EAS 595
Fundamental of Artificial Intelligence
Fall -2023
project 3
Total Marks: 100
Deadline: 12/8/2023, 11:59 pm

In this project you are going to perform time series analysis on a given data set. Time series analysis is a specific way of analyzing a sequence of data points collected over an interval of time. In time series analysis, analysts record data points at consistent intervals over a set period rather than just recording the data points intermittently or randomly.

Data set: https://archive.ics.uci.edu/dataset/360/air+quality

Air Quality dataset will be used to perform time series analysis. The dataset contains 9358 instances of hourly averaged responses from an array of 5 metal oxide chemical sensors embedded in an Air Quality Chemical Multisensor Device. The data set contains numerical values for different chemicals received by sensors.

Data were recorded from March 2004 to February 2005 (one year) representing the longest freely available recordings of on field deployed air quality chemical sensor devices responses. Ground Truth hourly averaged concentrations for CO, Non Methanic Hydrocarbons, Benzene, Total Nitrogen Oxides (NOx) and Nitrogen Dioxide (NO2) and were provided by a co-located reference certified analyzer.

Note: This project is focused on Time series analysis, and you are free to use any libraries.

Part 1: (Data Preprocessing) [20]

1. In your Jupyter Notebook or .py script, define a section as Part I: Data preprocessing.
2. Load the dataset, replace null or missing values if needed.
3. Drop 'Date' and 'Time' columns and plot correlation matrix for the rest of data.
4. 'PT08.S4(NO2)' is your label, decide any 4 features which highly correlates with the label.
5. Plot any 2 graphs on the final dataset.
6. Sequence is very important in time series analysis, never shuffle the data.
7. Apply 'MinMaxScaler()' transformation on the final data to keep it in a particular range.
8. Split data into train and test sets.
9. Create sequence for time series analysis, keep sequence length = 10. (Refer any online resources to understand and implement this, cite the reference in your report.)

In your report for Part 1

1. Briefly describe the nature of your dataset.
2. Include the correlation plot and mention the reason for choosing specific features.
3. Include all the graphs and write your understanding.
4. What is the purpose of breaking data into sequences.

Part 2: (Modeling and Evaluation) [RNN-20 + LSTM-20 = 40]

1. Implement RNN and LSTM on the prepared data using TensorFlow library.
2. Train the models for at least 20 epochs.
3. Save both trained models.
4. Read the saved model and then apply it to the test set.
5. Only use 30% of the test data to clearly visualize the output.
6. Your output graph should include both predicted and actual plots in different colors.

In your report for Part 2

1. In two lines write your understanding of RNN and LSTM models.
2. Include both the graphs in your report and write your analysis.
3. Which model do you think performed well and why?

Part 3: (Bias and Variance) [40]

1. The goal of this part is to help you determine the appropriate model based on bias and variance.
2. Import and read iris dataset from sklearn library and split the data into train and test sets.
3. Further split the train set into 10 subsets. (shuffle the data)
4. Train Decision Tree classifier model for each subset.
5. After the above step you will have 10 models for 10 subsets, now get train and test error for all the 10 models.
6. Plot a single graph showing the train and test errors for all the models.
7. Which model is the best and why?

In your report for Part 3

1. What is bias and variance in ML?
2. Include the graph and write your conclusion.

Academic Integrity:

The standing policy of the Department is that all students involved in any academic integrity violation (e.g., plagiarism in any way, shape, or form) will receive an **F grade** for the course. The catalog describes plagiarism as "Copying or receiving material from any source and submitting that material as one's own, without acknowledging and citing the particular debts to the source or in any other manner representing the work of another as one's own."

Submission

- Fully complete all parts of the assignment.
- Submit to UBLearns->Assignments.
- The code of your implementations should be written in Python. You can submit multiple files, but they all need to be labeled clearly.
- Multiple submissions are allowed, only the recent one is considered and graded.
- All assignment files should be packed in a ZIP file named: UBIT_project3.zip (e.g., anup1234_project3.zip). The zip file should contain the code (code.ipynb or code.py), saved models(lstm_model.h5, rnn_model.h5) and report(report.pdf). Your Jupyter notebook should be saved with the results. Suppose you submit python scripts after extracting the ZIP file and executing command python main.py in the first level directory. In that case, all the generated results and plots you used in your report should appear printed out clearly.
- Include all the references that have been used to complete the assignment.
