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

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

Abstract

Super cool abstract

1. Introduction

The goal of this paper is to present the main ideas of [ref], which describes[?] a Bayesian approach for structure learning of Bayesian networks. Furthermore, we'll show the contribution of the author to the relevant field, as well as provide additional experimental results, which we conducted on our own.

warum structure learning - bayesian approach - vorteile fuer sample likelihood

2. Related research

anderes paper gleicher author - neue papers researchen

3. Basics

In this chapter, we present the basics

3.1. 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 (RV) $X = \{X_1, ..., X_n\}$. Such a network is a Directed Acyclic Graph (DAG), which nodes represent the RV, and the edges describe the conditional dependencies between these RV. 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. This would result in $P(X_i|Pa_i)$. ?????

Figure[ref] shows a simple Bayesian network with three binary random variables, and the CPT for X_3.

Hauptseminar Machine learning, WS 13/14

3.2. Dirichlet probability distribution

The Dirichlet distribution is a multivariate continuous probability distribution, which depend on a 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. [[Note that equal a is uniform distribution and what ai stands for, and equivalent sample size + ref]]

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 x corresponds to a dataset D which contains Therefore, the posterior $P(\mu|D)$ is also Dirichlet distributed and is given by

$$P(\mu|D) \propto Dir(\mu|\alpha_1 + n_1, ..., \alpha_k + n_k)$$

where α_i is from the prior and n_i the number of occurences in D... For detailed information about the Dirichlet distribution, as well as the calculation of the posterior, we refer to[ref].

4. Structure learning

In order to learn the model of a Bayesian network from an observed dataset $D = \{d_1, ..., d_N\}$ where d_i is a full observation of X, the authors of [ref] 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.

4.1. Assumption

The authors of [ref] 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 for 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 authors also assumed that the state of X_i with a specific parent state combination pa_i^j is multinomial distributed with a parameter vector θ_{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 for a state k of X_i 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 as θ^m for a specific.

4.2. 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

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

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

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

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

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

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 authors. 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 on only 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 call these 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.

4.3. 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 information about the problem domain are available, the search space 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, as mentioned by the authors, is given by Buntine [ref]. 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 [ref].

4.4. 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 authors 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 the Dirichlet distribution

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

with $\alpha_i > 0$.

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

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 authors, 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. [[brauchen wir das?]]

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. [[data instead of x?]]

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. [[genauer?]]

The authors 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. Together we the assumption of parameter modularity, this implies a Dirichlet distribution even for non-complete models. [[man kann immer complete model construieren in dem parents wie im incomplete sind]]

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

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

4.5. Computation of the marginal likelihood

As shown in section [ref] the computation of the model posterior distribution $P(m|D)$ depends on the model prior $P(m)$, which was addressed in section [ref] as well as the marginal likelihood $P(D|m)$, given in equation [ref]. The marginal likelihood itself, depends on the parameter prior, which was addressed in [ref] 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 consists of a lot bal bla bla, which results in a high dimensional integration. But under the assumptions of a Dirichlet prior distribution, multinomial distributed data, and parameter independence, Cooper and Herskovits [ref] 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 results in

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

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 $N_{ij} = \sum_{k=1}^{r_i}$.

5. Heuristics

asdf

5.1. Score methods

5.2. Greedy hill climbing

5.3. Simulated annealing

6. evaluation results

7. relevance

8. conclusion

You can use footnotes[1] to provide readers with additional information about a topic without interrupting the flow of the paper. Indicate footnotes with a number in the text where the point is most relevant. Place the footnote in 9 point type at the bottom of the column in which it appears. Precede the first footnote in a column with a horizontal rule of 0.8 inches.[2]

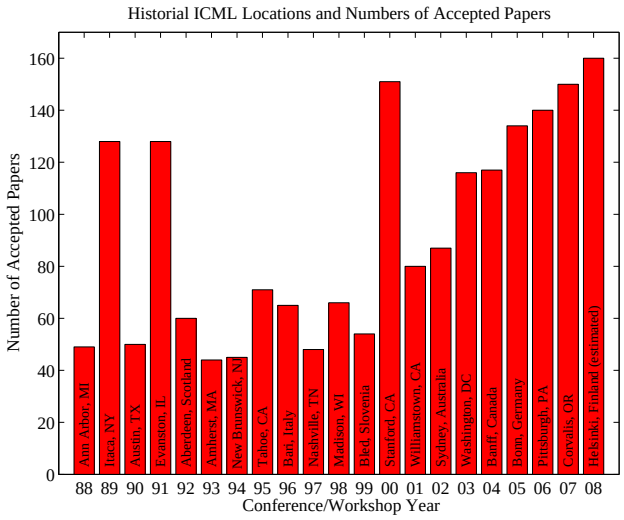

Figure 1. Historical locations and number of accepted papers for International Machine Learning Conferences (ICML 1993 – ICML 2008) and International Workshops on Machine Learning (ML 1988 – ML 1992). At the time this figure was produced, the number of accepted papers for ICML 2008 was unknown and instead estimated.

