Aprendizado de máquina combinada com o modelo epidemiológico SIR para predição de COVID-19 na região de Limeira-SP

Palavras-Chave: Sistemas de equações diferenciais, Aprendizado de máquina, COVID-19

Autores(as):
Diogo Matheus da Silva, FCA – UNICAMP
Prof. Dr. Cristiano Torezzan, coorientador FCA - UNICAMP
Prof. Dr. Washington Alves de Oliveira, orientador, FCA - UNICAMP

INTRODUÇÃO:

Seguindo o enorme impacto na sociedade, diversas áreas de pesquisa foram também afetadas e influenciadas pela pandemia de COVID-19. Neste período, houve grande interesse da comunidade científica na busca de métodos para controlar, entender e amenizar os efeitos da doença na população. Além dos problemas que afetam diretamente a saúde das pessoas, o controle do número de pessoas infectadas na sociedade é crucial devido à sobrecarga gerada nos sistemas de saúde. Para prever futuros casos de infectados e avaliar a dinâmica de evolução da pandemia é importante a compreensão de modelos que estimam o número real de casos. Neste contexto, o objetivo deste projeto é integrar métodos de aprendizado de máquina e uma modelagem matemática que descreve a dinâmica de uma epidemia para prever o número de novos casos.

A COVID-19, assim como outros vírus, faz as pessoas transitarem entre estados: suscetíveis ao vírus, infectadas e recuperadas ou falecidas. A modelagem deve, portanto, refletir essas transições de estados. Kermack e Mckendrick (1927) desenvolveram um sistema de equações diferenciais ordinárias (EDOs) que simulam tal situação, este nomeado como SIR (Suscetível-Infectado-Recuperado). O modelo trata-se de um sistema não linear de EDOs como descrito abaixo

$$\frac{dS(t)}{dt} = -\beta S(t)\frac{I(t)}{N}, \quad (1)$$

$$\frac{dI(t)}{dt} = \beta\frac{S(t)}{N}I(t) - \delta I(t), \quad (2)$$

$$\frac{dR(t)}{dt} = \delta I(t), \quad (3)$$

em que β e δ são respectivamente a taxa de contágio e a taxa de recuperação do vírus, $S(t)$, $I(t)$ e $R(t)$ são respectivamente o número de suscetíveis, infectados e recuperados/mortos, enquanto que N é o número fixo da população em estudo (Kermack e McKendrick, 1927).

Esta pesquisa busca entender o funcionamento e a dinâmica desse modelo, assim como aplicar métodos matemáticos e de aprendizado de máquina para de realizar predições de casos de infectados, devido à dificuldade da captação de banco de dados necessários para aplicar os métodos a região de Limeira-SP foi substituída para a região do Estado de São Paulo.

Metodologia:

Com o intuito de estudar e entender a dinâmica do modelo SIR, foram elaboradas simulações com diferentes valores fixos de δ, β, t e N para o sistema de EDOs. Uma vez que a solução analítica é complexa e possui limitações para usos práticos, é mais comumente encontrado na literatura a abordagem numérica de resolução (aproximada) desse sistema. Nesta pesquisa foi utilizado o método de Runge-Kutta de 4ª ordem (Liao et al., 2020).

O modelo SIR, como descrito anteriormente, pode ser muito útil para entender o funcionamento da epidemia em seu início, porém a longo prazo é necessário fazer alterações ou correções nos parâmetros, para que o modelo se ajuste melhor aos dados observados. Uma abordagem utilizada para esse propósito é variação temporal dos parâmetros β e δ de forma que, ao invés de serem parâmetros fixos, sejam variáveis no tempo, isto é, $\beta(t)$ e $\delta(t)$ (Chen et al.,2020). Desta forma a partir de uma discretização adequada do modelo SIR, obtêm-se

$$S(t+1) - S(t) = -\beta(t)S(t)\frac{I(t)}{N}, \quad (4)$$

$$I(t+1) - I(t) = \beta(t)\frac{S(t)}{N}I(t) - \delta(t)I(t), \quad (5)$$

$$R(t+1) - R(t) = \delta(t)I(t), \quad (6)$$

onde, no início da epidemia, considera-se $S(t) \approx N$, uma vez que inicialmente o número de pessoas suscetíveis em uma população é praticamente sua totalidade. Assim, ao manipular as equações (4), (5) e (6), obtemos:

$$I(t+1) = [1 + \beta(t) - \delta(t)]I(t) \quad (7)$$

$$\delta(t) = \frac{R(t+1) - R(t)}{I(t)}, \quad (8)$$

$$\beta(t) = \frac{I(t+1) - I(t) + R(t+1) - R(t)}{I(t)}. \quad (9)$$

Tendo um conjunto de dados com o número de pessoas recuperadas e infectadas diariamente em um período de tempo *T*, de determinada população, é possível determinar os valores das taxas de contágio e transmissão no tempo *t* para *t* ≤ *T-1*. Desta forma, pode-se ajustar uma linha de tendência que descreva o comportamento dos dados para então estimar seus valores para um período maior, devido a complexidade dos dados, aplicou-se um método de aprendizado de máquina supervisionado para regressão, comumente conhecido como Regressão de *Ridge*, em que se adiciona uma penalização controlada por um hiperparâmetro α na regressão como descrito abaixo:

$$min \sum_{i=1}^{n} [y_i - \widehat{y}_i]^2 + \alpha \sum_{i=1}^{n} w_i^2, \quad (10)$$

Em que w_i representa os valores dos coeficientes da função em um conjunto de n dados em treinamento. Neste trabalho foi utilizado a Regressão de Ridge através da biblioteca scikit-learn em Python. Assim podendo prever o número de infectados no tempo *t*.

