Probabilistic Reasoning about Possible Worlds

Brian Milch

Google
(based on work at UC Berkeley with Stuart Russell and his group)

IPAM Summer School
July 8, 2011

Probabilities on Possible Worlds

1.2 \times 10^{-12}

2.3 \times 10^{-13}

4.5 \times 10^{-13}

6.7 \times 10^{-15}

8.9 \times 10^{-15}

9.0 \times 10^{-17}
Traditional Approach

Data → Custom learning algorithm → Model in specific class → Evidence inference algorithm → Answer
Role of Representation Language

Data → Generic learning algorithm → Model in language → Generic inference algorithm → Evidence

Query → Answer
Outline

• Approaches to probability over possible worlds
• Bayesian Logic (BLOG)
  – Generalization of Bayesian networks
  – Unknown objects
• Inference: MCMC over possible worlds
• Application: Seismic monitoring
Levels of Expressivity

• **Atomic**: Enumerate all worlds (states)
  – Examples: histogram, Markov model

• **Propositional**: Describe worlds using propositions or random variables (RVs)
  – Examples: propositional logic, graphical models

• **First-order**: Parameterize propositions or RVs with logical variables
  – Examples: first-order logic, programming languages, plate models, BLOG, Markov Logic, …
Example: Rules of Chess

• 1 page in first-order logic
  \( \text{On}(\text{color}, \text{piece}, x, y, t) \)

• \(~100000\) pages in propositional logic
  \( \text{WhiteKingOnC4Move12} \)

• \(~10000000000000000000000000000000\) pages as atomic-state model
  \( \text{R.B.KB.RPPP..PPP..N..N…..PP…..q.pp..Q..n..n..ppp..pppr.b.kb.r} \)

(Note: Chess is still tiny compared to the real world)
First-Order Logical Languages

• For a given application, declare:
  – Constant symbols: Citation1, Ghahramani, …
  – Predicate symbols: IsProf(person),
    AuthorOf(person, publication), …
  – Function symbols: PubCited(citation),
    Name(person), Affiliation(person, date)…
First-Order Structures

- Formalization of possible worlds
- Each structure specifies domain of objects, and maps:
  - Constant symbols → objects
  - Predicate symbols → relations on objects
  - Function symbols → functions on objects
- Each sentence of logical language is true or false in each structure
Approach 1: Probabilities for First-Order Formulas

• Specify probabilities of sentences or formulas, as in:
  – 0.2: $\forall x \ IsProf(x)$ [Gaifman 1964]
  – $\forall x (\mu(\text{IsProf}(x)) = 0.7)$ [Halpern 1990]

• Drawback: hard to tell if there’s exactly one distribution over structures satisfying these constraints
Approach 2: Weights for Formulas

2.5: IsProf(x) → Busy(x)

<table>
<thead>
<tr>
<th></th>
<th>Busy(Smith)</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>e^{2.5}</td>
</tr>
<tr>
<td>F</td>
<td>e^{2.5}</td>
</tr>
</tbody>
</table>

• Weight of world x is exp(∑_{φ} w_{φ} #sat(φ ,x))
• Get probability distribution by normalizing
• Weights need to be learned jointly

(Markov logic [Richardson & Domingos 2006], relational Markov networks [Taskar et al. 2002])
Approach 3: Logic Plus Coin Flips

0.8: BeingProfCausesBusy(x)
IsProf(x) ∧ BeingProfCausesBusy(x) → Busy(x)

• Coin flips are **mutually independent**
• Coin flips yield a structure by logical implication; anything not implied is false

(probabilistic Horn abduction [Poole 1993], independent choice logic [Poole 1997], PRISM [Sato & Kameya 1997])
Approach 4: Conditional Probability Statements

Busy(x) {
    if IsProf(x) then ~ Bernoulli[0.8]
    else ~ Bernoulli[0.3]
}

• Specify conditional probability distribution (CPD) for each predicate and function
  – Defines directed graphical model
  – Model must be acyclic

(BUGS/plates [Thomas et al. 1992], relational Bayes nets [Jaeger 1997], probabilistic relational models [Friedman et al. 1999], Bayesian logic programs [Kersting & De Raedt 2001], Bayesian logic [Milch et al. 2005], multi-entity Bayes nets [Laskey 2008], many others)
Approach 5: Stochastic Programs

\[
\begin{align*}
\text{(define (Busy is_prof) (cond is_prof} \\
& \quad \text{(flip 0.8))} \\
& \quad \text{(flip 0.2)))}
\end{align*}
\]

