Probabilistic Modeling with Bayesian Networks

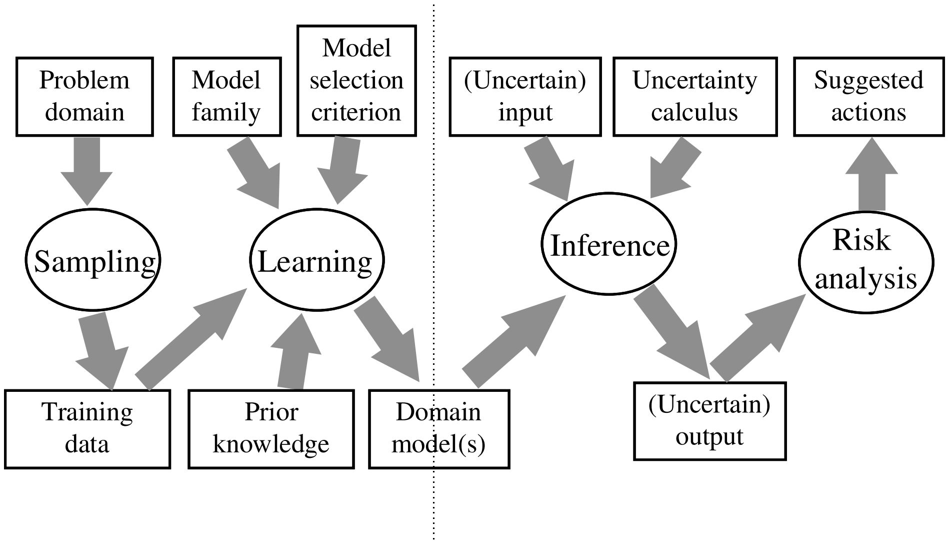
Petri Myllymäki
Complex Systems Computation Group (CoSCo)
P.O.Box 26, Department of Computer Science
FIN-00014 University of Helsinki, Finland

Petri.Myllymaki@cs.Helsinki.FI, http://www.cs.Helsinki.FI/~myllymak/

Outline

- Framework for building decision support systems
- Uncertainty measures
- Reasons for using probability theory
- Probabilistic reasoning
- Bayesian networks
- Probabilistic reasoning by Bayesian networks
- Constructing Bayesian networks
- Variations of standard Bayesian networks
- Connections to other models
- Applications
- Software
- Useful links
- Recommended reading

Framework for building decision support systems



Measures of uncertainty

- probability measure (Bayes 1763)
- Dempster-Shafer belief functions (Shafer 1976)
- fuzzy measures (Zadeh 1975)
- possibility measures (Dubois & Prade 1990)
- ε-semantics (Pearl 1989)
- κ-rankings (Goldszmidt & Pearl 1992)

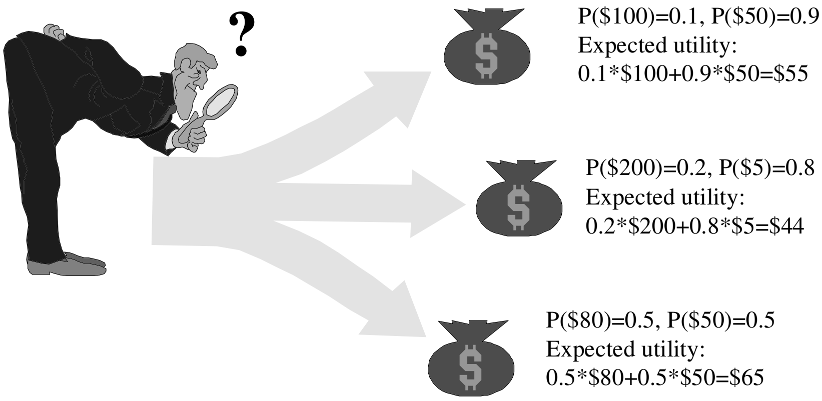
Bayesian networks: a billion dollar perspective

"Microsoft's competitive advantage, he [Gates] responded, was its expertise in "Bayesian networks". Ask any other software executive about anything "Bayesian" and you're liable to get a blank stare. Is Gates onto something? Is this alien-sounding technology Microsoft's new secret weapon?"

