Einführung in die Computerlinguistik
Hausaufgabe 5 (Wahrscheinlichkeitsrechnung und HMM-POS-Tagging), Abgabe 02.06.2014

Laura Kallmeyer

Sommersemester 2014, Heinrich-Heine-Universität Düsseldorf

Aufgabe 1 *Betrachten Sie das folgende Experiment: Sie haben ein geschlossenes Gefäß, in dem sich* 1 *schwarze,* 3 *weiße und* 4 *rote Kugeln befinden. Sie ziehen eine Kugel aus dem Gefäß. Die Ergebnismenge* Ω *der möglichen Farben ist* $\{s, w, r\}$ *(für schwarz, weiß, rot).*

Nehemen Sie an, dass für jede der 8 *Kugeln in dem Gefäß die Wahrscheinlichkeit, dass diese Kugel gezogen wird, gleich groß ist.*

1. *Definieren Sie ein Wahrscheinlichkeitsmaß auf* $\mathcal{P}(\Omega)$*, das diese Annahme widerspiegelt.*
2. *Gehen Sie jetzt davon aus, dass zweimal hintereinander* **ohne Zurücklegen** *gezogen wird. (Beim zweiten Ziehen sind also nur noch 7 Kugeln vorhanden.) Unser neues* Ω *ist also die Menge aller zweielementigen Folgen, die sich hier ergeben kann:* $\{rr, rw, rs, ww, wr, ws, sw, sr\}$*.*

   Geben Sie das Wahrscheinlichkeitsmaß P *für* Ω *an (es reicht, nur die Wahrscheinlichkeiten für die Elemente aus* Ω *zu definieren), das den Gegebenheiten entspricht.*
3. *Berechnen Sie jetzt die Wahrscheinlichkeit dafür, dass die ohne Zurücklegen gezogene zweielementige Folge zwei Kugeln verschiedener Farbe enthält.*
4. *Betrachten sie jetzt das Ereignis* A*, dass die erste Kugel in der zweielementigen Folge weiß ist. Wie sehen* A *und* $P(A)$ *aus?*

   Definieren Sie das bedingte Wahrscheinlichkeitsmaß $P(\cdot|A)$ *auf* $\mathcal{P}(\Omega)$*, indem Sie die Werte von* $P(\cdot|A)$ *für alle Elemente aus* $\{rr, rw, rs, ww, wr, ws, sw, sr\}$ *angeben.*

Lösung:

1. Wahrscheinlichkeitsmaß P:

Menge A	$P(A)$	Menge A	$P(A)$
$\emptyset$	0	$\{r, s\}$	$\frac{5}{8}$
$\{r\}$	$\frac{1}{2}$	$\{r, w\}$	$\frac{7}{8}$
$\{w\}$	$\frac{3}{8}$	$\{w, s\}$	$\frac{1}{2}$
$\{s\}$	$\frac{1}{8}$	$\{w, s, r\}$	1

2. Wahrscheinlichkeitsmaß P:

Menge A	$P(A)$	Menge A	$P(A)$
$\{rr\}$	$\frac{1}{2} \times \frac{3}{7}$	$\{ws\}$	$\frac{3}{8} \times \frac{1}{7}$
$\{rw\}$	$\frac{1}{2} \times \frac{3}{7}$	$\{wr\}$	$\frac{3}{8} \times \frac{4}{7}$
$\{rs\}$	$\frac{1}{2} \times \frac{1}{7}$	$\{sw\}$	$\frac{1}{8} \times \frac{3}{7}$
$\{ww\}$	$\frac{3}{8} \times \frac{2}{7}$	$\{sr\}$	$\frac{1}{8} \times \frac{4}{7}$

3. $P(\{rr, ww\}) = P(\{rr\}) + P(\{ww\}) = \frac{12+6}{56} = \frac{18}{56} = \frac{9}{28}$

   Wir suchen die Wahrscheinlichkeit der Komplementmenge, es ergibt sich also $1 - \frac{9}{28} = \frac{19}{28}$.
4. $A = \{wr, ww, ws\}$, $P(A) = \frac{3}{8} \times \frac{4+2+1}{7} = \frac{3}{8}$

