Methods for building belief functions

Thierry Denœux[1]

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

Spring School BFTA 2013
Carthage, Tunisia, May 20, 2013

Building belief functions

- The basic theory tells us how to reason and compute with belief functions, but it does not tell us where belief functions come from.
- We need methods for modeling evidence from
  - expert opinions or
  - statistical information.
- In this lecture, we will review some general methods and give some practical examples.

Outline

Least Commitment Principle

Definition

Definition (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.

- General approach:
  1. Express partial information (provided, e.g., by an expert) as a set of constraints on an unknown mass function;
  2. Find the least-committed mass function (according to some informational ordering), compatible with the constraints.
- Examples of partial information:
  - contour function;
  - conditional mass function.

Outline

LC mass function with given contour function

Problem statement and solution

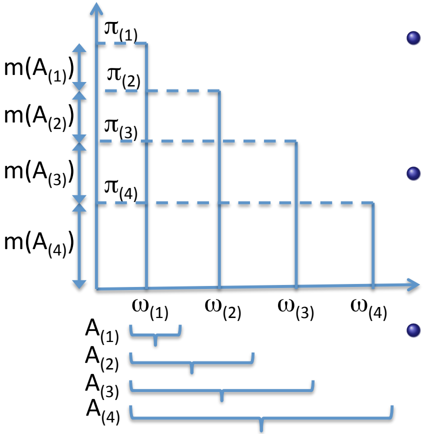
- Assume we ask an expert for the plausibility $\pi(\omega)$ of each $\omega \in \Omega$.
- We get a function $\pi : \Omega \rightarrow [0, 1]$. We assume that $\max_{\omega\in\Omega} \pi(\omega) = 1$.
- Let $\mathcal{M}(\pi)$ be the set of mass functions m such that $pl = \pi$.
- What is the least committed mass function in $\mathcal{M}(\pi)$?
- Taking $\sqsubseteq_q$ as the informational ordering, the least committed element in $\mathcal{M}(\pi)$ is the consonant mass function whose contour function is π.
- Its plausibility and commonality functions are defined as

$$Pl(A) = \max_{\omega\in A} \pi(\omega), \quad Q(A) = \min_{\omega\in A} \pi(\omega),$$

for all $A \subseteq \Omega$, $A \neq \emptyset$.

LC mass function with given contour function

Recovering the mass function



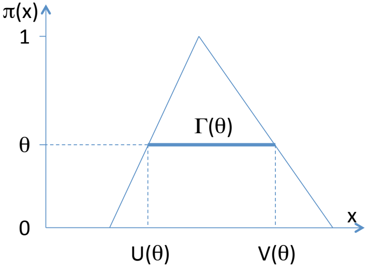
- Let $1=\pi_{(1)} \geq \pi_{(2)} \geq \ldots \geq \pi_{(K)}$ be the ordered values of π; $\omega_{(1)},\ldots,\omega_{(K)}$ the elements of Ω in the corresponding order, and $A_{(k)}=\{\omega_{(1)},\ldots,\omega_{(k)}\}$.
- We have

$$m(A_{(k)}) = \pi_{(k)} - \pi_{(k+1)},$$

for $k=1,\ldots,K-1$ and $m(\Omega)=\pi_{(K)}$.

- Random set: $\Theta=[0,1]$, P= Lebesgue measure, $\Gamma(\theta)=A_{(k)}$ if $\theta \in [\pi_{(k+1)},\pi_{(k)}]$.

LC mass function with given contour function
Continuous extension

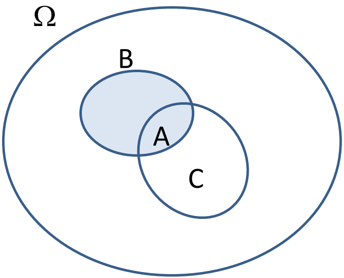


- Let $\pi : \mathbb{R} \rightarrow [0,1]$ be an upper semi-continuous function, $\Theta = [0,1]$, P the Lebesgue measure on $[0,1]$, and $\Gamma(\theta) = \{x \in \mathbb{R} | \pi(x) \geq \theta\}$.
- (Ω, P, Γ) defines a consonant random interval with contour function π and plausibility function

$$Pl(A) = \sup_{x \in A} \pi(x),$$

  for all $A \in \mathcal{B}(\mathbb{R})$
- The corresponding belief function is the q-least committed one among those for which $pl = \pi$.

Deconditioning

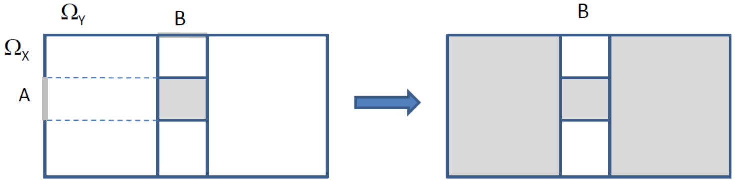


- Let m_0 be a mass function on Ω expressing our beliefs about X in a context where we know that $X \in B$.
- We want to build a mass function m verifying the constraint $m(\cdot|B) = m_0$.
- Any m built from m_0 by transferring each mass $m_0(A)$ to $A \cup C$ for some $C \subseteq \overline{B}$ satisfies the constraint.
- s-least committed solution: transfer $m_0(A)$ to the largest such set, which is $A \cup \overline{B}$:

$$m(D) = \begin{cases} m_0(A) & \text{if } D = A \cup \overline{B} \text{ for some } A \subseteq B, \\ 0 & \text{otherwise} \end{cases}$$

Deconditioning

Conditional embedding



- More complex situation: two frames Ω_X and Ω_Y.
- Let m_0^X be a mass function on Ω_X expressing our beliefs about X in a context where we know that $Y \in B$ for some $B \subseteq \Omega_Y$.
- We want to find m^{XY} such that $\left(m^{XY} \oplus (m_B^Y)^{\uparrow XY}\right)^{\downarrow X} = m_0^X$.
- s-least committed solution: transfer $m_0^X(A)$ to $(A \times \Omega_Y) \cup (\Omega_X \times \overline{B})$.
- Notation $m^{XY} = (m_0^X)^{\Uparrow XY}$ (conditional embedding).

