Validation of Models

Validation is a check of the operation correctness (predictive ability) of the analytical model trained on the machine learning basis. It also secures compliance of the analytical model with the requirements of the solved task.

It is performed on the basis of the independent validation set (namely, the set that has not been used for training and testing) after the model training and testing.

Validation Options

Without validation

Validation is not performed.

K-folds cross validation method

The method used for selection of training and test sets for the analytical model training under conditions of insufficient source data or nonuniform display of classes.

To provide the successful training of the analytical model, it is required to display classes in the training set approximately in the equal proportion. However, if there is no sufficient data or the "sampling" procedure was unsuccessfully performed when selecting the training set, one of the classes can become dominating. It can be considered as imbalance in the training process, and the dominating class will be considered as the most probable. The cross validation method enables to avoid it.

It is based on division of the source data set to k folds, for example, $k = 5$. Then to $k - 1$, namely, the model is trained on 4 blocks, and the 5th block is used for testing. The procedure is repeated k times. In this case, the new block is selected for check during each check, wheres the training is performed with the remaining ones.

The cross validation has two main advantages as compared with usage of one set for training and one set for testing of the model:

- distribution of classes is more uniform that improves the training quality;
- if the output model error is estimated for each check, and its average value is defined for all checks, its received estimate will be more reliable.

Practically $k = 10$ is more frequently selected (10 — 10-fold cross validation) when the model is trained using 9/10 of data and it is tested using 1/10. Research has shown that in this case the most reliable estimate of the output model error is provided.

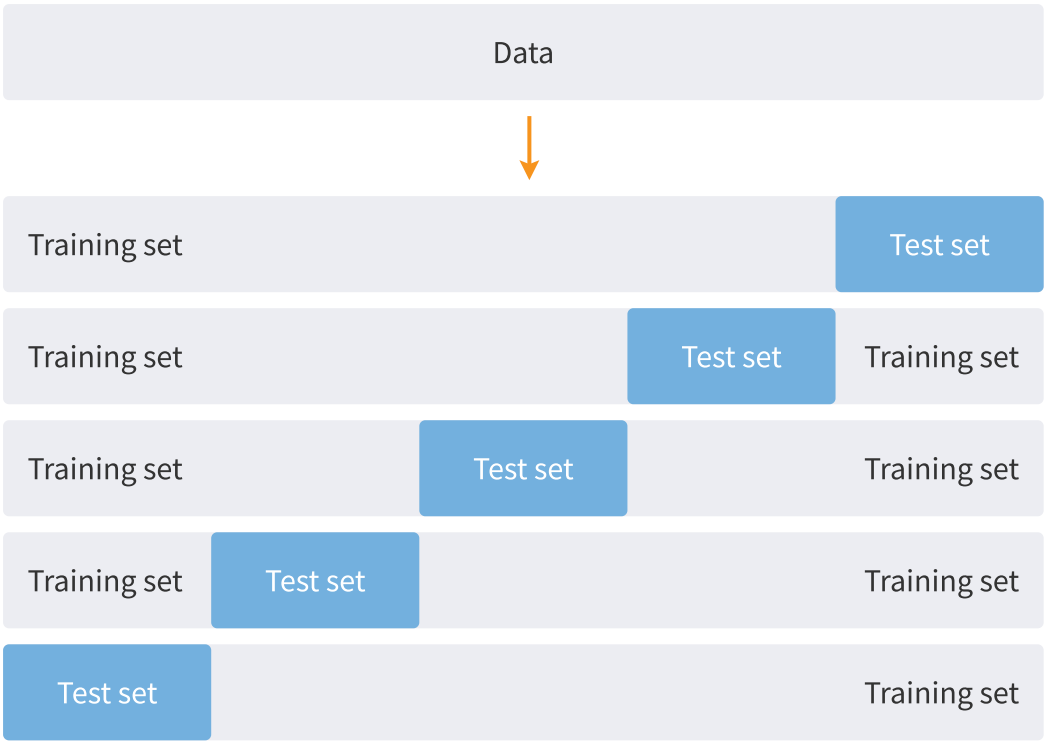


Figure 1. K-folds cross validation method

Monte Carlo method

According to the set proportion the data is randomly divided into the training and validation sets. The model is trainied on the training set basis, and it is checked using the validation set. The procedure is repeated N times where N is a number of iterations.