

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)

Đề bài tập lớn

“Mô hình SIR trong dự báo COVID-19”

GVHD: Nguyễn Tiến Thịnh
Nguyễn An Khương

SV thực hiện: Nguyễn Văn A – 22102134
Trần Văn B – 88471475
Lê Thị C – 36811334
Phạm Ngọc D – 97501334
Kiều Thị E – 12341334

Tp. Hồ Chí Minh, Tháng 07/2020



Mục lục

1	Giới thiệu đề tài	2
1.1	Bối cảnh	2
1.2	Mô hình SIR	2
1.2.1	Phát biểu mô hình	2
1.2.2	Phương pháp xấp xỉ Euler trong giải hệ SIR	3
1.2.3	Ước lượng hệ số β và γ	4
1.3	Kết hợp Mô hình Học Máy	7
1.3.1	Giới thiệu	7
1.3.2	Khó khăn trong xây dựng mô hình	7
1.3.3	Mô hình dự báo COVID-19 có Học Máy	8
1.4	Dữ liệu COVID-19	8
2	Hướng dẫn và yêu cầu	8
2.1	Hướng dẫn	8
2.2	Yêu cầu	8
2.3	Nộp bài	9
3	Đề bài	9
4	Cách đánh giá và xử lý gian lận	9
4.1	Đánh giá	9
4.2	Xử lý gian lận	9
	Tài liệu	10

1 Giới thiệu đề tài

1.1 Bối cảnh

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 25.9% và 26.24% số ca mắc và ca tử vong đã được thông báo trên toàn cầu.

Với tình hình phức tạp đó, các quốc gia trên thế giới đã đồng loạt 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

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

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

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

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 $I(t)$;
- $\gamma(t)$: Tỷ lệ hồi phục khi mắc bệnh;
- $N(t)$: Tổng số người trong cộng đồng bị cách ly được tính bằng

$$N(t) := S(t) + I(t) + R(t). \quad (4)$$

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 $t \geq t_0$. 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 $t \geq t_0$. 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 .

Ví dụ 1.1. Giả sử có một loại cúm đang lây lan trong một cộng đồng dân cư. Giả sử rằng:

- Cộng đồng này đang bị cách ly, không ai được ra và cũng không ai được vào;
- Loại cúm này có thời gian từ khi phát bệnh cho đến khi hồi phục là 2 tuần không đổi theo thời gian;
- Một người khi mắc bệnh và hồi phục thì không còn mắc bệnh này lại lần thứ hai;
- Sau một thời gian điều tra, tỷ lệ mắc bệnh khi có tiếp xúc với người bệnh ở mức 0.2% sau một tuần tiếp xúc và giả sử tỷ lệ này cũng không đổi theo thời gian.

Khi đó, với mỗi tuần, ta có thể sử dụng mô hình

$$\frac{dS}{dt} = -0.002IS, \quad (5)$$

$$\frac{dI}{dt} = 0.002IS - 0.5I, \quad (6)$$

$$\frac{dR}{dt} = 0.5I. \quad (7)$$

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

$$y' = f(t, y(t)). \quad (8)$$

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

$$y_{n+1} := y_n + f(t_n, y_n)\Delta t, \quad (9)$$

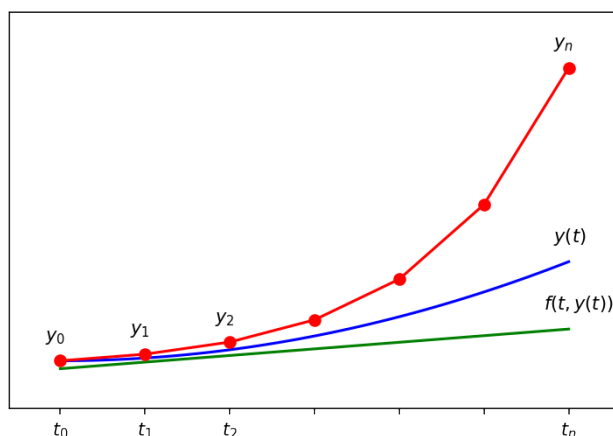
với Δt là bước xấp xỉ **đủ nhỏ** và $f(t, y(t))$ là độ dốc của đường cong y tính tại thời điểm t .

