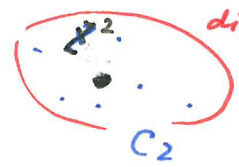


Clustering I

$x \in S \subseteq \mathbb{R}^p$ ← Anzahl von Merkmalen
 $n = |S|$ ← Anzahl von Daten
 Daten, z.B. Merkmale von Objekten

Cluster

Klassenzuordnung
 $S = C_1 \cup \dots \cup C_k$
 disjunkt



Repräsentant der l -ten Klasse ("Zentrum")
 $z^1, \dots, z^k$

$$\min_{z^e, l=1, \dots, k} \sum_{l=1}^k \sum_{x \in C_l} d(x, z^l)$$

$C_l, l=1, \dots, k$

Klassenpartition
 $S = C_1 \cup \dots \cup C_k$

Dissimilarity / Unähnlichkeit
 zwischen x und z^e

"Wähle Klassenpartition von S so, daß die Unähnlichkeit bzgl. der Klassenzentren minimal wird"

1) Objektclustering

- $d(x, z) = \|x - z\|^2$ - Euklidische Norm = $\sum_{j=1}^p (x_j - z_j)^2$
- $d(x, z) = \sum_{j=1}^p |x_j - z_j|$ - Summennorm
- $d(x, z) = \max_{1 \leq j \leq p} |x_j - z_j|$ - Maximumnorm

2) Dokumentenclustering

$TF(t, D) = \frac{\#(t, D)}{\max_{t'} \#(t', D)}$, wobei $\#(t, D)$ = "Vorkommenshäufigkeit von t in D "

↑ term ↑ document

term frequency

$IDF(t) = \ln \frac{|S|}{\sum_{D \ni t} 1}$

↑ inverse document frequency

Anzahl des Dokumente, die t enthalten

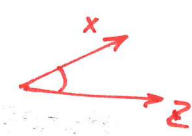
$TF-IDF(t, D) = TF(t, D) \cdot IDF(t)$ - Gewicht von t in D

$x = (TF-IDF(t, D) \mid t - \text{Terme}) \in \mathbb{R}^p$ ← Anzahl des Terme

Daten für Dokument D .

$$S(x, z) = \cos \angle(x, z) = \frac{x^T \cdot z}{\|x\| \cdot \|z\|} = \frac{\sum_{j=1}^P x_j \cdot z_j}{\sqrt{\sum_{j=1}^P x_j^2} \cdot \sqrt{\sum_{j=1}^P z_j^2}} \in [-1, 1]$$

↑
Cosine similarity



↑
"nicht ähnlich"
↑
"ähnlich"

↑
L = π
↑
L = 0

$$d(x, z) := \frac{1 - S(x, z)}{2} \quad \text{Dissimilarity in Text Mining}$$

(analog: Bewertungen von Nutzern vergleichen)

3) Warenclustering

$x \in \{0, 1\}^P$ ← Anzahl von Kunden (Binäre Daten)

nicht gekauft
↑
gekauft

← Gekauft gemeinsam

$$J(x, z) = \frac{\#(x_i = z_i = 1)}{\#(x_i \neq 0 \text{ oder } z_i \neq 0)} \quad \text{Jaccard-Koeffizient}$$

← Gekauft = $\#(x_i = 1 \text{ oder } z_i = 1)$

$$d(x, z) = 1 - J(x, z) = \frac{\#(x_i = 1 \text{ und } z_i = 0) + \#(x_i = 0 \text{ und } z_i = 1)}{\#(x_i = 1 \text{ oder } z_i = 1)}$$

