A Fresh Look at some Machine Learning Techniques from the Perspective of Dempster-Shafer Theory

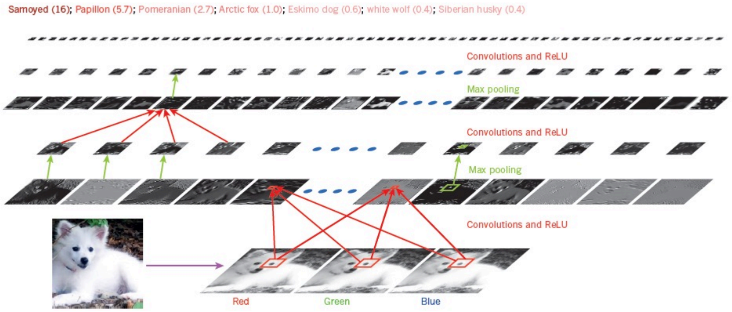
Thierry Denœux

Compiègne University of Technology
HEUDIASYC (UMR CNRS 7253)

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

CGCKD 2018, Chengdu, China
August 11, 2018

Machine Learning



(From Le Cun et al., *Nature*, 2015)

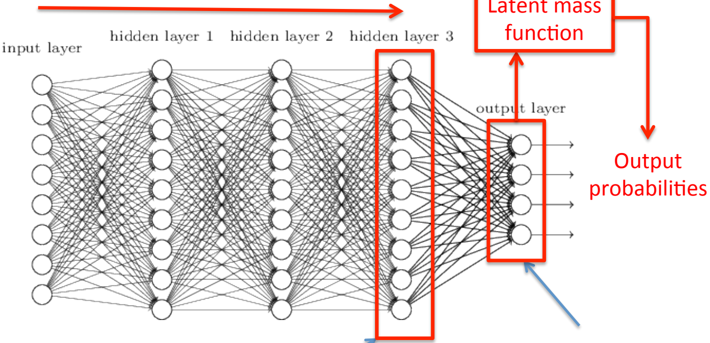
- In recent years, applications of Machine Learning (ML) have been flourishing following new developments in deep learning technology.
- A lot of progress has been made in extracting high-order features from data, so as to solve very complex classification problems.

Making Machine Learning more Transparent

- ML algorithms (and especially deep learning models) are essentially black boxes.
- Major challenge: make ML algorithms more transparent so that machine predictions can be interpreted (and trusted) by humans.
- To meet this challenge, we need new perspectives on how classification algorithms actually work.
- One such perspective is provided by the Dempster-Shafer (DS) theory of evidence.

The DS perspective

Construction of more and more informative features



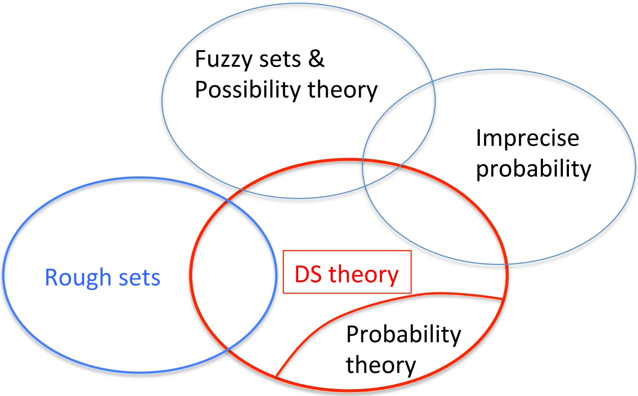
Highest-order features are pieces of evidence representable by DS mass functions

In the decision layer, pieces of evidence are combined using Dempster's rule

Outline

Outline

Uncertainty theories



Dempster-Shafer (DS) theory

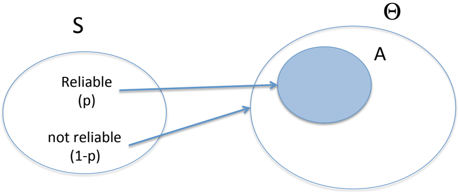
- Also referred to as evidence theory, theory of belief functions
- A formal framework for reasoning with partial (uncertain, imprecise) information.
- Originates from Arthur Dempster's seminal work of statistical inference in the late 1960's
- Formalized by Glenn Shafer in his seminal 1976 book
- Has been applied in may areas: statistical inference, knowledge representation, information fusion, machine learning, etc.

General philosophy

- We consider some question with (unknown) answer Y.
- We collect evidence about Y (measurements, expert opinions, observations, etc.)
- Each piece of evidence is modeled by a mass function.
- The mass functions are combined using Dempster's rule of combination.

Outline

Simple Mass Function



- Let Θ be the set of possible answers to some question (frame of discernment), Y the true answer.
- A source of information (sensor, expert, etc.) tells us that $Y \in A$, for some subset $A \subseteq \Theta$.
- There is probability p that the source is reliable.
- Representation: $m(A) = p$, $m(\Theta) = 1 - p$, $m(B) = 0$ for all other B.
- Meaning: with probability p we know that $Y \in A$, and with probability $1 - p$ we know nothing.

Mass Function

General Definition

Definition

A *mass function* *is a mapping* $m : 2^{\Theta} \rightarrow [0, 1]$ *such that*

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

and

$$m(\emptyset) = 0$$

- Every subset A of Θ such that $m(A) > 0$ is a focal set.
- Interpretation: $m(A)$ is the probability of knowing only that $Y \in A$, and nothing more specific.
- A simple mass function has at most two focal sets, one of which is Θ.

