Tính đúng đắn của mô hình thống kê truyền thống trong việc dự báo số ca nhiễm COVID-19

Nguyễn Ngọc Khôi Nguyên
Khoa Công nghệ thông tin
KHTN - ĐHQG TPHCM
19120106@student.hcmus.edu.vn

Lê Nhật Quang
Khoa Công nghệ thông tin
KHTN - ĐHQG TPHCM
19120340@student.hcmus.edu.vn

Vũ Văn Đô
Khoa Công nghệ thông tin
KHTN - ĐHQG TPHCM
19120007@student.hcmus.edu.vn

Tóm tắt nội dung*—Dịch COVID-19 đã tác động không chỉ đến hệ thống y tế, mà còn những vấn đề như việc làm, an sinh xã hội. Số ca nhiễm mắc COVID-19 mỗi ngày là một tiêu chí quan trọng để đánh giá cấp độ dịch của mỗi vùng trong nước. Dự báo số ca nhiễm, xu hướng dịch bùng phát ở Việt Nam là một vấn đề mà chúng tôi đang tìm hiểu và đánh giá bởi tính chất quan trọng của số ca nhiễm. Thông qua dự báo thì Chính phủ có thể đưa ra các chính sách liên quan đến giãn cách xã hội hay khoanh vùng dịch. Vấn đề dự báo số ca nhiễm đã xuất hiện trong các bài báo khoa học trên thế giới. Chúng tôi đã đưa ra một số mô hình có thể dự đoán số ca nhiễm COVID-19. Chúng tôi đã quyết định khảo sát mô hình hồi quy tuyến tính đa thức và mô hình đối với các hàm phi tuyến, sử dụng hàm loss là Root Mean Square Error (RMSE). Trước đó thì chúng tôi đã xử lý tiền dữ liệu bằng cách dùng một kỹ thuật là làm mịn dữ liệu (Data smoothing). Kết quả thí nghiệm cho thấy mô hình hồi quy đa thức khi kiểm trên tập số ca nhiễm mới đạt 2297 và số ca tử vong mới đạt 53.225. Đối với mô hình liên quan đến phi tuyến thì thấp nhất có chỉ số RMSE đạt 14231.991 khi dự đoán tổng số ca tích lũy. Khi chúng tôi khảo sát 2 mô hình trên, thì RMSE đều khá lớn, do đó các mô hình trên không thực sự hiệu quả, và chỉ mang tính chất tham khảo.

***Index Terms*—COVID-19, Dự báo, Số ca nhiễm, Hồi quy tuyến tính đa thức, hàm phi tuyến**

I. Giới thiệu

A. Động lực

Đại dịch COVID-19 được xem là đại dịch lớn nhất trong lịch sử và Việt Nam phải hứng chịu những biến thể mới như Delta và Omicron. Để hạn chế số ca mắc và số người chết, Chính phủ dã đưa ra nhiều tiêu chí để đánh giá mức độ nghiêm trọng của cấp độ dịch để từ đó đưa ra những giải pháp như giãn cách xã hội. Tổng số ca mắc tích lũy hằng ngày là một tiêu chí quan trọng trong đánh giá mức độ dịch. Bên cạnh việc đếm số ca nhiễm, việc dự báo xu hướng số ca nhiễm rất cần thiết bởi vì căn cứ vào đó mà Chính phủ có thể đưa ra những quyết định quan trọng như số ngày giãn cách xã hội hay phân bổ nguồn lực y tế bao gồm nhân lực, trang thiết bị y tế và vaccine. Động lực lớn nhất khi nghiên cứu đề tài này nằm ở việc xem xét xu hướng số ca nhiễm thông qua một số mô hình toán học đơn giản để từ đó đối sánh với dữ liệu hiện tại.

B. Khó khăn

Trong quá trình nghiên cứu đề tài, chúng tôi gặp không ít khó khăn khi giải quyết về bài toán liên quan đến dự đoán số ca nhiễm. Trước hết, đa số các nguồn dữ liệu đều dưới dạng các thông tin trên báo điện tử hoặc nếu có cũng chỉ là dưới dạng đã qua trực quan hóa thông quan các biểu đồ. Dữ liệu thô hoặc API về số liệu dịch COVID-19 rất hiếm, vì đa số dữ liệu không được công khai rộng rãi trên các trang web, nên việc này gặp một số khó khăn.

