Übungen Algorithmen der Bioinformatik II
Wintersemester 2007/08
Prof. Dr. Stefan Posch, Dr. Birgit Möller

Institut für Informatik
Universität Halle

Blatt 4

Aufgabe 4.1

(a) Beweisen Sie für die Bernoulli-Verteilung mit Parameter p:

$$\begin{aligned} E[X] &= p \\ Var[X] &= p(1-p) \end{aligned}$$

(b) Ein möglicher Schätzer für die Varianz einer gemäß einer Bernoulli-Verteilung verteilten Zufallsvariable X ist der folgende:

$$\hat{V}(X_1, \ldots, X_N) = \bar{X}(1-\bar{X})$$

$\bar{X}$ ist dabei wie üblich das arithmetische Mittel über N Realisierungen der Zufallsvariable X.

Überprüfen Sie, ob dies ein erwartungstreuer Schätzer ist.
Untermauern Sie Ihr Ergebnis durch Simulationen.

Aufgabe 4.2 Auf der Webseite zur Vorlesung finden Sie in der Datei "dnaSeq.txt" 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 $p_X = 1/4$.

(a) Bestimmen Sie die ML-Schätzwerte für die Modellparameter $\{p_A^L, p_C^L, p_G^L, p_T^L\}$, wobei L jeweils die ersten L Basen der Sequenz umfasst mit $L \in \{25, 50, 100\}$. Die zugehörigen Modelle seien jeweils 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: 09.11.07

Aufgabe 4.3 Wir betrachten N statistisch unabhängige Zufallsvariablen X_n, die alle einer Normalverteilung $\mathcal{N}(X|\mu, \sigma^2)$ folgen. Sie haben bereits die ML-Schätzer $\hat{\mu}$ und $\hat{\sigma}$ durch Maximieren der Log-Likelihood $\log P(\vec{X}|\mu, \sigma^2)$ bestimmt. Dabei ist $\vec{X} = (X_1, \ldots, X_N)$ wie üblich ein Vektor von N unabhängigen Zufallsvariablen mit dieser Normalverteilung.

Nun betrachten wir die folgendermaßen gewichtete Log-Likelihood:

$$\sum_{n=1}^{N} \gamma_n \log \mathcal{N}(x_n|\mu, \sigma^2)$$

(a) Bestimmen Sie diejenigen $\tilde{\mu}$ und $\tilde{\sigma}^2$, die diese Funktion maximieren

(b) Die Datenpunkte x_n liegen nun nicht mehr kontinuierlich vor, sondern sind äquidistant diskretisiert und liegen als Histgramm vor:

$$h[i] := |\{x_n | i \leq x_n < i + 1\}|$$

Wir sind wiederum an den ML-Schätzwerten $\hat{\mu}$ und $\hat{\sigma}$ interessiert, die wir nur mit Hilfe des Histogramms, d.h. ohne Rückgriff auf die Originaldatenpunkte, bestimmen wollen.

Abgabe: 09.11.07