Belief and plausibility functions

Definition

Given a mass function m on Θ, the belief and plausibility functions are defined, respectively, as

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

$$Pl(A) := \sum_{B \cap A \neq \emptyset} m(B) = 1 - Bel(\overline{A}),$$

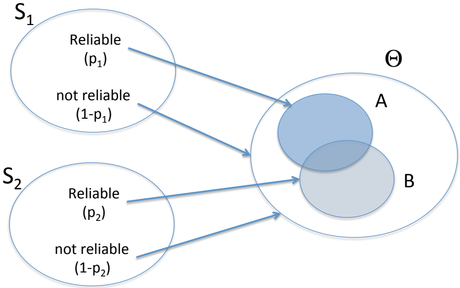
for all $A \subseteq \Theta$

- Interpretation:
  - $Bel(A)$ is a measure of the support in A
  - $Pl(A)$ is a measure of the lack of support in $\overline{A}$.
- Total ignorance: $Bel(A) = 0$ for all $A \neq \Theta$ and $Pl(A) = 1$ for all $A \neq \emptyset$.

Outline

Combining Mass Functions

Two independent sources:



What do we know?

		S_2 reliable $[p_2]$	S_2 not reliable $[1-p_2]$
S_1	reliable $[p_1]$	$A \cap B$ $[p_1 p_2]$	A $[p_1(1-p_2)]$
	not reliable $[1-p_1]$	B $[p_2(1-p_1)]$	Θ $[(1-p_1)(1-p_2)]$

Dempster's rule

Definition (Dempster's rule)

Let m_1 and m_2 be two mass functions. Their orthogonal sum is the mass function defined by

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

and $(m_1 \oplus m_2)(\emptyset) = 0$, where κ is the degree of conflict defined as

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

Remark: $m_1 \oplus m_2$ exists iff $\kappa < 1$.

Dempster's rule

Properties

Proposition

1. *The operator* $\oplus$ *is commutative, associative.*
2. *Let* $m_?$ *be the vacuous mass function* $m_?$ *defined by* $m_?(\Theta) = 1$*. For all mass function* m, $m \oplus m_? = m_? \oplus m = m$.

Weights of evidence

Dempster's rule can often be easily computed by adding weights of evidence.

Definition (Weight of evidence)

Given a simple mass function of the form

$$m(A) = s$$
$$m(\Theta) = 1 - s,$$

the quantity $w = -\ln(1-s)$ *is called the weight of evidence for* A*. Mass function* m *is denoted by* A^w.

Proposition

The orthogonal sum of two simple mass functions A^{w_1} *and* A^{w_2} *is*

$$A^{w_1} \oplus A^{w_2} = A^{w_1+w_2}$$

Outline

Outline

Binomial Logistic regression

- Consider a binary classification problem with d-dimensional feature vector $X = (X_1, \ldots, X_d)$ and class variable $Y \in \Theta = \{\theta_1, \theta_2\}$. Let $p(x)$ denote the probability that $Y = \theta_1$ given that $X = x$.
- (Binomial) Logistic Regression (LR) model:

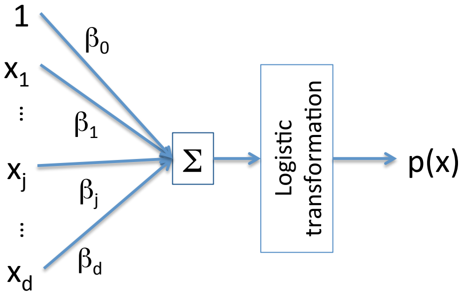
$$\ln \frac{p(x)}{1 - p(x)} = \boldsymbol{\beta}^T x + \beta_0,$$

with $\boldsymbol{\beta} \in \mathbb{R}^d$ and $\beta_0 \in \mathbb{R}$. Equivalently,

$$p(x) = \sigma(\boldsymbol{\beta}^T x + \beta_0),$$

where $\sigma(u) = (1 + \exp(-u))^{-1}$ is the logistic function.

Binomial Logistic Regression (continued)



Given a learning set $\{(x_i, y_i)\}_{i=1}^n$, parameters β and β_0 are usually estimated by minimizing the cross-entropy error function:

$$C(\beta, \beta_0) = -\sum_{i=1}^{n} \{I(y_i = \theta_1) \ln p(x_i) + I(y_i = \theta_2) \ln [1 - p(x_i)]\}$$

Multinomial Logistic Regression

- Multinomial logistic regression (MLR) extends binomial LR to $K > 2$ by assuming the following model:

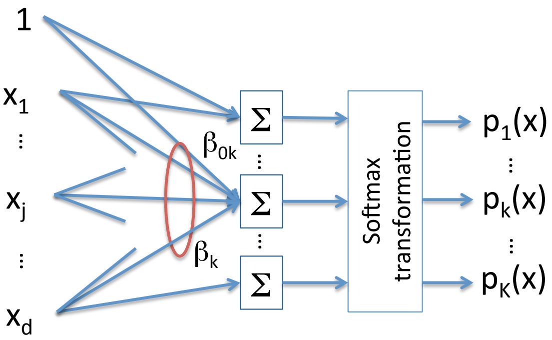
$$\ln p_k(x) = \boldsymbol{\beta}_k^T x + \beta_{k0} + \gamma,$$

where $p_k(x) = \mathbb{P}(Y = \theta_k | X = x)$, $\boldsymbol{\beta}_k \in \mathbb{R}^d$, $\beta_{k0} \in \mathbb{R}$ and $\gamma \in \mathbb{R}$ is a constant that does not depend on k.