Outline

Uncertainty measures

Motivation

- In some cases, the least committed mass function compatible with some constraints does not exist, or cannot be found, for any informational ordering.
- An alternative approach is then to maximize a measure of uncertainty, i.e., find the most uncertain mass function satisfying some constraints.
- Many uncertainty measures have been proposed, some of which generalize the Shannon entropy. They can be classified in 3 categories:
    1. Measures of imprecision;
    2. Measures of conflict;
    3. Measure of total uncertainty.

Uncertainty measures

Main measures

- Measures of imprecision:

$$I(m) = \sum_{\emptyset \neq A \subseteq \Omega} m(A) f(|A|)$$

with $f = Id$ (expected cardinality) or $f = \log_2$ (non-specificy).

- Measures of conflict:

$$C(m) = -\sum_{\emptyset \neq A \subseteq \Omega} m(A) \log_2 F(A)$$

with $F = Bel$ (confusion), Pl (dissonance) or P_m (discord).

- Measures of total uncertainty:

$$AU(m) = \max_{p \in \mathcal{P}(m)} \left(-\sum_{\omega \in \Omega} p(\omega) \log_2 p(\omega) \right)$$

$$EP(m) = -\sum_{\omega \in \Omega} p_m(\omega) \log_2 p_m(\omega)$$

Uncertainty measures

Example

- Assume we know $Pl(A_i) = \alpha_i$ for some $A_i \subseteq \Omega$, $i = 1, \ldots, n$.
- A maximally imprecise mass function can be defined as any solution of the following linear programming problem:

$$\max_m \sum_{\emptyset \neq A \subseteq \Omega} m(A)|A|$$

under the constraints

$$\sum_{B \cap A_i \neq \emptyset} m(B) = \alpha_i, \quad i = 1, \ldots, n$$

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

$$m(B) \geq 0, \quad \forall B \subseteq \Omega, B \neq \emptyset$$

$$m(\emptyset) = 0.$$

Outline

Problem statement

- Two variables $X \in \Omega$ et $\theta \in \Theta = \{\theta_1, \ldots, \theta_K\}$.
- Typically:
  - X is observed (sensor measurement),
  - θ is not observed (class, unknown parameter).
- Partial knowledge of X given each $\theta = \theta_k$:

$$m^\Omega(\cdot|\theta_k), \quad k = 1, \ldots, K.$$

- Prior knowledge about θ: $m_0^\Theta(\Theta)$ (may be vacuous).
- We observe $X \in A$.
- Belief function on Θ?

Solution

- Solution:

$$m^{\Theta}(\cdot|A) = \left(\bigoplus_{k=1}^{K} m^{\Omega}(\cdot|\theta_k)^{\Uparrow \Omega\times\Theta} \oplus m_A^{\Omega\uparrow\Omega\times\Theta} \oplus m_0^{\Theta\uparrow\Omega\times\Theta} \right)^{\downarrow\Theta}$$

- Expression:

$$m^{\Theta}(\cdot|A) = \bigoplus_{k=1}^{K} \overline{\{\theta_k\}}^{pl^{\Omega}(A|\theta_k)} \oplus m_0^{\Theta},$$

where $\overline{\{\theta_k\}}^{pl^{\Omega}(A|\theta_k)}$ is the simple mass function that assigns the mass $1 - pl^{\Omega}(A|\theta_k)$ to $\overline{\{\theta_k\}}$ and $pl^{\Omega}(A|\theta_k)$ to Θ.

Properties

- Property 1: Bayes' theorem is recovered as a special case when the conditional mass functions $m^{\Omega}(\cdot|\theta_k)$ and m_0^{Θ} are Bayesian.
- Property 2: If X and Y are cognitively independent conditionally on θ, i.e.:

$$pl^{XY}(A \times B|\theta_k) = pl^{X}(A|\theta_k) \cdot pl^{Y}(B|\theta_k),$$

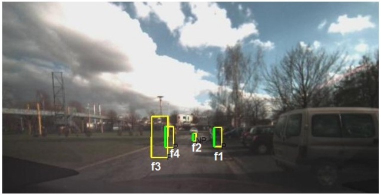
for all $A \subseteq \Omega_X$, $B \subseteq \Omega_Y$ and $\theta_k \in \Theta$, then

$$m^{\Theta}(\cdot|X \in A, Y \in B) = m^{\Theta}(\cdot|X \in A) \oplus m^{\Theta}(\cdot|Y \in B).$$

Outline

Problem description

- Let $E = \{e_1, \ldots, e_n\}$ and $F = \{f_1, \ldots, f_p\}$ be two sets of objects perceived by two sensors.
- Problem: find a matching between the two sets, in such a way that each object in one set is matched with at most one object in the other set.



