Pseudo-Likelihood for Relational Data

Oliver Schulte
School of Computing Science
Simon Fraser University
Vancouver, Canada

To appear at SIAM SDM conference on data mining.
The Main Topic

- In relational data, units are interdependent
  ⇒ no product likelihood function for model.
- How to do model selection?
- Proposal of this talk: use pseudo likelihood.
  - Unnormalized product likelihood.
  - Like independent-unit likelihood, but with event frequencies instead of event counts.
Overview

- Define pseudo log-likelihood for directed graphical models (Bayes Nets).
- Interpretation as expected log-likelihood of random small groups of units.
- Learning Algorithms:
  - MLE solution.
  - Model Selection.
- Simulations.
Outline

• Brief intro to relational databases.
• Statistics and Relational Databases.
• Briefer intro to Bayes nets.
• Relational Random Variables.
• Relational (pseudo)-likelihoods.
Relational Databases

• 1970s: Computers are spreading. Many organizations use them to store their data.
• Ad hoc formats
  ⇒ hard to build general data management systems.
  ⇒ lots of duplicated effort.
• The Standardization Dilemma:
  • Too restrictive: doesn’t fit users’ needs.
  • Too loose: back to ad-hoc solutions.
The Relational Format

- Codd (IBM Research 1970)
- The fundamental question: *What kinds of information do users need to represent?*
- Answered by 1\textsuperscript{st}-order predicate logic! (Russell, Tarski).
- The world consists of
  - Individuals/entities.
  - Relationships/links among them.
Tabular Representation

- Tables for Entity Types, Relationships.

<table>
<thead>
<tr>
<th>Student</th>
<th>Course</th>
<th>Professor</th>
<th>Registration</th>
</tr>
</thead>
<tbody>
<tr>
<td>s-id</td>
<td>c-id</td>
<td>p-id</td>
<td>s-id</td>
</tr>
<tr>
<td></td>
<td>Rating</td>
<td>Popularity</td>
<td>Salary</td>
</tr>
<tr>
<td>Jack</td>
<td>3</td>
<td>Oliver</td>
<td>High</td>
</tr>
<tr>
<td>Kim</td>
<td>2</td>
<td>Oliver</td>
<td>Low</td>
</tr>
<tr>
<td>Paul</td>
<td>1</td>
<td>Jim</td>
<td>Med</td>
</tr>
</tbody>
</table>

Pseudo-Likelihood for Relational Data - Statistics Seminar
Database Management Systems

- Maintain data in linked tables.
- Structured Query Language (SQL) allows fast data retrieval.
  - E.g., find all SFU students who are statistics majors with gpa > 3.0.
- Multi-billion dollar industry, $15+ bill in 2006.
- IBM, Microsoft, Oracle, SAP, Peoplesoft.
Relational Domain Models

- Visualizing Domain Ontology.
- Active Area of Research.
  - Unified Modelling Language (UML).
  - Semantic Web (XML).
- Classic Tool: The Entity-Relationship (ER) Diagram.
ER Diagram Example

Students
- name
- intelligence
- ranking

Registered
- grade
- satisfaction
- number

Courses
- rating
- difficulty

Professors
- name
- popularity
- teaching ability

Teaches
ER Model for Social Network

**Ring Diagram**
- **Actors**
  - Anna
  - Bob
- **Friend**
  - Smokes = true
  - Cancer = true
  - Smokes = true
  - Cancer = false

**Data Tables**

| Actors | | | |
|---|---|---|
| Name | Smokes | Cancer |
| Anna | T | T |
| Bob | T | F |

<table>
<thead>
<tr>
<th>Friend</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Name1</td>
<td>Name2</td>
<td></td>
</tr>
<tr>
<td>Anna</td>
<td>Bob</td>
<td></td>
</tr>
<tr>
<td>Bob</td>
<td>Anna</td>
<td></td>
</tr>
</tbody>
</table>
A single-relation social network is a simple special case of a relational database.

Converse also true if you allow:
- Different types of nodes ("actors").
- Labels on nodes.
- Different types of (hyper)edges.
- Labels on edges.


Observation A relational database is equivalent to a general network as described.
Outline

☑ Brief intro to relational databases.
  • *Statistics and Relational Databases*.
  • Briefer intro to Bayes nets.
  • Relational Random Variables.
  • Relational (pseudo)-likelihoods.
Beyond storing and retrieving data

- Much new interest in analyzing databases.
  - Data Mining.
  - Data Warehousing.
  - Business Intelligence.
  - Predictive Analytics.

