Maschinelles Übersetzen natürlicher Sprachen

Aufgabe 1

Wir betrachten folgendes Experiment: Ein Freund wirft insgesamt hundert mal ein Paar – möglicherweise gezinkter – Würfel. Statt der geworfenen Augenzahlen erfahren Sie jedoch nur die Summe der beiden Würfe. Es soll der EM-Algorithmus verwendet werden, um aus diesen unvollständigen Daten die wahrscheinlichsten Wurfwahrscheinlichkeiten für die beiden Würfel zu schätzen.

a) Bestimmen Sie für dieses Problem
- die Menge der unvollständigen Daten X,
- die Menge der vollständigen Daten Y sowie den Analysator $H\colon X \to \mathcal{P}(Y)$,
- und das Wahrscheinlichkeitsmodell $\mathcal{M} \subseteq \mathcal{M}(Y)$ auf den vollständigen Daten.

b) Werden die Anzahlen der geworfenen Augensummen in einem Korpus $h\colon X \to \mathbb{N}$ zusammengefasst, ergeben sich die folgenden Werte.

x	2	3	4	5	6	7	8	9	10	11	12
$h(x)$	2	3	5	9	10	12	16	20	13	7	3

Nutzen Sie, ausgehend von einer geeigneten Startverteilung, den EM-Algorithmus, um für diesen Korpus h die wahrscheinlichsten Parameter zu schätzen (Bestimmung des maximum likelihood estimators).

Aufgabe 2

Seien X_1, X_2 endliche Mengen. Gegeben $p_1\colon X_1 \to [0,1]$ und $p_2\colon X_2 \to [0,1]$, sei die Wahrscheinlichkeitsverteilung $p_1 \times p_2\colon X_1 \times X_2 \to [0,1]$ definiert durch

$$(p_1 \times p_2)(x_1, x_2) = p_1(x_1) \cdot p_2(x_2)$$

für alle $x_1 \in X_1$ und $x_2 \in X_2$.

Wir betrachten das Wahrscheinlichkeitsmodell $\mathcal{M} = \{p_1 \times p_2 \mid p_1 \in \mathcal{M}(X_1), p_2 \in \mathcal{M}(X_2)\}$. Es sei ein beliebiger Korpus $h\colon X_1 \times X_2 \to \mathbb{R}_{\geq 0}$ gegeben. Zeigen Sie, dass

$$\mathrm{mle}(h, \mathcal{M}) = \mathrm{rfe}(h_1) \cdot \mathrm{rfe}(h_2)$$

gilt, wobei die marginalisierten Korpora $h_1\colon X_1 \to \mathbb{R}_{\geq 0}$ und $h_2\colon X_2 \to \mathbb{R}_{\geq 0}$ für alle $x_1 \in X_1$, $x_2 \in X_2$ definiert sind durch

$$h_1(x_1) = \sum_{y \in X_2} h(x_1, y) \qquad \text{und} \qquad h_2(x_2) = \sum_{y \in X_1} h(y, x_2).$$