Logical and Probabilistic Knowledge Representation and Reasoning

Leopoldo Bertossi

Carleton University
School of Computer Science
Ottawa, Canada
A Bit of Knowledge Representation

- KR has to do with: [36]
  - Representing knowledge in computers
  - Enable computers to “reason” with that knowledge
- Traditional areas of KR (among others):
  1. Automated diagnosis
  2. Reasoning in the presence of uncertainty [16]
     Knowledge that is ambiguous, incomplete, inconsistent, etc.
  3. Commonsense reasoning (may fall under 2.)
  4. Reasoning about causality (in particular, action/effect)
  5. Probabilistic reasoning
- 1.-4. not necessarily or intrinsically of probabilistic nature
A “Logical” Approach:

Example: Commonsense reasoning

(*)

Knowledge: “Tweety is a bird”

Consequence: “Tweety flies”

• Possible interpretation of the use of (*): A “default rule” is applied

“In the absence of opposite information, conclude that the bird flies ...”

“In the absence of information that the bird is exceptional, conclude that the bird flies ...”

In more “logical” terms:

\[ KB = \{ \forall x (\text{Bird}(x) \land \neg\text{Abnormal}(x) \rightarrow \text{Flies}(x)), \text{Bird}(\text{tweety}) \} \]

“Since there is no evidence of Tweety’s abnormality, conclude it flies”

The extension of predicate \( \text{Abnormal} \) is minimized ...

Intuition formalized as “circumscription” (John McCarthy, 1980)
• A “non-monotonic”, non-classical logic ...

• Alternatively, “Default Logic” (Ray Reiter, 1980)

Deduction rule: 

\[
\frac{\text{Bird}(x) : \text{Flies}(x)}{\text{Flies}(x)}
\]

“Given that \( x \) is bird, if it can be consistently assumed that \( x \) flies, conclude \( x \) flies”

Example: Diagnosis Reasoning

Knowledge: \( \text{Flu} \Rightarrow \text{Fiver} \)

\( \text{Fiver} \) (evidence, observation)

Conclusion: \( \text{Flu} \)

• The diagnosis problem in AI has also been treated using logical methods [29]

For example, through abduction:
Given a knowledge base $K$, and an observation $O$, we want an explanation $E$, expressed in logical terms, such that:

$$K \land E \Rightarrow O$$  

(usually with an extra minimality assumption on $E$)

- Again, a non-monotonic, non-classical logic ...

**A “Probabilistic” Approach:**

- Problems 1.-4. may also have probabilistic elements, or be treated as such, even without explicit initial probabilities (cf. slide 2)

- For example, a default rule may be treated as a probabilistic or statistical statement (with high probability)

  And the consequences may be probabilistic too

- Or as conditional probabilities: $P(\text{flies} \mid \text{bird}) = 0.95$

  “the probability of flying being a bird is 0.95”
• How to represent conditional probabilities?

• “Bayesian networks” appear [5]

How to represent probabilistic knowledge; in particular, about the propagation of conditional probabilities?

The probabilistic model emerges from the network structure plus conditional probabilities, and other assumptions

• The model is a join probability distribution $P$ over all variables

• The structure conveys causal elements

• Conditional probabilities are initially given, and may be subjective

• Additional assumptions ...
Of the kind: the conditional probability on a node given its parent nodes does not depend upon its non-descendant nodes (locally Markovian network)

- A full approach to uncertain and causal reasoning is based on Bayesian networks [26, 27]

- And Markov networks: as above, but non-directed

  Instead of conditional probabilities they represent initial local join probability distributions

  Cliques in the NW have associated “potential functions”, which define join probability distributions (and factors of the global distribution)

- Using conditional probabilities allows attacking several KR problems

- In particular, maximum-likelihood reasoning

  For example, abductive diagnosis may also be probabilistic/statistical
Given a probability distribution, we may, e.g. find a maximum-likelihood explanation $E$:

$$\arg \max_E P(O \mid K \land E)$$

One that maximizes the probability of the observation, which was observed after all ...

