Theory of belief functions: Preliminaries

Thierry Denœux

Université de Technologie de Compiègne, France
HEUDIASYC (UMR CNRS 7253)
https://www.hds.utc.fr/˜tdenoeux

Beijing University of Technology,
Beijing, China, June 2016

Outline

1. Introduction

2. Set-based representation
   - Updating/combining information
   - Measuring information/uncertainty
   - Projection, extension

3. Probabilistic representation
   - Conditioning
   - Measuring uncertainty
   - Marginalization

Contents of this course

1. This course is about the theory of belief functions, a formal framework for reasoning and making decisions under uncertainty.
2. This framework originates from Arthur Dempster's seminal work on statistical inference with lower and upper probabilities.
3. It was then further developed by Glenn Shafer who showed that belief functions can be used as a general framework for representing and reasoning with uncertain information.
4. Also known as Evidence theory or Dempster-Shafer theory.
5. Many applications in several fields such as artificial intelligence, information fusion, pattern recognition, etc.
6. Recently, there has been a revived interested in its application to statistical inference.

Uncertainty

“There are some things that you know to be true, and others that you know to be false; yet, despite this extensive knowledge that you have, there remain many things whose truth or falsity is not known to you. We say that you are uncertain about them. You are uncertain, to varying degrees, about everything in the future; much of the past is hidden from you; and there is a lot of the present about which you do not have full information. Uncertainty is everywhere and you cannot escape from it.”

Dennis Lindley, “Understanding uncertainty”.

Uncertainty in statistics

- Uncertainty is a fundamental issue in statistics.
- Statistical inference:
  - Which statements can we make about a population, after observing a random sample?
  - How to predict the result of a random experiment?
- Approaches:
  - Frequentists confidence and prediction sets
  - Bayesian inference
  - Fiducial inference
  - Likelihood-based inference
  - ...

Formalization

- As a first step towards a formalization of uncertainty, one need to make some assumptions and define some notations.
- We suppose that we are concerned with some question Q, for which there exists one and only one correct answer X, which is assumed to be an element of some set Ω, called the frame of discernment.
- Assumptions:
  1. there is one and only one correct answer, and
  2. it is contained in Ω.
- The second assumption, in particular, is debatable for some problems. We will leave aside this difficulty for the moment.

Classical models of uncertainty

- Two classical models for expressing beliefs, or uncertain information:
  1. Sets (e.g. interval analysis), propositions;
  2. probabilities.
- Each of these models has methods for
  1. Updating beliefs based on new evidence;
  2. Measuring information/uncertainty;
  3. Converting belief representations from a frame to a finer, or a coarser frame.

Outline

1 Introduction

2 Set-based representation
- Updating/combining information
- Measuring information/uncertainty
- Projection, extension

3 Probabilistic representation
- Conditioning
- Measuring uncertainty
- Marginalization

Set-based representation of uncertainty

- Perhaps the simplest way of representing partial knowledge about some question is as a set $A \subseteq \Omega$ that certainly contains the true answer ω.
- There is a vast literature on set-membership approaches to uncertainty, with application, e.g., in computer science and automatic control.
- An important special case is interval arithmetics, which includes syntactic rules to compute with intervals, making it possible to produce rigorous enclosures of solutions to model equations.
- In statistics, a confidence set (region) contains the true value of the parameter of interest, for some fixed proportion of the samples.

Outline

1. Introduction

2. Set-based representation
   - Updating/combining information
   - Measuring information/uncertainty
   - Projection, extension

3. Probabilistic representation
   - Conditioning
   - Measuring uncertainty
   - Marginalization

Conjunctive combination

- Assume that two sources provide two subsets A and B of Ω, assumed to contain the answer to the question of interest. How to combine these pieces of information?

Conjunctive combination

- Assume that two sources provide two subsets A and B of Ω, assumed to contain the answer to the question of interest. How to combine these pieces of information?
- If both sources can be trusted, then it is reasonable to consider that the true answer is in the intersection of A and B, denoted by $A \cap B$, which is the set containing the elements of Ω that belong to both A and B.
- This mode of fusing information is called conjunctive; it is relevant when all information sources are assumed to be reliable.

Disjunctive combination

- However, when A and B are disjoint, i.e., $A \cap B = \emptyset$, this rule leads a contradiction. In that case, the assumption that the two sources can be trusted can no longer be valid.
- How to combine information in this case?

Disjunctive combination

- However, when A and B are disjoint, i.e., $A \cap B = \emptyset$, this rule leads a contradiction. In that case, the assumption that the two sources can be trusted can no longer be valid.
- How to combine information in this case?
- It is then more cautious to conclude that the true answer is in the union of A and B, denoted by $A \cup B$, which is the set containing the elements of Ω that belong to A or B.
- This is the simplest form of disjunctive rule for pooling information, which is suitable when at least one of the sources is assumed to be reliable.

