Monotonicity, a deep property in data science

Bernard DE BAETS, Ghent University, Belgium

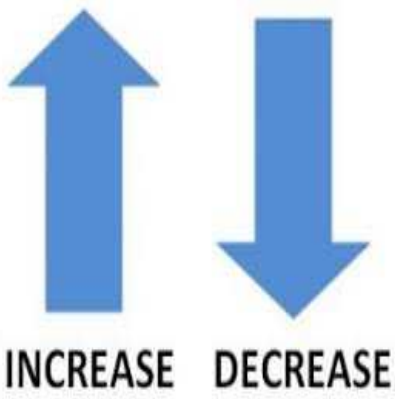
SFC2019

Nancy, France, 03/09/2019

The narrator

- **Training**: mathematician – computer scientist – knowledge engineer
- **Profession**: senior full professor in applied mathematics
- **Affiliation**: Faculty of Bioscience Engineering at Ghent University
- **Multi- and interdisciplinary research** in three interlaced threads: knowledge-based, predictive and spatio-temporal modelling
- **Ultimate aim**: innovative applications in the bio-engineering sciences

Today's main character in a three-act play: Monotonicity



Increasing/decreasing functions

A function $f : P \to P'$ between two **partially ordered sets** (posets) $(P, \leq)$ and $(P', \leq')$ is called

- **increasing** if $x \leq y$ implies $f(x) \leq' f(y)$
- **decreasing** if $x \leq y$ implies $f(y) \leq' f(x)$

Act I

DECEPTION

Example: Soil erosion

Phenomenon: loss of soil by erosion increases with increasing slope angle and decreasing soil coverage with vegetation

(Geoderma, Mitra et al., 1998)

slope angle class	forest	pasture
very large	moderately high	high
large	moderately low	moderately high
medium	low	moderately high
small	low	moderate
very small	low	low

land use class

Increasing, non-smooth rule base

Monotone fuzzy models?

Starting observations

- In many non-control applications (such as classification), fuzzy rule-based models are used for **one-shot decisions**
- At the level of linguistic terms, the underlying fuzzy rule base usually has some **flavor of monotonicity**
- However, is the resulting input-output function **effectively monotone?**

Fuzzy rule-based model

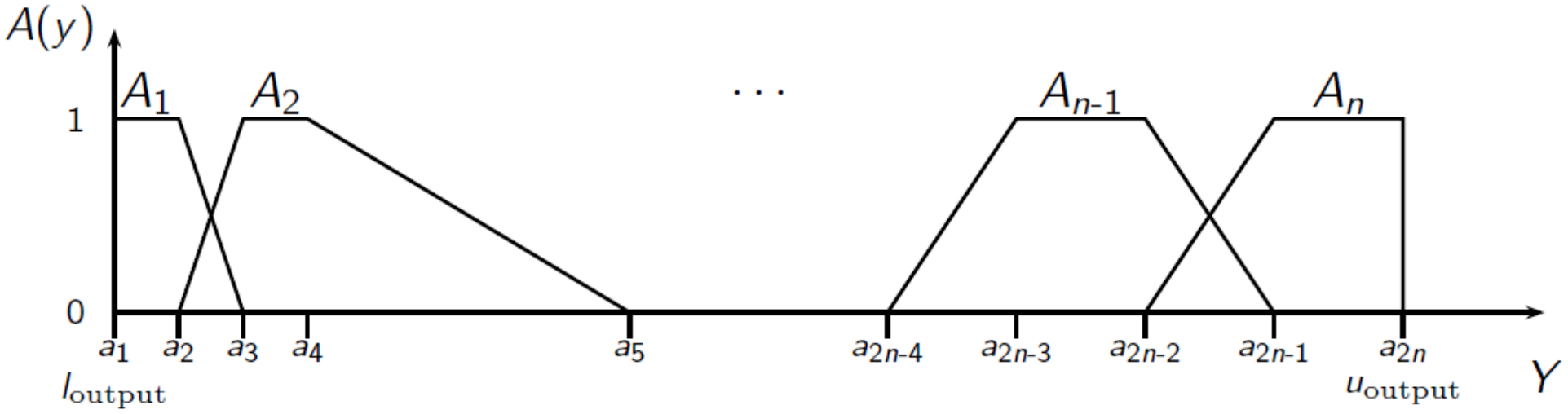
MISO model characteristics:

- m input variables X_ℓ and a single output variable Y
- rules of the form

$$R_s\text{: IF } X_1 \text{ IS } B^1_{j_{1,s}} \text{ AND } \ldots \text{ AND } X_m \text{ IS } B^m_{j_{m,s}} \text{ THEN } Y \text{ IS } A_{i_s}$$

- linguistic values $B^\ell_{j_{\ell,s}}$ of X_ℓ: **trapezoidal**; **Ruspini partition**
- linguistic values A_{i_s}: **trapezoidal**; **Ruspini partition** (bounded domain)
- **natural ordering** on the linguistic values of each variable

Ruspini partition



Mamdani–Assilian fuzzy models

Observation

Mamdani–Assilian fuzzy models with a monotone rule base do not necessarily result in a **monotone input-output mapping**

Monotone input-output behaviour under restrictive conditions only

If the original rule base is **complete** and **increasing**, then the input-output mapping can only be **increasing** in the following cases:

1. **Center-of-Gravity defuzzification**:
   - one input variable: basic t-norms $T_{\mathbf{M}}$, $T_{\mathbf{P}}$ and $T_{\mathbf{L}}$
   - two or three input variables: $T_{\mathbf{P}}$ and a **smooth** rule base
2. **Mean-of-Maxima defuzzification**:
   - one input variable: basic t-norms $T_{\mathbf{M}}$, $T_{\mathbf{P}}$ and $T_{\mathbf{L}}$
   - two or **more** input variables: $T_{\mathbf{M}}$ or $T_{\mathbf{P}}$, and a **smooth** rule base

Alternative approach

Trivial, yet crucial observations

Consider an **increasing** function $f : \mathbb{R} \to \mathbb{R}$ such that $f(0) = 0$, then:

- if $x \geq 0$, then $f(x) \geq 0$
- if $x \leq 0$, then $f(x) \leq 0$

Consequences for a fuzzy rule in an increasing rule base

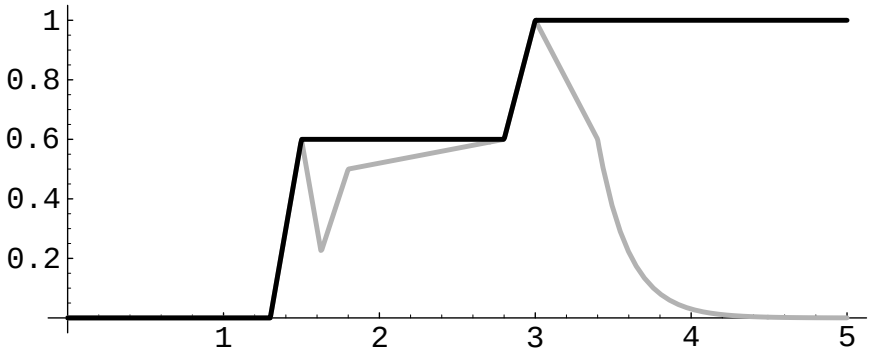
Consider a fuzzy rule "IF X IS C THEN Y IS D", then

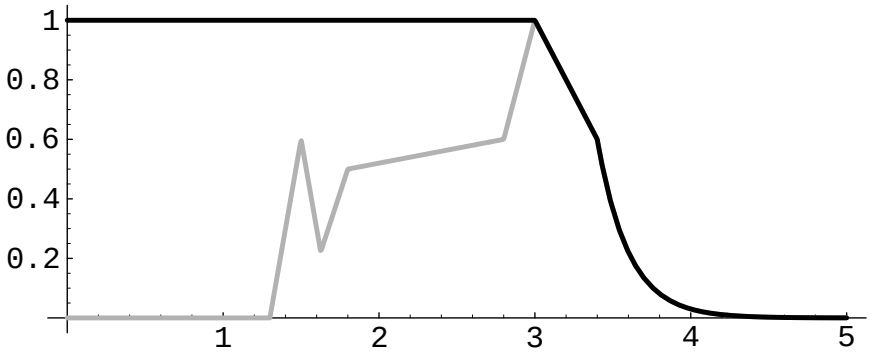
- IF X IS "**at least**" C THEN Y IS "**at least**" D
- IF X IS "**at most**" C THEN Y IS "**at most**" D

Cumulative modifiers of fuzzy sets

Cumulative modifiers

- **at-least modifier**: $\mathrm{ATL}(C)(x) = \sup\{C(t) \mid t \leq x\}$
- **at-most modifier**: $\mathrm{ATM}(C)(x) = \sup\{C(t) \mid t \geq x\}$





Implication-based fuzzy models (CRI)

Connectives: left-continuous t-norm and its residual implicator

Modifying an increasing rule base

- **ATL rule base**: applying ATL to all antecedents and consequents
- **ATM rule base**: applying ATM to all antecedents and consequents
- **ATLM rule base**: union of the above rule bases

Increasing input-output mapping

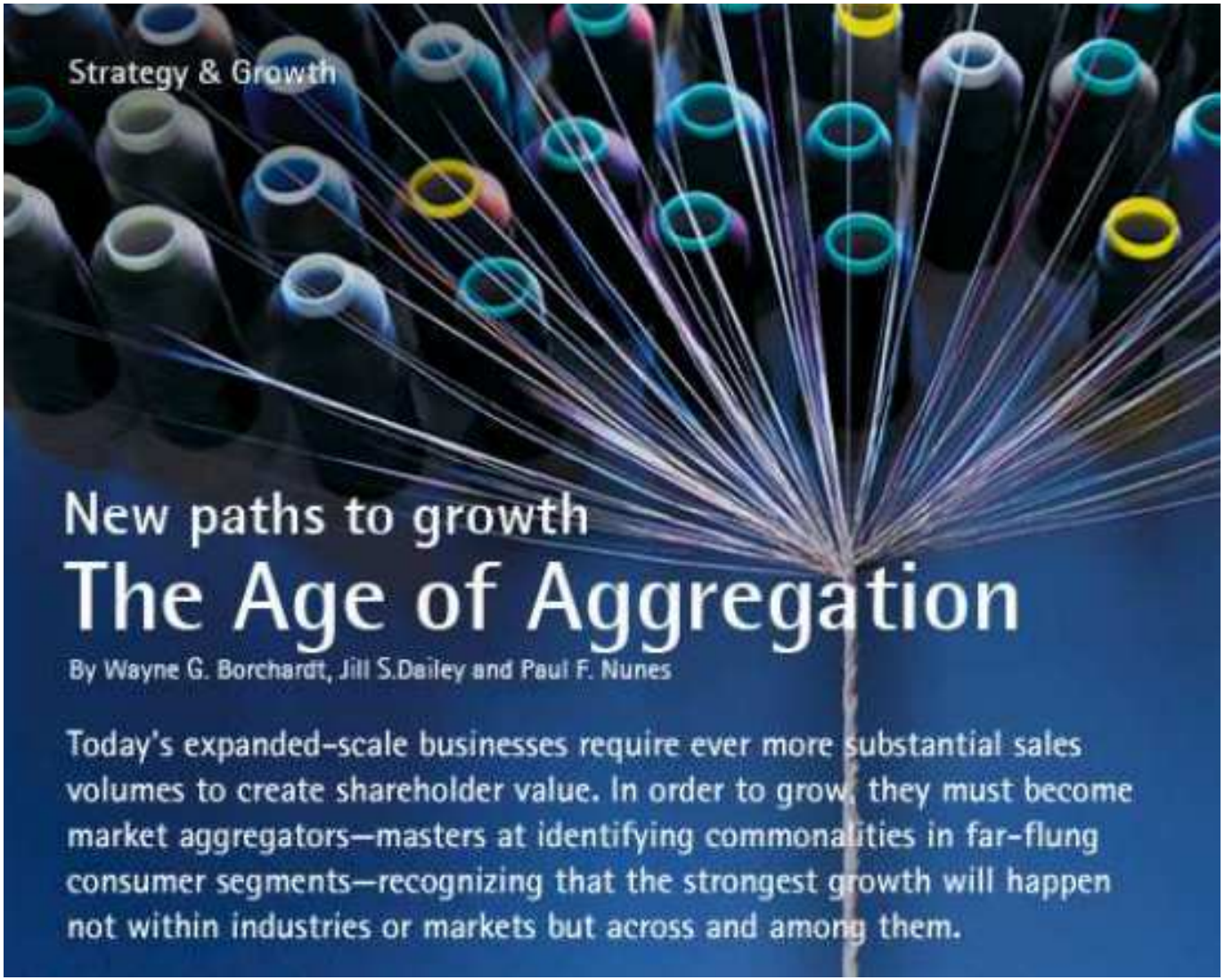
If the original rule base is **increasing**, then the input-output mapping is **increasing** in the following cases:

1. ATL rule base and **First-of-Maxima defuzzification**
2. ATM rule base and **Last-of-Maxima defuzzification**
3. ATLM rule base and **Mean-of-Maxima defuzzification**

Act II

OBSTRUCTION

The Age of Aggregation



The Age of Aggregation

Data aggregation has become a very successful business model

The battle is for the **customer interface**:

- **Uber**: the world's largest taxi company, owns no vehicles
- **Facebook**: the world's most popular media owner, creates no content
- **Alibaba**: the most valuable retailer, has no inventory
- **Airbnb:** the world's largest accommodation provider, owns no real estate

What processes do AGOP researchers study?

- Mathematically formalized by **aggregation functions**, formerly called **aggregation operators** (AGOPs)
- Historically, mostly **confined to real numbers**
- **Numerous examples** and **parametric families**: means, t-norms, t-conorms, uninorms, nullnorms, quasi-copulas, copulas, OWA operators, Sugeno integral, Choquet integral, ...
- Probably the **most important spin-off** of the fuzzy set community
- Monographs:
  - **Aggregation Functions** (2009) (Grabisch–Marichal–Mesiar–Pap)
  - **Aggregation Functions: A Guide for Practitioners** (2010) (Beliakov–Calvo–Pradera)
  - **A Practical Guide to Averaging Functions** (2015) (Beliakov–Bustince–Calvo)
  - etc.

Aggregation functions

Theory: Embarrassingly general

Consider a bounded poset $(P, \leq, \mathbf{0}, \mathbf{1})$ and $n \in \mathbb{N}$. A mapping $A : P^n \to P$ is called an n-ary **aggregation function** on $(P, \leq)$ if it satisfies:

1. $A(\mathbf{0}, \ldots, \mathbf{0}) = \mathbf{0}$ and $A(\mathbf{1}, \ldots, \mathbf{1}) = \mathbf{1}$
2. A is **increasing**: $\mathbf{x} \leq \mathbf{y} \Rightarrow A(\mathbf{x}) \leq A(\mathbf{y})$

Some comments

- Practice is embarrassingly narrow
- The poset context appears dogmatic
- Does not address data types of current interest

First example: Compositional data

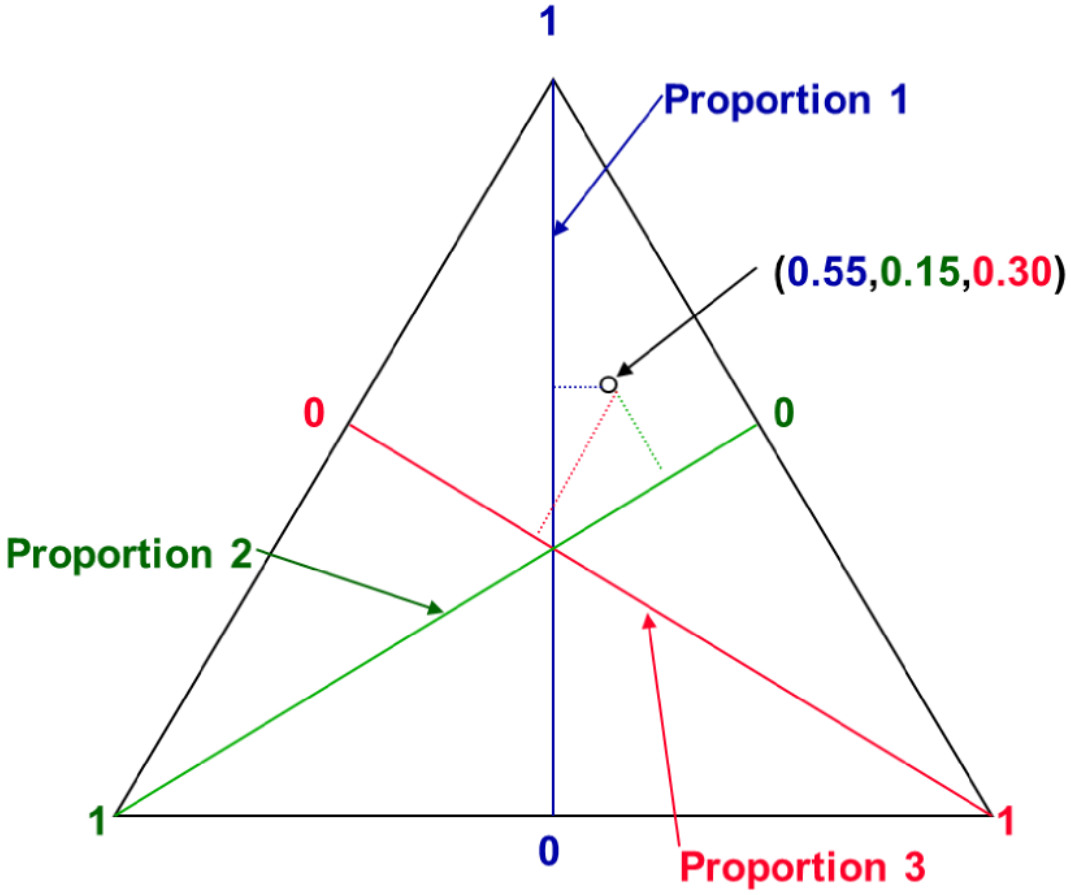
- k-dimensional compositional data vectors: simplex

$$\mathcal{S}_k = \{\mathbf{x} \in [0,1]^k \mid \sum_{i=1}^{k} x_i = 1\}$$

- Examples of application:
  - **soil science**: relative portions of sand, clay and silt in a soil sample
  - **chemistry**: compositions expressed as molar concentrations
  - **environmental science**: composition of air pollution
  - **mathematics**: weight vector of a weighted quasi-arithmetic mean
  - **fuzzy set theory**: vector of membership degrees in fuzzy c-means
  - **probability theory**: discrete probability distribution

Illustration: food composition (in %) ($k = 3$)

Food composition (% fat, % carbonates, % protein) in **barycentric coordinates**



Mixing compositions

We can "aggregate" compositional data vectors componentwisely resulting in a new compositional data vector: $\mathrm{C} : (\mathcal{S}_k)^n \to \mathcal{S}_k$

$$\mathrm{C}(\mathbf{x}^1, \ldots, \mathbf{x}^n)_j = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}^i)_j$$

- The set $\mathcal{S}_k$ is **not** a poset:
  - there is **no** natural ordering
  - there is **no** smallest or largest element

The function C can be written as

$$\mathrm{C}(\mathbf{x}^1, \ldots, \mathbf{x}^n)_j = \arg\min_y \sum_{i=1}^{n} ((\mathbf{x}^i)_j - y)^2$$

Second example: Ranking data

- Examples of application:
  - Traditionally: voting, decision making, preference modelling
  - Nowadays: high-throughput, omics-scale, biological data, e.g. ranking of genes
- Different problem settings:
  - full rankings
  - incomplete rankings; top-k lists
- The set of (full) rankings $\mathcal{L}(\mathcal{C})$ (briefly, $\mathcal{L}$) is **not** a poset:
  - there is **no** natural ordering
  - there is **no** smallest or largest element

Aggregation methods for full rankings

- **Borda methods**: apply aggregation functions to the ranks (possibly leading to ties, resulting in a weak order)
- **Distance-based methods**: consider n full rankings $\succ_i$

$$A(\succ_1, \ldots, \succ_n) = \arg\min_{\succ} \sum_{i=1}^{n} d(\succ_i, \succ)$$

where $d(\succ_i, \succ)$ is:

- **Kendall's distance function** K
  (number of pairwise discordances)

  or
- Spearman's footrule distance function S
  (sum of the absolute differences between the ranks)

Penalty functions

Penalty function

Let $I = [a, b] \subseteq \mathbb{R}$. A function $P : I \times I^n \to \mathbb{R}$ is a **penalty function** if

1. $P(y; \mathbf{x}) \geq 0$
2. $P(y; \mathbf{x}) = 0$ if and only if $\mathbf{x} = (y, \ldots, y)$
3. $P(\cdot; \mathbf{x})$ is quasi-convex and lower-semicontinuous

(The third condition implies that the set of minimizers of $P(\cdot; \mathbf{x})$ is either a singleton or an interval)

Penalty-based (aggregation) functions

Penalty-based function

Given a penalty function P, the corresponding **penalty-based function** is the function $f : I^n \to I$ defined by

$$f(\mathbf{x}) = \frac{\ell(\mathbf{x}) + r(\mathbf{x})}{2}$$

where $[\ell(\mathbf{x}), r(\mathbf{x})]$ is the interval closure of the set of minimizers of $P(\cdot; \mathbf{x})$

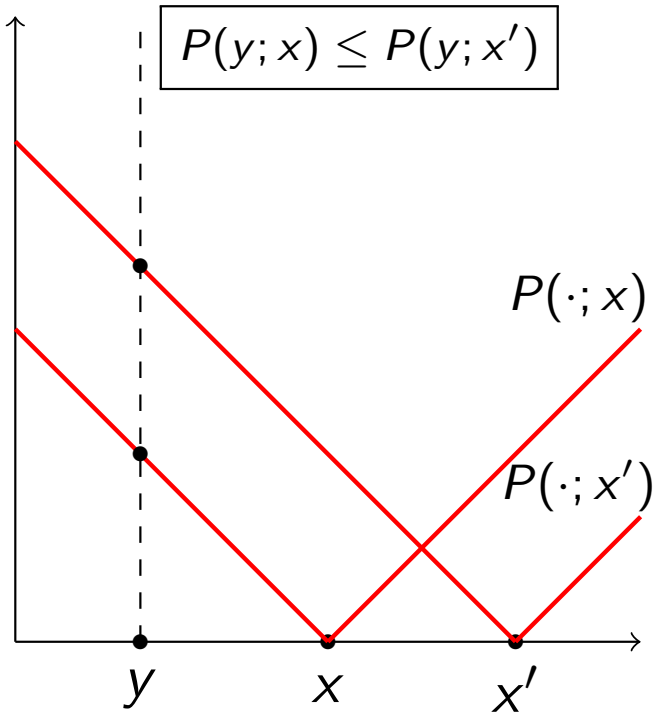
Aggregation function?

