Statistics and Machine Learning using belief functions

Lecture 1 – Representation and Combination of Evidence

Thierry Denœux

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

Beijing University of Technology
May 2017

Topic of this seminar

1. This course is about the theory of belief functions and its applications to Statistics and Machine Learning.
2. What is the Theory of Belief Functions?
   - A formal framework for reasoning and making decisions under uncertainty.
   - Originates from Arthur Dempster's seminal work on statistical inference with lower and upper probabilities.
   - 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.
   - Also known as Evidence theory or Dempster-Shafer theory.
3. Many applications in several fields such as artificial intelligence, information fusion, pattern recognition, etc.
4. Recently, there has been a revived interested in its application to Statistical Inference and Machine Learning (classification, clustering).

Outline of the seminar

1. Representation and combination of evidence
   Constructing Belief Functions from Sample Data Using Multinomial Confidence Regions. *International Journal of Approximate Reasoning* 42(3):228–252, 2006.
2. Decision-making and classification
   Analysis of evidence-theoretic decision rules for pattern classification. *Pattern Recognition* 30(7):1095–1107, 1997.
3. Clustering
   Evidential clustering of large dissimilarity data. *Knowledge-Based Systems* 106:179–195, 2016.
4. Learning from uncertain data
   Maximum likelihood estimation from Uncertain Data in the Belief Function Framework. *IEEE Trans. on Knowledge and Data Eng.* 25(1):119–130, 2013.
5. Estimation and prediction
   Prediction of future observations using belief functions: a likelihood-based approach. *International Journal of Approximate Reasoning* 72:71–94, 2016.

Outline

Outline

1. Representation of evidence
   - Mass functions
   - Belief and plausibility functions
2. Relations with alternative theories
   - Possibility theory
   - Imprecise probabilities
3. Combination of evidence
   - Dempster's rule
   - Disjunctive rule
   - Dubois-Prade rule
4. Predictive belief functions
   - Formalization
   - Method
   - Ordered data

Mass function

Definition

- Let X be a variable taking values in a finite set Ω (frame of discernment)
- Evidence about X may be represented by a mass function $m : 2^\Omega \to [0,1]$ such that

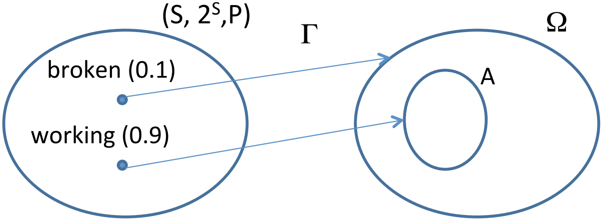
$$\sum_{A \subseteq \Omega} m(A) = 1$$

- Every A of Ω such that $m(A) > 0$ is a focal set of m
- m is said to be normalized if $m(\emptyset) = 0$. This property will be assumed hereafter, unless otherwise specified

Example: the broken sensor

- Let X be some physical quantity (e.g., a temperature), talking values in Ω.
- A sensor returns a set of values $A \subset \Omega$, for instance, $A = [20, 22]$.
- However, the sensor may be broken, in which case the value it returns is completely arbitrary.
- There is a probability $p = 0.1$ that the sensor is broken.
- What can we say about X? How to represent the available information (evidence)?

Analysis



- Here, the probability p is not about X, but about the state of a sensor.
- Let $S = \{\text{working}, \text{broken}\}$ the set of possible sensor states.
  - If the state is "working", we know that $X \in A$.
  - If the state is "broken", we just know that $X \in \Omega$, and nothing more.
- This uncertain evidence can be represented by a mass function m on Ω, such that

$$m(A) = 0.9, \quad m(\Omega) = 0.1$$

Source

- A mass function m on Ω may be viewed as arising from
  - A set $S = \{s_1, \ldots, s_r\}$ of states (interpretations)
  - A probability measure P on S
  - A multi-valued mapping $\Gamma : S \rightarrow 2^{\Omega}$
- The four-tuple $(S, 2^S, P, \Gamma)$ is called a source for m
- Meaning: under interpretation s_i, the evidence tells us that $X \in \Gamma(s_i)$, and nothing more. The probability $P(\{s_i\})$ is transferred to $A_i = \Gamma(s_i)$
- $m(A)$ is the probability of knowing that $X \in A$, and nothing more, given the available evidence

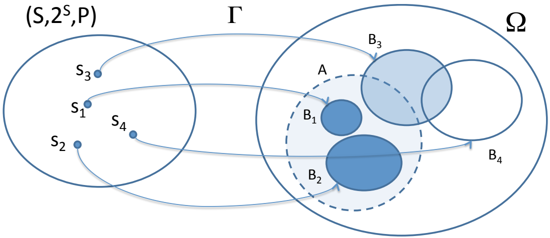
Special cases

- If the evidence tells us that $X \in A$ for sure and nothing more, for some $A \subseteq \Omega$, then we have a logical mass function m_A such that $m_A(A) = 1$
  - m_A is equivalent to A
  - Special case: $m_?$, the vacuous mass function, represents total ignorance
- If each interpretation s_i of the evidence points to a single value of X, then all focal sets are singletons and m is said to be Bayesian. It is equivalent to a probability distribution
- A Dempster-Shafer mass function can thus be seen as
  - a generalized set
  - a generalized probability distribution
- Total ignorance is represented by the vacuous mass function $m_?$ such that $m_?(\Omega) = 1$

Outline

Degrees of support and consistency

- Let m be a normalized mass function on Ω induced by a source $(S, 2^S, P, \Gamma)$.
- Let A be a subset of Ω.
- One may ask:
  1. To what extent does the evidence support the proposition $\omega \in A$?
  2. To what extent is the evidence consistent with this proposition?

