Clusteranalyse

Aufgabe 1:

Für fünf Filialen einer Supermarktkette erhält man für die Merkmale Umsatz und Verkaufsfläche, jeweils gemessen in geeigneten Einheiten, die folgende Datenmatrix:

Filiale	1	2	3	4	5
Umsatz	8	5	10	4	13
Verkaufsfläche	24	22	25	21	28

- Führen Sie eine hierarchische Klassifikation mit dem *Single Linkage* Verfahren durch. Verwenden Sie als zugrundeliegende Distanz zwischen einzelnen Objekten die quadrierte euklidische Distanz.
- Führen Sie eine hierarchische Klassifikation mit dem *Zentroid* Verfahren durch.
- Geben Sie für beide Verfahren das vollständige Dendrogramm an.

Aufgabe 2:

Der Datensatz `europa.txt` enthält Daten zu $n = 24$ europäischen Ländern. Folgende Variablen wurden erhoben: `ober` (Oberfläche in km^2), `einw` (Einwohner in Millionen), `brut` (BIP pro Kopf in \$) und `arb1` (Arbeitslosenquote in %).

- Lesen Sie den Datensatz in R ein und standardisieren Sie die Daten.
- Führen Sie mithilfe der Funktion `hclust()` eine hierarchische Klassifikation unter Einbeziehung aller vier Kovariablen mit dem *Single Linkage* Verfahren durch. Verwenden Sie als zugrundeliegende Distanz zwischen einzelnen Objekten die quadrierte euklidische Distanz.
- Führen Sie mithilfe der Funktion `hclust()` eine hierarchische Klassifikation unter Einbeziehung aller vier Kovariablen mit dem *Zentroid* Verfahren durch.
- Führen Sie mithilfe der Funktion `hclust()` eine hierarchische Klassifikation unter Einbeziehung aller vier Kovariablen mit dem *Complete Linkage* Verfahren durch. Verwenden Sie als zugrundeliegende Distanz zwischen einzelnen Objekten die Mahalanobis-Distanz.
Hinweis: Die Funktion `mahalanobis()` könnte hilfreich sein.
- Visualisieren und vergleichen Sie ihre Ergebnisse der Teilaufgaben b), c) und d) jeweils mithilfe eines Dendrogramms.

Aufgabe 3:

Der Übergang von der Partition $\mathcal{C}^{(\nu-1)}$ zur Folgepartition $\mathcal{C}^{(\nu)}$ sei durch die Vereinigung der Klassen C_m und $C_{\tilde{m}}$ aus $\mathcal{C}^{(\nu-1)}$ bestimmt. Zeigen Sie, dass sich für das *Single Linkage*, das *Complete Linkage* und das *Zentroid* Verfahren die Distanzen zwischen der neu entstehenden Klasse $C := C_m \cup C_{\tilde{m}}$ und den verbleibenden Klassen C_k , $k \neq m, \tilde{m}$ aus der folgenden Rekursionsformel berechnen lassen

$$D(C, C_k) = \alpha_m \cdot D(C_m, C_k) + \alpha_{\tilde{m}} \cdot D(C_{\tilde{m}}, C_k) + \beta \cdot D(C_m, C_{\tilde{m}}) + \gamma \cdot |D(C_m, C_k) - D(C_{\tilde{m}}, C_k)|,$$

mit

- $\alpha_m = \alpha_{\tilde{m}} = \frac{1}{2}$, $\beta = 0$ und $\gamma = -\frac{1}{2}$ für *Single Linkage*,
- $\alpha_m = \alpha_{\tilde{m}} = \frac{1}{2}$, $\beta = 0$ und $\gamma = \frac{1}{2}$ für *Complete Linkage*,