Proportional Graphical Models
Part 1: From Decision Making under uncertainty to MCMC

Andreas Holzinger
185.A83 Machine Learning for Health Informatics
2017S, VU, 2.0h, 3.0 ECTS
Module 02 - Week 13

http://hci.kdd.org/machine-learning-for-health-informatics-course

Red thread through the lecture today

- 01 Decision Making under uncertainty
- 02 Graphs – Networks
- 03 Example Medical Knowledge Representation
- 04 Graphical Models and Decision Making
- 05 Bayes Networks
- 06 Graphical Model Learning
- 07 Probabilistic Programming
- 08 Markov Chain Monte Carlo (MCMC)
- 09 Metropolis Hastings Algorithm

Decision trees are coming from Clinical Practice

Who is Who?

ML needs a concerted effort fostering integrated research

Interactive Data Mining

Knowledge Discovery

Privacy, Data Protection, Safety and Security

Machine Learning Jungle Top-Level View

Pre-Knowledge Quiz: Which concepts can you identify?
Temporal Decision Making under Uncertainty


Recursive reasoning: a case for probabilistic programming


What are Probabilistic Graphical Models?

- PGM can be seen as a combination between
- Graph Theory + Probability Theory + Machine Learning
- One of the most exciting advancements in AI in the last decades - with enormous future potential
- Compact representation for exponentially-large probability distributions
- Example Question: "Is there a path connecting two proteins?"
- Path \( X, Y \) := edge \( X, Y \)
- Path \( X, Y \) := edge \( X, Y \), path \( X, Y \)
- This can NOT be expressed in first-order logic
- Need a Turing-complete fully-fledged language

Key Challenges

- Medicine is an extremely complex application domain - dealing most of the time with uncertainties -> probable information!
- Key: Structure learning and prediction in large-scale biomedical networks with probabilistic graphical models
- Causality and Probabilistic Inference
- Uncertainties are present at all levels in health related systems
- Data sets from which ML learns are noisy, mislabeled, atypical, etc. etc.
- Even with data of high quality, groupings and a multitude of data sources and constraints in usually imperfect models of the world requires us to represent and process uncertain knowledge in order to make viable decisions in context and within reasonable time!
- In the increasingly complicated settings of modern science, model structure or causal relationships may not be known a-priori (1)
- Approximating probabilistic inference in Bayesian belief networks is NP-hard (2) -> here we need the "human-in-the-loop" (3)

02 Graphs=Networks

275 years later ... the “Nobel-prize in Computer Science”

http://acm.org/prizes/award/5231
http://www.nobelprize.org/nobel_prizes/chemistry/laureates/2013

First Question: Where does graphs come from?

- Graphs as models for networks
  - given as direct input (point cloud data sets)
  - Given as properties of a structure
  - Given as a representation of information (e.g., Facebook data, viral marketing, etc., ...)

- Graphs as nonparametric basis
  - we learn the structure from samples and infer
  - flat vector data, e.g., similarity graphs
  - encoding structural properties (e.g., smoothness, independence, ...)

We skip this interesting chapter for now ...

Our World In Data (1/2) – Macroscopic Structures

03 Network challenges

NGC 5139 Omega Centauri by Edmund Huell in 1977, ESO, Makana, Chile

Time
  - e.g. Entropy

Space
  - e.g. Topology

Complexity Problem: Time versus Space

P versus NP and the Computational Complexity Zoo, please have a look at https://www.complexityzoo.eu/
04 Graphical Models and Decision Making

\[ D \equiv \{ X_1^{(i)}, X_2^{(i)}, \ldots, X_m^{(i)} \}_{i=1}^N \]

---

**Classes of Graphical Models**

Graph models and probabilistic models

**Decision Making:** Learn good policy for selecting actions

**Goal:** Learn an optimal policy for selecting best actions within a given context

1. The world produces a "context" \( x_t \in \mathcal{X} \)
2. The learner selects an action \( a_t \in \{1, \ldots, K\} \)
3. The world reacts with a reward \( r_t(a_t) \in [0,1] \)

---

**Three types of Probabilistic Graphical Models**

- **Undirected:** Markov random fields, useful e.g. for computer vision (Details: Murphy 19)
  \[
  P(X) = \frac{1}{Z} \exp \left( \sum_{ij} w_{ij} x_i x_j + \sum_i x_i \right)
  \]

