1. Thuật toán liên quan đến Hidden Markov Model (HMM)

Các thuật toán liên quan đến HMM là trung tâm của việc áp dụng mô hình trong các bài toán thực tiễn. Dưới đây là ba thuật toán quan trọng, mỗi thuật toán giải quyết một trong ba bài toán cơ bản của HMM.

1.1. Thuật toán Forward và Backward

1.1.1. Mục đích:

Tính xác suất của một chuỗi quan sát $O=\{O_1,O_2,\ldots,O_T\}$ dựa trên một mô hình HMM $\lambda=(A,B,\pi)$.

1.1.2. Thuật toán Forward

Forward algorithm tính xác suất $P(O \mid \lambda)$ bằng cách sử dụng đệ quy.

- **Biến forward** $\alpha_t(i)$: Xác suất của chuỗi quan sát một phần $O_1,O_2,\ldots,O_t$ và hệ thống ở trạng thái S_i tại thời điểm t:

$$\alpha_t(i) = P(O_1, O_2, \ldots O_t, q_t = S_i|\lambda)$$

- **Quy trình tính toán:**
  1. **Khởi tạo:**

$$\alpha_t(i) = \pi_i b_i(O_1), 1 \leq i \leq N$$

  2. **Đệ quy:**

$$\alpha_{t+1}(j) = \sum_{i=1}^{N} \alpha_t(i)\alpha_{ij}b_j(O_{t+1}), 1 \leq j \leq N, 1 \leq t \leq T-1$$

  3. **Kết thúc:**

$$P(O \mid \lambda) = \sum_{i=1}^{N} \alpha_T(i)$$

- **Độ phức tạp:** $O(N^2T)$.

1.1.3. Thuật toán Backward

Backward algorithm hỗ trợ tính toán tương tự nhưng từ cuối chuỗi quan sát trở về đầu.

- **Biến backward** $\beta_t(i)$: Xác suất của chuỗi quan sát từ $O_{t+1},O_{t+2},\ldots,O_T$, với trạng thái $q_t=S_i$ tại thời điểm t:

$$\beta_t(i) = P(O_{t+1},O_{t+2},\ldots,O_T|q_t = S_i,\lambda)$$

- **Quy trình tính toán:**
  1. **Khởi tạo:**

$$\beta_T(i) = 1,\ 1 \le i \le N$$

  2. **Đệ quy:**

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(O_{t+1})\beta_{t+1}(j), 1 \le t \le T-1$$

  3. **Kết thúc:** Tính xác suất tổng quát:

$$P(O \mid \lambda) = \sum_{i=1}^{N} \pi_i b_i(O_1)\beta_1(i)$$

- **Độ phức tạp:** $O(N^2T)$.

1.2. Thuật toán Viterbi

1.2.1. Mục đích:

Tìm chuỗi trạng thái ẩn tối ưu $Q^*=\{q_1^*,q_2^*,\ldots,q_T^*\}$ giải thích tốt nhất chuỗi quan sát O.

1.2.2. Quy trình tính toán:

- **Biến trạng thái** $\delta_t(i)$: Xác suất lớn nhất của chuỗi trạng thái dẫn đến S_i tại thời điểm t:

$$\delta_t(i) = \max_{q_1,q_2,\dots,q_{t-1}} P(q_1, q_2, \dots, q_t = S_i, O_1, O_2, \dots, O_t \mid \lambda)$$

- **Bước thực hiện:**
  1. **Khởi tạo:**

$$\delta_1(i) = \pi_i b_i(O_1), \quad 1 \le i \le N$$

$$\psi_1(j) = 0, \quad 1 \le i \le N$$

  2. **Đệ quy:**

$$\delta_{t+1}(j) = \max_{i=1}^{N} \delta_t(i) a_{ij} b_j(O_{t+1}), \quad 1 \le j \le N, 1 \le t \le T-1$$

$$\psi_{t+1}(j) = \arg\max_{i=1}^{N} \delta_t(i) a_{ij}, 1 \le j \le N$$

  3. **Kết thúc:**

$$P(Q^*, O \mid \lambda) = \max_{i=1}^{N} \delta_T(i)$$

$$q_T^* = arg\ max_{i=1}^{N} \delta_T(i)$$

  4. **Truy vết trạng thái tối ưu:**

$$q_t^* = \psi_{t+1}(q_{t+1}^*), t = T-1, T-2, \dots, 1$$

- **Độ phức tạp:** $O(N^2T)$.

2. Các giả định của Hidden Markov Model (HMM)

Hidden Markov Model (HMM) dựa trên hai giả định cơ bản, giúp đơn giản hóa việc mô hình hóa và tính toán xác suất trong các bài toán thực tế. Mặc dù những giả định này có thể không hoàn toàn chính xác trong mọi trường hợp, chúng vẫn đủ mạnh để mô tả nhiều hệ thống thực tế một cách hiệu quả.

