Постановка задач и выбор моделей в машинном обучении

Вадим Викторович Стрижов

Московский физико-технический институт

Осенний семестр 2019

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.*

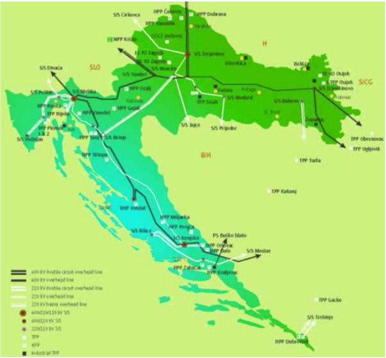
Examples

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 (T J)	Available net capacity (MW)	SO_2 (t)	NOx (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;
$\mathbf{y} = [y_1, \dots, y_m]^T$ is the vector of integral indicators;
$\mathbf{w} = [w_1, \dots, w_n]^T$ is the vector of feature importance weights;

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

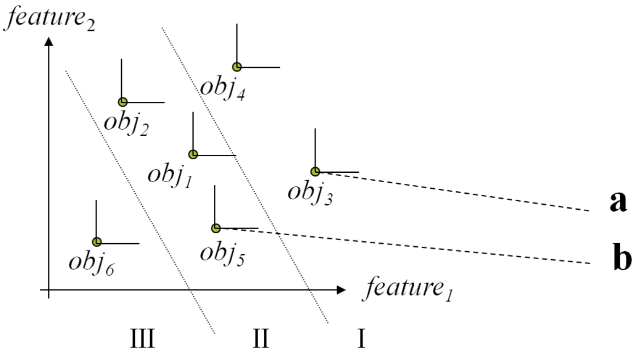
$$\begin{array}{c|c} & \mathbf{w}^T \\ \hline \mathbf{y} & X \end{array} \quad = \quad \begin{array}{c|cccc} & w_1 & w_2 & \dots & w_n \\ \hline y_1 & x_{11} & x_{12} & \dots & x_{1n} \\ y_2 & x_{21} & x_{22} & \dots & x_{2n} \\ \dots & \dots & \dots & \vdots & \dots \\ y_m & x_{m1} & x_{m2} & \dots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

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.

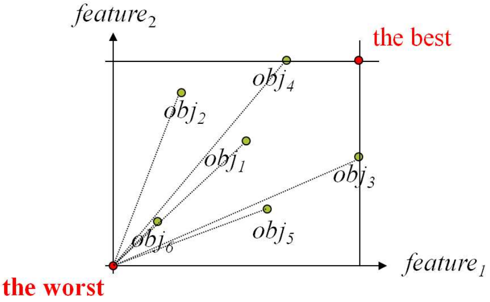
Metric algorithm

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.



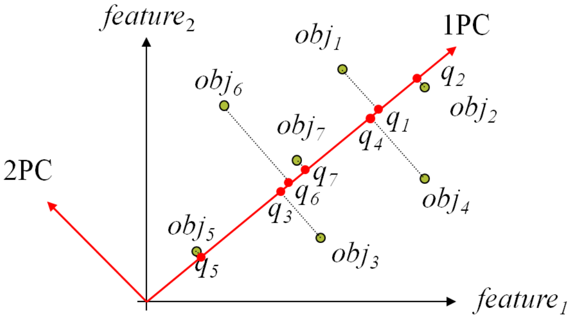
Weighted sum

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

$$\begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_m \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \dots & \dots & \vdots & \dots \\ x_{m1} & x_{m2} & \dots & x_{mn} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \dots \\ w_m \end{bmatrix}.$$

Principal Components Analysis

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

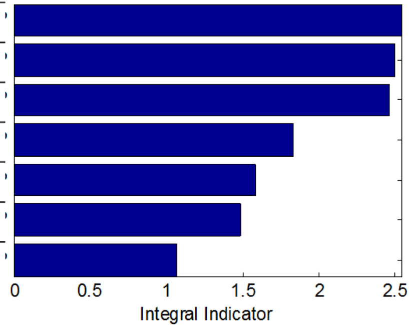


PCA gives minimum mean square error between objects and their projections.

