6.4.6 Parameterschätzung – Lernen der HMM (Überblick)

Schätzproblem

Gegeben: Beobachtungsfolge: $o = o_1, o_2, \ldots, o_T$

$1 \leq o_i \leq L$

Gesucht: Parameter eines HMM λ

das die Beobachtungsfolge am wahrscheinlichsten generiert

oder $P(o \mid \lambda)$ $\xrightarrow{\lambda}$ max

$P(o, s^* \mid \lambda)$

Parameterschätzung

- es ist kein Verfahren bekannt, das zu einer gegebenen Beobachtungsfolge ein optimales HMM in geschlossener Form liefern kann
- sofern es jedoch gelingt, eine geeignete Modellstruktur (Anzahl der Zustände, Art der Emissionsverteilungen) vorzugeben und deren freie Parameter mit sinnvollen initialen Werten zu belegen, kann eine iterative Optimierung des Modells in Bezug auf die betrachteten Daten erfolgen
- diese schrittweise Verbesserung der Modellparameter bezeichnet man als Training des Modells
- diese iterativen Methoden finden zumindest ein lokales Maximum

Methoden

- Baum – Welch – Algorithmus
  - Maximum – Likelihood – Schätzung

Iteration: $\lambda \rightarrow \hat{\lambda}$

$$P(o \mid \hat{\lambda}) \geq P(o \mid \lambda)$$

$P(o \mid \lambda)$ $\xrightarrow{\lambda}$ max

nichtlineares Optimierungsproblem unter linearen Nebenbedingungen

Methoden

- Viterbi – Training

Iteration: $\lambda \rightarrow \hat{\lambda}$

$$P(o, s^* \mid \hat{\lambda}) \geq P(o, s^* \mid \lambda)$$

$$P^*(o \mid \lambda) = P(o, s^* \mid \lambda)$$

λ → max

nichtlineares Optimierungsproblem unter linearen Nebenbedingungen

6.4.7 Baum – Welch – Algorithmus

Diskrete HMM

- iteratives Trainingsverfahren, welches ein gegebenes HMM schrittweise verbessert
- das gegebene kann ein Initial – HMM sein oder ein HMM aus einer vorherigen Trainingsiteration
- Anzahl N der Zustände und Anzahl der Beobachtungen L bleibt fest

Initial HMM

- gewünschte Topologie des HMM wählen
  - Anzahl der Zustände N
  - Anzahl der diskreten Beobachtungen L
  - welche Zustandsübergänge sind erlaubt
- Übergangswahrscheinlichkeiten (z.B. gleichverteilt)
- Beobachtungswahrscheinlichkeiten (z.B. gleichverteilt)

Hilfswahrscheinlichkeiten

$$\gamma_n(i) = P(S_n = i \mid o, \lambda)$$

Wahrscheinlichkeit zum Zeitpunkt n im Zustand i zu sein

$$\alpha_n(i) \cdot \beta_n(i) = P(o, S_n = i \mid \lambda)$$

$$1 \le n \le T$$

$$P(o \mid \lambda) = \sum_{i=1}^{N} \alpha_n(i) \cdot \beta_n(i)$$

$$\gamma_n(i) = \frac{P(S_n = i, o \mid \lambda)}{P(o \mid \lambda)} = \frac{\alpha_n(i)\beta_n(i)}{\sum_{i=1}^{N} \alpha_n(i)\beta_n(i)}$$

Hilfswahrscheinlichkeiten

$$\gamma_n(i, j) = P(S_n = i, S_{n+1} = j \mid o, \lambda)$$

Wahrscheinlichkeit zum Zeitpunkt n im Zustand i und zum Zeitpunkt n+1 im Zustand j zu sein

$$\gamma_n(i, j) = \frac{P(S_n = i, S_{n+1} = j, o \mid \lambda)}{P(o \mid \lambda)} = \frac{\alpha_n(i) a_{ij} b_{jo_{n+1}} \beta_{n+1}(j)}{\sum_{i=1}^{N} \alpha_n(i) \beta_n(i)}$$

$$1 \leq n < T$$

Hilfswahrscheinlichkeiten

$$\gamma_n(i) = \sum_{j=1}^{N} \gamma_n(i, j) \qquad 1 \leq n < T$$

Neue Parameter

$$\hat{e}_i = \gamma_1(i) = \frac{\alpha_1(i)\beta_1(i)}{\sum_{i=1}^{N} \alpha_1(i)\beta_1(i)}$$

$$\hat{a}_{ij} = \frac{\sum_{n=1}^{T-1} \gamma_n(i,j)}{\sum_{n=1}^{T-1} \gamma_n(i)} = \frac{\sum_{n=1}^{T-1} \alpha_n(i) a_{ij} b_{jo_{n+1}} \beta_{n+1}(j)}{\sum_{n=1}^{T-1} \alpha_n(i)\beta_n(i)}$$

