Learning Bayesian Networks from Data

—NIPS 2001 Tutorial—

Relevant 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 The seminal reference on Bayesian networks is [Pearl 1988]. A more recent book, which covers BN inference in more depth, as well as some of the recent developments in the area, is [Cowell et al. 1999]. 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 1998] 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.

Parameter Estimation

Learning parameters from complete data is discussed in [Spiegelhalter and Lauritzen 1990]. A more recent discussion can be found in [Buntine 1994].

Model Selection

The Bayesian score is originally discussed in [Cooper and Herskovits 1992] and further developed in [Buntine ; 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 ; Lam and Bacchus 1994; Suzuki]. 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] for search over equivalence network classes.

Several papers discuss the idea of doing structure discovery by approximating the full Bayesian model averaging. [Buntine ; Heckerman et al. 1995] discuss special cases where a full enumeration of models is possible. [Madigan and Raftery 1994] propose a heuristic approximation that restricts attention to only a subset of models. [Madigan and York 1995; Madigan et al. 1996; Giudici and Green 1999; Giudici et al. 2000] discuss the use of a Markov chain over the set of structures. [Friedman and Koller 2001] introduce the idea of a Markov chain over orderings.

Incomplete Data

Parameter Estimation 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 1994]. [Lauritzen 1995] describes how to apply EM to Bayesian networks. [Bauer et al.] describe methods for accelerating the convergence of EM. Learning using Gibbs sampling is discussed in [Gilks et al. 1996].

Model Selection [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. [P. and J. 1995] discuss Bayesian learning of mixture models with a single hidden variable. The structural EM approach was introduced in [Friedman 1997; Friedman 1998]. Other papers on structure learning with incomplete data include [Meila and Jordan ; Singh ; Thiesson et al.].

Causal Discovery

For different views of the relation of causality and Bayesian networks see [Spirtes et al. 1993; Heckerman and Shachter 1994; Pearl 2000]. [Pearl and Verma ; 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] for methods of learning a network that contains Gaussian distributions. [Hofmann and Tresp 1996; John and Langley] 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 a] learn Bayesian networks over continuous domains by discretizing the values of the continuous variables.

Learning Local Structure [Buntine ; Diez] 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.] extend these results to richer representations and discuss more advanced search procedures for learning both global and local structure.

Online & Active Learning See [Neal and Hinton 1994; Bauer et al.] for discussion on online parameter estimation for incomplete data, and [Buntine ; Friedman and Goldszmidt b; Lam and Bacchus] for sequential update of the structure as more data becomes available.

Active learning is a general framework where the learner can select additional samples that will best allow it to refine its learned model. [Tong and Koller 2001a] describes active learning for Bayesian networks with a fixed structure. [Tong and Koller 2001b] describes active learning for structure discovery.

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.] describe methods for learning the structure of dynamic Bayesian networks.

Hidden Variables [Elidan et al. ; Boyen et al. 1999] describe techniques for discovering a hidden variable from structural signatures in the learned model. [Elidan and Friedman 2001] describe a heuristic technique for picking the number of values for a hidden variable.

Probabilistic Relational Models *Probabilistic relational models* [Koller and Pfeffer] extend Bayesian networks to structured (relational) data. The basic framework for learning PRMs (parameters and structure) from data was discussed in [Friedman et al.]. [Taskar et al. 2001] shows how to deal with incomplete data in PRMs, and applies the framework to relational classification and clustering. [Getoor et al. 2001] shows how to learn PRMs which also include a probabilistic model about the presence of links.

Theory [Chickering 1996] 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 [P. and J. 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 ; Friedman et al. 1997; Singh and Provan] describe applications of more complex Bayesian network learning algorithms for classification. [Zweig and Russell] use Bayesian networks for speech recognition. [J. Breese 1998] discuss collaborative filtering methods that use Bayesian network learning algorithms. [Spirtes et al. 1993] describe several applications of causal learning in social sciences. The application of structural EM to phylogenetics is described in [Friedman et al. 2001]. [Segal et al. 2001] describes the application of probabilistic relational models to the analysis of gene microarray data.

References

Bauer, E., D. Koller, and Y. Singer. Update rules for parameter estimation in Bayesian networks. pp. 3–13.

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. Properties of Bayesian network learning algorithms. pp. 102–109.

Boyen, X., N. Friedman, and D. Koller (1999). Learning the structure of complex dynamic systems.

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

Buntine, W. L. Theory refinement on Bayesian networks. pp. 52–60.

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

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

Chickering, D. M. A transformational characterization of equivalent Bayesian network structures. pp. 87–98.

Chickering, D. M. (1996). 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. and D. Heckerman (1997). Efficient approximations for the marginal likelihood of bayesian networks with hidden variables. *Machine Learning 29*, 181–212.

Chickering, D. M., D. Heckerman, and C. Meek. A Bayesian approach to learning Bayesian networks with local structure. pp. 80–89.

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.

Cowell, R. G., A. P. Dawid, S. L. Lauritzen, and D. J. Spiegelhalter (1999). *Probabilistic Networks and Expert Systems*. Springer-Verlag.

Dasgupta, S. (1997). The sample complexity of learning fixed-structure Bayesian networks. *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. Parameter adjustment in Bayes networks: The generalized noisy or-gate. pp. 99–105.

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

Elidan, G. and N. Friedman (2001). Learning the dimensionality of hidden variables. In *Proc. Seventeenth Conf. on Uncertainty in Artificial Intelligence (UAI)*.

Elidan, G., N. Lotner, N. Friedman, and D. Koller. Discovering hidden variables: A structure-based approach.