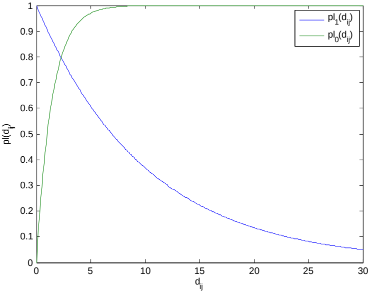
Formalization

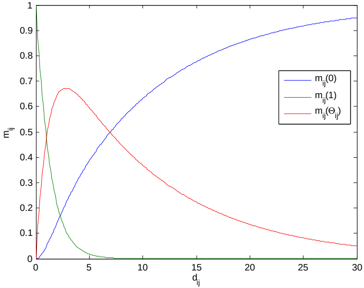
- Let R_{ij} be a binary variable equal to 1 if e_i and f_j are the same object, 0 otherwise.
- We know the distance d_{ij} between the positions of objects e_i and f_j.
- How to compute a mass function on $\Theta_{ij} = \{0, 1\}$ representing our knowledge of R_{ij}?
- We can use the GBT if we can assess the plausibility of observing d_{ij} given $R_{ij} = 1$ and given $R_{ij} = 0$.

Using the GBT

- Let $pl_1(d_{ij})$ and $pl_0(d_{ij})$ be the plausibilities that the distance between e_i and f_j is d_{ij} if $R_{ij} = 1$ and $R_{ij} = 0$, respectively.
- From the GBT, the mass function m_{ij} on $\Theta_{ij} = \{0, 1\}$ given d_{ij} is:

$$m_{ij} = \{0\}^{pl_1(d_{ij})} \oplus \{1\}^{pl_0(d_{ij})}.$$





Finding the most plausible matching

- Given the nm pairwise mass functions m_{ij}, how to match the two object sets?
- Approach:
  - Vacuously extend each m_{ij} in $\mathcal{R}$, the set of matching relations between sets E and F;
  - Combine the extended mass function using Dempster's rule;
  - Find the matching relation R with greatest plausibility.
- It can be shown that

$$pl(R) \propto \prod_{i,j} (1 - m_{ij}(0))^{R_{ij}} (1 - m_{ij}(1))^{1-R_{ij}},$$

- Maximizing $\log pl(R)$ is a linear assignment problem that can be solved in $o(\max(n, m)^3)$ time.

Outline

Discounting
Problem statement

- A source of information provides:
  - a value;
  - a set of values;
  - a probability distribution, etc..
- The information is:
  - not fully reliable or
  - not fully relevant.
- Examples:
  - Possibly faulty sensor;
  - Measurement performed in unfavorable experimental conditions;
  - Information is related to a situation or an object that only has some similarity with the situation or the object considered (case-based reasoning).

Discounting
Formalization

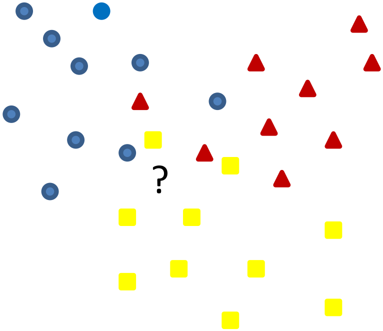
- A source S provides a mass function m_S^Ω.
- S may be reliable or not. Let $\mathcal{R} = \{R, NR\}$.
- Assumptions:
  - If S is reliable, we accept m_S^Ω as a representation of our beliefs:
$$m^\Omega(\cdot|R) = m_S^\Omega$$
  - If S is not reliable, we know nothing:
$$m^\Omega(\cdot|NR) = m_\Omega^\Omega$$
  - The source has a probability α of not being reliable:
$$m^\mathcal{R}(\{NR\}) = \alpha, \quad m^\mathcal{R}(\{R\}) = 1 - \alpha$$

Discounting
Solution

- Solution:
$$ {}^{\alpha}m^{\Omega} = \left(m^{\mathcal{R}\uparrow\Omega\times\mathcal{R}} \oplus m^{\Omega}(\cdot|R)^{\Uparrow\Omega\times\mathcal{R}}\right)^{\downarrow\Omega}. $$
- Simple expressions:
$$ {}^{\alpha}m^{\Omega} = (1-\alpha)m^{\Omega}_S + \alpha m^{\Omega}_{\Omega} $$
$$ {}^{\alpha}m^{\Omega}(A) = \begin{cases} (1-\alpha)m^{\Omega}_S(A) & \text{if } A \subset \Omega \\ (1-\alpha)m^{\Omega}_S(\Omega)+\alpha & \text{if } A = \Omega. \end{cases} $$
$$ {}^{\alpha}m^{\Omega} = m^{\Omega}_S \bigcirc\!\!\!\!\cup\, m^{\Omega}_0, \text{ with } m^{\Omega}_0(\Omega) = \alpha \text{ and } m^{\Omega}_0(\emptyset) = 1-\alpha. $$
- α is called the discount rate. It is the probability that the source is not reliable.
- ${}^{\alpha}m^{\Omega}$ is a s-less committed than (a generalization of) m^{Ω}_S:
$$ {}^{\alpha}m^{\Omega} \sqsupseteq_s m^{\Omega}_S. $$

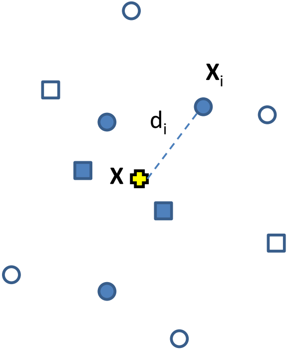
Outline

Classification



- A population is assumed to be partitioned in c groups or classes.
- Let $\Omega = \{\omega_1, \ldots, \omega_c\}$ denote the set of classes.
- Each instance is described by
  - A feature vector $\mathbf{x} \in \mathbb{R}^p$;
  - A class label $y \in \Omega$.
- Problem: given a learning set $\mathcal{L} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$, predict the class label of a new instance described by $\mathbf{x}$.

