

Veranstaltung: Wissensrepräsentation (»Grundlagen der KI«), SS08
Dozent: Nino Simunic M.A., Computerlinguistik, Campus DU

Übung (4), 09.06.2008 - 14.06.2008

Thema: Probabilistisches Schließen / Statistische Verfahren

Präsenzaufgaben (a)-(d),

Szenario

Es ist wieder soweit, und Prof. Humbug prüft wieder im Fach Mumpitzologie. Wie immer prüft er eine Person pro Tag – diesmal an drei aufeinanderfolgenden Tagen. Harry Krause, Student, wird am letzten Tag geprüft, und er will irgendwie sicherstellen, dass es auch ein Tag ist, an dem der Professor höchstwahrscheinlich bestehen lässt. (Gegebenenfalls will Harry sich dann lieber krank melden, falls dem nicht so ist.)

Harry erinnert sich daran, dass viele Studenten behaupteten: Aussichten, bei Prof. Humbug eine Prüfung zu bestehen, hängen von seinem Gewissen (»Ist schon oft passiert, dass, wenn er gestern hat durchfallen lassen, er heute viele Leute bestehen lässt«, ...) und seiner Laune ab (»Wenn er gut drauf ist, lässt er die meisten Leute an dem Tag bestehen.«, ...). Harry erinnert das an Wettervorhersage-Verfahren: Die Wahrscheinlichkeit, wie das Wetter wird, hängt nur vom Vortag ab. Zudem scheint es, dass auch Prof. Humbugs Laune an einem Tag eine Rolle bei der Bewertung spielt. Harry mutmaßt, dass die Kombination der beiden Informationen Laune und Gewissen eine gute Vorhersagegrundlage sein könnte, und will es nun genau wissen. Er erstellt noch zwei Wochen vor seiner Prüfung Statistiken von 100 Studierenden und Absolventen an Prof. Humbugs Lehrstuhl, um daraus zu lernen:

- **Gewissensstatistik** (Beeinflussung durch den Vortag):
 - An einem Prüfungstag ließ er 40-mal bestehen und 60-mal durchfallen, wenn am Vortag bestanden worden waren.
 - An einem Prüfungstag ließ er 90-mal bestehen und 10-mal durchfallen, wenn am Vortag durchgefallen worden war.
- **Laune-Statistik:**
 - Er war 70-mal gut und 30-mal schlecht drauf an einem Prüfungstag, an dem er hat bestehen lassen.
 - Er war 25-mal gut und 75-mal schlecht drauf an einem Prüfungstag, an dem er hat durchfallen lassen.
- Interessant vielleicht auch die **»Start«-Statistik**: Wie wahrscheinlich ist ein Nicht-/Bestehen am Start (d.h. am ersten Tag) der Prüfungssequenz?
 - 20-mal hat er die meisten Studenten am ersten der Prüfungstage bestehen und 80-mal durchfallen lassen.

Wenn Harry wüsste, wie die zwei Prüfungen vor seiner sind oder sein werden, könnte er noch rechtzeitig eine Entscheidung fällen. Dazu benötigt er sogar nur einen Teil der Statistiken:

	Bestanden	Durchgefallen
Bestanden	0.4	0.6
Durchgefallen	0.9	0.1

	Bestanden	Durchgefallen
Start ¹	0.2	0.8

¹ In der Vorlesung fehlte die Startwahrscheinlichkeit. Die Annahme war, dass sie für alle Wetterzustände gleich ist (also 1, und somit dort redundant).

Harry überlegt: Wüsste er nach dem zweiten Prüfungstag z.B., dass die beiden vorherigen Prüfungstage »Durchfallen-Tage« waren, könnte er die Wahrscheinlichkeiten dafür ermitteln, ob der dritte Tag ein Durchfallen- oder eher ein »Bestehen-Tag« ist. Dazu muss er -- analog zu dem ihm bekannten Wettervorhersage-Verfahren -- auf Basis einer **Markov-Kette (erster Ordnung)** die Wahrscheinlichkeiten für beide Möglichkeiten (ermitteln und) vergleichen.

(a) Wie sieht der endliche Automat / die Markov-Kette zu den Matrizen oberhalb aus?

