Resampling techniques: cross validation

Resampling is a fundamental part of statistical analysis, since it allows to overcome the lack of data usually faced. One of the main methods used for resampling is cross-validation, which also solves problems such as overfitting or selection bias.

The process starts by splitting the data in two sets: the training set, on which it is possible to build the model, and the test set, on which predictions are made, also evaluating the prediction performance of the model previously built. In theory, it is possible to divide the data in three parts: training, test and validation, where this last one is used to estimate the prediction errors.

In particular, k-fold cross-validation is a cross-validation technique which allows to solve the problem of unbalanced data splitting . The data is divided in k subsets, approximately equally sized, which need to be exhaustive and mutually exclusive. At each split, one of the subsets is considered as test set and the remaining (k-1) as training sets. The Mean Squared Error (MSE) is computed each time on a different test set. At the end of the k loop, the k-fold CV is estimated by computing the mean of these values