Evidential *k*-NN rule (1/2)



- Let $\mathcal{N}_k(\mathbf{x}) \subset \mathcal{L}$ denote the set of the k nearest neighbors of $\mathbf{x}$ in $\mathcal{L}$, based on some distance measure.
- Each $\mathbf{x}_i \in \mathcal{N}_k(\mathbf{x})$ can be considered as a piece of evidence regarding the class of $\mathbf{x}$.
- The strength of this evidence decreases with the distance d_i between $\mathbf{x}$ and $\mathbf{x}_i$.

Evidential *k*-NN rule (2/2)

- The evidence of $(\mathbf{x}_i, y_i)$, with $\mathbf{x}_i \in \mathcal{N}_k(\mathbf{x})$, tells us that $y = y_i$.
- Discounting this piece of evidence with a discount rate $\alpha(d_i)$, where $\alpha(\cdot)$ is an increasing function from $[0, +\infty)$ to $[0, 1]$, yields the following simple mass function:

$$m_i(\{y_i\}) = 1 - \alpha(d_i)$$

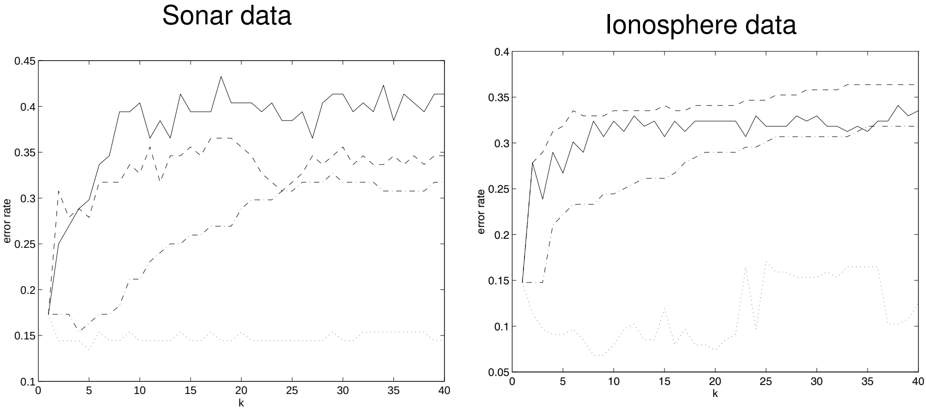
$$m_i(\Omega) = \alpha(d_i).$$

- The evidence of the *k* nearest neighbors of **x** is pooled using Dempster's rule of combination:

$$m = \bigoplus_{\mathbf{x}_i \in \mathcal{N}_k(\mathbf{x})} m_i.$$

- Function $\alpha(\cdot)$ can be fixed heuristically or selected among a family $\{\alpha_\theta | \theta \in \Theta\}$ using, e.g., cross-validation.

Performance comparison (UCI database)



Test error rates as a function of *k* for the voting (-), evidential (:), fuzzy (–) and distance-weighted (-.) *k*-NN rules.

Partially supervised data

- We now consider a learning set of the form

$$\mathcal{L} = \{(\mathbf{x}_i, m_i), i = 1, \ldots, n\}$$

  where
  - $\mathbf{x}_i$ is the attribute vector for instance i, and
  - m_i is a mass function representing uncertain expert knowledge about the class y_i of instance i.
- Special cases:
  - $m_i(\{\omega_k\}) = 1$ for all i: supervised learning;
  - $m_i(\Omega) = 1$ for all i: unsupervised learning;

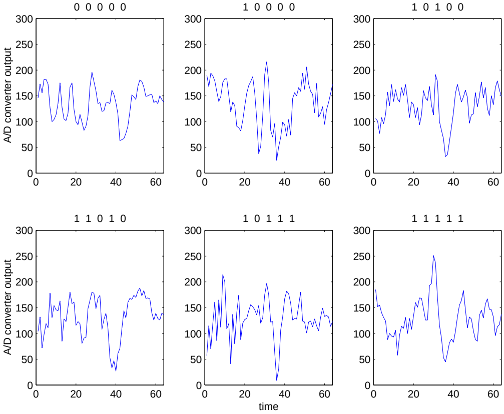
Evidential *k*-NN rule for partially supervised data

- Each instance $(\mathbf{x}_i, m_i)$, with $\mathbf{x}_i \in \mathcal{N}_k(\mathbf{x})$, is an item of evidence regarding y, which gives us a mass function m_i.
- The reliability of this piece of evidence decreases with the distance d_i between $\mathbf{x}$ and $\mathbf{x}_i$.
- Consequently, m_i is discounted with a discount rate $\alpha(d_i)$.
- The k discounted mass functions are combined using Dempster's rule:

$$m = \bigoplus_{\mathbf{x}_i \in \mathcal{N}_k(\mathbf{x})} {}^{\alpha(d_i)} m_i.$$

Example: EEG data

EEG signals encoded as 64-D patterns, 50 % positive (K-complexes), 50 % negative (delta waves), 5 experts.



