TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
ĐẠI HỌC QUỐC GIA TPHCM
KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO ĐỒ ÁN CUỐI KÌ - TÌM HIỂU CÁC MÔ HÌNH THỐNG KÊ ÁP DỤNG CHO COVID19

Môn học: **Toán ứng dụng và thống kê**

Thực hiện:

1712060 - Trần Vinh Hưng

1712702 - Nguyễn Hà Quang

MỤC LỤC

1. Các mô hình đã tìm hiểu

1.1. SIR

Mô hình SIR được sử dụng lần đầu bởi Kermack và McKendrick năm 1927 và được áp dụng để mô hình hóa rất nhiều căn bệnh truyền nhiễm, đặc biệt là các căn bệnh mà người nhiễm đạt được miễn dịch vĩnh viễn sau khi khỏi bệnh. SIR cho ta có được cái nhìn tổng quát về tốc độ truyền nhiễm, tỉ lệ dân số bị nhiễm, tỉ lệ tử vong hoặc khỏi bệnh. Để có được điều này, mô hình SIR chia dân số thành 3 thành phần tách biệt:

Susceptible: đối tượng chưa nhiễm bệnh đồng thời có khả năng bị nhiễm nếu tiếp xúc với người bị nhiễm

Infected: những người đang nhiễm bệnh và có khả năng lây nhiễm cho những người thuộc nhóm Susceptible

Recovered: đã nhiễm bệnh nhưng không còn khả năng lây nhiễm (tức là đã được chữa khỏi hoặc tử vong)

Trong đó: ***S + I + R = N = tổng dân số***

1.1.1 Mô hình SIR trên tập dân số không đổi

Giả sử chúng ta xem xét một tập dân số có N = 1000 người và biết rằng có 400 người đang nhiễm bệnh vào thời gian t (ví dụ t = 7 ngày kể từ ngày dịch bệnh bùng nổ). Theo mô hình SIR ta có S(7) = 400.

Mô hình SIR sẽ cho phép chúng ta chỉ cần đưa vào một số tham số và bộ giá trị $S(t_0)$, $I(t_0)$, $R(t_0)$, từ đó tính được tất cả các giá trị S(t), I(t), R(t) của tất cả các ngày sau t0. Chúng ta sẽ cùng xem xét các biến số và tham số của mô hình SIR thông qua một ví dụ đơn giản sau:

Chúng ta có một căn bệnh mới là X. Đối với X, tỉ lệ một người bị nhiễm bệnh lây nhiễm cho một người khỏe mạnh khi tiếp xúc là 20%. Trung bình 1 người tiếp xúc với 5 người khác mỗi ngày. Vì thế, mỗi ngày, mỗi cá thể đang nhiễm bệnh (người thuộc nhóm I) sẽ gặp 5 người và lây nhiễm cho β = 20% x 5 = 1 (người) trong số đó. Tuy nhiên không phải tất cả những người mà người này (nhóm I) gặp đều thuộc nhóm S. Ta sẽ phải bổ sung vào công thức một đại lượng biểu diễn xác suất 1 người mà người nhiễm gặp là người thuộc nhóm S, xác suất này là S/N.

Như vậy, chúng ta có công thức biểu thị số người bị nhiễm bệnh mới mỗi ngày sẽ là: I . β . S / N

Nói cách khác, mỗi ngày, số người thuộc nhóm S sẽ thay đổi một lượng là:

- I . β . S / N

Chú ý rằng, lượng thay đổi (giảm) này của nhóm S sẽ được chuyển vào nhóm I.

Tới đây ta chỉ cần xem xét thêm số lượng người khỏi bệnh mỗi ngày. Giả sử thời gian trung bình tính từ thời điểm một bệnh nhân nhiễm bệnh (bắt đầu có khả năng lây nhiễm) cho đến khi khỏi bệnh (không còn khả năng lây nhiễm) là D. Từ đó ta kì vọng rằng mỗi ngày, số người nhiễm chuyển thành khỏi bệnh sẽ là I / D = I . γ. Với γ = 1 / D.

