Prof. Dr. Eckhard Liebscher
Fachgruppe Mathematik

Aufgabenserie L1 zur Vorlesung "Maschinelles Lernen"

1. Für eine Stichprobe aus 6 Elementen $e_1, e_2, \ldots, e_6$ ist eine hierarchische Clusteranalyse durchzuführen. Vorgegeben ist dazu die Ähnlichkeitsmatrix

$$\begin{pmatrix} 0 & 11 & 27 & 51 & 31 & 23 \\ 11 & 0 & 3 & 77 & 98 & 16 \\ 27 & 3 & 0 & 7 & 18 & 84 \\ 51 & 77 & 7 & 0 & 12 & 26 \\ 31 & 98 & 18 & 12 & 0 & 22 \\ 23 & 16 & 84 & 26 & 22 & 0 \end{pmatrix}.$$

Vergleichen Sie die entstehenden Hierarchien von Gruppierungen bei der Methode des entferntesten bzw. nächsten Nachbarn und bei der Mittelwertmethode. Geben Sie die entsprechenden Dendrogramme an.

2. Gegeben ist eine Datentabelle, in der Lebensmittel mit entsprechenden Merkmalen versehen wurden:

	vorzubereiten	zu kaufen	zu kochen	einfrierbar	drinkbar
Hähnchen	1	1	1	1	0
Spaghetti	1	1	1	0	0
Salat	1	1	0	0	0
Saft	0	1	0	0	1
Wasser	0	0	0	1	1
Wein	0	1	0	0	1

Stellen Sie unter Verwendung des Jaccard-Maßes die Distanzmatrix auf. Führen Sie die hierarchische Clusteranalyse mit Hilfe des Nächste-Nachbar-Verfahren durch, um eine Lösung mit 3 Gruppen zu erstellen.

3. Wir betrachten die Situation gemischter Merkmale. Bestimmen Sie die Distanz der

Merkmalsvektoren

$$\vec{x} = \begin{pmatrix} 1 \\ 170 \\ 3 \\ \text{hell} \\ 4 \end{pmatrix}, \quad \vec{y} = \begin{pmatrix} 0 \\ 236 \\ 5 \\ \text{mittel} \\ 6 \end{pmatrix}.$$

Das erste Merkmal ist dichotom. Das zweite Merkmal ist metrisch mit Werten aus $\mathbb{R}$. Im Datensatz ist der kleinste Wert 142 und der größte gleich 248. Das dritte Merkmal beinhaltet eine Bewertung, wobei die Werte in $\{1, 2, 3, 4, 5\}$ liegen. Das vierte Merkmal gibt Helligkeitsstufen an. Mögliche Werte sind "hell", "mittel" und "dunkel". Die Distanzensind wie folgt definiert: d("hell","mittel") $= 0.4$, d("hell","dunkel") $= 1$, d("mittel","dunkel") $= 0.6$. Das fünfte Merkmal gibt die Nummer eines Teilkollektivs an. Jedes Merkmal geht in die Distanz mit gleichem Gewicht ein.

4. Gegeben sind die Punkte

$$Y_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, Y_2 = \begin{pmatrix} 3 \\ 5 \end{pmatrix}, Y_3 = \begin{pmatrix} 10 \\ 2 \end{pmatrix}, Y_4 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}, Y_5 = \begin{pmatrix} 7 \\ 6 \end{pmatrix}, Y_6 = \begin{pmatrix} 8 \\ 7 \end{pmatrix}.$$

Man führe den k-means-Algorithmus durch, wenn die Startgruppierung $\mathbb{G} = \{G_1, G_2, G_3\}, G_1 = \{1, 3\}, G_2 = \{2, 4, 5\}, G_3 = \{6\}$ ist. Benutzen Sie dabei die City-Block-Metrik.

5. Führen Sie für einen der folgenden Datensätze eine Clusteranalyse durch: "essen", "autod", "staedte", "mineralwasser", "monde", "waschmaschine", "essen2", "teigwaren", "galapagos", "naehrstoffinh", und "kinder". Dabei ist sowohl ein hierarchisches Verfahren als auch der k-means-Algorithmus anzuwenden. Standardisieren Sie die metrischen Daten vor der Analyse. Lassen Sie sich zu den Ergebnissen entsprechende Grafiken (Dendrogramm, Mittelwertplot) ausgeben. Wählen Sie beim hierarchischen Verfahren eine geeignete Anzahl von Clustern aus. Hängt dabei das Ergebnis von der Wahl der Distanz bzw. von der Wahl des Fusionsverfahrens ab?

Link zu k-means-Clusterverfahren:
https://www.naftaliharris.com/blog/visualizing-k-means-clustering/