# Sprachtechnologie

2. Übung

Prof. Tobias Scheffer
Uwe Dick

Sommer 2016

Ausgabe am: 25.04.16
Besprechung am: 02.05.16

**Aufgabe 1** *Binomialverteilung*

Bei der Produktion von Halbleiterelementen sind durchschnittlich 20% defekt. Wie groß sind die Wahrscheinlichkeiten, dass unter 11 Stück

- genau 6 Stück
- höchstens 3 Stück defekt sind?

**Aufgabe 2** *N-Gramme*

Angenommen, Sie wollen ein n-Gramm-Modell der deutschen Sprache erstellen. Dafür steht Ihnen folgendes Trainingskorpus zur Verfügung: *„Ich habe Hunger und Durst. Aber ich habe kein Essen und kein Getränk dabei. Ausserdem habe ich kein Geld. Ich muss zuerst Geld besorgen.“*. Außerdem ist ein Testkorpus gegeben: *„Ich habe kein Geld.“*.

- Bestimmen Sie die Vokabulargröße $k$ des Trainingskorpus (Hint: Zählen Sie die Wörter).
- Bestimmen Sie den ML- und MAP-Schätzer für die bedingten Wahrscheinlichkeiten $P(x_i|x_j), i, j = 1, ..., k$ aller Paare von Wörtern aus dem Trainingskorpus. Die Priorparameter der Dirichletverteilung sind gegeben als $\alpha_{x_i} = 2, i = 1, ..., k$. (Hint: Entsprechend der Definition aus der Vorlesung verwenden wir also ein 2-gramm-Model. Verwenden Sie die Likelihood-Funktion aus der Vorlesung. )
- Geben Sie mit Hilfe der jeweils geschätzen Parameter die Gesamtwahrscheinlichkeit für das Testkorpus an.

**Aufgabe 3** *Autovervollständigung*

Autovervollständigungssysteme wie sie beispielsweise bei Texteditoren Anwendung finden, sollen dem Benutzer Wörter vorschlagen bevor sie vollständig eingeben worden sind. Angenommen Sie haben die Aufgabe, eine derartig intelligente Eingabehilfe zu entwickeln. Diskutieren Sie Lösungsansätze für diese Aufgabenstellung.

**Zusatzaufgabe** *Binomialverteilung*

Beim Münzwurf-Experiment stellt sich die Frage, wie oft in einer Reihe von Münzwürfen Kopf geworfen wurde. Die Anzahl $N_K$ von beobachteteten Kopfwürfen ist dabei eine binomialverteilte Zufallsvariable (analog ist $N_Z$ die Anzahl von beobachteten Zahlwürfen). Wie in der Vorlesung gezeigt, besitzt diese Binomialverteilung einen Parameter $\theta$, der aus Daten gelernt werden kann. Bestimmen sie die ML-Schätzung für die Binomialverteilung, also genau den Parameter $\theta$, der die Likelihood der Binomialverteilung, gegeben durch:

$$\mathcal{L}(\theta) = \binom{N_K + N_Z}{N_K} \theta^{N_K} (1-\theta)^{N_Z}$$

maximiert.
*Hinweis:* Leiten Sie die Funktion $\ln \mathcal{L}(\theta)$ ab.