Logic, Probability and Computation: Statistical Relational AI and Beyond

David Poole

Department of Computer Science,
University of British Columbia

November 2015

Outline

1. Logic and Probability
   - Relational Probabilistic Models
   - Probabilistic Logic Programs
2. Lifted Inference
3. Undirected models, Directed models, and Weighted Formulae
4. Existence and Identity Uncertainty

First-order Predicate Calculus

The world (we want to represent) is made up of individuals (things) and relationships among individuals.

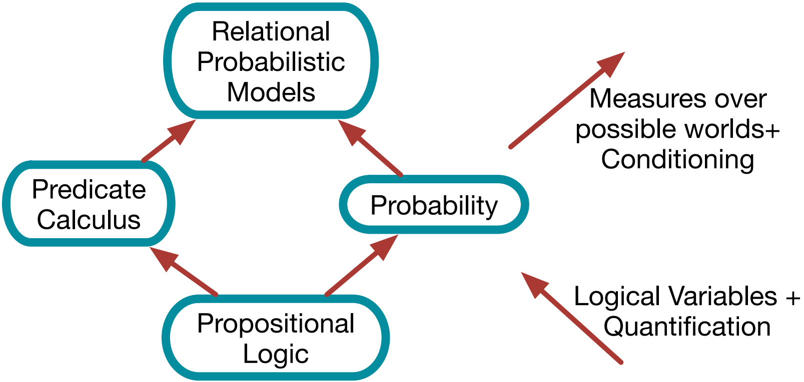
Classical (first order) logic lets us represent:

- individuals in the world
- relations amongst those individuals
- conjunctions, disjunctions, negations of relations
- quantification over individuals

Why Probability?

- There is lots of uncertainty about the world, but agents still need to act.
- Predictions are needed to decide what to do:
  - definitive predictions: you will be run over tomorrow
  - point probabilities: probability you will be run over tomorrow is 0.002 if you are not careful and 0.000001 if you are careful.
  - probability ranges: you will be run over with probability in range [0.001,0.34]
- Acting is gambling: agents who don't use probabilities will lose to those who do — Dutch books.
- Probabilities can be learned from data.
  Bayes' rule specifies how to combine data and prior knowledge.

Statistical Relational AI



Bayes' Rule

$$P(h|e) = \frac{\overbrace{P(e|h)}^{\text{Likelihood}}\ \overbrace{P(h)}^{\text{Prior}}}{\underbrace{P(e)}_{\text{Normalizing constant}}}$$

Bayes' Rule

Likelihood Prior

$$P(h|e) = \frac{P(e|h)\ P(h)}{P(e)}$$

Normalizing constant

- What if e is a patient's electronic health record?

Bayes' Rule

$$P(h|e) = \frac{\overbrace{P(e|h)}^{\text{Likelihood}}\ \ \overbrace{P(h)}^{\text{Prior}}}{\underbrace{P(e)}_{\text{Normalizing constant}}}$$

- What if e is a patient's electronic health record?
- What if e is the electronic health records for all of the people in the province?

Bayes' Rule

Likelihood Prior

$$P(h|e) = \frac{P(e|h)\ P(h)}{P(e)}$$

Normalizing constant

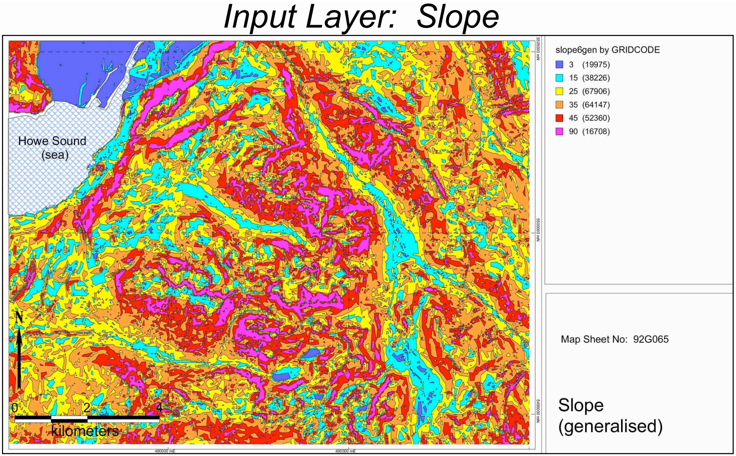
- What if e is a patient's electronic health record?
- What if e is the electronic health records for all of the people in the province?
- What if e is a collection of student records in a university?

Bayes' Rule

$$P(h|e) = \frac{\overbrace{P(e|h)}^{\text{Likelihood}}\ \ \overbrace{P(h)}^{\text{Prior}}}{\underbrace{P(e)}_{\text{Normalizing constant}}}$$

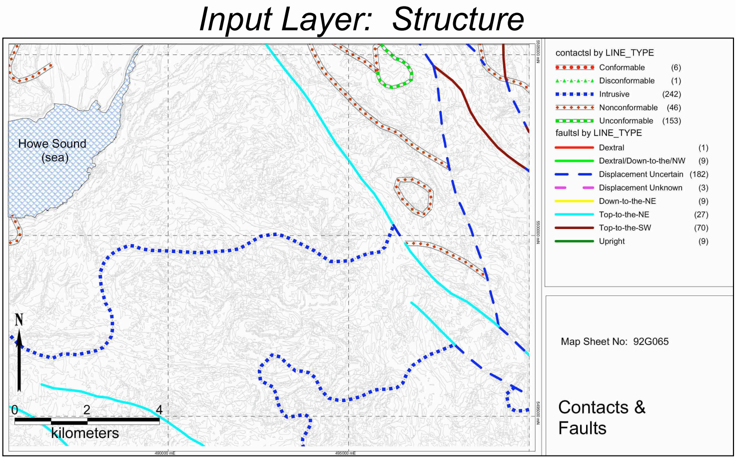
- What if *e* is a patient's electronic health record?
- What if *e* is the electronic health records for all of the people in the province?
- What if *e* is a collection of student records in a university?
- What if *e* is a description of everything known about the geology of Earth?

Example Observation, Geology



[Clinton Smyth, Georeference Online.]

Example Observation, Geology



[Clinton Smyth, Georeference Online.]

Outline

1. Logic and Probability
   - Relational Probabilistic Models
   - Probabilistic Logic Programs
2. Lifted Inference
3. Undirected models, Directed models, and Weighted Formulae
4. Existence and Identity Uncertainty

Relational Learning

- Machine learning typically assumes informative feature values. But often the values are names of individuals.
- It is the properties of these individuals and their relationship to other individuals that needs to be learned.
- Relational learning has been studied under the umbrella of "Inductive Logic Programming" as the representations were traditionally logic programs.

Example: trading agent

What does Joe like?

