Graphical models: Structure learning
Hauptseminar Machine learning, WS 13/14

Moritz Hamann MORITZ.HAMANN@GMX.COM
University of Stuttgart

Abstract

In this paper we present the main ideas of Heckerman's paper *Graphical models: Structure learning* (Heckerman, 2002). He formulates a Bayesian approach to learn the structure of a Bayesian network and introduces a the so called BDe-score, with can be employed by heuristic search methods, to find the most likely structure for a set of observed samples. In addition he addresses the problem of finding a prior distribution for the unknown parameters of the Bayesian network.

1. Introduction

A Bayesian network, is a common model to represent the conditional dependencies between random variables. If the dependencies and probability distributions are known, the network can be used to interfere the probability of new observed data samples. While the probability distributions can easily derived from an observed data set, the structure of the network has either be determined by the user with knowledge about the problem domain, or also be estimated from the observed data.

To learn the structure, Heckerman proposed in (Heckerman et al., 1995) a Bayesian approach. In this approach the posterior probability of a model under the observed data is computed, in order to find the model, which maximizes this probability. While several other researchers suggested a similar method (Buntine, 1991; Cooper & Herskovits, 1992; Spiegelhalter et al., 1993), Heckerman first examined the prior distribution of the probability parameters, and showed that certain assumptions about the model structure and the observed data lead to a Dirichlet distribution. Furthermore, he introduced the BDe-score, which allows an efficient assessment of the model structure.

Hauptseminar Machine learning, WS 13/14

2. Basics

The next section gives a brief overview about the concept of Bayesian networks, which are a probabilistic graphical model, as well as some important properties of the Dirichlet probability distribution. We assume a certain knowledge about basic probability theory of the user, and refer to one of the many books on this domain for more detail.

2.1. Bayesian network

A Bayesian network (sometimes also called a Bayes or belief network) is a probabilistic graphical model which encodes the conditional dependencies between a set of random variables $X = \{X_1, ..., X_n\}$. Such a network is a Directed Acyclic Graph (DAG), which nodes represent the variables in X, and the edges describe the conditional dependencies between these random variables. Therefore, each node in the Bayesian network can be seen as a conditional probability distribution of the random variable X_i under its parents Pa_i. The traditional notation in Bayesian probability theory for this conditional distribution would be $P(X_i|Pa_i)$.

2.2. Dirichlet probability distribution

The Dirichlet distribution is a multivariate continuous probability distribution, which depends on a hyperparameter vector α with positive entries. It is defined as

$$Dir(x_1, ..., x_m|\alpha_1, ..., \alpha_m) = \frac{1}{B(\alpha)} \prod_{i=1}^{m} x_i^{\alpha_i - 1}$$

where $\sum_i x_i = 1$, $x_i > 0$ and

$$B(\alpha) = \frac{\prod_{i=1}^{m} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{m} \alpha_i)}$$

with the Gamma function $\Gamma(x)$. Since $B(\alpha)$ is the multinomial extension of the Beta function, the Dirichlet distribution can be seen as the multivariate generalization of the Beta distribution.

The components of the hyperparameter α directly correspond to the components of x determine the expected val-

ues for x. In the special case of $\alpha_i = 1\ \forall i$, the Dirichlet distribution is actually the multidimensional uniform distribution. The variance is given by

$$Var[x_i] = \frac{\alpha_i(\alpha_* - \alpha_i)}{\alpha_*^2(\alpha_* + 1)}$$

where $\alpha_* = \sum_i \alpha_i$ which is in context of Bayesian probability theory sometimes called the *equivalent sample size*. Increasing this equivalent sample size decreases the variance of x.

The Dirichlet distribution is also the conjugated prior of the multinomial distribution. Therefore, it is often used in Bayesian probability theory to model the belief $P(\mu)$ about the parameter of a multinomial distribution $Multi(x|\mu)$ for a discrete random variable x. This yields the benefit of a simple calculation of the posterior $P(\mu|x)$. If D is a dataset of observations about x, the posterior $P(\mu|D)$ is given by

$$P(\mu|D) \propto Dir(\mu|\alpha_1 + N_1, ..., \alpha_k + N_k)$$

where α_i are the components of the hyperparameter from the prior distribution and N_i the number of occurrences of state x_i in the dataset D. For detailed information about the Dirichlet distribution, as well as the calculation of the posterior, we refer to (Heckerman et al., 1995).

