TECHNISCHE UNIVERSITÄT DRESDEN

Computational Logic ∴ Group

Foundations of Knowledge Representation

Uncertainty

Hannes Straß
Jonas Karge

DRESDEN concept
Exzellenz aus Wissenschaft und Kultur

Uncertainty – Outline

Motivation

Recall the limitations of first-order logic:

FOL is powerful, but still can't capture

- Transitive closure (Ancestor is the transitive closure of Parent)
- Defaults and exceptions (Birds fly by default; Penguins are an exception)
- Probabilistic knowledge (Children suffer from JRA with probability x)
- Vague knowledge (Ian is tall)
- ...

We will now focus on probabilistic and (some) vague knowledge.

We can distinguish the following, resembling, notions:

1. Ambiguity: A statement does not have a clear meaning, can be formally interpreted in several distinct ways.
   Visiting relatives can be exhausting.
2. Uncertainty: Lack of sufficient information about the state of the world, for determining whether a Boolean statement is true or false.
3. Imprecision: Refers to the contents of the considered statement and depends on the graunularity of the language.
   Robin is between 25 and 30 years old.
4. Incompleteness: Refers to sources with missing information, or that are not able to distinguish between several situations.
5. Vagueness: A vague statement contains vague or gradual predicates.
   Ian is tall.

Basic Terminology

Most representations of uncertainty start with a set of possible worlds:

> Possible Worlds: Intuitively, possible worlds are the worlds or outcomes that an agent considers possible.

Example: When tossing a die, we can consider six possible worlds, one for each outcome. This can be represented by a set $W = \{w_1, \ldots, w_6\}$ consisting of worlds w_i, for $i = 1, \ldots, 6$.

- For each w_i:

  The world w_i is the one where the die lands face i up.

- Sample space:

  The set W is often called a sample space.

Basic Terminology

More Terminology:

- Object of belief:

  The objects that are known (or considered likely or possible or probable) are events (or propositions).

- Event:

  Formally, an event (or proposition) is a set of possible worlds.

Example: The event that "*this die lands on an even number*" corresponds to the set $\{w_2, w_4, w_6\}$.

- The set of worlds that an agent considers possible can be viewed as a qualitative measure of her uncertainty.
- The more worlds she considers possible, the more uncertain she is as to the true state of affairs, and the less she knows.
- This is a very coarse-grained representation of uncertainty.

Probability Measures

Perhaps the best-known approach to getting a more fine-grained representation of uncertainty is probability.

Suppose that the agent's uncertainty is represented by the set $W = \{w_1, \ldots, w_n\}$ of possible worlds.

- A probability measure assigns to each of the worlds in W a number – a probability – that can be thought of as describing the likelihood of this world being the actual world.
- The set of objects of belief (the propositions) is the power set of W, denoted 2^W.
- A probability function pr is a function $pr : 2^W \to \mathbb{R}$, satisfying the probability axioms.

Probability Axioms

Given a set S of propositions (taken to be sets of possible worlds).

Kolmogorov axioms:

1. Non-Negativity: $pr(A) \geq 0$ for all $A \in S$.
2. Normalization: $pr(T) = 1$ for all necessary truths $T \in S$.
3. Finite Additivity: $pr(A \vee B) = pr(A) + pr(B)$ for all disjoint $A, B \in S$.

Closure

It is typically assumed that the set of subsets of W to which probability is assigned satisfies some closure properties.

Definition (Algebra)

An algebra over W is a set $\mathcal{F}$ of subsets of W that contains W and is closed under union and complementation, so that if U and V are in $\mathcal{F}$, then so are $U \cup V$ and $\overline{U}$.

A σ-algebra is closed under complementation and countable union, so that if $U_1, U_2, \ldots$ are all in $\mathcal{F}$, then so is $\bigcup \{U_1, U_2, \ldots\}$.

Note: If W is finite, every algebra is a σ-algebra.

Probability Measures

Consider again the example of tossing a die:

- If each of the six outcomes is considered equally likely, then it seems reasonable to assign to each of the six worlds the same number.
- With that, we get $pr(\{w_i\}) = \frac{1}{6}$ for every $w_i \in \{w_1, \ldots, w_6\}$.
- Applying the Kolmogorov axioms we then get $pr(W') = 1/6 \cdot |W'|$ for every $W' \subseteq \{w_1, \ldots, w_6\}$.

Justifying Probability

- If belief is quantified using probability, then we need to explain what the numbers represent.
- Without such explanation, it will not be clear how to assign probabilities in applications, nor how to interpret the results obtained by using probability.

