Samuel Scovronski
SpringBoard Capstone 2 Project Proposal
Gas Sensor Array Drift at Different Concentrations
September 2, 2021

What is the problem you want to solve?

- The problem that requires solving is twofold. First, we need to accurately predict/ determine what gas is being flowed and at what concentration using a model. Multiple models will need to be evaluated to see what the tradeoff is between the number of sensors used, and the number of runs used to create the model and the accuracy of the prediction. Second, we need to determine if the model stays accurate through the life of the sensor(s), or if there is deviation, determine why it is occurring and what can be done about it.

Who is your client and why do they care about this problem? What will they do or decide based on this analysis?

- The client cares about the first part of the problem because they need to know what gas is flowing and its concentration, but they want to use the minimum number of sensors and run as few experiments as possible, while maintaining sufficient accuracy. This would allow them to save money (reduced sensors) and time (# of experiments). Second, the client will want to be able to determine when a sensor(s) needs replacement either on their equipment or their customer's equipment so costly unexpected maintenance is not required or be able to compensate for any drift so they can save money by extending the usable life of the sensors.

What data are you using? How will you acquire it?

- All the data is available through the UCI Machine Learning Repository. It is provided courtesy of the BioCircutis Institute at the University of California San Diego. Acquiring requires downloading the dataset from the website.

Solution Approach

- Use the first half or third of the dataset (with respect to time) to create a "new" sensor model capable of classifying the chemical being run and at what concentration. Initially, the data will be split such that 70% is used for training a model and 30% for testing. The model creation process will be repeated for different training/test split percentages to create a tradeoff between the number of experiments required and the accuracy of the model. This will address the first aspect of the project that the client cares about. The best model will then be selected and used to predict the chemical and concentration on the second half of the dataset (with respect to time) with the expectation that there will be more error if drift is occurring.

Deliverables

- GitHub repo with a Jupyter notebook containing my code and appropriate comments, presentation on my findings, and a project report