$$\hat{b}_{jk} = \frac{\sum_{n:o_n=k} \gamma_n(j)}{\sum_{n=1}^{T} \gamma_n(j)} = \frac{\sum_{n:o_n=k} \alpha_n(j)\beta_n(j)}{\sum_{n=1}^{T} \alpha_n(j)\beta_n(j)}$$

$$P(o \mid \hat{\lambda}) \geq P(o \mid \lambda)$$

Bemerkung

- bisher wird nur eine einzelne Beobachtungssequenz verwendet
- es können aber mehrere voneinander unabhängige Sequenzen benutzt werden
- in den Formeln für die neuen Parameter summiert man einfach über alle Beobachtungssequenzen auf

CDHMM

$$b_j(\mathbf{x}) = \sum_{k=1}^{M_j} c_{jk} \cdot g_{jk}(\mathbf{x}) = \sum_{k=1}^{M_j} c_{jk} \cdot N(\mathbf{x}, \mu_{jk}, \Sigma_{jk})$$

$$\sum_{k=1}^{M_j} c_{jk} = 1, \qquad c_{jk} \geq 0$$

$$b_j(\mathbf{x}) = \sum_{k=1}^{M} c_{jk} \cdot g_k(\mathbf{x}) = \sum_{k=1}^{M} c_{jk} \cdot N(\mathbf{x}, \mu_k, \Sigma_k)$$

CDHMM

Wahrscheinlichkeit zum Zeitpunkt n im Zustand j
die k-te Mischverteilungskomponente zur Erzeugung
der kontinuierlichen Beobachtung $\mathbf{x}_n$ zu verwenden

$$\xi_n(j,k) = P(S_n = j, M_n = k \mid O, \lambda) = \frac{\sum_{i=1}^{N} \alpha_{n-1}(i) a_{ij} c_{jk} g_{jk}(\mathbf{x}_n) \beta_n(j)}{P(O \mid \lambda)}$$

$$n > 1$$

$$\xi_1(j,k) = \frac{e_j c_{jk} g_{jk}(\mathbf{x}_1) \beta_1(j)}{P(O \mid \lambda)}$$

Schätzung der Mischungsgewichte

$$b_j(\mathbf{x}) = \sum_{k=1}^{M_j} c_{jk} \cdot g_{jk}(\mathbf{x}) = \sum_{k=1}^{M_j} c_{jk} \cdot N(\mathbf{x}, \mu_{jk}, \Sigma_{jk})$$

$$\hat{c}_{jk} = \frac{\sum_{n=1}^{T} \xi_n(j,k)}{\sum_{n=1}^{T} \gamma_n(j)}$$

Schätzung der Mischverteilungskomponenten

Analog zum EM – Algorithmus
(Schätzung von Mischverteilungsmodellen)

$$b_j(\mathbf{x}) = \sum_{k=1}^{M_j} c_{jk} \cdot g_{jk}(\mathbf{x}) = \sum_{k=1}^{M_j} c_{jk} \cdot N(\mathbf{x}, \mu_{jk}, \Sigma_{jk})$$

$$\hat{\mu}_{jk} = \frac{\sum_{n=1}^{T} \xi_n(j,k)\mathbf{x}_n}{\sum_{n=1}^{T} \xi_n(j,k)} \qquad \hat{\Sigma}_{jk} = \frac{\sum_{n=1}^{T} \xi_n(j,k)(\mathbf{x}_n - \hat{\mu}_{jk})(\mathbf{x}_n - \hat{\mu}_{jk})^T}{\sum_{n=1}^{T} \xi_n(j,k)}$$

Schätzung der Mischverteilungskomponenten

$$\hat{\Sigma}_{jk} = \frac{\sum_{n=1}^{T} \xi_n(j,k)(\mathbf{x}_n - \hat{\mu}_{jk})(\mathbf{x}_n - \hat{\mu}_{jk})^T}{\sum_{n=1}^{T} \xi_n(j,k)}$$

$$\hat{\Sigma}_{jk} = \frac{\sum_{n=1}^{T} \xi_n(j,k)\mathbf{x}_n\mathbf{x}_n^T}{\sum_{n=1}^{T} \xi_n(j,k)} - \hat{\mu}_{jk}\hat{\mu}_{jk}^T$$

