Faculty of Mechanical Engineering and Robotics

Department of Robotics and Mechatronics

Python for Machine Learning and Data Science

Course for Mechatronic Engineering

Instruction 5:

Project datasets: Data preparation & regression/classification

You will learn: How to prepare datasets for machine learning applications and how to use machine learning models.

Additional materials:

- Course lectures 3 & 4 [*obligatory*] http://galaxy.agh.edu.pl/~zdw/Materials/Python/LectureNotes/
- Report template http://galaxy.agh.edu.pl/~zdw/Materials/Python/

Learning outcomes supported by this instruction:
IMA1A_U01, IMA1A_U05, IMA1A_U06, IMA1A_U07, IMA1A_K08

Course supervisor:
Ziemowit Dworakowski, zdw@agh.edu.pl

Instruction author:
Adam Machynia, machynia@agh.edu.pl

Introduction

During this lab, you will prepare your dataset for applying a machine learning model and then proceed with the initial application of a classification or regression model. You will discuss specific tasks with the teacher, but in general, all teams should complete the following tasks.

Remember to split the data into training, validation, and test sets.

Task 1: Data preparation. You should clean and organize the dataset. Consider the following points:

- If you have categorical variables, consider using *OrdinalEncoder* or *OneHotEncoder* from *sklearn.preprocessing*. Always carefully evaluate which approach is most appropriate for your specific case.
- If some values in the dataset are missing, consider using the *dropna()* function from Pandas.
- Evaluate which features are important and which might introduce ambiguity.
- Consider creating "new" features as combinations of existing ones. For example, you could use the ratio of volatile acidity or citric acid to fixed acidity as a new feature. In other cases, this might be GDP per capita or the price per square meter.
- You may need to handle outliers.
- Consider applying feature scaling.

The data preparation task will highly depend on your dataset, but it may be crucial to your project's success. Proper data preparation can simplify the actual task, allowing for the use of a simpler model or making its configuration more straightforward.

Task 2: Perform the classification or regression task outlined in your project scope using the chosen model. Draw conclusions.

Task 3: Repeat the previous task using different models.

- For regression: try a linear regressor and a neural network.
- For classification: try SVM and a neural network.

Also, experiment with various sets of features and compare the results.

Task 4: Explore other models not mentioned in Task 3.

Task 5: Prepare your data and code for further testing:

- Ensure clarity and transparency in your code.
- Organize it to facilitate easy modification of scoring metrics and model parameters.

Task 5: Encapsulate your processing path using a pipeline.

Task 6: Describe the implementation of the model used to solve the project task in a dedicated section of the report.