Sau khi thu thập được dữ liệu từ các báo cáo hằng ngày, có nhiều vấn đề phát sinh khiến cho dữ liệu gốc trở nên bất thường như tỉ lệ xét nghiệm khác nhau giữa các ngày, báo cáo chậm, dồn số ca bệnh hay không xử lý kịp mã số cho bệnh nhân. Có một số trường dữ liệu không rõ ràng về cách giải thích cũng như tiêu chí như số ca theo ngày tính theo đợt.

C. Phạm vi

Trong phạm vi đề tài tìm hiểu và nghiên cứu, chúng tôi sẽ dự đoán số ca nhiễm tính từ ngày 27/4/2021 đến 10/1/2022. Số ca nhiễm này chính là số ca nhiễm tích lũy lấy từ ngày 27/4/2021.

II. Phương pháp

A. Tập dữ liệu (datasets)

Nguồn tập dữ liệu được thu thập từ trang web của Bộ Y Tế (https://covid19.ncsc.gov.vn/dulieu) thông qua API. Dữ liệu được tổ chức theo dạng file csv. Chúng tôi xét dữ liệu từ ngày 27/04/2021 (Ngày bắt đầu đợt dịch thứ 4). Cấu trúc của dữ liệu như sau:

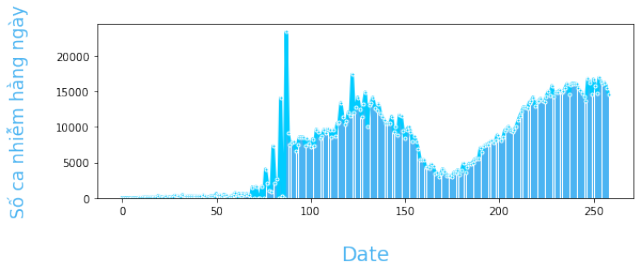
	date	case_by_day	case_by_time	death_by_day	once_injected_by_time	twice_injected_by_time
0		5	5	0	0	0
1		8	13	0	0	0
2		45	58	0	0	0
3		16	74	0	0	0
4		14	88	0	0	0

Hình 1: Dữ liệu

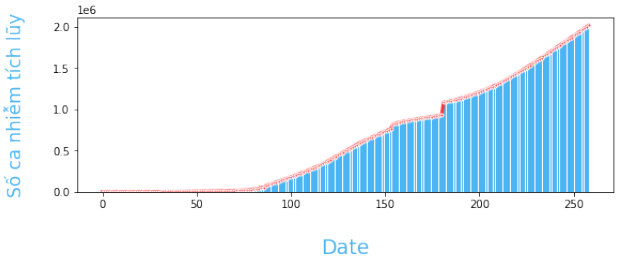
Các thông tin chúng tôi sử dụng bao gồm số ca nhiễm mới hằng ngày (case_by_day), số ca nhiễm tích lũy tính từ ngày 27/04/2021 (case_by_time), số ca tử vong hằng ngày (death_by_day), số người tiêm 1 mũi vaccine mũi (once_injected_by_time), số người tiêm 2 mũi vaccine (twice_injected_by_time). Cột ngày thể hiện số ngày từ ngày 27/04/2021 (ngày 0).

B. *Làm mượt dữ liệu*

Trong các dữ liệu thu thập được từ báo cáo hằng ngày COVID 19 ở Việt Nam, có nhiều vấn đề khiến dữ liệu gốc trở nên bất thường đã nhắc đến ở mục khó khăn. Do đó chúng tôi sẽ sử dụng dữ liệu được làm mịn để làm giảm bớt các sự bất thường này và phản ánh chính xác hơn về xu hướng dịch trong một khoảng thời gian.



