Statistics and Machine Learning using belief functions
Lecture 2 – Decision Analysis. Application to Classification

Thierry Denœux
Université de Technologie de Compiègne
HEUDIASYC (UMR CNRS 6599)
https://www.hds.utc.fr/~tdenoeux

Beijing University of Technology
May 2017
Outline

1. Classical decision theory
   - Decision-making under complete ignorance
   - Decision-making with probabilities
   - Savage’s theorem

2. Decision-making with belief functions
   - Upper and lower expected utility
   - Other approaches
   - Axiomatic justifications

3. Evidential classification
   - Evidential $K$-NN rule
   - Evidential neural network classifier
   - Decision analysis
Example of decision problem under uncertainty

<table>
<thead>
<tr>
<th>Act (Purchase)</th>
<th>Good Economic Conditions</th>
<th>Poor Economic Conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apartment building</td>
<td>50,000</td>
<td>30,000</td>
</tr>
<tr>
<td>Office building</td>
<td>100,000</td>
<td>-40,000</td>
</tr>
<tr>
<td>Warehouse</td>
<td>30,000</td>
<td>10,000</td>
</tr>
</tbody>
</table>
A decision problem can be seen as a situation in which a decision-maker (DM) has to choose a course of action (an act) in some set $\mathcal{F}$.

An act may have different consequences (outcomes), depending on the state of nature.

Denoting by $\Omega = \{\omega_1, \ldots, \omega_n\}$ the set of states of nature and by $\mathcal{C} = \{c_1, \ldots, c_r\}$ the set of consequences (or outcomes), an act can be formalized as a mapping $f$ from $\Omega$ to $\mathcal{C}$.

In this lecture, the three sets $\Omega$, $\mathcal{C}$ and $\mathcal{F}$ will be assumed to be finite.
Formal framework
Utilities

- The desirability of the consequences can often be modeled by a utility function \( u : C \rightarrow \mathbb{R} \), which assigns a numerical value to each consequence.
- The higher this value, the more desirable is the consequence for the DM.
- In some problems, the consequences can be evaluated in terms of monetary value. The utilities can then be defined as the payoffs, or a function thereof.
- If the actions are indexed by \( i \) and the states of nature by \( j \), we will denote by \( u_{ij} \) the quantity \( u[f_i(\omega_j)] \).
- The \( n \times r \) matrix \( U = (u_{ij}) \) will be called a payoff or utility matrix.
## Payoff matrix

<table>
<thead>
<tr>
<th>Act (Purchase)</th>
<th>Good Economic Conditions ($\omega_1$)</th>
<th>Poor Economic Conditions ($\omega_2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apartment building ($f_1$)</td>
<td>50,000</td>
<td>30,000</td>
</tr>
<tr>
<td>Office building ($f_2$)</td>
<td>100,000</td>
<td>-40,000</td>
</tr>
<tr>
<td>Warehouse ($f_3$)</td>
<td>30,000</td>
<td>10,000</td>
</tr>
</tbody>
</table>
Formal framework

Preferences

- If the true state of nature $\omega$ is known, the desirability of an act $f$ can be deduced from that of its consequence $f(\omega)$
- Typically, the state of nature is unknown. Based on partial information, it is usually assumed that the DM can express preferences among acts, which may be represented mathematically by a preference relation $\succeq$ on $\mathcal{F}$
- This relation is interpreted as follows: given two acts $f$ and $g$, $f \succeq g$ means that $f$ is found by the DM to be at least as desirable as $g$
- We also define
  - The strict preference relation as $f \succ g$ iff $f \succeq g$ and not $(g \succeq f)$ (meaning that $f$ is strictly more desirable than $g$) and
  - The indifference relation $f \sim g$ iff $f \succeq g$ and $g \succeq f$ (meaning that $f$ and $g$ are equally desirable)
Quite often, the decision problem is to construct a preference relation among acts, from a utility matrix and some description of uncertainty, and to find the maximal elements of this relation.

Depending on the nature of the available information, different decision problems arise:

1. Decision-making under ignorance
2. Decision-making with probabilities
3. Decision-making with belief functions
Outline

1. Classical decision theory
   - Decision-making under complete ignorance
   - Decision-making with probabilities
   - Savage’s theorem

2. Decision-making with belief functions
   - Upper and lower expected utility
   - Other approaches
   - Axiomatic justifications

3. Evidential classification
   - Evidential $K$-NN rule
   - Evidential neural network classifier
   - Decision analysis
Outline

1. Classical decision theory
   - Decision-making under complete ignorance
     - Decision-making with probabilities
     - Savage’s theorem

2. Decision-making with belief functions
   - Upper and lower expected utility
   - Other approaches
   - Axiomatic justifications

3. Evidential classification
   - Evidential $K$-NN rule
   - Evidential neural network classifier
   - Decision analysis
Problem and non-domination principle

- We assume that the DM is **totally ignorant of the state of nature**: all the information given to the DM is the utility matrix $U$.

- A act $f_i$ is said to be **dominated** by $f_k$ if the outcomes of $f_k$ are at least as desirable as those of $f_i$ for all states, and strictly more desirable for at least one state:

  $$\forall j, \quad u_{kj} \geq u_{ij} \quad \text{and} \quad \exists j, \quad u_{kj} > u_{ij}$$

- **Non-domination principle**: an act cannot be chosen if it is dominated by another one.
### Example of a dominated act

<table>
<thead>
<tr>
<th>Act</th>
<th>Good Economic Conditions ($\omega_1$)</th>
<th>Poor Economic Conditions ($\omega_2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apartment building ($f_1$)</td>
<td>50,000</td>
<td>30,000</td>
</tr>
<tr>
<td>Office building ($f_2$)</td>
<td>100,000</td>
<td>-40,000</td>
</tr>
<tr>
<td>Warehouse ($f_3$)</td>
<td>30,000</td>
<td>10,000</td>
</tr>
</tbody>
</table>
Criteria for rational choice

- After all dominated acts have been removed, there remains the problem of ordering them by desirability, and of finding the set of most desirable acts.

- Several criteria of “rational choice” that have been proposed to derive a preference relation over acts

  1. **Laplace criterion**
     \[ f_i \succeq f_k \text{ iff } \frac{1}{n} \sum_j u_{ij} \geq \frac{1}{n} \sum_j u_{kj}. \]

  2. **Maximax criterion**
     \[ f_i \succeq f_k \text{ iff } \max_j u_{ij} \geq \max_j u_{kj}. \]

  3. **Maximin (Wald) criterion**
     \[ f_i \succeq f_k \text{ iff } \min_j u_{ij} \geq \min_j u_{kj}. \]
### Example

<table>
<thead>
<tr>
<th>Act</th>
<th>$\omega_1$</th>
<th>$\omega_2$</th>
<th>ave</th>
<th>max</th>
<th>min</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apartment ($f_1$)</td>
<td>50,000</td>
<td>30,000</td>
<td><strong>40,000</strong></td>
<td>50,000</td>
<td><strong>30,000</strong></td>
</tr>
<tr>
<td>Office ($f_2$)</td>
<td>100,000</td>
<td>-40,000</td>
<td>30,000</td>
<td><strong>100,000</strong></td>
<td>-40,000</td>
</tr>
</tbody>
</table>
Hurwicz criteria

- Hurwicz criterion: \( f_i \succeq f_k \) iff

\[
\alpha \min_j u_{ij} + (1 - \alpha) \max_j u_{ij} \geq \alpha \min_j u_{kj} + (1 - \alpha) \max_j u_{kj}
\]

where \( \alpha \) is a parameter in \([0, 1]\), called the pessimism index.

- Boils down to
  - the maximax criterion if \( \alpha = 0 \)
  - the maximin criterion if \( \alpha = 1 \)

- \( \alpha \) describes the DM’s attitude toward ambiguity.
Minimax regret criterion

(Savage) Minimax regret criterion: an act $f_i$ is at least as desirable as $f_k$ if it has smaller maximal regret, where regret is defined as the utility difference with the best act, for a given state of nature.

