Statistical parameters

Here, we will briefly overview the statistical parameters used by OCHEM for evaluation of the predictive performance of QSAR models.

Regression models

In the formulae below, $\tilde{y}_i$ and y_i denote predicted and real values of the predicted property for i-th compound in the set. $E(\tilde{y})$ and $E(y)$ are the means of thepredicted and real property values; sigma (?) denotes the standard deviation.

RMSE

RMSE stands for Root Mean Squared Error and is calculated according to the formula:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (\tilde{y}_i - y_i)^2}{N}}$$

The R2 and Q2 indicators can be a subject to confusion, since they often mean different things in classical statistics and QSAR, depending on whether they are calculated for the training set or for the cross-validated set. All the formulae given here are calculated with respect to the selected validation protocol (bagging, cross-validation or no validation)

R2 (Pearson correlation coefficient)

$$r^2 = \frac{\sum_{i=1}^{N} (\tilde{y}_i - E(\tilde{y}))(y_i - E(y))}{\sigma(\tilde{y}) \cdot \sigma(y)}$$

Q2 (Coefficient of determination)

$$q^2 = 1 - \frac{RMSE}{\sigma(y)} = 1 - \frac{\sum_{i=1}^{N} (y_i - \tilde{y}_i)^2}{\sum_{i=1}^{N} (y_i - E(y))^2}$$

MAE

$$MAE = \frac{\sum_{i=1}^{N} |\tilde{y}_i - y_i|}{N}$$

Classification models

The most common case of classification models is binary classification, where the instances belong to either positive (active) or negative (inactive) class. Some statistical measures are applicable to binary classification models only.

For binary classification models the accepted notion is to discriminate:

TP = *true positives* - number of instances of active class, that were correctly predicted by the model as actives

FP = *false positives* - number of instances of inactive class, that were incorrectly predicted by the model as actives

TN = *true negatives* - number of instances of inactive class, that were correctly predicted by the model as inactives

FN = *false negatives* - number of instances of active class, that were incorrectly predicted by the model as inactives

Accuracy

"Accuracy" is merely the percentage of correctly classified samples. For binary classification accuracy can be calculated as follows:

ACC = (TP + TN) / (TP + FP + TN + FN)

Class hit rate

Hit rate is a measure that is applicable to a single class in a classification model and denotes a ratio of instances of a specific class that were correctly identified as belonging to this class.

For binary classification tasks class hit rate for positive class is called *sensitivity,* and for negative class - *specificity.*

Precision

Precision in the context of classification models is a measure applicable to a single class of a model and denotes a ratio between the instances correctly identified as belonging to a particular class and a total number of instances identified as belonging to this class.

For binary classification models precision for positive class is called *positive predictive value*, and for negative class - *negative predictive value.*

Sensitivity

Sensitivity is a measure applicable to binary classifications and denotes a ratio of positive instances that were correctly identified as such.

SENS = TP / (TP + FN)

Specificity

Specificity is a measure applicable to binary classifications and denotes a ratio of negative instances that were correctly identified as such.

SPEC = TN / (TN + FP)

Positive predictive value

Positive predictive value is a binary classification measure and shows a ratio of true positive to all instances that were classified as positive.

PPV = TP / (TP + FP)

Negative predictive value

Negative predictive value is a binary classification measure and shows a ratio of true positive to all instances that were classified as positive.

NPV = TN / (TN + FN)

Balanced accuracy

Balanced accuracy is the averaged accuracy for each class, e.g. (positive_class_accuracy + negative_class_accuracy) / 2.
This parameter is important for imbalanced datasets, which have significantly different number of samples in different classes.
If a classifier has a similar performance for both negative and positive classes, accuracy and balanced accuracy are also similar.

*BA = 0.5 * (TP / (TP + FN) + TN / (TN + FP)) = 0.5 * (SENS + SPEC)*

Matthews correlation coefficient (MCC)

MCC takes into account true and false positives and negatives and is generally regarded as a balanced measure which can be used even if the classes are of very different sizes.

MCC = (TP*TN - FP FN)/ SQRT((TP+FP)(TP+FN)(TN+FP)(TN+FN))

Area under the curve (AUC)

Receiver Operating Characteristic AUC (ROC-AUC) is calculated on the plot of Sensitivity vs Specificity, which is shown for each classification model

Confusion matrix

Confusion matrix shows the number of samples from a particular class classified as another particular class.

Prediction intervals

The prediction intervals (68%, i.e. approximately ± one standard deviation) for all statistical parameters are evaluated using bootstrap procedure with n = 1000 samples.
Following model calculation we get predicted values z_i for training (or test) samples with experimental values y_i, i.e. $\{z_i, y_i\}$ are pairs of values with $i = 1,..,N$.

We randomly sample pairs $\{z_i, y_i\}$ to form n = 1000 bootstrap sets of the same size as the analyzed set. For each bootstrap set we calculate
statistical parameters and use their respective distributions to determine respective confidence intervals.
Since the intervals are in general non-symmetric with respect to the values calculated for the analyzed sets, the average values are reported.