Hình 2: Dữ liệu số ca nhiễm hằng ngày gốc

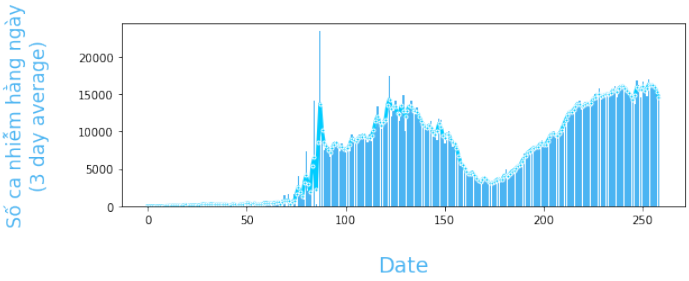


Hình 3: Dữ liệu số ca nhiễm tích lũy gốc

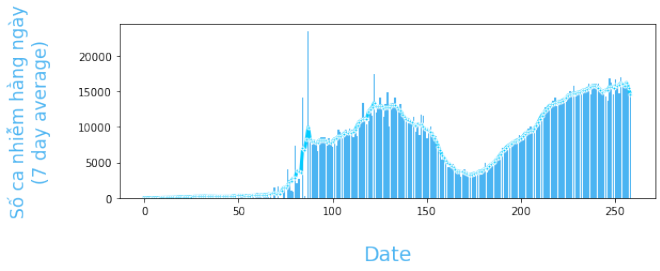
Ta sẽ làm mượt dữ liệu gốc bằng cách lấy trung bình số ca nhiễm trong 7 ngày theo công thức: Số ca nhiễm của ngày thứ n sẽ là số ca nhiễm trung bình của 7 ngày n-3, n-2, n-1, n, n+1, n+2, n+3 (các ngày đầu tiên và cuối cùng sẽ giữ nguyên). Lí do chúng tôi chọn khoảng thời gian là 7 vì:

- Sẽ có một số tỉnh, thành phố có xu hướng tổng hợp số ca nhiễm về 1, 2 ngày cố định trong tuần rồi mới báo cáo.
- 7 ngày là khoảng thời gian để bao đủ 7 ngày trong 1 tuần. Giả sử chọn khoảng thời gian là 3 ngày, và rơi vào thứ hai đến thứ tư. Khi đó ta sẽ thiếu khoảng dữ liệu của thứ 5 - chủ nhật.
- Nếu chọn một chu kỳ lớn hơn 7, như 11 ngày. Thì sẽ khiến khoảng thời gian số ca nhiễm bằng nhau quá dài. Mặt khác ví dụ, một số tỉnh có xu hướng báo cáo dịch vào chủ nhật, và trong chu kỳ 11 ngày đó có 2 ngày chủ nhật, và chu kỳ tiếp theo chỉ có 1 ngày, sẽ ảnh hưởng đến việc quan sát xu hướng dịch và dự đoán.

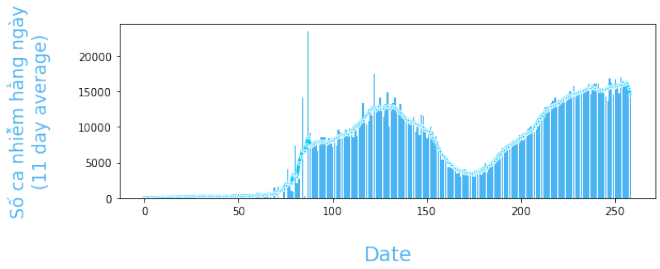
Dưới đây ví dụ về dữ liệu với các khoảng smooth lần lượt là 3, 7 và 11 ngày:



Hình 4: 3 ngày



Hình 5: 7 ngày



Hình 6: 11 ngày

Ta nhận thấy làm mịn với khoảng thời gian 3 ngày ít "mượt mà" hơn với khoảng thời gian 7 ngày. Trong khi nếu chọn lấy trung bình trong 11 ngày thì đồ thị có xu hướng trở nên quá mịn.

C. *Mô hình hồi quy tuyến tính đa thức*

Mô hình hồi quy tuyến tính chỉ hoạt động tốt khi mối quan hệ giữa các biến độc lập có quan hệ xấp xỉ tuyến tính. Trong trường hợp dữ liệu mang tính ngẫu nhiên hoặc bất thường thì mô hình đã nêu không hiệu quả. Mô hình hồi quy tuyến tính đa thức là một bản nâng cấp của hồi quy tuyến tính bởi vì mô hình này sử dụng ít nhất là hàm bậc hai để từ đó đưa ra câu trả lời xấp xỉ. Ở đây trong đề tài chúng tôi đã sử dụng hàm bậc ba được thể hiện dưới đây, trong đó X là biến độc lập và Y là biến phụ thuộc:

$$Y(x) = \theta_0 + \theta_1 X + \theta_2 X^2 + \theta_3 X^3 \quad (1)$$

Ta cũng có thể sử dụng hồi quy đa thức với nhiều biến, khi đó các số hạng ở vế trái (các đặc trưng) sẽ là các tổ hợp đa thức của các biến. Ví dụ, nếu hồi quy với 2 biến a và b thì các đặc trưng sẽ là $[1, a, b, a^2, ab, b^2]$.

