Dempster-Shafer Theory

Material used

- Halpern, chapter 2.4
- Frans Voorbraak: *Dempster-Shafer Theory* (www.blutner.de/uncert/DSTh.pdf)

General Motivation

Compare: Tossing a coin which is known to be fair
Tossing a coin which is not known to be fair

In both cases, we assign a probability of 0.5 to the proposition that the result is heads. In the first case this assignment is based on probabilistic knowledge, in the second case it is based on the absence of such knowledge.

- Generalizations of probability theory which do allow the representation of ignorance.
- E.g., there can be a medical test whose positive outcome supports some hypothesis *h* to degree 0.7 and to degree 0.3 it is ignorant (confirming $h \cup \sim h$ rather than confirming $\sim h$)

1 Overview – Generalizations of PT

Capacities

↓

Sets of probability functions
(upper & lower probability)

↓

Dempster-Shafer Theory

↓

Inner & outer measure

↓

Probability measure

Additivity

- For disjoint U, V: $\mu(U \cup V) = \mu(U) + \mu(V)$.
- "Experts" very often seem to use non-additive measures of degrees of belief.
- The Dutch book argument requires that degrees of belief are additive; thus, we have to reconsider this argument: Give up the unique breaking point!
- Skip *Additivity* and look for generalized measures. Consider finite sample spaces only!

Capacities

Definition: Let W be a sample space. A real-valued function F on all subsets of W is called a capacity iff the following conditions are satisfied:

1. $F(\varnothing) = 0$ (normalization)
2. $F(\Omega) = 1$ (normalization)
3. For all $U, V \subseteq W$, $U \subseteq V \Rightarrow F(U) \leq F(V)$

Exercise: Show that the upper (lower) measure is a capacity!

Remark: General definitions of the dual on the basis of capacities!

Dual

Definition: Let G and F be functions on 2^W.
G is the dual of F iff for every $U \subseteq W$, $F(U) = 1 - G(\sim U)$

- The dual of a capacity is also a capacity
- If G is the dual of F then F is the dual of G
- A probability function is its own dual

Sub- and superaddidivity

Definition: Let F be a capacity over Ω. Let be U, V disjoint, F is subadditive iff $F(U \cup V) \leq F(U) + F(V)$ for all disjoint events $U,V \subseteq W$. F is superadditive iff $F(U \cup V) \geq F(U) + F(V)$ for all disjoint events $U, V \subseteq W$.

- In the exercises we have shown that lower (and inner) measures are superadditive and upper (and outer) measures are subadditive.
- Upper (lower) measures can be characterized as subadditive (superadditive) capacities (+ a continuity property, cf. Halpern p. 31)

The inclusion-exclusion rule

How to characterize inner and outer measure?

For **probabilities** we have the following inclusion-exclusion rules (assuming $U_i \subseteq W$)

$$\mu(U_1 \cup U_2) = \mu(U_1) + \mu(U_2) - \mu(U_1 \cap U_2)$$

$$\mu(U_1 \cup U_2 \cup U_3) = \mu(U_1) + \mu(U_2) + \mu(U_3) - \mu(U_1 \cap U_2) - \mu(U_1 \cap U_3) - \mu(U_2 \cap U_3) + \mu(U_1 \cap U_2 \cap U_3)$$

$$\dots$$

Replacing = by $\geq$ we get an inclusion-exclusion rule that is necessary (but not sufficient) for **inner measures** μ_*. From duality it follows a corresponding condition necessary for outer measures μ^*. (cf. Halpern, p. 30 ff.)

