LOHUM

Introduction

- Lithium-ion batteries are becoming ubiquitous in our lives, from consumer electronics to EVs. In this assignment, you will work on time series data of these batteries in two different situations. This assignment comprises two problems.
- Problem 1 - Feature extraction from battery data
  - You are provided with the data collected from a battery that has been operational in an electric vehicle.
  - It is a time-series data sampled at a variable rate.
  - The data is provided as a csv file.
  - You are required to complete four tasks in this part.
- Problem 2 - Building a classifier
  - Build a classifier model that can accurately classify cells into three classes - A, B & C.
  - The data is provided as json files.
  - You are required to complete 3 tasks in this problem.
  - You will be scored based on the accuracy of the model. So ensure that you tweak your model to get maximum performance.
- We have provided the definitions of the basic concepts. Please feel free to use Google to understand them further.
- Submission :
  - Use Python and [Google Colaboratory] to write your code.
  - Submit your response as a link to your colab file.
  - Complete each task in a separate section and add comments to your code wherever necessary.
  - Use the standard packages for data processing and regression. Clearly document any external packages used by your code.
  - At the end of the notebook, include a section as a short write-up with subsections for each part and sub-subsections for each task, to present your answers and explanations or anything else of interest.

Problem 1 - Feature extraction

Definitions:

1. SOC (State Of Charge) : It is the level of charge in a battery relative to its capacity. A battery is fully charged when it is at 100% SOC and fully discharged when it is at 0% SOC.

2. FEC (Full Equivalent Cycles) : FEC can be defined as the number of **full** charge-discharge cycles that a battery has undergone. When a battery gets charged from 20% SOC to 70% SOC and then discharged back to 20% SOC, its FEC does not increase by 1. Instead, it takes two such "half" cycles for the FEC to go up by 1. This is because the FEC is measured based on the charge throughput of the battery; it increases by unity only when the entire capacity of the battery gets drained from it and then pumped back in again regardless of the number of steps. (Read this short description for more clarity - refer to section 2.2)

Description of the dataset:

1. There are 4 features : timestamp, soc, voltage, current.
2. The units of voltage and current are Volts and Amperes respectively.

Tasks:

1. Perform data preprocessing - remove null values, look for outliers and clean up the data.
2. Create appropriate visualizations for each feature using matplotlib. Comment on the distribution of values for each feature.
3. Based on SOC and current, create an algorithm to **obtain the FECs that the given battery has undergone**. Present the obtained FEC along with your reasoning in the write up section. Please note that the 'current' column does not discriminate between charging current and discharging current. (hint : add a new feature – 'charge current' by monitoring the change in SOC).
4. The capacity of a battery is measured in Ampere hours (Ah). Is it possible to obtain the capacity of the battery from the given information? If yes, calculate the capacity of the battery as an average over the first ten cycles.

Problem 2 - Model building

Definitions and concepts:

1. Discharge voltage profile : It refers to the measured voltage values while discharging batteries at a specific C-rate (Read this description of C-rate).
2. The typical voltage range for a battery while discharging starts from 4200 mV to 2700 mV. But, this may not always be the case as discharge profiles may vary depending on battery health, type etc. [Hint : Prepare the training data judiciously to ensure consistency between inputs and prevent unnecessary data loss]

Description of the dataset:

1. There are 3 features for each cell : timestamp, voltage and current.
2. The units of voltage and current are milliVolts and milliAmperes respectively.
3. The data is provided as nested json serialized files with integer keys for batteries and further key names provided as **'d_voltage', 'd_current', and 'd_time' for the three features.**
4. The label names are the file names themselves; for example, if the file is named 'A.json', the vehicle that those batteries belong to is 'A'.

Tasks:

1. Data preprocessing
    a. Perform data cleaning; treat missing/incomplete/faulty data or disregard if necessary.
    b. Normalize or scale the features if necessary.
    c. Convert timestamps into seconds, if necessary.
    d. Split the data into training (80%) and testing (20%) subsets.
2. Model selection and training
    a. **Build a classifier model to identify which vehicle each battery is from. The vehicles are A, B and C.**
    b. Describe the algorithm chosen and explain why they might be suitable for this task.
    c. You may try multiple algorithms and compare the performance of each algorithm on the test dataset.
    d. Evaluate each model on the test set and compare their performance using accuracy, precision, recall and F1-score.
    e. **Present your results in a table and discuss which model performed the best. Remember that this is the most important task based on which your solution will be evaluated. Please highlight the best model accordingly.**
3. Classification report and confusion matrix
    a. Create a classification report and a confusion matrix from the results.