Generalized Resubstitution for Error Estimation in Regression

Diego Marcondes[1] and Ulisses Braga-Neto[2]

[1]*Department of Computer Science, Institute of Mathematics and Statistics, University of São Paulo, Brazil*

[2]*Department of Electrical and Computer Engineering, Texas A&M University, USA*

A recent work [2] proposed the family of generalized resubstitution classification error estimators as the expected error under arbitrary empirical probability measures, and studied special cases such as bolstered, posterior-probability, Gaussian-process, Bayesian and bolstered posterior-probability error estimators. They showed that the generalized resubstitution error estimator is consistent and asymptotically unbiased for the two-class problem if the corresponding empirical probability measure converges uniformly to the standard empirical probability measure and the hypotheses space has a finite VC dimension. In this work, we extend the generalized resubstitution error estimators to the general statistical learning framework when the loss function has a moment of order $\alpha > 1$ uniformly bounded in the hypotheses space. In particular, we consider regression problems under the quadratic loss function and extend the special cases studied in [2] to them.

The generalized error estimators can be defined as the expectation of the loss function under an arbitrary empirical probability measure or as the expectation of a generalized loss function under the standard empirical probability measure. These representations allow us to establish many sufficient conditions for the consistency of these estimators that not only extend that of [2], but are also weaker than it. For instance, we show that if the generalized loss function converges uniformly to the original one, or if the expectation of the generalized loss function under the data-generating distribution converges to that of the original loss function, then the generalized resubstitution error estimator is consistent. In particular, this last condition allows for the variance of the empirical measure to not converge to zero. Sufficient conditions for consistency based on the variance of the empirical measure are also established for the case of twice-differentiable loss functions.

Finally, since consistency is attained under many conditions, there is a lot of room to choose the empirical measure, and we propose methods to estimate the parameters of the empirical measure from the data. We focus on bolstered error estimators and propose a method of moments and a maximum pseudo-likelihood estimator for the covariance matrix of the bolstered empirical measure. Our proposal justifies the heuristic approach proposed by [1].

References

[1] Ulisses Braga-Neto and Edward Dougherty. Bolstered error estimation. *Pattern Recognition*, 37(6):1267–1281, 2004.

[2] Parisa Ghane and Ulisses Braga-Neto. Generalized resubstitution for classification error estimation. *Journal of Machine Learning Research*, 23(280):1–30, 2022.