

Notation für Hidden-Markov-Modelle

Notationssammlung zur Vorlesung "Algorithmische Bioinformatik", Wintersemester 2003/04 an der FU Berlin. Author: Alexander Schliep.

Ein Hidden-Markov-Modell (HMM) besteht aus einer Markovkette über einer diskreten, endliche Menge von N Zuständen $\mathcal{S} = \{S_1, S_2, \dots, S_N\}$; häufig wird ein Zustand S_i einfach als i notiert. Eine Folge von Zuständen der Länge T schreiben wir als $Q = q_1 q_2 \dots q_T$. Die Wahrscheinlichkeit eines Überganges von einem Zustand i zur Zeit t zu einem Zustand j im nächsten Zeitschritt, wird als $a_{ij} := \mathbb{P}[q_{t+1} = j | q_t = i]$ bezeichnet und ergibt die $N \times N$ -Matrix $A := \{a_{ij}\}$. Aus der Definition folgt, daß A (zeilen-)stochastisch ist.

Diese Markovkette wird um einen weiteren stochastischen Prozess ergänzt. Sei $\Sigma = \{v_1, v_2, \dots, v_M\}$ ein sog. Ausgabealphabet der Kardinalität M . Die Ausgabewahrscheinlichkeit, $\mathbb{P}[o_t = v_m | q_t = j]$ wird mit $b_j(v_m)$ bezeichnet und ist unabhängig von t ; d.h. die Wahrscheinlichkeit einer Ausgabe hängt alleine vom Zustand ab, in dem sie erfolgt. Zusammengefasst werden die Ausgaben in der Ausgabematrix $B = \{b_j(v_m)\}_{1 \leq j \leq N, 1 \leq m \leq M}$.

Der Name Hidden-Markov-Modell ergibt sich, weil meist aus einer Sequenz von Ausgaben $O = o_1 o_2 \dots o_T$ (Observations-, Ausgabe-Sequenz) die dazugehörige Zustandssequenz $Q = q_1 q_2 \dots q_T$ nicht eindeutig bestimmen läßt; d.h. die Markovkette kann nicht beobachtet werden.

Mit $\lambda := (A, B, \pi)$ bezeichnen wir ein HMM.

$\mathbb{P}[O|\lambda]$ Berechnen: Forward-Variablen

Definition:

$$\alpha_t(i) := \mathbb{P}[o_1 o_2 \dots o_t, q_t = i | \lambda]. \quad (1)$$

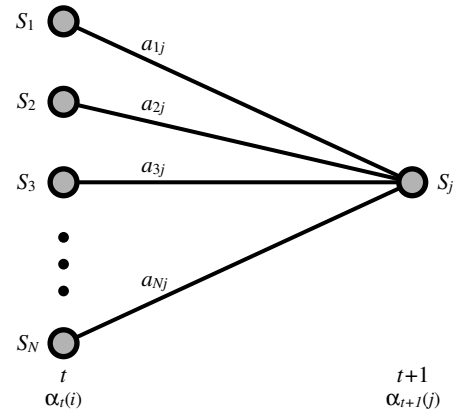
Zur Berechnung:

1. Initialisierung für $t = 1$:

$$\alpha_1(i) = \pi_i b_i(o_1), \quad \text{für alle } i. \quad (2)$$

2. Induktion für $t = 1, 2, \dots, T - 1$:

$$\alpha_{t+1}(i) = \sum_{j=1}^N \alpha_t(j) a_{ji} b_i(o_{t+1}), \quad \text{für alle } i. \quad (3)$$



Proposition:

$$\mathbb{P}[O|\lambda] = \sum_{i=1}^N \alpha_T(i) \quad (4)$$

Berechnung eines optimalen Q gegeben O : Der Viterbi-Algorithmus

Definition:

$$\delta_t(i) := \max_{q_1, q_2, \dots, q_t} \mathbb{P}[q_1 q_2 \dots q_{t-1}, q_t = i, o_1 o_2 \dots o_t | \lambda]. \quad (5)$$

Die $\psi_t(i)$ enthalten die für das Backtracking wesentlichen Informationen: Von welchem Zustand zur Zeit t ausgehend haben wir im $t + 1$ -ten Schritt das maximale $\delta_{t+1}(j)$ erhalten? Zur Berechnung:

1. Initialisierung für $t = 1$:

$$\begin{aligned} \delta_1(i) &= \pi_i b_i(o_1), \quad \text{für alle } i, \\ \psi_1(i) &= 0 \end{aligned}$$

2. Induktion für $t = 1, 2, \dots, T - 1$:

$$\begin{aligned} \delta_{t+1}(j) &= \max_{1 \leq i \leq N} (\delta_t(i) \cdot a_{ij}) \cdot b_j(o_{t+1}), \quad \text{für alle } j, \\ \psi_{t+1}(j) &= \arg \max_{1 \leq i \leq N} (\delta_t(i) \cdot a_{ij}). \end{aligned}$$

3. Berechnung von P^* , der Wahrscheinlichkeit des Viterbipfades:

$$\begin{aligned} P^* &= \max_{1 \leq i \leq N} \delta_T(i), \\ q_T^* &= \arg \max_{1 \leq i \leq N} \delta_T(i). \end{aligned}$$

4. Backtracking um den Viterbi-Pfad Q^* zu berechnen:

$$q_t^* = \psi_{t+1}(q_{t+1}^*), \quad t = T - 1, T - 2, \dots, 1. \quad (6)$$

Training von HMMs: Der Baum-Welch-Algorithmus

Problem: Gegeben O und λ , finde $\hat{\lambda}$ s.d. $\mathbb{P}[O|\hat{\lambda}]$ lokal maximal ist.

