Measuring our environment: Introducing physical sensors

Problem statement

Sensors play a critical role in generating signals that provide us with information about local systems and, in a collective sense, about the world. There is a large amount of data that can be gathered from our surroundings and used to analyze nearly anything, from the flow of people to the quality of public water and air to the stability of a bridge under load.

The recent proliferation of the Internet of Things (IoT) and the growing ubiquity of Smart devices provides the data scientist with an unprecedented perspective of the environment, its constituent elements, and behavior.

The high level goal of this project is for you to go thought the whole data science pipeline as defined in the first lectures with data that you have **collected yourself**. Hopefully you will appreciate that data does not always come from the 'Cloud' and that you can gather meaningful data yourself, taking into account any errors or experimental design issues.

READ THE USER GUIDE AND WATCH THE VIDEO

(click on links to open)

1. Video of Introduction:
   https://video.seas.harvard.edu/media/CS109a-Project-Sensors-10-10-19/1_aunv6hfu
2. Sensors User Guide:
   https://github.com/Harvard-IACS/2019-CS109A/blob/master/content/projects/CS109aSensorsUse
   pdf (file is in the same folder in public git)

Physical Device

You will be given two premade, predominantly single-ended embedded systems that feature sensors to track temperature, humidity, pressure, light, and motion. Each device will not only log the corresponding data against time, but also permit control over data collection parameters such as sampling rate.

Data collection in the sensors you were given is achieved through a process called transduction, wherein sensors act to convert a physical phenomenon into an electrical signal, which, depending on how it is manipulated, then constitutes one or another class of data. The complexity of the resulting data can be tied to the mechanistic basis for transduction, the parameters used in sampling or conditioning the signal, and the inherent stochasticity and interdependence of nature.

Sensor Package Specifications:

- Temperature (−20 to 60°C ±0.2)
- Humidity (20 to 95% RH with resolution 0.1%)
- Pressure (300 - 1100 hPa ±1 hPa)
- Light sensor (cadmium sulfide photoresistor AFEC)
- Motion sensor(s) (passive infrared detector with 150 cm over 100° range; records binary 0 or 1 pulses with 2sec resolution)
- Date and time (not recorded constantly but in starting points)

Each team is to place their sensors in two different spots of their choice inside the room of one of their team member's Harvard Dorm or place of residence. Please talk to your TF if you want to place the device outdoors or in a common area like a House dining room. Make sure you take care of having the device powered at all times. There are issues with the battery (see video) so it is preferred to have it constantly plugged in.

Set it to sample time, humidity, temperature, motion, and light in specific time intervals by setting the **sampling rate** and **duration**. Think of what **sampling rate** you want for each sensor. You may set different values for different sensors. For motion, for example it makes sense to sample every 1 sec, but for light every 5 min. Think of what the sampling rate for temperature should be (Hint: every 30 min or 1h?)

Data Resources

You will get the data off the sensor in an SD card. We think that a total data collection time of approximately **3 weeks** will give you enough data.

Make a csv with the following variables:

- **sensorcode**: "first 3 letters of location+CS109aGroupNumber+"-1" e.g. CAN025-1 (location should be Dorm name or other location such as CAM, SOM, ALL, etc). There should be a -1 or -2 at the end designating the number of sensor (each team gets two).
- **temp**: sampled at a rate of your choice
- **hum**: sampled at the same rate as temp
- **press**: sampled at the same rate as temp
- **light**: sampled at a rate of your choice
- **motion**: sampled at a rate of your choice
- **time_string**: only starting time is recorded. You will have to deduct each timestring from the initial time and the sampling rate. Careful with this!
- **your own custom variable (optional)**: you might want to include some other variable, other that the ones you are trying to predict, of course.

High-level project goals

We want you to go through the data science process: a) ask an interesting question, b) design the experiment and get the data, c) examine the data, d) make a baseline model to answer your question, e) test the model and try to make a better one, f) visualize your results.

The first step is to construct your dataset. Do not wait until the end of the collecting period. Download data every couple of days to check on the process. Once you are sure everything works you may allow longer recording periods. Also you do not have to wait until the end of the collecting period to start analyzing your data. You could do so with a small batch.

Do some EDA: e.g. plot the variables against time of day and day of the week, calculate min, max, and mean for each day for the variables for which this makes sense, eg. temp, pressure, and humidity. For measured light what would make sense? Note any extreme values as outliers and suggest ways of dealing with them. Take care of the missing values introduced by the different sampling rates.

Determine what data are important and what data might have noise or interference. For instance, how would you determine the difference between sunlight and artificial light?

There are a lot of questions that can be answered using the above data. We encourage you tho come up with your own questions. Below you will find ideas.

1. Plot averaged temperature curves during the day and look for patterns of overheating suggesting a suboptimal usage of the HVAC system and therefore wasted energy.
2. Does light correlate with activity, i.e. do students leave lights on when away?
3. Predict the outside temperature, or the time of day or day of the week, using indoor measurements.
4. Make a model to predict a) the outside temperature, and b) the time of day or day of the week.

Milestone 2 deliverables: what is mentioned in the assignment definition **plus** some data points from the SD card so we can make sure the process for the data collection is in place.

Authors

Eleni Kaxiras (SEAS CS109a), and Evan Smith (SEAS Active Learning Labs).