ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC BÁCH KHOA
KHOA KHOA HỌC - KỸ THUẬT MÁY TÍNH

MÔ HÌNH HOÁ TOÁN HỌC (mr)

Mô hình SIR trong dự báo COVID-19

GVHD: Nguyễn An Khương
SV thực hiện: **Tô Duy Hưng – 1810198**
Võ Hoàng Hải Nam – 1810340
Võ Thanh Phong – 1712633
Huỳnh Thị Uyên – 1810648
Lê Thành Sơn – 1810481

Tp. Hồ Chí Minh, Tháng 7/2020

Mục lục

Danh sách hình vẽ

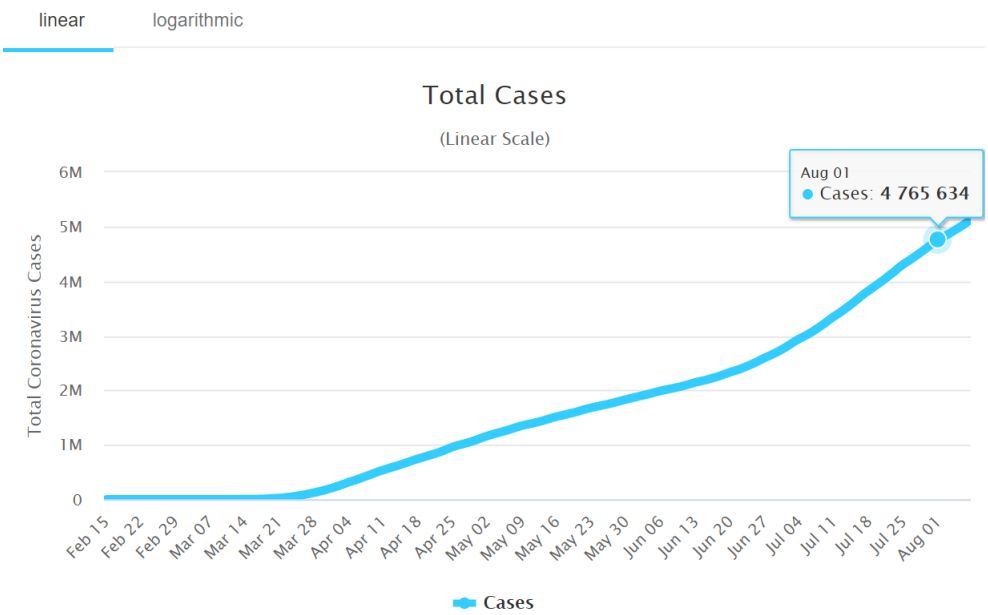
1 Kết hợp Mô hình Học máy

1.1 Giới thiệu

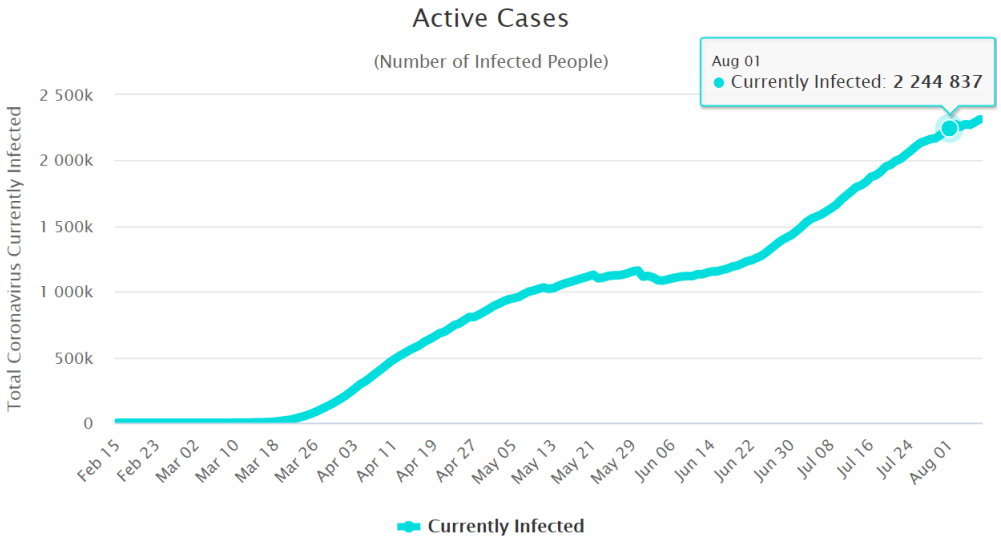
Trước diễn biến phức tạp của dịch bệnh COVID-19 trong thời gian gần đây, rất nhiều nhà nghiên cứu đã tiến hành sử dụng các mô hình Học máy (*machine learning*) để dự đoán tình hình dịch bệnh. Bằng cách tận dụng khả năng tính toán mạnh mẽ của máy tính, các mô hình Học máy có thể tìm ra các đặc tính và xu hướng phát triển chứa đựng trong dữ liệu. Một mô hình Học máy hiện đại có thể được mô tả tương tự như hình dạng của mạng lưới các tế bào Nơ-ron thần kinh nối liên tiếp với nhau, từ đó có thể rút trích đặc trưng của dữ liệu qua từng lớp của mạng lưới.

Tuy nhiên, Học máy vốn có trọng tâm là các bài toán tối ưu đi cực tiểu hóa các hàm chi phí, dùng để đo đạc sai số giữa giá trị thực tế và giá trị dự đoán của mô hình. Vậy nên, hai trong những vấn đề gặp phải trong quá trình xây dựng mô hình Học máy đó là việc thiết kế xây dựng các hàm chi phí phải phù hợp với dữ liệu đầu vào và việc tìm ra các điểm tối ưu của nó. Tùy vào từng loại dữ liệu mà mô hình được xây dựng rất khác nhau.