• Functions are \textit{stochastic}: each call may return a different value \textup{(but see “mem” in Church)}
  – Contrast with \texttt{IsProf(x)} in other approaches
  – Store results in program variables for re-use, implicitly define possible world

(Stochastic logic programs [Muggleton 1996], stochastic programs [Koller et al. 1997], IBAL [Pfeffer 2001], \(\lambda_O [\text{Frank et al. 2005}], \ Church [\text{Goodman et al. 2008}]\))
Summary of Approaches

First-order probabilistic model

Arbitrary formulas
- Formula probabilities
- Formula weights

Generative process
- Random-world semantics
- Logic plus coin flips
- CPD-based model
- Stochastic program

Markov logic

BLOG

Church

BLOG Model: Object Existence

type Researcher; type Pub; type Citation;
guaranteed Citation Cit1, Cit2, Cit3, Cit4;

#Researcher ~ NumResearchersPrior();

#Pub ~ NumPubsPrior();
BLOG Model: Relational Structure

type Researcher; type Pub; type Citation;
guaranteed Citation Cit1, Cit2, Cit3, Cit4;

#Researcher ~ NumResearchersPrior();

#Pub ~ NumPubsPrior();
random NaturalNum NumAuthors(Pub p) ~ NumAuthorsPrior();
random Researcher NthAuthor(Pub p, NaturalNum n)
  if (n < NumAuthors(p)) then ~ Uniform({Res r});

random Pub PubCited(Citation c) ~ Uniform({Pub p});
Full BLOG Model for Citations

type Researcher; type Pub; type Citation;
guaranteed Citation Cit1, Cit2, Cit3, Cit4;

#Researcher ~ NumResearchersPrior();
random String Name(Researcher r) ~ NamePrior();

#Pub ~ NumPubsPrior();
random NaturalNum NumAuthors(Pub p) ~ NumAuthorsPrior();
random Researcher NthAuthor(Pub p, NaturalNum n)
    if (n < NumAuthors(p)) then ~ Uniform({Res r});
random String Title(Pub p) ~ TitlePrior();

random Pub PubCited(Citation c) ~ Uniform({Pub p});
random String Text(Citation c)
    ~ FormatCPD(Title(PubCited(c)),
        {n, Name(NthAuthor(PubCited(c), n)) for
            NaturalNum n : n < NumAuthors(PubCited(c))});
Flexibility of Number Variables

• Can have whole family of number variables for a given type:

\[
\text{#Pub}(\text{FirstAuthor} = r) \sim \text{PubsPerAuthorPrior}();
\]

For each researcher \( r \), there is random variable indicating \( |\{ p : \text{FirstAuthor}(p) = r \}| \)

• Number can depend on other variables:

\[
\text{#Pub}(\text{FirstAuthor} = r)
\begin{align*}
& \text{if } \text{IsProf}(r) \text{ then } \sim \text{ProfPubsPrior}() \\
& \text{else } \sim \text{IndividualPubsCPD}(\text{Diligence}(r));
\end{align*}
\]
What Exactly Are the Objects?

#Researcher ~ NumResearchersPrior();

Objects are (Res, 1), (Res, 2), ...

#Pub(FirstAuthor = r) ~ PubsPerAuthorPrior();

Publications with FirstAuthor = (Res, 2) are:
(Pub, (FirstAuthor, (Res, 2)), 1)
(Pub, (FirstAuthor, (Res, 2)), 2)
...

Result: Given values of number variables, there’s no ambiguity about which publications have which first authors
BN Defined by BLOG Model

• BN includes
  – All variables that have values in any world
  – All potential dependencies

BN has no topological numbering

Previous results don’t guarantee that CPDs define joint distribution

#Pub unbounded → infinitely many Title nodes
Contingent Bayes Nets

• **Contingent Bayes net** (CBN)
  – Each edge labeled with condition
  – Each world yields subgraph of active edges

[Milch et al., AI/Stats 2005]
BLOG Semantics Using CBNs

- BLOG model $\mathcal{M}$ defines set of possible worlds $\Omega_M$, contingent Bayes net $\mathcal{B}_M$
- Theorem: If for each world in $\Omega_M$, the active subgraph of $\mathcal{B}_M$ has a topological numbering, then the model $\mathcal{M}$ fully defines a probability distribution on $\Omega_M$.