Individual	Property	Value
joe	*likes*	*resort_14*
joe	*dislikes*	*resort_35*
...	...	...
resort_14	*type*	*resort*
resort_14	*near*	*beach_18*
beach_18	*type*	*beach*
beach_18	*covered_in*	*ws*
ws	*type*	*sand*
ws	*color*	*white*
...	...	...

Example: trading agent

Possible hypothesis that could be learned:

$$
\begin{aligned}
& prop(joe, likes, R) \leftarrow \\
& \qquad prop(R, type, resort) \wedge \\
& \qquad prop(R, near, B) \wedge \\
& \qquad prop(B, type, beach) \wedge \\
& \qquad prop(B, covered\_in, S) \wedge \\
& \qquad prop(S, type, sand).
\end{aligned}
$$

"Joe likes resorts that are near sandy beaches."

Example: trading agent

Possible hypothesis that could be learned:

$$
\begin{aligned}
& prop(joe, likes, R) \leftarrow \\
& \qquad prop(R, type, resort) \wedge \\
& \qquad prop(R, near, B) \wedge \\
& \qquad prop(B, type, beach) \wedge \\
& \qquad prop(B, covered\_in, S) \wedge \\
& \qquad prop(S, type, sand).
\end{aligned}
$$

"Joe likes resorts that are near sandy beaches."

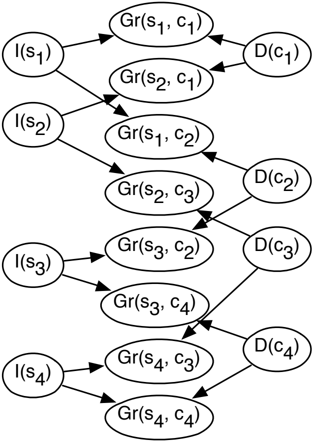
- But we want probabilistic predictions.

Example: Predicting Relations

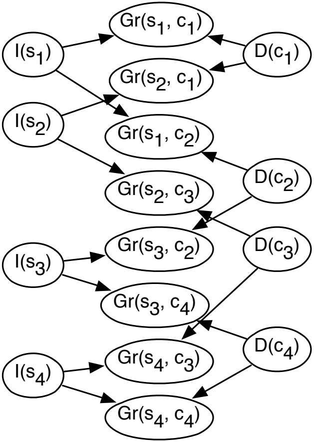
Student	*Course*	*Grade*
s_1	c_1	A
s_2	c_1	C
s_1	c_2	B
s_2	c_3	B
s_3	c_2	B
s_4	c_3	B
s_3	c_4	?
s_4	c_4	?

- Students s_3 and s_4 have the same averages, on courses with the same averages.
- Which student would you expect to better?

From Relations to Bayesian Belief Networks



From Relations to Bayesian Belief Networks



$I(S)$	$D(C)$	$Gr(S,C)$		
		A	B	C
true	*true*	0.5	0.4	0.1
true	*false*	0.9	0.09	0.01
false	*true*	0.01	0.09	0.9
false	*false*	0.1	0.4	0.5

$$P(I(S)) = 0.5$$
$$P(D(C)) = 0.5$$

"parameter sharing"

Example: Predicting Relations

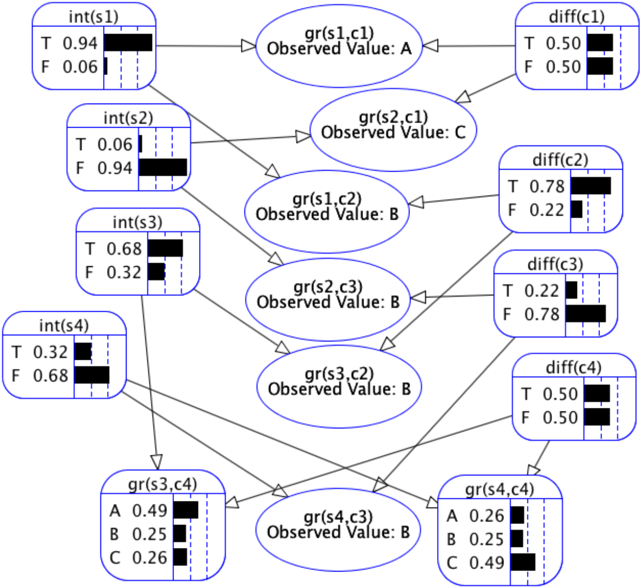
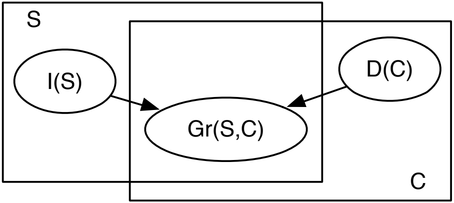
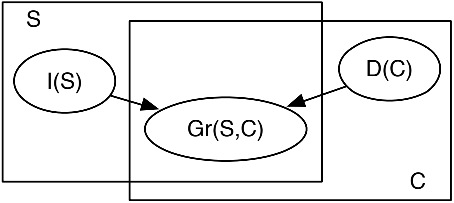


Plate Notation



- S, C logical variable representing students, courses
- the set of individuals of a type is called a population
- $I(S)$, $Gr(S, C)$, $D(C)$ are parametrized random variables

Plate Notation

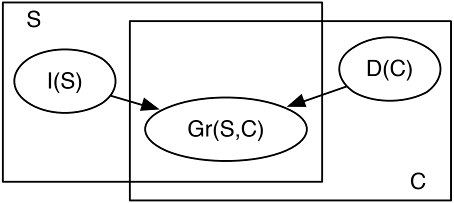


- S, C logical variable representing students, courses
- the set of individuals of a type is called a population
- $I(S)$, $Gr(S,C)$, $D(C)$ are parametrized random variables

Grounding:

- for every student s, there is a random variable $I(s)$
- for every course c, there is a random variable $D(c)$
- for every s, c pair there is a random variable $Gr(s,c)$
- all instances share the same structure and parameters

Plate Notation



- If there were 1000 students and 100 courses:
  Grounding contains
  - 1000 $I(s)$ variables
  - 100 $D(c)$ variables
  - 100000 $Gr(s, c)$ variables

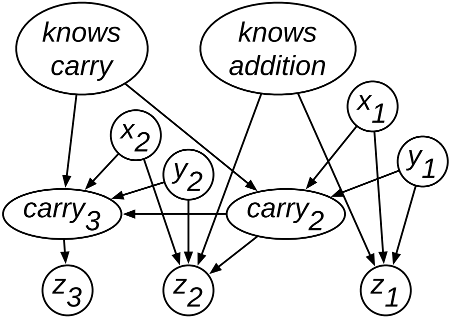
  total: 101100 variables
- Numbers to be specified to define the probabilities:
  1 for $I(S)$, 1 for $D(C)$, 8 for $Gr(S, C)$ = 10 parameters.