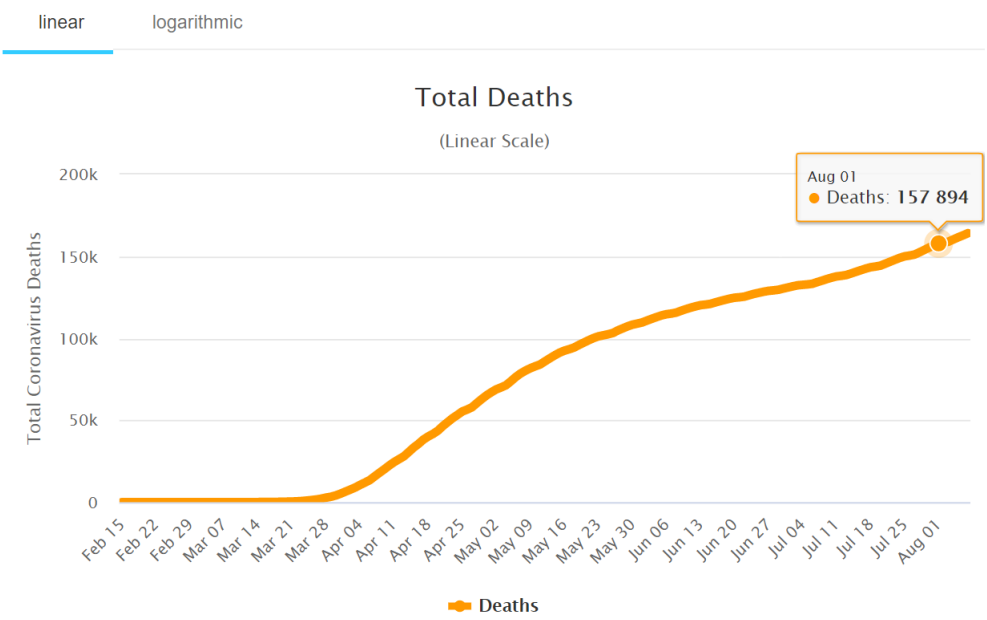
Trong bài báo cáo này, chúng ta sẽ xây dựng một mô hình dự báo COVID-19 có kết hợp Học máy. Dữ liệu được tham khảo từ CSSE của trường Đại học John Hopkins, được công bố tập hợp tại https://github.com/CSSEGISandData/COVID-19. Đối tượng mà bài báo cáo này tập trung hướng tới chính là nước Mỹ. Tính tới cuối ngày 01/08/2020, nước Mỹ đã có 4.765.634 ca nhiễm COVID-19 (trong đó có 2.244.837 ca được xét nghiệm là dương tính) 157.894 ca tử vong và 1.461.885 ca hồi phục trong tổng dân số 331.002.651 người. Mỹ là quốc gia có số ca nhiễm và tử vong vì COVID lớn nhất trên thế giới tại thời điểm. Những số liệu trên được phác thảo tại worldometers.info như Hình 1, Hình 2 và Hình 3



Hình 1: *Tổng số ca nhiễm COVID-19 ở Mỹ*



Hình 2: *Tổng số ca nhiễm được xét nghiệm là dương tính ở Mỹ*



Hình 3: *Tổng số ca tử vong vì dịch bệnh COVID-19 ở Mỹ*

1.2 Hiện thực

Sau khi lấy được dữ liệu, ta tiến hành tiền xử lý bằng cách loại bỏ các giá trị gây nhiễu, ví dụ như có những thời điểm dữ liệu được ghi nhận tuy nhiên lại không có bất kì ca nhiễm nào.

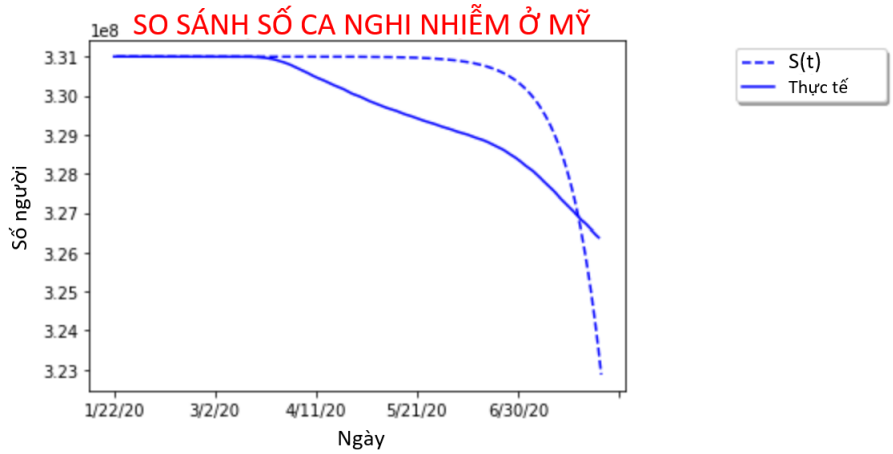
Tiếp theo, tiến hành tối ưu hóa giá trị của hàm RMSE. Đây là hàm sẽ ước lượng các giá trị S, I, R, D lần lượt tương ứng với số ca nghi nhiễm, số ca nhiễm, số ca hồi phục và số ca tử vong bằng phương pháp *solve_ivp* được cung cấp bởi thư viện *scipy*. Bởi vì có tới 4 giá trị được tính cho nên cách đơn giản nhất để đánh giá hàm RMSE đó là lấy trung bình cộng của 4 giá trị trên. Sau đó, dùng *optimize* để tìm điểm tối ưu của hàm RMSE đối với cả 3 biến β, γ, μ. Khi đó, ta

có các giá trị ước lượng lần lượt là:

- Tỉ lệ tiếp xúc của một người trong nhóm S với người trong nhóm I: $\beta = 0.11995893364421482$
- Tỉ lệ hồi phục khi mắc bệnh: $\gamma = 0.03804067192756672$
- Tỉ lệ tử vong khi mắc bệnh: $\mu = 0.004789802899428518$

