За какую задачу браться?

- Масштабность проблемы: решение проблемы должно касаться большого числа людей, специалистов, лиц принимающих решения.
- Заброшенность (популярность) проблемы. Общая ошибка: решать популярные проблемы.
- Решаемость проблемы. Выбор просто и элегантно решаемых проблем.
- Наша готовность к решению проблемы, квалификация: похожими проектами мы уже занимались.

Problem statement for machine learning

Formal problem statement, an analyst has to set

1) an algebraic structure for the dataset from measurements
2) a data generation hypothesis from 1)
3) a model, or a mixture from 2)
4) an error function (quality criteria with restrictions) from 2)
5) an optimization algorithm from 3) and 4)

The result of the model construction is a Cartesian product

$$\{\textbf{models} \times \textbf{data sets} \times \textbf{quality critea}\}.$$

Def: *Big data rejects the i.i.d. (independent and identically distributed random variables) data generation hypothesis from 2). It requests a mixture model.*

Quality criteria for model generation and selection

Three sources of quality criteria

1. Business: model operation productivity, agent impact to environment
2. Theory: statistical hypothesis, bayesian inference
3. Technology: optimization requirements, resources

The main criteria of model quality

- Precision: MAPE, AUC
- Stability (diversity): std deviation for prediction, covariance of parameters
- Complexity: structure complexity, MDL, evidence of model

Decision support and Integral indicator construction

The integral indicator is a measure

of object's quality. *It is a scalar, corresponded to an object.*

The integral indicator is an aggregation

of object's features that describe various components of the term "quality". *Expert estimation of object's quality could be an integral indicator, too.*

Examples

Index name	Objects	Features	Model
TOEFL exams	Students	Tests	Sum of scores
Eurovision	Singers	Televotes, Jury votes	Linear (weighted sum)
S&P500, NASDAQ	Time-ticks	Shares (prices, volumes)	Non-linear
Bank ratings	Banks	Requirements	By an expert commission
Integral Indicator of Croatian PP's	Power Plants	Waste measurements	Linear

There given a set of objects

Croatian Thermal Power Plants and Combined Heat and Power Plants

1. Plomin 1 TPP
2. Plomin 2 TPP
3. Rijeka TPP
4. Sisak TPP
5. TE-TO Zagreb CHP
6. EL-TO Zagreb CHP
7. TE-TO Osijek CHP
8. *Jetrovac TPP*

There given a set of features

Outcomes and Waste measurements

1. Electricity (GWh)

2. Heat (TJ)

3. Available net capacity (MW)

4. SO_2 (t)

5. NOX (t)

6. Particles (t)

7. CO_2 (kt)

8. Coal (kt)

9. Sulphur content in coal (%)

10. Liquid fuel (kt)

11. Sulphur content in liquid fuel (%)

12. Natural gas (10^6 m^3)

How to construct an index?

Assign a comparison criterion

Ecological footprint of the Croatian Power Plants

Gather a set of comparable objects

TPP and CHP (Jetrovac TPP excluded)

Gather features of the objects

Waste measurements

Make a data table: objects/features

See 7 objects and 10 features in the table below

Select a model

Linear model (with most informative coefficients)

Data table and feature optimums

N	Power Plant	Electricity (GWh)	Heat (TJ)	Available net capacity (MW)	SO_2 (t)	NO_x (t)	Particles (t)	CO_2 (kt)	Coal (kt)	Sulphur content in coal (%)	Liquid fuel (kt)	Sulphur content in liquid fuel (%)	Natural gas (10^6 m^3)
1	Plomin 1 TPP	452	0	98	1950	1378	140	454	198	0.54	0.43	0.2	0
2	Plomin 2 TPP	1576	0	192	581	1434	60	1458	637	0.54	0.37	0.2	0
3	Rijeka TPP	825	0	303	6392	1240	171	616	0	0	200	2.2	0
4	Sisak TPP	741	0	396	3592	1049	255	573	0	0	112	1.79	121
5	TE-TO Zagreb CHP	1374	481	337	2829	705	25	825	0	0	80	1.83	309
6	EL-TO Zagreb CHP		333	332	90	1259	900	19	355	0	0	39	2.1	126
7	TE-TO Osijek CHP		114	115	42	1062	320	35	160	0	0	37	1.1	24
				max	min	min	min	min	min	min	min	min	min

Notations

$X = \{x_{ij}\}$ is the $(n \times m)$ is the real matrix, the data set;
$\mathbf{y} = [y_1, \ldots, y_m]^T$ is the vector of integral indicators;
$\mathbf{w} = [w_1, \ldots, w_n]^T$ is the vector of feature importance weights;

$\mathbf{y}_0, \mathbf{w}_0$ are the expert estimations of the indicators and the weights;

$$
\begin{array}{c|c}
\mathbf{w}^T \\
\hline
\mathbf{y} & X
\end{array}
\quad = \quad
\begin{array}{c|cccc}
 & w_1 & w_2 & \ldots & w_n \\
\hline
y_1 & x_{11} & x_{12} & \ldots & x_{1n} \\
y_2 & x_{21} & x_{22} & \ldots & x_{2n} \\
\ldots & \ldots & \ldots & \vdots & \ldots \\
y_m & x_{m1} & x_{m2} & \ldots & x_{mn}
\end{array} \ .
$$

Usually, data prepared so that

- the minimum of each feature equals 0, while the maximum equals 1;
- the bigger value of each implies better quality of the index.

Pareto slicing

Find the non-dominated objects at each slicing level.
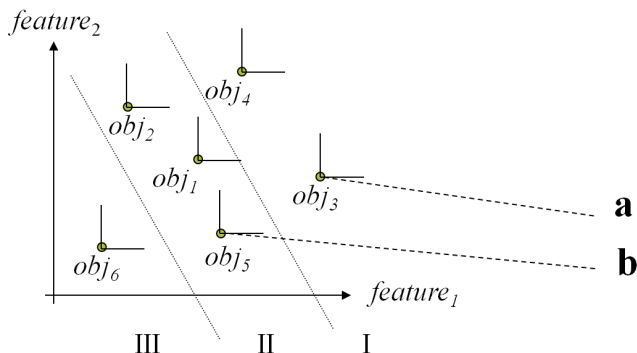
The object a is non-dominated

if there is no $\mathbf{b}_i$ such that $b_{ij} \geqslant a_j$ for all features index j.
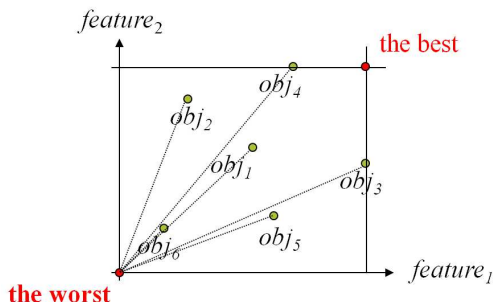
Metric algorithm

The best (worst) object is an object that contains the (maximum) minimum values of the features.

The index is

$$y_i = \sqrt[r]{\sum_{j=1}^{r} \left(x_{ij} - x_j^{\text{best}} \right)^r}$$

For r = 1, this algorithm coincides the weighted sum with equal weighs.