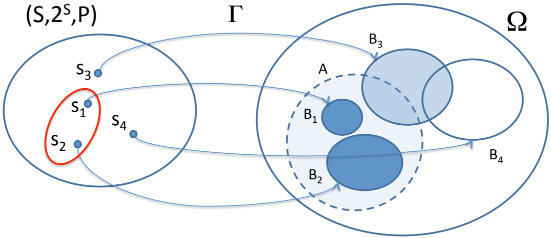


Belief function

Definition and interpretation

- For any $A \subseteq \Omega$, the probability that the evidence implies (supports) the proposition $X \in A$ is

$$Bel(A) = P(\{s \in S | \Gamma(s) \subseteq A\}) = \sum_{B \subseteq A} m(B).$$

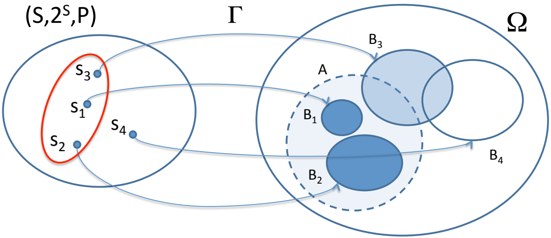


- The function $Bel : A \rightarrow Bel(A)$ is called a belief function.

Plausibility function

- The probability that the evidence is consistent with (does not contradict) the proposition $X \in A$

$$Pl(A) = P(\{s \in S | \Gamma(s) \cap A \neq \emptyset\}) = 1 - Bel(\overline{A})$$



- The function $Pl : \mathcal{A} \to Pl(A)$ is called a plausibility function.
- The function $pl : \omega \to Pl(\{\omega\})$ is called a contour function.

Two-dimensional representation

- The uncertainty about a proposition A is represented by two numbers: $Bel(A)$ and $Pl(A)$, with $Bel(A) \leq Pl(A)$
- The intervals $[Bel(A), Pl(A)]$ have maximum length when $m = m_?$ is vacuous: then, $Bel(A) = 0$ for all $A \neq \Omega$, and $Pl(A) = 1$ for all $A \neq \emptyset$.
- The intervals $[Bel(A), Pl(A)]$ have minimum length when m is Bayesian. Then, $Bel(A) = Pl(A)$ for all A, and Bel is a probability measure.

Broken sensor example

- From
$$m(A) = 0.9, \quad m(\Omega) = 0.1$$
we get
$$Bel(A) = m(A) = 0.9, \quad Pl(A) = m(A) + m(\Omega) = 1$$
$$Bel(\overline{A}) = 0, \quad Pl(\overline{A}) = m(\Omega) = 0.1$$
$$Bel(\Omega) = Pl(\Omega) = 1$$
- We observe that
$$Bel(A \cup \overline{A}) \geq Bel(A) + Bel(\overline{A})$$
$$Pl(A \cup \overline{A}) \leq Pl(A) + Pl(\overline{A})$$
- *Bel* and *Pl* are non additive measures.

Characterization of belief functions

- Function $Bel : 2^\Omega \rightarrow [0,1]$ is a completely monotone capacity: it verifies $Bel(\emptyset) = 0$, $Bel(\Omega) = 1$ and

$$Bel\left(\bigcup_{i=1}^{k} A_i\right) \geq \sum_{\emptyset \neq I \subseteq \{1,\ldots,k\}} (-1)^{|I|+1} Bel\left(\bigcap_{i \in I} A_i\right).$$

  for any $k \geq 2$ and for any family $A_1, \ldots, A_k$ in 2^Ω.
- Conversely, to any completely monotone capacity *Bel* corresponds a unique mass function m such that:

$$m(A) = \sum_{\emptyset \neq B \subseteq A} (-1)^{|A|-|B|} Bel(B), \quad \forall A \subseteq \Omega.$$

Relations between *m*, *Bel* et *Pl*

- Let m be a mass function, Bel and Pl the corresponding belief and plausibility functions
- For all $A \subseteq \Omega$,

$$Bel(A) = 1 - Pl(\overline{A})$$

$$m(A) = \sum_{\emptyset \neq B \subseteq A} (-1)^{|A|-|B|} Bel(B)$$

$$m(A) = \sum_{B \subseteq A} (-1)^{|A|-|B|+1} Pl(\overline{B})$$

- m, Bel et Pl are thus three equivalent representations of
  - a piece of evidence or, equivalently
  - a state of belief induced by this evidence

Least Commitment Principle

- It is sometimes interesting to compare two mass functions with respect to their information content.
- Let m_1 and m_2 be two mass functions on Ω. We say that m_1 is less committed than m_2 (noted $m_1 \sqsupseteq m_2$) if

$$Bel_1(A) \leq Bel_2(A), \quad \forall A \subseteq \Omega$$

  or, equivalently,

$$Pl_1(A) \geq Pl_2(A), \quad \forall A \subseteq \Omega$$

- Interpretation: m_1 and m_2 are consistent, but m_1 contains less information than m_2.
- Least Commitment Principle: when several belief functions are compatible with a set of constraints, the least informative according to some informational ordering (if it exists) should be selected.