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

$$y'_1 = f_1(t, y_1, \dots, y_N), \quad (10)$$

$$\vdots$$

$$y'_N = f_N(t, y_1, \dots, y_N), \quad (11)$$



Hình 1: Các điểm tròn màu đỏ chính là giá trị xấp xỉ đường cong y có độ dốc $f(t, y)$ nghĩa là $y' = f(t, y)$. Sai khác giữa y_n và y_{n-1} đúng bằng $f(t_{n-1}, y_{n-1})(t_n - t_{n-1})$

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 's với mọi $i \in \{1, \dots, N\}$. Phương pháp Euler khi đó được áp dụng cho từng y_i .

Giả sử tại thời điểm ban đầu, số người có khả năng bị nhiễm bệnh là 800 người, số người mắc bệnh là 5 người và số ca phục hồi khi ấy chưa có. Sử dụng phương pháp Euler, ta sẽ giải hệ SIR (5), (6) và (7) để tìm ra số người có khả năng bị lây nhiễm, số người mắc bệnh và số ca phục hồi sau hai tháng (8 tuần) tính từ thời điểm đầu ghi nhận số liệu.

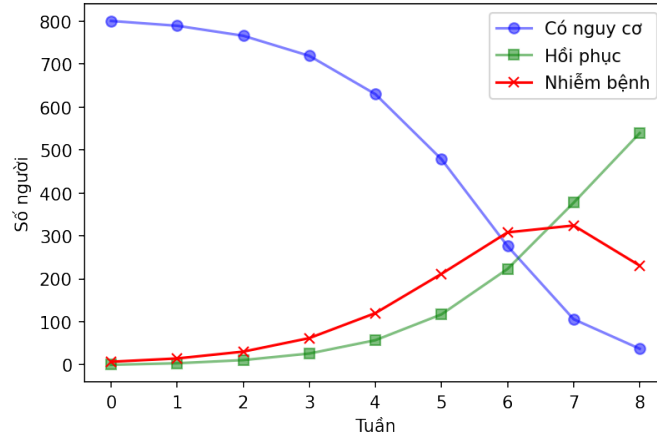
Kết quả thu được cho như dưới bảng sau.

Tuần	Có nguy cơ	Ca nhiễm	Ca hồi phục
0	800.000000	7.000000	0.000000
1	788.800000	14.700000	3.500000
2	765.609280	30.540720	10.850000
3	718.844763	62.034877	26.120360
4	629.657869	120.204332	57.137799
5	478.282662	211.477373	117.239965
6	275.990740	308.030609	222.978651
7	105.963549	324.042496	376.993955
8	37.290163	230.694633	539.015203

Xem [ESG93; EG96] cho thuật toán Euler và mở rộng như Runge–Kutta.

1.2.3 Ước lượng hệ số β và γ

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.



Hình 2: Số ca mắc bệnh tăng trong khoảng 6 tuần đầu tiên và giảm dần ở 2 tuần tiếp theo

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

$$\pi(\beta, \gamma|X) \propto \pi(X|\beta, \gamma)\pi(\beta, \gamma). \quad (12)$$

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 .