Weighted sum

$$\mathbf{y}_1 = X\mathbf{w}_0,$$

$$\left[\begin{array}{c} y_1 \\ y_2 \\ \dots \\ y_m \end{array} \right] = \left[\begin{array}{cccc} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \dots & \dots & \vdots & \dots \\ x_{m1} & x_{m2} & \dots & x_{mn} \end{array} \right] \left[\begin{array}{c} w_1 \\ w_2 \\ \dots \\ w_m \end{array} \right].$$

Principal Components Analysis

$Y = XV$, where V is the rotation matrix of the principal components. The indicators $\mathbf{y}_{PCA} = X\mathbf{w}_{1PC}$, where $\mathbf{w}_{1PC}$ is the 1st column vector of the matrix V in the singular values decomposition $X = ULV^{T}$.
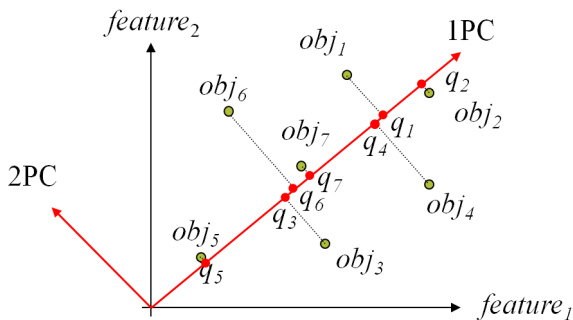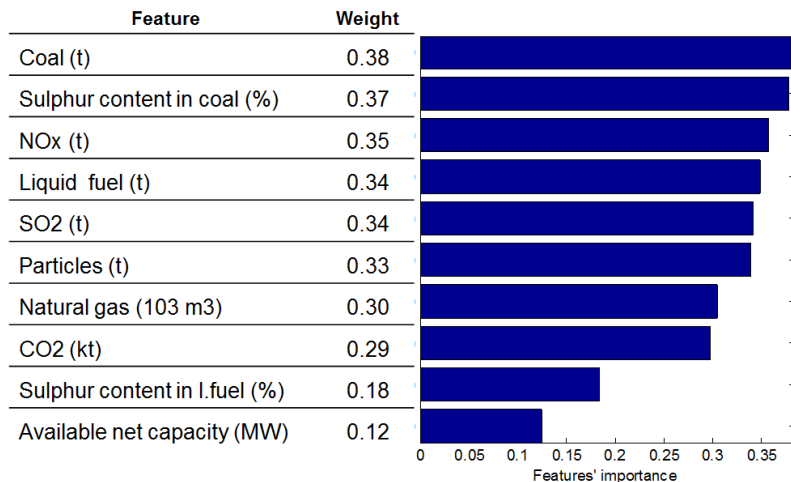

PCA gives minimum mean square error between objects and their projections.
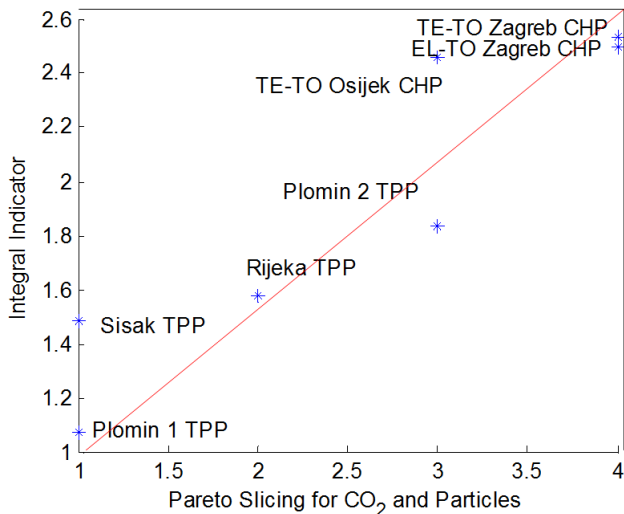
The Integral Indicator

Ecological Impact of the Croatian Power Plants

| Power Plant	Integral Indicator
TE-TO Zagreb CHP	2.53
EL-TO Zagreb CHP	2.49
TE-TO Osijek CHP	2.46
Plomin 2 TPP	1.83
Rijeka TPP	1.57
Sisak TPP	1.48
Plomin 1 TPP	1.07
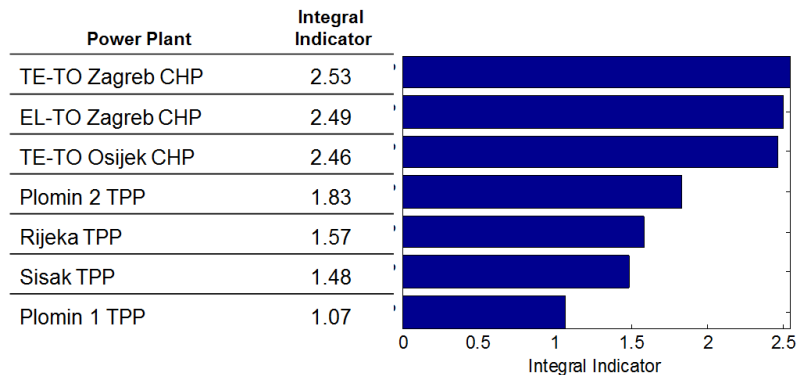
The Importance Weights of the Features

Feature	Weight
Coal (t)	0.38
Sulphur content in coal (%)	0.37
NOx (t)	0.35
Liquid fuel (t)	0.34
SO2 (t)	0.34
Particles (t)	0.33
Natural gas (103 m3)	0.30
CO2 (kt)	0.29
Sulphur content in l.fuel (%)	0.18
Available net capacity (MW)	0.12

Features' importance

The PCA Indicator versus Pareto Slicing

Pair-wise comparison, toy example

	a	s	p	i-c
apple	●	+	+	+
soup		●	+	−
porridge			●	−
ice-cream				●
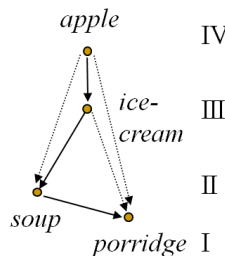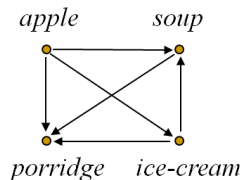
If an object in a row is better than the other one in a column then put "+", otherwise "-".

Make a graph, *row* + *column* means *row* ●——→● *column*.

Find the top and remove extra nodes.

The expert-statistical method

Having plan matrix $\mathbf{X}$ and expert-given target vector $\mathbf{y}_0$, compute optimal parameters

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w} \in \mathbb{R}^n} \|\mathbf{X}\mathbf{w} - \mathbf{y}_0\|^2.$$

Least squares:

$$\hat{\mathbf{w}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}_0.$$

The problem of specification

- **We have**
  expert estimations $\mathbf{y}_0, \mathbf{w}_0$,
  calculated weights and indicators $\mathbf{w}_1 = \mathbf{X}^+ \mathbf{y}_0$, $\mathbf{y}_1 = \mathbf{X} \mathbf{w}_0$.

- **Contradiction.** In general,

$$\mathbf{y}_1 \neq \mathbf{y}_0, \quad \mathbf{w}_1 \neq \mathbf{w}_0.$$

- **Concordance.** Call the estimations $\mathbf{y}$ and $\mathbf{w}$ *concordant* if the following conditions hold:
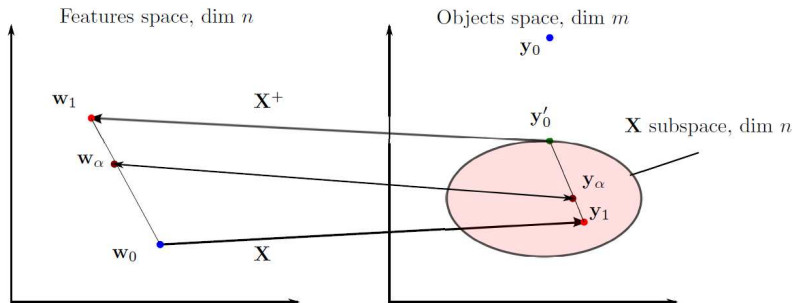
$$\mathbf{y} = \mathbf{X}\mathbf{w}, \quad \mathbf{w} = \mathbf{X}^+ \mathbf{y}.$$

Expert estimations concordance

- Denote by $\mathbf{y}'_0 = \mathbf{X}\mathbf{X}^+\mathbf{y}_0$ the projection of the vector $\mathbf{y}_0$ to the space of the columns of the matrix $\mathbf{X}$.