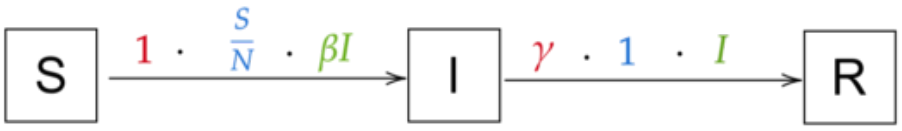
Tới đây, ta có thể kết luận rằng, mỗi ngày, số người thuộc nhóm R sẽ thay đổi một lượng là:

I . γ

Vì thế, mỗi ngày, số người thuộc nhóm I sẽ thay đổi một lượng là:

I . β . S / N − I . γ

Sơ đồ sau sẽ giúp ta hiểu rõ hơn về mô hình SIR:



Từ những phân tích ở trên, ta có hệ phương trình vi phân biểu diễn mô hình SIR:

$$\frac{dS}{dt} = -\frac{\beta SI}{N}$$
$$\frac{dI}{dt} = \frac{\beta SI}{N} - \gamma I$$
$$\frac{dR}{dt} = \gamma I$$

Xây dựng mô hình SIR cho dịch bệnh COVID19 là quá trình đưa vào các vector X(t) = (S(t), I(t), R(t)) dựa theo số liệu thống kê theo ngày của dịch bệnh COVID19 (số liệu lấy từ WHO) và trả về các tham số β và γ tối ưu nhất. Với các tham số đã tìm được, ta có thể một mô hình có thể dự đoán các giá trị S, I, R cho những ngày tiếp

theo. Độ chính xác của dự đoán phụ thuộc vào độ tối ưu của các tham số β và γ. Ta sẽ phân tích quá trình tìm tham số tối ưu ở phần sau.

1.1.2 Mô hình SIR trên tập dân số thay đổi

Trên thực tế, trên một tập dân số lớn, N không phải là một hằng số mà là một biến số thay đổi từng ngày. Sự thay đổi của N phụ thuộc vào tỉ lệ sinh μ và tỉ lệ tử υ, và do đó, hệ phương trình vi phân biểu diễn mô hình SIR trong trường hợp này sẽ là:

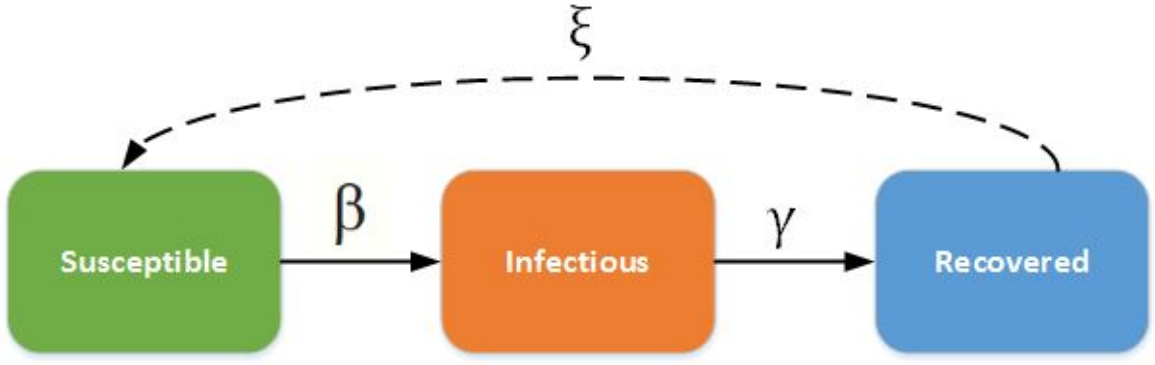
$$\frac{dS}{dt} = \mu N - \frac{\beta SI}{N} - \nu S$$
$$\frac{dI}{dt} = \frac{\beta SI}{N} - \gamma I - \nu I$$
$$\frac{dR}{dt} = \gamma I - \nu R$$

1.2. SIRS

SIRS là một biến thể của mô hình SIR dùng để mô hình hóa các căn bệnh mà người nhiễm sau khi khỏi bệnh không đạt được miễn dịch vĩnh viễn. Những người khỏi bệnh (nhóm R) chỉ đạt miễn dịch trong một thời gian E, sau đó sẽ bị mất miễn dịch và trở lại nhóm S. Vì thế, ta kì vọng rằng, mỗi ngày, số người thuộc nhóm R chuyển sang nhóm S sẽ là: R / E = ***ε*.R**

Với ε = 1 / E

Sơ đồ sau sẽ giúp ta hiểu rõ hơn về mô hình SIR:



Hệ phương trình vi phân biểu diễn mô hình SIRS trong trường hợp dân số không đổi là:

$$\frac{dS}{dt} = -\frac{\beta SI}{N} + \xi R$$
$$\frac{dI}{dt} = \frac{\beta SI}{N} - \gamma I$$
$$\frac{dR}{dt} = \gamma I - \xi R$$

