ML01 – Introduction to Machine Learning

Evidential machine learning

Thierry Denœux

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

Université de technologie de Compiègne

Spring 2021

Uncertainty in machine learning

- In ML, it is important to quantify uncertainty about
  - The predictions (classification, regression)
  - Knowledge extracted from the data (clustering)
- Most approaches are based on probability theory, but a current trend in ML is to investigate the use of other mathematical frameworks for modeling and reasoning with uncertainty.
- One of these frameworks is the theory of belief functions (also called evidence theory).
- ML based on evidence theory is called evidential ML. It is the topic of this chapter.

Overview

1. Theory of belief functions
   - Representation of evidence
   - Dempster's rule

2. Evidential classification
   - Evidential K-NN classifier
   - Evidential neural network classifier

3. Evidential clustering
   - Evidential clustering
   - ECM
   - EVCLUS

Theory of belief functions

- A mathematical formalism called
  - Dempster-Shafer (DS) theory
  - Evidence theory
  - Theory of belief functions
- This formalism was introduced by A. P. Dempster in the 1960's for statistical inference, and developed by G. Shafer in the late 1970's into a general theory for reasoning under uncertainty.
- DS generalizes probability theory.
- Many applications in engineering (information fusion, uncertainty quantification, risk analysis) and AI (expert systems, machine learning).

Overview

1. Theory of belief functions
   - Representation of evidence
   - Dempster's rule
2. Evidential classification
   - Evidential K-NN classifier
   - Evidential neural network classifier
3. Evidential clustering
   - Evidential clustering
   - ECM
   - EVCLUS

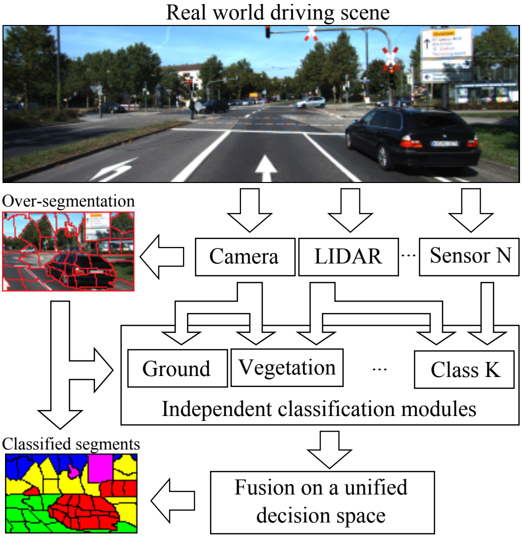
Mass function

- Let Y be a variable taking one and only one value in a finite set Ω, called the frame of discernment.
- Evidence (uncertain information) about Y can be represented by a mass function $m : 2^{\Omega} \rightarrow [0,1]$ such that

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

- Every subset 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 throughout most of this chapter, unless otherwise specified.

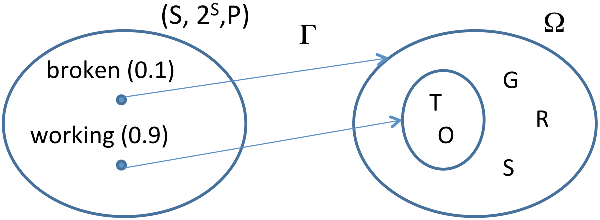
Example: road scene analysis



Example: road scene analysis (continued)

- Let Y be the type of object in some region of the image, and $\Omega = \{G, R, T, O, S\}$, corresponding to the possibilities **G**rass, **R**oad, **T**ree/Bush, **O**bstacle, **S**ky.
- Assume that a lidar sensor (laser telemeter) returns the information $Y \in \{T, O\}$, but we there is a probability $p = 0.1$ that the information is not reliable (because, e.g., the sensor is out of order).
- How to represent this information by a mass function?

Formalization



- Here, the probability p is not about Y, 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 $X \in \{T, O\}$.
  - 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

$$m(\{T, O\}) = 0.9, \quad m(\Omega) = 0.1$$

General framework

- A piece of evidence (information) about Y can be represented by
  - A set $S = \{s_1, \ldots, s_r\}$ of interpretations
  - A probability measure P on S
  - A multi-valued mapping $\Gamma : S \to 2^\Omega$
- Under interpretation $s \in S$, the evidence tells us that $Y \in \Gamma(s)$, and nothing more. The probability $P(\{s\})$ is transferred to the focal set $A = \Gamma(s)$ and we have

$$m(A) = P(\{s \in S : \Gamma(s) = A\})$$

- $m(A)$ is the probability of knowing that $Y \in A$, and nothing more, given the available evidence.

Special cases

Logical mass function If a mass function has only one focal set $A \subseteq \Omega$, it is said to be logical and it is denoted by m_A.

- Example: $m_{\{T,O\}}$ means the mass function such that $m_{\{T,O\}}(\{T,O\}) = 1$.
- Special case: m_Ω, the vacuous mass function, represents total ignorance.

Bayesian mass function If all focal sets of m are singletons, m is said to be Bayesian. It is equivalent to a probability distribution.

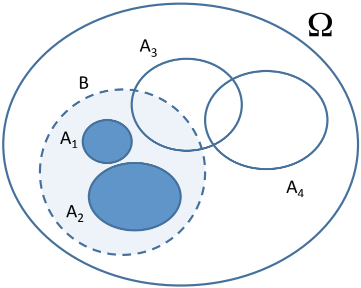
- Example: $m(\{T\}) = 0.5$, $m(\{O\}) = 0.5$.

A Dempster-Shafer mass function can thus be seen as

- a generalized set
- a generalized probability distribution

Belief function

- If the evidence tells us that the truth is in A, and $A \subseteq B$, we say that the evidence supports B.



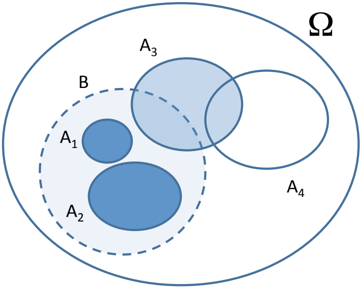
- Given a normalized mass function m, the probability that the evidence supports B is thus

$$\boxed{Bel(B) = \sum_{A \subseteq B} m(A)}$$

- The number $Bel(B)$ is called the credibility of B, or the degree of belief in B, and the function $B \rightarrow Bel(B)$ is called a belief function.

