Introduction of belief functions
Lecture 1: Uncertainty

T. Denœux

July 18, 2014

This course is about the theory of belief functions, a formal framework for reasoning and making decisions under uncertainty. This framework originates from Arthur Dempster's seminal work on statistical inference with lower and upper probabilities [2, 3]. It was then further developed by Glenn Shafer [4] who showed that belief functions can be used as a general framework for representing and reasoning with uncertain information, beyond the very important but limited confines of statistical inference. The theory of belief functions, also referred to as Evidence theory or Demspter-Shafer theory, has been widely used in several areas such as Artificial Intelligence, Information Fusion and Risk Analysis. Recently, there has been a revived interested in its application to statistical inference. This formalism seems particularly well suited to situations where we are facing limited information such as uncertain and low quality data, partially reliable and conflicting expert opinions, or both. There has been thousands of applications in many domains, including engineering, medicine, economics, etc.

In this introductory chapter, we will discuss the concept of uncertainty and review the main theories of uncertainty. As we shall see, these theories have shortcomings, which motivate the development of more general models, such as the theory of belief functions.

1 Sources of uncertainty

Uncertainty is ubiquitous in every area of human activity. Typically, we are interested in some question Q, such as: What is the mean value of some variable in a population? What will be the economic growth rate in Thailand next year? What was the amount of carbon dioxide emission in China in 2012? etc. In the following, we will denote by Ω the set of possible answers (one and only one is assumed to be true), and by $\boldsymbol{\omega}$ the true answer. If we know the exact value of $\boldsymbol{\omega}$, this is a situation of complete certainty. If we know nothing at all (except that $\boldsymbol{\omega}$ is in Ω), we have complete uncertainty. Actually, these two extreme situations are not frequent: usually, we have only partial knowledge of $\boldsymbol{\omega}$, based on limited evidence about the question of interest. The issue then arises of how to represent such partial information in such a way that it can be used for further reasoning, computation and rational decision making.

It has become customary in some areas (such as risk analysis) to distinguish between two main sources of uncertainty:

1. When the question of interest concerns some property of an object taken at random from a well-defined population (such as, e.g., the annual income of an household taken at random from the population of Thai households), we say that we have *random* or *aleatory* uncertainty. Such uncertainty cannot be reduced between it depends on the physical property of the population and of the random experiment.

2. In many situations, uncertainty does not arise from randomness but from lack of knowledge. For instance, the mean annual world temperature at the end of the 21th century is unknown, but it is not random because there is no notion of random experiment (in particular, the global warming process in the 21th century will happen only once). Such uncertainty is said to be *epistemic*. It can be reduced by acquiring further information related to the question of interest.

The two main classical formalisms for representing uncertainty are the set-based representation (such as interval analysis) and probability theory. These approaches will be discussed below, with greater emphasis on probability theory, which is by far the most general and widely used framework.

2 Set-based representations 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 $\boldsymbol{\omega}$.

When $\Omega \subseteq \mathbb{R}$, we usually restrict ourselves to intervals specifying the possible range of the quantity of interest. Propagating interval uncertainty in equations is usually much easier that propagating probability distributions. For instance, if we know that $X \in [a, b]$ and $Y \in [c, d]$, we can assert that $X + Y \in [a + c, b + d]$. A whole set of methods, known as Interval Analysis, has been developed to propagate interval uncertainty in equations while guaranteeing that the computed intervals always contain the true values of quantities of interest. When $\Omega \subseteq \mathbb{R}^p$ with $p > 1$, typical set representations are boxes (Cartesian products of intervals), unions of boxes, and ellipsoids.

The main limitation of set-based representations of uncertainty such as interval analysis is that they do not allow the expression of doubt. As a consequence, they favor a conservative approach, in which the sets have to be chosen very large to contain the true value with full certainty. A lot of information is usually lost in such a representation. For instance, if an expert is asked to give an interval that surely contains the annual inflation rate in Thailand in the next year, he will give a wide interval, even though he may actually believe that the inflation rate will be contained within narrower bounds. As we will show later, belief functions can be seen as extending the set-based representation of uncertainty by allowing us to provide different sets with attached degrees of support.

