Technische Universität Berlin
Fakultät IV – Elektrotechnik und Informatik

Künstliche Intelligenz: Grundlagen und Anwendungen

Albayrak, Fricke (AOT) – Opper, Thiel (KI)

Wintersemester 2016 / 2017

6. Aufgabenblatt

Abgabetermin: 18.01.2017 Musterlösung

Aufgabe 1 – Hidden Markov-Prozess (50%)

Hidden-Markov-Modelle werden in der Bioinformatik zur Analyse von DNA-Sequenzen eingesetzt. Eine Anwendung ist das Auffinden von *CpG-Inseln* in der beobachteten Sequenz $Y_t \in \{a, c, g, t\}$. Der verborgene Zustand $X_t \in \{w, f\}$ gibt an, ob das aktuelle Nukleotid zu einer CpG-Insel gehört ($X_t = w$) oder nicht ($X_t = f$). Für die Wahrscheinlichkeiten der einzelnen Nukleotide und die Übergangswahrscheinlichkeiten der Zustände gilt in einem stark vereinfachten Modell:

x	w	f
$P(X_{t+1} = x \mid X_t = w)$	0.7	0.3
$P(X_{t+1} = x \mid X_t = f)$	0.2	0.8

y	a	c	g	t
$P(Y_t = y \mid X_t = w)$	0.2	0.3	0.3	0.2
$P(Y_t = y \mid X_t = f)$	0.3	0.2	0.2	0.3

Als Anfangsbedingung wird $P(X_0 = w) = 0.5$ angenommen.

(a) Wie wahrscheinlich ist es, eine CpG-Insel der Länge k zu finden? Geben Sie $P(X_1 = \ldots = X_k = w, X_{k+1} = f \mid X_0 = f)$ für $k \geq 1$ an!

$$\begin{aligned} & P(X_1 = \ldots = X_k = w, X_{k+1} = f \mid X_0 = f) \\ = \; & P(X_1 = w \mid X_0 = f) P(X_{k+1} = f \mid X_k = w) \prod_{t=1}^{k-1} P(X_{t+1} = w \mid X_t = w) \\ = \; & 0.2 \cdot 0.3 \cdot 0.7^{k-1} = 0.06 \cdot 0.7^{k-1} \end{aligned}$$

(b) Sie wollen effizient CpG-Inseln in einer DNA-Sequenz finden und berechnen hierzu die Wahrscheinlichkeit $p_t = P(X_t = w \mid Y_1, \ldots, Y_t)$ aus p_{t-1} und Y_t. Wie sieht ein solcher Filter-Schritt für die Beobachtung $Y_t = g$ aus?

- Wegen der Markov-Annahme hängt p_t nur von der aktuellen Beobachtung Y_t und von $p_{t-1} = P(X_{t-1} = w \mid Y_1, \ldots, Y_{t-1})$ ab. Die früheren Beobachtungen $Y_1, \ldots, Y_{t-1}$ werden nicht explizit benötigt.

- Berechnung der Filter-Iteration für $Y_t = g$

$$\begin{aligned} p_t &= \alpha_t \, P(Y_t = g | X_t = w) \\ &\cdot \quad [P(X_t = w | X_{t-1} = w) \, p_{t-1} \\ &\quad + P(X_t = w | X_{t-1} = f)(1 - p_{t-1})] \\ &= 0.3 \, \alpha_t \, [0.7 \, p_{t-1} + 0.2(1 - p_{t-1})] = \alpha_t \, [0.06 + 0.15 \, p_{t-1}] \\ 1 - p_t &= \alpha \, P(Y_t = g | X_t = f) \\ &\cdot \quad [P(X_t = f | X_{t-1} = w) \, p_{t-1} \\ &\quad + P(X_t = f | X_{t-1} = f)(1 - p_{t-1})] \\ &= 0.2 \, \alpha_t \, [0.3 \, p_{t-1} + 0.8(1 - p_{t-1})] = \alpha_t \, [0.16 - 0.10 \, p_{t-1}] \end{aligned}$$

- Normierung

$$1 = \alpha_t \, [0.22 + 0.05 \, p_{t-1}] \Longleftrightarrow \alpha_t = \frac{1}{0.22 + 0.05 \, p_{t-1}}$$

- Filter-Iteration für $Y_t = c$ und $Y_t = g$

$$p_t = \frac{0.06 + 0.15 p_{t-1}}{0.22 + 0.05 p_{t-1}}$$

(c) Wie hoch ist die Wahrscheinlichkeit $P(X_t | Y_1 = c, Y_2 = g)$ für eine CpG-Insel an den Positionen $t = 3$ und $t = 4$, wenn Sie nur die ersten zwei Nukleotide der DNA-Sequenz kennen?