Outline

Outline

Consonant belief function

- When the focal sets of m are nested: $A_1 \subset A_2 \subset \ldots \subset A_r$, m is said to be consonant
- The following relations then hold, for all $A, B \subseteq \Omega$,

$$Pl(A \cup B) = \max\left(Pl(A), Pl(B)\right)$$

$$Bel(A \cap B) = \min\left(Bel(A), Bel(B)\right)$$

- *Pl* is this a possibility measure, and *Bel* is the dual necessity measure

Contour function

- The contour function of a belief function *Bel* is defined by

$$pl(\omega) = Pl(\{\omega\}), \quad \forall \omega \in \Omega$$

- When *Bel* is consonant, it can be recovered from its contour function,

$$Pl(A) = \max_{\omega \in A} pl(\omega).$$

- The contour function is then a possibility distribution
- The theory of belief function can thus be considered as more expressive than possibility theory

From the contour function to the mass function

- Let pl be a contour on the frame $\Omega = \{\omega_1, \ldots, \omega_n\}$, with elements arranged by decreasing order of plausibility, i.e.,

$$1 = pl(\omega_1) \geq pl(\omega_2) \geq \ldots \geq pl(\omega_n),$$

  and let A_i denote the set $\{\omega_1, \ldots, \omega_i\}$, for $1 \leq i \leq n$.
- Then, the corresponding mass function m is

$$\begin{aligned} m(A_i) &= pl(\omega_i) - pl(\omega_{i+1}), \quad 1 \leq i \leq n-1, \\ m(\Omega) &= pl(\omega_n). \end{aligned}$$

Example

- Consider, for instance, the following contour distribution defined on the frame $\Omega = \{a, b, c, d\}$:

ω	a	b	c	d
$pl(\omega)$	0.3	0.5	1	0.7

- The corresponding mass function is

$$
\begin{aligned}
m(\{c\}) &= 1 - 0.7 = 0.3 \\
m(\{c, d\}) &= 0.7 - 0.5 = 0.2 \\
m(\{c, d, b\}) &= 0.5 - 0.3 = 0.2 \\
m(\{c, d, b, a\}) &= 0.3.
\end{aligned}
$$

Outline

Credal set

- A probability measure P on Ω is said to be compatible with Bel if
$$Bel(A) \leq P(A)$$
for all $A \subseteq \Omega$
- Equivalently, $P(A) \leq Pl(A)$ for all $A \subseteq \Omega$
- The set $\mathcal{P}(m)$ of probability measures compatible with m is called the credal set of m
$$\mathcal{P}(Bel) = \{P : \forall A \subseteq \Omega, Bel(A) \leq P(A))\}$$

Construction of $\mathcal{P}(Bel)$

- An arbitrary element of $\mathcal{P}(Bel)$ can be obtained by distributing each mass $m(A)$ among the elements of A.
- More precisely, let $\alpha(\omega, A)$ be the fraction of $m(A)$ allocated to the element ω. We have
$$\sum_{\omega \in A} \alpha(\omega, A) = m(A).$$
- By summing up the numbers $\alpha(\omega, A)$ for each ω, we get a probability mass function on Ω,
$$p_\alpha(\omega) = \sum_{A \ni \omega} \alpha(\omega, A).$$
- It can be verified that
$$P_\alpha(A) = \sum_{\omega \in A} p_\alpha(\omega) \geq Bel(A),$$
for all $A \subseteq \Omega$.

Belief functions are coherent lower probabilities

- It can be shown (Dempster, 1967) that any element of the credal set $\mathcal{P}(Bel)$ can be obtained in that way.
- Furthermore, the bounds in the inequalities $Bel(A) \leq P(A)$ and $P(A) \leq Pl(A)$ are attained. We thus have, for all $A \subseteq \Omega$,

$$Bel(A) = \min_{P \in \mathcal{P}(Bel)} P(A)$$

$$Pl(A) = \max_{P \in \mathcal{P}(Bel)} P(A)$$

- We say that *Bel* is a coherent lower probability.
- Not all lower envelopes of sets of probability measures are belief functions!

A counterexample

- Suppose a fair coin is tossed twice, in such a way that the outcome of the second toss may depend on the outcome of the first toss.
- The outcome of the experiment can be denoted by $\Omega = \{(H,H),(H,T),(T,H),(T,T)\}$.
- Let $H_1 = \{(H,H),(H,T)\}$ and $H_2 = \{(H,H),(T,H)\}$ the events that we get Heads in the first and second toss, respectively.
- Let $\mathcal{P}$ be the set of probability measures on Ω which assign $P(H_1) = P(H_2) = 1/2$ and have an arbitrary degree of dependence between tosses.
- Let P_* be the lower envelope of $\mathcal{P}$.