- **Directed:** Bayes Nets, useful for designing models (Details: Murphy 10)
  \[
  p(x) = \prod_{k=1}^K p(x_k | \mathbf{pa}(x_k))
  \]

- **Factored:** useful for inference/learning
  \[
  p(x) = \prod_k f_k(x_k)
  \]

---

**Factor Graphs – learning at scale**

- **What is the advantage of factor graphs?**

<table>
<thead>
<tr>
<th>Dependency</th>
<th>Efficient Inference</th>
<th>Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bayesian Networks</td>
<td>Yes</td>
<td>Somewhat</td>
</tr>
<tr>
<td>Markov Networks</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Factor Graphs</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Table credit to Ralf Herbrich, Amazon

---

**Remember**

- Medicine is an extremely complex application domain – dealing most of the time with uncertainties -> probabilistic models!
- When we have big data but little knowledge automatic ML can help to gain insight:
- **Structure learning and prediction in large-scale biomedical networks with probabilistic graphical models**
- If we have little data and deal with NP-hard problems we still need the human-in-the-loop

---

**Naive Bayes classifier as DGM (single/nested plates)**

- Key idea: Conditional independence assumptions are very useful – however, Naive Bayes is extreme!
- \( X \) is conditionally independent of \( Y \) given \( Z \), if the P(\( X \)) governing \( X \) is independent of value \( y \), given value of \( Z \):
  \[
  (\forall i, j, k)(P(\mathbf{X} = x_j | Y = y_j, Z = z_k) = P(\mathbf{X} = x_j | Z = z_k))
  \]
  can be abbr. with \( P(\mathbf{X} | Y, Z) = P(\mathbf{X} | Z) \)
- Graphical models express sets of conditional independence assumptions via graph structure
- The graph structure plus associated parameters define joint probability distribution over the set of variables

---

**GM are amongst the most important ML developments**

Image credit to Anna Goldenberg, Toronto
05 Bayesian Networks

"Bayes’ Nets"

Bayesian Network (BN) - Definition

- a probabilistic model, consisting of two parts:
  1) a dependency structure and
  2) local probability models.

\[ p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i | Pa(x_i)) \]

Where \( Pa(x_i) \) are the parents of \( x_i \).

BN inherently model the uncertainty in the data. They are a successful marriage between probability theory and graph theory; allow to model a multidimensional probability distribution in a sparse way by searching independency relations in the data. Furthermore, this model allows different strategies to integrate two data sources.


Example: Directed Bayesian Network with 7 nodes

\[ p(X_1, \ldots, X_7) = p(X_1)p(X_2)p(X_3)p(X_4 | X_1, X_2, X_3)p(X_5 | X_1, X_3)p(X_6 | X_4)p(X_7 | X_4, X_5) \]

Clinical Case Example

Important in clinical practice \rightarrow prognosis

- the prediction of the future course of a disease conditional on the patient's history and a projected treatment strategy
- Danger: probabilistic information!
- Therefore, valid prognostic models can be of great benefit for clinical decision making and of great value to the patient, e.g., for notification and quality of life decisions


Breast cancer - big picture - state of 1999

- Alcohol & Smoking
- Hemorrhea
- Menopause
- Pregnancy
- Family History
- Architectural Distortion
- Nipple Discharge
- Skin Thickening
- Breast Pain
- have a lump
- Mass
- Microcalcifications


10 years later: Integration of microarray data

- Integrating microarray data from multiple studies to increase sample size;
- approach to the development of more robust prognostic tests


Example: BN with four binary variables

- Gene 1
- Gene 2

P(Gene 2 | Gene 1) = 0.6
P(Gene 1 | Gene 2) = 0.7


Dependency Structure \rightarrow first step (1/2)

- First, the structure is learned using a search strategy.
- Since the number of possible structures increases super exponentially with the number of variables,
- the well-known greedy search algorithm K2 can be used in combination with the Bayesian Dirichlet (BD) scoring metric:

\[ p(S|D) \propto p(S) \prod_{i=1}^{n} \frac{\Gamma(N_{ij})}{\Gamma(N_{ij} + N_{ij} + N_{ijk})} \frac{\Gamma(N_{ij} + N_{ijk})}{\Gamma(N_{ijk})} \]

\[ N_{ijk} \text{ - number of cases in the data set } D \text{ having variable } i \text{ in state } k \text{ associated with the } j \text{-th instantiation of its parents in current structure } S. \]

\[ n \text{ is the total number of variables.} \]
Often it is better to have a good solution within time – than an perfect solution (much) later ...