Bayesian Belief Networks

$$\begin{array}{rrrr} & & x_2 & x_1 \\ + & & y_2 & y_1 \\ \hline & z_3 & z_2 & z_1 \end{array}$$

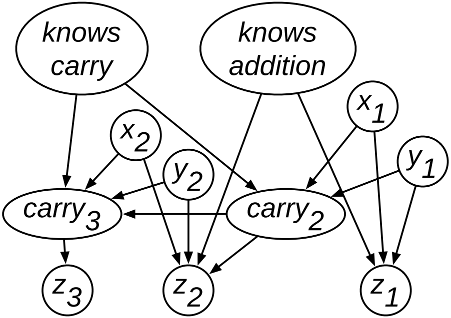
Bayesian Belief Networks

$$\begin{array}{cccc} & & x_2 & x_1 \\ + & & y_2 & y_1 \\ \hline & z_3 & z_2 & z_1 \end{array}$$



Bayesian Belief Networks

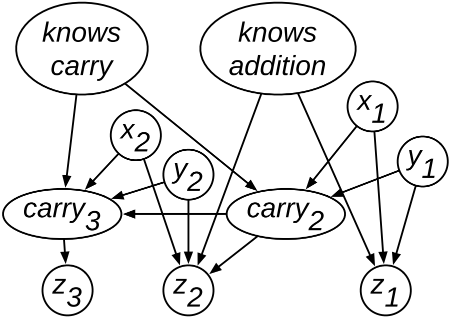
$$\begin{array}{rccc} & & x_2 & x_1 \\ + & & y_2 & y_1 \\ \hline & z_3 & z_2 & z_1 \end{array}$$



What if there were multiple digits

Bayesian Belief Networks

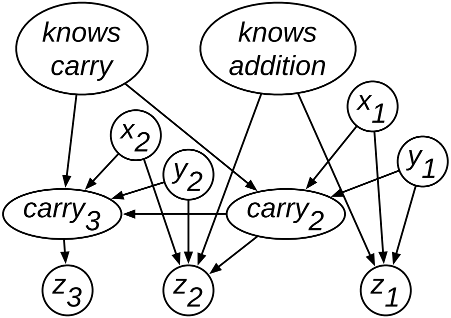
$$
\begin{array}{rccc}
 & & x_2 & x_1 \\
+ & & y_2 & y_1 \\
\hline
 & z_3 & z_2 & z_1
\end{array}
$$



What if there were multiple digits, problems

Bayesian Belief Networks

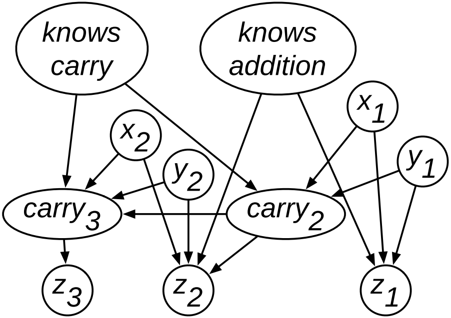
$$\begin{array}{rccc} & & x_2 & x_1 \\ + & & y_2 & y_1 \\ \hline & z_3 & z_2 & z_1 \end{array}$$



What if there were multiple digits, problems, students

Bayesian Belief Networks

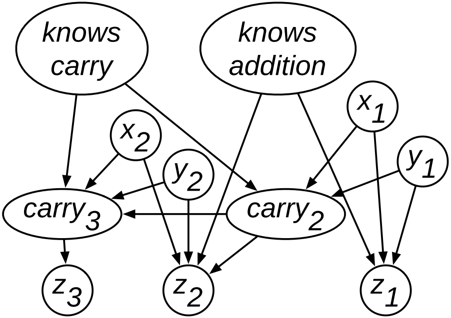
$$\begin{array}{rccc} & & x_2 & x_1 \\ + & & y_2 & y_1 \\ \hline & z_3 & z_2 & z_1 \end{array}$$



What if there were multiple digits, problems, students, times?

Bayesian Belief Networks

$$\begin{array}{cccc} & & x_2 & x_1 \\ + & & y_2 & y_1 \\ \hline & z_3 & z_2 & z_1 \end{array}$$

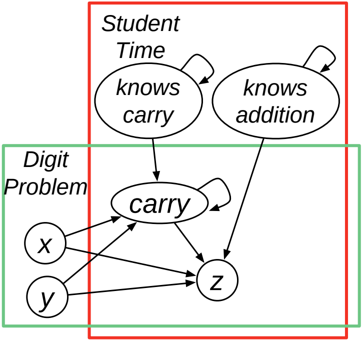


What if there were multiple digits, problems, students, times?

How can we build a model before we know the individuals?

Multi-digit addition with parametrized BNs / plates

$$
\begin{array}{ccccc}
 & x_{j_x} & \cdots & x_2 & x_1 \\
+ & y_{j_z} & \cdots & y_2 & y_1 \\
\hline
 & z_{j_z} & \cdots & z_2 & z_1
\end{array}
$$



Random Variables: $x(D, P)$, $y(D, P)$, $knowsCarry(S, T)$, $knowsAddition(S, T)$, $carry(D, P, S, T)$, $z(D, P, S, T)$
for each: digit D, problem P, student S, time T

Relational Probabilistic Models

Often we want random variables for combinations of individual in populations

- build a probabilistic model before knowing the individuals
- learn the model for one set of individuals
- apply the model to new individuals
- allow complex relationships between individuals

Exchangeability

- Before we know anything about individuals, they are indistinguishable, and so should be treated identically.

Representing Conditional Probabilities

- $P(gr(S, C) \mid int(S), diff(C))$ — parameter sharing — individuals share probability parameters.

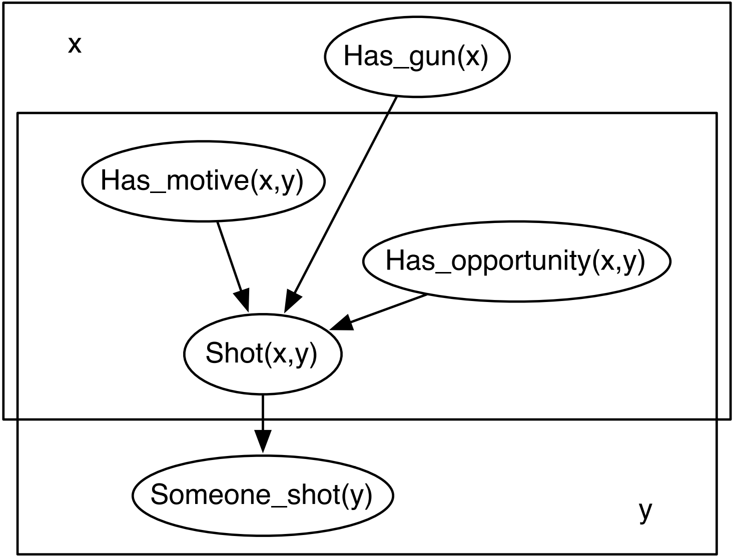
Representing Conditional Probabilities

- $P(gr(S, C) \mid int(S), diff(C))$ — parameter sharing — individuals share probability parameters.
- $P(happy(X) \mid friend(X, Y), mean(Y))$ — needs aggregation — $happy(a)$ depends on an unbounded number of parents.

