Sử dụng Poisson Hidden Markov Model để mô hình hóa dữ liệu đếm bị phân tán

Đặng Tiến Đạt, Nguyễn Tiến Dũng, Nguyễn Đức Anh[1]

Abstract

Bài báo cáo này tập trung vào việc xác định quá trình dựa vào các dữ liệu quan sát, mà cụ thể hơn là xác định quá trình đếm dựa trên các quan sát về lưu lượng xe trên một quãng đường trong một khoảng thời gian nhất định. Thông thường, đối với bộ dữ liệu đếm, mô hình Poisson là một mô hình rất phổ biến và thường được sử dụng để mô hình hóa các quá trình đếm. Mô hình Poisson tương đối đặc biệt khi có phương sai và kì vọng của mô hình mang cùng một giá trị. Tuy nhiên trong thực tế, rất nhiều bộ dữ liệu xảy ra hiện tượng phương sai mẫu lớn hơn một cách bất thường so với trung bình mẫu, dẫn đến những sai sót trong sử dụng mô hình Poisson để mô hình hóa quá trình này. Nhóm em đề xuất sử dụng phân phối Poisson kết hợp với xích Markov ẩn (HMM) thành Poisson Hidden Markov Model (PHMMs) để mô hình hóa dữ liệu đếm loại này. Mô hình Markov ẩn là một phương pháp thống kê rất phù hợp để giải thích sự phụ thuộc của chuỗi quan sát phụ thuộc thời gian bằng cách giả thiết các dữ liệu quan sát được xây dựng từ một hữu hạn hỗn hợp các phân phối, được điều chỉnh theo nguyên tắc của chuỗi Markov (MC). Các tham số mô hình của PHMMs được ước lượng bởi MLE, thông qua thuật toán Expectation – Maximiztion.

Keywords: overdispersion, poisson distribution, hidden Markov model, process identity

Mục lục

Danh sách hình vẽ

Danh sách bảng

---

1. Introduction

Mô hình Poisson là phương pháp phổ biến nhất để mô hình hóa chuỗi thời gian dữ liệu đếm. Một đặc điểm độc đáo của mô hình Poisson chính là kì vọng và phương sai mô hình bằng nhau, tuy nhiên đây là một điều kiện rất chặt mà không nhiều dữ liệu trong thực tế đáp ứng được. Thực tế, dữ liệu có biến động lớn này một phần do nguyên nhân đến từ mối quan hệ không rõ ràng giữa kì vọng và phương sai của mô hình Poisson[9]. Một số phương pháp đã được đề xuất để khắc phục trường hợp này đã được đưa ra bởi Greenwood & Yule (1920) và Neyman (1931). Wang & Famoye (1997) giới thiệu mô hình Poisson hồi qui tổng quát (Generalized Poisson Regression). Gần đây, Cepeda Cuervo và Cifuentes - Amado (2017) cũng đã phát triển các mô hình hồi quy trung bình và phân tán để phù hợp với dữ liệu bị phân tán dựa trên các mô hình nhị

thức âm, beta và sau đó Sebastian Georgea, Ambily Joseb (2019) trình bày về mô hình Poisson Markov ẩn tổng quát cho dữ liệu đếm bị phân tán[5], mô hình này khác phân phối Poisson khi cần tới 2 tham số thay vì 1 tham số của Poisson.

Trong trường hợp dữ liệu bị phân tán quá mức như vậy, chúng ta có thể giả sử cường độ Poisson không còn không đổi nhưng có một phân phối xác suất nhất định. Trong trường hợp này, chúng tôi có thể mô hình hóa quá trình đếm bằng các mô hình hỗn hợp Poisson, giả thuyết chúng ta có các quan sát là độc lập và mô hình hỗn hợp Markov tức là các mô hình Markov ẩn Poisson (PHMM).

PHMMs có nguồn gốc phát triển và được ứng dụng trong lĩnh vực sinh trắc học.

Trong bài báo cáo này, nhóm em đề xuất sử dụng PHMMs để mô hình hóa dữ liệu đếm phương tiện trên một đoạn đường tại Mỹ. Với giả thiết rằng quá trình ngẫu nhiên với thời gian rời rạc $\{(O_t; H_t)\}_{t\in\mathbb{N}}$ với O_t là chuỗi quan sát được và H_t là chuỗi các trạng thái ẩn với giả thiết $P[O_t \mid H_t] \sim \text{Poisson}(\lambda), \forall t$ và tham số λ phụ thuộc vào trạng thái của H_t.

Bài báo cáo được chia làm các phần với nội dung sau:

1. Cơ sở lí thuyết
2. Phương pháp ước lượng
3. Áp dụng phương pháp trên dữ liệu thực tế.
4. Kết luận.

2. Cơ sở lý thuyết

2.1. Overdispersed data

Trong thống kê, sự kiện dữ liệu bị phân tán quá mức (overdispesion) được thể hiện qua mức độ phân tán của một tập dữ liệu so với dự kiến dựa trên một mô hình thống kê nhất định. Sự phân tán quá mức xảy ra khi phương sai của tập mẫu cao hơn phương sai của mô hình theo lý thuyết. Sự phân tán quá mức xảy ra trong thực tế tương đối thường xuyên và thường được quan sát thấy trong phân tích dữ liệu rời rạc, ví dụ, dữ liệu đếm được phân tích theo mô hình Poisson hoặc dữ liệu dưới dạng tỉ lệ theo mô hình phân phối nhị thức.

Ngoài ra, ta đã biết một mục đích của thống kê ứng dụng là chọn một mô hình tham số để phù hợp với một tập hợp quan sát thực nghiệm nhất định. Điều này đòi hỏi phải đánh giá sự phù hợp của mô hình đã chọn. Thông thường có thể chọn các tham số mô hình theo cách sao cho trung bình lý thuyết của mô hình xấp xỉ bằng trung bình mẫu. Tuy nhiên, đặc biệt đối với các mô hình đơn giản có ít tham số (chẳng hạn như

phân phối Poisson khi chỉ có một tham số duy nhất, điều này không cho phép phương sai được điều chỉnh độc lập với giá trị trung bình) dẫn đến phương sai đo được có thể khác phương sai trong lý thuyết, từ đó dự đoán từ lý thuyết có thể không phù hợp với các quan sát thực nghiệm cho các thời điểm xa hơn và nếu phương sai đo được lớn hơn phương sai trong lý thuyết, ta có thể kết luận dữ liệu đã bị quá mức.

2.1.1. Kiểm định tính phân tán dữ liệu

Trong nhiều tình huống, sự phân tán quá mức của dữ liệu có thể biểu hiện rất rõ ràng thông qua sự vượt trội của thống kê mức độ phù hợp của Pearson hoặc độ lệch, ngay cả khi mô hình đã khớp hoàn toàn với dữ liệu[3]. Đã có rất nhiều tiêu chuẩn kiểm định sự phân tán quá mức được đưa ra và thảo luận trong một thời gian dài.