2.1. Giả định Markov (Markov Assumption)

2.1.1. Định nghĩa:

Giả định Markov phát biểu rằng trạng thái hiện tại qtq_tqt chỉ phụ thuộc vào trạng thái ngay trước đó q_{t-1}, không phụ thuộc vào các trạng thái trước đó trong chuỗi.

$$P(q_t|q_{t-1},q_{t-2},\ldots,q_1) = P(q_t \mid q_{t-1})$$

2.1.2. Ý nghĩa:

- Giả định này giảm độ phức tạp của mô hình, chỉ yêu cầu xét mối quan hệ giữa hai trạng thái liên tiếp thay vì toàn bộ chuỗi trạng thái.
- Trong thực tế, giả định Markov có thể hiểu là một hệ thống "có trí nhớ ngắn hạn", nơi trạng thái hiện tại chứa đủ thông tin để dự đoán trạng thái tiếp theo.

2.1.3. Hạn chế:

- Hệ thống thực tế có thể bị ảnh hưởng bởi nhiều trạng thái trong quá khứ, không chỉ bởi trạng thái ngay trước đó. Tuy nhiên, việc tăng bậc của mô hình Markov (Markov bậc cao hơn) có thể giúp giảm bớt hạn chế này, nhưng làm tăng độ phức tạp tính toán.

2.2. Giả định độc lập quan sát (Independence Assumption)

2.2.1. Định nghĩa:

Giả định này cho rằng mỗi quan sát OtO_tOt tại thời điểm ttt chỉ phụ thuộc vào trạng thái hiện tại qtq_tqt, không phụ thuộc vào các quan sát khác hoặc các trạng thái khác trong chuỗi.

$$P(O_t \mid q_t, q_{t-1}, O_{t-1}, \dots) = P(O_t \mid q_t)$$

2.2.2. Ý nghĩa:

- Giả định này cho phép ta mô hình hóa mối quan hệ giữa trạng thái ẩn và quan sát một cách độc lập, giảm đáng kể độ phức tạp khi tính toán xác suất.
- Đây là một trong những lý do HMM được áp dụng rộng rãi trong các bài toán như nhận dạng giọng nói và gắn thẻ từ loại.

2.2.3. Hạn chế:

- Trong thực tế, các quan sát thường có mối liên hệ phụ thuộc với nhau, đặc biệt trong các chuỗi dữ liệu có tính chất tuần tự cao. Giả định này có thể không hoàn toàn chính xác, nhưng thường được chấp nhận để đơn giản hóa mô hình.

Hai giả định Markov và độc lập quan sát là nền tảng của Hidden Markov Model, giúp mô hình này trở thành một công cụ đơn giản nhưng mạnh mẽ để mô tả các chuỗi dữ liệu tuần tự. Mặc dù có những hạn chế nhất định, chúng cho phép HMM áp dụng hiệu quả trong các bài toán thực tế với độ phức tạp tính toán thấp.

3. Ứng dụng của Hidden Markov Model (HMM) vào Gắn thẻ từ loại (POS Tagging)

Gắn thẻ từ loại (Part-of-Speech Tagging - POS Tagging) là một bài toán quan trọng trong xử lý ngôn ngữ tự nhiên (NLP), nhằm gán nhãn ngữ pháp (danh từ, động từ, tính từ,...) cho từng từ trong câu. Hidden Markov Model (HMM) là một phương pháp phổ biến để giải quyết bài toán này nhờ khả năng mô hình hóa chuỗi trạng thái ẩn (các nhãn từ loại) dựa trên chuỗi quan sát (các từ trong câu).

3.1. Mô hình HMM cho POS Tagging

Để áp dụng HMM vào bài toán POS Tagging, chúng ta cần xác định các thành phần của mô hình:

- **Tập trạng thái ẩn (S):**
  - Là tập các nhãn từ loại (POS tags), ví dụ: S={NN (danh từ),VB (động từ),JJ (tính từ),… }.
- **Tập quan sát (O):**
  - Là tập các từ trong câu, ví dụ: O={The, cat, runs, fast}
- **Phân phối xác suất ban đầu (π):**
  - Xác suất một từ trong câu bắt đầu với một từ loại cụ thể: $\pi_i=P(S_1=i)$ Ví dụ: Một câu thường bắt đầu bằng các nhãn như DT (mạo từ) hoặc NN (danh từ).
- **Ma trận chuyển trạng thái (A):**
  - Xác suất chuyển từ nhãn từ loại này sang nhãn từ loại khác: $a_{ij}=P(S_{t+1}=j|S_t=i)$ Ví dụ: Sau một danh từ (NN), khả năng cao sẽ là một động từ (VB) hoặc mạo từ (DT).
- **Ma trận xác suất phát xạ (B):**
  - Xác suất một nhãn từ loại phát sinh một từ cụ thể: $b_j(O_t)=P(O_t|S_t=j)$ Ví dụ: Xác suất từ "runs" thuộc nhãn động từ (VB) sẽ cao hơn các nhãn khác.