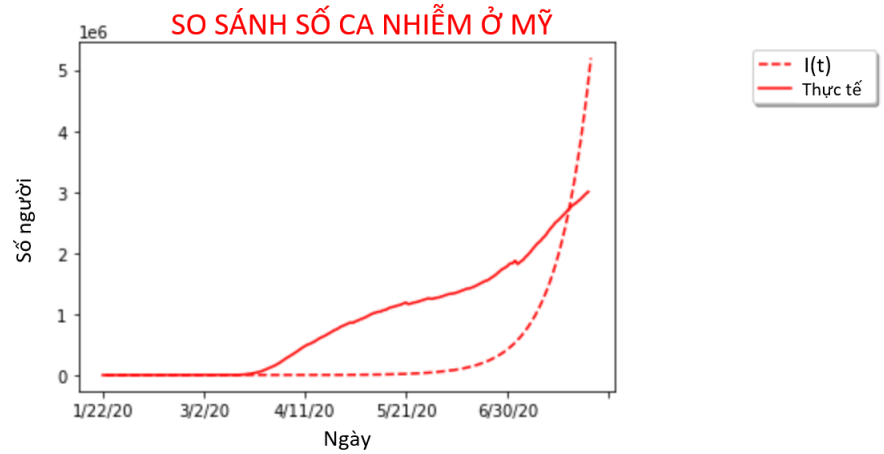
Áp dụng 3 giá trị β, γ, μ như trên vào mô hình SIRD ta có lần lượt:

- Đồ thị so sánh số ca nghi nhiễm bệnh ở Mỹ so với thực tế được phác thảo như Hình 4



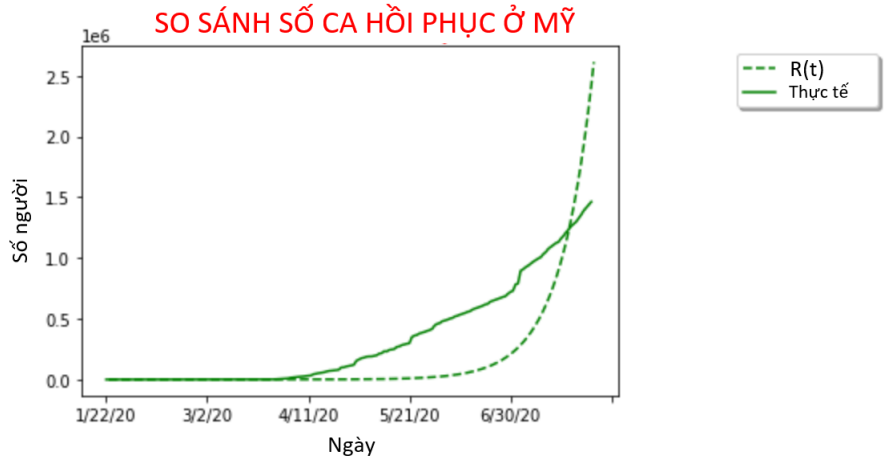
Hình 4: *So sánh giá trị của S(t) với thực tế*

- Đồ thị so sánh số ca nhiễm bệnh ở Mỹ so với thực tế được phác thảo như Hình 5

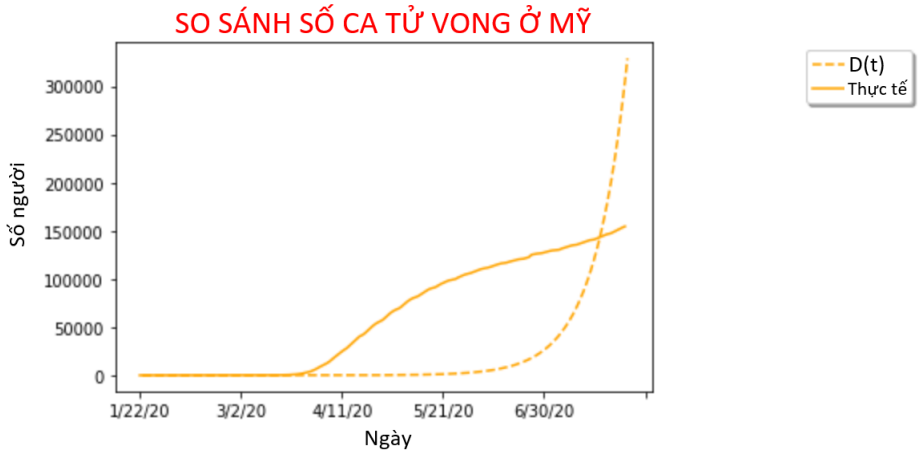


Hình 5: *So sánh giá trị của I(t) với thực tế*

- Đồ thị so sánh số ca hồi phục ở Mỹ so với thực tế được phác thảo như Hình 6
- Đồ thị so sánh số ca tử vong ở Mỹ so với thực tế được phác thảo như Hình 7



Hình 6: *So sánh giá trị của R(t) với thực tế*



Hình 7: *So sánh giá trị của D(t) với thực tế*

Ta có thể thấy, các giá trị tính toán trên đã đạt được độ chính xác cao hơn về mặt lý thuyết, tuy nhiên nếu so sánh với số liệu thực tế, sai số này vẫn còn là rất lớn. Ta sẽ tiếp tục tiến hành huấn luyện các thông số qua mạng neuron nhân tạo.

1.3 Xây dựng mô hình sử dụng mạng neuron (*Neural Network*)

1.3.1 Dữ liệu đầu vào (Input)

Ta sử dụng 80% tập dữ liệu được tham khảo từ CSSE của Đại học John Hopkins tương ứng với nước Mỹ làm dữ liệu đầu vào cho mô hình. Đó là những dữ liệu có giá trị của cột "Country/Region" là "US". Ngoài ra, để huấn luyện một mô hình, ta cần phải có các giá trị để được gán nhãn. Do đó, cần phải tính toán các giá trị β, γ, μ mỗi ngày để gán nhãn.

