Article:
King, RD, Orhobor, OI and Taylor, CC (2021) Cross-validation is safe to use. Nature Machine Intelligence, 3 (4). 276. p. 276.

https://doi.org/10.1038/s42256-021-00332-z

Cross-Validation is Safe to Use

Ross D. King[1,2,3], Oghenejokpeme I. Orhobor[1], Charles, C. Taylor[4]

1. Department of Chemical Engineering and Biotechnology, University of Cambridge, UK.

2. Alan Turing Institute, London, UK.

3. Department of Biology and Biological Engineering, Chalmers University of Technology, Sweden.

4. Department of Statistics, University of Leeds, UK.

The importance of machine learning (ML) to science is now widely recognized. For example, *Nature* in its last editorial of the last decade named ML as the 'breakthrough' of the decade: '*few fields are untouched by the machine-learning revolution, from material science to drug exploration; quantum physics to medicine'.* Despite the importance of ML, some basic ML ideas are still poorly understood in the general science community.

One such technique is cross-validation (Stone, 1974; Efron & Gong, 1983). Within ML cross-validation is so commonly used that it is rare to find a paper that doesn't use it. However, non ML scientists commonly misunderstand cross-validation, and avoid its use, thinking it unsafe. Two typical misunderstanding/concerns are quoted below:

'*Papers using machine learning must contain a dedicated subsection clearly describing the composition of the training dataset.* ***This should include information on the preparation of the cross-validation sets and of an independent test set that is not used in the training process. For data originating from biological sequences the description must furthermore address how homology between sequences is taken into account to ensure that the training and independent test sets do not have identical or near identical examples.*** *Papers using leave-one-out will be editorially rejected unless there is a special circumstance in which it can be argued that this procedure is meaningful for the problem addressed in the paper. Machine learning papers must report the performance on an independent test set. It is not sufficient to report the average error over the individual cross-validation sets*.' (bold from original). (Bioinformatics journal - Scope Guidelines, 2021)

'N fold cross validation is not a very tough test –or even the way the models are used-as QSAR (Quantitative Structure Activity Relation) models are most valuable if they can predict future compounds/activities – So with cross validation I am concerned there is leakage and independent reviewers may feel the same way, unless you can show them this is not a concern. Independent test sets are a more robust way of assessing the model – selected by date order – which is assessing the model ability to predict the future.' Drug design scientist 2020 (slightly edited for sense)

Both quoted statements seem to imply that cross-validation is unsafe and should be replaced by the alternative technique of train/test

This is very puzzling to ML researchers and statisticians, as cross-validation and train/test both share the exact same set of assumptions. So it is unreasonable to permit one technique and not the other. The main use of cross-validation and train/test are the same: to predict how well a predictive ML model will perform on new data from the same distribution.

The idea of train/test is to use one sample of data (the training data) to learn a predictive ML model. Then to use a second sample of data (the test data - whose true classifications are known but are not told to the predictor) to estimate the error rate of the predictive ML model. Note that there is a loss of efficiency here, as we do not use the full sample to train the ML model.

The idea of cross-validation is to divide the data into subsamples. Each sub-sample is predicted using the ML model learnt from the remaining subsamples, and the estimated error rate is the average error rate from these subsamples (Stone, 1974; Efron & Gong, 1983); Michie *et al.*, 1994). The ML model finally used is calculated from all the data. Cross-validation gives a better estimate than train/test at the cost of more computation. The leave-one-out method of Lachenbruch & Mickey (1968) is cross-validation with samples equal to the number of examples.

Given that cross-validation and train/test do the same job, and make the same assumptions, what could possibly be the reason for concerns about its use? The clue seems to be that both quotes refer to structure in the data: in the Bioinformatics journal case, that structure is the possible homologous relationship between examples; and in the drug design example, the temporal relationship between examples. Recall that cross-validation and train/test are used to predict how well a predictive ML model will perform on *new data from the same distribution*. This means that if cross-validation or train/test samples are selected with different distributions from future data, then the prediction of performance will be inaccurate. However, *this problem is exactly the same for cross-validation and train/test*. It is therefore irrational to trust cross-validation less than train/test.

In conclusion. ML is now a key technology in modern science. However, its techniques need to be better understood. We therefore call for a dialogue between ML and domain scientists in which ML methods, such as cross-validation, can be explained to domain scientists so that they can trust and benefit from them.

References

Scope Guidelines. [online] Available at: <https://academic.oup.com/bioinformatics/pages/scope_guidelines> [Accessed 21 January 2021].

Efron, B., & Gong, G. (1983) A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation. *The American Statistician*, **37**, 36-48.-

Lachenbruch, P. A. and Mickey, M. R. (1975). *Discriminant Analysis*. Hafner Press, New York.

Michie, D., Spiegelhalter, D.J., Taylor, C.C. (eds) (1994) *Machine Learning, Neural and Statistical Classification*. Ellis Horwood Ltd.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc.*, **36**, 111–147 (including discussion).