06 Graphical Model Learning

Learning the Structure of GM from data

1) Test if a distribution is decomposable with respect to a given graph.
   - This is the most direct approach. It is not bound to a graphical representation.
   - It can be carried out w.r.t. other representations of the set of subspaces to be used to compute the (candidate) decomposition of a given distribution.

2) Find a suitable graph by measuring the strength of dependencies.
   - This is a heuristic, but often highly successful approach, which is based on the frequently valid assumption that in a conditional independence graph an attribute is more strongly dependent on adjacent attributes than on attributes that are not directly connected to them.

3) Find an independence map by conditional independence tests.
   - This approach exploits the theorems that connect conditional independence and graphs and that represent decompositions.
   - It has the advantage that a single conditional independence test, if it fails, can exclude several candidate graphs. Beware, because wrong test results can thus have huge consequences.

Inference in Bayes Nets is intractable (NP-complete)

- For certain cases it is tractable if:
  - Just one variable is unobserved
  - We have singly connected graphs (no undirected loops) -> belief propagation
  - Assigning probability to fully observed set of variables
- Possibility: Monte Carlo Methods (generate many samples according to the Bayes Net distribution and then count the results)
- Otherwise: approximate solutions, NOTE:
  - Sometimes it is better to have an approximate solution to a complex problem - than a perfect solution to a simplified problem

Learning Graphical Models from data

- Remember: GM are a marriage between probability theory and graph theory and provide a tool for dealing with our two grand challenges in the biomedical domain:
  - Uncertainty and complexity
- The learning task is two-fold:
  1) Learning unknown probabilities
  2) Learning unknown structures


Relational Representation Learning and Prediction

- Asthma can be hereditary
- Friends may have similar smoking habits
- Augmenting graphical model with relations between the entities – Markov Logic

2.1 Asthma ⇒ Cough
2.5 Smokes ⇒ Cough
1.9 Smokes(x) ∧ Friends(y,x) ⇒ Smokes(y)
1.5 Asthma(x) ∧ Family(y,x) ⇒ Asthma(y)

Probabilistic Programming

- C → Probabilistic-C
- Scala → Figaro
- Scheme → Church
- Excel → Tabular
- Prolog → Problog
- Javascript → webPP
- → venture
- Python → PyMC

PyMC: Pythonic Markov chain Monte Carlo

08 Markov Chain Monte Carlo (MCMC)

Monte Carlo Method (MC)
Monte Carlo Sampling
Markov Chains (MC)
MCMC
Metropolis-Hastings

07 Probabilistic Programming

Example for probabilistic rule learning, in which probabilistic rules are learned from probabilistic examples: The ProMIS+ Algorithm learns this pattern by combining the principles of the rule learner C4.5 with the probabilistic Prolog called Prolog, von De Raedt, L., Drax, A., Thiel, L., Van Den Breek, D. & Gestel, T. 2015. Inducing probabilistic relational rules from probabilistic examples. International Joint Conference on Artificial Intelligence (IJCAI).

Medical Example

In real-world systems you have observable data $D$

- often we want to calculate characteristics of a high-dimensional probability distribution ...

$$p(h|d) \propto p(D|h) * p(h)$$

Posterior integration problem: (almost) all statistical inference can be deduced from the posterior distribution by calculating the appropriate sums, which involves an integration.

$$J = \int f(\theta) * p(D|\theta)d\theta$$
**Statistical physics:** computing the partition function – this is evaluating the posterior probability of a hypothesis and this requires summing over all hypotheses... remember:

\[
\mathcal{H} = \{H_1, H_2, ..., H_n\} \quad \forall (h, d)
\]

\[
P(h|d) = \frac{P(d|h) * P(h)}{\sum_{h' \in \mathcal{H}} P(d|h')P(h')}
\]

---

**Summary: What are Monte Carlo methods?**

- Class of algorithms that rely on **repeated random sampling**
- Basic idea: using **randomness** to solve problems with high uncertainty (Laplace, 1781)
- For solving **multidimensional integrals** which would otherwise intractable
- For simulation of systems with **many dof**
  - e.g., fluids, gases, particle collectives, **cellular structures** – see our last tutorial on Tumor growth simulation!