- The posterior probability of class θ_k can then be expressed using the softmax transformation as

$$p_k(x) = \frac{\exp(\boldsymbol{\beta}_k^T x + \beta_{k0})}{\sum_{l=1}^K \exp(\boldsymbol{\beta}_l^T x + \beta_{l0})}.$$

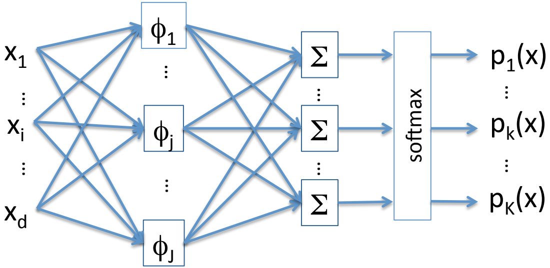
Multinomial Logistic Regression (continued)



Parameters $(\boldsymbol{\beta}_k, \beta_{k0})$, $k = 1 \ldots, K$ can be estimated by minimizing the cross-entropy as in the binomial case.

Outline

Nonlinear generalized LR classifiers



- LR can be applied to transformed features $\phi_j(x)$, $j = 1, \ldots, J$, where the ϕ_j's are nonlinear mappings from $\mathbb{R}^d$ to $\mathbb{R}$. We get nonlinear generalized LR classifiers.
- Both the new features $\phi_j(x)$ and the coefficients (β_k, β_{k0}) are usually learnt simultaneously by minimizing some cost function.

Generalized LR models

- Generalized additive models:

$$\phi_j(x) = \varphi_j(x_j)$$

- Radial basis function networks

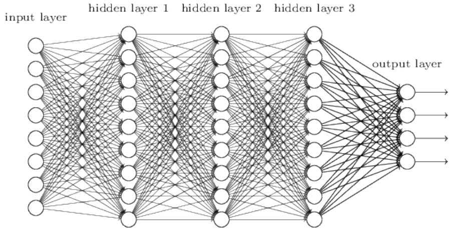
$$\phi_j(x) = \varphi(\|x - v_j\|)$$

- Support vector machines

$$\phi_j(x) = \mathcal{K}(x, x_j)$$

- Multilayer feedforward neural networks (NNs)

Multilayer feedforward neural networks



- Feedforward NNs are models composed of elementary computing units (or "neurons") arranged in layers. Each layer computes a vector of new features as functions of the outputs from the previous layer as

$$\phi_j^{(l)} = h\left(w_j^{(l)T}\phi^{(l-1)} + w_{j0}^{(l)}\right), \quad j = 1, \ldots, J_l,$$

where $\phi^{(l-1)} \in \mathbb{R}^{J_{l-1}}$ is the vector of outputs from the previous layer.
- For classification, the output layer is typically a softmax layer with K output units.

Relation with DS theory?

- LR and NN models seem totally unrelated to DS theory.
- Yet...

Outline

Outline

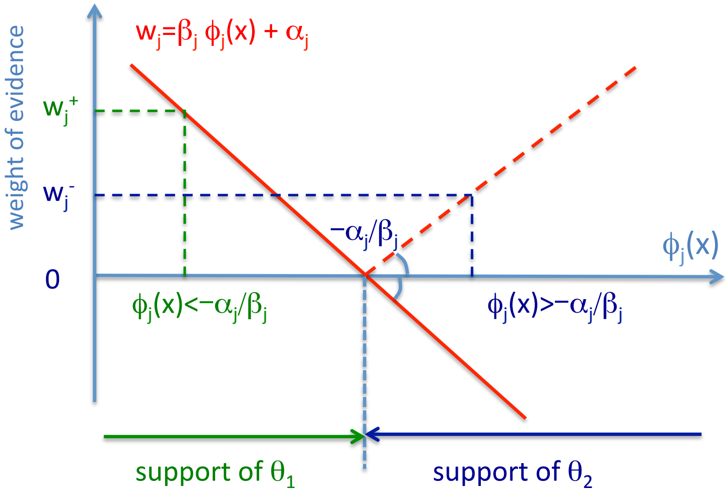
Feature values as evidence

- Consider a binary classification problem with $K = 2$ classes in $\Theta = \{\theta_1, \theta_2\}$. Let $\phi(x) = (\phi_1(x), \ldots, \phi_J(x))$ be a vector of J features.
- Each feature value $\phi_j(x)$ is a piece of evidence about the class $Y \in \Theta$ of the instance under consideration.
- Assume that this evidence points to θ_1 or θ_2 depending on the sign of

$$w_j := \beta_j \phi_j(x) + \alpha_j,$$

  where β_j and α_j are two coefficients:
  - If $w_j \geq 0$, feature ϕ_j supports class θ_1 with weight of evidence w_j
  - If $w_j < 0$, feature ϕ_j supports class θ_2 with weight of evidence $-w_j$

Feature values as evidence (continued)



Feature-based latent mass function

Under this model, the consideration of feature ϕ_j induces a feature-based latent mass function

$$m_j = \{\theta_1\}^{w_j^+} \oplus \{\theta_2\}^{w_j^-},$$

where

- $w_j^+ = \max(0, w_j)$ is the positive part of w_j and
- $w_j^- = \max(0, -w_j)$ is the negative part.

