h_da
HOCHSCHULE DARMSTADT
UNIVERSITY OF APPLIED SCIENCES
fbmn
FACHBEREICH MATHEMATIK
UND NATURWISSENSCHAFTEN

Antje Jahn
Multivariate Statistik
Sommersemester 2019

Arbeitsblatt 10

A 1 Für 4 Beobachtungen berechne sich die folgende Distanzmatrix

$$\begin{pmatrix} 0 & 0.3 & 0.4 & 0.7 \\ 0.3 & 0 & 0.5 & 0.8 \\ 0.4 & 0.5 & 0 & 0.45 \\ 0.7 & 0.8 & 0.45 & 0 \end{pmatrix}.$$

Zeichnen Sie jeweils ein Dendrogramm, das aus einer hierarchischen Clusteranalyse mit Complete- und Single-linkage-Verfahren resultieren würde. Welche Beobachtungen würden in einer 2-Cluster-Lösung jeweils zu einem Cluster zusammengefasst werden?

A 2 Im Datensatz Ex11.csv finden Sie für 40 Probanden Genexpressionsdaten zu jeweils 1000 Genen (Zeilen: Gene; Spalten: Probanden). Laden Sie die Daten in R (read.csv mit Option header=FALSE) und standardisieren Sie zunächst die Genexpressionsvariablen.

a) Clustern Sie die 40 Probanden mittels hierarchischer Clusteranalyse und euklidischem Distanzmaß (Funktion hclust). Verwenden Sie verschiedene Linkage-Verfahren und betrachten Sie die jeweiligen Dendrogramme (Methode `plot` der Klasse `hclust`). Betrachten Sie auch jeweils die Korrelation der euklidischen Distanzen mit den Distanzen aus dem Dendrogramm (R-Funktion `cophenetic`).
b) Für welche Anzahl an Clustern würden Sie sich jeweils entscheiden?
c) Teilen Sie für jedes Linkage-Verfahren die Probanden mithilfe der Clusteranalyse in zwei Cluster. Vergleichen Sie die Clusterzugehörigkeiten zwischen den verschiedenen Linkage-Verfahren (Kreuztabellen).
d) Die ersten 20 Beobachtungen stammen von gesunden, die letzten 20 Beobachtungen von erkrankten Probanden. Wie gut separieren die verschiedenen Clustermethoden diese beiden Gruppen? Erstellen Sie dazu eine Kreuztabelle für die Clusterzugehörigkeit vs Zustand (gesund/krank)

A 3

a) Zeigen Sie, dass das Single-Linkage Verfahren monoton und kontrahierend ist.
b) Finden Sie ein Zahlenbeispiel, in dem die Verschmelzungsniveaus des Zentroid-Verfahrens nicht monoton sind.

A 4 Ihnen liegen Daten von 6 Kunden bzgl. ihres Kaufverhaltens zu Fleisch, Gemüse und Fertigprodukten vor. In nachfolgender Tabelle sind die wöchentlichen durchschnittlichen Ausgaben in EUR dieser 6 Kunden angegeben.

Kunde	Gemüse	Fleisch	Fertigprodukte
A	3	6	13
B	2.5	4	10
C	8	14	22
D	0.5	10	8
C	3.5	24	20
D	3	20	14

a) Berechnen Sie die Distanzmatrix bzgl. der euklidischen Distanz.

b) Standardisieren Sie die Ausgaben pro Kunde (d.h. nicht die Standardisierung der drei Variablen, die die Ausgaben pro Produkt angeben, sondern Standardisierung der 6 Datenzeilen, die die Ausgaben pro Kunde angeben). Berechnen Sie dann die Distanzmatrix für die so standardisierten Werte bzgl. der euklidischen Distanz.

c) Berechnen Sie die paarweisen Korrelationen zwischen den Kunden bzgl. Ihres Kaufsverhaltens. Vergleichen Sie diese mit den Distanzen zwischen den Kunden nach Standardisierung. Was fällt Ihnen auf?

d) Erstellen Sie einen Linienplot mit einer Linie pro Kunde, die die Ausgaben des Kunden für die drei Produkte visualisiert

e) Führen Sie eine hierarchische Clusteranalyse für die Original-Daten und die standardisierten Daten durch. Interpretieren Sie den Unterschied.

A 5 In dieser Aufgabe sollen Sie Ihre Beobachtungen aus Aufgabe 4 analytisch belegen: Zeigen Sie, dass für die in Aufgabe 4 berechnete Korrelation zwischen zwei Kunden, ρ, und die euklidische Distanz zwischen den beiden Kunden bzgl. der standardisierten Werte, d, gilt

$$d^2 = 2 \cdot n \cdot (1 - \rho).$$

Hinweis: Sie können der Einfachheit halber davon ausgehen, dass bei der Standardisierung der Varianzschätzer $\frac{1}{n}\sum(x_i - \bar{x})^2$ gewählt wurde (bei Nutzung von $\frac{1}{n-1}\sum(x_i - \bar{x})^2$ gilt Prportionalität statt Gleichheit).