COVID-19 new cases prediction

097209 Machine Learning Project Report

Ben Nadler
Oded Salton

Abstract

Since December 2019 we are facing a global pandemic. The rate of infection and death caused by the virus worried government and citizens globally. We know today that those countries who took this threat seriously, evaluated the situation, and responded quickly were the ones succeeded the most in slowing down the infection rate of the virus and death rate of the population. Data has a crucial part in the fight against COVID-19. It is collected on a daily basis from most countries in the world and used to analyse and predict how the virus is spreading.
In this paper we will attempt to predict the daily status of new cases in terms of new cases per million people as a classification problem, using structured and unstructured machine and deep learning models and algorithms.
We will evaluate our models performance using numerous classification metrics and will try to optimize our models further and add variations to help us improve our performance.

Introduction

The question of what will be the status of new cases in a day can be complex, it can be affected by a large number of different variables which present how a country, which consist of the government intervention steps, the cooperation and willingness of the public to comply and their actions on daily basis. We believe that models who can perform well and predict daily cases with sufficient accuracy can help countries globally on this ongoing struggle to stop the infections.
Our main goal is to analyse the ability of machine and deep learning algorithms to predict the status of new cases per million people each day. We decided to represent the status of new cases per million people in a day by labelling our target "new cases per million" with the values {0,1,2,3,4}:

- 0 = no new cases today (resulting in no new cases per million people)
- 1 = 1-10 new cases per million people today
- 2 = 11-100 new cases per million people today
- 3 = 101-250 new cases per million people today
- 4 = more than 250 new cases per million today

The number of new cases for United States of America is far greater than the number of new cases for Israel for example and is proportional to the population of the country. Thus, we found the label "new cases per million" as a good indication for the situation in a country based on COVID-19 cases.
We assumed that the combination of neural nets and representing the data in sequences while using a structure will improve performance on our task. Furthermore, we wanted to experiment on how different sequence lengths can affect our predictions, under the assumption that the behaviour of the virus in a current day is affected by a week or two weeks that came before it. Finally, we wanted to find a good mitigation strategy to reduce the number of cases per day. Eventually, those assumptions yielded the best results on our BiLSTM/BiLSTM-CRF and we were able to find a winning strategy based on graph theory to reduce the number of cases per day by putting individuals with the highest degree in the graph on quarantine.

Models and Algorithms

For our unstructured approach, we implemented Logistic Regression, SVM and Random Forest learning algorithms and then compared their performances based on classification metrics.
For our structured approach we decided to implement an BiLSTM, following by a BiLSTM-CRF [2] which is a variation of BiLSTM with the help of a CRF layer, using Viterbi as our inference algorithm.

Long-Short-Term-Memory (LSTM) Networks are widely used in many sequence tagging tasks.
Our model will receive a sequence of days represented by feature vectors from our dataset which can be compared to the embedding of a word in an NLP task.
It is worth to note that while we use the output of our BiLSTM Network to predict the sequence of "new cases per million" labels, in our BiLSTM-CRF it is used only to learn a better representation of the data and then feed our CRF with it's features. Thus, the purpose of the BiLSTM layer in BiLSTM-CRF is only to feed features to CRF.

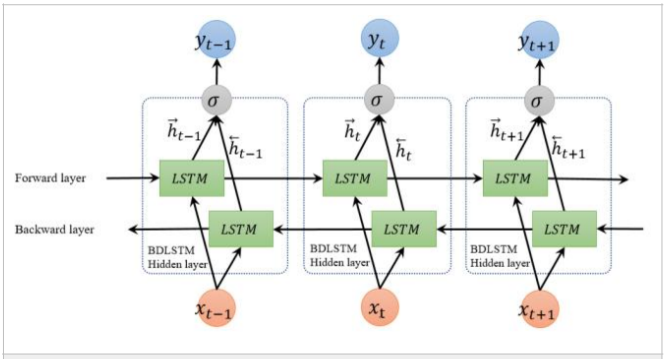


Figure 1. BiLSTM Architecture

For the BiLSTM model, after feeding it with a sequence of days features we then use a single linear layer to transform the out of the model into scores for each class {0,1,2,3,4}.