3 Probabilistic representation of uncertainty

3.1 Objective probabilities

Probability Theory (PT) is the most widely used mathematical model of uncertainty. It is clearly suitable to represent aleatory uncertainty, in which case the probability $P(A)$ for an event $A \subseteq \Omega$ is interpreted either as a *frequency* (actually, the limit of the frequency with which event A occurs, if the random experiment is repeated n times and $n \to +\infty$), or a as a *propensity* (i.e., the tendency of A to happen across a large number of repetitions of the random experiment). Since frequencies are additive, the additivity axiom of PT is well justified, i,e., for any two events A and B, we should have

$$P(A \cup B) = P(A) + P(B) - P(A \cap B). \quad (1)$$

Such probabilities can be considered as objective, because they describe physical properties of the chance setup. For instance, when tossing a coin, the probabilities $P(Heads) = P(Tails) = 1/2$ can be deduced from the symmetry of the coin.

3.2 Subjective probabilities

The use of probability measures to represent epistemic uncertainty (as advocated by the Bayesian school) is more problematic, because in this case probabilities can clearly no longer be interpreted as frequencies. In this context, they are usually interpreted as subjective (or personal) *degrees of belief*. However, we need to define more precisely the meaning of this notion and to explain why degrees of belief should be additive. This can be done in, at least, two ways: using a constructivist or a behavioral approach.

Constructivist approach In the constructivist approach, we construct a probability measure P by comparing our evidence (i.e., what we know) about Ω to a random experiment with known chances [5]. This allows us a construct a scale of degrees of belief, with canonical examples. For instance, in a coin tossing game, the chance for Heads is 1/2, which is taken as our degree of belief that Heads will come up. If our beliefs about the truth of some proposition A (e.g., "There is life on Mars") is comparable to our belief that Heads will come up when tossing a coin, we can say that our personal probability for A is 1/2.

Behavioral approach In the behavioral approach, we assume that the belief state of an agent can be deduced from observing its betting behavior. Assume that you have to enter a game where there is a player and a banker. The player gives an amount of money $\$p$ to the banker and the banker gives the player \$1 if an proposition A is true, and 0 otherwise. You do not know if you will be the banker or the player, and you are asked to fix p. By definition, your fair betting rate $P(A) = p$ is equated to your personal probability of proposition A. It is assumed to measure your belief in A: the more you believe in A, the more money you will be willing to give to enter the game. Now, the main point is that an opponent can compile a book of bets from your offer that assures a net gain from you (a Dutch book) if and only if P fails to be a probability function.

To show this, consider two disjoint events A and B and the three following bets:

1. Bet 1: you gain \$1 if A is true and 0 otherwise.
2. Bet 2: you gain \$1 if B is true and 0 otherwise.
3. Bet 3: you gain \$1 if $A \cup B$ is true and 0 otherwise.

Let $P(A)$, $P(B)$ and $P(A \cup B)$ be the fair prices you are willing to pay for the three tickets. Assume that $P(A \cup B) < P(A) + P(B)$. Then, the opponent can raise a Dutch book against you by deciding that you will be the player in the first two bets and the banker in the third bet. Similarly, if $P(A \cup B) > P(A) + P(B)$, you will lose if you are the banker in the first two bets and the player in the third bet.

If we interpret degrees of belief as betting rates, it can thus be argued that degrees of belief should be additive and our state of knowledge should be represented by a probability measure. However, this point of view is open to criticism:

1. First, the betting scheme just described is a highly idealized situation, and it is debatable if all situation of choice under uncertainty can fit this idealized picture (probabilities and utilities do not exist, they are a construction). Additionally, it is not obvious that the setting of betting rates in some particular betting scheme is the primary purpose of probability judgement.
2. Secondly, by slightly changing the story, we can arrive at different conclusions. For instance, if you are not obliged to enter the game and if you are not required to accept to be the banker and $P_*(A)$ is the highest price you are willing to pay for the lottery ticket, then a Dutch book can be raised against you iff P_* fails to be a lower probability function, i.e., the lower envelope of a family of probability measures.

3.3 Cox axioms

Some scholars have attempted to justify the use of probabilities to represent degrees of belief using an axiomatic approach. In particular, Cox axioms [1] and Savage's axioms are often invoked by Bayesians to argue that PT is the only "reasonable" formalism for reasoning with uncertainty. In this section, we will briefly discuss Cox axioms. Savage's axioms will be discussed in the chapter on decision making.

Let $Cr(A|B) \in \mathbb{R}$ be a measure of the "credibility" of proposition A, given that B is true, where A and B are non-empty subsets of Ω. Consider the following axioms:

A1 The credibility of the complement of A can be computed from the credibility of A:
$$Cr(\overline{A}|B) = S[Cr(A|B)]. \tag{2}$$

A2 $Cr(A \cap A'|B) = F[Cr(A'|A \cap B), Cr(A|B)]$.

