CSC 591/ECE 592 IoT Analytics - Project 2
- Data Acquisition and Analytics

Objectives

The objective of this project is to acquire a data set and develop a set of understanding insights with data analytics. For this analysis you can use any statistical package, such as MatLab, R, SAS, or use Python with all the available statistical functions. You may also use multiple packages for different tasks.

Task 1: Generation of Data Set

An IoT sensor is any sort of mechanism or tool, such as a camera or air quality monitor, integrated into a device. These sensors gather information, related to the environments in which they're deployed and transmit it to the cloud via Wi-Fi, bluetooth, 5G or other mobile network. Take smart thermostats, for instance, which are equipped with temperature sensors that measure fluctuations in a home's temperature. Because these thermostats are connected to the internet, users can log in via an app to monitor the temperature of their home and control the thermostat remotely.

What are IoT data sets?

- An IoT application with any one or two sensory devices can be selected based on the interests and availability of each student.
- A sensory device is a device that provides tactile sensations through vibration and sound to mimic the feeling of an object with a large mass. IoT sensors are pieces of hardware that detect changes in an environment and collect data. A set of sensory devices include follows, but not limited: Biomedical sensors Cameras Electric current sensors Flow sensors Gyroscopes Humidity sensors Motion sensors Pressure sensors Temperature sensors
- Data sets can be collected via any sensory devices, and they can also be obtained by existing APP, such as fitbit, Apple watch, MI motion, and more.
- For addition information, contact Ms. Mengning Li (mli55@ncsu.edu) for other possible settings and data collection.

What are the requirements of a data set?

Here are some details that students need to be considered in data collection:

- What are the *output* or *resulting* data? For instance of audio speaker, the volume or the acoustic signal strength can be used as the ouput data, which should be the consequential data resulting of a set of parameters.

- What are the *array* of factors or parameters that will affect the output data? Again, for the same example, there can be a large number of factors, such as the distance between the MIC and the acoustic receiver or measurements, the background noise, the objects and room size, the angle between the transmitter (i.e., the MIC) and receiver (i.e., the acoustic receiver).
- What is the total samples? We suggest to collect at least 200 data samples for the same settings, which will be used in the later projects. Up to 1,000 samples may be even better in training data sets. This in turn requires to consider other issues, such as the time period of each epoch, data storage size, the total time period of each setting, the design of settings, and variety of microphones, etc.
- What are the number of independeent variables that will affect the output data? We suggest to consider 3 to 5 variables, even though there can be much more than 20 factors to be considered.

What are IoT applications for data collection?

We have designed following IoT applications:

1. Acoustic signal: For acoustic signal data collection, the following equipment and software are required: (a) An audio recording device (e.g., Mobile Phone) or (b) A computing Device (Laptop/Desktop). For the **data sets**, our focus will be on recording and analyzing sound data. It is advised to save the recordings in widely-used digital audio formats, such as .mp3 or .wav, to facilitate easier processing and analysis.
2. WiFi signal (for mmWave-contact Mengning): The details of WiFi configuration can be found on Moodle site, under Project 2. The implementation requires following equipment and software.
   - Wi-Fi Adapter (capable of monitoring mode)
   - Directional and Omni-directional Antennas
   - A computing Device (Laptop/Desktop)
   - Network Monitoring Tools (e.g., Wireshark, tcpdump)
   - Signal Analysis Software (e.g., inSSIDer, WiFi Analyzer)

   For the **data sets**, we focus on essential Wi-Fi metrics, with an emphasis on Channel State Information (CSI). Students are expected to pick two other metrics from:

   - Channel State Information (CSI): Provides detailed information about the channel properties, such as fading, power decay, and temporal and spatial characteristics of the wireless channel, crucial for advanced Wi-Fi performance analysis and diagnostics.
   - Signal Strength (dBm): Measures the power level received by the Wi-Fi client, indicating the intensity of the Wi-Fi signal at the receiver's location.
   - Signal Quality (SNR): Represents the Signal-to-Noise Ratio, indicating the quality of the Wi-Fi signal by comparing the level of the desired signal to the level of background noise.
   - Throughput: The rate of successful data transmission over the Wi-Fi network, measuring how much data is successfully sent and received.
   - Network Traffic: Encompasses the analysis of data packets moving through the network, including packet sizes, transmission rates, and protocol types.
   - Channel Utilization: Involves assessing the frequency of use and congestion levels on different Wi- Fi channels, crucial for understanding network performance in crowded environments.

- Latency: The time it takes for a data packet to travel from the sender to the receiver, which is a critical factor in evaluating the responsiveness of a Wi-Fi network.

3. Brainwave (EEG) with MUSE: Required equipment and software for MUSE (EEG) data collection include: (a) MUSE Headband or Equivalent EEG Device, (b) EEG Data Analysis Software (e.g., MUSE Direct, OpenBCI), and (c) Computing Device (Laptop/Desktop). In the MUSE (EEG) **data set** collection, students need to choose and analyze three of the following key aspects:

   - Brainwave Patterns: Identifying and analyzing different brainwave frequencies, such as alpha, beta, theta, and delta waves.
   - Cognitive State Analysis: Interpreting various mental states (e.g., focus, relaxation, stress) based on EEG data patterns.
   - Event-Related Potentials (ERPs): Investigating the brain's electrical response to specific stimuli or events, such as sensory input or cognitive processes.
   - Sleep Patterns: Analyzing sleep stages and cycles, including REM and non-REM sleep, using EEG data.
   - Neural Connectivity: Studying the functional connectivity between different regions of the brain and how this relates to cognitive activities and states.
   - Attention and Focus Dynamics: Examining changes in brainwave patterns related to attention, concentration, and distractibility.