Để kiểm tra độ phân tán của dữ liệu, người ta đưa ra khái niệm điểm kiểm tra (Score Test) cho sự phân tán quá mức, sẽ so sánh phương sai mẫu với phương sai theo lý thuyết mô hình.

Đối với kiểm tra biến Poisson mở rộng, thống kê điểm kiểm tra để kiểm tra giả thuyết gốc: $H_1 : \tau = 0$ trong mô hình cùng với hàm phương sai quá mức $\mu_i + \tau\mu_i^2$ là:

$$T_{P1} = \frac{\sum_{i=1}^{n} \left\{ (y_i - \hat{\mu_i})^2 - \left(1 - \hat{h_i}\right) \hat{\mu_i} \right\}}{\left(2 \sum_{i=1}^{n} \hat{\mu_i^2}\right)^{1/2}}$$

Thống kê $T_{P1} \overset{n\to\infty}{\asymp} \mathcal{N}(0, 1)$ với giả thuyết H_1.

Đối với hàm phương sai $(1 + \tau)\mu_i$ ta sử dụng thống kê sau với giả thuyết gốc $H_2 : \tau = 0$:

$$T_{P2} = \frac{1}{\sqrt{2n}} \sum_{i=1}^{n} \left\{ \frac{(y_i - \hat{\mu_i})^2 - \left(1 - \hat{h_i}\right) \hat{\mu_i}}{\hat{\mu_i}} \right\}$$

Từ việc đánh giá các giá trị của T_{P1}, T_{P2} ta có thể kết luận là dữ liệu có bị phân tán quá mức hay không.

2.2. Poisson Hidden Markov Models

2.2.1. Mô hình Markov ẩn (Hidden Markov Models)

Definition 2.1. Mô hình Markov ẩn (HMM) là một mô hình phát hiện tín hiệu được giới thiệu lần đầu vào năm 1966 bởi Baum và Petrie. HMM giả sử rằng chuỗi quan sát (observation sequence) được xác định từ một chuỗi các trạng thái ẩn (hidden state sequence) của dữ liệu rời rạc và thỏa mãn các yêu cầu của quá trình Markov. HMM

được phát triển từ mô hình cho một biến quan sát thành mô hình cho nhiều biến quan sát.

Ứng dụng của mô hình HMM sau này được mở rộng tới rất nhiều lĩnh vực như nhận diện giọng nói, toán sinh học, chứng khoán,...

Definition 2.2. Các đại lượng trong mô hình

Các kí hiệu trong báo cáo là dùng đồng nhất, cho toàn bộ báo cáo.

Kí kiệu	Định nghĩa
$Q = \{q_1, q_2, ..., q_N\}$	không gian trạng thái của xích Markov.
$A = (a_{ij})_{N \times N}$	ma trận xác suất chuyển A
$O = (o_1, o_2, ..., o_T)$	dãy các quan sát độ dài T
$B = (b_t(i))$	$b_t(i) = P[o_t \mid q_t = i], t = 1, 2, ..., T, i = 1, 2, ..., N$ là các xác suất phụ thuộc trạng thái
$\pi = (\pi_1, \pi_2, ..., \pi_N)$	phân phối ban đầu của chuỗi xích Markov ẩn
$H^{(T)} = (h_1, h_2, ..., h_T)$	dãy các trạng thái của xích Markov ẩn độ dài T

Bảng 1: Bảng kí hiệu các đại lượng

Remark 2.3. Nhắc lại một số tính chất cơ bản

1. $a_{ij} = P(h_{t+1} = q_j | h_t = q_i), i, j = 1, 2, ..., N$
2. $\sum_{j=1}^{N} a_{ij} = 1 \ \forall i, j = 1, 2, ..., N$
3. $\sum_{1}^{N} \pi_i = 1$

Một mô hình Markov ẩn được xây dựng dựa trên hai giả thiết sau:

- Các trạng thái ẩn thoả mãn tính chất Markov:

$$P(h_t \mid h_1, h_2, ..., h_{t-1}) = P(h_t \mid h_{t-1})$$

- Các quan sát chỉ phụ thuộc vào trạng thái hiện tại của trạng thái ẩn và không phụ thuộc vào các dữ liệu trong quá khứ

$$P(o_t \mid h_1, h_2, ..., h_t, o_1 ... o_t) = P(o_t \mid h_t)$$

Khi đó, một mô hình Markov ẩn được xác định qua bộ tham số $\lambda = \{\pi, A, B\}$

2.2.2. Các bài toán cơ bản của mô hình Markov ẩn

Đối với một mô hình Markov ẩn, có ba vấn đề chính chúng ta cần quan tâm:

- Bài toán 1 (Likelihood): Cho mô hình Markov ẩn với bộ tham số $\lambda = \{\pi, A, B\}$ và chuỗi các quan sát O. Xác định $P(O \mid \lambda)$
- Bài toán 2 (Decoding): Cho chuỗi quan sát O và bộ tham số λ. Xác định chuỗi trạng thái ẩn H sao cho $P[H \mid O, \lambda]$ đạt cực đại.
- Bài toán 3 (Learning): Cho chuỗi quan sát được O và tập các trạng thái của xích ẩn trong mô hình Markov ẩn, tìm bộ tham số $\lambda = (\pi, A, B)$ phù hợp nhất của mô hình Markov ẩn.

Dưới đây, chúng ta sẽ đề xuất phương pháp giải quyết các bài toán trên.

Đối với bài toán thứ nhất, để tính $P[O \mid \lambda]$, ta giả sử rằng chuỗi trạng thái ẩn tương ứng với bộ (O, λ) trên là $H = (h_1, h_2, ..., h_T)$, trong đó T là độ dài của chuỗi quan sát.

Problem. Với mô hình HMM $\lambda = (\pi, A, B)$ và chuỗi quan sát O, tính $P[O \mid \lambda]$

Hiển nhiên giả thiết bên trên là đúng bởi trạng thái ẩn và quan sát là quan hệ 1 – 1.

Với chuỗi quan sát O và chuỗi trạng thái H, ta có:

$$P(O \mid H, \lambda) = P[o_1, o_{2,...,o_T} \mid H, \lambda]$$

Gọi Ω_T là không gian tất cả các chuỗi trạng thái ẩn độ dài T.

Theo công thức xác suất đầy đủ:

$$P(O \mid \lambda) = \sum_{H \in \Omega_T} P[O, H \mid \lambda]$$

Ta có

$$\begin{aligned} P(O \mid \lambda) &= \sum_{H \in \Omega_T} P[O, H \mid \lambda] \\ &= \sum_{H \in \Omega_T} \frac{P[O, H, \lambda]}{P[\lambda]} \\ &= \sum_{H \in \Omega_T} P[O \mid H, \lambda] P[H \mid \lambda] \end{aligned} \tag{1}$$

Theo phân phối hữu hạn chiều của một xích Markov, ta có

$$P\left[H \mid \lambda\right] = \pi_{h_1} a_{h_1,h_2} ... a_{h_{T-1},h_T}$$

Không gian trạng thái Ω có tới N^T phần tử, dẫn đến việc tính toán đối với biểu thức (1) có độ phức tạp tính toán là $O\left(TN^T\right)$

Thực tế, các giá trị N, T thường rất lớn, dẫn đến tính toán trực tiếp (1) tiêu tốn rất nhiều chi phí tính toán. Để giải quyết vấn đề trên, người ta đã giới thiệu thuật toán Forward (tạm dịch là chuyển tiếp) để tính $P\left(O \mid \lambda\right)$.

