Departamento de Informática

Relatório Modelação do Contágio/Propagação do Vírus SARS-CoV-2

André Ribeiro Martins, m10157
António José Marques Abreu, e10528

Maio 2020

Conteúdo

1 Introdução **2**

2 Tratamento dos Dados **2**

3 Visualização dos Dados **2**

4 Modelação e Previsão dos Dados **4**

5 Conclusão **7**

1 Introdução

Dada a situação mundial atual devido à COVID-19, achámos oportuno criar um modelo para modelar o contágio/propagação do mesmo, com recurso a algoritmos de Inteligência Artificial. O objetivo do modelo, é não só a modelação por aproximação dos dados já existentes, mas também a previsão de casos futuros, nomeadamente a previsão a sete dias. Iremos usar ferramentas gráficas para facilitar a compreensão dos resultados e dos processos realizados para obter os mesmos.

2 Tratamento dos Dados

Os dados que foram utilizados para fazer parte do *Time Series Covid19 Dataset*, disponibilizado pela *Johns Hopkins University Center for Systems Science and Engineering (JHU CCSE)*, dados estes, que se encontram na plataforma *The Humanitarian Data Exchange*. Os dados são atualizados diáriamente por volta das 9:00am (GMT).

Como os dados vêm separados em três subconjuntos (time_series_covid19_confirmed_global.csv, time_series_covid19_deaths_global.csv, time_series_covid19_recovered_global.csv) e de forma a permitir uma melhor manipulação dos mesmos em etapas futuras, criámos um *sript* (clean_data.py), que junta a informação dos três subconjuntos de dados num só (covid_19_clean_complete.csv). Como o conjunto de dados faz a divisão dos casos por Províncias/Estados de alguns países, como é o caso dos Estados Unidos e Austrália, os dados são mais uma vez manipulados, desta vez pelo *script* (data_country.py), que agrupa os dados das Províncias/Estados numa só linha com o nome do país a que pertencem. Ainda neste *srcipt* são criados um ficheiro CSV, para cada país no formato 'país.csv'.

Assim, depois de manipulados os dados pelos dois *scripts* mencionados, os dados estão no formato final e prontos para serem usados pelo modelo.

3 Visualização dos Dados

Para ajudar a perceber melhor os dados, decidimos criar representações gráficas dos mesmos. Decidimos criar um *heatmap* "animado", representado na Figura 1, sobreposto ao mapa mundo, que vai mostrando a evolução do número de casos ao longo do tempo. Para tal usamos a ferramenta *Plotly* e sobrepusemos a informação geográfica do conjunto de dados com o mapa. Sempre que o *slider* é movido, as cores do mapa mudam consoante o número de casos no dia seleccionado no *slider*.

Usando ainda as funcionalidades do *Plotly*, criámos mais três *heatmaps* estáticos (Figura 2) que representam o número de casos infetados, mortes e recuperados, para o último dia registado no conjunto de dados.

Por fim, e de forma a termos uma melhor noção da correlação entre casos confirmados, mortes e recuperados e mais uma vez com recurso à ferramenta *Plotly*, críamos três gráficos (Figura 3). O primeiro (da esquerda para a direita),

relaciona o número de mortes por cada 100 casos. O segundo, relaciona os casos recuperados por cada 100 casos infetados. Por último, o terceiro relaciona o número de mortes por cada 100 recuperados. De notar, que nestes três gráficos as relações são calculadas mundialmente, para todos os dias presentes no conjunto de dados, permitindo, assim, ver a sua evolução temporal.

Todos estes elementos visuais são gerados pelo *script* maps_n_graphs.py.

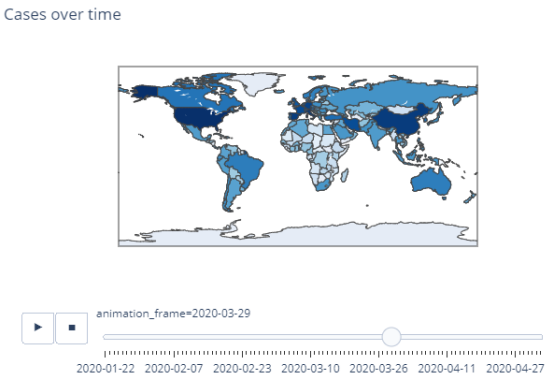

Figura 1: Mapa de casos ao longo do tempo.