1.3.2 Các thông số liên quan

- Thư viện hỗ trợ: Keras

- Lớp ẩn (*Hidden Layer*): Mạng neuron của ta là một mạng đơn giản gồm 2 Dense với số neuron lần lượt là 32 và 27, giống như Hình 8

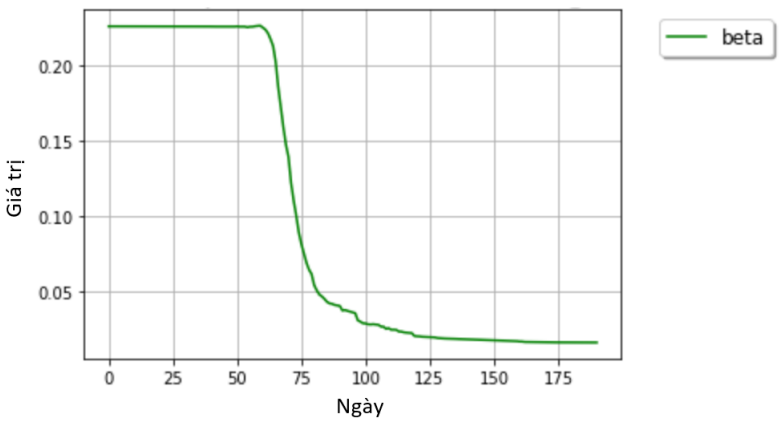
```
Model: "sequential_6"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_11 (Dense)             (None, 8)                 32
_________________________________________________________________
dense_12 (Dense)             (None, 3)                 27
=================================================================
Total params: 59
Trainable params: 59
Non-trainable params: 0
```

Hình 8: *Thông số về lớp ẩn của mạng neuron*

- epochs = 200
- batch_size = 1
- Hàm kích hoạt: Sigmoid
- Hàm tối ưu (*Optimizer*): Adam
- Hàm mất mát (*Loss function*): MSE (*Mean squared error*)
- Metrics: accuracy

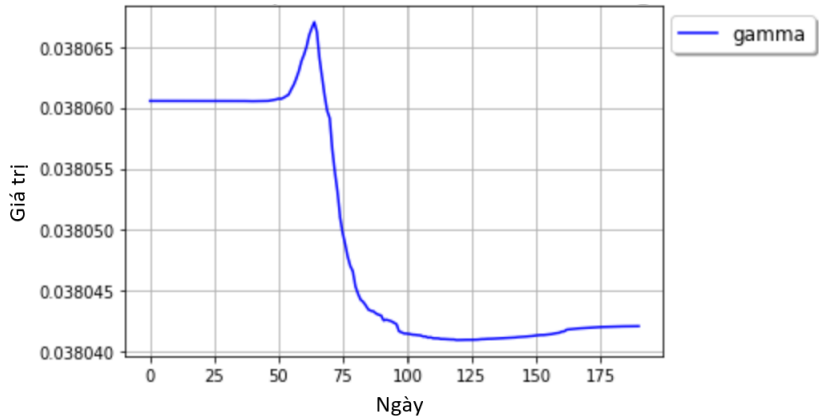
Sau khi đã huấn luyện, ta đưa vào tập giá trị SIRD thực tế để mô hình dự đoán các giá trị β, γ, μ tương ứng của ngày hôm đó. Kết quả thu được như sau:

- Giá trị của β được dự đoán như Hình 9

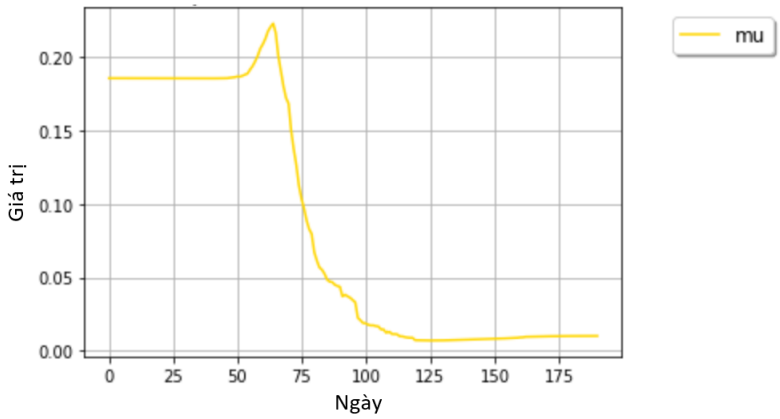


Hình 9: *Giá trị β dự đoán*

- Giá trị của γ được dự đoán như Hình 10
- Giá trị của μ được dự đoán như Hình 11



Hình 10: *Giá trị γ dự đoán*



Hình 11: *Giá trị μ dự đoán*

Cuối cùng, từ ba giá trị đã dự đoán ở mỗi ngày được như trên, ta có thể dự đoán giá trị của hệ số R_0 trong ngày hôm đó bằng công thức:

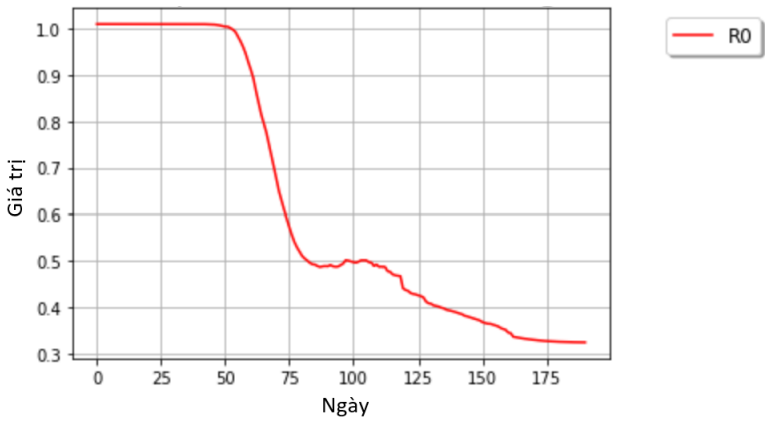
$$R_0 = \frac{\beta}{\gamma + \mu} \tag{1}$$