Độ phức tạp thuật toán là $O(N^2T)$, chấp nhận được nếu so với $O\left(TN^T\right)$.

Đặt $\alpha_t\left(i\right) = P\left(o_1, o_2, ..., o_t, H_t = q_i \mid \lambda\right)$

Khi đó ta có thể tính được $\alpha_t\left(i\right)$ một cách đệ quy như sau:

$$\alpha_{t+1}(j) = \sum_{i=1}^{N} \alpha_t(i) a_{ij} b_t(j)$$

Ta có:

$$\begin{aligned} P\left(O \mid \lambda\right) &= \sum_{i=1}^{N} P\left[o_1, o_2, ..., o_T, h_T = q_i \mid \lambda\right] \\ &= \sum_{i=1}^{N} \alpha_T\left(q_i\right) \end{aligned}$$

Tại thời điểm xuất phát $t = 1$

$$\alpha_1\left(i\right) = \pi_i b_t\left(i\right)$$

Vậy thuật toán chuyển tiếp có thể được mô tả đơn giản qua 3 công thức:

1. Khởi tạo: $\alpha_1(i) = \pi_i b_t(i)$ với $1 \leq i \leq N$
2. Đệ qui: $\alpha_{t+1}(j) = \sum_{i=1}^{N} \alpha_t(i) a_{ij} b_t(j)$ với $1 \leq j \leq N, 1 < t \leq T$
3. Giá trị cần tìm: $P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i)$

Đối với bài toán giải mã, ta cần tìm chuỗi trạng thái phù hợp nhất với quan sát. Đầu tiên, ta cần phải biết thế nào là tốt nhất

Problem. Với mô hình HMM $\lambda = (\pi, A, B)$ và chuỗi quan sát O, xác định chuỗi trạng thái ẩn tốt nhất.

Chuỗi ẩn H^* là chuỗi ẩn tốt nhất nếu:

$$P(H^*|O,\lambda) = \max_{H\in\Omega} P(H \mid \mathrm{O},\lambda)$$

Đối với bài toán này, người ta đưa ra thuật toán Viterbi [6] như sau:

- $v_t(j)$ biểu diễn xác suất HMM đang ở trạng thái j sau khi thấy tđiểm quan sát được đầu tiên và đi qua chuỗi trạng thái có thể xảy ra nhất $H^{(t-1)} = (h_1, h_2, ..., h_{t-1})$ với bộ tham số λ đã biết

$$v_t(j) = \max_{H^{(t-1)}\in\Omega_{t-1}} \left\{ P\left[H^{(t-1)}, h_t = q_j \mid O, \lambda\right]\right\} \text{ với } 1 \le j \le N, 1 < t \le T$$

- Mối liên hệ giữa $v_t(j)$ và $v_{t-1}(i)$: $v_t(j) = \max_{H^{(t)}} \{v_{t-1}(i)a_{ij}b_t(j)\}$
- Xác suất cao nhất : $P^* = \max_{1\le i\le N} \{v_T(i)\}$
- Giá trị khởi tạo : $v_1(j) = \pi_j b_1(j)$ với $1 \le j \le N$
- Do cần xác định chuỗi ẩn tốt nhất nên ta cần 1 đại lượng ghi lại đường đi tức chuỗi các trạng thái qua từng bước theo v_t, đại lượng đó gọi là the Viterbi backtrace :

$$\psi_t(j) = \operatorname*{argmax}_{i=1,2,...,N} \{v_{t-1}(i)a_{ij}b_t(j)\} \text{ với } 1 \le j \le N, 1 < t \le T$$

- Giá trị khởi tạo của backtrace: $\psi_1(j) = 0$ với $1 \le j \le N$
- Đặt h_t^* là trạng thái tốt nhất tại thời điểm t được xác định: $h_t^* = \psi_{t-1}\left(h_{t-1}^*\right)$
- Điểm bắt đầu của backtrace: $h_T^* = \operatorname*{argmax}_{i=1,2,...,N} \{v_T(i)\}$

Cuối cùng, ta đến với bài toán cuối cùng, tìm kiếm mô hình phù hợp nhất với chuỗi các quan sát O, cũng là bài toán quan trọng nhất trong bài báo cáo này.

Problem. Cho chuỗi quan sát được O và tập các trạng thái trong HMM, cần ước lượng bộ tham số $\lambda = (\pi, A, B)$ cho HMM.

Thuật toán tiêu chuẩn để đào tạo HMM là thuật toán Baum - Welch(1972), là 1 trường hợp đặc biệt của thuật toán Expectation-Maximization (thuật toán EM) - được giới thiệu vào năm 1977 bởi Dempster, Laird & Rubin [4].

Thuật toán cho phép chúng ta ước lượng các tham số là ma trận xác suất chuyển của xích ẩn A và xác suất phụ thuộc trạng thái B.

Kí hiệu $O^{(T)} = (o_1, o_2, ..., o_T)$

Mục đích của bài toán 3 chính là tối ưu hàm likelihood:

$$F_{L_T}(\lambda) = P(O^{(T)} \mid \lambda) = \sum_{h_1} \sum_{h_2} ... \sum_{h_T} (\pi_{h_1} b_1 (h_1)) \prod_{t=2}^{T} a_{h_{t-1} h_t} b_t (h_t) \tag{2}$$

thu được nhờ công thức:

$$\begin{aligned} P\left(O^{(T)}, h_1 = q_1, h_2 = q_2, ..., h_T = q_T\right) &= \pi_{q_1} a_{q_1 q_2} a_{q_2 q_3} ... a_{q_{T-1} q_T} b_1(q_1) b_2(q_2) ... b_T(q_T) \\ &= \pi_{q_1} b_1(q_1) \prod_{t=2}^{T} a_{q_{t-1} q_t} b_t(q_t) \end{aligned}$$

Vì log làm hàm số bảo toàn tính đơn điệu của hàm số và cho phép chuyển phép nhân thành phép cộng, dễ phân tích ta sẽ sử dụng hàm số này để xử lí biểu thức trên.

Đối với mô hình Markov ẩn bao gồm các dữ liệu ẩn, khiến cho bộ dữ liệu không phải là dữ liệu đầy đủ [8], thuật toán EM được triển khai bằng thuật toán Baum - Welch, là một trường hợp đặc biệt của thuật toán EM.

