Relational Probabilistic Models

David Poole

Department of Computer Science,
University of British Columbia

Work with: http://minervaintelligence.com, https://treatment.com/

February 25, 2020

"The mind is a neural computer, fitted by natural selection with combinatorial algorithms for causal and probabilistic reasoning about plants, animals, objects, and people."

...

"In a universe with any regularities at all, decisions informed about the past are better than decisions made at random. That has always been true, and we would expect organisms, especially informavores such as humans, to have evolved acute intuitions about probability. The founders of probability, like the founders of logic, assumed they were just formalizing common sense."

Steven Pinker, How the Mind Works, 1997, pp. 524, 343.

"The mind is a neural computer, fitted by natural selection with combinatorial algorithms for causal and probabilistic reasoning about plants, animals, objects, and people."

...

"In a universe with any regularities at all, decisions informed about the past are better than decisions made at random. That has always been true, and we would expect organisms, especially informavores such as humans, to have evolved acute intuitions about probability. The founders of probability, like the founders of logic, assumed they were just formalizing common sense."

Steven Pinker, How the Mind Works, 1997, pp. 524, 343.

Outline

1. Knowledge Graphs
   - Triples and Reification
   - Tensor Factorization and Neural Network Models
2. Representation Issues
   - Desiderata
   - How do relational models relate to probabilistic graphical models
3. Unique properties of relational models
   - Learning general knowledge vs learning about a data set
   - Varying Populations
   - What can be observed?
4. Conclusions and Challenges

Outline

1. Knowledge Graphs
   - Triples and Reification
   - Tensor Factorization and Neural Network Models
2. Representation Issues
   - Desiderata
   - How do relational models relate to probabilistic graphical models
3. Unique properties of relational models
   - Learning general knowledge vs learning about a data set
   - Varying Populations
   - What can be observed?
4. Conclusions and Challenges

Choosing Individuals and Relations in Logic

First-order logical languages allow many different ways of representing facts.

E.g., How to represent: "Pen #7 is red."

Choosing Individuals and Relations in Logic

First-order logical languages allow many different ways of representing facts.

E.g., How to represent: "Pen #7 is red."

- $red(pen_7)$. It's easy to ask "What's red?"
  Can't ask "what is the color of pen_7?"

Choosing Individuals and Relations in Logic

First-order logical languages allow many different ways of representing facts.

E.g., How to represent: "Pen #7 is red."

- $red(pen_7)$. It's easy to ask "What's red?"
  Can't ask "what is the color of pen_7?"
- $color(pen_7, red)$. It's easy to ask "What's red?"
  It's easy to ask "What is the color of pen_7?"
  Can't ask "What property of pen_7 has value red?"

Choosing Individuals and Relations in Logic

First-order logical languages allow many different ways of representing facts.

E.g., How to represent: "Pen #7 is red."

- $red(pen_7)$. It's easy to ask "What's red?"
  Can't ask "what is the color of pen_7?"
- $color(pen_7, red)$. It's easy to ask "What's red?"
  It's easy to ask "What is the color of pen_7?"
  Can't ask "What property of pen_7 has value red?"
- $prop(pen_7, color, red)$. It's easy to ask all these questions.

Choosing Individuals and Relations in Logic

First-order logical languages allow many different ways of representing facts.

E.g., How to represent: "Pen #7 is red."

- $red(pen_7)$. It's easy to ask "What's red?"
  Can't ask "what is the color of pen_7?"
- $color(pen_7, red)$. It's easy to ask "What's red?"
  It's easy to ask "What is the color of pen_7?"
  Can't ask "What property of pen_7 has value red?"
- $prop(pen_7, color, red)$. It's easy to ask all these questions.

With a single relation it can be implicit $\longrightarrow$ triples:
$\langle pen_7, color, red \rangle$.

Universality of *prop*

To represent “a is a parcel”

Universality of *prop*

To represent "a is a parcel"

- $prop(a, type, parcel)$, where *type* is a special property and *parcel* is a class.
- $prop(a, parcel, true)$, where *parcel* is a Boolean property (characteristic function of the class *parcel*).

Triples

- To represent $tutorial(\text{"}StarAI\text{"}, nips2017, 1045, hallC)$. "the Star AI tutorial at NIPS 2017 is scheduled at 10:45 in Hall C."

Triples

- To represent $tutorial(\text{"}StarAI\text{"}, nips2017, 1045, hallC)$. "the Star AI tutorial at NIPS 2017 is scheduled at 10:45 in Hall C."
- Let $t123$ name the offering of the tutorial:

$$prop(t123, type, tutorial).$$

$$prop(t123, title, \text{"}StarAI\text{"}).$$

$$prop(t123, event, nips2017).$$

$$prop(t123, time, 1045).$$

$$prop(t123, room, hallC).$$

Triples

- To represent $tutorial("StarAI", nips2017, 1045, hallC)$. "the Star AI tutorial at NIPS 2017 is scheduled at 10:45 in Hall C."
- Let $t123$ name the offering of the tutorial:

  $prop(t123, type, tutorial)$.

  $prop(t123, title, "StarAI")$.

  $prop(t123, event, nips2017)$.

  $prop(t123, time, 1045)$.

  $prop(t123, room, hallC)$.

- We have reified the booking.
- Reify means: to make into an individual.

Triples

- To represent $tutorial(\text{"}StarAI\text{"}, nips2017, 1045, hallC)$. "the Star AI tutorial at NIPS 2017 is scheduled at 10:45 in Hall C."
- Let $t123$ name the offering of the tutorial:

  $prop(t123, type, tutorial).$

  $prop(t123, title, \text{"}StarAI\text{"}).$

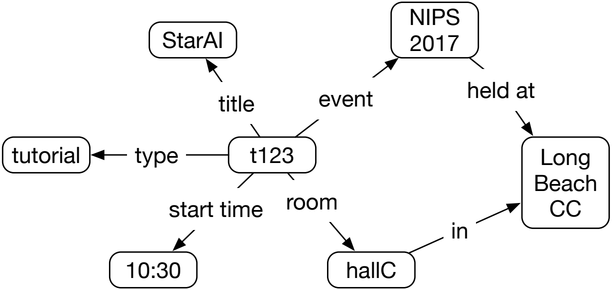
  $prop(t123, event, nips2017).$

  $prop(t123, time, 1045).$

  $prop(t123, room, hallC).$
- We have reified the booking.
- Reify means: to make into an individual.
- How can we add extra arguments (e.g., presenters, chair)?

Triples and Knowledge Graphs

- Subject–verb–object
  Individual–property–value
  triples can be depicted as edges in graphs



- A modeller or learner needs to invent (reified) objects.

Triples are universal representations of relations

All relations can be represented in terms of triples:

	...	P_j	...
	...	...	...
r_i	...	v_{ij}	...
	...	...	...

can be represented as

Triples are universal representations of relations

All relations can be represented in terms of triples:

	...	P_j	...
	...	...	...
r_i	...	v_{ij}	...
	...	...	...

can be represented as the triple $\langle r_i, P_j, v_{ij} \rangle$.

- r_i is either a primary key or a reified individual.

Triples are universal representations of relations

All relations can be represented in terms of triples:

	...	P_j	...
	...	...	...
r_i	...	v_{ij}	...
	...	...	...

can be represented as the triple $\langle r_i, P_j, v_{ij} \rangle$.

- r_i is either a primary key or a reified individual.
- Examples of reified individuals: a booking, a marriage, a talk, a lab report, an event, a party, a meeting, a drink

Triples are universal representations of relations

All relations can be represented in terms of triples:

	…	P_j	…
	…	…	…
r_i	…	v_{ij}	…
	…	…	…

can be represented as the triple $\langle r_i, P_j, v_{ij} \rangle$.

