Learning Bayesian Networks from Data

—AAAI 1998 Tutorial—

Additional Readings

The following is a list of references to the material covered in the tutorial and to more advanced subjects mentioned at various points. This list is far from being comprehensive and is intended only to provide useful starting points.

Background Material

Bayesian Networks A good reference on Bayesian networks is [Pearl 1988]. A more recent book, which covers Bayesian network inference in depth is [Jensen 1996]. A a short and gentle introduction can be found in [Charniak 1991].

Statistics, Pattern Recognition and Information Theory There are many books on statistics. We find [DeGroot 1970] to be a good introduction to statistics and Bayesian statistics in particular. A more recent book [Gelman et al. 1995] is also a good introduction to this field and also discusses recent advances, such as hierarchical priors. Books in pattern recognition, including the classic [Duda and Hart 1973] and the more recent [Bishop 1995], cover basic issues in density estimation and their use for pattern recognition and classification. A good introduction to information theory, and notions such as KL divergence and mutual information can be found in [Cover and Thomas 1991].

Tutorials and Surveys [Heckerman 1995] provides an in-depth tutorial on Bayesian methods in learning Bayesian networks. [Buntine 1996] surveys the literature. [Jordan 1998] is a collection of introductory surveys and papers discussing recent advances.

Learning Parameters

Learning parameters from complete data is discussed in [Spiegelhalter and Lauritzen 1990]. A more recent discussion can be found in [Buntine 1994]. An introduction to the possible problems with incomplete data and MAR assumptions can be found in [Rubin 1976]. Learning parameters from incomplete data using gradient methods is discussed by [Binder et al. 1997; Thiesson 1995]. The original EM paper is [Dempster et al. 1977]; an elegant alternative explanation of EM can found in [Neal and Hinton 1998]. [Lauritzen 1995] describes how to apply EM to Bayesian networks. [Bauer et al. 1997] describe methods for accelerating the convergence of EM. Learning using Gibbs sampling is discussed in [Thomas et al. 1992; Gilks et al. 1996].

Learning Structure

Complete Data The Bayesian score is originally discussed in [Cooper and Herskovits 1992] and further developed in [Buntine 1991; Heckerman et al. 1995]. The MDL score is based on the Minimal Description Length principle of [Rissanen 1989]; the application of this principle to Bayesian networks was developed by several authors [Bouckaert 1994; Lam and Bacchus 1994a; Suzuki 1993]. The method for learning trees was initially introduced in [Chow and Liu 1968] (see also the description in [Pearl 1988]). Learning structure using greedy hill-climbing and other variants is discussed and evaluated in [Heckerman et al. 1995]. See [Chickering 1996b] for search over equivalence network classes. [Buntine 1991; Heckerman et al. 1995; Madigan and Raftery 1994] discuss methods for approximating the full Bayesian model averaging.

Incomplete Data [Chickering and Heckerman 1997] discuss the problems with evaluating the score of networks in the presence of incomplete data and describe several approximation to the score. [Cheeseman and Stutz 1995] discuss Bayesian learning of mixture models with a single hidden variable. Recent works on learning structure in the presence of incomplete data include [Friedman 1997; Friedman 1998; Meila and Jordan 1998; Singh 1997; Thiesson et al. 1998].

Causal Discovery

For different views of the relation of causality and Bayesian networks see [Heckerman and Shachter 1995; Pearl 1993; Spirtes et al. 1993]. [Pearl and Verma 1991; Spirtes et al. 1993] describe constraint-based methods for learning causal relation from data. The Bayesian approach is discussed in [Heckerman et al. 1997].

Advanced Topics

Continuous Variables See [Heckerman and Geiger 1995] for methods of learning a network that contains Gaussian distributions. [Hofmann and Tresp 1996; John and Langley 1995] discuss learning Bayesian networks with non-parametric representations of density functions. [Monti and Cooper 1997] use neural networks to represent the conditional densities. [Friedman and Goldszmidt 1996] learn Bayesian networks over continuous domains by discretizing the values of the continuous variables.

