Übungen Algorithmen der Bioinformatik II
Wintersemester 2006/07

Dr. Ivo Große, Dipl.-Bioinf. Jan Grau

Institut für Informatik
Universität Halle

Blatt 8

Aufgabe 8.1 Auf der Webseite zur Vorlesung finden Sie in der Datei `dnaSeq.txt` eine hypothetische DNA-Sequenz D_{1000} 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.

Aufgabe 8.2 Betrachten Sie N statistisch unabhängige Zufallsvariablen X_n, die alle einer Normalverteilung $\mathcal{N}(x|\mu, \sigma^2)$ folgen. Leiten Sie den Erwartungswert des Maximum Likelihood Schätzers von μ und σ^2 her. Ist der Schätzer erwartungstreu? Falls nicht, wie könnte ein erwartungstreuer Schätzer aussehen? Generieren Sie 10 normalverteilte Zufallszahlen x_n mit der Dichtefunktion

$$P(X = x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x2}{2}},$$

und berechnen Sie aus dieser Stichprobe den Maximum Likelihood Schätzer von μ und σ^2. Wiederholen Sie das ganze 10^6 mal und erstellen Sie Histogramme von $\hat{\mu}$ und $\hat{\sigma^2}$. 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? Wiederholen Sie alles. Ändert sich etwas?
Hinweis (für octave): `help normal_rnd`

Abgabe: 12.12.06

Aufgabe 8.3 Leiten Sie die Maximum Likelihood Schätzer der Parameter p bzw. λ der geometrischen bzw. Poisson Verteilung her. Leiten Sie die Erwartungswerte dieser Schätzer wenn möglich analytisch her. Sind die Schätzer erwartungstreu? Überprüfen Sie Ihre Aussagen durch Simulationen mit einem Stichprobenumfang von $N = 10$ für $p = 0.5, 0.8, 0.95$ bzw. $\lambda = 1.0, 4.0, 19.0$.

Aufgabe 8.4 Der Datensatz `exons.txt` enthält Längen von protein-kodierenden Exons, und der Datensatz `introns.txt` enthält Längen von Introns. Führen Sie für beide Datensätze separat die folgenden Aufgaben durch.

(a) Stellen Sie das Histogramm der Längen grafisch dar.

(b) Stellen Sie für die geometrische bzw. die Poisson Verteilung die Log-Likelihood als Funktion von p bzw. λ grafisch dar.

(c) Berechnen Sie die Maximum Likelihood Schätzwerte $\hat{p}$ bzw. $\hat{\lambda}$ und deren Log-Likelihhod. Welcher der beiden Log-Likelihood-Werte ist größer? Welche der beiden Verteilungen scheint Ihnen (für diesen Datensatz) geeigneter?

(d) Stellen Sie die beiden Verteilungen $P(k|\hat{p})$ bzw. $P(k|\hat{\lambda})$ grafisch dar und vergleichen Sie sie mit dem oben erstellten Histogramm? Welche der beiden Verteilungen scheint Ihnen rein visuell (für diesen Datensatz) geeigneter?

Abgabe: 12.12.06