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

MÔ HÌNH HÓA TOÁN HỌC (CO2011)

Bài tập lớn

Mô hình SIR trong dự báo COVID-19

GVHD: Nguyễn An Khương
Nguyễn Tiến Thịnh
SV thực hiện: Đỗ Minh Tâm – 1813913
Lê Thanh Tân – 1813935

Tp. Hồ Chí Minh, Tháng 7/2020

Mục lục

1 Giới thiệu đề tài

1.1 Bối cảnh

[0.2]Dịch bệnh COVID-19 lần đầu tiên được ghi nhận tại Thành phố Vũ Hán (Trung Quốc) khoảng cuối năm 2019. Tính đến nay đã hơn 6 tháng trôi qua, dịch bệnh đã liên tục được ghi nhận trên khắp thế giới với số ca lây nhiễm đáng báo động trong các cộng đồng dân cư. Chỉ tính riêng tại Mỹ, tính đến ngày 18 tháng 06 năm 2020, số ca mắc COVID-19 là hơn hai triệu ca với hơn một trăm nghìn ca tử vong đã được xác nhận. Chiếm lần lượt 25.9vong đã được thông báo trên toàn cầu.

Với tình hình phức tạp đó, các quốc gia trên thế giới đã đồng loại thực hiện nhiều biện pháp mạnh mẽ nhằm kiểm soát được tình hình lây lan nhanh chóng của dịch bệnh. Hiện nay, biện pháp được cho là hiệu quả nhất là các biện pháp cách ly với thời gian cách ly 14 ngày. Tuy nhiên, để nâng cao hiệu quả phòng và chống dịch, nhiều mô hình dự báo đã được đưa ra để tiên đoán và có biện pháp kịp thời trước các đợt bùng phát dịch bệnh nghiêm trọng có thể xảy ra.

1.2 Mô hình SIR

1.2.1 Phát biểu mô hình

Mô hình SIR (Susceptible - Infectious - Recovered) là một trong các mô hình cách ly cơ bản được sử dụng nhiều nhất trong quá khứ và hiện tại để mô tả dịch bệnh. Mô hình được lần đầu phát biểu vào thế kỷ 20 (xem [KM27]). Mô hình thể hiện ba trạng thái (Có nguy cơ mắc bệnh - Mắc bệnh - Hồi phục) cho một nhóm người được cách ly với giả thiết rằng sẽ miễn dịch với bệnh nếu đã phục hồi. Mô hình SIR là một hệ động lực gồm ba phương trình vi phân sau

$$\frac{dS}{dt} = -\frac{\beta}{N} IS, \tag{1}$$

$$\frac{dI}{dt} = \frac{\beta}{N} IS - \gamma I, \tag{2}$$

$$\frac{dR}{dt} = \gamma I, \tag{3}$$

trong đó tại mỗi thời điểm $t \geq t_0 \geq 0$ với t_0 là thời điểm đầu ghi nhận.

- $S(t)$: Số người có nguy cơ mắc bệnh;
- $I(t)$: Số người nhiễm bệnh;
- $R(t)$: Số người phục hồi sau bệnh;
- $\beta(t)$: Tỷ lệ tiếp xúc của mỗi người trong nhóm $S(t)$ với người trong nhóm $I(t)$
- $\theta(t)$: Tỷ lệ hồi phục khi mắc bệnh;
- $N(t)$: Tổng số người trong cộng đồng bị cách ly được tính bằng

$$N(t) := S(t) + I(t) + R(t) \tag{4}$$

Hệ phương trình vi phân trên có thể được hiểu như sau

- Phương trình (1) thể hiện sự suy giảm số người có nguy cơ mắc bệnh tại thời điểm $t \geq t0$ Sự suy giảm được tính theo xác suất lây bệnh khi có tiếp xúc giữa nhóm $S(t)$ và nhóm $I(t)$;
- Phương trình (2) thể hiện độ biến thiên số người mắc bệnh tại thời điểm $t \geq t0$. Sự biến thiên này được tính bằng cách lấy số người ở nhóm $S(t)$ đã bị lây nhiễm sau khi tiếp xúc với người bệnh nhóm $I(t)$ và trừ đi số người ở nhóm $I(t)$ đã phục hồi với tỷ lệ $\gamma I(t)$;
- Phương trình (3) thể hiện số người đã hồi phục từ nhóm $I(t)$ theo tỷ lệ hồi phục là γ.