- r_i is either a primary key or a reified individual.
- Examples of reified individuals: a booking, a marriage, a talk, a lab report, an event, a party, a meeting, a drink
- Challenge for learning: each reified individual has limited data to learn from; (at most) one value for each property.

Triples are universal representations of relations

All relations can be represented in terms of triples:

	...	P_j	...
	...	...	...
r_i	...	v_{ij}	...
	...	...	...

can be represented as the triple $\langle r_i, P_j, v_{ij} \rangle$.

- r_i is either a primary key or a reified individual.
- Examples of reified individuals: a booking, a marriage, a talk, a lab report, an event, a party, a meeting, a drink
- Challenge for learning: each reified individual has limited data to learn from; (at most) one value for each property.

prop(*Individual*, *Property*, *Value*) is the only relation needed: $\langle Individual, Property, Value \rangle$ triples, Semantic network, entity relationship model, knowledge graphs, ...

Outline

1. Knowledge Graphs
   - Triples and Reification
   - Tensor Factorization and Neural Network Models
2. Representation Issues
   - Desiderata
   - How do relational models relate to probabilistic graphical models
3. Unique properties of relational models
   - Learning general knowledge vs learning about a data set
   - Varying Populations
   - What can be observed?
4. Conclusions and Challenges

Vector & Tensor Representations of Entities & Relations

- Knowledge graphs (semantic networks) are databases of triples
- We want to estimate $P(\langle h, r, t \rangle)$

Vector & Tensor Representations of Entities & Relations

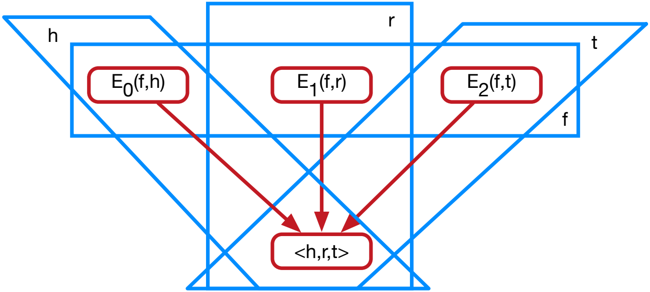
- Knowledge graphs (semantic networks) are databases of triples
- We want to estimate $P(\langle h, r, t \rangle)$
- Polyadic decomposition model (1927): two vector embeddings for each entity, and one for reach relation

$$P(\langle h, r, t \rangle) = sigmoid(\sum_f E_0(f, h) * E_1(f, r) * E_2(f, t))$$

Vector & Tensor Representations of Entities & Relations

- Knowledge graphs (semantic networks) are databases of triples
- We want to estimate $P(\langle h, r, t\rangle)$
- Polyadic decomposition model (1927): two vector embeddings for each entity, and one for reach relation

$$P(\langle h, r, t\rangle) = sigmoid(\sum_f E_0(f, h) * E_1(f, r) * E_2(f, t))$$



Polyadic decomposition variations

- Polyadic decomposition doesn't work very well...
  - Consider $\langle p123, likes, m53\rangle$ and $\langle m53, directed\_by, p534\rangle$.

Polyadic decomposition variations

- Polyadic decomposition doesn't work very well...
  - Consider $\langle p123, \mathit{likes}, m53\rangle$ and $\langle m53, \mathit{directed\_by}, p534\rangle$.
  - Requires two embeddings per entity, but head embeddings and tail embeddings do not interact.

Polyadic decomposition variations

- Polyadic decomposition doesn't work very well...
  - Consider $\langle p123, likes, m53\rangle$ and $\langle m53, directed\_by, p534\rangle$.
  - Requires two embeddings per entity, but head embeddings and tail embeddings do not interact.
- DistMult: share same embedding for head and tail.

Polyadic decomposition variations

- Polyadic decomposition doesn't work very well...
  - Consider $\langle p123, \mathit{likes}, m53\rangle$ and $\langle m53, \mathit{directed\_by}, p534\rangle$.
  - Requires two embeddings per entity, but head embeddings and tail embeddings do not interact.
- DistMult: share same embedding for head and tail.
  Problem:

Polyadic decomposition variations

- Polyadic decomposition doesn't work very well...
    - Consider $\langle p123, likes, m53\rangle$ and $\langle m53, directed\_by, p534\rangle$.
    - Requires two embeddings per entity, but head embeddings and tail embeddings do not interact.
- DistMult: share same embedding for head and tail.
  Problem: can only represent symmetric relations.

Polyadic decomposition variations

- Polyadic decomposition doesn't work very well...
  - Consider $\langle p123, likes, m53\rangle$ and $\langle m53, directed\_by, p534\rangle$.
  - Requires two embeddings per entity, but head embeddings and tail embeddings do not interact.
- DistMult: share same embedding for head and tail.
  Problem: can only represent symmetric relations.
- CompleX: the embeddings are complex numbers, tail is the conjugate of the head embedding

Polyadic decomposition variations

- Polyadic decomposition doesn't work very well...
  - Consider $\langle p123, \mathit{likes}, m53\rangle$ and $\langle m53, \mathit{directed\_by}, p534\rangle$.
  - Requires two embeddings per entity, but head embeddings and tail embeddings do not interact.
- DistMult: share same embedding for head and tail.
  Problem: can only represent symmetric relations.
- CompleX: the embeddings are complex numbers, tail is the conjugate of the head embedding
- SimpleE: have an embedding for r^{-1} and learn to predict both $\langle h, r, t\rangle$ and $\left\langle t, r^{-1}, h\right\rangle$

Learning general knowledge vs learning about a data set

- Suppose you want to create a model of who is friends with whom.

Learning general knowledge vs learning about a data set

- Suppose you want to create a model of who is friends with whom. Options:
  - learn general knowledge, e.g., transitivity, how male and female friendships work, how location affect friendship...

Learning general knowledge vs learning about a data set

- Suppose you want to create a model of who is friends with whom. Options:
  - learn general knowledge, e.g., transitivity, how male and female friendships work, how location affect friendship...
  - learn specific knowledge about who is friends with who; e.g., which particular group of people are generally friends with each other.

Learning general knowledge vs learning about a data set

- Suppose you want to create a model of who is friends with whom. Options:
  - learn general knowledge, e.g., transitivity, how male and female friendships work, how location affect friendship...
  - learn specific knowledge about who is friends with who; e.g., which particular group of people are generally friends with each other.
- The specific knowledge will tend to be more accurate on that population, but doesn't generalize to different populations.
- The general knowledge will tend to transfer better.
- Which is better depends on the goals and how success is measured.

Learning general knowledge vs learning about a data set

- Suppose you want to create a model of who is friends with whom. Options:
  - learn general knowledge, e.g., transitivity, how male and female friendships work, how location affect friendship...
  - learn specific knowledge about who is friends with who; e.g., which particular group of people are generally friends with each other.
- The specific knowledge will tend to be more accurate on that population, but doesn't generalize to different populations.
- The general knowledge will tend to transfer better.
- Which is better depends on the goals and how success is measured.
- Many of the methods try to do both; learn about specific individuals and general knowledge.

Challenges of learning knowledge graphs

- Evaluating predictions when only positive examples are provided
  Consider the following relations:
  - Married to
  - Friend of
  - Knows about
  - Would get along with

Challenges of learning knowledge graphs

- Evaluating predictions when only positive examples are provided
  Consider the following relations:
  - Married to — each person related to 0 or 1 other persons (with a few exceptions)
  - Friend of — each person related to tens or hundreds of others
  - Knows about — each person might know about hundreds or thousands of others. Some people my be known by millions or billions of others.
  - Would get along with — almost everyone gets along with almost everyone else, but with some exceptions.