Summary

Capacities

↓

Sets of probability functions (upper & lower probability)

sub / superadditivity

↓

Dempster-Shafer Theory

inclusion-exclusion rule (with ≥; resp. dual)

↓

Inner & outer measure

(see later)

↓

Probability measure

inclusion-exclusion rule with =

2 Dempster-Shafer Belief Functions

A general (abstract) formulization sees Belief functions as a special case of upper probabilities:

Definition: A belief function Bel defined on a space W satisfies the following three properties:

B1. $\text{Bel}(\varnothing) = 0$ (normalization)

B2. $\text{Bel}(W) = 1$ (normalization)

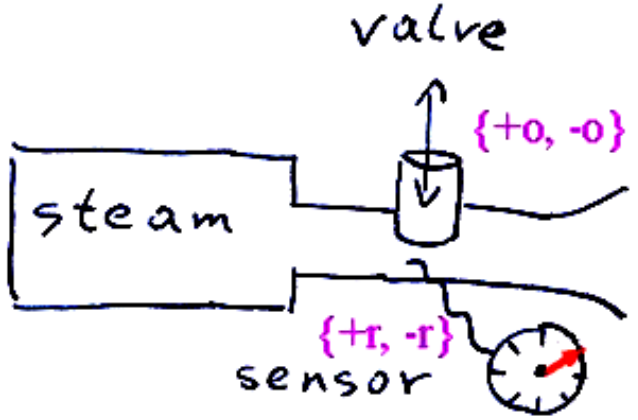
B3. $\text{Bel}(U_1 \cup U_2) \geq \text{Bel}(U_1) + \text{Bel}(U_2) - \text{Bel}(U_1 \cap U_2)$

$\text{Bel}(U_1 \cup U_2 \cup U_2) \geq \text{Bel}(U_1) + \text{Bel}(U_2) + \text{Bel}(U_3) - \text{Bel}(U_1 \cap U_2) - \text{Bel}(U_1 \cap U_3) - \text{Bel}(U_2 \cap U_3) + \text{Bel}(U_1 \cap U_2 \cap U_3)$

… (inclusion-exclusion rule)

Dempster's scenario

Suppose one is interested in the question whether the ***valve*** is closed or open. The only information about the state of the valve is provided by a ***sensor.*** It is known that the sensor is unreliable in exactly 20 % of the cases (represented by a variable *r* - *hidden parameter*). Suppose the sensor indicates "valve open".



$W=\{+o, -o\}$; $H=\{+r, -r\}$

$\mu(\{+r\}) = 0.8$, $\mu(\{-r\}) = 0.2$

Mapping Γ: H => 2^W-$\{\varnothing\}$; $\Gamma(+r)=\{o\}$, $\Gamma(-r)=\{+o, -o\}$

$\mathrm{Bel}(U) =_{\mathrm{def}} \mu(\{h \in H: \Gamma(h) \subseteq U\})$

$\mathrm{Pl}(U) =_{\mathrm{def}} \mu(\{h \in H: \Gamma(h) \cap U \neq \varnothing\})$

	Bel	Pl
{+o}	0.8	1
{–o}	0	0.2

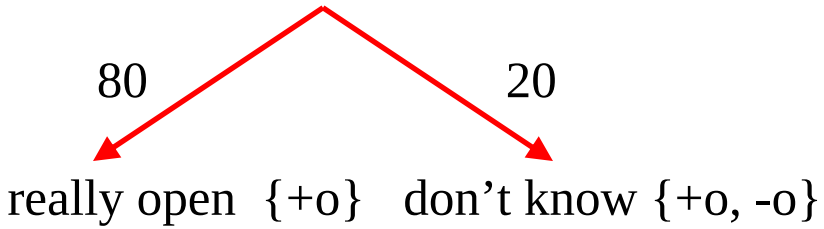
Shafer's interpretation

- In Dempster's scenario belief functions are constructed by means of multi-valued mappings.
- Bel and its dual, Pl (plausibility), are special kind of lower/upper probability functions: You can see it by defining $\mathcal{P}_{\text{Bel}} = \{\mu: \mu(U) \geq \text{Bel}(U) \text{ for all } U \subseteq W\}$ and showing that $\text{Bel} = (\mathcal{P}_{\text{Bel}})_*$ and $\text{Pl} = (\mathcal{P}_{\text{Bel}})^*$
- Shafer gave a somewhat different interpretation of these ideas (given in the book *A Mathematical Theory of Evidence*). In his theory, belief functions are part of a theory of *evidence*.

