Kharkevich Institute For Information Transmission Problems, RAS

DATADVANCE
AN EADS COMPANY

On the criterion for selection of regression model

Burnaev E.V., Prihodko P.V., Panin I.I.

We present a method for selection of regression model (The Hybrid criterion) which allows to detect inaccurate regression models by inspecting their variability and deviation from piece-wise linear approximation.

- An unknown dependency $Y=f(X), X\in \boldsymbol{X}\subset \boldsymbol{R}^n, Y\in \boldsymbol{R}^1$ is given by training sample: $\boldsymbol{Train}=\{(X_i, Y_i=f(X_i)), \mathrm{i}=1,..,N_{Train}\}$. **The general problem** is to construct approximation function: $\hat{f}:\ \boldsymbol{X}\to Y,\ \hat{f}\approx f$. Quality of the approximation constructed is estimated by the mean error of approximation on the test set (particularly, MAE):

$$Err_{Test}(\hat{f})=\|Y-\hat{f}(X)\|^m_{Test}, \boldsymbol{Test}=[(X_i, Y_i=f(X_i)),\ \mathrm{i}=1,...,N_{test}],\quad \|Y-\hat{f}(X)\|^{MAE}_{Test}=\frac{1}{N_{test}}\sum_{i=1}^{N_{test}}|Y_i-\hat{f}(X_i)|$$

Hybrid criterion

- **Assume that there is a set of models** which was constructed for one training sample:
  $\boldsymbol{H}(\boldsymbol{Train}, param)=[\hat{f}_j(X), j=1,2,...,J]$.

  Our purpose is to find the best (or at least not the worst) model in terms of mean error on the test set.

- **The method proposed:**

  Does not construct additional approximations $\hat{f}$.

  Makes no assumptions about the distribution of input data.

  Does not depend on the particular type of model approximation.

- **The idea of the Hybrid criterion is to apply successively piece-wise linear criterion, criterion based on model variability and criterion based on train error taking into account their significance.**

Piece-wise linear criterion

- Let $Lin_{err} = \|\hat{f}_{lin}(X) - \hat{f}(X)\|^m, X \in \boldsymbol{X}_{valid}$

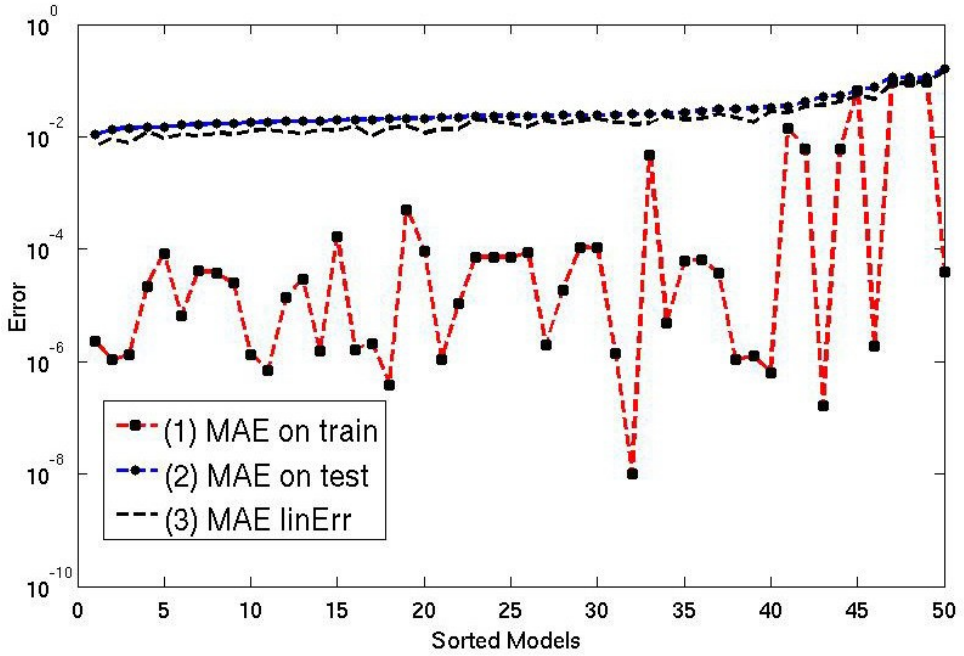
  $\hat{f}_{lin}$ - piece-wise linear approximation constructed by means of multi-dimensional Delaunay triangulation of $\boldsymbol{X}_{train}$ and linear regression.

