Statistical prediction using belief functions

Thierry Denœux

Université de Technologie de Compiègne
HEUDIASYC (UMR CNRS 6599)
http://www.hds.utc.fr/˜tdenoeux

MDAI 2015
Skövde, Sweden, September 21, 2015

Prediction

- Prediction (forecasting): making statements about not-yet observed variables (quantities) to be generated from a random/stochastic process
- Part of data science (using past data to anticipate future events)
- Many applications:
  - Pattern recognition / machine learning
  - Economic/financial times series forecasting
  - Sales forecasting
  - Reliability analysis (predicting the life-time of a piece of equipment)
  - Environmental sciences: predicting the time of occurence of a natural phenomenon
  - etc.
- Main problems:
  1. Provide predictions as accurate as possible (usually, problem-dependent)
  2. Describe the uncertainty/reliability of the prediction as faithfully as possible (focus of this talk)

Uncertainty

- Simple example:
  - Urn containing an unknown proportions θ of black balls, and $1-\theta$ of white balls
  - y black balls obtained out of n draws with replacement
  - What is color of the next ball?
- Two sources of uncertainty
  - The ball is drawn at random (aleatory uncertainty)
  - We don't know θ (epistemic uncertainty)
- Classical approaches
  - Frequentist: gives an answer that is correct most the time (over infinitely many replications of the random experiment)
  - Bayesian: assumes prior knowledge on θ and computes a posterior predictive probability $P(black|y)$

Criticism of the frequentist approach

- The frequentist approach makes a statement that is correct, say, for 95% of the samples
- However, 95% is not a correct measure of the confidence in the statement for a particular sample
- Example:
  - Let the prediction be $\{black, white\}$ with probability 0.95 and $\emptyset$ with probability 0.05 (irrespective of the data). This is a 95% prediction set.
  - This prediction is either know for sure to be true, or known for sure to be false.
- Also, the frequentist approach does not allow us to easily
  - Use additional information on θ, if it is available
  - Combine predictions from several sources/agents

Criticism of the Bayesian approach

- The mainstream approach in AI
- Principle: compute $P(black|y)$ as

$$P(black|y) = \int P(black|\theta) f(\theta|y) d\theta$$

  with

$$f(\theta|y) \propto P(y|\theta) f(\theta)$$

- $P(black|y)$ makes sense as a measure of confidence in the statement "the next ball will be black"
- Problem: we need to specify a prior $f(\theta)$ even if we have no prior knowledge at all
- Usual solution: uniform prior. But the prior on $1/\theta$ is not uniform! (when does the knowledge on $1/\theta$ come from?)

Main ideas of this talk

- None of the classical approaches to prediction (frequentist and Bayesian) is conceptually satisfactory
- Proposal of a new approach based on belief functions
- The new approach boils down to Bayesian prediction when a probabilistic prior is available, but it does not require the user to provide such a prior
- Applications:
  1. Calibration of SVM classifiers
  2. Forecasting sales of innovative products

Outline of the new approach (1/2)

- Let us come back to the urn example
- Let $Z \sim \mathcal{B}(\theta)$ be defined as

$$Z = \begin{cases} 1 & \text{if next ball is black} \\ 0 & \text{otherwise} \end{cases}$$

- We can write Z as a function of θ and a pivotal variable $W \sim \mathcal{U}([0,1])$,

$$\begin{aligned} Z &= \begin{cases} 1 & \text{if } W \leq \theta \\ 0 & \text{otherwise} \end{cases} \\ &= \varphi(\theta, W) \end{aligned}$$

Outline of the new approach (2/2)

- The equality

$$Z = \varphi(\theta, W)$$

  allows us to separate the two sources of uncertainty on Z
  1. uncertainty on W (random/aleatory uncertainty)
  2. uncertainty on θ (epistemic uncertainty)
- Two-step method:
  1. Represent uncertainty on θ using a likelihood-based belief function Bel_y^{Θ} constructed from the observed data y (estimation problem)
  2. Combine Bel_y^{Θ} with the probability distribution of W to obtain a predictive belief function $Bel_y^{\mathbb{Z}}$

Outline

1. Reminder on belief functions
   - Introductory example
   - General definitions
2. Prediction method
   - Step 1: likelihood-based belief function
   - Step 2: Predictive belief function
3. Applications
   - Evidential calibration of SVM classifiers
   - Innovation diffusion forecasting

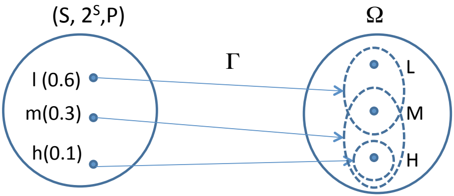
Outline

1. Reminder on belief functions
   - Introductory example
   - General definitions
2. Prediction method
   - Step 1: likelihood-based belief function
   - Step 2: Predictive belief function
3. Applications
   - Evidential calibration of SVM classifiers
   - Innovation diffusion forecasting

Example
Jaffray and Wakker, 1994

- At closing time, a TV set retailer has to decide whether or not to serve a last customer. If he does, he will miss the concert he plans to attend, but he is certain to sell one more TV
- The retailer's profit depends on the price category, L(ow), M(edium), or H(igh), of the new TV bought by the customer
- The prevision of the TV set retailer concerning the type of TV that the customer will buy is based on the following evidence:
  - 60% of the customers own a low (l) price TV, 30% a medium (m) price TV, 10% a high (h) price TV
  - People, when buying a new TV, either remain in the same price range as in the previous purchase or move to the price range directly above

Example
Source



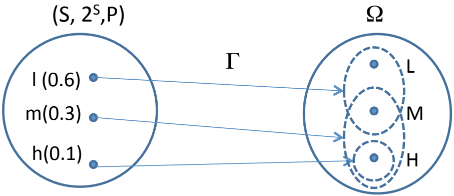
- Let $\Omega = \{L, M, H\}$ be the set of answers to the question of interest (price category of the customer's purchase)
- The four-tuple $(S, 2^S, \mathbb{P}, \Gamma)$ is called a source (random set). It defines the following mass function on Ω:

$$m(\{L, M\}) = 0.6, \quad m(\{M, H\}) = 0.3, \quad m(\{H\}) = 0.1$$

and $m(A) = 0$ for all other A

Example

Belief and plausibility



How to quantify the uncertainty of the proposition "the customer will buy a High price TV"?