---

**MC connects Computer Science with Cognitive Science**

- for solving problems of probabilistic inference involved in developing computational models
- as a source of hypotheses about how the human mind might solve problems of inference
- For a function \( f(x) \) and distribution \( P(x) \), the expectation of \( f \) with respect to \( P \) is generally the average of \( f \), when \( x \) is drawn from the probability distribution \( P(x) \)

\[
\mathbb{E}_P(f(x)) = \sum_x f(x)P(x)dx
\]

---

**Physical simulation via MC**

- Physical simulation
- estimating neutron diffusion time
- Computing expected utilities and best responses toward Nash equilibria
- Computing volumes in high-dimensions
- Computing eigen-functions and values of operators (e.g., Schrödinger)
- Statistical physics
- Counting many things as fast as possible

---

**Equation of State Calculations by Fast Computing Machines**

*The Monte Carlo Method*

A new algorithm is described for the determination of the equation of state of many-body systems (e.g., liquid, solid, plasma). The method is based on Monte Carlo simulations, which involve random sampling from a large number of possible states to obtain statistical averages over the entire phase space. This approach allows for the calculation of thermodynamic properties without the need for solving complex differential equations. The method is particularly useful for systems with long-range interactions, such as liquid and plasma, where traditional methods are computationally prohibitive.

---

**5,223 citations as of 26.03.2017**

---

**34,140 citations (as of 26.03.2017)**
**09 Metropolis-Hastings Algorithm**

- **Importance sampling**
  - Importance sampling is a technique to approximate averages with respect to an intractable distribution \( p(x) \).
  - The term ‘sampling’ is arguably a misnomer since the method does not attempt to draw samples from \( p(x) \).
  - Rather the method draws samples from a simpler importance distribution \( q(x) \) and then reweights them.
  - such that averages with respect to \( p(x) \) can be approximated using the samples from \( q(x) \).

- **Gibbs Sampling**
  - The Gibbs Sampler is an interesting special case of MH:

**Remember**

- Expectation of a function \( f(x, y) \) with respect to a random variable \( x \) is denoted by \( \mathbb{E}_x[f(x, y)] \).
- In situations where there is no ambiguity as to which variable is being averaged over, this will be simplified by omitting the suffix, for instance \( \mathbb{E}_x \).
- If the distribution of \( x \) is conditioned on another variable \( z \), then the corresponding conditional expectation will be written \( \mathbb{E}_x[f(x) | z] \).
- Similarly, the variance is denoted \( \text{var}[f(x)] \), and for vector variables the covariance is written \( \text{cov}[x, y] \).

**Global optimization: What is the main problem?**

\[
\text{argmax}_x f(x)
\]

- **Normalization:**
  \[
p(x|y) = \frac{p(y|x) * p(x)}{\int_X p(y|x) * p(x) dx}
\]

- **Marginalization:**
  \[
p(x) = \int_Z p(x, z) dz
\]

- **Expectation:**
  \[
\mathbb{E}_p(x)[f(x)] = \int_X f(x) p(x) dx
\]
MCMC based DPFM outperforms other approaches

Still ... there are a lot of open problems and challenges to solve ... no chance to retire!

Thank you!

Questions

- What is the main difference between the ideas of Pierre Simon de Laplace and Lady Lovelace?
- What is medical action consiting most of the time?
- How does a human make a decision - as far as we know?
- What is the main idea of a probabilistic programming language?
- Why did Judea Pearl receive the Turing Award (Noble Prize in Computer Science)?
- What fieds are coming together in PGM?
- What are the challenges in network structures?
- Give a classification of Graphical Models!
- What are plates and nested plates?
- Provide corresponding examples of metabolic networks!
Appendix


Where do the data come from?


http://www.ebi.ac.uk/intact/