(b) Ist Bestanden oder Durchgefallen der wahrscheinlichere Zustand, gegeben der beobachteten Zustandsfolge D D?

$$P(D | D D) = P(D|S) * P(D|D) * P(D|D) = 0.8 * 0.1 * 0.1 = 0,008$$

$$P(B | D D) = P(D|S) * P(D|D) * P(B|D) = 0.8 * 0.1 * 0.9 = 0,072$$

Problem: Erstens: Harry kann mit dieser Vorhersage nicht viel anfangen kann, da die Prüfungsergebnisse erst eine Woche nach den Prüfungen bekannt gegeben werden, und Kontakt zu anderen Prüflingen bis dahin unter Androhung von Exmatrikulation untersagt ist. Er kann also nicht beobachten, ob es sich bei den Prüfungstagen vorher um »Bestehen-« oder »Durchfallen-Tage« handelte. Zweitens: Harry gehen wertvolle Informationen bei dieser Berechnung verloren, denn er hat ja die starke Vermutung, dass die Kombination von Gewissen und Laune Prof. Humbug's Verhalten kombiniert beeinflusst.

Was er aber beobachten darf und kann, ist die Laune von Prof. Humbug an den drei Prüfungstagen, da Harry ihn täglich sieht.² Ideal wäre also, die über einen Zeitraum (die Prüfungstage) beobachtbaren Launen des Professors zusammen mit den vorhandenen Statistiken zur Vorhersage des Wissenszustands von Prof. Humbug an Harrys Prüfungstag zu nehmen. Anscheinend muss also versucht werden, auf Basis einer Beobachtungssequenz (Laune) die dazugehörige, versteckte Zustandssequenz (Gewissen) zu ermitteln.

Hidden Markov Modelle (Erweiterung des (sichtbaren) Markov Modells)

Harry fällt dazu eine Vorlesung ein, in der ein ähnliches Problem vorlag: Das automatische Auszeichnen von Wörtern mit Wortarten (Automatisches POS-Tagging) mittels HMM (*Hidden Markov Model*). Beim HMM-Tagging-Prozess sind dabei nur die Wörter sichtbar, versteckt sind die Wortarten, die es automatisch für die vorliegenden Wörter bzw. eine Sequenz dessen zu ermitteln gilt. Im Vorfeld des eigentlichen Taggings werden (ähnlich wie bei Harry, Wettervorhersage) Wahrscheinlichkeiten antrainiert. Das Lernmaterial ist dabei von Menschenhand erstellt worden und enthält neben Text auch die richtigen, intellektuell ermittelten Wortarten. Das Erlernte im Überblick:

- (1) Zustandsübergangswahrscheinlichkeiten bzw. Zustandsemissionswahrscheinlichkeiten für (Sequenzen von) Wortarten.
 - Z.B. $P(DET|N)$, $P(V|N)$
 - Bei Harry z.B. die Übergänge zwischen Wissenszuständen: $P(Bestehen|Durchgefallen)$
- (2) Symbolemissionswahrscheinlichkeiten³ für die Wörter und mit welcher Wortart sie auftreten.
 - Wörter können verschiedene Wortarten haben, Z.B. $P(N|book)$, $P(V | book)$
 - Bei Harry entspricht das den Laune-Statistiken, z.B. $P(Bestehen|Gute Laune)$
- (3) Startzustandswahrscheinlichkeiten: Wie häufig fängt ein Satz mit einer Wortart an.
 - Z.B. $P(DET|START)$
 - Bei Harry z.B. $P(Bestehen|START)$

Die spätere Beobachtung (der beobachtbare Teil des Modells) entspricht dann den Wörtern in noch unbearbeiteten Texten, deren wahrscheinlichste Wortartenfolge zugewiesen werden sollen. Die Wortarten entsprechen dem versteckten Teil des Hidden Markov Modells, da sie ja beim Tagging-Prozess eben nicht sichtbar sondern versteckt sind. Ziel ist es auch, das versteckte zu »erraten«.

Harrys HMM

Harry sieht, dass er ein HMM sehr gut für seine Zwecke einsetzen kann. Ihm liegen insgesamt die folgenden (»erlernten«) Informationen vor:

² Harrys Prüfung ist erst um 16.00 Uhr – er kann also noch vorher sehen oder sich durchgeben lassen, wie der Humbug an dem Tag drauf ist.

³ In der Vorlesung wurde ein visible (*dt.* sichtbares) Markov-Modell bzw. eine solche Kette gezeigt

(c) Wie sieht der endliche (doppelt-stochastische) Automat zu den Matrizen aus?