(Leslie Helms, Los Angeles Times, October 28, 1996.)

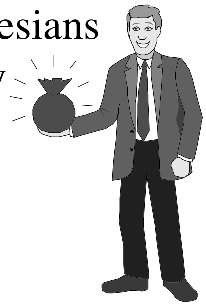
Reasons for using probability theory

- Decision theory offers a theoretical framework for optimal decision making



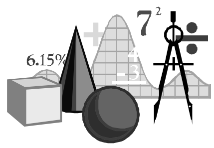
Optimal actions

- Optimal policy: choose the action with **maximal expected utility**
- The Dutch book argument: the Bayesians always end up with the most money
- Where to get the utilities?
  - ➔ decision **support** systems



More reasons for using probability theory

- The input and output variables (the inference task) does not have to be determined in advance
  - – probabilistic models can be used for solving both discriminative tasks (classification), and regression problems (configuration problems and prediction)
  - – cf. decision trees



More reasons for using probability theory

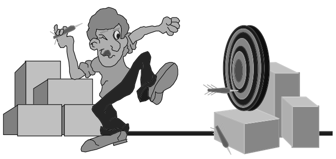
- consistent calculus
  - creating a consistent calculus for uncertain inference is not easy (the Cox theorem)
  - cf. fuzzy logic
- Probabilistic models can handle both discrete and continuous variables at the same time
- Various approaches for handling missing data (both in learning and in reasoning)

Nice theory, but...

- "isn't probabilistic reasoning counter-intuitive, something totally different from human reasoning?"
  - Cause for confusion: the old frequentist interpretation. But probabilities does NOT have to be thought of as frequencies, but as measures of belief
    - consider e.g. the probability of World War III starting at the year 2000.
  - The so called paradoxes are often misleading
    - A: P($1.000.000)=1.0
    - B: P($1.000.000)=0.25, P($4.000.000)=0.25, P($0)=0.5
  - Even if that were true, maybe that would be a good thing!
  - Probabilistic vs. fuzzy reasoning: (Cheeseman 1985, 1986)

Nice theory, but...

- "Where do all the numbers come from?"
  - Bayesian networks: small number of parameters
  - the numbers does not have to be accurate (Henrion et al 1996).
  - probability theory offers a framework for constructing models from sample data, from expert knowledge, or from their combination



Probabilistic reasoning

- n (discrete) random variables $X_1,...,X_n$
- joint probability distribution $P(X_1,...,X_n)$
- Input: a partial value assignment E,

$E=< X_1, X_2=x_2, X_3, X_4=x_4, X_5=x_5, X_6,...,X_n>$

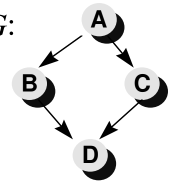
- Probabilistic reasoning:
  - compute $P(X=x|E)$ for all X not instantiated in E, and for all values of X (marginal distribution).
  - find a MAP (maximum a posterior probability) assignment consistent with E

Bayesian networks: the textbook definition

- A Bayesian (belief) network representation for a probability distribution *P* on a domain (X_1,...,X_n) is a pair (*G*,θ), where *G* is a directed acyclic graph whose nodes correspond to the variables X_1,...,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 $P(X|F_X)$.