   $P(\{rr\}|A) = P(\{rw\}|A) = P(\{rs\}|A) = P(\{sw\}|A) = P(\{sr\}|A) = 0$

$P(\{wr\}|A) = \frac{P(\{wr\})}{P(A)} = \frac{4}{7}$
$P(\{ww\}|A) = \frac{P(\{ww\})}{P(A)} = \frac{2}{7}$
$P(\{ws\}|A) = \frac{P(\{ws\})}{P(A)} = \frac{1}{7}$

Aufgabe 2 *Angenommen, wir untersuchen Wortanfänge im Englischen. Wir haben anhand von Textuntersuchungen festgestellt, dass* 70% *der Vorkommen von Artikeln (*the, a, *...) in den untersuchten Texten mit* th- *beginnen. Außerdem haben unsere Untersuchungen ergeben, dass im Durchschnitt jedes 14. Wort ein Artikel ist und dass* $\frac{1}{16}$ *der Wörter in den untersuchten Texten mit* th- *beginnen.*

Wie hoch ist die Wahrscheinlichkeit, dass ein Wort, das mit th- *beginnt, ein Artikel ist?*

Lösung:

A ist das Ergebnis, dass es sich um einen Artikel handelt, B das Ergebnis, dass ein Wort mit *th-* beginnt. Wir wissen $P(A) = \frac{1}{14}$ und $P(B) = \frac{1}{16}$. Und $P(B|A) = 0.7$. Damit lässt sich mit der Formel von Bayes berechnen

$$P(\text{Artikel}|\textit{th-}) = P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{0.7 \times \frac{1}{14}}{\frac{1}{16}} = \frac{7 \times 16}{10 \times 14} = 0.8$$

Aufgabe 3 *Nehmen Sie an, Sie haben einen HMM-POS Tagger, mit dem folgende Eingabe getaggt werden soll:* learning changes thoroughly. *Dem Tagger liegen folgende Wahrscheinlichkeiten zugrunde:*

Emissionswahrscheinlichkeiten:

$P(learning|V) = 3 \cdot 10^{-3}$ $P(changes|V) = 4 \cdot 10^{-3}$ $P(thoroughly|Adv) = 2 \cdot 10^{-3}$
$P(learning|N) = 1 \cdot 10^{-3}$ $P(changes|N) = 3 \cdot 10^{-3}$

Alle anderen Emissionswahrscheinlichkeiten für unsere Eingabewörter seien 0.

Relevante Übergangswahrscheinlichkeiten:

$P(N|N) = 1 \cdot 10^{-1}$ $P(N|V) = 4 \cdot 10^{-1}$ $P(Adv|V) = 4 \cdot 10^{-1}$
$P(Adv|N) = 1 \cdot 10^{-1}$ $P(V|N) = 3 \cdot 10^{-1}$ $P(V|V) = 1 \cdot 10^{-1}$

Angenommen, die Wahrscheinlichkeit, dass ein N am Satzanfang steht ist $1 \cdot 10^{-1}$*, die, dass ein V am Satzanfang steht* $2 \cdot 10^{-1}$*. Die, dass ein Satzende auf Adv folgt, ist* $1 \cdot 10^{-1}$*.*

1. *Geben Sie die Viterbi Matrix an, die sich bei diesen Wahrscheinlichkeiten für die Eingabe* learning changes thoroughly *ergibt. Es reicht, die Einträge anzugeben, die* $\neq 0$ *sind.*
2. *Welche POS Tags ermittelt der Tagger für die Eingabe?*

Lösung:

1.

q_F				$192 \cdot 10^{-13}$, Adv
Adv			$192 \cdot 10^{-12}$, V	
V	$6 \cdot 10^{-4}$, q_0	$24 \cdot 10^{-8}$, V		
N	$1 \cdot 10^{-4}$, q_0	$72 \cdot 10^{-8}$, V		
	1 learning	2 changes	3 thoroughly	

2. Die beste POS-TAG Folge ist V V Adv.