- α-concordance method: vectors $\mathbf{w}_\alpha, \mathbf{y}_\alpha$,

$$\mathbf{w}_\alpha = \alpha\mathbf{w}_0 + (1-\alpha)\mathbf{X}^+\mathbf{y}'_0, \quad \mathbf{y}_\alpha = (1-\alpha)\mathbf{y}'_0 + \alpha\mathbf{X}\mathbf{w}_0,$$
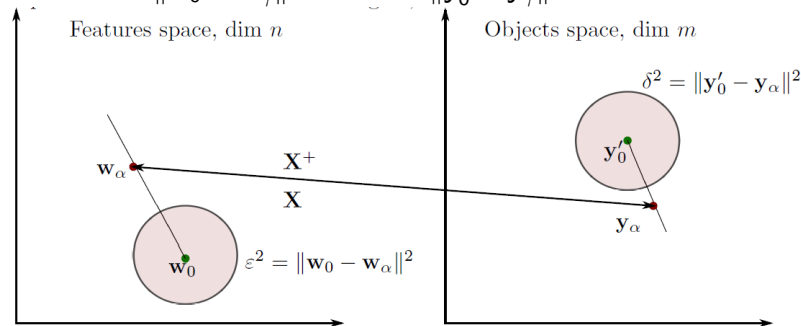
are concordant for $\alpha \in [0; 1]$.

γ-**concordance**

The γ-concordance method finds concordant estimations in the neighborhoods of the vectors $\mathbf{w}_0, \mathbf{y}_0'$ as a solution of the following optimization problem,

$$\mathbf{w}_\gamma = \arg \min_{\mathbf{w} \in \mathbb{R}^n} (\varepsilon^2 + \gamma^2 \delta^2),$$

where $\varepsilon^2 = \|\mathbf{w}_0 - \mathbf{w}_\gamma\|^2$ and $\delta^2 = \|\mathbf{y}_0' - \mathbf{y}_\gamma\|^2$.
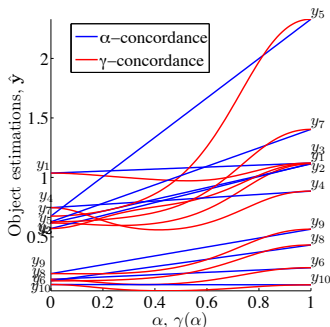

Features space, dim n

Objects space, dim m

$\delta^2 = \|\mathbf{y}_0' - \mathbf{y}_\alpha\|^2$

$\mathbf{y}_0'$

$\mathbf{w}_\alpha$

$\mathbf{X}^+$

$\mathbf{X}$

$\mathbf{y}_\alpha$

$\mathbf{w}_0$ $\varepsilon^2 = \|\mathbf{w}_0 - \mathbf{w}_\alpha\|^2$

Concordance methods comparison

The x-axis shows the values of the parameter α changing from 0 to 1, whereas parameter γ is the function of α,

$$\gamma = \frac{\alpha}{1-\alpha},$$

so γ changes from 0 to ∞.

Ordinal-scaled expert estimations

Experts make estimations in the ordinal scales:

$$\begin{cases} y_1 \geqslant ... \geqslant y_m \geqslant 0, \\ w_1 \geqslant ... \geqslant w_n \geqslant 0. \end{cases}$$

In matrix notations:

$$\begin{cases} J_m \mathbf{y} \geqslant 0, \\ J_n \mathbf{w} \geqslant 0, \end{cases} \quad \text{where } J = \begin{pmatrix} 1 & -1 & 0 & \dots & 0 \\ 0 & 1 & -1 & \dots & 0 \\ \dots\dots\dots\dots\dots\dots\dots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix}.$$

Consider **two cones** instead of two vectors:

$$\mathcal{Y} = \{\mathbf{y} \,|\, J_m \mathbf{y} \geqslant 0\},$$

$$\mathcal{W} = \{\mathbf{w} \,|\, J_n \mathbf{w} \geqslant 0\}.$$

Ordinal specification

- The linear operator $\mathbf{X}$ maps the cone $\mathcal{W}_0$ of the expert estimations of the criteria weights $\mathbf{w}_0$ to the computed cone $\mathbf{X}\mathcal{W}_0$.

- The linear operator $\mathbf{X}\mathbf{X}^+$ maps the cone $\mathcal{Y}_0$ of the expert estimations of the objects $\mathbf{y}_0$ to the cone $\mathcal{Y}_0' = \mathbf{X}\mathbf{X}^+\mathcal{Y}_0$.
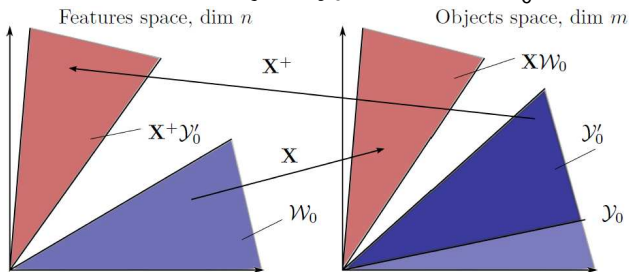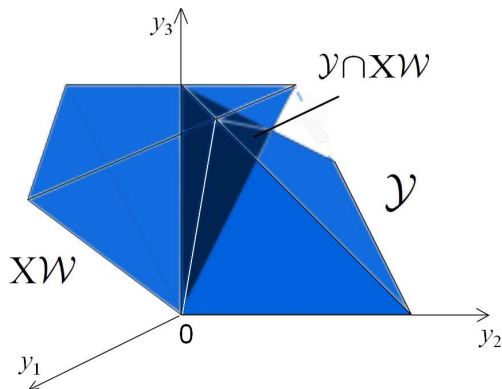
Cones intersection: specification is needed

The cones $\mathcal{Y}, \mathbf{X}\mathcal{W}$ do not intersect: the expert estimations **contradict** each other.
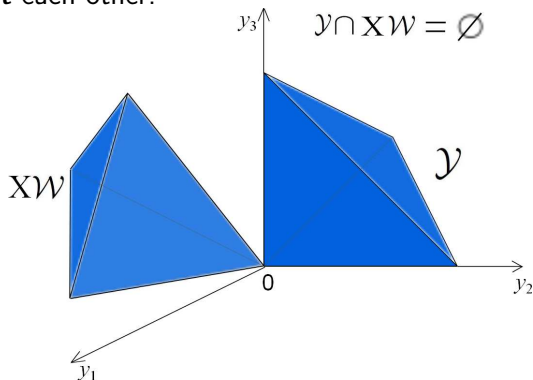
Cones intersection: no specification is needed

The cones $\mathcal{Y}, \mathbf{X}\mathcal{W}$ intersect: the expert estimations **do not contradict** each other.

How to specify the expert estimations?

We propose 2 ways to specify the rank-scaled expert estimations.
Searching of the nearest vectors in the cones by:

- Distance minimization,
- Correlation maximization,
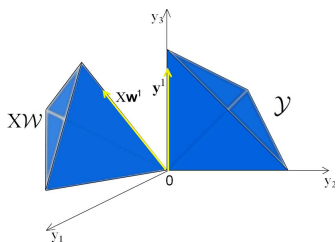
and Nearly-isotonic regression.

Nearest vectors in the cones

Distance minimization:

$$(\mathbf{y}^1, \mathbf{w}^1) = \min_{\mathbf{y} \in \mathcal{Y}, \ \mathbf{w} \in \mathcal{W}} \|\mathbf{y} - X\mathbf{w}\|_2 \text{ subject to } \|X\mathbf{w}\|_2 = 1, \|\mathbf{y}\|_2 = 1.$$

Correlation maximization (ρ is the Spearman rank-correlation coefficient):

$$(\mathbf{y}^1, \mathbf{w}^1) = \max_{\mathbf{y} \in \mathcal{Y}, \ \mathbf{w} \in \mathcal{W}} \rho(\mathbf{y}, X\mathbf{w}) \text{ subject to } \|X\mathbf{w}\|_2 = 1, \|\mathbf{y}\|_2 = 1.$$

Alternative approach: Nearly-Isotonic Regression

Again, the expert estimations:

1. $w_1 \geqslant ... \geqslant w_n \geqslant 0$,

2. $\widetilde{\mathbf{w}} = X^+ \mathbf{y}_0$,
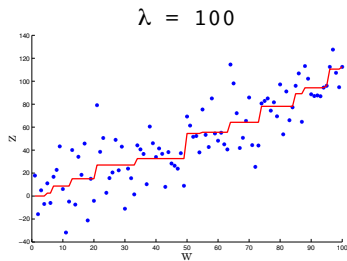
The problem of specification in rank scales:

$$
\widehat{\mathbf{w}} = \arg \min_{\mathbf{w} \in \mathbb{R}^n} \left(\underbrace{\frac{1}{2} \sum_{j=1}^{n} (\widetilde{w}_j - w_j)^2}_{\text{ref. to } \mathbf{y}_0} + \underbrace{\lambda \sum_{j=1}^{n-1} (w_j - w_{j+1})_+}_{\text{ref. to expert estimations of } \mathbf{w}} \right),
$$

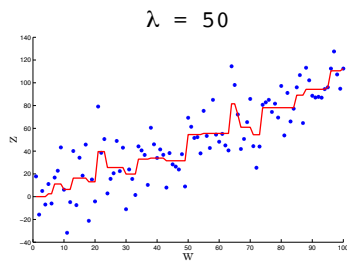where λ is a regularizer.