Results on EEG data
(Denoeux and Zouhal, 2001)

- $c = 2$ classes, $p = 64$
- For each learning instance $\mathbf{x}_i$, the expert opinions were modeled as a mass function m_i.
- $n = 200$ learning patterns, 300 test patterns

k	k-NN	w k-NN	Ev. k-NN (crisp labels)	Ev. k-NN (uncert. labels)
9	0.30	0.30	0.31	0.27
11	0.29	0.30	0.29	0.26
13	0.31	0.30	0.31	0.26

Outline

1. Least Commitment Principle
   - Examples
   - Uncertainty measures
2. Generalized Bayes Theorem
   - Derivation and properties
   - Example: object association
3. Discounting
   - Problem statement and solution
   - Example: evidential *k* nearest neighbor rule
   - Contextual discounting
4. Fitting mass functions to data
   - General approach
   - Evidential *c*-means
5. Using the likelihood function
   - Principle
   - Handling low-quality data

Generalization: Contextual Discounting
Formalization

- A more general model allowing us to take into account richer meta-information about the source.
- Let $\Theta = \{\theta_1, \ldots, \theta_L\}$ be a partition of Ω, representing different contexts.
- Let $m^{\mathcal{R}}(\cdot|\theta_k)$ denote the mass function on $\mathcal{R}$ quantifying our belief in the reliability of source S, when we know that the actual value of X is in θ_k.
- We assume that:

$$m^{\mathcal{R}}(\{R\}|\theta_k) = 1 - \alpha_k, \quad m^{\mathcal{R}}(\{NR\}|\theta_k) = \alpha_k.$$

  for each $k \in \{1, \ldots, L\}$.
- Let $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_L)$.

Contextual Discounting

Example

- Let us consider a simplified aerial target recognition problem, in which we have three classes: airplane ($\omega_1 \equiv a$), helicopter ($\omega_2 \equiv h$) and rocket ($\omega_3 \equiv r$).
- Let $\Omega = \{a, h, r\}$.
- The sensor provides the following mass function: $m_S^\Omega(\{a\}) = 0.5$, $m_S^\Omega(\{r\}) = 0.5$.
- We assume that
  - The probability that the source is reliable when the target is an airplane is equal to $1 - \alpha_1 = 0.4$;
  - The probability that the source is reliable when the target is either a helicopter, or a rocket is equal to $1 - \alpha_2 = 0.9$.
- We have $\Theta = \{\theta_1, \theta_2\}$, with $\theta_1 = \{a\}$, $\theta_2 = \{h, r\}$, and $\boldsymbol{\alpha} = (0.6, 0.1)$.

Contextual Discounting

Solution

- Solution:

$$^{\alpha}m^{\Omega} = \left(\bigoplus_{k=1}^{L} m^{\mathcal{R}}(\cdot|\theta_k)^{\Uparrow\Omega\times\mathcal{R}} \oplus m^{\Omega}(\cdot|R)^{\Uparrow\Omega\times\mathcal{R}} \right)^{\downarrow\Omega}.$$

- Result:

$$^{\alpha}m^{\Omega} = m_S^{\Omega} \cup\!\!\!\!\bigcirc\, m_1^{\Omega} \cup\!\!\!\!\bigcirc \ldots \cup\!\!\!\!\bigcirc\, m_L^{\Omega}$$

with $m_k^{\Omega}(\theta_k) = \alpha_k$ and $m_k^{\Omega}(\emptyset) = 1 - \alpha_k$.

- Standard discounting is recovered as a special case when $\Theta = \{\Omega\}$.

Contextual Discounting

Example (continued)

- The discounted mass function can be obtained by combining disjunctively 3 mass functions:
  - $m_S^\Omega(\{a\}) = 0.5$, $m_S^\Omega(\{r\}) = 0.5$;
  - $m_1^\Omega(\{a\}) = 0.6$, $m_1^\Omega(\emptyset) = 0.4$;
  - $m_1^\Omega(\{h, r\}) = 0.1$, $m_1^\Omega(\emptyset) = 0.9$.
- Result:

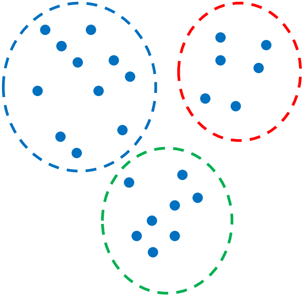
A	h	a	r	h, a	h, r	a, r	Ω
$m_S^\Omega(A)$	0	0.5	0.5	0	0	0	0
$^\alpha m^\Omega(A)$	0	0.45	0.18	0	0.02	0.27	0.08

Outline

Fitting mass functions to data

- In some cases, we have n objects described by data $D = \{d_1, \ldots, d_n\}$ and we want to find n mass functions $M = \{m_1, \ldots, m_n\}$ that fit the data in some way.
- The mass functions can then be found by minimizing a cost function $C(M, D)$ with respect to M.
- Example: evidential clustering.

Clustering



- n objects described by
  - Attribute vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ (attribute data) or
  - Dissimilarities (proximity data).
- Goal: find a meaningful structure in the data set, usually a partition into c crisp or fuzzy subsets.
- Belief functions may allow us to express richer information about the data structure.

Different partition concepts

- Hard partition: each object belongs to one and only one group. Group membership is expressed by binary variables u_{ik} such that $u_{ik} = 1$ if object i belongs to group k and $u_{ik} = 0$ otherwise.
- Fuzzy partition: each object has a degree of membership $u_{ik} \in [0, 1]$ to each group, with $\sum_{k=1}^{c} u_{ik} = 1$. The membership degrees $(u_{i1}, \ldots, u_{ic})$ define a probability distribution over the set Ω of groups.
- Credal partition: the group membership of each object is described by a mass function m_i over Ω.

Credal partition

Example

A	$m_1(A)$	$m_2(A)$	$m_3(A)$	$m_4(A)$	$m_5(A)$
$\emptyset$	0	0	0	0	0
$\{\omega_1\}$	0	0	0	0.2	0
$\{\omega_2\}$	0	1	0	0.4	0
$\{\omega_1, \omega_2\}$	0.7	0	0	0	0
$\{\omega_3\}$	0	0	0.2	0.4	0
$\{\omega_1, \omega_3\}$	0	0	0.5	0	0
$\{\omega_2, \omega_3\}$	0	0	0	0	0
Ω	0.3	0	0.3	0	1

Hard and fuzzy partitions are recovered as special cases when all mass functions are certain or Bayesian, respectively.