8.1. Figures

You may want to include figures in the paper to help readers visualize your approach and your results. Such artwork should be centered, legible, and separated from the text. Lines should be dark and at least 0.5 points thick for purposes of reproduction, and text should not appear on a gray background.

Label all distinct components of each figure. If the figure takes the form of a graph, then give a name for each axis and include a legend that briefly describes each curve. Do not include a title inside the figure; instead, the caption should serve this function.

Number figures sequentially, placing the figure number and caption *after* the graphics, with at least 0.1 inches of space before the caption and 0.1 inches after it, as in Figure 1. The figure caption should be set in 9 point type and centered unless it runs two or more lines, in which case it should be flush left. You may float figures to the top or bottom of a column, and you may set wide figures across both columns (use the environment `figure*` in LaTeX), but always place two-column figures at the top or bottom of the page.

Algorithm 1 Bubble Sort

Input: data x_i, size m
repeat
  Initialize $noChange = true$.
  for $i = 1$ **to** $m - 1$ **do**
    if $x_i > x_{i+1}$ **then**
      Swap x_i and x_{i+1}
      $noChange = false$
    end if
  end for
until $noChange$ is $true$

8.2. Algorithms

If you are using LaTeX, please use the "algorithm" and "algorithmic" environments to format pseudocode. These require the corresponding stylefiles, algorithm.sty and algorithmic.sty, which are supplied with this package. Algorithm 1 shows an example.

8.3. Tables

You may also want to include tables that summarize material. Like figures, these should be centered, legible, and numbered consecutively. However, place the title *above* the table with at least 0.1 inches of space before the title and the same after it, as in Table 2. The table title should be set in 9 point type and centered unless it runs two or more lines, in which case it should be flush left.

8.4. Citations and References

Please use APA reference format regardless of your formatter or word processor. If you rely on the LaTeX bibliographic facility, use `natbib.sty` and `icml2014.bst`

[1] For the sake of readability, footnotes should be complete sentences.

[2] Multiple footnotes can appear in each column, in the same order as they appear in the text, but spread them across columns and pages if possible.

Table 2. Classification accuracies for naive Bayes and flexible Bayes on various data sets.

DATA SET	NAIVE	FLEXIBLE	BETTER?
BREAST	95.9± 0.2	96.7± 0.2	√
CLEVELAND	83.3± 0.6	80.0± 0.6	×
GLASS2	61.9± 1.4	83.8± 0.7	√
CREDIT	74.8± 0.5	78.3± 0.6	
HORSE	73.3± 0.9	69.7± 1.0	×
META	67.1± 0.6	76.5± 0.5	√
PIMA	75.1± 0.6	73.9± 0.5	
VEHICLE	44.9± 0.6	61.5± 0.4	√

included in the style-file package to obtain this format.

Citations within the text should include the authors' last names and year. If the authors' names are included in the sentence, place only the year in parentheses, for example when referencing Arthur Samuel's pioneering work (**?**). Otherwise place the entire reference in parentheses with the authors and year separated by a comma (**?**). List multiple references separated by semicolons (**???**). Use the 'et al.' construct only for citations with three or more authors or after listing all authors to a publication in an earlier reference (**?**).

Authors should cite their own work in the third person in the initial version of their paper submitted for blind review. Please refer to Section **??** for detailed instructions on how to cite your own papers.

Use an unnumbered first-level section heading for the references, and use a hanging indent style, with the first line of the reference flush against the left margin and subsequent lines indented by 10 points. The references at the end of this document give examples for journal articles (**?**), conference publications (**?**), book chapters (**?**), books (**?**), edited volumes (**?**), technical reports (**?**), and dissertations (**?**).

Alphabetize references by the surnames of the first authors, with single author entries preceding multiple author entries. Order references for the same authors by year of publication, with the earliest first. Make sure that each reference includes all relevant information (e.g., page numbers).