Machine Learning for Information Networks

Oliver Schulte
School of Computing Science
Simon Fraser University
Outline

- What are information networks/multi-relational data?
- Why machine learning for information networks?
- Unifying logic and statistics: learning first-order Bayesian networks
- Applications
  - Frequency Modelling/Density Estimation
  - Relational Exception Mining
- How is relational learning different from non-relational learning?
What Are Information Networks?

Representing Relational Data
Definition

An information network (Sun and Han 2012) is a graph with

- **nodes** (aka entities)
- **edges** (aka relationships)
  - can be hyperedges
- Nodes and edges
  - can be of different types ➔ heterogeneity
  - can have attributes (aka features)

Toy Example

gender = Man, country = U.S.
$500,000, runtime = 98 min, country = U.S.

gender = Man, country = U.S.
$5,000,000, runtime = 111 min, country = U.S.

gender = Woman, country = U.S.
$2,000,000
Different Communities Use Different Formats for Information Network


Nodes and edges in heterogenous network
(Sun and Han 2012)

Data Format

graphical

Logical Facts
- Knowledge Graph/ RDFTriples
  (Nickel et al. 2015)
- Literals

logical

tabular

Database Tables
SQL

arrays

Matrices Tensors

Gammatial Facts
- Knowledge Graph/ RDFTriples
  (Nickel et al. 2015)
- Literals
### Table Representation

#### One table for each type of entity/link

#### Actors

<table>
<thead>
<tr>
<th>Name</th>
<th>gender</th>
<th>country</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brad_Pitt</td>
<td>M</td>
<td>U.S.</td>
</tr>
<tr>
<td>Lucy_Liu</td>
<td>W</td>
<td>U.S.</td>
</tr>
<tr>
<td>Steve_Buscemi</td>
<td>M</td>
<td>U.S.</td>
</tr>
<tr>
<td>Uma_Thurman</td>
<td>W</td>
<td>U.S.</td>
</tr>
</tbody>
</table>

#### ActsIn

<table>
<thead>
<tr>
<th>Name</th>
<th>Title</th>
<th>salary (M$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lucy_Liu</td>
<td>Kill_Bill</td>
<td>2</td>
</tr>
<tr>
<td>Steve_Buscemi</td>
<td>Fargo</td>
<td>0.5</td>
</tr>
<tr>
<td>Uma_Thurman</td>
<td>Kill_Bill</td>
<td>5</td>
</tr>
</tbody>
</table>
Plug: The Prague Relational Learning Repository

- 80+ relational databases Repository
- Can search for different dataset properties.
- Write-up and connection details are available http://arxiv.org/abs/1511.03086
Why Machine Learning for Information Networks?
Enterprise Data Are Relational

- Most organizations maintain data in a relational database management system.
- Structured Query Language (SQL) allows fast data retrieval.
  - E.g., find all movie ratings > 4 where the user is a woman.
- Multi-billion dollar industry, $Bn 15+ in 2006.
- IBM, Microsoft, Oracle, SAP, Peoplesoft.
Impedance Mismatch

- Standard machine learning packages (R, SciKit, Weka,..) accept a single data table as input.
- In a database with multiple tables, which table do we input?
- SAP data scientist: “When our customers want to use machine learning, they spend 80% of their time getting the data into the right format”.

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Name</th>
<th>gender</th>
<th>country</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Brad_Pitt</td>
<td>M</td>
<td>U.S.</td>
</tr>
<tr>
<td></td>
<td>Lucy_Liu</td>
<td>W</td>
<td>U.S.</td>
</tr>
<tr>
<td></td>
<td>Steve_Buscemi</td>
<td>M</td>
<td>U.S.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name</th>
<th>Title</th>
<th>salary (M$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lucy_Liu</td>
<td>Kill_Bill</td>
<td>2</td>
</tr>
<tr>
<td>Steve_Buscemi</td>
<td>Fargo</td>
<td>0.5</td>
</tr>
<tr>
<td>Uma_Thurman</td>
<td>Kill_Bill</td>
<td>5</td>
</tr>
</tbody>
</table>
AI Motivation: Expressive Power