Topological numbering can vary from world to world

Flexibility in modeling unknown objects, relations
Markov Chain Monte Carlo (MCMC) over Possible Worlds

- Random walk that converges to distribution defined by model
- Proposed move $w \rightarrow w' \text{ accepted or rejected based on } P(w') / P(w)$
Partial Worlds

- Instantiate varying subsets of the random variables
- When is this correct?
MCMC over Sets of Worlds

• Think of partial world as description satisfied by whole set of worlds

• Probability of partial world is probability of this whole set

• **Theorem:** MCMC over partial worlds is correct if:
  • Each partial world is specific enough to affirm the given evidence and answer the query
  • Partial worlds define non-overlapping sets of possible worlds

[Milch & Russell, UAI 2006]
Probabilities of Partial Worlds

• What’s the probability of the set of worlds satisfying this description?

• As in BN, it’s product of CPDs:

\[
p(#\text{Pub} = 100) \times p(\text{Title}(\text{Pub37}) = \ldots\text{analysers} \mid #\text{Pub} = 100) \\
\times p(\text{PubCited}(\text{Cit1}) = \text{Pub37} \mid #\text{Pub} = 100) \\
\times p(\text{Text}(\text{Cit1}) = \ldots\text{analyzers} \mid \text{PubCited}(\text{Cit1}) = \text{Pub37}, \\
\text{Title}(\text{Pub37}) = \ldots\text{analysers})
\]

Probability of partial world is **product of CPDs** if the world **includes all the active parents** of the variables it instantiates.
Numbering of Objects?

• Specifying publication numbers is unnecessary and inefficient

Reducing #Pub to 80 would be rejected here
Abstraction over Objects

- Can use existential quantifiers, describe larger set of worlds
- Probabilities include factorials
MCMC over Partial Worlds

- Derived conditions under which generic MCMC engine can use partial worlds
- Engine avoids spending time changing irrelevant aspects of world
Experiment: Citation Data Set

- Citeseer created four sets of 300-500 citations by searching for certain strings ("face", "reinforcement", "reasoning", "constraint")
- Lots of noise in citation strings


Experiment: Accuracy Results

Four data sets of ~300-500 citations, referring to ~150-300 papers
Seismological Monitoring

- Comprehensive Nuclear Test Ban Treaty (CTBT) organization responsible for detecting nuclear explosions
- Detect seismological events, then figure out if earthquakes or nuclear tests
- Vertically Integrated Seismic Analysis (VISA) project using BLOG-style model [Arora et al., NIPS 2010]
254 monitoring stations
Start: 3/26/02 3:45:05 UTC (L) Station: Edmonds WA 47.849N 122.328W Samples: 179975  SPS: 25
Comment: M6.5 9724 Km from Edmonds WA, SW RYUKYU ISL., JAPAN  Max/Min: 4746/-5112  X: 1:15:00 Y: x1
Event Time: 03/26 03:45:48.0 Lat/Long: 23.54N 123.91E Depth: 33km 20.5mi Mag: M6.5
Org: 3:45:47.9  Diff: 10:36.9min Dist: 87.467deg 9724.3km 6038.8mi  Mag: M1?? JB: 33
Vertically Integrated Seismic Analysis

• The problem is hard:
  – ~10000 “detections” per day, 90% false
  – CTBT system (SEL3) finds 69% of significant events plus about twice as many spurious (nonexistent) events
  – 16 human analysts find more events, correct existing ones, throw out spurious events, generate LEB (“ground truth”)
  – Unreliable below magnitude 4 (1kT)

• Solve it by **global** probabilistic inference
  – NET-VISA finds around 88% of significant events
Generative model for detections

- Events occur in time and space with magnitude
  - Natural spatial distribution a mixture of Fisher-Binghams
  - Man-made spatial distribution uniform
  - Time distribution Poisson with given spatial intensity
  - Magnitude distribution Gutenberg-Richter (exponential)
  - Aftershock distribution (not yet implemented)
- Travel time according to IASPEI91 model plus Laplacian error distribution for each of 14 phases
- Detection depends on magnitude, distance, station
- Detected azimuth, slowness have Laplacian error
- False detections with station-dependent distribution
# Seismic Events

SeismicEvents \sim \text{Poisson}[\text{TIME\_DURATION} \times \text{EVENT\_RATE}];