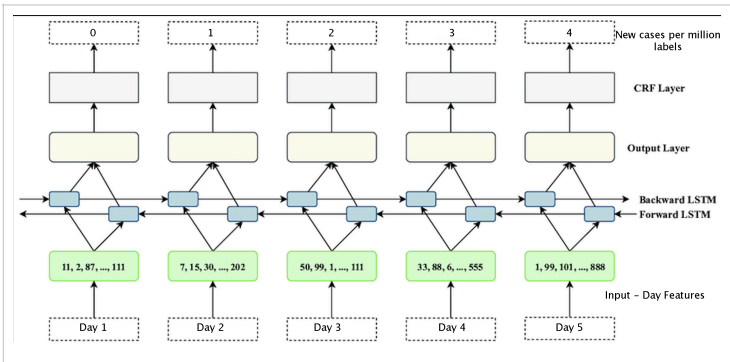


Figure 2. BiLSTM-CRF Architecture

For our advanced part we decided to use our BiLSTM as a feature feeding layer and use a CRF layer to improve our performance. The combination of both is thought to be strong on sequence tagging tasks.
We thought that the capabilities of BiLSTM to refine our data and keep the relevant parts of it will result in a better feature representation to the CRF layer than our dataset features could achieve alone.
The scores which we mentioned before, after using a linear layer to transform the BiLSTM output are known as emission scores, which represent the likelihood of the day being a certain class.
A CRF calculates not only the emission scores but also the transition scores, which are the likelihood of a day being a certain class considering the previous day was a certain class. Therefore, the transition scores measure how likely it is to transition from one class to another.
The last part is the Viterbi decoding, this is a way to construct the most optimal class sequence while considering not only the likelihood of a class at a certain day (emission scores), but also the likelihood of a class considering the previous and next classes (transition scores).
For the creative part we chose to use the SEIR (Susceptible-Exposed-Infected-Recovered) model on which we will elaborate more later.

Experiments and Data

We chose to construct our dataset based on multiple data tables we gathered from ourworldindata.org/coronavirus. We took data from March to July as the training data and from then on we treated the data of August as test data. Furthermore, we decided that taking all countries in the data source might not be beneficial to us due to the fact that some countries are deemed as unreliable in terms of the data they publish such as China or Russia. As a result, we chose 11 countries which we thought can represent the virus spread behaviour and actions taken against it globally and generalize our model.

The countries we chose were: Israel, United States of America, Italy, Germany, Mexico, Brazil, United Arab Emirates, Japan, Republic of Korea, Saudi Arabia and India.

The features in which we chose to use for our models are from the following list:

Daily Deaths (int numeric)	Cumulative Deaths (int numeric)	Population in millions (float)	Daily Deaths per million (float)
International travel policy (int categorical)	Testing Policy (int categorical)	Workplace closure policy (int categorical)	Workplace visitor change (%)
Public event cancellation policy (int categorical)	School closure policy (int categorical)	Staying Home Instructions (int categorical)	Transportation Closure (int categorical)
Public gathering policy (int categorical)	Campaigns policy (int categorical)	Contact Tracing Efforts (int categorical)	Covid Relief (int categorical)
Internal Movement (int categorical)	Retail & Recreation (%)	Grocery & Pharmacy Stores (%)	Residential (%)
Transit Stations (%)	Parks visitor change (%)	Workplaces attendance change (%)	

For the evaluation of our models and algorithms we chose the following classification metrics:

- Accuracy - Percentage of correct labelling
- F1 - Weighted average of the precision and recall metrics, which contribute equally
- Precision - Ratio between true classified days with label i and number of days which were classified as i, where $i \in \{0,1,2,3,4\}$
- Recall - Ratio between the correct classified days with label i and number of days which are true labeled as i, where $i \in \{0,1,2,3,4\}$

In order to evaluate our learning models performance, we ran our algorithms on our dataset. Our target is “new cases per million” and is self explanatory, we received the following results:

