Prof. Dr. Annika Hoyer, Marc Schneble, Cornelia Gruber

Lösung - Clusterverfahren

Aufgabe 1: Hierarchisches Clustering in R

Lösen Sie folgende Aufgaben in `R`. Verwenden Sie dabei den Datensatz `mtcars`.

a) Warum ist es sinnvoll, die Variablen zu standardisieren (zentrieren und skalieren), oder zu normalisieren (Maximum abziehen und durch Spannweite/Range teilen) bevor man die euklidische Distanzmatrix berechnet?

Lösung Aufgabe 1a:

LQ-Normen sind nicht skaleninvariant, die Distanzen hängen also von der Maßeinheit des jeweiligen Merkmals ab. Vor der Berechnung der LQ-Distanzen sollte also auf gemeinsame Maßeinheit standardisiert bzw. normalisiert werden, damit Variablen mit größerer Spannweite keinen größeren Einfluss haben.

Standardisieren:

$$x^* = \frac{x - \bar{x}}{\sqrt{Var(x)}}$$

- wenn X normalverteilt ist: $X \sim N(\mu, \sigma^2)$
- nach Standardisierung: $X^* \sim N(0, 1)$

Normalisieren:

$$x^{**} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

- wenn X möglicherweise nicht normalverteilt ist
- behält Form der ursprünglichen Verteilung
- Vergleichbarkeit, da alle x^{**} im Intervall [0, 1] sind

---

b) Laden Sie das Paket `cluster` und machen Sie sich mit den Funktionen `agnes()` und `diana()` vertraut.

c) Berechnen Sie die euklidische Distanzmatrix für die standardisierten und nicht standardisierten Variablen.

d) Führen Sie ein agglomeratives Clustering mit `agnes()` durch. Verwenden Sie hierbei die Manhattan-Metrik und die euklidische Metrik für das Complete Linkage-Verfahren und vergleichen Sie die Dendrogramme für beide Metriken.

e) Führen Sie ein divisives Clustering mit `diana()` durch. Verwenden Sie hierbei die Manhattan-Metrik und die euklidische Metrik und vergleichen Sie beide Dendrogramme.

Aufgabe 2: k-means Clustering

Die folgende Tabelle enthält die Anzahl der Mitarbeiter und den Umsatz (in 10 000 Euro) für fünf verschiedene Unternehmen.

Unternehmen	Mitarbeiter	Umsatz (in 10 000 Euro)
U1	2	10
U2	2	5
U3	8	4
U4	5	8
U5	7	5

Führen Sie ein k-means Clustering durch, um die fünf Unternehmen in zwei verschiedene Cluster einzuordnen. Als Distanzmaß soll die quadrierte euklidische Distanz verwendet werden. Die beiden Unternehmen U1 und U4 sollen dabei die Startpartition für je ein Cluster sein. Man erhält für Iteration 0 folgende Distanzen:

$$x_1^{(0)} = (2, 10)^T, \quad x_2^{(0)} = (5, 8)^T$$

$$\mathbf{D}^{(0)} = \begin{pmatrix} 0 & 25 & 72 & 13 & 50 \\ 13 & 18 & 25 & 0 & 13 \end{pmatrix} \begin{matrix} \rightarrow \text{ Distanz zu } x_1^{(0)} \\ \rightarrow \text{ Distanz zu } x_2^{(0)} \end{matrix}$$
$$U1 \quad U2 \quad U3 \quad U4 \quad U5$$

a) Geben Sie $\mathcal{C}^{(0)}$ an und berechnen Sie die neuen Schwerpunkte für jede Klasse.

b) Führen Sie das k-means Clusterverfahren zu Ende.

c) Überlegen Sie sich, ob eine andere Startpartition zu einem anderen Ergebnis geführt hätte.