3. Structure learning

In order to learn the structure - in the further text also called model - of a Bayesian network from an observed dataset $D = \{d_1, ..., d_N\}$ where d_i is a full observation of X, Heckerman (2002) proposed a Bayesian approach and introduced the random variable m. It has the states $m_1, ...m_M$ which correspond to the possible models of a Bayesian network for the set of random variables X.

3.1. Assumption

As a simplification, Heckerman considered discrete random variables for X, which means every X_i has a finite number of states r_i. We use the notation x_i^k if the random variable X_i is in state k with $k = 1...r_i$. Since every X_i has a finite number of parents Pa_i, there exists a finite amount of possible combinations of the parents states $q_i = \prod_{X_m \in Pa_i} r_m$. We denote a specific configuration j of Pa_i with pa_i^j and $j = 1...q_i$.

In addition to discrete random variables, the author also assumed that the state of each X_i is multinomial distributed. That means for every possible parent state pa_i^j of X_i, there exists a parameter vector θ_{ij} with r_i components, such that $P(x_i|pa_i^j) = Multi(x_i|\theta_{ij})$.

This simplifies the construction and inference in the Bayesian network, because the probability distribution for each node can now be stored as a conditional probability table (CPT). In this CPT exists a parameter vector θ_{ij} for every random variable and every possible parent state combination. To denote the probability that X_i is in state k with parent state pa_i^j, we use the notation θ_{ijk}. Since the states of X_i are multinomial distributed it is clear that

$$\sum_{k=1}^{r_i} \theta_{ijk} = 1$$

In the following sections we refer to the full set of parameters for a specific model m as θ^m.

3.2. Bayesian approach

In order to find the optimal model m for an observation D, one has to maximize the posterior of m under D. Using the Bayes' rules this yields

$$P(m|D) = \frac{P(D|m)P(m)}{P(D)} = \frac{P(D|m)P(m)}{\sum_m P(D|m)P(m)} \quad (1)$$

for the posterior of m. Similar, one can compute the posterior for the parameter set θ^m dependent on the observed data

$$P(\theta^m|D, m) = \frac{P(D|\theta^m, m)P(\theta^m|m)}{P(D|m)} \quad (2)$$

In both equations it is necessary to compute the likelihood of the dataset D under a specific model m. Heckerman refers to it as the *marginal likelihood*, which is given as an integral over all possible values for θ^m

$$P(D|m) = \int P(D|\theta^m, m)P(\theta^m|m)d\theta^m \quad (3)$$

Before going into detail how to calculate the marginal likelihood, or how to choose the model and parameter priors, we want to focus on the benefits of the Bayesian approach as pointed out by the author. In contrast to other methods [[find references]], which learn only the most probable model, the Bayesian approach yields a probability distribution over all possible models. This allows a comparison of the probability between different models or the selection of models which have a similar probability than the best.

Another important benefit is the ability to determine the probability of a hypothesis, i.e. the likelihood of a new data sample d_{N+1}, over all possible models instead the most likely one. The probability of the new data sample is then

$$P(d_{N+1}|D) = \sum_m P(m|D) \int P(d_{N+1}|\theta^m, m)P(\theta^m|m)d\theta^m$$

The author calls this a *full* Bayesian approach, since the probability is determined as an average over all possible models. Unfortunately, the number of possible models in

a DAG with n nodes grows super exponentially with n. Therefore, the averaging over all possible models is impractical and one often chooses a fixed number of the most likely models and pretend that these are exhaustive.

3.3. Model prior

The most simple choice for the model prior $P(m)$ is a uniform distribution. This represents the belief that no information about the model structure is available and thus every model is same likely. If some informations about the problem domain are available, the number of models can be reduced by excluding specific models or model families (e.g. if some random variables cant have parents or children). This is achieved by setting the prior $P(m)$ for these model to zero and assume an uniform distribution over the remaining models.