Local Structure [Buntine 1991; Diez 1993] discuss learning the "noisy-or" conditional probability. [Meek and Heckerman 1997] discuss how to learn a several extensions of this local model. [Friedman and Goldszmidt 1998] describe how to learn tree-like representations of local structure and why this helps in learning global structure. [Chickering et al. 1997] extend these results to richer representations and discuss more advanced search procedures for learning both global and local structure.

Online Learning & Updates See [Buntine 1991; Friedman and Goldszmidt 1997; Lam and Bacchus 1994b] for discussion on how to sequentially update the structure of a network as more data is available.

Temporal Processes *Dynamic Bayesian networks* [Dean and Kanazawa 1989] is an extension of Bayesian networks for representing stochastic models. [Smyth et al. 1997] discussed how this representation generalizes hidden Markov networks, and how methods from both fields are related. [Ghahramani and Jordan 1997] describe methods for learning parameters for complex dynamic Bayesian networks with non-trivial unobserved state. [Friedman et al. 1998] describe methods for learning the structure of dynamic Bayesian networks.

Theory [Chickering 1996a] shows that finding the structure that maximizes the Bayesian score is NP-hard. [Dasgupta 1997; Friedman and Yakhini 1996] discuss the *sample complexity*—that is, how many examples are required to achieve a desired accuracy—for learning parameters and structure.

Applications

The AutoClass system [Cheeseman and Stutz 1995] is an unsupervised clustering program that the simple "naive" Bayesian network. This program has been used in numerous applications. The "naive" Bayesian classifier has been used since the early days of pattern recognition [Duda and Hart 1973]. [Ezawa and Schuermann 1995; Friedman et al. 1997; Singh and Provan 1995] describe applications of more complex Bayesian network learning algorithms for classification. [Zweig and Russell 1998] use Bayesian networks for speech recognition. [Breese et al. 1998] discuss collaborative filtering methods that use Bayesian network learning algorithms. [Spirtes et al. 1993] describe several applications of causal learning in social sciences.

References

Bauer, E., D. Koller, and Y. Singer (1997). Update rules for parameter estimation in bayesian networks. In D. Geiger and P. Shanoy (Eds.), *Proc. Thirteenth Conference on Uncertainty in Artificial Intelligence (UAI '97)*, San Francisco, Calif., pp. 3–13. Morgan Kaufmann.

Binder, J., D. Koller, S. Russell, and K. Kanazawa (1997). Adaptive probabilistic networks with hidden variables. *Machine Learning 29*, 213–244.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford, U.K.: Oxford University Press.

Bouckaert, R. R. (1994). Properties of Bayesian network learning algorithms. In R. López de Mantarás and D. Poole (Eds.), *Proc. Tenth Conference on Uncertainty in Artificial Intelligence (UAI '94)*, pp. 102–109. San Francisco, Calif.: Morgan Kaufmann.

Breese, J. S., D. Heckerman, and C. Kadie (1998). Empirical analysis of predictive algorithms for collaborative filtering. In G. F. Cooper and S. Moral (Eds.), *Proc. Fourteenth Conference on Uncertainty in Artificial Intelligence (UAI '98)*. San Francisco, Calif.: Morgan Kaufmann.

Buntine, W. (1991). Theory refinement on Bayesian networks. In B. D. D'Ambrosio, P. Smets, and P. P. Bonissone (Eds.), *Proc. Seventh Annual Conference on Uncertainty Artificial Intelligence (UAI '92)*, pp. 52–60. San Francisco, Calif.: Morgan Kaufmann.

Buntine, W. (1994). Operations for learning with graphical models. *J. of Artificial Intelligence Research 2*, 159–225.

Buntine, W. (1996). A guide to the literature on learning probabilistic networks from data. *IEEE Trans. on Knowledge and Data Engineering 8*, 195–210.

Charniak, E. (1991). Bayesian networks without tears. *AI Magazine 12*, 50–63.