Plausibility function

- If the evidence tells us that the truth is in A, and $A \cap B \neq \emptyset$, we say that the evidence is consistent with B.



- The probability that the evidence is consistent with B is, thus,

$$\boxed{Pl(B) = \sum_{A \cap B \neq \emptyset} m(A)}$$

- The number $Pl(B)$ is called the plausibility of B, and the function $B \rightarrow Pl(B)$ is called a plausibility function.

Interpretation and elementary properties

- Properties:
  1. $Bel(A) \leq Pl(A)$ for all $A \subseteq \Omega$
  2. $Bel(\emptyset) = Pl(\emptyset) = 0$
  3. $Bel(\Omega) = Pl(\Omega) = 1$
  4. For all $A \subseteq \Omega$,
$$Bel(A) = 1 - Pl(\overline{A})$$
$$Pl(A) = 1 - Bel(\overline{A})$$
- Interpretation:
  - $Bel(A)$ is the probability that A is supported by the evidence
  - $Bel(\overline{A})$ is the probability that $\overline{A}$ is supported by the evidence
  - $Pl(A) = 1 - Bel(\overline{A})$ is the probability that $\overline{A}$ is not supported by the evidence, i.e., that A is consistent with the evidence

Relations between m, Bel and Pl

- Let m be a mass function, Bel and Pl the corresponding belief and plausibility functions
- Thanks to the following equations, given any one of these functions, we can recover the other two: for all $A \subseteq \Omega$,

$$Bel(A) = \sum_{B \subseteq A} m(B)$$

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

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

- m, Bel et Pl are thus three equivalent representations of a piece of evidence.

Overview

1. Theory of belief functions
   - Representation of evidence
   - Dempster's rule
2. Evidential classification
   - Evidential K-NN classifier
   - Evidential neural network classifier
3. Evidential clustering
   - Evidential clustering
   - ECM
   - EVCLUS

Road scene example continued

- Variable Y was defined as the type of object in some region of the image, and the frame was $\Omega = \{G, R, T, O, S\}$, corresponding to the possibilities **G**rass, **R**oad, **T**ree/Bush, **O**bstacle, **S**ky
- A lidar sensor gave us the following mass function:

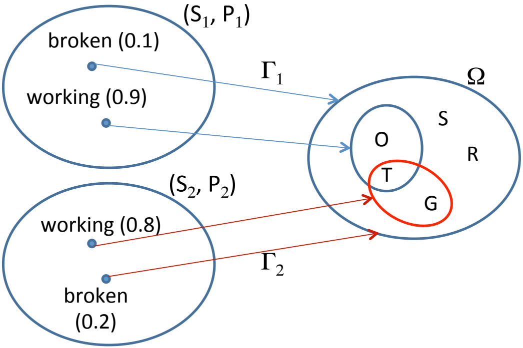
$$m_1(\{T, O\}) = 0.9, \quad m_1(\Omega) = 0.1$$

- Now, assume that a camera returns the mass function:

$$m_2(\{G, T\}) = 0.8, \quad m_2(\Omega) = 0.2$$

- 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 $P_1(\{s_1\})P_2(\{s_2\})$

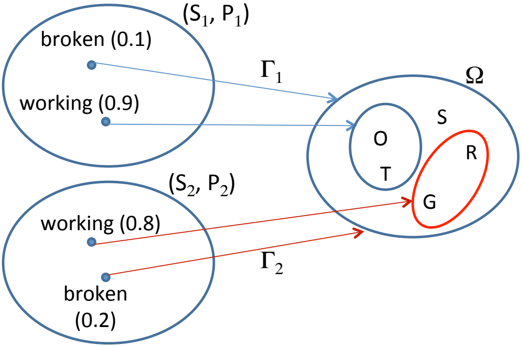
Computation

$m_1 \backslash m_2$	$\{T, G\}$ (0.8)	Ω (0.2)
$\{O, T\}$ (0.9)	$\{T\}$ (0.72)	$\{O, T\}$ (0.18)
Ω (0.1)	$\{T, G\}$ (0.08)	Ω (0.02)

We then get the following combined mass function,

$$
\begin{aligned}
m(\{T\}) &= 0.72 \\
m(\{O, T\}) &= 0.18 \\
m(\{T, G\}) &= 0.08 \\
m(\Omega) &= 0.02
\end{aligned}
$$

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

$m_1 \backslash m_2$	$\{G, R\}$ (0.8)	Ω (0.2)
$\{O, T\}$ (0.9)	$\emptyset$ (0.72)	$\{O, T\}$ (0.18)
Ω (0.1)	$\{G, R\}$ (0.08)	Ω (0.02)

We then get the following combined mass function,

$$
\begin{aligned}
m(\emptyset) &= 0 \\
m(\{O, T\}) &= 0.18/0.28 = 9/14 \\
m(\{G, R\}) &= 0.08/0.28 = 4/14 \\
m(\Omega) &= 0.02/0.28 = 1/14
\end{aligned}
$$

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

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

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

- $m_1 \oplus m_2$ is called the orthogonal sum of m_1 and m_2
- This rule can be used to combine mass functions induced by independent pieces of evidence

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

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

Properties

1. Commutativity: $\forall m_1, m_2,\ m_1 \oplus m_2 = m_2 \oplus m_1$
2. Associativity: $\forall m_1, m_2, m_3,\ (m_1 \oplus m_2) \oplus m_3 = m_1 \oplus (m_2 \oplus m_3)$.
3. Neutral element: $\forall m,\ m \oplus m_\Omega = m$.
4. Generalization of intersection: if m_A and m_B are logical mass functions and $A \cap B \neq \emptyset$, then
$$m_A \oplus m_B = m_{A\cap B}$$
5. Let $Pl_{1\oplus 2}$ be the plausibility function corresponding to $m_1 \oplus m_2$. Then,
$$\forall \omega \in \Omega, \quad Pl_{1\oplus 2}(\{\omega\}) = \frac{Pl_1(\{\omega\})Pl_2(\{\omega\})}{1-\kappa}$$

Overview

1. Theory of belief functions
   - Representation of evidence
   - Dempster's rule
2. Evidential classification
   - Evidential K-NN classifier
   - Evidential neural network classifier
3. Evidential clustering
   - Evidential clustering
   - ECM
   - EVCLUS