Combined latent mass function

Assuming that the values of the J features can be considered as independent pieces of evidence, the feature-based latent mass functions can be combined by Dempster's rule:

$$
\begin{aligned}
m &= \bigoplus_{j=1}^{J} \left(\{\theta_1\}^{w_j^+} \oplus \{\theta_2\}^{w_j^-} \right) \\
&= \left(\bigoplus_{j=1}^{J} \{\theta_1\}^{w_j^+} \right) \oplus \left(\bigoplus_{j=1}^{J} \{\theta_2\}^{w_j^-} \right) \\
&= \{\theta_1\}^{w^+} \oplus \{\theta_2\}^{w^-},
\end{aligned}
$$

where

- $w^+ := \sum_{j=1}^{J} w_j^+$ is the total weight of evidence supporting θ_1
- $w^- := \sum_{j=1}^{J} w_j^-$ is the total weight of evidence supporting θ_2.

Expression of m

$$m(\{\theta_1\}) = \frac{[1-\exp(-w^+)]\exp(-w^-)}{1-\kappa}$$

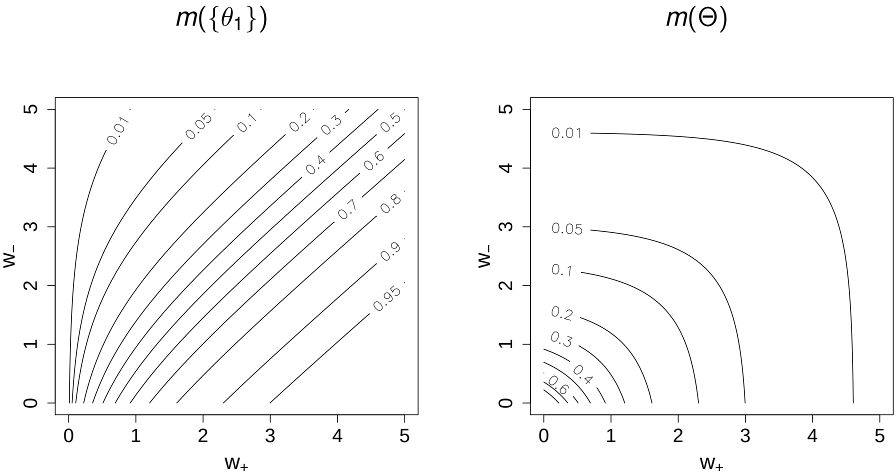
$$m(\{\theta_2\}) = \frac{[1-\exp(-w^-)]\exp(-w^+)}{1-\kappa}$$

$$m(\Theta) = \frac{\exp(-w^+ - w^-)}{1-\kappa}$$

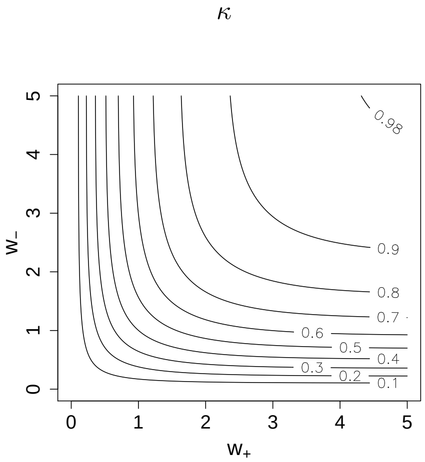
where κ is the degree of conflict:

$$\kappa = [1-\exp(-w^+)][1-\exp(-w^-)]$$

$m(\{\theta_1\})$ and $m(\Theta)$ vs. weights of evidence



Degree of conflict vs. weights of evidence



Normalized plausibilities

The normalized plausibility of class θ_1 as

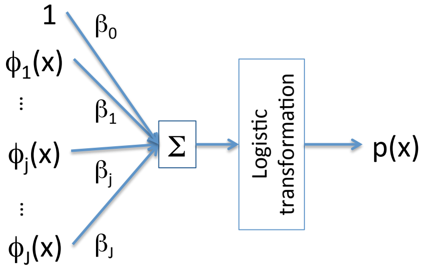
$$\frac{Pl(\{\theta_1\})}{Pl(\{\theta_1\}) + Pl(\{\theta_2\})} = \frac{m(\{\theta_1\}) + m(\Theta)}{m(\{\theta_1\}) + m(\{\theta_2\}) + 2m(\Theta)}$$
$$= \frac{1}{1 + \exp[-(\boldsymbol{\beta}^T \phi(x) + \beta_0)]}$$
$$= p(x)$$

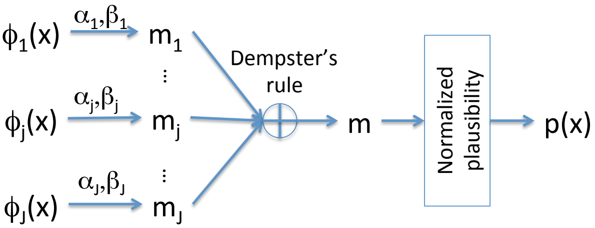
with $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_J)$ and $\beta_0 = \sum_{i=1}^J \alpha_j$.

Proposition

The normalized plausibilities are equal to the posterior class probabilities of the binomial LR model: the two models are equivalent.

Two Views of Binomial Logistic Regression





Parameter identification

- As explained before, parameters $\beta_0, \beta_1, \ldots, \beta_J$ can be estimated by maximizing the likelihood. Let $\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_J$ be the corresponding MLEs.
- However, the DS model has J more additional parameters $\alpha_1, \ldots, \alpha_J$ linked to β_0 by the relation $\sum_{j=1}^J \alpha_j = \beta_0$: the problem is underdetermined.
- Solution: find the parameter values $\alpha_1^*, \ldots, \alpha_J^*$ that will give us the least informative mass function.
- The least informative mass function is defined as the one based on the smallest weights of evidence.