↑ Jaccard-Metrik

4) Zeilenclustering

$x, z \in A$
Zeilen Alphabet

(i) Einfügen: $\varepsilon \rightarrow x$

(ii) Entfernen: $x \rightarrow \varepsilon$

(iii) Ersetzen: $x \rightarrow y \quad (x \neq y)$

$d(x, z)$ - minimale Anzahl von Operationen, um aus z die Zeile x zu bekommen
Levenshtein-Metrik

z. B. DNA = {Cytosin, Guanin, Adenin, Thymin} - Sequenz

$$x = \text{ACCGAT} \Rightarrow d(x, z) = 4$$

$$z = \text{AGCAT}$$

$$x = \text{ACCGAT} \Rightarrow d(x, z) = 2$$

$$z = \text{AGCAT}$$

5) Zeitreihenclustering

$x, z \in \mathbb{R}^P$ ← beobachtbare Zeitpunkte

Mittelwerte

$$\rho(x, z) = \frac{\text{cov}(x, z)}{\sigma(x) \cdot \sigma(y)} = \frac{\sum_{j=1}^P (x_j - \bar{x}) \cdot (z_j - \bar{z})}{\sqrt{\sum_{j=1}^P (x_j - \bar{x})^2} \cdot \sqrt{\sum_{j=1}^P (z_j - \bar{z})^2}}$$

Korrelationskoeffizient
↑
negativ korreliert
↑
positiv korreliert

$$d(x, z) = \frac{1 - \rho(x, z)}{2} \in [0, 1]$$

↑ pos. Korrelation

↑ $d(x, z)$ ist innerhalb einer Klasse klein

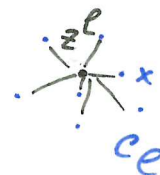
z. B. Aktienclustering über Renditenstatistik → Klassen mit positiv korrelierten Aktienrenditen (Risikodiversifikation)

K-means Algorithmus

$z^1, \dots, z^k$ - Zentren von $C_1, \dots, C_k$

① $C_\ell := \{x \in S \mid d(x, z^\ell) \leq d(x, z^{\ell'}) \forall \ell' = 1, \dots, k\}$
 $\ell = 1, \dots, k$ "Elemente mit kleinster Unähnlichkeit zu z_ℓ "

② $z^\ell \in \arg \min_z \sum_{x \in C_\ell} d(x, z)$ "Zentrum von C_ℓ "



Zeitschritt: Clustering/Centers
 Analysis: $t \rightarrow z^\ell(t), C^\ell(t)$
 $t+1 \rightarrow z^\ell(t+1), C^\ell(t+1)$

$$\underline{d(C(t+1), z(t+1))} = \sum_{\ell=1}^k \sum_{x \in C_\ell(t+1)} d(x, z^\ell(t+1)) \leq \sum_{\ell=1}^k \sum_{x \in C_\ell(t+1)} d(x, z^\ell(t))$$

$$\leq \sum_{\ell'=1}^k \sum_{x \in C_{\ell'}(t)} d(x, z^{\ell'}(t)) = \underline{d(C(t), z(t))}$$

② $\leq \sum_{x \in C_\ell(t+1)} d(x, z^\ell(t))$
 für $x \in C_{\ell'}(t)$ gilt:
 ① $d(x, z^\ell(t)) \leq d(x, z^{\ell'}(t))$
 ($z^\ell(t)$ in ① für $C_\ell(t+1)$)

$\Rightarrow d(C(t), z(t))$ ist monoton fallend bzgl. $t \Rightarrow$
 ↑ Dissimilarity

Konvergenz
 gegen lokale Minima
 von $d(C, z)$

z. B. $d(x, z) = \|x - z\|^2$ - Euklidische Norm

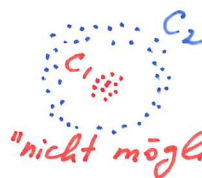
• Struktur von Clustern:

$$C_\ell := \{x \in S \mid \|x - z\|^2 \text{ ist minimal für } z^\ell\}$$

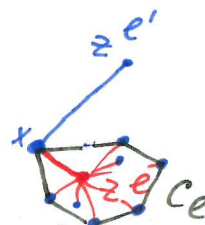
• Struktur von Zentren:

$$\min_z \sum_{x \in C_\ell} \|x - z\|^2$$

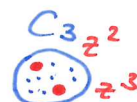
Optimalität: $\sum_{x \in C_\ell} 2 \cdot (x - z) = 0 \Rightarrow \sum_{x \in C_\ell} x = |C_\ell| \cdot z \Rightarrow z = \frac{1}{|C_\ell|} \sum_{x \in C_\ell} x$
 $\nabla_z(\dots) = 0$



Cluster - konvexe Menge



• Lokale Konvergenz:



1. "lokales Optimum"