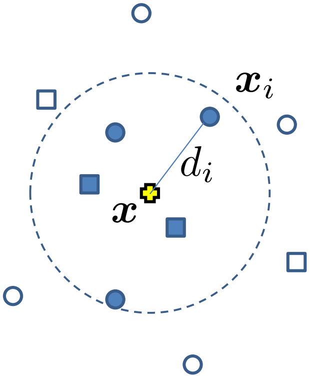
Evidential classifier

- Sometimes, the class cannot be predicted from the feature vector with high certainty.
- Assessing the uncertainty in the classification is an important issue.
- Most traditional classifiers represent uncertainty by computing a conditional probability distribution $P(\cdot \mid \mathbf{x})$
- An evidential classifier represents classification uncertainty using belief functions.
- There are several methods to construct evidential classifiers. We will see two of them:
  1. The evidential K nearest neighbor (EK-NN) classifier
  2. The evidential neural network classifier

Overview

1. Theory of belief functions
   - Representation of evidence
   - Dempster's rule
2. Evidential classification
   - Evidential K-NN classifier
   - Evidential neural network classifier
3. Evidential clustering
   - Evidential clustering
   - ECM
   - EVCLUS

Principle



- 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 $\mathbf{x}_i$

EK-NN classifier

Modeling evidence from the each NN

- Let $\mathbf{x}_i \in \mathcal{N}_K(\mathbf{x})$ and assume that $y_i = k$.
- The evidence of $(\mathbf{x}_i, y_i)$ can be represented by the mass function

$$m_i(\{\omega_k\}) = \varphi_k(d_i)$$
$$m_i(\Omega) = 1 - \varphi_k(d_i)$$

  where φ_k is a decreasing function from $[0, +\infty)$ to $[0, 1]$ such that $\lim_{d \to +\infty} \varphi_k(d) = 0$. (When $d \to +\infty$, m_i tends to the vacuous mass function).
- Common choice for φ_k:

$$\varphi_k(d) = \alpha \exp(-\gamma_k d^2)$$

  where α and $(\gamma_1, \ldots, \gamma_c)$ are parameters.

EK-NN classifier

Combination of evidence from the K NN

- The evidence of the K nearest neighbors of $\mathbf{x}$ is pooled using Dempster's rule of combination

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

- The focal sets of m are the singletons $\{\omega_k\}$, $k = 1, \ldots, c$ and Ω.
- A decision can be made by selecting the class with the highest plausibility:

$$C(\mathbf{x}) = \arg\max_k Pl(\{\omega_k\})$$

Learning

- Assume $\varphi_k(d) = \alpha \exp(-\gamma_k d^2)$.
- Parameter $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_c)$ can be learnt from the data by minimizing the following loss function

$$J(\gamma) = \sum_{i=1}^{n} \sum_{k=1}^{c} (Pl_{(-i)}(\{\omega_k\}) - y_{ik})^2,$$

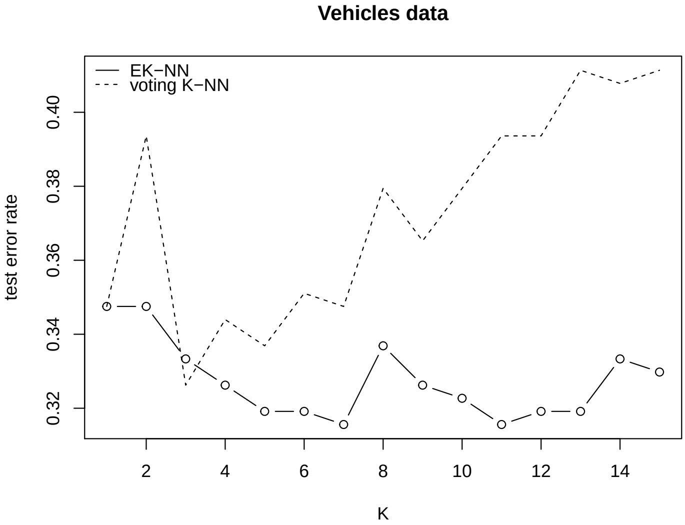
where $Pl_{(-i)}$ is the plausibility function obtained by classifying $\mathbf{x}_i$ using its K nearest neighbors in the learning set, and $y_{ik} = I(y_i = k)$

- Function $J(\gamma)$ can be minimized by an iterative nonlinear optimization algorithm.

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.