Challenges of learning knowledge graphs

- Evaluating predictions when only positive examples are provided
  Consider the following relations:
  - Married to — each person related to 0 or 1 other persons (with a few exceptions)
  - Friend of — each person related to tens or hundreds of others
  - Knows about — each person might know about hundreds or thousands of others. Some people my be known by millions or billions of others.
  - Would get along with — almost everyone gets along with almost everyone else, but with some exceptions.
- Often we compare rankings (ordering), but what if the answer is "no"?

Challenges of learning knowledge graphs

- Evaluating predictions when only positive examples are provided
  Consider the following relations:
  - Married to — each person related to 0 or 1 other persons (with a few exceptions)
  - Friend of — each person related to tens or hundreds of others
  - Knows about — each person might know about hundreds or thousands of others. Some people my be known by millions or billions of others.
  - Would get along with — almost everyone gets along with almost everyone else, but with some exceptions.
- Often we compare rankings (ordering), but what if the answer is "no"?
- Difficult to learn about reified entities.

Predicting Properties

- Tensor factorization models work well for predicting relations, but not for predicting properties.
  Tensor factorization relies on lower-dimensional representations, and there isn't one for properties.

Predicting Properties

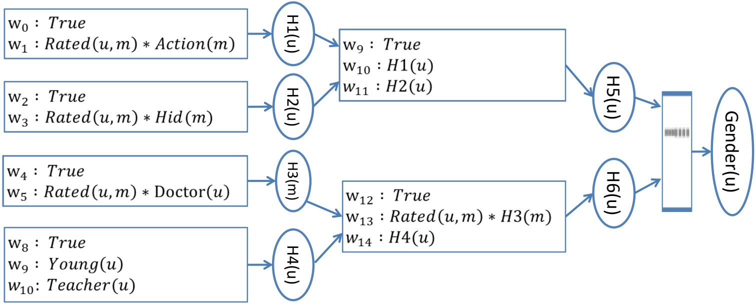
- Tensor factorization models work well for predicting relations, but not for predicting properties.
  Tensor factorization relies on lower-dimensional representations, and there isn't one for properties.
- Imagine trying to predict $age(P)$, the age of person P, and $rating(P, M)$ the rating of person P on movie M.
  One of the embeddings of each person can just memorize the age — no generalization!
  — there are too many parameters

Predicting Properties

- Tensor factorization models work well for predicting relations, but not for predicting properties.
  Tensor factorization relies on lower-dimensional representations, and there isn't one for properties.
- Imagine trying to predict $age(P)$, the age of person P, and $rating(P, M)$ the rating of person P on movie M.
  One of the embeddings of each person can just memorize the age — no generalization!
  — there are too many parameters
- We can build relational neural networks to solve this

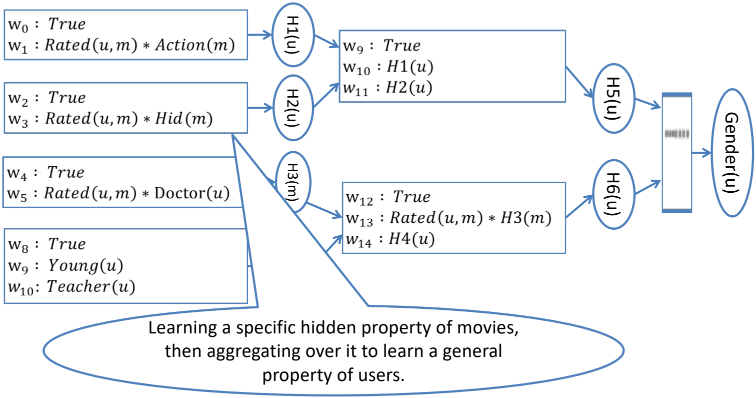
[Kazemi & Poole, AAAI 2017]

Relational Neural Nets (RelNNs)



[Kazemi & Poole AAAI 2017]

Relational Neural Nets (RelNNs)



[Kazemi & Poole AAAI 2017]

Outline

1. Knowledge Graphs
   - Triples and Reification
   - Tensor Factorization and Neural Network Models
2. Representation Issues
   - Desiderata
   - How do relational models relate to probabilistic graphical models
3. Unique properties of relational models
   - Learning general knowledge vs learning about a data set
   - Varying Populations
   - What can be observed?
4. Conclusions and Challenges

Desiderata for a Representation

- Expressiveness:
  Is it expressive enough to solve problem at hand?

Desiderata for a Representation

- Expressiveness:
  Is it expressive enough to solve problem at hand?
- Efficient Inference:
  Is it efficient in the worst case or average case?
  Can it exploit structure (e.g., independencies and symmetries)

Desiderata for a Representation

- Expressiveness:
  Is it expressive enough to solve problem at hand?
- Efficient Inference:
  Is it efficient in the worst case or average case?
  Can it exploit structure (e.g., independencies and symmetries)
- Understandability or explainability:
  Can people understand the model?
  Can a particular prediction be explained?

Desiderata for a Representation

- Expressiveness:
  Is it expressive enough to solve problem at hand?
- Efficient Inference:
  Is it efficient in the worst case or average case?
  Can it exploit structure (e.g., independencies and symmetries)
- Understandability or explainability:
  Can people understand the model?
  Can a particular prediction be explained?
- Learnability: Can it be learned from:
  - heterogenous data
  - prior knowledge

Desiderata for a Representation

- Expressiveness:
  Is it expressive enough to solve problem at hand?
- Efficient Inference:
  Is it efficient in the worst case or average case?
  Can it exploit structure (e.g., independencies and symmetries)
- Understandability or explainability:
  Can people understand the model?
  Can a particular prediction be explained?
- Learnability: Can it be learned from:
  - heterogenous data
  - prior knowledge
- Modularity:
  Can independently developed parts be combined to form larger model?
  Can a larger model be decomposed into smaller parts?

Outline

1. Knowledge Graphs
   - Triples and Reification
   - Tensor Factorization and Neural Network Models
2. Representation Issues
   - Desiderata
   - How do relational models relate to probabilistic graphical models
3. Unique properties of relational models
   - Learning general knowledge vs learning about a data set
   - Varying Populations
   - What can be observed?
4. Conclusions and Challenges

Directed vs Undirected Probabilistic Graphical Models

- Undirected models (Markov networks, factor graphs) represent probability distributions in terms of factors.
  - a factor is a non-negative function of a set of variables
  - variables in a factor are neighbours of each other
  - each variable in independent of its non-neighbours given its neighbours.

Directed vs Undirected Probabilistic Graphical Models

- Undirected models (Markov networks, factor graphs) represent probability distributions in terms of factors.
  - a factor is a non-negative function of a set of variables
  - variables in a factor are neighbours of each other
  - each variable in independent of its non-neighbours given its neighbours.
- In directed models, factors represent conditional probabilities:
  - each variable in independent of its non-descendents given its parents.