The Integral Indicator

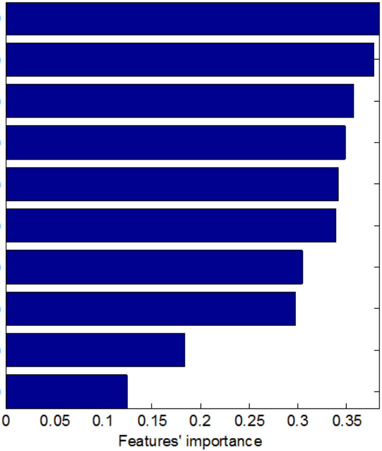
Ecological Impact of the Croatian Power Plants

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

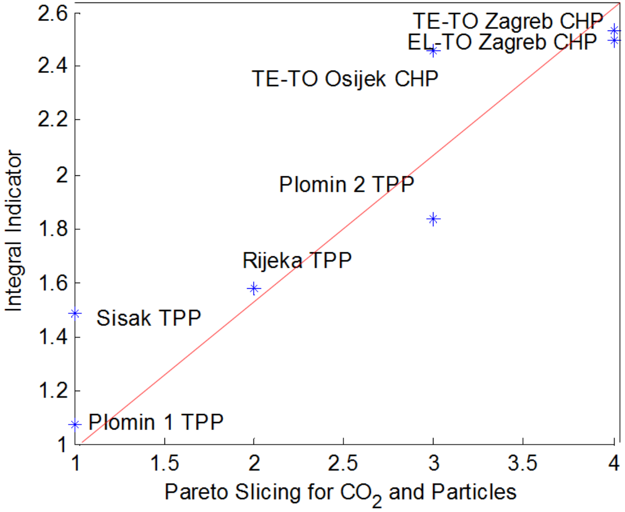


The Importance Weights of the Features

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

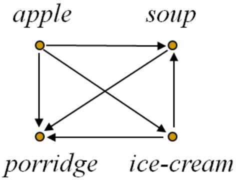


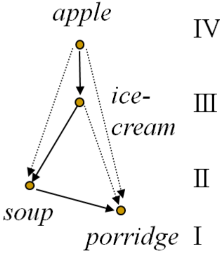
The PCA Indicator versus Pareto Slicing



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* $\longrightarrow$ *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:

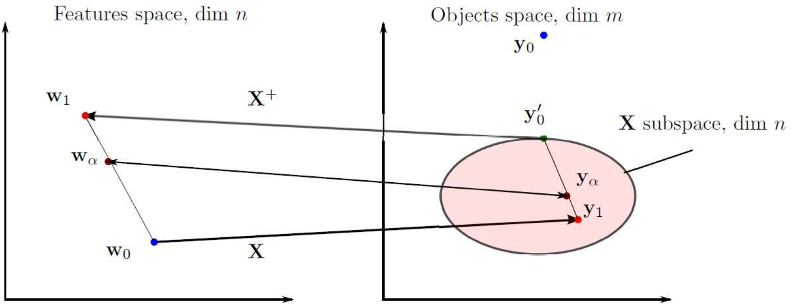
$$\mathbf{y} = \mathbf{X}\mathbf{w}, \quad \mathbf{w} = \mathbf{X}^+\mathbf{y}.$$

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 $\alpha \in [0; 1]$.