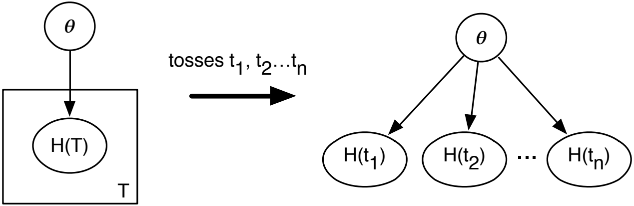
Representing Conditional Probabilities

- $P(gr(S, C) \mid int(S), diff(C))$ — parameter sharing — individuals share probability parameters.
- $P(happy(X) \mid friend(X, Y), mean(Y))$ — needs aggregation — $happy(a)$ depends on an unbounded number of parents.
- There can be more structure about the individuals
  - the carry of one digit depends on carry of the previous digit
  - probability that two authors collaborate depends on whether they have a paper authored together

Example: Aggregation



Example Plate Notation for Learning Parameters



Outline

1. Logic and Probability
   - Relational Probabilistic Models
   - Probabilistic Logic Programs
2. Lifted Inference
3. Undirected models, Directed models, and Weighted Formulae
4. Existence and Identity Uncertainty

Independent Choice Logic (ICL)

- A language for relational probabilistic models.
- Idea: combine logic and probability, where all uncertainty in handled in terms of Bayesian decision theory, and logic specifies consequences of choices.
- An ICL theory consists of a choice space with probabilities over choices and a logic program that gives consequences of choices.
- History: parametrized Bayesian belief networks, abduction and default reasoning $\longrightarrow$ probabilistic Horn abduction (IJCAI-91); richer language (negation as failure + choices by other agents $\longrightarrow$ independent choice logic (AIJ 1997) $\longrightarrow$ Problog (probabilistic programming language)

Independent Choice Logic

- An atomic hypothesis is an atomic formula.
  An alternative is a set of atomic hypotheses.
  $\mathcal{C}$, the choice space is a set of disjoint alternatives.
- $\mathcal{F}$, the facts is an acyclic logic program that gives consequences of choices (can contain negation as failure).
  No atomic hypothesis is the head of a rule.
- P_0 a probability distribution over alternatives:

$$\forall A \in \mathcal{C} \ \sum_{a \in A} P_0(a) = 1.$$

Meaningless Example

$$\mathcal{C} = \{\{c_1, c_2, c_3\}, \{b_1, b_2\}\}$$

$$\begin{aligned}
\mathcal{F} = \{ \ & f \leftarrow c_1 \wedge b_1, && f \leftarrow c_3 \wedge b_2, \\
& d \leftarrow c_1, && d \leftarrow \ \sim c_2 \wedge b_1, \\
& e \leftarrow f, && e \leftarrow \ \sim d\}
\end{aligned}$$

$$\begin{array}{lll}
P_0(c_1) = 0.5 & P_0(c_2) = 0.3 & P_0(c_3) = 0.2 \\
P_0(b_1) = 0.9 & P_0(b_2) = 0.1 &
\end{array}$$

Semantics of ICL

- There is a possible world for each selection of one element from each alternative.
- The logic program together with the selected atoms specifies what is true in each possible world.
- The elements of different alternatives are probabilistically independent.

Meaningless Example: Semantics

$$\mathcal{F} = \{ \; f \leftarrow c_1 \wedge b_1, \quad f \leftarrow c_3 \wedge b_2, \\ d \leftarrow c_1, \quad d \leftarrow \sim c_2 \wedge b_1, \\ e \leftarrow f, \quad e \leftarrow \sim d \}$$

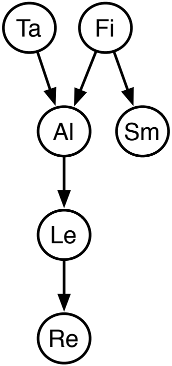
$$P_0(c_1) = 0.5 \quad P_0(c_2) = 0.3 \quad P_0(c_3) = 0.2$$
$$P_0(b_1) = 0.9 \quad P_0(b_2) = 0.1$$

		selection		logic program			
w_1	$\models$	c_1	b_1	f	d	e	$P(w_1) = 0.45$
w_2	$\models$	c_2	b_1	$\sim f$	$\sim d$	e	$P(w_2) = 0.27$
w_3	$\models$	c_3	b_1	$\sim f$	d	$\sim e$	$P(w_3) = 0.18$
w_4	$\models$	c_1	b_2	$\sim f$	d	$\sim e$	$P(w_4) = 0.05$
w_5	$\models$	c_2	b_2	$\sim f$	$\sim d$	e	$P(w_5) = 0.03$
w_6	$\models$	c_3	b_2	f	$\sim d$	e	$P(w_6) = 0.02$

$$P(e) = 0.45 + 0.27 + 0.03 + 0.02 = 0.77$$

Belief Networks, Decision trees and ICL rules

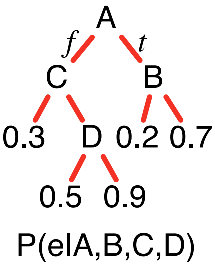
- There is a local mapping from Bayesian belief networks into ICL.



prob ta : 0.02.
prob $fire$: 0.01.
$alarm \leftarrow ta \land fire \land atf$.
$alarm \leftarrow \sim ta \land fire \land antf$.
$alarm \leftarrow ta \land \sim fire \land atnf$.
$alarm \leftarrow \sim ta \land \sim fire \land antnf$.
prob atf : 0.5.
prob $antf$: 0.99.
prob $atnf$: 0.85.
prob $antnf$: 0.0001.
$smoke \leftarrow fire \land sf$.
prob sf : 0.9.
$smoke \leftarrow \sim fire \land snf$.
prob snf : 0.01.