A counterexample – continued

- It is clear that $P_*(H_1) = 1/2$, $P_*(H_2) = 1/2$ and $P_*(H_1 \cap H_2) = 0$ (as the occurrence Heads in the first toss may never lead to getting Heads in the second toss).
- Now, in the case of complete positive dependence, $P(H_1 \cup H_2) = P(H_1) = 1/2$, hence $P_*(H_1 \cup H_2) \leq 1/2$.
- We thus have

$$P_*(H_1 \cup H_2) < P_*(H_1) + P_*(H_2) - P_*(H_1 \cap H_2),$$

  which violates the complete monotonicity condition for $k = 2$.

Two different theories

- Mathematically, the notion of coherent lower probability is thus more general than that of belief function.
- However, the definition of the credal set associated with a belief function is purely formal, as these probabilities have no particular interpretation in our framework.
- The theory of belief functions is not a theory of imprecise probabilities.

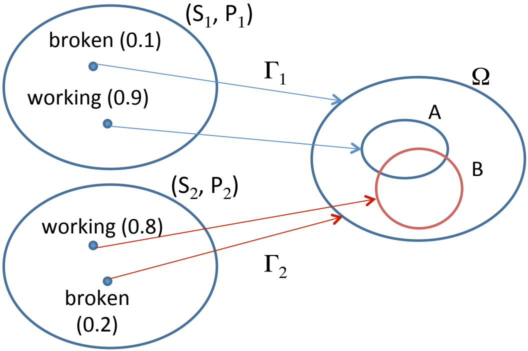
Outline

Outline

Broken sensor example continued

- The first item of evidence gave us: $m_1(A) = 0.9$, $m_1(\Omega) = 0.1$.
- Another sensor returns another set of values B, and it is in working condition with probability 0.8.
- This second piece if evidence can be represented by the mass function: $m_2(B) = 0.8$, $m_2(\Omega) = 0.2$
- How to combine these two pieces of evidence?

Analysis



- If interpretations $s_1 \in S_1$ and $s_2 \in S_2$ both hold, then $X \in \Gamma_1(s_1) \cap \Gamma_2(s_2)$
- If the two pieces of evidence are independent, then the probability that s_1 and s_2 both hold is $P_1(\{s_1\})P_2(\{s_2\})$

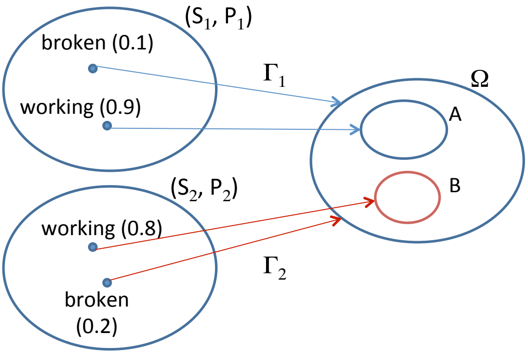
Computation

	S_2 working (0.8)	S_2 broken (0.2)
S_1 working (0.9)	$A \cap B$, 0.72	A, 0.18
S_1 broken (0.1)	B, 0.08	Ω, 0.02

We then get the following combined mass function,

$$\begin{aligned} m(A \cap B) &= 0.72 \\ m(A) &= 0.18 \\ m(B) &= 0.08 \\ m(\Omega) &= 0.02 \end{aligned}$$

Case of conflicting pieces of evidence



- If $\Gamma_1(s_1) \cap \Gamma_2(s_2) = \emptyset$, we know that s_1 and s_2 cannot hold simultaneously
- The joint probability distribution on $S_1 \times S_2$ must be conditioned to eliminate such pairs

Computation

	S_2 working (0.8)	S_2 broken (0.2)
S_1 working (0.9)	$\emptyset$, 0.72	A, 0.18
S_1 broken (0.1)	B, 0.08	Ω, 0.02

We then get the following combined mass function,

$$
\begin{aligned}
m(\emptyset) &= 0 \\
m(A) &= 0.18/0.28 \approx 0.64 \\
m(B) &= 0.08/0.28 \approx 0.29 \\
m(\Omega) &= 0.02/0.28 \approx 0.07
\end{aligned}
$$

