

14. Mai 2007

Übung zur Vorlesung

Data Mining mit CI-Methoden
(SS 2007)

Blatt 2

Aufgabe 2 (10 Punkte)

Gegeben seien die Datenpunkte $x_1, \dots, x_N \in \mathbb{R}^n$. Wir nehmen an, dass die Datenpunkte von einer Mischverteilung aus K multivariaten Normalverteilungen mit Wahrscheinlichkeitsdichten

$$f_{C_k}(x) = \frac{1}{\pi^{n/2} \sqrt{\det \Sigma_k}} \exp \left[-\frac{1}{2} (x - \mu_k)' \Sigma_k^{-1} (x - \mu_k) \right]$$

für $k = 1, \dots, K$ generiert wurden.

- Verwenden Sie den EM-Algorithmus für $K = 2, 4, 6$ Zufallsvariable bzgl. der Daten in den Dateien `uniform.txt`, `uniform4.txt` und `olympic.txt` (von Übungsblatt 1). Geben Sie eine Zusammenfassung Ihrer Resultate an.
- Setzen Sie zur Maximierung der Log-Likelihoodfunktion

$$L(\pi_k, \mu_k, \Sigma_k : k = 1, \dots, K) = \sum_{i=1}^N \log \left[\sum_{k=1}^K \pi_k \cdot f_{C_k}(x_i) \right]$$

bzgl. der zu schätzenden Parameter $\theta = (\pi_k, \mu_k, \Sigma_k : k = 1, \dots, K)$ nicht den EM-Algorithmus sondern einen evolutionären Algorithmus Ihrer Wahl ein. Verwenden Sie die gleichen Datensätze wie in der vorherigen Teilaufgabe für $K = 2, 4, 6$. Geben Sie eine Zusammenfassung Ihrer Resultate an.

Hinweis: Es genügt, den Spezialfall $\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_n^2)$ zu betrachten.

Abgabetermin: Dienstag, 22. Mai 2007, 20:00 Uhr per email.

Bitte dokumentieren Sie (für sich), wieviel Zeit Sie für die Aufgaben benötigt haben.