



$$P(D|M)=\int P(D|M,\theta)P(\theta)d\theta=\prod_{i=1}^{n}\prod_{j=1}^{q_i}\frac{\Gamma(N_{ij}^{'})}{\Gamma(N_{ij}^{'}+N_{ij})}\prod_{k=1}^{r_i}\frac{\Gamma(N_{ijk}^{'}+N_{ijk})}{\Gamma(N_{ijk}^{'})},$$

where: n is the number of variables in M,
q_i is the number of predecessors of X_i
r_i is the number of possible values for X_i
N_{ijk} is the number of cases in D, where $X_i=x_{ik}$ and $F_i=f_{ij}$
N_{ij} is the number of cases in D where $F_i=f_{ij}$
$N_{ijk}^{'}$ is the Dirichlet exponent of θ_{ijk}, "a prior number of cases " identical to the N_{ijk} in D.
$N_{ij}^{'}$ is the "prior number of cases" identical to the N_{ij} in D.

B-Course: An Interactive Tutorial on Bayesian Networks
http://b-course.cs.helsinki.fi

Petri Myllymäki, Henry Tirri: Bayes-verkkojen mahdollisuudet (Tekesin Teknologiaraportti 58/98)

Copies of these slides, the above report and other relevant material can be found at http://www.cs.helsinki.fi/research/cosco/Bnets