RESULTADOS E DISCUSSÃO:

Observou-se o comportamento das soluções do modelo SIR através das simulações utilizando Runge-Kutta de 4ª ordem, e percebeu-se que quando a taxa de recuperação δ é maior do que a taxa de contágio β no início, a epidemia não ocorre. Entretanto, quando a taxa de contágio é maior, então a pandemia se desenvolve, no qual o número de infectados em uma população possui uma crescente devido a taxa de contágio e depois apresenta um comportamento decrescente devido a diminuição no número de suscetíveis. Um exemplo da simulação pode ser observado na Figura 1.

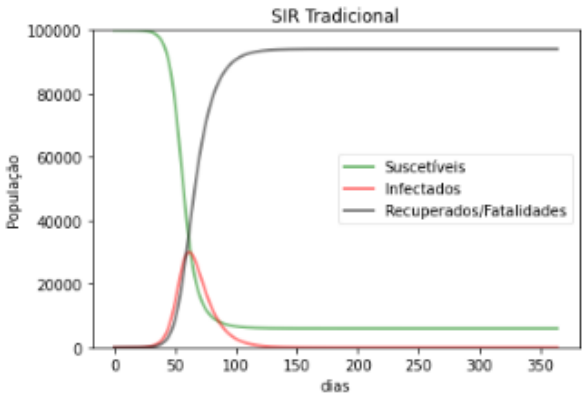

Figura 1: Simulação do modelo SIR. Fonte: criado pelo autor

Utilizando dados abertos da giscard, do dia 28 de setembro de 2021 até dia 1 de janeiro de 2022 para o estado de São Paulo. A Figura 2 ilustra os valores de beta e gama obtidos a partir das equações (7) e (8).

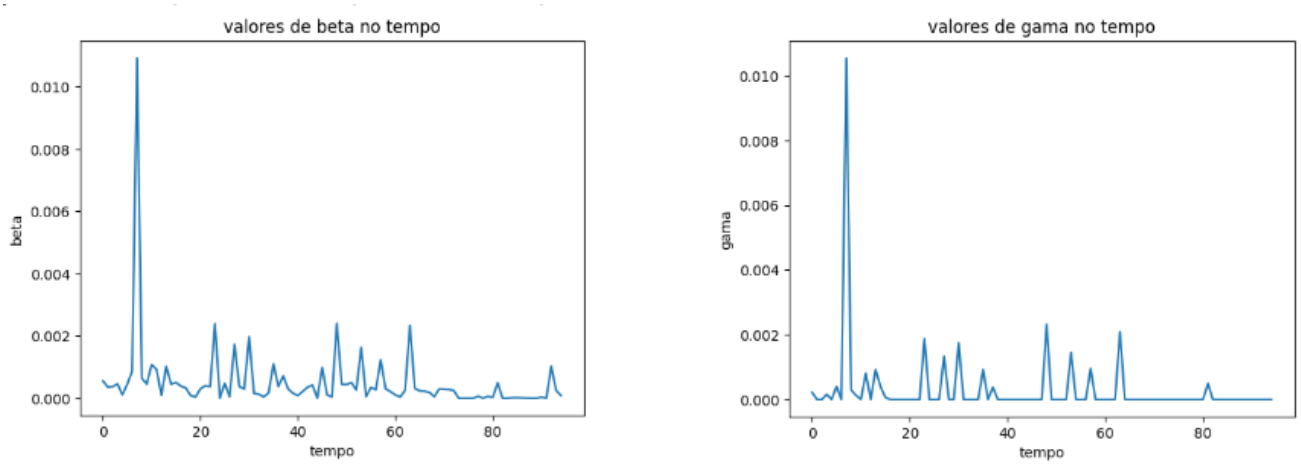

Figura 2: Gráfico dos valores de beta e gama em função do tempo. Fonte: criado pelo autor

Então, aplicando o método de aprendizado de máquina, regressão de ridge, sendo 0,7 o valor de α , obtêm-se valores de $\widehat{\beta}(t)$ e $\widehat{\delta}(t)$ previstos para o tempo *t*. A Figura 3 ilustra os gráficos e as linhas de tendências das regressões para os valores de beta e gama.