Lösung Aufgabe 2:

a) Ordne alle Unternehmen dem Cluster zu, zu dessen Clusterschwerpunkt sie die kleinere Distanz haben:

$$\mathcal{C}^{(0)} = \Big\{\{U1\}, \{U2, U3, U4, U5\}\Big\}.$$

Anhand dieser Clusterzuordnung werden neue Clusterschwerpunkte berechnet:

$$x_1^{(1)} = (2, 10)^T$$
$$x_2^{(1)} = \frac{1}{4}\big((2 + 8 + 5 + 7), (5 + 4 + 8 + 5)\big)^T = (5.5, 5.5)^T$$

b)

$$\mathbf{D}^{(1)} = \begin{pmatrix} 0 & 25 & 72 & 13 & 50 \\ 32.5 & 12.5 & 8.5 & 6.5 & 2.5 \end{pmatrix}$$

Die oberen Einträge der Distanzmatrix sind dieselben wie in Iteration 0. Die unteren Einträge ergeben sich durch

$$d(U1, x_2^{(1)}) = (2-5.5)^2 + (10-5.5)^2 = 12.25 + 20.25 = 32.5$$
$$d(U2, x_2^{(1)}) = (2-5.5)^2 + (5-5.5)^2 = 12.25 + 0.25 = 12.5$$
$$d(U3, x_2^{(1)}) = (8-5.5)^2 + (4-5.5)^2 = 6.25 + 2.25 = 0.5$$
$$d(U4, x_2^{(1)}) = (5-5.5)^2 + (8-5.5)^2 = 0.25 + 6.25 = 6.5$$
$$d(U5, x_2^{(1)}) = (7-5.5)^2 + (5-5.5)^2 = 2.25 + 0.25 = 2.5$$

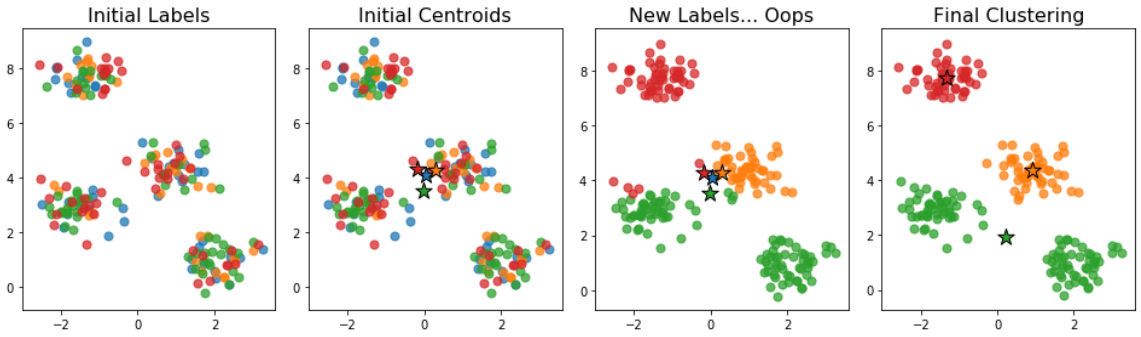
Als neue Clusterzuordnung erhält man daher

$$\mathcal{C}^{(1)} = \Big\{\{U1\}, \{U2, U3, U4, U5\}\Big\} = \mathcal{C}^{(0)}$$

$\Rightarrow$ Konvergenz!

c) Ja, andere Startpartitionen können zu anderen Ergebnissen führen! U1 beeinflusst bei dieser Initialisierung die Clusterwahl stark, vor allem wegen des hohen Umsatzes.

weiteres Beispiel:



Quelle: `http://www.salientiastuff.com/tag/k-means.html`

Kein Punkt ist dem blauen Zentroid am nächsten. In der Clusterzuordnung im nächsten Schritt wird dementsprechend kein Punkt diesem Cluster zugeteilt. Der Algorithmus konvergiert mit nur 3 Clustern, obwohl 4 Cluster erwünscht sind.

Aufgabe 3: modell-basiertes Clustering

a) Nennen Sie Unterschiede zwischen modellbasiertem Clustering, hierarchischem Clustering und k-means Clustering.