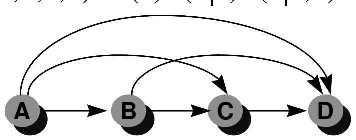
θ = {P(+a), P(+b|+a), P(+b|-a), P(+c|+a), P(+c|-a), P(+d|+b,+c), P(+d|-b,+c), P(+d|+b,-c), P(+d|-b,-c)}

G:

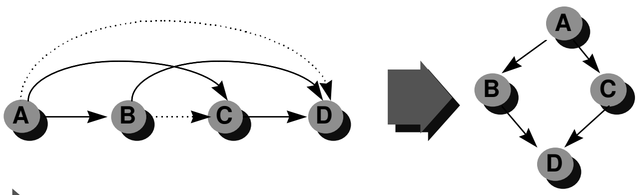


A more intuitive description

- From Bayes' rule, it follows that P(a,b,c,d)=P(a)P(b|a)P(c|a,b)P(d|a,b,c)



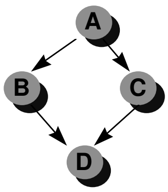
- Assume: P(c|a,b)=P(c|a) and P(d|a,b,c)=P(d|b,c)



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

Advantages of using BNs

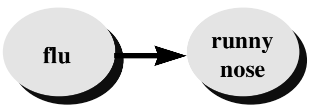
- simple conditional probabilities are easier to determine than the full joint probabilities
- in many domains, the networks are relatively sparse, so only a small number of conditional probabilities is needed



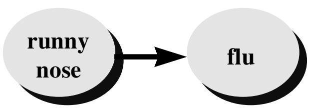
$$P(+a,+b,+c,+d)=P(+a)P(+b|+a)P(+c|+a)P(+d|+b,+c)$$
$$P(-a,+b,+c,+d)=P(-a)P(+b|-a)P(+c|-a)P(+d|+b,+c)$$
$$P(-a,-b,+c,+d)=P(-a)P(-b|-a)P(+c|-a)P(+d|-b,+c)$$
$$P(-a,-b,-c,+d)=P(-a)P(-b|-a)P(-c|-a)P(+d|-b,-c)$$
$$P(-a,-b,-c,-d)=P(-a)P(-b|-a)P(-c|-a)P(-d|-b,-c)$$
$$P(+a,-b,-c,-d)=P(+a)P(-b|+a)P(-c|+a)P(-d|-b,-c)$$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

Are the links causal?

- Not necessarily, but causality makes it easier to determine the conditional probabilities.



$$P(flu, rn) = P(flu)P(rn \mid flu)$$

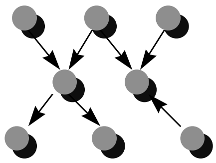


$$P(flu, rn) = P(rn)P(flu \mid rn)$$

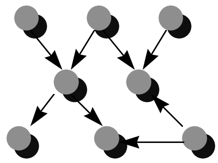
- Equivalence class=set of BN structures which can used for representing exactly the same set of probability distributions.

Singly-connected BNs

- a singly connected BN = polytree (disregarding the arc directions, no two nodes can be connected with more than one path).

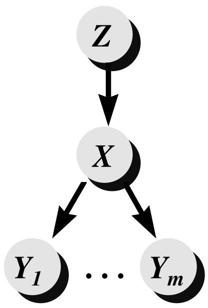


singly-connected



multi-connected

Probabilistic reasoning in singly-connected BNs



$$P(X = x|E) \propto P(E_{X-}|X = x)P(X = x|E_{X+})$$

$$P(E_{X-}|X = x) = \prod_{Y} P(E_{Y-}|X = x)$$

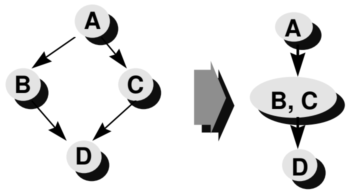
$$P(E_{Y-}|X = x) = \sum_{y} P(Y = y|X = x)P(E_{Y-}|Y = y)$$

$$P(X = x|E_{X+}) = \sum_{z} P(X = x|Z = z)P(Z = z|E_{Z+})$$

- a computationally efficient message-passing scheme: time requirement linear in the number of conditional probabilities in θ.

