

Assignment #5

Course: *Machine learning*
Due date: *December 15th, 2023*

Assignment

In this assignment, you will learn about gradient boosting.

Write the code for gradient boosting of trees for solving a binary classification problem from scratch.

You can use the tree implementation you programmed in Assignment 4 or the Scikit-learn implementation. The trees have to be weak learners, so their depth has to be set to a small value. You can choose their final depth arbitrarily. Think about overfitting.

Download the dataset on Učilnica "wine-quality.csv".

This is the same dataset as used in Assignment 3, but now we are solving a binary classification problem (quality is the target and is either high or low). Be careful about continuous and categorical variables.

Test different learning rates. What is a good learning rate?

Test different numbers of trees that are built during gradient boosting and comment on the results. Does your model overfit? If yes, try to prevent overfitting.

You have to explain your solution.

Compare the cross-validation results from your implementation with the "Gradient-BoostingClassifier" classifier implemented in Scikit-learn.

Be careful to use the same cross-validation splits in both implementations.

Try modelling your data also with XGBoost, LightGBM and CatBoost models (fit all three models). Compare all the models you built. Be careful to evaluate all three models on the same test dataset.

Be careful to use the same test data for all three models. If you make hyperparameter tuning be careful to use a validation set.