Multivariate Verfahren SoSe 2020 **Tutorium 3**

Prof. Dr. Annika Hoyer, Marc Schneble, Cornelia Gruber

Aufgabe 1: Hierarchisches Clustering in R

Lösen Sie folgende Aufgaben in `R`. Verwenden Sie dabei den Datensatz `mtcars`.

a) Warum ist es sinnvoll, die Variablen zu standardisieren (zentrieren und skalieren), oder zu normalisieren (Maximum abziehen und durch Spannweite/Range teilen) bevor man die euklidische Distanzmatrix berechnet?

b) Laden Sie das Paket `cluster` und machen Sie sich mit den Funktionen `agnes()` und `diana()` vertraut.

c) Berechnen Sie die euklidische Distanzmatrix für die standardisierten und nicht standardisierten Variablen.

d) Führen Sie ein agglomeratives Clustering mit `agnes()` durch. Verwenden Sie hierbei die Manhattan-Metrik und die euklidische Metrik für das Complete Linkage-Verfahren und vergleichen Sie die Dendrogramme für beide Metriken.

e) Führen Sie ein divisives Clustering mit `diana()` durch. Verwenden Sie hierbei die Manhattan-Metrik und die euklidische Metrik und vergleichen Sie beide Dendrogramme.

Aufgabe 2: k-means Clustering

Die folgende Tabelle enthält die Anzahl der Mitarbeiter und den Umsatz (in 10 000 Euro) für fünf verschiedene Unternehmen.

Unternehmen	Mitarbeiter	Umsatz (in 10 000 Euro)
U_1	2	10
U_2	2	5
U_3	8	4
U_4	5	8
U_5	7	5

Führen Sie ein k-means Clustering durch, um die fünf Unternehmen in zwei verschiedene Cluster einzuordnen. Als Distanzmaß soll die quadrierte euklidische Distanz verwendet werden. Die beiden Unternehmen U_1 und U_4 sollen dabei die Startpartition für je ein Cluster sein.

Man erhält für Iteration 0 folgende Distanzen:

$$x_1^{(0)} = (2, 10)^T, \quad x_2^{(0)} = (5, 8)^T$$

$$\mathbf{D}^{(0)} = \begin{pmatrix} 0 & 25 & 72 & 13 & 50 \\ 13 & 18 & 25 & 0 & 14 \end{pmatrix} \begin{matrix} \rightarrow \text{ Distanz zu } x_1^{(0)} \\ \rightarrow \text{ Distanz zu } x_2^{(0)} \end{matrix}$$
$$\begin{matrix} \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ U_1 & U_2 & U_3 & U_4 & U_5 \end{matrix}$$

a) Geben Sie $\mathcal{C}^{(0)}$ an und berechnen Sie die neuen Schwerpunkte für jede Klasse.

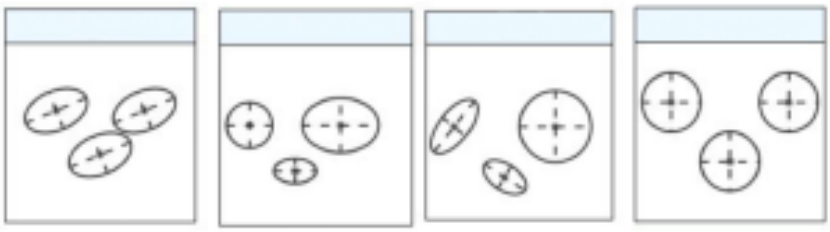
b) Führen Sie das k-means Clusterverfahren zu Ende.

c) Überlegen Sie sich, ob eine andere Startpartition zu einem anderen Ergebnis geführt hätte.

Aufgabe 3: modellbasiertes Clustering

a) Nennen Sie Unterschiede zwischen modellbasiertem Clustering, hierarchischem Clustering und k-means Clustering.

b) i) Lesen Sie sich die Hilfe zu der Funktion `mclustModelNames` aus dem Paket `mclust` durch und ordnen Sie den unten stehenden Abbildungen jeweils ein Modell zu.



Ellipses of isodensity for 4 Gaussian models obtained by eigen-decomposition (Scrucca et al., 2016)

ii) Beschreiben Sie anhand der Kovarianzmatrix, wie Distribution, Volume, Shape und Orientation ausgeprägt sind.

c) Laden Sie den Datensatz `wine` aus dem `gclus` Paket. Der Datensatz enthält 13 Messungen einer chemischen Analyse von 178 italienischen Weinsorten aus drei verschiedenen Kultursorten (Barolo, Grignolino, Barbera).

i) Führen Sie ein modellbasiertes Clustering durch, das automatisch das BIC optimale Modell ausgibt und interpretieren Sie den R Output und die Abbildung.

ii) Beschreiben Sie den `summary` Output.

iii) Vergleichen Sie die Zuordnung aus dem modellbasierten Clustering mit der wahren Partition.

iv) Betrachten und beschreiben Sie den adjustierten Rand Index. Erläutern Sie den Rand Index in eigenen Worten.

Quellen:

Luca Scrucca, Michael Fop, T Brendan Murphy, and Adrian E Raftery. mclust 5: Clustering, classification and density estimation using gaussian finite mixture models. *The R journal*, 8(1):289, 2016.