Übungen Expressionsdatenanalyse I
Sommersemester 2007

Dr. Ivo Große

Institut für Informatik
Universität Halle

Blatt 3

Aufgabe 3.1 (6 Punkte)

(a) Finden Sie ein Gegenbeispiel, welches zeigt, dass auch der verallgemeinerte k-means Algorithmus (Kapitel III.4.2.2) die verallgemeinerte *Intra-Cluster-Streuung* $W_d(\underline{k})$ nicht monoton fallen lässt. Wählen Sie hierfür einen Abstand d, der verschieden vom Euklidischen Abstand ist.

(b) Finden Sie ein Gegenbeispiel, welches zeigt, dass der k-median Algorithmus (Kapitel III.4.2.3) die verallgemeinerte *Intra-Cluster-Streuung* $W_d(\underline{k})$ nicht monoton fallen lässt, wobei d hier den Manhattan Abstand bezeichnet.

(c) Finden ein Gegenbeispiel, welches zeigt, dass der k-medoids Algorithmus (Kapitel III.4.2.4) die verallgemeinerte *Intra-Cluster-Streuung* $W_d(\underline{k})$ nicht monoton fallen lässt.

Aufgabe 3.2 (8 Punkte)

Der Datensatz `gauss_2` enthält 100 reellwertige Datenpunkte, die durch ein Gaußsches Mischmodell mit $K = 2$ Klassen generiert wurden. Die beiden Klassenwahrscheinlichkeiten $\pi_1 = \pi_2 = 0.5$ sowie die beiden Varianzen $\sigma_1^2 = \sigma_2^2 = 1$ sind extern vorgegeben. Die einzigen zu schätzenden Parameter dieses Modells sind die Mittelwerte μ_1 und μ_2.

(a) Schätzen Sie die Parameter μ_1 und μ_2 mittels Maximum Likelihood Prinzip unter Nutzung der gegebenen Klassenzugehörigkeiten.

(b) Ignorieren Sie für die folgenden drei Teilaufgaben die Klassenzugehörigkeiten, d. h. betrachten Sie die Klassenzugehörigkeiten im folgenden als nicht gegeben. Plotten Sie die Log-Likelihood als Funktion von μ_1 und μ_2.

(c) Bestimmen Sie die Maxima und Maximalstellen dieser Funktion mit geringem Aufwand durch ein Verfahren Ihrer Wahl (gitterbasierte Abrasterung, Maximumssuche per Auge, Gradientenanstieg, etc.). Vergleichen Sie die Maximalstellen mit den in Aufgabe 3.2 (a) geschätzten Werten, und diskutieren Sie die Unterschiede.

(d) Versuchen Sie, die Maximalstellen analytisch zu bestimmen, indem Sie die Log-Likelihood nach π_1 und π_2 ableiten und beide Ableitungen Null setzen. Worin liegt das Problem, dieses Gleichungssystem (mit lediglich zwei Gleichungen und zwei Unbekannten) analytisch zu lösen?

Aufgabe 3.3 (3 Punkte)

Beweisen Sie, dass für zwei verschiedene Wahrscheinlichkeitsvektoren $\underline{p} \neq \underline{q}$ mit Komponenten p_i bzw. q_i die relative Entropie $\sum_i p_i \ln \frac{p_i}{q_i}$ stets positiv ist.

Aufgabe 3.4 (3 Punkte)

Beweisen Sie, dass aus der Identität

$$P(x_1, \ldots, x_N, u_1, \ldots, u_N) = \prod_{i=1}^{N} P(x_i, u_i)$$

die Identitäten

$$\begin{aligned} P(x_1, \ldots, x_N | u_1, \ldots, u_N) &= \prod_{i=1}^{N} P(x_i | u_i) \\ P(u_1, \ldots, u_N | x_1, \ldots, x_N) &= \prod_{i=1}^{N} P(u_i | x_i) \end{aligned}$$

folgen.

Abgabetermin: 29. Mai