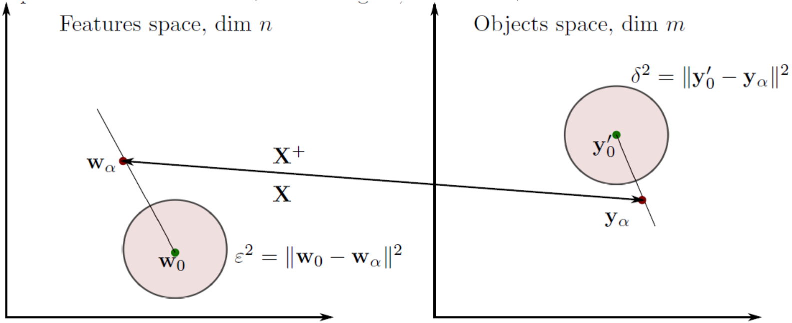


γ-concordance

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

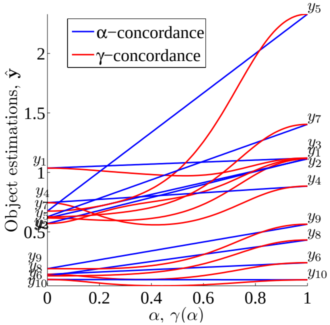


Concordance methods comparison

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:

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

In matrix notations:

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

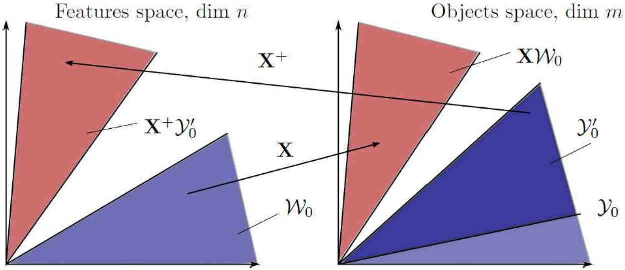
Consider **two cones** instead of two vectors:

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

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

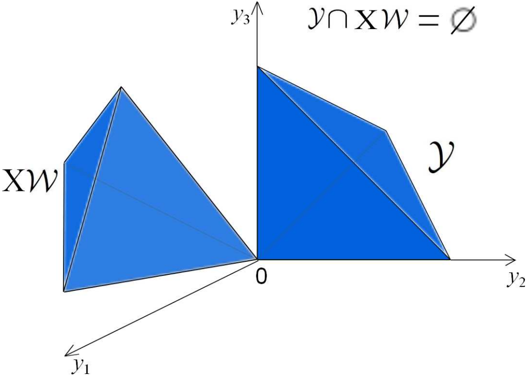
Ordinal specification

- 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 $\mathcal{Y}_0' = \mathbf{X}\mathbf{X}^+\mathcal{Y}_0$.



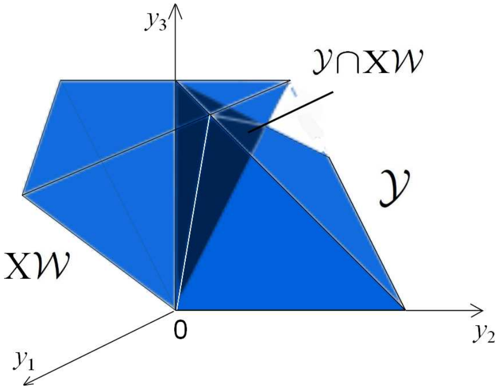
Cones intersection: specification is needed

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.



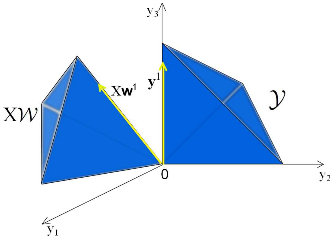
Nearest vectors in the cones

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. $w_1 \geqslant \ldots \geqslant w_n \geqslant 0$,
2. $\widetilde{\mathbf{w}} = X^+ \mathbf{y}_0$,

The problem of specification in rank scales:

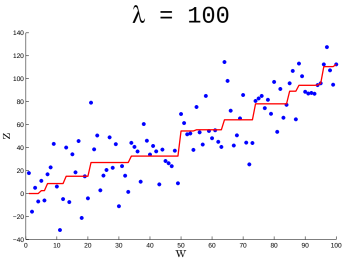
$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w} \in \mathbb{R}^n} \left(\underbrace{\frac{1}{2} \sum_{j=1}^{n} (\widetilde{w}_j - w_j)^2}_{\text{ref. to } \mathbf{y_0}} + \underbrace{\lambda \sum_{j=1}^{n-1} (w_j - w_{j+1})_+}_{\text{ref. to expert estimations of } \mathbf{w}} \right),$$

where λ is a regularizer.