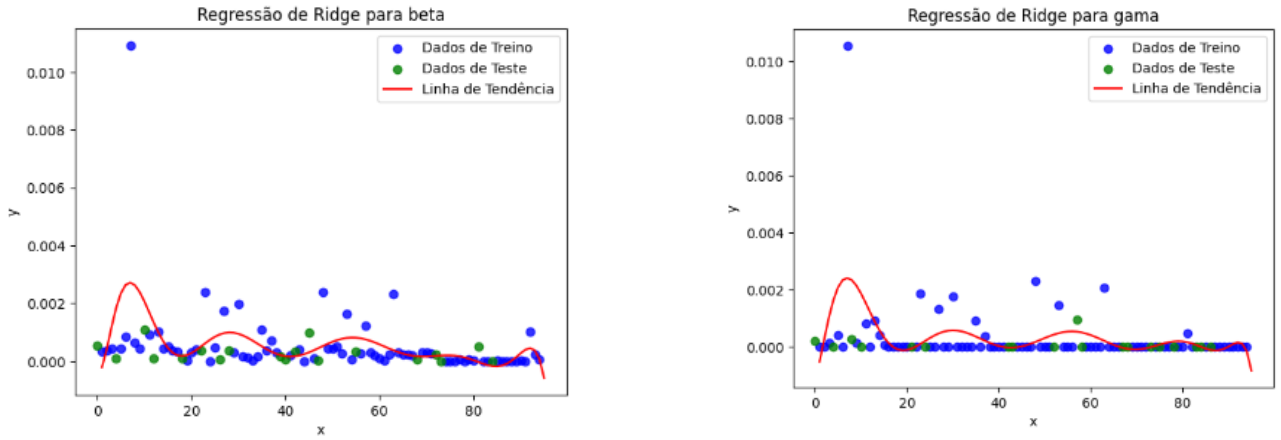

Figura 3: Gráficos dos valores de beta e gama, incluindo a regressão de ridge. Fonte: criado pelo autor

Voltando para equação (7) é possível obter o número previsto de infectados utilizando os valores $\widehat{\beta}(t)$ e $\widehat{\delta}(t)$ obtidos. Uma comparação do número de casos reais e os valores obtidos a partir da equação (7) pode ser observada na Figura 4, em que a previsão realizada para um dia, isto é, no tempo $t = T$, dia 1 janeiro de 2022, foi de 4458380 infectados totais, sendo 0,0429% maior do que o valor real, segundo o conjunto de dados o número de infectados é de 4456469 no estado de São Paulo.

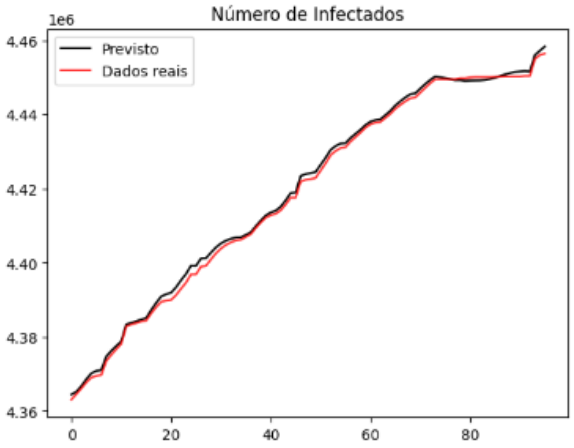

Figura 4: Gráfico do número real de pessoas infectadas e valores previstos no tempo t. Fonte: criado pelo autor

CONCLUSÕES:

A modelagem matemática do modelo epidemiológico SIR é crucial para descrever a transição de estados entre suscetíveis, infectados e recuperados ou falecidos, e o modelo SIR é utilizado para simular essa dinâmica. Além disso, a previsão de futuros casos de infectados é importante para controlar uma epidemia e avaliar a eficiência do modelo. Com a utilização de aprendizado de máquina e o uso de uma base histórica de dados, é possível aprimorar as previsões e contribuir para o desenvolvimento de estratégias de combate à COVID-19. É relevante ressaltar a importância de um controle quanto às informações dos números de pessoas infectadas e recuperadas por parte das autoridades públicas, uma vez que a qualidade dos dados são fundamentais para a realização de estudos e aplicação de políticas públicas no combate à epidemia e no planejamento do atendimento das unidades de saúde.

BIBLIOGRAFIA

William Ogilvy Kermack and Anderson G McKendrick. **A contribution to the mathematical theory of epidemics**. Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character, 115(772):700–721, 1927.

STEPHANOU, G. Painel Covid-19 -> Estatísticas do Coronavírus - Coronavirus Statistics by Giscard Stephanou. Disponível em: <http://www.giscard.com.br/coronavirus/estatisticas-coronavirus-estado.php?cod_pais=1025>. Acesso em: 25 jul. 2023.

Chen, Y.C.; Lu, P.E.; Chang, C.S.; Liu, T.H. **A time-dependent SIR model for COVID-19 with undetectable infected persons**. *IEEE Trans. Netw. Sci. Eng.* 2020, *7*, 3279–3294.

Zhifang Liao, Peng Lan, Zhining Liao, Yan Zhang, and Shengzong Liu. **Tw-sir: time-window based sir for covid-19 forecasts**. Scientific reports, 10(1):1–15, 2020.