If a student is not interested in collecting IoT data in real systems or settings. There are two options:

1. **Driving Monitoring Datasets (DMD):** This is one of the most comprehensive online source that provides a large set of data for driving-monitoring with a good level of multi-channel data and 'clean' data sets of variables. The web site is `https://dmd.vicomtech.org/#activities`: The student needs to use google (e.g., ncsu) account to associate with an INDIVIDUAL account on BOX. Then any one of the examples with relevant 'variable' can be downloaded. One suggestion is to watch out **file size** before downloading. As the number of data sets (e.g., videos more than camera, etc.) and files are large, make sure you download one by one in a *quite* place to avoid interruption. Also, be prepared to have a 1TB USB disk to save the data, if necessary

2. Simulation data sets: students can either contact the instructor or Ms. Mengning Li for simulated data or existing data sets.

3. Benchmark data sets: There are millions of available data sets online, such as the temperature in Raleigh over years, number of students or populations in North Carolina, and more. The **important** factors need to be considered is the number of *columns* in the data set, e.g., there are 3 to 5 variables that are included in the data set.

Task 2: Basic Statistics Analysis

This task is centered on conducting fundamental statistical analysis of a dataset comprised of variables X, with an output variable Y as a function of these variables. The analysis is divided into **two** stages: initially, a small dataset of 20 samples is analyzed manually, followed by a more complex analysis of 100 samples using coding techniques.

For each variable, name them as X_1, X_2, X_3, X_4 and X_5, where the output data Y is a function of these variables. The following tasks can be obtained by using a set of 20 samples with hand-calculation, and 100 samples with code.

2.1 **Descriptive statistics:** For each variable X_i, i.e., column in the data set corresponding to X_i, find the mean, median, and variance to understand the central tendencies and variability. Also find the range and interquartile range (IQR) to assess data spread and the central 50%.

2.2 **Probability Mass Function (PMF):** For each variable X_i, (a) segment the range of the dataset into 50 equal intervals, ensuring that the minimum and maximum values of the dataset are included; (b) Plot the PMF for each variable X_i . The PMF will exhibit the probability distribution of the discrete variables in the dataset.

2.3 **Analysis Report: SOCS Framework:**

1. **Shape:** Describe and analyze the distribution shape of each variable X_i , such as symmetric, skewed, etc., as inferred from the PMF and descriptive statistics.
2. **Outliers:** Identify and discuss any outliers present in the dataset. Evaluate their potential impact on the overall analysis.
3. **Center:** Discuss the central tendency measures (mean, median) for each variable X_i and interpret what these measures reveal about the dataset.
4. **Spread:** Analyze the variability of the data, using measures such as range, IQR, and variance, for each variable X_i . The report should elucidate a comprehensive understanding of each variable's statistical properties and their implications on the output variable Y .

Task 3. Data Visualization

This task emphasizes the importance of data visualization for understanding and interpreting complex datasets. Students will create various plots to effectively illustrate the data, revealing underlying relationships and distributions.

3.1 Create a boxplot (or box-and-whisker plot) for each variable (X_1, X_2, X_3, X_4, X_5). These plots will highlight the median, quartiles, and outliers, providing a concise overview of each variable's distribution.

3.2 Plot Scatter plots to display the relationship between the output data column Y with respect to each variable (i.e., the column).

3.3 Plot the density curve of the output data.

3.4 Analysis and interpretation: (a) (a) Analyze the findings from the boxplots, scatter plots, and density curve. (b) Discuss key observations, such as variable correlations, outliers, or data distribution trends. (c) Reflect on how these visualizations enhance the understanding of the dataset and the analysis process.

What to Submit?

1. For each task 1, 2, and 3 submit the following:

   The code you used for the task. It does not have to run on eos, and it may be a number of different pieces of code from different packages.

Sharing code is not allowed and constitutes cheating, in which case both students (the one that aids and the one that receives) will get a zero for the project and will be reported to the student conduct office.

2. Your results (graphs, tables, etc) and your conclusions.

You will receive a bad grade if you submit results without substantive conclusions, or conclusions that are not backed by sufficient results.

Grading:

- Task 1: 25 points: (i) 15 points for Equipment Setup and Software Configuration, and (ii) 10 points for effective configuration of the router and execution of the data collection process.
- Task 2: 35 points: (i) 5 points for coding, which refers to the accuracy in statistical calculations and correct implementation of the PMF construction in the coding. ii) 15 points for data visualization, which refers to the clarity and precision in the visual representation of PMFs and other statistical plots. Each visualization should be appropriately labeled and effectively convey the statistical characteristics of the dataset. (iii) 15 points for analysis, which refers to the in-depth analysis of each X_i using SOCS, highlighting their influence on Y.
- Task 3: 40 points: (i) 20 points for evaluation based on the accuracy, clarity, and effectiveness of the boxplots, scatter plots, and density curve. (ii) 20 points for assessment based on the depth of analysis, and coherence in interpreting the visual data and the implications it holds.

Remember that you will be graded mostly on your ability to interpret the results