Outline

1 Introduction

2 Set-based representation

- Updating/combining information
- Measuring information/uncertainty
- Projection, extension

3 Probabilistic representation

- Conditioning
- Measuring uncertainty
- Marginalization

Comparing information content

- Consider two statements: "$X \in A$" and "$X \in B$", where $A, B \subseteq \Omega$.
- The statement "$X \in A$" is more precise/specific/informative than "$X \in B$" if $A \subseteq B$.
- We thus have a way to compare the information content of two pieces of information about the same variable X.
- How to measure quantitatively the information content of such statements?

Measuring uncertainty

- Assume we receive some piece of information of the form $X \in A$ for some non-empty subset A of Ω.
- The amount of uncertainty associated with that statement can be measured by the amount of information needed to remove the uncertainty.
- Such a measure should naturally be a function of the cardinality of A. Let $h : \mathbb{N} \to [0, +\infty)$ be such a function.
- Requirements:

  (H1) Monotonicity $h(s) < h(s+1)$.
  (H2) Normalization $h(2) = 1$.
  (H3) Additivity $h(r \cdot s) = h(r) + h(s)$.

Measuring uncertainty

Meaning of H_3

- Consider a partition of Ω into r subsets of s elements.
- Characterizing an element of Ω requires the amount $h(r \cdot s)$ of information.
- However, we could also proceed in two steps: first, we could characterize the subset to which the element belongs (requiring an amount $h(r)$ of information, and then characterize the element in this subset (with required information $h(s)$).
- The equivalence of the two methods leads to Axiom $H3$.

Hartley function

- It can be shown that the only function h verifying these three axioms is defined by
$$h(n) = \log_2 n.$$
- The function $H : 2^\Omega \setminus \emptyset \to [0, +\infty)$ defined by
$$H(A) = \log_2 |A|$$
is called the Hartley function.
- It is a measure of the uncertainty of the statement "$X \in A$". Its range is $[0, \log_2 |\Omega|]$.

Outline

1 Introduction

2 Set-based representation
- Updating/combining information
- Measuring information/uncertainty
- Projection, extension

3 Probabilistic representation
- Conditioning
- Measuring uncertainty
- Marginalization

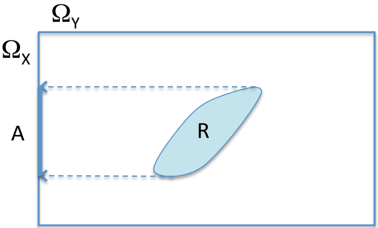
Cartesian products, relations

- Let us now assume that we have two questions of interest, whose true answers are denoted by X and Y (X and Y may be called variables).
- Let Ω_X and Ω_Y be the sets of possible values for X and Y. To represent information about the values that X and Y may take jointly, we need to place ourselves in the Cartesian product $\Omega_X \times \Omega_Y$, denoted more concisely by Ω_{XY}, and defined as the set of ordered pairs (x, y) of an element of Ω_X and an element of Ω_y.
- A subset of R of Ω_{XY} is called a relation. It can be used to represent a constraint on the values that X and Y may take jointly.

Projection

- Let R be a relation on Ω_{XY}.
- The projection of R onto Ω_X, denoted by $R\downarrow\Omega_X$, is the subset of Ω_X containing all $x\in\Omega_X$ such that $(x,y)\in\Omega_{XY}$ for some y:

$$R\downarrow\Omega_X=\{x\in\Omega_X|\exists y\in\Omega_Y,(x,y)\in R\}.$$

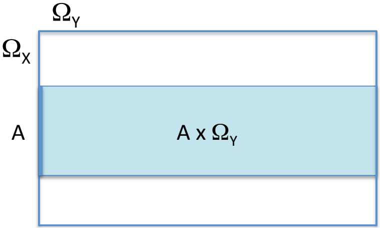


Cylindrical extension

- Conversely, let A be a subset of Ω_X. There are many subsets of Ω_{XY} whose projection on Ω_X is A. The largest one is

$$A \uparrow \Omega_{XY} = A \times \Omega_Y = \{(x,y) \in \Omega_{XY} | x \in A\}.$$

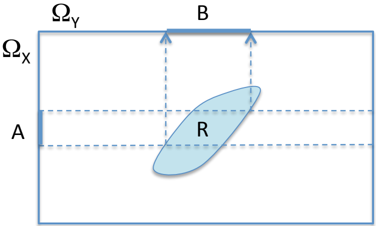
- It is called the cylindrical extension of A in Ω_{XY}.