The regret $r_{ij}$ for act $f_i$ and state $\omega_j$ is

$$r_{ij} = \max_{\ell} u_{\ell j} - u_{ij}$$

The maximum regret for act $f_i$ is $R_i = \max_j r_{ij}$

$f_i \succeq f_k$ iff $R_i \leq R_k$
Example

- **Pay-off matrix**

<table>
<thead>
<tr>
<th>Act</th>
<th>$\omega_1$</th>
<th>$\omega_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apartment ($f_1$)</td>
<td>50,000</td>
<td>30,000</td>
</tr>
<tr>
<td>Office ($f_2$)</td>
<td>100,000</td>
<td>-40,000</td>
</tr>
</tbody>
</table>

- **Regret matrix**

<table>
<thead>
<tr>
<th>Act</th>
<th>$\omega_1$</th>
<th>$\omega_2$</th>
<th>max regret</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apartment ($f_1$)</td>
<td>50,000</td>
<td>0</td>
<td><strong>50,000</strong></td>
</tr>
<tr>
<td>Office ($f_2$)</td>
<td>0</td>
<td>70,000</td>
<td>70,000</td>
</tr>
</tbody>
</table>
The Laplace, maximax, maximin and Hurwicz criteria correspond to different ways of aggregating the utilities resulting each act, using, respectively, the average, the maximum, the minimum, and a convex sum of the minimum and the maximum.

These four operators belong to a family of operators called Ordered Weighted Average (OWA) operators (Yager, 1988).
OWA operators

- An OWA operator of dimension $n$ is a function $F : \mathbb{R}^n \rightarrow \mathbb{R}$ of the form

$$F(x_1, \ldots, x_n) = \sum_{i=1}^{n} w_i x(i)$$

where $x(i)$ is the $i$-th largest element in the collection $x_1, \ldots, x_n$, and $w_1, \ldots, w_n$ are positive weights verifying $\sum_{i=1}^{n} w_i = 1$

- The four previous operators are obtained for different choices of the weights:
  - Average: $(1/n, 1/n, \ldots, 1/n)$
  - Maximum: $(1, 0, \ldots, 0)$
  - Minimum: $(0, \ldots, 0, 1)$
  - Hurwicz: $(1 - \alpha, 0, \ldots, 0, \alpha)$
Setting the weights of an OWA operator

- In a decision-making context, each weight $w_i$ may be interpreted as a probability that the $i$-th best outcome will happen.
- Yager (1988) defines the degree of optimism of an OWA operator with weight vector $w$ as
  \[
  \text{OPT}(w) = \sum_{i=1}^{n} \frac{n - i}{n - 1} w_i
  \]
- $\text{OPT}(w) = 1$ for the maximum, $\text{OPT}(w) = 0$ for the minimum, $\text{OPT}(w) = 0.5$ for the mean, $\text{OPT}(w) = 1 - \alpha$ for Hurwicz.
- Given a degree of optimism $\beta$, we can then choose the OWA operator that maximizes the entropy
  \[
  \text{ENT}(w) = -\sum_{i=1}^{n} w_i \log w_i
  \]
  under the constraint $\text{OPT}(w) = \beta$. 
**Axioms of rational choice**

- Let $F^*$ denote the choice set, defined as a set of optimal acts.
- Arrow and Hurwicz (1972) have proposed four axioms a choice operator $F \rightarrow F^*$ should verify:
  1. **Axiom A1**: if $F_1 \subset F_2$ and $F_2^* \cap F_1 \neq \emptyset$, then $F_1^* = F_2^* \cap F_1$.
  2. **Axiom A2**: Relabeling actions and states does not change the optimal status of actions.
  3. **Axiom A3**: Deletion of a duplicate state does not change the optimality status of actions.
  4. **Axiom A4** (dominance): If $f \in F^*$ and $f'$ dominates $f$, then $f' \in F^*$. If $f \notin F^*$ and $f$ dominates $f'$, then $f' \notin F^*$.

- Under some regularity assumptions, Axioms A1 – A4 imply that the choice set depends only on the worst and the best consequences of each act.
- In particular, these axioms rule out the Laplace and minimax regret criteria.
### Violation of Axiom A3 by the Laplace criterion

<table>
<thead>
<tr>
<th>Act</th>
<th>$\omega_1$</th>
<th>$\omega_2$</th>
<th>ave</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apartment ($f_1$)</td>
<td>50,000</td>
<td>30,000</td>
<td><strong>40,000</strong></td>
</tr>
<tr>
<td>Office ($f_2$)</td>
<td>100,000</td>
<td>-40,000</td>
<td>30,000</td>
</tr>
</tbody>
</table>

Let us split the state of nature $\omega_1$ in two states: “Good economic conditions and there is life on Mars” ($\omega'_1$) and “Good economic conditions and there is no life on Mars” ($\omega''_1$)

<table>
<thead>
<tr>
<th>Act</th>
<th>$\omega'_1$</th>
<th>$\omega''_1$</th>
<th>$\omega_2$</th>
<th>ave</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apartment ($f_1$)</td>
<td>50,000</td>
<td>50,000</td>
<td>30,000</td>
<td><strong>43,333</strong></td>
</tr>
<tr>
<td>Office ($f_2$)</td>
<td>100,000</td>
<td>100,000</td>
<td>-40,000</td>
<td><strong>53,333</strong></td>
</tr>
</tbody>
</table>
Violation of Axiom A1 by minimax regret

- Pay-off matrix

<table>
<thead>
<tr>
<th>Act</th>
<th>$\omega_1$</th>
<th>$\omega_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apartment ($f_1$)</td>
<td>50,000</td>
<td>30,000</td>
</tr>
<tr>
<td>Office ($f_2$)</td>
<td>100,000</td>
<td>-40,000</td>
</tr>
<tr>
<td>$f_4$</td>
<td>130,000</td>
<td>-45,000</td>
</tr>
</tbody>
</table>

- Regret matrix

<table>
<thead>
<tr>
<th>Act</th>
<th>$\omega_1$</th>
<th>$\omega_2$</th>
<th>max regret</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apartment ($f_1$)</td>
<td>80,000</td>
<td>0</td>
<td>80,000</td>
</tr>
<tr>
<td>Office ($f_2$)</td>
<td>30,000</td>
<td>70,000</td>
<td>70,000</td>
</tr>
<tr>
<td>$f_4$</td>
<td>0</td>
<td>75,000</td>
<td>75,000</td>
</tr>
</tbody>
</table>

We had $\mathcal{F}_1 = \{f_1, f_2\}$ and $\mathcal{F}_1^* = \{f_1\}$. Now, $\mathcal{F}_2 = \{f_1, f_2, f_4\}$ and $\mathcal{F}_2^* = \{f_2\}$. So, $\mathcal{F}_1^* \neq \mathcal{F}_2^* \cap \mathcal{F}_1$.
Outline

1. Classical decision theory
   - Decision-making under complete ignorance
   - Decision-making with probabilities
   - Savage’s theorem

2. Decision-making with belief functions
   - Upper and lower expected utility
   - Other approaches
   - Axiomatic justifications

3. Evidential classification
   - Evidential $K$-NN rule
   - Evidential neural network classifier
   - Decision analysis
Let us now consider the situation where uncertainty about the state of nature is quantified by probabilities $p_1, \ldots, p_n$ on $\Omega$.

These probabilities can be objective (decision under risk) or subjective.

