Technische Universität Berlin
Fakultät IV – Elektrotechnik und Informatik

Künstliche Intelligenz: Grundlagen und Anwendungen

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

Wintersemester 2014 / 2015

7. Aufgabenblatt

Musterlösung

Aufgabe 1 – Modellauswahl (50%)

Durch den Vergleich zweier Markov-Modelle lässt sich in der Bioinformatik entscheiden, ob eine DNA-Sequenz zu einer *CpG-Insel* gehört ($X = w$) oder nicht ($X = f$). Für die Übergangswahrscheinlichkeiten $P(Y_{t+1}|Y_t, X = f)$ zwischen aufeinander folgenden Nukleotiden gilt:

	$X = f$				$X = w$			
	$Y_t = a$	$Y_t = c$	$Y_t = g$	$Y_t = t$	$Y_t = a$	$Y_t = c$	$Y_t = g$	$Y_t = t$
$Y_{t+1} = a$	0.30	0.32	0.25	0.18	0.18	0.17	0.16	0.08
$Y_{t+1} = c$	0.20	0.30	0.25	0.24	0.27	0.37	0.34	0.36
$Y_{t+1} = g$	0.29	0.08	0.29	0.29	0.43	0.27	0.37	0.38
$Y_{t+1} = t$	0.21	0.30	0.21	0.29	0.12	0.19	0.13	0.18

Der Anfang $Y_1 \in \{a, c, g, t\}$ einer DNA-Sequenz wird als gleichverteilt angenommen: $P(Y_1 = a) = P(Y_1 = c) = P(Y_1 = g) = P(Y_1 = t) = 0.25$.

(a) Berechnen Sie die Likelihood der DNA-Sequenz „TCGCGA“ für beide Modelle! Für welches Modell würden Sie sich ohne weitere Informationen gemäß der Maximum-Likelihood-Methode entscheiden?

- Likelihood für $X = f$

$$\begin{aligned} P(Y|X = f) &= P(Y_1 = t)P(Y_2 = c|Y_1 = t, X = f) \\ &\cdot\ P(Y_3 = g|Y_2 = c, X = f)P(Y_4 = c|Y_3 = g, X = f) \\ &\cdot\ P(Y_5 = g|Y_4 = c, X = f)P(Y_6 = a|Y_5 = g, X = f) \\ &= 0.25 \cdot 0.24 \cdot 0.08 \cdot 0.25 \cdot 0.08 \cdot 0.25 \\ &= 0.08^2 \cdot 0.24 \cdot 0.25^3 \\ &\approx 2.400 \cdot 10^{-5} \end{aligned}$$

- Likelihood für $X = w$

$$\begin{aligned} P(Y|X=w) &= P(Y_1=t)P(Y_2=c|Y_1=t,X=w) \\ &\cdot\ P(Y_3=g|Y_2=c,X=w)P(Y_4=c|Y_3=g,X=w) \\ &\cdot\ P(Y_5=g|Y_4=c,X=w)P(Y_6=a|Y_5=g,X=w) \\ &= 0.25\cdot 0.36\cdot 0.27\cdot 0.34\cdot 0.27\cdot 0.16 \\ &= 0.16\cdot 0.25\cdot 0.27^2\cdot 0.34\cdot 0.36 \\ &\approx 3.569\cdot 10^{-4} \end{aligned}$$

- Das Modell $X = w$ hat die größere Likelihood. Deshalb nimmt die Maximum-Likelihood-Methode an, dass die beobachtete DNA-Sequenz $Y =$ „TCGCGA“ zu einer *CpG-Insel* gehört.

(b) Wie hoch ist die Posterior-Wahrscheinlichkeit, dass diese DNA-Sequenz zu einer CpG-Insel gehört? Verwenden Sie $P(X = w) = 0.2$ als Prior-Wahrscheinlichkeit!

$$\begin{aligned} P(X=w|Y) &= \frac{P(Y|X=w)P(X=w)}{P(Y|X=w)P(X=w)+P(Y|X=f)P(X=f)} \\ &= \frac{3.569\cdot 10^{-4}\cdot 0.2}{3.569\cdot 10^{-4}\cdot 0.2+2.400\cdot 10^{-5}\cdot 0.8} \approx 0.7880 \end{aligned}$$

(c) Berechnen Sie die Wahrscheinlichkeit $P(Y_7 = g|$„TCGCGA“$)$, dass das nächste Nukleotid $Y_7 = g$ ist! Berücksichtigen Sie dabei beide Modelle.

$$\begin{aligned} P(Y_7=g|Y) &= P(Y_7=g|Y_6=a,X=f)P(X=f|Y) \\ &+\ P(Y_7=g|Y_6=a,X=w)P(X=w|Y) \\ &\approx 0.29\cdot 0.2120+0.43\cdot 0.7880 \approx 0.4003 \end{aligned}$$

(d) Ändert sich diese Vorhersage, wenn Sie nur das wahrscheinlichste Modell gemäß der MAP-Methode berücksichtigen?