Đồ thị dự đoán giá trị R_0 được phác thảo như Hình 12

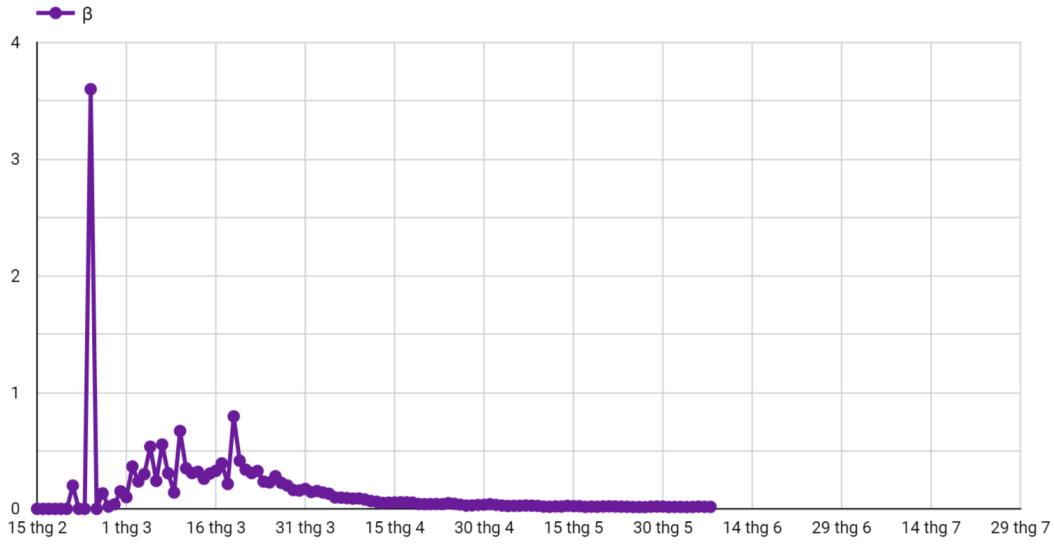
1.3.3 So sánh với thực tế

Số liệu thực tế về các giá trị β, γ, μ và R_0 tại Mỹ được tham khảo từ `covid19sa.org` như sau:

- Giá trị của β thực tế được phác thảo như Hình 13
- Giá trị của γ thực tế được phác thảo như Hình 14
- Giá trị của μ thực tế được phác thảo như Hình 15
- Giá trị của $_0$ thực tế được phác thảo như Hình 16



Hình 12: *Giá trị R_0 dự đoán*



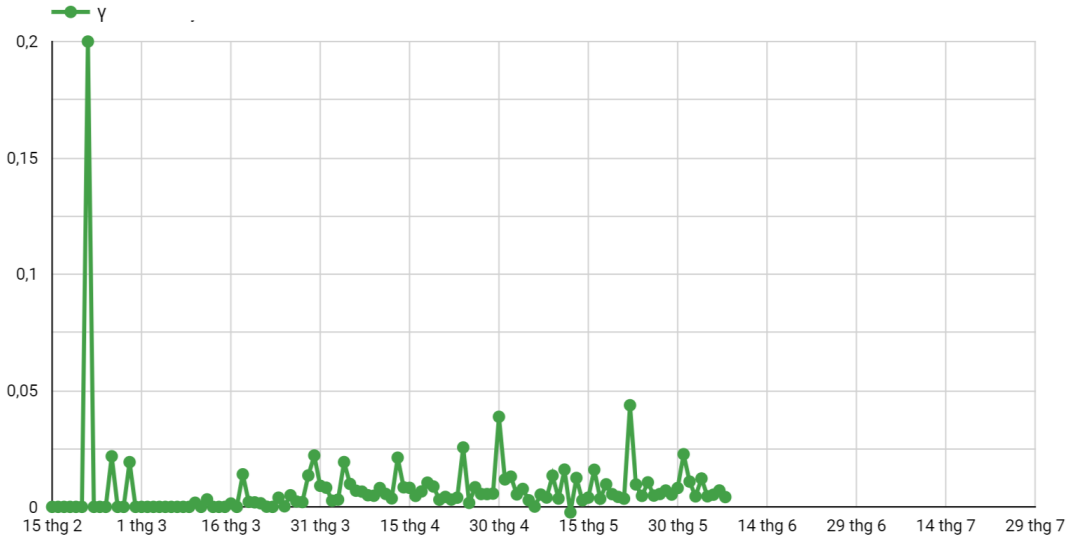
Hình 13: *Giá trị β thực tế ở Mỹ*

Nhìn chung, kết quả huấn luyện đạt được tốt hơn so với truyền thống và phù hợp với thực trạng đang diễn ra ở Mỹ. Người đọc có thể tham khảo chi tiết thực hiện trong tại file *SIR_neural_network.ipynb* gửi kèm hoặc tại https://github.com/ToDuyHung/assignment_with_SIRD_model