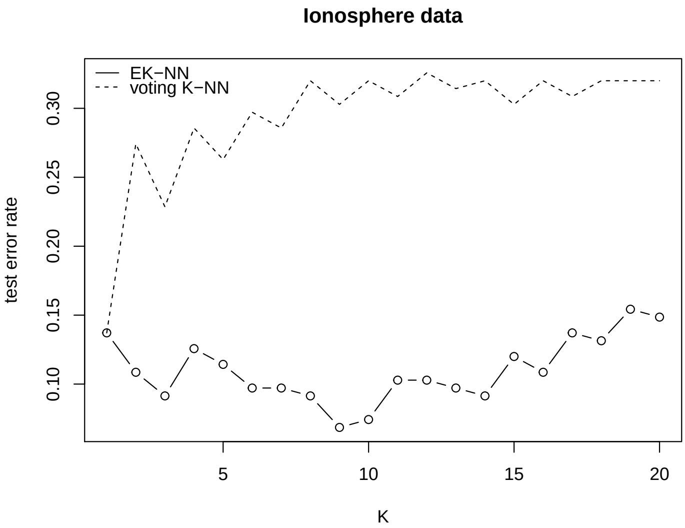
Vehicles datasets: result



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

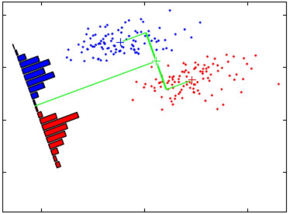


Implementation in R

```
library("evclass")

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

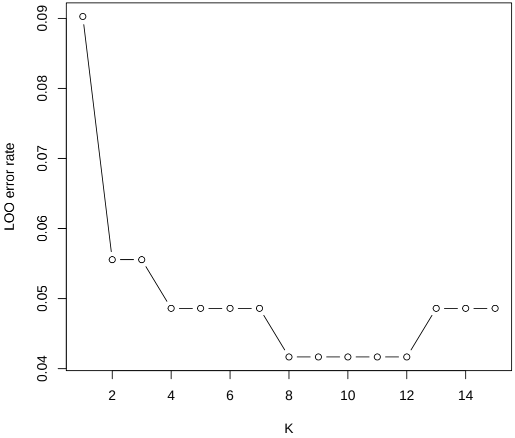
Face data

Projection in a 5D subspace (LDA) → Logistic Regression → decision



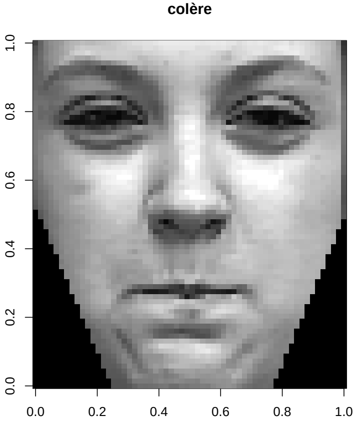
- 216 images 70 × 60 (36 per expression)
- 144 for learning, 72 for testing
- 5 features extracted by linear discriminant analysis

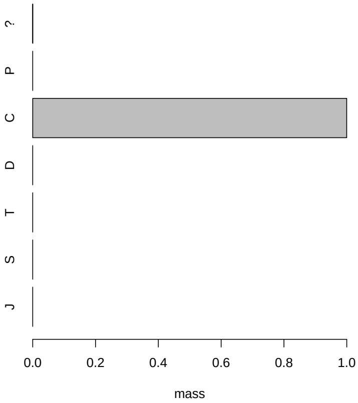
Face data: training



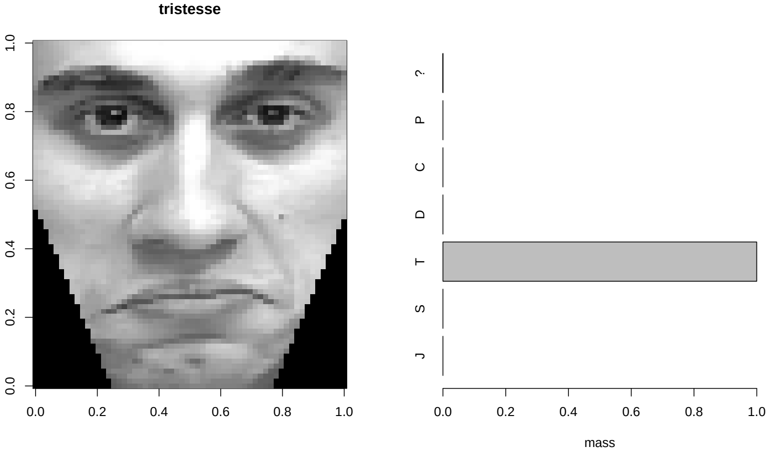

```
> print(val$err)
0.1527778
> table(ytst,val$ypred)