- Fundamental Question: how to combine logic and probability?
  - Domingos (U of W, CS): “Logic handles complexity, probability represents uncertainty.”
Typical Tasks for Statistical-Relational Learning (SRL)

- **Link-based Classification**: given the links of a target entity and the attributes of related entities, predict the class label of the target entity.

- **Link Prediction**: given the attributes of entities and their other links, predict the existence of a link.
Link-based Classification

- Predict Attributes given Links, other Attributes
- E.g., $P(\text{diff}(101))$?

<table>
<thead>
<tr>
<th>s-id</th>
<th>Intelligence</th>
<th>Ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jack</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Kim</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Paul</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>c-id</th>
<th>Rating</th>
<th>Difficulty</th>
</tr>
</thead>
<tbody>
<tr>
<td>101</td>
<td>3</td>
<td>???</td>
</tr>
<tr>
<td>102</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>p-id</th>
<th>Popularity</th>
<th>Teaching-a</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oliver</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Jim</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>s-id</th>
<th>p-id</th>
<th>Salary</th>
<th>Capability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jack</td>
<td>Oliver</td>
<td>High</td>
<td>3</td>
</tr>
<tr>
<td>Kim</td>
<td>Oliver</td>
<td>Low</td>
<td>1</td>
</tr>
<tr>
<td>Paul</td>
<td>Jim</td>
<td>Med</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>c-id</th>
<th>Grade</th>
<th>Satisfaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>101</td>
<td>A</td>
<td>1</td>
</tr>
<tr>
<td>102</td>
<td>B</td>
<td>2</td>
</tr>
<tr>
<td>102</td>
<td>A</td>
<td>1</td>
</tr>
<tr>
<td>101</td>
<td>B</td>
<td>1</td>
</tr>
</tbody>
</table>
Link prediction

- Predict links given links, attributes.
- E.g., P(Registered(jack, 101))?

<table>
<thead>
<tr>
<th>Student</th>
<th>Course</th>
<th>Professor</th>
</tr>
</thead>
<tbody>
<tr>
<td>s-id</td>
<td>Intelligence</td>
<td>Ranking</td>
</tr>
<tr>
<td>Jack</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Kim</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Paul</td>
<td>1</td>
<td>2 RA</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>c-id</th>
<th>Rating</th>
<th>Difficulty</th>
</tr>
</thead>
<tbody>
<tr>
<td>101</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>p-id</th>
<th>Popularity</th>
<th>Teaching-a</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oliver</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>s-id</th>
<th>p-id</th>
<th>Salary</th>
<th>Capability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jack</td>
<td>Oliver</td>
<td>High</td>
<td>3</td>
</tr>
<tr>
<td>Kim</td>
<td>Oliver</td>
<td>Low</td>
<td>1</td>
</tr>
<tr>
<td>Paul</td>
<td>Jim</td>
<td>Med</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>s-id</th>
<th>c.id</th>
<th>Grade</th>
<th>Satisfaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jack</td>
<td>101</td>
<td>A</td>
<td>1</td>
</tr>
<tr>
<td>Jack</td>
<td>102</td>
<td>B</td>
<td>2</td>
</tr>
<tr>
<td>Kim</td>
<td>102</td>
<td>A</td>
<td>1</td>
</tr>
<tr>
<td>Paul</td>
<td>101</td>
<td>B</td>
<td>1</td>
</tr>
</tbody>
</table>
Generative Models

- Model the joint distribution over links and attributes.
- Today’s Topic.
- We’ll use Bayes nets as the model class.
What is a Bayes (belief) net?

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
- A solution to the relevance problem: Easy to compute “Is X relevant to Y given Z”.
- Nice UBC Demo.
Outline

- Brief intro to relational databases.
- Statistics and Relational Databases.
- Briefer intro to Bayes nets.
  - Relational Random Variables.
  - Relational (pseudo)-likelihoods.
Relational Data: what are the random variables?

• Intuitively, the attributes and relationships in the database.
  • i.e., the columns plus link existence.
  • i.e., the components of the ER diagrams.

• Proposal from David Poole (CS UBC): apply the concept of **functors** from Logic Programming.

• I’m combining this with Halpern (CS Cornell) and Bacchus’ (CS U of T) random selection probabilistic semantics for logic.
Population Variables

Russell: “A good notation thinks for us”.