1.2.2 Phương pháp xấp xỉ Euler trong giải hệ SIR

Phương pháp Euler là một phương pháp bậc một thường được sử dụng trong việc giải các phương trình vi phân thường. Phương pháp được đặt tên theo Leonhard Euler, người đã giới thiệu phương pháp trong quyển sách Institutionum Calculi Integralis cùng tên xuất bản trong khoảng thời gian 1768 đến 1770.
Giả sử ta có phương trình vi phân bậc nhất

$$y' = f(t, y(t)). \tag{5}$$

Khi đó, ý tưởng của phương pháp Euler là xấp xỉ nghiệm y bằng dãy y_n sao cho

$$y_{n+1} := y_n + f(t_n, y_n)\Delta t, \tag{6}$$

với Δt là bước xấp xỉ đủ nhỏ và $f(t, y(t))$ là độ dốc của đường cong y tính tại thời điểm t.

Ở dạng tổng quát, một hệ phương trình vi phân bậc một được viết dưới dạng

$$y' = f_1(t, y_1, ..., y_N), \tag{7}$$

$$\vdots$$

$$y'_N = f_N(t, y_1, ..., y_N), \tag{8}$$

trong đó y_i là các hàm số thực phụ thuộc vào biến $t \geq 0$ và f_i là các hàm số thực phi tuyến phụ thuộc vào biến $t \geq 0$ và các y'_i với mọi $i \in 1, ..., N$. Phương pháp Euler khi đó được áp dụng cho từng y_i.

1.2.3 Ước lượng hệ số β và α

Trong bối cảnh hiện nay, phương án cách ly từng nhóm người đã từng tiếp xúc trực tiếp hoặc gián tiếp được các quốc gia trên thế giới xem như một cách hữu hiệu nhất để giảm thiểu số ca mắc bệnh. Như vậy các mô hình cách ly dạng SIR có thể được sử dụng trong trường hợp này. Tuy nhiên ta sẽ xét các hệ số β và γ có thể biến đổi theo thời gian do có sự điều chỉnh trong các lệnh cách ly theo thời gian. Từ đó, cơ hội tiếp xúc gần với người nhiễm virus tăng hoặc giảm dẫn đến xác suất lây nhiễm β thay đổi. Khi tình trạng các bệnh viện trở nên quá tải, sự tập trung của các y bác sĩ cho từng bệnh nhân có thay đổi và dẫn đến tỷ lệ phục hồi cũng khác nhau theo thời gian.
Việc ước lượng các hệ số β và γ phụ thuộc vào dữ liệu về COVID-19 đã được công bố. Cụ thể là số ca mắc bệnh và phục hồi tích lũy theo thời gian. Ở đây, ta sẽ sử dụng phương pháp suy luận Bayes. Gọi

- X : biến ngẫu nhiên quan sát số ca mắc bệnh và số ca hồi phục tại từng thời điểm $t \geq t_0$;

- $\pi(\beta, \gamma|X)$: phân bố xác suất hậu nghiệm của β và γ khi có dữ liệu quan sát;
- $\pi(X|\beta\gamma)$: phân bố xác suất của số ca mắc bệnh và số ca phục hồi khi β và γ cho trước;
- $\pi(\beta, \gamma)$: phân bố xác suất tiên nghiệm khi chưa có dữ liệu ghi nhận về số ca mắc bệnh và số ca phục hồi.

Định lý Bayes được phát biểu như sau

$$\pi(\beta, \gamma|X) \propto \pi(X|\beta, \gamma)\pi(\beta, \gamma). \tag{9}$$

Nghĩa là phân bố xác suất hậu nghiệm của β và γ có thể được tính bằng cách lấy phân bố xác suất của số ca mắc bệnh và số ca phục hồi khi β và γ cho trước nhân với phân bố xác suất tiên nghiệm của β và γ.