- Vorhersage-Iteration ohne Beobachtung

$$\begin{aligned} p_t &= P(X_t = w | X_{t-1} = w) \, p_{t-1} + P(X_t = w | X_{t-1} = f)(1 - p_{t-1}) \\ &= 0.7 \, p_{t-1} + 0.2(1 - p_{t-1}) \; = \; 0.5 \, p_{t-1} + 0.2 \end{aligned}$$

- Berechnung der Wahrscheinlichkeiten

$$\begin{aligned} p_0 &= 0.500 \\ p_1 &= \frac{0.06 + 0.15 \cdot 0.500}{0.22 + 0.05 \cdot 0.500} = 0.551 \\ p_2 &= \frac{0.06 + 0.15 \cdot 0.551}{0.22 + 0.05 \cdot 0.551} \approx 0.576 \\ p_3 &\approx 0.5 \cdot 0.576 + 0.2 \approx 0.488 \\ p_4 &\approx 0.5 \cdot 0.488 + 0.2 \approx 0.444 \end{aligned}$$

- Marginale Zustandsverteilung

t	0	1	2	3	4
p_t	50.0%	55.1%	57.6%	48.8%	44.4%

(d) Wie hoch ist die Wahrscheinlichkeit $P(X_1 = w|Y_1 = c, Y_2 = g)$, dass bereits das erste Nukleotid der DNA-Sequenz $Y_1 = c, Y_2 = g, \ldots$ zu einer CpG-Insel gehört?

- Filter-Verteilung für $t = 1$:

$$P(X_1 = w|Y_1 = c) = 0.551$$

- Likelihood für $Y_2 = g$

$$\begin{aligned} P(Y_2 = g|X_1 = w) &= P(Y_2 = g|X_2 = w)P(X_2 = w|X_1 = w) \\ &+ P(Y_2 = g|X_2 = f)P(X_2 = f|X_1 = w) \end{aligned}$$

$$0.3 \cdot 0.7 + 0.2 \cdot 0.3 = 0.27$$

$$\begin{aligned} P(Y_2 = g|X_1 = f) &= P(Y_2 = g|X_2 = w)P(X_2 = w|X_1 = f) \\ &+ P(Y_2 = g|X_2 = f)P(X_2 = f|X_1 = f) \end{aligned}$$

$$0.3 \cdot 0.2 + 0.2 \cdot 0.8 = 0.22$$

- Marginale Posterior-Verteilung

$$\begin{aligned} & P(X_1 = w|Y_1 = c, Y_2 = g) \\ = & \frac{P(Y_2 = g|X_1 = w)P(X_1 = w|Y_1 = c)}{\sum_{x\in\{w,f\}} P(Y_2 = g|X_1 = x)P(X_1 = x|Y_1 = c)} \\ = & \frac{0.27 \cdot 0.551}{0.27 \cdot 0.551 + 0.22 \cdot (1 - 0.551)} \\ = & \frac{0.149}{0.149 + 0.099} \approx 0.601 \end{aligned}$$

(e) Verwenden Sie den Viterbi-Algorithmus, um die wahrscheinlichste Folge von X_t für die DNA-Sequenz $Y_1 = a, Y_2 = c, Y_3 = g, Y_4 = t$ zu finden!