Then, if S is twice differentiable and if F is twice differentiable with a continuous derivative, then Cr is isomorphic to a probability distribution, in the sense that there exists a one-to-one mapping $g : \mathbb{R} \to \mathbb{R}$ such that $g \circ Cr$ is a probability measure, and

$$g[Cr(A|B)] \cdot g[Cr(B)] = g[Cr(A \cap B)] \tag{3}$$

for any A and non-empty B, with $Cr(B) = Cr(B|\Omega)$.

Significant as it may be, this result can hardly be considered as a final justification of probabilities for representing degrees of belief. Indeed, close inspection of the axioms shows that they can be seriously questioned.

The first assumption is that the credibility of a proposition can be represented by a single number. This condition is not assumed in some alternative theories of uncertainty, such as the theory of belief functions. Axiom $A1$ is also quite debatable. If degrees of credibility are identified with degrees of support, the degree of support for some proposition is not a function of the degree of support for its negation (if A is not supported, $\overline{A}$ may be supported or not), and $Cr(A|\Omega)$ will not be determined by $Cr(\overline{A}|\Omega)$.

Cox merely justifies axiom A2 by an example. If A is the proposition that some athlete can run to some point, given the conditions of the race expressed by B, and if A' denotes the proposition that he can come back, then the probability that he can run to the point and come back depends on the probability that he can come back, given that he has already reached the point, and the probability that he can reach the point. Yet, as noted by Shafer, even admitting that $Cr(A \cap A'|B)$ should be a function of $Cr(A'|A \cap B)$ and $Cr(A|B)]$, it is not obvious that the same function F should always be used.

3.4 Two paradoxes

As shown in the previous section, attempts to justify the use of probabilities to represent degrees of belief have not settled the question. In contrast, there appears to be some serious arguments against the use of PT as a model of epistemic uncertainty (Bayesian model) In particular, the use of a probability distribution to represent ignorance may lead to some inconsistencies, and PT does not seem to be a plausible model of how people make decisions based on weak information. These arguments are exemplified by the following two paradoxes.

The wine/water paradox Assume that all we now about some quantity X is that it belongs to some set A. According to Laplace's principle of indifference (PI) – and also according to the principle of maximal entropy, this state of knowledge should be represented by assigning equal probabilities to any possible values of X. However, consider the following paradox, attributed to Von Mises.

Consider a certain quantity of liquids. All we know is that this liquid 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?

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:

$$P(X \leq 2) = (2 - 1/3)/(3 - 1/3) = 5/8. \tag{4}$$

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:

$$P(Y \geq 1/2) = (3 - 1/2)/(3 - 1/3) = 15/16. \tag{5}$$

By comparing (4) and (5), we can see that we have a paradox, as the propositions $X \leq 2$ and $Y \geq 1/3$, being logically equivalent, should receive the same probability.

The reason for this paradox is that, if X has a uniform distribution on some set A, and if f is a non linear 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.

Ellsberg's paradox Suppose you have an urn containing 30 red balls and 60 balls, either black or yellow. You are given a choice between two gambles:

- f_1: You receive 100 euros if you draw a red ball;
- f_2: You receive 100 euros if you draw a black ball.

Also, you are given a choice between these two gambles (about a different draw from the same urn):

- f_3: You receive 100 euros if you draw a red or yellow ball;
- f_4: You receive 100 euros if you draw a black or yellow ball.

Most people strictly prefer f_1 to f_2, hence $P(red) > P(black)$, but they strictly prefer f_3 to f_4, hence $P(black) > P(red)$.

This famous paradox shows that PT is not a plausible descriptive model of how people make decisions under ambiguity (i.e., when objective probabilities are not given).

4 Conclusions

The two main formalisms for representing uncertain information are set-based representations and probability theory. We have shown in this lecture that none of these two formalisms seems to be sufficient to represent all kinds of uncertainties. In the next lecture, we will introduce the theory belief functions, which can be seen as generalizing the two classical frameworks outlined above.

References

[1] R. T. Cox. Probability, frequency and reasonable expectation. *American Journal of Physics*, 14(1):1–12, 1946.

[2] A. P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, 38:325–339, 1967.

[3] A. P. Dempster. A generalization of Bayesian inference (with discussion). *J. R. Statistical Society B*, 30:205–247, 1968.

[4] G. Shafer. *A mathematical theory of evidence.* Princeton University Press, Princeton, N.J., 1976.

[5] G. Shafer. Constructive probability. *Synthese*, 48(1):1–60, 1981.