Trong trường hợp dân số thay đổi, hệ phương trình vi phân biểu diễn mô hình SIRS là:

$$\frac{dS}{dt} = \mu N - \frac{\beta SI}{N} + \xi R - \nu S$$
$$\frac{dI}{dt} = \frac{\beta SI}{N} - \gamma I - \nu I$$
$$\frac{dR}{dt} = \gamma I - \xi R - \nu R$$

1.3. SIRD

SIRD là một phiên bản đầy đủ hơn so với mô hình SIR. Trong mô hình này, trạng thái R của SIR sẽ được tách thành 2 trạng thái là R (recovered - những bệnh nhân đã được chữa khỏi) và D (dead - những bệnh nhân đã tử vong). Tóm lại, mô hình SIRD chia dân số thành 4 trạng thái:

Susceptible: đối tượng chưa nhiễm bệnh đồng thời có khả năng bị nhiễm nếu tiếp xúc với người bị nhiễm

Infected: những người đang nhiễm bệnh và có khả năng lây nhiễm cho những người thuộc nhóm Susceptible

Recovered: nhiễm bệnh nhưng đã được chữa khỏi, không có khả năng tái nhiễm.

Dead: bệnh nhân đã tử vong

Trong đó: ***S + I + R + D = N = tổng dân số***

Hệ phương trình vi phân biểu diễn mô hình SIRD là:

$$\frac{dS}{dt} = -\frac{\beta IS}{N},$$
$$\frac{dI}{dt} = \frac{\beta IS}{N} - \gamma I - \mu I,$$
$$\frac{dR}{dt} = \gamma I,$$
$$\frac{dD}{dt} = \mu I,$$

Tương tự như ở mô hình SIR, ý nghĩa của các tham số trong phương trình vi phân trên lần lượt là:

- β - hệ số lây truyền
- γ - tỉ lệ chữa khỏi bệnh
- μ - tỉ lệ tử vong

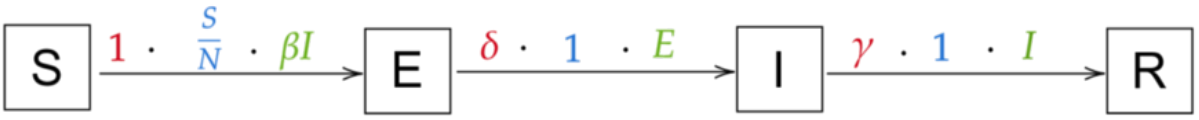
1.4. SEIR

Nhiều bệnh truyền nhiễm có thời gian ủ bệnh trước khi lây nhiễm, trong thời gian đó vật chủ không thể truyền bệnh. Mô hình SEIR xem xét những cá nhân như vậy bởi một biến số mới: **E**xposed - người phơi nhiễm.

Giả sử thời gian ủ bệnh trung bình là F, kỳ vọng số người chuyển từ trạng thái E sang I mỗi ngày sẽ là: E / F = **δ.E**

Với δ = 1 / F

Nếu đã biết về mô hình SIR, ta có thể dễ dàng hiểu được mô hình SEIR thông qua sơ đồ sau:



Hệ phương trình vi phân biểu diễn mô hình SEIR là:

$$\frac{dS}{dt} = -\beta \cdot I \cdot \frac{S}{N}$$
$$\frac{dE}{dt} = \beta \cdot I \cdot \frac{S}{N} - \delta \cdot E$$
$$\frac{dI}{dt} = \delta \cdot E - \gamma \cdot I$$
$$\frac{dR}{dt} = \gamma \cdot I$$

1.5. Sử dụng mô hình SIR và SIRD

- Dễ cài đặt: mô hình SIR trên tập dân số không đổi có số lượng tham số và biến số nhỏ, vì thế nó dễ cài đặt hơn so với các mô hình còn lại.

- Ngoài ra, ta không xem xét mô hình trên tập dân số thay đổi bởi vì đối với bối cảnh dịch bệnh COVID19, số ngày kể từ lúc bùng nổ dịch bệnh chưa đủ lớn để những thay đổi dân số (tỉ lệ sinh tử) gây ảnh hưởng đến mô hình dự đoán.

- Có nhiều nghiên cứu cho rằng người nhiễm COVID19 sau khi khỏi bệnh sẽ không có được miễn dịch vĩnh viễn, yếu tố này gợi ý ta rằng sử dụng mô hình SIRS cho COVID19 là phù hợp hơn so với SIR. Tuy nhiên, thời gian kể từ lúc bùng nổ dịch bệnh đến hiện tại là chưa đủ lớn để những bệnh nhân đã khỏi bệnh bị mất miễn dịch nên mô hình SIRS là không cần thiết.

