ME 592: Data Analytics and Machine Learning for Cyber-Physical Systems

Homework 1

Homework Assigned on: February 6, 2025 Homework Due on: February 18, 2025

This homework serves as an introduction to python and data preparation/exploration. This first homework will be completed individually, but feel free to discuss with your classmates.

1 Git Repository & Code Management

All homework assignments must be submitted via Git commits, including all relevant code files. You need to commit the `*.ipynb` file from your JS2 instance to the Git repository. While these repositories can be made **public** on either BitBucket or GitHub, we recommend creating a repository on GitHub. Please send us the link to your Git repository and ensure that it is set to **public**.

2 Simple Programming & Exploratory Analytics

The following problems must be coded using specified tool only in order to get everyone to use python.

2.1 Images

Task:

1. Download the MNIST dataset using the torchvision library and create a subset of the dataset containing 1,000 samples. Each sample will include a 28 × 28 pixel image, along with a single integer value denoting the sample's respective class (number).
2. Plot a histogram of classes in your subset of MNIST.
3. Use the einops python package to 'batch' the subset of MNIST **images**. Each batch should contain 25 different samples.
   Hint: The subset of MNIST image data should change dimensions from [1000, 28, 28] → [Number of Batches, Batch Size, 1, 28, 28]
   Context: Deep learning models operate over batches of samples. This rapidly speeds up computation as the forward pass runs parallel across the batch dimension. Batching also can improve optimization and provides the *S* in SGD (Stochastic Gradient Descent). Additionally, neural networks expect a channel dimension for image data. MNIST is grayscale so it only contains one channel, but other images can contain color, in which case they will have three channels, RGB.
   Once you have completed this, Congrats! You've built your own dataloader!
4. Randomly select an MNIST image sample and, using matplotlib, plot it in three dimensions. The x and y-axis' should be the respective pixel locations, and the z-axis should be the pixel intensity.

2.2 Time Series

Data: Experimental data used to analyze appliances energy use in an energy efficient building. The data set is collected at a frequency of 10 min for about 4.5 months. The house temperature and humidity conditions were monitored with a ZigBee wireless sensor network. Each wireless node transmitted the temperature and humidity conditions at a period of around 3.3 min. Then, the wireless data was averaged for 10 minute periods. The energy data was logged every 10 minutes with m-bus energy meters. Weather from the nearest airport weather station (Chievres Airport, Belgium) was downloaded from a public data set from Reliable Prognosis (rp5.ru), and merged together with the experimental data sets using the date and time column. Random variable is included in the data set for testing the regression models and to filter out non predictive attributes (parameters). This data is adopted from the UCI machine learning repositories [1] and several aspects of the data were analyzed [2]. Our motivation is to explore some aspects of this time series data.

Task: Load the data (energydata_complete.csv) and perform the following analysis.

1. Plot the appliances energy consumption for whole period and a closer look at any one week of consumption.
2. Plot heatmap of hourly consumption of appliances for a week. An example heatmap looks like Figure 1.
3. Plot the histogram of energy consumption of appliances.
4. Construct a feature variable NSM (no. of seconds from midnight) and plot energy consumption vs. NSM.
5. Plot appliances energy consumption vs. Press_mm_Hg.
6. It is observed that the major contributing factors for the energy consumption among all other features is NSM and Press_mm_Hg. Comment on it.

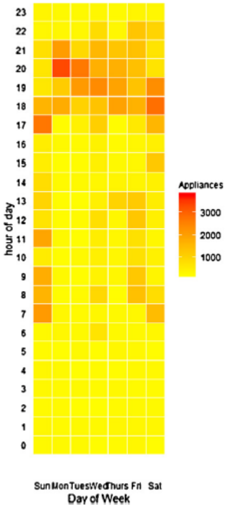

Figure 1: Example heatmap of hourly energy consumption of appliances over a week

2.3 Multi-variate

Data: The NASA data set comprises different sizes of NACA 0012 airfoils at various wind tunnel speeds and angles of attack. The span of the airfoil and the observer position were the same in all of the experiments. This problem has the following inputs:

1. Frequency, in Hz.
2. Angle of attack, in degrees.
3. Chord length, in meters.
4. Free-stream velocity, in meters per second.
5. Suction side displacement thickness, in meters.

The only output is Scaled sound pressure level, in decibels.

Task: Load the data and Compute the following descriptive statistics of the data:

1. Mean
2. Variance (or Standard Deviation)
3. Median
4. Kurtosis
5. Skewness
6. Range

Data

The following hyperlinks direct you to the URLs of the publicly available data.

1. Energy Data
2. Airfoil Data - Use the airfoil_self_noise.dat file.

Expected Outcome

Final code committed with results for Section 2 should be pushed before the deadline.

References

1. https://github.com/LuisM78/Appliances-energy-prediction-data

2. Data driven prediction models of energy use of appliances in a low-energy house. Luis M. Candanedo, Véronique Feldheim, Dominique Deramaix. Energy and Buildings, Volume 140, 1 April 2017, Pages 81-97, ISSN 0378-7788

3. https://archive.ics.uci.edu/dataset/291/airfoil+self+noise

4. T.F. Brooks, D.S. Pope, and A.M. Marcolini. Airfoil self-noise and prediction. Technical report, NASA RP-1218, July 1989.