Minimizing the sum of squared weights of evidence

- Let $\{(x_i, y_i)\}_{i=1}^n$ be the learning set and let $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_J)$.
- The values α_j^* minimizing the sum of squared weights of evidence can be found by solving the following minimization problem:

$$\min f(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \sum_{j=1}^{J} \left(\widehat{\beta}_j \phi_j(x_i) + \alpha_j \right)^2$$

subject to $\sum_{j=1}^J \alpha_j = \widehat{\beta}_0$.

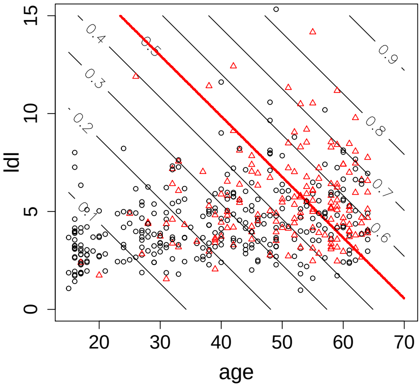
- Solution:

$$\alpha_j^* = \frac{\widehat{\beta}_0}{J} + \frac{1}{J} \sum_{q=1}^{J} \widehat{\beta}_q \mu_q - \widehat{\beta}_j \mu_j$$

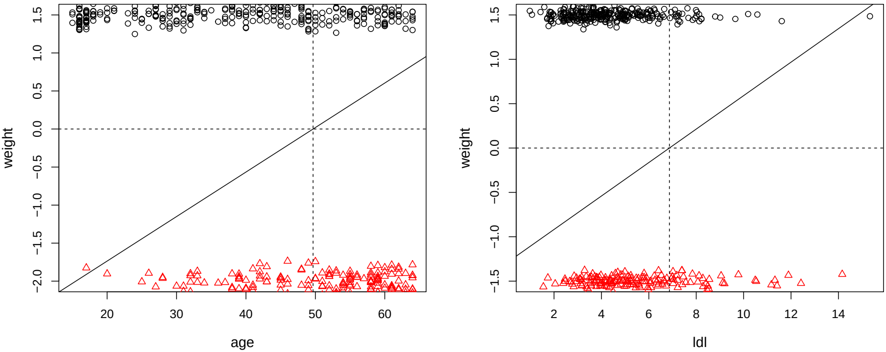
with $\mu_j = \frac{1}{n} \phi_j(x_i)$.

Example

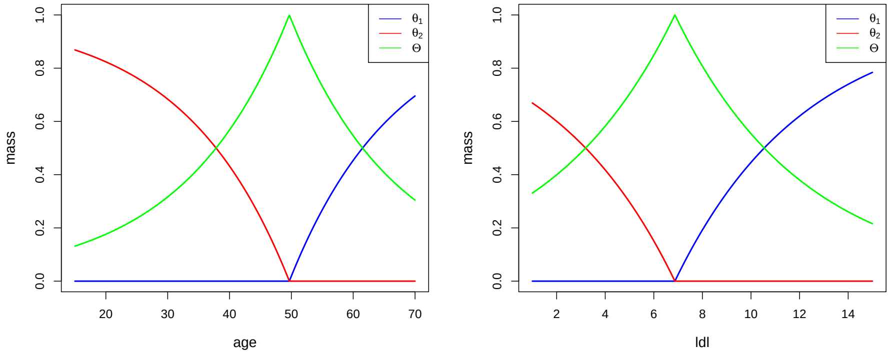
- Data about the intensity of ischemic heart disease risk factors in a rural area of South Africa. Population: white males between 15 and 64. Response variable: presence or absence of myocardial infarction (MI).
- Two variables: age and LDL (“bad” cholesterol).