- Không sử dụng SEIR vì hạn chế về dữ liệu: một cách trực quan, chúng ta có thể thấy rằng mô hình SEIR tốt hơn SIR trong việc mô hình hóa dịch bệnh COVID19. Tuy nhiên, để thiết lập mô hình SEIR, ta cần có dữ liệu người phơi nhiễm - Exposed, tức là những người đã mắc bệnh nhưng chưa có khả năng lây nhiễm. Dữ liệu hiện tại chỉ có số liệu những người đang mắc bệnh, hạn chế này buộc nhóm quyết định không sử dụng mô hình SEIR mà chỉ dùng SIR.

- Mô hình SIRD tách trạng thái R trong SIR thành R và D, vì thế nó mô phỏng dịch bệnh một cách đầy đủ hơn so với SIR. **Nhóm em sẽ phân tích cách tìm tham số của mô hình SIR và tiến hành cài đặt thực nghiệm cả 2 mô hình SIR và SIRD trên tập dữ liệu COVID19.**

2. Các phương pháp tìm tham số trong mô hình SIR

Nhắc lại, hệ phương trình vi phân biểu diễn mô hình SIR:

$$\frac{dS}{dt} = -\frac{\beta SI}{N}$$
$$\frac{dI}{dt} = \frac{\beta SI}{N} - \gamma I$$
$$\frac{dR}{dt} = \gamma I$$

Ta có thể xem $dt = 1$ (ngày), từ đó ta có hệ mới như sau:

$$S(t+1) = S(t) - \frac{\beta S(t)I(t)}{N}$$
$$I(t+1) = I(t) + \frac{\beta S(t)I(t)}{N} - \gamma I(t)$$
$$R(t+1) = R(t) + \gamma I(t)$$

Ở đây, β và γ là 2 ẩn số mà ta cần tìm. Các giá trị $S(t+1),\ S(t),\ I(t+1),\ I(t)$ và $R(t+1),\ R(t)$ là các tham số (nhận vào từ dataset) của hệ phương trình.

Giả sử input dataset gồm có $(n+1)$ ngày, bài toán của chúng ta là tìm 2 giá trị β và γ sao cho thỏa mãn "tốt nhất" tất cả các hệ phương trình trên với mọi giá trị của t từ 0 đến $(n-1)$.

2.1. Trung bình

Giả sử input của chúng ta gồm có 2 ngày t = 0 và t + 1 = 1, khi đó chúng ta chỉ cần giải hệ phương trình sau (với β và γ là ẩn số):

$$S(1) = S(0) - \frac{\beta S(0)I(0)}{N} \quad (1)$$
$$I(1) = I(0) + \frac{\beta S(0)I(0)}{N} - \gamma I(0) \quad (2)$$
$$R(1) = R(0) + \gamma I(0) \quad (3)$$

Hệ trên có 2 ẩn nhưng chứa 3 phương trình, rất có thể nó sẽ vô nghiệm. Tuy nhiên, ta dễ dàng nhận thấy phương trình (2) chỉ là hệ quả của hai phương trình (1) và (3), vì: $(2) = -(1) - (3)$

Vì thế, hệ trên có nghiệm là: $\beta = \frac{(S(0) - S(1)) \cdot N}{S(0) \cdot I(0)}$; $\gamma = \frac{R(1) - R(0)}{I(0)}$

Tới đây, ta có thể đưa ra một cách tiếp cận vô cùng đơn giản cho việc tìm giá trị của β và γ là: giá trị trung bình.

Gọi $\beta(t)$ và $\gamma(t)$ là nghiệm của hệ:

$$S(t+1) = S(t) - \frac{\beta(t)S(t)I(t)}{N}$$
$$R(t+1) = R(t) + \gamma(t)I(t)$$

Với input dataset gồm có $(n+1)$ ngày, thông qua việc giải n hệ phương trình, ta tìm được n bộ giá trị $\beta(t)$, $\gamma(t)$ (với $t = 0..n$)

Tới đây, ta chọn giá trị β và γ bằng cách tính trung bình các giá trị $\beta(t)$ và $\gamma(t)$.