Reasoning with relations

- Assume that we get
  - Evidence that X belongs to a subset A of Ω_X;
  - Evidence about the values that X and Y can take jointly, represented by a relation $R \subseteq \Omega_{XY}$.
- What can we deduce about Y?
- Let B denote the set of possible values for Y. It is clear that

$$B = \{y \in \Omega_Y | \exists x \in A, (x, y) \in R\} = (R \cap (A \uparrow \Omega_{XY})) \downarrow \Omega_Y$$



Outline

1. Introduction

2. Set-based representation
   - Updating/combining information
   - Measuring information/uncertainty
   - Projection, extension

3. Probabilistic representation
   - Conditioning
   - Measuring uncertainty
   - Marginalization

Basic definition

- Let us first assume the frame Ω to be finite. A probability mass function on Ω is a mapping $p : \Omega \to [0,1]$ such that

$$\sum_{\omega \in \Omega} p(\omega) = 1.$$

- The mass $P(A)$ assigned to $A \subseteq \Omega$ is called the probability of A,

$$P(A) = \sum_{\omega \in A} p(\omega),$$

  and the mapping $P : 2^\Omega \to [0,1]$ is called a probability measure.

- It verifies the following properties,
  1. $P(\Omega) = 1$;
  2. For all elements A and B of 2^Ω such that $A \cap B = \emptyset$,

$$P(A \cup B) = P(A) + P(B).$$

Finitely additive probabilities (general case)

- Let Ω be a set and $\mathcal{A} \subseteq 2^{\Omega}$ an algebra of subsets of Ω, defined as non-empty collection of subsets of Ω (called events), closed under complementation and finite union, i.e., for all A and B in $\mathcal{A}$, $A \cup B \in \mathcal{A}$. We can remark that Ω necessarily belongs to $\mathcal{A}$.
- A finitely additive probability measure on $(\Omega, \mathcal{A})$ is a function P from $\mathcal{A}$ to $[0, 1]$ such that
  1. $P(\Omega) = 1$;
  2. For all elements A and B of $\mathcal{A}$ such that $A \cap B = \emptyset$,
  $$P(A \cup B) = P(A) + P(B).$$
- Stricter notions: σ-algebra, countably additive probability measure.

Interpretations of probabilities

- The mathematical model briefly described above may be used to represent different aspects of the real world.
- In particular, it can be used to represent
  - objective properties of random experiments (limits of frequences), or
  - subjective degrees of belief.
- Question: why should degrees of belief be additive?
- There are some arguments, but not very compelling. Alternative theories of uncertainty exist, in which the additivity axiom is replaced by some weaker axiom.

Outline

1 Introduction

2 Set-based representation
- Updating/combining information
- Measuring information/uncertainty
- Projection, extension

3 Probabilistic representation
- Conditioning
- Measuring uncertainty
- Marginalization

Bayes' conditioning

- Let P be a probability measure on Ω representing your beliefs about some question. Assume that you learn that the truth lies for sure in some subset $E \subset \Omega$. How should you update your beliefs to account for this new information?
- If $P(E) > 0$, then you can construct a new probability measure P^* represent your new belief state after receiving the new information.
- Assume that you impose the following requirements on P^*
  1. $P^*(E) = 1$.
  2. For all $A, B \subseteq \Omega$, $P^*(B) = 0$ if $P(B) = 0$, and

$$\frac{P^*(A)}{P^*(B)} = \frac{P(A)}{P(B)} \quad \text{if } P(B) > 0.$$

- The only probability measure verifying these requirements is

$$P^*(A) = P(A|E) = \frac{P(A \cap E)}{P(E)}, \quad \forall A \subseteq \Omega.$$

Jeffrey's conditioning

- Jeffrey (1965) proposed to extend Bayes' rule as follows. Let $E_1, \ldots, E_n$ be a partition of Ω.
- Assume that you receive some new evidence, and the effect of that evidence does not go beyond changing your degrees of beliefs about $E_1, \ldots, E_n$. Let $q_1, \ldots, q_n$ be these degrees of belief.
- We thus want to construct an updated probability measure P^* such that $P^*(E_i) = q_i$, and $P^*(A|E_i) = P(A|E_i)$ for all $A \subseteq \Omega$ and $i = 1, \ldots, n$.

Jeffrey's conditioning (continued)

- The unique solution is

$$P^*(A) = \sum_{i=1}^{n} P(A|E_i) q_i.$$

- Bayes' rule is recovered as a special case with $(E_1, E_2) = (E, \overline{E})$, $q_1 = 1$ and $q_2 = 0$.
- This operation is called Jeffrey's conditioning. It is a mechanism for updating probabilistic knowledge in light of uncertain evidence.