- Consider a model with multiple populations.
- Let $X_1, X_2, Y_1, Y_2, \ldots$ be **population variables**.
- Each variable represents a random draw from a population.
- Population variables are jointly independent.
- A **functor** $f$ is a function of one or more population variables.
- A **functor random variable** is written as $f_1(X)$ or $f_2(X, Y)$ or $f_3(X, Y, Z)$. 
Unary Functors = Descriptive Attributes of Entities

- Population of Students, Professors.
- Population variables $S, P$.
- Attributes r.v.s $age(S), gpa(S), age(P), rank(P)$.
- Can have several selections $age(S_1), age(S_2)$.
- If $S$ is uniform over students in the database:
  - $P(gpa(S) = 3.0) = \text{empirical or database frequency}$ of 3.0 gpa in student population.
- Can instantiate or ground functors with constants.
  - E.g., $gpa(jack)$ returns the gpa of Jack.
Binary Functors = Relationships


- If \( S, C \) uniformly distributed over observed population:
  - \( P(\text{Registered}(S, C) = 1) = \frac{\#(s, c) \text{ s.t. Student } s \text{ is registered in course } c}{\#\text{Students} \times \#\text{Courses}}. \)
  - = Database Frequency of Registration.

- Can also form chains:
  - \( P(\text{grade}(S, C) = A, \text{Teaches}(C, P) = 1). \)
Functor Bayes Nets

Poole IJCAI 2003: A **functor Bayes Net** is a Bayes net whose nodes are functor random variables.

\[
P(S(Y) = T | S(X) = T, F(X,Y) = T) = 70% \\
P(S(Y) = T | S(X) = T, F(X,Y) = F) = 75%
\]

...
Likelihood Functions for Functor Bayes Nets: Latent Variables

- Problem: Given a database $D$ and an FBN model $B$, how to define $P(D | B)$?
- Fundamental Issue: interdependent units, not iid.
- One approach: introduce latent variables such that units are independent conditional on hidden “state” (e.g., Kersting et al. IJCAI 2009).
- Cf. social network analysis Hoff, Rafferty (U of W Stats), Linkletter SFU Stats.
- Cf. nonnegative matrix factorization----Netflix challenge.
Likelihood Function for Single-Table Data

For single table $T$:

$$\ln[P(T \mid B)] = L(T \mid B) = \sum_{\text{nodes } i} \sum_{\text{values } a} \sum_{\text{parent } - \text{state } j} n_T(a, j) \times \ln(P_B(a \mid j))$$

where $n_T(a, j)$ is the table count of co-occurrences of child node value $a$ and parent state $j$, and $P_B(a \mid j)$ is the parameter of the Bayes net for node $a$ given parent state $j$.
For database D:

\[ \ln[P(T \mid B)] = L(T \mid B) = \sum_{\text{nodes } i} \sum_{\text{values } a} \sum_{\text{parent state } j} p_D(a, j) \times \ln(P_B(a \mid j)) \]

Database joint frequency of child node value and parent state

Parameter of Bayes net

### Actors

<table>
<thead>
<tr>
<th>Name</th>
<th>Smokes</th>
<th>Cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anna</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>Bob</td>
<td>T</td>
<td>F</td>
</tr>
</tbody>
</table>

### Friend

<table>
<thead>
<tr>
<th>Name1</th>
<th>Name2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anna</td>
<td>Bob</td>
</tr>
<tr>
<td>Bob</td>
<td>Anna</td>
</tr>
</tbody>
</table>
Random Selection Log-Likelihood

1. Randomly select instances \( X_1 = x_1, \ldots, X_n = x_n \) for each variable in FBN.
2. Look up their properties, relationships in database.
3. Compute log-likelihood for the FBN assignment obtained from the instances.
4. \( L^R = \) expected log-likelihood over uniform random selection of instances.

<table>
<thead>
<tr>
<th>Hyperentity</th>
<th>Hyperfeatures</th>
<th>( P_B^R )</th>
<th>( \ln(P_B^R) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \gamma_1 ) Anna Bob</td>
<td>T T T T F</td>
<td>0.105</td>
<td>-2.254</td>
</tr>
<tr>
<td>( \gamma_2 ) Bob Anna</td>
<td>T T F T T</td>
<td>0.245</td>
<td>-1.406</td>
</tr>
<tr>
<td>( \gamma_3 ) Anna Anna</td>
<td>F T T T T</td>
<td>0.263</td>
<td>-1.338</td>
</tr>
<tr>
<td>( \gamma_4 ) Bob Bob</td>
<td>F T F T F</td>
<td>0.113</td>
<td>-2.185</td>
</tr>
</tbody>
</table>

\[ L^R = -(2.254 + 1.406 + 1.338 + 2.185)/4 \approx -1.8 \]