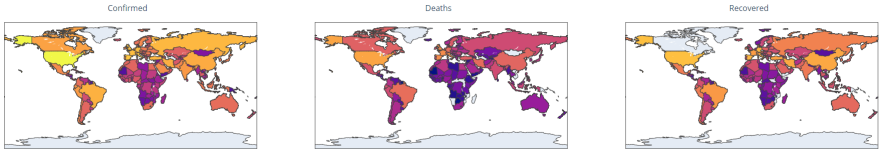

Figura 2: *Heatmaps* de Infetados, Mortes e Recuperados.

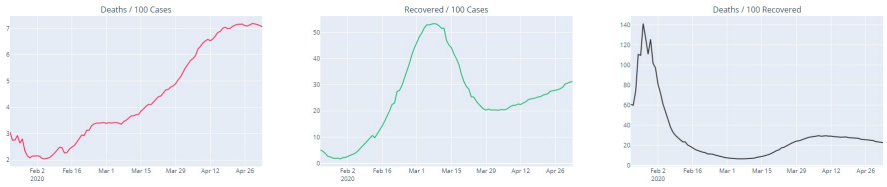

Figura 3: Gráficos de mortes e recuperados por cada 100 casos e mortes por cada 100 recuperados.

4 Modelação e Previsão dos Dados

Inicialmente, pensámos em usar um modelo baseado nos algoritmos clássicos de regressão linear, contudo, como queremos que o modelo tenha a capacidade de realizar previsões de casos futuros, concluímos que regressão linear não seria a melhor opção. Os modelos de regressão linear são bons estimadores quando operam dentro de uma gama de dados conhecida, ou seja, estimar um valor alvo, com base num conjunto de características para qual o modelo tenha sido previamente "treinado". Como a informação relevante do conjunto de dados para o efeito são o número de casos confirmados, mortes e recuperados, não é possível separar os dados em valores alvo e características. Mais ainda, para prever casos futuros, o algoritmo teria de estimar valores com base em caraterísticas fora da gama de caraterísitcas com que fora treinado, o que iria comprometer a qualidade dos resultados.

Após alguma pesquisa, encontrámos o *fbprophet*, uma API *OpenSource* baseada em *sklearn*, criada pela *Facebook Inc.*, inicialmente desenhada para prever flutuações no valor de criptomoeda. Esta API, usa dados na forma *Time Series*, que consiste num conjunto de dados no qual cada linha representa uma data e a respetiva evolução de determinados parâmetros ao longo do período em que os dados foram observados. Esta, é precisamente a forma na qual os dados da *JHU* se encontram. O modelo "estuda"a evolução dos dados ao longo do tempo e tenta encontrar padrões, nomeadamente as flutuações dos dados ao longo de uma semana. Depois de estudados os dados, faz uma modelação por aproximação não linear dos dados existentes, e com base nessa aproximação e no estudo das flutuações calcula uma previsão da evolução dos dados no futuro.

O *script* no qual fazemos uso desta API é o data_processing.py. O modelo, gera um ficheiro CSV com todos os dados relevantes, resultantes dos cálculos que efetuou. Desse ficheiro (forecast_.csv) extraímos os dados mais relevantes, que mencionaremos mais à frente, dados esses que são usados para criar um gráfico em *Plotly* (uma biblioteca *OpenSource*, que serve para criar gráficos interativos), para poderem ser visualizados e mais facilmente interpretados.

O valor mais relevante calculado pelo modelo é a tendência, isto é, os valores estimados se o modelo continuar a seguir a tendência que tem verificado até agora. No entanto, o *fbprophet*, tem um parâmetro designado *interval_width*, que assume valores entre 0 e 1. Esse parâmetro, serve para adicionar à tendência alguma variação mais acentuada em relação à que se tem verificado. Quanto mais perto de 1 for o valor do parâmetro, maior a variação adicionada. Daqui resultam mais quatro conjuntos de valores que podem ser agrupados em dois pares, e que representam uma tendência, para o caso de estudo em concreto, para um cenário mais otimista e um mais pessimista. Dito de outra forma, à tendência esperada, são adicionadas mais duas, uma para o caso em que a situação do contágio do SARS-CoV-2 piore em relação ao que se tem vindo a registar, e outra, para o caso em que melhore.