D. *Mô hình các hàm toán học phi tuyến*

Trong phần này, chúng tôi sẽ xấp xỉ đồ thị ca nhiễm bằng các hàm toán học để có thể dự đoán được xu hướng của dịch trong tương lai. Tiêu chí là chọn các hàm thể hiện được tốc độ gia tăng phi tuyến, có ít tham số (đơn giản cho việc xác định) và có thể đã được áp dụng cho dữ liệu về dịch bệnh COVID-19 ở các quốc gia, vùng lãnh thổ khác [1].

*1) **Mô hình hàm mũ (Exponential model)**:* Mô hình này giả sử sự gia tăng của các ca nhiễm tích lũy theo quy luật hàm mũ.

$$Y(x) = N_0 * (1+p)^{(x-x_0)} \quad (2)$$

Với N_0 là số ca nhiễm khởi đầu, $1+p$ là tốc độ gia tăng ca nhiễm hằng ngày theo phần trăm. x_0 là hằng số để điều chỉnh đường cong của mô hình.

*2) **Mô hình logistics (Generalized logistics)**:* Mô hình hàm mũ thể hiện sự gia tăng của COVID theo cấp số mũ, tuy nhiên trên thực tế thì số ca nhiễm tích lũy sẽ dần đi đến trạng thái bảo hòa do nhiều nguyên nhân như: miễn dịch toàn dân, giới hạn của dân số, các biện pháp giãn cách,... Do đó ta cần một mô hình phức tạp hơn một chút để thể hiện điều này. Mô hình mà chúng tôi chọn là Generalized logistics:

$$Y(x) = \frac{c}{1+e^{-a(x-b)}} \quad (3)$$

Trong đó c đại diện cho giới hạn tổng số ca nhiễm.

Ta có một phiên bản khác của hàm này để khiến mô hình trở nên linh hoạt hơn:

$$Y(x) = \frac{c}{(1+e^{-a(x-b)})^{\alpha}} \quad (4)$$

*3) **Mô hình Gompertz**:* Một mô hình được đặt theo tên của Benjamin Gompertz, một hàm toán học đơn giản chỉ với 3 tham số. Đây là một mô hình đã được thử nghiệm với số liệu về dịch bệnh COVID ở các quốc gia khác và cho kết quả khá tốt. Về cơ bản, đồ thị của nó cũng tương tự như Generalized logistics. Điểm khác biệt là ban đầu thì tốc độ tăng của hàm Gompertz nhanh hơn Generalized logistics, tuy nhiên khi số ca nhiễm tích lũy càng gần bảo hòa thì tốc độ sẽ càng chậm, và đạt đến đỉnh điểm chậm hơn hàm logistics.

$$Y(x) = ae^{-be^{-cx}} \quad (5)$$

III. THỰC NGHIỆM

A. *Nội dung*

Dữ liệu sẽ được chia thành 2 phần:

- Tập train: Gồm 240 ngày (tính từ ngày 27/04/2021)
- Tập test: Gồm 19 ngày (19 ngày cuối cùng cho đến ngày 10/01/2022)

Chúng tôi đã thực hiện các thí nghiệm dự đoán sau đây:

1) Hồi quy đa thức: Dự đoán số ca nhiễm mới và số ca tử vong dựa trên ngày và tổng số ca mắc tích lũy.

Chúng tôi cho rằng (giả sử) tổng số ca mắc hiện tại có ảnh hưởng đến số ca nhiễm mới và số ca tử vong. Hay nói đơn giản hơn, nếu bạn đang sống trong cộng đồng có càng nhiều ca nhiễm thì tỷ lệ bạn nhiễm bệnh, hay dẫn đến tử vong cũng sẽ tăng lên theo. Nhiệm vụ là tìm ra mối liên hệ này bằng mô hình hồi quy đa thức.

Bài toán

- Input:
  - t (số ngày tính từ ngày 27/04/2021 - ngày 0)
  - c (Số ca nhiễm tích lũy đến ngày t).
- Ouput:
  - Mô hình dự đoán số ca nhiễm mới: Số ca nhiễm mới dự đoán của ngày thứ t
  - Mô hình dự đoán số ca tử vong mới: Số ca tử vong dự đoán của ngày thứ t

Hai mô hình sẽ được train trên tập train để đưa ra tham số phù hợp.