- If the customer owns a high price TV, he will certainly buy another one:

$$Bel(\{H\}) = \mathbb{P}(\{s \in S | \Gamma(s) \subseteq \{H\}\}) = \mathbb{P}(\{h\}) = 0.1$$

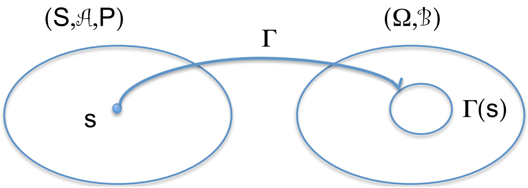
- If the customer owns a medium or high price TV, he may possibly buy a High price one

$$Pl(\{H\}) = \mathbb{P}(\{s \in S | \Gamma(s) \cap \{H\} \neq \emptyset\}) = \mathbb{P}(\{m, h\}) = 0.3+0.1 = 0.4$$

Outline

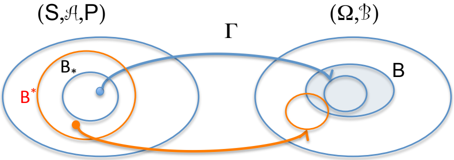
1. Reminder on belief functions
   - Introductory example
   - General definitions
2. Prediction method
   - Step 1: likelihood-based belief function
   - Step 2: Predictive belief function
3. Applications
   - Evidential calibration of SVM classifiers
   - Innovation diffusion forecasting

Source



- Let S be a state space, $\mathcal{A}$ an algebra of subsets of S, $\mathbb{P}$ a finitely additive probability on $(S, \mathcal{A})$
- Let Ω be a set and $\mathcal{B}$ an algebra of subsets of Ω
- Γ a multivalued mapping from S to $2^\Omega \setminus \{\emptyset\}$
- The four-tuple $(S, \mathcal{A}, \mathbb{P}, \Gamma)$ is called a source

Belief and plausibility functions



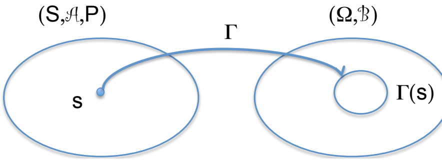
- Under some measurability conditions, the source $(S, \mathcal{A}, \mathbb{P}, \Gamma)$ induces belief and plausibility functions on $(\Omega, \mathcal{B})$:

$$Bel(B) = \mathbb{P}(\{s \in S | \Gamma(s) \subseteq B\})$$

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

- Mathematically, *Bel* and *Pl* are, respectively, completely monotone and completely alternating capacities

Interpretation



- Typically, Ω is the domain of an unknown quantity ω, and S is a set of interpretations of a given piece of evidence about ω
- If $s \in S$ holds, then the evidence tells us that $\omega \in \Gamma(s)$, and nothing more
- Then
  - $Bel(B)$ is the probability that the evidence supports B
  - $Pl(B)$ is the probability that the evidence is consistent with B

Special case I

Belief function on a finite set

- When Ω is finite, *Bel* can be represented by a mass function $m : 2^\Omega \rightarrow [0,1]$ such that $m(\emptyset) = 0$ and

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

- We then have

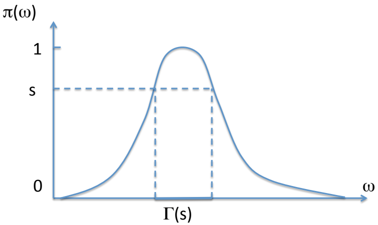
$$Bel(B) = \sum_{A \subseteq B} m(A)$$

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

for all $B \subseteq \Omega$

Special case II

Consonant belief function



- Let π be a mapping from Ω to $S = [0, 1]$ s.t. $\sup \pi = 1$
- Let Γ be the multi-valued mapping from S to 2^Ω defined by

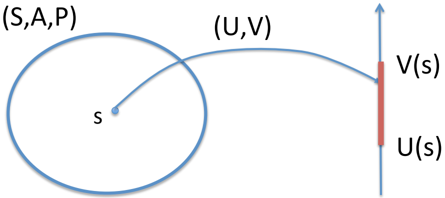
$$\forall s \in [0, 1], \quad \Gamma(s) = \{\omega \in \Omega | \pi(\omega) \geq s\}$$

- The source $(S, \mathcal{B}(S), \lambda, \Gamma)$ defines a consonant BF on Ω, such that $pl(\omega) = \pi(\omega)$ (contour function)
- The corresponding plausibility function is a possibility measure

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

Special case III

Random closed interval

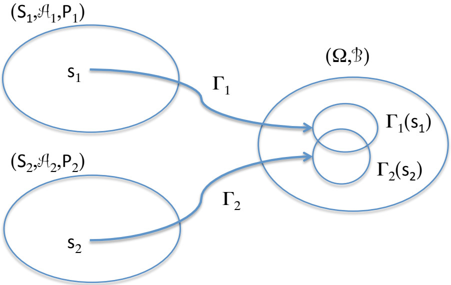


- Let (U, V) be a bi-dimensional random vector from a probability space $(S, \mathcal{A}, \mathbb{P})$ to $\mathbb{R}^2$ such that $U \leq V$ a.s.
- Multi-valued mapping:

$$\Gamma : s \rightarrow \Gamma(s) = [U(s), V(s)]$$

- The source $(S, \mathcal{A}, \mathbb{P}, \Gamma)$ is a random closed interval. It defines a BF on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$