Ví dụ 1.2. Giả sử biến quan sát là độc lập trong các lần đo đạc và có phân bố xác suất $X \sim \Gamma(\beta, \gamma)$. Ta cũng giả sử phân bố xác suất tiên nghiệm của các hệ số β và γ là Gamma như sau

$$\beta \sim \Gamma(\lambda_\beta, \nu_\beta) \quad (13)$$

$$\gamma \sim \Gamma(\lambda_\gamma, \nu_\gamma) \quad (14)$$

Khi đó, $\pi(\beta, \gamma|X)$ có thể được tính bằng tích của ba phân bố xác suất sau

$$\pi(X|\beta, \gamma) = \prod_{i=1}^n f(X(t_i)|\beta, \gamma) = \prod_{i=1}^n \frac{\gamma^\beta}{\Gamma(\beta)} X(t_i)^{\beta-1} \exp\{-\gamma X(t_i)\}, \quad (15)$$

$$\pi(\beta) = \frac{\nu_\beta^{\lambda_\beta}}{\Gamma(\lambda_\beta)} \beta^{\lambda_\beta-1} \exp\{-\nu_\beta \beta\} \quad (16)$$

và

$$\pi(\gamma) = \frac{\nu \gamma^{\lambda_\gamma}}{\Gamma(\lambda_\gamma)} \gamma^{\lambda_\gamma - 1} \exp\{-\nu \gamma\}, \quad (17)$$

trong đó n là số lần ghi nhận giá trị của biến quan sát X và Γ là hàm Gamma

$$\Gamma(y) = \int_0^\infty z^{y-1} \exp(-z) dz. \quad (18)$$

Ước lượng này có vai trò rất quan trọng và được thể hiện trên hệ số

$$R_0 := \frac{\beta}{\gamma}. \quad (19)$$

Khi hệ số $R_0 < 1$ thì không có đợt bùng phát dịch bệnh xảy ra vì tỷ lệ tiếp xúc người mắc bệnh β nhỏ hơn tốc độ phục hồi. Khi hệ số $R_0 > 1$ thì các đợt bùng phát dịch bệnh sẽ xảy ra trong tương lai vì tỷ lệ tiếp xúc với người mắc bệnh cao hơn tốc độ phục hồi sau bệnh. Đặc biệt giá trị trung bình của hệ số này đúng bằng

$$E(R_0) = \int \pi(\beta, \gamma | X) R_0(\beta, \gamma) d(\beta, \gamma) \quad (20)$$

Trong đó X là dữ liệu về số ca mắc bệnh và phục hồi quan sát được. Giá trị trung bình này có thể ước lượng được vì phân bố xác suất $\pi(\beta, \gamma | X)$ có thể tính được nhờ vào công thức Bayes (12).

Tuy nhiên, tích phân (20) không thể được tính toán một cách trực tiếp. Thay vào đó chúng ta sẽ sử dụng công thức xấp xỉ

$$E(R_0) = \int \pi(\beta, \gamma | X) R_0(\beta, \gamma) d(\beta, \gamma) \propto \int \pi(X | \beta, \gamma) \pi(\beta, \gamma) R_0(\beta, \gamma) d(\beta, \gamma) \approx \sum_{i=1}^m \pi(X | \beta_i, \gamma_i) \frac{\beta_i}{\gamma_i}, \quad (21)$$

trong đó (β_i, γ_i) được lấy ra dựa trên phân bố xác suất tiên nghiệm $\pi(\beta, \gamma)$ và m là kích cỡ mẫu.

Nghĩa là, chúng ta có thể lấy một mẫu kích cỡ m đủ lớn gồm các giá trị của biến ngẫu nhiên (β, γ) dựa trên phân phối xác suất tiên nghiệm $\pi(\beta, \gamma)$ và tính tổng các giá trị $\pi(X | \beta, \gamma) \frac{\beta}{\gamma}$ trên mẫu này. Thuật toán Metropolis–Hastings khi đó có thể được sử dụng để tạo một mẫu (β, γ) dựa trên phân bố xác suất $\pi(\beta, \gamma)$.