Como é espectável que a "curva"do contágio diminua ao longo do tempo, por via dos esforços que as várias nações têm vindo a fazer para tal efeito, bem como os avanços na investigação de uma cura e de forma a que o modelo faça uma re-

presentação mais aproximada da realidade, fizemos uma pequena alteração aos valores da tendência digamos que central (a que é calculada como sendo a mais provável). Em vez de representarmos no gráfico a tendência central, tal como calculada pelo modelo, os valores da tendência central, passam a ser a média dos valores da tendência central original e a tendência otimista. Ficamos assim com uma curva que se situa entre a tendência central original e a tendência otimista. Após várias observações, verificámos que esta nova curva, e agora desculpe a redundância, é uma representação mais realista da realidade, do que a tendência central original. Como as mortes e recuperados estão diretamente relacionados com o número de casos infetados, o mesmo procedimento foi aplicado à tendência central destes dois dados. Até porque se assim não fosse, poderia haver a situação em que o modelo previa um número total de casos recuperados superior ao total de infetados, o que é de todo descabido. Garantimos assim, que as curvas se assemelhem a uma sigmóide, tal como é espectável que venha a acontecer.

Existe variação do número de testes diários e eventuais erros na contagem dos casos, como aconteceu em Portugal no dia 1 de maio em que se registou um número de casos totais 161 casos inferiores ao do dia anterior, de forma a corrigir um erro que ocorreu na contabilização do número de casos registados nessa semana. Estas e outras variâncias que ocorrem em todos os países infetados, levam a que para certos países o modelo tenha uma aproximação e previsão mais erráticas do que em outros países onde as variações são menos recorrentes. No entanto, como se tratam de 187 países infetados até á data, essas variações pouco afetam as medições mundiais. Desta forma, o modelo tem o seu melhor desempenho quando modela os dados mundiais.

Desta forma, nas Figuras 4 e 5 estão representadas a aproximação calculada pelo nosso modelo aos dados (do dia 22 de Janeiro a 5 de Março) e a previsão de casos para os 7 dias seguintes.

A figura 6 é a representação gráfica da variação média da tendência ao longo da semana. Isto é, a tendência média do aumento ou diminuição do número de casos, em relação ao dia anterior da semana. Podemos observar, que no decorrer da semana, a quarta é tipicamente o dia com menos casos e sexta e sábado os dias com mais.

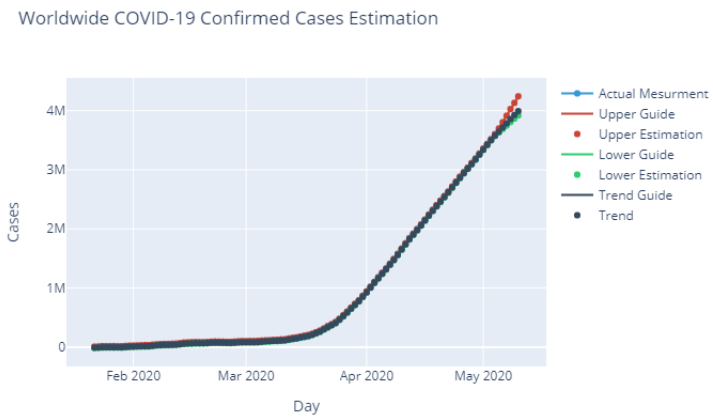

Figura 4: Aproximação do Modelo aos dados e previsão (22/01 a 10/05).

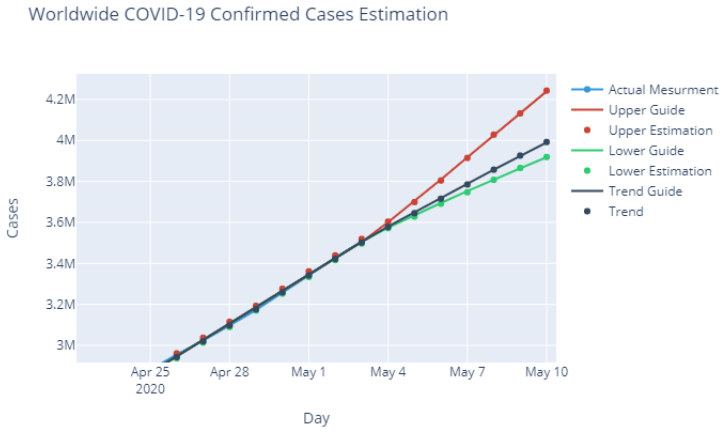

Figura 5: Aproximação do Modelo aos dados e previsão (25/04 a 10/05).