Thuật toán Baum - Welch được mô tả như sau:

- E step : thực hiện sự tính toán kỳ vọng. Hàm $L(\lambda, \lambda')$ được định nghĩa tại bước E, bằng cách lấy :

$$L\left(\lambda, \lambda'\right) = E\left[\log P\left(O \mid H, \lambda\right) \mid H, \lambda'\right]$$

- M step: tìm λ sao để hàm L đạt cực đại.

Hàm L là một hàm không giảm, tuy nhiên việc tìm kiếm được nghiệm tối ưu toàn cục thường rất khó thực hiện[8].

Ta có:

$$L(\lambda, \lambda') = E\left[\log P(O, H|\lambda) | H, \lambda'\right] = \sum_{H \in \Omega_T} \log P\left(O, H|\lambda\right) P\left(H|O, \lambda'\right)$$

trong đó λ' là tham số khởi tạo (hoặc có thể là tham số của vòng lặp trước), Ω_T là không gian của mọi chuỗi xích ẩn có độ dài là T.

Giả sử gọi $\lambda^{(k)}$ là bộ tham số tại vòng lặp thứ k . Khi đó, tại vòng lặp thứ $(k+1)$, các bước E và M sẽ được thực hiện:

- Bước E : đã biết $\lambda^{(k)}$, tính $L\left(\lambda, \lambda^{(k)}\right) = E\left[\log P(O, H|\lambda) | H, \lambda^{(k)}\right]$

- Bước M: tìm kiếm $\lambda^{(k+1)}$ sao cho hàm $L\left(\lambda, \lambda^{(k)}\right)$ đạt cực đại:

$$L\left(\lambda^{(k+1)}, \lambda^{(k)}\right) \geq L\left(\lambda, \lambda^{(k)}\right), \forall \lambda$$

Bước E và M sẽ cần lặp lại qua các vòng lặp cho đến khi chuỗi $\left\{\log F_{L_T}\left(\lambda^{(k)}\right)\right\}$hội tụ. Ta sử dụng đánh giá sau để dừng thuật toán:

$$\log F_{L_T}\left(\lambda^{(k+1)}\right) - \log F_{L_T}\left(\lambda^{(k)}\right) \leq \text{tol}$$

với tol là một giá trị cho trước.
Với chuỗi trạng thái ẩn $H \in \Omega_T$ bất kì, ta có công thức:

$$P(O, H|\lambda') = \pi_{h_1} \prod_{t=2}^{T} a_{h_{t-1}h_t} b_t(h_t)$$

Khi đó L được biểu diễn như sau:

$$\begin{aligned} L(\lambda, \lambda') = &\sum_{H \in \Omega_T} \log \pi_{h_1} P\left(H|O, \lambda'\right) + \sum_{H \in \Omega_T} \left(\sum_{t=2}^{T} \log a_{h_{t-1}h_t}\right) P(H|O, \lambda') \\ &+ \sum_{H \in \Omega_T} \left(\sum_{t+1}^{T} \log b_t(h_t)\right) P\left(H|O, \lambda'\right) \end{aligned}$$

Từ đây, ta có cập nhật cho các tham số của mô hình như sau:

- Ma trận xác suất chuyển A:

$$a_{ij} = \frac{\sum_{t=1}^{T} P\left(h_{t-1} = q_i, h_t = q_j|O, \lambda'\right)}{\sum_{t=1}^{T} P\left(h_{t-1} = q_i|O, \lambda'\right)}$$

- Xác suất phụ thuộc trạng thái:

$$b_k(i) = \frac{\sum_{t=1}^{T} P\left(h_t = q_i|O, \lambda'\right) \delta_{o_t, v_k}}{\sum_{t=1}^{T} P\left(h_t = q_i|O, \lambda'\right)}$$

- Phân phối ban đầu:

$$\pi_i = \frac{P\left(h_1 = q_i|O, \lambda'\right)}{P\left(O, \lambda'\right)}$$

Như ta đã đề cập ở trên, thuật toán Baum - Welch là trường hợp đặc biệt của thuật toán EM để phù hợp với bối cảnh áp dụng trong mô hình Markov ẩn.

Kí hiệu $O_t^T = (o_{t+1}, o_{t+2}, ..., o_T)$

Ta đặt ra $\beta_t(i) = P\left(O_t^T | h_t = q_i, \lambda\right)$, được gọi là xác suất lùi (backward - probability)

- Khởi tạo : $\beta_T(i) = 1, 1 \leq i \leq N$
- Đệ quy: $\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_{t+1}(j) \beta_{t+1}(j), 1 \leq i \leq N, 1 \leq t < T$
- $F_{L_T}(O \mid \lambda) = P(O \mid \lambda) = \sum_{j=1}^{N} \pi_j b_1(j) \beta_1(j) = \sum_{j=1}^{N} \alpha_t(j) \beta_t(j), \forall 1 \leq t \leq T$

Kí hiệu $\gamma_t(i) = P(h_t = q_i | O, \lambda)$ là xác suất chuỗi ẩn ở trạng thái q_i tại thời điểm t.

Chú ý rằng:

$$P(h_t = q_i | O, \lambda) = \frac{P(O, h_t = q_i | \lambda)}{P(O|\lambda)} = \frac{P(O, h_t = q_i | \lambda)}{\sum_{j=1}^{N} P(O, h_t = q_j | \lambda)} = \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{N} \alpha_t(j)\beta_j(t)}$$

Kí hiệu $\xi_t(i,j) = P(h_t = q_i, h_{t+1} = q_j | O, \lambda)$ là xác suất ở trạng thái q_i tại thời điểm t và chuyển sang trạng thái q_j ở thời điểm $t+1$.

Ta có :

$$\begin{aligned} \xi_t(i,j) &= \frac{P(h_t = q_i, h_{t+1} = q_j, O|\lambda)}{P(O|\lambda)} \\ &= \frac{\alpha_t(i) a_{ij} b_{t+1}(j) \beta_{t+1}(j)}{\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_t(i) a_{ij} b_{t+1}(j) \beta_{t+1}(j)} \end{aligned}$$

hoặc

$$\begin{aligned} \xi_t(i,j) &= \frac{P(h_t = q_i | O) P(O_{t+1}^T, h_{t+1} = q_j | h_t = q_i, \lambda)}{P(O_{t+1}^T | h_t = q_i, \lambda)} \\ &= \frac{\gamma_t(i) a_{ij} b_{t+1}(j) \beta_{t+1}(j)}{\beta_t(i)} \end{aligned}$$

Khi đó ta có:

- $\sum_{t=1}^{T} \gamma_t(i)$ là số lần trung bình xích ở trạng thái q_i tính từ thời điểm $t = 1$ tới $t = T$
- $\sum_{t=1}^{T-1} \xi_t(i,j)$ là số lần trung bình xích chuyển từ trạng thái q_i sang trạng thái q_j tính từ thời điểm $t = 1$ tới $t = T - 1$

Với V là tập giá trị của các quan sát $V = \{v_1, v_2, \ldots, v_m, ...\}$, ta có thể chuyển các công thức cập nhật các tham số ở trên qua γ_t và ξ_t :

$$a_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)}$$
$$b_k(i) = \frac{\sum_{t=1}^{T} \delta_{o_t,v_k} \gamma_t(i)}{\sum_{t=1}^{T} \gamma_t(i)}$$
$$\pi_i = \gamma_1(i)$$

2.2.3. Poisson Hidden Markov Models

Mô hình Poisson Markov ẩn (PHMMs) là mô hình Markov ẩn đặc biệt, trong đó xác suất phụ thuộc trạng thái tuân theo phân phối Poisson.