A penalty-based function f is not necessarily increasing

Remark

Originally, the following condition has been required for a (local) penalty function ($n = 1$):

$$\text{if } x' \leq x \leq y \text{ or } y \leq x \leq x', \text{ then } P(y;x) \leq P(y;x')$$



Betweenness relations instead of order relations

Betweenness relation

A **ternary** relation B on X is called a **betweenness relation** (BR) if:

1. **Symmetry in the end points**: $(a, b, c) \in B \Leftrightarrow (c, b, a) \in B$
2. **Closure**: $\big((a, b, c) \in B \wedge (a, c, b) \in B\big) \Leftrightarrow b = c$
3. **End-point transitivity**:

$$((o, a, b) \in B \wedge (o, b, c) \in B) \Rightarrow (o, a, c) \in B$$

Product betweenness relation on X^n

The ternary relation $B^{(n)}$ on X^n defined by

$$(\mathbf{a}, \mathbf{b}, \mathbf{c}) \in B^{(n)} \quad \Leftrightarrow \quad (\forall i \in \{1, \ldots, n\})((a_i, b_i, c_i) \in B)$$

Examples

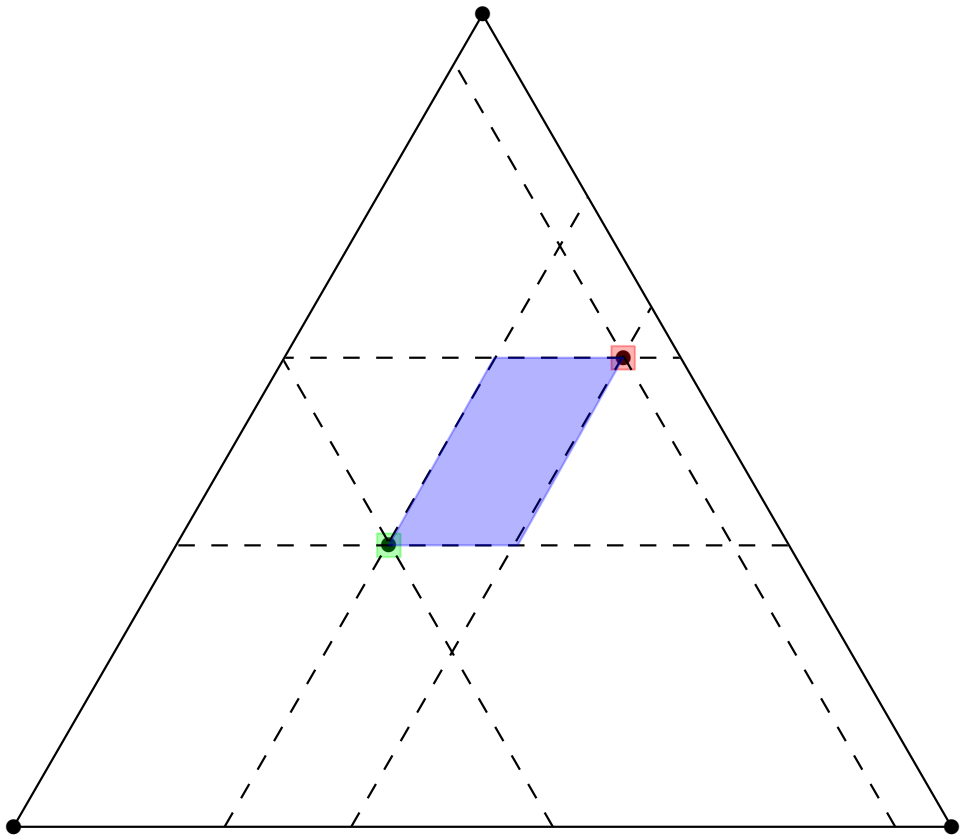
Examples (order relation $\leq$, distance function d)

1. $B_0 = \{(x, y, z) \in X^3 \mid x = y \vee y = z\}$ (trivial BR)
2. $B_{\leq} = B_0 \cup \{(x, y, z) \in X^3 \mid (x \leq y \leq z) \vee (z \leq y \leq x)\}$
3. $B_d = \{(x, y, z) \in X^3 \mid d(x, z) = d(x, y) + d(y, z)\}$

A betweenness relation on compositional data

A natural betweenness relation: $B_{\mathcal{S}_k} := (B_{[0,1]})^{(k)} \cap (\mathcal{S}_k)^3$

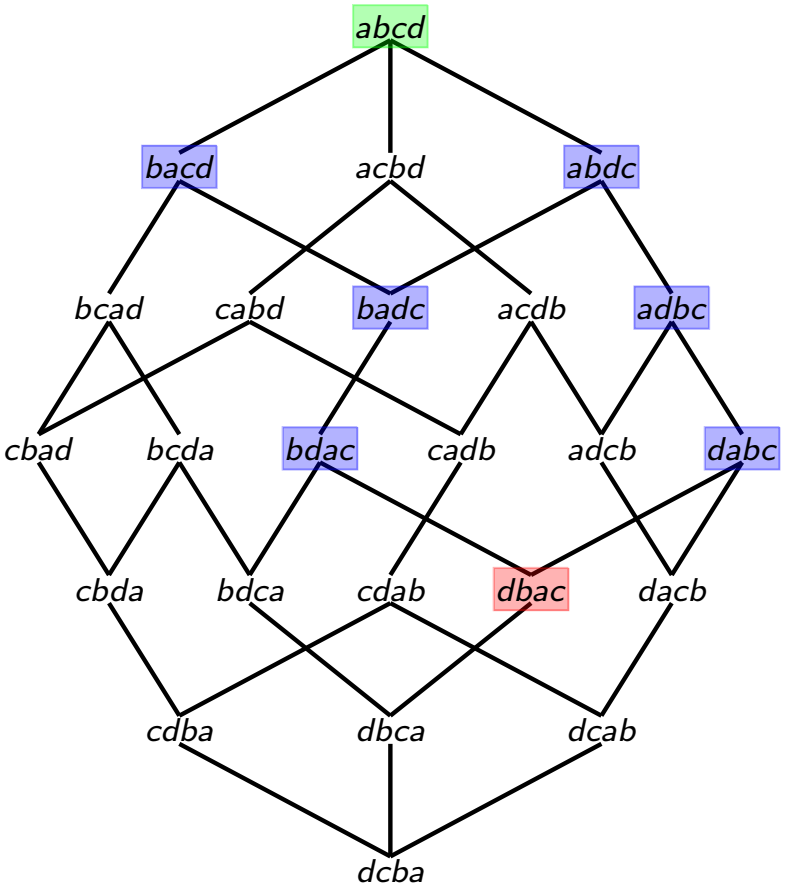
$$(\mathbf{x}, \mathbf{y}, \mathbf{z}) \in (B_{[0,1]})^{(k)} \Leftrightarrow (\forall i \in \{1, \ldots, k\})(\min(x_i, z_i) \leq y_i \leq \max(x_i, z_i))$$



A betweenness relation on rankings

Betweenness relation based on Kendall's d.f.:

$$(\succ_1, \succ_2, \succ_3) \in B_K \quad \Leftrightarrow \quad K(\succ_1, \succ_3) = K(\succ_1, \succ_2) + K(\succ_2, \succ_3)$$



Generalized penalty functions

Definition

Consider $n \in \mathbb{N}$, a set X and a BR $\mathbf{B}$ on X^n. A function $P : X \times X^n \to \mathbb{R}$ is called a **penalty function** (compatible with $\mathbf{B}$) if

(P1) $P(y; \mathbf{x}) \geq 0$

(P2) $P(y; \mathbf{x}) = 0$ if and only if $\mathbf{x} = (y, \ldots, y)$

(P3) The set of minimizers of $P(\cdot; \mathbf{x})$ is always non-empty

(P4) $P(y; \mathbf{x}) \leq P(y; \mathbf{x}')$, whenever $((y, \ldots, y), \mathbf{x}, \mathbf{x}') \in \mathbf{B}$

Generalized penalty functions

Optional conditions for fixed x

(P5) For any minimizer $z \in X$ of $P(\cdot; \mathbf{x})$ such that

$$((z, \ldots, z), (y, \ldots, y), (y', \ldots, y')) \in \mathbf{B}$$

it holds that $P(y; \mathbf{x}) \leq P(y'; \mathbf{x})$

(P6) For any two minimizers $z, z' \in X$ of $P(\cdot; \mathbf{x})$ such that

$$((z, \ldots, z), (y, \ldots, y), (z', \ldots, z')) \in \mathbf{B}$$

it holds that $P(y; \mathbf{x}) = P(z; \mathbf{x})$

Penalty-based function

Given a penalty function P, the corresponding **penalty-based function** is the function $f : X^n \to \mathcal{P}(X)$ such that $f(\mathbf{x})$ is the set of minimizers of $P(\cdot; \mathbf{x})$

How to create penalty functions?

Monometric

A mapping $M : X^2 \to \mathbb{R}$ is called a **monometric** w.r.t. a betweenness relation B on X if it satisfies

1. **Non-negativity**: $M(x, y) \geq 0$
2. **Coincidence**: $M(x, y) = 0 \Leftrightarrow x = y$
3. **Compatibility**: if $(x, y, z) \in B$, then $M(x, y) \leq M(x, z)$

Proposition

A distance function $d : X^2 \to \mathbb{R}$ is a monometric w.r.t.

$$B_d = \{(x, y, z) \in X^3 \mid d(x, z) = d(x, y) + d(y, z)$$

Monometric-based penalty functions

Monometric M on X w.r.t. B

The function $P : X^{n+1} \to \mathbb{R}^+$ defined by

$$P(y; \mathbf{x}) = A(M(y, x_1), \ldots, M(y, x_n)),$$

is a penalty function (compatible with the betweenness relation $B^{(n)}$ on X^n) if A is an n-ary **increasing** function such that $A(x_1, \ldots, x_n) = 0$ iff $x_i = 0$ for $i = 1, \ldots, n$

Particular cases: addition and maximum

$$P(y; \mathbf{x}) = \sum_{i=1}^{n} M(y, x_i) \qquad P(y; \mathbf{x}) = \max_{i=1}^{n} M(y, x_i)$$

Key examples of penalty-based aggregation

- **Averaging compositional data**:
  - satisfies (P5) and (P6)
- **Method of Kemeny for rankings**:
  - satisfies (P5)
  - satisfies (P6) for those profiles of rankings for which there exists a **Condorcet ranking**
- **The center string procedure**:

$$C(S_1, \ldots, S_n) = \arg\min_S \max_{i=1}^{n} d_H(S_i, S)$$

where d_H is the Hamming distance between strings of the same length

  - neither satisfies (P5) nor (P6)

Trends in a related area: Machine Learning

- Originally: shared interest with **statistics** in classification and regression problems (focus on generalization abilities rather than inference)
- Currently: focus on a **broad range of problem settings** involving more and more complex data (at the input as well as the output side)
  - **classification** (multi-label, hierarchical, extreme)
  - **regression** (ordinal, monotone)
  - **structured prediction** or structured (output) learning
  - **preference learning** (label ranking, instance ranking)
  - **pairwise learning**
  - **relational learning**
  - **multi-task learning**
  - and so on
- One commonality: all models (i.e. functions) are the result of solving a mathematical **optimization problem**

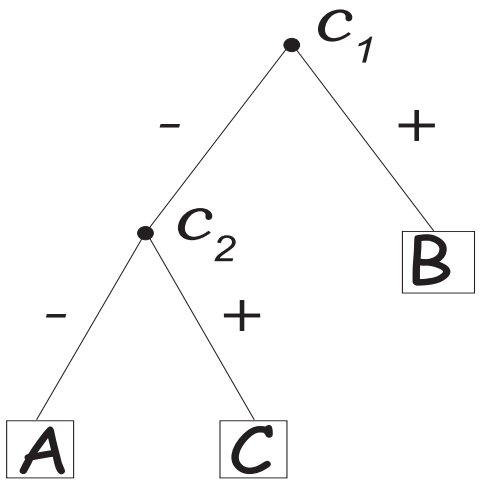
Act III

SHORT-SIGHTEDNESS

Toy example

Classification problem:

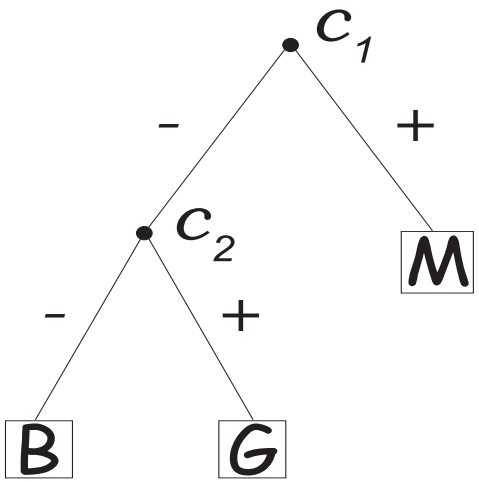
	c_1	c_2	c_3	class label
a_1	$-$	$-$	$+$	A
a_2	$+$	$-$	$-$	B
a_3	$-$	$+$	$+$	C
a_4	$+$	$+$	$-$	B



Toy example

Monotone classification problem:

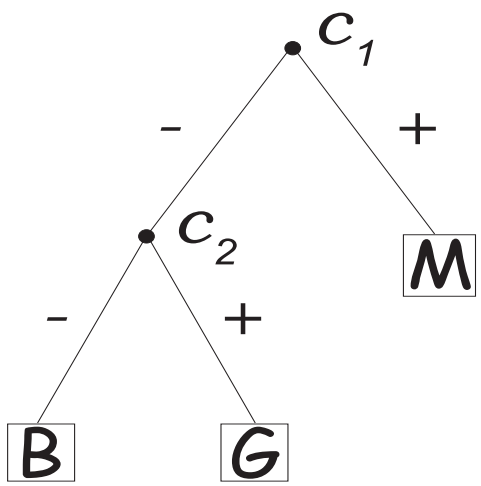
	c_1	c_2	c_3	evaluation
a_1	$-$	$-$	$+$	Bad
a_2	$+$	$-$	$-$	Moderate
a_3	$-$	$+$	$+$	Good
a_4	$+$	$+$	$-$	Moderate



Toy example

Monotone classification problem:

	c_1	c_2	c_3	evaluation
a_1	$-$	$-$	$+$	Bad
a_2	$+$	$-$	$-$	Moderate
a_3	$-$	$+$	$+$	Good
a_4	$+$	$+$	$-$	Moderate
a_5	$-$	$+$	$-$	**Good**
a_6	$+$	$+$	$+$	**Moderate**



If monotonicity applies, any violation of it is simply unacceptable

How to produce guaranteed **monotone** classification results, even when the set of learning examples is **not monotone**?

Multi-class classification

- Problem: to assign labels from a finite set $\mathcal{L}$ to the elements of some set of **objects** Ω
- Each object $a \in \Omega$ is represented by a **feature vector**

$$\mathbf{a} = (c_1(a), c_2(a), \ldots, c_n(a))$$

  in the **feature space** $\mathcal{X}$
- **Collection of learning examples**: **multiset**

$$(\mathcal{S}, d) \equiv \{\langle \mathbf{a}, d(a) \rangle \mid a \in \mathcal{S}\}$$

  where:
  - $\mathcal{S} \subseteq \Omega$ is a given set of objects
  - $d \colon \mathcal{S} \to \mathcal{L}$ is the associated decision function
  - **multiset**: the same entry can occur more than once, usually giving this entry more importance: we do **not** write $\langle \mathbf{a}, d(\mathbf{a}) \rangle$
  - notation: $\mathcal{S}_{\mathcal{X}} = \{\mathbf{a} \mid a \in \mathcal{S}\}$

Multi-class classification

- **Goal** of supervised classification algorithms:
  - extend the function d to Ω in the most reasonable way
  - concentrate on finding a function $\lambda\colon \mathcal{X} \to \mathcal{L}$ that **minimizes the expected loss** on **an independent set of test examples**
- Different approaches:
  - **instance-based**, such as nearest neighbour methods
  - **model-based**, such as classification trees
- **Distribution classifiers**: output is a PMF over $\mathcal{L}$
  - mathematically: $\tilde{\lambda}\colon \mathcal{X} \to \mathcal{F}(\mathcal{L})$
  - selecting a single label: Bayesian decision
    (label with the highest probability is returned)

Multi-criteria evaluation

- In many cases, $\mathcal{L}$ exhibits a natural ordering and could be treated as an ordinal scale (chain): **ordinal classification/regression**
- Often, objects are described by (true) **criteria** $(c_i, \leq_{c_i})$ (chains)
- The **product ordering** turns $\mathcal{X}$ into a **partially ordered set** $(\mathcal{X}, \leq_{\mathcal{X}})$ (poset)
- Multi-criteria evaluation: quality assessment, environmental data, social surveys, etc.

Natural monotonicity constraint

An object a that scores at least as good on all criteria as an object b must be classified (ranked) at least as good as object b

Monotone classification

Monotone classifier

Classifier + **basic monotonicity constraint**:

$$\mathbf{x} <_{\mathcal{X}} \mathbf{y} \Rightarrow \lambda(\mathbf{x}) \leq_{\mathcal{L}} \lambda(\mathbf{y})$$

(supervised ranking/ordered sorting, monotone ordinal regression)

Monotone distribution classifier

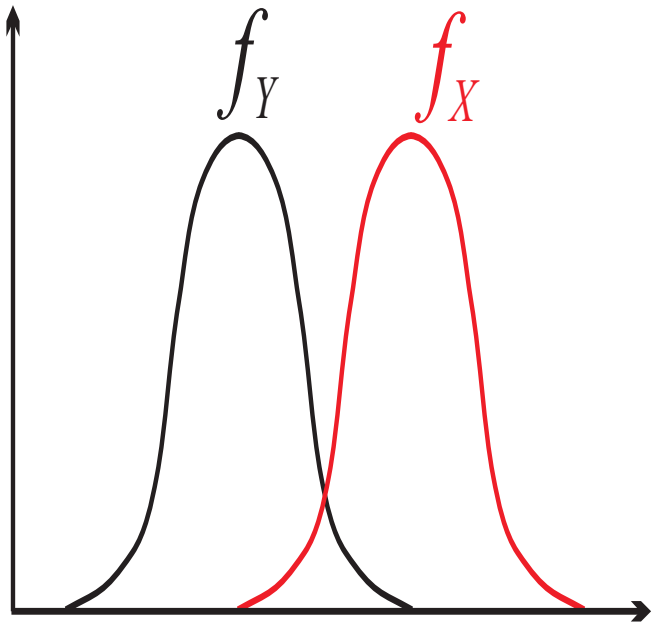
Distribution classifier + **stochastic monotonicity constraint**:

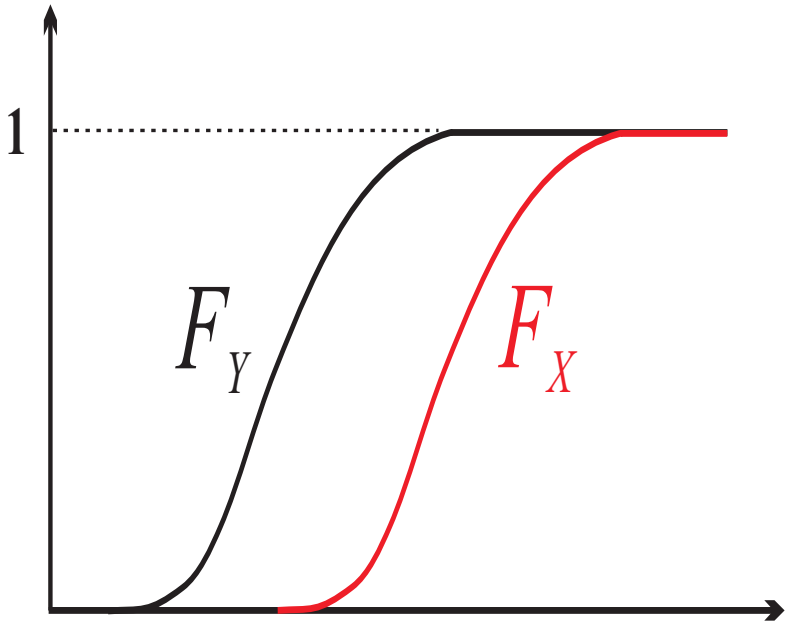
$$\mathbf{x} <_{\mathcal{X}} \mathbf{y} \Rightarrow \tilde{\lambda}(\mathbf{x}) \preceq_{\mathrm{SD}} \tilde{\lambda}(\mathbf{y})$$

(First order) Stochastic Dominance (SD):

$$f_X \preceq_{\mathrm{SD}} f_Y \Leftrightarrow F_X \geq F_Y$$

Stochastic dominance





Selecting a single label

- Bayesian decision potentially breaks the desired monotonicity and is **no longer acceptable** in this case
- The well-known relationship

$$f_X \preceq_{\mathrm{SD}} f_Y \Rightarrow \mathrm{E}[f_X] \leq \mathrm{E}[f_Y]$$

  cannot be used as it requires the transformation of the ordinal scale into a numeric scale
- **Set of medians** (interval) of f_X:

$$\mathrm{med}(f_X) = \{\ell \in \mathcal{L} \,|\, \mathcal{P}\{X \leq \ell\} \geq 1/2 \wedge \mathcal{P}\{X \geq \ell\} \geq 1/2\}$$

  - reduces in the continuous case to the median m: $\mathcal{P}\{X \leq m\} = 1/2$
  - only endpoints of the interval have non-zero probability

Selecting a single label from the set of medians

- The set of medians reduces the PMF to an interval. Does there exist an ordering on intervals that is compatible with FSD?

$$[k_1, \ell_1] \preceq_{\mathcal{L}}^{[2]} [k_2, \ell_2] \Leftrightarrow \left(k_1 \preceq_{\mathcal{L}} k_2 \wedge \ell_1 \preceq_{\mathcal{L}} \ell_2\right)$$

- New relationship:

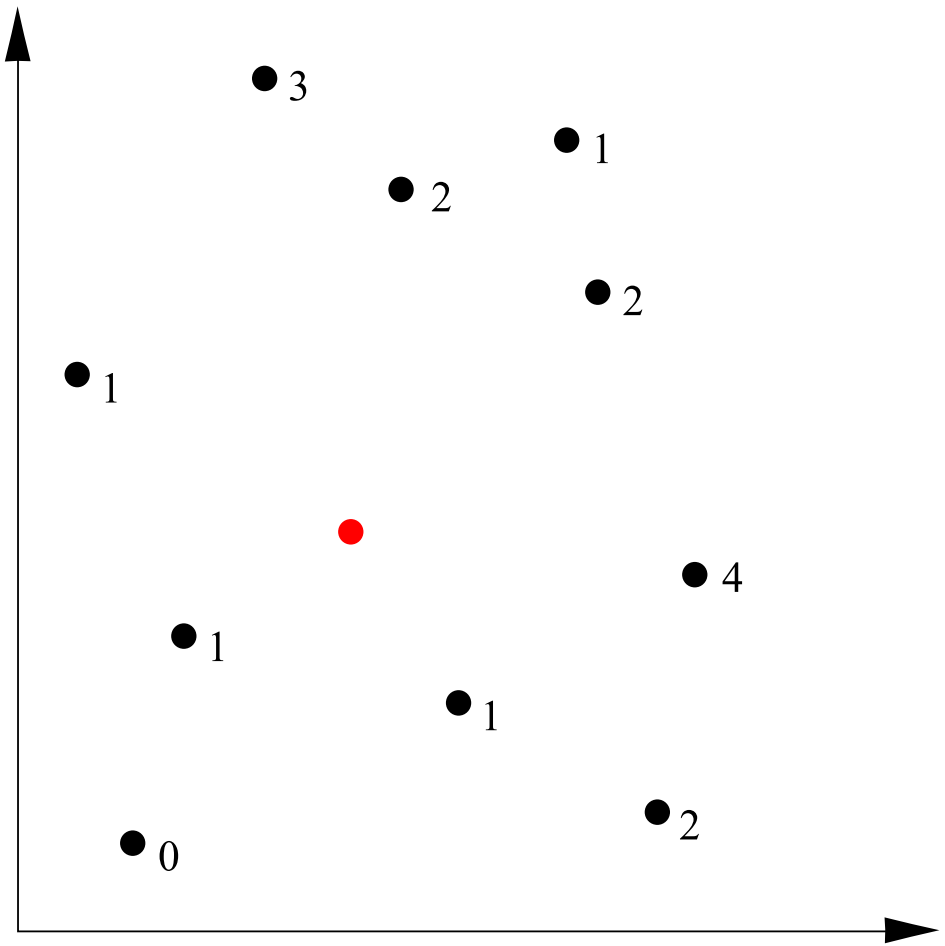
$$f_X \preceq_{\mathrm{SD}} f_Y \Rightarrow \mathrm{med}(f_X) \preceq_{\mathcal{L}}^{[2]} \mathrm{med}(f_Y)$$

Selecting a single label

1. **Pessimistic median** (lower)
2. **Optimistic median** (upper)
3. Midpoint (or smaller/greater of the two midpoints) [not meaningful]

turn a monotone distibution classifier into a monotone classifier

How to label a new point?