Classifier	Accuracy	F1	Precision	Recall	# of iterations
Logistic Regression	0.443	0.441	0.512	0.513	1
SVM	0.463	0.468	0.439	0.474	1
Random Forest	0.538	0.517	0.486	0.536	1
BiLSTM	0.598	0.612	0.682	0.598	50
BiLSTM-CRF-7 day	0.643	0.651	**0.859**	0.643	100
BiLSTM-CRF-14 day	**0.664**	**0.669**	0.769	**0.664**	50

We can see that the unstructured Random Forest classifier outperforms all other unstructured algorithms in Accuracy, F1 and Recall. Logistic Regression achieved the highest Precision score out of all unstructured algorithms. These results indicate that those classifiers have a hard time predicting our classes correctly, possibly because the data is not lineary separable.

Classifier	input_dim	hidden_dim	Optimizer	Learning Rate	LSTM Layers	output_dim	batch_size
BiLSTM	24	16	SGD	0.01	1	5	1
BiLSTM-CRF	24	10	SGD	0.01	1	5	1

As for the structured algorithms we can clearly see that they outperforms by quite a large margin from their structured counterparts, showing the power neural networks and structured approach possess in sequence classification tasks.
We can see that the addition of the CRF layer to our BiLSTM model did improve our performance. [5]
Due to the fact that we used data based on time, we chose 133 day samples through March to July as train and 21 day samples at the end of July to August as test for each of our 11 countries, 1463 train samples and 231 test samples.
To further expand our data we chose to implement a sliding window of 1 day where we look at a certain day and the 6 days before it thus creating a sequence of 7 days (13 days before for 14-day sequences). Using these sequences we were able to expand our data.
For our structured models we had to choose the sequence length, we tried a sequence length of 7 days or 14 days and eventually chose the 14 day sequence as it gave the better results with the exception of precision which achieved a better score for the 7 day sequence.

Creative Part

While our learning models could predict cases they cannot offer strategies to reduct them. In this part we were trying to prove a concept which says that if we have the relation matrix in the population we can optimize the mitigation of the virus spread by converting the government policy into graph theory mitigation strategy and check what is the most efficient one. [6][7][8]

First, we used two models to generate the community graph, each family represented by clique and the relations between individuals that are not in the same clique is represented by either work or friend relation type, we assume that these relations distribute according to power law distribution. [1(chapter:1.3)]

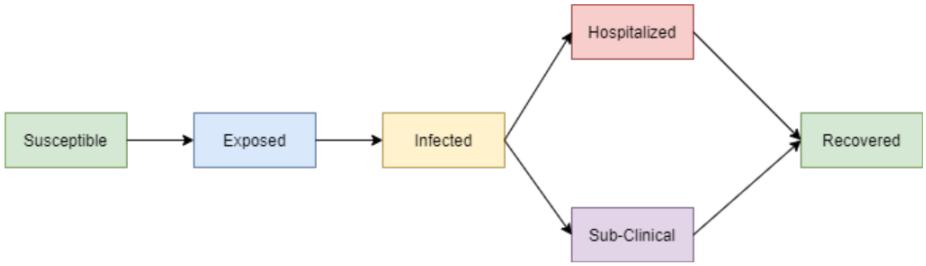
Secondly the problem was modelled as stochastic markovian S,E,I,R model according to the (DTM- Dynamic Transmission model) [3] where our set of states are $\{S, E, I, H, Sc, R\}$ presented in fig 3.
We choose the following assumptions:

1. Exposed means the incubation time of the disease, we assume that that during the individual is not infectious during that time (4 days) and then it moves to state I.
2. After two days that the individual is infectious, we assume that the healthcare system confirms that he is sick and then he is either hospitalized or subclinical and in both he is in quarantine and does not expose more individuals to the virus (moves from state I to Sc or H).
3. After 14 days in the hospital or 7 days in quarantine we assume that the person is recovered (or dies) which is the same outcome to the network since it is immune and can't get exposed to the virus again.
4. $R_0 \sim N(3,1)$ with no interventions according to [3] (figure S2) from the daily R_0 we derive the probabilities of each relation to expose individuals in status S to the virus (state S to E).

