Technical description of the permutation test comparing two learning algorithms by their cross-validated predictions

Jussi Tohka

October 2020

We will present the permutation test for testing the hypothesis that mean absolute error (MAE) by a prediction method 1 is better than the MAE by a prediction method 2. It might be that one of the methods (say, method 2) is a simple baseline method and then our aim is to demonstrate that method 1 indeed results in useful predictions.

This test was originally developed for [LET+18] and subsequently used in [PAZ+20]. However, only related bootstrap confidence interval computation was formally described in [LET+18] and the permutation test displaying a strong resemblance to the bootstrap estimation was not formally described. This document aims to correct for this omission. Finally, we note that only the use of repeated cross-validation separates this test from the standard permutation test.

For each CV run and each prediction method, we get a predicted value for each subject. We will denote these predicted values by $\hat{y}_{i,r}(m)$ where i indexes the subject, r CV run, and m method, $m = 1, 2$. The true value for the subject i is y_i and N is the number of subjects. We assume that y_i are independent.

We compute the MAE as

$$MAE(m) = \frac{1}{NR} \sum_{i=1}^{N} \sum_{r=1}^{R} |y_i - \hat{y}_{i,r}(m)|.$$

We will define $z_{i,r}(m) = \frac{|y_i - \hat{y}_{i,r}(m)|}{NR}$ and write

$$MAE(m) = \sum_{i=1}^{N} \sum_{r=1}^{R} z_{i,r}(m).$$

We want to test the hypothesis that a population level MAEs are equal between the two methods. For this, we take as the null hypothesis that the errors $z_{i,r}(m)$ are drawn from the same distribution for both methods. The obvious test statistic is

$$T = MAE(1) - MAE(2).$$

We can implement this test as a permutation test by writing

$$T = \sum_{i=1}^{N}(\sum_{r=1}^{R} z_{i,r}(1) - \sum_{r=1}^{R} z_{i,r}(2)) = \sum_{i=1}^{N}(u_i(1) - u_i(2)).$$

The method labels $1, 2$ are now exchangeable under the null hypothesis in u_i and therefore we can construct a permutation test by randomly swapping the labels of u_i [Goo05].

References

[Goo05] Phillip Good. Testing hypotheses. *Permutation, Parametric and Bootstrap Tests of Hypotheses*, pages 33–65, 2005.

[LET+18] John D Lewis, Alan C Evans, Jussi Tohka, Brain Development Co-operative Group, et al. T1 white/gray contrast as a predictor of chronological age, and an index of cognitive performance. *Neuroimage*, 173:341–350, 2018.

[PAZ+20] Mithilesh Prakash, Mahmoud Abdelaziz, Linda Zhang, Bryan A Strange, and Jussi Tohka. Quantitative longitudinal predictions of alzheimer's disease by multi-modal predictive learning. *bioRxiv*, 2020.