  $\boldsymbol{X}_{valid}$ - validation set constructed by means of multi-dimensional Delaunay triangulation with with a subsequent generation of random points at each *simplex.*

- According to this criterion, models with the biggest deviation from piece-wise linear approximation are supposed to be 'bad'.

An example of good correlation between errors on the test and piecewise-linear approximation error (Lin_err's).

As one can see, the bigger test error the bigger Lin_err.



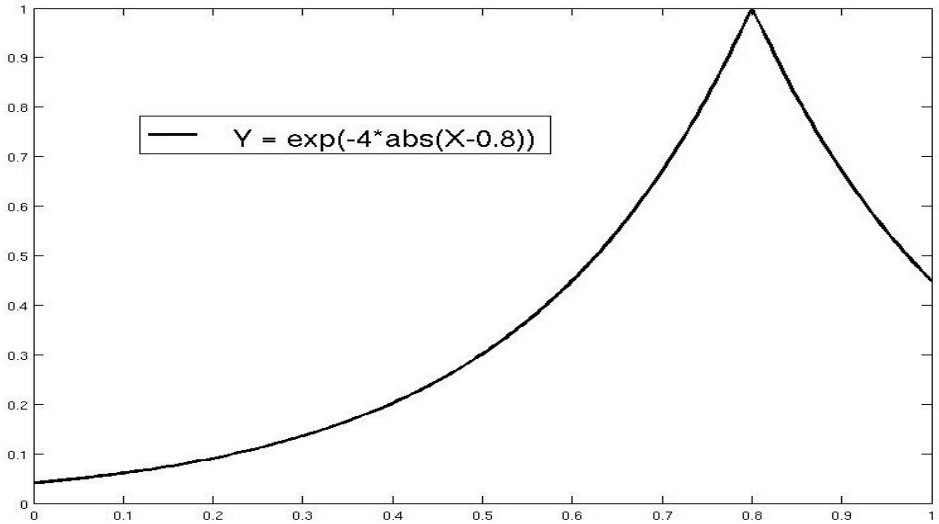
Criterion based on model variability

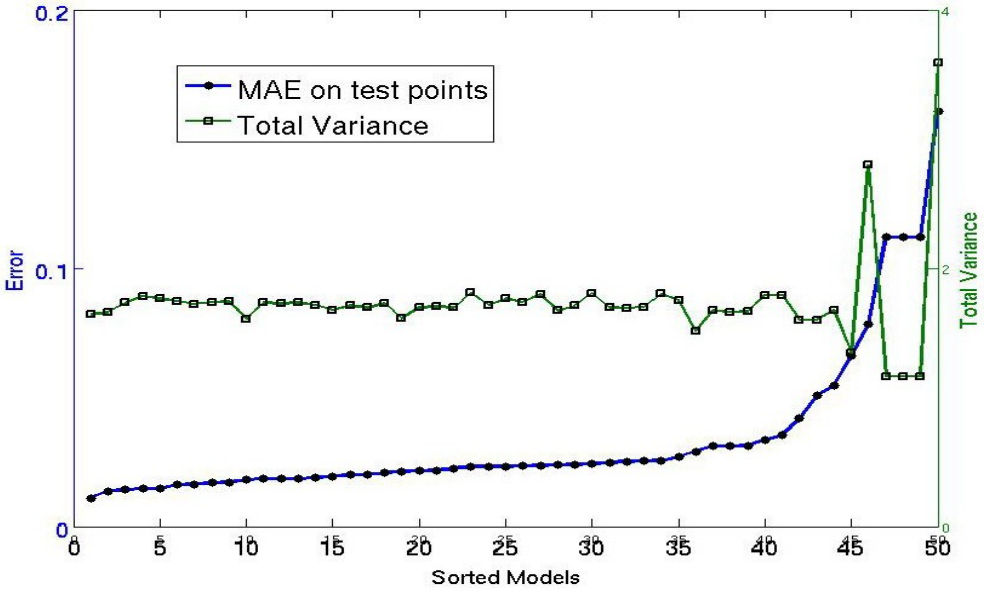
- Let **model variability**:

$$TVar = \frac{\sum_{l=1}^{N_{valid}} |grad \ \ \hat{f}(X_l)|}{N_{valid}} \ , X_l \in \boldsymbol{X}_{valid}$$

X_{valid} - validation set constructed by means of multi-dimensional Delaunay triangulation with a subsequent generation of random points at each *simplex.*

- Shown in pictures are a function and TVar's of approximation models sorted by error on test set. As one can see, the bigger test error the bigger change in TVar's.
- According to this criterion, models with the smallest and the biggest TVar are supposed to be 'bad'.

