Постановка задач и выбор моделей в машинном обучении

Вадим Викторович Стрижов

Московский физико-технический институт

Осенний семестр 2019

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 $\mathbf{x}_i = [x_{i1}, \dots, x_{in}]^\mathsf{T}$,
x_{ij} is an element of a non-numeric matrix $\mathbf{X}$ with columns X_j:

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

Elements

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

belong to sets with given partial order relations.

The problem

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

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

- **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

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

Partial order cone

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

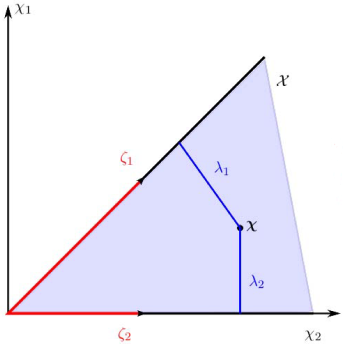
$$\mathcal{X}_0 = \{\chi \mid \mathbf{A}\chi \le \mathbf{0}\},$$

such that a string of matrix $\mathbf{A}$ 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,

$$\mathcal{X}_0 = \{\sum_{k=1}^{r} \lambda_k \zeta_k \mid \lambda_k \ge 0\}.$$



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 $f : \mathbb{L}_1 \times \ldots \times \mathbb{L}_n \to \mathbb{R}$, 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 $\mathcal{X}_1, \ldots, \mathcal{X}_n$:

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

Parametrization theorem [Kuznetsov: 2013]

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

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

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:

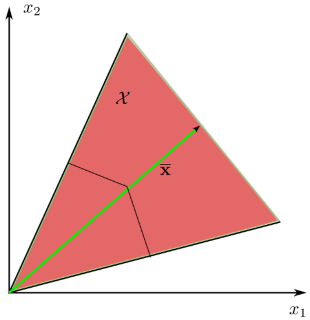
$$\hat{\boldsymbol{\lambda}} = \arg\min_{\boldsymbol{\lambda}_j \geq 0} \|\mathbf{y} - \sum_{j=1}^{n} \mathbf{Z}_j \boldsymbol{\lambda}_j\|_2^2.$$

Reduction of parameter space

Polyhedral model:

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

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

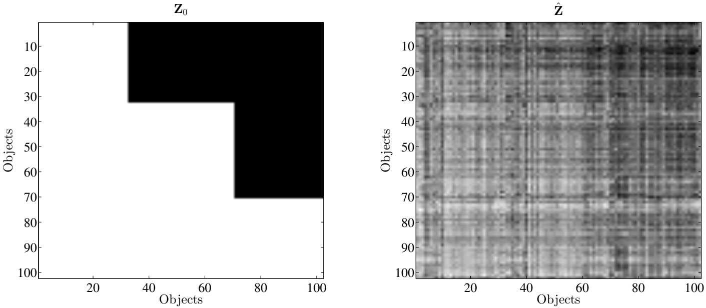
$$\mathbf{f}(\mathbf{X}) = \hat{\mathbf{Z}} \boldsymbol{\lambda},$$

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 $\mathbf{Z}_1, \ldots, \mathbf{Z}_n, \mathbf{Z}_Y$.
2. Estimate a matrix $\hat{\mathbf{Z}}$: $\hat{\mathbf{Z}}(i,k) = \sum_{j=1}^{n} w_j \mathbf{Z}_j(i,k)$.



3. Estimate parameters $\boldsymbol{\lambda}$: $\hat{\boldsymbol{\lambda}} = \arg\min_{\boldsymbol{\lambda}_j \geq \mathbf{0}} \|\mathbf{y} - \hat{\mathbf{Z}}\boldsymbol{\lambda}\|_2^2$.

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 $2k+1$:
$\hat{\boldsymbol{\lambda}}_y = \min\limits_{\boldsymbol{\lambda}_y \geq \mathbf{0}} \Vert \mathbf{Z}_y \boldsymbol{\lambda}_y - \mathbf{X}_w \hat{\boldsymbol{\lambda}}_w \Vert_2^2$	$\hat{\boldsymbol{\lambda}}_w = \min\limits_{\boldsymbol{\lambda}_w \geq \mathbf{0}} \Vert \mathbf{Z}_y \hat{\boldsymbol{\lambda}}_y - \mathbf{X}_w \boldsymbol{\lambda}_w \Vert_2^2$

- 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	Hablitat 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