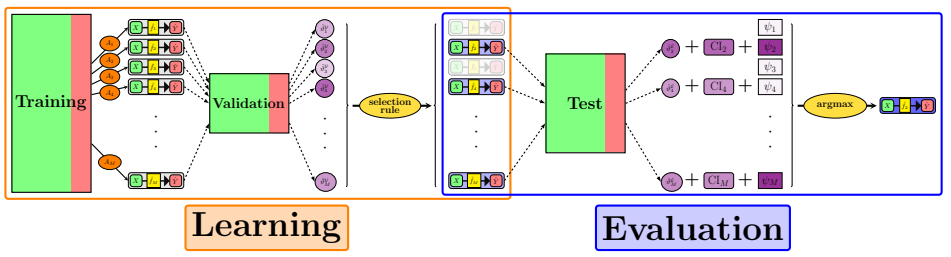
PhD projects R3

Model Selection and Evaluation in Supervised Machine Learning
Post-Selection Inference for Prediction Performance

Dr. Max Westphal, Pascal Rink, Prof. Werner Brannath

RTG 2224

Model Selection and Evaluation in Supervised ML



The predominant recommendation in the machine learning literature is to strictly separate model development (training and selection) and evaluation by means of data splitting. Following this *default* strategy to evaluate only a single final model on a distinct test dataset has two main advantages: (a) The generalization performance (or error) can be estimated without bias and (b) it is easy to test statistical hypotheses regarding the final performance. However, this approach is inflexible as one cannot alter the final model choice after the performance assessment without compromising the statistical inference (e.g. the type I error rate). We explore new model evaluation strategies where multiple models are deliberately assessed simultaneously to utilize the test data to improve model selection.[1,2,3,4,5]

Methods and Results

We consider evaluation strategies $\psi = (\kappa, \varphi)$ which consist of a selection rule κ based on the hold-out- or cross-validation ranking and a multiple test φ depending on the test data.
As a generic simultaneous test procedure φ, we employ the so-called maxT-approach which relies on a multivariate normal approximation and is applicable for most performance measures.[1,2] As an alternative Bayesian procedure for the analysis of multiple correlated proportions (e.g. classification accuracies), we propose a multivariate Beta-Binomial model.[3] Both methods can be extended to the so-called co-primary endpoint analysis whereby sensitivity and specificity of multiple binary classifiers or assessed simultaneously.[4,5] In extensive simulation studies, we compared different selection rules κ on various synthetic binary classification tasks for which initial candidate models were trained via prominent learning algorithms (CART, EN, SVM, XGBoost). Besides the *default* approach (selection of single best validation model), we implemented the heuristic *within 1 SE* rule (selection of all models within 1 standard error of best validation model). Moreover, we assessed the *optimal EFP* rule derived in the framework of Bayesian decision theory.[4] Our results indicate that the test data can and should be used to improve model selection in supervised machine learning unless an unbiased estimation is deemed to be the main study goal. Moreover, statistical power can also be increased while still controlling the type I error rate of overoptimistic performance claims when a suitable adjustment for multiple comparisons is employed.

Post-Selection Inference for Prediction Performance

In order to assess the prediction performance of a statistical model it is key to obtain a reliable estimate of its ability to predict the outcome of future observations. In many cases, hypotheses testing and confidence intervals for performance measures could facilitate the predictor's credibility by overcoming an overly optimistic performance estimation. Thus, an at least asymptotically valid statistical inference on prediction performance measures post model selection is at least desirable. The aim of this research project is the development of tools and methodology for this post-selection inference for prediction performance. Besides, as the assessment of the prediction performance is usually deferred to a separate and independent validation study which is both costly and time consuming, we aim to simplify and improve the validity of the study in which the predictor has been developed.