Gewissen

	B	D
B	0.4	0.6
D	0.9	0.1

Start

	B	D
S	0.2	0.8

Laune

	☺	☹
B	0.7	0.3
D	0.25	0.75

B: Bestehen-Zustand; D: Durchfallen-Zustand; S: Start(-zustand);
☺: Gute Laune; ☹: Schlechte Laune

Berechnung der Wahrscheinlichkeit einer Beobachtungssequenz/-folge:

Die Wahrscheinlichkeit einer Beobachtungsfolge O gegeben einer Zustandsfolge T berechnet sich nun aus dem Gesamtprodukt der Einzelprodukte:

$$P(O | T) = P(T_1 * O_1 * T_2 * O_2 * T_3 * O_3 * T_4 * O_4 * \dots * T_n * O_n)$$

Mit O_1 ist die Wahrscheinlichkeit gemeint, dass O_1 mit T_1 auftritt, also bspw. dass ☺ an einem Bestehen-Tag auftritt. Mit T_2 z.B. ist die Wahrscheinlichkeit von T_2 gegeben T_1 gemeint. T_1 am Anfang der Multiplikation oberhalb ist in der Regel die Startwahrscheinlichkeit des dieses Zustands.

Beispiel im Harry-Kontext. Beobachtung ist ☺ ☹. Wie ist die Wahrscheinlichkeit, dass die Sequenz B D für diese Beobachtung »verantwortlich« ist?

$$\begin{aligned} P(\text{☺ ☹} | B D) &= P(B | S) * P(\text{☺} | B) * P(D | B) * P(\text{☹} | D) \\ &= 0.2 * 0.7 * 0.6 * 0.75 \\ &= \underline{0,063} \end{aligned}$$

(Harrys) Vorhersage der wahrscheinlichsten Zustandsfolge im HMM

Es ist soweit: Harrys Prüfungstag. Harry beobachtete bis dahin, dass Prof. Humbug an den zwei Prüfungstagen vorher jeweils schlechte Laune (☹ ☹) hatte, er heute aber vor Freude nur so strahlt (☺). Mit dieser Folge von Symbolemissionen (☹ ☹ ☺) wird nun die wahrscheinlichste Sequenz von Gewissen-Zuständen gesucht: $P(? ? ? | \text{☹ ☹ ☺})$. Im Idealfall ist eine Zustandssequenz am wahrscheinlichsten, deren letzter Zustand »Bestehen« ist. Dann würde sich Harry sicherlich zur Prüfung trauen. Zu berechnen sind also die Wahrscheinlichkeiten für alle Zustandsfolgen, die die Beobachtung ☹ ☹ ☺ erzeugen, damit verglichen und die wahrscheinlichste Zustandsfolge ermittelt werden kann.

(d) Welche ist die wahrscheinlichste (hier versteckte) Sequenz (Gewissen-Zustände)?

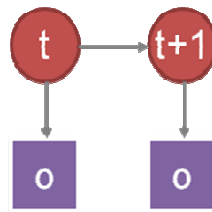
O1	O2	O3	T1	T2	T3	T1	O1	T2	O2	T3	O3		
P(☹☹☺ D B B)	=	0.8	*	0.75	*	0.9	*	0.3	*	0.4	*	0.7	<u>=0.045360000000000004</u>
P(☹☹☺ D D B)	=	0.8	*	0.75	*	0.1	*	0.75	*	0.9	*	0.7	=0.028350000000000007
P(☹☹☺ D B D)	=	0.8	*	0.75	*	0.9	*	0.3	*	0.6	*	0.25	=0.024300000000000006
P(☹☹☺ D D D)	=	0.8	*	0.75	*	0.1	*	0.75	*	0.1	*	0.25	=0.001125000000000003
P(☹☹☺ B B B)	=	0.2	*	0.3	*	0.4	*	0.3	*	0.4	*	0.7	=0.002016
P(☹☹☺ B D B)	=	0.2	*	0.3	*	0.6	*	0.75	*	0.9	*	0.7	=0.017009999999999997
P(☹☹☺ B B D)	=	0.2	*	0.3	*	0.4	*	0.3	*	0.6	*	0.25	=0.00108
P(☹☹☺ B D D)	=	0.2	*	0.3	*	0.6	*	0.75	*	0.1	*	0.25	=0.0006749999999999999

Nach den vorliegenden Berechnungen stehen die Chancen gut, dass am dritten Tag ein »Bestehen-Tag« ist.

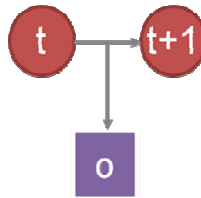
ANHANG 1: FORMALES ZU HMM

Ein HMM ist ein doppelt-stochastische-r/s Prozess/Modell. Im vorliegenden Fall handelt es sich um ein HMM erster Ordnung. Allgemein: Zahl drückt Anzahl der Zustände aus, von denen der aktuelle Zustand abhängt.