Dempster's rule

- Let m_1 and m_2 be two mass functions and

$$\kappa = \sum_{B \cap C = \emptyset} m_1(B) m_2(C)$$

their degree of conflict

- If $\kappa < 1$, then m_1 and m_2 can be combined as

$$(m_1 \oplus m_2)(A) = \frac{1}{1-\kappa} \sum_{B \cap C = A} m_1(B) m_2(C), \quad \forall A \neq \emptyset$$

and $(m_1 \oplus m_2)(\emptyset) = 0$

Another example

A	$\emptyset$	$\{a\}$	$\{b\}$	$\{a,b\}$	$\{c\}$	$\{a,c\}$	$\{b,c\}$	$\{a,b,c\}$
$m_1(A)$	0	0	0.5	0.2	0	0.3	0	0
$m_2(A)$	0	0.1	0	0.4	0.5	0	0	0

		m_2		
		$\{a\}, 0.1$	$\{a,b\}, 0.4$	$\{c\}, 0.5$
	$\{b\}, 0.5$	$\emptyset, 0.05$	$\{b\}, 0.2$	$\emptyset, 0.25$
m_1	$\{a,b\}, 0.2$	$\{a\}, 0.02$	$\{a,b\}, 0.08$	$\emptyset, 0.1$
	$\{a,c\}, 0.3$	$\{a\}, 0.03$	$\{a\}, 0.12$	$\{c\}, 0.15$

The degree of conflict is $\kappa = 0.05 + 0.25 + 0.1 = 0.4$. The combined mass function is

$$
\begin{aligned}
(m_1 \oplus m_2)(\{a\}) &= (0.02 + 0.03 + 0.12)/0.6 = 0.17/0.6 \\
(m_1 \oplus m_2)(\{b\}) &= 0.2/0.6 \\
(m_1 \oplus m_2)(\{a,b\}) &= 0.08/0.6 \\
(m_1 \oplus m_2)(\{c\}) &= 0.15/0.6.
\end{aligned}
$$

Dempster's rule

Properties

- Commutativity, associativity. Neutral element: $m_?$
- Generalization of intersection: if m_A and m_B are logical mass functions and $A \cap B \neq \emptyset$, then
$$m_A \oplus m_B = m_{A\cap B}$$
- If either m_1 or m_2 is Bayesian, then so is $m_1 \oplus m_2$ (as the intersection of a singleton with another subset is either a singleton, or the empty set).

Dempster's conditioning

- Conditioning is a special case, where a mass function m is combined with a logical mass function m_A. Notation:

$$m \oplus m_A = m(\cdot|A)$$

- It can be shown that

$$Pl(B|A) = \frac{Pl(A \cap B)}{Pl(A)}.$$

- Generalization of Bayes' conditioning: if m is a Bayesian mass function and m_A is a logical mass function, then $m \oplus m_A$ is a Bayesian mass function corresponding to the conditioning of m by A

Commonality function

- Commonality function: let $Q : 2^\Omega \to [0,1]$ be defined as

$$Q(A) = \sum_{B \supseteq A} m(B), \quad \forall A \subseteq \Omega$$

- Conversely,

$$m(A) = \sum_{B \supseteq A} (-1)^{|B \setminus A|} Q(B)$$

- Q is another equivalent representation of a belief function.

Commonality function and Dempster's rule

- Let Q_1 and Q_2 be the commonality functions associated to m_1 and m_2.
- Let $Q_1 \oplus Q_2$ be the commonality function associated to $m_1 \oplus m_2$.
- We have

$$(Q_1 \oplus Q_2)(A) = \frac{1}{1-\kappa} Q_1(A) \cdot Q_2(A), \quad \forall A \subseteq \Omega, A \neq \emptyset$$

$$(Q_1 \oplus Q_2)(\emptyset) = 1$$

- In particular, $pl(\omega) = Q(\{\omega\})$. Consequently,

$$pl_1 \oplus pl_2 \propto (1-\kappa)^{-1} pl_1 pl_2.$$

Outline

Disjunctive rule

Definition and justification

- Let (S_1, P_1, Γ_1) and (S_2, P_2, Γ_2) be sources associated to two pieces of evidence
- If interpretation $s_k \in S_k$ holds and piece of evidence k is reliable, then we can conclude that $X \in \Gamma_k(s_k)$
- If interpretation $s \in S_1$ and $s_2 \in S_2$ both hold and we assume that at least one of the two pieces of evidence is reliable, then we can conclude that $X \in \Gamma_1(s_1) \cup \Gamma_2(s_2)$
- This leads to the TBM disjunctive rule:

$$(m_1 \cup m_2)(A) = \sum_{B \cup C = A} m_1(B) m_2(C), \quad \forall A \subseteq \Omega$$

Disjunctive rule
Example

A	$\emptyset$	$\{a\}$	$\{b\}$	$\{a,b\}$	$\{c\}$	$\{a,c\}$	$\{b,c\}$	$\{a,b,c\}$
$m_1(A)$	0	0	0.5	0.2	0	0.3	0	0
$m_2(A)$	0	0.1	0	0.4	0.5	0	0	0

		m_2		
		$\{a\}, 0.1$	$\{a,b\}, 0.4$	$\{c\}, 0.5$
	$\{b\}, 0.5$	$\{a,b\}, 0.05$	$\{a,b\}, 0.2$	$\{b,c\}, 0.25$
m_1	$\{a,b\}, 0.2$	$\{a,b\}, 0.02$	$\{a,b\}, 0.08$	$\{a,b,c\}, 0.1$
	$\{a,c\}, 0.3$	$\{a,c\}, 0.03$	$\{a,b,c\}, 0.12$	$\{a,c\}, 0.15$