An other possibility for the choice of the model prior mentioned by the authors, is given by (Buntine, 1991). In this case the prior distribution can be computed under the assumption that the random variables can be ordered (e.g. through time precedence). For detailed information we refer to the original paper.

3.4. Parameter prior

Another important choice is the prior distribution for the parameters $P(\theta^m|m)$. To simplify the computation the authors assumed parameter independence, which means that the joint probability distribution can be computed with

$$P(\theta^m|m) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} P(\theta_{ij}|m)$$

The parameter independence also holds for the posterior $P(\theta_{ij}|D, m)$, which means that each θ_{ij} can be updated individually.

As mentioned before, a common choice in Bayesian probability theory for unknown parameter distributions is to use the conjugated prior distribution of the likelihood. Since the author assumed a multinomial distribution for X_i, the likelihood $P(D|\theta_{ij}, m)$ is also multinomial distributed, and hence the conjugated prior would be a Dirichlet distribution

$$P(\theta_{ij}|m) = Dir(\theta_{ij}|\alpha)$$

with $\alpha_i > 0$.

An important contribution of the author is the proof that certain assumptions actually imply a Dirichlet distribution of the parameter prior $P(\theta^m|m)$. The following section briefly presents three key concepts and assumptions of that proof. The complete proof, as well as detailed information on these assumptions, is given in (Heckerman et al., 1995). [[bessere formulierung]]

Markov equivalence: Two models m_1 and m_2 for a set of random variables X are called *markov equivalent*, if they encode the same conditional independence relation of X. For example, in the case of $X = \{X_1, X_2, X_3\}$, the models $X_1 \rightarrow X_2 \rightarrow X_3$, $X_1 \leftarrow X_2 \leftarrow X_3$ and $X_1 \leftarrow X_2 \rightarrow X_3$ are markov equivalent, since they all encode that X_1 and X_3 are independent, given X_2. As shown by the author, the set of complete models for X is also markov equivalent. A complete model is a DAG in which every X_i has either an incoming or outgoing edge to every other X_j.

Distribution equivalence: Two models m_1 and m_2 are called *distribution equivalent* (with respect to some distribution family $\mathcal{F}$), if they represent the same joint probability distribution. This is the case if for every θ^{m_1} exists a θ^{m_2} so that $P(X|\theta^{m_1}, m_1) = P(X|\theta^{m_2}, m_2)$. Distribution equivalence is closely related to markov equivalence, and in fact, distribution equivalence implies markov equivalence, where the opposite may not hold.

Parameter modularity: The assumption of *parameter modularity* implies that $P(\theta_{ij}|m_1) = P(\theta_{ij}|m_2)$ if the random variable X_i has the same parents in the model m_1 and m_2.

Furthermore, the author argued that if two models m_1, m_2 are distribution equivalent, it is unlikely that data can discriminate them. Therefore they assumed $P(D|m_1) = P(D|m_2)$ and called this property *likelihood equivalence*.

Under the assumption of likelihood equivalence, and parameter independence, the authors were able to show that the parameters of every complete model have to be Dirichlet distributed. The hyperparameter α for the prior distribution is restricted and can be computed with

$$\alpha_{ijk} = \alpha_* \cdot P(x_i^k, pa_i^j|m_c)$$

where $P(x_i^k, pa_i^j|m_c)$ is computed from the observed data D and is the joint probability of X_i in state k with parent configuration j for an arbitrary complete model m_c. The author called α_* the *equivalent sample size* which can be chosen by the user. Increasing α_* means increasing the sum $\sum_{k=1}^{r_i} \alpha_{ijk}$, which decreases the variance. Therefore, the equivalent sample size can be interpreted as a level of confidence about the prior distribution.

The joint probability $P(x_i, pa_i^j|m_c)$ is the same for every complete model. Therefore, under the assumption of parameter modularity, the prior distribution for a non-complete model m_{NC} can be computed. One simply constructs a complete network, in which the random variable X_i has the same parents as in m_{NC}, and computes the joint probability in this complete network. Parameter modularity then implies that $P(\theta_{ij}|m_c) = P(\theta_{ij}|m_{NC})$.