ytst 1 2 3 4 5 6
1 10 0 0 0 0 0
2 0 14 0 0 0 0
3 0 0 11 0 4 0
4 0 1 1 7 0 0
5 0 0 0 0 11 0
6 2 0 1 0 2 8
```

Results

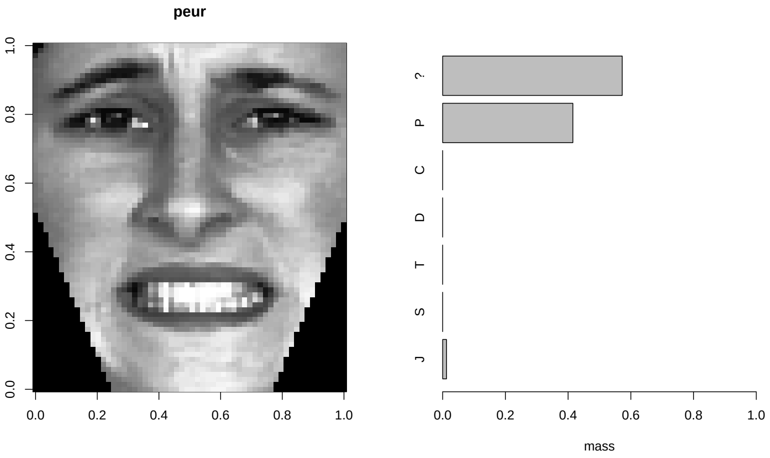




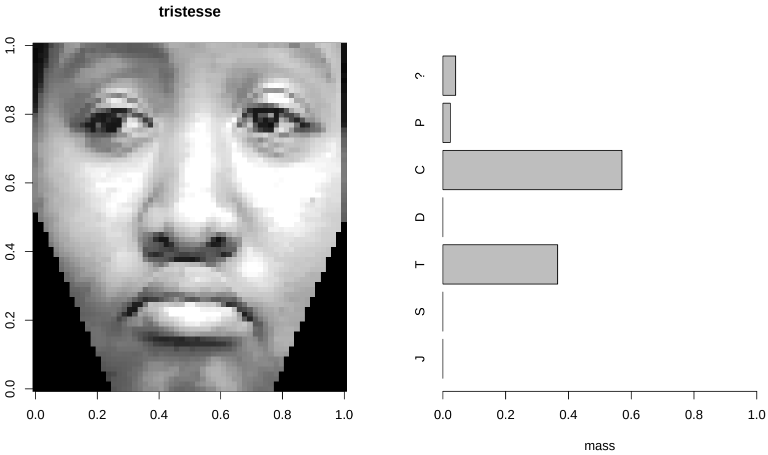
Results



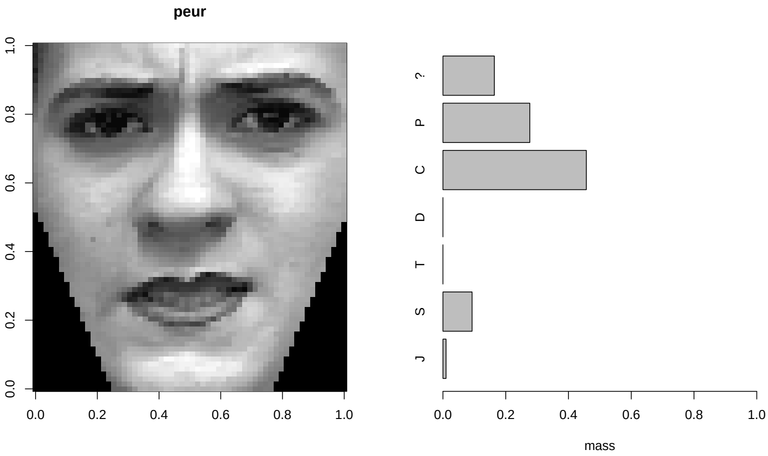
Results



Results



Results



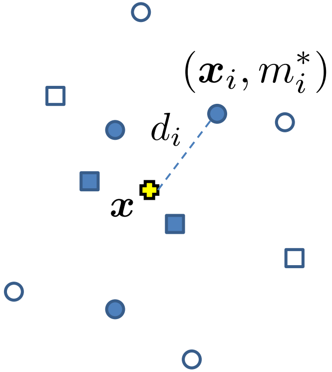
Data with soft labels

- 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 (soft label)
- Special cases:
  - $m_i^*(\{\omega_k\}) = 1$ for all i: supervised learning
  - $m_i^*(\Omega) = 1$ for all i: unsupervised learning
  - general case: partially supervised learning

Evidential k-NN rule with soft labels



- Each mass function m_i^* is discounted with a rate depending on the distance d_i

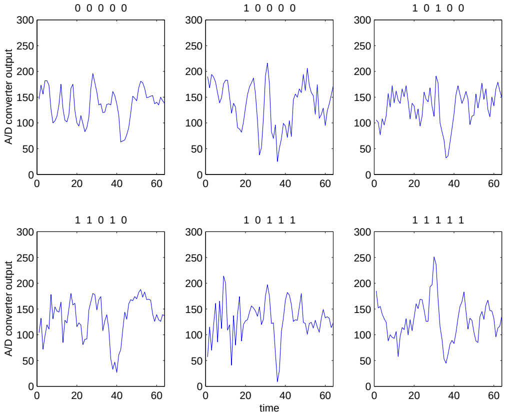
$$m_i(A) = \varphi(d_i)\, m_i^*(A), \quad \forall A \subset \Omega$$
$$m_i(\Omega) = 1 - \sum_{A \subset \Omega} m_i^*(A)$$

- The K mass functions m_i are combined using Dempster's rule

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

Example: EEG data

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



Results on EEG data

(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	weighted K-NN	EK-NN (crisp labels)	EK-NN (soft 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

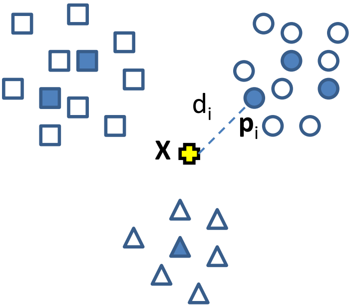
Overview

1. Theory of belief functions
   - Representation of evidence
   - Dempster's rule

2. Evidential classification
   - Evidential K-NN classifier
   - Evidential neural network classifier

3. Evidential clustering
   - Evidential clustering
   - ECM
   - EVCLUS

Principle



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

Propagation equations

- Mass function induced by prototype $\mathbf{p}_i$:

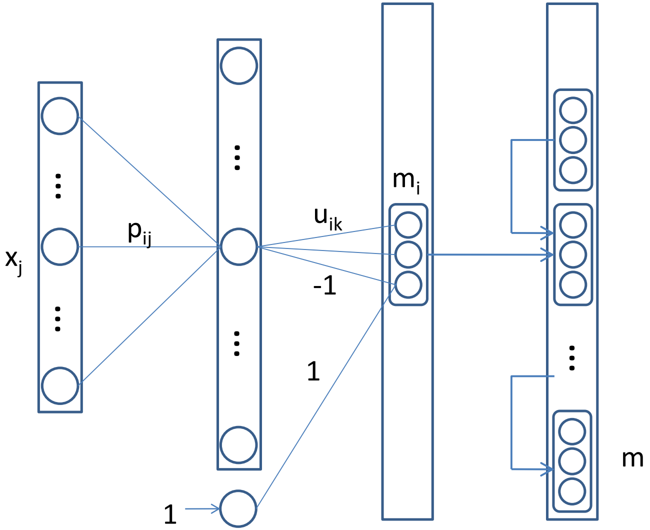
$$
\begin{aligned}
m_i(\{\omega_k\}) &= \alpha_i u_{ik} \exp(-\gamma_i d_i^2), \quad k = 1, \ldots, c \\
m_i(\Omega) &= 1 - \alpha_i \exp(-\gamma_i d_i^2)
\end{aligned}
$$

- Combination:

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

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

Neural network implementation



Learning

- The parameters are the
  - The prototypes $\mathbf{p}_i$, $i = 1, \ldots, r$ (rp parameters)
  - The membership degrees u_{ik}, $i = 1, \ldots, r$, $k = 1 \ldots, c$ (rc parameters)
  - The α_i and γ_i, $i = 1 \ldots, r$ ($2r$ parameters).
- Let $\boldsymbol{\theta}$ denote the vector of all parameters. It can be estimated by minimizing a loss function such as

$$J(\boldsymbol{\theta}) = \underbrace{\sum_{i=1}^{n} \sum_{k=1}^{c} (pl_{ik} - y_{ik})^2}_{\text{error}} + \mu \underbrace{\sum_{i=1}^{r} \alpha_i}_{\text{regularization}}$$

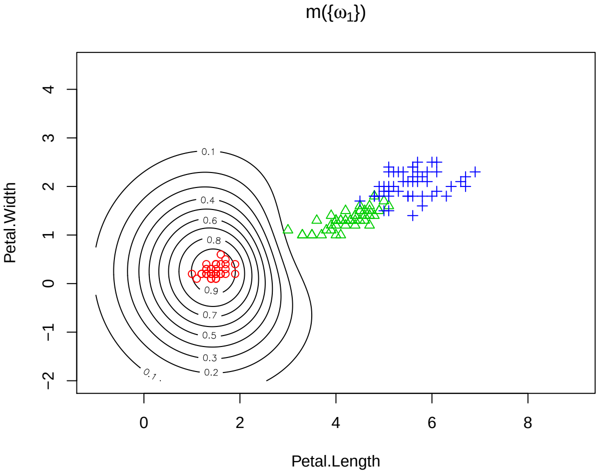
  where pl_{ik} is the output plausibility for instance i and class k, and μ is a regularization coefficient (hyperparameter).
- The hyperparameter μ can be optimized by cross-validation.

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
```

