Improving Model Selection
by Employing the Test Data

Max Westphal, Werner Brannath

University of Bremen, Germany
Institute for Statistics
Research Training Group $\pi^3$

mwestphal@uni-bremen.de
https://github.com/maxwestphal/

ICML 2019, Long Beach
June 11, 2018
Train-Validation-Test Split

Train

\[ X \xrightarrow{f_1} \hat{Y} \]
\[ X \xrightarrow{f_2} \hat{Y} \]
\[ X \xrightarrow{f_3} \hat{Y} \]
\[ X \xrightarrow{f_4} \hat{Y} \]
\[ \vdots \]
\[ X \xrightarrow{f_M} \hat{Y} \]

Validation

\[ \hat{\vartheta}_V \]
\[ \hat{\vartheta}_V \]
\[ \hat{\vartheta}_V \]
\[ \hat{\vartheta}_V \]
\[ \hat{\vartheta}_V \]

Learning

Max Westphal

Improving Model Selection
Train-Validation-Test Split

Training

Validation

Max Westphal
Improving Model Selection

Learning

\[ \hat{Y} \]

\[ \hat{f}_1 \]

\[ \hat{f}_2 \]

\[ \hat{f}_3 \]

\[ \hat{f}_4 \]

\[ \hat{f}_M \]

\[ A_1 \]

\[ A_2 \]

\[ A_3 \]

\[ A_4 \]

\[ A_M \]

\[ \hat{\varphi}_1 \]

\[ \hat{\varphi}_2 \]

\[ \hat{\varphi}_3 \]

\[ \hat{\varphi}_4 \]

\[ \hat{\varphi}_M \]
Train-Validation-Test Split

Training

Validation

\[ \text{argmax} \]

Learning

Max Westphal

Improving Model Selection

Max Westphal
Particularly in regulated environments, we need a reliable performance assessment before implementing a prediction model in practice.

Example: disease diagnosis / prognosis based on clinical data

Usually recommended strategy:

Evaluate a single final model on independent test data.
Train-Validation-Test Split

Max Westphal

Improving Model Selection

Evaluation

\[
\begin{align*}
X \xrightarrow{f_1} \hat{Y} \\
X \xrightarrow{f_2} \hat{Y} \\
X \xrightarrow{f_3} \hat{Y} \\
X \xrightarrow{f_4} \hat{Y} \\
\vdots \\
\vdots \\
X \xrightarrow{f_M} \hat{Y}
\end{align*}
\]

\[
\delta_4^E + \text{CI}_4 + \varphi_4
\]
Particularly in regulated environments, we need a reliable performance assessment before implementing a prediction model in practice.

Example: disease diagnosis / prognosis based on clinical data

Usually recommended strategy:

Evaluate a single final model on independent test data!

Easy-to-use strategy, allowing for a reliable performance assessment and simple inference.

However, we have no way to fix a bad model selection after having observed the test data.
Simultaneous Model Evaluation

Training

Validation

Learning

Max Westphal

Improving Model Selection
Simultaneous Model Evaluation

Test

Evaluation

Max Westphal

Improving Model Selection
**Simulation study**

**Learning**
- **Idea:** simulate data and train, select and evaluate binary classifiers in different scenarios
  - 24 artificial classification tasks
  - 72,000 replications of complete ML pipeline
  - 28,800,000 distinct models (EN, CART, SVM, XGB)

**Evaluation**
- **Goal:** comparison of different evaluation strategies
  - **default:** best validation model only
  - **within 1 SE:** all models within 1 SE of best validation model
Simulation Results

\[ n_{\text{learn}} = 400 \]

\[ n_{\text{learn}} = 800 \]

Performance gain vs. \( n_{\text{test}} \) for different values of \( n_{\text{learn}} \). The box plots show the distribution of performance gain for each \( n_{\text{test}} \) value.
Simulation Results

False & true positive test decisions

n_{test} = 200
n_{test} = 400
n_{test} = 800

FPR  TPR

Selection rule: default, within1SE
When in doubt, delay the final model choice to the test data.

- Improvements in model performance and probability to correctly identify a good model in all investigated scenarios.
- Adjustment for multiple comparisons via approximate parametric procedure taking into account model similarity (maxT-approach).

Questions & feedback welcomed!
- mwestphal@uni-bremen.de
- https://github.com/maxwestphal/

POSTER #123 (Pacific Ballroom)