3.5. Computation of the marginal likelihood

As shown in section 3 the computation of the model posterior distribution $P(m|D)$ depends on the model prior $P(m)$, which was addressed in section 3.3 as well as the marginal likelihood $P(D|m)$, given in equation (3). The marginal likelihood itself, depends on the parameter prior, which was addressed in section 3.4 and was shown to be a Dirichlet distribution with hyperparameter α, and on the likelihood of the data D under these parameters.

Although both can easily evaluated, the computation of $P(D|m)$ yields the difficulty of integrating over all possible values for θ^m. Since θ^m is a set of parameter vectors, the integral is actually an integration over all θ_{ijk}. Even for small networks, this results in a highly dimensional integration, which is computational dificult. But under the assumptions of a Dirichlet prior distribution, multinomial distributed data, and parameter independence, Cooper and Herskovits (1992) were able to derive a closed form expression for the integral, which is sometimes called a *Dirichlet multinomial compound distribution*. The closed form expression for the marginal likelihood is

$$P(D|m) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + N_{ij})} \cdot \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})}$$

where α_{ijk} is the k-th component of the hyperparameter in the prior distribution for θ_{ij}, and $\alpha_{ij} = \sum_{k=1}^{r_i} a_{ijk}$. Furthermore, N_{ijk} is the number of times the random variable X_i was observed in state k with parent configuration j and $N_{ij} = \sum_{k=1}^{r_i}$.

4. Search methods

Section 3 addressed the question how to find the posterior probability distribution of the network structure. Assumption of a Dirichlet prior distribution for the parameters and multinomial distributed random variables allowed the efficient closed form solution for the marginal likelihood, and therefore an easy computation of the posterior probability of a model m.

A remaining problem is the number of models. As mentioned earlier, the amount of possible models for a set of n random variables grows super exponentially with n. As an example, table 1 shows the number of models up until $n = 6$. If the goal is to find the most likely model instead of the whole probability distribution, a heuristic search algorithm can be used to address this problem. Heckerman suggested a score method, which allows an efficient assessment of the model, together with the greedy hill climbing algorithm to find a model that maximizes this score.

Table 1. Number of possible DAG structures for a network with n nodes

N	1	2	3	4	5	6
NO. OF MODELS	1	3	25	543	29281	3781503

4.1. Score methods

In order to use a heuristic search method, a function has to be defined, which maps a graph G together with the observed data D to a real valued number. In the reminder of this section we will refer to this function as the score $S(m, D)$ of a model. Naturally, this score should assess the model according to the model posterior defined in section 3.

An other preferable property of the score function is the [[factorable score]] This is, if the score function for the model can be divided into the product of local scores for each random variable

$$S(m, D) = \prod_{i=1}^{n} S(X_i, Pa_i, d)$$

where d is the data D restricted to X_i and its parents. This has the benefit that for local changes in the model, the score can be updated locally without the computation of the score for the whole model. A possible score is given with the numerator of the posterior distribution $P(m|D)$, where the marginal likelihood is calculated with the closed form expression (eqn. (3)). In the literature this score is called the *BDe-score* (Bayesian Dirichlet equivalent) (Heckerman, 2002).

4.2. Greedy hill climbing

The authors suggested the greedy hill climbing algorithm, as s simple heuristic search method to maximize the BDe-score. Starting from a initial model, this algorithm tries to improve the model score in each iteration by making small local changes to the model. In specific, it considers every possible arc change, meaning addition, deletion or reversing the direction. Due to the factorable score, the differences in the score for a single arc change, can be computed efficiently from the local score of the two involved random variables.

A problem with greedy hill climbing, as well as many other heuristic algorithms is the convergence to a local maxima. This means the algorithm terminates with a non-optimal model. One possible solution is the utilization of a set of different initial models, hoping at least one reaches the

global maxima. Another possible approach is the method of random restarts. After a (possibly local) maxima is found by greedy hill climbing, this method makes random changes in the found model. After that, it restart the search method on the perturbed model, in order to escape a local maxima.