Lösung: Iteriere die folgenden Re-estimierungen bis Konvergenz erreicht wird.

$$\bar{a}_{ij} := \frac{\text{Erwartete Anzahl der Übergänge } S_i \rightarrow S_j}{\text{Erwartete Anzahl der Übergänge aus } S_i}$$

$$\bar{b}_{jm} := \frac{\text{Erwartete Anzahl der Ausgaben } v_m \text{ in } S_j}{\text{Erwartete Anzahl der Übergänge nach } S_j}$$

$$\bar{\pi}_i := \text{Wahrscheinlichkeit von } q_1 = S_i.$$

Backward-Variablen

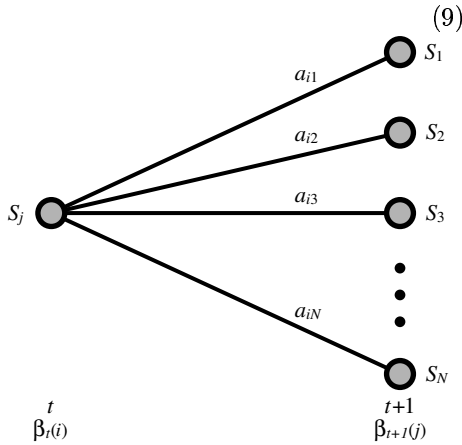
$$\beta_t(i) := \mathbb{P}[o_{t+1}, o_{t+2}, \dots, o_T | q_t = i, \lambda]. \quad (7)$$

1. Initialization for $t = T$:

$$\beta_T(i) = 1, \quad \text{für alle } i. \quad (8)$$

2. Induction for $t = T - 1, T - 2, \dots, 1$:

$$\beta_t(i) = \sum_{j=1}^N a_{ij} \cdot b_j(o_{t+1}) \cdot \beta_{t+1}(j), \quad \text{für alle } i. \quad (9)$$



Reestimierungsformeln

Mit den Backward-Variablen können wir die Erwartungswerte oben direkt ausrechnen.

$$\gamma_t(i) := \frac{\mathbb{P}[q_t = i, O | \lambda]}{\mathbb{P}[O | \lambda]} \quad (10)$$

$$= \frac{\alpha_t(i) \beta_t(i)}{\sum_{k=1}^N \alpha_t(k) \beta_t(k)}, \quad (11)$$

$$\xi_t(i, j) := \frac{\alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}{\sum_{k=1}^N \alpha_t(k) \beta_t(k)}, \quad (12)$$

Es gilt dann

$$\sum_{t=1}^T \gamma_t(i) = \text{Erwartete Anzahl der Übergänge aus } S_i$$

und

$$\sum_{t=1}^{T-1} \xi_t(i, j) = \text{Erwartete Anzahl der Übergänge } S_i \rightarrow S_j$$

Damit

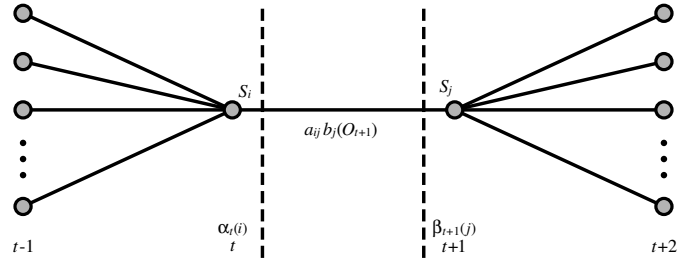
$$\bar{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i, j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \quad (13)$$

$$\bar{b}_{jm} = \frac{\sum_{t=1, O_t=v_m}^T \gamma_t(j)}{\sum_{t=1}^T \gamma_t(j)} \quad (14)$$

$$\bar{\pi}_i = \gamma_1(i) \quad (15)$$

Zur Interpretation hilft es, die Definition von $\xi_t(i, j)$ und $\gamma_t(i)$ auszunutzen und sich die "Trellis" (s.u.) anzuschauen

$$\bar{a}_{ij} = \frac{\sum_{t=1}^{T-1} \alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(i)}{\sum_{t=1}^{T-1} \alpha_t(i) \beta_t(j)}$$



Training mit mehreren Sequenzen

Gegeben $\mathcal{O} = \{O^{(1)}, O^{(2)}, \dots, O^{(K)}\}$. Betrachte die Sequenzen als unabhängig und optimiere

$$\mathbb{P}[\mathcal{O} | \lambda] = \prod_{k=1}^K \mathbb{P}[O^{(k)} | \lambda]$$

Es folgen die modifizierten Reestimierungsformeln wie z.B.

$$\bar{a}_{ij} = \frac{\sum_{k=1}^K \sum_{t=1}^{T-1} \xi_t^k(i, j)}{\sum_{k=1}^K \sum_{t=1}^{T-1} \gamma_t^k(i)}$$