Ezawa, K. J. and T. Schuermann. Fraud/uncollectable debt detection using a Bayesian network based learning system: A rare binary outcome with mixed data structures. pp. 157–166.

Friedman, N. (1997). Learning belief networks in the presence of missing values and hidden variables. pp. 125–133.

Friedman, N. (1998). The Bayesian structural EM algorithm.

Friedman, N., D. Geiger, and M. Goldszmidt (1997). Bayesian network classifiers. *29*, 139–164.

Friedman, N., L. Getoor, D. Koller, and A. Pfeffer. Learning probabilistic relational models.

Friedman, N. and M. Goldszmidt. Discretization of continuous attributes while learning Bayesian networks. pp. 157–165.

Friedman, N. and M. Goldszmidt. Sequential update of Bayesian network structure. To appear.

Friedman, N. and M. Goldszmidt (1998). Learning Bayesian networks with local structure. In M. I. Jordan (Ed.), *Learning in Graphical Models*, pp. 421–460. Dordrecht, Netherlands: Kluwer.

Friedman, N. and D. Koller (2001). Being Bayesian about Bayesian network structure: A Bayesian approach to structure discovery in Bayesian networks. *Machine Learning*. Accepted for publication. Earlier version appeared in UAI'2000.

Friedman, N., K. Murphy, and S. Russell. Learning the structure of dynamic probabilistic networks. pp. 139–147.

Friedman, N., M. Ninio, I. Pe'er, and T. Pupko (2001). A structural EM algorithm for phylogentic inference. In *Proc. RECOMB*.

Friedman, N. and Z. Yakhini (1996). On the sample complexity of learning Bayesian networks.

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

Getoor, L., N. Friedman, D. Koller, and B. Taskar (2001). Learning probabilistic models of relational structure. In *Eighteenth International Conference on Machine Learning (ICML)*.

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

Gilks, W., S. Richardson, and D. Spiegelhalter (1996). *Markov Chain Monte Carlo Methods in Practice*. CRC Press.

Giudici, P. and P. Green (1999, December). Decomposable graphical gaussian model determination. *Biometrika 86*(4), 785–801.

Giudici, P., P. Green, and C. Tarantola (2000). Efficient model determination for discrete graphical models. *Biometrika*. To appear.

Heckerman, D. (1998). A tutorial on learning with Bayesian networks. In M. I. Jordan (Ed.), *Learning in Graphical Models*. Dordrecht, Netherlands: Kluwer.

Heckerman, D. and D. Geiger. Learning Bayesian networks: a unification for discrete and Gaussian domains. pp. 274–284.

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 (1994). A decision-based view of causality. In *Proceedings of Tenth Conference on Uncertainty in Artificial Intelligence*, pp. 302–310. Morgan Kaufmann.

Hofmann, R. and V. Tresp (1996). Discovering structure in continuous variables using bayesian networks.

J. Breese, D. Heckerman, C. K. (1998). Empirical analysis of predictive algorithms for collaborative filtering.

John, G. H. and P. Langley. Estimating continuous distributions in Bayesian classifiers. pp. 338–345.

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

Koller, D. and A. Pfeffer. Probabilistic frame-based systems.

Lam, W. and F. Bacchus. Using new data to refine a Bayesian network. pp. 383–390.

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

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

Madigan, D., S. Andersson, M. Perlman, and C. Volinsky (1996). Bayesian model averaging and model selection for Markov equivalence classes of acyclic graphs. *Communications in Statistics: Theory and Methods 25*, 2493–2519.

Madigan, D. and E. Raftery (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal Americal Statistical Association 89*, 1535–1546.

Madigan, D. and J. York (1995). Bayesian graphical models for discrete data. *International statistical Review 63*, 215–232.

Meek, C. and D. Heckerman (1997). Structure and parameter learning for causal independence and causal interaction models. pp. 366–375.

Meila, M. and M. I. Jordan. Estimating dependency structure as a hidden variable.

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 (1994). A new view of the EM algorithm that justifies incremental and other variants. unpublished manuscript.

P., C. and S. J. (1995). Bayesian classification (AutoClass): Theory and results. In F. U., P.-S. G., S. P., and U. R. (Eds.), *Advances in Knowledge Discovery and Data Mining*, pp. 153–180. Menlo Park, CA: AAAI Press.

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

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge Univ. Press.

Pearl, J. and T. S. Verma. A theory of inferred causation. pp. 441–452.

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.

Segal, E., B. Taskar, A. Gasch, N. Friedman, and D. Koller (2001). Rich probabilistic models for gene expression. *Bioinformatics 17*(Suppl 1), S243–52.

Singh, M. Learning bayesian networks from incomplete data. pp. 27–31.

Singh, M. and G. M. Provan. A comparison of induction algorithms for selective and non-selective Bayesian classifiers. pp. 497–505.

Smyth, P., D. Heckerman, and M. I. 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. A construction of Bayesian networks from databases based on an MDL scheme. pp. 266–273.

Taskar, B., E. Segal, and D. Koller (2001). Probabilistic classification and clustering in relational data. In *Seventeenth International Joint Conference on Artificial Intelligence*, pp. 870–876.

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)*, pp. 306–311. AAAI Press.

Thiesson, B., C. Meek, D. M. Chickering, and D. Heckerman. Learning mixtures of Bayesian networks.

Tong, S. and D. Koller (2001a). Active learning for parameter estimation in Bayesian networks. In *Proc. NIPS 13*. To appear.

Tong, S. and D. Koller (2001b). Active learning for structure in bayesian networks. In *Proc. Seventeenth International Joint Conference on Artificial Intelligence*, pp. 863–869.

Zweig, G. and S. J. Russell. Speech recognition with dynamic Bayesian networks.