Vì đầu vào cần số ca tích lũy cho đến ngày cần dự đoán. Cho nên với cả 2 mô hình, để dự đoán thông số của ngày t, mô hình cần chạy dự đoán số ca nhiễm của tất cả những ngày trước ngày t để cộng tổng tích lũy lên từ đó tìm được tổng số ca tích lũy ở ngày thứ t.

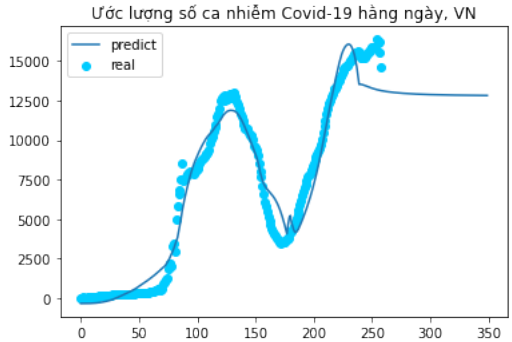
2) Các mô hình phi tuyến: Dự đoán số ca tích lũy theo thời gian và thời điểm đạt đỉnh dịch.

Bài toán: Chúng tôi cố gắng tìm một hàm xấp xỉ tốt với đồ thị số ca nhiễm COVID-19 tích lũy ở Việt Nam trong đợt dịch thứ 4 từ các hàm phi tuyến đã đề cập ở phần Method. Sau đó ta có thể dự đoán số ca nhiễm trong tương lai cũng xấp xỉ với phần còn lại của đồ thị.

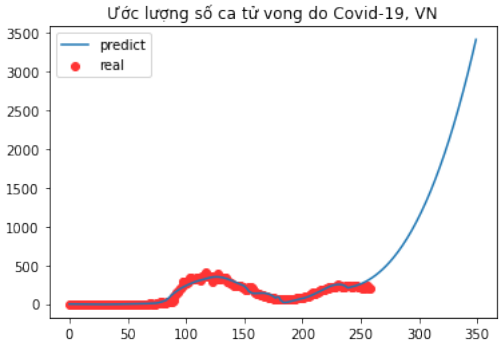
Chúng tôi sử dụng hàm $scipy.optimize.curve\_fit()$ trong python để tìm ra tham số của hàm xấp xỉ với dữ liệu của tập train đưa vào. Các mô hình sẽ được xem xét và đánh giá ở mục Kết quả.

B. *Kết quả*

1) Hồi quy đa thức: Với mô hình hồi quy, chúng tôi thu được đồ thị kết quả dự đoán như sau (bậc tối đa của đa thức là bậc 3)



Hình 7: Dự đoán số ca nhiễm mới theo ngày COVID-19, VN



Hình 8: Dự đoán số ca tử vong theo ngày COVID-19, VN

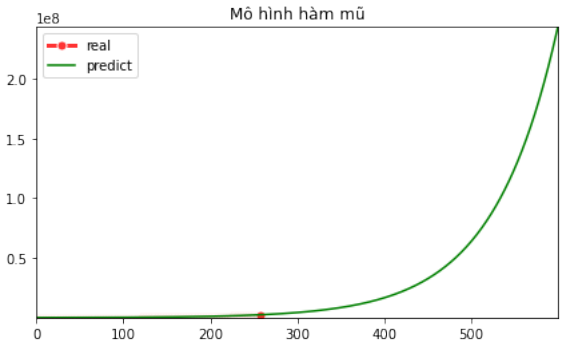
Chúng tôi thu được giá trị RMSE trên tập test của 2 mô hình như sau.

Mô hình	Test - RMSE
Số ca nhiễm mới	2297.252
Số ca tử vong mới	53.225

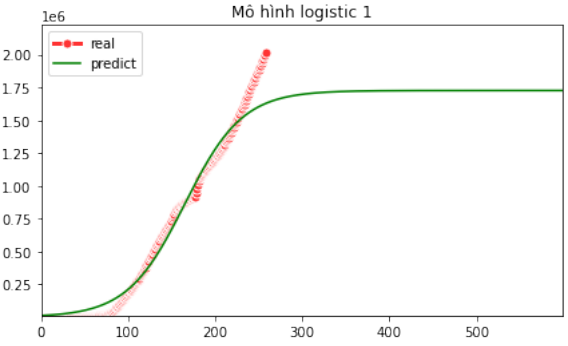
Bảng I: RMSE của mô hình hồi quy đa thức

Từ hình vẽ và sai số RMSE, ta nhận thấy 2 kết quả không hoạt động tốt, sai số còn lớn và có xu hướng overfitting.