Mô hình PHMM là một trường hợp riêng của Poisson hỗn hợp (mixed Poisson model).

Xác suất phụ thuộc trạng thái đối với mô hình PHMM được xác định như sau:

$$S_k(i) = P(O_t = k | h_t = q_i) = e^{-\lambda_i} \frac{\lambda_i^k}{k!}, \forall k, i = 1, 2, ..., N$$

Ta kí hiệu $\pi_t(i) = P(h_t = i)$ là xác suất của xích Markov ẩn ở trạng thái i vào thời điểm t. Khi đó:

$$\sum_{i=1}^{N} \pi_t(i) = 1$$
$$P(O_t = k) = \sum_{i=1}^{N} P(h_t = q_i) P(O_t = k | h_t = q_i) = \sum_{i=1}^{N} \pi_t(i) S_k(i)$$

Biểu diễn dưới dạng ma trận như sau:

$$P(O_t = k) = \pi(t) P(k) 1'$$

với $P(k)$ được định nghĩa là ma trận đường chéo, phần tử trên đường chéo thứ i là $b_k(i)$ và $1'$ là véc tơ một N chiều, còn $\pi(t) = \pi(1) A^{t-1}$ với A là ma trận xác suất chuyển.

Ngoài ra, ta có:

$$E(O_t) = \sum_{i=1}^{N} \pi_1(i) \lambda_i$$

Cuối cùng, chú ý là dữ liệu bị phân tán quá mức, nghĩa là phương sai sẽ lớn hơn kỳ

vọng, điều đó cũng được thể hiện qua công thức:

$$V(O_t) = M'DM + \pi_1'M - (\pi_1'M)^2 > E(O_t) = \pi_1'M$$

trong đó $M = (\lambda_i)_{i=1,2,..,N} = (\lambda_1....\lambda_N)$ là véc tơ gồm N giá trị tham số và $D = \text{diag}\,(\pi_1)$.

Đối với mô hình PHMM, việc ước lượng tham số sẽ có đôi chút khác biệt.

Tập tham số của PHMM bao gồm:

- Phân phối ban đầu : $\pi_1 = (\pi_1(1), \pi_1(2), ..., \pi_1(N))$
- Ma trận xác suất chuyển : $A = (a_{ij})_{i,j=1,..,N}$
- Xác suất phụ thuộc trạng thái : $S_k(i)$

Chúng ta cần tìm kiếm ước lượng cho các tham số trên. Cụ thể hơn, chúng ta cần ước lượng $N^2 - N$ phần tử của ma trận xác suất chuyển A và N tham số Poisson λ_i của xác suất phụ thuộc trạng thái $S_k(i)$[5].

Bằng cách sử dụng ma trận A, ta có thể ước lượng phân phối ban đầu khi phân phối ban đầu là phân phối ổn định : $\pi_1' = \pi_1'A$.

Đặt Θ là véc tơ của các tham số chưa biết dùng để ước lượng hàm hợp lí cực đại

$$\Theta = \left(a_{11}, a_{12}, ..., a_{1(N-1)}, ..., a_{N1}, ..., a_{N(N-1)}, \lambda_1, \lambda_2, ..., \lambda_N\right)$$

và đặt Φ là không gian các tham số.

Áp dụng các kết quả vừa trình bày ở phần mô hình Markov ẩn cùng một số chú ý:

- Tuy tập các tham số là khác nhau nhưng xác suất phụ thuộc trạng thái $b_t(i)$ của HMMs cũng như xác suất phụ thuộc trạng thái của PHMMs, đều tính được nhờ tham số λ_i , $i = 1, 2, ..., N$ nên ta hoàn toàn có thể có được công thức cập nhật các λ_i qua công thức cập nhật của $b_t(i)$.
- Khi đã xác định được A tốt để ước lượng hàm hợp lí cực đại thì ta hiển nhiên tính được phân phối ban đầu π_1. Tuy nhiên, với độ dài của chuỗi quan sát T lớn thì ảnh hưởng của phân phối ban đầu π_1 là không đáng kể
- Tổng số lượng tham số ta cần ước lượng là N^2.

Các tham số được ước lượng qua các vòng lặp như sau[7]:

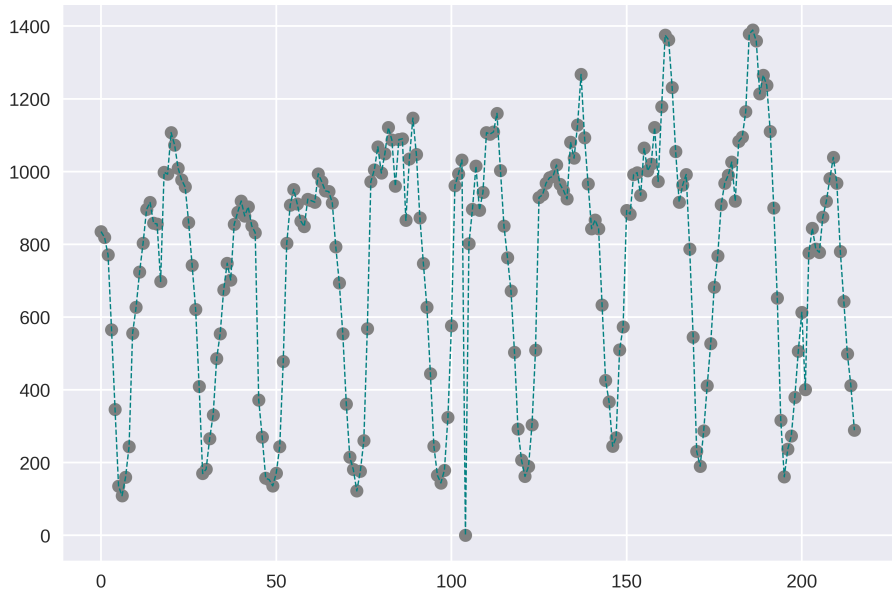
$$a_{ij}^{(k+1)} = \frac{\sum_{t=1}^{T-1} \alpha_t^{(k)}(i) a_{ij}^{(k)} b_{t+1}^{(k)}(j) \beta_{t+1}^{(k)}(j)}{\sum_{t=1}^{T-1} \alpha_t^{(k)}(i) \beta_t(i)}$$

$$\lambda_i^{(k+1)} = \frac{\sum_{t=1}^{T} \alpha_t^{(k)}(i) \beta_t(i) o_t}{\sum_{t=1}^{T} \alpha_t^{(k)}(i) \beta_t(i)}$$

3. Áp dụng mô hình trên bộ dữ liệu thực tế

Mô hình được trình bày ở phần trên được áp dụng trên bộ dữ liệu đếm lượng phương tiện lưu thông trên Đại lộ 4, từ địa chi East 11 street đến East 12 street trong khoảng thời gian từ 2014 đến 2017, được thu thập bởi DOT cho NYMTC (Source: Traffic volume counts collected by DOT for New York Metropolitan Transportation Council (NYMTC) to validate the New York Best Practice Model (NYBPM)).