Nearly-isotonic regression algorithm: illustration

A blue dot is a feature weight.

$$z(w_j) = \widetilde{w}_j, \quad n = 100.$$

(a)   (b)

Notations

The sample $\mathfrak{D} = (\mathbf{X}, \mathbf{y})$ is given.

- $\mathbf{X} = [\chi_1, ..., \chi_n]$ is a plan matrix,

$$\text{features } \chi_j \subset \mathbb{R}^m, \quad j \in \mathcal{J} = \{1, ..., n\}.$$

- $\mathbf{y} \subset \mathbb{R}^m$ is a set of integral indicators.

- Linear scale: $\quad \chi_j \in \mathbb{R}^m, \quad j \in \mathcal{J}.$

- Ordinal scale (total order):

$$\chi_j = \{\mathbf{x}_j \in \mathbb{R}_+^m | \quad x_{j1} \geqslant x_{j2} \geqslant ... \geqslant x_{jm}\}.$$

- Ordinal scale (partial order):

$$\chi_j = \{\mathbf{x}_j \in \mathbb{R}_+^m | \quad x_{jk_1} \geqslant x_{jk_2}, ...\}.$$

Ordinal scale

- Ordinal-scaled feature χ_j,

$$\chi_j = \{\mathbf{x}_j \in \mathbb{R}^m_+ | \quad x_{j1} \geqslant x_{j2} \geqslant ... \geqslant x_{jm}\},$$

  is a convex polyhedral cone in $\mathbb{R}^m$.

- The cone χ_j corresponds to a partial order defined on the set of values of the j-th feature.

Cone point decomposition

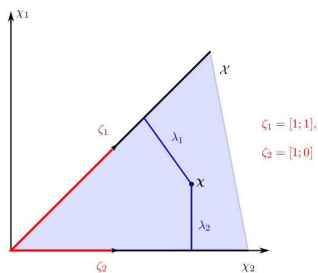For every point $\mathbf{x}$ of the cone χ the following theorem holds:

$$\mathbf{x} = w \sum_{k=1}^{L} \lambda_k \zeta_k, \quad w \geqslant 0, \quad \sum_{k=1}^{L} \lambda_k = 1, \quad \lambda_k \geq 0,$$

where L is a number of different values for the feature χ,
ζ_k is a cone generator corresponding to the feature χ:

$$\zeta_k(i) = \begin{cases} 1, & \text{if } x_i \geqslant x_k, \\ 0, & \text{otherwise.} \end{cases}$$
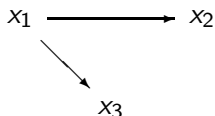
Cone point decomposition: illustration

$$\mathbf{x} = w \sum_{k=1}^{L} \lambda_k \boldsymbol{\zeta}_k, \quad w \geqslant 0, \quad \sum_{k=1}^{L} \lambda_k = 1, \quad \lambda_k \geq 0.$$

Ordinal scale binarization illustration

Give an example of a partially ordered set:

$$\chi = \{x_1, x_2, x_3 \big| \quad x_1 \geqslant x_2, x_1 \geqslant x_3\}.$$

$$x_1 \longrightarrow x_2$$

$$x_3$$

A matrix $Z = [\zeta_1, \zeta_2, \zeta_3]$, corresponding to this graph:

$$
\begin{array}{ccc}
\zeta_1 & \zeta_2 & \zeta_3
\end{array}
$$
$$
Z = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},
$$

$$\chi \ni \mathbf{x} = \boldsymbol{\lambda} Z, \quad \boldsymbol{\lambda} \geqslant \mathbf{0}, \quad \|\boldsymbol{\lambda}\| = 1.$$

Integral indicator construction model

- Model of the integral indicator construction:

$$\mathbf{y}_1 = f([\chi_1, ..., \chi_n], \mathbf{w}).$$

- Linear model generalization for the case of ordinal features:

$$\mathbf{y}_1 = w_1 \boldsymbol{\lambda}_1 Z_1 + ... + w_n \boldsymbol{\lambda}_n Z_n,$$

$$w_j \in \mathbb{R}, \qquad \boldsymbol{\lambda}_j \in \Lambda = \{\boldsymbol{\lambda} | \boldsymbol{\lambda} \geqslant \mathbf{0}, \|\boldsymbol{\lambda}\|_1 = 1\}.$$

Preference learning problem

The goal of research

To provide a method for preference estimation on a set of objects described by a set of ordinal features

The challenges

1. To propose a partial order cone concept to describe a finite partially ordered set.
2. To investigate properties of a composition of partial order cones.
3. To develop a method of ordinal-scaled dependency recovering.

History of problem

1. **Social choice theory** (K. Arrow, 1951)
2. Preference aggregation (J. Kemeny, 1959)
3. Ordinal regression (P. McCullagh, 1980)
4. Expert estimations (B. Litvak, 1996)
5. **Learning to rank** (W. Cohen, R. Shapiro, 1999)
6. Ranking SVM (R. Herbrich et al., 1999)
7. **Cones usage** (V. Strijov, 2006)
8. Criteria importance theory (V. Podinovskiy, 2007)
9. Decision theory (F. Aleskerov, 2007)
10. **Preference learning** (J. Fuernkranz, 2011)

Illustrative example: categorization of threatened species

There are given an ordinal description and categories of species

Species	Population size	Habitat structure	Structure variation	Category
Green sturgeon	small	spotty	stable	extinct in the wild
Ladoga coregonus	small	solid	segmented	critically endangered
long-finned charr	high	dispersion	segmented	endangered
Polar bear	high	solid	unknown	vulnerable

The goal: to construct a model of categorization by the expert-given ordinal object description

Problem setting

Input data

There given a set of pairs $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ such that an object $\mathbf{x}_i = [x_{i1}, ..., x_{in}]^\mathsf{T}$,
x_{ij} is an element of a non-numeric matrix $\mathbf{X}$ with columns X_j:

$$\mathbf{X} = [X_1, ..., X_n].$$

Elements
$$x_{ij} \in \mathbb{L}_j \ \ \text{and} \ \ y_i \in \mathbb{Y}$$

belong to sets with given partial order relations.

