Niklaus Parcell - CS5350 - Final Project Report

May 4, 2019

Problem Definition and Motivation

Currently, there are numerous factors contributing to global warming and climate change, i.e. greenhouse gasses/ emissions, agriculture, etc [1]. In Salt Lake city, it is common for locals to recognize the inversion that settles in the valley, especially in the winter and summer seasons. In addition, there are atmospheric temperature differences, where polluted air "sinks" and stays closer to the valley floor in the winter [2]. In summer, particulates from surrounding fires pollute the area. In efforts to study the behavior of pollutants in the Salt Lake Valley, the Division of Air Quality (DAQ) has mounted sensors in various local areas that all have different observed measured levels of pollutants. These sensors also take other measurements such as temperature, barometric pressure, NOx concentration, etc. However, although accurate, these sensors can be expensive and therefore limiting on how many can be mounted across the valley. Thus, the Department of Chemical Engineering at the University of Utah has engaged a project to mount numerous cost-effective AirU sensors around the valley in commitment to quantify air quality data [3]. The sensors work by using a semi-conductor (SnO_2) that experiences changes in resistance as gases adsorb to or release from the surface [3]. However, although the AirU sensors are cost-effective, they have some measurement imprecision that needs to be fitted to match the data collected from the DAQ sensors. Ideally, these could be used to predict what factors are affecting air quality the most.

Solution

For this project, the objective is to use the data collected from both DAQ and AirU types of sensors in parallel, and fit the collected AirU data to DAQ data to have this cost-effective sensor "be smarter." By using machine learning techniques, Objective 1) is to create a predictable model with a neural net, and analyze the error of predictions. Objective 2) is to see which factors are affecting air quality the most.

- Objective 1) Creates a neural net using TensorFlow, Keras, and Pandas modules. Using data from the DAQ and AirU sensors, the ozone levels in ppm can be predicted. One training method is to train on roughly 90% of the data, and use the remaining 10% as testing data with test predictions from the neural net to see how accurate the model is. The second method of training would be to use two datasets and compare test predictions versus actual labels to visualize error.
- Objective 2) involves performing a gradient descent method that returns a weighted vector that shows which variables are affecting air quality the most; i.e. which environmental factors are most important/unimportant. Both of these steps could, in effect, both apply a smarter method to recording data with cost-effective sensors,

and then see which factors are affecting air quality the most. After these steps are completed, perhaps more ideas of how to fight air quality in the Salt Lake valley can be postulated and pursued in efforts to relieve certain areas of poor air quality and have cleaner air in Utah.

Thus far in the project for AirU, the professors and students have gone through a process of installing numerous sensors across the valley in numerous locations [3]. Also, a "scoping" experiment (in a controlled environment in the lab) was performed to determine what data might be expected to be seen under certain controlled and varied conditions. However, since the sensor is cheaper than those from the DAQ, the resulting outputs are a little skewed from what the Department of Chemical Engineering expects to see in the sensors versus "actual" data. Given other measured variables such as wind speed, wind direction, relative humidity, concentrations of gasses such as CO or NOx, and more data categories, it is uncertain which factors affect the semiconductive sensor the most.

From this information, it is possible to deduce which variables might be interacting with harmful substances such as ozone; a highly sensitive gas species. Thus, a smarter way of analyzing the problem is needed in order to determine which variables are affecting the sensors the most. Apparently, when observing the data "by eye", it is somewhat apparent what variables might be affecting the data collected by the AirU sensors. Therefore, a machine learning method such as gradient descent could precisely show which variables are most important/unimportant in affecting ozone concentrations.

In addition to seeing which factors affect the air quality sensor the most, it is also uncertain what ozone levels to expect with a random set of values for each factor. Although data has been collected under varied conditions, the way the sensor will react in nature is still somewhat uncertain. In order to have a more predictable way of knowing how the sensor will react in nature, a regression can be performed with a predictable model that can be made using machine learning techniques. Using machine learning modules in Python such as TensorFlow, Keras, and Pandas, a neural net can be made to make a predictable model.

Experimental Evaluation

Data from both DAQ and AirU sensors have been collected for one of the locations. The first step in this process consisted of making sure the data is "true". This means identifying and eliminating certain data points that should not be part of the process. Data is collected hourly for the numerous categories in Table #1. At approximately 7:30am and 3:30pm each day, the sensor is heated to remove gases that have adsorbed to the sensor surface, and thus these datapoints should be removed. The reason for doing this, is because there are only so many "activation sites" available for gas particles to adhere to, and thus, they need to be removed in order to have accurate data that is collected during each experimentation period. After ensuring that data points are "true", some analysis can be performed using machine learning techniques.