Cheeseman, P. and J. Stutz (1995). Bayesian classification (AutoClass): Theory and results. In U. Fayyad, G. Piatesky-Shapiro, P. Smyth, and R. Uthurusamy (Eds.), *Advances in Knowledge Discovery and Data Mining*. Menlo Park, California: AAAI Press.

Chickering, D. M. (1996a). Learning Bayesian networks is NP-complete. In D. Fisher and H.-J. Lenz (Eds.), *Learning from Data: Artificial Intelligence and Statistics V*. Springer Verlag.

Chickering, D. M. (1996b). Learning equivalence classes of Bayesian-network structure. In E. Horvitz and F. Jensen (Eds.), *Proc. Twelfth Conference on Uncertainty in Artificial Intelligence (UAI '96)*. San Francisco, Calif.: Morgan Kaufmann.

Chickering, D. M. and D. Heckerman (1997). Efficient approximations for the marginal likelihood of incomplete data given a Bayesian network. *Machine Learning 29*, 181–212.

Chickering, D. M., D. Heckerman, and C. Meek (1997). A Bayesian approach to learning Bayesian networks with local structure. In D. Geiger and P. Shanoy (Eds.), *Proc. Thirteenth Conference on Uncertainty in Artificial Intelligence (UAI '97)*, San Francisco, Calif., pp. 80–89. Morgan Kaufmann.

Chow, C. K. and C. N. Liu (1968). Approximating discrete probability distributions with dependence trees. *IEEE Trans. on Info. Theory 14*, 462–467.

Cooper, G. F. and E. Herskovits (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning 9*, 309–347.

Cover, T. M. and J. A. Thomas (1991). *Elements of Information Theory*. New York: John Wiley & Sons.

Dasgupta, S. (1997). The sample complexity of learning fixed-structure Bayesian networks. *Machine Learning 29*, 165–180.

Dean, T. and K. Kanazawa (1989). A model for reasoning about persistence and causation. *Computational Intelligence 5*, 142–150.

DeGroot, M. H. (1970). *Optimal Statistical Decisions*. New York: McGraw-Hill.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B 39*, 1–39.

Diez, F. J. (1993). Parameter adjustment in Bayes networks: The generalized noisy or-gate. In D. Heckerman and A. Mamdani (Eds.), *Proc. Ninth Conference on Uncertainty in Artificial Intelligence (UAI '93)*, pp. 99–105. San Francisco, Calif.: Morgan Kaufmann.

Duda, R. O. and P. E. Hart (1973). *Pattern Classification and Scene Analysis*. New York: John Wiley & Sons.

Ezawa, K. J. and T. Schuermann (1995). Fraud/uncollectable debt detection using a Bayesian network based learning system: A rare binary outcome with mixed data structures. In P. Besnard and S. Hanks (Eds.), *Proc. Eleventh Conference on Uncertainty in Artificial Intelligence (UAI '95)*, pp. 157–166. San Francisco, Calif.: Morgan Kaufmann.

Friedman, N. (1997). Learning Bayesian networks in the presence of missing values and hidden variables. In D. Fisher (Ed.), *Proceedings of the Fourteenth International Conference on Machine Learning*. San Francisco, Calif.: Morgan Kaufmann.

Friedman, N. (1998). The Bayesian structural EM algorithm. In G. F. Cooper and S. Moral (Eds.), *Proc. Fourteenth Conference on Uncertainty in Artificial Intelligence (UAI '98)*. San Francisco, Calif.: Morgan Kaufmann.

Friedman, N., D. Geiger, and M. Goldszmidt (1997). Bayesian network classifiers. *Machine Learning 29*, 131–163.

Friedman, N. and M. Goldszmidt (1996). Discretization of continuous attributes while learning Bayesian networks. In L. Saitta (Ed.), *Proceedings of the Thirteenth International Conference on Machine Learning*, pp. 157–165. San Francisco, Calif.: Morgan Kaufmann.