Algorithms

- EVCLUS (Denoeux and Masson, 2004):
  - Proximity (possibly non metric) data,
  - Multidimensional scaling approach.
- Evidential *c*-means (ECM): (Masson and Denoeux, 2008):
  - Attribute data,
  - HCM, FCM family (alternate optimization of a cost function).
- Relational Evidential *c*-means (RECM): (Masson and Denoeux, 2009): ECM for proximity data.
- Constrained Evidential *c*-means (CECM) (Antoine et al., 2011): ECM with pairwise constraints.

Outline

Principle

- Problem: generate a credal partition $M = (m_1, \ldots, m_n)$ from attribute data $X = (\mathbf{x}_1, ..., \mathbf{x}_n)$, $\mathbf{x}_i \in \mathbb{R}^p$.
- Generalization of hard and fuzzy *c*-means algorithms:
  - Each class represented by a prototype;
  - Alternate optimization of a cost function with respect to the prototypes and to the credal partition.

Fuzzy c-means (FCM)

- Minimize

$$J_{\text{FCM}}(U, V) = \sum_{i=1}^{n} \sum_{k=1}^{c} u_{ik}^{\beta} d_{ik}^{2}$$

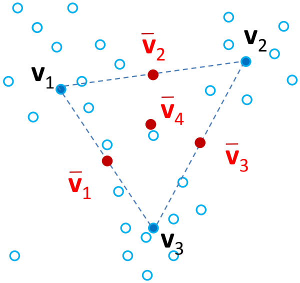
with $d_{ik} = ||\mathbf{x}_i - \mathbf{v}_k||$ under the constraints $\sum_k u_{ik} = 1, \forall i$.

- Alternate optimization algorithm:

$$\mathbf{v}_k = \frac{\sum_{i=1}^{n} u_{ik}^{\beta} \mathbf{x}_i}{\sum_{i=1}^{n} u_{ik}^{\beta}} \quad \forall k = 1, \ldots, c,$$

$$u_{ik} = \frac{d_{ik}^{-2/(\beta-1)}}{\sum_{\ell=1}^{c} d_{i\ell}^{-2/(\beta-1)}}.$$

ECM algorithm
Principle



- Each class ω_k represented by a prototype $\mathbf{v}_k$.
- Each non empty set of classes A_j represented by a prototype $\bar{\mathbf{v}}_j$ defined as the center of mass of the $\mathbf{v}_k$ for all $\omega_k \in A_j$.
- Basic ideas:
  - For each non empty $A_j \in \Omega$, $m_{ij} = m_i(A_j)$ should be high if $\mathbf{x}_i$ is close to $\bar{\mathbf{v}}_j$.
  - The distance to the empty set is defined as a fixed value δ.

ECM algorithm

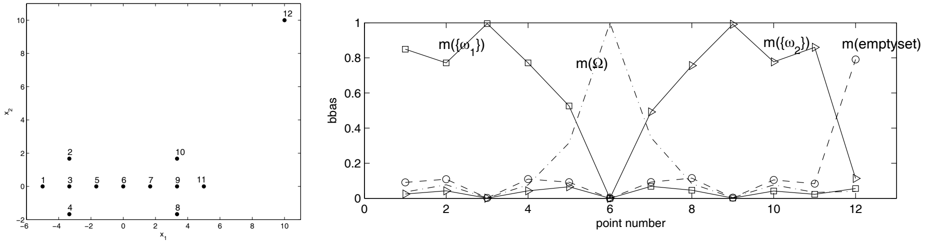
Objective criterion

- Criterion to be minimized:

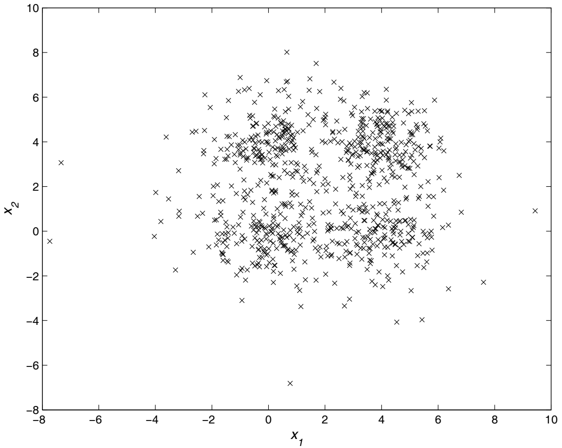
$$J_{\text{ECM}}(M, V) = \sum_{i=1}^{n} \sum_{\{j/A_j \neq \emptyset, A_j \subseteq \Omega\}} |A_j|^{\alpha} m_{ij}^{\beta} d_{ij}^2 + \sum_{i=1}^{n} \delta^2 m_{i\emptyset}^{\beta},$$

- Parameters:
  - α controls the specificity of mass functions;
  - β controls the hardness of the evidential partition;
  - δ controls the amount of data considered as outliers.
- $J_{\text{ECM}}(M, V)$ can be iteratively minimized with respect to M and V using an alternate optimization scheme.