Tức là:

$$\beta = \frac{1}{n}.\sum_{t=0}^{n}\beta(t) \quad ; \qquad \gamma = \frac{1}{n}.\sum_{t=0}^{n}\gamma(t)$$

2.2. Gradient descent

Trong các bài toán máy học, việc cần phải tìm công thức cho điểm tối ưu trong hàm mất mát là một công việc phức tạp, và cần phải tính toán lại với mỗi công thức hàm lỗi cũng như hàm truyền tới khác nhau. Do đó, để tiện cho việc mở rộng công thức, chúng em còn cài đặt thêm các giải thuật cho việc tối ưu hàm lỗi một cách tùy biến.

Dù trước đây có rất nhiều giải thuật tối ưu khác nhau, nhưng bọn em sẽ cài đặt giải thuật đơn giản và phổ biến nhất hiện nay, Gradient Descent.

Đặt f là hàm mất mát. Ta có:
$f'(x) > 0$ => Hàm f hiện đang tăng khi x tăng => Cần giảm x để giảm f
$f'(x) < 0$ => Hàm f hiện đang giảm khi x tăng => Cần tăng x để giảm f

=> Để tìm cực trị, ta cần tăng x ngược chiều với f'.

Do đó, ta có công thức: $x = x - \alpha * f'(x)$, với hệ số học α là một số rất nhỏ.

Giải thuật Gradient Descent có nhiều nhược điểm khi phụ thuộc vào α, cách bọn em chọn α bằng cách tính ra gradient tại x khởi tạo, và xấp xỉ

$$\alpha = 10^{-2} * f'(x\ khởi\ tạo).$$

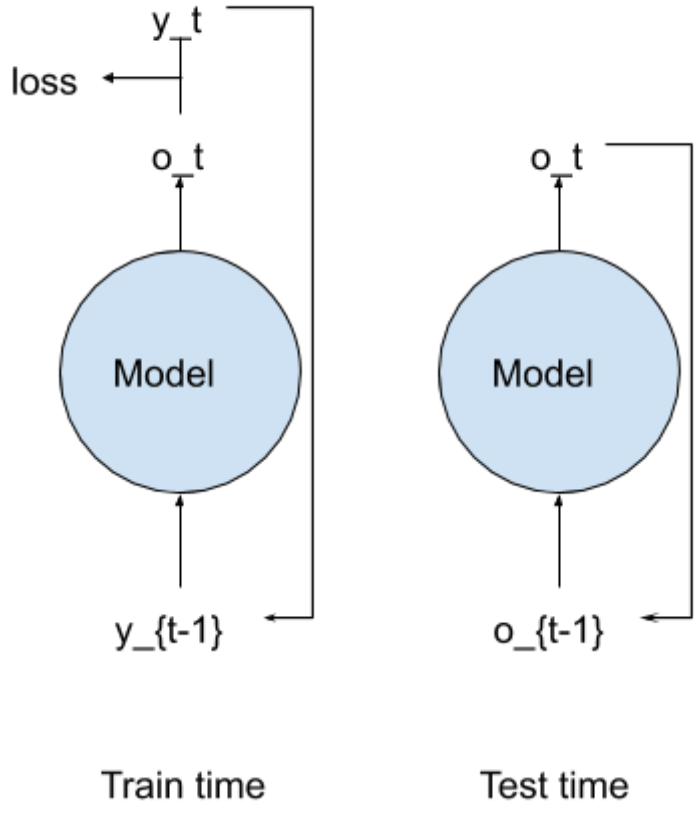
Trong bài toán SIR, hàm tối ưu bọn em chọn là khoảng cách Euclid giữa kết quả dự đoán và chân trị.

2.3. Tìm cực trị bằng đạo hàm

Gọi $\hat{S}, \hat{I}, \hat{R}$ là các dãy giá trị dự đoán (S, I, R là giá trị thực).

Ta có: $\hat{S}_t \approx S_t$.

Ta có thể tạm xem trong quá trình học, mô hình sẽ nhận đầu vào là giá trị thực tế. (Phương pháp Teacher Forcing trong giải quyết bài toán Timeseries).



Khi đó, mô hình có thể xem như nhận các đầu vào độc lập nhau và xem đây như bài toán hồi quy bình thường.

