Übungen Algorithmen der Bioinformatik II
Wintersemester 2004/05
Stefan Posch, Birgit Möller, Denis Williams

Institut für Informatik
Universität Halle

Blatt 3

Aufgabe 3.1 Betrachten Sie N statistisch unabhängige Zufallsvariablen X_n, die alle einer Normalverteilung $\mathcal{N}(x|\mu, \sigma^2)$ folgen.

(a) Bestimmen Sie die Maximum Likelihood Schätzer für μ und σ.

(b) Vergleichen Sie dies mit den erwartungstreuen Schätzern.

Aufgabe 3.2 Generieren Sie 10 normalverteilte Zufallszahlen x_n mit der Dichtefunktion

$$P(X = x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$$

(a) Berechnen Sie nun aus dieser "Stichprobe" die MLEs $\hat{\mu}$ und $\hat{\sigma^2}$.

(b) Wiederholen Sie das Ganze 10^6 mal und erstellen Sie Histogramme für $\hat{\mu}$ und $\hat{\sigma^2}$.

(c) Welchen Zusammenhang gibt es zwischen den Histogrammen und den Erwartungswerten $E(\hat{\mu})$ und $E(\hat{\sigma^2})$. Welchen Zusammenhang gibt zwischen den Histogrammen und den Schätzern, wenn wir diese als Zufallsvariable betrachten?

(d) Wiederholen Sie alles. Ändert sich etwas?

Hinweis (für octave): `help normal_rnd`

Aufgabe 3.3 Unter

`http://www.informatik.uni-halle.de/~posch/AG/Lehre/AdB-II_WS04/uebungen/dnaSeq.txt`

ist eine hypothetische DNA-Sequenz D mit 1000 Basen als String gegeben. Jedes Zeichen im String wurde, unabhängig von den anderen Zeichen, mit dem gleichen DNA-Würfel gewürfelt. Dabei hatte jede Seite $X \in \{A, C, G, T\}$ des Würfels die gleiche Wahrscheinlichkeit $p_X = 1/4$. Wir bezeichnen dieses Würfelmodell mit M_0.

(a) Bestimmen Sie anhand der ersten L Basen der DNA-Sequenz mögliche Modellparameter $\{p_A^L, p_C^L, p_G^L, p_T^L\}$ durch Auswertung der relativen Häufigkeiten von A, C, G, T für Werte $L \in \{25, 50, 100\}$. Die zugehörigen Modelle seien mit M_L bezeichnet.

(b) Betrachten Sie im folgenden Modelle $M(p_C)$ wobei $p_A = 1/4$, $p_C \in [0, 1/2]$, $p_G = 1/2 - p_C$ und $p_T = 1/4$. Es existiert also ein freier Parameter (p_C) im Modell. Wählen Sie ein Gitter für diesen Parameter und berechnen Sie für alle Parameterwerte die Likelihood-Werte $P(D_n|M(p_C))$ für $n \in \{50, 100, 500, 1000\}$. Für jeden Wert n normalisieren Sie die erhaltenen Funktionswerte $P(D_n|M(p_C))$, so dass das Maximum eins beträgt, und zeichnen die normalisierten Werte als Funktion von p_C in ein Diagramm. Diskutieren Sie das Diagramm.

Abgabe: 4.11.04