Dữ liệu được minh họa ở hình **??**



Hình 1: Dữ liệu về số lượng phương tiện tham gia giao thông

Bộ dữ liệu này có phương sai mẫu $\mu = 679.05$ và phương sai mẫu $\sigma^2 = 103585.03$.

Kết quả kiểm định theo các tiêu chuẩn[3, 9] cho thấy bộ dữ liệu đếm này bị phân tán.

3.1. Ước lượng các tham số cho mô hình

Với bộ dữ liệu như trên, ta thực hiện việc ước lượng mô hình phù hợp tương ứng. Bởi mô hình sau ước lượng phụ thuộc vào các tham số ban đầu nên nghiệm thu được là

nghiệm tối ưu địa phương, do đó ta sẽ thực hiện ước lượng mô hình với một vài tham số ban đầu khác nhau ngẫu nhiên để lựa chọn ra được bộ tham số phù hợp nhất.

Thực hiện ước lượng cho với các giá trị khác nhau của N là kích thước không gian trạng thái của mô hình Markov ẩn. Với $N = 1, 2, 3, 4, 5$ ta có các mô hình tương ứng cho bởi bảng 2.

N	λ	π	A
1	738.92	(1)	1
2	265.4, 911.17	$(0, 1)$	0.8127 0.1873 0.0687 0.9313
3	970.88, 590.96, 219.26	$(0, 1, 0)$	0.9106 0.0712 0.0182 0.2686 0.5199 0.2115 0.0257 0.2052 0.7691
4	199.47, 833.05, 1036.99, 499.11	$(0, 0, 1, 0)$	0.7362 0.0295 0 0.2343 0 0.6962 0.1493 0.1545 0.0151 0.1324 0.8525 0 0.2957 0.2011 0.0946 0.4086
5	729.91, 195.73, 1120.02, 933.07, 470.58	$(0, 0, 1, 0, 0)$	0.4168 0 0 0.2493 0.3339 0.0303 0.7273 0 0 0.2424 0 0.0344 0.6264 0.3392 0 0.1283 0 0.1488 0.7229 0 0.1664 0.3334 0 0.1669 0.3333

Bảng 2: Các tham số của mô hình với các kích thước của không gian trạng thái

3.2. Lựa chọn mô hình phù hợp

Một trong những vấn đề cơ bản của thống kê là chọn ra mô hình phù hợp nhất với bộ quan sát hiện tại, và là vấn đề rất phổ biến đối với những mô hình chứa nhiều tham số [2]

Ta sẽ sử dụng hai tiêu chuẩn là AIC[2] và BIC[1] để lựa chọn ra N tối ưu cho bộ dữ liệu. Các tiêu chuẩn AIC, BIC giúp ta tránh khỏi việc chọn phải một mô hình bị underfitting hoặc overfitting. Cụ thể hơn, với k là số lượng tham số cần ước lượng của mô hình $\Theta = (\pi, A, B), O$ là dãy quan sát, ta có:

1. $\text{AIC} = 2k - 2\log\left(L_T\left(O|\Theta\right)\right)$
2. $\text{BIC} = k\log\left(n\right) - 2\log\left(L_T\left(O|\Theta\right)\right)$ trong đó n là số lượng quan sát,

Ta cần chọn N sao cho AIC và BIC đạt cực tiểu. Ta đã biết $0 < N$ và không có chặn trên, do đó để tìm được N là nghiệm tối ưu toàn cục là điều rất khó. Mặt khác với N lớn, mô hình dần trở nên quá phức tạp. Do đó ta sẽ xét N trong khoảng từ 1 đến 8.

Các chỉ số AIC, BIC của từng mô hình cho bởi bảng 3

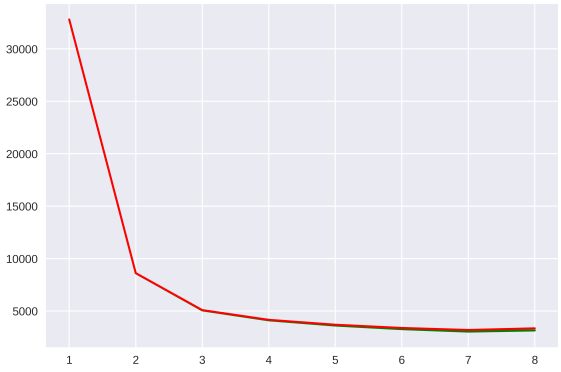
N	1	2	3	4	5	6	7	8
AIC	32792	8605	5062	4112	3616	3273	3038	3147
BIC	32795	8618	5090	4163	3696	3388	3195	3351

Bảng 3: Bảng giá trị AIC và BIC tương ứng

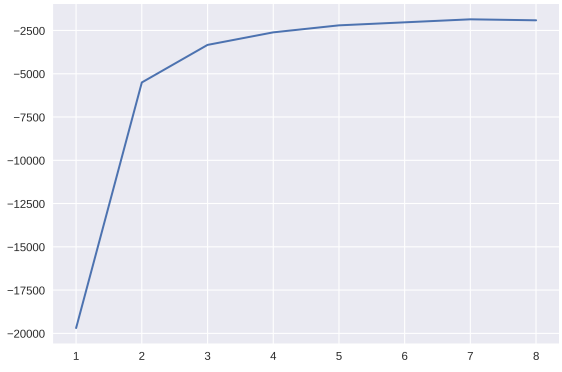
Với mỗi $N = 1, 2, ..., 8$ ta có đồ thị biểu diễn giá trị AIC, BIC tương ứng được cho bởi đồ thị 2a

Hình 2: Biểu diễn các chỉ số AIC, BIC và CDLL của các mô hình

(a) Đồ thị biểu diễn giá trị AIC, BIC tương ứng.



