3.Tutorium Multivariate Verfahren

- Clusteranalyse -

Nicole Schüller:
30.05.2016 und 06.06.2016

Hannah Busen:
31.05.2016 und 07.06.2016

Institut für Statistik, LMU München

Gliederung

Gliederung

Problemstellung

- Einteilung einer Menge von n Individuen $\{a_1, \ldots, a_n\}$ in Teilmengen, sogenannte **Cluster**!
- Die Einteilung soll so erfolgen, dass
  - sich die Individuen innerhalb eines Clusters möglichst ähnlich sind (Homogenität innerhalb der Cluster)
  - sich die Cluster untereinander möglichst unterscheiden (Heterogenität über die Cluster hinweg)
- **Beachte:** Die Klassen/Gruppen sind vorab nicht bekannt und werden gesucht! (im Gegensatz zur Diskriminanzanalyse)

Datensituation

- Gegeben sind n Individuen mit zugehörigen Merkmalsvektoren $\boldsymbol{x}_i$, $i = 1, \ldots, n$
- $\mathfrak{C} = \{C_1, \ldots, C_k\}$ sei eine Partition der Individuen in k Cluster
- Gesucht ist eine disjunkte Zerlegung $\{C_1, \ldots, C_k\}$ mit folgenden Eigenschaften:
  a) $\bigcup_{i=1}^{k} C_i = \{a_1, \ldots, a_n\}$
  b) $C_i \cap C_j = \emptyset \; \forall i \neq j$

Gliederung

Distanz zwischen den Individuen

- Für Distanzmaße $d : \Omega \times \Omega \to \mathbb{R}$ postuliert man:
  - $d(a_i, a_j) = d(a_j, a_i)$
  - $d(a_i, a_i) = 0$
  - $d(a_i, a_j) \geq 0 \; \forall i, j$
  - $d(a_i, a_j) \leq d(a_i, a_r) + d(a_r, a_j)$ (Dreiecksungleichung)

- **Typische Distanzmaße:**
  - $d_q(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sqrt[q]{\sum_{l=1}^{p}(x_{il} - x_{jl})^q}$ (L_q-Metrik)
  - $d_2(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sqrt{\sum_{l=1}^{p}(x_{il} - x_{jl})^2}$ (euklidische Distanz)
  - $d_1(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_{l=1}^{p} |x_{il} - x_{jl}|$ (Manhattan-Metrik)

Gliederung

Grundstruktur

- Konstruktion einer Hierarchie von Partitionen $\mathfrak{C} = \{C_1, \ldots, C_k\}$. Die Anzahl der Cluster variiert dabei von 1 bis zur Anzahl der Individuen!
- **Agglomerative** Verfahren: zu Beginn bildet jedes Individuum einen eigenen Cluster
- **Divisive** Verfahren: zu Beginn ein großer Cluster, der alle Individuen enthält
- **Merke:** Die Hierarchie enthält das Klassifikationsergebnis für jede mögliche Anzahl an Clustern (beliebig wählbar)
- Hierarchischen Klassifikationen erfordern Definition der
  - Distanz zwischen Individuen (vgl. Metriken Folie 7)
  - Distanz zwischen Clustern $\rightarrow$ **Linkage**

Linkage-Methoden

- C_r, C_s: Cluster
- **Single Linkage:** $D(C_r, C_s) = \min_{a_i \in C_r, a_j \in C_s} d(a_i, a_j)$
- **Complete Linkage:** $D(C_r, C_s) = \max_{a_i \in C_r, a_j \in C_s} d(a_i, a_j)$
- **Average Linkage:** $D(C_r, C_s) = \frac{1}{|C_r||C_s|} \sum_{a_i \in C_r, a_j \in C_s} d(a_i, a_j)$
- **Zentroid-Verfahren:** $D(C_r, C_s) = \|\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s\|^2$
- **Ward:** $D(C_r, C_s) = \frac{|C_r||C_s|}{|C_r|+|C_s|} \|\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s\|^2$,
  $\bar{\mathbf{x}}_r, \bar{\mathbf{x}}_s$: Mittelwert der Individuen in Cluster C_r, C_s

