CS 174 Project Proposal - Air Quality Prediction

Hanli Li, Yue Song, Ziyang Wang

Problem The project is about air quality prediction. The ultimate goal of the project is to build a predictive model that is capable of accurately predicting dangerous levels of metropolitan air pollutants on an hourly basis. Basically, this is a supervised learning problem. At different time chunks of a year, levels of different pollutants at a couple of testing sites were recorded for 11 consecutive days on an hourly basis. Simultaneously, the weather condition, air pressure and temperature at that point were also recorded. What we need to do is to build a model, train it and predict the levels of these pollutants at a given time location point. The prediction results can be used by patients with respiratory diseases for their outdoor activities.

Method Thinking of prediction method, especially when response variables and predictor variables are given, the most common algorithm comes to mind is linear regression. In this project, there are 39 response variables, which means 39 linear regression models need to be trained. But on one hand, the feasibility seems to be vague. On the other hand, more importantly, there are correlations between pollutants types, and there are associations and patterns needed to be discovered. Besides, the linearity assumption may also fail to be satisfied. To model the dynamic and the potential nonlinear characteristic of the problem, a neural network needs to be trained. For a reference of neural network algorithm and applications it can be used for, please check `http://ufldl.stanford.edu/wiki/index.php/Neural_Networks`.

Data Dataset used here is provided by the Cook County, Illinois, local government. There are 210 time slices of 11 days pollutants measurements. Among these, first 8 days' measurements are used as training set, the following 3 days are used as testing set. For each day in each time slice, data of weather, air pressure, temperature, location were recorded from hour to hour. The corresponding level of pollutants at different sites were also recorded. The link to training set is: `http://www.kaggle.com/c/dsg-hackathon/data`.

Goal By milestone due date, we will have trained a neural network model with proper number of hidden layers. Prediction will be made on the testing set. We will use cross validation to tune the model parameters. The correlation between air pollutants will also be explored.