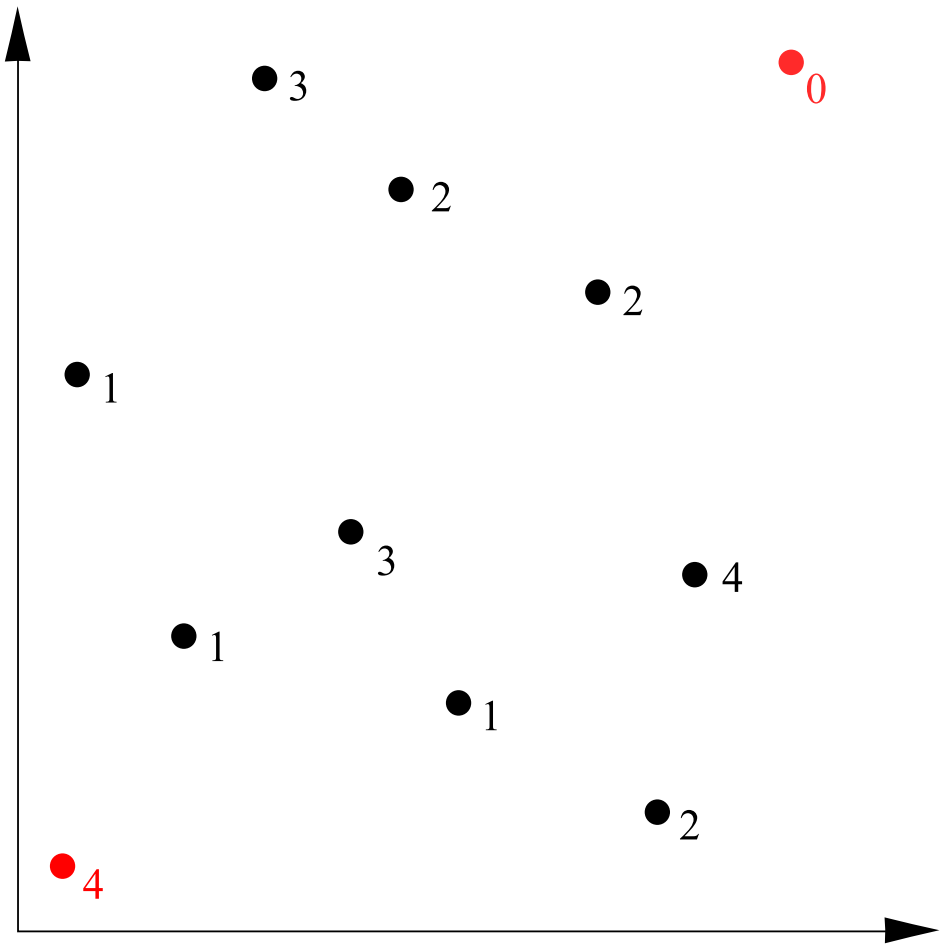
Minimal and maximal extensions

1. **Minimal Extension**: $\lambda_{\min} : \mathcal{X} \to \mathcal{L}$
   - assigns best label of **"objects below"**:
   $$\lambda_{\min}(\mathbf{x}) = \max\{d(s) \mid \mathbf{s} \in \mathcal{S}_{\mathcal{X}} \wedge \mathbf{s} \leq_{\mathcal{X}} \mathbf{x}\}$$
   - if no such object: $\lambda_{\min}(\mathbf{x}) = \min(\mathcal{L})$
2. **Maximal Extension**: $\lambda_{\max} : \mathcal{X} \to \mathcal{L}$
   - assigns worst label of **"objects above"**:
   $$\lambda_{\max}(\mathbf{x}) = \min\{d(s) \mid \mathbf{s} \in \mathcal{S}_{\mathcal{X}} \wedge \mathbf{x} \leq_{\mathcal{X}} \mathbf{s}\}$$
   - if no such object: $\lambda_{\max}(\mathbf{x}) = \max(\mathcal{L})$

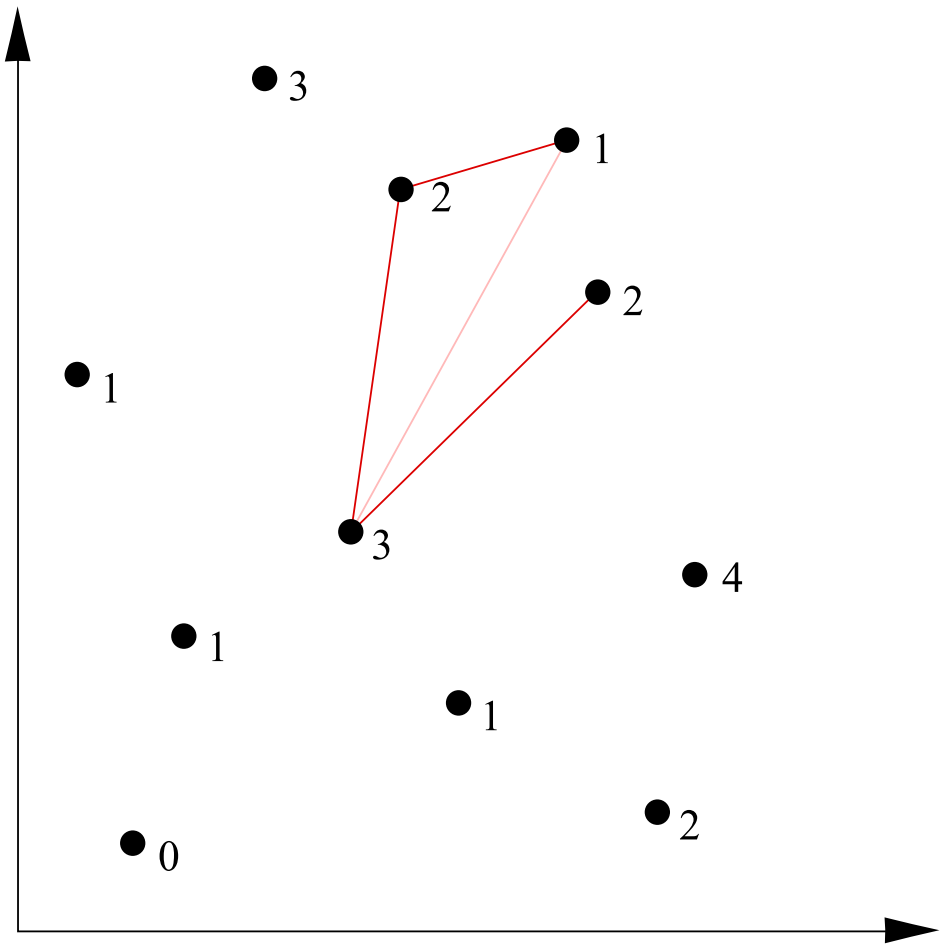
Monotone classifiers

1. $\lambda_{\min}$ and $\lambda_{\max}$ are monotone classifiers
2. **Interpolation**: midpoint leads to a monotone classifier

Things can go dead wrong



A more realistic non-monotone data set



Noise in multi-criteria evaluation

- $(\mathcal{S}, d)$ is called **monotone** if for all x and y in $\mathcal{S}$

$$\mathbf{x} = \mathbf{y} \Rightarrow d(x) = d(y)$$

(absence of doubt/ambiguity)

and

$$\mathbf{x} <_{\mathcal{X}} \mathbf{y} \Rightarrow d(x) \leq_{\mathcal{L}} d(y)$$

(absence of reversed preference)

- Non-monotonicity defines a **symmetric** and **transitive** relation on $\mathcal{S}$

Monotone extensions

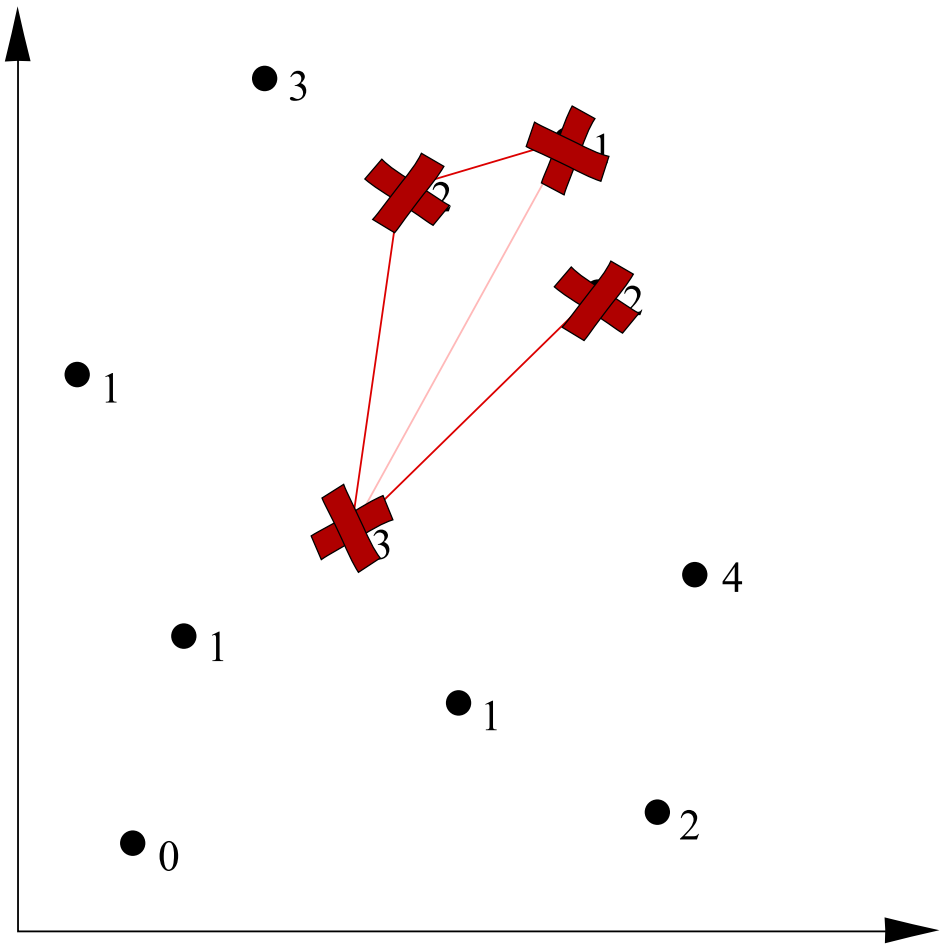
If the data set is monotone, then

1. $\lambda_{\min}$ and $\lambda_{\max}$ are monotone **extensions** of d to $\mathcal{X}$
2. any monotone extension λ of d to $\mathcal{X}$: $\lambda_{\min} \leq_{\mathcal{L}} \lambda \leq_{\mathcal{L}} \lambda_{\max}$