We can then compute, for each act $f_i$, its expected utility as

$$EU(f_i) = \sum_j u_{ij}p_j$$

Maximum Expected Utility (MEU) principle: an act $f_i$ is more desirable than an act $f_k$ if it has a higher expected utility: $f_i \succeq f_k$ iff $EU(f_i) \geq EU(f_k)$.
### Example

<table>
<thead>
<tr>
<th>Act</th>
<th>$\omega_1$</th>
<th>$\omega_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apartment ($f_1$)</td>
<td>50,000</td>
<td>30,000</td>
</tr>
<tr>
<td>Office ($f_2$)</td>
<td>100,000</td>
<td>-40,000</td>
</tr>
</tbody>
</table>

Assume that there is 60% chance that the economic situation will be poor ($\omega_2$). The expected utilities of acts $f_1$ and $f_2$ are

$$EU(f_1) = 50,000 \times 0.4 + 30,000 \times 0.6 = 38,000$$

$$EU(f_2) = 100,000 \times 0.4 - 40,000 \times 0.6 = 16,000$$

Act $f_1$ is thus more desirable according to the maximum expected utility criterion.
Axiomatic justification of the MEU principle

- The MEU principle was first axiomatized by von Neumann and Morgenstern (1944).
- Given a probability distribution on $\Omega$, an act $f : \Omega \rightarrow C$ induces a probability measure $P$ on the set $C$ of consequences (assumed to be finite), called a lottery.
- We denote by $\mathcal{L}$ the set of lotteries on $C$.
- If we agree that two acts providing the same lottery are equivalent, then the problem of comparing the desirability of acts becomes that of comparing the desirability of lotteries.
- Let $\succeq$ be a preference relation among lotteries. Von Neumann and Morgenstern argued that, to be rational, a preference relation should verify three axioms.
### Von Neumann and Morgenstern’s axioms

1. **Complete preorder:** the preference relation is a complete and non trivial preorder (i.e., it is a reflexive, transitive and complete relation) on $L$

2. **Continuity:** for any lotteries $P$, $Q$ and $R$ such that $P \succ Q \succ R$, there exists a probabilities $\alpha$ and $\beta$ in $[0, 1]$ such that

$$\alpha P + (1 - \alpha) R \succ Q \succ \beta P + (1 - \beta) R$$

where $\alpha P + (1 - \alpha) R$ is a compound lottery, which refers to the situation where you receive $P$ with probability $\alpha$ and $Q$ with probability $1 - \alpha$. This axiom implies, in particular, that there is no lottery $R$ that is so undesirable that it cannot become desirable if mixed with some very desirable lottery $P$

3. **Independence:** for any lotteries $P$, $Q$ and $R$ and for any $\alpha \in (0, 1]$

$$P \succeq Q \iff \alpha P + (1 - \alpha) R \succeq \alpha Q + (1 - \alpha) R$$
Von Neumann and Morgenstern’s theorem

The two following propositions are equivalent:

1. The preference relation $\succeq$ verifies the axioms of complete preorder, continuity, and independence

2. There exists a utility function $u : C \rightarrow \mathbb{R}$ such that, for any two lotteries $P = (p_1, \ldots, p_r)$ and $Q = (q_1, \ldots, q_r)$

$$P \succeq Q \iff \sum_{i=1}^{r} p_i u(c_i) \geq \sum_{i=1}^{r} q_i u(c_i)$$

Function $u$ is unique up to a strictly increasing affine transformation
Outline

1. Classical decision theory
   - Decision-making under complete ignorance
   - Decision-making with probabilities
   - Savage’s theorem

2. Decision-making with belief functions
   - Upper and lower expected utility
   - Other approaches
   - Axiomatic justifications

3. Evidential classification
   - Evidential $K$-NN rule
   - Evidential neural network classifier
   - Decision analysis
Savage’s theorem

- We have reviewed some criteria for decision-making under complete ignorance, i.e., when uncertainty cannot be probabilized.
- Some researchers have defended the view that a rational DM always maximizes expected utility, for some subjective probability measure and utility function.
- **Savage’s theorem (1954):** A preference relation \( \succsim \) among acts verifies some rationality requirements iff there is a finitely additive probability measure \( P \) and a utility function \( u : \mathcal{C} \rightarrow \mathbb{R} \) such that

  \[
  \forall f, g \in \mathcal{F}, \quad f \succsim g \iff \int_{\Omega} u(f(\omega))dP(\omega) \geq \int_{\Omega} u(g(\omega))dP(\omega)
  \]

  Furthermore, \( P \) is unique and \( u \) is unique up to a positive affine transformation.
- A strong argument for probabilism, but Savage’s axioms can be questioned!
Savage’s axioms

- Savage has proposed seven axioms, four of which are considered as meaningful (the other three are technical)
- Axiom 1: $\succeq$ is a total preorder (complete, reflexive and transitive)
- Axiom 2 [Sure Thing Principle]. Given $f, h \in \mathcal{F}$ and $E \subseteq \Omega$, let $fEh$ denote the act defined by

$$
(fEh)(\omega) = \begin{cases} 
  f(\omega) & \text{if } \omega \in E \\
  h(\omega) & \text{if } \omega \notin E 
\end{cases}
$$

Then the Sure Thing Principle states that $\forall E, \forall f, g, h, h'$

$$
fEh \succeq gEh \Rightarrow fEh' \succeq gEh'
$$

- This axiom seems reasonable, but it is not verified empirically!
Ellsberg’s paradox

Suppose you have an urn containing 30 red balls and 60 balls, either black or yellow. Consider the following 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
- $f_3$: 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 $f_2$, but they strictly prefer $f_4$ to $f_3$

<table>
<thead>
<tr>
<th></th>
<th>$R$</th>
<th>$B$</th>
<th>$Y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_1$</td>
<td>100</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$f_2$</td>
<td>0</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>$f_3$</td>
<td>100</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>$f_4$</td>
<td>0</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

Now, $f_1 = f_1\{R, B\}0$, $f_2 = f_2\{R, B\}0$, $f_3 = f_1\{R, B\}100$, $f_4 = f_2\{R, B\}100$

The Sure Thing Principle is violated!
Summary

Classically, we distinguish two kinds of decision problems:

1. Decision under ignorance: we only know, for each act, a set of possible outcomes
2. Decision under risk: we are given, for each act, a probability distribution over the outcomes

It has been argued that any decision problem under uncertainty should be handled as a problem of decision under risk. However, the axiomatic arguments are questionable.

In the next part: decision-making when uncertainty is described by a belief functions
Outline

1. Classical decision theory
   - Decision-making under complete ignorance
   - Decision-making with probabilities
   - Savage’s theorem

2. Decision-making with belief functions
   - Upper and lower expected utility
   - Other approaches
   - Axiomatic justifications

3. Evidential classification
   - Evidential $K$-NN rule
   - Evidential neural network classifier
   - Decision analysis
Belief functions become of component of a decision problem in any of the following two situations (or both)

1. The decision maker’s subjective beliefs concerning the state of nature may be described by a belief function $Bel_\Omega$ on $\Omega$

2. The DM may not be able to precisely describe the outcomes of some acts under each state of nature
Case 1: uncertainty described by a belief function

- Let $m^\Omega$ be a mass function on $\Omega$
- Any act $f : \Omega \to C$ then carries $m^\Omega$ to the set $C$ of consequences, yielding a mass function $m^C_f$, which quantifies the DM's beliefs about the outcome of act $f$
- Each mass $m^\Omega(A)$ is transferred to $f(A)$

$$m^C_f(B) = \sum_{\{A \subseteq \Omega | f(A) = B\}} m^\Omega(A)$$

for any $B \subseteq C$

- $m^C_f$ is a credibilistic lottery corresponding to act $f$
In that case, an act may formally be represented by a multi-valued mapping $f : \Omega \rightarrow 2^C$, assigning a set of possible consequences $f(\omega) \subseteq C$ to each state of nature $\omega$.

Given a probability measure $P$ on $\Omega$, $f$ then induces the following mass function $m^C_f$ on $C$,

$$m^C_f(B) = \sum_{\{\omega \in \Omega | f(\omega) = B\}} p(\omega)$$

for all $B \subseteq \Omega$. 