IsEarthQuake(e) \sim \text{Bernoulli}(0.999);

EventLocation(e) \sim \begin{cases} 
\text{EarthQuakeDistribution()} & \text{if IsEarthQuake(e)} \\
\text{UniformEarthDistribution()} & \text{else}
\end{cases}

Magnitude(e) \sim \text{Exponential}(\log(10)) + \text{MIN\_MAG};

Distance(e,s) = \text{GeographicalDistance(EventLocation(e), SiteLocation(s))};

IsDetected(e,p,s) \sim \text{Logistic}[\text{SITE\_COEFFS}(s,p)](\text{Magnitude(e)}, \text{Distance(e,s)});

#Arrivals(site = s) \sim \text{Poisson}[\text{TIME\_DURATION} \times \text{FALSE\_RATE}(s)];

#Arrivals(event=e, phase = p, site = s) = \begin{cases} 
1 & \text{if IsDetected(e,p,s)} \\
0 & \text{else}
\end{cases};

Time(a) \sim \begin{cases} 
\text{Uniform}(0, \text{TIME\_DURATION}) & \text{if (event(a) = null)} \\
\text{IASPEI(EventLocation(event(a)), SiteLocation(site(a)), Phase(a))} + \text{TimeRes(a)} & \text{else}
\end{cases};

TimeRes(a) \sim \text{Laplace}(\text{TIMLOC}(site(a)), \text{TIMSCALE}(site(a)));

Azimuth(a) \sim \begin{cases} 
\text{Uniform}(0, 360) & \text{if (event(a) = null)} \\
\text{GeoAzimuth(EventLocation(event(a)), SiteLocation(site(a)))} + \text{AzRes(a)} & \text{else}
\end{cases};

AzRes(a) \sim \text{Laplace}(0, \text{AZSCALE}(site(a)));

Slow(a) \sim \begin{cases} 
\text{Uniform}(0, 20) & \text{if (event(a) = null)} \\
\text{IASPEI-SLOW(EventLocation(event(a)), SiteLocation(site(a)))} + \text{SlowRes(site(a))} & \text{else}
\end{cases};
Prior on Event Locations

\[ \text{EventLocation}(e) \sim \text{If IsEarthQuake}(e) \text{ then EarthQuakeDistribution}() \]

Prior Density of Events
Detection Probability

\[
\text{IsDetected}(e,p,s) \sim \text{Logistic}[\text{SITE\_COEFFS}(s,p)](\text{Magnitude}(e), \text{Distance}(e,s));
\]
Travel time

\[ Time(a) \sim IASPEI(EventLocation(event(a)), SiteLocation(site(a)), Phase(a)) + TimeRes(a); \]
Azimuth\textsubscript{(a)} \sim \text{GeoAzimuth}(\text{EventLocation}(\text{event}(a)), \text{SiteLocation}(\text{site}(a)) + \text{AzRes}(a);
Fraction of LEB events missed

NET-VISA

SEL3

NET-VISA
Precision-Recall Curve

Precision-Recall curve with LEB as ground truth

- SEL3
- SEL3 extrapolation
- NET-VISA
- NETVISA (last year)
Event distribution: LEB vs SEL3

LEB (yellow) and SEL3 (red)
Event distribution: LEB vs NET-VISA

LEB(yellow) and NET-VISA(blue)
Why does NET-VISA work?

- Empirically calibrated *seismological* models
  - Improving model structure and quality improves the results
- Sound Bayesian combination of evidence
- Measured arrival times, phase labels, azimuths, etc., NOT taken as truth
- *Absence* of detections provides negative evidence
- More detections per event than SEL3 or LEB
Why does NET-VISA not work (perfectly)?

• Needs hydroacoustics for mid-ocean events
• Weaknesses in model:
  – Travel time residuals for all phases along a single path are assumed to be uncorrelated
  – Each phase arrival is assumed to generate at most one detection; in fact, multiple detections occur
• Arrival detectors use high SNR thresholds, look only at local signal to make hard decisions
Detection-based and signal-based monitoring

events

detections

waveform signals

SEL3  NET-VISA  SIG-VISA
Summary

• **CPD-based** first-order probabilistic models
  – Describe generative processes for worlds
  – Generalize Bayesian networks
• **BLOG** adds support for **unknown objects**
• Can run MCMC over **partial worlds**
• Strong results when **CPDs come from science**, as in seismological monitoring