The resulting mass function is

$$\begin{aligned}
(m_1 \cup m_2)(\{a,b\}) &= 0.05 + 0.2 + 0.02 + 0.08 = 0.35 \\
(m_1 \cup m_2)(\{b,c\}) &= 0.25 \\
(m_1 \cup m_2)(\{a,c\}) &= 0.03 + 0.15 = 0.18 \\
(m_1 \cup m_2)(\Omega) &= 0.1 + 0.12 = 0.22.
\end{aligned}$$

Disjunctive rule

Properties

- Commutativity, associativity.
- No neutral element.
- $m_?$ is an absorbing element.
- Expression using belief functions:

$$Bel_1 \cup Bel_2 = Bel_1 \cdot Bel_2$$

Outline

Definition

- In general, the disjunctive rule may be preferred in case of heavy conflict between the different pieces of evidence.
- An alternative rule, which is somehow intermediate between the disjunctive and conjunctive rules, has been proposed by Dubois and Prade (1988). It is defined as follows:

$$(m_1 \uplus m_2)(A) = \sum_{B \cap C = A} m_1(B) m_2(C) + \sum_{\{B \cap C = \emptyset, B \cup C = A\}} m_1(B) m_2(C),$$

for all $A \subseteq \Omega$, $A \neq \emptyset$, and $(m_1 \uplus m_2)(\emptyset) = 0$.

Example

A	$\emptyset$	$\{a\}$	$\{b\}$	$\{a,b\}$	$\{c\}$	$\{a,c\}$	$\{b,c\}$	$\{a,b,c\}$
$m_1(A)$	0	0	0.5	0.2	0	0.3	0	0
$m_2(A)$	0	0.1	0	0.4	0.5	0	0	0

		m_2		
		$\{a\}, 0.1$	$\{a,b\}, 0.4$	$\{c\}, 0.5$
m_1	$\{b\}, 0.5$	$\{a,b\}, 0.05$	$\{b\}, 0.2$	$\{b,c\}, 0.25$
	$\{a,b\}, 0.2$	$\{a\}, 0.02$	$\{a,b\}, 0.08$	$\{a,b,c\}, 0.1$
	$\{a,c\}, 0.3$	$\{a\}, 0.03$	$\{a\}, 0.12$	$\{c\}, 0.15$

$$
\begin{aligned}
(m_1 \uplus m_2)(\{a,b\}) &= 0.05 + 0.08 = 0.13\\
(m_1 \uplus m_2)(\{b\}) &= 0.2\\
(m_1 \uplus m_2)(\{b,c\}) &= 0.25\\
(m_1 \uplus m_2)(\{a\}) &= 0.02 + 0.03 + 0.12 = 0.17\\
(m_1 \uplus m_2)(\{c\}) &= 0.15\\
(m_1 \uplus m_2)(\Omega) &= 0.1.
\end{aligned}
$$

Properties

- The DP rule boils down to the conjunctive and disjunctive rules when, respectively, the degree of conflict is equal to zero and one.
- In other cases, it has some intermediate behavior.
- It is not associative. If several pieces of evidence are available, they should be combined at once using an obvious *n*-ary extension of the above formula.

Outline

Introductory example

- Consider an urn with white (ξ_1), red (ξ_2) and black (ξ_3) balls in proportions p_1, p_2 and p_3.
- Let $X \in \mathcal{X} = \{\xi_1, \xi_2, \xi_3\}$ be the color of a ball that will be drawn from the urn: belief on X?
- Two cases:
  1. We know the proportions p_k: then $bel^{\mathcal{X}}(\{\xi_k\}) = p_k$ (Hacking's Principle);
  2. We have observed the result of n drawings from the urn with replacement, e.g. 5 white balls, 3 red balls and 2 black balls.
- How to build a belief function from data in the 2nd case ?
- A solution was described in

  T. Denoeux. Constructing Belief Functions from Sample Data Using Multinomial Confidence Regions. *International Journal of Approximate Reasoning* 42(3):228-252, 2006.

Outline

Formalization

- Discrete variable $X \in \mathcal{X} = \{\xi_1, \ldots, \xi_K\}$ defined as the result of a random experiment.
- X is characterized by an unknown frequency (probability) distribution $\mathbb{P}_X$.
- $\mathbb{P}_X(A)$: limit frequency of the event $A \subseteq \mathcal{X}$ in an infinite sequence of trials.
- We have observed a realization $\mathbf{x}_n$ of an iid random sample $X_n = (X_1, \ldots, X_n)$ with parent distribution $\mathbb{P}_X$.
- Problem: build a belief function $bel^{\mathcal{X}}[\mathbf{x}_n]$ with well-defined properties with respect to the unknown frequency distribution $\mathbb{P}_X \rightarrow$ predictive belief function.

