Udacity Machine Learning Engineer - Capstone proposal

A. Domain Background

As of March 2020, it seems clear that epidemic will be a growing concern worldwide.

However other epidemic than COVID-19 (technically a pandemic) have occurred for centuries, and many are still occurring today. Although epidemics such as COVID-19 are next to impossible to predict because the causality is similar to the butterfly effect, other epidemics have been ongoing for years with seasonality and others have been linked to climate changes.

This is thought to be the case for mosquito-transmitted diseases as explained in [1] and [2], although the relationship between the two is complex: > climate variable may increase dengue transmission potential through one aspect of the system while simultaneously decreasing transmission potential through another. [2]

In this project I will take part in a DrivenData competition: DengAI: Predicting Disease Spread where the goal is: > to predict the number of dengue cases each week (in each location) based on environmental variables describing changes in temperature, precipitation, vegetation, and more.

Dengue disease comes from dengues viruses transmitted by several types of mosquitoes that are also known to transmit other diseases such as Zika and chikungunya. According to [3]: > * Dengue is common in more than 100 countries > * Forty percent of the world's population, about 3 billion people, live in areas with a risk of dengue > * Each year, up to 400 million people get infected with dengue. Approximately 100 million people get sick from infection, and 22,000 die from severe dengue.

A quick look on Wikipedia and we can see that 15 of the 63 epidemics listed for the 21st century are dengue related, and there was a recent dengue fever outbreak in 2019-2020 [5], with 2019 being a record year for Latin America > with more than 2.7 million cases and 1206 deaths during the first 10 months of 2019.

B. Problem Statement

The goal will be to predict the number of dengue cases for two cities, San Juan and Iquitos, per week over several years.

More precisely, the test cases (and thus prediction span) will be:

- 2010 to 2013 for Iquitos
- 2008 to 2013 for San Juan

These predictions will be made based on precedent year information (features and labels) as well as test sets features.

The test set is non-overlapping with the training set.

C. Datasets and Inputs

The datasets are those provided by DrivenData.

a. Labels

The labels will be the number of dengue cases per week for each city.

b. Features

There are 24 features based on weather information and vegetation for each city and week, gather from the National Oceanic and Atmoshperic Administration [6].

NOAA's GHCN daily climate data weather station measurements

- `station_max_temp_c` – Maximum temperature
- `station_min_temp_c` – Minimum temperature
- `station_avg_temp_c` – Average temperature
- `station_precip_mm` – Total precipitation

- `station_diur_temp_rng_c` – Diurnal temperature range

PERSIANN satellite precipitation measurements (0.25x0.25 degree scale)

- `precipitation_amt_mm` – Total precipitation

NOAA's NCEP Climate Forecast System Reanalysis measurements (0.5x0.5 degree scale)

- `reanalysis_sat_precip_amt_mm` – Total precipitation
- `reanalysis_dew_point_temp_k` – Mean dew point temperature
- `reanalysis_air_temp_k` – Mean air temperature
- `reanalysis_relative_humidity_percent` – Mean relative humidity
- `reanalysis_specific_humidity_g_per_kg` – Mean specific humidity
- `reanalysis_precip_amt_kg_per_m2` – Total precipitation
- `reanalysis_max_air_temp_k` – Maximum air temperature
- `reanalysis_min_air_temp_k` – Minimum air temperature
- `reanalysis_avg_temp_k` – Average air temperature
- `reanalysis_tdtr_k` – Diurnal temperature range

Satellite vegetation - Normalized difference vegetation index (NDVI) - NOAA's CDR Normalized Difference Vegetation Index (0.5x0.5 degree scale) measurements

- `ndvi_se` – Pixel southeast of city centroid
- `ndvi_sw` – Pixel southwest of city centroid
- `ndvi_ne` – Pixel northeast of city centroid
- `ndvi_nw` – Pixel northwest of city centroid

c. Spatiality and temporality

The data is indexex by year, week of the year and city.

c. Training/Test

The training data set spans from: * 2000 to 2010 for Iquitos * 1990 to 2008 for San Juan

The test data set spans from * 2010 to 2013 for Iquitos * 2008 to 2013 for San Juan

D. Solution Statement

This problem will be solved by training a Recurrent Neural Network to predict the number of dengue cases per week based on this week's features and prior week information.

E. Benchmark Model

The people at DrivenData has conveniently proposed a benchmark model using a Negative Binomial model with a final Mean Absolute Error of 6.47, and a score of submission of 25.8173.

F. Evaluation Metrics

We will use the evaluation of this competition to evaluate our model against our test set which is the Mean Absolute Error [7].

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

We will also report the score of the submission which is done on a private test set.

G. Project Design

We will follow a classical data science approach: 1. Gather all data 2. Exploratory Analysis to gain first insights on possible feature design 3. Feature design 4. Model design and training 5. Evaluation

I will first focus on having a working solution to be able to submit my work on DataDriven platform., and from here improve the model.

I will focus on Recurrent Neural Network, firstly to improve my understanding and experience on this, and also because it seems to be one of the of state-of-the-art solution to address time series prediction based on external features.

Resources

- [1]: Lindsay P. Campbell, Caylor Luther, David Moo-Llanes, Janine M. Ramsey, Rogelio Danis-Lozano and A. Townsend Peterson. 2015. Climate change influences on global distributions of dengue and chikungunya virus vectors. Phil. Trans. R. Soc. B370: 20140135. https://doi.org/10.1098/rstb.2014.0135
- [2]: Morin CW, Comrie AC, Ernst KC. 2013. Climate and dengue transmission: evidence and implications. Environ Health Perspect 121:1264–1272. https://doi.org/10.1289/ehp.1306556
- [3]: Centers for Disease Control and Prevention, National Center for Emerging and Zoonotic Infectious Diseases (NCEZID), Division of Vector-Borne Diseases (DVBD) https://www.cdc.gov/dengue/about/index.html
- [4]: List of epidemics https://en.wikipedia.org/wiki/List_of_epidemics#21st_century
- [5]: 2019–2020 dengue fever epidemic https://en.wikipedia.org/wiki/2019%E2%80%932020_dengue_fever_
- [6]: National Oceanic and Atmoshperic Administration https://www.noaa.gov/
- [7]: Mean absolute error https://en.wikipedia.org/wiki/Mean_absolute_error