In this experiment, there are numerous variables that could possibly be affecting the sensor, which are shown in the table below:

Categories measured and (potentially) affecting air quality
Wind Speed
Wind Direction
Temperature
SO2
Relative Humidity (RH)
PM2.5
O3
NOX
NO2
NO
CO

Table #1: List of variables measured that could potentially affect air quality.

The above categories contain data measured from the Department of Air Quality. Since the AirU sensors are mounted next to the DAQ sensors, the resistance measured on the AirU sensors can be compared to the data at each timestamp, and a regression can be performed to see which variables are affecting the O_3 concentration the most by returning a weighted vector of values corresponding to each variable category.

So far, I have performed a regression using TensorFlow in Python to analyze this data. In order to make the predictable model, a neural net with 3 layers, of 5, 4, and 3 neurons respectively was made to train on 90% of the data, and predict O_3 concentration for the remaining 10% with computed error. The first two layers with 5 and 4 neurons respectively use tf.nn.relu, one of the general layer settings for a neural net in TensorFlow. The last layer with 3 neurons uses a sigmoid which accounts for some error in predictability. Also, the small amount of neurons (5, 4, and 3) is used in order to reduce overfitting in the model. Although the method has been shown to be powerful in other experiments [4], the current regression that I have done has produced error of 0.0046ppm, which is ~3% of the maximum measured O_3 concentration. Although the percentage sounds okay, this is only on 17 predictions/tests. In order to ensure that the model is actually performing well, the neural net needs to be tested on more data points. The following graph shows an example of a regression performed with predictions thus far:

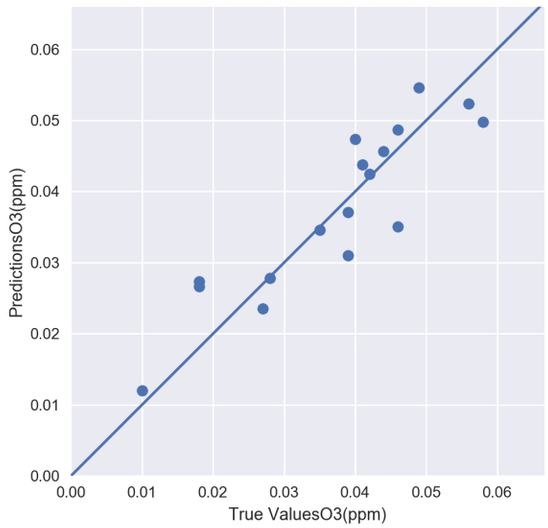

Figure #1: True values versus predictions of $O_3(ppm)$. Average prediction error is 0.0046ppm.

In the figure above, with limited data, the predictions are still somewhat accurate within a somewhat acceptable accuracy of 0.0046ppm. After knowing that the neural net that was made is somewhat capable of producing a predictable model, it also brings to question which factors may be affecting O_3 concentration the most. In order to find this answer, further machine learning techniques can be used. Using Tensorflow again, an additional neural net was made by using one sensor's data as training data, and another sensor's data as testing data. The model contained a Xavier activation function and a neural net with 3 layers, 20 neurons per layer, a dropout rate of 0.5 to prevent overfitting, and 1000 epochs. The resulting test predictions versus test labels are shown in the figure below:

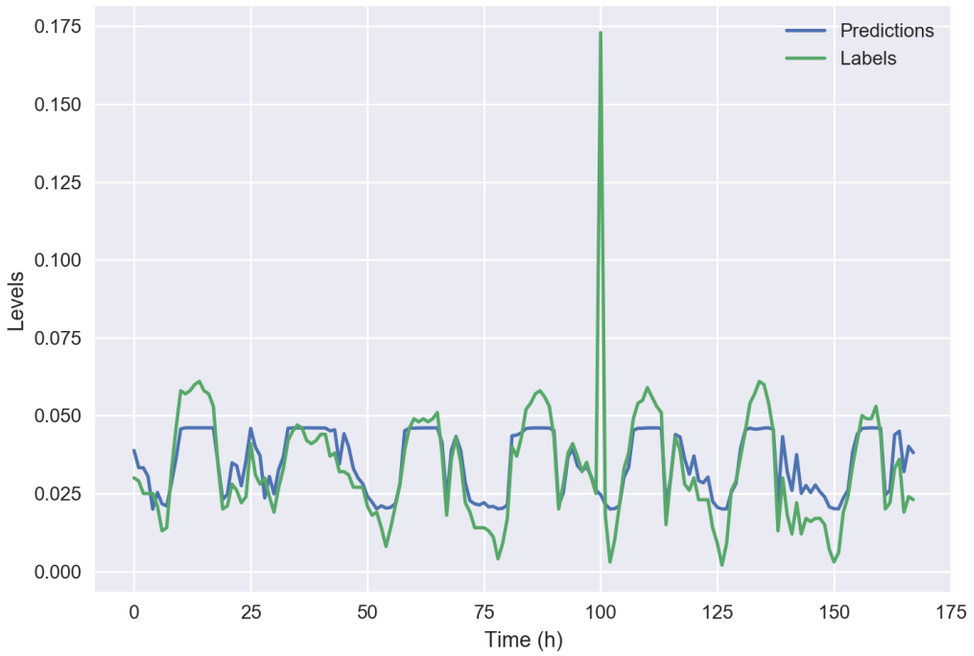

Figure 2: Test predictions vs. test labels of ozone levels (ppm) with error of 0.0074ppm.