Directed vs Undirected Probabilistic Graphical Models

- Undirected models (Markov networks, factor graphs) represent probability distributions in terms of factors.
  - a factor is a non-negative function of a set of variables
  - variables in a factor are neighbours of each other
  - each variable in independent of its non-neighbours given its neighbours.
- In directed models, factors represent conditional probabilities:
  - each variable in independent of its non-descendents given its parents.
- $\{directed\_models\} \subset \{undirected\_models\}$

Directed vs Undirected Probabilistic Graphical Models

- Undirected models (Markov networks, factor graphs) represent probability distributions in terms of factors.
  - a factor is a non-negative function of a set of variables
  - variables in a factor are neighbours of each other
  - each variable in independent of its non-neighbours given its neighbours.
- In directed models, factors represent conditional probabilities:
  - each variable in independent of its non-descendents given its parents.
- $\{directed\_models\} \subset \{undirected\_models\}$
- Directed (and undirected) models are universal: can represent any probability distribution over a finite set of variables

Directed vs Undirected Probabilistic Graphical Models

- Undirected models (Markov networks, factor graphs) represent probability distributions in terms of factors.
  - a factor is a non-negative function of a set of variables
  - variables in a factor are neighbours of each other
  - each variable in independent of its non-neighbours given its neighbours.
- In directed models, factors represent conditional probabilities:
  - each variable in independent of its non-descendents given its parents.
- $\{directed\_models\} \subset \{undirected\_models\}$
- Directed (and undirected) models are universal: can represent any probability distribution over a finite set of variables
- Algorithms developed for undirected models work for both.

Directed vs Undirected Probabilistic Graphical Models

- Undirected models (Markov networks, factor graphs) represent probability distributions in terms of factors.
  - a factor is a non-negative function of a set of variables
  - variables in a factor are neighbours of each other
  - each variable in independent of its non-neighbours given its neighbours.
- In directed models, factors represent conditional probabilities:
  - each variable in independent of its non-descendents given its parents.
- $\{directed\_models\} \subset \{undirected\_models\}$
- Directed (and undirected) models are universal: can represent any probability distribution over a finite set of variables
- Algorithms developed for undirected models work for both.
- That does not mean that representations for undirected models can represent directed models.

Example

Weighted formulae about a social situation:

$$-5 : \mathit{funFor}(X)$$

$$10 : \mathit{funFor}(X) \land \mathit{knows}(X, Y) \land \mathit{social}(Y)$$

Example

Weighted formulae about a social situation:

$$-5 : \mathit{funFor}(X)$$

$$10 : \mathit{funFor}(X) \land \mathit{knows}(X, Y) \land \mathit{social}(Y)$$

If Π includes observations for all $\mathit{knows}(X, Y)$ and $\mathit{social}(Y)$:

$$P(\mathit{funFor}(X) \mid \Pi) = \mathit{sigmoid}(-5 + 10 n_T)$$

n_T is the number of individuals Y for which $\mathit{knows}(X, Y) \land \mathit{social}(Y)$ is *True* in Π.

$$\mathit{sigmoid}(x) = \frac{1}{1 + e^{-x}}$$

Example

Weighted formulae about a social situation:

$$-5 : \mathit{funFor}(X)$$

$$10 : \mathit{funFor}(X) \wedge \mathit{knows}(X, Y) \wedge \mathit{social}(Y)$$

If Π includes observations for all $\mathit{knows}(X, Y)$ and $\mathit{social}(Y)$:

$$P(\mathit{funFor}(X) \mid \Pi) = \mathit{sigmoid}(-5 + 10 n_T)$$

n_T is the number of individuals Y for which $\mathit{knows}(X, Y) \wedge \mathit{social}(Y)$ is *True* in Π.

$$\mathit{sigmoid}(x) = \frac{1}{1 + e^{-x}}$$

- Using wighted formulae to define conditional probabilities $\rightarrow$ relational logistic regression (RLR).
- Using wighted formulae to define distributions $\rightarrow$ Markov logic networks (MLNs).

Abstract Example

$\alpha_0 : q$
$\alpha_1 : q \wedge \neg r(x)$
$\alpha_2 : q \wedge r(x)$
$\alpha_3 : r(x)$

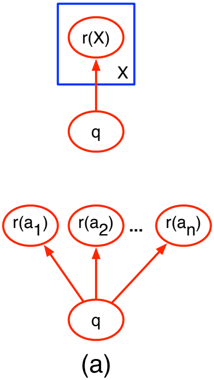
If $r(x)$ for every individual x is observed:

$$P(q \mid obs) = sigmoid(\alpha_0 + n_F \alpha_1 + n_T \alpha_2)$$

n_T is number of individuals for which $r(x)$ is true
n_F is number of individuals for which $r(x)$ is false

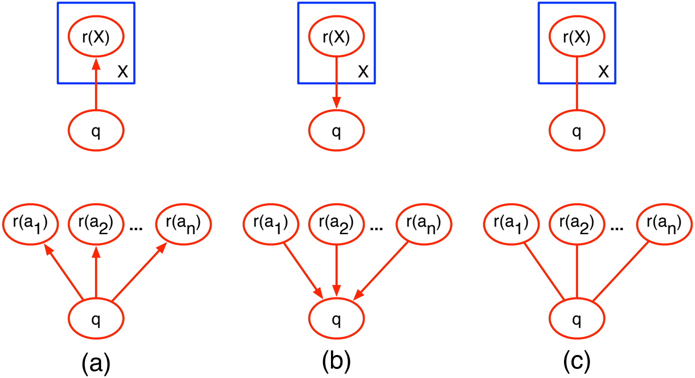
$$sigmoid(x) = \frac{1}{1 + e^{-x}}$$

Three Elementary Models



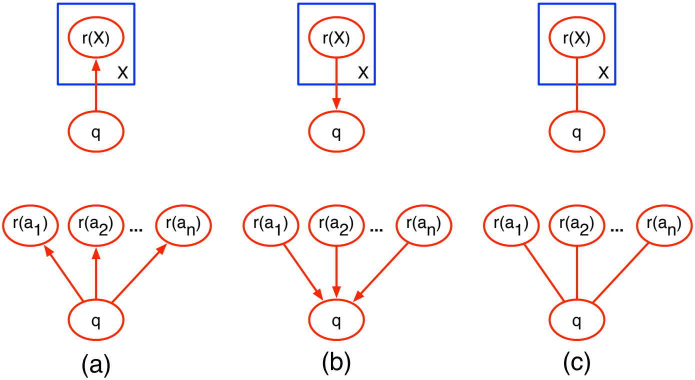
(a)

Three Elementary Models



(a) Naïve Bayes
(b)

Three Elementary Models

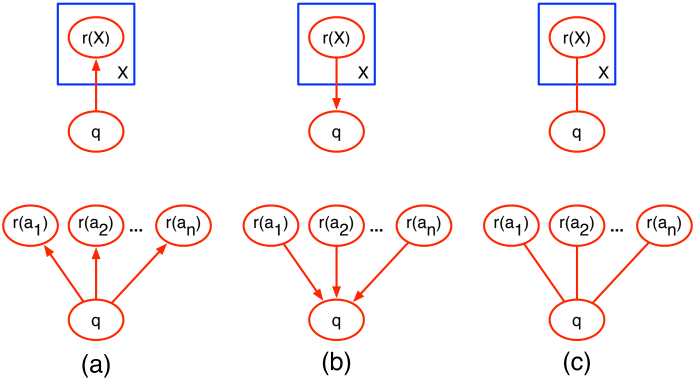


(a) Naïve Bayes

(b) (Relational) Logistic Regression

(c)