**Putting Things Together:**

- Research that unifies or connects different probabilistic approaches are of crucial relevance
- Traditionally, the “logical-” and “probabilistic schools” have been separate and competitors
- In the last few years they have become complementary approaches
- Today, KR problems are attacked with mathematical models/techniques that involve simultaneously logic (rather classical) and probability
Logic + Probability in AI

- Different forms of KR combine logic and probability
  Each of them with a domain of intended applications
  Just a couple of approaches, to give the idea ...

**ME Distributions:**

- Logical consequences from a knowledge base $KB$?

- In $KB$ there is:
  - “hard” knowledge, e.g. $emu \rightarrow bird$
  - “soft”, conditional, probabilistic rules, of the form $r: (\alpha|\beta)[p]$  
    
    E.g. $r_v : (flies|bird)[0.9]$
• There is a collection $C$ of worlds, $\mathcal{W}$, associated to $KB$ that satisfy the hard knowledge, but not necessarily $\beta \rightarrow \alpha$
  
  (conditionals as classical implications)

• There are also several probability distributions (candidates) $P$ over $C$: $\mathcal{W} \mapsto P(\mathcal{W})$

We restrict ourselves to the collection $\mathcal{P}$ of distributions $P$ that satisfy the $KB$, in particular, the conditionals rules $r$: $(\alpha|\beta)[p]$

For example, it must hold:\footnote{$P(\alpha|\beta):= \frac{P(\alpha \land \beta)}{P(\beta)} := \frac{P(\{W \mid W|=\alpha \land \beta\})}{P(\{W \mid W|=\beta\})}$} $P(\alpha|\beta) = p$ (and $P(\beta) > 0$)

• The logical-probabilistic consequences of $KB$ are those with high probability

• But, under which of the distributions in $\mathcal{P}$?

There may be some that are better than others ...
We choose one having maximum entropy:

\[ P^* := \text{arg max}_{P \in \mathcal{P}} \text{Entropy}(P) \]

\[ = \text{arg max}_{P \in \mathcal{P}} -\sum_{\mathcal{W} \in \mathcal{C}} P(\mathcal{W}) \times \ln(P(\mathcal{W})) \]

Distribution with less arbitrary assumptions/structure, maximum disorder, maximum independence, less information for free:

- The logical-probabilistic consequences of KB are those with high probability under \( P^* \)

\[ KB \models_{\text{lp}} \varphi \iff P^*(\varphi) \geq 1 - \epsilon \]

- The computation of consequences from KB via the maximum entropy distribution can be automated \[18\]

- It is a philosophically interesting choice \[25, 3\]
Markov Logic Networks: (MLNs) [30, 10]

- MLNs combine FO logic and Markov Networks (MNs) in the same representation

- Coming from Bayesian, Belief and MNs, they are used not only for KR, but also in machine learning (ML)

  Networks can be learned from data for producing KR models and new inferences

- We have a knowledge base $KB$ in FO logic

- The interpretations (models, worlds, ...) have associated probabilities

- There may be “models” that are more likely than others

- A world that violates a formula is not invalid (non-model), but only less probable
• Weights $w_i$ are assigned (real-valued) to each of the $F$ formulae $F_i$ in $KB$

The higher the weight, the higher is the difference between a world that satisfies the formula and another that doesn’t (with everything else the same)

<table>
<thead>
<tr>
<th>Formula</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\forall x : \text{Steal}(x) \Rightarrow \text{Prison}(x)$</td>
<td>3</td>
</tr>
<tr>
<td>$\forall x \forall y : \text{CrimePartners}(x, y) \land \text{Steal}(x) \Rightarrow \text{Prison}(y)$</td>
<td>1.5</td>
</tr>
</tbody>
</table>

• The weight of a formula captures the way the probability decreases when an “instance” of the formula is violated

• There is a fixed, finite, domain of constants, with which predicates and formulae can be instantiated, e.g. $Dom = \{\text{bob, anna, ...}\}$