Thuật toán Metropolis–Hastings được biết đến là một phương pháp xích Markov Monte Carlo (xem [Has70]) có các bước như sau

1. Khởi tạo β_0 và γ_0 từ phân bố xác suất tiên nghiệm $\pi(\beta, \gamma)$.
2. Gán $\beta := \beta_0$ và $\gamma := \gamma_0$.
3. Khởi tạo β^* và γ^* ngẫu nhiên từ phân phối xác suất bất kỳ $p(\beta, \gamma)$.
4. Nếu $p(\beta, \gamma)$ là đối xứng, nghĩa là $p(\beta^*, \gamma^* | \beta, \gamma) = p(\beta, \gamma | \beta^*, \gamma^*)$, gán cho r là xác suất giữ lại β^* và γ^* bằng công thức

$$r := \min \left(1, \frac{\pi(\beta^*, \gamma^*)}{\pi(\beta, \gamma)} \right) \quad (22)$$

và đi đến Bước 6.

5. Nếu không đối xứng, gán

$$r := \min \left(1, \frac{\pi(\beta^*, \gamma^*) p(\beta, \gamma | \beta^*, \gamma^*)}{\pi(\beta, \gamma) p(\beta^*, \gamma^* | \beta, \gamma)} \right) \quad (23)$$

và đi đến Bước 6.

6. Khởi tạo giá trị q ngẫu nhiên từ phân phối đều liên tục $U(0, 1)$.

7. Nếu $q < r$, tạo $\beta_{i+1} := \beta^*$ và $\gamma_{i+1} := \gamma^*$ với i là chỉ số phần tử trong mẫu và đi đến Bước 9.

8. Ngược lại, tạo $\beta_{i+1} := \beta_i$ và $\gamma_{i+1} := \gamma_i$ và đi đến bước 9.

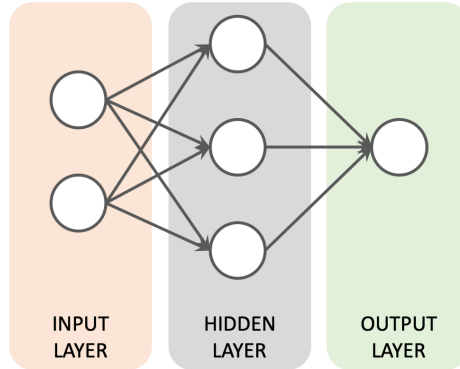
9. Lặp lại từ Bước 2 với $\beta := \beta_i$ và $\gamma := \gamma_i$ cho đến khi đủ kích cỡ mẫu.

Chi tiết về phương pháp Monte Carlo và thuật toán Metropolis–Hastings xem trong [Bro+11].

1.3 Kết hợp Mô hình Học Máy

1.3.1 Giới thiệu

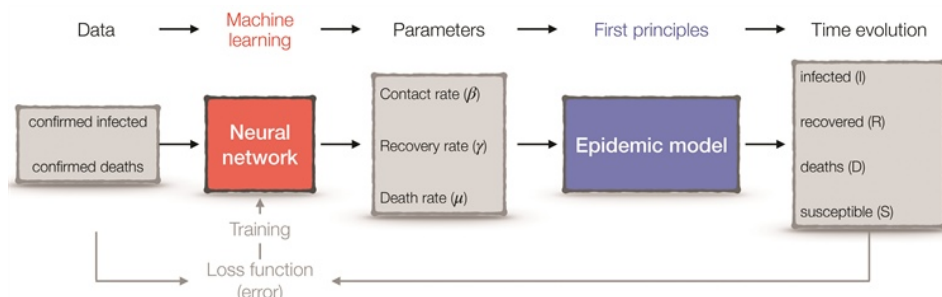
Bên cạnh các phương pháp ước lượng bằng suy luận Bayes, các mô hình Học Máy cũng được đưa vào dự đoán tình hình dịch bệnh COVID-19 trong các tháng vừa qua. Điểm nổi bật nhất của các mô hình Học Máy là khả năng tính toán mạnh mẽ của máy tính để tìm ra các đặc tính và xu hướng phát triển chứa đựng trong 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 Nơon 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.



Hình 3: Sơ đồ mạng lưới của một mô hình Học Máy

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 đặc 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.



