MACHINE LEARNING TECHNIQUES FOR STRUCTURED AND UNSTRUCTURED DATA FUSION

LJK - PhD students half day

Thibaud Rahier, first year PhD student

December 17, 2015

PHD supervisors:
Florence Forbes and Stephane Girard (MISTIS team at INRIA)
Sylvain Marié (Analytics for solutions team at Schneider Electric)
Research project presentation

Overview of Probabilistic Graphical Models

Next steps
Structured data

- Time series from housing data (power consumption, temperature...)
- Metadata surrounding these time series (location and type of measurement device, category of physical quantity measured...)

Unstructured data

- Diverse web information sources (tweets, weather, satellite pictures...)
- Expert (human) knowledge on the field
Industrial applications:

**Regression**

- Virtual sensing
- Forecasting
- Key performance indicator (KPI) prediction

**Classification**

- Fault detection (one class classification), diagnosis (multiclass classification)
- Predictive maintenance, predictive diagnosis

**Clustering**

- e.g. Typical days profiling
CURRENT EXPLORATION: PROBABILISTIC GRAPHICAL MODELS

Technical advantages of PGMs

- Broad category of models (Hidden Markov Models, Markov Random Fields, Bayesian Networks, Deep Belief Networks...)
- Easy handling of missing data / heterogeneous data (metadata)
- Several ways to incorporate human knowledge
- Generative models: same model can be used for diverse applications

A very active research field

- Recent reference books:
  - Probabilistic Graphical Models (Koller and Friedman, 2012)
  - A Probabilistic Approach on Machine Learning (Murphy, 2012)
- Benchmarks keep getting better in several recent papers
- Programming libraries are currently being developed
OVERVIEW OF PROBABILISTIC GRAPHICAL MODELS
THEORETICAL FRAMEWORK AND VOCABULARY

Data
$M$ instances of r.v. $(X_1, \ldots, X_n)$

Model (PGM)
directed: BN
undirected: MRF

Knowledge
from field expert

Application
marginal estimation
value prediction

Learning
ELICITATION

Inference
Inference

- **Marginal inference**: For $i \in \{1, \ldots, n\}$, evaluate $P(X_i)$.
- **MAP inference**: For $i \in \{1, \ldots, n\}$, evaluate $\arg\max_{x_i \in \text{Val}(X_i)} P(X_i)$

Learning

- **Parameter learning** (for a given structure): maximize likelihood
- **Structure learning**:
  - structure scoring: likelihood-based
  - structure exploration (optimisation in super-exponential space)

Knowledge incorporation

- **Priors in Bayesian approach** for both learning tasks
- **Starting points proposals** for heuristic algorithms (particularly interesting in super-exponential structure spaces)
NEXT STEPS
AIMS FOR THE NEXT WEEKS:

Programming:
Define a simple application to be run on a real dataset

Bibliographic exploration:
Sum-product networks (sometime inference is possible in linear time), cutting-edge variational methods

Short term goal:
Build a PGM that incorporates expert knowledge and metadata, and that competes with Schneider Electric’s current benchmark algorithms (SVM...) on a given problem (KPIs prediction?)
Thank you
Two main categories of graphical models:

**Directed (Bayesian networks)** Joint distribution decomposes in a product of local conditional distributions:

\[
P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | Pa_{X_i})
\]

Ex: Dynamic Bayesian networks, Kalman filters, Hidden Markov models

**Undirected (Markov random fields)** Joint distribution decomposes in a product of local potentials:

\[
P(X_1, \ldots, X_n) = \frac{1}{Z} \prod_{c} \phi_c(X_c)
\]

Ex: Ising model, Deep Belief Networks

We often use the Gibbs representation: for each clique c, \( \phi_c(x_c) = \exp(-E_c(x_c)) \) where \( E_c \) is the energy function of clique c
What is inference

What we want to infer

- Marginal inference. For \( i \in \{1, \ldots, n\} \), evaluate

\[
P(X_i) \\
P(X_i \mid E = e)
\]

- MAP inference For \( i \in \{1, \ldots, n\} \), evaluate

\[
\arg\max_{x_i \in \text{Val}(X_i)} P(x_i) \\
\arg\max_{x_i \in \text{Val}(X_i)} P(x_i \mid E = e)
\]

Complexity

For both directed and undirected graphs: \textbf{NP-Hard} problem in general: number of computations is exponential in number of variables
Different inference algorithms:

**Exact**
- Var Elimination
- Junction tree

**Approximate**
- Stochastic
  - MCMC (Gibbs, MH)

**Deterministic**
- Belief propagation

**Variational**
- Mean field
Log-likelihood function

- Bayesian networks

$$l(\theta) = \sum_{m=1}^{M} \sum_{i=1}^{n} \log (P(X_i = x_i [m] | Pa_{X_i} = u_i [m]))$$

- Markov random fields

$$l(\theta) = \sum_{m=1}^{M} \sum_{c} \log (\phi_c(x_c [m])) - M \log(Z(\theta))$$
**Table:** Maximum likelihood estimation in 4 main situations

<table>
<thead>
<tr>
<th>Complete data</th>
<th>Incomplete data</th>
</tr>
</thead>
<tbody>
<tr>
<td>unimodal likelihood</td>
<td>multimodal likelihood</td>
</tr>
<tr>
<td><strong>BNs</strong></td>
<td><strong>Specific algorithms</strong> (Expectation-Maximization) <strong>inference</strong></td>
</tr>
<tr>
<td>sufficient statistics tractable optimization in closed form</td>
<td>both challenges part. funct. &amp; missing data inference</td>
</tr>
<tr>
<td><strong>MRFs</strong></td>
<td><strong>Inference</strong></td>
</tr>
<tr>
<td>no closed-form optimization (partition function) <strong>inference</strong></td>
<td></td>
</tr>
</tbody>
</table>

16
Priors in Bayesian approach

for both learning tasks (\(\theta\) is the vector of parameters, \(l(\theta)\) the corresponding log-likelihood, \(D\) the data and \(G\) the graph structure)

- Parameter learning: maximum a posteriori:
  \[
  \hat{\theta}^{MAP} = \arg\max_{\theta} P(\theta|D) = \arg\max_{\theta} l(\theta) + \log(P(\theta))
  \]

- Structure learning (Bayesian score):
  \[
  \text{score}_B(G : D) = \log\left( \frac{P(D|G)}{\int P(D|\theta,G)P(\theta|G)d\theta} \right) + \log(P(G))
  \]

Starting points in optimization algorithms

Starting point proposals for heuristic algorithms (particularly interesting in super-exponential structure spaces) can be very useful