The problem

To construct a model $f : \mathbb{L}_1 \times ... \times \mathbb{L}_n \to \mathbb{R}$ such that

- f is a monotone function, $\mathbf{x}_i \succeq \mathbf{x}_k \ \to \ f(\mathbf{x}_i) \geq f(\mathbf{x}_k)$.
- $\mathbf{f}$ optimally estimates a preference relation given by a vector $\mathbf{y}$. The optimality condition is

$$y_i \succeq y_k \ \to \ f(\mathbf{x}_i) \geq f(\mathbf{x}_k).$$

Basic concepts

- **Partial order relation** $\succeq$ is a reflexive, antisymmetric, and transitive binary relation.
- **Partial order matrix Z** is a matrix describing pairwise dominance relation of objects:

$$\mathbf{Z}(i, k) = \begin{cases} 1, & \text{if} \quad x_i \succeq x_k, \\ 0, & \text{if} \quad x_i \not\succeq x_k. \end{cases}$$

- $\mathcal{X}_0$ is a **partial order cone** coresponding to a finite poset X, if

$$\mathcal{X}_0 = \{\chi \in \mathbb{R}_+^m | \quad x_i \succeq x_k \to \chi_i \geq \chi_k \quad i, k = 1, ..., m\}.$$

Partial order cone

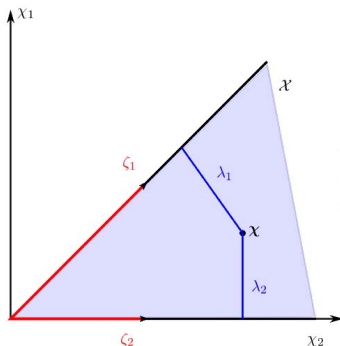$\mathcal{X}_0$ is a polyhedral partial order cone given by a matrix **A** of size $m^2 \times m$:

$$\mathcal{X}_0 = \{\chi \mid \mathbf{A}\chi \leq \mathbf{0}\},$$

such that a string of matrix **A** has the form $[0, ..., 0, -1_i, 0, ..., 0, 1_k, 0, ..., 0]$ and corresponds ta an inequality $x_i \succeq x_k$.

Theorem [Minkowski, Weyl]

Each polyhedral cone $\mathcal{X}_0$ is non-empty and finitely generated,

$$\mathcal{X}_0 = \{\sum_{k=1}^{r} \lambda_k \boldsymbol{\zeta}_k \mid \lambda_k \geq 0\}.$$

Explicit form of cone generators

Let a cone $\mathcal{X}$ be generated with the columns of matrix $\mathbf{Z}$:

$$\mathcal{X} = \{\mathbf{Z}\boldsymbol{\lambda} \mid \boldsymbol{\lambda} \geq \mathbf{0}\}.$$

Theorem [Kuznetsov: 2013]

The following statements are valid for the cone $\mathcal{X}$ and the partial order cone $\mathcal{X}_0 = \{\boldsymbol{\chi} \mid \mathbf{A}\boldsymbol{\chi} \leq \mathbf{0}\}$.

1. The cone $\mathcal{X}$ is a subset of the cone $\mathcal{X}_0$,

$$\mathcal{X} \subseteq \mathcal{X}_0.$$

2. In the case of linear order on X, the cones $\mathcal{X}$ and $\mathcal{X}_0$ are equal:

$$\mathcal{X} = \mathcal{X}_0$$

Polyhedral model with ordinal features

The problem

To construct a monotone function $f : \mathbb{L}_1 \times ... \times \mathbb{L}_n \to \mathbb{R}$, optimally estimating a preference relation given by a vector $\mathbf{y}$,

$$y_i \succeq y_k \;\to\; f(\mathbf{x}_i) \geq f(\mathbf{x}_k).$$

To solve a problem, define a class of models.

Linear polyhedral model

A set of values for a model $\mathbf{f}(\mathbf{X})$ on the sample D is a Minkowski sum of the cones $\mathcal{X}_1, ..., \mathcal{X}_n$:

$$\mathbf{f}(\mathbf{X}) \in \mathcal{X} = \mathcal{X}_1 \oplus ... \oplus \mathcal{X}_n.$$

Polyhedral model parameter optimization

Parametrization theorem [Kuznetsov: 2013]

A point $\mathbf{f}(\mathbf{X})$ of the cone $\mathcal{X}$ is defined by a formula:

$$\mathbf{f}(\mathbf{X}) = \sum_{j=1}^{n} \mathbf{Z}_j \lambda_j, \quad \lambda_j \geq \mathbf{0},$$

where $\mathbf{Z}_j$ is a partial order matrix for the cone $\mathcal{X}_j$.

Optimal parameters

We construct a solution $\hat{\mathbf{f}}$ as a projection of a vector $\mathbf{y} \in D$ to the cone $\mathcal{X}$:

$$\hat{\mathbf{f}} = P_{\mathcal{X}}(\mathbf{y}).$$

Optimal parameters $\hat{\boldsymbol{\lambda}}$ minimize the expression:

$$\hat{\boldsymbol{\lambda}} = \arg \min_{\lambda_j \geq 0} \| \mathbf{y} - \sum_{j=1}^{n} \mathbf{Z}_j \lambda_j \|_2^2.$$

Reduction of parameter space

Polyhedral model:

$$\mathbf{f}(\mathbf{X}) = \sum_{j=1}^{n} \mathbf{Z}_j \lambda_j, \quad \lambda_j \geq \mathbf{0}.$$

In the cone $\mathcal{X}$ we consider a central point $\bar{\mathbf{x}} = \frac{1}{n} \sum_{j=1}^{n} \mathbf{z}_j$.
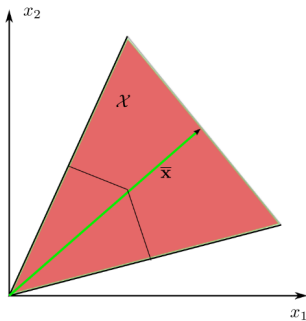
Theorem [Kuznetsov: 2014]

By replacing each cone $\mathcal{X}_k = \{\sum \lambda_{jk} \mathbf{z}_{jk} \mid \boldsymbol{\lambda}_k \geq \mathbf{0}\}$ with its central point, the polyhedral model can be expressed as

$$\mathbf{f}(\mathbf{X}) = \hat{\mathbf{Z}} \boldsymbol{\lambda},$$

where $\hat{\mathbf{Z}} = \sum_{j=1}^{n} w_j \mathbf{Z}_j$ is a matrix of pairwise object dominance.
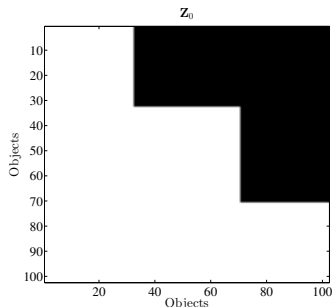
Algotihm of parameter optimization

Input: a sample $D = \{(\mathbf{x}_i, y_i)\}$.
Output: optimal parameters $\hat{\mathbf{w}}, \hat{\boldsymbol{\lambda}}$.
Algorithm:

1. Construct matrices $\mathbf{Z}_1, ..., \mathbf{Z}_n, \mathbf{Z}_Y$.
2. Estimate a matrix $\hat{\mathbf{Z}}$: $\hat{\mathbf{Z}}(i, k) = \sum\limits_{j=1}^{n} w_j \mathbf{Z}_j(i, k)$.
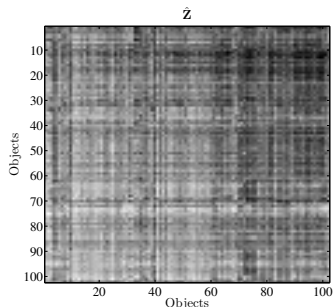
3. Estimate parameters $\boldsymbol{\lambda}$: $\hat{\boldsymbol{\lambda}} = \arg\min\limits_{\lambda_j \geq \mathbf{0}} \|\mathbf{y} - \hat{\mathbf{Z}}\boldsymbol{\lambda}\|_2^2$.

Concordance of ordinal expert estimations

The problem

To construct a monotone function $f : \mathbb{R}^n \to Y$ such that

$$y_i \succ y_k \quad \to \quad f(\mathbf{x}_i) \succ f(\mathbf{x}_k),$$

with consideration of preference relation over the set of features,

$$X_s \succeq X_r.$$

Set of linear models

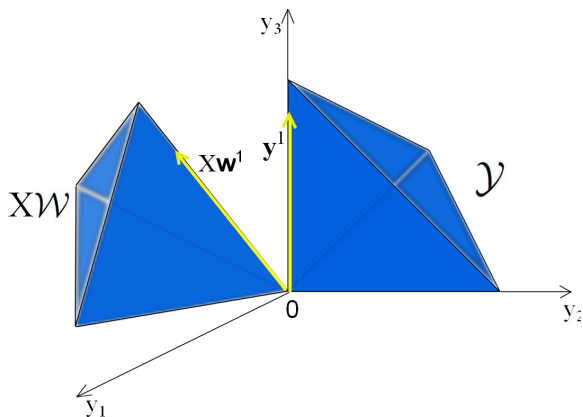We consider a linear model with a vector of parameters $\mathbf{w} \in \mathcal{W}$,

$$f(\mathbf{x}) = \sum_{i=1}^{n} w_i x_i,$$

where $\mathcal{W} = \{\mathbf{w} \mid w_1 \geq ... \geq w_k\}$ is a partial order cone on the set of features.