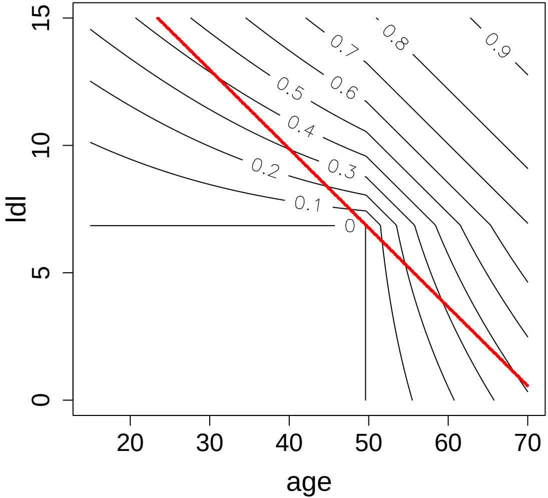
Weights of evidence



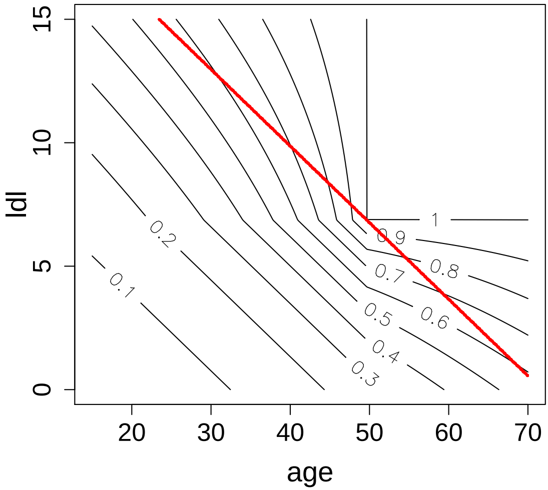
Feature mass functions



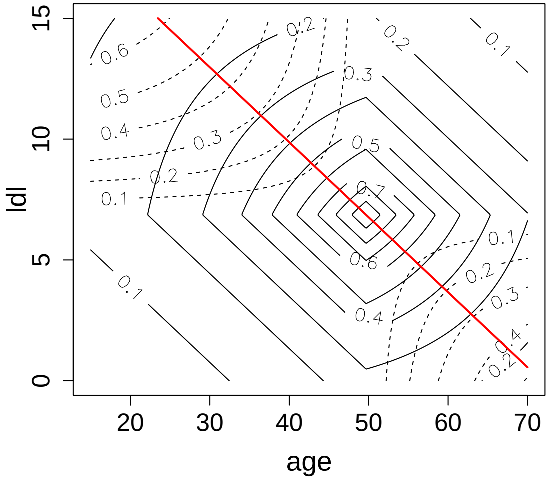
Degrees of belief (positive class)



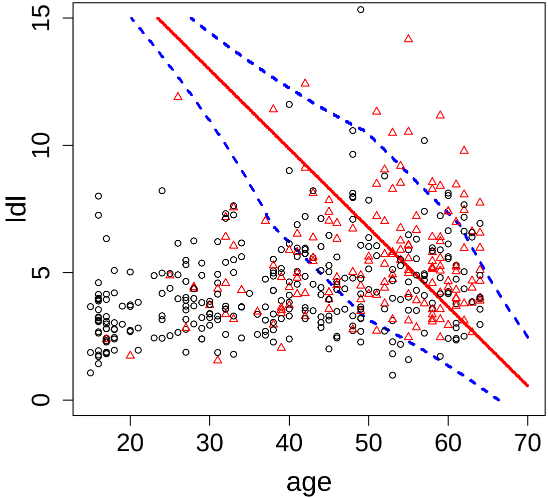
Degrees of Plausibility (positive class)



Mass on Θ and degree of conflict



Decision regions



Outline

Model

- Let $\Theta = \{\theta_1, \ldots, \theta_K\}$ with $K > 2$.
- Each feature ϕ_j now induces K mass functions $m_{j1}, \ldots, m_{jK}$.
- Mass function m_{jk} points either to the singleton $\{\theta_k\}$ or to its complement $\overline{\{\theta_k\}}$, depending on the sign of

$$w_{jk} = \beta_{jk}\phi_j(x) + \alpha_{jk},$$

  where $(\beta_{jk}, \alpha_{jk})$, $k = 1, \ldots, K$, $j = 1, \ldots, J$ are parameters.
- Expression of m_{jk}:

$$m_{jk} = \{\theta_k\}^{w_{jk}^+} \oplus \overline{\{\theta_k\}}^{w_{jk}^-}$$

- The latent mass function induced by feature ϕ_j is

$$m_j = \bigoplus_{k=1}^{K} \left(\{\theta_k\}^{w_{jk}^+} \oplus \overline{\{\theta_k\}}^{w_{jk}^-} \right).$$

Combined latent mass function

- We thus have JK elementary mass functions $m_{jk} = \{\theta_k\}^{w_{jk}^+} \oplus \overline{\{\theta_k\}}^{w_{jk}^-}$.
- The combined mass function can be written as

$$
\begin{aligned}
m &= \bigoplus_{j=1}^{J} \bigoplus_{k=1}^{K} \left(\{\theta_k\}^{w_k^+} \oplus \overline{\{\theta_k\}}^{w_k^-} \right) \\
&= \bigoplus_{k=1}^{K} \left(\{\theta_k\}^{w_k^+} \oplus \overline{\{\theta_k\}}^{w_k^-} \right),
\end{aligned}
$$

where

- $w_k^+ = \sum_{j=1}^J w_{jk}^+$ is the total weight of evidence for class θ_k
- $w_k^- = \sum_{j=1}^J w_{jk}^-$ is the total weight of evidence against class θ_k

Link with multinomial logistic regression

The normalized plausibility of class θ_k is:

$$\frac{Pl(\{\theta_k\})}{\sum_{l=1}^{K} Pl(\{\theta_l\})} = \frac{\exp\left(\sum_{j=1}^{J} \beta_{jk}\phi_j(x) + \beta_{0k}\right)}{\sum_{l=1}^{K} \exp\left(\sum_{j=1}^{J} \beta_{jl}\phi_j(x) + \beta_{0l}\right)}$$