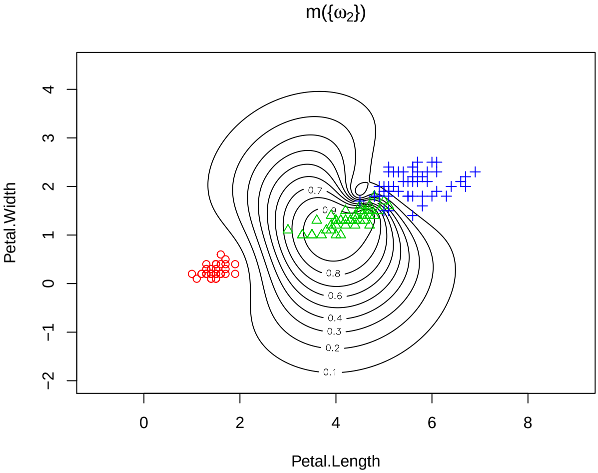
Results on the Iris data

Mass on $\{\omega_1\}$



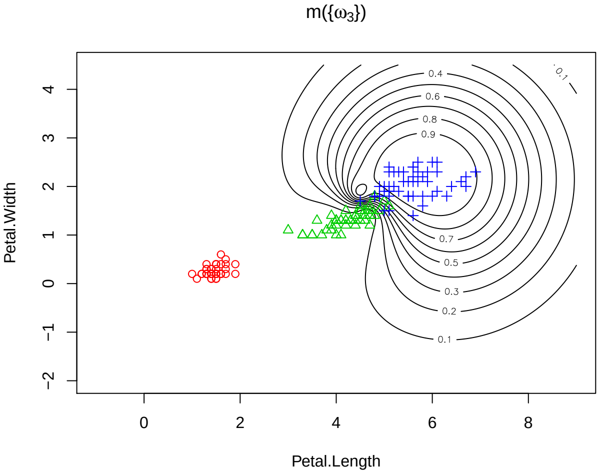
Results on the Iris data

Mass on $\{\omega_2\}$



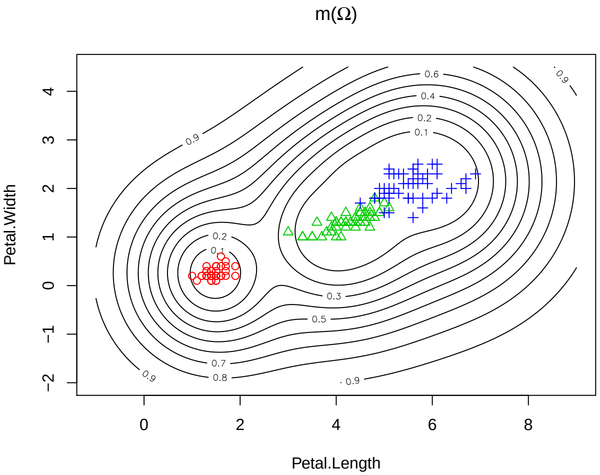
Results on the Iris data

Mass on $\{\omega_3\}$



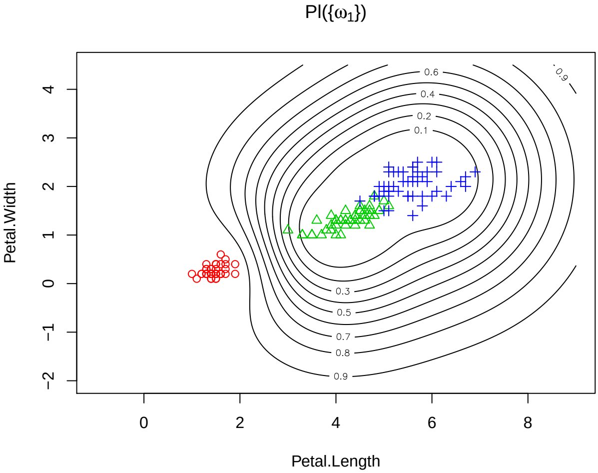
Results on the Iris data

Mass on Ω



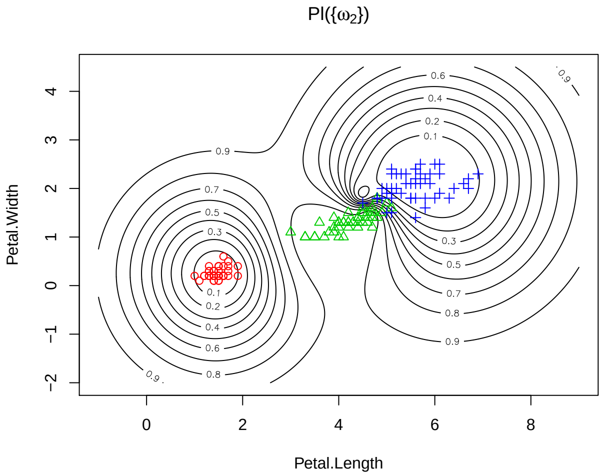
Results on the Iris data

Plausibility of $\{\omega_1\}$



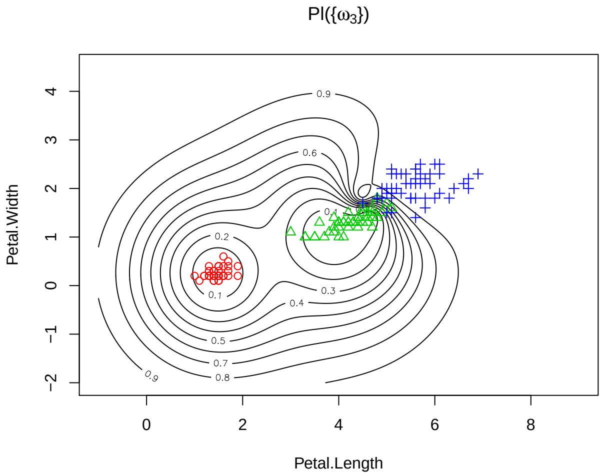
Results on the Iris data

Plausibility of $\{\omega_2\}$



Results on the Iris data

Plausibility of $\{\omega_3\}$



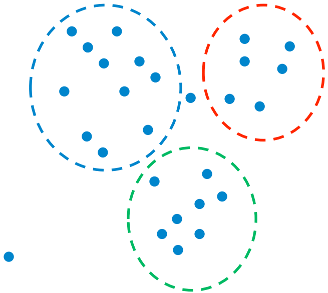
Overview

1 Theory of belief functions
- Representation of evidence
- Dempster's rule

2 Evidential classification
- Evidential K-NN classifier
- Evidential neural network classifier

3 Evidential clustering
- Evidential clustering
- ECM
- EVCLUS

Evidential clustering



- n objects described by
  - Attribute vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ (attribute data) or
  - Dissimilarities (proximity data)
- Goals:
  1. Discover groups in the data
  2. Assess the uncertainty in group membership