---

Case 2: partial knowledge of outcomes
Example

- Let $\Omega = \{\omega_1, \omega_2, \omega_3\}$ and $m^\Omega$ the following mass function
  
  \[
  m^\Omega(\{\omega_1, \omega_2\}) = 0.3, \quad m^\Omega(\{\omega_2, \omega_3\}) = 0.2 \\
  m^\Omega(\{\omega_3\}) = 0.4, \quad m^\Omega(\Omega) = 0.1
  \]

- Let $C = \{c_1, c_2, c_3\}$ and $f$ the act
  
  \[
  f(\omega_1) = \{c_1\}, \quad f(\omega_2) = \{c_1, c_2\}, \quad f(\omega_3) = \{c_2, c_3\}
  \]

- To compute $m^C_f$, we transfer the masses as follows
  
  \[
  m^\Omega(\{\omega_1, \omega_2\}) = 0.3 \rightarrow f(\omega_1) \cup f(\omega_2) = \{c_1, c_2\} \\
  m^\Omega(\{\omega_2, \omega_3\}) = 0.2 \rightarrow f(\omega_2) \cup f(\omega_3) = \{c_1, c_2, c_3\} \\
  m^\Omega(\{\omega_3\}) = 0.4 \rightarrow f(\omega_3) = \{c_2, c_3\} \\
  m^\Omega(\Omega) = 0.1 \rightarrow f(\omega_1) \cup f(\omega_2) \cup f(\omega_3) = \{c_1, c_2, c_3\}
  \]

- Finally, we obtain the following mass function on $C$
  
  \[
  m^C(\{c_1, c_2\}) = 0.3, \quad m^C(\{c_2, c_3\}) = 0.4, \quad m^C(C) = 0.3
  \]
Decision problem

- In the two situations considered above, we can assign to each act $f$ a credibilistic lottery, defined as a mass function on $C$.
- Given a utility function $u$ on $C$, we then need to extend the MEU model.
- Several such extensions will now be reviewed.
Outline

1. Classical decision theory
   - Decision-making under complete ignorance
   - Decision-making with probabilities
   - Savage’s theorem

2. Decision-making with belief functions
   - Upper and lower expected utility
   - Other approaches
   - Axiomatic justifications

3. Evidential classification
   - Evidential $K$-NN rule
   - Evidential neural network classifier
   - Decision analysis
Upper and lower expectations

- Let $m$ be a mass function on $\mathcal{C}$, and $u$ a utility function $\mathcal{C} \rightarrow \mathbb{R}$.
- The **lower and upper expectations** of $u$ are defined, respectively, as the averages of the minima and the maxima of $u$ within each focal set of $m$

\[
\bar{E}_m(u) = \sum_{A \subseteq \mathcal{C}} m(A) \min_{c \in A} u(c) \\
\underline{E}_m(u) = \sum_{A \subseteq \mathcal{C}} m(A) \max_{c \in A} u(c)
\]

- It is clear that $\bar{E}_m(u) \leq \underline{E}_m(u)$, with the inequality becoming an equality when $m$ is Bayesian, in which case the lower and upper expectations collapse to the usual expectation.
- If $m = m_A$ is logical with focal set $A$, then $\bar{E}_m(u)$ and $\underline{E}_m(u)$ are, respectively, the minimum and the maximum of $u$ in $A$. 

The lower and upper expectations are lower and upper bounds of expectations with respect to probability measures compatible with $m$:

\[
\mathbb{E}_m(u) = \min_{P \in \mathcal{P}(m)} \mathbb{E}_P(u)
\]

\[
\overline{E}_m(u) = \max_{P \in \mathcal{P}(m)} \mathbb{E}_P(u)
\]

The mean of minima (res., maxima) is also the minimum (resp., maximum) of means with respect to all compatible probability measures.
Corresponding decision criteria

- Having defined the notions of lower and upper expectations, we can define two preference relations among credibilistic lotteries as

\[ m_1 \succ m_2 \text{ iff } \underline{E}_{m_1}(u) \geq \underline{E}_{m_2}(u) \]

and

\[ m_1 \preceq m_2 \text{ iff } \overline{E}_{m_1}(u) \leq \overline{E}_{m_2}(u) \]

- Relation \( \succ \) corresponds to a pessimistic (or conservative) attitude of the DM. When \( m \) is logical, it corresponds to the maximin criterion.

- Symmetrically, \( \preceq \) corresponds to an optimistic attitude and extends the maximax criterion.

- Both criteria boil down to the MEU criterion when \( m \) is Bayesian.
Back to Ellsberg’s paradox

Here, $\Omega = \{R, B, Y\}$ and $m^\Omega(\{R\}) = 1/3$, $m^\Omega(\{B, Y\}) = 2/3$

The mass functions on $C = \{0, 100\}$ induced by the four acts are

$m_1(\{100\}) = 1/3$, $m_1(\{0\}) = 2/3$
$m_2(\{0\}) = 1/3$, $m_2(\{0, 100\}) = 2/3$
$m_3(\{100\}) = 1/3$, $m_3(\{0, 100\}) = 2/3$
$m_4(\{0\}) = 1/3$, $m_4(\{100\}) = 2/3$

Corresponding lower and upper expectations

<table>
<thead>
<tr>
<th></th>
<th>R</th>
<th>B</th>
<th>Y</th>
<th>$\underline{E}_m(u)$</th>
<th>$\overline{E}_m(u)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_1$</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>$u(100)/3$</td>
<td>$u(100)/3$</td>
</tr>
<tr>
<td>$f_2$</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>$u(0)$</td>
<td>$2u(100)/3$</td>
</tr>
<tr>
<td>$f_3$</td>
<td>100</td>
<td>0</td>
<td>100</td>
<td>$u(100)/3$</td>
<td>$u(100)$</td>
</tr>
<tr>
<td>$f_4$</td>
<td>0</td>
<td>100</td>
<td>100</td>
<td>$2u(100)/3$</td>
<td>$2u(100)/3$</td>
</tr>
</tbody>
</table>
Interval dominance

- If we drop the requirement that the preference relation among acts be complete, then we can consider the interval dominance relation,

  \[ m_1 \succ_{ID} m_2 \text{ iff } \underline{E}_{m_1}(u) \geq \underline{E}_{m_2}(u) \]

- Given a collection of credibilistic lotteries, we can then compute the set of maximal (i.e., non dominated) elements of \( \succ_{ID} \)

- Imprecise probability view

  \[ m_1 \succ_{ID} m_2 \iff \forall P_1 \in \mathcal{P}(m_1), \forall P_2 \in \mathcal{P}(m_2), \underline{E}_{P_1}(u) \geq \underline{E}_{P_2}(u) \]

- The justification for this preference relation is not so clear from the point of view of belief function theory (i.e., if one does not interpret a belief function as a lower probability)
Outline

1. Classical decision theory
   - Decision-making under complete ignorance
   - Decision-making with probabilities
   - Savage’s theorem

2. Decision-making with belief functions
   - Upper and lower expected utility
   - Other approaches
   - Axiomatic justifications

3. Evidential classification
   - Evidential $K$-NN rule
   - Evidential neural network classifier
   - Decision analysis
The Hurwicz criterion can be generalized as

$$E_{m, \alpha}(u) = \sum_{A \subseteq C} m(A) \left( \alpha \min_{c \in A} u(c) + (1 - \alpha) \max_{c \in A} u(c) \right)$$

$$= \alpha E_m(u) + (1 - \alpha) E(u)$$

where $\alpha \in [0, 1]$ is a pessimism index.

This criterion was introduced and justified axiomatically by Jaffray (1988).