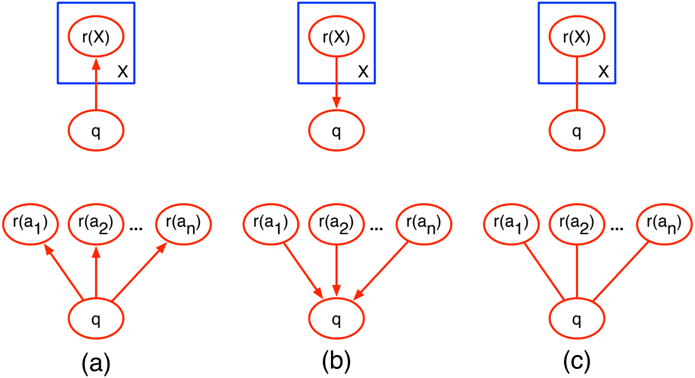
Three Elementary Models



(a) Naïve Bayes
(b) (Relational) Logistic Regression
(c) Markov (Logic) network

} described by unary and pairwise factors

Three Elementary Models

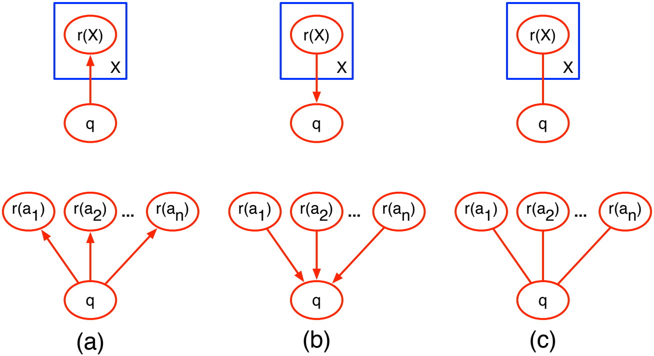


(a) Naïve Bayes
(b) (Relational) Logistic Regression
(c) Markov (Logic) network

} described by unary and pairwise factors

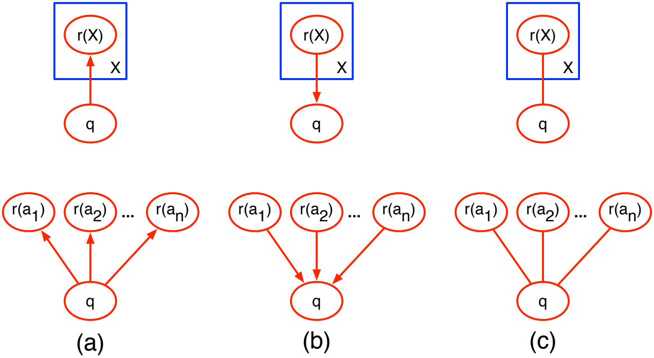
— They are identical models when all r's are observed.

Independence Assumptions



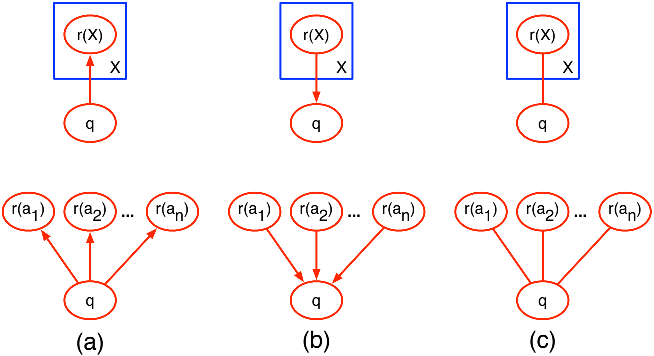
- Naïve Bayes (a) and Markov network (c): $r(a_i)$ and $r(a_j)$
  - are independent

Independence Assumptions



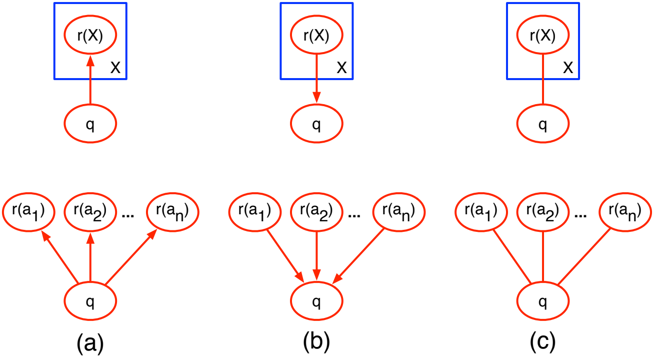
- Naïve Bayes (a) and Markov network (c): $r(a_i)$ and $r(a_j)$
  - are independent given Q
  - are dependent

Independence Assumptions



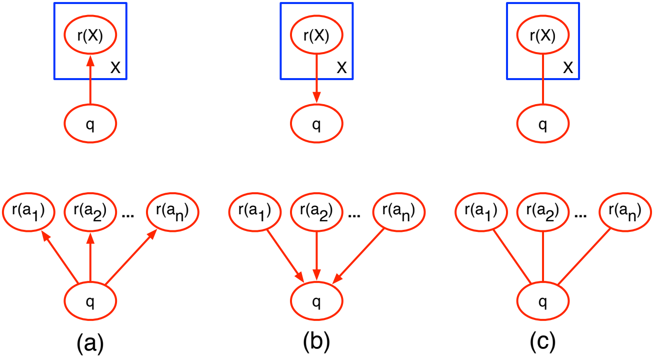
- Naïve Bayes (a) and Markov network (c): $r(a_i)$ and $r(a_j)$
  - are independent given Q
  - are dependent not given Q.
- Directed model with aggregation (b): $r(a_i)$ and $r(a_j)$
  - are dependent

Independence Assumptions



- Naïve Bayes (a) and Markov network (c): $r(a_i)$ and $r(a_j)$
  - are independent given Q
  - are dependent not given Q.
- Directed model with aggregation (b): $r(a_i)$ and $r(a_j)$
  - are dependent given Q,
  - are independent

Independence Assumptions



- Naïve Bayes (a) and Markov network (c): $r(a_i)$ and $r(a_j)$
  - are independent given Q
  - are dependent not given Q.
- Directed model with aggregation (b): $r(a_i)$ and $r(a_j)$
  - are dependent given Q,
  - are independent not given Q.

Modularity

- Directed models are inherently modular.
  $P(q \mid r(X))$ is defined so that distribution over $r(a_1) \ldots r(a_n)$ is not affected when q is summed out (and not observed)

Modularity

- Directed models are inherently modular.
  $P(q \mid r(X))$ is defined so that distribution over $r(a_1) \dots r(a_n)$ is not affected when q is summed out (and not observed)
- MLNs are provably not modular: If there is a distribution over $r(a_1) \dots r(a_n)$ (e.g., they are independent),
  $P(q \mid r(X))$ cannot be defined in an MLN so that
  - q depends on the r's ($P(q \mid r(X)) \neq P(q)$) and
  - if q is summed out, the distribution over $r(a_1) \dots r(a_n)$ is not changed.

Modularity

- Directed models are inherently modular.
  $P(q \mid r(X))$ is defined so that distribution over $r(a_1) \ldots r(a_n)$ is not affected when q is summed out (and not observed)
- MLNs are provably not modular: If there is a distribution over $r(a_1) \ldots r(a_n)$ (e.g., they are independent),
  $P(q \mid r(X))$ cannot be defined in an MLN so that
  - q depends on the r's ($P(q \mid r(X)) \neq P(q)$) and
  - if q is summed out, the distribution over $r(a_1) \ldots r(a_n)$ is not changed.
  - Why? requires factors on arbitrary subsets of $r(a_1) \ldots r(a_k)$
    — pairwise (or 3-wise or ...) interactions are not adequate
    — can't marry the parents

Cyclic Models

Whether people smoke depends on whether their friends smoke.

- MLN:

$$w : smokes(X) \leftarrow friends(X, Y) \wedge smokes(Y)$$

Cyclic Models

Whether people smoke depends on whether their friends smoke.