Belief Networks, Decision trees and ICL rules

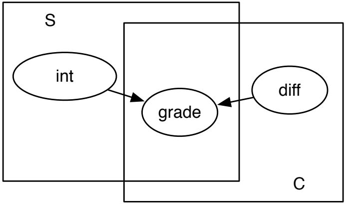
- Rules can represent decision tree with probabilities:



$$
\begin{array}{ll}
e \leftarrow a \wedge b \wedge h_1. & P_0(h_1) = 0.7 \\
e \leftarrow a \wedge \sim b \wedge h_2. & P_0(h_2) = 0.2 \\
e \leftarrow \sim a \wedge c \wedge d \wedge h_3. & P_0(h_3) = 0.9 \\
e \leftarrow \sim a \wedge c \wedge \sim d \wedge h_4. & P_0(h_4) = 0.5 \\
e \leftarrow \sim a \wedge \sim c \wedge h_5. & P_0(h_5) = 0.3
\end{array}
$$

Predicting Grades

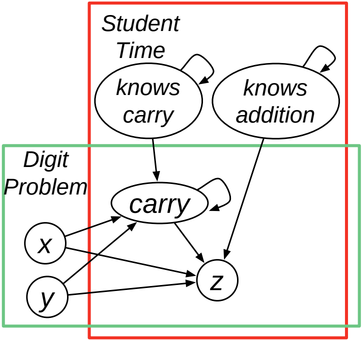
Plates correspond to logical variables.



prob $int(S) : 0.5.$
prob $diff(C) : 0.5.$
$grade(S, C, G) \leftarrow int(S) \wedge diff(C) \wedge idg(S, C, G).$
prob $idg(S, C, a) : 0.5,\ idg(S, C, b) : 0.4,\ idg(S, C, c) : 0.1.$
$grade(S, C, G) \leftarrow int(S) \wedge \sim diff(C) \wedge indg(S, C, G).$
prob $indg(S, C, a) : 0.9,\ indg(S, C, b) : 0.09,\ indg(S, C, c) : 0.01.$
...

Multi-digit addition with parametrized BNs / plates

$$
\begin{array}{ccccc}
 & x_{j_x} & \cdots & x_2 & x_1 \\
+ & y_{j_z} & \cdots & y_2 & y_1 \\
\hline
 & z_{j_z} & \cdots & z_2 & z_1
\end{array}
$$



Random Variables: $x(D, P)$, $y(D, P)$, $knowsCarry(S, T)$, $knowsAddition(S, T)$, $carry(D, P, S, T)$, $z(D, P, S, T)$
for each: digit D, problem P, student S, time T

☛ parametrized random variables

ICL rules for multi-digit addition

$$
\begin{aligned}
z(D,P,S,T) = V \leftarrow\; & \\
& x(D,P) = Vx \wedge \\
& y(D,P) = Vy \wedge \\
& carry(D,P,S,T) = Vc \wedge \\
& knowsAddition(S,T) \wedge \\
& \neg mistake(D,P,S,T) \wedge \\
& V \text{ is } (Vx + Vy + Vc) \text{ div } 10.
\end{aligned}
$$

$$
\begin{aligned}
z(D,P,S,T) = V \leftarrow\; & \\
& knowsAddition(S,T) \wedge \\
& mistake(D,P,S,T) \wedge \\
& selectDig(D,P,S,T) = V. \\
z(D,P,S,T) = V \leftarrow\; & \\
& \neg knowsAddition(S,T) \wedge \\
& selectDig(D,P,S,T) = V.
\end{aligned}
$$

Alternatives:

$\forall DPST \{noMistake(D,P,S,T), mistake(D,P,S,T)\}$

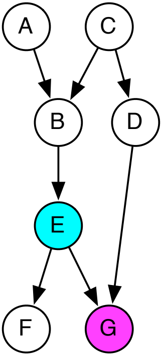
$\forall DPST \{selectDig(D,P,S,T) = V \mid V \in \{0..9\}\}$

Outline

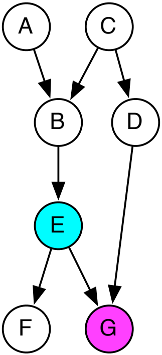
1. Logic and Probability
   - Relational Probabilistic Models
   - Probabilistic Logic Programs
2. Lifted Inference
3. Undirected models, Directed models, and Weighted Formulae
4. Existence and Identity Uncertainty

Bayesian Belief Network Inference

$$P(E \mid g) = \frac{P(E \wedge g)}{\sum_E P(E \wedge g)}$$



Bayesian Belief Network Inference

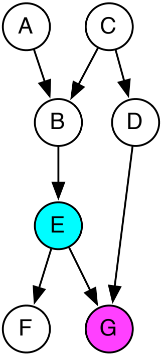


$$P(E \mid g) = \frac{P(E \wedge g)}{\sum_E P(E \wedge g)}$$

$$P(E \wedge g) = \sum_F \sum_B \sum_C \sum_A \sum_D P(A)P(B \mid AC)$$

$$P(C)P(D \mid C)P(E \mid B)P(F \mid E)P(g \mid ED)$$

Bayesian Belief Network Inference



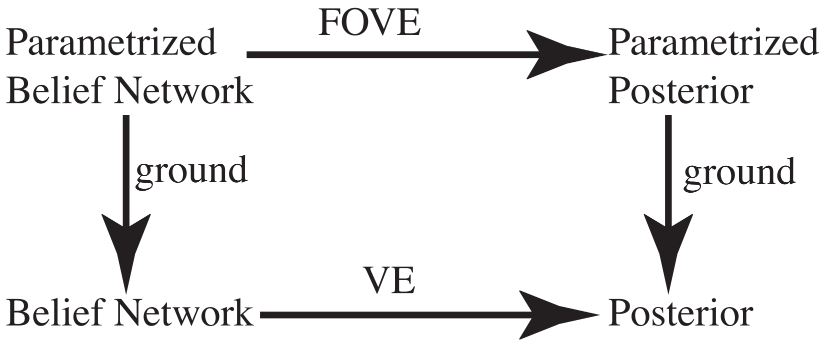
$$P(E \mid g) = \frac{P(E \wedge g)}{\sum_E P(E \wedge g)}$$

$$P(E \wedge g) = \sum_F \sum_B \sum_C \sum_A \sum_D P(A)P(B \mid AC)$$
$$P(C)P(D \mid C)P(E \mid B)P(F \mid E)P(g \mid ED)$$
$$= \left(\sum_F P(F \mid E)\right)$$
$$\sum_B P(e \mid B) \sum_C P(C) \left(\sum_A P(A)P(B \mid AC)\right)$$
$$\left(\sum_D P(D \mid C)P(g \mid ED)\right)$$

Lifted Inference

- Idea: treat those individuals about which you have the same information as a block; just count them.
- Use the ideas from lifted theorem proving - no need to ground.
- Potential to be exponentially faster in the number of non-differentialed individuals.
- Relies on knowing the number of individuals (the population size).

First-order probabilistic inference



Queries depend on population size

Suppose we observe:

- Joe has purple hair, a purple car, and has big feet.
- A person with purple hair, a purple car, and who is very tall was seen committing a crime.

What is the probability that Joe is guilty?

Theorem Proving and Unification

In 1965, Robinson showed how unification allows many ground steps with one step:

$$\underbrace{f(X,Z) \vee h(X,a) \qquad \neg h(b,Y) \vee g(Y,W)}_{f(b,Z) \vee g(a,W)}$$

Substitution $\{X/b, Y/a\}$ is the most general unifier of $h(X,a)$ and $h(b,Y)$.

