



Genomanalyse

11. Übungsblatt WS 2009

Abgabe: Dienstag 19.01.2010 vor der Übung, Besprechung: in der Übung

Bitte geben Sie auf Ihren abgegebenen Lösungen Name und Matrikelnummer an und arbeiten Sie in Gruppen von 2 oder 3 Personen zusammen.

Aufgabe 23: [(5+2+1) + (5+5+1+3+1) + 0 = 23 Punkte] Hidden Markov Models

Consider the HMM with transition probabilities A and emission probabilities e :

$$A :=$$

from \ to	0	P	Q
0	0	0.7	0.3
P	0.8	0	0.2
Q	0.6	0.2	0.2

$$e :=$$

	x	y
P	0.2	0.8
Q	0.4	0.6

For the sequence $s = s_1s_2 = xy$:

- a)
 - i) Apply the Viterbi algorithm.
 - ii) What is the Viterbi path?
 - iii) What is the Viterbi probability?
 - b)
 - i) Apply the forward algorithm.
 - ii) Apply the backward algorithm.
 - iii) What is the full probability that xy is emitted by the HMM?
 - iv) Compute the posterior decoding path.
 - v) What is the probability of the posterior decoding path?
 - c) Explain (in words) why the two paths differ.
-

Aufgabe 24: [5 Punkte] Training Hidden Markov Models

You are given the sequences $xyzyzyxyzx$, $xyyz$, $xzzyzyzxxz$ and their respective underlying paths $PRQPQPRRRP$, $PQQQ$, $PQQRPRRQP$. Using these as training data, determine the parameters for an HMM using maximum likelihood.

zu Aufgabe 23:

Viterbi algorithm

	\emptyset	x	y	\emptyset
0	1	0	0	$\max\{A(P, 0) \cdot V(P, s_2), A(Q, 0) \cdot V(Q, s_2)\} = 0.01536$
P	0	$e_P(x) \cdot \frac{A(0, P) \cdot V(0, 0)}{= 0.2 \cdot 0.7 \cdot 1 = 0.14}$	$e_P(y) \cdot \max\{A(P, P) \cdot V(P, s_1), A(Q, P) \cdot V(Q, s_1)\} = 0.0192$	0
Q	0	$e_Q(x) \cdot \frac{A(0, Q) \cdot V(0, 0)}{= 0.4 \cdot 0.3 \cdot 1 = 0.12}$	$e_Q(y) \cdot \max\{A(P, Q) \cdot V(P, s_1), A(Q, Q) \cdot V(Q, s_1)\} = 0.0168$	0

Der wahrscheinlichste Pfad, um die Sequenz s zu erzeugen, ist $0 \rightarrow Q \rightarrow P \rightarrow 0$, mit einer Wahrscheinlichkeit von $\approx 1.5\%$

Forward algorithm

	\emptyset	x	y	\emptyset
0	1	0	0	$A(P, 0) \cdot F(P, s_2) + A(Q, 0) \cdot F(Q, s_2) = 0.03408$
P	0	$e_P(x) \cdot A(0, P) = 0.14$	$e_P(y) \cdot A(Q, P) \cdot F(Q, s_1) = 0.0192$	0
Q	0	$e_Q(x) \cdot A(0, Q) = 0.12$	$e_Q(y) \cdot (A(P, Q) \cdot F(P, s_1) + A(Q, Q) \cdot F(Q, s_1)) = 0.0312$	0

Backward algorithm

	\emptyset	x	y	\emptyset	
0		$e_P(x) \cdot A(0, P) \cdot B(P, s_1) + e_Q(x) \cdot A(0, Q) \cdot B(Q, s_1) = 0.03408$	0	$A(0, 0) = 0$	1
P	0		$e_P(y) \cdot A(P, P) \cdot B(P, s_2) + e_Q(y) \cdot A(P, Q) \cdot B(Q, s_2) = 0.072$	$A(P, 0) = 0.8$	0
Q	0		$e_P(y) \cdot A(Q, P) \cdot B(P, s_2) + e_Q(y) \cdot A(Q, Q) \cdot B(Q, s_2) = 0.2$	$A(Q, 0) = 0.6$	0

Die totale Wahrscheinlichkeit, die Sequenz xy zu generieren, ist $\approx 3.4\%$.

Posterior-Decoding

	\emptyset	x	y	\emptyset
0	1	0	0	1
P	0	$F(P, s_1) \cdot B(P, s_1) / Pr(s) = 0.295774647887$	$F(P, s_2) \cdot B(P, s_2) / Pr(s) = 0.450704225352$	0
Q	0	$F(Q, s_1) \cdot B(Q, s_1) / Pr(s) = 0.704225352113$	$F(Q, s_2) \cdot B(Q, s_2) / Pr(s) = 0.549295774648$	0

Posterior Decoding ergibt den Pfad $0 \rightarrow Q \rightarrow Q \rightarrow 0$. Die Wahrscheinlichkeit, dass dieser Pfad durchlaufen wird und dabei xy emittiert wird, ist $0.3 \cdot 0.4 \cdot 0.2 \cdot 0.6 \cdot 0.6 = 0.00864 \approx 0.8\%$, also deutlich kleiner als die Viterbi-Wahrscheinlichkeit.

Für den Pfad $0 \rightarrow P \rightarrow Q \rightarrow 0$ erhält man übrigens die Wahrscheinlichkeit $0.01008 \approx 1\%$.

Der Viterbi-Pfad berechnet nur den wahrscheinlichsten Pfad. Es kann aber sein, dass die Summe der weniger wahrscheinlichen Pfade für ein bestimmtes s_i , die Aufenthaltswahrscheinlichkeit für einen anderen Zustand größer machen, als es dem Viterbi-Pfad entspräche. Dies ist hier bei $s_2 = y$ der Fall.