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.

Proceedings of the 31st International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

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 shows two key concepts [[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. A special set of markov equivalent models, is the set of complete models for a set of random variables X. In a complete model, every X_i has either an incoming or outgoing edge to every other X_j.

Distribution equivalence: bla bla bla

Dirchichlet distribution for complete model

Parameter modularity: bla bla bla

4.5. Computation of the marginal likelihood

As seen in the previous section, the model prior is

closed loop evaluation

5. Heuristics

6. evaluation results

7. relevance

8. conclusion

9. Electronic Submission

As in the past few years, ICML will rely exclusively on electronic formats for submission and review.

9.1. Templates for Papers

Electronic templates for producing papers for submission are available for LaTeX . Templates are accessible on the World Wide Web at:
http://icml.cc/2014/

Send questions about these electronic templates to `program@icml.cc`.

The formatting instructions below will be enforced for initial submissions and camera-ready copies.

- The maximum paper length is 8 pages excluding references, and 9 pages including references.
- Do not alter the style template; in particular, do not compress the paper format by reducing the vertical spaces.
- Do not include author information or acknowledgments in your initial submission.
- Place figure captions *under* the figure (and omit titles from inside the graphic file itself). Place table captions *over* the table.
- References must include page numbers whenever possible and be as complete as possible. Place multiple citations in chronological order.

Please see below for details on each of these items.

9.2. Submitting Papers

Submission to ICML 2014 will be entirely electronic, via a web site (not email). The URL and information about the submission process are available on the conference web site at

http://icml.cc/2014/

Paper Deadline: The deadline for paper submission to ICML 2014 is at 23:59 Universal Time (3:59 Pacific Daylight Time) on the due dates (October 4, or January 31, depending on the review cycle). If your full submission does not reach us by this time, it will not be considered for publication. There is no separate abstract submission.

Anonymous Submission: To facilitate blind review, no identifying author information should appear on the title page or in the paper itself. Section 10.3 will explain the details of how to format this.

Simultaneous Submission: ICML will not accept any paper which, at the time of submission, is under review for another conference or has already been published. This policy also applies to papers that overlap substantially in technical content with conference papers under review or previously published. ICML submissions must not be submitted to other conferences during ICML's review period. Authors may submit to ICML substantially different versions of journal papers that are currently under review by the journal, but not yet accepted at the time of submission. Informal publications, such as technical reports or papers in workshop proceedings which do not appear in print, do not fall under these restrictions.

To ensure our ability to print submissions, authors must provide their manuscripts in **PDF** format. Furthermore, please make sure that files contain only Type-1 fonts (e.g., using the program `pdffonts` in linux or using File/DocumentProperties/Fonts in Acrobat). Other fonts (like Type-3) might come from graphics files imported into the document.

Authors using **Word** must convert their document to PDF. Most of the latest versions of Word have the facility to do this automatically. Submissions will not be accepted in Word format or any format other than PDF. Really. We're not joking. Don't send Word.

Those who use **LaTeX** to format their accepted papers need to pay close attention to the typefaces used. Specifically, when producing the PDF by first converting the dvi output of LaTeX to Postscript the default behavior is to use non-scalable Type-3 PostScript bitmap fonts to represent the standard LaTeX fonts. The resulting document is difficult to read in electronic form; the type appears fuzzy. To avoid this problem, dvips must be instructed to use an alternative font map. This can be achieved with something like the following commands:

dvips -Ppdf -tletter -G0 -o paper.ps paper.dvi
ps2pdf paper.ps

Note that it is a zero following the "-G". This tells dvips to use the config.pdf file (and this file refers to a better font mapping).

Another alternative is to use the **pdflatex** program instead of straight LaTeX. This program avoids the Type-3 font problem, however you must ensure that all of the fonts are embedded (use `pdffonts`). If they are not, you need to configure pdflatex to use a font map file that specifies that the fonts be embedded. Also you should ensure that images are not downsampled or otherwise compressed in a lossy way.