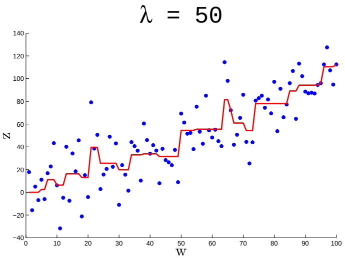
Nearly-isotonic regression algorithm: illustration

A blue dot is a feature weight.

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



(a)



(b)

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 $\mathfrak{D} = \{(\mathbf{x}_i, y_i)\}$, $i \in \mathcal{I} = \{1, \dots, m\}$.

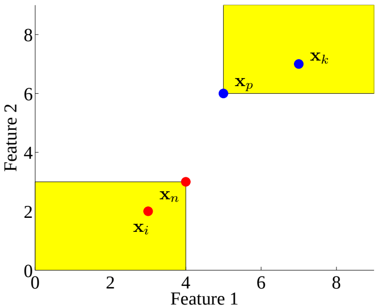
Ordinal scales and class labels

Every object $\mathbf{x} = [\chi_1, \dots, \chi_j, \dots, \chi_d]^T$, is described by ordinal-scaled features $\chi_j \in \mathbb{L}_j = \{1 \prec \dots \prec k_j\}$. 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 $\varphi\colon \mathbf{x} \mapsto \hat{y}$

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

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: $\mathbf{x} \not\succ_n \mathbf{x}$, $\mathbf{x} \not\succ_p \mathbf{x}$.

Dominance relation

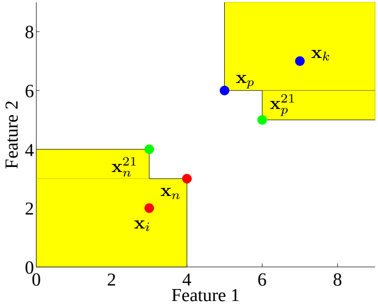
With features hierarchy

Leat a feature r be more important than t.

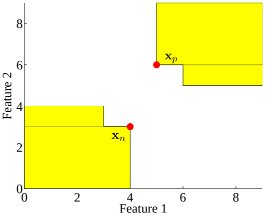
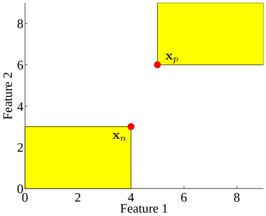
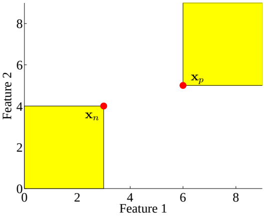
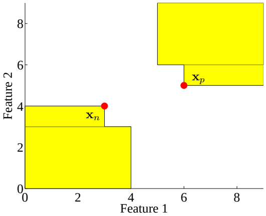
$\mathbf{x}_n \succ_{\tilde{n}} \mathbf{x}_i$, if $\mathbf{x}_n \succ_n \mathbf{x}_i$
or $x_{nr} > x_{nt}$ and $\mathbf{x}_n^{rt} \succ_n \mathbf{x}_i$.