Variable Elimination and Unification

- Multiplying parametrized factors:

$$\underbrace{[f(X,Z),h(X,a)] \quad \times \quad [h(b,Y),g(Y,W)]}_{[f(b,Z),h(b,a),g(a,W)]}$$

  Doesn't quite work because the first parametrized factor can't subsequently be used for $X = b$ but can be used for other instances of X.

- We split $[f(X,Z),h(X,a)]$ into

  $[f(b,Z),h(b,a)]$
  $[f(X,Z),h(X,a)]$ with constraint $X \neq b$,

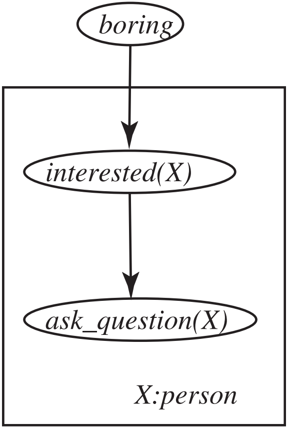
Parametric Factors

A parametric factor is a triple $\langle C, V, t\rangle$ where

- C is a set of inequality constraints on parameters,
- V is a set of parametrized random variables
- t is a table representing a factor from the random variables to the non-negative reals.

$$\left\langle \{X \neq sue\}, \{interested(X), boring\}, \begin{array}{|ll|l|} \hline interested & boring & Val \\ \hline yes & yes & 0.001 \\ yes & no & 0.01 \\ & \dots & \\ \hline \end{array} \right\rangle$$

Removing a parameter when summing



$|people| = n$

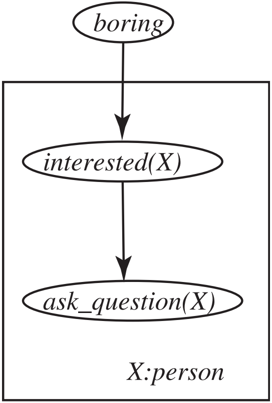
we observe no questions

Eliminate *interested*:

$$\langle \{\}, \{boring, interested(X)\}, t_1 \rangle$$
$$\langle \{\}, \{interested(X)\}, t_2 \rangle$$
$$\downarrow$$
$$\langle \{\}, \{boring\}, (t_1 \times t_2)^n \rangle$$

$(t_1 \times t_2)^n$ is computed pointwise; constant time (to fixed precision).

Counting Elimination



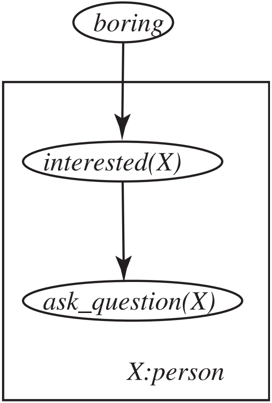
$|people| = n$

Eliminate *boring*:

VE: factor on $\{interested(p_1), \ldots, interested(p_n)\}$

Size is $O(d^n)$ where d is size of range of interested.

Counting Elimination



$|people| = n$

[de Salvo Braz et al. 2007] and [Milch et al. 08]

Eliminate *boring*:

VE: factor on $\{interested(p_1), \ldots, interested(p_n)\}$

Size is $O(d^n)$ where d is size of range of interested.

Exchangeable: only the number of interested individuals matters.

Counting Formula:

#interested	Value
0	v_0
1	v_1
...	...
n	v_n

Complexity: $O(n^{d-1})$.

Potential of Lifted Inference

- Lifting reduces complexity:

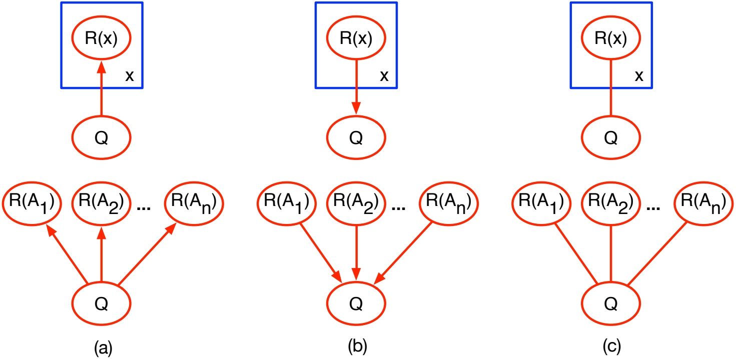
  $$\textit{polynomial} \longrightarrow \textit{logarithmic}$$

  $$\textit{exponential} \longrightarrow \textit{polynomial}$$

  in the population size of undifferentiated individuals compared to grounding
- We can now lift all unary relations, but we know we can't do all binary relations [Guy Van den Broeck, 2013].
  Always exponentially faster.
- Current most efficient algorithm compile to secondary representations. (E.g. Mehran Kazemi compiles to C++).
- Great potential for approximate inference

Outline

Three Elementary Models

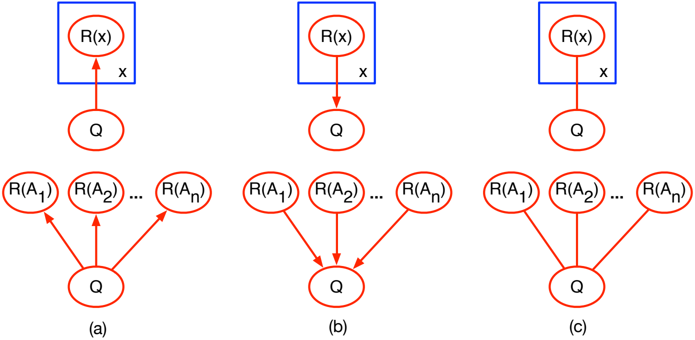


(a) Naïve Bayes

(b) (Relational) Logistic Regression

(c) Markov network

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.

Logistic Regression

Logistic Regression, write $R(a_i)$ as R_i:

$$P(Q|R_1,\ldots,R_n) = sigmoid(w_0 + w_1 R_1 + \cdots + w_n R_n)$$

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

Logistic Regression

Logistic Regression, write $R(a_i)$ as R_i:

$$P(Q|R_1, \ldots, R_n) = sigmoid(w_0 + w_1 R_1 + \cdots + w_n R_n)$$

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

If all of the R_i are exchangeable $w_1, \ldots, w_n$ must all be the same:

$$P(Q|R_1, \ldots, R_n) = sigmoid(w_0 + w_1 \sum_i R_i))$$

Logistic Regression

Logistic Regression, write $R(a_i)$ as R_i:

$$P(Q|R_1,\ldots,R_n) = sigmoid(w_0 + w_1 R_1 + \cdots + w_n R_n)$$

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

If all of the R_i are exchangeable $w_1,\ldots,w_n$ must all be the same:

$$P(Q|R_1,\ldots,R_n) = sigmoid(w_0 + w_1 \sum_i R_i))$$

If we learn the parameters for $n=10$ the prediction for $n=20$ depends on how values R_i are represented numerically:

- If $True=1$ and $False=0$ then $P(Q|R_1,\ldots,R_n)$ depends on the number of R_i that are true.