Classical approach: Reduce a situation to a number of elementary outcomes.

- A natural assumption (principle of indifference): All elementary outcomes are equally likely.
- Intuitively: In the absence of any other information, there is no reason to consider one more likely than another.

More Interpretations of Probabilities

The relative-frequency interpretation: Takes probability to be an objective property of a situation.

The (extreme) subjective viewpoint: Argues that there is no such thing as an objective notion of probability.

- Probability is a number assigned by an individual representing his or her subjective assessment of likelihood.
- This assessment is valid as long as it satisfies the probability axioms.

Problems with Probabilities

Despite its widespread acceptance, using probabilities to represent uncertainty is not without problems. Two examples:

Problem 1: Suppose that a coin is tossed once. There are two possible worlds, *heads* and *tails*.

- If the coin is known to be fair, it seems reasonable to assign probability 1/2 to each of these worlds.
  However, suppose that the coin has an unknown bias. How should this be represented?
- One approach is to continue to take heads and tails as the elementary outcomes and to apply the principle of indifference.
- Still, there seems to be a significant difference between a fair coin and a coin of unknown bias.

Problems with Probabilities

Problem 2: Suppose that a bag contains 100 marbles. 30 are known to be red, and the remainder are known to be either blue or yellow, although the exact proportion of blue and yellow is not known.

- What is the likelihood that a marble taken out of the bag is yellow?
- This can be modeled with three possible worlds: red, blue, and yellow (one for each outcome).
- It seems reasonable to assign probability 0.3 to the outcome of choosing a red marble, and thus probability 0.7 to choosing either blue or yellow. But what probability should be assigned to the other two outcomes?

Summary – Problems with Probabilities

In a nutshell. . .

Problem 1: Probability is not good at representing severe uncertainty.

Problem 2: While an agent may be prepared to assign probabilities to some sets, she may not be prepared to assign probabilities to all sets.

Dempster-Shafer Theory

The Dempster-Shafer theory of evidence provides another approach to attaching likelihood to events.

It is based on two ideas:

- The idea of obtaining degrees of belief for one question from subjective probabilities for a related question.
- Dempster's rule for combining such degrees of belief when they are based on independent items of evidence.

This approach starts with a belief function:

> Belief Function: Given a set W of possible worlds and $U \subseteq W$, the belief in U, denoted $Bel(U)$, is a number in the real interval $[0, 1]$.

Example – Belief Functions

To illustrate this view, consider the following example:

- Let T be the proposition that *It will snow in Dresden on New Year's day in 2026* and suppose our agent's belief function assigns a value of 0.6 to this claim.
- We represent this by writing $Bel(T) = 0.6$.

Note the differences to a probability function:

- If the agent's belief function was a probability function, then it would follow that: $Bel(\neg T) = 0.4$.
- However, our belief function can assign any value less than or equal to 0.4 to $\neg T$.

Belief Functions versus Probability Functions

More generally:

- The value assigned to the disjunction of two disjoint propostitions A and B, $Bel(A \cup B)$ does not necessarily equal the sum of $Bel(A)$ and $Bel(B)$.
- The requirement is only that the value assigned to the disjunction must be at least as big as the sum of the values assigned to the disjuncts.
- Thus the belief function is not a probability function, as the third probability axiom does not apply.

Theory of Evidence

Belief functions are part of a theory of evidence.
Intuitively, evidence supports events to varying degrees.

Consider as example again an urn that contains 100 marbles:

- The information that there are exactly 30 red marbles provides support in degree 0.3 for red.
- The information that there are 70 yellow and blue marbles does not provide any positive support for either blue or yellow.
- But it does provide support 0.7 for {blue, yellow}.

Theory of Evidence

In general, evidence provides some degree of support (possibly 0) for each subset of W.

- The total amount of support is 1.
- The belief that U holds, $Bel(U)$, is then the sum of all the supports on subsets of U.

Formally, this is captured by defining belief functions and plausibility functions based on so-called mass functions:

Basic Terminology

A mass function (sometimes called basic probability assignment) on W is a function $m : 2^W \to [0, 1]$ satisfying the following properties:

$$m(\emptyset) = 0 \tag{M1}$$

$$\sum_{U \subseteq W} m(U) = 1 \tag{M2}$$

Intuitively, $m(U)$ describes the extent to which the evidence supports U.

Let us see how we could motivate M1 and M2:

Mass Function – Explanation

Suppose U is an observation:

> Accurate Observation:
> Say that an observation is *accurate* if the following holds:
> If U is observed, the actual world is in U.