Friedman, N. and M. Goldszmidt (1997). Sequential update of Bayesian network structure. In D. Geiger and P. Shanoy (Eds.), *Proc. Thirteenth Conference on Uncertainty in Artificial Intelligence (UAI '97)*. San Francisco, Calif.: Morgan Kaufmann. To appear.

Friedman, N. and M. Goldszmidt (1998). Learning Bayesian networks with local structure. In M. I. Jordan (Ed.), *Learning in Graphical Models*. Dordrecht, Netherlands: Kluwer. A preliminary version appeared in E. Horvitz and F. Jensen eds., *Proc. Twelfth Conference on Uncertainty in Artificial Intelligence*, 1996, pp. 252–262.

Friedman, N., K. Murphy, and S. Russell (1998). Learning the structure of dynamic probabilistic networks. In G. F. Cooper and S. Moral (Eds.), *Proc. Fourteenth Conference on Uncertainty in Artificial Intelligence (UAI '98)*. San Francisco, Calif.: Morgan Kaufmann.

Friedman, N. and Z. Yakhini (1996). On the sample complexity of learning Bayesian networks. In E. Horvitz and F. Jensen (Eds.), *Proc. Twelfth Conference on Uncertainty in Artificial Intelligence (UAI '96)*. San Francisco, Calif.: Morgan Kaufmann.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (1995). *Bayesian Data Analysis*. London: Chapman & Hall.

Ghahramani, Z. and M. I. Jordan (1997). Factorial hidden Markov models. *Machine Learning 29*, 245–274.

Gilks, W., S. Richerdson, and D. Spiegelhalter (1996). *Markov Chain Monte Carlo in Practice*. Chapman and Hall.

Heckerman, D. (1995). A tutorial on learning with Bayesian networks. Technical Report MSR-TR-95-06, Microsoft Research, Redmond, Washington. Available from `http://www.research.microsoft.com/research/dtg/heckerma/heckerma.html`.

Heckerman, D. and D. Geiger (1995). Learning Bayesian networks: a unification for discrete and Gaussian domains. In P. Besnard and S. Hanks (Eds.), *Proc. Eleventh Conference on Uncertainty in Artificial Intelligence (UAI '95)*, pp. 274–284. San Francisco, Calif.: Morgan Kaufmann.

Heckerman, D., D. Geiger, and D. M. Chickering (1995). Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning 20*, 197–243.

Heckerman, D., C. Meek, and G. Cooper (1997). A bayesian approach to causal discovery. Technical report. Technical Report MSR-TR-97-05, Microsoft Research.

Heckerman, D. and R. Shachter (1995). Decision-theoretic foundations for causal reasoning. *Journal of A.I. Research 3*, 405–430.

Hofmann, R. and V. Tresp (1996). Discovering structure in continuous variables using Bayesian networks. In *Advances in Neural Information Processing Systems 8*, pp. 500–506.

Jensen, F. (1996). *An Introduction to Bayesian Networks*. Springer.

John, G. H. and P. Langley (1995). Estimating continuous distributions in Bayesian classifiers. In P. Besnard and S. Hanks (Eds.), *Proc. Eleventh Conference on Uncertainty in Artificial Intelligence (UAI '95)*, pp. 338–345. San Francisco, Calif.: Morgan Kaufmann.

Jordan, M. I. (Ed.) (1998). *Learning in Graphical Models*. Dordrecht, Netherlands: Kluwer.

Lam, W. and F. Bacchus (1994a). Learning Bayesian belief networks: An approach based on the MDL principle. *Computational Intelligence 10*, 269–293.

Lam, W. and F. Bacchus (1994b). Using new data to refine a Bayesian network. In R. López de Mantarás and D. Poole (Eds.), *Proc. Tenth Conference on Uncertainty in Artificial Intelligence (UAI '94)*, pp. 383–390. San Francisco, Calif.: Morgan Kaufmann.

Lauritzen, S. L. (1995). The EM algorithm for graphical association models with missing data. *Computational Statistics and Data Analysis 19*, 191–201.

