**Applied Machine Learning CSCI-5939, Summer 2017**

**Hurricane Path Prediction**

Viral Thakkar

University of Houston Clear Lake

Houston, USA

Email: [**thakkarv0626@UHCL.edu**](mailto:thakkarv0626@UHCL.edu)

Instructor

Dr. Ahmed Abukmail

Associate Professor of Computer Science

University of Houston – Clear Lake

Houston, USA

Email: Abukmail@uhcl.edu

**Introduction**

A hurricane is a storm that occurs in the Atlantic Ocean and the Pacific Ocean, their effect on human is often devastating. Every year during the season of the hurricane, many places are affected by it. But, if we somehow predict the hurricane that may occur in the specific region from the previous data, then we can take necessary steps. NOAA (National Oceanic and Atmospheric Administration) is a government agency which focuses on the conditions of the oceans and the atmosphere. National Hurricane Center which is part of NOAA conducts a post-storm analysis of each tropical cyclone in its area of responsibility to determine the official assessment of the cyclone’s history. This Analysis makes use of all available observations, including those that may not have been available in real time. This report is known as HURDAT database. This database contains all storm that appeared in 1851-2016. Each storm has data points from start to end in latitude and longitude format. It has many more important results related to storm and location. That’s means if we plot this points in the graph will give you the path for the particular storm. We are trying to predict storm path if we found in any suspecting activity in the ocean and our program will give a list of points where the storm can occur. So, the government takes some necessary steps to nearer that area.

**Problem Definition**

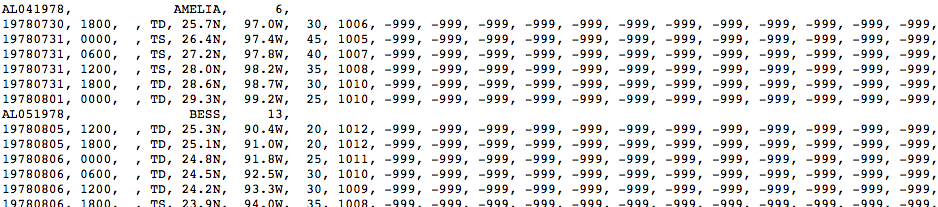
Each Tropical Storm contains start to end points. That’s means data set has each location points where it affected.

Let’s say Tropical Strom named PQR started from (X1, Y1) and then it affected on (X2, Y2), (X3 Y3), (X4, Y4), etc. It went up to (X49 Y49). That means total 49 locations are affected due to PQR storm. NOAA has more than 1000s storms data.

Every time storm starts from any location and It affects to any random number of locations. Now, weather conditions can give the idea what is going to happen in particular location. But, it will not predict if a storm starts in location, where it will go next. So, our goal is to use all storm occurred in past to predict the path of a future storm if we find any unusual activity in ocean related to the storm.

**Data Exploration**

NOAA provides a data set of the hurricane into two different part: The Atlantic Ocean and the Pacific Ocean. They have provided this dataset publicly which is known as HURDAT2. Dataset is in normal text file Which has the following format.



**Fig.1** **Atlantic Ocean Data**

We can see that there are two types of lines are available in the dataset. Let’s get deeper into both types of lines.

**First Line**

AL041978, AMELIA, 6,

AL – Ocean Name (Atlantic in this Case)

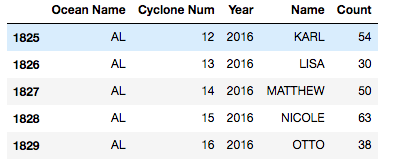
04 – Cyclone Number for particular year

1978 – Year

AMELIA – Name of the Cyclone

6 – Number of Rows Follow

So, the first line will give an idea to follow the next number of lines for a particular storm. So, here as per example storm named **AMELIA** and after that, it is showing **6**. From that, we can interpret into following way: **AMELIA** named cyclone which occurs in the Atlantic Ocean in 1978 that affected 6 locations and which was the 4th Cyclone of that year. **NOAA** has this type of data set for more than 1000s storms.