Probabilistic reasoning in multi-connected BNs

- generally not computationally feasible as the problem has been shown to be NP-hard (Cooper 1990, Shimony 1994).
- exact methods:
  - clustering
  - conditioning
  - algebraic methods



- approximative methods:
  - stochastic sampling algorithms
  - deterministic approximations with bounded accuracy

Constructing Bayesian networks

- For constructing probabilistic models, you can use either
  - domain expert knowledge (knowledge aquisition),
  - statistical data (machine learning),
  - or both
- For learning, you need to define
  - the model space
  - a criterion for good models
  - a search algorithm for finding good models

Learning Bayesian networks

- Model space:
  - the space of all the possible DAGs, or alternatively, the space of all the possible equivalence classes of the DAGs (E-space, Chickering 1996).
- Search algorithm:
  - use your favorite
  - natural operations are adding, removing, and reversing an arc (somewhat different for the E-space)
  - when the networks structure has been fixed, determining the parameters (the probability tables) is straightforward

Bayesian metric for learning

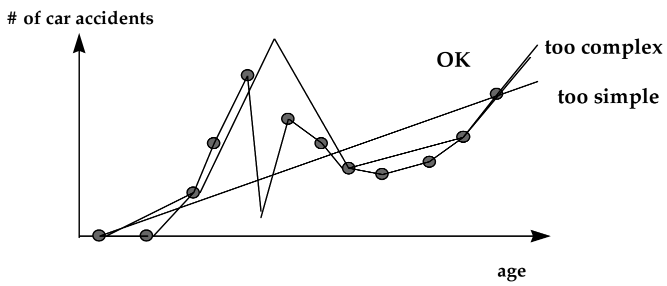
$$P(M|D) = \frac{P(D|M)P(M)}{P(D)} \propto P(D|M)P(M)$$

- P(D) is constant with respect to different models, so it can be considered constant
- The *evidence* (marginal likelihood) criterion: P(D|M) (data+expert knowledge)

  $P(D|M) = \int P(D|M,\theta)P(\theta|M)d\theta$
- Prior P(M) (expert knowledge)
- can be used for combining expert knowledge with statististical data

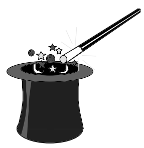
The Occam's razor principle

- The Bayesian learning framework allows methods for determining the optimal model size
  - automatic Occam's razor, avoids overfitting!



Computing the evidence

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

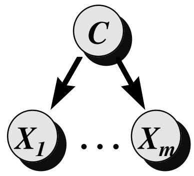


$$P(D|M) = \int P(D|M,\theta)P(\theta|M)d\theta = \prod_{i=1}^{n}\prod_{j=1}^{q_i}\frac{\Gamma(N'_{ij})}{\Gamma(N'_{ij}+N_{ij})}\prod_{k=1}^{r_i}\frac{\Gamma(N'_{ijk}+N_{ijk})}{\Gamma(N'_{ijk})},$$

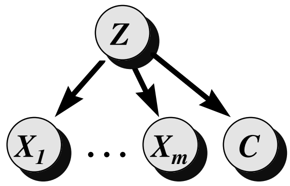
where: n is the number of variables in M,
q_i is the number of predecessors of X_i
r_i is the number of possible values for X_i
N_{ijk} is the number of cases in D, where $X_i=x_{ik}$ and $F_i=f_{ij}$
N_{ij} is the number of cases in D where $F_i=f_{ij}$
N_{ijk}' is the Dirichlet exponent of θ_{ijk}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

Variations of standard BNs

- mixed variables
- noisy-or: P(a | b,c,d)=1-(P(-a|b)P(-a|c)P(-a|d))
- qualitative BNs
- the naive Bayes model
- finite mixture models
- more general graphical models (chain graphs etc.)



the naive Bayes model



the finite mixture model

Information theoretic view of the evidence measure

- The *stochastic complexity* measure (Rissanen 1989) can be seen as an information theoretic formulation for the evidence:

  $$SC(D) = - \log P(D|M).$$

- SC(D) is the *shortest possible codelength* needed for coding the data D, by using the model class M
- An alternative coding scheme:

  see (Rissanen, 1996, Kontkanen et al., 1997)