In Figure 2 above, there are some minor discrepancies in test predictions vs. test labels, with testing error of 0.0074ppm. There is a random spike for the label around the 100h timestamp, which could be due to imprecision in measurement data. However, the predictions somewhat follow the same trend as the test labels. Perhaps with more data, these predictions could be more precise.

In addition to the neural net and regression that was performed, I also set up a gradient descent algorithm with a returned weighted vector to see which variables are affecting each sensor the most over each experiment period outside. Using modified code that was previously written for the CS5350 course on assignment #2 for gradient descent, a weighted vector was returned which corresponds to how much each variable is affecting O_3 concentration. After, a sorting algorithm was used to see which variables are affecting each of the three sensors at the Hawthorne location the most. The resulting variables are shown in Table 2 below:

Rank	HW137	HW103	HW016
1	Temp	Temp	Temp
2	Wind Dir.	Wind Dir.	Wind Sp.
3	Wind Sp.	Wind Sp.	Wind Dir.
4	CO	CO	CO
5	AirU	NO_2	AirU
6	NO_2	RH	RH
7	RH	PM2.5	NO_2
8	PM2.5	AirU	NOX
9	NOX	NOX	PM2.5
10	NO	NO	SO_2
11	SO_2	SO_2	NO

Table #2: Ranked categories of importance in affecting O_3 concentrations. 1 is the most-important, and 11 is the least important.

In Table #2, I was later asked to remove SO_2 from the dataset because some of the measured values were not correct. Therefore, this leaves 10 rankings with NO in position 10 for each category. It should also be noted that each sensor has a highest ranked variable of importance as Temperature. Scientifically, this makes sense for two immediate reasons: 1) temperature affects reactivity of substances, and 2) the sensor being heated is the method used for removing gases. It should also be noted, with great importance, that each sensor has different "most-important" variables that affect O_3 concentration. Another notable observation of this test, is that the "scoping" experiment performed by the AirU group showed temperature as one of the factors least affecting ozone levels [3]. This shows that nature reacts differently than the lab, and that this is a localized problem. This is why machine learning is the best technique to analyze this issue; in each area in the Salt Lake valley, there are different factors that are going to be more responsible than others in affecting air quality.

Future Plan

As discussed previously, a regression has been performed using a neural net, as well as error analysis of test predictions versus test labels with results that only show a somewhat predictable model. This is most-likely due to a small amount of training and testing data points that were used. As said before, only 168 examples were used for each sensor in testing and training total, and would not give a very predictable model on such little data. This was only done on one location over one experimentation period. In order to obtain a more predictable model, more data will have to be used in order to ensure the neural net's validity. It was also worthwhile to look in to which variables are affecting the semiconductive sensor the most. Using gradient descent, the returned weighted vector for the dataset based on the output values shows which variables are most improtant/unimportant in affecting air quality.

The continuing goal from here, would be to establish that the machine learning algorithms for both objectives 1) and 2) are accurate by collecting and testing on more data. Although the machine learning techniques have shown potential, further validating them will increase their worth so the sensors can be more widely used. With further experimentation and modeling, the data from this project should be further validated in the future. There are also ongoing parts of the AirU project that may include the mounting of sensors on drones to measure air

quality data. The machine learning models would be useful here in fitting to the correct levels of toxins in the air during experimentation periods.

References

1. Cox, Peter M., et al. "Acceleration of Global Warming Due to Carbon-Cycle Feedbacks in a Coupled Climate Model." Nature News, Nature Publishing Group, 9 Nov. 2000, www.nature.com/articles/35041539.

2. "Long-Term Winter Inversion Properties in a Mountain Valley of the Western United States and Implications on Air Quality*." The Abrupt Shift of Typhoon Activity in the Vicinity of Taiwan and Its Association with Western North Pacific–East Asian Climate Change: Journal of Climate: Vol 22, No 13, journals.ametsoc.org/doi/full/10.1175/JAMC-D-15-0172.1.

3. Garff, Alicia. "Investigation of Metal Oxide Air Quality Sensor For Measurement of Ozone." Dec. 2018.

4. Chollet, Francois. "Regression: Predict Fuel Efficiency." TensorFlow Tutorials, 2017.