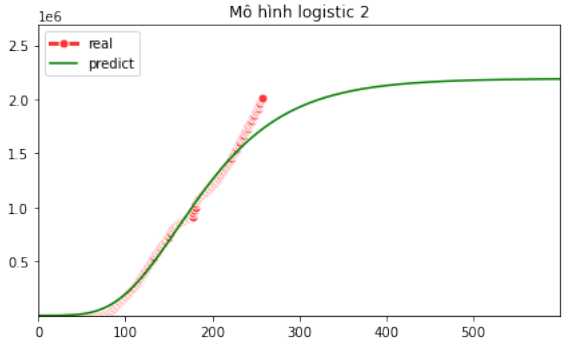
2) Các mô hình phi tuyến: Chúng tôi thu được kết quả của các mô hình với các hàm (2), (3), (4), (5) như hình bên dưới



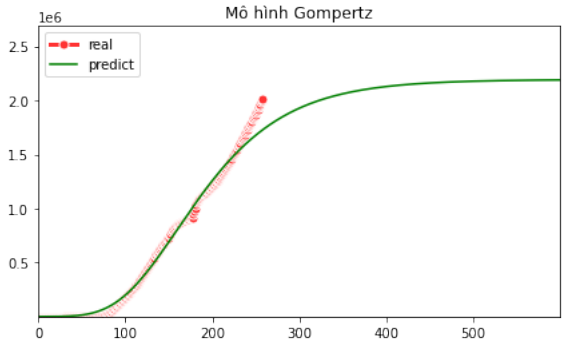
Hình 9: Xấp xỉ với hàm mũ, số ca tích lũy COVID-19, VN



Hình 10: Xấp xỉ với hàm Generalized logistics 1, số ca tích lũy COVID-19, VN



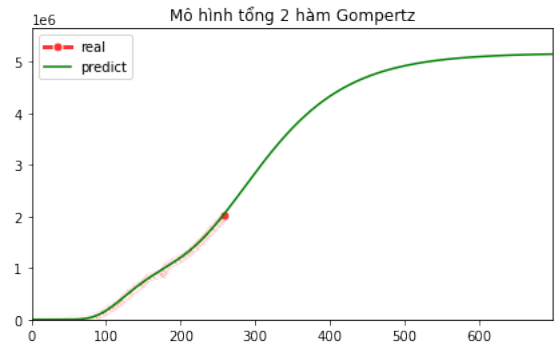
Hình 11: Xấp xỉ với hàm Generalized logistics 2, số ca tích lũy COVID-19, VN



Hình 12: Xấp xỉ với hàm Gompertz, số ca tích lũy COVID-19, VN

Chúng tôi nhận thấy khoảng giữa có đồ thị số ca nhiễm có 1 thời điểm tốc độ tăng trở nên nhanh hơn và sau đó bắt đầu bình ổn trở lại. Theo Gompertz thì đây sẽ là thời điểm số ca nhiễm bắt đầu bình ổn. Tuy nhiên sau đó lại có một đợt bùng phát của số ca nhiễm làm tốc độ tăng thay đổi một lần nữa. Do đó cả mô hình logistics và Gompertz đều không thể "thích ứng" với sự thay đổi này.

Chúng tôi đã thử thay đổi mô hình Gompertz dựa trên công thức gốc bằng cách lấy tổng của 2 hàm Gompertz, với hy vọng tạo xấp xỉ được việc số ca nhiễm bùng phát 2 lần. Kết quả thu được thể hiện ở hình dưới:



Hình 13: Xấp xỉ với tổng 2 hàm Gompertz, số ca tích lũy COVID-19, VN

Bảng đánh giá các giá trị RMSE trên tập test của 2 mô hình như sau:

Mô hình	Train - RMSE	Test - RMSE
Mô hình hàm mũ	126762.138	341906.641
Mô hình hàm Generalized Logistic 1	56297.754	281543.21
Mô hình hàm Generalized Logistic 2	36871.978	206776.743
Mô hình hàm Gompertz	36870.552	206754.822
Mô hình tổng 2 hàm Gompertz	14231.991	14027.122

Bảng II: RMSE của các mô hình phi tuyến

Ta nhận thấy mô hình tổng 2 hàm Gompertz có sai số RMSE nhỏ nhất xấp xỉ tốt nhất với dữ liệu số ca tích lũy bệnh nhân COVID-19 ở Việt Nam.

