Heinrich Heine Universität Düsseldorf

HEIDERLAB

Exercise 9 on Machine Learning WS 2023/24
Prof. Dr. Dominik Heider
M. Sc. Arsam (Mohammad) Tajabadi
Submission: Until 10.01.2024, 23:59 on Ilias

Task 1. Random Forest (7 Points)

Create a random forest in Python (you can use any packages) for dataset *bank.csv*. (This comes from the publication S. Moro, P. Cortez and P.Rita. *A Data-Driven Approach to Predict the Success of Bank Telemarketing*. Decision Support Systems (2014), doi:10.1016/j.dss.2014.03.001.). Split the data into train and test sets with ratios of 0.8 and 0.2. Perform any preprocessing steps needed to prepare the data for training.

a) Determine the out-of-bag-error and the error for the test dataset. (1.5 P.)

b) What does the out-of-bag-error describe? (0.5 P.)

c) Determine the mean decrease in Gini for all features and plot it. (1.5 P.)

d) Use the seven best attributes and create a new random forest. (1.5 P.)

e) Plot an ROC curve for each of the two classifiers and determine the AUC value. (1.5 P.)

f) Which classifier provides better predictions and why? (0.5 P.)

Task 2. Performance Prediction (3 Points)

a) Would it have been better to calculate precision and recall for task 1? Briefly explain your answer. (0.5 P.)

b) Calculate Precision and Recall (in Python). Create a Precision-Recall Curve for the two classifiers. (You are allowed to use any packages.) Compare your results with the ROC Curve. (1.5 P.)

c) Determine the Accuracy, F1-Score, and Matthews Correlation Coefficient on the complete test set. Interpret the difference. (1 P.)