Dempster's rule of combination



- Let $(S_i, \mathcal{A}_i, \mathbb{P}_i, \Gamma_i)$, $i = 1, 2$ be two sources representing independent items of evidence, inducing BF Bel_1 and Bel_2
- The combined BF $Bel = Bel_1 \oplus Bel_2$ is induced by the source $(S_1 \times S_2, \mathcal{A}_1 \otimes \mathcal{A}_2, \mathbb{P}_1 \otimes \mathbb{P}_2, \Gamma_\cap)$ with

$$\Gamma_\cap(s_1, s_2) = \Gamma_1(s_1) \cap \Gamma_2(s_2)$$

Monte Carlo approximation

Require: Desired number of focal sets N
  $i \leftarrow 0$
  while $i < N$ **do**
    Draw s_1 in S_1 from $\mathbb{P}_1$
    Draw s_2 in S_2 from $\mathbb{P}_2$
    $\Gamma_\cap(s_1, s_2) \leftarrow \Gamma_1(s_1) \cap \Gamma_2(s_2)$
    if $\Gamma_\cap(s_1, s_2) \neq \emptyset$ **then**
      $i \leftarrow i + 1$
      $B_i \leftarrow \Gamma_\cap(s_1, s_2)$
    end if
  end while
  $\widehat{Bel}(B) \leftarrow \frac{1}{N}\#\{i \in \{1, \ldots, N\} | B_i \subseteq B\}$
  $\widehat{Pl}(B) \leftarrow \frac{1}{N}\#\{i \in \{1, \ldots, N\} | B_i \cap B \neq \emptyset\}$

Outline

1. Reminder on belief functions
   - Introductory example
   - General definitions

2. Prediction method
   - Step 1: likelihood-based belief function
   - Step 2: Predictive belief function

3. Applications
   - Evidential calibration of SVM classifiers
   - Innovation diffusion forecasting

Parameter estimation

- Let $\boldsymbol{y} \in \mathbb{Y}$ denote the observed data and $f_{\boldsymbol{\theta}}(\boldsymbol{y})$ the probability mass or density function describing the data-generating mechanism, where $\boldsymbol{\theta} \in \Theta$ is an unknown parameter
- Having observed $\boldsymbol{y}$, how to quantify the uncertainty about Θ, without specifying a prior probability distribution?
- Likelihood-based solution (Shafer, 1976; Wasserman, 1990; Denœux, 2014)

Likelihood-based belief function

Requirements

Let $Bel_{\boldsymbol{y}}^{\Theta}$ be a belief function representing our knowledge about θ after observing $\boldsymbol{y}$. We impose the following requirements:

1. Likelihood principle: $Bel_{\boldsymbol{y}}^{\Theta}$ should be based only on the likelihood function
$$\theta \rightarrow L_{\boldsymbol{y}}(\theta) = f_{\theta}(\boldsymbol{y})$$
2. Compatibility with Bayesian inference: when a Bayesian prior P_0 is available, combining it with $Bel_{\boldsymbol{y}}^{\Theta}$ using Dempster's rule should yield the Bayesian posterior:
$$Bel_{\boldsymbol{y}}^{\Theta} \oplus P_0 = P(\cdot|\boldsymbol{y})$$
3. Principle of minimal commitment: among all the belief functions satisfying the previous two requirements, $Bel_{\boldsymbol{y}}^{\Theta}$ should be the least committed (least informative)

Likelihood-based belief function

Solution (Denœux, 2014)

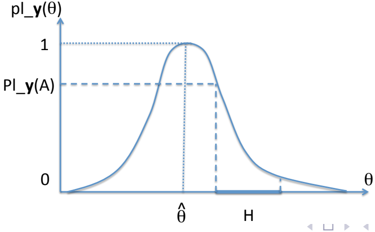
- $Bel_{\mathbf{y}}^{\Theta}$ is the consonant belief function induced by the relative likelihood function

$$pl_{\mathbf{y}}(\boldsymbol{\theta}) = \frac{L_{\mathbf{y}}(\boldsymbol{\theta})}{L_{\mathbf{y}}(\widehat{\boldsymbol{\theta}})}$$

where $\widehat{\boldsymbol{\theta}}$ is a MLE of $\boldsymbol{\theta}$, and it is assumed that $L_{\mathbf{y}}(\widehat{\boldsymbol{\theta}}) < +\infty$

- Corresponding plausibility function

$$Pl_{\mathbf{y}}^{\Theta}(H) = \sup_{\theta \in H} pl_{\mathbf{y}}(\theta), \quad \forall H \subseteq \Theta$$

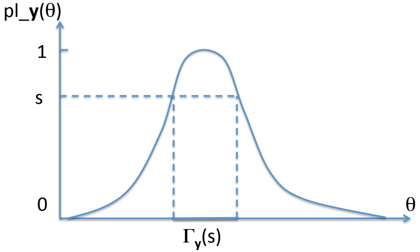


Source

- Corresponding random set:

$$\Gamma_{\boldsymbol{y}}(s) = \left\{ \theta \in \Theta | \frac{L_{\boldsymbol{y}}(\theta)}{L_{\boldsymbol{y}}(\widehat{\theta})} \geq s \right\}$$

with s uniformly distributed in $[0, 1]$



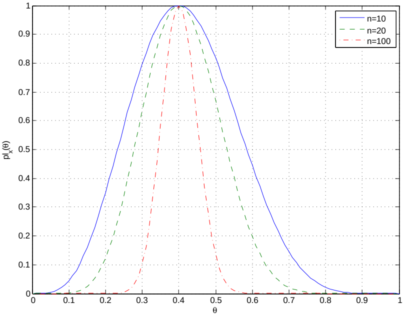
- If $\Theta \subseteq \mathbb{R}$ and if $L_{\boldsymbol{y}}(\theta)$ is unimodal and upper-semicontinuous, then $Bel_{\boldsymbol{y}}^{\Theta}$ corresponds to a random closed interval

Binomial example

In the urn model, $Y \sim \mathcal{B}(n, \theta)$ and

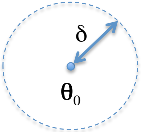
$$pl_y(\theta) = \frac{\theta^y (1-\theta)^{n-y}}{\widehat{\theta}^y (1-\widehat{\theta})^{n-y}} = \left(\frac{\theta}{\widehat{\theta}}\right)^{n\widehat{\theta}} \left(\frac{1-\theta}{1-\widehat{\theta}}\right)^{n(1-\widehat{\theta})}$$

for all $\theta \in \Theta = [0, 1]$, where $\widehat{\theta} = y/n$ is the MLE of θ.



