Rhonald Reside

DS670 - Assignment 3

Contribution and State of the Art

**Contribution**

The data set I was assigned was the weather data. The source of the data set is from CityPulse. The website contains a collection of smart city data sets from Aarhus, Denmark. Aarhus is the second largest city in Denmark, located east coast of the Jutland peninsula. The population of Aarhus in 2016 was about 1.378 million. By utilizing smart city data collection of Aarhus, Data Scientist are able “find effective and sustainable solutions to the challenges faced by many cities today.”

I was first interested in working with the .tar format file. Since this is my first exposure to .tar, I was really interested in seeing its data structure and content. I first thought working with .tar would be a nice contribution to the study. I have never worked with .tar, nor ever heard of it. Working with it would help other students who come across .tar an understanding of how to tackle it and convert it to a format that can be used on other tools and software. At first I tried to research what programs or tools can read this .tar file. Unfortunately, I was not successful to open the .tar file. From my research, it seemed to be just a proprietary file only readable to certain machines. As much as I wanted to work with .tar, I had to abandon this format.

On the CityPulse Smart City website, the weather data set is available in two types of format. One format is in .tar format and the other format is in JSON format. The data is separated by two sets of dates. The first range of the weather data set is February 2014 to June of 2014. The second set of the weather data set is in August 2014 to September 2014.

I will use the JSON format and I plan to use logistic regression for my analysis. Statistically, the use of logistic regression to find the relationship between temperature, humidity and dew point. It will use binary logic to estimate the probability of a binary response based on our independent features, i.e humidity and dew point. I will use R Programming language to model the data and then fit it into the general linear model or glm() function. The library that has this is the ISLR library in R. I will use temperature as my Y-Variable, also known as my dependent variable. I label my variables as dew point and humidity; we would also know this as independent variables or our features.

**State-of-the-Art**

My competitor’s name is Mark F Lawrence. His article’s name is “The Relationship between Relative Humidity and the Dew point Temperature in the Mount Air.” His article focuses on a three things. First, he is interested in defining the relationship between Relative Humidity with Temperature, humidity and dew point. He discusses two formulas that he believes explain this relationship. In the first formula, he says relative humidity is the ratio of the actual water vapor pressure e to the equilibrium vapor pressure over a plane of water . It is the “saturation” vapor pressure. Here is his formula for this explanation:

RH =

Another way to understand this is it is the point at percentage of the air can hold before it cannot hold any more water and then condenses and in some places, turns into rain. This explains the sudden thunderstorms that occurs in Florida. It is that “real feel” weather you feel when standing in line at Disney in Florida where your shirt is full of perspiration in the summer months of June to August. That occurrence is relative humidity.

Lawrence also shows us another equation that describes the relationship between temperature, dew point, and pressure. He states it is the ratio of actual water vapor dry mass mixing ratio *w* to the equilibrium mixing ratio at the ambient temperature and pressure. He further explains it in the following formula:

RH =

He then explains the relationship between the variables through Linear regression for moist air. He states that relative humidity must be greater than 50 percent for it to become linear.

RH =

In response to his work, with the other variable I have available, i.e. wind direction, wind speed, and air pressure. I would be interested to know if these other variables have any kind of effect on relative humidity. As we work with the data, we will have error rates for each predictor and classification accuracy to help build a better picture of the relationship to temperature.