Note that the 2014 style files use the `hyperref` package to make clickable links in documents. If this causes problems for you, add `nohyperref` as one of the options to the `icml2014` usepackage statement.

9.3. Reacting to Reviews

We will continue the ICML tradition in which the authors are given the option of providing a short reaction to the initial reviews. These reactions will be taken into account in the discussion among the reviewers and area chairs.

9.4. Submitting Final Camera-Ready Copy

The final versions of papers accepted for publication should follow the same format and naming convention as initial submissions, except of course that the normal author information (names and affiliations) should be given. See Section 10.3.2 for details of how to format this.

The footnote, "Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute." must be modified to "*Proceedings of the 31^{st} International Conference on Machine Learning*, Beijing, China, 2014. JMLR: W&CP volume 32. Copyright 2014 by the author(s)."

For those using the **LaTeX** style file, simply change `\usepackage{icml2014}` to

```
\usepackage[accepted]{icml2014}
```

Authors using **Word** must edit the footnote on the first page of the document themselves.

Camera-ready copies should have the title of the paper as running head on each page except the first one. The running title consists of a single line centered above a horizontal rule which is 1 point thick. The running head should be centered, bold and in 9 point type. The rule should be 10 points above the main text. For those using the **LaTeX** style file, the original title is automatically set as running head using the `fancyhdr` package which is included in the ICML 2014 style file package. In case that the original title exceeds the size restrictions, a shorter form can be supplied by using

```
\icmltitlerunning{...}
```

just before `\begin{document}`. Authors using **Word** must edit the header of the document themselves.

10. Format of the Paper

All submissions must follow the same format to ensure the printer can reproduce them without problems and to let readers more easily find the information that they desire.

10.1. Length and Dimensions

Papers must not exceed eight (8) pages, including all figures, tables, and appendices, but excluding references. When references are included, the paper must not exceed nine (9) pages. Any submission that exceeds this page limit or that diverges significantly from the format specified herein will be rejected without review.

The text of the paper should be formatted in two columns, with an overall width of 6.75 inches, height of 9.0 inches, and 0.25 inches between the columns. The left margin should be 0.75 inches and the top margin 1.0 inch

(2.54 cm). The right and bottom margins will depend on whether you print on US letter or A4 paper, but all final versions must be produced for US letter size.

The paper body should be set in 10 point type with a vertical spacing of 11 points. Please use Times typeface throughout the text.

10.2. Title

The paper title should be set in 14 point bold type and centered between two horizontal rules that are 1 point thick, with 1.0 inch between the top rule and the top edge of the page. Capitalize the first letter of content words and put the rest of the title in lower case.

10.3. Author Information for Submission

To facilitate blind review, author information must not appear. If you are using LaTeX and the `icml2014.sty` file, you may use `\icmlauthor{...}` to specify authors. The author information will simply not be printed until `accepted` is an argument to the style file. Submissions that include the author information will not be reviewed.

10.3.1. Self-Citations

If your are citing published papers for which you are an author, refer to yourself in the third person. In particular, do not use phrases that reveal your identity (e.g., "in previous work (**?**), we have shown ...").

Do not anonymize citations in the reference section by removing or blacking out author names. The only exception are manuscripts that are not yet published (e.g. under submission). If you choose to refer to such unpublished manuscripts (**?**), anonymized copies have to be submitted as Supplementary Material via CMT. However, keep in mind that an ICML paper should be self contained and should contain sufficient detail for the reviewers to evaluate the work. In particular, reviewers are not required to look a the Supplementary Material when writing their review.

10.3.2. Camera-Ready Author Information

If a paper is accepted, a final camera-ready copy must be prepared. For camera-ready papers, author information should start 0.3 inches below the bottom rule surrounding the title. The authors' names should appear in 10 point bold type, electronic mail addresses in 10 point small capitals, and physical addresses in ordinary 10 point type. Each author's name should be flush left, whereas the email address should be flush right on the same line. The author's physical address should appear flush left on the ensuing line, on a single line if possible. If successive authors have the same affiliation, then give their physical address only once.

