3. Tutorium Multivariate Verfahren
- Clusteranalyse -

Cornelia Gruber
26.05.2020

Institut für Statistik, LMU München

Gliederung

1. Idee der Clusteranalyse
2. Distanzmaße
3. Hierarchische Klassifikationsverfahren
4. Nichthierarchische Verfahren

Gliederung

1. Idee der Clusteranalyse
2. Distanzmaße
3. Hierarchische Klassifikationsverfahren
4. Nichthierarchische Verfahren

Problemstellung

- Einteilung einer Menge von n Beobachtungen $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ in Teilmengen, sogenannte **Cluster**!
- Die Einteilung soll so erfolgen, dass sich
  - die Beobachtungen innerhalb eines Clusters möglichst ähnlich sind
    $\rightarrow$ Homogenität innerhalb eines Cluster
  - die Cluster untereinander möglichst stark unterscheiden
    $\rightarrow$ Heterogenität zwischen verschiedenen Cluster
- **Beachte:** Die Klassen/Gruppen sind vorab nicht bekannt und werden gesucht! (im Gegensatz zur Diskriminanzanalyse)
  $\rightarrow$ unsupervised learning

Datensituation

- Gegeben sind n Beobachtungen mit zugehörigen Merkmalsvektoren $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$, $i = 1, \ldots, n$
- $\mathfrak{C} = \{C_1, \ldots, C_k\}$ sei eine Partition der Beobachtungen in k Cluster
- Gesucht ist eine disjunkte Zerlegung $\{C_1, \ldots, C_k\}$ mit folgenden Eigenschaften:
  a) $\bigcup_{i=1}^{k} C_i = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\}$
  b) $C_i \cap C_j = \emptyset \; \forall i \neq j$

Gliederung

1 Idee der Clusteranalyse

2 Distanzmaße

3 Hierarchische Klassifikationsverfahren

4 Nichthierarchische Verfahren

Distanz zwischen den Beobachtungen

- Für Distanzmaße $d : \Omega \times \Omega \to \mathbb{R}$ gilt:
  - $d(x_i, x_j) = d(x_j, x_i)$ (Symmetrie)
  - $d(x_i, x_i) = 0$
  - $d(x_i, x_j) \geq 0, \forall i, j$
  - $d(x_i, x_j) \leq d(x_i, x_r) + d(x_r, x_j)$ (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

1 Idee der Clusteranalyse

2 Distanzmaße

3 Hierarchische Klassifikationsverfahren

4 Nichthierarchische Verfahren

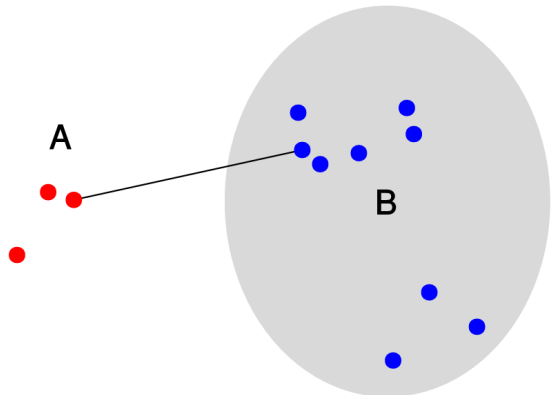
Grundstruktur

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

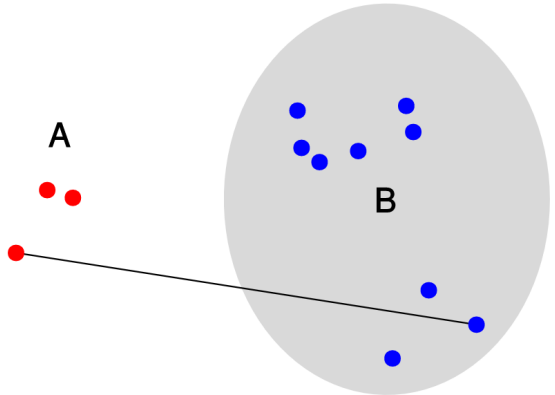
Linkage-Methoden

- C_r, C_s: Cluster
- **Single Linkage:** $D(C_r, C_s) = \min_{x_i \in C_r, x_j \in C_s} d(x_i, x_j)$
- **Complete Linkage:** $D(C_r, C_s) = \max_{x_i \in C_r, x_j \in C_s} d(x_i, x_j)$
- **Average Linkage:** $D(C_r, C_s) = \frac{1}{|C_r||C_s|} \sum_{x_i \in C_r, x_j \in C_s} d(x_i, x_j)$
- **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 Beobachtungen in Cluster C_r, C_s