Producing atoms and instantiated formulae (of the propositional kind), e.g. $\text{CrimePartners}(\text{bob, anna}), \text{Steal}(\text{bob}) \rightarrow \text{Prison}(\text{bob})$, resp.
A world is a set of atoms (a Herbrand structure), e.g.

\[ W_1 = \{\text{CrimePartners}(bob, anna), \text{Steal}(bob)\} \]

\[ W_2 = \{\text{CrimePartners}(bob, anna), \text{Steal}(bob), \text{Prison}(anna)\} \]

The instantiated atoms are the nodes of the network (not directed)

Each node \( N \) takes the value 0/1 depending on whether it is false or true (in a world \( W \))

Each node \( N \) is a Bernoulli random variable \( X^N \)

\( W_1 \) represented by random vector: \( \chi = \langle X^{N_1}, X^{N_2}, X^{N_3}, \ldots \rangle = \langle 1, 1, 0, \ldots, 0 \rangle \)

Edge between two nodes if they appear in a same instantiated formula
• As in ML, the representation has associated “features”, which are measurable variables (functions, i.e. random variables)

• Each instantiation of a formula generates a “feature”, with value 1 if true in the world \( \mathcal{W} \), and 0, otherwise

• All these ingredients plus a log-linear “potential function” (to generate join probability distributions) give the probability of world \( \mathcal{W} \) associated to \( x \in \{0, 1\}^M \):

\[
P(X = x) = \frac{1}{Z} \exp(\sum_{i=1}^{F} w_i \times n_i(x))
\]

\( n_i(x) \): number of instantiations of \( F_i \) true in world \( x \)

\( Z \) normalizes considering all possible worlds:

\[
Z = \sum_{z \in \{0, 1\}^M} \exp(\sum_{i=1}^{F} w_i \times n_i(z))
\]

• This is just the start ...

• How to learn an MLN? (ML) \[12, 20, 31, 21\]

• How to do inference with MLNs?
Inference in Markov Logic Networks:

- Inference under MLNs is a probabilistic nature

A basic inference task is computing the probability of a world, as in slide 15, but it could be more interesting: for a logical sentence $\varphi$

$$P(\varphi) := P(x \in \{0, 1\}^M \mid \varphi \text{ is true in } x) = \sum_{x \in \{0, 1\}^M \mid \varphi} P(x)$$

If this probability is sufficiently high, we can accept it as an inference from the MLN

- Computing the probabilities amounts, directly or not, to counting models (possibly with specific properties) (see slide 15)

And combining the resulting numbers with the weights; to use the probabilistic formulas above
• An active area of research in these days

  Different areas converge: model counting in logic (around SAT-related problems), graph theory, and data management [23, 2, 13]

• Instantiating and counting next may lead to combinatorial explosion

  Can we do better? Counting models without instantiating?

• Can we approximate model counting (and probabilities), also without instantiating?

• This is what is called “lifted” inference [19, 14, 15, 35]

  Lifted up to the FO representation ...
Some Lifted Inference: In the end, it attacks the problem of weighted (first-order) model counting [34]

- Probabilistic inference based on lifted weighted model counting runs in polynomial time in the number of objects in the domain → data complexity!

  (Not necessarily in the number of predicates, formulas, logical variables)

- Lifted inference avoids grounding and counting by exploiting:
  - Stochastic independence
    Related to the (underlying) graph structure, in particular tree-width
    Markovian assumptions and conditional independence
• **Symmetries, exchangeability**

A set of random variables $\mathcal{X} = \{X^1, \ldots, X^n\}$ is fully exchangeable if and only if

$$P(X^1 = x_1, \ldots, X^n = x_n) = P(X^1 = x_{\pi(1)}, \ldots, X^n = x_{\pi(n)})$$

for all permutations $\pi$ of $\{1, \ldots, n\}$

Best understood thinking of Bernoulli random variables indicating head/tail

Exchangeability means that it is only the number of heads that matters and not their particular order

