Building Bayesian Influence Ontologies
Literature Review

Tamlin Love
1438243

March 26, 2019

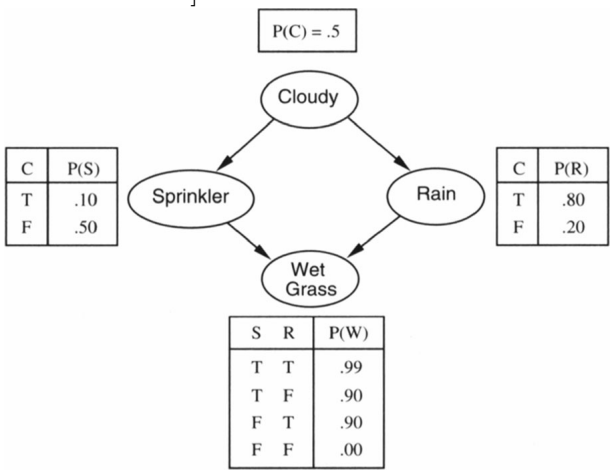
1 Introduction

In section 2, we discuss the concept of a Bayesian network and the properties of this type of model. In section 3, we discuss methods by which Bayesian networks are constructed given a set of observations. We briefly discuss constraint-based methods in section 3.1 before discussing score-based methods in greater depth in section 3.2. In section 4, we discuss parameter estimation techniques in the case of partially observed data. Finally, methods for the construction of networks of independently learned models are discussed in section 5.

2 Bayesian Networks

When considering a joint probability distribution across n random variables, classical probability states that the number of parameters needed to represent the distribution grows exponentially in n [Koller and Friedman 2009]. Even in the simple case of binary variables, we would still need $2^n - 1$ parameters to describe the distribution. This is clearly unfeasible for practical applications, in which the number of random variables can grow very large.
Bayesian networks, originally developed by Pearl [1988], present a way of reducing the number of parameters needed to represent a joint distribution. A Bayesian network is a directed acyclic graph (DAG) whose nodes represent random variables and whose edges represent influence of one variable on another. This structure can also be thought of as a representation of the conditional independencies between the random variables [Koller and Friedman 2009]. Indeed, it is through the exploitation of these independency assumptions that a Bayesian network can more compactly represent a joint distribution.

Figure 1: A famous example of a Bayesian network, showing how a complete representation of any random variable X requires considering only those variables who are parents of X in the graphical representation [Norvig and Russell 1994].



An important notion in Bayesian networks is that of d-separation, first presented by Pearl [1986], which is used to find the set $\mathcal{I}(\mathcal{G})$ of conditional independencies in the graph $\mathcal{G}$. $\mathcal{I}(\mathcal{G})$ is used as the basis for an equivalence relation, I-equivalence, for which any two I-equivalent graphs represent the same independency assumptions [Verma and Pearl 1991]. An important developement by Pearl [1986] is that any I-equivalence class can be represented as a partially directed acyclic graph (PDAG) in which undirected edges represent edges that can be oriented any way and still result in a graph belonging to the same class.

3 Structure Learning

The manual construction of networks is generally unfeasible for a large number of variables [Koller and Friedman 2009]. Fortunately, strategies exist to learn model structures from data $\mathcal{D}$.

3.1 Constraint-Based Structure Learning

One approach to the construction of model structures is the constraint-based approach, in which dependencies between variables are first queried and then, based on these dependencies, a PDAG is constructed [Koller and Friedman 2009]. This strategy can be traced back to Verma and Pearl [1991].
However, this approach is generally not preferred, as failure in the individual independence queries can lead to the construction of a network which poorly matches the data [Koller and Friedman 2009].

3.2 Score-Based Structure Learning

A more popular approach to the problem is score-based structure learning, in which entire networks are constructed and then evaluated and modified based on some scoring metric [Koller and Friedman 2009]. Two areas of interest in this approach are the choice of scoring function and the method of structure search.