Asymptotic consistency

- $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ iid from $f_{\boldsymbol{\theta}}(y)$, $\boldsymbol{\theta}_0 =$ true value
- Let $B_\delta(\boldsymbol{\theta}_0) = \{\boldsymbol{\theta} \in \Theta | \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leq \delta\}$ be a ball centered on $\boldsymbol{\theta}_0$, with radius δ



- Under mild assumptions, for all $\delta > 0$,

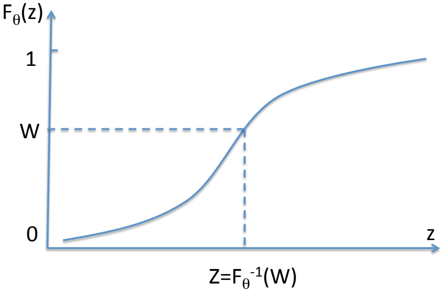
$$Bel_{\boldsymbol{Y}}^{\Theta}(B_\delta(\boldsymbol{\theta}_0)) \xrightarrow{a.s.} 1$$

Outline

Prediction problem

- Observed (past) data: $\boldsymbol{y}$ from $\boldsymbol{Y} \sim f_\theta(\boldsymbol{y})$
- Future data: $Z|\boldsymbol{y} \sim F_{\theta,\boldsymbol{y}}(z)$ (real random variable)
- Problem: quantify the uncertainty of Z using a predictive belief function

φ-equation



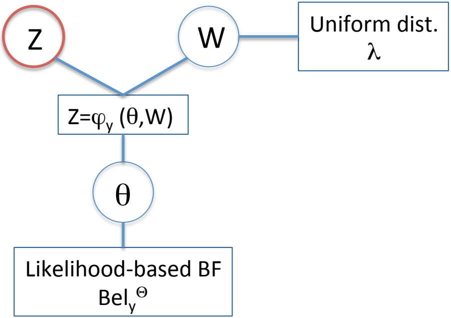
We can always write Z as a function of $\boldsymbol{\theta}$ and W as

$$Z = F^{-1}_{\boldsymbol{\theta},\boldsymbol{y}}(W) = \varphi_{\boldsymbol{y}}(\boldsymbol{\theta}, W)$$

where $W \sim \mathcal{U}([0,1])$ and $F^{-1}_{\boldsymbol{\theta},\boldsymbol{y}}$ is the generalized inverse of $F_{\boldsymbol{\theta},\boldsymbol{y}}$,

$$F^{-1}_{\boldsymbol{\theta},\boldsymbol{y}}(W) = \inf\{z | F_{\boldsymbol{\theta},\boldsymbol{y}}(z) \geq W\}$$

Main result

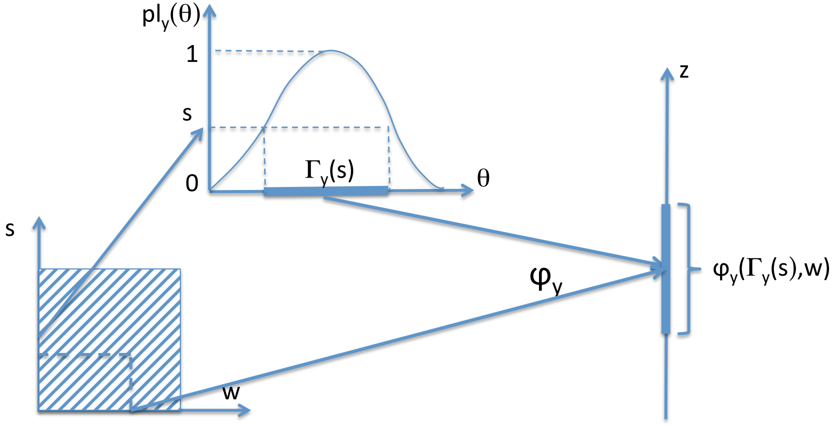


After combination by Dempster's rule and marginalization on $\mathbb{Z}$, we obtain the predictive BF on Z induced by the multi-valued mapping

$$(s, w) \rightarrow \varphi_{\boldsymbol{y}}(\Gamma_{\boldsymbol{y}}(s), w).$$

with (s, w) uniformly distributed in $[0, 1]^2$

Graphical representation



Practical computation

- Analytical expression when possible (simple cases), or
- Monte Carlo simulation:
  1. Draw N pairs (s_i, w_i) independently from a uniform distribution
  2. compute (or approximate) the focal sets $\varphi_{\mathbf{y}}(\Gamma_{\mathbf{y}}(s_i), w_i)$
- The predictive belief and plausibility of any subset $A \subseteq \mathbb{Z}$ are then estimated by

$$\widehat{Bel}_{\mathbf{y}}^{\mathbb{Z}}(A) = \frac{1}{N}\#\{i \in \{1, \ldots, N\} | \varphi_{\mathbf{y}}(\Gamma_{\mathbf{y}}(s_i), w_i) \subseteq A\}$$

$$\widehat{Pl}_{\mathbf{y}}^{\mathbb{Z}}(A) = \frac{1}{N}\#\{i \in \{1, \ldots, N\} | \varphi_{\mathbf{y}}(\Gamma_{\mathbf{y}}(s_i), w_i) \cap A \neq \emptyset\}$$

Example: the urn model

- Here, $Y \sim \mathcal{B}(n, \theta)$. The likelihood-based BF is induced by a random interval

$$\Gamma(s) = \{\theta : pl_y(\theta) \geq s\} = [\underline{\theta}(s), \overline{\theta}(s)]$$