**Fig.2 Number of locations affected in each storm. Total number of storms are 1829. Actually, it**

**is 1830 because index starts from 0**.

**Second Line**

From above observation, we need to follow next 6 lines for that cyclone. Each line has the following type of information: Year, Month, Day, Hours, Minutes, Type of Storm, Latitude, Longitude, Wind Speed, Water Pressure. But, our goal to predict the path of the storm. For that, we need only latitude and longitude data. So, we removed other data from a dataset and converted each storm data into single row data.



**Fig.3 Converted each storm data into row wise**

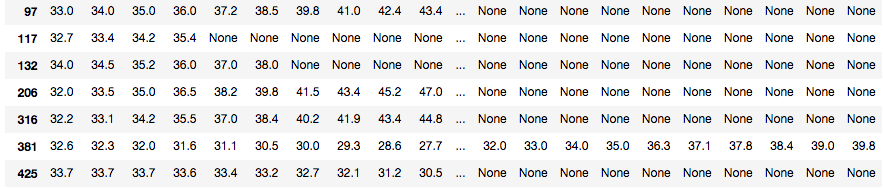
We are predicting longitude and longitude data separately. So, we will have two types of new data set as we mention in above screenshot: One for Longitude and One for Latitude. Now, here is one question why do we have so many columns? Because a maximum number of locations has been affected by the single storm is 133 and minimum is 1. So, each storm has a different count. If a storm doesn’t have next location in a dataset, then it will have none value.

**Algorithm**

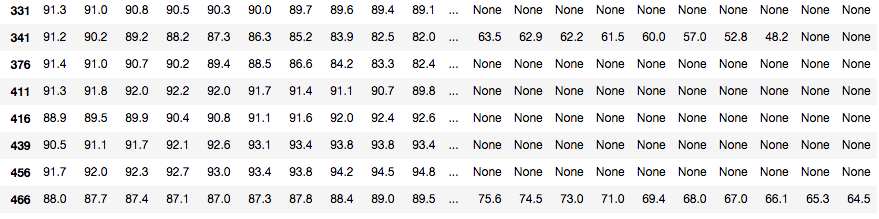
Predicting the path from the previous dataset, we can clearly give the answer that it is problem of Linear Regression. We have used linear regression in some different way. We have used a concept of rolling regression.

So, let’s say if storm just started at the following location (34.1, 90.5).

Based on the input we fetched nearer data. Suppose here latitude is 34.1 and longitude is 90.5. So, we have fetched data of storm which latitude is between 32 to 36 and which longitude is between 88 to 92. Here, our **scaling factor is 2**. You can take any factor like 0.5,1 or 1.5 It depends on how accurate prediction you want and how many locations you want. Depending on the factor you will have a number of locations which are either more or less.



**Fig.4 Latitude Sample Data**



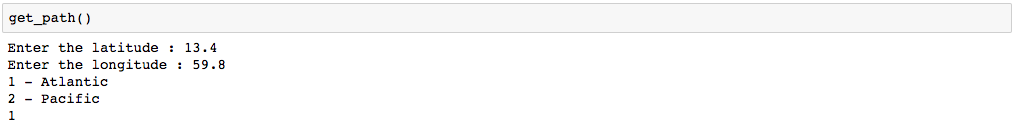
**Fig.5 Longitude Sample Data**

Let’s say we have 50 columns in above figure. Let’s label it as X1, X2, X3, … X50. As we said that we are using rolling regression. First, we take the input P0 and will make X1 as X-Label and X2 as Y-Label. It will give predicated value P1. Now we will make X2 as X-Label and X3 as Y-Label and will get predicted value P2. We will do this procedure so on for (N -1), where N is a number of columns. It will give us P0, P1, P2, …P49. We have to do the same procedure for Latitude and Longitude Data.

Let’s Label Latitude data as Lt1, Lt2, ... Lt49 and Longitude data as Lo1, Lo2, … Lo49. Now, we will combine this data into (Lt1, Lo1), (Lt2, Lo2), (Lt3, Lo3), … (Lt49, Lo49). Now, you can plot this coordinates in a graph.