Lösung Aufgabe 3a:
Modellbasiert

- finde Verteilung, die Daten generiert hat → Inferenz möglich
- Wahrscheinlichkeitsverteilung (W'keit für jeden Punkt zu Cluster zu gehören)
- iterative Lösung durch EM

Hierarchisch

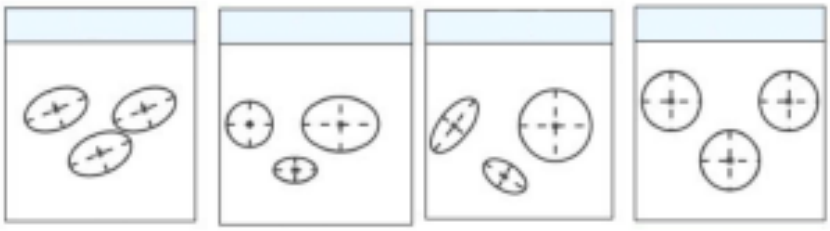
- heuristisch
- Kategoriale Variablen möglich
- Reihenfolge an Clusterlösungen → k nicht festgelegt
- Korrelation zwischen Attributen wird meist nicht berücksichtigt

k-Means

- heuristisch
- Kategoriale Variablen möglich
- k fest
- abhängig von Initialisierung der Startpunkte
- Korrelation zwischen Attributen wird meist nicht berücksichtigt

---

b) i) Lesen Sie sich die Hilfe zu der Funktion `mclustModelNames` aus dem Paket `mclust` durch und ordnen Sie den unten stehenden Abbildungen jeweils ein Modell zu.



Ellipses of isodensity for 4 Gaussian models obtained by eigen-decomposition in case of three groups of two dimensions (Scrucca et al., 2016)

ii) Beschreiben Sie anhand der Kovarianzmatrix, wie Distribution, Volume, Shape und Orientation ausgeprägt sind.

Lösung Aufgabe 3b:

i) 1. EEE: ellipsodial; equal shape, volume and orientation
2. VVI: diagonal; varying volume and shape
3. VVV: ellipsoidal; varying volume, shape and orientation
4. EII: spherical; equal volume

ii) Spherical: unabhängige Attribute und gleiche Varianz:

$$\begin{pmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_1^2 & 0 \\ 0 & 0 & \sigma_1^2 \end{pmatrix}$$

Diagonal: unabhängige Attribute, dürfen unterschiedliche Varianz haben:

$$\begin{pmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{pmatrix}$$

Ellipsoid: abhängige Attribute mit unterschiedlicher Varianz:

$$\begin{pmatrix} \sigma_1^2 & \sigma_2^2 & \sigma_3^2 \\ \sigma_4^2 & \sigma_5^2 & \sigma_6^2 \\ \sigma_7^2 & \sigma_8^2 & \sigma_9^2 \end{pmatrix}$$

$\Rightarrow$ Volume, Shape und Orientation lassen sich aus der Zerlegung der Kovarianzmatrix erklären. Jeder Cluster hat eine Kovarianzmatrix

$$\Sigma_k = \lambda_k D_k A_k D_k^T,$$

wobei gilt:

* λ_k gibt das Volumen an
* A_k ist eine Diagonalmatrix mit Determinante 1, die die Shape angibt
* D_k ist eine Orthogonalmatrix, die die Orientation angibt

---

c) Laden Sie den Datensatz `wine` aus dem `gclus` Paket. Der Datensatz enthält 13 Messungen einer chemischen Analyse von 178 italienischen Weinsorten aus drei verschiedenen Kultursorten (Barolo, Grignolino, Barbera).

i) Führen Sie ein modell-basiertes Clustering durch, das automatisch das BIC optimale Modell ausgibt und interpretieren Sie den R Output und die Abbildung.

ii) Beschreiben Sie den `summary` Output.

iii) Vergleichen Sie die Zuordnung aus dem modell-basierten Clustering mit der wahren Partition.

iv) Betrachten und beschreiben Sie den adjustierten Rand Index. Erläutern Sie den Rand Index in eigenen Worten.

Quellen:

Luca Scrucca, Michael Fop, T Brendan Murphy, and Adrian E Raftery. mclust 5: Clustering, classification and density estimation using gaussian finite mixture models. *The R journal*, 8(1):289, 2016.