The problem is to find vectors $\hat{\mathbf{y}} \in \mathcal{Y}$ и $\hat{\mathbf{w}} \in \mathcal{W}$ such that:

$$(\hat{\mathbf{y}}, \hat{\mathbf{w}}) = \underset{\mathbf{y} \in \mathcal{Y}, \ \mathbf{w} \in \mathcal{W}}{\arg\min} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2, \quad \|\mathbf{y}\|_2^2 = 1, \ \|\mathbf{w}\|_2^2 = 1.$$

Parameter optimization for a concordance problem [Kuznetsov: 2012]

- To find optimal parameters we solve consecutive problems of non-negative least squares with a matrix $\mathbf{X}_w = \mathbf{X}\mathbf{Z}_w$:

Problem $2k$:	Problem $2k + 1$:
$\hat{\boldsymbol{\lambda}}_y = \min\limits_{\boldsymbol{\lambda}_y \geq \mathbf{0}} \|\mathbf{Z}_y \boldsymbol{\lambda}_y - \mathbf{X}_w \hat{\boldsymbol{\lambda}}_w\|_2^2$	$\hat{\boldsymbol{\lambda}}_w = \min\limits_{\boldsymbol{\lambda}_w \geq \mathbf{0}} \|\mathbf{Z}_y \hat{\boldsymbol{\lambda}}_y - \mathbf{X}_w \boldsymbol{\lambda}_w\|_2^2$

- Theorem: an algorithm finds an optimal solution for not more than $m + n$ iterations.

Partial order matrix properties

$\mathbf{r}_1$, $\mathbf{r}_2$ — ordinal vectors, $\mathbf{Z}_1, \mathbf{Z}_2$ — corresponding matrices.

Theorem: ordinal metrics generalization

The following expressions hold.

1. Spearman correlation:

$$\rho_s(\mathbf{r}_1, \mathbf{r}_2) \propto \sum_i^m \left(\sum_k^m \mathbf{Z}_1(i,k) - \mathbf{Z}_2(i,k) \right)^2.$$

2. Kendall correlation:

$$\tau(\mathbf{r}_1, \mathbf{r}_2) \propto \sum_i \sum_k \left(\mathbf{Z}_1(i,k) - \mathbf{Z}_2(i,k) \right)^2.$$

3. Multiclass AUC:

$$\text{AUC} = \frac{1}{M} \sum_{k,i=1}^m [\mathbf{Z}(k,i) = 0][\hat{\mathbf{Z}}(k,i) = 0], \ M = \sum_{k_1 \prec k_2} m_{k_1} m_{k_2}.$$

Categorization of threatened species of the IUCN Red List

Expert-given data

Species	Population size	habitat square	Genetic diversity	Category
Green sturgeon	2	2	0	1
Ladoga whitefish	0	2	1	2
Long-finned charr	3	1	0	3
Polar bear	3	3	0	4
Buff-breasted sandpiper	2	1	0	3
Azovian beluga	1	3	1	1
Water chestnut	3	3	2	2
Omphalina hudsoniana	2	2	0	3
Sakhalin sturgeon	1	2	1	1
Dinnik's viper	3	3	2	2
Siberian tiger	2	2	1	2
Tropical lichens	2	1	1	5

Features description

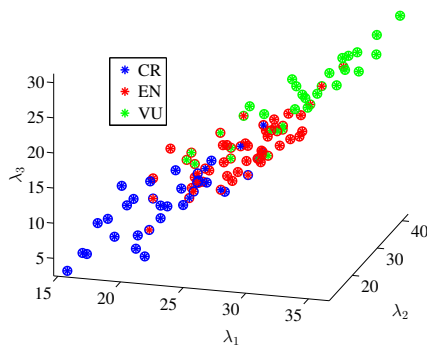| Feature	scale
Population size	3 — large 2 — small 1 — critically small 0 — unknown
Habitat square	3 — big 2 — limited 1 — very limited 0 — unknown
Genetic diversity	3 — high 2 — low 1 — unknown
Category	5 — least concern 4 — vulnerable 3 — endangered 2 — critically endangered 1 — extinct in the wild

Pairwise feature dominance

	Population size	Habitat square	Genetic diversity
Population size	1	1	1
Hablitat square	0	1	0
Genetic diversity	0	0	1

Categorization results

Error function – Hamming loss, $S(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{m} \sum_{i=1}^{m} |y_i - \hat{y}_i|$.

Algorithm	Learning error	Tessting error
Conic	0.29	**0.58**
Decision tree	0.25	0.69
GLM	0.57	0.71

Basic statements

The goal:

to construct a model of the IUCN Red List threatened species categorization using expert estimations of the features.

The model must:

1. use ordinal scales of expert estimations,
2. obtain optimal complexity,
3. rely on expert-given categorization.

Features assumptions

The following assumptions about features structure are considered:

1. the given set of features is sufficient to construct an adequate model;

2. the complete order relation is defined on the feature values;

3. the rule "the bigger the better" is valid, that is the greater feature value causes the greater preference by an object;

4. different expert estimations of the same object are allowed.

List of features

1. Population size.
2. Growth rate.
3. Occurency/density.
4. Physiological state.
5. Habitat state.
6. Population structure trend.
7. Monitoring.
8. New populations.
9. Capacity build.

Input data

A data fragment.

Species: **Russian desman**

Feature	Condition	Change trend
Population size	3 – high; **2** – low; 1 – critical	4 – grows; 3 – stable; **2** – decreases slowly; 1 – decreases rapidly
Population structure	2 – complex; **1** – simple	**2** – stable; 1 – local populations disappear

A partial order is defined over the set of features.

Problem statement

There is given

a set of pairs $\mathfrak{D} = \{(\mathbf{x}_i, y_i)\}$, $i \in \mathcal{I} = \{1, \ldots, m\}$.

Ordinal scales and class labels

Every object $\mathbf{x} = [\chi_1, \ldots, \chi_j, \ldots, \chi_d]^T$, is described by ordinal-scaled features $\chi_j \in \mathbb{L}_j = \{1 \prec \cdots \prec k_j\}$. A partial order is set over the set of features.

Over the set $\mathbb{Y} = \{1, 2, 3\}$ of the class labels y it is given a strict order relation: $1 \prec 2 \prec 3$.

The goal is to construct a monotone function $\varphi \colon \mathbf{x} \mapsto \hat{y}$

$$\varphi_{opt} = \arg\min_{\varphi} S(\varphi) = \arg\min_{\varphi} \frac{1}{m} \sum_{i \in \mathcal{I}} r(y_i, \varphi(\mathbf{x}_i)).$$

Dominance relation
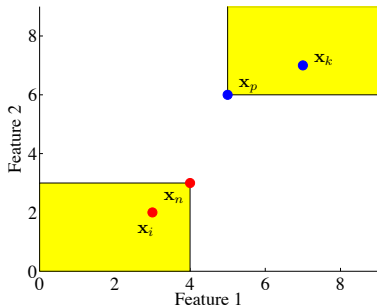
Without features hierarchy

$\mathbf{x}_n \succ_n \mathbf{x}_i$,
if $x_{nj} \geq x_{ij}$ for each $j \in \mathcal{J}$.

$\mathbf{x}_p \succ_p \mathbf{x}_k$,
if $x_{pj} \leq x_{kj}$ for each $j \in \mathcal{J}$.

Any object doesn't dominate
itself: $\mathbf{x} \not\succ_n \mathbf{x}$, $\quad \mathbf{x} \not\succ_p \mathbf{x}$.