Shafer's interpretation: Example

- W={+o, –o} *Frame of discernment*
- m({+o}) = 0.8, m({+o, –o}) = 0.2, m({–o}) = 0, m(∅) = 0. *Mass function* or *basic probability assignment.* Intuitively, m(U) describes the extent to which the evidence supports U.
- $Bel(U) = \Sigma_{U' \subseteq U}\, m(U')$; $Pl(U) = \Sigma_{U' \cap U \neq \varnothing}\, m(U')$

Sensor says "valve open" (100 events)



	Bel	Pl
{+o}	0.8	1
{–o}	0	0.2

Definition (mass function)
A mass function on W is a function m: $2^W \rightarrow [0, 1]$ such that the following two conditions hold:

$$m(\varnothing) = 0.$$
$$\Sigma_{U \subseteq W}\, m(U) = 1$$

Definition (belief/plausibility function based on m)
Let m be a mass function on W. Then for every $U \subseteq W$:

$$Bel(U) =_{def} \Sigma_{U' \subseteq U}\, m(U')$$
$$Pl(U) =_{def} \Sigma_{U' \cap U \neq \varnothing}\, m(U')$$

Facts

- *Bel* and *Pl* are dual.

$$\Sigma_{U' \subseteq \sim U}\ \mathrm{m}(U') + \Sigma_{U' \cap U \neq \emptyset}\ \mathrm{m}(U') =$$
$$\Sigma_{U' \cap U = \emptyset}\ \mathrm{m}(U') + \Sigma_{U' \cap U \neq \emptyset}\ \mathrm{m}(U') = 1$$

- If *Bel* is a belief function on *W*, then there is a unique mass function m over Ω such that *Bel* is the belief function based on m. This mass function is given by the following equation:

$$\text{For all } U \subseteq W,\ \mathrm{m}(U) = \Sigma_{U' \subseteq U}\ (-1)^{|\,U \setminus U'\,|}\ Bel(U')$$

Comments

- The complete information about the measure of belief in U can be represented by the interval [Bel(U), Pl(U)], where Pl(U) - Bel(U) is a natural expression of the ignorance concerning U
- It is tempting to consider Bel(U) resp. Pl(U), as lower, resp. upper, bound of the “true” probability of U.
- Not every belief function over W is an inner measure extension over W. This follows from the fact that for inner measure extensions the *focal elements* are pairwise disjoint.

Safecracker example

Important documents were stolen from a safe. Sherlock Holmes comes with the following two clues:

1. Examination of the safe suggests, with a 70% degree of certainty, that the safecracker was left-handed (and with 30% we don't know)
   [finding a *hanky* on the left hand site of the safe]
2. Since the door giving entrance to the room with the safe has not been forced, it can be concluded, with a certainty of 80%, that it was an inside job (with 20% we don't know)

What is the belief function (concerning possible thieves) in case of using clue 1 only?

Safecracker example

Answer: W is the set of possible safecrackers (exactly one of them is the actual safecracker); L is the subset of left-handed persons in W.

$m_1(L) = 0.7$, $m_1(W) = 0.3$

$$\text{Bel}_1(U) = \begin{cases} 1 & \text{if } U = W \\ 0.7 & \text{if } L \subseteq U \neq W \\ 0 & \text{otherwise} \end{cases}$$

Remark: If m_1 had been an ordinary probability distribution then you would have expected $m_1(R) = 0.3$, which would have meant, with a 30% degree of certainty, that the thief was right-handed. So DS probability assignments distribute the remaining belief over the universal hypothesis, whereas classical probability distributions distribute it over the complement of the current hypothesis.

Bayesian belief function

- A belief function *Bel* is called *Bayesian* if *Bel* is a probability function.
- The following conditions are equivalent
  - **Bel** is Bayesian
  - All the focal elements of **Bel** are singletons
    [$U \subseteq W$ is called a *focal element of* **Bel** iff $m(U) > 0$]
  - For every $U \subseteq W$, $Bel(U) + Bel(\sim U) = 1$
- The inner measure can be characterized by the condition that the focal elements are pairwise disjoint.

