Cross-Validation

@wikipedia

Synonym: Cross-Validation = Blind Testing

A specific technique for estimating how accurately the given model is capable to predict the Source Dataset.

It assumes that Source Dataset is split into two subsets: Training dataset and Validation dataset.

The Training dataset is used to calibrate the model parameters.

The discrepancy between model values and Training dataset values can be low but it does not mean that predicting accuracy of the model on the data outside the Training dataset will be the same low.

This may happen because the model is not unique and a given model realization may not be the best predictor.

In order to assess predictability of the model it should bee validated on the data outside the Training dataset, which is called Validation dataset.

If model discrepancy on Validation dataset is close to model discrepancy on Training dataset one can say that a given model has a good predictability within the Source Dataset range.

If model discrepancy on Validation dataset is not close to model discrepancy on Training dataset then this phenomenon is called overtraining and means that a given model realization has "remembered" the Training dataset bu t can not accurately predict on the data points outside the Training dataset.

Splitting the Source Dataset into Training dataset and Validation dataset can be done in different ways.

It can be done manually or randomly (see Bootstrapping).

It should be noted though that Source Dataset may not hold enough of representative events/occurrences to provide the opportunity for Cross-Validation and in this case the Goodness of fit over the Training dataset (which is the whole S ource Dataset in this case) will be the only one available, thus increasing the risk of future Model Prediction.

See also

Natural Science / System / Model / Model Validation

Formal science / Mathematics / Statistics

[Source Dataset] [Training dataset] [Validation dataset]