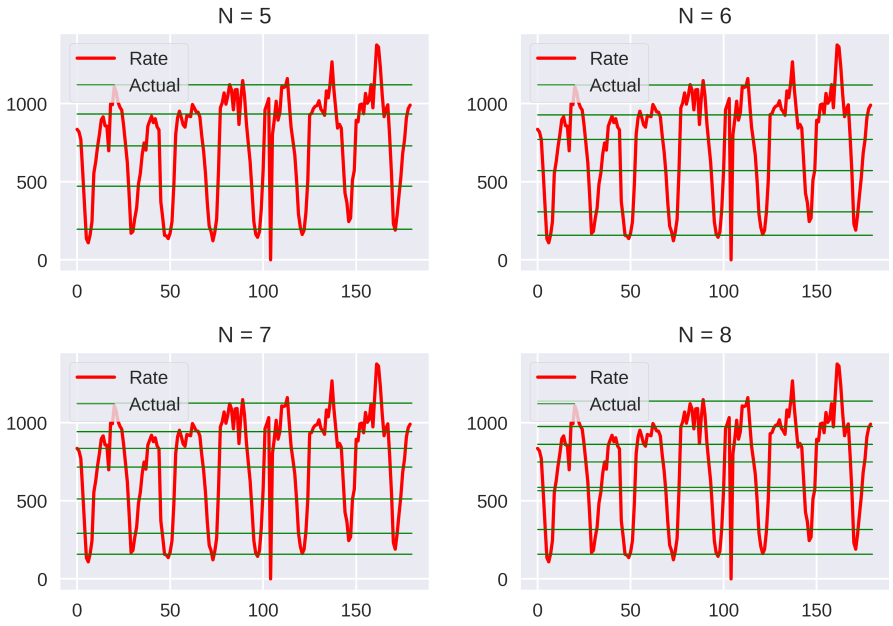
(b) CDLL của các mô hình



Các log-likelihood của các mô hình được cho như hình 2b. Dựa vào đồ thị, nhận thấy với $N = 4, 5, 6, 7, 8$, các giá trị AIC, BIC cũng như CDLL tương ứng với các mô hình không có nhiều khác biệt.

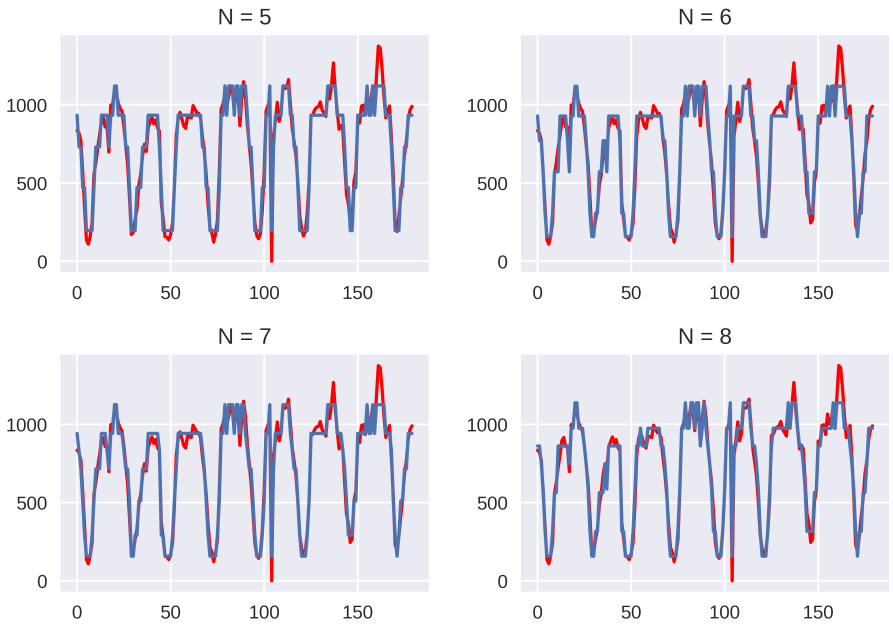
Từ đây có thể dễ dàng thấy rằng $P\left(O \mid \Theta_{(N)}\right)$ không chênh lệch nhau đáng kể với $N = 4, 5, 6, 7, 8$. Để dễ theo dõi, ta sẽ biểu diễn các mô hình tương ứng với giá trị thực của chúng qua hình 3

Hình 3: Biểu đồ minh họa so sánh giữa các mô hình



Các đường nằm ngang trên hình 3 tương ứng với các λ của mỗi mô hình. Nhìn vào đồ thị, có thể thấy rằng các mô hình thu được đều phù hợp với bộ số liệu đưa ra. Sử dụng thuật toán Viterbi[6] và nối các điểm sinh bởi thuật toán này và dữ liệu thực tế cũng cho kết quả rất khả quan, xem ở hình 4.

Hình 4: So sánh giữa giá trị thực và kết quả phân lớp của các mô hình



Từ hình 4 ta thấy mặc dù $N = 7$ cho ra mô hình tốt nhất, tuy nhiên khác biệt này là không quá lớn.

Mô hình với $N = 7$ trạng thái có các tham số cho bởi 3

$$\pi = (0, 1, 0, 0, 0, 0, 0)$$

$$A_7^{\text{PHMM}} = \begin{pmatrix} 0.1668 & 0.2227 & 0.0556 & 0 & 0 & 0.2217 & 0.3333 \\ 0 & 0.7048 & 0 & 0.1367 & 0.1586 & 0 & 0 \\ 0 & 0 & 0.6183 & 0 & 0 & 0.0477 & 0.3340 \\ 0.1111 & 0 & 0 & 0 & 0 & 0.7778 & 0.1111 \\ 0 & 0.3630 & 0.0367 & 0 & 0.6003 & 0 & 0 \\ 0.3007 & 0.2991 & 0 & 0 & 0 & 0.4002 & 0 \\ 0.4435 & 0 & 0.3308 & 0 & 0 & 0 & 0.2257 \end{pmatrix} \tag{3}$$

$$\lambda = (510.75, 941.66, 157.09, 834.76, 1125.32, 715.05, 291.75)$$

3.3. Sử dụng mô hình

Với mô hình đã xây dựng như trên, ta có thể phân loại các trạng thái dựa vào các dữ liệu thu được bằng cách sử dụng thuật toán Viterbi.

Với mỗi chuỗi quan sát mới tại thời điểm $T + \Delta t$, ta sử dụng thuật toán Viterbi để tìm ra chuỗi trạng thái ẩn phù hợp nhất với các quan sát

$$H^* = \underset{H \in \Omega_T}{\text{argmax}} \{P[H \mid O, \lambda]\}$$

Kết quả thu được cho ta biết tình trạng giao thông đang có lưu lượng xe ra sao, từ đó đưa ra các quyết định hợp lí.