A sample file (in PDF) with author names is included in the ICML2014 style file package.

10.4. Abstract

The paper abstract should begin in the left column, 0.4 inches below the final address. The heading 'Abstract' should be centered, bold, and in 11 point type. The abstract body should use 10 point type, with a vertical spacing of 11 points, and should be indented 0.25 inches more than normal on left-hand and right-hand margins. Insert 0.4 inches of blank space after the body. Keep your abstract brief and self-contained, limiting it to one paragraph and no more than six or seven sentences.

10.5. Partitioning the Text

You should organize your paper into sections and paragraphs to help readers place a structure on the material and understand its contributions.

10.5.1. Sections and Subsections

Section headings should be numbered, flush left, and set in 11 pt bold type with the content words capitalized. Leave 0.25 inches of space before the heading and 0.15 inches after the heading.

Similarly, subsection headings should be numbered, flush left, and set in 10 pt bold type with the content words capitalized. Leave 0.2 inches of space before the heading and 0.13 inches afterward.

Finally, subsubsection headings should be numbered, flush left, and set in 10 pt small caps with the content words capitalized. Leave 0.18 inches of space before the heading and 0.1 inches after the heading.

Please use no more than three levels of headings.

10.5.2. Paragraphs and Footnotes

Within each section or subsection, you should further partition the paper into paragraphs. Do not indent the first line of a given paragraph, but insert a blank line between succeeding ones.

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

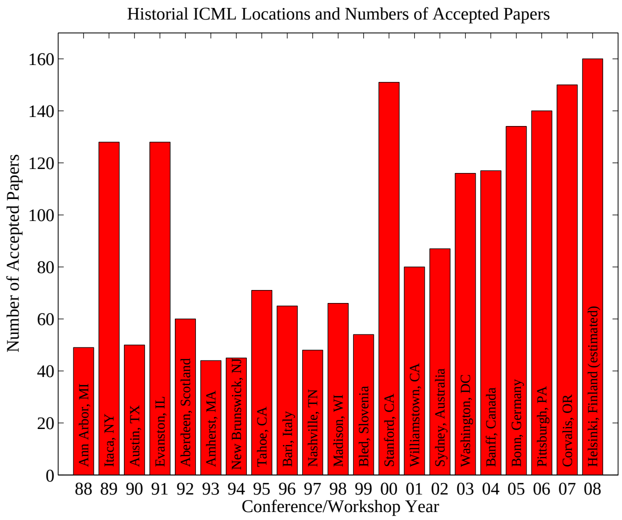

Figure 1. Historical locations and number of accepted papers for International Machine Learning Conferences (ICML 1993 – ICML 2008) and International Workshops on Machine Learning (ML 1988 – ML 1992). At the time this figure was produced, the number of accepted papers for ICML 2008 was unknown and instead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

of 0.8 inches.[2]

10.6. 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.

[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.

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 $true$

Table 1. 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	√

10.7. 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.

10.8. 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 1. 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.

Tables contain textual material that can be typeset, as contrasted with figures, which contain graphical material that must be drawn. Specify the contents of each row and column in the table's topmost row. Again, you may float tables to a column's top or bottom, and set wide tables across both columns, but place two-column tables at the top or bottom of the page.

10.9. 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` 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 10.3 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).

10.10. Software and Data

We strongly encourage the publication of software and data with the camera-ready version of the paper whenever appropriate. This can be done by including a URL in the camera-ready copy. However, do not include URLs that reveal your institution or identity in your submission for review. Instead, provide an anonymous URL or upload the material as "Supplementary Material" into the CMT reviewing system. Note that reviewers are not required to look a this material when writing their review.

Acknowledgments

Do not include acknowledgements in the initial version of the paper submitted for blind review.

If a paper is accepted, the final camera-ready version can (and probably should) include acknowledgements. In this case, please place such acknowledgements in an unnumbered section at the end of the paper. Typically, this will include thanks to reviewers who gave useful comments, to colleagues who contributed to the ideas, and to funding agencies and corporate sponsors that provided financial support.