Übungen Sequenzanalyse I
Sommersemester 2007

Dr. Ivo Große

Institut für Informatik
Universität Halle

Blatt 4

Aufgabe 4.1 (4 Punkte)

Führen Sie den Beweis zur folgenden effizienten Auswertung von $Q(\vec{\theta}; \vec{\theta}^{(t)})$ im EM-Algorithmus für unabhängige Beobachtungen:

$$\begin{aligned} Q(\vec{\theta}; \vec{\theta}^{(t)}) &= \sum_{i=1}^{N} \sum_{\vec{u} \in \mathcal{U}} \left(\alpha_{iu_i} \left(\prod_{j=1}^{N} \gamma_{ju_j}^{(t)} \right) \right) \\ &= \sum_{i=1}^{N} \sum_{u_i=1}^{K_i} \alpha_{iu_i} \gamma_{ju_j}^{(t)} \\ &= \sum_{i=1}^{N} \sum_{u=1}^{K_i} \log P_i(U_i = u, \vec{X}_i = \vec{x}_i | \vec{\theta}) P_i(U_i = u | \vec{X}_i = \vec{x}_i, \vec{\theta}^{(t)}) \end{aligned}$$

Aufgabe 4.2 (8 Punkte)

Leiten Sie jeweils die drei Varianten (π gegeben, q zu schätzen; q gegeben, π zu schätzen; π und q zu schätzen) des M-Schrittes des EM-Algorithmus für Mischungen aus homogenen und inhomogenen Markov Modellen nullter und erster Ordnung her.

Aufgabe 4.3 (4 Punkte)

Schätzen Sie die Parameter q_1 und q_2 für den Datensatz aus Aufgabe 3.3 mit Hilfe des EM-Algorithmus. Definieren Sie Ihr Initialisierungs- und Ihr Abbruchkriterium, und plotten Sie die Log-Likelihood für jeden Iterationsschritt. Wiederholen Sie den EM-Algorithmus 100-mal, und bestimmen Sie das Maximum der erreichten Log-Likelihoods. In wie vielen der 100 EM-Läufe wurde diese maximale Log-Likelihood erreicht? Vergleichen Sie diese maximale Log-Likelihood mit der in Aufgabe 3.3 bestimmten. Vergleichen Sie die dazugehörigen Maximalstellen, d. h. die dazugehörigen Schätzwerte.

Aufgabe 4.4 (8 Punkte)

Benutzen Sie die Datensätze `seq_1` und `seq_3` zum Trainieren eines PWM-Mischmodells mit $K = 2$ Klassen und eines WAM-Mischmodells mit $K = 2$ Klassen. Wie lauten die beiden PWMs bzw. die beiden Dinukleotid-PWMs? Vergleichen Sie diese PWMs bzw. Dinukleotid-PWMs mit der PWM bzw. Dinukleotid-PWM aus Aufgabe 2.2.

Plotten Sie die ROC Kurven für die beiden Modellkombinationen (PWM-Mischmodell für Donor und PWM-Mischmodell für Nicht-Donor, WAM-Mischmodell für Donor und WAM-Mischmodell für Nicht-Donor) für die Datensätze `seq_2` und `seq_4`. Welche Modellkombination liefert die genauere Klassifizierung von Donorstellen und Nicht-Donorstellen?

Vertauschen Sie die Datensätze `seq_1` und `seq_3` mit den Datensätzen `seq_2` und `seq_4`, und wiederholen Sie die Analyse. Wie robust sind die beiden ROC Kurven und die beiden AUC Werte?

Wiederholen Sie die Analyse für $K = 1$, $K = 3$, $K = 4$ und $K = 5$. Testen Sie, dass Ihre Ergebnisse für $K = 1$ identisch sind zu Ihren Ergebnissen aus Aufgabe 2.2. Welche der zehn Modellkombinationen liefert die genaueste Klassifizierung von Donorstellen und Nicht-Donorstellen?

Hinweis: Als Mass für die Genauigkeit der Klassifikation eignet sich die Fläche unter der ROC Kurve (AUC).

Abgabetermin: 6. Juni