Übungen zur Vorlesung „Hidden Markov Models“

Blatt 2

Abgabetermin: Dienstag, 31.10.2023, bis 14:15 Uhr in der Vorlesung.
(Bitte nur maximal zu zweit abgeben.)

Aufgabe 1 (4 Punkte)

Sei $(\varepsilon_k)_{k\geq 1}$ eine reellwertige iid-Folge. Für $p \in \mathbb{N}$ seien $a_1, \ldots, a_p \in \mathbb{R}$, $a_p \neq 0$. Einen Prozess $(X_k)_{k\geq 0}$ mit $X_0 = \cdots = X_{p-1} = 0$, welcher der Rekursion

$$X_n = \sum_{k=1}^{p} a_k X_{n-k} + \varepsilon_n \quad (n \geq p)$$

genügt, bezeichnen wir als AR(p)-Prozess mit Startwert $X_0 = \cdots = X_{p-1} = 0$. Beweisen Sie:

a) Für $p > 1$ besitzt ein AR(p)-Prozess nicht die Markov-Eigenschaft.

b) Durch geeignete Vegrößerung des Zustandsraumes kann man den AR(p)-Prozess dennoch als Markovprozess darstellen.

Aufgabe 2 (4 Punkte)

Seien $(X_k)_{k\geq 0}$ und $(Y_k)_{k\geq 0}$ zwei unabhängige AR(1)-Prozesse mit Startwert $X_0 = Y_0 = 0$. Beweisen Sie, dass $(X_k, Y_k)_{k\geq 0}$ zwar eine bivariate Markovkette aber kein Hidden-Markov-Modell ist.

Aufgabe 3 (4 Punkte)

DNA-Sequenzen kodieren genetische Information in vier Buchstaben, A, C, G, T. Jedoch kann die Art und Weise, in der ein und dasselbe Merkmal kodiert ist, bei verschiedenen Individuen deutlich variieren. Beispielsweise kodieren die folgenden fünf Abfolgen dasselbe Merkmal: $ACAATG$, $AGAATC$, $ACACAGC$, $ACCGATC$ $TCAATGATC$. Um das ihnen gemeinsame Muster hervorzubringen, sind sie in folgender Tabelle angeordnet:

l1	l2	l3	li	li	li	l4	l5	l6
A	C	A	-	-	-	A	T	G
A	G	A	-	-	-	A	T	C
A	C	A	C	-	-	A	G	C
A	C	C	G	-	-	A	T	C
T	C	A	A	T	G	A	T	C.

Offenbar variiert das Basismuster $ACAATC$ auf zwei Weisen: Die individuelle Abfolge $l1 - l6$ kann eine Mutation enthalten und zusätzliche beliebige Symbole li können in die Mitte der Abfolge eingefügt sein.

> Modellieren Sie diese Muster als Hidden-Markov-Modell mit den Zustandsräumen $Y = \{A, C, G, T\}$ und $X = \{l1, \ldots, l6, li, le\}$, wobei $Q(l6, \{le\}) = Q(le, \{le\}) = 1$ gelte (le ist ein sogenannter terminaler Zustand). Sei weiter $G(le, \{y\}) = 1/4$ für alle $y \in Y$. Bestimmen Sie den Übergangskern Q des Signalprozesses und den Beobachtungskern G.