Dominance relation

With features hierarchy
Leat a feature r be more important than t.

$\mathbf{x}_n \succ_{\tilde{n}} \mathbf{x}_i$, if $\mathbf{x}_n \succ_n \mathbf{x}_i$
or $x_{nr} > x_{nt}$ and $\mathbf{x}_n^{rt} \succ_n \mathbf{x}_i$.

$\mathbf{x}_p \succ_{\tilde{p}} \mathbf{x}_k$, if $\mathbf{x}_p \succ_p \mathbf{x}_k$
or $x_{pr} < x_{pt}$ and $\mathbf{x}_p^{rt} \succ_p \mathbf{x}_k$.

Any object doesn't dominate
itself: $\mathbf{x} \not\succ_{\tilde{n}} \mathbf{x}, \quad \mathbf{x} \not\succ_{\tilde{p}} \mathbf{x}$.
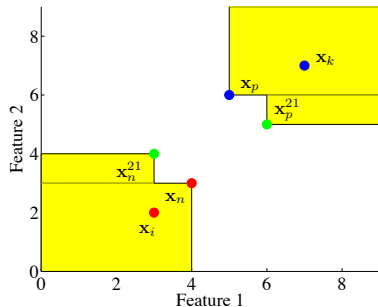
Dominance areas

	The feature 1 is more important than 2	The feature 2 is more important than 1
$x_{n1} > x_{n2}$, $x_{p1} < x_{p2}$		
$x_{n1} < x_{n2}$, $x_{p1} > x_{p2}$		

Optimal Pareto fronts

POF_n, POF_p

A set of objects x, if for each element doesn't exist any other element x' such that
$POF_n : \quad x' \succ_n x \ (x' \succ_{\tilde{n}} x); \qquad POF_p : \quad x' \succ_p x \ (x' \succ_{\tilde{p}} x).$
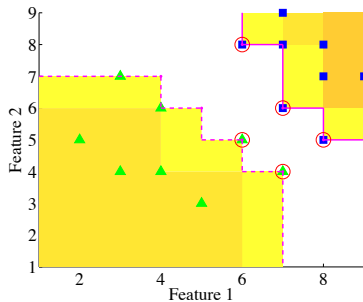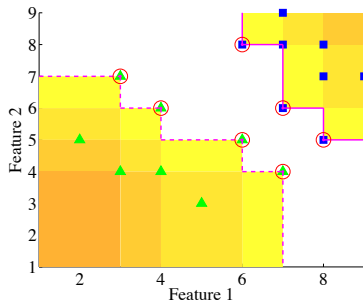
Two-class classification

x — a classified object
$f(\cdot)$ — a classifier function

$$
f(\mathbf{x}) = \begin{cases} 0, & \mathbf{x}_n \succ_n \mathbf{x}; \\ 1, & \mathbf{x}_p \succ_p \mathbf{x}; \\ f\left(\underset{\mathbf{x}' \in \overline{\mathrm{POF}}_n \cup \overline{\mathrm{POF}}_p}{\arg\min} \big(\rho(\mathbf{x}, \mathbf{x}') \big) \right), & \text{otherwise.} \end{cases}
$$

$\overline{\mathrm{POF}}_n, \overline{\mathrm{POF}}_p$ are boundaries of dominance spaces for the corresponding optimal Pareto fronts.
ρ is a distance function between objects,

$$
\rho(\mathbf{x}, \mathbf{x}') = \sum_{j=1}^{d} r(x_j, x'_j).
$$

Two-class classification example
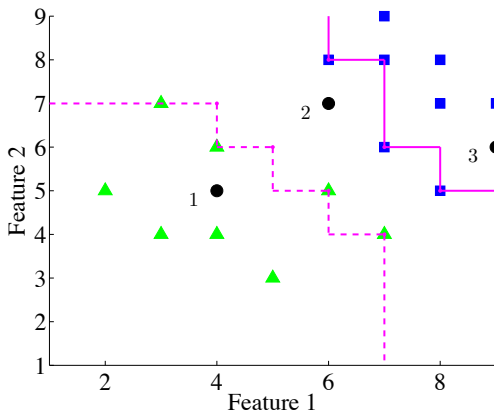
№	Object x	$f(\mathbf{x})$
1	(4,5)	0
2	(6,7)	1
3	(9,6)	1

Separable sample construction
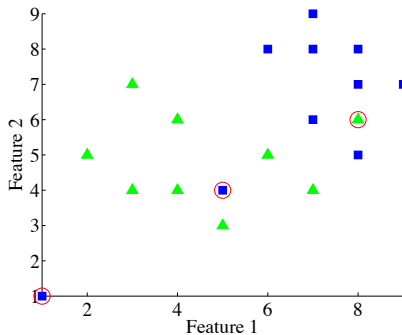
(c) With defective objects
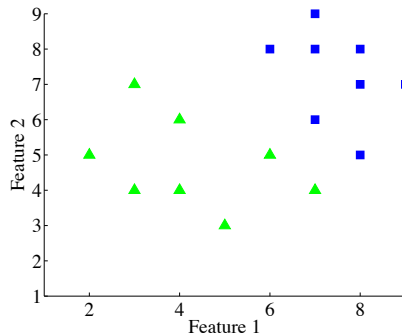
(d) Without defective objects

Monotone classifier definition

$\{1 \prec \cdots \prec u \prec u + 1 \prec \cdots \prec z\} = \mathbb{Z}$ — class labels

$f_{u,u+1} \colon \mathbf{x} \mapsto \hat{y} \in \{0, 1\}$ — two-class classifier for a pair of adjacent classes

«0» — classes with labels $y \preceq u$
«1» — classes with labels $y \succeq u + 1$

$$\varphi(\mathbf{x}) = \begin{cases} \min_{u \in \mathbb{Z}} \{u \mid f_{u,u+1}(\mathbf{x}) = 0\}, & \text{if } \{u \mid f_{u,u+1}(\mathbf{x}) = 0\} \neq \emptyset; \\ z, & \text{if } \{u \mid f_{u,u+1}(\mathbf{x}) = 0\} = \emptyset. \end{cases}$$

1, 2	...	$u-1$, u	u, $u+1$	...	$z-1$, z
1	...	1	0	...	0

Multiclass classification example
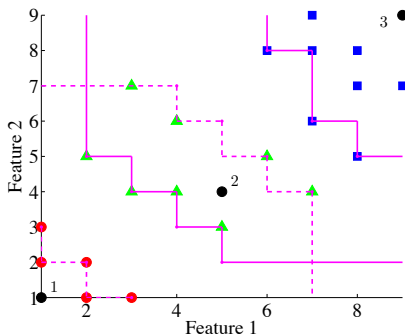
№	Object $\mathbf{x}$	$f_{12}(\mathbf{x})$	$f_{23}(\mathbf{x})$	$\varphi(\mathbf{x})$
1	(1,1)	0	0	1
2	(5,4)	1	0	2
3	(9,9)	1	1	3

Fronts extension for monotone classification
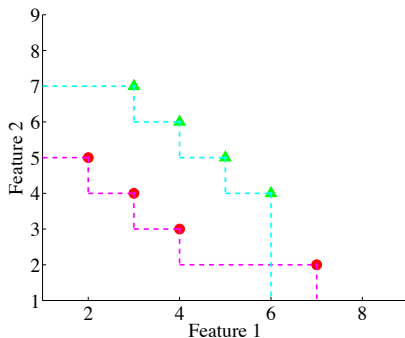
(e) Without extension
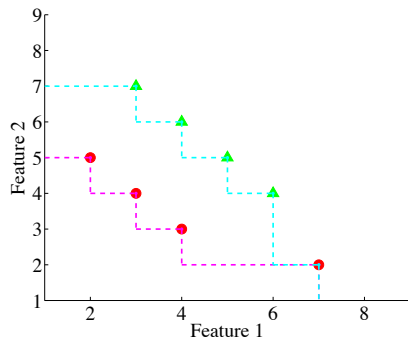
(f) With extension