- Russell and Norvig: Hierarchy of environment representations
- The more information an agent has about its environment, the better its performance

Logic and Probability

• Russell (UC Berkeley): “Their unification holds enormous promise for AI”
• Domingos (U of Washington): “Logic handles complexity, probability represents uncertainty.”
Unifying Logic and Statistics

Poole, D. (2003), First-order probabilistic inference, 'IJCAI'.
Function Representation

- The attributes and relationships in an information network can mathematically be represented using *functions*, e.g.
  - gender
  - ActsIn
  - salary
Example Function Representation

gender = Man
country = U.S.
False
n/a
runtime = 98 min
drama = true
$500K

gender = Man
country = U.S.
False
n/a
runtime = 111 min
drama = false
$5M

gender = Woman
country = U.S.
False
salary

gender = Woman
country = U.S.
False
salary

ActsIn

runtime = 98 min
drama = true
$500K

runtime = 111 min
drama = false
$5M

runtime = 111 min
drama = false
$2M
First-Order Logic: Terms

- A **constant** refers to an individual
  - “Fargo”
- A **first-order variable** refers to a class of individuals
  - “Movie” refers to Movies

Terms

- A constant or first-order variable is a term.
- The result of applying a function to a term is a term.

contains first-order variables?

**first-order term**
e.g. salary(Actor, Movie)

**ground term**
e.g. salary(UmaThurman, Fargo)

Relational Random Variables

- *First-order random variable* = *First-order term* + probabilistic semantics (Wang et al. 2008)
- Both complex terms and complex random variables are built by function application

<table>
<thead>
<tr>
<th>Statistics</th>
<th>Logic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apply function to random variable(s)</td>
<td>Apply function to term(s)</td>
</tr>
<tr>
<td>➔ new random variable</td>
<td>➔ new term</td>
</tr>
</tbody>
</table>

Formulas

• A (conjunctive) formula is a joint assignment
  \( \text{term}_1 = \text{value}_1, \ldots, \text{term}_n = \text{value}_n \)
  • e.g., ActsIn(Actor, Movie) = T, gender(Actor) = W
• A ground formula contains only constants
  • e.g., ActsIn(UmaThurman, KillBill) = T, gender(UmaThurman) = W
What is a Bayesian network?

Compact representation of joint probability distributions via conditional independence

Qualitative part:
Directed acyclic graph (DAG)
- Nodes - random vars.
- Edges - direct influence

Together:
Define a unique distribution in a factored form

Quantitative part:
Set of conditional probability distributions

\[ P(B,E,A,C,R) = P(B)P(E)P(A | B,E)P(R | E)P(C | A) \]
Why are Bayes nets useful?

- Graph structure supports
- Modular representation of knowledge
- Local, distributed algorithms for inference and learning
- Intuitive (possibly causal) interpretation
- Easy to compute “Is X relevant to Y given Z”.
- [UBC Demo](#)
Bayesian networks for relational data

- A first-order Bayesian network is a Bayesian network whose nodes are first-order terms (Wang et al. 2008)
- AKA parametrized Bayesian network (Poole 2003, Kimmig et al. 2014)

Frequency Semantics for First-Order Bayesian Networks


Random Selection Semantics for First-Order Bayesian Networks

- We can compute joint probabilities from a FOBN, e.g.
  \[ P(\text{gender(Actor)} = W, \text{ActsIn(Actor,Movie)} = T, \text{Drama(Movie)} = F) = 2/8 \]
- But what does this represent?