- Vorwärtsiteration

$$\begin{aligned} m_t^+ &= P(Y_t|X_t = w) \\ &\cdot \max[P(X_t = w|X_{t-1} = w)m_{t-1}^+; P(X_t = w|X_{t-1} = f)m_{t-1}^-] \\ m_t^- &= P(Y_t|X_t = f) \\ &\cdot \max[P(X_t = f|X_{t-1} = w)m_{t-1}^+; P(X_t = f|X_{t-1} = f)m_{t-1}^-] \end{aligned}$$

- Wahrscheinlichkeit der wahrscheinlichsten Zustandsfolge

$$\begin{aligned} m_1^+ &= 0.2\max[0.7 \cdot 0.5000;\ 0.2 \cdot 0.5000] &&= 0.0700 && (X_0 = w) \\ m_1^- &= 0.3\max[0.3 \cdot 0.5000;\ 0.8 \cdot 0.5000] &&= 0.1200 && (X_0 = f) \\ m_2^+ &= 0.3\max[0.7 \cdot 0.0700;\ 0.2 \cdot 0.1200] &&= 0.0147 && (X_1 = w) \\ m_2^- &= 0.2\max[0.3 \cdot 0.0700;\ 0.8 \cdot 0.1200] &&= 0.0192 && (X_1 = f) \\ m_3^+ &= 0.3\max[0.7 \cdot 0.0147;\ 0.2 \cdot 0.0192] &&\approx 0.0031 && (X_2 = w) \\ m_3^- &= 0.2\max[0.3 \cdot 0.0147;\ 0.8 \cdot 0.0192] &&\approx 0.0031 && (X_2 = f) \\ m_4^+ &= 0.2\max[0.7 \cdot 0.0031;\ 0.2 \cdot 0.0031] &&\approx 0.0004 && (X_3 = w) \\ m_4^- &= 0.3\max[0.3 \cdot 0.0031;\ 0.8 \cdot 0.0031] &&\approx 0.0007 && (X_3 = f) \end{aligned}$$

Die Angabe (X_{t-1} =?) jeweils am Ende der Zeile gibt an, mit welchem Vorgängerzustand die Maximumsnachricht berechnet wurde.

- Die Zustandsfolge mit der maximalen Wahrscheinlichkeit $m_4^+ \approx 0.0007$ endet mit dem Zustand $X_4 = f$. Durch Rückwärtsiteration über die Vorgängerzustände kann nun die wahrscheinlichste Folge ermittelt werden. Hier wurde m_4^- aus m_3^-, m_3^- aus m_2^- und m_2^- aus m_1^- berechnet. Daraus folgt:

$$X_1 = f, X_2 = f, X_3 = f, X_4 = f.$$

Aufgabe 2 – Hidden-Markov-Modell **(50%)**

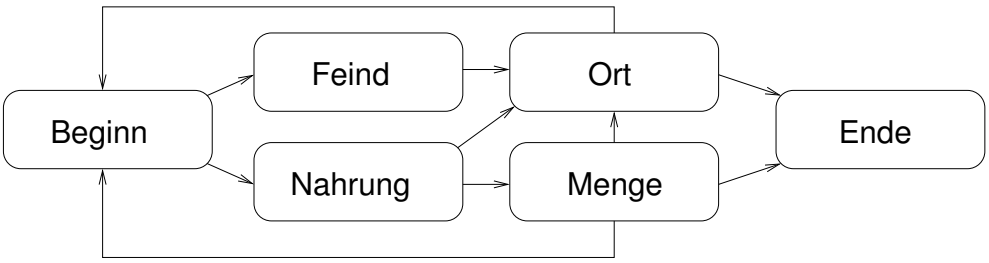
Eine neuentdeckte Chamäleonart nutzt ihre Hautfarbe, um komplexe Botschaften zu kommunizieren. Wir unterscheiden zwischen einer Folge von Segmenten $x_1, x_2, x_3, \ldots$ und der tatsächlich beobachteten Folge von Farben $y_1, y_2, y_3, \ldots$.

x_i	x_{i+1}	$P(x_{i+1}\|x_i)$
Beginn	Feind	0.4
Beginn	Nahrung	0.6
Feind	Ort	1.0
Nahrung	Ort	0.2
Nahrung	Menge	0.8
Ort	Beginn	0.3
Ort	Ende	0.7
Menge	Ort	0.3
Menge	Beginn	0.2
Menge	Ende	0.5

x_i	y_i	$P(y_i\|x_i)$
Beginn	weiß	1.0
Feind	rot	0.6
Feind	blau	0.4
Nahrung	rot	0.7
Nahrung	grün	0.3
Ort	blau	0.8
Ort	orange	0.2
Menge	blau	0.1
Menge	grün	0.9
Ende	schwarz	1.0

Die Markovkette beginnt immer mit $x_1 = \textit{Beginn}$ und endet mit $x_k = \textit{Ende}$. Alle nicht angegebenen Wahrscheinlichkeiten $P(x_{i+1}|x_i)$ und $P(y_i|x_i)$ sind Null. Sie können die Segmenttypen mit großen und die Farben mit kleinen Anfangsbuchstaben abkürzen, um Platz zu sparen.

