Mô tả Đồ án

Mô hình Markov ẩn

Ngày 8 tháng 3 năm 2022

1 Giới thiệu

Mô hình Markov ẩn (Hidden Markov model) là một mô hình máy học cổ điển thông dụng trong việc xử lý chuỗi (sequence processing). Cụ thể, mô hình này thường được dùng cho các bài toán phân loại các thành phần trong một chuỗi. Có nhiều thể hiện khác nhau của loại bài toán này, bao gồm xác định từ loại của các từ trong một câu (part-of-speech tagging) hoặc nhận dạng tiếng nói (speech recognition).
Trong đồ án này, các bạn sẽ tìm hiểu sâu hơn các thành phần và thuật toán liên quan đến mô hình này. Bên cạnh đó, các bạn sẽ được cài đặt cũng như là sử dụng các cài đặt thư viện sẵn có để áp dụng vào các tình huống thực tế.

2 Yêu cầu

2.1 Lý thuyết (6đ)

Phần Lý thuyết yêu cầu các bạn tìm hiểu và trình bày lại các thành phần và cơ chế hoạt động của mô hình Markov ẩn. Phần trình bày của các bạn cần trả lời được các câu hỏi sau:

1. (0.5đ) Các thành phần của một mô hình Markov ẩn là gì? Chúng khác gì với mô hình Markov?

2. (0.5đ) Các giả thiết (assumption) đặt ra cho mô hình Markov ẩn là gì? Tìm ví dụ các bài toán mà các giả thiết này hợp lý và bất hợp lý.

3. (1.5đ) Cho một mô hình Markov ẩn với các tham số đã biết, thuật toán tiến trước (forward algorithm) được dùng để xác định độ hợp lý (likelihood) của một chuỗi quan sát (observation). Mô tả và đánh giá độ phức tạp của thuật toán tiến trước.

4. (1.5đ) Cho một mô hình Markov ẩn với các tham số đã biết, thuật toán Viterbi được dùng để xác định chuỗi trạng thái (state) khả dĩ nhất. Mô tả và đánh giá độ phức tạp của thuật toán Viterbi.

5. (2đ) Cho một chuỗi quan sát, giả sử ta cho rằng chuỗi quan sát này được sinh ra từ một mô hình Markov ẩn với tham số chưa biết, thuật toán Baum-Welch được dùng để ước lượng các tham số này. Thuật toán Baum-Welch là trường hợp đặc biệt của thuật toán

Kỳ vọng-Tối ưu (Expectation-Maximization, hay EM). Thuật toán này gồm 2 bước: bước E (Expectation, hay kỳ vọng) và bước M (Maximization, hay tối ưu).

(a) (1đ) Mô tả thuật toán Kỳ vọng-Tối ưu tổng quát.

(b) (1đ) Mô tả và đánh giá độ phức tạp của bước E và bước M của thuật toán Baum-Welch.

Lưu ý: Khi mô tả một thuật toán, các bạn cần trình bày được đầu vào và đầu ra của thuật toán. Ngoài ra, cần phải viết được mã giả của thuật toán.

2.2 Cài đặt (3đ)

Phần Cài đặt yêu cầu các bạn cài đặt các thuật toán tìm hiểu trong phần Lý thuyết, và áp dụng nó vào một ví dụ đơn giản.

1. (1.5đ) Cài đặt thuật toán tiến trước, thuật toán Viterbi, và thuật toán Baum-Welch.

2. (1.5đ) Khi làm quản trò, anh Huy thường sử dụng 2 viên xúc xác khác nhau. Viên đầu tiên là một viên xúc xắc cân bằng, mọi mặt đều có cùng xác suất. Viên thứ hai là một viên xúc xắc lỗi, khi tung sẽ có 50% xác suất ra mặt số 6 và 10% xác suất ra mỗi mặt còn lại. Mỗi lần tung, anh sẽ chọn 1 trong 2 viên xúc xắc này để tung. Người chơi không thể biết anh đã tung viên nào, chỉ biết được lần tung đó ra mặt nào.