Strat (1990) who proposed to interpret $\alpha$ as the DM’s subjective probability that the ambiguity will be resolved unfavorably.
Transferable belief model

- A completely different approach to decision-making with belief function was advocated by Smets, as part of the **Transferable Belief Model**
- Smets defended a two-level mental model
  1. a **credal level**, where an agent’s belief are represented by belief functions, and
  2. a **pignistic level**, where decisions are made by maximizing the EU with respect to a probability measure derived from a belief function
- The rationale for introducing probabilities at the decision level is the avoidance of **Dutch books**
- Smets argued that the belief-probability transformation $T$ should be **linear**, i.e., it should verify

$$T(\alpha m_1 + (1 - \alpha) m_2) = \alpha T(m_1) + (1 - \alpha) T(m_2),$$

for any mass functions $m_1$ and $m_2$ and for any $\alpha \in [0, 1]$
Pignistic transformation

- The only linear belief-probability transformation \( T \) is the **pignistic transformation**, with \( p_m = T(m) \) given by

\[
p_m(c) = \sum_{\{A \subseteq C | c \in A\}} \frac{m(A)}{|A|}, \quad \forall c \in C
\]

- The pignistic probability \( p_m \) is mathematically identical to the **Shapley value** in cooperative game theory.

- The expected utility w.r.t. the pignistic probability is

\[
E_p(u) = \sum_{c \in C} p_m(c)u(c) = \sum_{A \subseteq C} m(A) \left( \frac{1}{|A|} \sum_{c \in A} u(c) \right)
\]

- The maximum pignistic expected utility criterion thus extends the **Laplace criterion**
Generalized OWA criteria

- A more general family of expected utility criteria can be defined by aggregating the utilities $u(c)$ within each focal set $A$ using OWA operators.
- To determine the weights of the OWA operators, Yager (1992) proposes to fix the degree of optimism $\beta$ and to use the maximum-entropy operators, for each cardinality $|A|$

$$E_{m, \beta}^{\text{owa}} = \sum_{A \subseteq C} m(A) F_{|A|, \beta} \left( \{ u(c) | c \in A \} \right)$$

where $F_{|A|, \beta}$ is the maximum-entropy OWA operator with degree of optimism $\beta$ and arity $|A|$
- Parameter $\beta$ has roughly the same interpretation as one minus the pessimism index $\alpha$ in the Hurwicz criterion.
- However, each $F_{|A|, \beta} \left( \{ u(c) | c \in A \} \right)$ depends on all the values $u(c)$ for all $c \in A$, and not only on the minimum and the maximum.
Generalized minimax regret

- Yager (2004) also extended the minimax regret criterion to belief functions.
- We need to consider $n$ acts $f_1, \ldots, f_n$, and we write $u_{ij} = u[f_i(\omega_j)]$.
- The regret if act $f_i$ is selected, and state $\omega_j$ occurs, is $r_{ij} = \max_k u_{kj} - u_{ij}$.
- For a non-empty subset $A$ of $\Omega$, the maximum regret of act $f_i$ is
  $$R_i(A) = \max_{\omega_j \in A} r_{ij}$$

- The expected maximal regret for act $f_i$ is
  $$\overline{R}_i = \sum_{\emptyset \neq A \subseteq \Omega} m^\Omega(A) R_i(A)$$

- Act $f_i$ is preferred over act $f_k$ if $\overline{R}_i \leq \overline{R}_k$.
- The minimax regret criterion is recovered when $m^\Omega$ is logical.
- The MEU model is recovered when $m^\Omega$ is Bayesian.
## Summary

<table>
<thead>
<tr>
<th>non-probabilized</th>
<th>belief functions</th>
<th>probabilized</th>
</tr>
</thead>
<tbody>
<tr>
<td>maximin</td>
<td>←→ lower expectation</td>
<td></td>
</tr>
<tr>
<td>maximax</td>
<td>←→ upper expectation</td>
<td></td>
</tr>
<tr>
<td>Laplace</td>
<td>←→ pignistic expectation</td>
<td></td>
</tr>
<tr>
<td>Hurwicz</td>
<td>←→ generalized Hurwicz</td>
<td></td>
</tr>
<tr>
<td>OWA</td>
<td>←→ generalized OWA</td>
<td></td>
</tr>
<tr>
<td>minimax regret</td>
<td>←→ generalized minimax regret</td>
<td></td>
</tr>
</tbody>
</table>
Outline

1. Classical decision theory
   - Decision-making under complete ignorance
   - Decision-making with probabilities
   - Savage’s theorem

2. Decision-making with belief functions
   - Upper and lower expected utility
   - Other approaches
   - Axiomatic justifications

3. Evidential classification
   - Evidential K-NN rule
   - Evidential neural network classifier
   - Decision analysis
Linear utility of credibilistic lotteries

- Except for the generalized minimax regret criterion, the previous decision criteria are of the form

$$m_1 \succeq m_2 \text{ iff } U(m_1) \geq U(m_2)$$

with

$$U(m) = \sum_{\emptyset \neq A \subseteq C} m(A)U(m_A)$$

where $m_A$ is the logical mass function with focal set $A$

- Writing $U(A)$ in place of $U(m_A)$, and $u(c)$ for $U(\{c\})$
  - $U(A) = \min_{c \in A} u(c)$ for the lower expectation criterion
  - $U(A) = \max_{c \in A} u(c)$ for the upper expectation criterion
  - $U(A) = \alpha \min_{c \in A} u(c) + (1 - \alpha) \max_{c \in A} u(c)$ for the Hurwicz criterion
  - $U(A) = (1/|A|) \sum_{c \in A} u(c)$ for the pignistic criterion
  - $U(A) = F_{|A|, \beta}(\{u(c) | c \in A\})$ for the OWA criterion
Jaffray’s theorem

Jaffray (1989) showed that a preference relation among credibilistic lotteries is representable by a linear utility function if and only if it verifies the Von Neumann and Morgenstern axioms extended to credibilistic lotteries, i.e.,

1. **Transitivity and Completeness:** $\succsim$ is a transitive and complete relation (i.e., is a weak order)

2. **Continuity:** for all $m_1$, $m_2$ and $m_3$ such that $m_1 \succ m_2 \succ m_3$, there exists $\alpha$, $\beta$ in $(0, 1)$ such that

$$\alpha m_1 + (1 - \alpha) m_3 \succ m_2 \succ \beta m_1 + (1 - \beta) m_3$$

3. **Independence:** for all $m_1$ and $m_2$ and for all $\alpha$ in $(0, 1)$, $m_1 \succ m_2$ implies

$$\alpha m_1 + (1 - \alpha) m_3 \succ \alpha m_2 + (1 - \alpha) m_3$$
Consequences of Jaffray’s theorem

- Under the previous requirements, we thus have
  \[ U(m) = \sum_{\emptyset \neq A \subseteq C} m(A) U(A) \]

- The EU is recovered when \( m \) is Bayesian
  \[ U(m) = \sum_{c \in C} m(\{ c \}) u(c) \]

- Jaffray’s theorem does not tell us how to compute \( U(A) \). In the general case, we need to elicit the utility values \( U(A) \) for each subset \( A \subseteq C \) of consequences, which limits the practical use of the method

- However, Jaffray (1989) showed that a major simplification can be achieved by introducing an additional axiom
Dominance axiom

Let us write \( c_1 \succeq c_2 \) whenever \( m\{c_1\} \succeq m\{c_2\} \)

Furthermore, let \( c_A \) and \( \overline{c}_A \) denote, respectively, the worst and the best consequence in \( A \)

**Dominance axiom:** for all non-empty subsets \( A \) and \( B \) of \( C \), if \( c_A \succeq c_B \) and \( \overline{c}_A \succeq \overline{c}_B \), then \( m_A \succeq m_B \)

