Sprachtechnologie

4. Übung

Prof. Tobias Scheffer
Paul Prasse
Michael Großhans
Sebastian Arzt
Philipp Schmidt

Sommer 2014

Ausgabe am: 12.05.14
Besprechung am: 19.05.14

Aufgabe 1 *Markov-Ketten*

Auf einer Ufologen-Konferenz erklärt Nina Hagen die Regeln des Hochmarsianisch (die offizielle Sprache der MarsianerInnen): Zuerst einmal gibt es nur drei Laute: *argh*, *bob* und *zonk*. Desweiteren gilt:

- Auf 80% der *argh*-Laute folgt wieder ein *argh*, während in 10% darauf *zonk* gesagt wird.
- Ein *zonk*-Laut wird zu 70% wiederholt, während in 10% ein *bob*-Laut folgt.
- 60% der *bob*-Laute werden von einem weiteren *bob* gefolgt, in 20% der Fälle folgt ein *argh*.
- Frau Hagen informiert weiter, dass ein Marsianer jeden Tag zu 40% Wahrscheinlichkeit mit einem *argh* beginnt und mit je 30% entweder *bob* oder *zonk* zuerst sagt.

1. Stellen Sie die Matrix mit den Übergangswahrscheinlichkeiten $A = (a_{ij})_{i,j=1,\ldots N}$ auf.
2. Können Sie Frau Hagen verraten wie die Verteilung der zweiten und dritten Laute am Tag aussieht? (Könnte Ihnen eine Matrixmultiplikation hier weiterhelfen?)
3. Welche Bedingung muss gelten, damit sich die Verteilung der Laute über die Zeit nicht mehr verändert? (Optional: Welche Gleichung müssten Sie lösen um den stationären Zustand auszurechnen?)

Aufgabe 2 *HMM*

Auf derselben Konferenz verliest Erich von Däniken einen selbstgeschriebenen Aufsatz, der die Kommunikationsstrukturen der Marsianer zum Thema hat. Zusammenfassend beinhaltet er folgende Punkte:

- Es kommunizieren immer drei Marsianer miteinander.
- Nachdem ein Laut gesagt wurde sagt einer der beiden anderen den nächsten Laut mit einer Wahrscheinlichkeit von je 40%, in 20% der Fälle fügt der aktuelle Redner noch einen weiteren Laut hinzu.
- Wer das Gespräch beginnt ist gleichwahrscheinlich.

Sie erinnern sich an Herrn von Däniken als Sie an einem Vorhang vorbeigehen, hinter dem die drei einzigen Marsianer auf der Konferenz gerade ein Gespräch beginnen. Zufällig wissen Sie, dass die drei aus unterschiedlichen Regionen kommen und verschiedene Dialekte sprechen. Der erste benutzt nur die Laute *zonk* und *bob* im Verhältnis $\frac{2}{5}$ zu $\frac{3}{5}$, der zweite nur *argh* und *zonk* im Verhältnis $\frac{1}{11}$ zu $\frac{10}{11}$ und der dritte eine Kombination aller drei Laute *bob*, *argh* und *zonk* in den Verhältnissen $\frac{4}{9} : \frac{2}{9} : \frac{3}{9}$. Sie hören die ersten drei Laute des Gespräches: *zonk*, *bob* und *argh*. Können Sie Herrn von Däniken sagen, welcher Marsianer welchen Laut gesagt hat?

1. Modellieren Sie die Aufgabenstellung mit einem HMM. Identifizieren Sie dazu die folgenden Parameter: Welche Zustände gibt es? Was sind die möglichen Beobachtungen? Welche Start-, Übergangs- und Beobachtungswahrscheinlichkeiten nehmen Sie an?

2. Zeichnen Sie Ihr Hidden Markov Model.

3. Bestimmen Sie die Likelihood des Modells.

4. Welcher Zustand ist der wahrscheinlichste (gegeben die Sequenz) für den zweiten Laut (also zum Zeitpunkt $t = 2$)?

Aufgabe 3 *Viterbi-Algorithmus*

Angenommen Sie hätten die folgenden Wahrscheinlichkeiten aus einem Korpus geschätzt:

t\ t+1	JJ	VBG	VBP	PRP	VBD	π	schlau	werden	wir
JJ	0.3	0.0	0.1	0.3	0.3	0.3	0.2	0.0	0.0
VBG	0.2	0.2	0.2	0.2	0.2	0.2	0.0	0.3	0.0
VBP	0.2	0.3	0.1	0.3	0.1	0.1	0.0	0.5	0.0
PRB	0.0	0.4	0.3	0.1	0.2	0.3	0.0	0.0	0.2
VBD	0.2	0.1	0.2	0.2	0.3	0.1	0.4	0.0	0.0

Zeichnen Sie das entstandene HMM und bestimmen Sie mit Hilfe des Viterbi-Algorithmuses die wahrscheinlichste Zustandsfolge zu der Beobachtung „*Wir werden schlau*“. Visualisieren Sie anschliessend Ihre Berechnung und Lösung im Trellis. Geben Sie eine probabilistische reguläre Grammatik an, welche die gleiche Wahrscheinlichkeitsverteilung modelliert.

Aufgabe 4 *Trigramm-Tagger*

Bei der Wortarterkennung mit HMMs stellt jeder Zustand ein Pos-Tag dar. Um größere Kontextabhängigkeiten zu erkennen, müssen die Zustände mehrere Pos-Tags kodieren. Wie groß ist der Zustandsraum für einen Trigramm-Tagger bei m möglichen Pos-Tags?