   Ngoài ra, nếu ở lần tung này, anh Huy sử dụng viên xúc xắc cân bằng, thì có 80% khả năng anh sẽ tiếp tục sử dụng viên xúc xắc này cho lần tung tiếp theo (20% còn lại anh sẽ đổi sang dùng viên lỗi). Con số này là 30% đối với viên lỗi (70% đổi sang dùng viên cân bằng).

   (a) (0.25đ) Mô hình hóa tình huống trên bằng một mô hình Markov ẩn. Cho biết các tham số của mô hình này.

   (b) (0.25đ) Sinh ngẫu nhiên một chuỗi $T = 100$ lần tung đúng theo mô tả trên.

   (c) (0.5đ) Sử dụng thuật toán Viterbi để dự đoán viên xúc xắc được dùng cho mỗi lần tung. Độ chính xác của dự đoán này là bao nhiêu? Hãy lặp lại thí nghiệm này nhiều lần nếu cần thiết. Báo cáo và nhận xét kết quả thu được.

   (d) (0.5đ) Giả sử bạn là một người chơi, hãy sử dụng thuật toán Baum-Welch để ước lượng các tham số cho mô hình Markov ẩn. Hãy lặp lại thí nghiệm nhiều lần nếu cần thiết. Báo cáo và nhận xét kết quả thu được.

Lưu ý:

- Chỉ sử dụng các thư viện nhập xuất và tính toán số cơ bản (ví dụ NumPy).
- Đảm bảo các cài đặt chạy trong thời gian hợp lý.

2.3 Vận dụng (3đ)

Tìm một ứng dụng của mô hình Markov ẩn trong thực tế. Một số ví dụ bao gồm: POS tagging, speech recognition, v.v.
Bạn cần báo cáo:

- Mô tả bài toán (đầu vào và đầu ra kỳ vọng)
- Mô tả các thành phần của mô hình (tập các trạng thái ẩn, các quan sát có thể, v.v.) Các giả thiết của mô hình Markov ẩn có phù hợp với tình huống này hay không?
- Tập dữ liệu đã dùng là gì? Bạn đã áp dụng các bước tiền xử lý nào?
- Bạn đã đánh giá mô hình như thế nào? Kết quả đánh giá sau cùng là gì?
- Nhận xét về kết quả và đề xuất một số hướng cải tiến.

Lưu ý:

- Bạn có thể dùng các cài đặt sẵn có của mô hình Markov ẩn.
- Đồ án không yêu cầu kết quả đánh giá phải tốt. Điều quan trọng là cách xây dựng thí nghiệm và đánh giá mô hình.

3 Quy định

- Đồ án này có thể được thực hiện bởi **cá nhân** hoặc **nhóm không quá 03 người**.
- Ưu tiên sử dụng ngôn ngữ lập trình **Python**. Ngoài ra, cũng có thể sử dụng C/C++ hoặc các ngôn ngữ khác.
- Yêu cầu bài nộp: ba phần cần được tách biệt thư mục
  - Phần Lý thuyết: 01 tập tin báo cáo dạng PDF;
  - Phần Cài đặt: mã nguồn và báo cáo (có thể kết hợp trong 01 tập tin Jupyter Notebook);
  - Phần Vận dụng: mã nguồn và báo cáo (có thể kết hợp trong 01 tập tin Jupyter Notebook).
- **Không gian lận** dưới bất kì hình thức nào. Mọi hành vi gian lận đều sẽ bị đánh giá 0đ. Trong mọi tình huống, quyền quyết định cuối cùng thuộc về các trợ giảng và giảng viên.

  Các hình thức gian lận bao gồm (nhưng chưa đầy đủ):
  - Chép một phần hoặc hoàn toàn bài của nhóm khác, bài làm của khóa trên, và/hoặc tài liệu tham khảo;
  - Nhờ người làm hộ.

Tài liệu

[1] Appendix A: Hidden Markov Model of "Speech and Language Processing" by Daniel Jurafsky & James H. Martin [PDF]