Trees / Clusteranalyse

Aufgabe 1:
Verwenden Sie für die folgenden Aufgaben das R-Paket `rpart` sowie zur Visualisierung das Paket `partykit`.

a) Der Datensatz `airquality` in R enthält Daten von Messungen der Luftqualität in New York.

   i) Fitten Sie ein lineares Modell mit `Ozone` als Responsevariable und den Variablen `Solar.R`, `Wind` und `Temp` (mit und ohne Interaktionen) als erklärende Variablen.

   ii) Fitten Sie anschließend einen Regressionsbaum mit `Ozone` als Responsevariable und den Variablen `Solar.R`, `Wind` und `Temp` als erklärende Variablen und vergleichen Sie die Ergebnisse.

b) Der Datensatz `iris` in R enthält Messungen zu den Größen von Kelch- und Blütendaten von 3 Iris-Arten.

   i) Fitten Sie ein einen Klassifikationsbaum mit den Einflussgrößen `Sepal.Length` und `Sepal.Width` und veranschaulichen sie die Klassifikation graphisch.

   ii) Fitten Sie anschließend einen Klassifikationsbaum mit allen vorhandenen Kovariablen und vergleichen Sie die Ergebnisse.

Aufgabe 2:
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

a) Führen Sie eine hierarchische Klassifikation mit dem *Single Linkage* Verfahren durch. Verwenden Sie als zugrundeliegende Distanz zwischen einzelnen Objekten die quadrierte euklidische Distanz.

b) Führen Sie eine hierarchische Klassifikation mit dem *Zentroid* Verfahren durch.

c) Geben Sie für beide Verfahren das vollständige Dendrogramm an.

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* und das *Complete Linkage* 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

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

mit

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

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