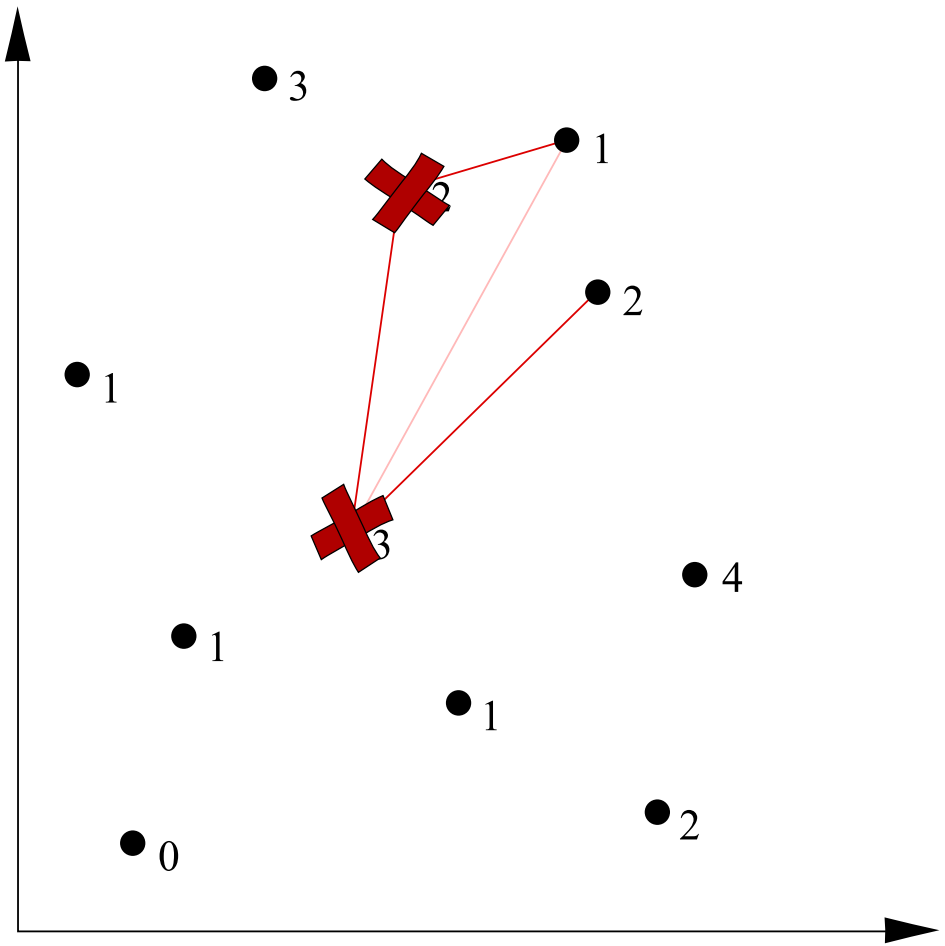
How to handle noise?

1. **Data reduction**: **identify** the noisy objects and **delete** them
2. **Data relabelling**: **identify** the noisy objects and **relabel** them
3. **Non-invasive approach**: keep the data set as is
   - excludes the use of some monotone classification algorithms
   - restricts the accuracy of any monotone classifier (**independence number**)

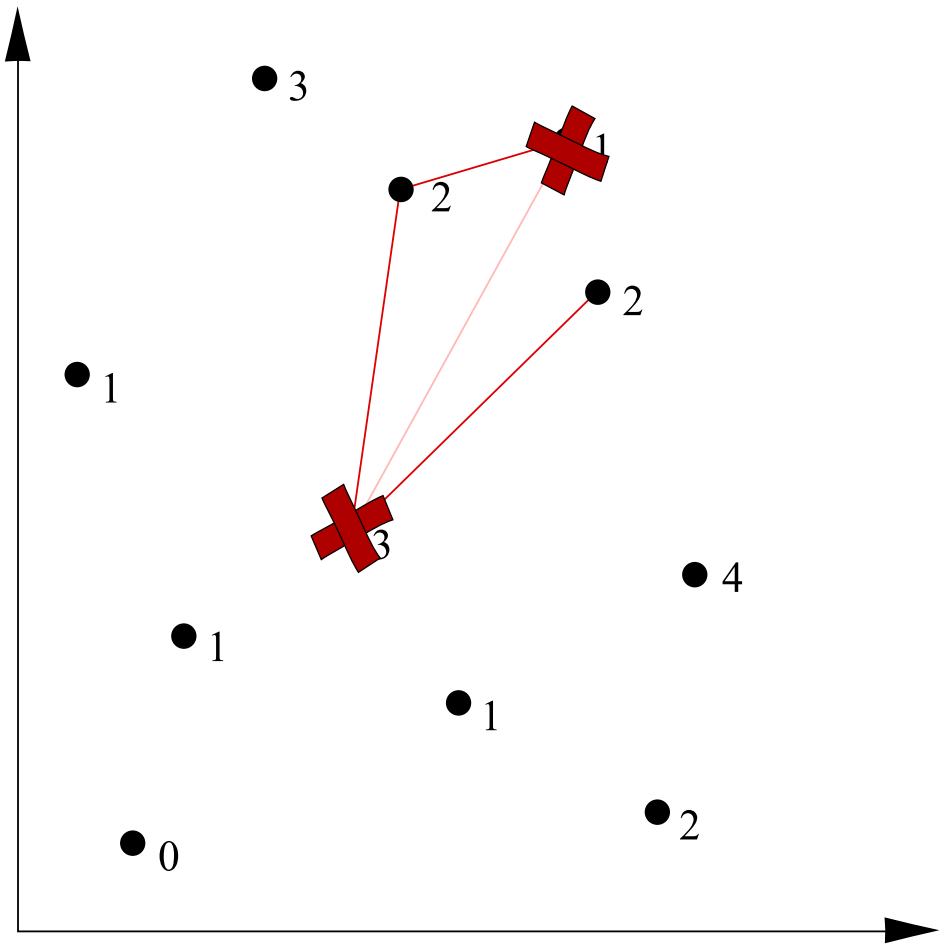
Option 1, Data reduction: A non-monotone data set



Option 1, Data reduction: A non-monotone data set



Option 1, Data reduction: A non-monotone data set

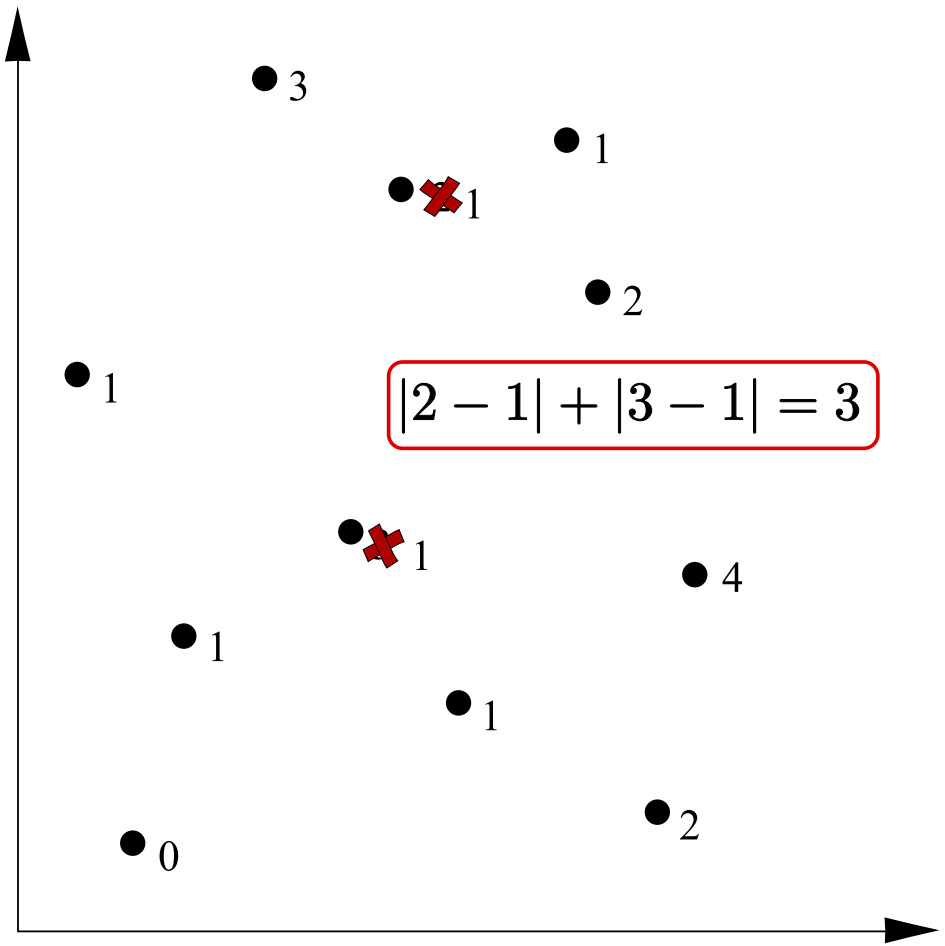


The maximum independent set problem

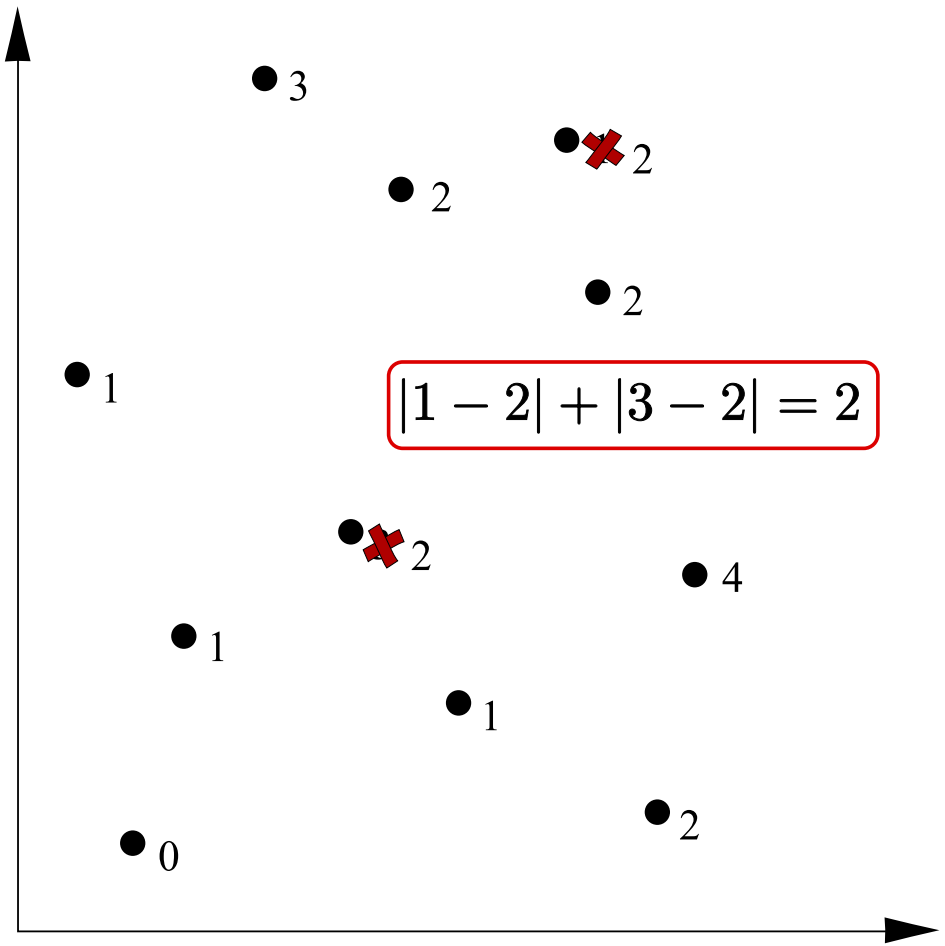
The non-monotonicity relation corresponds to a **comparability graph**:

- A monotone subset corresponds to an **independent set** of this graph
- **Maximal independent set** = independent set that is not a subset of any other independent set
- **Maximum independent set** (MIS) = independent set of biggest cardinality (= **independence number** α)
- A MIS in a comparability graph can be determined using **network flow theory** (cubic time complexity)

Option 2, Data relabelling: which MIS to select?