$$p_k(x)$$

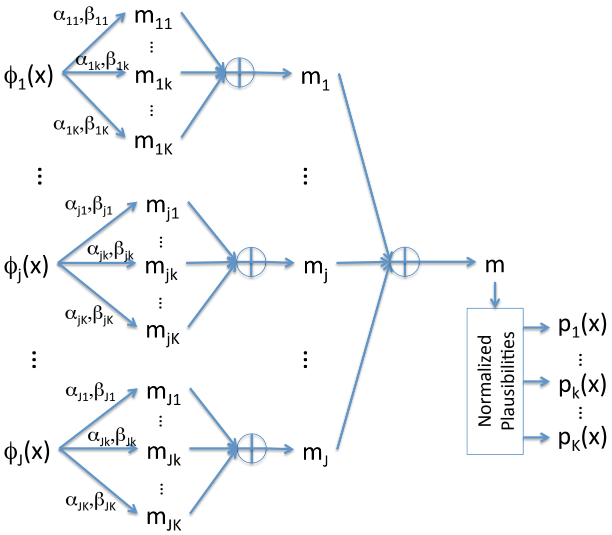
with

$$\beta_{0k} = \sum_{j=1}^{J} \alpha_{jk}.$$

Proposition

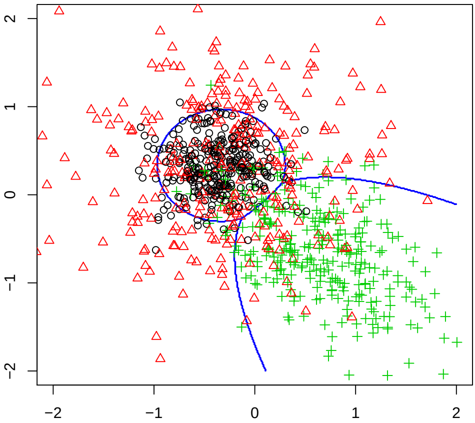
The normalized plausibilities are equal to the posterior class probabilities of the multinomial LR model: the two models are equivalent.

Multinomial Logistic Regression: DS view



Example

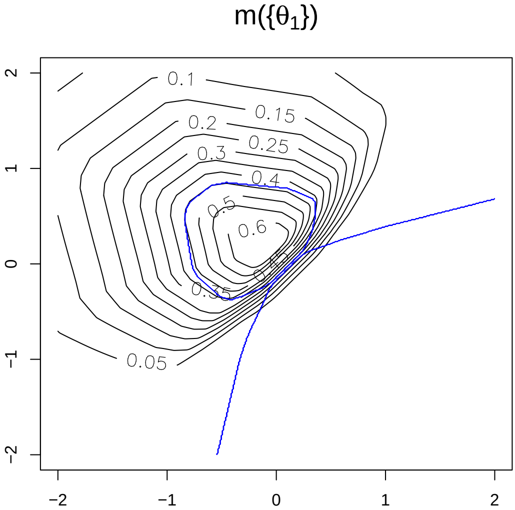
Dataset: 900 instances, 3 equiprobable classes with Gaussian distributions



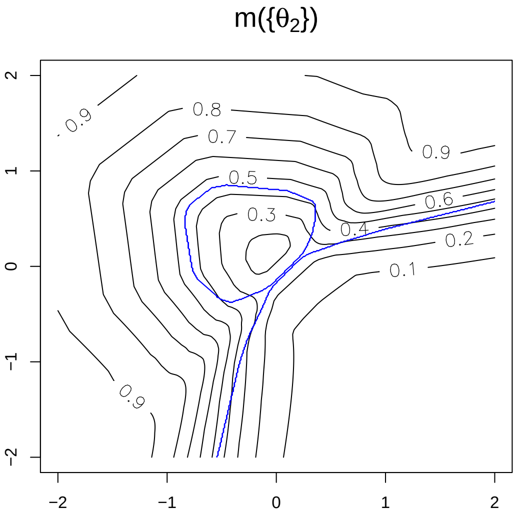
NN model

- NN with 2 layers of 20 and 10 neurons
- ReLU activation functions in hidden layers, softmax output layer
- Batch learning, minibatch size=100
- L_2 regularization in the last layer ($\lambda = 1$).