(a) Stellen Sie das Modell für die Ausdrücke in einem Übergangsdiagramm graphisch dar! Sie brauchen keine Wahrscheinlichkeiten einzutragen.



(b) Mit welcher Wahrscheinlichkeit tritt die Farbfolge „weiß-rot-blau-orange-schwarz“ in diesem Modell auf?

Es gibt nur eine Segmentfolge die diese Farben erzeugen kann:

$$\begin{aligned} P(\text{„w-r-b-o-s“}) &= P(N|B)P(M|N)P(O|M)P(E|O) \\ &\cdot\ P(w|B)P(r|N)P(b|M)P(o|O)P(s|E) \\ &= 0.6 \cdot 0.8 \cdot 0.3 \cdot 0.7 \\ &\cdot\ 1.0 \cdot 0.7 \cdot 0.1 \cdot 0.2 \cdot 1.0 \\ &\approx 0.0014 \end{aligned}$$

(c) Sie beobachten die Folge „weiß-rot-orange-schwarz“. Ist es wahrscheinlicher, dass es um Nahrung oder Feinde geht? Wie sicher ist dies?

- Wenn es sich um einen Feind handelt:

$$\begin{aligned} P(x_2 = F|\text{„w-r-o-s“}) &\propto P(F|B)P(O|F)P(E|O) \\ &\cdot\ P(w|B)P(r|F)P(o|O)P(s|E) \\ &= 0.4 \cdot 1.0 \cdot 0.7 \\ &\cdot\ 1.0 \cdot 0.6 \cdot 0.2 \cdot 1.0 \\ &= 0.0336 \end{aligned}$$

- Wenn es sich um Nahrung handelt:

$$\begin{aligned} P(x_2 = N|\text{„w-r-o-s“}) &\propto P(N|B)P(O|N)P(E|O) \\ &\cdot\ P(w|B)P(r|N)P(o|O)P(s|E) \\ &= 0.6 \cdot 0.2 \cdot 0.7 \\ &\cdot\ 1.0 \cdot 0.7 \cdot 0.2 \cdot 1.0 \\ &= 0.01176 \end{aligned}$$

- Normierung:

$$\frac{0.0336}{0.0336 + 0.01176} \approx 0.74$$

- Die Nachricht handelt zu ungefähr 74% von einem Feind.

(d) Wie wahrscheinlich ist eine Nachricht aus 4 Segmenten?

- Es gibt 3 Möglichkeiten für eine Nachricht aus 4 Segmenten: Beginn-Feind-Ort-Ende, Beginn-Nahrung-Ort-Ende und Beginn-Nahrung-Menge-Ende.
- Beginn-Feind-Ort-Ende:

$$\begin{aligned} P(BFOE) &= P(F|B)P(O|F)P(E|O) \\ &= 0.4 \cdot 1.0 \cdot 0.7 \\ &= 0.28 \end{aligned}$$

- Beginn-Nahrung-Ort-Ende

$$\begin{aligned} P(BNOE) &= P(N|B)P(O|N)P(E|O) \\ &= 0.6 \cdot 0.2 \cdot 0.7 \\ &= 0.084 \end{aligned}$$

- Beginn-Nahrung-Menge-Ende

$$\begin{aligned} P(BNME) &= P(N|B)P(M|N)P(E|M) \\ &= 0.6 \cdot 0.8 \cdot 0.5 \\ &= 0.24 \end{aligned}$$

- Gesamtwahrscheinlichkeit:

$$\begin{aligned} P(BFOE) + P(BNOE) + P(BNME) &= 0.28 + 0.084 + 0.24 \\ &= 0.604 \end{aligned}$$

- Eine Nachricht mit vier Segmenten tritt mit der Wahrscheinlichkeit 60.4% auf.