“if we randomly select an actor and a movie, the probability is 2/8 that the actor appears in the movie, the actor is a woman, and the movie is a drama”
Random Selection Semantics

Population
Actors

Population
(first-order)
variables

First-Order
Random Variables
(Terms)

Actor
Random Selection
from Actors.
P(Actor = brad_pitt) = 1/4

gender(Actor)
Gender of selected actor.
P(gender(Actor) = W) = 1/2

Movie
Random Selection
from Movies.
P(Movie = Fargo) = 1/2

ActsIn(Actor,Movie) = T if selected actor appears in selected movie, F otherwise
P(ActsIn(Actor,Movie) = T) = 3/8

Drama(Movie)
Is the selected movie a drama?
P(Drama(Movie) = T) = 1/2
Real-World Examples

- To illustrate frequency semantics, learn and evaluate on the training set

  - ground truth about frequencies

- We discuss generalization later
IMDb Data Format

data with two relationships

Learning Bayesian Networks for Complex Relational Data
Learned Bayes Net for Full IMDB
Learned Bayes Net for IMDb

With only 1 relationship HasRated(User,Movie).
Bayes Net Query

Query Results

Query Results for P(e) [Action(Movie)=1] [HasRated(User,Movie)=T] [Gender(User)=W]

P(e) = 0.00284

OK
Data Query

<table>
<thead>
<tr>
<th>Num Movies</th>
<th>3883</th>
</tr>
</thead>
<tbody>
<tr>
<td>Num Users</td>
<td>6039</td>
</tr>
<tr>
<td>Num Movie-User Pairs</td>
<td>$3883 \times 6039 = 23449437$</td>
</tr>
</tbody>
</table>

movie-user pairs with action movie, woman user

Action(Movie) = T, HasRated(User,Movie) = T, gender(User) = W

<table>
<thead>
<tr>
<th>Frequency</th>
<th>66642</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\frac{66642}{23449437} = 0.0028$</td>
</tr>
</tbody>
</table>

More Examples in spreadsheet on website

Learning Bayesian Networks for Complex Relational Data
Mondial Data Format

```
Borders
Name1 VARCHAR(35)
Name2 VARCHAR(35)
Indexes

Country
Name VARCHAR(35)
Continent VARCHAR(20)
DiscreteArea VARCHAR(35)
DiscretePopulation VARCHAR(35)
Indexes
```

Learning Bayesian Networks for Complex Relational Data
Learned Bayes Net for Mondial
Bayes Net query

Click on a node to stop or start monitoring its probability.

Learning Bayesian Networks for Complex Relational Data
### Data Query

<table>
<thead>
<tr>
<th>Number of Europe-Europe Borders</th>
<th>156</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of *-Europe Borders</td>
<td>166</td>
</tr>
<tr>
<td>(P(\text{continent}(\text{country1}) = \text{Europe}</td>
<td>\text{Borders}(\text{country1}, \text{country2}) = T, \text{continent}(\text{country2} = \text{Europe})))</td>
</tr>
</tbody>
</table>

- BN was learned with frequency smoothing (Laplace correction)
- More Examples in spreadsheet on tutorial website
Bayesian Networks are Excellent Estimators of Network Frequencies

- Queries Randomly Generated
- Example: $P(\text{gender}(A) = W | \text{ActsIn}(A, M) = \text{true}, \text{Drama}(M) = T)$?
- Learn Bayesian network and test on entire database as in Getoor et al. 2001


Relational Exception Mining

Random Individuals vs. Specific Individuals
Profile-Based Outlier Detection for Relational Data

Population Database
e.g. IMDB

Individual Database
Profile, Interpretation, egonet
e.g. Brad Pitt’s movies

Goal: Identify exceptional individual databases

Example: population data

gender = Man
country = U.S.
False
n/a
runtime = 98 min
drama = true
action = true

gender = Man
country = U.S.
False
n/a
runtime = 111 min
drama = false
action = true

 gender = Woman
country = U.S.
False
n/a

gender = Woman
country = U.S.
False
n/a
salary
$500K

$5M

$2M

 ActsIn
Example: individual data

- **gender**: Man
- **country**: U.S.
- **runtime**: 98 min
- **drama**: True

[Image of Brad Pitt]
Compare Random Individual to Target Individual

Outlierness Metric (quality measure) = Measure of dissimilarity between class and individual BN e.g. KLD, ELD (new)

Example: class and individual Bayesian network parameters

\[ P(\text{gender}(A) = M) = 0.5 \]
\[ P(\text{Drama}(M) = T) = 0.5 \]

