Assignment #2

Course: *Machine learning*
Date: *October 20h, 2023*

Assignment

In this assignment, you will learn about resampling methods for model evaluation and attribute selection. You have to predict the crime rate based on a large number of different attributes.
You can use the SciKit-learn library to fit the linear regressions. The rest of the code has to be programmed from scratch.

Be careful not to use the same data for training (any stage of training) and testing your model.

Download the "Communities and Crime" dataset, preprocess the data, and preprocess them so you can use them for linear regression.

The last column of the dataset (ViolentCrimesPerPop) is your target variable. Remove the attributes state, county, community, community name, and fold (columns 1 to 5).
`https://archive.ics.uci.edu/dataset/183/communities+and+crime`

Implement the cross-validation method and the leave-one-out method.

Inside use your own implementation of cross-validation.

Implement forward attribute selection. Fit linear regression.

Use the attribute selection you implemented to select a reasonable set of attributes for your linear model.

You have to decide which metric is used for the inclusion of attributes (the metric that decides which attribute is most important at each step) and the criteria to stop adding attributes.

Test your model using your independent test set and report the results.

Implement the bootstrap method and apply it to the train set to generate 1000 different train sets and train 1000 different linear models.

Use only the attributes you selected before.

Use the bootstrapped results to assess the confidence intervals of the performance of the linear model on your independent train set.

You can use different metrics to assess the performance of the linear model (e.g. MAE, RMS). Calculate the confidence intervals for the metric you selected.