Sprachtechnologie

3. Übung

Prof. Tobias Scheffer
Uwe Dick

Sommer 2015

Ausgabe am: 04.05.15
Besprechung am: 11.05.15

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 $A = (a_{ij})_{i,j=1,\dots N}$ 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 Marsianer 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 *T9*

Eine naive Implementierung der Eingabehilfe T9 (siehe Übung 2), bei der alle möglichen Buchstabenfolgen $w_1, \dots w_T$ durchsucht werden, hat eine exponentielle Laufzeit in T (Siehe Folie 45 der Sprachmodelle-Vorlesung). In der Vorlesung wurde erklärt, dass man durch eine geeignete Dekodierung eine Laufzeit von $\mathcal{O}(TV^N)$ erreichen kann, wobei V die Größe des Alphabets ist und N der Parameter des N-Gram-Modells. Geben Sie einen konkreten Algorithmus mit dieser Laufzeit an.

Zusatzaufgabe *ML-Schätzer*

Zeigen Sie, dass für den ML-Schätzer des n-Gramm-Modells

$$\theta^{ML}_{v_{i_n}|\dots v_{i_1}} = \frac{x_{v_{i_n}\dots v_{i_1}}}{x_{v_{i_{n-1}}\dots v_{i_1}}}$$

gilt.

Hinweis: Leiten Sie die log-Likelihood von $\theta^{ML}_{v_{i_n}|\dots v_{i_1}} = \arg\max_{\theta_{v_{i_n}|\dots v_{i_1}}} \prod_{g=v_1\dots v_1}^{v_K\dots v_K} \left(\prod_{w=v_1}^{v_K} \theta_{w|g}^{x_{wg}}\right)$ ab und betrachten Sie gegebenfalls nur den Fall $K = 2$.