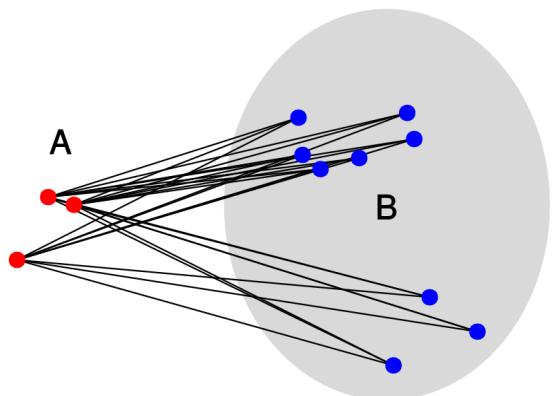
Single Linkage



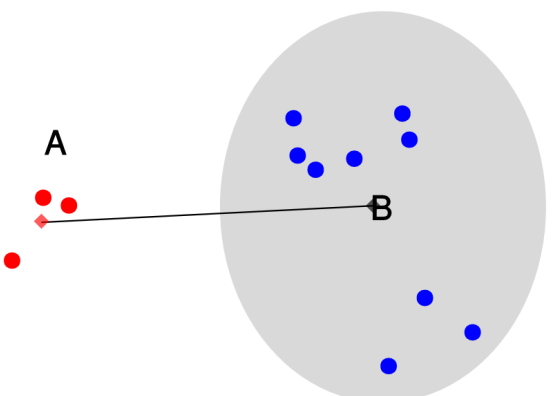
Complete Linkage



Average Linkage



Centroid Linkage



Agglomerative Verfahren

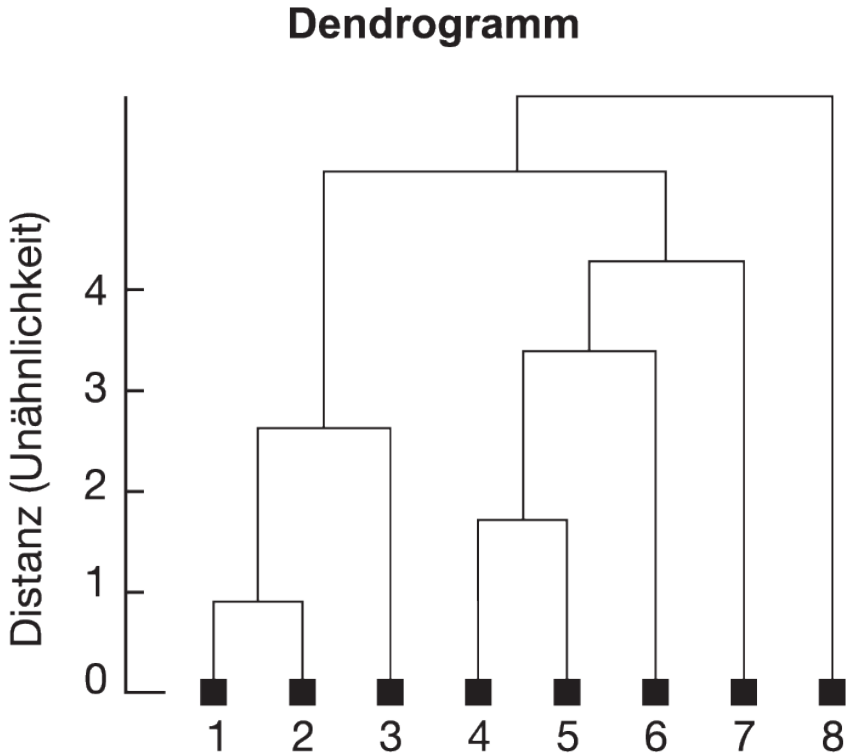
- **Algorithmus:**
  1. **Start:** Feinste Partition: Jede Beobachtung bildet ein eigenes Cluster: $\mathfrak{C} = \{\{x_1\}, \ldots, \{x_n\}\}$
  2. Vereinige im ν-ten Schritt zwei Cluster C_r, C_s mit dem kleinsten Abstand und berechne den Homogenitätsindex: $h_\nu = \min_{r \neq s} D(C_r, C_s)$.
  3. Wiederhole 2. bis ein großes Cluster entsteht, das alle Beobachtungen enthält: $\mathfrak{C} = \{x_1, \ldots, x_n\}$
- Die Distanzen der Beobachtungen werden durch die festgelegte *Metrik*, die Distanzen der Cluster durch die festgelegte *Linkage-Methode* bestimmt!

Divisive Verfahren

- **Algorithmus:**
  1. Bestimmte Cluster mit dem größten Durchmesser (Durchmesser = größte Distanz zwischen zwei Objekten im Cluster)
  2. Splitte zunächst das Objekt vom Cluster ab, das zu den anderen Objekten im Cluster die größte mittlere Unähnlichkeit aufweist. Dieses Objekt initialisiert die "Splittergruppe"
  3. Ordne Objekte aus dem geteilten Cluster der Splittergruppe zu, wenn sie dieser ähnlicher sind als dem "alten" Cluster
- **Beachte:** Pro Schritt gibt es $O(2^n)$ Möglichkeiten zu splitten $\Rightarrow$ Für großes n ist agglomerativ einfacher ($O(n^2)$ Möglichkeiten zum Zusammenfassen)!

Dendrogramm

- Graphische Darstellung eines hierarchischen Clusterings
- y–Achse: Homogenitätsmaß h
  $\rightarrow$ je kleiner h, desto homogener ist der Cluster!
- Sukzessives Zusammenfassen bzw. Aufteilen erkennbar



Gliederung

1 Idee der Clusteranalyse

2 Distanzmaße

3 Hierarchische Klassifikationsverfahren

4 Nichthierarchische Verfahren

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 Ausgangspartion aus k Beobachtungen
  2. Prüfe in der Partition $\mathfrak{C}^{alt}$, ob die Zuordnung jeweils einer Beobachtung 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 (= k-means clustering)

$$H(\mathfrak{C}) = \sum_{r=1}^{k} \sum_{x_i \in C_r} \|\mathbf{x}_i - \bar{\mathbf{x}}_r\|^2$$

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 Beobachtungen bzw. die zugehörigen Merkmalsvektoren $\boldsymbol{x}$ als Clusterschwerpunkte $Z_r, r = 1, \ldots, k$. Die Clusteranzahl k ist fest!
2. Ordne jede Beobachtung 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 Beobachtungen im Cluster!
4. Wiederhole 2. und 3. bis zur Konvergenz.

siehe Animation:
`https://de.wikipedia.org/wiki/K-Means-Algorithmus#/media/Datei:K-means_convergence.gif`