Performance metrics

Evaluation of regression models

`R²`

> This implementation of R² is specifically for out-of-sample evaluation and it measures the proportion of variance in the dependent variable that is predictable from the independent variable(s). This R² is used as it penalizes systematic bias as well as poor correlation by focusing on the error and not the correlation. It is calculated as one minus the ratio of the MSE of the model predictions over the MSE of the trivial model predictions (constant number). A value of 1 indicates perfect predictions, while values around 0 indicate that the regression results are no different than the trivial approach of constantly predicting the average value of the outcome. Negative values indicate that the model is worse than the trivial approach.

`Classic R²`

> This implementation of R² measures the proportion of the outcome variance explained by the predictions. A value of 1 indicates perfect predictions while values around 0 indicate that the regression has no predicting power beyond just predicting the outcome mean. Negative values indicate that the model delivers predictions that are negatively correlated with the outcome.

`Mean absolute error`

> The average difference between the actual outcome and the predicted outcome. Higher values indicate poorer prediction quality, while a value of 0 indicates perfect predictions.

`Mean squared error`

> The average of the squared difference between the actual outcome and the predicted outcome. Higher values indicate poorer prediction quality, while a value of 0 indicates perfect predictions.

`Mean squared logarithmic error`

> The average of the squared difference between the logarithms of theactual outcome and the predicted outcome. Higher values indicate poorer prediction quality, while a value of 0 indicates perfect predictions.

`Pearson correlation coefficient`

The correlation between actual outcome and the predicted outcome, computed according to Pearson's formula. Use this coefficient if a strong relation between the target parameter and molecular structure is expected. A value of 1 indicates perfect concordance between predictions and actual values, -1 means perfect discordance, while 0 indicates no relation whatsoever.

`Relative absolute error`

The average percentage difference between the actual outcome and the predicted values. Higher values indicate poorer prediction quality, while a value of 0 indicates perfect predictions. For example, a value of 0.1 indicates that the predictions are on average 10% off with respect to the actual outcome.

`Relative squared error`

The average of the squared difference between the actual outcome and the predicted outcome, reported as a percentage. Higher values indicate poorer prediction quality, while a value of 0 indicates perfect predictions.

`Spearman correlation coefficient`

The correlation between actual outcome and the predicted outcome, computed according to Spearman's formula. Use this coefficient if there is only a vague relation between the target parameter and molecular structure. A value of 1 indicates perfect concordance between predictions and actual values, -1 means perfect discordance, while 0 indicates no relation whatsoever.

Evaluation of categorical models

`Area under the curve`

Also known as area-under-the-ROC (receiver operating characteristic) curve. Use this measure to evaluate balanced datasets with near equal data points in each bin (use the balance and entropy scores found under the "Target Parameters" tab in ChemX as guides if needed). A measure of how well a model can distinguish between two groups (e.g. cases/controls). An area of 1 represents a perfect model, an area of 0.5 represents random guessing.

`Average precision`

Also known as the area-under-the-precision-recall curve. Use this measure to evaluate imbalanced datasets (use the balance and entropy scores found under the "Target Parameters" tab in ChemX as guides if needed). The precision-Recall curve shows the tradeoff between precision and recall for different thresholds. A high area under the curve represents both high recall and high precision, where high precision relates to a low false positive rate, and high recall relates to a low false negative rate.

`Accuracy`

The fraction of predictions that a classification model got right. In multi-class classification, accuracy is defined as: correct predictions/total number of observations.

`Average F1 Score`

The average F1 score across all classes.

`Average MCC`

Average Matthews correlation criterion across all classes.

`Balanced accuracy`

The average of Sensitivity and Specificity, obtained by thresholding the prediction values at zero. It is defined as Balanced Accuracy = (Sensitivity + Specificity) / 2.

`F1 Score`

Harmonic mean of precision and recall. It is defined as F1 score = 2 · (Recall · Precision) / (Recall + Precision).

`False negative ratio`

The number of false negatives divided by the total number of predictions.

`False Positive ratio`

The number of false positives divided by the total number of predictions.

`Matthews correlation criterion (phi coefficient)`

The coefficient 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. The MCC is in essence a correlation coefficient between the observed and predicted binary classifications; it returns a value between -1 and +1. A coefficient of +1 represents a perfect prediction, 0 no better than random prediction and -1 indicates total disagreement between prediction and observation.

`Precision`

A measure how many positives were predicted correctly out of the total number of positive predictions. It is defined as Precision = TP / (TP + FP).

`Specificity`

Also known as the True Negative Rate. A measure of how many of the negative predictions were correct. It is defined as Specificity = TN / (TN + FP).

`True negative ratio`

The number of true negatives divided by the total number of predictions.

`True positive rate`

Also known as Recall or Sensitivity. A measure of how many of the positive predictions were correct. It is defined as True Positive Rate = TP / (TP + FN).

`True positive ratio`

The number of true positives divided by the total number of predictions.