Statistik IV für Nebenfachstudierende

6. Hierarchische Clusterverfahren

Prof. Dr. Andreas Mayr

Institut für Statistik, LMU München

Sommersemester 2017

Motivation

- *Problem*: n Objekte sollen in Klassen eingeteilt werden, so dass:
  - innerhalb der Klassen große Homogenität
  - zwischen den Klassen große Heterogenität

  vorliegt.
- Unterschied zur Diskriminanzanalyse: Keine vorgegebenen Klassen.
- Deswegen auch: *unsupervised-learning*.

Beispiel:

- Ausgangslage:

$$a_1 \ , \ \dots \ , \ a_n$$

$$\underline{x}_1 \ , \ \dots \ , \ \underline{x}_n$$

- Gesucht:

$$C_1 \ , \ \dots \ , \ C_g$$

$$\cup_{j=1}^{g} C_j = \{a_1 \ , \ \dots \ , \ a_n\}$$

  - Disjunkte Cluster
  - Überlappende Cluster
  - Fuzzy Cluster

Übersicht

1. Ähnlichkeit
2. Distanzmaße
3. Hierarchische Clusterverfahren
4. Optimale Partition
5. Mischverteilungsansätze
6. Stochastische Partition

Ähnlichkeit bzw. Distanzmaße

- Grundlage für Homogenität / Heterogenität
- *Ähnlichkeitsmaß* hat die Form:

$$s: \quad \Omega \times \Omega \longrightarrow \mathbb{R}^+$$

$$(a_i, a_j) \longrightarrow s(a_i, a_j)$$

mit

$$s(a_i, a_j) = s(a_j, a_i)$$

$$s(a_i, a_i) \geq s(a_j, a_i)$$

- *Distanzsmaß* hat die Form:

$$d: \quad \Omega \times \Omega \longrightarrow \mathbb{R}^+$$

$$(a_i, a_j) \longrightarrow d(a_i, a_j)$$

mit

$$\begin{aligned} d(a_i, a_i) &= 0 \\ d(a_i, a_j) &\geq 0 \\ d(a_i, a_j) &= d(a_j, a_i) \\ d(a_i, a_j) &\leq d(a_i, a_k) + d(a_k, a_j) \end{aligned}$$

Distanzmaße

- *Metrische Merkmale*

$$\underline{x}_i = (x_{i1}, \ldots, x_{ip}) \quad , \quad \underline{y}_i = (y_{i1}, \ldots, y_{ip})$$

L_q– Metrik

$$d(\underline{x}_i, \underline{y}_i) = \left(\sum_{k=1}^{p} \mid x_{ik} - y_{ik} \mid^q \right)^{1/q}$$

$q = 1$ *City–Block–Metrik*

$q = 2$ *Euklidische Metrik*

- $q = 2$ Euklidische Metrik (Distanz):

- $q = 1$ City–Block–Metrik (Distanz)

Einheitskreis für L_1 und L_2:

- *Binäre Merkmale*

$$\underline{x}_i^\top = (x_{i1}, \ldots, x_{ip}) \quad \text{mit} \quad x_{ik} \in \{0, 1\}$$

$h_{ij}(y, z)$ *bezeichnet die Anzahl der übereinstimmenden Komponenten in* $\underline{x}_i$ *und* $\underline{x}_j$ *mit Ausprägung* y *in* $\underline{x}_i$ *und* z *in* $\underline{x}_j$*.*

$h_{ij}(y, z) = \mid \{k \mid (x_{ik}, x_{jk}) = (y, z)\} \mid$ *für* $(y, z) \in \{(0,0)\ ,\ (0,1)\ ,\ (1,0)\ ,\ (1,1)\}$.

Matching–Koeffizient

$$s_{ij} = \frac{h_{ij}(1,1) + h_{ij}(0,0)}{p}$$

- *Mehrkategorielle Variablen*

$$\underline{x}_i^\top = (x_{i1}, \ldots, x_{ip}) \text{ mit } x_{ij} \in \{1, ..., k\}$$

Genauso wie binäre Merkmale, nur mit Dummy Variablen:

$$x_{i,j}^{(r)} = \begin{cases} 1 & x_{i,j} = r \\ 0 & x_{i,j} \neq r \end{cases}$$

Hierarchische Clusterverfahren

- **Agglomeratives Clustern:**

  Start mit Partition

  $$C_1 = \{a_1\}, ..., C_n = \{a_n\}$$

  $\rightarrow$ sukzessives Zusammenfassen

- **Divisives Clustern:**

  Starte mit einem Cluster

  $$C = \{a_1, ..., a_n\}$$

  $\rightarrow$ sukzessives Aufteilen

- Prinzip agglomerative Verfahren:
  - Ausgangsposition:

$$\mathcal{E}^{[0]} = \{\{a_1\}, ..., \{a_n\}\} = \{C_1^{[0]}, ..., C_n^{[0]}\}$$

  - Aus Partition

$$\mathcal{E}^{[i-1]} = \{C_1^{[i-1]}, ..., C_n^{[i-1]}\}$$

    gewinnt man $\mathcal{E}^{[i]}$ durch Vereinigung der beiden Cluster für die das Heterogenitätsmaß minimal ist:

$$D(C_r^{[i-1]}, C_s^{[i-1]})$$

  - Wir brauchen ein Distanzmaß $D(C_r, C_s)$ zwischen Clustern.

Single Linkage

$$D(C_r, C_s) = \min_{\substack{a_i \in C_r \\ a_j \in C_s}} d(a_i, a_j)$$

Beispiel

Complete Linkage

$$D(C_r, C_s) = \max_{\substack{a_i \in C_r \\ a_j \in C_s}} d(a_i, a_j)$$

Average Linkage

$$
\begin{aligned}
D(C_r, C_s) &= \frac{1}{n_r n_s} \sum_{a_i \in C_r} \sum_{a_j \in C_s} d(a_i, a_j) \\
\text{mit } n_i &= \mid C_i \mid
\end{aligned}
$$

Zentroid Verfahren

$$D(C_r, C_s) \quad = \quad \| \bar{\underline{x}}_r - \bar{\underline{x}}_s \|^2 \qquad \text{mit } \bar{\underline{x}}_i = \frac{1}{n_i} \sum_{\underline{x}_j \in C_i} \underline{x}_j$$

Optimale Partitionierung (nicht-hierarchisch)

- Grundidee: Berechne alle möglichen Partitionierungen und wähle die beste.
- Im Allgemeinen kein hierarchisches Vorgehen.
- Formal:
  - $H(\mathcal{E})$ ist Heterogenität der Partition.
  - Gesucht wird optimale Partitionierung $\mathcal{E}_{\text{opt}}$.

$$\mathcal{E}_{\text{opt}} = \min_{\mathcal{E}} H(\mathcal{E})$$

- Wie findet man ε_{opt}?
- Ansatz: *Try all, keep best*
- Hauptproblem: Dimensionalität
  - Schon bei $n = 10$ und 3 Clustern gibt es 9330 Möglichkeiten.
  - Bei $n = 100$ und 3 Clustern $\approx 6 \cdot 10^{29}$