A metalearning study for robust nonlinear regression

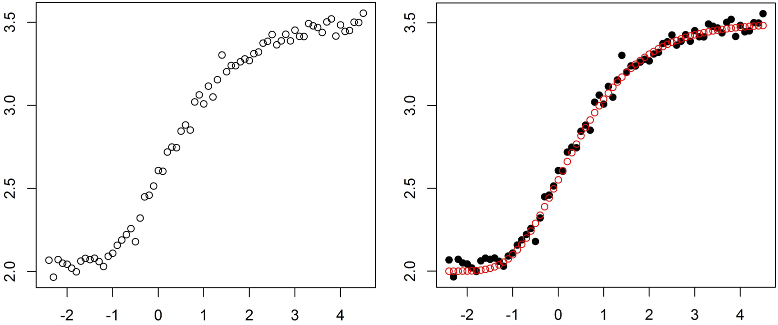
Jan Kalina & Petra Vidnerová

The Czech Academy of Sciences, Institute of Computer Science

Metalearning: motivation, principles

- Transfer learning for automatic method selection
- Automatic algorithm selection
- Empirical approach for (black-box) comparison of methods
- Attempt to generalize information across datasets
- Learn prior knowledge from previously analyzed datasets and exploit it for a given dataset
- A dataset (instance) viewed as a point in a high-dimensional space

Nonlinear regression



- Model

$$Y_i = f(\beta_1 X_{i1}, \ldots, \beta_p X_{ip}) + e_i, \quad i = 1, \ldots, n$$

- Nonlinear least squares (NLS)
- Minimal sum of squares
- Vulnerability to outliers

Nonlinear least weighted squares estimator (NLWS)

$$Y_i = f(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}) + e_i, \quad i = 1, \ldots, n$$

- $Y = (Y_1, \ldots, Y_n)^T$ = continuous outcome
- f = given nonlinear function
- Nonlinear least squares: sensitive to outliers

Residuals for a fixed $\mathbf{b} = (b_0, b_1, \ldots, b_p)^T \in \mathbb{R}^{p+1}$:

$$u_i(b) = Y_i - f(b_0 - b_1 X_{i1} - \cdots - b_p X_{ip}), \quad i = 1, \ldots, n$$

Squared residuals arranged in ascending order:

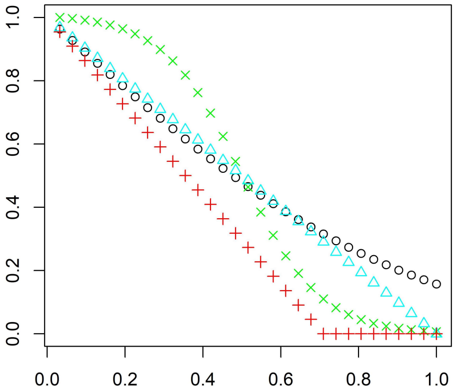
$$u^2_{(1)}(\mathbf{b}) \leq u^2_{(2)}(\mathbf{b}) \leq \cdots \leq u^2_{(n)}(\mathbf{b}).$$

Nonlinear least weighted squares (NLWS):

$$\mathbf{b}_{LWS} = \arg\min \sum_{i=1}^{n} w_i u^2_{(i)}(b) \quad \text{over} \quad \mathbf{b} = (b_0, b_1, \ldots, b_p)^T \in \mathbb{R}^{p+1},$$

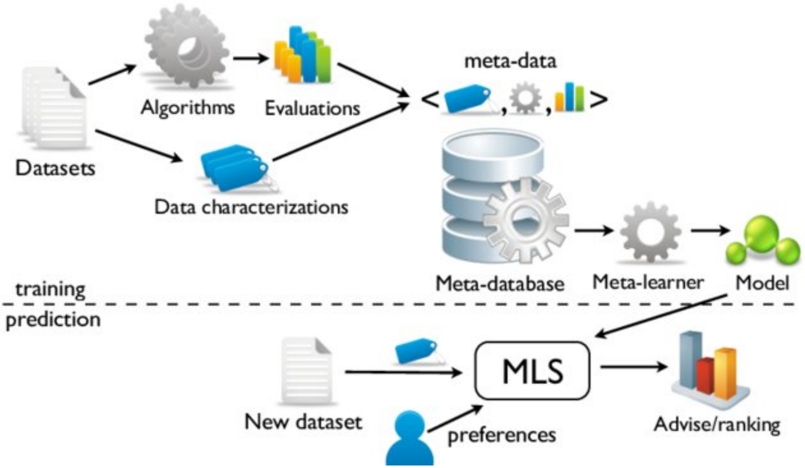
where $w_1, \ldots, w_n$ are given magnitudes of weights.