3.2. Thuật toán Viterbi để giải bài toán POS Tagging

POS Tagging sử dụng thuật toán Viterbi để tìm chuỗi nhãn từ loại tối ưu $S^*=\{S_1^*,S_2^*,\ldots,S_T^*\}$ tương ứng với chuỗi quan sát $O=\{O_1,O_2,\ldots,O_T\}$.

Quy trình thực hiện:

B1: Khởi tạo: Tại thời điểm t=1:

$$\delta_1(i)=\pi_i \cdot b_i(O_1), \quad \psi 1(i)=0$$

- $\delta_1(i)$: Xác suất lớn nhất khi bắt đầu với trạng thái S_i.
- $\psi_1(i)$: Truy vết trạng thái trước đó, tại thời điểm khởi đầu, giá trị này bằng 0.

B2: Đệ quy: Từ t=2 đến T (số lượng từ trong câu):

$$\delta_t(j) = \max_i \left[\delta_{t-1}(i) \cdot a_{ij} \cdot b_j(O_t)\right], \psi_t(j) = arg\max_i \left[\delta_{t-1}(i) \cdot a_{ij}\right]$$

- $\delta_t(j)$: Xác suất lớn nhất dẫn đến trạng thái S_j tại thời điểm t.
- $\psi_t(j)$: Truy vết trạng thái S_i tốt nhất trước S_j.

B3: Kết thúc: Tại thời điểm cuối T:

$$S_T^* = arg\max_i \delta_T(i)$$

B4: Truy vết: Từ t=T−1 đến t=1:

$$S_t^* = \psi_{t+1}(S_{t+1}^*)$$

St∗=ψt+1(St+1∗)S_t^* = \psi_{t+1}(S_{t+1}^*)St∗=ψt+1(St+1∗)

- Kết quả là chuỗi nhãn từ loại tối ưu $S^*=\{S_1^*,S_2^*,\ldots,S_T^*\}$.

3.3. Ví dụ minh họa

Đề bài: Cho câu quan sát:

O={"The", "cat", "runs"}

Với tập nhãn từ loại:

S={DT (mạo từ),NN (danh từ),VB (động từ)}

Các tham số mô hình:

- π={P(DT)=0.6,P(NN)=0.3,P(VB)=0.1}.
- Ma trận chuyển trạng thái:

$$A = \begin{bmatrix} P(DT \rightarrow DT) & P(DT \rightarrow NN) & P(DT \rightarrow VB) \\ P(NN \rightarrow DT) & P(NN \rightarrow NN) & P(NN \rightarrow VB) \\ P(VB \rightarrow DT) & P(VB \rightarrow NN) & P(VB \rightarrow VB) \end{bmatrix} = \begin{bmatrix} 0 & 0.7 & 0.3 \\ 0.1 & 0.4 & 0.5 \\ 0.6 & 0.3 & 0.1 \end{bmatrix}$$

- Ma trận phát xạ:

$$B = \begin{bmatrix} P(O \mid DT) \\ P(O \mid NN) \\ P(O \mid VB) \end{bmatrix} = \begin{bmatrix} P(\text{"The"}) = 0.5, P(\text{"cat"}) = 0.1, P(\text{"}runs\text{"}) = 0.1 \\ P(\text{"The"}) = 0.1, P(\text{"cat"}) = 0.6, P(\text{"}runs\text{"}) = 0.1 \\ P(\text{"The"}) = 0.1, P(\text{"cat"}) = 0.1, P(\text{"}runs\text{"}) = 0.8 \end{bmatrix}$$

Giải:

- **Khởi tạo:**

$$\delta_1(DT) = \pi_{DT} \cdot b_{DT}(\text{"The"}) = 0.6 \cdot 0.5 = 0.3$$

$$\delta_1(NN) = \pi_{NN} \cdot b_{NN}(\text{"The"}) = 0.3 \cdot 0.1 = 0.03$$

$$\delta_1(VB) = \pi_{VB} \cdot b_{VB}(\text{"The"}) = 0.1 \cdot 0.1 = 0.01$$

- **Đệ quy (tại** t=2**):**

$$\delta_2(NN) = max[\delta_1(DT) \cdot a_{DT \to NN}, \delta_1(NN) \cdot a_{NN \to NN}, \delta_1(VB) \cdot a_{VB \to NN}] \cdot b_{NN}("cat")$$

- **Tiếp tục:**
  Lặp lại các bước trên cho đến t=3 để tìm chuỗi nhãn tối ưu.