Summary

Capacities

↓

Sets of probability functions (upper & lower probability) — sub / superadditivity

↓

Dempster-Shafer Theory — inclusion-exclusion rule (≥)

↓

Inner & outer measure — focal elements are disjoint

↓

Probability measure — focal elements are singletons

3 Combining the Evidence

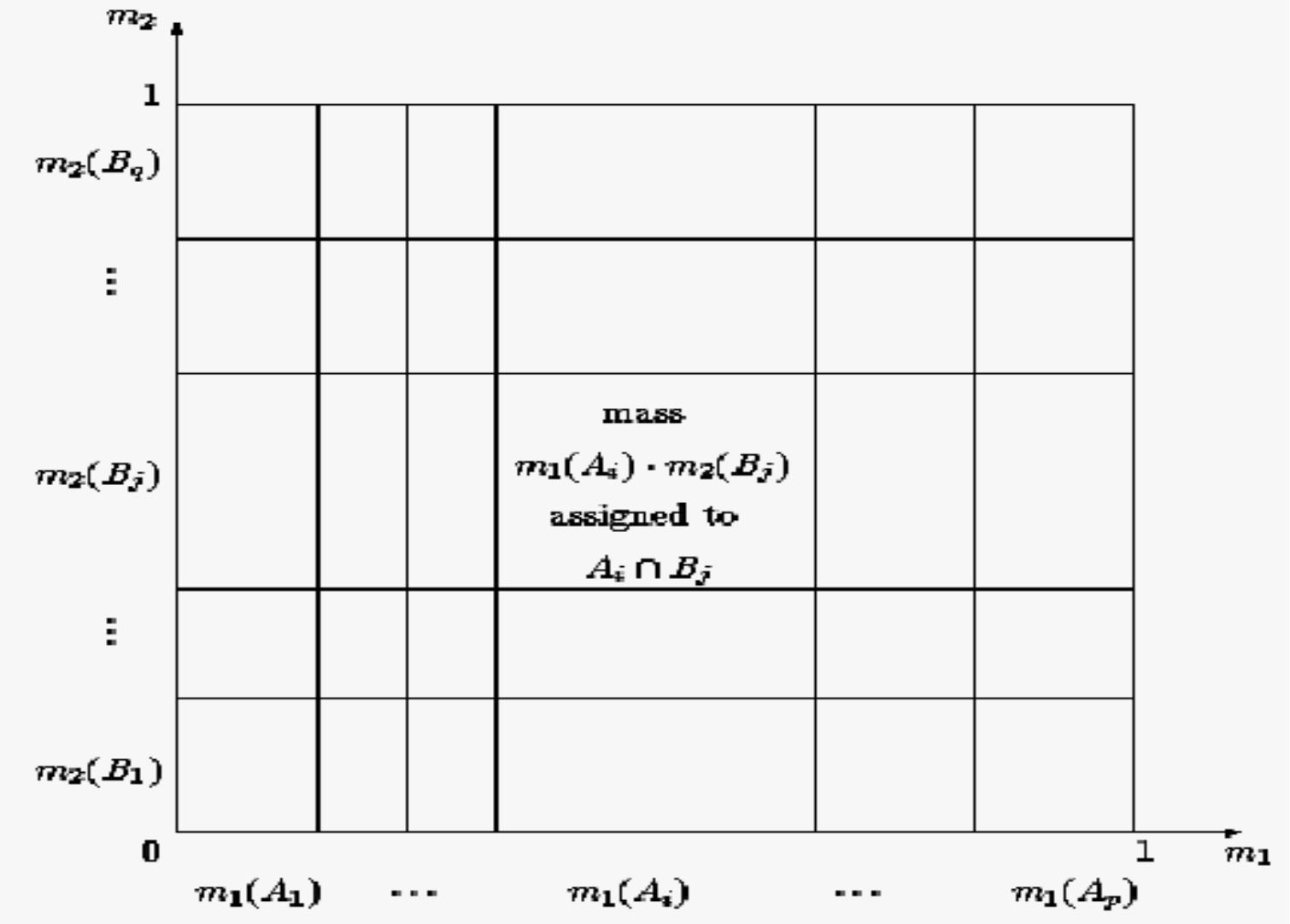
- Dempster-Shafer Theory as a theory of evidence has to account for the combination of different sources of evidence
- Dempster & Shafer's Rule of Combination is a essential step in providing such a theory
- This rule is an intuitive axiom that can best be seen as a heuristic rule rather than a well-grounded axiom.