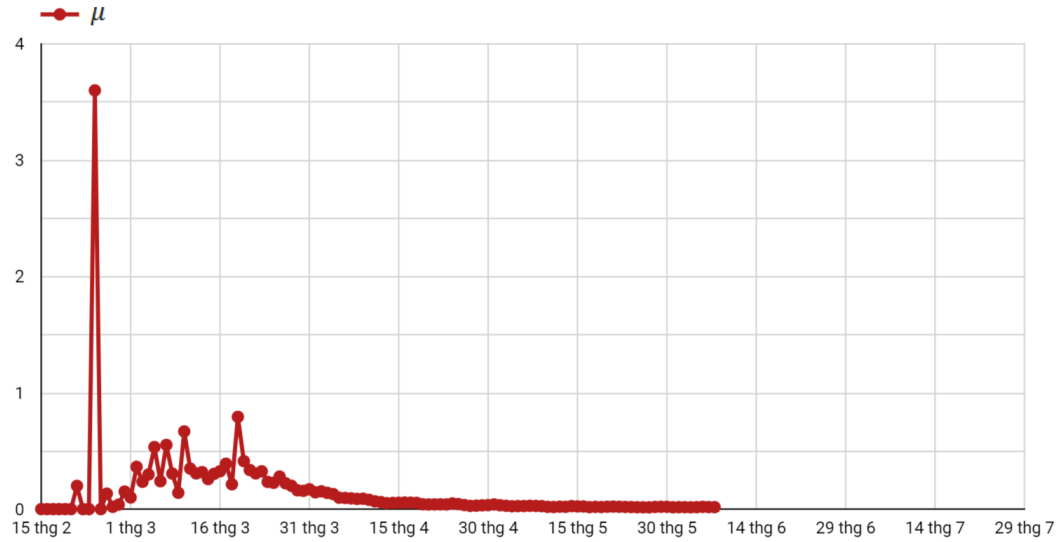
1.4 Phân tích

Trường hợp đầu tiên được xác nhận về sự bùng phát đại dịch toàn cầu của COVID-19 tại Mỹ đã được công bố vào ngày 21 tháng 1 năm 2020. Các ca nhiễm bệnh đã được xác nhận ở tất cả 50 tiểu bang của Mỹ, quận Columbia và tất cả các vùng lãnh thổ có người ở Mỹ, ngoại trừ Samoa và Quần đảo Bắc Mariana. Tính đến ngày 12 tháng 7 năm 2020, Mỹ là quốc gia có số ca nhiễm bệnh và tổng số ca tử vong nhiều nhất thế giới do virus này, với các bang New York và New Jersey là trung tâm của dịch bệnh với hơn một nửa số ca.

Ca đầu tiên nhiễm COVID-19 được biết đến ở Mỹ đã được xác nhận ngày 20 tháng 1 năm 2020, ở một người đàn ông 35 tuổi trở về từ Vũ Hán, Trung Quốc năm ngày trước đó. Lực lượng đặc nhiệm nhằm ứng phó với virus corona của Nhà Trắng đã được thành lập vào ngày 29 tháng



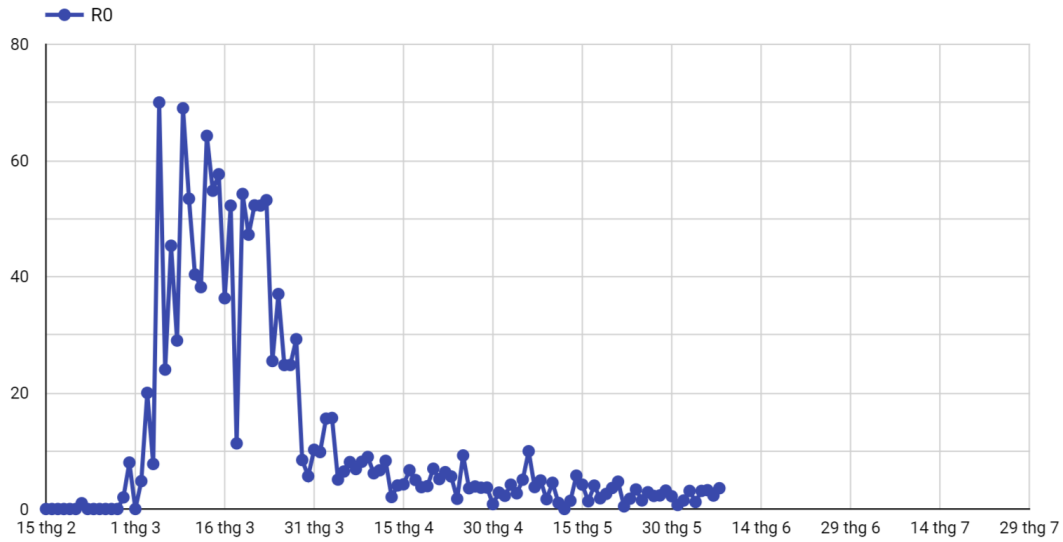
Hình 14: *Giá trị γ thực tế ở Mỹ*



Hình 15: *Giá trị μ thực tế ở Mỹ*

1. Hai ngày sau, chính quyền Donald Trump đã tuyên bố tình trạng khẩn cấp y tế công cộng và tuyên bố hạn chế đối với các du khách đến từ Trung Quốc. Sau đó, vào ngày 26 tháng 2 đã xảy ra trường hợp tử vong đầu tiên ở Mỹ vì COVID-19 được xác nhận bởi Trung tâm kiểm soát và phòng ngừa dịch bệnh (Centers for Disease Control, viết tắt là CDC) tại Bắc California.

Từ tháng 1 đến giữa tháng 3 năm 2020, Mỹ đã khởi đầu tương đối chủ quan và chậm chạp trong việc tiến hành xét nghiệm COVID-19 trong lãnh thổ quốc gia của mình. Trong giai đoạn đó, chính phủ liên bang Mỹ đã từ chối các bộ dụng cụ thử nghiệm được sử dụng trên phạm vi quốc tế. Thay vào đó họ tự phát triển bộ dụng cụ thử nghiệm của chính phủ tuy nhiên kết quả là chúng lại bị lỗi. Chỉ có các bộ dụng cụ thử nghiệm phi chính phủ được phê duyệt từ cuối tháng 2 và có hướng dẫn kiểm tra đủ điều kiện hạn chế cho đến đầu tháng 3 mới cho ra kết quả chính xác hơn. Điều này làm cho quá trình ứng phó với dịch bệnh COVID-19 ở Mỹ diễn ra rất trễ. Mặc dù ngay lập tức sau đó, chính phủ Mỹ đã phải công bố một loạt các biện pháp nhằm



Hình 16: *Giá trị R_0 thực tế ở Mỹ*

tăng tốc độ xét nghiệm. Nhưng tính đến ngày 20 tháng 3, 100.000 bài kiểm tra đã được tiến hành, con số này vẫn chưa là đủ. CDC cảnh báo rằng việc lây truyền rộng rãi có thể buộc một số lượng lớn người dân Mỹ phải nhập viện và chăm sóc sức khỏe, khiến các hệ thống chăm sóc sức khỏe ở quốc gia này trở nên quá tải. Đó là lý do giải thích tại sao giá trị R_0 thực tế ở Mỹ trong khoảng thời gian này lại cao và biến động liên tục với biên độ lớn đến như vậy.