<table>
<thead>
<tr>
<th>Gender (A)</th>
<th>Drama (M)</th>
<th>Cond. Prob. of ActsIn(A,M) = T</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>T</td>
<td>1/2</td>
</tr>
<tr>
<td>M</td>
<td>F</td>
<td>0</td>
</tr>
<tr>
<td>W</td>
<td>T</td>
<td>0</td>
</tr>
<tr>
<td>W</td>
<td>F</td>
<td>1</td>
</tr>
</tbody>
</table>

\[ P(\text{gender}(\text{BradPitt}) = M) = 1 \]
\[ P(\text{Drama}(M) = T) = 0.5 \]

<table>
<thead>
<tr>
<th>Gender (BradPitt)</th>
<th>Drama (M)</th>
<th>Cond. Prob. of ActsIn(A,M) = T</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>T</td>
<td>0</td>
</tr>
<tr>
<td>M</td>
<td>F</td>
<td>0</td>
</tr>
</tbody>
</table>
Case Study: Strikers and Movies

Data are from Premier League Season 2011-2012.

<table>
<thead>
<tr>
<th>Player Name</th>
<th>Position</th>
<th>KLD Rank</th>
<th>KLD Max Node</th>
<th>Feature Max Value</th>
<th>Individual Probability</th>
<th>Class Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Edin Dzeko</td>
<td>Striker</td>
<td>1</td>
<td>Dribble Efficiency</td>
<td>DE = Low</td>
<td>0.16</td>
<td>0.50</td>
</tr>
<tr>
<td>Paul Robinson</td>
<td>Goalie</td>
<td>2</td>
<td>SavesMade</td>
<td>SM = Medium</td>
<td>0.30</td>
<td>0.04</td>
</tr>
</tbody>
</table>

Striker = Normal

<table>
<thead>
<tr>
<th>Movie Title</th>
<th>Genre</th>
<th>KLD Rank</th>
<th>KLD Max Node</th>
<th>Feature Max Value</th>
<th>Individual Probability</th>
<th>Class Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brave Heart</td>
<td>Drama</td>
<td>1</td>
<td>Actor_Quality</td>
<td>a_quality=4</td>
<td>0.93</td>
<td>0.42</td>
</tr>
<tr>
<td>Austin Powers</td>
<td>Comedy</td>
<td>2</td>
<td>Cast_position</td>
<td>cast_num=3</td>
<td>0.78</td>
<td>0.49</td>
</tr>
<tr>
<td>Blue Brothers</td>
<td>Comedy</td>
<td>3</td>
<td>Cast_position</td>
<td>cast_num=3</td>
<td>0.88</td>
<td>0.49</td>
</tr>
</tbody>
</table>
How is Relational Learning Different From IID Learning?

Challenges and Solutions
IID Data vs. Relational Data

Traditional Data Matrix represents independent and identically distributed data points (i.i.d.)

- special case of relational data with 0 relationships
- unary functors

```
gender = Man
country = U.S.
```
```
gender = Man
country = U.S.
```
```
gender = Woman
country = U.S.
```
```
gender = Woman
country = U.S.
```

Relational Data Are Not Independent

<table>
<thead>
<tr>
<th>Name</th>
<th>Title</th>
<th>Salary (M$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lucy_Liu</td>
<td>Kill_Bill</td>
<td>2</td>
</tr>
<tr>
<td>Uma_Thurman</td>
<td>Kill_Bill</td>
<td>5</td>
</tr>
<tr>
<td>Uma_Thurman</td>
<td>Be_Cool</td>
<td>9</td>
</tr>
</tbody>
</table>

- Uma Thurman’s salary in Kill Bill carries information about her salary in Be Cool
- Also carries information about Lucy Liu’s salary in Kill Bill
Difficulty #1: Likelihood Function

- Most Bayesian network learning methods are based on a **score function**
- Key component: the likelihood function $P(\text{data} | \text{model})$
  1. Measure how likely each datapoint is according to the Bayesian network
  2. **Multiply** datapoint probabilities to define likelihood for whole dataset – assumes independence and single table
Solution #1: The Random Selection Likelihood Score

1. Randomly select a grounding/instantiation for all first-order variables in the first-order Bayesian network
2. Compute the log-likelihood for the attributes of the selected grounding
3. Log-likelihood score = expected log-likelihood for a random grounding

Theoretical Validation #1

- **Proposition** (Schulte 2011) The random selection log-likelihood score is maximized by setting the conditional probabilities to the \textit{frequencies observed in the network}.