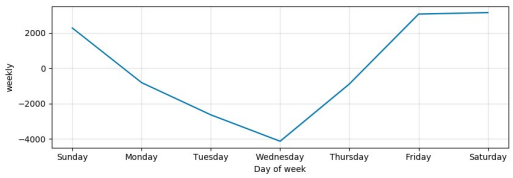

Figura 6: Variação média da tendência ao longo da semana.

5 Conclusão

Tal como tínhamos proposto, criámos um modelo que representa o contágio/ propagação do novo vírus SARS-CoV-2 e que consegue estimar com alguma precisão, os casos futuros. Os gráficos que criámos ao longo do trabalho, serviram e servem para compreender de forma mais simples e clara a realidade e a severidade da situação em que nos encontramos.

Os centros urbanos com elevada densidade populacional, a grande percentagem de população envelhecida e debilitada à priori e a inexistência de um tratamento eficaz, tem tornado a COVID-19 uma das mais graves pandemias do mundo moderno. Tal como os dados refletem, temos de continuar a adotar comportamentos que reduzam o risco de contágio, pelo menos até à descoberta de uma possível vacina. Os países que mais cedo adotaram medidas, são aqueles que melhor têm conseguido conter o vírus, embora que com algumas excepções.

Apesar de simples, o modelo mostrou-se capaz de aproximar e prever a evolução da pandemia. Como é óbvio fica aquém de outros modelos computacionais desenvolvidos por grandes instituições e grupos científicos (como é o caso da investigação realizada pela OMS, JHU, Paul Ehrlich Institute, entre outros) para este e outros casos. Todo o trabalho de investigação realizada por estas entidades, é muito mais minucioso e conta com a opinião e saber de peritos e com uma variedade de dados muito mais alargada e mais especifica para o caso de estudo em questão. Tudo isto, leva a um resultado final mais correto do ponto de vista científico, resultados esses que são de facto usados par impulsionar a criação de uma vacina e outros tratamentos. No entanto, nunca foi nosso objetivo "competir"com tais modelos. O objetivo fundamental, e que foi alcançado, era perceber a evolução de um vírus, neste caso concreto o SARS-CoV-2 e as relações entre os vários parâmetros (infetados, mortos, recuperados). Tão importante como os estudos realizados por grandes comunidades cientificas, é a informação que é passada à população em geral, pois o trabalho de investigação realizado, nada é sem o esforço de todos nós de seguirmos as recomendações dadas por tais organismos, de forma a combatermos este e outros vírus. Este trabalho reflete isso mesmo, uma fonte de informação simples e fácil de compreender por todos.

Futuramente, este modelo pode ser usado para outras pandemias e para aproximar o contágio de outros vírus. Este e outros modelos do mesmo género, são fundamentais para percebermos a forma como os vírus se propagam e quais as formas de nos anteciparmos a estes, para que os seus impactos sejam minimizados.

Todos os *scripts* criados e usados no âmbito deste trabalho, estarão anexados ao presente relatório para consulta.

Referências

[1] Novel Coronavirus (COVID-19) Cases Data [Online], Último acesso: 11/05/2020
https://data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases

[2] HDX [Online], Último acesso: 11/05/2020
https://data.humdata.org/

[3] Documentação fbprophet 0.6 [Online], Último acesso: 11/05/2020
https://pypi.org/project/fbprophet/

[4] Dataset - COVID-19 Case Study - Analysis, Viz Comparisons [Online], Último acesso: 11/05/2020
https://www.kaggle.com/tarunkr/covid-19-case-study-analysis-viz-comparisons

[5] A Guide to Time Series Forecasting with Prophet in Python 3 [Online], Último acesso: 11/05/2020
https://www.digitalocean.com/community/tutorials/a-guide-to-time-series-forecasting-with-prophet-in-python-3

[6] Forecasting the Spread of Coronavirus (COVID-19) Using Python [Online], Último acesso: 11/05/2020
https://laconicml.com/coronavirus-prediction-covid-19/

[7] Documentação plotly 4.7.1 [Online], Último acesso: 11/05/2020
https://pypi.org/project/plotly/