

# Sprachtechnologie

## 7. Übung

Prof. Tobias Scheffer  
Christoph Sawade

Sommer 2009

Ausgabe am: 15.06.09  
Besprechung am: 22.06.09

### Aufgabe 1

*k-Means*

Nina Hagen möchte ihre alten Marsianertexte nach ihrer Ähnlichkeit kategorisieren. Einer spontanen Eingebung folgend sollen dabei nur die beiden Laute *argh* und *zonk* berücksichtigt werden, die in 79 bzw. 90 von 100 Texten vorkommen. Es werden 4 repräsentative Texte ausgegeben, anhand derer zwei Kategorien ermittelt werden sollen. Die Termfrequenzen in den 4 Texten sind:

ID	<i>argh</i>	<i>zonk</i>
1	25	10
2	2	13
3	6	14
4	17	21

1. Geben Sie die normierte TF-IDF-Repräsentationen der Texte an.
2. Wenden Sie den *k*-means Algorithmus mit  $k = 2$  an. Welche Initialisierung benutzen Sie? Welche Cluster entstehen? Veranschaulichen Sie Ihre Lösung in einem Koordinatensystem.
3. Geben Sie eine neue Initialisierung an, mit der der *k*-means Algorithmus andere Cluster erzeugt.
4. Bestimmen Sie das Optimum der beiden erhaltenen Lösungen.

### Aufgabe 2

*Global optimales Clustering*

In der Vorlesung wurde erklärt, dass der EM-Algorithmus nur ein lokales Optimum liefert. Wie könnte ein naiver Algorithmus aussehen, der das Minimum der Funktion

$$J = \sum_{n=1}^N \sum_{k=1}^K r_{nk} \|\mathbf{x}_n - \mu_k\|^2$$

findet? Analysieren Sie den Aufwand des Verfahrens. Zeigen Sie, dass es exponentiell viele Aufteilungen von  $n$  Beispielen in  $k$  (auch leere) Cluster gibt.

**Aufgabe 3***Mixture of Gaussians*

Der EM-Algorithmus besteht aus zwei sich wiederholenden Schritten: Die Berechnung der Clusterzugehörigkeiten für die Beispiele (*expectation*) und die darauf folgende Maximierung der Modell-Parametern  $\Theta = (\pi, \mu, \Sigma)$  (*maximization*).  $\pi_k$  ist dabei der geschätzte Anteil von Punkten im Cluster  $k$ . Zeigen Sie für den binären Fall ( $k = 2$ ), dass

$$\pi_k = \frac{N_k}{N}$$

die Zielfunktion  $\mathcal{Q}(\Theta, \Theta_t)$  maximiert.

**Aufgabe 4***EM-Algorithmus*

Der EM-Algorithmus maximiert die Likelihood-Funktion  $\mathcal{Q}(\Theta, \Theta_t)$ . Angenommen Sie möchten eine Hypothese lernen, die die A-Posterior Verteilung  $P(\Theta|\mathbf{X})$  maximiert. Wie muss die modifizierte Zielfunktion  $\mathcal{Q}'(\Theta, \Theta_t)$  aussehen?