A common object for two *n*-fronts

Admissible classifiers

Transitivity condition

$$\begin{cases} f_{u,u+1}(\mathbf{x}) = 0 \Rightarrow f_{(u+s)(u+1+s)}(\mathbf{x}) = 0 & \text{for each } s\colon (u+1+s) \leqslant z, \\ f_{u,u+1}(\mathbf{x}) = 1 \Rightarrow f_{(u-s)(u+1-s)}(\mathbf{x}) = 1 & \text{for each } s\colon (u-s) \geqslant 1. \end{cases}$$
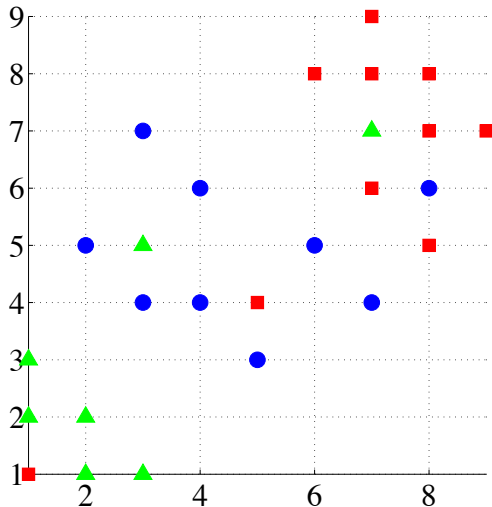
Definition

Classifier φ is called *admissible*, if for every classifier function $f_{u,u+1}$ the transitivity condition holds.

Theorem

If the Pareto optimal fronts $POF_n(u)$ and $POF_p(u+1)$ don't intersect for each $u = 1, \ldots, z-1$, then the transitivity condition holds for any classified object.
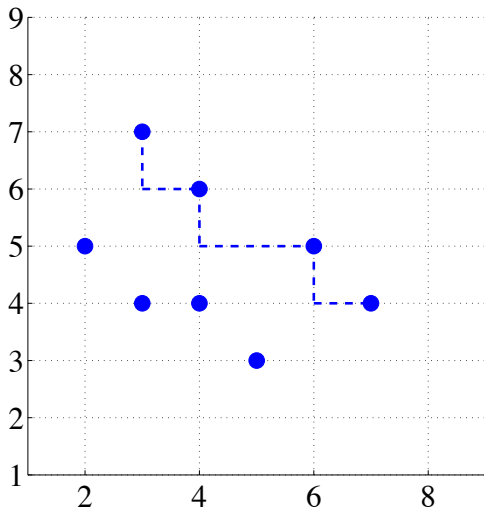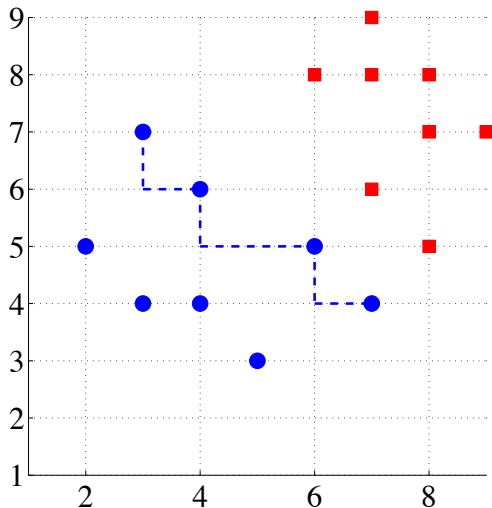
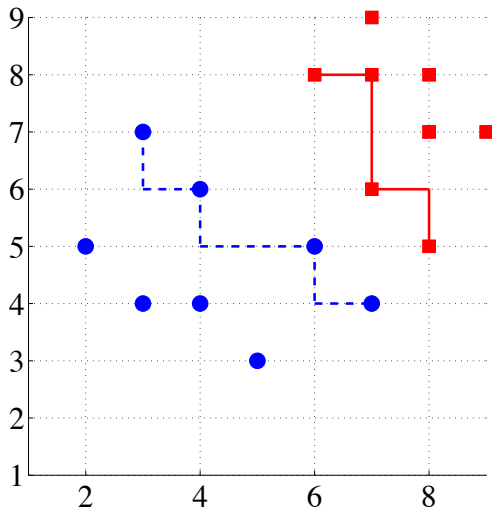# Initial sample of objects

Objects of the category 2
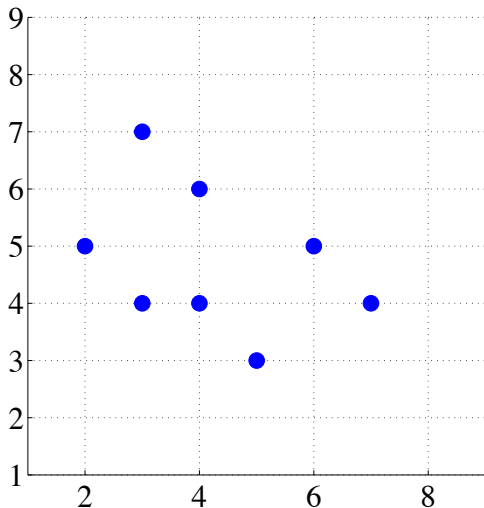
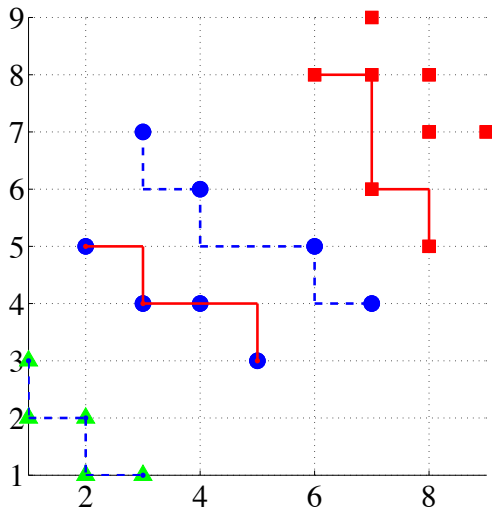## Optimal Pareto front (POF$_n$)
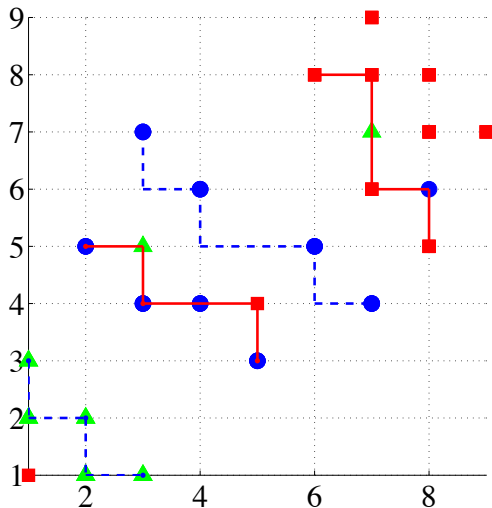
Objects of the category 2 and 3

Optimal Pareto fronts (POF$_n$, POF$_p$)

Model with all fronts

Excluded defective objects

Algorithms comparison

Algorithm	Mean error on test	LOO	Time of model construction, sec
POF (proposed)	0.22	0.56	2.1
Decision trees	0.25	0.69	0.4
Curvilinear regression [1]	0.57	0.71	3.6
Cones [2]	0.29	0.58	1.2
Copulas [3]	0.57	0.61	0.25

[1] 5. M.P. Kuznetsov, V.V. Strijov, M.M. Medvednikova Multiclass classification algorithm of the ordinal scaled objects // St. Petersburg State Polytechnical University Journal. Computer Science. Telecommunication and Control Systems, 2012. №. 5. C. 92-95.

[2] 1. M.P. Kuznetsov and V.V. Strijov. Methods of expert estimations concordance for integral quality estimation Expert Systems with Applications, 41(4):1988-1996, March 2014.

[3] Kuznetsov M.P. Integral indicator construction using copulas // Journal of Machine Learning and Data Analysis. 2012. V. 1, № 4. Pp. 411-419.