Stochastische Lernalgorithmen in der CL

Gerhard Jäger

30. April 2003

Hidden Markov Models

Das Konzept

- Erste Sichweise: HMMs sind probabilistische Versionen von regulären Grammatiken
- Bsp.:
  - Non-terminals: *S, H, C, F*
  - Terminals: *1, 2, 3*
  - Startsymbol: *S*
  - Regeln:

$S \rightarrow H$	.5
$S \rightarrow C$	.5
$H \rightarrow H, 1$	.08
$H \rightarrow H, 2$	.16
$H \rightarrow H, 3$	.56
$H \rightarrow C, 1$	.01
$H \rightarrow C, 2$	.02
$H \rightarrow C, 3$	.07
$C \rightarrow H, 1$	.07
$C \rightarrow H, 2$	.02
$C \rightarrow H, 3$	.01
$C \rightarrow C, 1$	.56
$C \rightarrow C, 2$	.16
$C \rightarrow C, 3$	.08
$H \rightarrow F, 1$	.01
$H \rightarrow F, 2$	.02
$H \rightarrow F, 3$	.07
$C \rightarrow F, 1$	.07
$C \rightarrow F, 2$	.02
$C \rightarrow F, 3$	.01

  - Jede Regel ist mit Gewicht (Wahrscheinlichkeit) annotiert
  - Gewichte der Regeln mit identischer LHS summieren sich zu 1.0
- Alternative (äquivalente) Sichweise: HMMs sind probabilistische endliche Automaten
- Jeder Zustand ist mit bestimmter Wahrscheinlichkeit Anfangszustand (summiert sich fr alle Zustnde auf 1.0)
- Jeder Zustandübergang hat Gewicht/Wahrscheinlichkeit
- Summe der Gewichte aller Transitionen aus einem Zustand = 1.0
- Jeder Zustand emmitiert ein Terminalsymbol
- Wahl des Emissionssymbols ebenfalls probabilistisch
- zu obiger PRG äquivalenter Automat:
  - Zustände: *H, C, F*
  - Terminalsymbole: *1, 2, 3*
  - Übergangs- und Emissionswahrscheinlichkeiten:

	$p(...\|C)$	$p(...\|H)$	$p(...\|\text{START})$
$p(1\|...)$	.7	.1	
$p(2\|...)$	.2	.2	
$p(3\|...)$	.1	.7	
$p(C\|...)$	.8	.1	.5
$p(H\|...)$	.1	.8	.5
$p(F\|...)$	.1	.1	.0

- Allgemeine Definition: Ein HMM ist ein 6-tupel, bestehend aus:
  - endliche Menge von Zuständen $S = \{s_1, ..., s_N\}$
  - Endzustand $F \in S$
  - endliches Output-Alphabet ("Terminalsymbole") $K = \{k_1, ..., k_M\} = \{1, ..., M\}$
  - Anfangszustand-Wahrscheinlichkeiten: $\Pi = \{\pi_i\}, i \in S$
  - Übergangswahrscheinlichkeiten $A = \{a_{ij}\}, i, j \in S$
  - Emissionswahrscheinlichkeiten $B = \{b_{ik}\}, i \in S, k \in K$
- Drei fundamentale Fragen für HMMs:

1. Wie wahrscheinlich ist eine Sequenz von Output-Symbolen für ein gegebenes HMM?
2. Wenn das HMM und ein Output-Sequenz bekannt ist, welche korrespondierende Zustands-Sequenz ist am wahrscheinlichsten?
3. Welches HMM erklärt eine beobachtete Sequenz von Output-Symbolen am besten?

Wahrscheinlichkeit einer Beobachtungs-Sequenz (Forward-Algorithmus, Backward-Algorithmus

- Gegeben:
  - Beobachtungs-Sequenz $O = (o_1, \ldots, o_T)$
  - HMM $\mu = (A, B, \Pi)$
- Gesucht:
  - $P(O|\mu)$
- Es gilt für jede Zustand-Sequenz $X = (X_1, \ldots, X_T)$ (mit $X_T = F$)

$$\begin{aligned} P(O|X,\mu) &= \Pi_{t=1}^{T} P(o_t|X_t,\mu) \\ &= b_{X_1 o_1} b_{X_2 o_2} \cdots b_{X_T o_T} \end{aligned}$$

- Weiterhin gilt:

$$P(X|\mu) = \pi_{X_1} a_{X_1 X_2} a_{X_2 X_3} \cdots a_{X_{T-1} X_T}$$

- Deshalb

$$\begin{aligned} P(O,X|\mu) &= P(O|X,\mu) P(X|\mu) \\ &= \pi_{X_1} b_{X_1 o_1} \Pi_{t=1}^{T-1} a_{X_t X_{t+1}} b_{X_t o_t} \\ P(O|\mu) &= \sum_{X} \pi_{X_1} b_{X_1 o_1} \Pi_{t=1}^{T-1} a_{X_t X_{t+1}} b_{X_t o_t} \end{aligned}$$