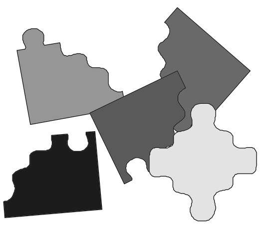
Connections to other models

- The finite mixture model can be seen as a common framework for
  - kernel density estimators/radial basis functions
  - probabilistic neural networks
  - case-based reasoning
  - nearest neighbor methods

  See (Tirri et al. 1996) and other publications of the CoSCo group at http://www.cs.Helsinki.FI/research/cosco/

…more connections

- Sigmoid belief networks (Jordan 1995)
- Bayesian learning of neural networks (Mackay 1992)
- Mapping Bayesian networks to Boltzmann machines (Myllymäki 1995)



Applications

- Microsoft: AnswerWizard, software and system troubleshooters, MS Home Healthcare System, "adaptive OS" Lumiere,...
- NASA: Shuttle mission control (VISTA), automated classification of satellite images (LandSat, IRAS)
- Intel: Processor fault diagnosis
- Lockheed: Autonomous underwater vehicle control
- GE: Power generator monitoring
- Mitre: Weapons scheduling
- Several medical diagnostic systems (Intellipath, DXPLAIN, ILIAD, ACORN, ...)
- ...
- More: see the AUAI home page.

Available software

- Commercial:
  - Baron, Demos, DXpress, Ergo, GRAPHICAL-BELIEF 2.0, HUGIN, Netica, Strategist
- Free:
  - BAYES, BELIEF 1.2, IDEAL, MacEvidence, Microsoft Bayes Networks (MSBN), Pulcinella, S-ElimBel, SPI, TresBel, XBaies
- More info:

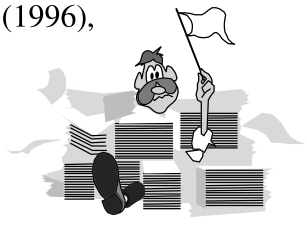
http://bayes.stat.washington.edu/almond/belief.html

Useful links

- http://www.auai.org/ (Association for Uncertainty in AI)
- http://www.vuse.vanderbilt.edu/~dfisher/ai-stats/society.html (Society for Artificial Intelligence and Statistics)
- http://www.cs.Helsinki.FI/research/cosco/ (CoSCo group home page)
- http://bayes.stat.washington.edu/bayes_people.html (Bayesians worldwide)

Recommended reading:

- Uncertainty measures: Shafer & Pearl (1990), Friedman & Halpern (1996), Paris (1994), Walley (1996)
- Probability theory: Gelman et al (1994), Bernardo & Smith (1994), Jaynes (199?)
- Decision theory: Gärdenfors & Sahlin (1988), Lindley (1971), DeGroot (1970)
- Bayesian networks: Pearl (1988), Neapolitan (1990), Jensen (1996), Castillo et al. (1996), Charniak (1991)
- More general graphical models: Lauritzen (1996), Whittaker (1990), Schafer (1996)
- Bayesian learning: Heckerman (1995), Buntine (1994), MacKay (1992)