Schätzung der Mischverteilungskomponenten

$$b_j(\mathbf{x}) = \sum_{k=1}^{M} c_{jk} \cdot g_k(\mathbf{x}) = \sum_{k=1}^{M} c_{jk} \cdot N(\mathbf{x}, \mu_k, \Sigma_k)$$

$$\hat{c}_{jk} = \frac{\sum_{n=1}^{T} \xi_n(j,k)}{\sum_{n=1}^{T} \gamma_n(j)}$$

Schätzung der Mischverteilungskomponenten

$$b_j(\mathbf{x}) = \sum_{k=1}^{M} c_{jk} \cdot g_k(\mathbf{x}) = \sum_{k=1}^{M} c_{jk} \cdot N(\mathbf{x}, \mu_k, \Sigma_k)$$

$$P(M_n = k \mid O, \lambda) = \xi_n(k) = \sum_j \xi_n(j,k)$$

$$\hat{\mu}_k = \frac{\sum_{n=1}^{T} \xi_n(k)\mathbf{x}_n}{\sum_{n=1}^{T} \xi_n(k)}$$

$$\hat{\Sigma}_k = \frac{\sum_{n=1}^{T} \xi_n(k)\mathbf{x}_n\mathbf{x}_n^T}{\sum_{n=1}^{T} \xi_n(k)} - \hat{\mu}_k\hat{\mu}_k^T$$

6.4.8 Viterbi – Training

Prinzip

$$P^*(o \mid \lambda) = P(o, s^* \mid \lambda)$$

Iteration: $\lambda \rightarrow \hat{\lambda}$

$$P^*(o \mid \hat{\lambda}) \geq P^*(o \mid \lambda)$$

Viterbi – Training für diskrete HMM

$$\chi_n(i) = \begin{cases} 1 & \mathrm{s}_\mathrm{n}^* = i,\ s^* = \underset{\mathrm{s}}{\operatorname{argmax}}\, P(s, o \mid \lambda) \\ 0 & \text{sonst} \end{cases}$$

1. Initialisierung: λ

2.Segmentierung: s^* (6.4.5)

Gegeben: Beobachtungsfolge: $o = o_1, o_2, \ldots, o_T \quad 1 \le o_i \le L$

HMM λ

Neue Parameter

$$\chi_n(i) = \begin{cases} 1 & s_n^* = i,\ s^* = \underset{s}{\operatorname{argmax}}\, P(s, o \mid \lambda) \\ 0 & \text{sonst} \end{cases}$$

Sinnvolle Schätzungen für Startwahrscheinlichkeiten erhält man mit Hilfe des Viterbi – Trainings nicht

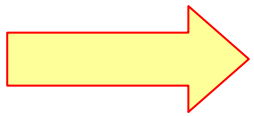
$$\hat{e}_i = \gamma_1(i) \qquad \Rightarrow \qquad \hat{e}_i = \chi_1(i)$$

Da aber die Startwahrscheinlichkeiten in praktischen Anwendungen kaum einen Einfluss ausüben, bedeutet dies keine wesentliche Einschränkung des Verfahrens

Neue Parameter

$$\chi_n(i) = \begin{cases} 1 & \mathrm{s}_n^* = i,\ s^* = \underset{s}{\operatorname{argmax}}\, P(s, o \mid \lambda) \\ 0 & \text{sonst} \end{cases}$$

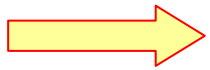
$$\hat{a}_{ij} = \frac{\sum_{n=1}^{T-1} \gamma_n(i,j)}{\sum_{n=1}^{T-1} \gamma_n(i)}$$



$$\hat{a}_{ij} = \frac{\sum_{n=1}^{T-1} \chi_n(i)\chi_{n+1}(j)}{\sum_{n=1}^{T-1} \chi_n(i)}$$

Neue Parameter

$$\hat{b}_{jk} = \frac{\sum_{n:o_n=k}^{T} \gamma_n(j)}{\sum_{n=1}^{T} \gamma_n(j)}$$



$$\hat{b}_{jk} = \frac{\sum_{n:o_n=k}^{T} \chi_n(j)}{\sum_{n=1}^{T} \chi_n(j)}$$

CDHMM

- Neuschätzung kontinuierlicher Mischverteilungsmodelle hier deutlich schwieriger
- segmental k – means - Algorithmus

6.4.9 Verwendung von Logarithmen

Idee

- die Werte der Wahrscheinlichkeiten werden mit wachsender Beobachtungssequenzlänge exponentiell kleiner
- deshalb Verwendung von Logarithmen
- Multiplikationen in den Rekursionsgleichungen werden in Additionen umgewandelt

Kingsbury – Rayner – Formel

Logarithmus einer Summe:

$$\log_b(x+y) = \log_b\left(x\left(1+\frac{y}{x}\right)\right)$$

$$= \log_b x + \log_b\left(1+\frac{y}{x}\right)$$

$$= \log_b x + \log_b\left(1+b^{\log_b y - \log_b x}\right)$$