Ta sẽ tối ưu kết quả trên khoảng cách euclide giữa đầu ra và giá trị thực tế.

$$L(x, y \mid \theta) = (f(x \mid \theta) - y)^2$$

$$L(S, I, R \mid \beta, \gamma) = \frac{1}{2N} \sum_{t=1}^{N} ((S_t - \hat{S}_t)^2 + (I_t - \hat{I}_t)^2 + (R_t - \hat{R}_t)^2)$$

$$L(S,\ I,\ R \mid \beta,\ \gamma)\ = \frac{1}{2N} \sum_{t=0}^{N-1} ((S_{t+1} - S_t + \beta \frac{S_t I_t}{N})^2\ +\ (I_{t+1}\ - I_t - \frac{\beta}{N} S_t\, I_t + \gamma I_t)^2 + (R_{t+1} - R_t - \gamma I_t)^2)$$

Ta cần tìm β và γ sao cho L nhỏ nhất.

Do đó β và γ là nghiệm của hệ phương trình:

$$\frac{\delta L}{\delta \beta}\ =\ 0$$

$$\frac{\delta L}{\delta \gamma}\ =\ 0$$

Giải hệ trên, ta thu được kết quả:

$$\gamma\ = \frac{BC + AE}{B^2 + AD}$$

$$\beta\ = \frac{C - B(BC+AE)/(B^2+AD)}{A}$$

Với:

$$A\ =\ -2 \sum_{t=1}^{N} (\frac{S_t I_t}{N})^2$$

$$B\ =\ \sum_{t=1}^{N} \frac{S_t I_t^2}{N}$$

$$C\ =\ \sum_{t=1}^{N} \frac{S_t I_t}{N} (S_{t+1} - S_t - I_{t+1} + I_t)$$

$$D\ =\ 2 \sum_{t=1}^{N} I_t^2$$

$$E\ =\ \sum_{t=1}^{N} I_t (R_{t+1} - R_t - I_{t+1} + I_t)$$

3. Thí nghiệm với mô hình SIR

3.1. Bố trí thí nghiệm: US và cả thế giới

3.1.1 Thu thập & tiền xử lí dữ liệu

- Thu thập dữ liệu số ca nhiễm từ Internet (Link đính kèm bên dưới). Dữ liệu mà nhóm em chọn là dữ liệu chính thức được cung cấp bởi WHO.
- Để tiện cho việc xử lý, bọn em tách file dữ liệu thô cung cấp bởi WHO ra định dạng riêng như sau:
    - Mỗi nước là một file csv mang tên nước
    - Mỗi dòng tượng trưng cho một ngày
    - Dữ liệu gồm 3 cột chính: Confirmed (số ca đã xác nhận), Death (Số ngươi chết), Recovered (số ca khỏi).
- Sau đó, bọn em thu thập dữ liệu dân số các nước thông qua nguồn dữ liệu [3]
- Đối với dữ liệu thế giới, bọn em gom ngày của các nước sau đó tính tổng.

    Ta có công thức như sau:

    $$R = recover + death$$
    $$I = confirmed - R$$
    $$S = population - I - R$$

- Bọn em đều chọn khung ngày 100 đến 130 làm tập huấn luyện và từ ngày 131 đến 160 là tập kiểm nghiệm mô hình.

3.1.2 Đánh giá thí nghiệm

Bọn em đánh giá kết quả mô hình thông qua:

i. Độ lệch trung bình tuyệt đối: (Mean Absolute Error)

$$MAE(y, \hat{y}) = \frac{1}{n} \sum_{t}^{\{s,i,r\}} \sum_{j=1}^{n} \left| y_{tj} - \hat{y}_{tj} \right|$$

Với $y, \hat{y}$ lần lượt là chân trị và kết quả dự đoán.

ii. Độ lệch trung bình tuyệt đối đã được bình thường hóa (Normalized Mean Absolute Error):

$$NMAE(y, \hat{y}) = \frac{1}{n} \sum_{t}^{\{s,i,r\}} \sum_{j=1}^{n} \left| \frac{y_{tj} - \hat{y}_{tj}}{y_{tj}} \right|$$