3.2.1 Scoring Function

One possible scoring function would be the maximum likelihood function (most often in its logarithm form), finding graph $\mathcal{G}$ that maximises

$$score_L(\mathcal{G} : \mathcal{D}) = l(\hat{\theta}_{\mathcal{G}} : \mathcal{D})$$

which decomposes to

$$score_L(\mathcal{G} : \mathcal{D}) = M \sum_{i=1}^{n} [\mathbb{I}_{\hat{P}}(X_i; Pa_{X_i}^{\mathcal{G}}) - H_{\hat{P}}(X_i)]$$

for number of variables n, number of samples M, mutual information $\mathbb{I}_{\hat{P}}$ and entropy $H_{\hat{P}}$ [Koller and Friedman 2009]. However, this score always prefers a more connected network, and is thus prone to overfitting.
Other scores designed to balance fit to data with network complexity are the Akaike Information Criterion (AIC), proposed by Akaike [1998]

$$score_{AIC}(\mathcal{G} : \mathcal{D}) = l(\hat{\theta}_{\mathcal{G}} : \mathcal{D}) - Dim(\mathcal{G})$$

and the Bayesian Information Criterion (BIC), proposed by Schwarz [1978]

$$score_{BIC}(\mathcal{G} : \mathcal{D}) = l(\hat{\theta}_{\mathcal{G}} : \mathcal{D}) - \frac{logM}{2} Dim(\mathcal{G})$$

where $Dim(\mathcal{G})$ denotes the dimension of $\mathcal{G}$. In particular, Schwarz [1978] shows that the BIC is an assymptotic approximation of the Bayesian score under the assumptions of independent, identically distributed observations with a density function of the form

$$f(x, \theta) = exp(\theta \cdot y(x) - b(\theta))$$

where y is the sufficient statistic, and where it is also assumed that the penalty for guessing an incorrect model is fixed. The $\frac{logM}{2}$ term in the BIC ensures that, as M grows, more consideration is placed in models of greater complexity [Koller and Friedman 2009].

3.2.2 Structure Search

The problem of structure search is to find the graph $\mathcal{G}$ that maximises the chosen scoring function for the given data $\mathcal{D}$. In general, this problem is NP-hard for a graph whose variables have at most $d \geq 2$ parents [Chickering 1996]. Fortunately, there exist heuristic algorithms which can find assist in this regard. Some of the earliest of these algorithms include the K2 algorithm of Cooper and Herskovits [1992], which relied on a predetermined ordering of variables, and the local search algorithms proposed by Heckerman *et al.* [1995].
These algorithms define a search space of graphs, where each graph can be transformed into another by a set of operators [Koller and Friedman 2009]. These operators commonly include edge addition, edge deletion and edge reversal.
A search procedure is then required to traverse the search space and select an optimal graph. A common choice is the greedy hill-climbing algorithm, which applies only the operations which maximise the score [Koller and Friedman 2009]. This technique is prone to local maxima and the plateaus in score caused by I-equivalent graphs. Methods which work around this problem include the tabu search, proposed by Glover [1986], which keeps track of recent operations and does not allow them to be reversed until a certain number of iterations has passed, and the method of random restarts, which restarts the search with random initial conditions [Koller and Friedman 2009].

4 Partially Observed Data

The methods discussed in section 3 require data that are fully observed, in that every datum assigns a value to each variable [Koller and Friedman 2009]. In practice, this is not always possible. Sometimes, these latent or hidden variables are simply impractical to measure. In other cases, these variables represent factors that cannot be measured quantitatively.
Unfortunately, the introduction of hidden variables complicates the structure learning process by introducing a number of terms to the maximum likelihood estimate, with the number of terms growing exponentially as more variables are hidden [Koller and Friedman 2009]. Furthermore, the introduction of latent variables results in a likelihood function which is not locally decomposable, and this renders the techniques discussed in section 3 useless in general. Thus different methods must be applied for both the learning of parameters and the learning of structure. For the sake of relevance, we will discuss only the problem of parameter estimation, as the problem of structure learning is beyond the scope of this paper.
Here we discuss two popular methods for parameter estimation. The first is gradient ascent, in which parameters are chosen by iteratively moving in the direction of the gradient of the likelihood function [Koller and Friedman 2009]. Early applications of gradient methods to the likelihood function include the work of Thiesson [1995] and Binder *et al.* [1997]. In these algorithms, the gradient of the likelihood function is given as

