technische universität dortmund

UNIVERSITÄT DUISBURG ESSEN
Offen im Denken

Bioinformatik, Fakultät für Informatik

Genominformatik, Medizinische Fakultät

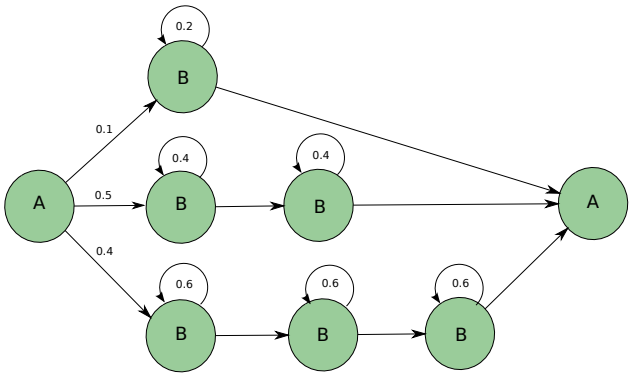
Prof. Dr. Sven Rahmann

Wintersemester 2019/20

Algorithmische Bioinformatik
Übungsblatt 5

Ausgabe: 12. November 2019 · **Besprechung:** 26. November (2 Wochen)

Aufgabe 5.1 Gegeben ist folgender Ausschnitt aus einem HMM. Die Emissionen sind deterministisch (A oder B). Offensichtlich werden Sequenzen der Form $\mathtt{AB}^+\mathtt{A}$ erzeugt. Berechne für alle sinnvollen Werte von ℓ die Wahrscheinlichkeit für das Ereignis, dass die Länge des erzeugten B-Laufs genau ℓ beträgt.



Aufgabe 5.2 Wir erinnern uns an das Casino von Blatt 4. Wir hatten Würfelwurfreihen mit Zustandsinformation (fair, unfair, anlocken) simuliert und deren log-Gesamtwahrscheinlichkeit berechnet, sowie für jede Reihe den Viterbi-Pfad berechnet und mit der wahren Zustandsfolge verglichen.

Nun soll zusätzlich *posterior decoding* angewendet werden, d.h. der Forward-Backward-Algorithmus. Welcher Zustand ist in jeder Reihe am wahrscheinlichsten? Gestalte die Ausgabe so, dass der Zustand nur ausgegeben wird, wenn seine Wahrscheinlichkeit $\geq 2/3$ ist, sonst gib ? aus. Wie oft entspricht diese Methode dem wahren Zustand, und wie oft wird ? ausgegeben?

Aufgabe 5.3 Zuletzt untersuchen wir, ob wir die drei Würfeltypen und ihre Wahrscheinlichkeiten aus beobachteten Sequenzen schätzen können. Dazu gehen wir, wie in der Vorlesung beschrieben, iterativ vor (Baum-Welch-Algorithmus, eine Variante des expectation maximization Algorithmus).

Wir setzen dazu voraus, dass Startparameter vorliegen, und zwar sowohl für die Startwahrscheinlichkeiten und die Übergänge als auch für die Emissionen. Diese können initial zufällig gewählt werden, aber eine bessere Wahl nutzt die Information, dass es einen fairen, einen unfairen und einen Anlockwürfel gibt, indem die Wahrscheinlichkeiten in etwa korrekt gesetzt werden, sich zumindest in die richtige Richtung voneinander unterscheiden.

Mit dem Forward-Backward-Algorithmus berechnen wir die bedingten Wahrscheinlichkeiten, zum Zeitpunkt t in Zustand q zu sein, sowie die *gemeinsame* Wahrscheinlichkeit, zum Zeitpunkt t in q' und zum Zeitpunkt t in q zu sein. Daraus werden die neuen Emissions- und Übergangswahrscheinlichkeiten geschätzt.

Wie gut lassen sich die Übergangs- und die Emissionswahrscheinlichkeiten rekonstruieren, wenn die initialen Parameter (a) zufällig und (b) approximativ korrekt gewählt werden?