Option 2, Data relabelling: which MIS to select?



Option 2, Data relabelling: options

Universal tool: weighted MIS problems and **network flow theory**

1. **Optimal ordinal relabelling**: relabelling a minimum number of objects, of which all corona objects are relabelled to a minimum extent
2. **Optimal cardinal relabelling** (identifying $\mathcal{L}$ with the first n integers): minimal relabelling loss
   - zero-one loss: MIS
   - broad class of loss functions, including L1 loss and squared loss
3. **Optimal hierarchical cardinal relabelling** (single pass):
   - minimizing loss while relabelling a minimal number of objects
   - relabelling a minimal number of objects while minimizing loss

Distribution representation of a data set

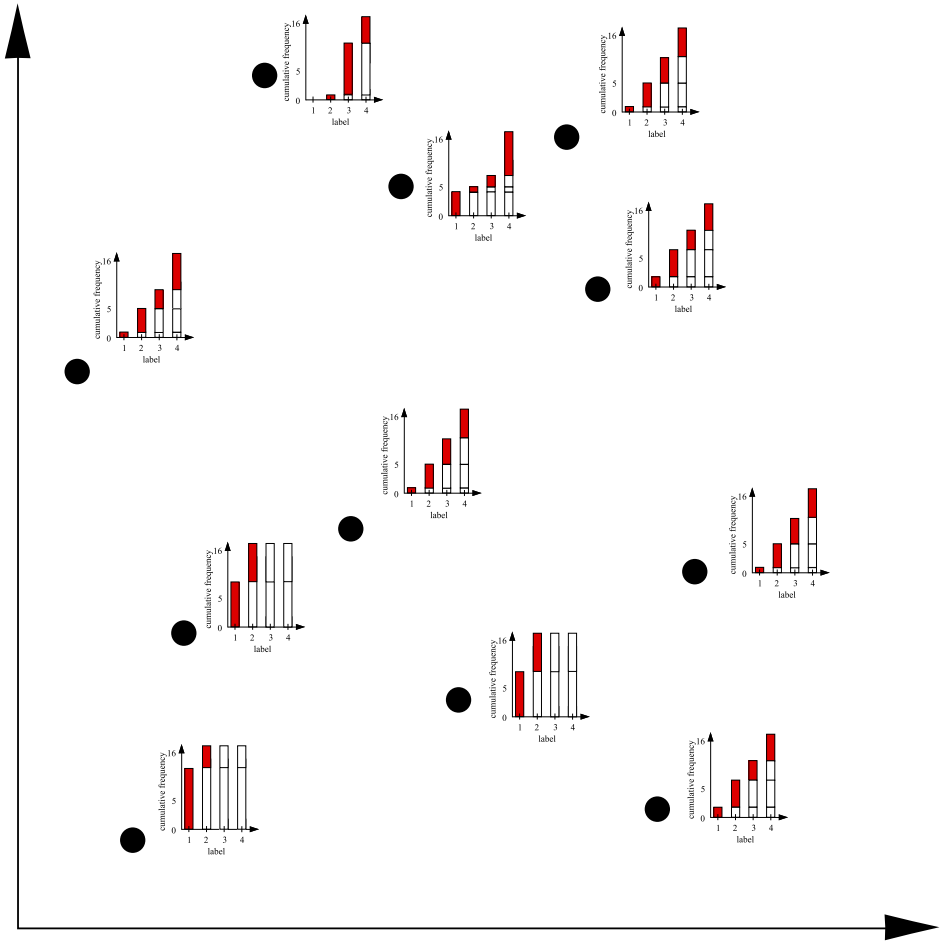
- Collection of learning examples $(\mathcal{S}, d)$
- For each $\mathbf{x} \in \mathcal{S}_{\mathcal{X}}$, a CDF $\hat{F}(\mathbf{x}, \cdot)\colon \mathcal{L} \to [0,1]$ is built from the collection of learning examples

$$\hat{F}(\mathbf{x}, \ell) = \frac{|\{a \in \mathcal{S} \,|\, \mathbf{a} = \mathbf{x} \wedge d(a) \leq_{\mathcal{L}} \ell\}|}{|\{a \in \mathcal{S} \,|\, \mathbf{a} = \mathbf{x}\}|}$$

(cumulative relative frequency distribution)

- The distribution data set $(\mathcal{S}_{\mathcal{X}}, \hat{F})$

A distribution data set



Stochastic minimal and maximal extensions

1. **Minimal Extension**: $F_{\min} : \mathcal{X} \times \mathcal{L} \to [0,1]$

$$F_{\min}(\mathbf{x}, \ell) = \min\{\hat{F}(\mathbf{s}, \ell) \mid \mathbf{s} \in \mathcal{S}_{\mathcal{X}} \wedge \mathbf{s} \leq_{\mathcal{X}} \mathbf{x}\}$$

   - if no such object: $f_{\min}(\mathbf{x}, \min(\mathcal{L})) = 1$

2. **Maximal Extension**: $F_{\max} : \mathcal{X} \times \mathcal{L} \to [0,1]$

$$F_{\max}(\mathbf{x}, \ell) = \max\{\hat{F}(\mathbf{s}, \ell) \mid \mathbf{s} \in \mathcal{S}_{\mathcal{X}} \wedge \mathbf{x} \leq_{\mathcal{X}} \mathbf{s}\}$$

   - if no such object: $f_{\max}(\mathbf{x}, \max(\mathcal{L})) = 1$

Monotone distribution classifiers

1. $F_{\min}$ and $F_{\max}$ are monotone distribution classifiers
2. **Interpolation**: for any $S \in [0,1]$, the mapping

$$\tilde{F} = S\, F_{\min} + (1 - S)\, F_{\max}$$

is also a monotone distribution classifier

Monotone distribution data sets

- $(\mathcal{S}_{\mathcal{X}}, \hat{F})$ is called **monotone** if for all $\mathbf{x}$ and $\mathbf{y}$ in $\mathcal{S}_{\mathcal{X}}$

$$\mathbf{x} <_{\mathcal{X}} \mathbf{y} \Rightarrow \hat{F}(\mathbf{x}, \cdot) \preceq_{\mathrm{SD}} \hat{F}(\mathbf{y}, \cdot)$$

- **Reversed preference**:

$$\mathbf{x} <_{\mathcal{X}} \mathbf{y} \text{ while not } \hat{F}(\mathbf{x}, \cdot) \preceq_{\mathrm{SD}} \hat{F}(\mathbf{y}, \cdot)$$

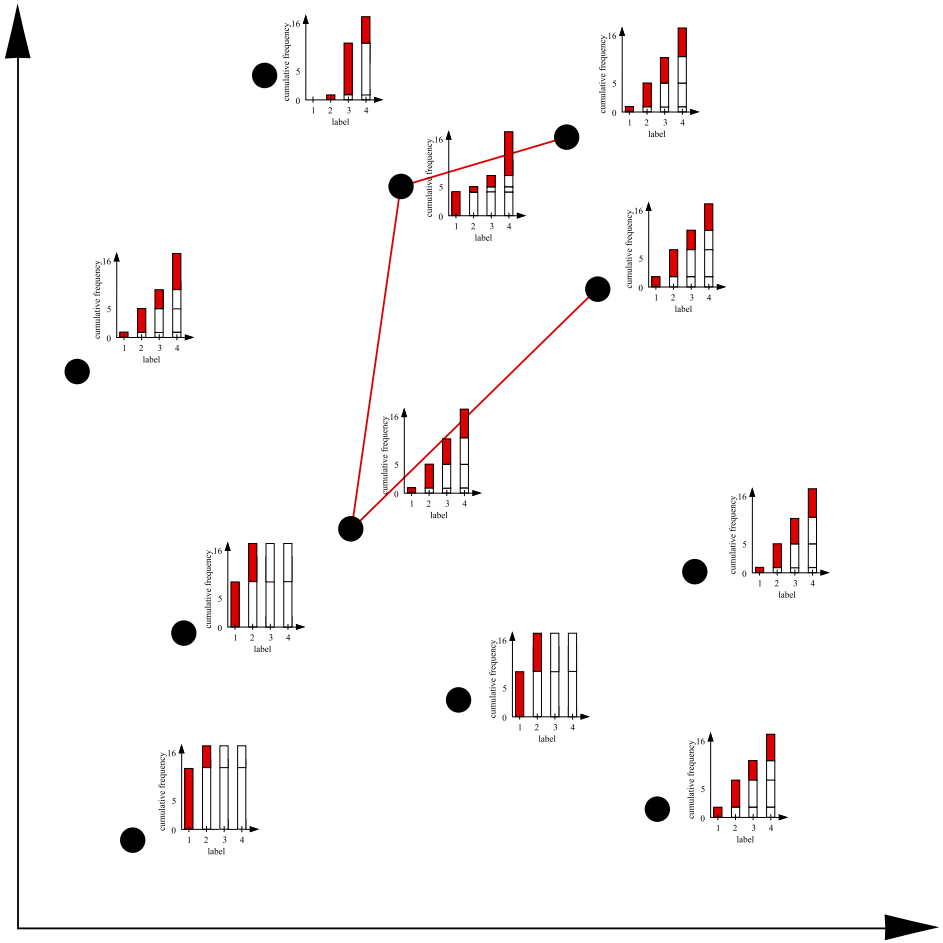
Monotone extensions

If the distribution data set is monotone, then

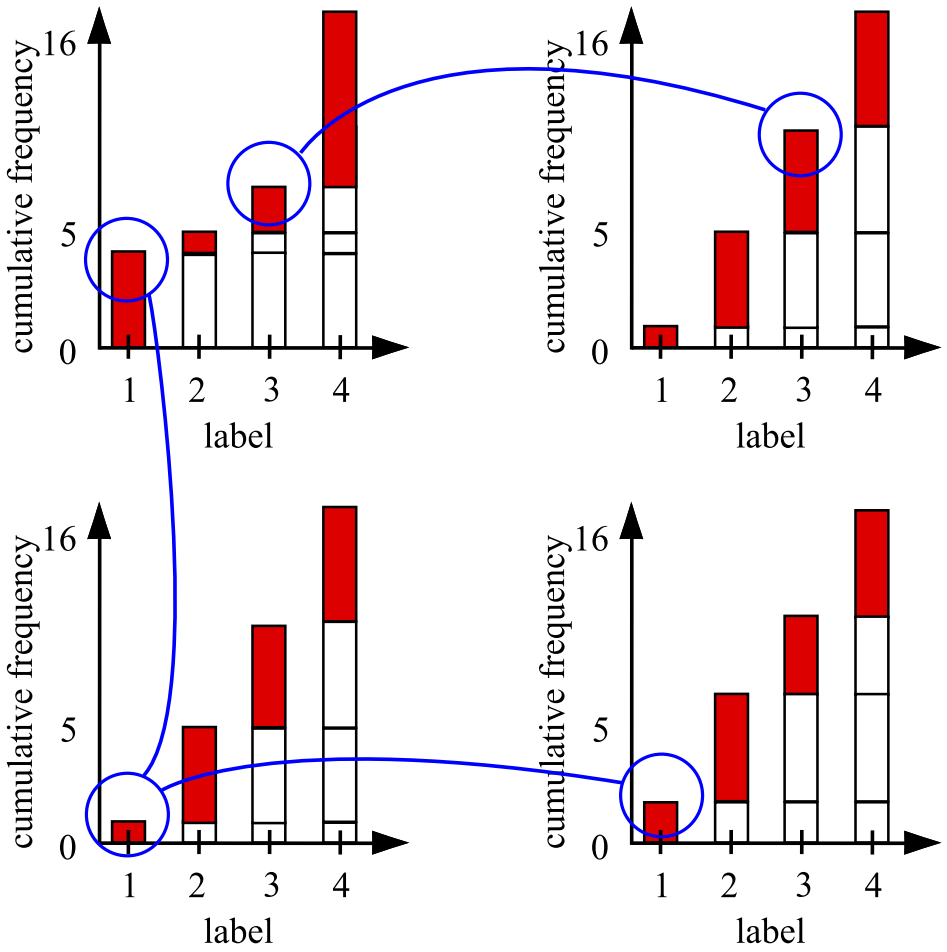
1. F_{min} and F_{max} are monotone **extensions** of $\hat{F}$ to $\mathcal{X}$
2. any monotone extension F of d to $\mathcal{X}$:

$$F_{\mathsf{min}}(\mathbf{y}, \cdot) \preceq_{\mathrm{SD}} F(\mathbf{y}, \cdot) \preceq_{\mathrm{SD}} F_{\mathsf{max}}(\mathbf{y}, \cdot)$$

A non-monotone distribution data set



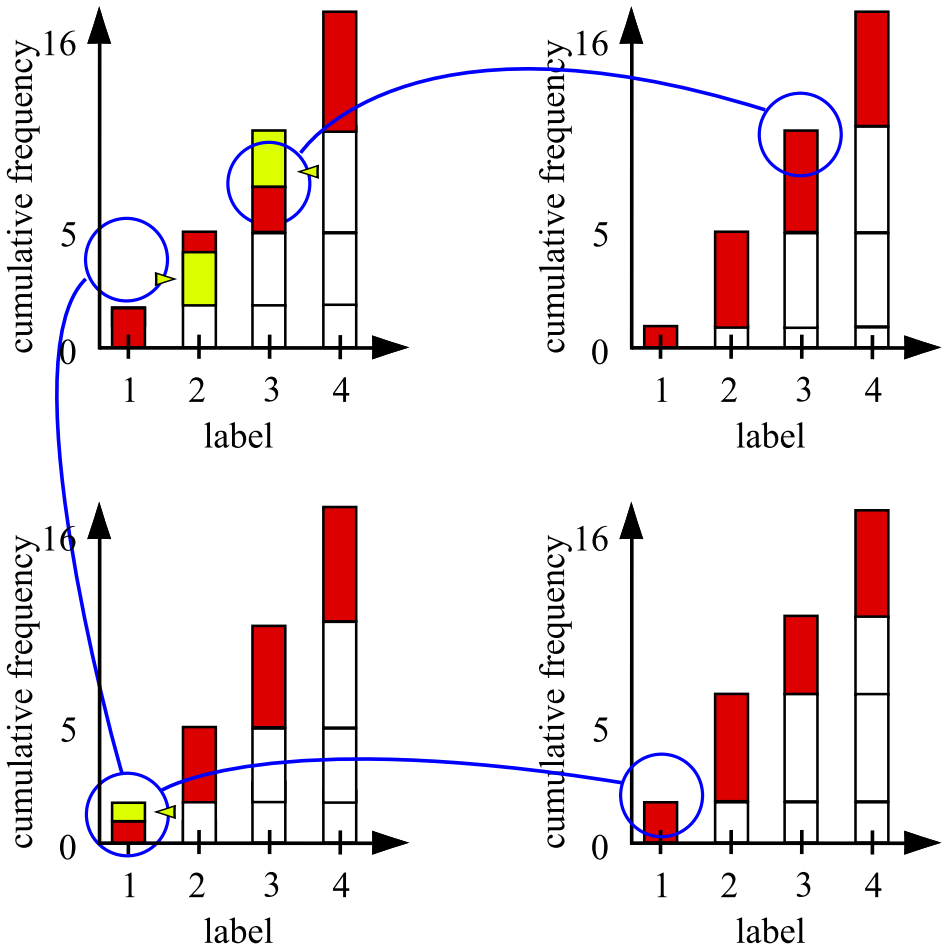
A non-monotone distribution data set



How to handle noise?

1. **Data reduction**: **identify** the noisy distributions and **delete** them
   - the non-monotonicity relation is **not transitive** (MIS problem is NP-complete)
   - deleting entire distributions is quite invasive
   - deleting a single instance affects the entire distribution and is hard to realize
2. **Data relabelling**: **identify** the noisy distributions and **modify** them
   - transitivity of non-monotonicity still holds at the label level
   - **L1-optimal relabelling** is possible using network flow algorithms
   - does not affect the frequency of feature vectors
3. **Non-invasive approach**: keep the data set as is

After relabelling: a monotone distribution data set



A non-invasive approach

- Aim: to build a monotone distribution classifier from a possibly non-monotone distribution data set
- Weighted sums of $F_{\min}$ and $F_{\max}$ are solutions to this problem
- Aim: to identify more general interpolation schemes, depending on both the element $\mathbf{x}$ and the label ℓ
- For given $\mathbf{x}$ and ℓ:
  - **monotone situation**: $F_{\min}(\mathbf{x}, \ell) \geq F_{\max}(\mathbf{x}, \ell)$
  - **reversed preference situation**: $F_{\min}(\mathbf{x}, \ell) < F_{\max}(\mathbf{x}, \ell)$

The main theorem

OSDL generic theorem

Given two $\mathcal{X} \times \mathcal{L} \to [0,1]$ mappings s and t, the mapping $\tilde{F} : \mathcal{X} \times \mathcal{L} \to [0,1]$

$$\tilde{F}(\mathbf{x},\ell) = \begin{cases} s(\mathbf{x},\ell)F_{\min}(\mathbf{x},\ell) + \big(1 - s(\mathbf{x},\ell)\big)F_{\max}(\mathbf{x},\ell) \\ \qquad\qquad\qquad\qquad \text{if } F_{\min}(\mathbf{x},\ell) \geq F_{\max}(\mathbf{x},\ell) \\ t(\mathbf{x},\ell)F_{\min}(\mathbf{x},\ell) + \big(1 - t(\mathbf{x},\ell)\big)F_{\max}(\mathbf{x},\ell) \\ \qquad\qquad\qquad\qquad \text{if } F_{\min}(\mathbf{x},\ell) < F_{\max}(\mathbf{x},\ell) \end{cases}$$

is a **monotone distribution classifier** if and only if

1. s is decreasing in 1st and increasing in 2nd argument
2. t is increasing in 1st and decreasing in 2nd argument

The main theorem: realizations

Several realizations

1. **OSDL**: if one does not want to distinguish between the monotone and the reversed preference situation (s and t are identical), then the simple interpolation scheme is the only one
2. **Balanced and Double-balanced OSDL**: use as weighing functions **measures of support** that count:
   - the number of instances that indicate that $\mathbf{x}$ should receive a label strictly greater than ℓ
   - the number of instances that indicate that $\mathbf{x}$ should receive a label at most ℓ

Epilogue

Concluding observations

1. In many modelling problems, there exists a **monotone relationship** between some or all of the **input variables** and the **output variable** that has to be accounted for
2. **Mamdani–Assilian fuzzy models** for one-shot decisions should be **abandoned**
3. Aggregation theory needs a **reboost**
4. **Resolution of non-monotonicity** can be translated into an optimization problem (network flow theory)
5. Loyalty to the credo of fuzzy set theory (**"First process the data, then defuzzify"**) urges us to develop new mathematics

References

Fuzzy modelling

1. E. Van Broekhoven and B. De Baets, *Fast and accurate center of gravity defuzzification of fuzzy system outputs defined on trapezoidal fuzzy partitions*, Fuzzy Sets and Systems **157** (2006), 904–918.
2. E. Van Broekhoven and B. De Baets, *Monotone Mamdani–Assilian models under Mean of Maxima defuzzification*, Fuzzy Sets and Systems **159** (2008), 2819–2844.
3. E. Van Broekhoven and B. De Baets, *Only smooth rule bases can generate monotone Mamdani–Assilian models under COG defuzzification*, IEEE Trans. Fuzzy Systems **17** (2009), 1157–1174.
4. M. Štěpnička and B. De Baets, *Implication-based models of monotone fuzzy rule bases*, Fuzzy Sets and Systems **232** (2013), 134–155.

Aggregation theory

1. R. Pérez-Fernández, M. Rademaker and B. De Baets, *Monometrics and their role in the rationalisation of ranking rules*, Information Fusion **34** (2017), 16–27.
2. R. Pérez-Fernández, M. Sader and B. De Baets, *Joint consensus evaluation of multiple objects on an ordinal scale: an approach driven by monotonicity*, Information Fusion **42** (2018), 64–74.
3. M. Sader, R. Pérez-Fernández, L. Kuuliala, F. Devlieghere and B. De Baets, *A combined scoring and ranking approach for determining overall food quality*, Internat. J. of Approximate Reasoning **100** (2018), 161–176.
4. R. Pérez-Fernández and B. De Baets, *On the role of monometrics in penalty-based data aggregation*, IEEE Trans. on Fuzzy Systems **27** (2019), 1456–1468.
5. R. Pérez-Fernández, B. De Baets and M. Gagolewski, *A taxonomy of monotonicity properties for the aggregation of multidimensional data*, Information Fusion **52** (2019), 322–334.
6. M. Gagolewski, R. Pérez-Fernández and B. De Baets, *An inherent difficulty in the aggregation of multidimensional data*, IEEE Trans. on Fuzzy Systems, to appear.
7. R. Pérez-Fernández and B. De Baets, *Aggregation theory revisited*, IEEE Trans. Fuzzy Systems, submitted.

Machine learning

1. K. Cao-Van and B. De Baets, *Growing decision trees in an ordinal setting*, Internat. J. Intelligent Systems **18** (2003), 733–750.
2. S. Lievens, B. De Baets and K. Cao-Van, *A probabilistic framework for the design of instance-based supervised ranking algorithms in an ordinal setting*, Annals of Operations Research **163** (2008), 115–142.
3. S. Lievens and B. De Baets, *Supervised ranking in the WEKA environment*, Information Sciences **180** (2010), 4763–4771.

Machine learning

Relabelling:

1. M. Rademaker, B. De Baets and H. De Meyer, *Optimal monotone relabelling of partially non-monotone ordinal data*, Optimization Methods and Software **27** (2012), 17–31.
2. M. Rademaker, B. De Baets and H. De Meyer, *Loss optimal monotone relabelling of noisy multi-criteria data sets*, Information Sciences **179** (2009), 4089–4096.
3. M. Rademaker and B. De Baets, *Optimal restoration of stochastic monotonicity w.r.t. cumulative label frequency loss functions*, Information Sciences **181** (2011), 747–757.

Monotone data set generation:

1. K. De Loof, B. De Baets and H. De Meyer, *On the random generation and counting of weak order extensions of a poset with given class cardinalities*, Information Sciences **177** (2007), 220–230.
2. K. De Loof, B. De Baets and H. De Meyer, *On the random generation of monotone data sets*, Information Processing Letters **107** (2008), 216–220.

Merci pour votre attention