Universität Trier
LDV / Computerlinguistik
PS Quantitative Linguistik, Sommersemester 2005
Leitung: Prof. Dr. Reinhard Köhler
Referent: Kai Kugler

24.5.2005

Markov-Ketten

Markov-Kette:

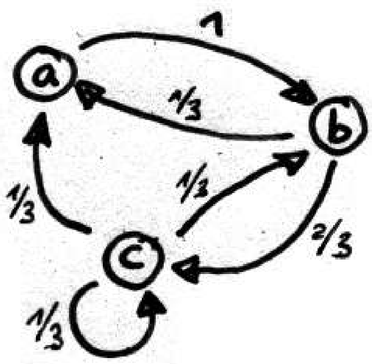
- nach Andrej Andreevič Markov, russischer Mathematiker, 1856-1922
- Eine <u>Markov-Kette</u> (auch „Markov-Prozeß“) ist ein stochastischer Prozeß, also eine Familie von Zufallsvariablen (=Zufallsgrößen) $\{X_0, X_1, ..., X_t, ...\}$ (X_t ist hierbei die Zufallsvariable zum „Zeitpunkt“ t), der die <u>Markov-Eigenschaft</u> erfüllt.
- Bei linguistischen Untersuchungen ist sehr oft die lineare Verkettbarkeit von Einheiten -oder Klassen von Einheiten- von Interesse, Markov-Ketten können hierzu als Modell dienen.
  - Beispiele: Alliterationen in Texten, Text 'als Kette von Wortarten', Folgen von Phonemen (anhand bestimmter Merkmale),...

Simulation einer Markov-Kette:

<u>wir brauchen dazu:</u>

1. Eine Menge von Werten S, die die Zufallsvariablen X_i annehmen können (=<u>Zustände</u>), z.B. $S = \{1,...,N\}$
   Für unsere Simulation beschränken wir uns auf drei Zustände: $S=\{a,b,c\}$
2. Eine Anfangsverteilung, also die Wahrscheinlichkeitsverteilung für die Markov-Kette zum Zeitpunkt t=0.
   Für unsere Simulation reicht es, einen Anfangszustand zu wählen, wir starten bei b (also $X_0=b$), die Anfangsverteilung ist die zweite Zeile der Matrix.
3. Übergangswahrscheinlichkeiten, um von einem Zustand zum nächsten zu gelangen. Ist $S=\{1,..,N\}$, so bilden die Übergangswahrscheinlichkeiten eine NxN-Matrix. Die Einträge der Matrix p_{ij} beschreiben die Wahscheinlichkeit, daß die Kette zum Zeitpunkt t auf dem Zustand i steht und bei t+1 zu Zustand j wechselt:
   also $p_{ij} = P\{X_{t+1} = j \mid X_t = i\}$
   (die Einträge dieser Matrix sind alle nichtnegativ, die Zeilensummen jeweils 1)
   Für die Simulation gehen wir von folgenden Übergangswahrscheinlichkeiten aus:

$$\begin{array}{cc} & \begin{array}{ccc} a & b & c \end{array} \\ \begin{array}{c} a \\ b \\ c \end{array} & \begin{pmatrix} 0 & 1 & 0 \\ \frac{1}{3} & 0 & \frac{2}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix} \end{array} = P$$



4. Um von einem Zustand zum nächsten zu kommen brauchen wir je eine Zufallszahl U zwischen 0 und 1.

jetzt geht es los:

- Wir starten mit X_0=b. Sei U_1=0,84... eine Zufallszahl. Für X_1 gilt:
  X_1 = a für $0<U_1<0{,}3$
  X_1 = c für $0{,}3<U_1<1$ (vergleiche die zweite Zeile der Matrix)
  Wir landen also für X_1 auf Zustand c.

- Sei U_2=0,123... eine Zufallszahl. Für X_2 gilt:
  X_2 = a für $0<U_2<0{,}3$
  X_2 = b für $0{,}3<U_2<0{,}6$
  X_2 = c für $0{,}6<U_2<1$ (vergleiche die dritte Zeile der Matrix)
  Wir landen also für X_2 auf Zustand a.

- Von Zustand a kann man nur zu Zustand b übergehen, demnach ist X_3=b.

- ...nun könnten wir immer so weitermachen, tun wir aber nicht.
- Unsere Kette bisher: X_0=b, X_1=c, X_2=a, X_3=b

- Die Übergangswahrscheinlichkeit für X_4 ist somit

$$P\{X_4=j \mid X_0=b, X_1=c, X_2=a, X_3=b\} = \begin{matrix} 0.3 & \text{für } j=a \\ 0 & \text{für } j=b \\ 0.6 & \text{für } j=c \end{matrix}$$

  man sieht, daß die Werte der Variablen X_0, X_1 und X_2 hier irrelevant sind, es gilt also:
  $P\{X_4=j \mid X_0=b, X_1=c, X_2=a, X_3=b\} = P\{X_4=j \mid X_3=b\}$
  Diese Eigenschaft ist die sog. Markov-Eigenschaft oder auch „Gedächtnislosigkeit".