C. Phân tích

Từ kết quả thực nghiệm, chúng tôi thấy mô hình tuyến tính không hoạt động tốt trong việc dự đoán số ca nhiễm cũng như số ca tử vong mới do COVID-19 gây ra ơ Việt Nam. Lý do chính dẫn đến việc này là:

- Chúng tôi giả định rằng số ca nhiễm mới và số ca tử vong mới phụ thuộc vào số ca tích lũy và số ngày bùng phát dịch. Tuy nhiên thực tế, còn phụ thuộc và nhiều yếu tố như: biến chủng mới, các chính sách phòng chống dịch của từng tỉnh, các ngày dịch bùng phát như lễ tết, sự quá tải của bênh viện, khu cách ly,...
- Mô hình tuyến tính không thể hiện được các biện pháp can thiệp của chính phủ, cơ quan chức năng trong các trường hợp khác nhau
- Thiếu các yếu tổ về lây nhiễm, dịch tể học.

Trong các mô hình toán phi tuyến, ta nhận thấy chỉ có mô hình xấp xỉ bằng tổng 2 hàm Gompertz (Hình 12) là có vẻ phù hợp với số ca nhiễm tích lũy ở Việt Nam, và đưa ra một dự đoán về thời điểm số ca nhiễm bắt đầu bảo hòa khoảng vào ngày thứ 550, đó là thời điểm sẽ bắt đầu kết thúc dịch (khoảng tháng 10/2022). Mô hình hàm Gompertz cũng đã được cho thấy phù hợp với số liệu ở các vùng lãnh thổ khác.

Tuy nhiên mô hình này vẫn gặp các vấn đề như mô hình hồi quy:

- Không phù hợp dự đoán cho các trường hợp bất thường như biến chủng mới, sự can thiệp trong chống dịch của nhà nước (Nếu có ta phải giả định xem sự can thiệp xảy ra vào thời điểm nào, thêm các tham số không chắc chắn)
- Thiếu các yếu tố, nguyên lý về dịch tể học.

Cho nên mô hình Gompertz chỉ có tính chất tham khảo. Không nên quá tin tưởng áp dụng vào thực tế.

IV. KẾT LUẬN

Mặc dù các mô hình thống kê truyền thống được sử dụng như hồi quy tuyến tính đa thức hay logistics bởi ưu điểm là tính đơn giản và ít tham số tinh chỉnh, nhưng nhược điểm lớn nhất thể hiện rõ trong đại dịch có các biến thể mới xuất hiện cùng với những yếu tố khác tác động dến dịch, bên cạnh đó không nắm rõ yếu tố liên quan đến dịch tễ, dẫn đến việc mô hình trên không hiệu quả trong việc dự báo [2]. Hai mô hình được tìm hiểu bao gồm liên quan đến hồi quy và phi tuyến. Đối với hồi quy đa thức, sai số còn khá lớn, RMSE của số ca nhiễm mới nằm ở 2297.252, còn đối với các mô hình về phi tuyến, trong đó tổng 2 hàm Gompertz có RMSE đạt 14027.122 đối với trường hợp tổng số ca nhiễm. Khi phân tích các mô hình trên, thì các mô hình nói chung không xét những yếu tố đi kèm với dịch COVID như các bệnh đi kèm, độ tuổi, yếu tố nguy cơ,.. Do đó, các mô hình học thống kê trên mang tính chất tham khảo đối với người đọc.

TÀI LIỆU

[1] M. Villalobos-Arias, "Using generalized logistics regression to forecast population infected by covid-19," *arXiv preprint arXiv:2004.02406*, 2020.

[2] N. V. Tuấn, "Tất cả mô hình tiên lượng covid đều sai," https://nguyenvantuan.info/2021/08/24/tat-ca-mo-hinh-tien-luong-covid-deu-sai/.