- Nach der MAP-Methode ist das Modell $X = w$ zutreffend. Somit wird für die Vorhersage angenommen, dass die DNA-Sequenz zu einer *CpG-Insel* gehört. Daraus folgt dann sofort $P(Y_7 = g|$„TCGCGA“$) = 0.43$.
- Im Gegensatz dazu berücksichtigt die Mittelung auch das andere, nicht so wahrscheinliche Modell $X = f$. Dies fürt zu unterschiedlichen Vorhersagen, insbesondere weil ca. 21% Wahrscheinlichkeit für eine Lage außerhalb einer *CpG-Insel* nicht zu vernachlässigen sind.

Aufgabe 2 – Parameterschätzung **(50%)**

Die Berliner S-Bahn ist dafür bekannt, dass die Züge häufig zu spät ankommen. Als einfaches Modell nehmen Sie an, dass jede Zugfahrt unabhängig ist und eine Verspätung mit der Wahrscheinlichkeit p auftreten kann. Allerdings hatten Sie in diesem Monat Glück und bei bisher 10 Fahrten mit der S-Bahn waren die Züge alle pünktlich.

(a) Welche Wahrscheinlichkeitsverteilung hat das Auftreten von mindestens einer Verspätung bei insgesamt n Zugfahrten, wenn die Wahrscheinlichkeit hierfür bei jeder Fahrt p beträgt?

$$P(n|p) = 1 - (1-p)^n$$

(b) Die S-Bahn plant ein neues Entschädigungsmodell für Käufer von 4-Fahrten-Karten. Diese sollen 1 Euro zurückerhalten, falls es auf mindestens einer der 4 Fahrten zu einer Verspätung kommt. Wie hoch ist die zu erwartende Entschädigung für eine 4-Fahrten-Karte, wenn Sie $p = 0.1$ annehmen?

$$\langle E \rangle = 1 - (1-0.1)^4 = 1 - 0.9^4 = 1 - 0.6561 = 0.3439$$

(c) In der Zeitung lesen Sie, dass zwei von fünf S-Bahn-Zügen verspätet am Ziel ankommen. Sie wollen diese Information als Vorwissen in Form einer Beta-Verteilung $\text{Beta}(p;\alpha,\beta) = B(\alpha,\beta)\, p^{\alpha-1}(1-p)^{\beta-1}$ in ihrer Schätzung von p nutzen. Wie sollten Sie die Hyperparameter α und β wählen? Begründen Sie!

- Die Nachricht entspricht einer Beobachtung von 2 Verspätungen bei 5 Fahrten.
- In diesem Fall ist die zugehörige Likelihood durch die Binomialverteilung

  $$P(k=2|n=5,p) = \binom{5}{2} p^2(1-p)^3$$

  gegeben.
- Um diese Information im neuen Prior zu verwenden, sollte dieser proportional zur oben berechneten Likelihood gewählt werden:

  $$\begin{aligned} & P(p) \propto P(k=2|n=5,p) \\ \iff\ & p^2(1-p)^3 = p^{\alpha-1}(1-p)^{\beta-1} \\ \iff\ & \alpha = 3 \wedge \beta = 4 \end{aligned}$$

  Das entspricht den Hyperparametern $\alpha = 3$ und $\beta = 4$.

(d) Zeigen Sie, dass die Maximum-a-posteriori Hypothese für eine Folge von n pünktlichen Zügen durch $p = (\alpha - 1)/(n + \alpha + \beta - 2)$ gegeben ist!

- Der Logarithmus des Posteriors ist durch

$$\begin{aligned}\log P(p|n) &= \log\left[\frac{P(n|p)\mathrm{Beta}(p;\alpha,\beta)}{P(n)}\right] \\ &= n\log(1-p) + (\alpha-1)\log p + (\beta-1)\log(1-p) + \log C \\ &= (\alpha-1)\log p + (n+\beta-1)\log(1-p) + \log C\end{aligned}$$

  mit der Normierungskonstanten $C = B(\alpha,\beta)/P(n)$ gegeben.

- Berechnung der Ableitung:

$$\frac{\mathrm{d}}{\mathrm{d}p}\log P(p|n) = \frac{\alpha-1}{p} - \frac{n+\beta-1}{1-p}$$

- Bedingung für ein Maximum

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}p}\log P(p|n) = 0 &\iff \frac{\alpha-1}{p} = \frac{n+\beta-1}{1-p} \\ &\iff (\alpha-1)(1-p) = (n+\beta-1)p \\ &\iff p = \frac{\alpha-1}{n+\alpha+\beta-2}\end{aligned}$$

(e) Welchen Wert p hat die Maximum-a-posteriori-Hypothese für $n = 10$, wenn Sie die Hyperparameter auf $\alpha = 2$ und $\beta = 6$ setzen?

$$p = \frac{\alpha-1}{n+\alpha+\beta-2} = \frac{2-1}{10+2+6-2} = \frac{1}{16} \approx 0.0625$$