Sau đó, kể từ ngày 19 tháng 3 năm 2020, Bộ Ngoại giao Mỹ đã phải khuyên công dân của mình hạn chế tối đa tất cả các chuyến du lịch quốc tế, khuyến khích không nên tập trung hơn 10 người tại một địa điểm. Sau giữa tháng 3 năm 2020, Chính phủ Liên bang đã thực hiện một thay đổi đáng kể, tất nhiên là sử dụng quân đội Mỹ để khởi xướng và nỗ lực phát triển nhanh chóng các cơ sở chăm sóc đặc biệt COVID-19 toàn quốc. Công binh Lục quân, dưới quyền lực pháp lý hiện có đến từ Quốc hội ủy quyền và quyền hạn của Fema, sẽ nhanh chóng cho thuê một số lượng lớn các tòa nhà, khách sạn trên khắp quốc gia để ngay lập tức tăng số lượng phòng và giường với khả năng điều trị tích cực (ICU) cho các bệnh nhân của đại dịch COVID-19. bên cạnh đó, Chính phủ cũng đưa ra các phản ứng trước sự bùng phát của dịch bệnh bao gồm các lệnh cấm và hủy bỏ các cuộc tụ họp quy mô lớn, đóng cửa các trường học và các tổ chức giáo dục, hủy bỏ triển lãm thương mại, hội nghị, âm nhạc các lễ hội, hủy bỏ và đình chỉ các sự kiện thể thao và giải đấu. Điều này góp phần giúp cho tỉ lệ tử vong μ và nguy cơ bùng phát dịch R_0 từ tháng 4 năm 2020 đến nay ở Mỹ có dấu hiệu giảm mạnh. Tuy nhiên, Mỹ vẫn đang là nước có số ca nhiễm và tử vong lớn nhất trên thế giới hiện nay.

2 Ước lượng bằng suy luận Bayes

2.1 Giới thiệu

Bên cạnh các mô hình Học máy, phương pháp ước lượng bằng suy luận Bayes cũng được đưa vào dự đoán tình hình dịch bệnh COVID-19 rất phổ biến trong các tháng vừa qua.

Nhắc đến suy luận Bayes, đây là phương pháp quy nạp sử dụng một ước lượng bằng số về mức độ tin tưởng vào một giả thuyết trước khi quan sát được bằng chứng. Sau đó khi đã quan sát được bằng chứng, ta sẽ đi tính toán lại một ước lượng bằng số khác về mức độ tin tưởng vào giả thuyết trên. Trong quá trình quy nạp, suy luận Bayes thường dựa vào các mức độ tin tưởng,

hay là các xác suất chủ quan, và không nhất thiết khẳng định về việc cung cấp một phương pháp quy nạp khách quan.

Phương pháp này điều chỉnh các xác suất khi được cho bằng chứng mới theo công thức của định lý Bayes như sau:

$$P(H_0|E) = \frac{P(E|H_0)P(H_0)}{P(E)} \tag{2}$$

Trong đó:

- H_0 đại diện cho một giả thuyết (*hypothesis*)
- $P(H_0)$ là xác suất tiền nghiệm của H_0
- $P(E|H_0)$ được gọi là xác suất có điều kiện của việc quan sát thấy bằng chứng E nếu biết rằng giả thuyết H_0 là đúng
- $P(E)$ là xác suất biên của E
- $P(H_0|E)$ là xác suất hậu nghiệm của H_0 nếu biết E

2.2 Hiện thực

Để có các được thông số β, γ, μ khác đi một chút so với khi hiện thực ở phần 1.2, thay vì chỉ xét số liệu trên phần lãnh thổ đất liền của Mỹ ta sẽ cả xét những phần lãnh thổ khác bao gồm cả những địa phận hải ngoại và đảo.