Examples of weight functions:



- $\sum_{i=1}^{n} w_i = 1$
- Nonlinear least trimmed squares (NLTS): 0-1 weights

Meta-learning process



- We start with 2000 real publicly available datasets (github)
- Diversity of domains
- Automatic downloading
- Pre-processing in Python
- Reducing n
- Missing values
- Categorical variables
- Reducing p
- $$Y_i \longmapsto \frac{Y_i - \min_i Y_i}{\max_i Y_i - \min_i Y_i}, \quad i = 1, \ldots, n$$
- Standardizing continuous regressors

Description of the metalearning study)

- Datasets
  - 721 real datasets
- Algorithms
  - Fully automatic, including finding suitable parameters
  - Least squares & 6 robust nonlinear estimators (NLTS, NLWS with various weights, nonlinear regression median)
- Prediction measure
  - Mean square error (MSE) evaluated within a cross validation
  - Robust versions: trimmed MSE (TMSE), weighted MSE (WMSE)

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n} r_i^2, \quad \mathrm{TMSE}(\alpha) = \frac{1}{h}\sum_{i=1}^{h} r_{(i)}^2, \quad \mathrm{WMSE} = \sum_{i=1}^{n} w_i r_{(i)}^2$$

- Features of the datasets
  - 9 features
- Metalearning (performed over metadata)
  - Classification by means of various classifiers

Selected 10 features of the datasets

1. The number of observations n
2. The number of variables p
3. The ratio n/p
4. Normality of residuals (p-value of Shapiro-Wilk test)
5. Skewness of residuals
6. Kurtosis of residuals
7. Coefficient of determination R^2,
8. Percentage of outliers (estimated by the LTS) – important!
9. Heteroscedasticity (p-value of Breusch-Pagan test)
10. Donoho-Stahel outlyingness measure of X

- Model

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} + \sum_{j=1}^{p} \beta_{p+j} (X_{ij} - \bar{X}_j)^2 + e_i, \quad i = 1, \dots, n$$

- Leave-one-out cross validation
- MSE:
  - NLS yields the minimal prediction error for 23 % of the datasets,
  - NLTS 26 %
  - any of the versions of the **NLWS** 31 %
  - nonlinear median 20 %
- TMSE:
  - **NLTS** best for 39 % of datasets
  - NLWS 35 %
- WMSE:
  - NLTS best for 34 % of datasets
  - **NLWS** 45 %
- Weights for the NLWS: no choice uniformly best

Results of metalearning evaluated as the classification accuracy in a leave-one-out cross validation study. Three different prediction error measures are compared.

Classification method	Classification accuracy MSE	TMSE	WMSE
Classification tree	0.35	0.45	0.47
k-nearest neighbor ($k = 3$)	0.56	0.61	0.64
LDA	0.60	0.68	0.65
SCRDA	0.60	0.68	0.66
Linear MWCD-classification	0.60	0.68	0.66
Multilayer perceptron	0.56	0.66	0.66
Logistic regression	0.56	0.67	0.69
SVM (linear)	0.60	0.69	0.70
SVM (Gaussian kernel)	0.64	0.71	0.70

Conclusions

- First comparison of robust nonlinear regression estimates
- 721 datasets
- Arguments in favor of the NLWS estimator (robustness & efficiency)
- Metalearning is useful
- Future work: **robust** metalearning

Limitations of metalearning:

- No theory
- Number of methods/algorithms/features
- Choice of datasets
- Too automatic
- Correct pre-processing (incl. variable selection) of data needed!