- We have

$$Z = \varphi(\theta, W) = \begin{cases} 1 & \text{if } W \leq \theta \\ 0 & \text{otherwise} \end{cases}$$

- Consequently,

$$\varphi(\Gamma(s), W) = \varphi\left(\left[\underline{\theta}(s), \overline{\theta}(s)\right], W\right) = \begin{cases} \{1\} & \text{if } W \leq \underline{\theta}(s) \\ \{0\} & \text{if } \overline{\theta}(s) < W \\ \{0, 1\} & \text{otherwise} \end{cases}$$

Example: the urn model

Analytical formula

We have

$$m_y^{\mathbb{Z}}(\{1\}) = \mathbb{P}\left(\varphi\left(\Gamma(s), W\right) = \{1\}\right) = \hat{\theta} - \frac{\underline{B}(\hat{\theta}; y+1, n-y+1)}{\hat{\theta}^y (1-\hat{\theta})^{n-y}}$$

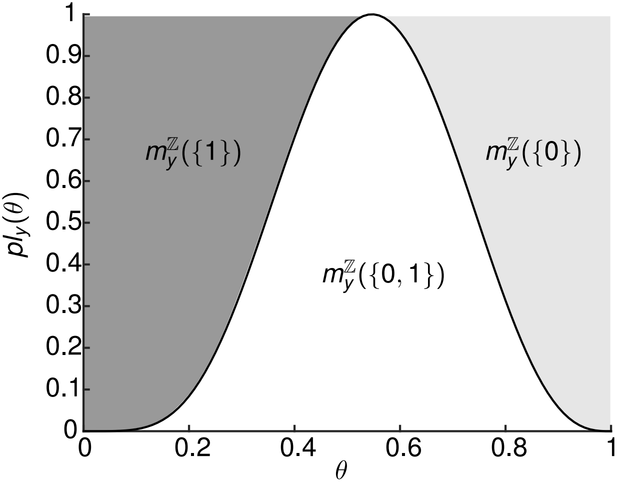
$$m_y^{\mathbb{Z}}(\{0\}) = \mathbb{P}\left(\varphi\left(\Gamma(s), W\right) = \{0\}\right) = 1 - \hat{\theta} - \frac{\underline{B}(1-\hat{\theta}; n-y+1, y+1)}{\hat{\theta}_j^y (1-\hat{\theta})^{n-y}}$$

$$m_y^{\mathbb{Z}}(\{0,1\}) = 1 - m_y^{\mathbb{Z}}(\{0\}) - m_y^{\mathbb{Z}}(\{1\})$$

where $\underline{B}(z; a, b) = \int_0^z t^{a-1}(1-t)^{b-1} dt$ is the incomplete beta function

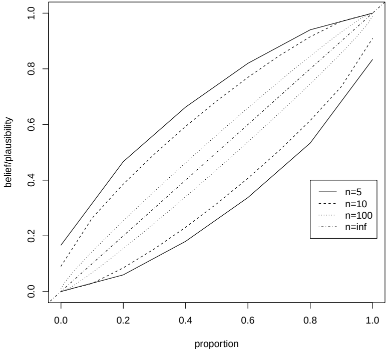
Example: the urn model

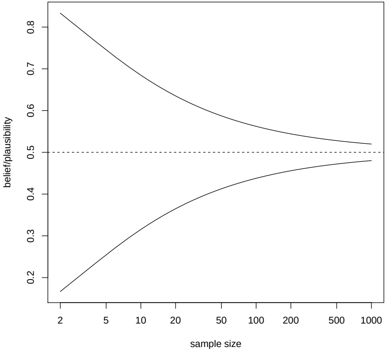
Geometric representation



Example: the urn model

Belief/plausibility intervals





Consistency

- Here, it is easy to show that

$$m_y^{\mathbb{Z}}(\{1\}) \xrightarrow{P} \theta_0 \quad \text{and} \quad m_y^{\mathbb{Z}}(\{0\}) \xrightarrow{P} 1-\theta_0$$

as $n \to \infty$, i.e., the predictive belief function converges to the true distribution of Z

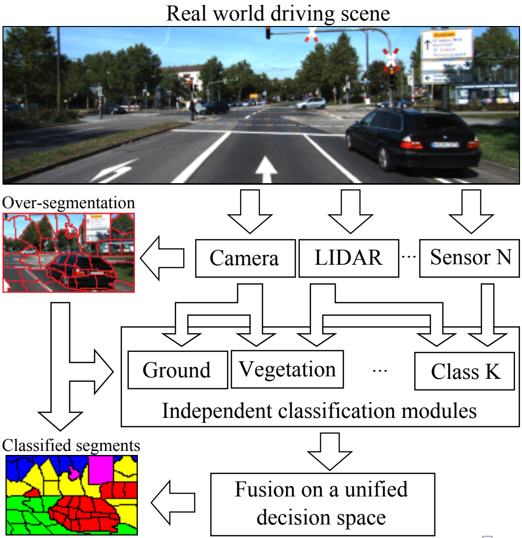
- When the predictive belief function is induced by a random interval $[\underline{Z}, \overline{Z}]$, we can show that, under mild conditions,

$$\underline{Z} \xrightarrow{d} Z \quad \text{and} \quad \overline{Z} \xrightarrow{d} Z$$

- The consistency remains to be proved in the general case

Outline

Classifier fusion



Classifier calibration

- Binary classification problem: predict the class $Y \in \{0,1\}$ of an instance described by a feature vector $\boldsymbol{x}$, based on a training set $\mathcal{X} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$
- Here, we consider the case where a classifier such as an SVM has already been trained to provide a score s, such that the predicted class is

$$\widehat{Y} = \begin{cases} 1 & \text{if } s > s_0 \\ 0 & \text{if } s \leq s_0 \end{cases}$$

- Problem: quantify the uncertainty on Y, to
  - postpone the decision when the uncertainty is high, or
  - combine several classifiers
- Classical approach: estimate the probability $P(y|s)$ (probabilistic calibration)
- Our approach: compute a predictive belief function on Y (evidential calibration)