- MLN:

$$w : smokes(X) \leftarrow friends(X, Y) \land smokes(Y)$$

(where $\leftarrow$ is material implication) is equivalent to

$$\begin{aligned} w :& true(X) \land true(Y) \\ -w :& \neg smokes(X) \land friends(X, Y) \land smokes(Y) \end{aligned}$$

Cyclic Models

Whether people smoke depends on whether their friends smoke.

- MLN:

$$w : smokes(X) \leftarrow friends(X, Y) \wedge smokes(Y)$$

(where $\leftarrow$ is material implication) is equivalent to

$$w : true(X) \wedge true(Y)$$
$$-w : \neg smokes(X) \wedge friends(X, Y) \wedge smokes(Y)$$

- ICL/Problog

$$w : smokes(X) \leftarrow \quad friends(X, Y) \wedge smokes(Y)$$

Cyclic Models

Whether people smoke depends on whether their friends smoke.

- MLN:

$$w : smokes(X) \leftarrow friends(X, Y) \wedge smokes(Y)$$

(where $\leftarrow$ is material implication) is equivalent to

$$w : true(X) \wedge true(Y)$$
$$-w : \neg smokes(X) \wedge friends(X, Y) \wedge smokes(Y)$$

- ICL/Problog

$$w : smokes(X) \leftarrow \exists Y \; friends(X, Y) \wedge smokes(Y)$$

Cyclic Models

Whether people smoke depends on whether their friends smoke.

- MLN:

$$w : smokes(X) \leftarrow friends(X, Y) \wedge smokes(Y)$$

(where $\leftarrow$ is material implication) is equivalent to

$$\begin{aligned} w :& true(X) \wedge true(Y) \\ -w :& \neg smokes(X) \wedge friends(X, Y) \wedge smokes(Y) \end{aligned}$$

- ICL/Problog

$$w : smokes(X) \leftarrow \exists Y \ friends(X, Y) \wedge smokes(Y)$$

  - Probability of smokes goes up as the number of friends increases!

Cyclic Models

Whether people smoke depends on whether their friends smoke.

- MLN:

$$w : smokes(X) \leftarrow friends(X, Y) \land smokes(Y)$$

(where $\leftarrow$ is material implication) is equivalent to

$$w : true(X) \land true(Y)$$
$$-w : \neg smokes(X) \land friends(X, Y) \land smokes(Y)$$

- ICL/Problog

$$w : smokes(X) \leftarrow \exists Y \ friends(X, Y) \land smokes(Y)$$

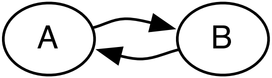
  - Probability of smokes goes up as the number of friends increases!
  - ICL/Problog cannot represent negative effects: someone is less likely to smoke if their friends smoke (without there being a non-zero probability of logical inconsistency)

Cyclic Directed Models

- Make model acyclic, by totally ordering variables.
  Destroys exchangeability. Symmetries are not preserved.

Cyclic Directed Models

- Make model acyclic, by totally ordering variables.
  Destroys exchangeability. Symmetries are not preserved.
- (Relational) dependency networks: directed model,



  - $P(A, B)$ has 3 degrees of freedom,
  - $P(A \mid B), P(B \mid A)$, uses 4 numbers; typically inconsistent.
  - resulting distribution means stationary (equilibrium) distribution of Markov chain.

Outline

1. Knowledge Graphs
   - Triples and Reification
   - Tensor Factorization and Neural Network Models
2. Representation Issues
   - Desiderata
   - How do relational models relate to probabilistic graphical models
3. Unique properties of relational models
   - Learning general knowledge vs learning about a data set
   - Varying Populations
   - What can be observed?
4. Conclusions and Challenges

Learning general knowledge vs learning about a data set

- Suppose you want to create a model of who is friends with whom.

Learning general knowledge vs learning about a data set

- Suppose you want to create a model of who is friends with whom. Options:
  - learn general knowledge, e.g., transitivity, how male and female friendships work, how location affect friendship...

Learning general knowledge vs learning about a data set

- Suppose you want to create a model of who is friends with whom. Options:
  - learn general knowledge, e.g., transitivity, how male and female friendships work, how location affect friendship...
  - learn specific knowledge about who is friends with who; e.g., which particular group of people are generally friends with each other.

Learning general knowledge vs learning about a data set

- Suppose you want to create a model of who is friends with whom. Options:
  - learn general knowledge, e.g., transitivity, how male and female friendships work, how location affect friendship...
  - learn specific knowledge about who is friends with who; e.g., which particular group of people are generally friends with each other.
- The specific knowledge will tend to be more accurate on that population, but doesn't generalize to different populations.
- The general knowledge will tend to transfer better.
- Which is better depends on the goals and how success is measured.

Learning general knowledge vs learning about a data set

- Suppose you want to create a model of who is friends with whom. Options:
  - learn general knowledge, e.g., transitivity, how male and female friendships work, how location affect friendship...
  - learn specific knowledge about who is friends with who; e.g., which particular group of people are generally friends with each other.
- The specific knowledge will tend to be more accurate on that population, but doesn't generalize to different populations.
- The general knowledge will tend to transfer better.
- Which is better depends on the goals and how success is measured.
- Many of the methods try to do both; learn about specific individuals and general knowledge.

Canonical polyadic tensor factorization model

- Tensor factorization models — which learn vectors for individuals — tend to not learn generalized knowledge but about the particular individuals

Canonical polyadic tensor factorization model

- Tensor factorization models — which learn vectors for individuals — tend to not learn generalized knowledge but about the particular individuals
- Lifted graphical models (MLNs, RLR, Problog) learn general knowledge through training weights (and structure) and specific knowledge through conditioning.

Canonical polyadic tensor factorization model

- Tensor factorization models — which learn vectors for individuals — tend to not learn generalized knowledge but about the particular individuals
- Lifted graphical models (MLNs, RLR, Problog) learn general knowledge through training weights (and structure) and specific knowledge through conditioning.
- Still open research problem

Outline

1. Knowledge Graphs
   - Triples and Reification
   - Tensor Factorization and Neural Network Models
2. Representation Issues
   - Desiderata
   - How do relational models relate to probabilistic graphical models
3. Unique properties of relational models
   - Learning general knowledge vs learning about a data set
   - Varying Populations
   - What can be observed?
4. Conclusions and Challenges

What happens as Population size n Changes: Simplest case

$$\begin{aligned}
\alpha_0 &: q \\
\alpha_1 &: q \wedge \neg r(x) \\
\alpha_2 &: q \wedge r(x) \\
\alpha_3 &: r(x)
\end{aligned}$$

Weighted formulae define distribution:

$$P_{MLN}(q \mid n) = sigmoid(\ \alpha_0 + n \log(e^{\alpha_2} + e^{\alpha_1 - \alpha_3})\ )$$

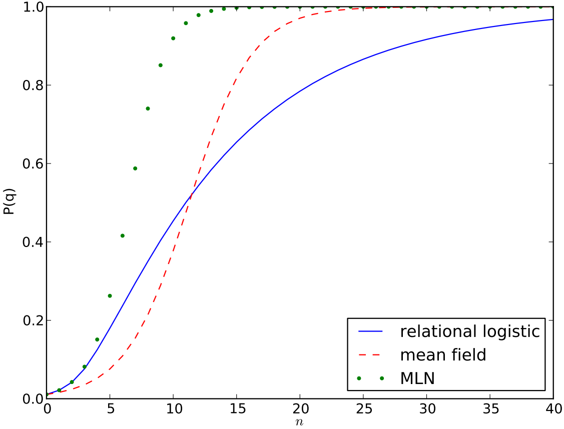
Weighted formulae define conditionals:

$$P_{RLR}(q \mid n) = \sum_{i=0}^{n} \binom{n}{i} sigmoid(\alpha_0 + i\alpha_1 + (n-i)\alpha_2)(1-p_r)^i p_r^{n-i}$$

