Machine Learning - CSL 508

Assignment on Covid-19 Trend Prediction

Submitted by:
Abdul Sattar Mapara (BT16CSE053)
Saket Chopade (BT16CSE018)

Guided by:
Dr. R.B. Keskar Sir

Last Updated on: 15th April, 2020

Problem Statement

Predict the number of Covid-19 cases in India per day for the next fortnight considering the current trend and a number of other parameters that include lockdown strictness, weather conditions and density of population.

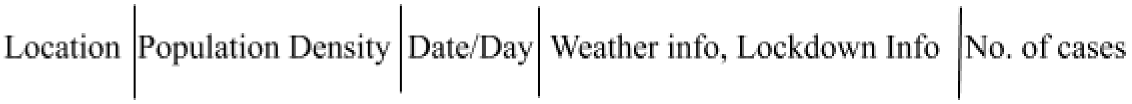
Approach and Design

Using the **time-series data of various countries** that include China (where the pandemic curve has become almost flat), Italy, USA, India from the day the first confirmed case was reported till a certain date, we plan to predict the number of cases in India per day using along with the current trend, the **following parameters**:

1. **Lockdown -** [none, partial (only in areas with large number of cases), full] + [less strict, very strict]
2. **Weather conditions** - minimum, maximum temperature of the day
3. **Population Density** of the location

Data Format

Data will be prepared in the following format-



Location	Population Density	Date/Day	Weather info, Lockdown Info	No. of cases

- Data from multiple locations is considered to allow the machine learning model to learn about the trend of the spread. For example, in China, after strict lockdown the pandemic curve has flattened. So, the model can learn to predict that if the lockdown is strict in India, then there is a good chance that the curve will flatten. Similarly, the data from other countries helps the model to learn about the future predictions (based on other parameters too)
- Number of tests conducted per day is removed as a parameter because -
    - It is difficult to get the number of tests conducted for each day for every country
    - The data available was not from end of January onwards for every country

- Moreover, for some countries the data was not at all available

Details of the design

Technique to be used -
The problem is a multivariate time-series forecasting/prediction problem.
We decide to use **Long Short Term Memory** (LSTM) Networks for the following reasons -

- **Inductive bias**: It is biased towards preserving information over long sequences
- The data is time series based which requires memorization. Since, the Artificial Neural Networks do not have memory associated with them, they are not suitable for this particular problem.
- The inductive bias for the naive bayes algorithm - Each input depends only on the output class or label; the inputs are independent from each other, which is not completely true in this case as location, population density and number of tests are dependent on each other. Also, the number of tests conducted and weather conditions are dependent on the location considered.

Input to the model -
For each location/country and for each day from the date the first confirmed case was reported till a certain date -

- Population density of the location
- Weather information (min./max. temperature averaged over the area, etc.) for each day
- Lockdown information for each day (to be represented by a numeric value)- Whether the lockdown has been imposed on the particular day, and if imposed, how strict it is
- Number of Covid-19 cases reported for each day

Output to be given by the model -

- Prediction of the number of cases per day for the next 15 days

Libraries for implementation -
We will be using the Python programming language and hence for implementing the machine learning model, we plan to use keras library, which is a high-level neural networks API, capable of running on top of TensorFlow.
Along with keras, pandas library (for pre-processing and manipulating data), matplotlib (for plotting the data), scikit-learn, numpy are used for implementation.

Data Sources -

1. For time series data of India, the data was obtained from the API - https://api.covid19india.org/data.json
2. For time series data of other countries, the data was obtained from https://github.com/ulklc/covid19-timeseries/tree/master/countryReport/country
3. The weather data was collected from the API - https://www.worldweatheronline.com/developer/api/
4. Lockdown information obtained from news articles

References

1. https://machinelearningmastery.com/multivariate-time-series-forecasting-lstms-keras/
2. https://machinelearningmastery.com/time-series-prediction-lstm-recurrent-neural-networks-python-keras/
3. https://keras.io/getting-started/sequential-model-guide/
4. https://github.com/keras-team/keras/issues/4446
5. https://machinelearningmastery.com/gentle-introduction-long-short-term-memory-networks-experts/