• **Reasoning at FO level**

• **Reasoning about groups of objects**
Example: (weighted FO model counting, not necessarily for MLNs)$^2$

A schema

Possible worlds

Logical interpretations

A logical theory:

$$\forall xy(Smokes(x) \land Friends(x, y) \rightarrow Smokes(y))$$

Models of the theory

$\text{FO model count} \sim \#\text{SAT}$

$^2$Borrowed from Guy van den Broeck
Now a logical theory and weights for predicates:

\[ \text{Smokes} \mapsto 1 \]
\[ \text{¬Smokes} \mapsto 2 \]
\[ \text{Friends} \mapsto 4 \]
\[ \text{¬Friends} \mapsto 1 \]

Weighted FO model count \( \sim \) partition function
Example:

1. Logical sentence: 
   \[ \forall x (\text{Stress}(x) \rightarrow \text{Smokes}(x)) \]

   Domain: 
   Alice

<table>
<thead>
<tr>
<th>Stress(Alice)</th>
<th>Smokes(Alice)</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

3 models!

2. Logical sentence: 
   \[ \forall x (\text{Stress}(x) \rightarrow \text{Smokes}(x)) \]

   Domain: 
   \( n \) people

3\(^n\) models!
3. Logical sentence:  
\[
\forall y (\text{ParentOf}(y) \land \text{Female} \rightarrow \text{MotherOf}(y)) \quad n \text{ people}
\]

- If \text{Female}: \quad \forall y (\text{ParentOf}(y) \rightarrow \text{MotherOf}(y))

- If not \text{Female}: \quad \text{true}

\((3^n + 4^n)\) models!

4. Logical sentence:  
\[
\forall xy (\text{ParentOf}(x, y) \land \text{Female}(x) \rightarrow \text{MotherOf}(x, y)) \quad n \text{ people}
\]

\((3^n + 4^n)^n\) models!
5. Logical sentence:

\[ \forall xy (\text{Smokes}(x) \land \text{Friends}(x, y) \rightarrow \text{Smokes}(y)) \quad n \text{ people} \]

- If we know precisely who smokes, and there are \( k \) smokers:

Database:

\[
\begin{align*}
\text{Smokes}(Alice) &= 1 \\
\text{Smokes}(Bob) &= 0 \\
\text{Smokes}(Charlie) &= 0 \\
\text{Smokes}(Dave) &= 1 \\
\text{Smokes}(Eve) &= 0 \\
\end{align*}
\]
- If we know precisely who smokes, and there are $k$ smokers
  \[ \longrightarrow 2^{n^2 - k(n-k)} \text{ models} \]

- If we know that there are $k$ smokers
  \[ \longrightarrow \binom{n}{k} 2^{n^2 - k(n-k)} \text{ models} \]

- In total
  \[ \sum_{k=0}^{n} \binom{n}{k} 2^{n^2 - k(n-k)} \text{ models} \]
Some Conclusions

- The conception and use of probability and statistical methods have direct and practical relevance in AI and KR, in particular.

  There are models and methods that start from certain specific positions in this regard.

- The schools of logic and probability in AI followed parallel paths for a long time.

  Now we see a convergence/combination of the approaches in KR.

- New kinds of “KBs” are being developed/used: MLN KBs.

  There are also the new and related area of “probabilistic logic programming” [6, 4], among others.

  Established connections with statistical relational learning [7, 8].
• Apart from inference, there is the problem of learning (the parameters of) those KBs

• New methods have been developed for the integration of FO logical and probabilistic reasoning, e.g. lifted inference

• Some of them appeal to existing and long-standing technology, e.g. DMBSs, with all their built-in and optimized techniques

Some of them appeal to newer developments in data management, e.g. probabilistic DBs for approximate logical/probabilistic reasoning

• There are many applications we have not discussed here, some of them to ER and link mining [33, 32, 11, 28] and Semantic Web [9, 1]

• The future is full of interesting scientific challenges ...

http://www.scs.carleton.ca/~bertossi
Email: bertossi@scs.carleton.ca
References


