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

Institut für Informatik
Universität Halle

Blatt 12

Aufgabe 12.1 Wir betrachten L diskrete Zufallsvariablen X_l über dem identischen Wertevorrat $\mathcal{A}$ und einen DAG auf der Knotenmenge $\{X_1, \ldots, X_L\}$.

Gegeben seinen nun bedingte Wahrscheinlichkeiten

$$\begin{array}{lll} \forall l : |\mathrm{Pa}_l| > 0, & \forall a \in \mathcal{A}, b \in \mathcal{A}^{|\mathrm{Pa}_l|} & : P(X_l = a | \mathrm{Pa}_l = b) \\ \forall l : |\mathrm{Pa}_l| = 0, & \forall a \in \mathcal{A} & : P(X_l = a) \end{array}$$

Zeigen Sie, daß

$$P(\vec{X} = \vec{x}) := \prod_{l:|\mathrm{Pa}_l|>0} P(X_l = x_l | \mathrm{Pa}_l = \vec{x}_{\mathrm{Pa}_l}) \prod_{l:|\mathrm{Pa}_l|=0} P(X_l = x_l) \qquad , \forall \vec{x} \in \mathcal{A}^L$$

eine Verbundwahrscheinlichkeit auf $(X_1, \ldots, X_L)$ definiert.

($\vec{x}_{\mathrm{Pa}_l}$ bezeichnet dabei die Elemente des Vektors $\vec{x}$, die zu den Eltern Pa_l von X_l gehören.)

Aufgabe 12.2

(a) Betrachten Sie den linkesten Graphen aus der Aufgabe 11.2 als Bayesnetz mit den dort die angegebene Verbundwahrscheinlichkeit, die allerdings korrekterweise wie rechts angegeben lauten müßte. Nehmen Sie an, daß diese Verbundwahrscheinlichkeit der ML-Schätzwert einer Menge von Beobachtunge sei. Geben Sie eine mögliche Menge solcher Beobachtungen an.

V	C	S	P(V,C,S)
1	1	1	$\frac{2}{13}$
1	1	0	$\frac{1}{13}$
1	0	1	$\frac{1}{13}$
1	0	0	$\frac{1}{13}$
0	1	1	$\frac{4}{13}$
0	1	0	$\frac{2}{13}$
0	0	1	$\frac{1}{13}$
0	0	0	$\frac{1}{13}$

(b) Bestimmen Sie für diese Menge von Beobachtungen die ML-Schätzwerte der Parameter des Bayesnetzes zum rechtesten Graphen.

(c) Berechnen Sie für beide Bayesnetze die Likelihood des Datensatzes.

Aufgabe 12.3 Der Datensatz

`http://www.informatik.uni-halle.de/~posch/AG/Lehre/AdB-II_WS04/uebungen/sigma70.txt`
besteht aus 238 Sigma-70 Bindungsstellen der Länge 12.

Unter `http://www.informatik.uni-halle.de/~posch/AG/Lehre/AdB-II_WS04/uebungen/bg.txt`
ist ein Hintergrunddatensatz von ebenfalls 238 Sequenzen der Länge 12.

Teilen Sie beide Datensätze jeweils in zwei Hälften, wir erhalten so zwei TFBS-Datensätze F_1 und F_2 und zwei Hintergrunddatensätze B_1 und B_2.

(a) Konstruieren Sie zwei Bayes-Klassifikatoren K_i, indem Sie jeweils auf B_i und $F_i, i \in \{1, 2\}$, ein inhomogenes bzw. homogenes MM(0) für die TFBS bzw. den Hintergrund trainieren.

Machen Sie sich Gedanken zur Schätzung der a priori Wahrscheinlichkeiten der beiden Klassen!

(b) Berechnen Sie für alle vier Markov-Modell jeweils das Sequenz-Logo.

(c) Wenden Sie beide Klassifikatoren auf beide Datensätze an und bestimmen Sie die Fehlerrate.

Abgabe: 20.1.05