5. Evaluation results

Unfortunately the original paper did not provide any evaluation result of the BDe score or a comparison with other structure learning approaches. The authors presented a small example in which they examined the dependencies of factors, which influence the intention of US students to attend college. Since only five factors were considered (sex, socioeconomic status, IQ, parental encouragement and college plans) they were able to perform a full search over all possible models, instead of using a heuristic search method. The found model showed a high posterior probability over a wide range for the choice of the equivalent sample size. But since no ground truth is available, a qualitative assessment of the example is impossible.

Heckermann also co-authored (Heckerman et al., 1994) where the BDe-score was tested in a 36 node alarm network. From a ground truth network, they sampled a database with 100 samples, and used these to learn the network. They showed, that the quality of the learned model depends on the equivalent sample size α_*.

Silander et al. (2007) and Liu et al. (2012) evaluated the *BDeu-score*, which is a variant of the BDe-score. In the BDeu-score the hyperparameter α_{ijk} do not depend on the joint probability, and are distributed equally with

$$\alpha_{ijk} = \alpha_* \frac{1}{q_i r_i}$$

Their findings suggests that the quality of the learned model with BDeu-score also depends on the choice of the equivalent sample size α_*. If α_* is too low, the learned network may be too sparse, if α_* is too high, it may be too dense.

We are not sure, if the findings for the BDeu-score can be transferred to the BDe-score, since the parameter prior distribution is different, which can have an impact on the learned model. But we assume that similar to the BDeu-score the BDe-score also depends heavily on the equivalent sample size, which is difficult to determine without knowledge about the problem domain or the dataset.

6. Conclusion

In this paper we presented the Bayesian approach by Heckermann (Heckerman, 2002) for structure learning of Bayesian networks. This approach not only yields the most likely model, but also complete probability distribution over all models. This probability distribution can be used to calculate the likelihood of a new observation averaged over all models.

Two main problems were addressed by Heckermann, which are the prior distribution for the parameters as well as a score method to efficiently employ heuristic search methods. It can be shown, that the assumption of parameter independence and likelihood equivalence implies a Dirichlet distribution for the parameter prior, which depends on the so called equivalent sample size. Furthermore, the closed form solution for the marginal likelihood, together with the model prior, can be used by heuristic search methods to efficiently find an optimal model. The author called this score the BDe-score.

Unfortunately, no performance comparison or evaluation result of the presented method was given by the author. But a recent overview about different structure learning approaches, including *BDeu-score, Minimum description length, Akaike's information criterion* and *factorized normalized maximum likelihood*, are given by Lie et al. in [ref Bio], which suggest that the performance depends on the right choice of the equivalent sample size, as well on the problem domain.

This paper (Heckerman, 2002) as well as the related (Heckerman et al., 1995) had a great impact on the topic of structure learning. On the one hand, the introduced BDe/BDeu-score is today still seen as one of the standard methods of structure learning (Liu et al., 2012). On the other hand the paper serves as a complete and detailed introduction to the bayesian approach to structure learning.

References

Buntine, Wray L. Theory refinement on bayesian networks. In *UAI*, pp. 52–60, 1991.

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

Heckerman. Graphical models: Structure learning. In *The Handbook of Brain Theory and Neural Networks (2nd edition)*. MIT Press, 2002.

Heckerman, David, Geiger, Dan, and Chickering, David Maxwell. Learning bayesian networks: The combination of knowledge and statistical data. In *UAI*, pp. 293–301, 1994.

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

Liu, Zhifa, Malone, Brandon, and Yuan, Changhe. Empirical evaluation of scoring functions for bayesian network model selection. *BMC Bioinformatics*, 13(15):1–16, 2012.

Parr, Ronald and van der Gaag, Linda C. (eds.). *UAI 2007, Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence, Vancouver, BC, Canada, July 19-22, 2007*, 2007. AUAI Press. ISBN 0-9749039-3-0.

Spiegelhalter, David J., Dawid, A. Philip, Lauritzen, Steffen L., and Cowell, Robert G. Bayesian analysis in expert systems. *Statistical Science*, 8(3):219–247, 08 1993. doi: 10.1214/ss/1177010888. URL http://dx.doi.org/10.1214/ss/1177010888.