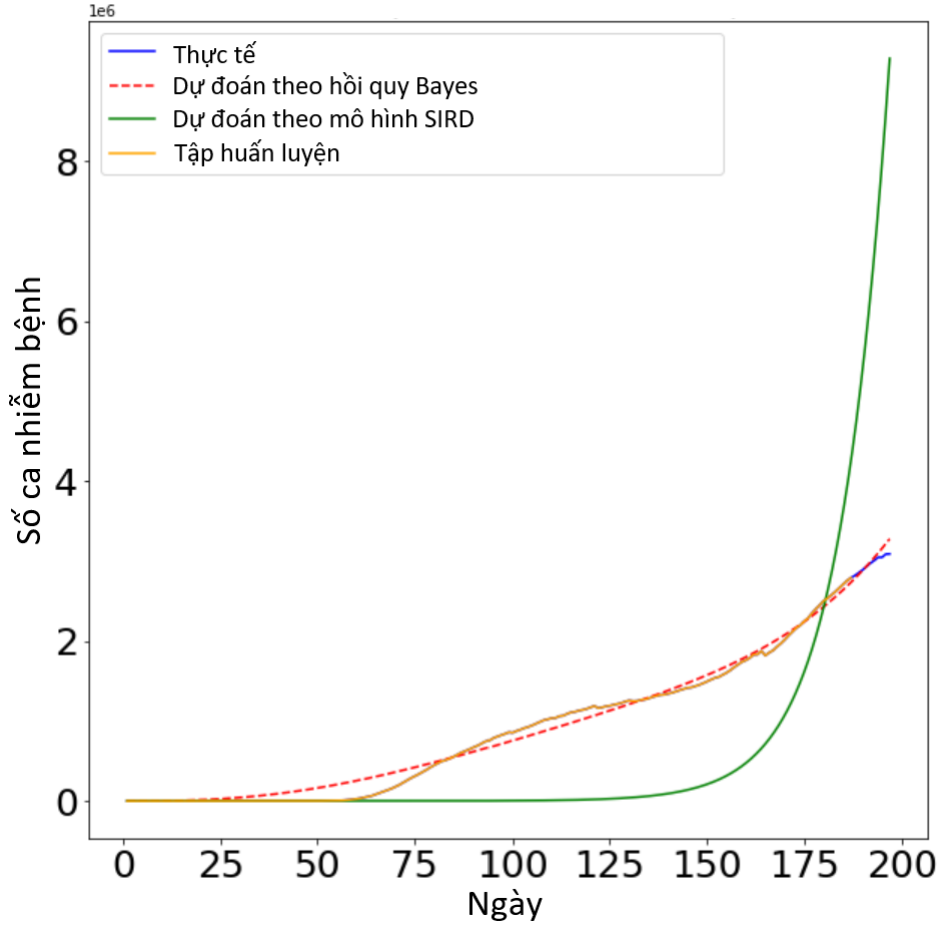
Sau đó, ta cũng tiến hành tiền xử lý các giá trị gây nhiễu và tối ưu hóa giá trị của hàm RMSE như trên. Ta có các giá trị ước lượng mới lần lượt là:

- $\beta = 0.1275683213242846$
- $\gamma = 0.040063523380211796$
- $\mu = 0.005291731933752999$

2.3 Xây dựng mô hình sử dụng hồi quy tuyến tính Bayes đa biến (*Bayes Ridge Polynomial regression*)

2.3.1 Các thông số liên quan

- Thư viện hỗ trợ: Scikit-learn
- Hàm tối ưu: minimize, differential_evolution
- Hàm tiền xử lý (*preprocessing*): PolynomialFeatures
- Hàm kích hoạt: BayesianRidge
- Hàm mất mát: MSE



Hình 17: *So sánh kết quả thu được bằng phương pháp hồi quy tuyến tính Bayes với dữ liệu thực tế*

2.4 Kết quả

Xét số ca nhiễm bệnh COVID-19 ở Mỹ, ta có kết quả thu được từ dự đoán bằng hồi quy tuyến tính Bayes đa biến và dữ liệu thực tế trong 200 ngày được phác thảo như Hình 17

Nhìn Hình 17, ta kết luận rằng kết quả huấn luyện đạt được bởi mô hình này cũng tương đối phù hợp với thực trạng hiện tại đang diễn ra ở Mỹ. Người đọc có thể tham khảo chi tiết thực hiện tại file *SIR_ BayesLinear.ipynb* gửi kèm hoặc tại https://github.com/ToDuyHung/assignment_with_SIRD_model

3 Kết luận

Mô hình SIR/SIRD được sử dụng trong bài báo cáo là mô hình tương đối tiêu chuẩn và đơn giản để mô phỏng lại đại dịch COVID-19. Với các số liệu được cập nhật hằng ngày, mô hình có thể đưa ra con số tương đối để phỏng đoán tình hình trong các ngày tiếp theo. Tuy nhiên, do ở mức đơn giản, không xét đến các biến số quan trọng khác như tình hình cách ly của người dân, tỉ lệ trang bị y tế được cung cấp, các nguồn hỗ trợ,... nên mô hình chỉ mang tính tham khảo hoặc cảnh báo, không nên sử dụng như số liệu chính thống cho các việc quan trọng như ban hành các điều luật, ban bố chính sách từ các Đảng, Nhà nước.

Hiện nay, với các công cụ hiện đại hơn như Học máy và sức mạnh tính toán lớn của các siêu máy tính, việc thêm và cập nhật các biến số trở nên khả thi thì mô hình thu được sẽ có sai số rất nhỏ so với thực tế. Đây là triển vọng rất lớn trong việc phân tích và đưa ra các chính sách phù hợp hơn nhằm đẩy lùi dịch bệnh, khôi phục nền kinh tế và xã hội về lại trạng thái bình thường.

Tài liệu

[Dal] Dalgaard, P. *Introductory Statistics with R.* Springer 2008.

[K-Z] Kenett, R. S. and Zacks, S. *Modern Industrial Statistics: with applications in R, MINITAB and JMP,* 2nd ed., John Wiley and Sons, 2014.

[Ker] Kerns, G. J. *Introduction to Probability and Statistics Using R,* 2nd ed., CRC 2015.

[VINIF] Nguyễn Hoàng Thạch, Phan Thị Hà Dương *Tìm hiểu về một Mô hình dự báo dịch Covid-19 từ Vũ Hán*

[A-J] Abhijit Chakraborty, Jiaying Chen *Analyzing Covid-19 Data using SIRD Models*

[F-S] Frank R. Giordano, William P. Fox, Steven B. Horton *A First Course in Mathematical* 5th ed., 2013.

[Muk] Mukesh Jakhar, P. K. Ahluwalia and Ashok Kumar *COVID-19 Epidemic Forecast in Different States of India using SIR Model*, May 14, 2020

[NDH] NDH COVID-19 Corona Virus South African Resource The Johns Hopkins University *https://www.covid19sa.org*

[TTO] Báo Tuổi trẻ *Mỹ: Ổ dịch COVID-19 lớn nhất thế giới cả về ca nhiễm lẫn tử vong*, April 13, 2020