Adaptive Resampling in a Parallel World

Max Kuhn[1⋆]

1. Pfizer Global R&D, Groton CT
⋆Contact author: max.kuhn@pfizer.com

Keywords: Machine Learning, Classification, Regression, Parameter Tuning

Many predictive models require parameter tuning. For example, a classification tree requires the user to specify the depth of the tree. This type of "meta parameter" or "tuning parameter" cannot be estimated directly from the training data. Resampling (e.g. cross-validation or the bootstrap) is a common method for finding reasonable values of these parameters (Kuhn and Johnson, 2013). Suppose B resamples are used with M candidate values of the tuning parameters. This can quickly increase the computational complexity of the task.

Some of the M models could be disregarded early in the resampling process due to poor performance. Maron and Moore (1997) and Shen *el at* (2011) describe methods to adaptively filter which models are evaluated during resampling and reducing the total number of model fits. However, model parameter tuning is an "embarrassingly parallel" task; model fits can be calculated across multiple cores or machines to reduce the total training time. With the availability of parallel processing is it still advantageous to adaptively resample?

This talk will briefly describe adaptive resampling methods and characterize their effectiveness using parallel processing via simulations.

References

Kuhn, M. and Johnson, K. (2013). *Applied Predictive Modeling*. Springer Verlag.

Maron, O. and Moore, A. (1997). The Racing Algorithm: Model selection for lazy learners. In D. Aha (Ed.), *Lazy Learning*, 193225. Springer.

Shen, H., Welch, W. J. and Hughes–Oliver, J. M. (2011). Efficient, adaptive cross–validation for tuning and comparing models, with application to drug discovery. *The Annals of Applied Statistics*, 5(4), 26682687.