Approach and Further Steps

We have started with assuming the situation of a validation study where, based on a collection of independently pre-fitted models, an optimal value for the hyperparameter is selected and the model is refitted with this hyperparameter now based on the validation data. This extends Westphal & Brannath (2019a)[1] where no model refitting is done. Inference is then sought for a prediction performance measure ϑ of the finally selected model. Here we want to apply weighted bootstrap methods by nonparametric tilting (cf. Efron 1981)[6] to resample the distribution of the performance estimate under the null $\vartheta = \vartheta_0$. The weights $\mathbf{w} = (w_i)$ are determined numerically from a high-dimensional and constrained optimization problem such that the Kullback-Leibler distance $D(\mathbf{w}, \mathbf{w^0}) = \sum w_i \log(n w_i)$ between the reweighted distribution and the empirical distribution, with weights $\mathbf{w}$ and $\mathbf{w}^0$, respectively, is minimized. In a first approach, we will use the boot R package[7] and apply its tilt.boot function to apply bootstrap tilting in order to evaluate the prediction accuracy rate of a binary classifier. This function estimates the empirical influence values $U_i(\mathbf{w}) = \lim_{\epsilon \to 0} \epsilon^{-1}\{\vartheta([1-\epsilon]\mathbf{w} + \epsilon \boldsymbol{\delta}_i) - \vartheta(\mathbf{w})\}$ using regression of bootstrap replicates[8] of the prediction accuracy, and yields weights $w_i = \exp(\tau U_i(\mathbf{w}^\tau)) / \sum \exp(\tau U_j(\mathbf{w}^\tau))$ that minimize $D(\mathbf{w}, \mathbf{w^0})$ among all $\mathbf{w}$ satisfying $\vartheta(\mathbf{w}) = \vartheta_0$.[6] In the second step we will then consider refitting all models before hyperparameter selection. In the final step we will aim to take the entire complex model building process into account and to control the multiple type I error rate and simultaneous coverage probabilities, at least asymptotically. At this stage we might need to use more powerful optimization methods to cope with this complex optimization problem and adapt methods from the WORHP software library (cf. Kuhlmann & Büskens, 2018)[9] in coorperation with the working group Optimization and Optimal Control of Prof. Büskens at the Center for Industrial Mathematics.

[1] Westphal, Max and Werner Brannath (2019a). Evaluation of multiple prediction models: A novel view on model selection and performance assessment. In: Statistical Methods in Medical Research. Advance online publication. DOI: 10.1177/ 0962280219854487. URL: https://doi.org/10.1177/0962280219854487.
[2] Westphal, Max and Werner Brannath (2019b). Improving Model Selection by Employing the Test Data. In: Proceedings of the 36th International Conference on Machine Learning. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. Long Beach, California, USA: PMLR, pp. 6747–6756. URL: http://proceedings. mlr.press/v97/westphal19a.html.
[3] Westphal, Max (2020). Simultaneous Inference for Multiple Proportions: A Multivariate Beta-Binomial Model. arXiv: 1911.00098 [stat.ME].
[4] Westphal, Max, Antonia Zapf, and Werner Brannath (2020). A multiple testing framework for diagnostic accuracy studies with co-primary endpoints. arXiv: 1911.02982 [stat.ME].
[5] Westphal, Max (2020). Model Selection and Evaluation in Supervised Machine Learning. (Doctoral dissertation, University of Bremen, Bremen, Germany). URL: https://doi.org/10.26092/elib/16
[6] Efron, Bradley (1981). Nonparametric standard errors and confidence intervals. In: Canadian Journal of Statistics (Vol. 9, No. 2, pp. 139-158). DOI:10.2307/3314608.
[7] Canty, Angelo and Brian Ripley (2015). R package 'boot'. https://cran.r-project.org/web/packages/boot/.
[8] Davison, Anthony C. and David V. Hinkley (1997) Bootstrap methods and their application. Cambridge University Press. DOI:10.1017/CBO9780511802843.
[9] Kuhlmann, Renke & Christof Büskens (2018). A primal-dual augmented Lagrangian penalty-interior-point filter line search algorithm. In: Mathematical Methods of Operations Research (Vol. 87, pp. 451-483). DOI:10.1007/s00186-017-0625-x.

DFG

Universität Bremen