UNIVERSIDADE DE AVEIRO

Master Program in Computational Engineering

Applications of Artificial Intelligence

ML PROJECT 2022/2023 - Instructions

Some guidelines for Project 2

- Evaluation by a written report;
- Submission by e-mail for sonia.gouveia@ua.pt, deadline: 8th January 2023;
- Students are expected to work in groups of two students;
- Same specifications and evaluation criteria as those of Project 1;

I. PROJECT GOALS

The goal of this project is to apply suitable machine-learning algorithms learned in class or self-learned to solve a specific data science problem (classification, regression or forecasting, association/similarity, synchrony/delay) in time series data, data streams or data sequences. Represent the results in graphical/table formats and make analysis and conclusions.

II. PROJECT PROPOSALS

Within the scope of the project "Time series analysis for price recommendation in the telecommunications market" the students can explore

- **Topic A**. Forecasting the price for a given product of a company in a univariate or multivariate setting (with information respecting the market competitors).
  Technical approaches based on e.g. RNN, LSTM, Stacked LSTM, Bidirectional LSTM, CNN LSTM, ConvLSTM, GRU, Wavenet, Facebook's Prophet, SARIMA framework, ARIMA Hyndman-Khandakar algorithm, … and multivariate or multi-step approaches.

- **Topic B**. Association and synchrony/delay between time series to quantify a measure for profiling competitors as reactive, proactive or passive. Briefly, a reactive competitor primarily changes its prices in reaction to changes made by other competitors. A proactive competitor initiates price changes without them being triggered by changes from other companies. Finally, a passive competitor is mostly indifferent to other competitors' price changes and rarely changes its prices.
  Technical approaches based on e.g. Pearson correlation, time-lagged cross-correlations, Dynamic Time Warping, instantaneous phase synchrony, …

III. AVAILABLE DATA

Each data file contains time series data of daily prices (€) for several pieces of equipment and different commercial competitors/companies. The name of each file is in the format "xxxx_product_group_id_xx" where xxxx={"long", "wide"} corresponds to the format and xx=number corresponding to the product id. The "wide" format presents data on one competitor/company per column whereas the "long" format presents a column that identifies the company, another column to identify the time and another reporting the price.

The data files are organized into three folders, namely

- time_series_1 – products for which the longest series has > 1 year of samples;
- time_series_2 – products for which the intersection of all time series is > half a year long;
- time_series_3 – products for which the intersection of all time series is > one year long;

There are two additional files: "product_information" contains the information (product name and offer type) associated with a product group, and "read_file" is a Jupyter Notebook demonstrating how to read a time series group file and its information.

IV. PROJECT ASSESSMENT

This project will be graded by

- the ongoing development and interaction with the instructor during the classes
- a report, to be delivered at the end of the project.

The report is evaluated based on a submitted paper (IEEE Latex format). The work done by each student has to be explicitly specified. All files related to the project (pdf and Latex files of the report, the code implementing the algorithms) are sent to the course instructor (sonia.gouveia@ua.pt) in a compressed format having the following name

P2_ML2022_XXXXX_YYYYY

where XXXXX and YYYYY are substituted by the academic (mechanographic) number of each student. If the file is too large to email as an attached document, feel free to use any big file transfer option you may know (wetransfer, dropbox, link in a cloud, etc).

IV. EVALUATION CRITERIA (total score 20)

[1] Report content (12/20):

- Data description and preprocessing (if necessary normalization, feature selection, transformation, etc.). Motivation for choosing the particular problem.
- Data visualization (histograms, box plots, other plots).
- Short description of the implemented ML models.

- Model training (data splitting – train, validate, test, k-fold Cross-validation). Visualize graphically the cost function trajectory over iterations. Training with regularized and nonregularized cost functions, if applicable.
- Model hyper-parameter selection - regularization parameter lambda, number of NN hidden layer units, number of hidden layers (if necessary), etc. Systematic approach instead of just one or several randomly chosen values.
- Performance comparison between the models.
- Results in graphical or table formats.
- Conclusions.
- Problem complexity.

[2] Report formatting (3/20):

- IEEE Latex format, affiliation (Department, University, subject, course instructor), abstract, keywords, workload per student.
- Sufficiently detailed report.
- References, reference citation in the report.
- Clear figures (title, legends, axis labels) and tables referred in the text.

[3] Novelty and contributions (5/20):

- Compare different solutions to solve the problem as well as with other literature approaches (published references), trying to propose a better solution, e.g. improve the performance of the ML model in solving the problem at hand.