Technische Universität Berlin
Fakultät IV – Elektrotechnik und Informatik

Künstliche Intelligenz: Grundlagen und Anwendungen

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

Wintersemester 2016 / 2017

7. Aufgabenblatt

Abgabetermin: 25.01.2017

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?

(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!

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

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

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?

(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?

(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!

(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!

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