Outline

1. Introduction

2. Set-based representation
   - Updating/combining information
   - Measuring information/uncertainty
   - Projection, extension

3. Probabilistic representation
   - Conditioning
   - Measuring uncertainty
   - Marginalization

Shannon entropy

- Let $I(p)$ be a measure of how much information is acquired due to the observation of an event with probability p. Requirements:
  - $I(p) \geq 0$ – information is a non-negative quantity
  - $I(1) = 0$ – sure events do not communicate information
  - $I(p_1 p_2) = I(p_1) + I(p_2)$ – information due to independent events is additive
- The unique solution is $I(p) = \log(1/p)$.
- For a mass function $p : \Omega \rightarrow [0, 1]$, the Shannon entropy is the expected information,

$$S(p) = -\sum_{\omega \in \Omega} p(\omega) \log p(\omega)$$

- It is a measure of the uncertainty in p.
- It ranges from 0 to $\log(|\Omega|)$ (uniform distribution).

Maximum entropy principle

- The maximum entropy (ME) principle is, when a probability distribution is partially specified, to choose the one with the largest entropy.
- It is a principle of maximum uncertainty
- In particular, we we do not know anything, we get the uniform distribution.
- This is also known as the Principle of Indifference (PI).

Outline

1. Introduction

2. Set-based representation
   - Updating/combining information
   - Measuring information/uncertainty
   - Projection, extension

3. Probabilistic representation
   - Conditioning
   - Measuring uncertainty
   - Marginalization

Marginalization

- Assume that we have two finite frames of discernment Ω_X and Ω_Y for two different questions. Let P_{XY} be a probability measure on $\Omega_{XY} = \Omega_X \times \Omega_Y$ with probability mass function p_{XY}.
- The marginal probability mass function on Ω_X is $p_{XY\downarrow X}$defined by

$$p_{XY\downarrow X}(x) = P_{XY}(\{x\} \times \Omega_Y) = \sum_{y\in\Omega_Y} p_{XY}(x,y).$$

- Marginalization plays the same role as projection in the set-membership setting.

Recovering a joint distribution from marginals

- The inverse operation of marginalization is recovering a joint distribution from a marginal.
- Specifically, let p_X be a probability mass function on Ω_X. Problem: find a joint distribution p_{XY} on Ω_{XY} such that $p_X = p_{XY\downarrow X}$.
- There are infinitely many solutions. The solution given by the ME principle is obtained by maximizing $S(p_{XY})$ under the constraints

$$\sum_{y\in\Omega_Y} p_{XY}(x,y) = p_X(x), \quad \forall x \in \Omega_X.$$

- Solution:

$$p_{XY}(x,y) = \frac{p_X(x)}{|\Omega_Y|}.$$

Expression of set-valued information using probabilities

- Assume that we receive information in the form $X \in A$.
- Can we express this information as a probability mass function p_X?
- PI: the best translation is the uniform distribution,

$$p_A(x) = \frac{1_A(x)}{|A|} \quad \forall x \in \Omega.$$

- For any $A \subseteq \Omega$, the Hartley measure $H(A)$ is equal to the Shannon entropy of the uniform probability distribution p_A on A.
- However, we face some paradoxes.

The wine/water paradox

There is a certain quantity of liquids. All that we know about the liquid is that it is composed entirely of wine and water, and the ratio of wine to water is between 1/3 and 3.

What is the probability that the ratio of wine to water is less than or equal to 2?

The wine/water paradox (continued)

- Let X denote the ratio of wine to water. All we know is that $X \in [1/3, 3]$. According to the PI, $X \sim \mathcal{U}_{[1/3,3]}$. Consequently

$$P(X \leq 2) = (2 - 1/3)/(3 - 1/3) = 5/8$$

- Now, let $Y = 1/X$ denote the ratio of water to wine. All we know is that $Y \in [1/3, 3]$. According to the PI, $Y \sim \mathcal{U}_{[1/3,3]}$. Consequently

$$P(Y \geq 1/2) = (3 - 1/2)/(3 - 1/3) = 15/16$$

- However, $P(X \leq 2) = P(Y \geq 1/2)$!

Need for a unified framework

- The reason for the wine/water paradox is that, if X has a uniform distribution on some set A, and if f is a nonlinear mapping, $f(X)$ does not have, in general, a uniform distribution on $f(A)$.
- However, if we only know that X is in A, we only know that $f(X)$ is in $f(A)$.
- This argument shows that set-valued information cannot be adequately represented by a probability measure.
- We thus need a formalism that allows us to represent both set-valued information and probabilistic information in the same setting.
- Two approaches:
  1. Define sets of probability measures (Imprecise Probability)
  2. Assign probabilities to sets (Belief Functions)

Theory of belief functions

Main idea

- The theory of belief functions extends both the set-membership approach and Probability Theory
  - A belief function may be viewed both as a generalized set and as a non additive measure
  - The theory includes extensions of probabilistic notions (conditioning, marginalization) and set-theoretic notions (intersection, union, inclusion, etc.)
- Dempter-Shafer reasoning produces the same results as probabilistic reasoning or interval analysis when provided with the same information
- However, the greater expressive power of the theory of belief functions allows us to represent what we know in a more faithful way.