Agglomerative Verfahren

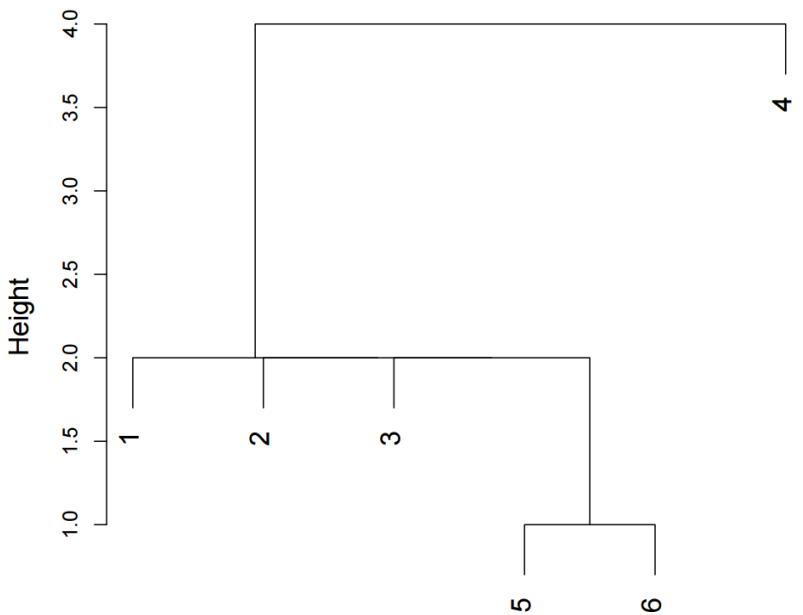
- **Algorithmus:**
  1. **Start:** Feinste Partition: Jedes Individuum bildet einen eigenen Cluster: $\mathfrak{C} = \{\{a_1\}, \ldots, \{a_n\}\}$
  2. Vereinige im ν-ten Schritt zwei Cluster C_r, C_s mit dem kleinsten Abstand und berechne den Homogenitätsindex: $h_\nu = \min\limits_{r \neq s} D(C_r, C_s)$.
  3. Wiederhole 2. bis ein großer Cluster entsteht, der alle Individuen enthält: $\mathfrak{C} = \{a_1, \ldots, a_n\}$
- Die Distanzen der Individuen werden durch die festgelegte *Metrik*, die Distanzen der Cluster durch die festgelegte *Linkage-Methode* bestimmt!

Divisive Verfahren

- **Algorithmus:**
  1. **Start:** Gröbste Partition: Ein großer Cluster, der alle Individuen enthält: $\mathfrak{C} = \{a_1, \ldots, a_n\}$.
  2. Sukzessive Unterteilung der Partition bis die feinste Partition entsteht: $\mathfrak{C} = \{\{a_1\}, \ldots, \{a_n\}\}$
- **Beachte:** $2^n - 1$ Möglichkeiten zu splitten
  $\Rightarrow$ Für großes n ist agglomerativ einfacher!

Dendrogramm

- Graphische Darstellung einer hierarchischen Clusterung
- y–Achse: Homogenitätsmaß h
  $\rightarrow$ je kleiner h, desto homogener ist der Cluster!
- Sukzessives Zusammenfassen bzw. sukzessive Aufteilung im Plot erkennbar.



Gliederung

Grundprinzip

- **Ziel:** Finde die Partition $\mathfrak{C} = \{C_1, \ldots, C_k\}$ bestehend aus k Clustern, die bezüglich eines Gütekriteriums optimal ist!
- Man betrachtet für jede Partition ein **Optimalitätskriterium**, das die Heterogenität erfasst:

$$H(\mathfrak{C}_{opt}) = \min_{\mathfrak{C}} H(\mathfrak{C})$$

- **Austauschverfahren:**
  1. Wähle eine zufällige Ausgangspartionen aus k Individuen
  2. Prüfe in der Partition $\mathfrak{C}^{alt}$, ob die Zuordnung jeweils eines Individuums in einen anderen Cluster, das betrachtete Optimalitätskriterium minimiert.
- **Beachte:**
  - Die Anzahl der Cluster k muss vom Anwender fest vorgegeben werden!
  - Je nach gewählter Ausgangspartition, können unterschiedliche Cluster entstehen.

Optimalitätskriterien

1. Varianzkriterium

$$H(\mathfrak{C}) = \sum_{r=1}^{k} \sum_{x_i \in C_r} \|\mathbf{x}_i - \bar{\mathbf{x}}_r\|^2 \quad \hat{=} \text{„k-means clustering“}$$

2. Determinantenkriterium

$$H(\mathfrak{C}) = |\mathbf{W}(\mathfrak{C})| \underset{\mathfrak{C}}{\rightarrow} \min$$

3. Verallgemeinertes Determinantenkriterium

$$H(\mathfrak{C}) = \sum_{r=1}^{k} n_r \log\left(\frac{1}{n_r}\mathbf{W}(\mathfrak{C}_r)\right) \underset{\mathfrak{C}}{\rightarrow} \min$$

Merke: Die Determinante entspricht einer verallgemeinerten Varianzmatrix!

k-means clustering

Vorgehen:

1. Wähle zufällig k Individuen bzw. die zugehörigen Merkmalsvektoren $\boldsymbol{x}$ als Clusterschwerpunkte $Z_r, r = 1, \ldots, k$. Die Clusteranzahl k ist fest!
2. Ordne jedes Individuum dem Clusterzentrum zu, zu dem die geringste Distanz $d_r, r = 1, \ldots, k$ besteht.
3. Berechne neue Clusterschwerpunkte $Z_r, r = 1, \ldots, k$ als Mittelwertsvektoren der Merkmalsvektoren der Individuen im Cluster!
4. Wiederhole 2. und 3. bis zur Konvergenz.