Hình 4: Mô hình Học Máy kết hợp SIRD được công bố bởi Luca Luca Magri và Nguyen Anh Khoa Doan [MD20]

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.

1.4 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 có thể lấy từ đường dẫn “COVID-19/csse_covid_19_data/csse_covid_19_time_series/”.

2 Hướng dẫn và yêu cầu

2.1 Hướng dẫn

Đọc kỹ thông tin về mô hình trong các TLTK [GFH13, Example 5 on page 50 and Example 3 on page 564] và [JKG20] (xem bản dịch tiếng Việt tại <https://tinyurl.com/y8bd13hn>).

SV có thể chia nhóm giữa các lớp L01-04 với nhau chứ không nhất thiết cùng một lớp. Mỗi nhóm từ 3-5 người.

2.2 Yêu cầu

- Hạn nộp bài: 24/7/2020. Đối với mỗi câu hỏi, yêu cầu sinh viên trình bày bố cục rõ ràng, mạch lạc.
- Viết báo cáo theo đúng **bố cục như trong file mẫu** bằng LaTeX.
- Mỗi nhóm khi nộp bài **cần phải nộp theo file log (nhật ký)** ghi rõ: tiến độ công việc, phân công nhiệm vụ, trao đổi của các thành viên,...

2.3 Nộp bài

- SV chỉ nộp bài qua hệ thống BK-eLearning: nén tất cả các file cần thiết (file .tex, file .py, ...) thành một file tên là “*BTL-CO2011-MT192-Cac-MSSV.zip*” và nộp trong mục Assignment trên trang BK-eLearning.
- Lưu ý: mỗi nhóm **chỉ cần một thành viên là nhóm trưởng nộp bài**.

3 Đề bài

Sinh viên thực hiện các yêu cầu sau.

Bài toán 1 (*Bắt buộc*). 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.

Bài toán 2 (*Bắt buộc*). Viết chương trình sử dụng thuật toán Euler tìm nghiệm của hệ SIR hoặc mở rộng của nó với tham số đầu vào gồm biến thời gian t , các hệ số tiếp xúc β , hệ số phục hồi γ và điều kiện đầu là số ca mắc bệnh $I(t_0)$ và số ca phục hồi $R(t_0)$ của mô hình tính tại thời điểm đầu tiên ghi nhận được các ca nhiễm bệnh. Giá trị trả về là mảng chứa số người nhiễm bệnh $I(t)$ và số người đã hồi phục $R(t)$ tính tại thời điểm $t \geq t_0$. Cho một số ví dụ về điều kiện đầu và các hệ số trong mô hình và dùng chương trình đã viết để tìm nghiệm xấp xỉ. Biểu diễn nghiệm xấp xỉ bằng cách vẽ đồ thị. Trường hợp là hệ SIR mở rộng như SIRD thì cần trả về $I(t)$, $R(t)$ và $D(t)$ là số ca tử vong tại thời điểm $t \geq t_0$. Trình bày chi tiết kết quả trong báo cáo.

Bài toán 3 (*Bắt buộc*). Viết chương trình theo ngôn ngữ tự chọn để lấy mẫu sử dụng thuật toán Metropolis–Hastings với tham số đầu vào là phân bố xác suất tiên nghiệm $\pi(\beta, \gamma)$ cho trước. Giá trị trả về là một mẫu gồm các cặp β và γ có phân bố xác suất $\pi(\beta, \gamma)$. Vẽ biểu đồ thể hiện quá trình chọn mẫu. Trình bày chi tiết kết quả trong báo cáo.

Bài toán 4 (*Bắt buộc*). Mỗi nhóm hãy tự chọn lấy một khu vực gồm một số quốc gia, dùng các chương trình ở Bài tập 3 để ước lượng giá trị trung bình hệ số R_0 trong (20) ở khu vực này. Phân tích rõ chính sách hạn chế đi lại và cách ly đã ảnh hưởng đến hệ số R_0 ở khu vực này như thế nào. Nêu rõ dẫn chứng. Xem tham khảo trong [JKG20]. Chương trình mẫu viết bằng ngôn ngữ R có thể tham khảo trong [LM05]. Trình bày chi tiết kết quả trong báo cáo.

