Tutorium 3

Montag, 25. Mai 2020 15:37

1 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?

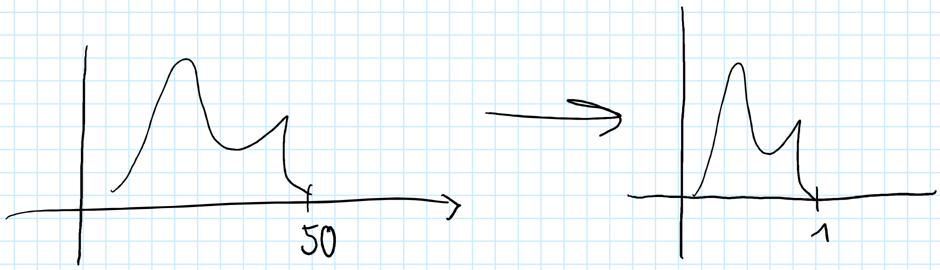
Standardisieren: $X^* = \frac{x - \bar{x}}{\sqrt{Var(x)}}$

→ danach $E(X^*) = 0$ & $Var(X^*) = 1$

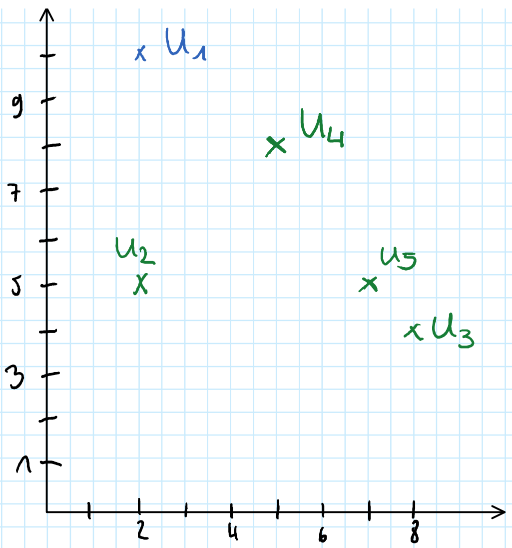
Normalisieren: $X^{**} = \frac{x - x_{min}}{x_{max} - x_{min}}$

→ behält Form der Verteilung

→ $X^{**} \in [0, 1]$



2a) a) Geben Sie $\mathcal{C}^{(0)}$ an und berechnen Sie die neuen Schwerpunkte für jede Klasse.



$C^0 = \{ \{U_1\}, \{U_4, U_2, U_3, U_5\} \}$

neue Schwerpunkte: $x_1^{(1)} = \begin{pmatrix} 2 \\ 10 \end{pmatrix}$, $x_2^{(1)} = \begin{pmatrix} \frac{2+8+5+7}{4} \\ \frac{5+4+8+5}{4} \end{pmatrix} = \begin{pmatrix} 5,5 \\ 5,5 \end{pmatrix}$

2b) b) Führen Sie das k-means Clusterverfahren zu Ende.

$$D^{(1)} = \begin{array}{c} \\ \\ \\ \end{array}\begin{array}{ccccc} \text{min} & & & & \\ \left(\begin{matrix} 0 & 25 & 72 & 13 & 50 \\ 32{,}5 & 12.5 & 8.5 & 6.5 & 2.5 \end{matrix}\right. & & & & \\ \uparrow u_1 & \uparrow u_2 & \uparrow u_3 & \uparrow u_4 & \uparrow u_5 \end{array}\left.\right) \begin{array}{l} \leftarrow x_1 \text{ Distanz zu Schwerpunkt 1} \\ \leftarrow x_2 \text{ Distanz zu Schwerpunkt 2} \end{array}$$

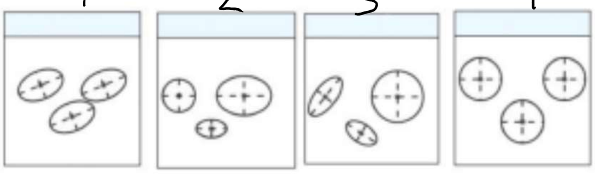
$$d(u_1, x_2^{(1)}) = (2 - 5{,}5)^2 + (10 - 5{,}5)^2 = 12.25 + 20.25 = 32{,}5$$

$$C^{(1)} = \{\{u_1\}, \{u_2, u_3, u_4, u_5\}\} = C^{(0)} \quad \rightarrow \text{Konvergenz}$$

neue Partition = alte Partition

3b) i) 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.

1 2 3 4



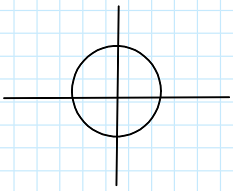
1: EEE ellipsoidal, equal volume, shape

2. VII diagonal, varying volume, shape

3. VVV ellipsoidal, varying volume, shape, & orientation

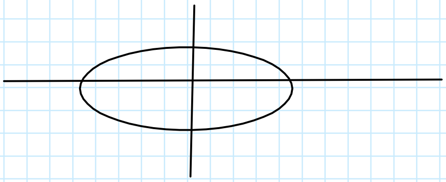
4. EII spherical, equal volume (→ ähnliche Ergebnisse wie k-means)

spherical: keine Korrelationen unabhängige Variablen
gleiche Varianzen der Variablen

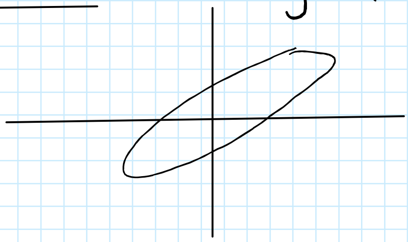


diagonal: Streckung in X & Y Richtung → unterschiedliche Varianzen
→ keine Korrelationen

→ keine Korrelationen
(↦ unabhängige Variablen)



ellipsoidal: Streckung in beliebige Richtung → Korrelationen möglich
→ unterschiedliche Varianzen



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

$$\left.\begin{array}{ll}
\text{spherical} & \Sigma_i = \begin{pmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_1^2 \end{pmatrix} \\[2em]
\text{diagonal} & \Sigma_i = \begin{pmatrix} \sigma_1^2 & & 0 \\ & \sigma_2^2 & \\ 0 & & \ddots \end{pmatrix} \\[2em]
\text{ellipsoidal} & \Sigma_i = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots \\ \sigma_{21} & \sigma_2^2 & \\ \vdots & & \ddots \end{pmatrix}
\end{array}\right\} \text{Kovarianz für } \underline{\text{ein}} \text{ Cluster}$$

$$\Sigma_k = \lambda_k D_k A_k D_k^T$$ → Zerlegung der Kovarianzmatrix

λ_k: Volumen

A_k: Diagonalmatrix mit Determinante 1 → shape

D_k: Orthogonalmatrix → orientation