Mean-field approximation:

$$P_{MF}(q \mid n) = sigmoid(\alpha_0 + np_r\alpha_1 + n(1-p_r)\alpha_2)$$

Population Growth: $P(q \mid n)$



All: $P(q \mid n) \rightarrow$ 0 or 1 as $n \rightarrow \infty$

Challenges of varying populations

- Example: The Movielens 100k dataset contains data about $rated(P, M, R, T)$ meaning person P gave movie M a rating of R at time T.
  Plus user demographic and movie information.
- Number of ratings per user is between 20 (arbitrary threshold) and 737; average of 106
- Number of ratings per movie is between 1 and 583; average of 59

Challenges of varying populations

- Example: The Movielens 100k dataset contains data about $rated(P, M, R, T)$ meaning person P gave movie M a rating of R at time T.
  Plus user demographic and movie information.
- Number of ratings per user is between 20 (arbitrary threshold) and 737; average of 106
- Number of ratings per movie is between 1 and 583; average of 59
- Predicting age from ratings is difficult:
  - With some high counts we can't assume movies produce independent evidence

Challenges of varying populations

- Example: The Movielens 100k dataset contains data about $rated(P, M, R, T)$ meaning person P gave movie M a rating of R at time T.
  Plus user demographic and movie information.
- Number of ratings per user is between 20 (arbitrary threshold) and 737; average of 106
- Number of ratings per movie is between 1 and 583; average of 59
- Predicting age from ratings is difficult:
  - With some high counts we can't assume movies produce independent evidence
  - With many low counts we need to regularize (and can't measure dependence)

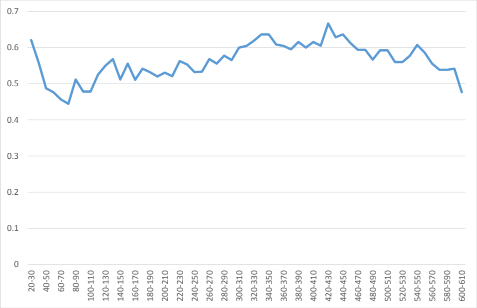
Challenges of varying populations

- Example: The Movielens 100k dataset contains data about $rated(P, M, R, T)$ meaning person P gave movie M a rating of R at time T.
  Plus user demographic and movie information.
- Number of ratings per user is between 20 (arbitrary threshold) and 737; average of 106
- Number of ratings per movie is between 1 and 583; average of 59
- Predicting age from ratings is difficult:
  - With some high counts we can't assume movies produce independent evidence
  - With many low counts we need to regularize (and can't measure dependence)
  - There is 100 times as much ratings information as age information

Challenges of varying populations

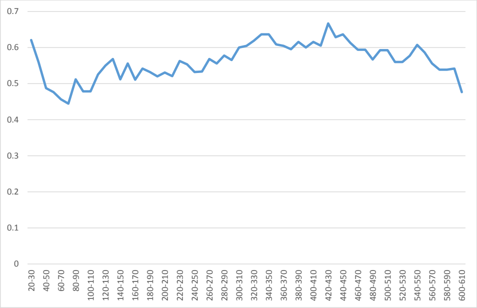
- Example: The Movielens 100k dataset contains data about *rated*(P, M, R, T) meaning person P gave movie M a rating of R at time T.
  Plus user demographic and movie information.
- Number of ratings per user is between 20 (arbitrary threshold) and 737; average of 106
- Number of ratings per movie is between 1 and 583; average of 59
- Predicting age from ratings is difficult:
  - With some high counts we can't assume movies produce independent evidence
  - With many low counts we need to regularize (and can't measure dependence)
  - There is 100 times as much ratings information as age information
- Bigger datasets have even more variability.

Real Data



Observed $P(25 < Age(u) < 45 \mid n)$, where n is number of movies watched from the Movielens dataset.

Real Data



Observed $P(25 < Age(u) < 45 \mid n)$, where n is number of movies watched from the Movielens dataset.
Dont use:

$$w : middle\_age(U) \leftarrow rated(U, M) \land foo(M)$$

then $P(middle\_age(U)) \rightarrow$ 0 or 1 as number of movies increases.

Example of polynomial dependence of population

$\alpha_0 : q$

$\alpha_1 : q \wedge true(X)$

$\alpha_2 : q \wedge r(X)$

$\alpha_3 : true(X)$

$\alpha_4 : r(X)$

$\alpha_5 : q \wedge true(X) \wedge true(Y)$

$\alpha_6 : q \wedge r(X) \wedge true(Y)$

$\alpha_7 : q \wedge r(X) \wedge r(Y)$

Example of polynomial dependence of population

$$\alpha_0 : q$$
$$\alpha_1 : q \wedge true(X)$$
$$\alpha_2 : q \wedge r(X)$$
$$\alpha_3 : true(X)$$
$$\alpha_4 : r(X)$$
$$\alpha_5 : q \wedge true(X) \wedge true(Y)$$
$$\alpha_6 : q \wedge r(X) \wedge true(Y)$$
$$\alpha_7 : q \wedge r(X) \wedge r(Y)$$

In RLR and in MLN, if all $r(a_i)$ are observed:

Example of polynomial dependence of population

$$
\begin{aligned}
\alpha_0 &: q \\
\alpha_1 &: q \wedge true(X) \\
\alpha_2 &: q \wedge r(X) \\
\alpha_3 &: true(X) \\
\alpha_4 &: r(X) \\
\alpha_5 &: q \wedge true(X) \wedge true(Y) \\
\alpha_6 &: q \wedge r(X) \wedge true(Y) \\
\alpha_7 &: q \wedge r(X) \wedge r(Y)
\end{aligned}
$$

In RLR and in MLN, if all $r(a_i)$ are observed:

$$P(q \mid obs) = sigmoid(\alpha_0 + n\alpha_1 + n_1\alpha_2 + n^2\alpha_5 + n_1 n\alpha_6 + n_1^2\alpha_7)$$

$r(X)$ is true for n_1 individuals out of a population of n.

Danger of fitting to data without understanding the model

- MLN/RLR can fit sigmoid of any polynomial.
- Consider sigmoid of polynomials of degree 2:

$$sigmoid(-0.01n^2 - 0.2n + 8)$$

$$sigmoid(0.01n^2 - n + 16)$$

Both go from ≈ 1 at $n = 10$ to ≈ 0 at $n = 30$.
What happens as $n \to \infty$?

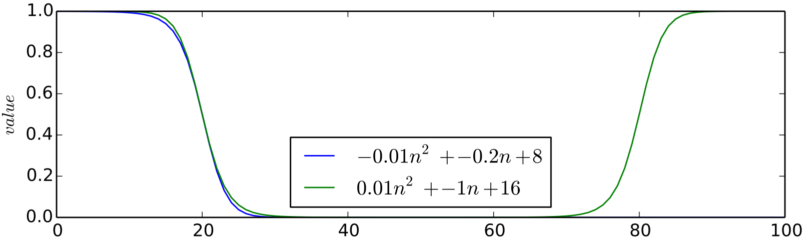
Danger of fitting to data without understanding the model

- MLN/RLR can fit sigmoid of any polynomial.
- Consider sigmoid of polynomials of degree 2:

$$sigmoid(-0.01n^2 - 0.2n + 8)$$

$$sigmoid(0.01n^2 - n + 16)$$

Both go from ≈ 1 at $n = 10$ to ≈ 0 at $n = 30$.
What happens as $n \to \infty$?