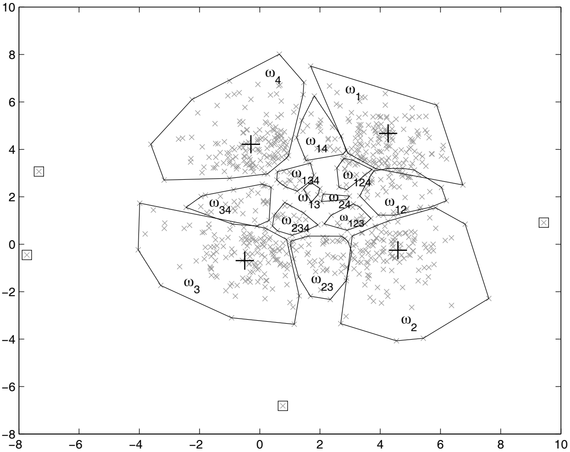
Butterfly dataset



4-class data set

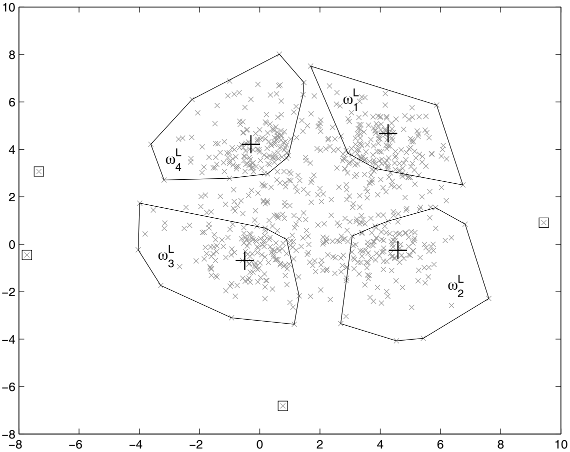


4-class data set
Hard credal partition



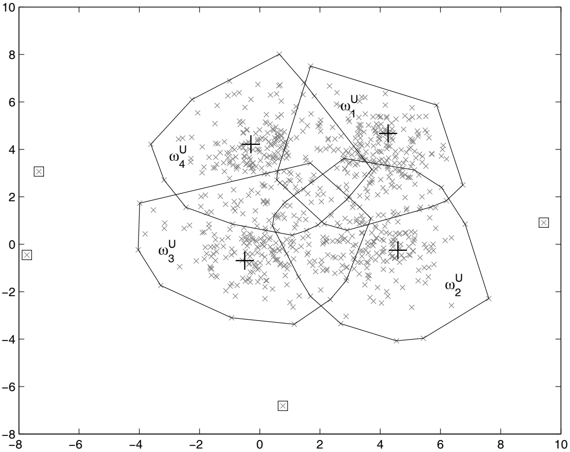
4-class data set

Lower approximation



4-class data set

Upper approximation



Outline

The problem

- We consider a statistical model $\{f(x,\theta), x \in \mathcal{X}, \theta \in \Theta\}$, where $\mathcal{X}$ is the sample space and Θ the parameter space.
- Having observed x, how to quantify the uncertainty about Θ, without specifying a prior probability distribution?
- Example:
  - We have observed 3 white balls out of 10 drawings from an urn with replacement. What does this observation tell us about the proportion θ of white balls?
  - In that case, $\mathcal{X} = \{0, \ldots, 10\}$, $\Theta = [0, 1]$ and $f(x,\theta) = C_n^x \theta^x (1-\theta)^{n-x}$.
- Two solutions using belief functions:
  1. Dempster's solution based an auxiliary variable with a pivotal probability distribution (Dempster, 1967);
  2. Likelihood-based approach (Shafer, 1976).

Likelihood-based belief function

Requirements

1. Likelihood principle: $Bel_\Theta(\cdot; x)$ should be based only on the likelihood function $L(\theta; x) = f(x; \theta)$.
2. Compatibility with Bayesian inference: when a Bayesian prior P_0 is available, combining it with $Bel_\Theta(\cdot, x)$ using Dempster's rule should yield the Bayesian posterior:

$$Bel_\Theta(\cdot, x) \oplus P_0 = P(\cdot|x).$$

3. Least commitment principle: among all the belief functions satisfying the previous two requirements, $Bel_\Theta(\cdot, x)$ should be the least committed (least informative).

Likelihood-based belief function
Solution

- From Requirements 1 and 2, the contour function of $Bel_\Theta(\cdot; x)$ should be proportional to $L(\theta; x)$:

$$pl(\theta; x) = cL(\theta; x)$$

  for some $c > 0$ depending only on the likelihood function $L(\theta; x)$.

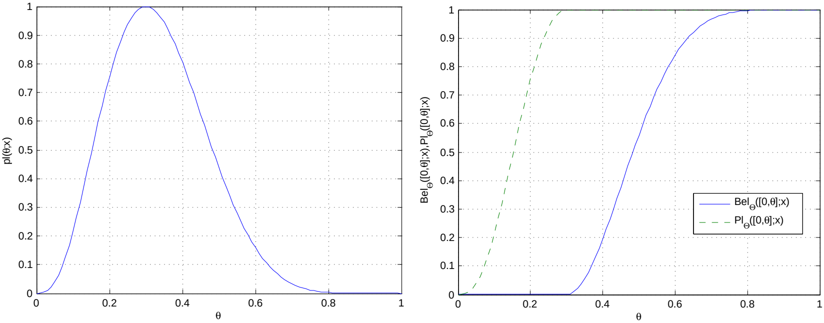
- From Requirement 3 with $\sqsubseteq_q$ as informational ordering, the unique solution is the consonant belief function $Bel_\Theta(\cdot; x)$ with contour function equal to the normalized likelihood:

$$pl(\theta; x) = \frac{L(\theta; x)}{\sup_{\theta' \in \Theta} L(\theta'; x)},$$

- The corresponding plausibility function is:

$$Pl_\Theta(A; x) = \sup_{\theta \in A} pl(\theta; x) = \frac{\sup_{\theta \in A} L(\theta; x)}{\sup_{\theta \in \Theta} L(\theta; x)}, \quad \forall A \subseteq \Theta.$$