**Justification:**

- If \( c_A \succeq c_B \) and \( \overline{c}_A \succeq \overline{c}_B \), it is possible to construct a set \( \Omega \) of states of nature, and two acts \( f : \Omega \to A \) and \( f' : \Omega \to B \), such that, for any \( \omega \in \Omega \),
  \[
  f(\omega) \succ f'(\omega)
  \]
- As act \( f \) dominates \( f' \), it should be preferred whatever the information on \( \Omega \)
- Hence, \( f \) should be preferred to \( f' \) when we have a vacuous mass function on \( \Omega \), in which case \( f \) and \( f' \) induce, respectively, the logical mass function \( m_A \) and \( m_B \) on \( C \)

**Consequence:** \( U(A) \) can be written as \( U(A) = u(c_A, \overline{c}_A) \)
Example

- Assume that $c_1 \succeq c_2 \succeq c_3 \succeq c_4 \succeq c_5 \succeq c_6$
- Let $A = \{c_1, c_4, c_5\}$ and $B = \{c_2, c_3, c_6\}$
- Consider the two acts

<table>
<thead>
<tr>
<th></th>
<th>$\omega_1$</th>
<th>$\omega_2$</th>
<th>$\omega_3$</th>
<th>$\omega_4$</th>
<th>$\omega_5$</th>
<th>$\omega_6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f$</td>
<td>$c_1$</td>
<td>$c_4$</td>
<td>$c_5$</td>
<td>$c_1$</td>
<td>$c_1$</td>
<td>$c_1$</td>
</tr>
<tr>
<td>$g$</td>
<td>$c_6$</td>
<td>$c_6$</td>
<td>$c_6$</td>
<td>$c_2$</td>
<td>$c_3$</td>
<td>$c_6$</td>
</tr>
</tbody>
</table>

- $f$ dominates $g$: it should be preferred whatever the information on $\Omega$
- With $m^\Omega$ vacuous, we get $m_f^C = m_A^C$ and $m_g^C = m_B^C$
- Hence, $m_A^C \succeq m_B^C$
Local Hurwicz criterion

- Adding the dominance axiom to the three previous ones,

\[ U(m) = \sum_{\emptyset \neq A \subseteq C} m(A) u(c_A, \overline{c}_A) \]

- We can write

\[ u(c, \overline{c}) = \alpha(c, \overline{c}) u(c) + (1 - \alpha(c, \overline{c})) u(\overline{c}) \]

where \( \alpha(c, \overline{c}) \) is a local pessimism index, defined as the value of \( \alpha \) which makes the DM indifferent between:

1. Receiving at least \( c \) and at most \( \overline{c} \), with no further information, and
2. Receiving either \( c \) with probability \( \alpha \) or \( \overline{c} \) with probability \( 1 - \alpha \).

- We then have

\[ U(m) = \sum_{\emptyset \neq A \subseteq C} m(A) [\alpha(c_A, \overline{c}_A) u(c_A) + (1 - \alpha(c_A, \overline{c}_A)) u(\overline{c}_A)] \]

- The generalized Hurwicz criterion corresponds to the case where \( \alpha(c, \overline{c}) \) is equal to a constant \( \alpha \)
Summary

- Several criteria for decision-making with belief functions have been reviewed.
- These criteria mix criteria for decision-making under ignorance, and the MEU principle.
- A general form of the Hurwicz principle can be justified axiomatically, assuming that uncertainty is quantified by belief functions.
- There is no counterpart of Savage’s theorem for belief functions.
References on decision

J.-Y. Jaffray.
Linear utility theory for belief functions.

P. Smets.
Decision making in a context where uncertainty is represented by belief functions.

J.-Y. Jaffray and P. Wakker.
Decision making with belief functions: compatibility and incompatibility with the sure-thing principle.

I. Gilboa.
Expected utility with purely subjective non-additive probabilities.
Outline

1. Classical decision theory
   - Decision-making under complete ignorance
   - Decision-making with probabilities
   - Savage’s theorem

2. Decision-making with belief functions
   - Upper and lower expected utility
   - Other approaches
   - Axiomatic justifications

3. Evidential classification
   - Evidential $K$-NN rule
   - Evidential neural network classifier
   - Decision analysis
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 \( x \in \mathbb{R}^p \)
- A class label \( y \in \Omega \)

Problem: given a learning set \( \mathcal{L} = \{(x_1, y_1), \ldots, (x_n, y_n)\} \), predict the class label of a new instance described by \( x \).
Outline

1. Classical decision theory
   - Decision-making under complete ignorance
   - Decision-making with probabilities
   - Savage’s theorem

2. Decision-making with belief functions
   - Upper and lower expected utility
   - Other approaches
   - Axiomatic justifications

3. Evidential classification
   - Evidential $K$-NN rule
   - Evidential neural network classifier
   - Decision analysis
Let $N_K(x) \subset \mathcal{L}$ denote the set of the $K$ nearest neighbors of $x$ in $\mathcal{L}$, based on some distance measure. Each $x_i \in N_K(x)$ can be considered as a piece of evidence regarding the class of $x$. The strength of this evidence decreases with the distance $d_i$ between $x$ and $x_i$.
Definition

- If \( y_i = \omega_k \), the evidence of \((x_i, y_i)\) can be represented by

\[
m_i(\{\omega_k\}) = \varphi_k(d_i) \\
m_i(\{\omega_\ell\}) = 0, \quad \forall \ell \neq k \\
m_i(\Omega) = 1 - \varphi(d_i)
\]

where \( \varphi_k, k = 1, \ldots, c \) are decreasing functions from \([0, +\infty)\) to \([0, 1]\) such that \( \lim_{d \to +\infty} \varphi_k(d) = 0 \)

- The evidence of the \( K \) nearest neighbors of \( x \) is pooled using Dempster's rule of combination

\[
m = \bigoplus_{x_i \in N_K(x)} m_i
\]

- Decision: any of the decision rules mentioned in the first part.
- With 0-1 losses and no rejection, the optimistic, pessimistic and pignistic rules yield the same decisions (see below).
Learning

- Choice of functions $\varphi_k$: for instance, $\varphi_k(d) = \alpha \exp(-\gamma_k d^2)$.
- Parameters $\gamma_1, \ldots, \gamma_c$ can be optimized (see below).
- Parameter $\gamma = (\gamma_1, \ldots, \gamma_c)$ can be learnt from the data by minimizing the following cost function

$$C(\gamma) = \sum_{i=1}^{n} \sum_{k=1}^{c} (pl_{(-i)}(\omega_k) - t_{ik})^2,$$

where

- $pl_{(-i)}$ is the contour function obtained by classifying $x_i$ using its $K$ nearest neighbors in the learning set.
- $t_{ik} = 1$ is $y_i = k$, $t_{ik} = 0$ otherwise.
- Function $C(\gamma)$ can be minimized by an iterative nonlinear optimization algorithm.
Computation of $pl_{(-i)}$

- Contour function from each neighbor $x_j \in \mathcal{N}_K(x_i)$:

$$pl_j(\omega_k) = \begin{cases} 1 & \text{if } y_j = \omega_k \\ 1 - \varphi_k(d_{ij}) & \text{otherwise} \end{cases}, \quad k = 1, \ldots, c$$

- Contour function of the combined mass function

$$pl_{(-i)}(\omega_k) \propto \prod_{x_j \in \mathcal{N}_K(x_i)} (1 - \varphi_k(d_{ij}))^{1-t_{jk}}$$

where $t_{jk} = 1$ if $y_j = \omega_k$ and $t_{jk} = 0$ otherwise

- It can be computed in time proportional to $K|\Omega|$
Example 1: Vehicles dataset

- The data were used to distinguish 3D objects within a 2-D silhouette of the objects.
- Four classes: bus, Chevrolet van, Saab 9000 and Opel Manta.
- 846 instances, 18 numeric attributes.
- The first 564 objects are training data, the rest are test data.
Vehicles datasets: result