**Proposition** The random selection log-likelihood equals the pseudo log-likelihood.
Parameter Estimation

**Proposition** For a given database D, the parameter values that maximize the pseudo likelihood are the empirical conditional frequencies.
Model Selection

- New model selection algorithm (Khosravi, Schulte et al. AAAI 2010).
- Level-wise search through table join lattice.
Running time on benchmarks

<table>
<thead>
<tr>
<th>Dataset</th>
<th>JBN</th>
<th>MLN</th>
<th>CMLN</th>
</tr>
</thead>
<tbody>
<tr>
<td>University</td>
<td>0.03+0.032</td>
<td>5.02</td>
<td>11.44</td>
</tr>
<tr>
<td>MovieLens</td>
<td>1.2+120</td>
<td>NT</td>
<td>NT</td>
</tr>
<tr>
<td>MovieLens Subsample 1</td>
<td>0.05 + 0.33</td>
<td>44</td>
<td>121.5</td>
</tr>
<tr>
<td>MovieLens Subsample 2</td>
<td>0.12 + 5.10</td>
<td>2760</td>
<td>1286</td>
</tr>
<tr>
<td>Mutagenesis</td>
<td>0.5 +NT</td>
<td>NT</td>
<td>NT</td>
</tr>
<tr>
<td>Mutagenesis subsample 1</td>
<td>0.1 +5</td>
<td>3360</td>
<td>900</td>
</tr>
<tr>
<td>Mutagenesis subsample 2</td>
<td>0.2 +12</td>
<td>NT</td>
<td>3120</td>
</tr>
</tbody>
</table>

- Time in Minutes. NT = did not terminate.
- $x + y = \text{structure learning} + \text{parametrization (with Markov net methods)}$.
- JBN: Our join-based algorithm.
- MLN, CMLN: standard programs from the U of Washington (Alchemy)
Accuracy

Basically, leave-one-out average.
Future Work: Inference

Prediction is usually based on *knowledge-based model construction* (Ngo and Haddaway, 1997; Koller and Pfeffer, 1997; Haddaway, 1999).

- Basic Idea: instantiate population variables with all population members. Predict using instantiated model.
- With Bayes nets, can lead to cycles.
- My conjecture: cycles can be handled with a normalization constant that has a closed form.
- Help?!
Summary: Likelihood for relational data.

- Combining relational databases and statistics.
  - Very important in practice.
  - Combine logic and probability.
- Interdependent units → hard to define model likelihood.
- Proposal: Consider a randomly selected small group of individuals.
- Pseudo log-likelihood = expected log-likelihood of randomly selected group.
Summary: Statistics with Pseudo-Likelihood

- Theorem: Random pseudo log-likelihood equivalent to standard single-table likelihood, replacing table counts with database frequencies.
- Maximum likelihood estimates = database frequencies.
- Efficient Model Selection Algorithm based on lattice search.
- In simulations, very fast (minutes vs. days), much better predictive accuracy.
Thank you!

- Any questions?
Choice of Functors

- Can have complex functors, e.g.
  - Nested: \( \text{wealth}(\text{father}(\text{father}(X))) \).
  - Aggregate: \( \text{AVG}_C \{ \text{grade}(S,C) : \text{Registered}(S,C) \} \).
- In remainder of this talk, use functors corresponding to
  - Attributes (columns), e.g., \( \text{intelligence}(S) \), \( \text{grade}(S,C) \)
  - Boolean Relationship indicators, e.g. \( \text{Friend}(X,Y) \).
• Assign unobserved values \( u(\text{jack}) \), \( u(\text{jane}) \).
• Probability that Jack and Jane are friends depends on their unobserved “type”.
• In ground model, \( \text{rich(jack)} \) and \( \text{rich(jane)} \) are correlated given that they are friends, but neither is an ancestor.
• $1M prize in Netflix challenge.
• Also for multiple types of relationships (Kersting et al. 2009).
• Computationally demanding.
Typical Tasks for Statistical-Relational Learning (SRL)

- **Link-based Classification**: given the links of a target entity and the attributes of related entities, predict the class label of the target entity.
- **Link Prediction**: given the attributes of entities and their other links, predict the existence of a link.