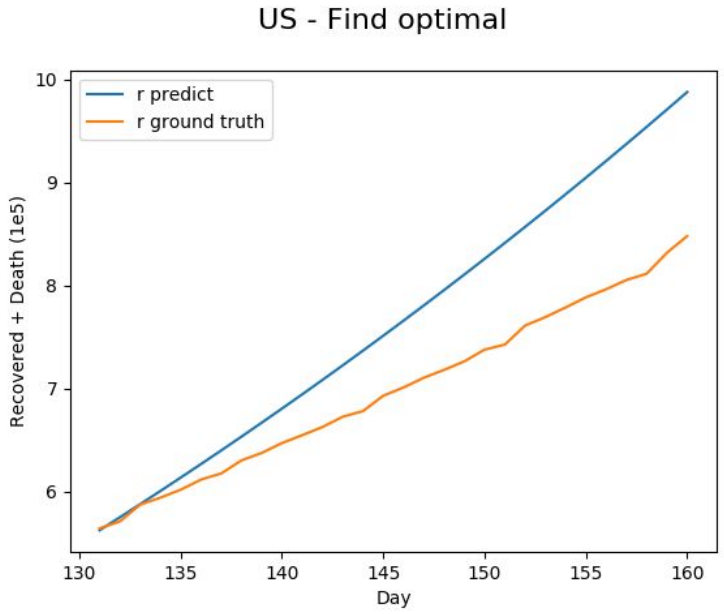
Vì độ chênh lệch lớn giữa S, I, R, để công bằng hơn cho việc đánh giá, bọn em thêm một trọng số để bình thường hóa kích cỡ của S, I, R.

3.2. Kết quả tính

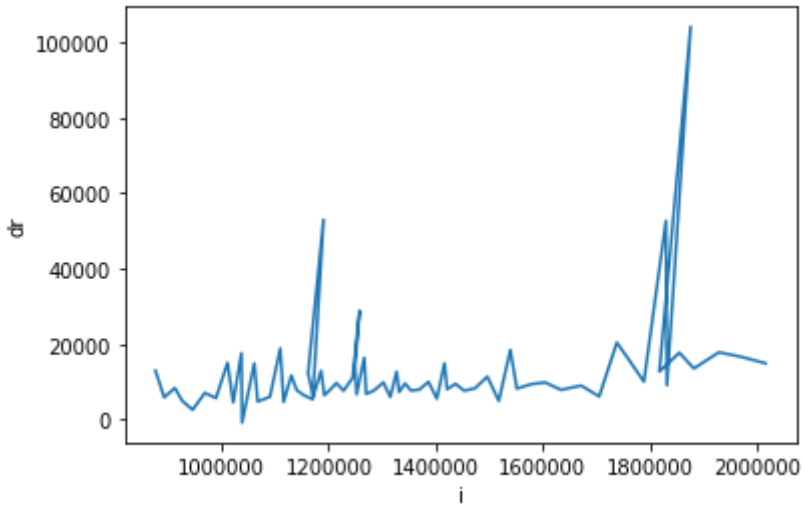
3.2.1. Kết quả tính của US:

Tên phương pháp	MAE	NMAE
Trung bình	2063273.0	1.1142
Gradient Descent	1613623.0	0.9726
Tìm cực trị bằng đạo hàm	1799687.666	1.0971

Trong lúc tính toán, bọn em để ý thấy đối với dữ liệu của US, khả năng dự đoán R không tốt lắm mặc dù đã dùng phương pháp tìm điểm xấp xỉ R tốt nhất của mô hình SIR bằng công thức được chứng minh cụ thể qua đạo hàm.



Do đó, bọn em quyết định kiểm tra sâu hơn về dữ liệu. Tại đây, bọn em đã hiểu tại sao lại có hiện tượng bất thường như thế:



Đồ thị tương quan giữa I và dR.

Ta có:

$dR = \beta * I$, hay dR tuyến tính với I. Tuy nhiên, đồ thị trên không thể hiện điều đó
=> Mô hình SIR không thích hợp với dữ liệu COVID19 tại US.

3.2.2. Kết quả tính của thế giới:

Tên phương pháp	MAE	NMAE
Trung bình	3621413.0	0.78424
Gradient Descent	3160206.67	0.6620
Tìm cực trị bằng đạo hàm	3161338.0	0.6603

3.3. Kết luận

Ta có thể thấy kết quả dự đoán của thế giới tốt hơn so với các kết quả dự đoán của US. Theo nhóm em, sự chênh lệch này xuất phát từ dữ liệu.

Dữ liệu của thế giới có tính bao quát hơn nếu so với dữ liệu đặc thù tại US. Tại US có rất nhiều tham số trong dịch bệnh lần này như:

- Quá tải về bệnh viện trong thời gian đầu
  (Chỉ người giàu mới được khám bệnh)

- Người dân chưa ý thức được dịch bệnh
- Sự sai lệch trong ghi chép giữa các bang

Những nguyên nhân này đã ảnh hưởng xấu đến vận tốc lành bệnh của bệnh nhân cũng như dữ liệu.

Có lẽ chúng ta sẽ cần một mô hình với nhiều đầu vào hơn để phù hợp hơn cho đặc thù tại nước Mỹ.