With that, $m(U)$ can be viewed as the likelihood of observing U.

Motivation, M1: It is impossible to observe $\emptyset$.

Motivation, M2: Something must be observed.

Given a mass function m, the likelihood of the actual world being in U can be approximated from below using a belief function, and from above using a plausibility function.

Belief Function and Plausibility Function

Given a mass function m, define the belief function based on m, Bel_m, by taking:

$$Bel_m(U) = \sum_{U' \subseteq U} m(U') \tag{1}$$

Intuitively, $Bel_m(U)$ is the sum of the probabilities of the evidence or observations that guarantees that the actual world is in U.

The corresponding plausibility function $Plaus_m$ is defined as:

$$Plaus_m(U) = \sum_{\substack{U' \subseteq W, \\ U' \cap U \neq \emptyset}} m(U') \tag{2}$$

$Plaus_m(U)$ can be thought of as the sum of the probabilities of the evidence that is compatible with the actual world being in U.

Belief Function – Interpretation

One way to interpret the idea of a belief function, is as a measure of the weight of evidence for each proposition.

Consider again agent A's belief function that assigns a value of 0.6 to proposition S.

- Suppose that A asked a friend whether it will snow in Dresden that day who is sure that it will.
- A considers this friend to be reliable in 60% of cases of this sort (this is why A's belief function assigns a value of 0.6 to this claim).

Belief Function – Interpretation

- If this is all evidence A has, her belief function assigns a value of 0 to $\neg S$.
- This zero does not mean that A is sure that $\neg S$ is not the case but that her friend's testimony gives no reason to believe that $\neg S$.
- In cases where she has evidence from two different sources (another friend gives her opinion on S), then the belief functions that result from these different bodies of evidence need to be combined.

Let us see how such a combination is carried out:

Combination Rules

Combination rules are special types of aggregation methods for data obtained from multiple sources.

In Dempster-Shafer theory we assume that these sources are independent.

> Central Question: How do we aggregate our data from multiple sources?

From a set theoretic standpoint, these rules can potentially occupy a continuum between conjunction (set intersection) and disjunction (set union).

Combination Rules

We can distinguish the following situations:

- In the situation where all sources are considered reliable, a conjunctive operation is appropriate (A and B and C . . .).
- In the case where there is one reliable source among many, we can justify the use of a disjunctive combination operation (A or B or C . . .).
- However, many combination operations lie between these two extremes (A and B or C, A and C or B, etc.).
- Dempster's rule strongly emphasizes the agreement between multiple sources and ignores all the conflicting evidence through a normalization factor.
- This can be considered a strict AND-operation.

The Rule of Combination

How does the Rule of Combination work intuitively?

- Suppose that an agent obtains evidence from two sources, one characterized by m_1 and the other by m_2.
- An observation U_1 from the first source and an observation U_2 from the second source can be viewed as together providing evidence for $U_1 \cap U_2$.
- The evidence for a set U_3 should consist of all the ways of observing sets U_1 from the first source and U_2 from the second source such that $U_1 \cap U_2 = U_3$.
- Assuming that the two sources are independent, the likelihood of observing both U_1 and U_2 is the product of the likelihood of observing each one, namely, $m_1(U_1)m_2(U_2)$.
- This suggests that the contribution of U_1 and U_2 to the mass of U_3 according to $m_1 \oplus m_2$ should be $m_1(U_1)m_2(U_2)$.

The Rule of Combination

Dempster's Rule of Combination provides a way of constructing a new mass function $m_1 \oplus m_2$, provided there are at least two sets U_1 and U_2 such that $U_1 \cap U_2 \neq \emptyset$ and $m_1(U_1)m_2(U_2) > 0$.

Definition (Rule of Combination)

For $U = \emptyset$ clearly $(m_1 \oplus m_2)(\emptyset) = 0$; for $U \neq \emptyset$:

$$(m_1 \oplus m_2)(U) = \frac{1}{c} \sum_{\substack{U_1, U_2 \subseteq W, \\ U_1 \cap U_2 = U}} m_1(U_1)m_2(U_2)$$

where the normalisation constant c is defined by

$$c = \sum_{\substack{U_1, U_2 \subseteq W, \\ U_1 \cap U_2 \neq \emptyset}} m_1(U_1)m_2(U_2)$$

Rule of Combination – Example

Suppose that a physician sees a case of jaundice.

She considers four possible hypotheses regarding its cause: hepatitis (hep), cirrhosis (cirr), gallstone (gall), and pancreatic cancer (pan).

