Karlsruher Institut für Technologie

Lehrstuhl für Analytics and Statistics
Prof. Dr. Oliver Grothe
SS 2022

Aufgaben zur Vorlesung

Multivariate Verfahren

Übungsblatt VII

---

1. Beantworten Sie folgende Fragen bzw. bearbeiten Sie folgende Arbeitsanweisungen:
   (a) Was ist der Zweck einer Clusteranalyse?
   (b) Wie kann man bei einer Clusteranalyse vorgehen?
   (c) Wieso ist es schön, wenn die Distanzen zwischen den Objekten im Rahmen einer Clusteranalyse sehr unterschiedlich sind? Ist das wirklich schön?
   (d) Geben Sie ein Beispiel aus der Praxis, bei dem eine Clusteranalyse eine methodische Hilfe sein kann.

2. Gegeben ist folgende Distanzmatrix:

$$D = \begin{pmatrix} 0 & 4 & 5 & 8 \\ 4 & 0 & 2 & 4 \\ 5 & 2 & 0 & 3 \\ 8 & 4 & 3 & 0 \end{pmatrix}$$

   Erinnern Sie sich an die Lance-Williams Formel und
   (a) führen Sie den ersten Schritt eines Clusteringverfahrens durch mit $\{\alpha_i, \alpha_j, \beta, \gamma\} = \{1/2, 1/2, 0, -1/2\}$ und bestimmen Sie die neue Distanzmatrix.
   (b) führen Sie den ersten Schritt eines Clusteringverfahrens durch mit $\{\alpha_i, \alpha_j, \beta, \gamma\} = \{1/2, 1/2, 0, 1/2\}$ und bestimmen Sie die neue Distanzmatrix.
   (c) führen Sie den ersten Schritt eines Clusteringverfahrens durch mit $\{\alpha_i, \alpha_j, \beta, \gamma\} = \{\frac{n_i}{n_i+n_j}, \frac{n_j}{n_i+n_j}, 0, 0\}$ und bestimmen Sie die neue Distanzmatrix.

3. Gegeben ist eine Folge von Partitionen in Abhängigkeit des Abstands d:

$$\begin{array}{rl} 0 \le d < 2 & \{\{1\},\{2\},\{3\},\{4\},\{5\}\} \\ 2 \le d < 5 & \{\{1,3\},\{2\},\{4\},\{5\}\} \\ 5 \le d < 7 & \{\{1,2,3\},\{4\},\{5\}\} \\ 7 \le d < 15 & \{\{1,2,3\},\{4,5\}\} \\ 15 \le d & \{1,2,3,4,5\} \end{array}$$

   (a) Skizzieren Sie das zugehörige Dendrogramm.
   (b) Bestimmen Sie die zugehörige kophenetische Matrix D^*.

4. Gegeben sei ein Datensatz mit zwei Merkmalen von Nagetieren: Der Anzahl Kinder pro Wurf und der Anzahl Würfe während der Lebensspanne des jeweiligen Nagetieres. In der folgenden Tabelle sind die Daten gegeben:

Laufende Nummer	Anzahl Kinder/Wurf	Anzahl Würfe/Leben
1	1	1
2	1	2
3	2	1
4	3	8
5	4	7
6	4	9
7	5	1
8	6	1

   Führen Sie das K-Means Verfahren mit $K = 3$ durch. Benutzen Sie als Zentroide die Mittelwerte der Punktwolken und verwenden Sie als Startkonfiguration $x_1 = (0, 0)$, $x_2 = (2, 10)$ und $x_3 = (6, 3)$.

5. Rechnerübung: Verwenden Sie den Datensatz *iris* aus Ilias, den Sie bereits bei der Multiplen Linearen Regression kennen gelernt haben. Dieser enthält vier Merkmale von Blütenblättern unterschiedlicher Schwertlilienarten (jeweils Länge und Breite des Sepalum (Kelchblatt) und Petalum(Kronblatt)). Führen Sie eine hierarchische Clusteranalyse durch. Wie viele unterschiedliche Schwertlilienarten sind Ihrer Meinung nach im Datensatz vertreten?