1.3 Kết hợp mô hình học máy

1.3.1 Giới thiệu

Bên cạnh các phương pháp ước lượng bằng suy luận Bayes, các mô hình Học Máy cũng được đưa vào dự đoán tình hình dịch bệnh COVID-19 trong các tháng vừa qua. Điểm nổi bật nhất của các mô hình Học Máy là khả năng tính toán mạnh mẽ của máy tính để 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 Nơron thần kinh nối liên tiếp với nhau để rút trích đặc trưng của dữ liệu qua từng lớp của mạng lưới. Học Máy vốn có bắt nguồn từ Thống kê và lần đầu tiên được Marvin Minsky và Dean Edmonds xây dựng nên vào năm 1951 với sự trợ giúp của các máy tính trong quá trình huấn luyện.

1.3.2 Khó khăn trong xây dựng mô hình

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. Việc thiết kế xây dựng các hàm chi phí phù hợp với dữ liệu đầu vào và việc tìm ra các điểm tối ưu của nó là một trong những vấn đề gặp phải. Tùy vào từng loại dữ liệu mà mô hình được xây dựng rất khác nhau.

1.3.3 Mô hình dự báo COVID-19 có Học Máy

Một mô hình Học Máy có kết hợp các mô hình cách ly cổ điển sẽ được thiết kế như sau. Thứ nhất, khởi tạo SIR hay SIRD với các hệ số đầu vào ban đầu β (hệ số lây nhiễm), γ (hệ số phục hồi) và μ (hệ số tử vong nếu là SIRD) để tính ra số ca mắc bệnh dự đoán $I(t)$, số ca phục hồi dự đoán $R(t)$ và số ca tử vong dự đoán $D(t)$ tại thời điểm $t \geq t_0$. Sau đó sử dụng một mạng lưới Nơron và dùng các số liệu đã xác nhận từ các quốc gia về số ca mắc bệnh, số ca tử vong và số ca phục hồi thật sự tại các quốc gia đó để ước tính lại các chỉ số hồi phục γ, chỉ số lây nhiễm β và chỉ số tử vong μ tại mỗi thời điểm công bố của dịch bệnh COVID-19. Mô hình và hàm chi phí cụ thể xem trong [MD20]. Phương pháp cực tiểu hóa hàm chi phí có thể tham khảo [SL04] và các sách về Học Máy và Học sâu.

1.3.4 Dữ liệu COVID-19

Trong khuôn khổ của bài tập lớn này, ta sẽ sử dụng dữ liệu đã được công bố tập hợp tại https://github.com/CSSEGISandData/COVID-19. Dữ liệu dạng chuỗi thời gian.

2 Đề bài

2.1 Bài toán 1

Để bài: Trình bày lại chi tiết cách xây dựng mô hình SIR (cả trường hợp rời rạc lẫn liên tục) hoặc mở rộng của nó và những vấn đề liên quan.
Mô hình SIR biểu diễn cho dịch bệnh COVID-19 bắt đầu từ thành phố Vũ Hán(Wuhan). Giả sử:

- Cộng đồng này được cách ly, không ai có thể ra vào.
- Mỗi người được chia thành có thể bị nhiễm(S), đã bị nhiễm(I) và số người không còn nhiễm bệnh(R tính cả số người đã mất).
- Lúc bắt đầu mọi người đều được cho là S hoặc I, R = 0.
- Một người đã mắc bệnh và hồi phục thì cơ thể sẽ có kháng thể nên sẽ không bị mắc bệnh lần nữa.
- Thời gian trung bình một người hồi phục sau khi mắc bệnh là 2 tuần.
- Đơn vị thời gian chúng ta tính là tuần.

2.1.1 Mô hình SIR rời rạc

Chúng ta sẽ bắt đầu mô hình hóa bằng R(t). Khoảng thời gian mà môt người mắc bệnh là 2 tuần. Nên sau mỗi tuần số người không còn nhiễm nữa là 50

$$R(t+1) = R(t) + 0.5I(t) \tag{10}$$

Tỉ lệ 0.5 trên được gọi là tỉ lệ người không còn nhiễm bệnh sau mỗi tuần β. I(t) sẽ có cả biểu thức tăng và giảm sau một khoảng thời gian. Nó được giảm dựa trên số người không nhiễm bệnh sau mỗi tuần: 0.5I(t). Và được tăng lên bởi số người có thể nhiễm bệnh(S) tiếp xúc với số người đã nhiễm bệnh(I).: αS(t)I(t). Ta định nghĩa α là tốc độ người nhiễm bệnh. Ta coi như α là 1 hằng số có thể tìm được từ những ca nhiễm đầu tiên.
Ví dụ: Chúng ta có 1 khu dân cư gồm 2000 người, ban đầu có 10 người nhiễm bệnh: I(0) = 10 và S(0) = 1990. Sau 1 tuần tổng số người bị nhiễm bệnh là 20. Ta tính *alpha* như sau:

$$\begin{aligned} I(0) &= 10 \\ I(1) &= I(0) - 0.5 * I(0) + \alpha * I(0) * S(0) \\ 20 &= 10 - 5 + \alpha * 10 * 1990 \\ 15 &= 19900 * \alpha \\ \alpha &= 0.00075376884 \end{aligned} \tag{11}$$

Ta xét S(t). Con số này sẽ giảm dần theo thời gian khi những người có thể nhiễm bệnh bị nhiễm

$$S(t+1) = S(t) - \alpha S(n)I(n) \tag{12}$$

Từ đó ta rút ra được mô hình:

$$\begin{aligned} R(t+1) &= R(t) + 0.5I(t) \\ I(t+1) &= I(t) - 0.5I(t) + 0.00075376884I(t)S(t) \\ S(t+1) &= S(t) - 0.00075376884S(t)I(T) \\ I(0) &= 10, S(0) = 1990, R(0) = 0 \end{aligned} \tag{13}$$

2.1.2 Mô hình SIR liên tục

Chúng ta sẽ bắt đầu mô hình hóa bằng R(t). Khoảng thời gian mà một người mắc bệnh là 2 tuần. Nên sau mỗi tuần số người không còn nhiễm nữa là 50

$$\frac{dR}{dt} = 0.5 * I(t) \tag{14}$$

Tỉ lệ 0.5 trên được gọi là tỉ lệ người không còn nhiễm bệnh sau mỗi tuần β. I(t) sẽ có cả biểu thức tăng và giảm sau một khoảng thời gian. Nó được giảm dựa trên số người không nhiễm bệnh sau mỗi tuần: 0.5I(t). Và được tăng lên bởi số người có thể nhiễm bệnh(S) tiếp xúc với số người đã nhiễm bệnh(I).: αS(t)I(t). Ta định nghĩa α là tốc độ người nhiễm bệnh. Ta coi như α là 1 hằng số có thể tìm được từ những ca nhiễm đầu tiên. Ta sẽ ước lượng α là: 0.00075376884
Ta xét S(t). Con số này sẽ giảm dần theo thời gian khi những người có thể nhiễm bệnh bị nhiễm. Ta sẽ dùng tỉ lệ α để dựng mô hình

$$\frac{dS}{dt} = -0.00075376884 * S(t) * I(t) \tag{15}$$

Từ đó ta suy ra được mô hình SIR sau:

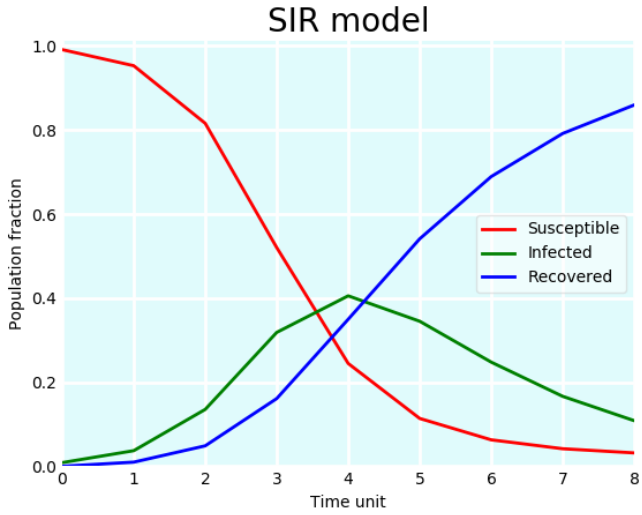
$$\begin{aligned}
&\frac{dR}{dt} = 0.5I(t) \\
&\frac{dI}{dt} = -0.5I(t) + 0.00075376884S(t)I(t) \\
&\frac{dS}{dt} = 0.00075376884S(t)I(t) \\
&I(0) = 10, S(0) = 1990, R(0) = 0
\end{aligned} \tag{16}$$