- Suppose that these are the only causes of jaundice, and that a patient with jaundice suffers from exactly one of those problems.
- Thus, the physician can take the set W of possible worlds to be $\{hep, cirr, gall, pan\}$.
- Only subsets of 2^W are of diagnostic significance.

Rule of Combination – Example

Suppose there are two types of test:

- There are tests that support each of the individual hypotheses,
- and tests that support intrahepatic cholestasis, $\{hep, cirr\}$, and extrahepatic cholestasis, $\{gall, pan\}$;
- the latter two tests do not provide further support for the individual hypotheses.

Rule of Combination – Example

To begin with, suppose that a single test is carried out that provides evidence for intrahepatic cholestasis to degree 0.6.

That is, no combination of evidences has to take place:

- This can be represented by a mass function that assigns 0.6 to $\{hep, cirr\}$ and the remaining 0.4 to W.
- The fact that the test provides support only 0.6 $\{hep, cirr\}$ does not mean that it provides support 0.4 for its complement, $\{gall, pan\}$.
- Rather, the remaining 0.4 is viewed as uncommitted. As a result $Bel(\{hep, cirr\}) = 0.6$ and $Plaus(\{hep, cirr\}) = 1$.

Rule of Combination – Example

Now, suppose that two tests are carried out:

- The first confirms hepatitis to degree 0.8 and says nothing about the other hypotheses;
- this is captured by the mass function m_1 such that $m_1(\{hep\}) = 0.8$ and $m_1(W) = 0.2$.
- The second confirms intrahepatic cholestasis to degree 0.6; it is captured by the mass function m_2 such that $m_2(\{hep, cirr\}) = 0.6$ and $m_2(W) = 0.4$).

A straightforward computation shows that $c = 1$, and

$$(m_1 \oplus m_2)(\{hep\}) = 0.8$$
$$(m_1 \oplus m_2)(\{hep, cirr\}) = 0.12$$
$$(m_1 \oplus m_2)(W) = 0.08$$

Fuzzy Logic – A Very Brief Outlook

- Fuzzy logic may be viewed as an extension of classical logical systems.
- It provides an effective conceptual framework for dealing with applications in KR in an environment of uncertainty and imprecision.

> Example: Most experts believe that the likelihood of a severe earthquake in the near future is very low.

FOL and classical probability theory lack the means for representing the meaning of fuzzy concepts.

Basic Idea

A fuzzy set is a mapping F from $\mathcal{S}$ (the set of possible states of affairs) to a totally ordered set $\mathcal{L}$ often chosen to be the unit interval $[0, 1]$.

- The value F(s) is the membership degree of the element s in F.
- It evaluates the compatibility between the situation s and the predicate F.
- The membership degree can be seen as a degree of truth of a proposition.
- $\mathcal{L}$ has a natural ordering $\leq$, ranging from total falsity (represented by 0) to total truth (represented by 1) through a continuum of intermediate truth degrees.

Basic Idea – Formally

Connectives are to be interpreted truth-functionally over the set of truth-degrees. Such truth-functions are assumed to behave classically on the extremal values 0 and 1.

A very natural behavior of conjunction and disjunction is achieved by imposing $x \wedge y = min\{x, y\}$ and $x \vee y = max\{x, y\}$ for each $x, y \in [0, 1]$.

Another, non-idempotent, conjunction is typically added:

- It is interpreted by a binary operation on [0,1], which is still associative, commutative, non-decreasing in both arguments and has 1 as neutral element.
- It is based on the idea that applying partially true hypothesis twice might lead to a different degree of truth than using it only once.

A Dangerous Confusion

Notice the difference between a degree of truth (1) and a degree of certainty (2):

(1) John is very young.

(2) John is probably young.

(1) expresses the fact that the degree of membership of $age(John)$ to the fuzzy set of young ages is for sure high.

For instance, take $age(John) = 22$. The degree of membership $F(s)$ represents the degree of adequacy of a fuzzy category, here $F =$ young, to a state of affairs, here $s = 22$.

(2) Here, it is not ruled out that John is not young at all.

Choosing a Representation

Summary of our models of uncertainty:

- Probability has the advantage of being well understood. However, probability theory has some drawbacks when there is uncertainty about the likelihood.
- Belief functions may prove useful as a model of evidence, especially when combined with Dempster's Rule of Combination. It has the resources designed to model severe uncertainty.
- Fuzzy logic is of great use for approximate reasoning: when information is not only uncertain but also lexically imprecise.