Probabilistic calibration

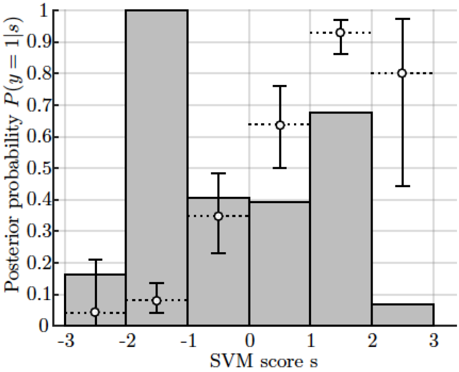
1. Binning:
   - Partition the score space into bins $[\underline{s}_j, \overline{s}_j)$, $j = 1, \ldots, J$, and assume that $P(y|s)$ is (approximately) constant in each bin
   - For the j-th bin, count the number of positive examples k_j over all the n_j training examples in this bin
   - If $\underline{s}_j \leq s < \overline{s}_j$, $P(y = 1|s)$ is estimated by $\hat{\theta}_j = k_j / n_j$
2. Logistic regression: consider a model

$$P(y = 1|s) = h_s(\theta) = \frac{1}{1 + \exp(\theta_0 + \theta_1 s)}$$

and compute the MLE $\widehat{\theta}$ of $\theta = (\theta_0, \theta_1) \in \mathbb{R}^2$

Limitations of probabilistic calibration

Probabilistic calibration does not take adequately quantify the uncertainty of the prediction, as it does not take into account the size of the training set



Evidential calibration: binning approach

- We assume that

$$\mathbb{P}(Y = 1 | s \in [\underline{s}_j, \overline{s}_j)) = \theta_j$$

- Let n_j be the number of instances in bin j, and k_j the number of positive instances
- If s falls in bin j, the predictive mass function is

$$m_j^{\mathbb{Y}}(\{1\}) = \hat{\theta}_j - \frac{\underline{B}(\hat{\theta}_j; k_j + 1, n_k - k_j + 1)}{\hat{\theta}_j^{k_j}(1 - \hat{\theta}_j)^{n_j - k_j}}$$

$$m_j^{\mathbb{Y}}(\{0\}) = 1 - \hat{\theta}_j - \frac{\underline{B}(1 - \hat{\theta}_j; n_j - k_j + 1, k_j + 1)}{\hat{\theta}_j^{k_j}(1 - \hat{\theta}_j)^{n_j - k_j}}$$

$$m_j^{\mathbb{Y}}(\{0, 1\}) = 1 - m_j^{\mathbb{Y}}(\{1\}) - m_j^{\mathbb{Y}}(\{0\})$$

Evidential calibration: logistic regression

- We assume that $\mathbb{P}(Y = 1|s)$ is

$$\tau = h_s(\theta) = \frac{1}{1 + \exp(\theta_0 + \theta_1 s)}$$

- The likelihood function is

$$L_{\mathcal{X}}(\theta) = \prod_{i=1}^{n} h_s(\theta)^{y_i} (1 - h_s(\theta))^{1-y_i}$$

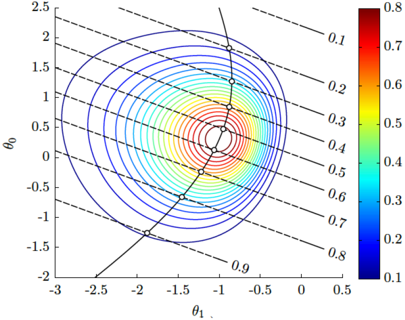
- The likelihood-based belief function on θ is defined by

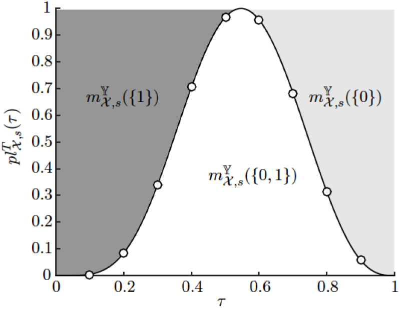
$$pl_{\mathcal{X}}(\theta) = L_{\mathcal{X}}(\theta)/L_{\mathcal{X}}(\hat{\theta})$$

- The corresponding belief function on τ is defined by

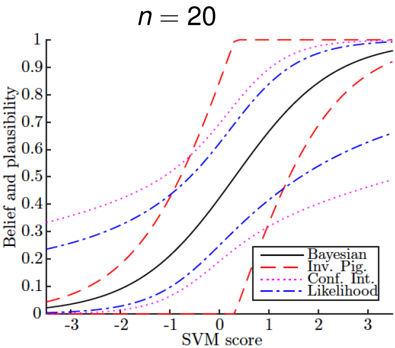
$$\begin{aligned} pl_{\mathcal{X},s}(\tau) &= Pl_{\mathcal{X}}^{\Theta}(\{\theta \in \Theta \mid \tau = h_s(\theta)\}) \\ &= \sup_{\theta_1 \in \mathbb{R}} pl_{\mathcal{X}}^{\Theta}\left(\ln\left(\tau^{-1} - 1\right) - \theta_1 s, \theta_1\right) \end{aligned}$$

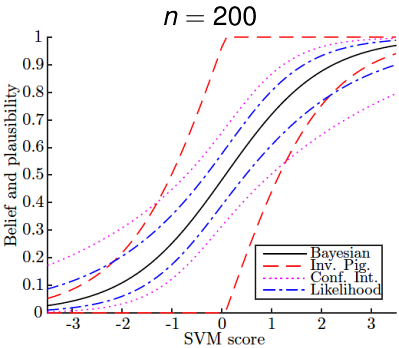
Evidential calibration: logistic regression





Evidential calibration: logistic regression





Evidential calibration: logistic regression

Experiment

10 classifiers trained on 10 overlapping subsets of the training data:

(a) (b) (c)

Three scenarios:

(a) All classifiers use 10% of the training data
(b) 5 classifiers use $1/6^{\text{th}}$ of the data and the other 5 use the rest
(c) One classifier uses $2/3^{\text{rd}}$ of the data, another one uses $1/5^{\text{th}}$ and the eight other classifiers use the rest

Evidential calibration: logistic regression

Classification error rates

	Adult #train=600, #test=16,281			Australian #train=300, #test=390		
Scenario	(a)	(b)	(c)	(a)	(b)	(c)
Bayesian	**16.76**%	17.30%	19.10%	<u>**14.87**</u>%	**14.10**%	14.10%
Likelihood	<u>**16.71**</u>%	<u>**16.97**</u>%	<u>**18.35**</u>%	<u>**14.87**</u>%	<u>**13.33**</u>%	<u>**11.54**</u>%

	Diabetes #train=300, #test=468		
Scenario	(a)	(b)	(c)
Bayesian	**21.58**%	**22.86**%	46.58%
Likelihood	<u>**20.94**</u>%	<u>**22.65**</u>%	<u>**31.84**</u>%

The best results are underlined and those that are not significantly different are in bold

Outline

Innovation diffusion

- Forecasting the diffusion of an innovation has been a topic of considerable interest in marketing research
- Typically, when a new product is launched, sale forecasts have to be based on little data and uncertainty has to be quantified to avoid making wrong business decisions based on unreliable forecasts
- Our approach uses the Bass model (Bass, 1969) for innovation diffusion together with past sales data to quantify the uncertainty on future sales using the formalism of belief functions

Bass model

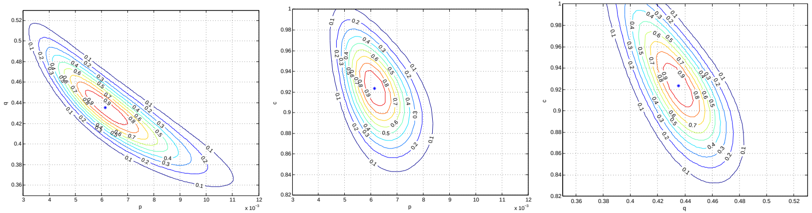
- Fundamental assumption (Bass, 1969): for eventual adopters, the probability $f(t)$ of purchase at time t, given that no purchase has yet been made, is an affine function of the number of previous buyers
$$\frac{f(t)}{1-F(t)} = p + qF(t)$$
where p is a coefficient of innovation, q is a coefficient of imitation and $F(t) = \int_0^t f(u)du$.
- Solving this differential equation, the probability that an individual taken at random from the population will buy the product before time t is
$$\Phi_\theta(t) = cF(t) = \frac{c(1-\exp[-(p+q)t])}{1+(p/q)\exp[-(p+q)t]}$$
where c is the probability of eventually adopting the product and $\theta = (p, q, c)$

Parameter estimation

- Given $y_1, \ldots, y_{T-1}$, where y_i = observed number of adopters in time interval $[t_{i-1}, t_i)$, we can compute the likelihood function

$$L_{\mathbf{y}}(\theta) \propto \prod_i p_i^{y_i}$$

- The belief function on θ is defined by $pl_{\mathbf{y}}(\theta) = L_{\mathbf{y}}(\theta)/L_{\mathbf{y}}(\widehat{\theta})$



Sales forecasting

- Let us assume we are at time t_{T-1} and we wish to forecast the number Z of sales between times τ_1 and τ_2, with $t_{T-1} \leq \tau_1 < \tau_2$
- Z has a binomial distribution $\mathcal{B}(Q, \pi_\theta)$, where
  - Q is the number of potential adopters at time $T-1$
  - π_θ is the probability of purchase for an individual in $[\tau_1, \tau_2]$, given that no purchase has been made before t_{T-1}

$$\pi_\theta = \frac{\Phi_\theta(\tau_2) - \Phi_\theta(\tau_1)}{1 - \Phi_\theta(t_{T-1})}$$

- Z can be written as $Z = \varphi(\theta, \boldsymbol{W}) = \sum_{i=1}^{Q} \mathbb{1}_{[0,\pi_\theta]}(W_i)$ where

$$\mathbb{1}_{[0,\pi_\theta]}(W_i) = \begin{cases} 1 & \text{if } W_i \leq \pi_\theta \\ 0 & \text{otherwise} \end{cases}$$

and $\boldsymbol{W} = (W_1, \ldots, W_Q)$ has a uniform distribution in $[0,1]^Q$.

Predictive belief function

Multi-valued mapping

- The predictive belief function on Z is induced by the multi-valued mapping $(s, \boldsymbol{w}) \rightarrow \varphi(\Gamma_{\boldsymbol{y}}(s), \boldsymbol{w})$ with

$$\Gamma_{\boldsymbol{y}}(s) = \{\theta \in \Theta : pl_{\boldsymbol{y}}(\theta) \geq s\}$$

- When θ varies in $\Gamma_{\boldsymbol{y}}(s)$, the range of π_θ is $[\underline{\pi}_\theta(s), \overline{\pi}_\theta(s)]$, with

$$\underline{\pi}_\theta(s) = \min_{\{\theta | pl_{\boldsymbol{y}}(\theta) \geq s\}} \pi_\theta, \quad \overline{\pi}_\theta(s) = \max_{\{\theta | pl_{\boldsymbol{y}}(\theta) \geq s\}} \pi_\theta$$

- We have

$$\varphi(\Gamma_{\boldsymbol{y}}(s), \boldsymbol{w}) = [\underline{Z}(s, \boldsymbol{w}), \overline{Z}(s, \boldsymbol{w})],$$

where $\underline{Z}(s, \boldsymbol{w})$ and $\overline{Z}(s, \boldsymbol{w})$ are, respectively, the number of w_i's that are less than $\underline{\pi}_\theta(s)$ and $\overline{\pi}_\theta(s)$
- For fixed s, $\underline{Z}(s, \boldsymbol{W}) \sim \mathcal{B}(Q, \underline{\pi}_\theta(s))$ and $\overline{Z}(s, \boldsymbol{W}) \sim \mathcal{B}(Q, \overline{\pi}_\theta(s))$