2.2 Bài toán 2

Để tìm nghiệm cho hệ SIR bằng thuật toán Euler, ta sử dụng hàm odeint từ thư viện scipy của ngôn ngữ Python.

def SIR(t: np.linspace, beta, gamma, S0, I0, R0, N):
    y0 = S0, I0, R0
    ret = odeint(func, y0, t, args=(N, beta, gamma))
    S, I, R = ret.T
    return S,I,R

Hàm odeint sử dụng các phương trình vi phân của mô hình SIR cơ bản để tìm các giá trị S, I, R trên một khoảng thời gian với bước nhảy nhất định. Sau đây là đồ thị biểu diễn mô hình SIR với các biến đầu vào là $\beta = 2, \gamma = 0.5, S0 = 800, I0 = 0, R0 = 7$

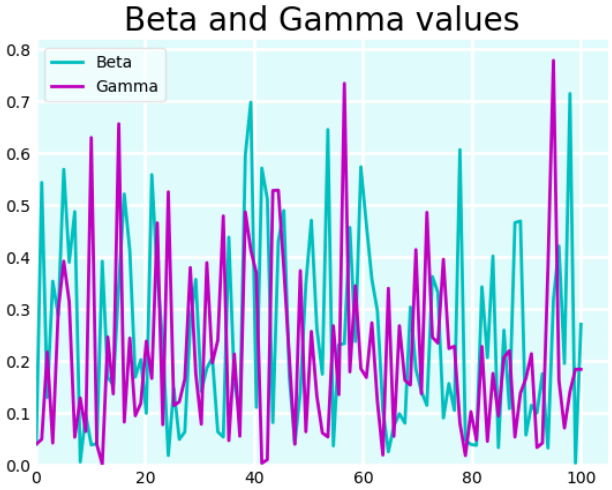


Các xu hướng theo thời gian trên đồ thị:

- Tỉ lệ người có khả năng mắc bệnh giảm dần
- Tỉ lệ người có khả năng nhiễm bệnh tăng đến cực đại rồi giảm
- Tỉ lệ người phục hồi tăng dần

2.3 Bài toán 3

Để tìm ra các giá trị β và γ, ta bắt đầu ước lượng β_0 và γ_0 bằng giá trị dương của phân phối chuẩn với trung bình là 0 và độ lệch chuẩn 0.3. Giả sử chúng ta tính 100 cặp giá trị β và γ. Ta sử dụng phương pháp Metropolis-Hastings để tính các giá trị này và lưu vào hai mảng lb, lg tương ứng. Sau đây là đồ thị biểu diễn một ví dụ về các giá trị β, γ:



Các giá trị β và γ dao động trong khoảng từ 0 đến 0.8 và hoàn toàn ngẫu nhiên giúp cho kết quả của bài toán đa dạng hơn.

2.4 Bài toán 4

Chúng ta sẽ lấy dữ liệu của khu vực Mỹ để ước lượng giá trị R_0. Thư viện pandas của python sẽ giúp ta lấy dữ liệu từ `https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/archived_data/archived_time_series/time_series_19-covid-Confirmed_archived_0325.csv` và `https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/archived_data/archived_time_series/time_series_19-covid-Confirmed_archived_0325.csv` và lưu vào hai DataFrame có tên d và d1. Mảng c sẽ được tạo để lưu các ngày từ 22/01/20 đến 23/03/20. Số liệu Infected và Recovered của từng ngày sẽ được lưu vào hai mảng lI và LR tương ứng. Sử dụng công thức (20) trong đề bài cùng với số lượng mẫu β và γ là 100, chúng ta tính các giá trị R_0 từ ngày 22/01/20 đến 23/03/20 của nước Mỹ. Do trong khoảng thời gian dịch chưa bùng phát mạnh nên các giá trị $\pi(X|\beta,\gamma)$ được tính ra dạng vô định (0^0).

3 Kết luận

Trong báo cáo này, nhóm đã trình bày cách xây dựng mô hình SIR của dịch bệnh COVID-19 và mô hình được đưa ra để tiên đoán và có biện pháp kịp thời trước các đợt bùng phát dịch bệnh nghiêm trọng có thể xảy ra.

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.