Vehicles data

\begin{center}
\begin{tikzpicture}
\begin{axis}[
    title={Vehicles data},
    xlabel={K},
    ylabel={test error rate},
    xmin=2, xmax=15,
    ymin=0.32, ymax=0.40,
    legend style={at={(0.5,0.85)},anchor=north},
]
\addplot[->, dashed, mark=o, mark size=2pt, color=black]
coordinates {
(2,0.33) (4,0.34) (6,0.32) (8,0.34) (10,0.40) (12,0.38) (14,0.36)
};
\addlegendentry{EK-NN}
\addplot[->, dashdotted, mark=o, mark size=2pt, color=black]
coordinates {
(2,0.32) (4,0.34) (6,0.36) (8,0.38) (10,0.40) (12,0.38) (14,0.36)
};
\addlegendentry{voting K-NN}
\end{axis}
\end{tikzpicture}
\end{center}
Example 2: Ionosphere dataset

- This dataset was collected by a radar system and consists of phased array of 16 high-frequency antennas with a total transmitted power of the order of 6.4 kilowatts.
- The targets were free electrons in the ionosphere. "Good" radar returns are those showing evidence of some type of structure in the ionosphere. "Bad" returns are those that do not.
- There are 351 instances and 34 numeric attributes. The first 175 instances are training data, the rest are test data.
Ionosphere datasets: result

Ionosphere data

Test error rate

EK-NN
voting K-NN

Thierry Denœux (UTC/HEUDIASYC)
Belief Functions Seminar
BJUT, May 2017
Implementation in R

```r
library("evclass")

data("ionosphere")
xapp<-ionosphere$x[1:176,]
yapp<-ionosphere$y[1:176]
x tst<-ionosphere$x[177:351,]
y tst<-ionosphere$y[177:351]

opt<-EkNNfit(xapp,yapp,K=10)
class<-EkNNval(xapp,yapp,xtst,K=10,ytst,opt$param)

> class$err
0.07428571
> table(ytst, class$ypred)
ytst 1 2
1 106 6
2 7 56
```
Partially supervised data

We now consider a learning set of the form

$$\mathcal{L} = \{(x_i, m_i), i = 1, \ldots, n\}$$

where

- \(x_i\) 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 mass function \( m_i \) is discounted (weakened) with a rate depending on the distance \( d_i \)

\[
m'_i(A) = \varphi(d_i) \ m_i(A) , \quad \forall A \subset \Omega
\]

\[
m'_i(\Omega) = 1 - \sum_{A \subset \Omega} m'_i(A)
\]

- The \( K \) mass functions \( m'_i \) are combined using Dempster’s rule

\[
m = \bigoplus_{x_i \in \mathcal{N}_K(x)} m'_i
\]
Example: EEG data

EEG signals encoded as 64-D patterns, 50 % positive (K-complexes), 50 % negative (delta waves), 5 experts.
Results on EEG data
(De-noeux and Zouhal, 2001)

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

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>0.30</td>
<td>0.30</td>
<td>0.31</td>
<td>0.27</td>
</tr>
<tr>
<td>11</td>
<td>0.29</td>
<td>0.30</td>
<td>0.29</td>
<td>0.26</td>
</tr>
<tr>
<td>13</td>
<td>0.31</td>
<td>0.30</td>
<td>0.31</td>
<td>0.26</td>
</tr>
</tbody>
</table>
Outline

1. Classical decision theory
   - Decision-making under complete ignorance
   - Decision-making with probabilities
   - Savage’s theorem

2. Decision-making with belief functions
   - Upper and lower expected utility
   - Other approaches
   - Axiomatic justifications

3. Evidential classification
   - Evidential $K$-NN rule
   - Evidential neural network classifier
   - Decision analysis
Principle

- The learning set is summarized by \( r \) prototypes.
- Each prototype \( p_i \) has membership degree \( u_{ik} \) to each class \( \omega_k \), with \( \sum_{k=1}^{c} u_{ik} = 1 \).
- Each prototype \( p_i \) is a piece of evidence about the class of \( x \), whose reliability decreases with the distance \( d_i \) between \( x \) and \( p_i \).
Propagation equations

- Mass function induced by prototype $p_i$:

$$m_i(\{\omega_k\}) = \alpha_i u_{ik} \exp(-\gamma_i d_i^2), \quad k = 1, \ldots, c$$

$$m_i(\Omega) = 1 - \alpha_i \exp(-\gamma_i d_i^2)$$

- Combination:

$$m = \bigoplus_{i=1}^{r} m_i$$

The computation of $m_i$ requires $O(rp)$ arithmetic operations (where $p$ denotes the number of inputs), and the combination can be performed in $O(rc)$ operations. Hence, the overall complexity is $O(r(p + c))$ operations to compute the output for one input pattern.

- The combined mass function $m$ has as focal sets the singletons $\{\omega_k\}$, $k = 1, \ldots, c$ and $\Omega$. 

Neural network implementation
Learning

- The parameters are the
  - The prototypes $p_i$, $i = 1, \ldots, r$ ($rp$ parameters)
  - The membership degrees $u_{ik}$, $i = 1, \ldots, r$, $k = 1 \ldots, c$ ($rc$ parameters)
  - The $\alpha_i$ and $\gamma_i$, $i = 1 \ldots, r$ ($2r$ parameters).

- Let $\theta$ denote the vector of all parameters. It can be estimated by minimizing a cost function such as

$$C(\theta) = \sum_{i=1}^{n} (pl_{ik} - t_{ik})^2 + \mu \sum_{i=1}^{r} \alpha_i$$

where $pl_{ik}$ is the output plausibility for instance $i$ and class $k$, $t_{ik} = 1$ if $y_i = k$ and $t_{ik} = 0$ otherwise, and $\mu$ is a regularization coefficient (hyperparameter).

- The hyperparameter $\mu$ can be optimized by cross-validation.
Implementation in R

library("evclass")

data(glass)
xtr<-glass$x[1:89,]
ytr<-glass$y[1:89]
xtst<-glass$x[90:185,]
ytst<-glass$y[90:185]

param0<-proDSinit(xtr,ytr,nproto=7)
fit<-proDSfit(x=xtr,y=ytr,param=param0)
val<-proDSval(xtst,fit$param,ytst)

> print(val$err)
0.3333333

> table(ytst,val$ypred)
ytst 1 2 3 4
1 30 6 4 0
2 6 27 1 3
3 4 3 1 0
4 0 5 0 6
Results on the Iris data
Mass on $\{\omega_1\}$
Results on the Iris data

Mass on $\{\omega_2\}$
Results on the Iris data

Mass on $\{\omega_3\}$
Results on the Iris data

Mass on $\Omega$
Results on the Iris data

Plausibility of $\{\omega_1\}$
Results on the Iris data

Plausibility of \{\omega_2\}
Results on the Iris data

Plausibility of \( \{ \omega_3 \} \)
**Results on classical data**

<table>
<thead>
<tr>
<th>Classifier</th>
<th>test error rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multi-layer perceptron (88 units)</td>
<td>0.49</td>
</tr>
<tr>
<td>Radial Basis Function (528 units)</td>
<td>0.47</td>
</tr>
<tr>
<td>Gaussian node network (528 units)</td>
<td>0.45</td>
</tr>
<tr>
<td>Nearest neighbor</td>
<td>0.44</td>
</tr>
<tr>
<td>Linear Discriminant Analysis</td>
<td>0.56</td>
</tr>
<tr>
<td>Quadratic Discriminant Analysis</td>
<td>0.53</td>
</tr>
<tr>
<td>CART</td>
<td>0.56</td>
</tr>
<tr>
<td>BRUTO</td>
<td>0.44</td>
</tr>
<tr>
<td>MARS (degree=2)</td>
<td>0.42</td>
</tr>
<tr>
<td>Evidential NN (33 prototypes)</td>
<td>0.38</td>
</tr>
<tr>
<td>Evidential NN (44 prototypes)</td>
<td>0.37</td>
</tr>
<tr>
<td>Evidential NN (55 prototypes)</td>
<td>0.37</td>
</tr>
</tbody>
</table>