5. We assume that $P_{family} = 3P_{work} = 6P_{friends}$ according to [1] chapter 2.1(the odds of exposure of individuals to the virus depending on their relations).
6. We divide the individuals into ages where 50% of population is young - 0-24, 30% adults - 35-55, 10% mid age 55-65 and 10% elders above 65 and we set probabilities for getting to hospital according to the age [4] (Israel age structure).

Finally, our community size is of 35343 individuals(nodes) and 225357 internal relations (edges).
We ran 88 simulations where each time we randomly infected 5 individuals and each simulation represents a time frame of 75 days. Each exposure iteration took one day.

Mitigation Strategies



In order to try to simulate the effects of the mitigation strategies we have decided that as a threshold we can remove only 25000 edges from our network (11%) . We have simulated 4 different strategies and tried to see what is the effect on the virus spread. The mitigation strategies are:
Quarantine (reduce node degree to 0) of all the elders, quarantine of the most connected individuals(highest degree),quarantine the individuals with the lowest clustering coefficient score ,remove edges with the lowest neighborhood overlap score.

Results (Mean of individuals and 90% confidence interval for the mean in braces):

	Exposed	Infected	Hospitalize	subclinical	recovered
No intervention	951 (712-1191)	479 (300-619)	288 (211,361)	1240 (971-1569)	4593 (2819-5906)
Quarantine elders	713 (531-912)	372 (175-539)	161 (110-216)	1014 (730-1296)	3982 (1611-5951)
Quarantine most connected	**89 (24-138)**	**38 (11-52)**	**22 (8-31)**	**100 (30-162)**	**515 (239-803)**
Quarantine C.C	599 (378-789)	294 (153-401)	156 (84-236)	707 (427-1023)	2396 (914-3677)
Quarantine N.O	715 (533-917)	349 (197-513)	215 (289-157)	936 (728-1223)	3765 (1863-5591)

Conclusions

- We faced a difficult learning objective, with lack of data in general and reliable data in particular due to the fact that the pandemic began only 10 months ago. More reliable data is needed.
- Random Forest proved to be the best unstructured model on most metrics by quite a large margin.
- Neural Networks are powerful and with LSTM can greatly improve performance for sequence labelling tasks.
- The assumption that we can rely on a sequence of 14 days provided the best results on most metrics followed by 7 days, with a large margin from the unstructured models.
- We showed that with a network which models a community with relations between individuals we can reduce up to 90% of the number of cases by removing 11% of the graph relations with a winning strategy of quarantine for most connected individuals.

References

[1] https://citalid.com/blog/covid-19-containing-the-epidemic/#conclusion

[2] Z. Huang, W. Xu, et al. “Bidirectional LSTM-CRF Models for Sequence Tagging”. *arXiv:1508.01991v1 [cs.CL] (9 Aug 2015)*

[3] Davies, Nicholas G., et al. "Effects of non-pharmaceutical interventions on COVID-19 cases, deaths, and demand for hospital services in the UK: a modelling study." *The Lancet Public Health* (2020)

[4] https://www.indexmundi.com/israel/age_structure.html

[5] N. Reimers, I. Gurevych. “Optimal Hyperparameters for Deep LSTM-Networks for Sequence Labeling Tasks”. *arXiv:1707.06799v2 [cs.CL] (16 Aug 2017)*

[6] Ventresca, Mario, and Dionne Aleman. "Evaluation of strategies to mitigate contagion spread using social network characteristics." *Social Networks* 35.1 (2013): 75-88

[7] Halloran, M. Elizabeth, et al. "Modeling targeted layered containment of an influenza pandemic in the United States." *Proceedings of the National Academy of Sciences* 105.12 (2008): 4639-4644

[8] Easley, D., & Kleinberg, J. (2010). *Networks, crowds, and markets* (Vol. 8). Cambridge: Cambridge university press