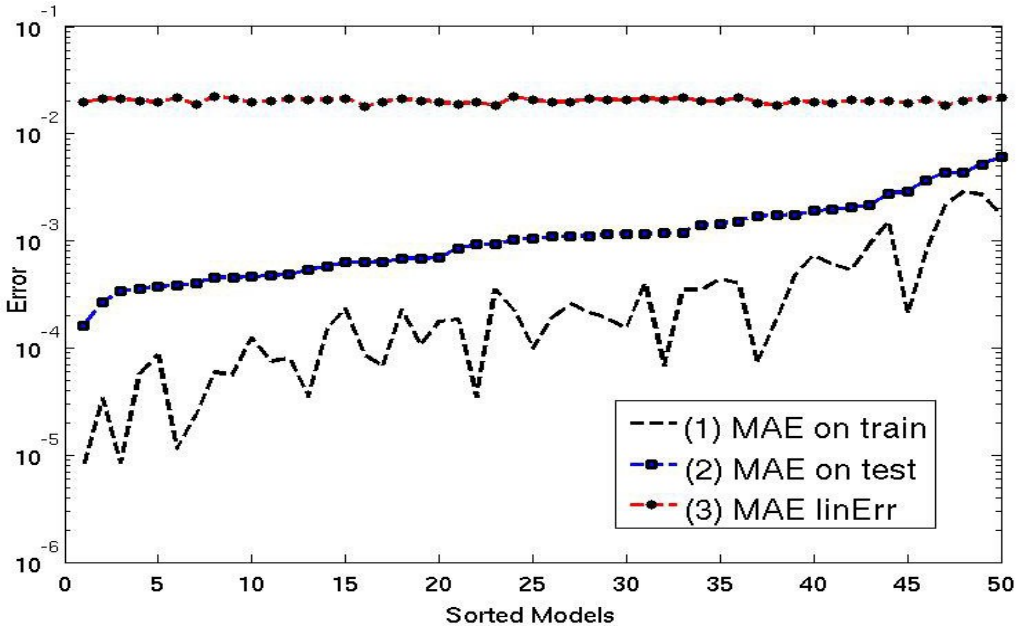




Separability

- In some cases criteria values for different models can be poorly separable. It is often due to the fact that these criteria can be too rough and their values can be poorly correlated with the error on the test.
- Let **separability of criterion**:
$$V(\boldsymbol{Q}) = \frac{max(\boldsymbol{Q}) - min(\boldsymbol{Q})}{max(\boldsymbol{Q}) + min(\boldsymbol{Q})},$$
where Q – a set of criterion values on models.
- Introduce a threshold separability and assume that if $V < V_{0}$ then a criterion did not work.

Pic. An example of poor correlation between errors on the test and piecewise-linear approximation error.



Hybrid criterion algorithm

- **If separability of piece-wise linear criterion is bigger than some threshold α then this criterion is applied.**
- **Else if separability of criterion based on model variability is bigger than some threshold β , then this criterion is applied.**
- **If the first two criteria do not work, then we choose a model that has minimal error on the training sample.**

Remark. Values of the parameters α and β can be selected empirically.

Computational Experiment

- It was considered 14 one-dimensional functions (60 samples for each function) and 19 two-dimensional functions (20 samples for each function). There was constructed 10 approximators for each sample.
- We tested the efficiency of finding not the worst model using different criteria. The measure of the efficiency was **error of determination** — a proportion of the cases when the criterion chose the worst out of 10 model (in terms of MAE on test).
- The parametras α and β was equal to 0.3 and 0.3 respectively.

Results

Errors of determination. 1d functions.

Errors of determination for two cases I **Criteria**	**Lin_err**	**Err_Train**	**TVar**	**Hybrid**
Without threshhold	6.2%	3.5%	6.8%	1.8%
In Hybrid criterion	2.6%	0.44%	5.9%	

Errors of determination. 2d functions.

Errors of determination for two cases I **Criteria**	**Lin_err**	**Err_Train**	**TVar**	**Hybrid**
Without threshhold	6.3%	4.7%	6.6%	1.6%
In Hybrid criterion	1.0%	1.3%	4.4%	

Conclusion

- The introduction of a threshold level for the piece-wise linear criterion and the criterion based on model variability increased the efficiency of their work.
- The Hybrid criterion allows to select bad models more effectively than separate criteria proposed.

Thank you for your attention!