Outline

1. Knowledge Graphs
   - Triples and Reification
   - Tensor Factorization and Neural Network Models
2. Representation Issues
   - Desiderata
   - How do relational models relate to probabilistic graphical models
3. Unique properties of relational models
   - Learning general knowledge vs learning about a data set
   - Varying Populations
   - What can be observed?
4. Conclusions and Challenges

Observation Protocols

- Example: we want to predict the probability that Sam will like a apartment.
- Observation: there is a pink bedroom.

Observation Protocols

- Example: we want to predict the probability that Sam will like a apartment.
- Observation: there is a pink bedroom.

The protocol of how the observation was obtained matters:

- Was it a pink sensor that reported what was pink?

Observation Protocols

- Example: we want to predict the probability that Sam will like a apartment.
- Observation: there is a pink bedroom.

The protocol of how the observation was obtained matters:

- Was it a pink sensor that reported what was pink?
- Was it reporting the first thing it observed?

Observation Protocols

- Example: we want to predict the probability that Sam will like a apartment.
- Observation: there is a pink bedroom.

The protocol of how the observation was obtained matters:

- Was it a pink sensor that reported what was pink?
- Was it reporting the first thing it observed?
- Was it reporting the most unusual feature of the apartment?

Observation Protocols

- Example: we want to predict the probability that Sam will like a apartment.
- Observation: there is a pink bedroom.

The protocol of how the observation was obtained matters:

- Was it a pink sensor that reported what was pink?
- Was it reporting the first thing it observed?
- Was it reporting the most unusual feature of the apartment?
- Was it telling us the most positive aspect of the apartment?

Observation Protocols

- Example: we want to predict the probability that Sam will like a apartment.
- Observation: there is a pink bedroom.

The protocol of how the observation was obtained matters:

- Was it a pink sensor that reported what was pink?
- Was it reporting the first thing it observed?
- Was it reporting the most unusual feature of the apartment?
- Was it telling us the most positive aspect of the apartment?
- Did it know that Sam was interested in whether there was a pink bedroom?

Observation Protocols

- Example: we want to predict the probability that Sam will like a apartment.
- Observation: there is a pink bedroom.

The protocol of how the observation was obtained matters:

- Was it a pink sensor that reported what was pink?
- Was it reporting the first thing it observed?
- Was it reporting the most unusual feature of the apartment?
- Was it telling us the most positive aspect of the apartment?
- Did it know that Sam was interested in whether there was a pink bedroom?
- Was it reporting the colour of each bedroom?

Observation Protocols

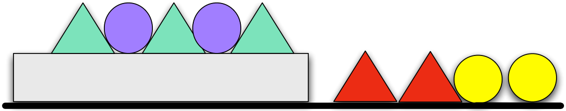
- Example: we want to predict the probability that Sam will like a apartment.
- Observation: there is a pink bedroom.

The protocol of how the observation was obtained matters:

- Was it a pink sensor that reported what was pink?
- Was it reporting the first thing it observed?
- Was it reporting the most unusual feature of the apartment?
- Was it telling us the most positive aspect of the apartment?
- Did it know that Sam was interested in whether there was a pink bedroom?
- Was it reporting the colour of each bedroom?

There are unboundedly many possible relations in a real-world object such as a house.

Observation Protocols

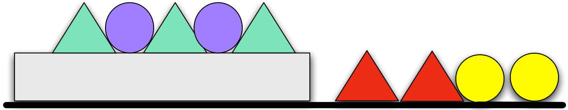


Observe a triangle and a circle touching. What is the probability the triangle is green?

$$P(green(x) \mid triangle(x) \land \exists y \; circle(y) \land touching(x, y))$$

The answer depends on how the x and y were chosen!

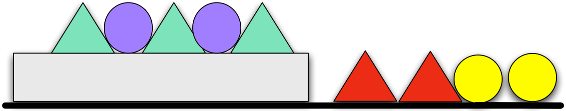
Protocol for Observing



$$P(green(x) \mid triangle(x) \wedge \exists y \; circle(y) \wedge touching(x, y))$$

$select(x)$	$select(y)$	$select(x, y)$
$select(y)$	$select(x)$	
$3/4$	$2/3$	$4/5$

Protocol for Observing



$$P(green(x) \mid triangle(x) \land \exists y \; circle(y) \land touching(x, y))$$

$select(x)$	$select(y)$	$select(x, y)$
$select(y)$	$select(x)$	
$3/4$	$2/3$	$4/5$

A logical formula does not provide enough information to determine the probabilities.

Data

Real data is messy!

- Multiple levels of abstraction
- Multiple levels of detail
- Sometimes observations are abstract and lifted
  e.g., "3 people out of 300 in the audience asked a question".
- Uses the vocabulary from many ontologies
- Rich meta-data:
  - Who collected each datum? (identity and credentials)
  - Who transcribed the information?
  - What was the protocol used to collect the data? (Chosen at random or chosen because interesting?)
  - What were the controls — what was manipulated, when?
  - What sensors were used? What is their reliability and operating range?
  - What is the provenance of the data; what was done to it when?
- Errors, forgeries, . . .

Outline

1 Knowledge Graphs
- Triples and Reification
- Tensor Factorization and Neural Network Models

2 Representation Issues
- Desiderata
- How do relational models relate to probabilistic graphical models

3 Unique properties of relational models
- Learning general knowledge vs learning about a data set
- Varying Populations
- What can be observed?

4 Conclusions and Challenges

Take Home

- Exchangeability and dependence on population size distinguish relational models from non-relational models

Take Home

- Exchangeability and dependence on population size distinguish relational models from non-relational models
- Relational models are different from normal graphical models

Take Home

- Exchangeability and dependence on population size distinguish relational models from non-relational models
- Relational models are different from normal graphical models

Challenges that relational models tackle:

- Heterogeneity: information about individuals varies greatly in kind and amount

Take Home

- Exchangeability and dependence on population size distinguish relational models from non-relational models
- Relational models are different from normal graphical models

Challenges that relational models tackle:

- Heterogeneity: information about individuals varies greatly in kind and amount
- Representations should
  - let people state their prior knowledge,
  - let them understand what they stated, and
  - let them understand the posterior models (given evidence).

Take Home

- Exchangeability and dependence on population size distinguish relational models from non-relational models
- Relational models are different from normal graphical models

Challenges that relational models tackle:

- Heterogeneity: information about individuals varies greatly in kind and amount
- Representations should
  - let people state their prior knowledge,
  - let them understand what they stated, and
  - let them understand the posterior models (given evidence).
- Learn general knowledge as well as about particular individuals

Take Home

- Exchangeability and dependence on population size distinguish relational models from non-relational models
- Relational models are different from normal graphical models

Challenges that relational models tackle:

- Heterogeneity: information about individuals varies greatly in kind and amount
- Representations should
  - let people state their prior knowledge,
  - let them understand what they stated, and
  - let them understand the posterior models (given evidence).
- Learn general knowledge as well as about particular individuals
- Use the meta-data of how data was collected
- Model protocol used to generate the observations
- Also model what is not observed (e.g., because it was redundant information, unimportant, false or unknown)

What is now required is to give the greatest possible development to mathematical logic, to allow to the full the importance of relations, and then to found upon this secure basis a new philosophical logic, which may hope to borrow some of the exactitude and certainty of its mathematical foundation. If this can be successfully accomplished, there is every reason to hope that the near future will be as great an epoch in pure philosophy as the immediate past has been in the principles of mathematics. Great triumphs inspire great hopes; and pure thought may achieve, within our generation, such results as will place our time, in this respect, on a level with the greatest age of Greece.

– Bertrand Russell 1917