USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA
UNIVERSIDADE DA CORUÑA
Universidade Vigo

P2.1: Recurrent Neural Networks (RNNs).

- Deadline: April 13, 16:00.
- Objective. To develop different deep recurrent neural networks to solve a prediction problem for time series.
- Dataset.
    - We will use the Walmart Sales Dataset of 45 Stores (Kaggle dataset). The file has information about the Weekly Sales of 45 stores for the year 2010-2012, including the factors affecting Sales such as Holidays, Temperature, Fuel Price, CPI (Consumer Price Index), and Unemployment.
    - It has 6,435 data rows (143 for each store). The rows are ordered by store and, for each store, by date.
    - The file generateWalmartDataset.ipnb creates the time series for training and testing. Test data is the 20% most recent data of all stores. The time series is generated with the method generateTrainTestData, which has as arguments the name of the data file, the percentage of test data, the length of the sequences to be generated, and the batch size.
    - The values of the variables in the time series have been normalized by the generateTrainTestData method. Thus, for a variable x, the variable has been transformed to $x'=(x-\mu_x)/\sigma_x$, where μ_x is the mean of the variable and σ_x is the standard deviation.
    - The only parameter that can be modified for this method is the length of the sequences to be generated. The values of the other parameters must be kept as they are.
    - The method returns the train and test data, the standard deviation of the weekly sales, and the number of input features.
    - The output feature (the feature to be predicted) is the weakly sales in 3 weeks.
    - As an example, if the sequence length is set to 2, each training/test sample will consist of: Weekly_Sales (t-1), Holiday_Flag(t-1), Temperature(t-1), Fuel_Price(t-1), CPI(t-1), Unemployment(t-1), Weekly_Sales (t), Holiday_Flag(t), Temperature(t), Fuel_Price(t), CPI(t), Unemployment(t), Weekly_Sales (t+3). The value to be predicted is Weekly_Sales (t+3).
- Evaluation. The evaluation metric that must be used is MAE. As data is normalized, you must denormalize MAE to provide the final result. To denormalize MAE you must multiply the normalized MAE by the standard deviation of the weekly sales.
- Task to be carried out.
    - Use the training set to train the network, and the test set to provide the final result.

- You model can have any type of Keras layer (Dense, SimpleRNN, LSTM, etc.), including combinations of them. However, **at least one of the layers of the model must be any kind of RNN**.
- You can try any hyperparameters for your model: number of layers, number of units, activation function, regularization, etc. Also, the sequence length hyperparameter can be modified.

- Submission.
  - The exercise will be developed using a Jupyter Notebook and the Keras framework.
- The notebook should include:
  - The practice can be carried out alone or in pairs, so the first cell of the notebook must be the full names of the authors.
  - The code for each of the models developed should be included and it should be a complete ML process: data loading and manipulation, network creation, training, and results.
  - The notebook will be saved with the results of its execution included.
  - The code shall be accompanied by cells with an explanatory report containing a description of the process followed, detailing the problems encountered and justifying the decisions taken. It should also include a section on results and discussion of them.
- Submission process.
  - The exercises will be submitted using the virtual campus of each university:
    - Universidade da Coruña: https://campusvirtual.udc.gal/
    - Universidade de Vigo: https://moovi.uvigo.gal/
    - Universidade de Santiago de Compostela: https://cv.usc.es/
  - Each member of the practice group must submit the notebook in their corresponding Moodle task.
  - There is a strict deadline for each assignment. Past due submissions will be rejected.
- Evaluation criteria.
  - Quality of the predictions made. Take as a reference:
    - A Mean Absolute Error (MAE) of less than 68,000.0 on the test set.
  - Quality of the design.
    - The network design follows the recommendations on how to create RNNs.

- Quality of explanations:
  - The process is adequately detailed, and the decisions made are justified.
  - The results are commented and interpreted correctly.