$$\frac{\partial l(\boldsymbol{\theta} : \mathcal{D})}{\partial P(x|\boldsymbol{u})} = \frac{1}{P(x|\boldsymbol{u})} \sum_{m=1}^{M} P(x, \boldsymbol{u}|\boldsymbol{o}[m], \boldsymbol{\theta})$$

where $\boldsymbol{\theta}$ are the parameters, $\mathcal{D} = \{\boldsymbol{o}[1], ..., \boldsymbol{o}[M]\}$ is the set of partially observed data, $x \in Val(X)$ for a variable $X \in \mathcal{X}$ and $\boldsymbol{u} \in Val(Pa_X)$ [Koller and Friedman 2009].
The second method to be discussed is Expectation Maximisation (EM), introduced by Dempster *et al.* [1977] as a generalisation of several earlier methods such as those presented by Baum *et al.* [1970] and Orchard and Woodbury [1972]. EM consists of two steps: Expectation (E-step) and Maximisation (M-step), which are repeated until convergence, starting from some initial $\boldsymbol{\theta}^0$ [Koller and Friedman 2009].
In the E-step, we compute the expected sufficient statistic for each $x \in X$ and $u \in U$

$$\bar{M}_{\boldsymbol{\theta}^t}[x, \boldsymbol{u}] = \sum_{m=1}^{M} P(x, \boldsymbol{u}|\boldsymbol{o}[m], \boldsymbol{\theta}^t)$$

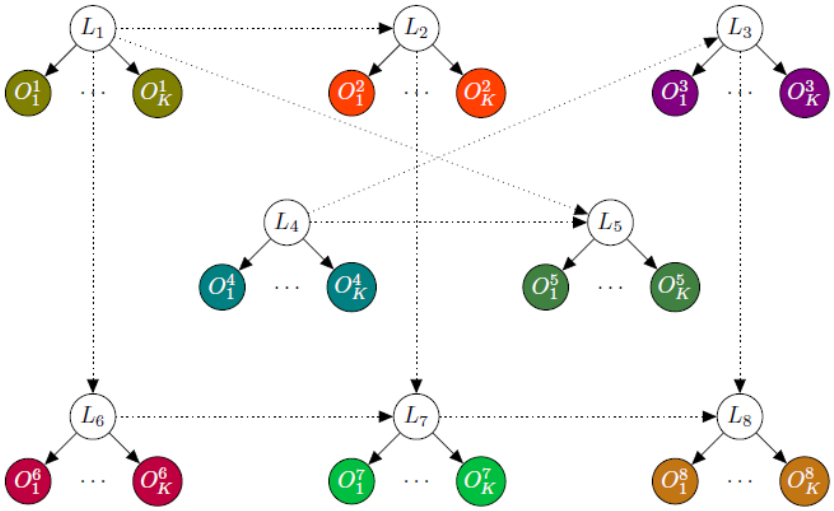
where $\boldsymbol{\theta}^t$ denotes the value of $\boldsymbol{\theta}$ at iteration t.
In the M-step, the $\bar{M}$ is treated as the observed sufficient statistic and is used to calculate $\boldsymbol{\theta}^{t+1}$ using maximum likelihood estimation.
An important property of EM is that it is guaranteed to improve $l(\boldsymbol{\theta^t} : \mathcal{D})$ monotonically as t increases [Koller and Friedman 2009].

5 Influence between Models

We now turn to the specific problem of tracking influence between models. A number of different approaches have been developed to suit different models. For example, Pan *et al.* [2005] use Jeffrey's rule [Pearl 1990] to propagate beliefs on variables from one Bayesian network to another and thus map concepts between ontologies.
However, the methods that are most relevant to this paper are those presented by Ajoodha and

Rosman [2017]. In this work, the authors track influence between a set of naïve Bayes models (NBMs). They do so by first partitioning the observable data into k sets and learning each of the k NBMs independently through the use of EM. They then compute the score of the overarching network (see figure 5) using the BIC, relearn model parameters for the new independence assertions using EM, use the search operators discussed in section 3.2.2 to try and improve the network's score, and then repeat this process until no more improvement can be made to the score. This process is thus a greedy hill-climbing, although in their final implementation, the authors also made use of tabu lists and random restarts to avoid the pitfalls discussed in section 3.2.2.