Overview

1. Theory of belief functions
   - Representation of evidence
   - Dempster's rule
2. Evidential classification
   - Evidential K-NN classifier
   - Evidential neural network classifier
3. Evidential clustering
   - Evidential clustering
   - ECM
   - EVCLUS

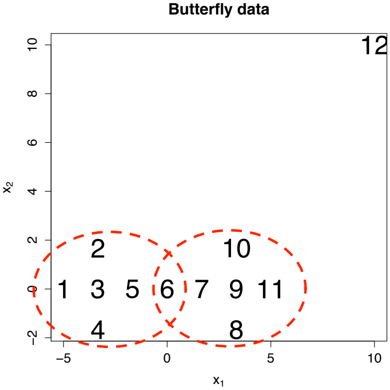
Evidential partition

- Let $\{o_1, \ldots, o_n\}$ be a set of n objects and $\Omega = \{\omega_1, \ldots, \omega_c\}$ be a set of c groups (clusters).
- Each object o_i is assumed to belong to at most one group.
- Evidence about the group membership of object o_i is represented by a mass function m_i on Ω.
- To account for the possibility that an object may not belong to any of the c groups, we use unnormalized mass functions m_i such that $m_i(\emptyset) \geq 0$.

Definition

The n-tuple $M = (m_1, \ldots, m_n)$ is called an evidential partition.

Example



Evidential partition:

	$\emptyset$	$\{\omega_1\}$	$\{\omega_2\}$	$\{\omega_1, \omega_2\}$
m_3	0	1	0	0
m_5	0	0.5	0	0.5
m_6	0	0	0	1
m_{12}	0.9	0	0.1	0

Evidential clustering algorithms

- An evidential clustering algorithm computes an evidential partition for a set of attribute or proximity data.
- There are several such algorithms. We will study two of them:
  - The Evidential c-Means (ECM) algorithm (for attribute data)
  - The EVCLUS algorithm (for attribute and proximity data)

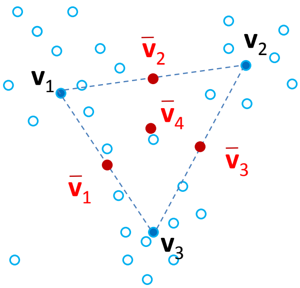
Overview

1. Theory of belief functions
   - Representation of evidence
   - Dempster's rule
2. Evidential classification
   - Evidential K-NN classifier
   - Evidential neural network classifier
3. **Evidential clustering**
   - Evidential clustering
   - **ECM**
   - EVCLUS

ECM algorithm

- The ECM algorithm is based on the representation of clusters by prototypes, and the iterative minimization of a loss function.
- It belongs to the same family of algorithms as the Hard c-Means (HCM) and the Fuzzy c-Means (FCM) algorithms.
- We start by recalling these "classical" algorithms before introducing ECM.

ECM algorithm: principle



- Each cluster ω_k is represented by a prototype $\mathbf{v}_k$.
- Each meta-cluster (=nonempty set of clusters) A_j is 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 nonempty $A_j \subseteq \Omega$, $m_{ij} = m_i(A_j)$ 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: cost function

- Define the nonempty focal sets $\mathcal{F} = \{A_1, \ldots, A_f\} \subseteq 2^\Omega \setminus \{\emptyset\}$.
- Minimize

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

  subject to the constraints $\sum_{j=1}^{f} m_{ij} + m_{i\emptyset} = 1$ for all i.
- Parameters:
  - α controls the specificity of mass functions (default: 1)
  - β controls the hardness of the evidential partition (default: 2)
  - δ controls the proportion of data considered as outliers
- $J_{\mathrm{ECM}}(M, V)$ can be iteratively minimized with respect to M and to V.

ECM algorithm: update equations

Update of M:

$$m_{ij} = \frac{c_j^{-\alpha/(\beta-1)} d_{ij}^{-2/(\beta-1)}}{\sum_{k=1}^{f} c_k^{-\alpha/(\beta-1)} d_{ik}^{-2/(\beta-1)} + \delta^{-2/(\beta-1)}},$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, f$, and

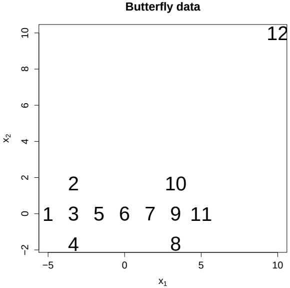
$$m_{i\emptyset} = 1 - \sum_{j=1}^{f} m_{ij}, \quad i = 1, \ldots, n$$

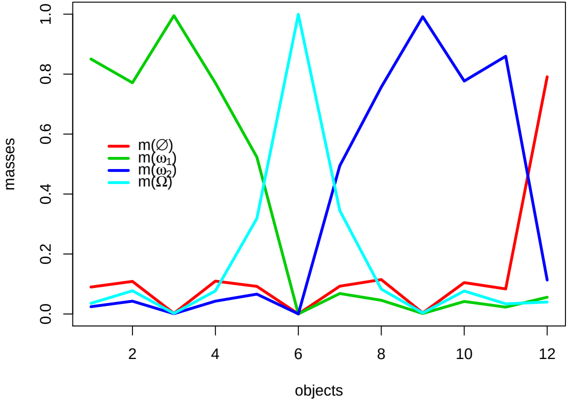
Update of V: solve a linear system of the form

$$HV = B,$$

where B is a matrix of size $c \times p$ and H a matrix of size $c \times c$.

Butterfly dataset

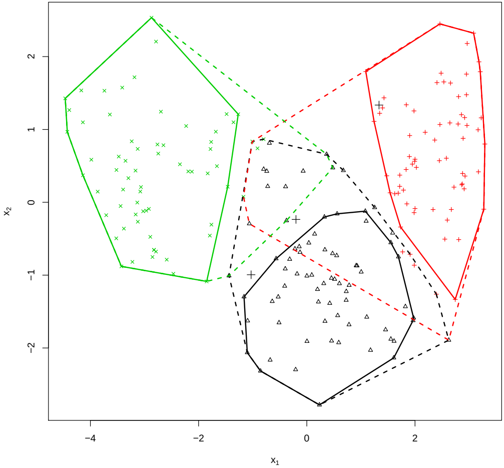




Example in R (Seeds data)

```
library(evclust)

em<-ecm(x, c,type="simple")

plot(em,z[,1:2])
```

Result



Meaning of this graph: see next slide.