Bài toán 5 (*Nâng cao và là bắt buộc đối với hệ kỹ sư tài năng, tùy chọn đối với hệ CQ*). Sử dụng hàm chi phí và mô hình ước lượng như trong bài báo [MD20], huấn luyện lại mô hình theo dữ liệu từ khu vực các quốc gia đã chọn. Phân tích kết quả và trực quan hóa kết quả bằng cách vẽ đồ thị. Trình bày chi tiết kết quả trong báo cáo.

4 Cách đánh giá và xử lý gian lận

4.1 Đánh giá

Mỗi bài làm sẽ được đánh giá như sau.

4.2 Xử lý gian lận

Bài tập lớn phải được sinh viên (nhóm) TỰ LÀM. Sinh viên (nhóm) sẽ bị coi là gian lận nếu:



Nội dung	Tỉ lệ điểm (%)
Các chương trình được viết gọn gàng và thực thi được	30%
Phân tích mạch lạc, có tính hệ thống, đúng trọng tâm câu hỏi	30%
Biểu đồ và đồ thị đúng, rõ ràng và trực quan	20%
Trình bày kiến thức chuẩn bị rõ ràng, phù hợp	15%
Trình bày văn bản đẹp, đúng chuẩn	5%

- Có sự giống nhau bất thường giữa các bài thu hoạch (nhất là phần kiến thức chuẩn bị). Trong trường hợp này, **TẤT CẢ** các bài nộp có sự giống nhau đều bị coi là gian lận. Do vậy sinh viên (nhóm) phải bảo vệ bài làm của mình.
- Sinh viên (nhóm) không hiểu bài làm do chính mình viết. Sinh viên (nhóm) có thể tham khảo từ bất kỳ nguồn tài liệu nào, tuy nhiên phải đảm bảo rằng mình hiểu rõ ý nghĩa của tất cả những gì mình viết.

Bài bị phát hiện gian lận thì sinh viên sẽ bị xử ý theo quy định của nhà trường.

Tài liệu

- [Bro+11] Steve Brooks et al. *Handbook of Markov Chain Monte Carlo*. CRC press, 2011.
- [EG96] Hairer Ernst and Wanner Gerhard. *Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems*. Springer, 1996.
- [ESG93] Hairer Ernst, P. Nørsett Syvert, and Wanner Gerhard. *Solving Ordinary Differential Equations I: Nonstiff Problems*. Springer, 1993.
- [GFH13] Frank Giordano, William P Fox, and Steven Horton. *A first course in mathematical modeling*. Nelson Education, 2013.
- [Has70] W. K. Hastings. “Monte Carlo Sampling Methods Using Markov Chains and Their Applications”. In: *Biometrika* 57 (1) (1970), pp. 97–109.
- [JKG20] T. Wu Joseph, Leung Kathy, and Leung Gabriel. “Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study”. In: 395 (2020).
- [KM27] William Ogilvy Kermack and A. G. McKendrick. “A contribution to the mathematical theory of epidemics”. In: *Proc. R. Soc. Lond. A* 115 (1927), pp. 700–721.
- [LM05] S. T. Ho Lam and A. Suchard Marc. *Simple MCMC under SIR*. 2005. URL: <https://cran.r-project.org/web/packages/MultiBD/vignettes/SIR-MCMC.pdf>.
- [MD20] Luca Magri and Nguyen Anh Khoa Doan. “First-principles Machine Learning for COVID-19 Modeling”. In: *arXiv preprint arXiv:2004.09478* (2020).
- [SL04] Boyd Stephan and Vandenberghe Lieven. *Convex Optimization*. Cambridge University Press, 2004.