Figure 2: A simple example of an influence network between a set of independently learned NBMs, where each latent variable L_i is learned from a set of observations $\{O_1^i, ..., O_K^i\}$ and is related to other latent variables via the high-level influence network [Ajoodha and Rosman 2017]



Interesting extensions to this work include the work of Ajoodha and Rosman [2018], who extend the above process to temporal models. In essence, they replace NBMs with hidden Markov models (HMMs) in order to model stochastic processes. They then use very similar techniques to construct a delayed dynamic influence network that models the high-level influence between the HMMs.

6 Conclusion

References

[Ajoodha and Rosman 2017] Ritesh Ajoodha and Benjamin Rosman. Tracking influence between naïve bayes models using score-based structure learning. In *2017 Pattern Recognition Association of South Africa and Robotics and Mechatronics (PRASA-RobMech)*. IEEE, November 2017.

[Ajoodha and Rosman 2018] Ritesh Ajoodha and Benjamin Rosman. Learning the influence structure between partially observed stochastic processes using iot sensor data. In *Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence*. AAAI Publications, 2018.

[Akaike 1998] Hirotogu Akaike. *Information Theory and an Extension of the Maximum Likelihood Principle*, pages 199–213. Springer New York, New York, NY, 1998.

[Baum *et al.* 1970] Leonard E. Baum, Ted Petrie, George Soules, and Norman Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The Annals of Mathematical Statistics*, 41(1):164–171, 1970.

[Binder *et al.* 1997] John Binder, Daphne Koller, Stuart Russell, and Keiji Kanazawa. Adaptive probabilistic networks with hidden variables. *Machine Learning*, 29(2):213–244, Nov 1997.

[Chickering 1996] David Maxwell Chickering. *Learning Bayesian Networks is NP-Complete*, pages 121–130. Springer New York, New York, NY, 1996.

[Cooper and Herskovits 1992] Gregory F. Cooper and Edward Herskovits. A bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9(4):309–347, Oct 1992.

[Dempster *et al.* 1977] Arthur Dempster, Nan Laird, and Donald Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[Glover 1986] Fred Glover. Future paths for integer programming and links to artificial intelligence. *Computers & Operations Research*, 13(5):533 – 549, 1986. Applications of Integer Programming.

[Heckerman *et al.* 1995] David Heckerman, Dan Geiger, and David M. Chickering. Learning bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20(3):197–243, Sep 1995.

[Koller and Friedman 2009] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models - Principles and Techniques*. MIT Press, 2009.

[Norvig and Russell 1994] Peter Norvig and Stuart J. Russell. *Artificial Intelligence: A Modern Approach*. Prentice Hall, December 1994.

[Orchard and Woodbury 1972] Terence Orchard and Max A. Woodbury. A missing information principle: theory and applications. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Theory of Statistics*, pages 697–715, Berkeley, Calif., 1972. University of California Press.

[Pan *et al.* 2005] Rong Pan, Zhongli Ding, Yang Yu, and Yun Peng. A Bayesian network approach to ontology mapping. In *The Semantic Web – ISWC 2005*, pages 563–577. Springer Berlin Heidelberg, 2005.

[Pearl 1986] Judea Pearl. Fusion, propagation, and structuring in belief networks. *Artificial Intelligence*, 29(3), September 1986.

[Pearl 1988] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

[Pearl 1990] Judea Pearl. *Jeffrey's Rule, Passage of Experience, and Neo-Bayesianism*, pages 245–265. Springer Netherlands, Dordrecht, 1990.

[Schwarz 1978] Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 03 1978.

[Thiesson 1995] Bo Thiesson. *Accelerated quantification of Bayesian networks with incomplete data*. Aalborg Universitetsforlag, 1995.

[Verma and Pearl 1991] Thomas Verma and Judea Pearl. Equivalence and synthesis of causal models. In *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, UAI '90, pages 255–270, New York, NY, USA, 1991. Elsevier Science Inc.