- **Theorem** (Xiang and Neville 2011) The random selection log-likelihood score is \textit{consistent} (asymptotically correct).

Distance between correct and maximum-likelihood parameter values

\# of entities
Difficulty #2: No global sample size

• What is the sample size - #Users, #Movies, #Ratings?

• Typical model selection scores are of the form
  \[ \text{score(model, data)} = \log\text{-likelihood(data | model)} - f(\#model parameters, sample size) \]

• e.g. for BIC we have
  \[ f = \frac{\log(N)}{2} \times \#\text{parameters} \]

already discussed  
penalize complex models
Solution #2

- Use local sample sizes = number of possible child-parent instantiations
- When comparing two models, normalize both penalty terms by the larger local sample size.

\[
f = \frac{\log(\text{#Users} \times \text{#Movies})}{2 \times \text{#parameters}} \times \frac{1}{\text{#Users} \times \text{#Movies}}
\]

Schulte, O. & Gholami, S. (2017), Locally Consistent Bayesian Network Scores for Multi-Relational Data, IJCAI 2017
Theoretical Validation #2

- **Theorem** (Schulte and Gholami 2017) If a score is consistent for i.i.d. data, then the normalized score is consistent for relational data:
  - converges to a model of the network frequencies
  - with a minimum number of edges

Distance between network frequencies and FOBN joint probabilities

Schulte, O. & Gholami, S. (2017), Locally Consistent Bayesian Network Scores for Multi-Relational Data, *IJCAI 2017*
Summary: Information Networks

- Heterogeneous information networks are ubiquitous, go by several names:
  - relational database
  - first-order model
  - matrixes/tensors

- Unifying logic and statistics:
  - Relational random variable = first-order term
  - First-order Bayesian network = BN whose nodes are first-order terms
Summary: Applications of FOBNs

- Modelling correlations and frequencies in relational data
  - applies classic random selection semantics for probabilistic logic
- Exception Mining and Anomaly Detection
Summary: Learning Challenges

- Network nodes and links are not independent
- Difficult to define likelihood for entire network
- Solution: apply random selection semantics to define expected log-likelihood from random instances
- There is no global sample size $N$
- Difficult to define model selection score
- Normalize score by (max) local sample size
- Theoretical and extensive empirical validation
There’s More (In Tutorial)

- https://oschulte.github.io/srl-tutorial-slides/
- Scalable Algorithms:
  - for counting relational frequencies
  - for relational model structure search
- Latent variable models for clustering, community detection, matrix factorization, relational deep learning
- Applications:
  - link-based classification
  - link prediction
  - feature extraction
References

- Github https://github.com/sfu-cl-lab
  - Code and names of collaborators (thank you thank you!)
The Bayes Net Likelihood Function for IID data

1. For each row, compute the log-likelihood for the attribute values in the row.
2. Log-likelihood for table = sum of log-likelihoods for rows.

Assumes independence of rows (data points)
## IID Example

<table>
<thead>
<tr>
<th>Title</th>
<th>Drama</th>
<th>Action</th>
<th>Horror</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fargo</td>
<td>T</td>
<td>T</td>
<td>F</td>
</tr>
<tr>
<td>Kill_Bill</td>
<td>F</td>
<td>T</td>
<td>F</td>
</tr>
</tbody>
</table>