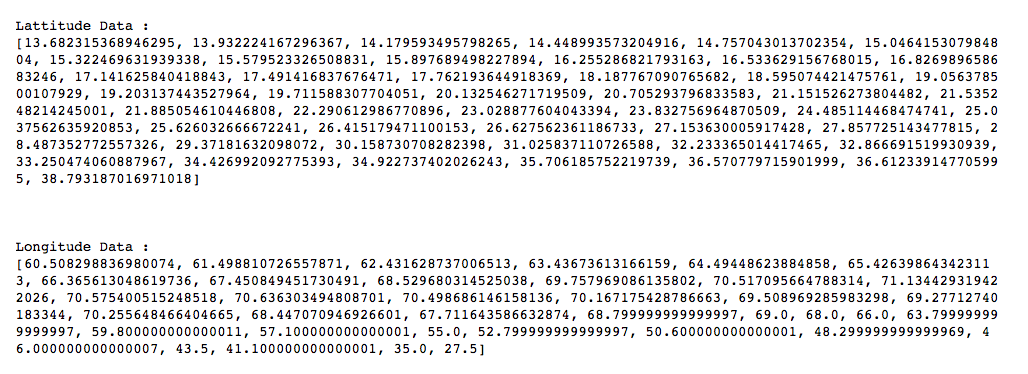
**Results**

You have to give storm starting location input to program in terms of longitude and latitude. You also have to specify for which ocean you want to predict path: The Atlantic and The Pacific.



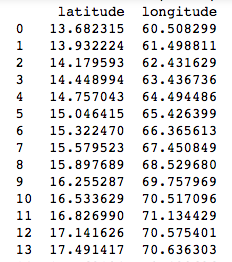
**Fig.6 User Screen for inputting latitude and longitude data.**

It will predict the Latitude and Longitude data separately. Following screenshot is the result of above input.

****

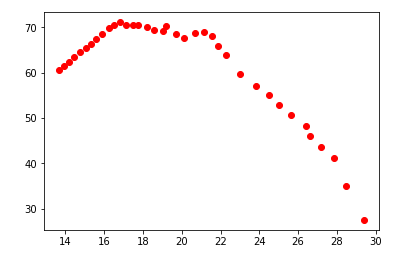
**Fig.7 Predicted Result**

Now, let’s combine the data.

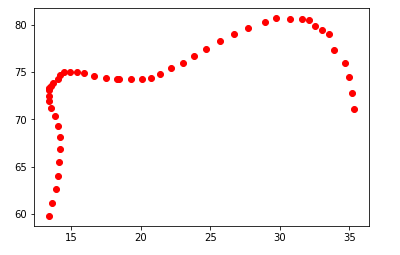


**Fig.8 Combine Result**

Let’s plot above data set and also plot if hurricane is already started at same location previously.



**Fig.9 Predicted Graph**



**Fig.10 Original graph of Dataset (13.4,59.8)**

**Tools**

Program is mainly written in python. But, for different functionalities we have used different computation libraries. We have used following libraries of python.

* **NumPy** – It supports for large and Multi-dimensional arrays and matrices. It also supports high-level mathematical functions. We have used to create multi-dimensional array of latitude and longitude data.
* **Pandas** – It is providing similar kinds of functionality like SQL. It is powerful library for Data Analysis. It is providing fast, flexible, and expressive data structure designed to make working with “relational” or “labeled” data both easy and intuitive. We have used Pandas for following purpose:
  + Handle missing Data
  + Slicing, Sub-setting, and Indexing Data
  + Merging and Joining different sliced data
  + Reshaping the datasets
* **Scikit-Learn –** It is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy, Pandas, and SciPy. We have used Linear Regression Library to predict each value based on our input.
* **Mataplotlib –** This library is extension of NumPy library. It is useful to draw plots. We have used to plot Latitude and Longitude Data.

**Future Work**

Currently, we are predicting path based on the Latitude and Longitude Data. We can consider more feature to predict path based on the direction of wind and water pressure.

**References**

[1] <https://stats.stackexchange.com/questions/207533/the-right-way-to-use-machine-learning-to-predict-latitude-and-longitude>

[2] <http://www.nhc.noaa.gov/modelsummary.shtml>

[3] <http://www.aoml.noaa.gov/hrd/hurdat/Data_Storm.html>

[4] <https://docs.scipy.org/doc/numpy-dev/user/quickstart.html>

[5] <http://pandas.pydata.org/pandas-docs/stable/>

[6] <http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html>

[7] <https://matplotlib.org/users/pyplot_tutorial.html>