Metalearning for Robust Regression: Sensitivity and Robustification

Petra Vidnerová, Jan Kalina, Aleš Neoral

The Czech Academy of Sciences, Institute of Computer Science

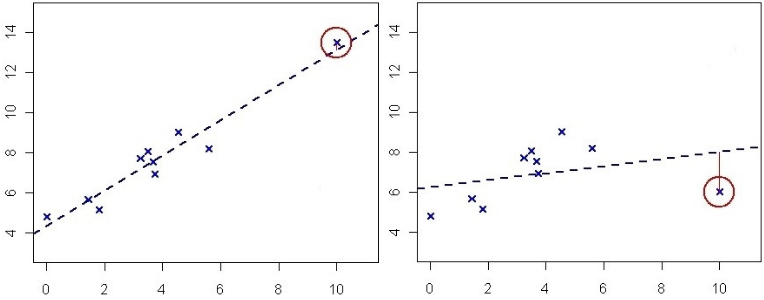
AMISTAT 2019

Robust regression

1. **Jurečková** J., Sen P.K., Picek J. *Methodology in robust and nonparametric statistics.* CRC Press, Boca Raton, 2013.
2. Huber P.J. *Robust statistics.* Wiley, New York, 1981.
3. Rousseeuw P.J., Leroy A.M. *Robust regression and outlier detection.* Wiley, New York, 1987.
4. Víšek J.Á. (2011): Consistency of the least weighted squares under heteroscedasticity. *Kybernetika* **47** (2), 179 – 206.
5. Čížek P. (2011): Semiparametrically weighted robust estimation of regression models. *Computational Statistics and Data Analysis* **55**, 774 – 788.

Outliers in linear regression

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

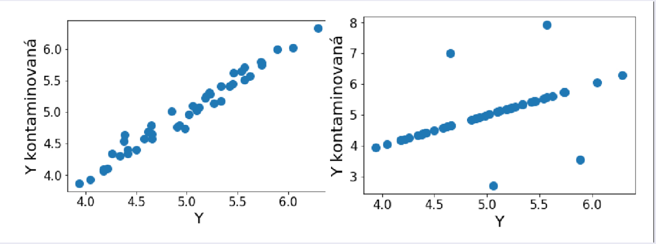


- Outliers vs. leverage points
- Outlier detection: masking and swamping effects

Robust regression

Contamination

Local vs. global contamination



Methods

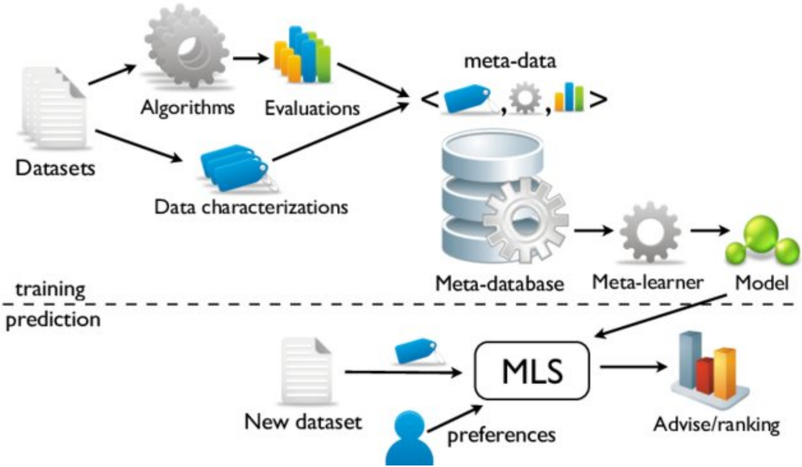
- M-estimators (Huber's estimator, Hampel's estimator)
- Least trimmed squares
- Least weighted squares

Standard metalearning

1. Brazdil P., Giraud-Carrier C., Soares C., Vilalta E. (2009): *Metalearning: Applications to data mining*. Springer, Berlin.
2. Rice, J.R. (1976): The algorithm selection problem. Advances in Computers **15**, 65 – 118.
3. Smith-Miles K., Baatar D., Wreford B., Lewis R. (2014): Towards objective measures of algorithm performance across instance space. *Computers and Operations Research* **45**, 12 – 24.
4. Smith-Miles K.A. (2009): Cross-disciplinary perspectives on meta-learning for algorithm selection. *ACM Computing Surveys* **41**, Article 6.