### Log-likelihood Calculation

**Fargo**
- \( P(\text{Drama} | \text{Action} = \text{T}) = 1/2 \)
- \( P(\text{Horror} | ...) = 1 \)

**Kill_Bill**
- \( P(\text{Drama} | \text{Action} = \text{T}) = 1/2 \)
- \( P(\text{Horror} | ...) = 1 \)

### Total Log-likelihood Score

<table>
<thead>
<tr>
<th>Title</th>
<th>Drama</th>
<th>Action</th>
<th>Horror</th>
<th>( P_B )</th>
<th>( \ln(P_B) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fargo</td>
<td>T</td>
<td>T</td>
<td>F</td>
<td>1x1/2x1 = 1/2</td>
<td>-0.69</td>
</tr>
<tr>
<td>Kill_Bill</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>1x1/2x1 = 1/2</td>
<td>-0.69</td>
</tr>
</tbody>
</table>

**Total Log-likelihood Score for Table = -1.38**
Theoretical Validation #1

- **Proposition** (Schulte 2011) The random selection log-likelihood score is maximized by setting the conditional probabilities to the frequencies observed in the network.

- **Theorem** (Xiang and Neville 2011) The random selection log-likelihood score is *consistent* (asymptotically correct).

Distance between correct and maximum-likelihood parameter values

# of entities
Likelihood Function for Relational Data
Wanted: a likelihood score for relational data

Log-Likelihood, e.g. -3.5

Problems
- Multiple Tables.
- Dependent data points

Learning Bayesian Networks for Complex Relational Data
### Example

- **gender(A)**
- **ActsIn(A,M)**

#### Database Table:

<table>
<thead>
<tr>
<th>Prob</th>
<th>A</th>
<th>M</th>
<th>gender(A)</th>
<th>ActsIn(A,M)</th>
<th>P_B</th>
<th>ln(P_B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/8</td>
<td>Brad_Pitt</td>
<td>Fargo</td>
<td>M</td>
<td>F</td>
<td>3/8</td>
<td>-0.98</td>
</tr>
<tr>
<td>1/8</td>
<td>Brad_Pitt</td>
<td>Kill_Bill</td>
<td>M</td>
<td>F</td>
<td>3/8</td>
<td>-0.98</td>
</tr>
<tr>
<td>1/8</td>
<td>Lucy_Liu</td>
<td>Fargo</td>
<td>W</td>
<td>F</td>
<td>2/8</td>
<td>-1.39</td>
</tr>
<tr>
<td>1/8</td>
<td>Lucy_Liu</td>
<td>Kill_Bill</td>
<td>W</td>
<td>T</td>
<td>2/8</td>
<td>-1.39</td>
</tr>
<tr>
<td>1/8</td>
<td>Steve_Buscemi</td>
<td>Fargo</td>
<td>M</td>
<td>T</td>
<td>1/8</td>
<td>-2.08</td>
</tr>
<tr>
<td>1/8</td>
<td>Steve_Buscemi</td>
<td>Kill_Bill</td>
<td>M</td>
<td>F</td>
<td>3/8</td>
<td>-0.98</td>
</tr>
<tr>
<td>1/8</td>
<td>Uma_Thurman</td>
<td>Fargo</td>
<td>W</td>
<td>F</td>
<td>2/8</td>
<td>-1.39</td>
</tr>
<tr>
<td>1/8</td>
<td>Uma_Thurman</td>
<td>Kill_Bill</td>
<td>W</td>
<td>T</td>
<td>2/8</td>
<td>-1.39</td>
</tr>
</tbody>
</table>

**Probabilities:****

- $P(g(A)=M) = 1/2$
- $P(ActsIn(A,M)=T|g(A)=M) = 1/4$
- $P(ActsIn(A,M)=T|g(A)=W) = 2/4$
Proposition The random selection log-likelihood score is maximized by setting the Bayesian network parameters to the observed conditional frequencies

\[
\begin{align*}
\text{gender}(A) & \quad \text{P}(g(A)=M) = 1/2 \\
\text{ActsIn}(A,M) & \quad \text{P}(\text{ActsIn}(A,M)=T \mid g(A)=M) = 1/4 \\
& \quad \text{P}(\text{ActsIn}(A,M)=T \mid g(A)=W) = 2/4
\end{align*}
\]