- **State Emission Model:** zur Zeit t emittiertes Symbol **hängt ab vom Zustand zur Zeit t**



- **Arc Emission Model:** Zur Zeit t emittiertes Symbol **hängt ab vom Zustand zur Zeit t und vom Zustand zur Zeit $t+1$**



Ein HMM ist formal durch ein Fünf-Tupel spezifiziert:

$$(S, K, \Pi, A, B)$$

$\Leftrightarrow$

(Menge der Zustände, Ausgabealphabet, Wahrscheinlichkeiten der Startzustände,
Wahrscheinlichkeiten der Zustandsübergänge, Wahrscheinlichkeiten der Symbolemissionen)

$$S = \{s_1, \dots, s_N\}$$

$$K = \{k_1, \dots, k_M\} = \{1, \dots, M\}$$

$$\Pi = \{\pi_i\}, i \in S$$

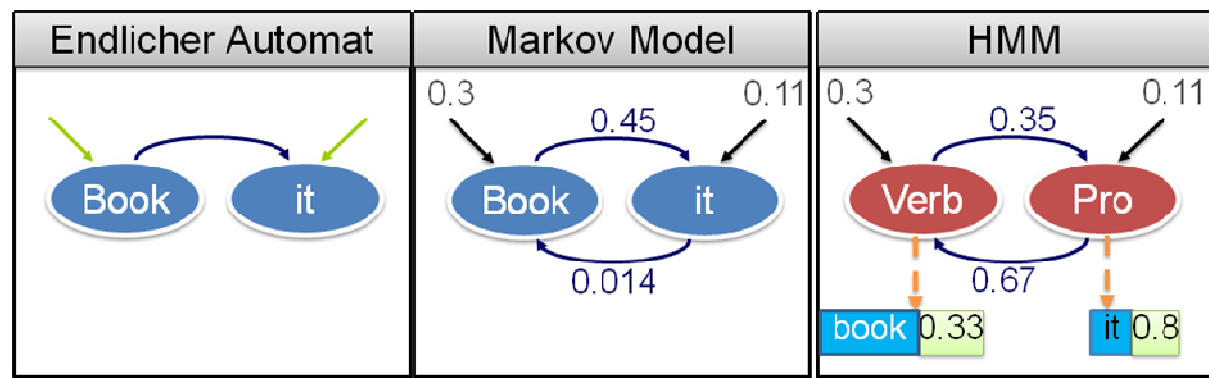
$$A = \{a_{ij}\}, \quad i, j \in S \quad \sum_{j=1}^N a_{ij} = 1$$

$$B = \{b_{ijk}\}, \quad i, j \in S, \quad k \in K \quad \sum_{k=1}^M b_{ijk} = 1$$

HMM sind statistische Verfahren (der KI) und sind in den verschiedensten Bereichen weit verbreitet: Computerlinguistik (POS-Tagging, Erkennung gesprochener Sprache), Medizin (Symptom, Ursache), Glücksspiel, Finanzen, Wetter, etc.

⁴ Stochastische Bedingung: Summe aller Wahrscheinlichkeiten (Zeile in Matrix, ausgehende Kanten eines Zustands in Endl. Automaten) muss =1 ergeben.

ANHANG 2: ENDLICHER AUTOMAT, MARKOV MODEL, HMM



ANHANG 3: STATISTIKEN EINES HMM FÜRS POS-TAGGING

Aus [MAN_SCHÜT:1999, 342 ff.]:

First tag	Second tag					
	AT	BEZ	IN	NN	VB	PERIOD
AT	0	0	0	48636	0	19
BEZ	1973	0	426	187	0	38
IN	43322	0	1325	17314	0	185
NN	1067	3720	42470	11773	614	21392
VB	6072	42	4758	1476	129	1522
PERIOD	8016	75	4656	1329	954	0

	AT	BEZ	IN	NN	VB	PERIOD
<i>bear</i>	0	0	0	10	43	0
<i>is</i>	0	10065	0	0	0	0
<i>move</i>	0	0	0	36	133	0
<i>on</i>	0	0	5484	0	0	0
<i>president</i>	0	0	0	382	0	0
<i>progress</i>	0	0	0	108	4	0
<i>the</i>	69016	0	0	0	0	0
	0	0	0	0	0	48809

REFERENZEN

[MAN_SCHÜT:1999] Manning, Christopher D.; Schütze, Hinrich. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, Mass., London: The MIT Press