Approach

- Let $bel^{\mathcal{X}}[\mathbf{x}_n]$ be the BF on X after observing a realization $\mathbf{x}_n$ of random sample $\mathbf{X}_n = (X_1, \ldots, X_n)$.
- Which properties should $bel^{\mathcal{X}}[\mathbf{x}_n]$ verify with respect to $\mathbb{P}_X$?
- Hacking's principle (1965): if $\mathbb{P}_X$ is know, then $bel^{\mathcal{X}}[\mathbf{x}_n] = \mathbb{P}_X$.
- Weak version:

$$\forall A \subseteq \mathcal{X}, \quad bel^{\mathcal{X}}[\mathbf{X}_n](A) \xrightarrow{P} \mathbb{P}_X(A), \text{ as } n \to \infty.$$

(Requirement R_1)

Approach (continued)

- Least Commitment Principle: for fixed n, $bel^{\mathcal{X}}[\mathbf{x}_n]$ should be less informative that $\mathbb{P}_X$:

$$bel^{\mathcal{X}}[\mathbf{x}_n](A) \leq \mathbb{P}_X(A), \quad \forall A \subseteq \mathcal{X}.$$

- This condition is too restrictive (it leads to the vacuous BF).
- Weaker condition ((Requirement R_2):

$$\mathbb{P}(bel^{\mathcal{X}}[\mathbf{X}_n] \leq \mathbb{P}_X) \geq 1 - \alpha,$$

for some $\alpha \in (0, 1)$.

Meaning of Requirement R_2

$$\mathbf{x}_n = (x_1, \ldots, x_n) \rightarrow bel^{\mathcal{X}}[\mathbf{x}_n]$$
$$\mathbf{x}'_n = (x'_1, \ldots, x'_n) \rightarrow bel^{\mathcal{X}}[\mathbf{x}'_n]$$
$$\mathbf{x}''_n = (x''_1, \ldots, x''_n) \rightarrow bel^{\mathcal{X}}[\mathbf{x}''_n]$$
$$\vdots$$

- As the number of realizations of the random sample tends to ∞, the proportion of belief functions less committed than $\mathbb{P}_X$ should tend to $1 - \alpha$.
- To achieve this property: use of a multinomial confidence region.

Outline

1. Representation of evidence
   - Mass functions
   - Belief and plausibility functions
2. Relations with alternative theories
   - Possibility theory
   - Imprecise probabilities
3. Combination of evidence
   - Dempster's rule
   - Disjunctive rule
   - Dubois-Prade rule
4. Predictive belief functions
   - Formalization
   - Method
   - Ordered data

Multinomial Confidence Region

- Let $N_k = \#\{i | X_i = \xi_k\}$. Vector $\mathbf{N} = (N_1, \ldots, N_K)$ has a multinomial distribution $\mathcal{M}(n, p_1, \ldots, p_K)$, with $p_k = \mathbb{P}_X(\{\xi_k\})$.
- Let $\mathcal{S}(\mathbf{N}) \subseteq [0,1]^K$ a random region of $[0,1]^K$. It is a confidence region for $\mathbf{p}$ at level $1-\alpha$ if
$$\mathbb{P}(\mathcal{S}(\mathbf{N}) \ni \mathbf{p}) \geq 1-\alpha.$$
- $\mathcal{S}(\mathbf{N})$ is an asymptotic confidence region if the above inequality holds in the limit as $n \to \infty$.
- Simultaneous confidence intervals: $\mathcal{S}(\mathbf{N}) = [P_1^-, P_1^+] \times \ldots \times [P_K^-, P_K^+]$

Multinomial Conf. Region (cont.)

- Goodman's simultaneous confidence intervals:

$$P_k^- = \frac{b + 2N_k - \sqrt{\Delta_k}}{2(n+b)},$$

$$P_k^+ = \frac{b + 2N_k + \sqrt{\Delta_k}}{2(n+b)},$$

with $b = \chi^2_{1;1-\alpha/K}$ and $\Delta_k = b\left(b + \frac{4N_k(n-N_k)}{n}\right)$.

Example

- 220 psychiatric patients categorized as either neurotic, depressed, schizophrenic or having a personality disorder.
- Observed counts: $\mathbf{n} = (91, 49, 37, 43)$.
- Goodman' confidence intervals at confidence level $1 - \alpha = 0.95$:

Diagnosis	N_k/n	P_k^-	P_k^+
Neurotic	0.41	0.33	0.50
Depressed	0.22	0.16	0.30
Schizophrenic	0.17	0.11	0.24
Personality disorder	0.20	0.14	0.27

From Conf. Regions to Lower Probabilties

- To each $\mathbf{p} = (p_1, \ldots, p_K)$ corresponds a probability measure $\mathbb{P}_X$.
- Consequently, $\mathcal{S}(\mathbf{N})$ may be seen as defining a family of probability measures, uniquely defined by the following lower probability measure:

$$P^-(A) = \max\left(\sum_{\xi_k \in A} P_k^-, 1 - \sum_{\xi_k \notin A} P_k^+\right)$$

- P^- satisfies requirements R_1 and R_2:
  - $P^-(A) \xrightarrow{P} \mathbb{P}_X(A)$ as $n \to \infty$, for all $A \subseteq \mathcal{X}$,
  - $\mathbb{P}(P^- \leq \mathbb{P}_X) \geq 1 - \alpha$.