Logistic Regression

Logistic Regression, write $R(a_i)$ as R_i:

$$P(Q|R_1,\ldots,R_n) = sigmoid(w_0 + w_1 R_1 + \cdots + w_n R_n)$$

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

If all of the R_i are exchangeable $w_1,\ldots,w_n$ must all be the same:

$$P(Q|R_1,\ldots,R_n) = sigmoid(w_0 + w_1 \sum_i R_i))$$

If we learn the parameters for $n = 10$ the prediction for $n = 20$ depends on how values R_i are represented numerically:

- If $True = 1$ and $False = 0$ then $P(Q|R_1,\ldots,R_n)$ depends on the number of R_i that are true.
- If $True = 1$ and $False = -1$ then $P(Q|R_1,\ldots,R_n)$ depends on how many more of R_i are true than false.

Logistic Regression

Logistic Regression, write $R(a_i)$ as R_i:

$$P(Q|R_1, \ldots, R_n) = sigmoid(w_0 + w_1 R_1 + \cdots + w_n R_n)$$

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

If all of the R_i are exchangeable $w_1, \ldots, w_n$ must all be the same:

$$P(Q|R_1, \ldots, R_n) = sigmoid(w_0 + w_1 \sum_i R_i))$$

If we learn the parameters for $n = 10$ the prediction for $n = 20$ depends on how values R_i are represented numerically:

- If $True = 1$ and $False = 0$ then $P(Q|R_1, \ldots, R_n)$ depends on the number of R_i that are true.
- If $True = 1$ and $False = -1$ then $P(Q|R_1, \ldots, R_n)$ depends on how many more of R_i are true than false.
- If $True = 0$ and $False = -1$ then $P(Q|R_1, \ldots, R_n)$ depends on the number of R_i that are false.

Directed and Undirected models

- Weighted formula (WF): $\langle L, F, w \rangle$
  - L is a set of logical variables,
  - F is a logical formula: $\{$free logical variables in $F\} \subseteq L$
  - w is a real-valued weight.
- Instances of weighted formule obtained by assigning individuals to variables in L.

Directed and Undirected models

- Weighted formula (WF): $\langle L, F, w \rangle$
    - L is a set of logical variables,
    - F is a logical formula: $\{$free logical variables in $F\} \subseteq L$
    - w is a real-valued weight.
- Instances of weighted formule obtained by assigning individuals to variables in L.
- A world is an assignment of a value to each ground instance of each atom.
- Markov logic network (MLN): "undirected model" weighted formulae define measures on worlds.

Directed and Undirected models

- Weighted formula (WF): $\langle L, F, w\rangle$
  - L is a set of logical variables,
  - F is a logical formula: $\{$free logical variables in $F\} \subseteq L$
  - w is a real-valued weight.
- Instances of weighted formule obtained by assigning individuals to variables in L.
- A world is an assignment of a value to each ground instance of each atom.
- Markov logic network (MLN): "undirected model" weighted formulae define measures on worlds.
- Relational logistic regression (RLR): "directed model" weighted formulae define conditional probabilities.

Weighted formulae for conditionals $\rightarrow$ logistic regression

Weighted formulae:

$$\langle\{x\}, \mathit{funFor}(x), -5\rangle$$
$$\langle\{x, y\}, \mathit{funFor}(x) \wedge \mathit{friends}(x, y) \wedge \mathit{social}(y), 10\rangle$$
$$\langle\{x, y\}, \mathit{funFor}(x) \wedge \mathit{friends}(x, y) \wedge \neg\mathit{social}(y), -3\rangle$$

If *obs* includes observations for all $\mathit{friends}(x, y)$ and $\mathit{social}(y)$:

$$P(\mathit{funFor}(x) \mid \mathit{obs}) = \mathit{sigmoid}(-5 + 10n_s(x) - 3n_a(x))$$

$$n_s(x) = |\{y \mid \mathit{friends}(x, y) \wedge \mathit{social}(y)\}|$$

$$n_a(x) = |\{y \mid \mathit{friends}(x, y) \wedge \neg\mathit{social}(y)\}|$$

Weighted formulae for conditionals → logistic regression

Weighted formulae:

$$\langle \{x\}, funFor(x), -5 \rangle$$
$$\langle \{x,y\}, funFor(x) \wedge friends(x,y) \wedge social(y), 10 \rangle$$
$$\langle \{x,y\}, funFor(x) \wedge friends(x,y) \wedge \neg social(y), -3 \rangle$$

If *obs* includes observations for all $friends(x,y)$ and $social(y)$:

$$P(funFor(x) \mid obs) = sigmoid(-5 + 10 n_s(x) - 3 n_a(x))$$

$$n_s(x) = |\{y \mid friends(x,y) \wedge social(y)\}|$$

$$n_a(x) = |\{y \mid friends(x,y) \wedge \neg social(y)\}|$$

- Weighted formulae give arbitrary polynomials of counts.

Representation Issues

- Probabilities of directed model can be interpreted locally

Representation Issues

- Probabilities of directed model can be interpreted locally
- Directed models are modular — e.g., adding a dependent variable without side effects is straightforward, but impossible for MLNs
  [Buchman and Poole, AAAI 2015]

Representation Issues

- Probabilities of directed model can be interpreted locally
- Directed models are modular — e.g., adding a dependent variable without side effects is straightforward, but impossible for MLNs
  [Buchman and Poole, AAAI 2015]
- Directed models allow for pruning in inference.

Representation Issues

- Probabilities of directed model can be interpreted locally
- Directed models are modular — e.g., adding a dependent variable without side effects is straightforward, but impossible for MLNs
  [Buchman and Poole, AAAI 2015]
- Directed models allow for pruning in inference.
- Directed models require the structure of the conditional probabilities to be acyclic. Or maybe not...

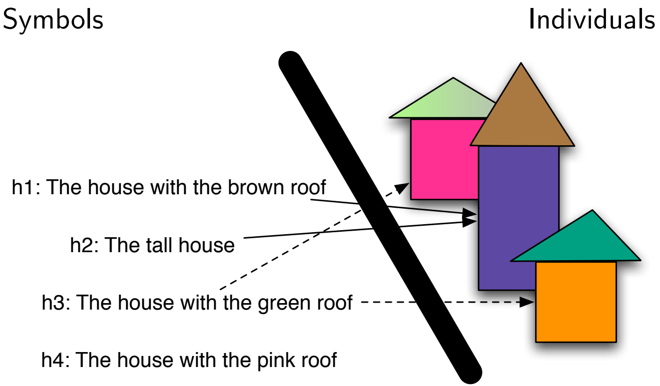
Representation Issues

- Probabilities of directed model can be interpreted locally
- Directed models are modular — e.g., adding a dependent variable without side effects is straightforward, but impossible for MLNs
  [Buchman and Poole, AAAI 2015]
- Directed models allow for pruning in inference.
- Directed models require the structure of the conditional probabilities to be acyclic. Or maybe not...
- Noisy-or aggregation corresponds to logic programs.
  With layered relational logistic regression, can we get relational neural networks?