Example: Binomial sample



Discussion

- The likelihood-based method is much simpler to implement than Dempster's method, even for complex models.
- By construction, it boils down to Bayesian inference when a Bayesian prior is available.
- It is compatible with usual likelihood-based inference:
  - Assume that $\theta = (\theta_1, \theta_2) \in \Theta_1 \times \Theta_2$ and θ_2 is a nuisance parameter. The marginal contour function on Θ_1

    $$pl(\theta_1; x) = \sup_{\theta_2 \in \Theta_2} pl(\theta_1, \theta_2; x) = \frac{\sup_{\theta_2 \in \Theta_2} L(\theta_1, \theta_2; x)}{\sup_{(\theta_1, \theta_2) \in \Theta} L(\theta_1, \theta_2; x)}$$

    is the relative profile likelihood function.
  - Let $H_0 \subset \Theta$ be a composite hypothesis. Its plausibility

    $$Pl(H_0; x) = \frac{\sup_{\theta \in H_0} L(\theta; x)}{\sup_{\theta \in \Theta} L(\theta; x)}.$$

    is the usual likelihood ratio statistics $\Lambda(x)$.

Outline

Motivation

- Classical statistical procedures address idealized situations where the data are precisely observed and can be considered as being drawn from a well defined population described by some parameter of interest θ.
- There are situations, however, where this simple model does not apply.
- For instance, some of data may collected from a population that is only known to “resemble” the population of interest (because, e.g., there were collected at different times or places) $\rightarrow$ partially relevant data.

Problem statement

- Assume that we are interested in a parameter $\theta \in \Theta$ related to a certain population and we observe a random variable X with probability density or mass function $f(x;\theta')$, where $\theta' \in \Theta$ is a parameter believed to be "close" to θ.
- For instance, θ might be the death rate in some hospital, and X the number of deaths in a neighboring hospital.
- Having observed $X = x$, our belief about θ' is represented by the contour function

$$pl'(\theta';x) = \frac{L(\theta';x)}{\sup_{\theta'} L(\theta';x)}.$$

- What does x tell us about θ?

Solution

- Assume that the statement "θ' is close to θ" can be formalized as $d(\theta, \theta') \leq \delta$, where d is a distance measure defined on Θ and δ is a known constant.
- This piece of information can be modeled by a logical belief function with focal set $S_\delta = \{(\theta, \theta') | d(\theta, \theta') \leq \delta\} \subset \Theta^2$.
- Combining it with $pl'(\theta'; x)$ using Dempster's rule yields a consonant belief function on $\Theta \times \Theta'$, with contour function

$$pl(\theta, \theta'; x) = pl'(\theta'; x) \mathbb{1}_{S_\delta}(\theta, \theta').$$

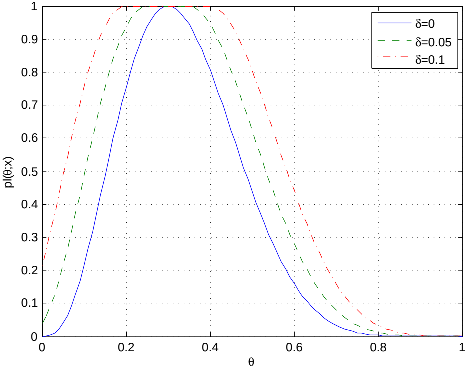
- Marginalizing out θ' yields:

$$pl(\theta; x) = \sup_{\theta'} pl(\theta, \theta'; x) = \sup_{\theta' \in B_\delta(\theta)} pl'(\theta'; x),$$

where $B_\delta(\theta) = \{\theta' \in \Theta | d(\theta, \theta') \leq \delta\}$.

Example

Assume we have observed 3 white balls out of 10 drawings with replacement from an urn with a proportion θ' of white balls. We are interested in the proportion θ of white balls in another urn. We know that $|\theta - \theta'| \leq \delta$. What do we know about θ?



Summary

- Developing pratical applications using the Dempster-Shafer framework requires modeling expert knowledge and statistical information using belief functions.
- Systematic and principled methods now exist:
  - Least-commitment principle;
  - GBT ;
  - Discounting;
  - Likelihood-based belief functions;
  - etc.
- Specific methods will be studied in following lectures (classification, etc.).
- More research on expert knowledge elicitation and statistical inference is needed.

References I

cf. https://www.hds.utc.fr/~tdenoeux

Ph. Smets.
Belief functions: the disjunctive rule of combination and the generalized Bayesian theorem.
International Journal of Approximate Reasoning, 9:1-35, 1993.

T. Denœux.
A k-nearest neighbor classification rule based on Dempster-Shafer theory.
IEEE Transactions on SMC, 25(05):804-813, 1995.

D. Mercier, B. Quost and T. Denœux.
Refined modeling of sensor reliability in the belief function framework using contextual discounting.
Information Fusion, Vol. 9, Issue 2, pages 246-258, 2008.

M.-H. Masson and T. Denœux.
ECM: An evidential version of the fuzzy c-means algorithm.
Pattern Recognition, 41(4):1384-1397, 2008.

References II

cf. https://www.hds.utc.fr/~tdenoeux

T. Denoeux.
Maximum likelihood estimation from Uncertain Data in the Belief Function Framework.
IEEE Transactions on Knowledge and Data Engineering, 25(1):119-130, 2013.

T. Denoeux.
Likelihood-based belief function: justification and some extensions to low-quality data
International Journal of Approximate Reasoning, accepted for publication, 2013.
Available at: http://hal.archives-ouvertes.fr/hal-00813021