From Lower Probabilities to Belief Functions

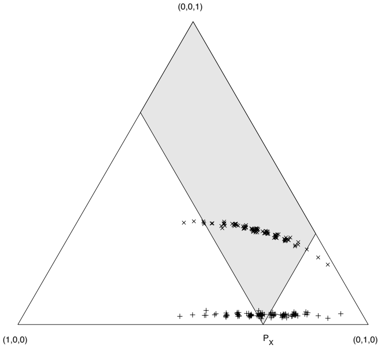
- Is P^- a belief function ?
- If $K = 2$ or $K = 3$, P^- is a belief function.
- Case $K = 2$:

$$m^{\mathcal{X}}(\{\xi_1\}) = P_1^-, \quad m^{\mathcal{X}}(\{\xi_2\}) = P_2^-$$
$$m^{\mathcal{X}}(\mathcal{X}) = 1 - P_1^- - P_2^-.$$

- If $K > 3$, P^- is not a belief function in general. We can find the most committed belief function satisfying $bel^{\mathcal{X}} \leq P^-$ by solving a linear optimization problem.
- The solution satisfies requirements R_1 and R_2: it is a predictive belief function (at confidence level $1 - \alpha$).

Example 1

- $K = 2$, $p_1 = \mathbb{P}_X(\{\xi_1\}) = 0.3$. 100 realizations of a random sample of size $n = 30 \rightarrow 100$ predictive belief functions at level $1 - \alpha = 0.95$.



Example 2: Psychiatric Data

A	$P^{-}(A)$	$bel^{\mathcal{X}*}(A)$	$m^{\mathcal{X}*}(A)$
$\{\xi_1\}$	0.33	0.33	0.33
$\{\xi_2\}$	0.16	0.14	0.14
$\{\xi_1, \xi_2\}$	0.50	0.50	0.021
$\{\xi_3\}$	0.11	0.097	0.097
$\{\xi_1, \xi_3\}$	0.45	0.45	0.020
$\{\xi_2, \xi_3\}$	0.28	0.28	0.036
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\{\xi_1, \xi_3, \xi_4\}$	0.70	0.66	0.038
$\{\xi_2, \xi_3, \xi_4\}$	0.50	0.48	0.019
$\mathcal{X}$	1	1	0

Outline

Case of ordered data

- Assume $\mathcal{X}$ is ordered: $\xi_1 < \ldots < \xi_K$.
- The focal sets of $bel^{\mathcal{X}}[\mathbf{x}_n]$ can be constrained to be intervals $A_{k,r} = \{\xi_k, \ldots, \xi_r\}$.
- Under this additional constraint, an analytical solution to the previous optimization problem can be found:

$$m^{\mathcal{X}*}(A_{k,k}) = P_k^-,$$

$$m^{\mathcal{X}*}(A_{k,k+1}) = P^-(A_{k,k+1}) - P^-(A_{k+1,k+1}) - P^-(A_{k,k}),$$

$$m^{\mathcal{X}*}(A_{k,r}) = P^-(A_{k,r}) - P^-(A_{k+1,r}) - P^-(A_{k,r-1}) + P^-(A_{k+1,r-1})$$

for $r > k+1$, and $m^{\mathcal{X}*}(B) = 0$, for all $B \notin \mathcal{I}$.

Example: rain data

- January precipitation in Arizona (in inches), recorded during the period 1895-2004.

class ξ_k	n_k	n_k/n	p_k^-	p_k^+
< 0.75	48	0.44	0.32	0.56
$[0.75, 1.25)$	17	0.15	0.085	0.27
$[1.25, 1.75)$	19	0.17	0.098	0.29
$[1.75, 2.25)$	11	0.10	0.047	0.20
$[2.25, 2.75)$	6	0.055	0.020	0.14
≥ 2.75	9	0.082	0.035	0.18

- Degree of belief that the precipitation in Arizona next January will exceed, say, 2.25 inches?

Rain data: Result

$m(A_{k,r})$	1	2	3	4	5	6
1	0.32	0	0	0.13	0.11	0
2	-	0.085	0	0	0.012	0.14
3	-	-	0.098	0	0	0
4	-	-	-	0.047	0	0
5	-	-	-	-	0.020	0
6	-	-	-	-	-	0.035

- We get $bel^{\mathcal{X}}(X \geq 2.25) = bel^{\mathcal{X}*}(\{\xi_5, \xi_6\}) = 0.055$ and $pl(X \geq 2.25) = 0.317$.
- In 95 % of cases, the interval $[bel^{\mathcal{X}}(A), pl^{\mathcal{X}}(A)]$ computed using this method contains $\mathbb{P}_X(A)$.

Conclusions

- A "frequentist" approach, based on multinomial confidence regions, for building a belief function quantifying the uncertainty about a discrete random variable X with unknown probability distribution, based on observed data.
- Two "reasonable" properties of the solution with respect to the true frequency distribution $\mathbb{P}_X$:
  - it is less committed than $\mathbb{P}_X$ with some user-defined probability, and
  - it converges towards $\mathbb{P}_X$ in probability as the size of the sample tends to infinity.