Outline

1. Logic and Probability
   - Relational Probabilistic Models
   - Probabilistic Logic Programs
2. Lifted Inference
3. Undirected models, Directed models, and Weighted Formulae
4. Existence and Identity Uncertainty

Correspondence Problem



c symbols and i individuals $\longrightarrow c^{i+1}$ correspondences

Clarity Principle

Clarity principle: probabilities must be over well-defined propositions.

- What if an individual doesn't exist?
  - $house(h4) \wedge roof\_colour(h4, pink) \wedge \neg exists(h4)$

Clarity Principle

Clarity principle: probabilities must be over well-defined propositions.

- What if an individual doesn't exist?
  - $house(h4) \land roof\_colour(h4, pink) \land \neg exists(h4)$
- What if more than one individual exists? Which one are we referring to?
  —In a house with three bedrooms, which is the second bedroom?

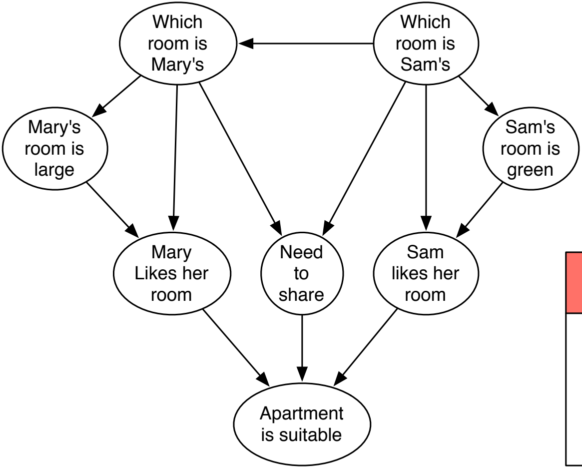
Role assignments

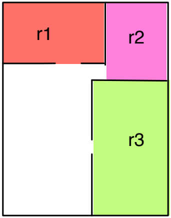
Hypothesis about what apartment Mary would like.

Whether Mary likes an apartment depends on:

- Whether there is a bedroom for daughter Sam
- Whether Sam's room is green
- Whether there is a bedroom for Mary
- Whether Mary's room is large
- Whether they share

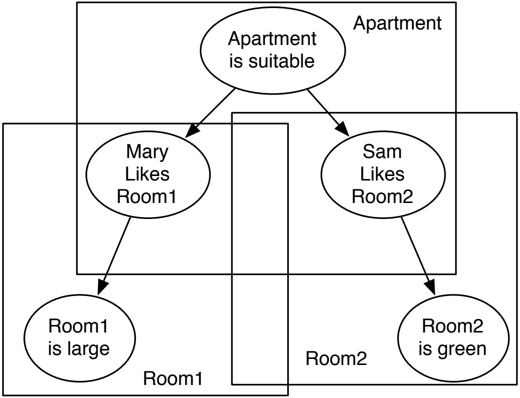
Bayesian Belief Network Representation

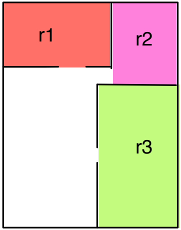




How can we condition on the observation of the apartment?

Naive Bayes representation





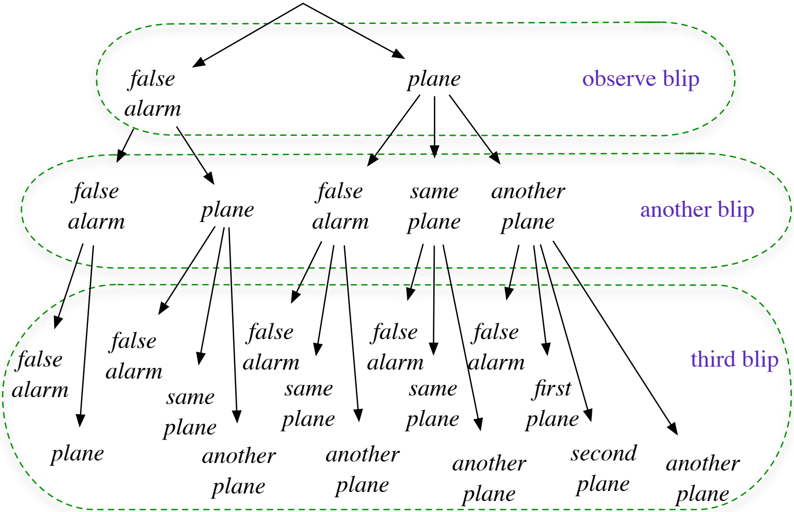
How do we specify that Mary chooses a room?
What about the case where they (have to) share?

Number and Existence Uncertainty

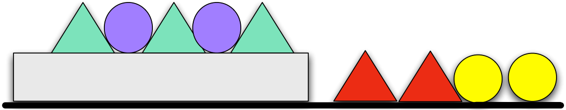
- PRMs (Pfeffer et al.), BLOG (Milch et al.): distribution over the number of individuals. For each number, reason about the correspondence.
- NP-BLOG (Carbonetto et al.): keep asking: is there one more?

  e.g., if you observe a radar blip, there are three hypotheses:
  - the blip was produced by plane you already hypothesized
  - the blip was produced by another plane
  - the blip wasn't produced by a plane

Existence Example



Observation Protocols

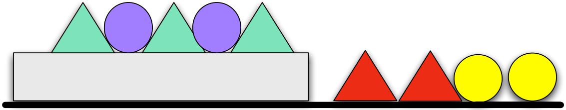


Observe a triangle and a circle touching. What is the probability the triangle is green?

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

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

Protocol for Observing



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

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

Other Issues

- Probabilistic programming
- Much data is being published with respect to formal ontologies.
  How can probabilistic models interact with such data?
- We'd like to publish hypotheses that make probabilistic predictions so they interoperate with data.
- Identity uncertainty. Probability of equality.
  Do these citations refer to the same publication?
- To make decisions, probabilistic models need to interact with utility models.
- Representing actions, time,...

Conclusion

- The field of "statistical relational AI" looks at how to combine first-order logic and probabilistic reasoning.

Challenges

- Representation: heuristically and epistemologically adequate representations for probabilistic models + observations (+ causation + actions + utilities + ontologies)
- Inference: exploit structure + exchangeability
  compute posterior probabilities (or optimal actions) quickly enough to be useful
- Learning: find best hypotheses conditioned on all observations
  ....just inference?

Age of Relations (100 years later)

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]

AI: computational agents that act intelligently