- Berechnungskomplexität steigt exponentiell mit Länge der Beobachtung
- dynamische Berechnung:
  - Vorwärts-Variable (per Konvention, $s_N = F$)

$$\alpha_i(t) = P(o_1 o_2 \cdots o_{t-1}, X_t = i|\mu)$$

1. Initialisierung:
$$\alpha_i(1) = \pi_i$$
2. Induktion:
$$\alpha_j(t+1) = \sum_{i=1}^{N-1} \alpha_i(t) a_{ij} b_{io_t}$$
3. Abschluss:
$$P(O|\mu) = \alpha_F(T) b_{Fo_T}$$

– Berechnungskomplexität ist linear in Länge der Beobachtung

– Analog: Rückwärtsvariable
$$\beta_i(t) = P(o_t \cdots o_T | X_t = i, \mu)$$
1. Initialisierung:
$$\beta_i(T) = b_{Fo_T}$$
2. Induktion:
$$\beta_i(t) = \sum_{j=1}^{N} a_{ij} b_{io_t} \beta_j(t+1)$$
3. Abschluss:
$$P(O|\mu) = \sum_{i=1}^{N-1} \pi_i \beta_i(1)$$

- Kombination Vorwärts-Rückwärts:
$$P(O, X_t = i) = \alpha_i(t)\beta_i(t)$$

Wahrscheinlichste Zustands-Sequenz (Viterbi-Algorithmus)

siehe M&S 331-333, Charniak 53-56, ...

Überwachtes Lernen

- Gegeben:
  – Beobachtungs-Sequenz $O = (o_1 \cdots o_T)$
  – zugehörige Zustands-Sequenz $X = (X_1 \cdots X_T F)$
- Gesucht:
$$\arg_{\mu} \max P(O, X|\mu)$$

- Spezialfall von *Maximum Likelihood Estimation* (wenn 8 mal Kopf und 2 mal Zahl gefallen ist, beträgt die MLE für die Wahrscheinlichkeit von Kopf 0.8)

$$\begin{aligned}
\pi_{o_1} &= 1.0 \\
\pi_i &= 0 \text{ wenn } i \neq o_1 \\
a_{ij} &= |\{n|X_n = i \ \& \ X_{n+1} = j\}| \\
b_{ik} &= |\{n|X_n = i \ \& \ o_n = k\}|
\end{aligned}$$

Unüberwachtes Lernen (Forward-Backward-Algorithmus)

- Gegeben: Beobachtungs-Sequenz $O = (o_1 \cdots o_T)$
- Gesucht:

$$\arg_{\ \mu} \max P(O|\mu)$$

- Strategie:
  - Wähle ein HMM als Ausgangshypothese μ_0
  - Berechne mit Hilfe der Vorwärts- und Rückwärts-Variablen die Wahrscheinlichkeitsverteilung der verborgenen Zustände (gegeben μ_0)
  - Benutze diese rekonstruierten Werte als Basis für Parameterabschätzung im Sinne des überwachten Lernens
  - Ersetze μ_0 durch das so gewonnene Modell μ_1
  - Iteriere diesen Vorgang, bis sich $P(O|\mu_n)$ nicht mehr ändert
- Formal:

$$\begin{aligned}
\gamma_i(t) &= P(X_t = i|O, \mu) \\
&= \frac{P(X_t = i, O|\mu)}{P(O|\mu)} \\
&= \frac{\alpha_i(t)\beta_i(t)}{P(O|\mu)} \\
p_t(i,j) &= P(X_t = i, X_{t+1} = j|O, \mu) \\
&= \frac{P(X_t = 1, X_{t+1} = j, O|\mu)}{P(O|\mu)} \\
&= \frac{\alpha_i(t)a_{ij}b_{io_t}\beta_j(t+1)}{P(O|\mu)} \\
\widehat{\pi}_i &= \gamma_i(1) \\
\widehat{a}_{ij} &= \frac{\text{erwartete Anzahl von Transitionen von } i \text{ nach } j}{\text{erwartete Anzahl von Transitionen von } i}
\end{aligned}$$

$$= \frac{\sum_{t=1}^{T-1} p_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)}$$

$$\widehat{b}_{ik} = \frac{\text{erwartete Anzahl von Emissionen des Symbols } k \text{ im Zustand } i}{\text{erwartete Anzahl der in } i \text{ beginnenden Transitionen}}$$

$$= \frac{\sum_{\{t:o_t=k\}} \gamma_i(t)}{\sum_{t=1}^{T} \gamma_i(t)}$$

- Wir legen fest:

$$\widehat{\mu} = (\widehat{A}, \widehat{B}, \widehat{\Pi})$$

- Theorem:

$$P(O|\widehat{\mu}) \geq P(O|\mu)$$

Der Forward-Backward-Algorithmus ist eine Form des **Erwartungs-Maximierung (EM)**

Hausaufgabe

Manning & Schütze, Aufgabe 9.5 Seite 336. Beachte, dass die dort verwendete Definition von HMMs leicht von der hier vorgestellten abweicht!