Building Bayesian Influence Ontologies
Literature Review

Tamlin Love
1438243

March 15, 2019

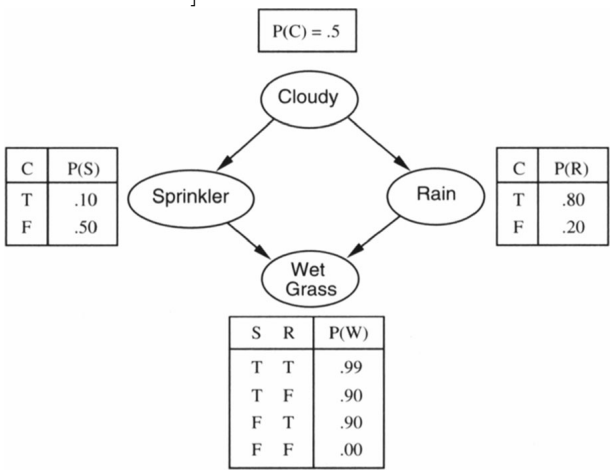
1 Introduction

2 Similarity Metrics

3 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.

4 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 $\mathcal{D}$.

4.1 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].

4.2 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.

4.2.1 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, greater importance is placed on reducing the dimensionality of the model [Koller and Friedman 2009].

4.2.2 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].

5 Bayesian Similarity

References

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

[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.

[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.

[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.

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

[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.