$\mathbf{x}_p \succ_{\tilde{p}} \mathbf{x}_k$, 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

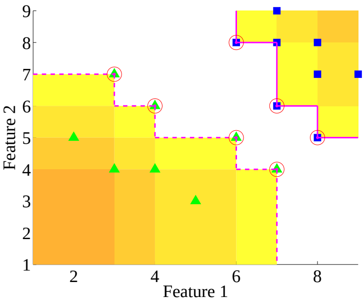
	The feature 1 is more important than 2	The feature 2 is more important than 1
$x_{n1} > x_{n2}$, $x_{p1} < x_{p2}$		
$x_{n1} < x_{n2}$, $x_{p1} > x_{p2}$		

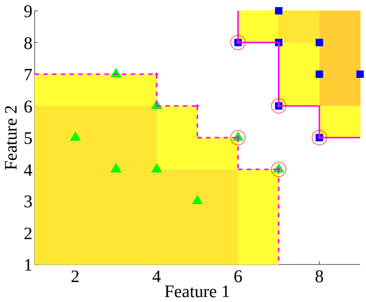
Optimal Pareto fronts

POF_n, POF_p

A set of objects $\mathbf{x}$, if for each element doesn't exist any other element $\mathbf{x}'$ such that

$\mathrm{POF}_n: \quad \mathbf{x}' \succ_n \mathbf{x}\ (\mathbf{x}' \succ_{\tilde{n}} \mathbf{x});\qquad \mathrm{POF}_p: \quad \mathbf{x}' \succ_p \mathbf{x}\ (\mathbf{x}' \succ_{\tilde{p}} \mathbf{x}).$





Two-class classification

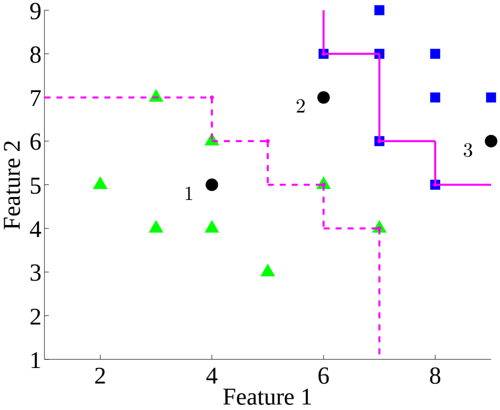
$\mathbf{x}$ — a classified object
$f(\cdot)$ — a classifier function

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

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

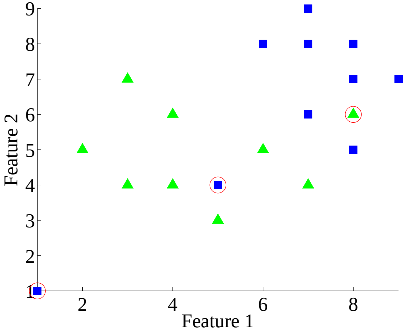
$$\rho(\mathbf{x}, \mathbf{x}') = \sum_{j=1}^{d} r(x_j, x'_j).$$

Two-class classification example

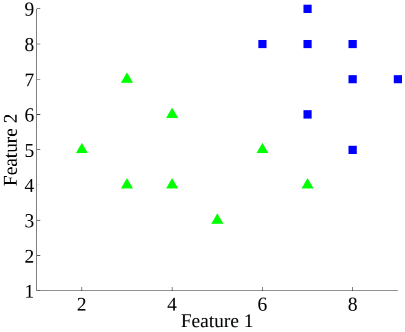


№	Object x	$f(x)$
1	(4,5)	0
2	(6,7)	1
3	(9,6)	1

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

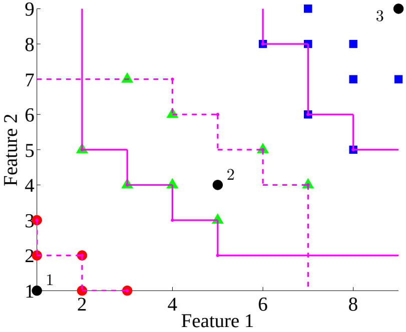
$f_{u,u+1} \colon \mathbf{x} \mapsto \hat{y} \in \{0, 1\}$ — two-class classifier for a pair of adjacent classes