Markov-Eigenschaft:

- Die Wahrscheinlichkeitfür einen Zustand zum Zeitpunkt t+1 hängt nur von den n vorhergehenden Zuständen ab. (in unserem Beispiel war n=1)

- für n=1: $P\{X_{t+1}=i_{t+1} \mid X_t=i_t, X_{t-1}=i_{t-1}, \ldots, X_0=i_0\} = P\{X_{t+1}=i_{t+1} \mid X_t=i_t\}$
- für n>1: $P\{X_{t+1}=i_{t+1} \mid X_t=i_t, X_{t-1}=i_{t-1}, \ldots, X_0=i_0\} =$
  $P\{X_{t+1}=i_{t+1} \mid X_t=i_t, \ldots, X_{t-(n+1)}=i_{t-(n+1)}\}$

- Man spricht von Markov-Ketten der Ordnung n.
- In der Linguistik sind sehr oft Abhängigkeiten wichtig, die über den direkten Vorgänger hinausgehen => Modell: Markov-Kette n-ter Ordnung.
- Beispiel: Untersucht wird eine Phonemkette auf das Merkmal vokalisch (/i:, i, e:, a:,.../) oder konsonantisch (/p, t, k, b, d, g, .../) (hier: v oder k) mit einer Markov-Kette zweiter Ordnung (n=2). Zustandskombinationen: vv, vk, kv, kk können übergehen zu vvk, vvv, vkk, vkv, kvk, kvv, kkk, kkv. Hier haben wir eigentlich dreidimensionale Ergeignisse, die wir aber zweidimensional wie folgt abbilden:

	kk	kv	vk	vv
kk	0,22	0,78	0	0
kv	0	0	0,56	0,44
vk	0,32	0,68	0	0
vv	0	0	0,6	0,4

z.B. ist die Wahrscheinlichkeit, daß nach vv ein k folgt (also das Auftauchen der Kombination vk nach vv) 0,6 (vgl. Köhler, S. 141).

Pfade in Markov-Ketten:

- für jeden Pfad, den die Markov-Kette beschreiten kann, folgt:
  $P\{X_0=i_0, X_1=i_1, X_2=i_2, \ldots, X_t=i_t\} =$

  $P\{X_0=i_0\}P\{X_1=i_1 \mid X_0=i_0\}P\{X_2=i_2 \mid X_1=i_1, X_0=i_0\}\ldots P\{X_t=i_t \mid X_{t-1}=i_{t-1},\ldots,X_0=i_0\}$
  ...ist einfach...
  $= P\{X_0=i_0\}P\{X_1=i_1 \mid X_0=i_0\}P\{X_2=i_2 \mid X_1=i_1\}\ldots P\{X_t=i_t \mid X_{t-1}=i_{t-1}\}$

- Beispiel: Folge von betonten (Zustand 1) und unbetonten (Zustand 0) Silben eines geg. Textes ($S=\{0,1\}$). Die Wahrscheinlichkeiten für den übernächsten Zustand bei Übergangswahrscheinlichkeiten P ist P^2.

Häufigkeitstabellen und Übergangswahrscheinlichkeiten:

- Man erhält eine Matrix mit Übergangswahrscheinlichkeiten aus einer Häufigkeitstabelle, indem ,man jeden Eintrag der Tabelle durch die jeweilige Zeilensumme teilt.

	a	b	c	Σ
a	1	0	1	2
b	1	1	1	3
c	0	1	1	2

	a	b	c
a	0,5	0	0,5
b	$0,\overline{3}$	$0,\overline{3}$	$0,\overline{3}$
c	0	0,5	0,5

- Meist hat man jedoch Übergangswahrscheinlichkeiten, die z.B. aus theoretischen Überlegungen stammen und möchte diese These anhand von gezählten Merkmalshäufigkeiten überprüfen („sind die Daten eine Realisation einer Markov-Kette mit geg. Matrix?"). Hierzu eignen sich Verfahren wie z.B. der Likelihood-Ratio-Test.

Literatur:

Köhler, R.: Markov-Ketten und Autokorrelation in der Sprach- und Textanalyse. In: Glottometrika 5. Hg. von R. Köhler und J. Boy. (=Quantitative Linguistics 20) Bochum 1983.