Safecracker example, combining clues

$m_1(L) = 0.7$, $m_1(W) = 0.3$
$m_2(I) = 0.8$, $m_2(W) = 0.2$
$m(L \cap I)=0.56$, $m(L)=0.14$, $m(I)=0.24$, $m(W)=0.06$

$Bel(L) = 0.56 + 0.14 = 0.7$ (as before)
$Bel(L \cap I) = 0.56$ (new!)
$Bel(I) = 0.56 + 0.24 = 0.8$ (as before)

Mass assignment for combined evidences



Three Problems

- A subset A of W may be the combination of different pairs A_i and B_j.
- There can be focal elements A_i of m_1 and B_j of m_2 such that $A_i \cap B_j = \varnothing$.
- Mass functions are not always *combinable*. For example, they are not combinable if $A_i \cap B_j = \varnothing$ for each i and j.

Dempster's rule of combination

Suppose m_1 and m_2 are basic probability functions over W. Then $m_1 \oplus m_2$ is given by

$$m_1 \oplus m_2(A) = \begin{cases} 0 & \text{if } A = \emptyset \\ \dfrac{\sum_{A_i \cap B_j = A} m_1(A_i) \cdot m_2(B_j)}{\sum_{A_i \cap B_j \neq \emptyset} m_1(A_i) \cdot m_2(B_j)} & \text{if } \emptyset \neq A \subseteq W \end{cases}$$

The factor $[\Sigma_{Ai \cap Bj \neq \varnothing}\ m_1(A_i) \cdot m_2(B_j)]^{-1}$ is called renormalization constant.

Justification of DS theory

- An important property that in general is not true is *idempotency*: Bel⊕Bel = Bel

  (check it for the safecracker example)

- Main requirement for the proper working of the combination rule:

  The belief functions to be combined are actually based on entirely distinct bodies of evidence.

Complexity

- The task of finding all pairs *y* and *z* of subsets of θ such that y∩z = x is $o(2^{|\Theta|-|x|} \times 2^{|\Theta|-|x|})$. This is a painfully large number.

- Gordon & Shortliffe (*Artificial Intelligence* 26) describe how you can improve on this complexity by compromising with the rule of combination.

Dempster’s rule vs. MYCIN

Gordon & Shortliffe also compare Dempster's rule with some *ad hoc* rules that they used in the medical expert system MYCIN, coming to the following conclusions:

(i) Dempster's rule seems rather cleaner and better behaved than their own rules

(ii) If you have good expert rules then your program will behave well even with unclear unprincipled rules of combination, if you have poor expert rules then your program will behave poorly even with clear principled rules of combination.

Advantages of DS theory

(i) The difficult problem of specifying priors can be avoided

(ii) In addition to uncertainty, also ignorance can be expressed

(iii) It is straightforward to express pieces of evidence with different levels of abstraction

(iv) Dempster's combination rule can be used to combine pieces of evidence

Disadvantages

(i) Potential computational complexity problems

(ii) It lacks a well-established decision theory (whereas Bayesian decision theory maximizing expected utility is almost universally accepted.

(iii) Experimental comparisons between DS theory and probability theory seldom done and rather difficult to do; no clear advantage of DS theory shown.