Madigan, D. and A. Raftery (1994). Model selection and accounting for model uincertainty in graphical models using Occam's window. *Journal of the American Statiscal Association 89*, 1535–1546.

Meek, C. and D. Heckerman (1997). Structure and parameter learning for causal independence and causal interaction models. In D. Geiger and P. Shanoy (Eds.), *Proc. Thirteenth Conference on Uncertainty in Artificial Intelligence (UAI '97)*, San Francisco, Calif., pp. 366–375. Morgan Kaufmann.

Meila, M. and M. I. Jordan (1998). Estimating dependency structure as a hidden variable. In *NIPS 10*.

Monti, S. and G. F. Cooper (1997). Learning Bayesian belief networks with neural network estimators. In *Advances in Neural Information Processing Systems 9*, pp. 579–584.

Neal, R. M. and G. E. Hinton (1998). A new view of the EM algorithm that justifies incremental and other variants. In M. I. Jordan (Ed.), *Learning in Graphical Models*. Dordrecht, Netherlands: Kluwer.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. San Francisco, Calif.: Morgan Kaufmann.

Pearl, J. (1993). Graphical models, causality and intervention. *Statistical Science 8*, 266–273.

Pearl, J. and T. S. Verma (1991). A theory of inferred causation. In J. A. Allen, R. Fikes, and E. Sandewall (Eds.), *Principles of Knowledge Representation and Reasoning: Proc. Second International Conference (KR '91)*, pp. 441–452. San Francisco, Calif.: Morgan Kaufmann.

Rissanen, J. (1989). *Stochastic Complexity in Statistical Inquiry*. River Edge, NJ: World Scientific.

Rubin, D. R. (1976). Inference and missing data. *Biometrica 63*, 581–592.

Singh, M. (1997). Learning bayesian networks from incomplete data. In *Proc. National Conference on Artificial Intelligence (AAAI '97)*, pp. 27–31. Menlo Park, CA: AAAI Press.

Singh, M. and G. M. Provan (1995). A comparison of induction algorithms for selective and non-selective Bayesian classifiers. In A. Prieditis and S. Russell (Eds.), *Proceedings of the Twelfth International Conference on Machine Learning*, pp. 497–505. San Francisco, Calif.: Morgan Kaufmann.

Smyth, P., D. Heckerman, and M. Jordan (1997). Probabilistic independence networks for hidden Markov probability models. *Neural Computation 9*(2), 227–269.

Spiegelhalter, D. J. and S. L. Lauritzen (1990). Sequential updating of conditional probabilities on directed graphical structures. *Networks 20*, 579–605.

Spirtes, P., C. Glymour, and R. Scheines (1993). *Causation, Prediction and Search*. Number 81 in Lecture Notes in Statistics. New York: Springer-Verlag.

Suzuki, J. (1993). A construction of Bayesian networks from databases based on an MDL scheme. In D. Heckerman and A. Mamdani (Eds.), *Proc. Ninth Conference on Uncertainty in Artificial Intelligence (UAI '93)*, pp. 266–273. San Francisco, Calif.: Morgan Kaufmann.

Thiesson, B. (1995). Accelerated quantification of Bayesian networks with incomplete data. In *Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95)*, Montreal, Canada, pp. 306–311. AAAI Press.

Thiesson, B., C. Meek, D. M. Chickering, and D. Heckerman (1998). Learning mixtures of Bayesian networks. In G. F. Cooper and S. Moral (Eds.), *Proc. Fourteenth Conference on Uncertainty in Artificial Intelligence (UAI '98)*. San Francisco, Calif.: Morgan Kaufmann.

Thomas, A., D. Spiegelhalter, and W. Gilks (1992). Bugs: A program to perfrom Bayesian inference using Gibbs sampling. In J. Bernardo, J. Berger, A. Dawid, and A. Smith (Eds.), *Bayesian Statistics 4*, pp. 837–842. Oxford Univ. Press.

Zweig, G. and S. J. Russell (1998). Speech recognition with dynamic Bayesian networks. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI-98)*.