References

- Bernardo & Smith, 1994: Bayesian Theory. John Wiley & Sons, Chichester.
- Buntine, 1994: Operations for learning with graphical models. Journal of Artificial Intelligence Research 2, 159-225.
- Castillo, Gutiérrez, and Hadi, 1996: Expert Systems and Probabilistic Network Models. Springer-Verlag .
- Charniak, 1991: Bayesian Networks without Tears. AI Magazine 12 (4), 50-63.
- Cheeseman, 1985: In Defense of Probability. Pp. 1002-1009 in Proceedings of IJCAI'85.
- Cheeseman, 1986: Probabilistic versus Fuzzy Reasoning. Pp. 85-102 in Proceedings of UAI'86.
- Chickering, 1996: Learning equivalence classes of Bayesian network structures. Pp. 150-157 in Proceedings of UAI'96.
- Cooper, 1990: The computational complexity of probabilistic inference using Bayesian belief networks. Artificial Intelligence 42, 393-405.
- DeGroot, 1970. Optimal Statistical Decisions. McGraw-Hill Book Company, New York.
- Dubois & Prade, 1990: An introduction to possibilistic and fuzzy logics. Pp. 742-762 in Shafer & pearl, 1990.
- Friedman & Halpern, 1996: Plausibility Measures and Default Reasoning. Pp.1297-1304 in Proceedings of AAAI'96.
- Gelman, Carlin, Stern, and Rubin, 1995: Bayesian Data Analysis. Chapman & Hall, London.
- Goldszmidt & Pearl, 1992: Rank-based systems: a simple approach to nonmonotonic reasoning. Pp. 661-672 in Proceedings of KR'92.
- Gärdenfors & Sahlin, 1988: Decision, Probability, and Utility. Cambridge University Press, New York.
- Heckerman, Geiger, and Chickering, 1995: Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. Machine Learning 20, 197-243.
- Henrion, Pradhan, Del Favero, Huang, Provan, and O'Rourke, 1996: Why is diagnosis using belief networks insensitive to imprecision in probabilities? Pp. 307-314 in Proceedings of UAI'96.
- Jaynes, 199?: Probability Theory: The Logic of Science. Under construction, a draft copy can be obtained from ftp://bayes.wustl.edu/Jaynes.book
- Jensen, 1996: An Introduction to Bayesian Networks. UCL Press, London.
- Jordan, 1995: Why the logistic function? A tutorial discussion on probabilities and neural networks. Technical report 9503, Computational Cognitive Science, MIT.

- Kontkanen, Myllymäki, Silander, Tirri, and Grünwald, 1997: Comparing Predictive Inference Methods for Discrete Domains. Proc. Sixth Int. Workshop on Artificial Intelligence and Statistics, 311-318.
- Lauritzen, 1996: Graphical Models. Oxford University Press.
- Lindley, 1971: Making Decisions. John Wiley & Sons, London.
- MacKay, 1992: Bayesian Methods for Adaptive Models. Ph.D. Dissertation, California Institute of Technology.
- Myllymäki, 1995: Mapping Bayesian Networks to Stochastic Neural Networks: A Foundation for Hybrid Bayesian-Neural Systems. Ph.D. Dissertation, Report A-1995-1, Department of Computer Science, University of Helsinki.
- Neapolitan, 1990: Probabilistic Reasoning in Expert Systems: Theory and Algorithms. John Wiley & Sons, New York.
- Paris, 1994: The Uncertain Reasoner's Companion. University Press, Cambridge.
- Pearl, 1988: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann Publishers, San Mateo, CA.
- Rissanen, 1989: Stochastic Complexity in Statistical Inquiry. World Scientific Publishing Company, New Jersey.
- Rissanen, 1996: Fisher Information and Stochastic Complexity. IEEE Trans. on Inf. Theory, 42 (1), 40-47.
- Pearl, 1989: Probabilistic semantics for nonmonotonic reasoning: a survey. Pp. 505-516 in Proceedings of KR'89.
- Shafer, 1976: A Mathematical Theory of Evidence. Princeton University Press.
- Schafer, 1996: Probabilistic Expert Systems. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA.
- Shafer & Pearl, 1990: Readings in Uncertain Reasoning. Morgan Kaufmann Publishers, San Mateo, CA..
- Shimony, 1994: Finding MAPs for belief networks is NP-hard. Artificial Intelligence 68, 399-410.
- Tirri, Kontkanen, and Myllymäki, 1996: Probabilistic Instance-Based Learning. Pp. 507-515 in Machine Learning: Proceedings of the Thirteenth International Conference, edited by L.Saitta. Morgan Kaufmann Publishers, San Francisco, CA.
- Walley, 1996: Measures of uncertainty in expert systems. Artificial Intelligence 83, 1-58.
- Whittaker, 1990: Graphical Models in Applied Multivariate Statistics. John Wiley & Sons, Chichester.
- Zadeh, 1975: Fuzzy logic and approximate reasoning. Synthese (30), 407-425.