Inner and outer approximations

- For each object i, let $A_i \subseteq \Omega$ such that
$$m_i(A_i) = \max_{A \subseteq \Omega} m_i(A)$$
- The inner approximation of cluster $\omega \in \Omega$ is the set of objects that surely belong to ω:
$$\underline{\omega} = \{o_i : A_i = \{\omega\}\}$$
- The outer approximation of cluster $\omega \in \Omega$ is the set of objects that possibly belong to ω:
$$\overline{\omega} = \{o_i : \omega \in A_i\}$$
- The outliers are the objects for which $A_i = \emptyset$ (they do not belong to any outer approximation).

Overview

1. Theory of belief functions
   - Representation of evidence
   - Dempster's rule
2. Evidential classification
   - Evidential K-NN classifier
   - Evidential neural network classifier
3. Evidential clustering
   - Evidential clustering
   - ECM
   - EVCLUS

Learning an evidential partition from proximity data

- Problem: given the dissimilarity matrix $D = (d_{ij})$, how to build a "reasonable" evidential partition ?
- We need a model that relates cluster membership to dissimilarities.
- Basic idea: "The more similar two objects, the more plausible it is that they belong to the same group".
- How to formalize this idea?

Formalization

- Let m_i and m_j be mass functions regarding the group membership of objects o_i and o_j.
- We can show that the plausibility that objects o_i and o_j belong to the same group is

$$pl_{ij}(S_{ij}) = \sum_{A \cap B \neq \emptyset} m_i(A)m_j(B) = 1 - \kappa_{ij}$$

  where κ_{ij} = degree of conflict between m_i and m_j.
- Problem: find an evidential partition $M = (m_1, \ldots, m_n)$ such that larger degrees of conflict κ_{ij} correspond to larger dissimilarities d_{ij}.

Cost function

- Approach: minimize the discrepancy between the dissimilarities d_{ij} and the degrees of conflict κ_{ij}.
- Example of a cost (stress) function:

$$J(M) = \sum_{i<j} (\kappa_{ij} - \varphi(d_{ij}))^2$$

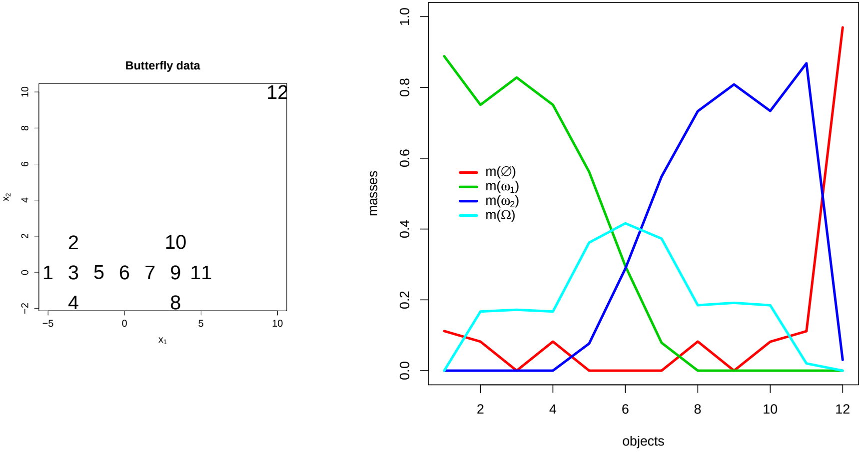
where φ is an increasing function from $[0, +\infty)$ to $[0, 1]$, for instance

$$\varphi(d) = 1 - \exp(-\gamma d^2),$$

where γ is a scaling coefficient.

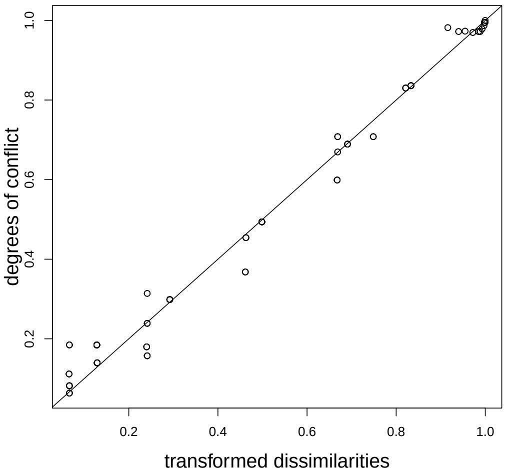
Butterfly example

evidential partition



Butterfly example

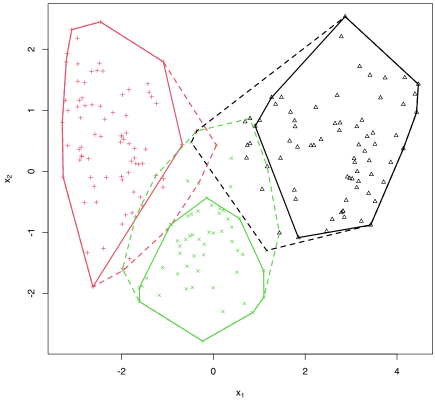
Shepard diagram

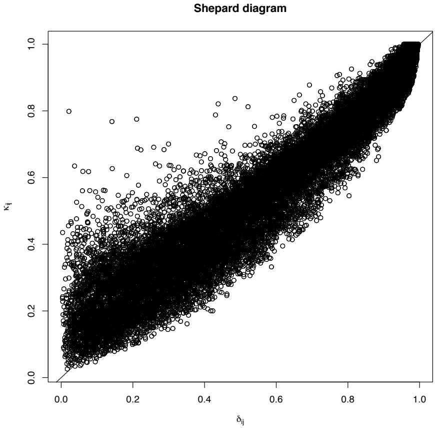


Example in R (Seeds data)

```
library(evclust)

clus<-kevclus(x,c=3)

plot(clus,z[,1:2])
```

Result





Advantages of EVCLUS

- Conceptually simple, clear interpretation.
- EVCLUS can handle nonmetric dissimilarity data (even expressed on an ordinal scale).
- It was also shown to outperform some of the state-of-the-art clustering techniques on proximity datasets.

Summary

- The theory of belief functions makes it possible to implement "cautious" approaches to classification and clustering that provide faithful representations of prediction uncertainty.
- The techniques presented in this chapter belong to an emerging field of Evidential Machine Learning.
- This field is still largely uncharted, which makes it a research topic of choice for Master and PhD students!