**Vowel data**

- \( c = 11 \)
- \( p = 10 \)
- \( n = 568 \)
- test: 462 ex.
- (different speakers)
Data fusion example

- \( c = 2 \) classes
- Learning set (\( n = 60 \)): \( \mathbf{x} \in \mathbb{R}^5, \mathbf{x}' \in \mathbb{R}^3 \), Gaussian distributions, conditionally independent
- Test set (real operating conditions): \( \mathbf{x} \leftarrow \mathbf{x} + \mathbf{\epsilon}, \mathbf{\epsilon} \sim \mathcal{N}(0, \sigma^2 I) \)
## Results

Test error rates: uncorrupted data

<table>
<thead>
<tr>
<th>Method</th>
<th>$x$ alone</th>
<th>$x'$ alone</th>
<th>$x$ and $x'$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evidental NN</td>
<td>0.106</td>
<td>0.148</td>
<td>0.061</td>
</tr>
<tr>
<td>MLP</td>
<td>0.113</td>
<td>0.142</td>
<td>0.063</td>
</tr>
<tr>
<td>RBF</td>
<td>0.133</td>
<td>0.159</td>
<td>0.083</td>
</tr>
<tr>
<td>QUAD</td>
<td>0.101</td>
<td>0.141</td>
<td>0.049</td>
</tr>
<tr>
<td>BAYES</td>
<td>0.071</td>
<td>0.121</td>
<td>0.028</td>
</tr>
</tbody>
</table>
Results

Test error rates: $x + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2 I)$

![Graph showing error rates for different classifiers](image-url)
Outline

1. Classical decision theory
   - Decision-making under complete ignorance
   - Decision-making with probabilities
   - Savage’s theorem

2. Decision-making with belief functions
   - Upper and lower expected utility
   - Other approaches
   - Axiomatic justifications

3. Evidential classification
   - Evidential $K$-NN rule
   - Evidential neural network classifier
   - Decision analysis
To formalize the decision problem, we need to define:

1. The set of consequences
2. The set of acts
3. The utility function

Let

- \( C = \{\text{correct}, \text{error}\} \)
- \( \mathcal{F} = \{f_1, \ldots, f_c\} \) with \( f_k \) = assignment to class \( \omega_k \),

\[
f_k(\omega_k) = \text{correct}, \quad f_k(\omega_\ell) = \text{error}, \quad \forall \ell \neq k
\]

- \( u(\text{correct}) = 1, \; u(\text{error}) = 0 \)

In classification, we more often use the notion of loss, to be minimized. Here, the loss function can be defined as

\[
\lambda(\text{correct}) = 0, \quad \lambda(\text{error}) = 0.
\]

The expected loss is called the risk.
Simple decision setting (continued)

- Given a mass function $m$ on $\Omega$, act $f_k$ induces the following mass $m_k$ on $C$:

\[
\begin{align*}
  m_k(\{\text{correct}\}) &= m(\{\omega_k\}) = Bel(\{\omega_k\}) \\
  m_k(\{\text{error}\}) &= \sum_{\omega_k \notin A} m(A) \quad (1 - Pl(\{\omega_k\})) \\
  m_k(C) &= \sum_{\omega_k \in A, |A| > 1} m(A) 
\end{align*}
\]

- The lower and upper risk are

\[
\begin{align*}
  \underline{E}_{m_k}(\lambda) &= m_k(\{\text{correct}\}) \times 0 + m_k(\{\text{error}\}) \times 1 + m_k(C) \times 0 \\
  &= 1 - Pl(\{\omega_k\}) \\
  \overline{E}_{m_k}(\lambda) &= m_k(\{\text{correct}\}) \times 0 + m_k(\{\text{error}\}) \times 1 + m_k(C) \times 1 \\
  &= 1 - Bel(\{\omega_k\}) 
\end{align*}
\]

- When the focal sets of $m$ are $\{\omega_k\}$, $k = 1, \ldots, c$ and $\Omega$, the different decision rules (optimistic, pessimistic, Hurwicz, pignistic) are equivalent.
Implementation in R

\[
\begin{align*}
\text{param0} & \leftarrow \text{proDSinit}(x, y, 6) \\
\text{fit} & \leftarrow \text{proDSfit}(x, y, \text{param0}) \\
\text{val} & \leftarrow \text{proDSval}(\text{xtst}, \text{fit}\$\text{param}) \\
L & \leftarrow 1 - \text{diag}(c) \\
D & \leftarrow \text{decision}(\text{val}\$\text{m}, L=L, \text{rule}=\text{’upper’})
\end{align*}
\]
Decision regions (Iris data)
Let us now assume

- \( C = \{ \text{correct}, \text{error}, \text{reject} \} \)
- \( F = \{ f_0, f_1, \ldots, f_c \} \), where \( f_0 \) denotes rejection,

\[
    f_0(\omega_k) = \text{reject}, \quad \forall k
\]

and \( f_k \) = assignment to class \( \omega_k \), as before.

\( \lambda(\text{correct}) = 0, \lambda(\text{error}) = 1, \lambda(\text{reject}) = \lambda_0 \)

We can carry out the analysis as before. In this case, the different decision rules generally lead to different decisions.
param0<-proDSinit(x,y,6)
fit<-proDSfit(x,y,param0)

val<-proDSval(xtst,fit$param)
L<-cbind(1-diag(c),rep(0.3,c))
D1<-decision(val$m,L=L,rule='upper')
D2<-decision(val$m,L=L,rule='lower')
D3<-decision(val$m,L=L,rule='pignistic')
D4<-decision(val$m,L=L,rule='hurwicz',rho=0.5)
Decision regions (Iris data)

Lower risk

Thierry Denœux (UTC/HEUDIASYC)
Decision regions (Iris data)

Upper risk
Decision regions (Iris data)
Pignistic risk
Decision regions (Iris data)
Hurwicz strategy ($\rho = 0.5$)
Decision with rejection and novelty detection

- Assume that there exists an unknown class $\omega_u$, not represented in the learning set.
- Let the acts now be
  - $f_k = \text{assignment to class } \omega_k$, $k = 1, \ldots, c$
  - $f_u = \text{assignment to class } \omega_u$
  - $f_0 = \text{rejection}$
- Assume the loss function is defined by the following matrix

<table>
<thead>
<tr>
<th></th>
<th>$f_1$</th>
<th>$f_2$</th>
<th>$f_3$</th>
<th>$f_0$</th>
<th>$f_u$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\omega_1$</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>$\lambda_0$</td>
<td>$\lambda_u$</td>
</tr>
<tr>
<td>$\omega_2$</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>$\lambda_0$</td>
<td>$\lambda_u$</td>
</tr>
<tr>
<td>$\omega_3$</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>$\lambda_0$</td>
<td>$\lambda_u$</td>
</tr>
<tr>
<td>$\omega_u$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>$\lambda_0$</td>
<td>0</td>
</tr>
</tbody>
</table>
Implementation in R

```r
param0 <- proDSinit(x, y, 6)
fit <- proDSfit(x, y, param0)

val <- proDSval(xtst, fit$param)
L <- cbind(1 - diag(c), rep(0.3, c), rep(0.32, c))
L <- rbind(L, c(1, 1, 1, 0.3, 0))
D1 <- decision(val$m, L=L, rule='lower')
D2 <- decision(val$m, L=L, rule='pignistic')
D3 <- decision(val$m, L=L, rule='hurwicz', rho=0.5)
```
Decision regions (Iris data)

Lower risk
Decision regions (Iris data)
Pignistic risk
Decision regions (Iris data)

Hurwicz, $\rho = 0.5$
References on classification

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

T. Denœux.  
A k-nearest neighbor classification rule based on Dempster-Shafer theory.  

T. Denœux.  
A neural network classifier based on Dempster-Shafer theory.  

T. Denœux.  
Analysis of evidence-theoretic decision rules for pattern classification.  