Tham khảo

Nguồn dữ liệu:

[1] https://www.kaggle.com/imdevskp/corona-virus-report

[2] https://github.com/CSSEGISandData/COVID-19

[3] https://www.worldometers.info/world-population/population-by-country/

Tài liệu về các mô hình:

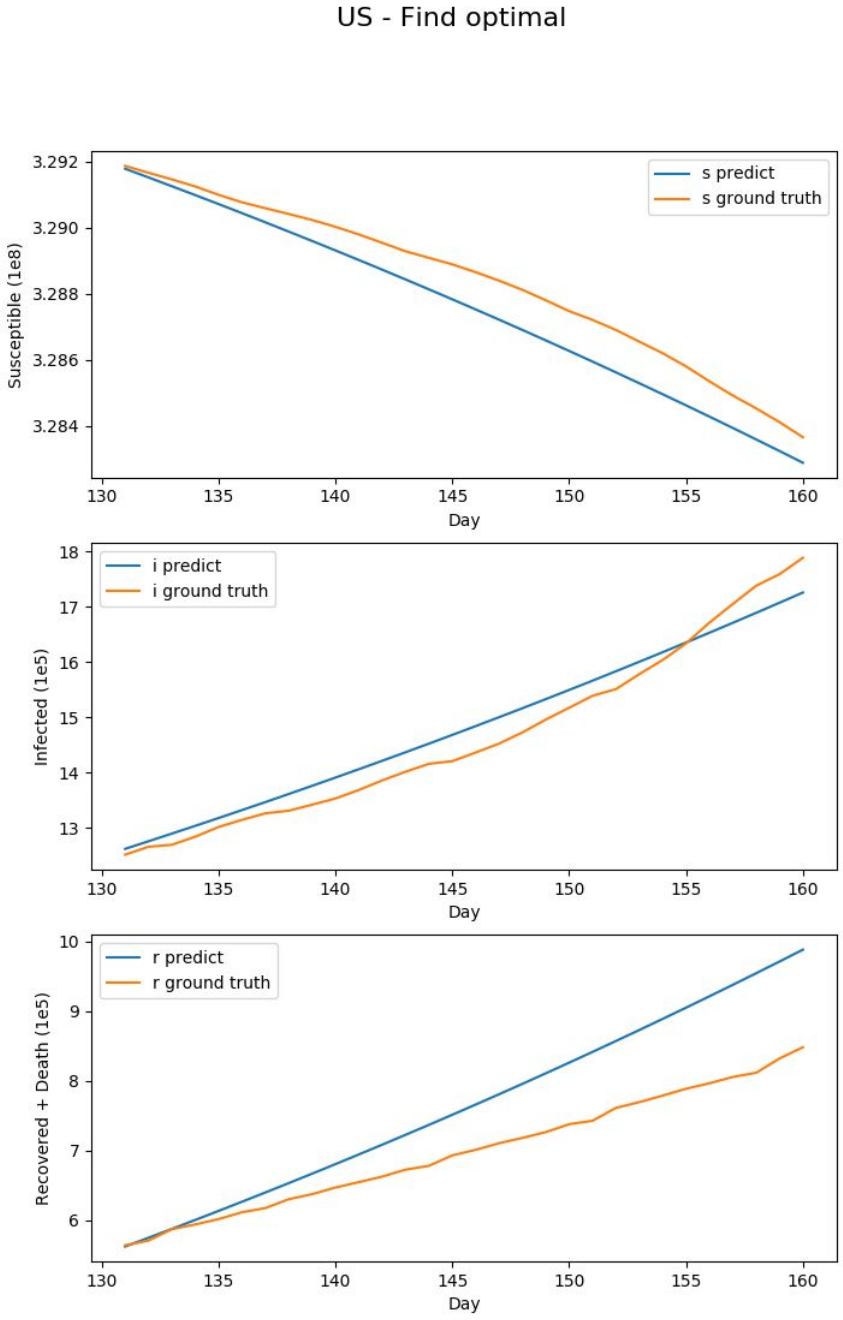
[4] https://idmod.org/docs/general/model-sir.html

[5] https://towardsdatascience.com/infectious-disease-modelling-part-i-understanding-sir-28d60e29fdfc

[6] https://towardsdatascience.com/infectious-disease-modelling-beyond-the-basic-sir-model-216369c584c4

[7] https://medium.com/topos-ai/disease-models-differential-equations-connecting-geographies-with-time-series-clustering-8f91d3545876

Phụ lục: Các biểu đồ so sánh



World Wide - Gradient