Übungen Expressionsdatenanalyse I
Sommersemester 2007

Dr. Ivo Große

Institut für Informatik
Universität Halle

Blatt 4

Aufgabe 4.1 (8 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 die sieben Varianten des M-Schrittes des EM-Algorithmus für univariate Gaußsche Mischmodelle her.

Aufgabe 4.3 (8 Punkte)

Schätzen Sie die Parameter μ_1 und μ_2 für den Datensatz aus Aufgabe 3.2 mit Hilfe des EM-Algorithmus. Definieren Sie Ihr Initialisierungsverfahren 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 3.2 bestimmten. Vergleichen Sie die dazugehörigen Maximalstellen, d. h. die dazugehörigen Schätzwerte.

Aufgabe 4.4 (16 Punkte)

Der Datensatz `gauss_3` wurde durch ein bivariates Gaußsches Mischmodell mit $K = 3$ Klassen generiert, wobei in jeder Klasse die beiden Komponenten der bivariaten Gaußverteilung statistisch unabhängig sind.

(a) Bekannt seien die Parameter $\pi_1 = 0.5$, $\mu_{21} = 0$, $\sigma_{21}^2 = 1$, $\mu_{32} = 1$ und $\sigma_{31}^2 = 16$. Schätzen Sie die restlichen Parameter π_2, π_3, μ_{11}, μ_{12}, μ_{22}, μ_{31}, σ_{11}^2, σ_{12}^2, σ_{22}^2

und σ_{32}^2 per EM-Algorithmus, und geben Sie die Schätzwerte sowie die erreichte Log-Likelihood an. Definieren Sie außerdem Ihr Initialisierungsverfahren und Ihr Abbruchkriterium.

(b) Bekannt sei jetzt nur noch der Parameter $\pi_1 = 0.5$. Außerdem sei nun aber bekannt, dass $\mu_{21} = \mu_{31}$ und $\sigma_{11}^2 = \sigma_{21}^2 = \sigma_{32}^2$. Schätzen Sie für diesen Fall alle Parameter außer π_1, und geben Sie die Schätzwerte sowie die erreichte Log-Likelihood an. Definieren Sie außerdem Ihr Initialisierungsverfahren und Ihr Abbruchkriterium.

Wiederholen Sie beide Teilaufgaben mit dem stochastischen EM-Algorithmus. Definieren Sie hierbei wiederum Ihr Initialisierungsverfahren und Ihr Abbruchkriterium. Vergleichen Sie die Parameter und dazugehörigen Log-Likelihood Werte mit denen des EM-Algorithmus. Welche Schlussfolgerungen können Sie aus diesem Vergleich ableiten?

Wiederholen Sie beide Teilaufgaben mit dem aus dem EM-Algorithmus abgeleiteten k-means Algorithmus. Definieren Sie hierbei wiederum Ihr Initialisierungsverfahren und Ihr Abbruchkriterium. Vergleichen Sie die Parameter und dazugehörigen Log-Likelihood Werte mit denen des EM-Algorithmus und denen des stochastischen EM-Algorithmus. Welche Schlussfolgerungen können Sie aus diesem Vergleich ableiten?

Abgabetermin: 12. Juni