Project Assignment

Machine Learning 2023 – LECD, LEEC
Informatics Engineering Department

1 Background

In this assignment, you must apply the methods you have learned throughout the semester in a predictive maintenance dataset available at https://www.kaggle.com/datasets/shivamb/machine-predictive-maintenance-classification. In this assignment the goal is to tackle two problems:

1) Predict if there was a failure or not (`Target`)
2) Predict the type of failure (`Failure Type`)

2 Dataset Description (from Kaggle)

Since real predictive maintenance datasets are generally difficult to obtain and in particular difficult to publish, this dataset presents and provides a synthetic dataset that reflects real predictive maintenance encountered in the industry.

The dataset consists of 10000 data points stored as rows with 14 features in columns:

- **UID**: unique identifier ranging from 1 to 10000
- **productID**: consisting of a letter L, M, or H for low (50% of all products), medium (30%), and high (20%) as product quality variants and a variant-specific serial number
- **air temperature** [K]: generated using a random walk process later normalized to a standard deviation of 2 K around 300 K
- **process temperature** [K]: generated using a random walk process normalized to a standard deviation of 1 K, added to the air temperature plus 10 K.
- **rotational speed** [rpm]: calculated from powepower of 2860 W, overlaid with a normally distributed noise
- **torque** [Nm]: torque values are normally distributed around 40 Nm with an Ïƒ = 10 Nm and no negative values.
- **tool wear** [min]: The quality variants H/M/L add 5/3/2 minutes of tool wear to the used tool in the process. and a 'machine failure' label that indicates, whether the machine has failed in this particular data point for any of the following failure modes are true.

3 Objectives

3.1 **Scenario A (Binary Classifier)** - The objective of this scenario is to predict each instance as a failure (1) or not (0).

3.2 **Scenario B (Multi-Class Problem)** – The objective of this scenario is to classify the type of failure as one of six: 'No Failure', 'Power Failure', 'Tool Wear Failure', 'Overstrain Failure', 'Random Failures', 'Heat Dissipation Failure'].

4 Practical Assignment

4.1 Data import

Develop scripts for feature data import. Organize data into sub-sets, relating to each source type you intend to test, e.g: create training, validation, and testing sets. Remove or find ways to handle missing data (for example doing the mean or the median of a certain feature).

4.2 Data Analysis

You need to understand the data you are working with, to that end you should explore it. Use different types of data visualization tools (histograms, pie charts, box plots, correlation plots) and analyse the data. Consider using feature selection technics and see how they affect the performance of the machine learning algorithms. Notice the class imbalance and the different features' domains in this dataset and find strategies to cope. Make sure you know your features! Do not forget to present your findings in the final report.

4.2 Experimental Analysis

You should be able to design experiences in order to run the machine learning algorithms in the given data and evaluate their results. Define the appropriate performance metrics and justify your choices!

Run the experiments multiple times! To be able to present average results and standard deviations (of the metrics used), you can decide to cross-validation. In the end, you should be able to choose the best classifier and evaluate them in a testing set (hold out).

Do not forget that manually inspecting the predictions of your algorithms can give you precious insights into where they might be failing (and why), and what you can do to improve them (e.g., what makes the algorithm fail in this particular case? what special characteristic does it have that makes it so hard? how can I make the algorithm better deal with those cases?). Go back and forward to the hyperparameter choice of the different machine learning algorithms until you are satisfied with the results. It is a good idea to keep track of the evolution of the performance of your algorithm during this process. Try to show these trends in your final report, to be able to justify all the issues involved (choosing parameters, model fit, etc.). You should try to understand how they perform differently in your data.

In this project, you must use **3 different machine learning methods**: Support Vector Machines (SVM) and Neural Networks (NN) are **mandatory** (**if you don't do this you will have a severe penalty in your final grade)**. The third method is up to you to choose, justify the choice.

4.3 Libraries and Language

Your code needs to be in python. You are free to use Colab, Jupiter, or create a script. For Neural Networks we recommend you use Pytorch or Keras.

4.4 Results and Discussion

Present and discuss the final results obtained in your Project assignment. This problem was already studied by other authors. Compare your results with the results from other sources, compare, not copy. In this problem, one important aspect is to evaluate among the data available the more appropriate for the different scenarios.

5 Documentation

5.1 Description

Write documentation (in Portuguese or English) about your project. The documentation should include a cover page where the course name, project title, date, names, and student numbers of the authors are mentioned.

Describe the methods used for classification in such detail that the reader would be able to implement the same functions for feature extraction and classification based on your documentation and some basic background in pattern recognition. Always justify your choices, even when they are based on intuition. Do not forget to verify your assumptions! Include classification results with the given data in your documentation. At the end of your documentation, you should have a list of all references used.

5.2 Requirements, Submission, Discussion

The practical assignment is meant to be done in groups of two. If someone wants to work alone, this is also possible. Larger groups are in principle not allowed.

Final Project Deadline: May 12th, 2023

Deliverables:

- Exploratory Data Analysis
- Experimental design and analysis for both Scenarios, e.g., several models for the different datasets
- Final Report and Code

Project discussion: along the week of May, 15th