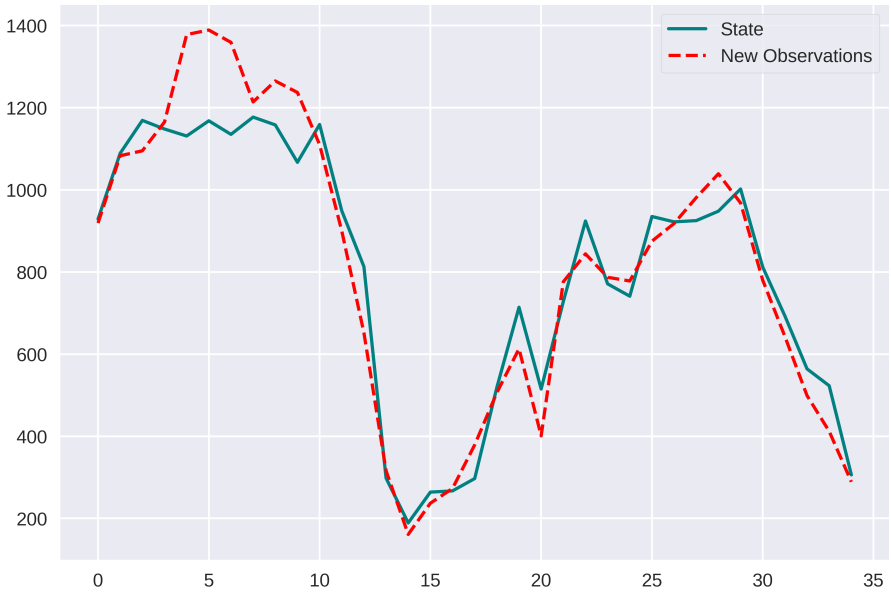
Kết quả phân loại dữ liệu được mô tả bởi hình 5

Bên cạnh đó, ta có thể ước lượng được thời gian lưu lượng xe sẽ ở các mức độ đông xe ra sao. Tuy nhiên, ta đặc biệt quan tâm đến thời gian tắc đường sẽ diễn ra trong bao lâu.

Vấn đề có thể được giải quyết như sau.

Với mô hình λ thu được, ta có ma trận chuyển trạng thái tương ứng với các phân phối phụ thuộc trạng thái đươc cho bởi bảng 4.

Hình 5: Phân loại dữ liệu chuỗi thời gian



CS1	CS2	CS3	CS4	CS5	CS6	CS7
0.1668	0.2227	0.0555	0	0	0.2217	0.3333
0	0.7047	0	0.1367	0.1586	0	0
0	0	0.6183	0	0	0.0477	0.3340
0.1111	0	0	0	0	0.7778	0.1111
0	0.3630	0.0367	0	0.6003	0	0
0.3007	0.2991	0	0	0	0.4002	0
0.4435	0	0.3308	0	0	0	0.2257

(a) Ma trận xác suất chuyển

λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7
510.75	941.66	157.09	834.76	1125.32	715.05	291.75

(b) Tham số của phân phôi phụ thuộc trạng thái

Bảng 4: Các tham số mô hình của dữ liệu quan sát

Dĩ nhiên, ta sẽ chỉ quan tâm đến thời điểm nhiều xe nhất, tức tham số tại thời điểm đó là lớn nhất, là thời điểm xích ở trạng thái 5 (CS5). Ta muốn biết trung bình, quá trình sẽ ở trạng thái này bao lâu trước khi chuyển sang trạng thái khác.

Gọi k_{5A} là số bước trung bình xích xuất phát từ trạng thái $CS5$ và chạm vào tập

trạng thái $A = \{CS1, CS2, CS3, CS4, CS6, CS7\}$. Khi đó, ta có thể tính như sau

$$k_{5A} = 1 + p_{55}k_{5A} + \sum_{j \in A} p_{5j}k_{jA} \tag{4}$$

Hiển nhiên rằng nếu $j \in A$ thì $k_{jA} = 0$. Do đó 4 tương đương với

$$\begin{aligned} k_{5A} &= 1 + p_{55}k_{5A} \\ \Longleftrightarrow k_{5A} &= \frac{1}{1 - p_{55}} \sim 2.5 \end{aligned}$$

Như vậy, trung bình đường sẽ ở trạng thái lưu lượng cao trong khoảng 2.5 đơn vị thời gian trước khi chuyển sang trạng thái có lưu lượng giao thông thấp hơn.

4. Kết luận và tổng kết

Trong báo cáo này, nhóm chúng em đã trình bày về phương pháp xác định quá trình dựa trên dữ liệu đếm quan sát bằng mô hình Poisson Markov ẩn. Dựa vào những kết quả thu được, có thể thấy rằng mô hình hóa dữ liệu đếm bằng PHMMs là một cách tiếp cận khá hiệu quả, bởi cách tiếp cận gần gũi và tính toán không quá phức tạp. Mô hình được triển khai bằng ngôn ngữ Python (Dat T. Dang & Dung T. Nguyen) cho thời gian chạy rất ổn định. Chúng em sẽ phát triển module này trong tương lai để cải thiện năng lực của mô hình này.

Nhóm em đề xuất sử dụng mô hình Poisson Markov ẩn để xử lí các vấn đề về quá mức của dữ liệu đếm. Mô hình Poisson Markov ẩn (PHMM) là phương pháp được sử dụng nhiều nhất để mô hình hóa những dữ liệu như vậy. Cụ thể với bộ dữ liệu thực là đếm số lượng phương tiện đi qua trên một đoạn đường, trước tiên, ta cần kiểm định tính phân tán của dữ liệu, và nếu bộ dữ liệu này là quá mức, nhóm em tiếp tục sử dụng PHMMs để mô hình hóa dữ liệu. Với thuật toán EM, những tham số mô hình đã liên tục được cập nhật và tìm ra bộ tham số đủ tốt. Trong bài báo cáo này, nhóm em sử dụng xích Markov ẩn (HMM) để mô hình hóa các tác nhân ảnh hưởng tới số lượng phương tiện trên đoạn đường. Việc lựa chọn mô hình được thực hiện qua các tiêu chuẩn kiểm định AIC và BIC. Mô hình sau cùng thu được cho kết quả rất tốt trên các tập dữ liệu thu được sau này.

Tài liệu

[1] *Bayesian information criteria*, pp. 211–237, Springer New York, New York, NY, 2008.

[2] Hamparsum Bozdogan, *Model selection and akaike's information criterion (aic): The general theory and its analytical extensions*, Psychometrika **52** (1987), 345–370.

[3] Charmaine B. Dean and Erin R. Lundy, *Overdispersion*, pp. 1–9, American Cancer Society, 2016.

[4] A. P. Dempster, N. M. Laird, and D. B. Rubin, *Maximum likelihood from incomplete data via the em algorithm*, Journal of the Royal Statistical Society: Series B (Methodological) **39** (1977), no. 1, 1–22.

[5] Sebastian George and Ambily Jose, *Generalized poisson hidden markov model for overdispersed or underdispersed count data*, Revista Colombiana de Estadistica **43** (2020), 71–82.

[6] Stan Z. Li and Anil Jain (eds.), *Viterbi algorithm*, pp. 1376–1376, Springer US, Boston, MA, 2009.

[7] Roberta Paroli, *Poisson hidden markov models for time series of overdispersed insurance counts*, 2002.

[8] Nima Sammaknejad, Yujia Zhao, and Bo Huang, *A review of the expectation maximization algorithm in data-driven process identification*, Journal of Process Control **73** (2019), 123–136.

[9] Zhao Yang Yang, James W. Hardin, Cheryl L. Addy, and Quang Hong Vuong, *Testing approaches for overdispersion in poisson regression versus the generalized poisson model.*, Biometrical journal. Biometrische Zeitschrift **49 4** (2007), 565–84.