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



[fig by Joaquin Vanschoren]

Description of standard metalearning (Smith-Miles, 2009)

- Datasets
  - Typically not very many
  - Real datasets (simulated datasets are biased)
  - We consider 271 datasets
- Algorithms
  - Fully automatic, including finding suitable parameters
  - Least squares, Huber's M, Hampel's M, LTS ($h = \lfloor n/2 \rfloor$ and $h = \lfloor 3n/4 \rfloor$)
- Prediction measure
  - Mean square error (MSE) evaluated within a cross validation
- Features of the datasets
  - How many (there should not be too many)
  - Relevant for the model selection
  - Their choice requires to understand the primary task
- Metalearning (performed over metadata)
  - Typically a classification task

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

Results of primary learning

Data set	Ranks according to MSE				
	(1)	(2)	(3)	(4)	(5)
Aircraft	5	3	4	1	2
Ammonia	5	3	4	2	1
Auto MPG	3	2	1	4	5
Cirrhosis	2.5	1	2.5	5	4
Coleman	1	2	4	5	3
Delivery	5	4	2	3	1
⋮	⋮	⋮	⋮	⋮	⋮

- Leave-one-out cross validation
- (1) Least squares
- (2) Huber's M-estimator
- (3) Hampels's M-estimator
- (4) LTS with $h = \lfloor n/2 \rfloor$
- (5) LTS with $h = \lfloor 3n/4 \rfloor$
- Most often: LTS is the best

Results of metalearning

Method	Clas. accuracy
LDA	0.30
SVM (linear)	0.40
SVM (polynomial)	**0.43**
SVM (radial)	**0.43**
SVM (sigmoid)	0.40
k-NN (k=1)	0.30
k-NN (k=3)	0.30
k-NN (k=5)	0.33

- Classification accuracy in a leave-one-out cross validation
- Methods (and their principles):
  - LDA: linear discriminant analysis
  - SVM: support vector machine
  - k-NN: k-nearest neighbor

Study 1

- Implementation in Python
- At first, we downloaded about 2000 datasets
- `https://vincentarelbundock.github.io/Rdatasets/datasets.html`
- Pre-processing
  - Categorial variables
  - Missing values
  - Make the datasets homogeneous
- Finally: 721 real datasets
- Least squares, Hampel's M-estimator, LTS with $h = \lfloor 3n/4 \rfloor$, LWS
- 10 features
- classification accuracy evaluated in a leave-one-out cross validation: 59%
- better than random choice
- better than choosing the most frequent winner

Study 2

Classification accuracy in crossvalidation study, if using **MSE**:

	Contamination		
Classifier	None	Local	Global
SVM	0.59	0.49	0.33
Logistic Regression	0.59	0.50	0.36
LDA	0.59	0.50	0.35
KNN	0.59	0.48	0.36

Classification accuracy in crossvalidation study, if **trimmed MSE**:

	Contamination		
Classifier	None	Local	Global
SVM	0.45	0.36	0.36
Logistic Regression	0.53	0.43	0.41
LDA	0.53	0.43	0.41
KNN	0.53	0.40	0.41

Study 3

- Improving metalearning
- Automatic dimensionality reduction by means of t-tests

Classification accuracy in crossvalidation study, if **trimmed MSE**:

	Dimensionality reduction	
Classifier	No	Yes
SVM	0.45	0.50
Logistic Regression	0.53	0.58
LDA	0.53	0.57
KNN	0.53	0.56

Pros and cons of metalearning

Advantages of metalearning:

- Extracting knowledge from previously analyzed datasets
- No theoretical analysis needed
- Clear, simple, comprehensible
- Computationally feasible
- Popular in computer science

Limitations:

- No theory
- Number of methods/algorithms/features
- Choice of datasets
- Too automatic
- The problem itself is unstable and the whole process should be robustified

Conclusion

What we recommend for application of metalearning:

- Avoid the choice of very different datasets
- Choose carefully the prediction measure (MSE vs. TMSE)
- Classification instead of regression
- Correct pre-processing (incl. variable selection) of data needed!

Thank you!
Questions?