ML-SPARD: a dataset for Machine Learning Performance Analysis in Small-samples Regression Problem

Rosalvo Ferreira de Oliveira Neto[1]

Abstract
A set of databases has been created to cater to the academic community's research needs in the realm of Machine Learning algorithm performance, particularly in scenarios with limited sample sizes and regression problems. These databases encompass varying sample sizes, data dimensionality, and the linearity of the response variable.

Keywords
Dataset — Small dataset — Regression datasets

[1] *Engenharia da Computação, Universidade Federal do Vale do São Francisco, Juazeiro-BA, Brasil*
*__Rosalvo F. de O. Neto__:rosalvo.oliveira@univasf.edu.br

Contents

1 Introduction 1
2 Description 1
3 Files 2
4 Setup Experimental 2
References 2

1. Introduction

Researchers can access a wide range of datasets to investigate the robustness of algorithms when facing data with different characteristics, be it small or large sample sizes, low or high-dimensional data, and linear or nonlinear response variables [1] e [2]. This versatility enables the academic community to derive insights into the strengths and limitations of various algorithms, aiding in the development of more accurate and efficient models for regression tasks. Whether it's fine-tuning existing algorithms or innovating new ones, this dataset collection serves as a valuable asset for advancing the field of Machine Learning and enhancing its applicability in real-world scenarios.

The primary objective of this collection is to empower researchers with diverse datasets to explore the behavior and effectiveness of Machine Learning algorithms when dealing with small-scale data and regression tasks. The databases offer a valuable resource for testing, benchmarking, and refining algorithms that are crucial for predictive modeling in various fields, such as economics, biology, and engineering.

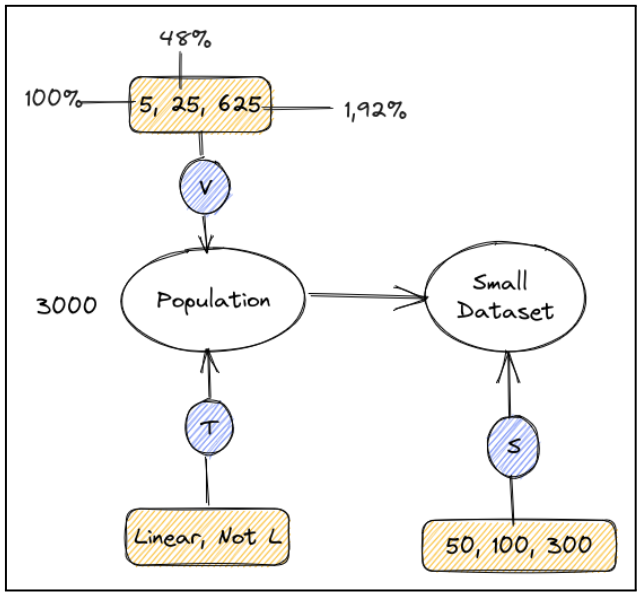

Figure 1. Datasets generation process

2. Description

The datasets were generated taking into account three factors, which are: 1) whether the target variable (T) has a linear relationship with the input variables or not. 2) the number of input variables (V), which can be 5, with all of them being relevant (100%), 25, with only 5 (48%) being relevant, or 625, with only 5 (1.92%) being relevant. 3) the sample size (S), which can be 50, 100, or 300, all generated from the same population with 3000 examples. Figure 1 illustrates the data generation process.

3. Files

Three zip files were provided: 1) F5 containing files with five columns. 2) F25 containing files with five columns. 3) F625 containing files with six hundred and twenty-five columns. Each file was generated three times by changing the random number generator seed, so within the F5, F25, and F625 folders, there are three subdirectories: 001, 002, and 003. Tables 1, 2, and 3 display the names of the files generated in each folder.

Table 1. Five variables files (F5.zip)

File Name	Description
F5_P_YL.csv	Population with 3000 elements with 5 input variables and linear target.
F5_P_YNL.csv	Population with 3000 elements with 5 input variables and non-linear target.
F5_S50_YL.csv	Sample with 50 elements with 5 input variables and linear target.
F5_S50_YNL.csv	Sample with 50 elements with 5 input variables and non-linear target.
F5_S100_YL.csv	Sample with 100 elements with 5 input variables and linear target.
F5_S100_YNL.csv	Sample with 100 elements with 5 input variables and non-linear target.
F5_S300_YL.csv	Sample with 300 elements with 5 input variables and linear target.
F25_S300_YNL.csv	Sample with 300 elements with 5 input variables and non-linear target.

Table 2. Twenty five variables files (F25.zip)

File Name	Description
F25_P_YL.csv	Population with 3000 elements with 25 input variables and linear target.
F25_P_YNL.csv	Population with 3000 elements with 25 input variables and non-linear target.
F25_S50_YL.csv	Sample with 50 elements with 25 input variables and linear target.
F25_S50_YNL.csv	Sample with 50 elements with 25 input variables and non-linear target.
F25_S100_YL.csv	Sample with 100 elements with 25 input variables and linear target.
F25_S100_YNL.csv	Sample with 100 elements with 25 input variables and non-linear target.
F25_S300_YL.csv	Sample with 300 elements with 25 input variables and linear target.
F25_S300_YNL.csv	Sample with 300 elements with 25 input variables and non-linear target.

Table 3. Six hundred and twenty-five variables files (F625.zip)

File Name	Description
F625_P_YL.csv	Population with 3000 elements with 625 input variables and linear target.
F625_P_YNL.csv	Population with 3000 elements with 625 input variables and non-linear target.
F625_S50_YL.csv	Sample with 50 elements with 625 input variables and linear target.
F625_S50_YNL.csv	Sample with 50 elements with 625 input variables and non-linear target.
F625_S100_YL.csv	Sample with 100 elements with 625 input variables and linear target.
F625_S100_YNL.csv	Sample with 100 elements with 625 input variables and non-linear target.
F625_S300_YL.csv	Sample with 300 elements with 625 input variables and linear target.
F625_S300_YNL.csv	Sample with 300 elements with 625 input variables and non-linear target.

4. Setup Experimental

The dataset was generated using the **make_regression** function from the **sklearn.datasets** library. The non-linear target was generated by applying the abs function to the linear target variable. The input variables were normalized using the **MinMaxScaler** class from the **sklearn.preprocessing** library.

References

[1] Jing Yao, Danfeng Hong, Lianru Gao, and Jocelyn Chanussot. Multimodal remote sensing benchmark datasets for land cover classification. In *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*, pages 4807–4810, 2022.

[2] Adeel Yousaf, M. Jaleed Khan, M. Imran, and Khurram Khurshid. Benchmark dataset for offline handwritten character recognition. In *2017 13th International Conference on Emerging Technologies (ICET)*, pages 1–5, 2017.