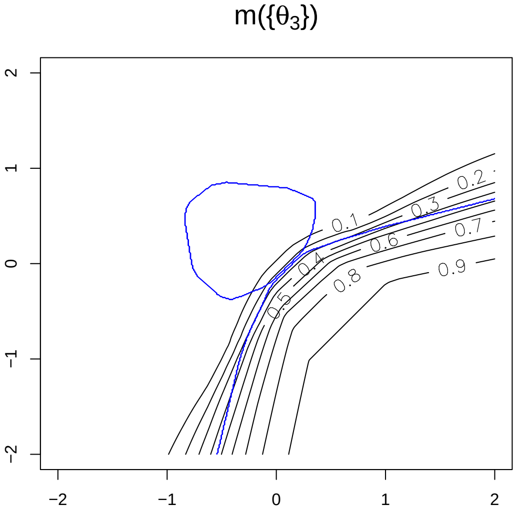
Mass on $\{\theta_1\}$



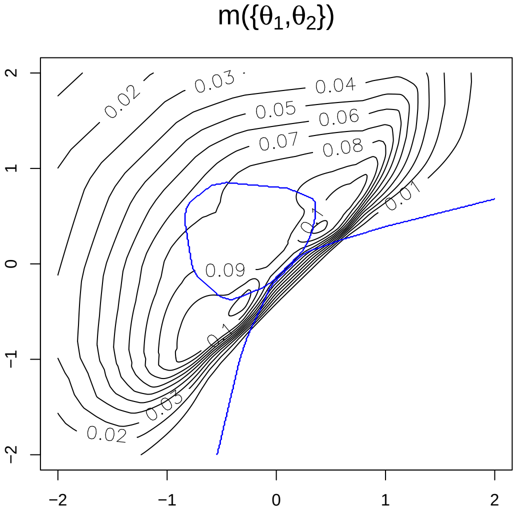
Mass on $\{\theta_2\}$



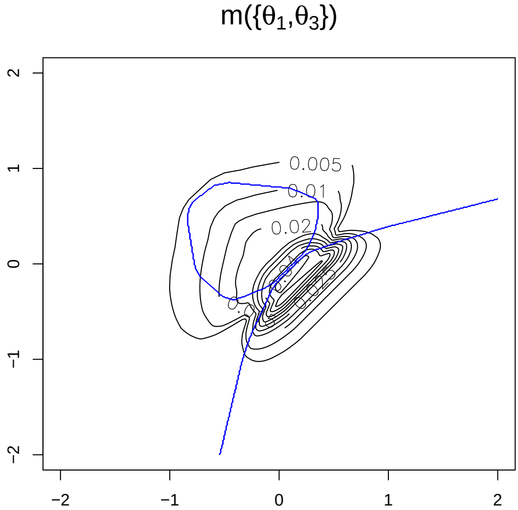
Mass on $\{\theta_3\}$



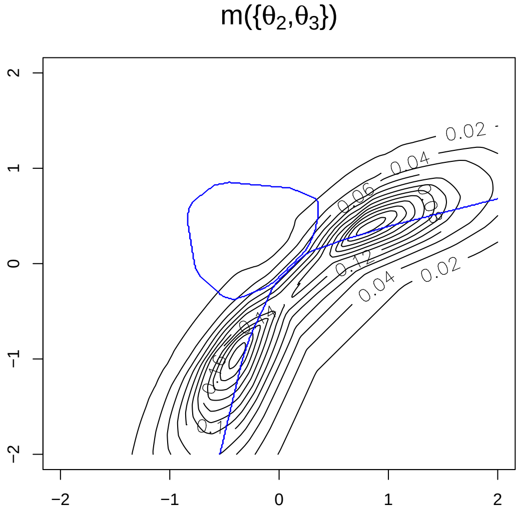
Mass on $\{\theta_1, \theta_2\}$



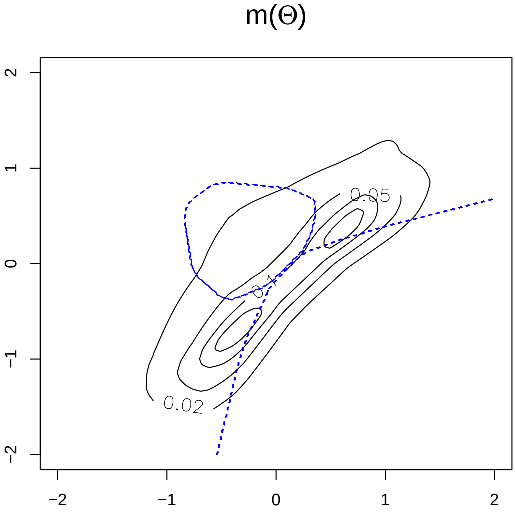
Mass on $\{\theta_1, \theta_3\}$



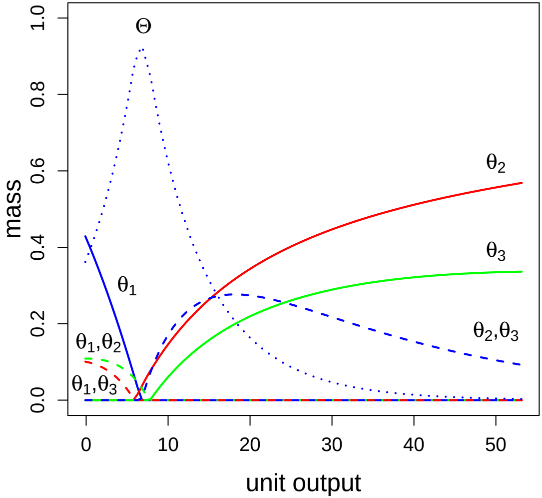
Mass on $\{\theta_2, \theta_3\}$



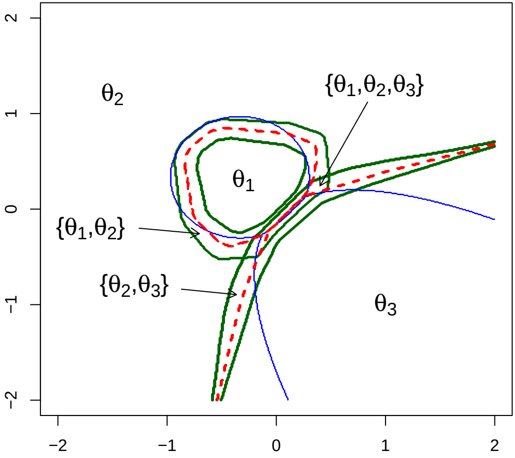
Mass on Θ



Hidden unit 2



Decision regions



Summary

- The theory of belief functions has great potential in machine learning to
  - combine classifiers
  - design specific classifiers, called evidential classifiers
- Logistic regression, neural networks, and other nonlinear classifiers such as SVMs can be viewed as evidential classifiers: they are based on
  - a model relating feature values to weights of evidence, and
  - Dempster's rule of combination.
- Viewing neural network classifiers as evidential classifiers has important implications in terms of
  - interpretation
  - decision strategies
  - classifier fusion
  - handling missing or uncertain inputs, etc.

  These implications are currently being investigated.

References

cf. https://www.hds.utc.fr/˜tdenoeux

T. Denœux.
Logistic regression revisited: belief function analysis.
5th International Conference on Belief Functions and Applications, Compiègne, France, September 2018.

T. Denœux.
Logistic Regression, Neural Networks and Dempster-Shafer Theory: a New Perspective
Preprint, arXiv:1807.01846, May 2018.