Predictive belief function

Calculation

- The belief and plausibilities that Z will be less than z are

$$Bel_{\mathbf{y}}^{\mathbb{Z}}([0,z]) = \int_0^1 F_{Q,\underline{\pi}_\theta(s)}(z)ds$$

$$Pl_{\mathbf{y}}^{\mathbb{Z}}([0,z]) = \int_0^1 F_{Q,\overline{\pi}_\theta(s)}(z)ds$$

where $F_{Q,p}$ denotes the cdf of the binomial distribution $\mathcal{B}(Q,p)$

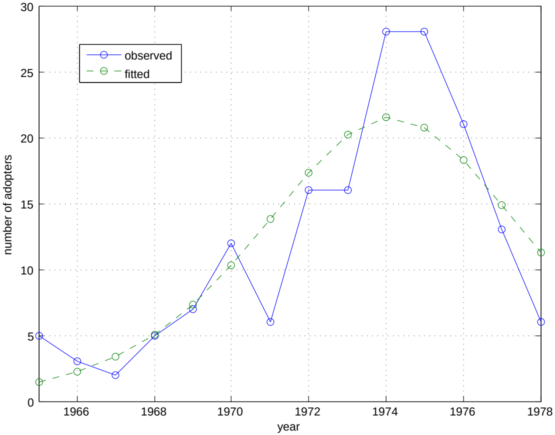
- The contour function of Z is

$$pl_{\mathbf{y}}(z) = \int_0^1 \left(F_{Q,\underline{\pi}_\theta(s)}(z) - F_{Q,\overline{\pi}_\theta(s)}(z-1)\right) ds$$

- Theses integrals can be approximated by Monte-Carlo simulation

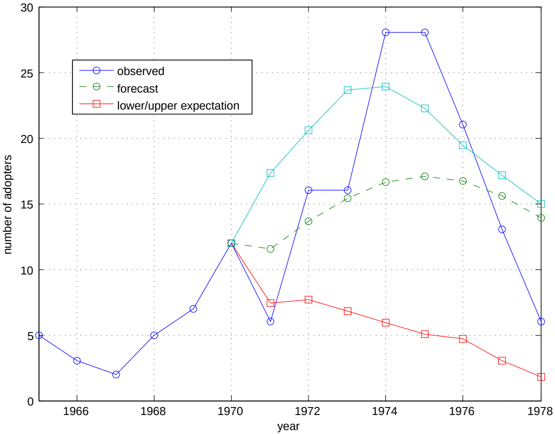
Ultrasound data

Data collected from 209 hospitals through the U.S.A. (Schmittlein and Mahajan, 1982) about adoption of an ultrasound equipment



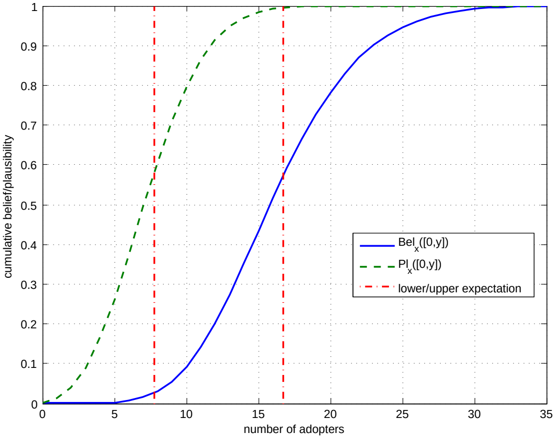
Forecasting

Predictions made in 1970 for the number of adopters in the period 1971-1978, with their lower and upper expectations



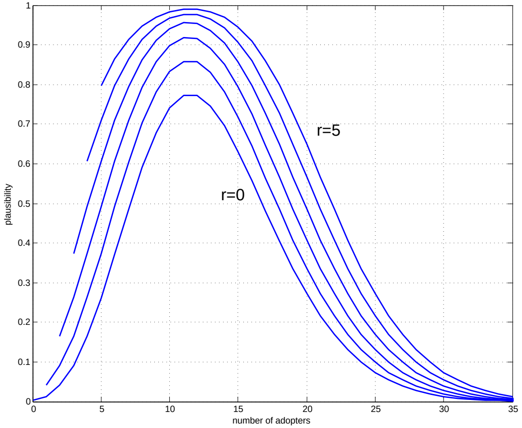
Cumulative belief and plausibility functions

Lower and upper cumulative distribution functions for the number of adopters in 1971, forecasted in 1970



Pl-plot

Plausibilities $Pl_{\boldsymbol{y}}^{\mathbb{Y}}([z-r, z+r])$ as functions of z, from $r=0$ (lower curve) to $r=5$ (upper curve), for the number of adopters in 1971, forecasted in 1970:



Conclusions

- Uncertainty quantification is an important component of any forecasting methodology. The approach introduced in this paper allows us to represent forecast uncertainty in the belief function framework, based on past data and a statistical model
- The proposed method is conceptually simple and computationally tractable
- The belief function formalism makes it possible to combine information from several sources (such as expert opinions and statistical data)
- The Bayesian predictive probability distribution is recovered when a prior on θ is available
- The consistency of the method has been established under some conditions

References

cf. `http://www.hds.utc.fr/~tdenoeux`

- T. Denœux.
  Likelihood-based belief function: justification and some extensions to low-quality data.
  International Journal of Approximate Reasoning, 55(7):1535–1547, 2014.

- O. Kanjanatarakul, S. Sriboonchitta and T. Denœux
  Forecasting using belief functions. An application to marketing econometrics.
  International Journal of Approximate Reasoning, 55(5):1113–1128, 2014.

- Ph. Xu, F. Davoine, H. Zha, T. Denœux
  Evidential calibration of binary SVM classifiers
  International Journal of Approximate Reasoning (accepted)