**HIDDEN MARKOV MODEL**

1. **Tìm hiểu và chuẩn bị dữ liệu**
2. **Truy cập dữ liệu**

* Link dữ liệu: [Time-Series-Analysis-and-Forecasting-/Data/DailyTotalFemaleBirths.csv at main · indrachk/Time-Series-Analysis-and-Forecasting-](https://github.com/indrachk/Time-Series-Analysis-and-Forecasting-/blob/main/Data/DailyTotalFemaleBirths.csv)

1. **Khám phá dữ liệu**

* Date: Ngày (định dạng MM/DD/YYYY)
* Births: Số ca sinh

1. **Tổng quan**
2. **Giới thiệu**

Bộ dữ liệu ‘DailyTotalFemaleBirths.csv’ được sử dụng cùng với mô hình Hidden Markov Model (HMM) để phân tích dữ liệu chuỗi thời gian về số ca sinh nữ hàng ngày trong năm 1959. Tập trung vào việc xây dựng và sử dụng HMM để mô hình hóa các trạng thái ẩn (số ca sinh thấp, trung bình và cao) và suy ra các trạng thái này từ dữ liệu quan sát.

1. **Các mô hình, thuật toán sử dụng**

HMM là một mô hình thống kê được sử dụng để mô tả các chuỗi quan sát được tạo ra từ các trạng thái ẩn, với giả định Markov rằng trạng thái hiện tại chỉ phụ thuộc vào trạng thái liền trước. Trong bài tập lần này, HMM có 3 trạng thái ẩn, ma trận chuyển trạng thái (A), ma trận phát sinh (B) cho 3 nhóm quan sát (Thấp, trung bình, cao) và phân phối trạng thái ban đầu (pi).

Để Training, triển khai thuật toán Baum – Welch để huấn luyện HMM trên bộ dữ liệu ‘DailyTotalFemaleBirths.csv’. Dữ liệu này được phân loại thành 3 nhóm. Thuật toán lặp tối đa 100 vòng và dừng lại khi tham số thay đổi nhỏ hơn 1e-4. Kết quả của các tham số A, B và pi được lưu vào tệp hmm\_model.json

Thuật toán Baul – Welch là một phương pháp học không giám sát thuộc Expectation Maxmization (EM), được sử dụng để ước lượng các tham số của HMM từ dữ liệu quan sát mà không cần biết trạng thái ẩn. Cách triển khai:

* Bước E (Expectation): Sử dụng thuật toán Forward và Backward để tính xác suất các trạng thái ẩn và các chuyển đổi trạng thái dựa trên dữ liệu quan sát
* Bước M (Maximization): Cập nhật các tham số A, B và pi
  + Pi: xác suất trạng thái ban đầu
  + A: Ma trận chuyển trạng thái
  + B: Ma trận phát sinh

Thuật toán Viterbi là một phương pháp lập trình động để tìm chuỗi trạng thái ẩn có xác suất cao nhất tương ứng với mỗi chuỗi quan sát, dựa trên các tham số HMM (A, B, pi). Vai trò của thuật toán này sẽ suy ra chuỗi trạng thái ẩn (0: Thấp, 1: Trung bình, 2: Cao) cho mỗi ngày trong file dữ liệu dựa trên mô hình đã được huấn luyện

1. **Tổng kết:**

Sau khi chạy code, ta sẽ thu được kết quả sau:

A screenshot of a computer

AI-generated content may be incorrect.

* Tham số tối ưu đã được tìm thấy sau 3 vòng lặp. Sau khi chạy các thuật toán và train mô hình, ta xuất mô hình ra file json ‘hmm\_model.json’.

0.46684726

0.5186941

0.01445846

0.25056814

0.39519625

0.33207463

0.27246933

0.39533742

0.07886567

1

0

2

* Ma trận chuyển trạng thái (A): Biểu thị xác suất chuyển động từ một trạng thái ẩn này sang trạng thái ẩn khác . Với 3 trạng thái ẩn là 0,1,2 tương ứng với Thấp, Trung bình và Cao và A[i][j] sẽ là xác suất chuyển từ trạng thái này sang trạng thái khác. Ma trận A cho thấy cách số ca sinh chuyển đổi giữa các trạng thái.
* Ma trận phát sinh (B): là ma trận trong HMM, biểu thị xác suất phát sinh một quan sát từ một trạng thái ẩn. Trong mã, số ca sinh sẽ được phân loại làm 3 nhóm quan sát tương ứng với số ca sinh thấp, trung bình, cao. Ma trận cho biết mỗi trạng thái ẩn có khả năng tạo ra số ca sinh ở mức nào.
* Phân phối trạng thái ban đầu (pi): là một vector trong HMM, biểu thị xác suất bắt đầu ở mỗi trạng thái ẩn tại thời điểm ban đầu và tổng các phần tử bằng 1. Vector xác định trạng thái ban đầu của chuỗi thời gian.