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

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

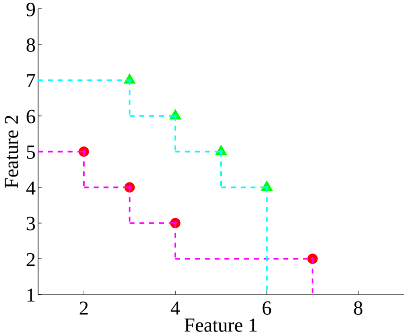
1, 2	...	$u-1$, u	u, $u+1$	...	$z-1$, z
1	...	1	0	...	0

Multiclass classification example

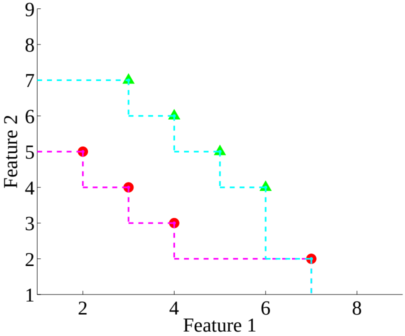


№	Object x	$f_{12}(x)$	$f_{23}(x)$	$\varphi(x)$
1	(1,1)	0	0	1
2	(5,4)	1	0	2
3	(9,9)	1	1	3

Fronts extension for monotone classification



(e) Without extension



(f) With extension

A common object for two n-fronts

Admissible classifiers

Transitivity condition

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

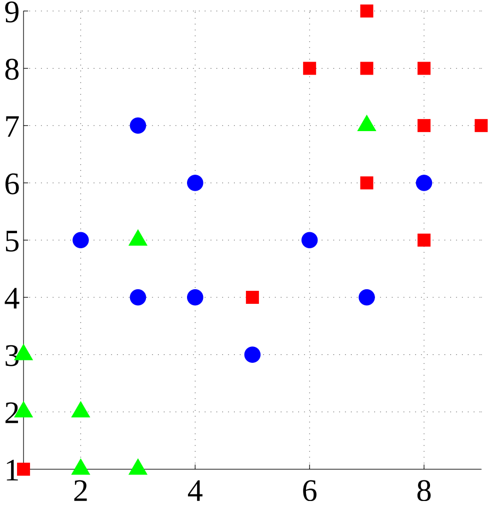
Definition

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

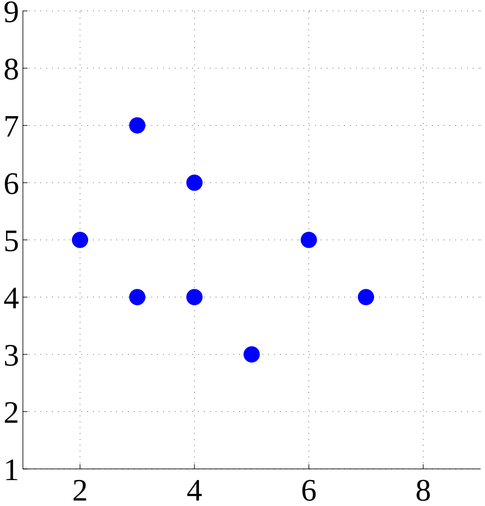
Theorem

If the Pareto optimal fronts $\mathrm{POF}_n(u)$ and $\mathrm{POF}_p(u+1)$ don't intersect for each $u = 1, \dots, z-1$, then the transitivity condition holds for any classified object.

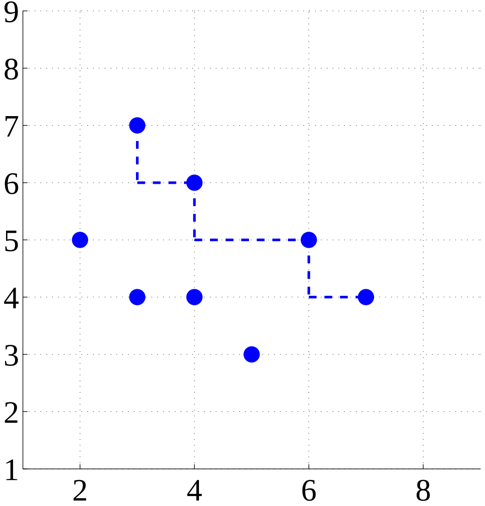
Initial sample of objects



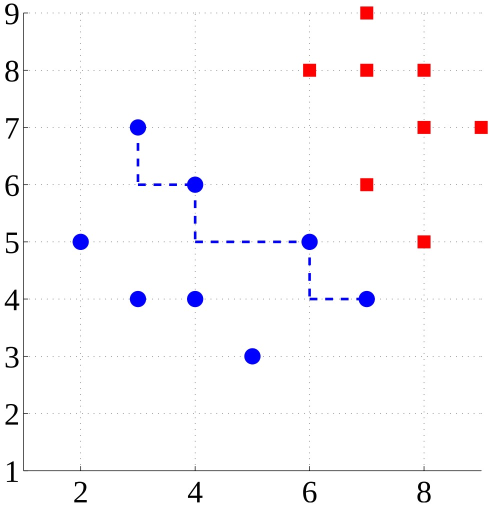
Objects of the category 2



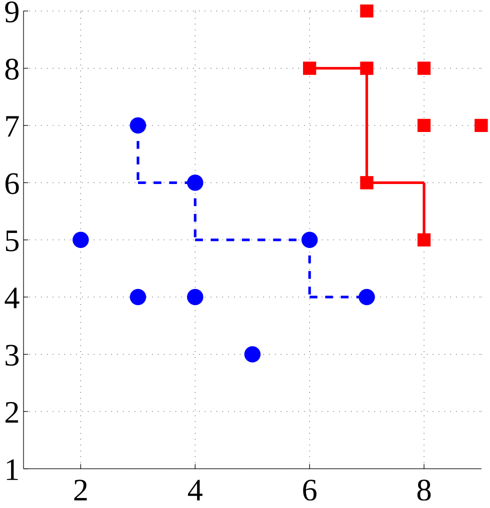
Optimal Pareto front (POF$_n$)



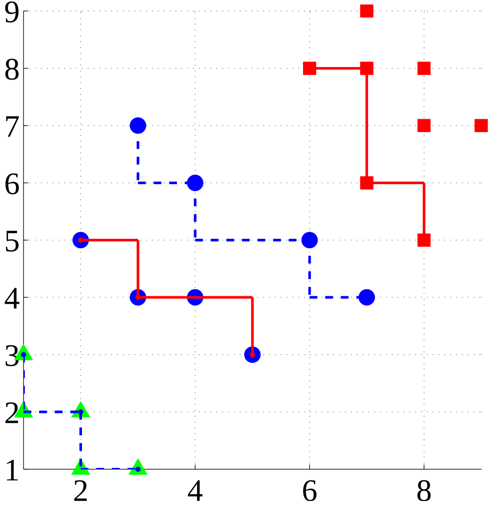
Objects of the category 2 and 3



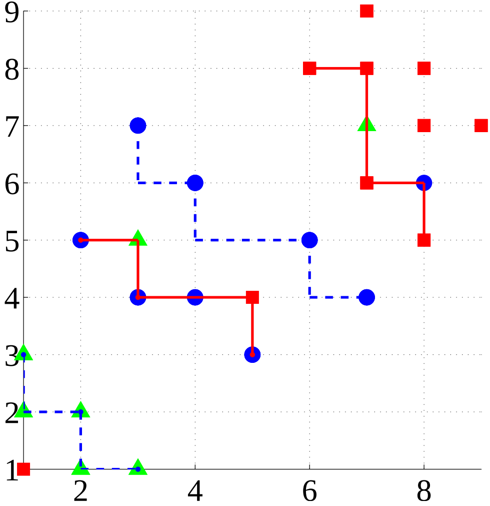
Optimal Pareto fronts (POF_n, POF_p)



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.