Lernzettel

Unüberwachtes Lernen: Clustering, Dimensionsreduktion, Mustererkennung

Universität: Technische Universität Berlin
Kurs/Modul: Informationssysteme und Datenanalyse

Erstellungsdatum: September 19, 2025

Lernzettel: Unüberwachtes Lernen: Clustering, Dimensionsreduktion, Mustererkennung

(1) Grundidee und Ziel
Beim unüberwachten Lernen liegen die Daten ohne zugehörige Labels vor. Ziel ist es, Struktur in den Daten zu entdecken, z. B. Gruppen (Clustering), relevante Repräsentationen (Dimensionsreduktion) oder Muster/Regeln in den Daten (Mustererkennung).

(2) Distance- und Ähnlichkeitsmaße
Für Vektoren x, y gilt häufig:

$$d_E(x, y) = \|x - y\|$$

$$d_M(x, y) = \sum_j |x_j - y_j| \quad \text{(Manhattan-Distanz)}$$

$$d_{\text{Cos}}(x, y) = 1 - \frac{x \cdot y}{\|x\|\|y\|}$$

Distance-Maße beeinflussen Clustering signifikant.

(3) Clustering (Ziel und grobe Idee)
Ziel: Daten in Gruppen so zu ordnen, dass ähnliche Objekte zusammenliegen und Unterschiede zwischen Gruppen möglichst groß sind.

(3.1) K-Means
Ziel-Funktion (Gauß-ähnliche, flache Gruppen):

$$J = \sum_{i=1}^{n} \|x_i - \mu_{c(i)}\|^2$$

Dabei ist $\mu_{c(i)}$ der Zentroid der zuordnung von x_i zu Cluster c(i).

(3.2) Hierarchisches Clustering
- Agglomerativ: Start mit einzelnen Punkten, schrittweise Zusammenführen von Klasterpaaren.
- Verknüpfungskriterien: single-link, complete-link, average-link.
- Ergebnis: Dendrogramm, Schnittpunkt-Indikation für Cluster-Größen.

(3.3) Dichtedichte-basiertes Clustering
DBSCAN: Cluster entstehen dort, wo dichtebereiche sind; Randklassen als Rauschen.

(4) Dimensionsreduktion
Ziel: Repräsentationen in weniger Dimensionen erhalten, die wesentliche Struktur bewahren.

(4.1) PCA (Principal Component Analysis)
Zentralidee: Projektion der Daten auf die Top-k-Eigenrichtungen der Kovarianzmatrix.

$$C = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^\top$$

$$W = [w_1, \ldots, w_k] \quad \text{mit } Cw_j = \lambda_j w_j \text{ und } \lambda_1 \geq \lambda_2 \geq \ldots$$

$$z = W^\top (x - \bar{x})$$

$$\text{var}_{explained} = \frac{\lambda_j}{\sum_{m=1}^{d} \lambda_m}$$

(4.2) Nichtlineare Dimensionsreduktion (Hinweis)
Methoden wie t-SNE, UMAP erzeugen oft sinnvolle Visualisierungsebenen, nutzen aber komplexe Optimierungen. Formeln hier nur grob skizziert.

(5) Mustererkennung im unüberwachten Kontext
Ziel ist es, Muster, Strukturen oder Anomalien zu identifizieren, ohne vorgegebene Labels.

(5.1) Dichte-Schätzung
- Kernel-Dichte-Schätzung (KDE) schätzt eine Wahrscheinlichkeitsdichte als Summe von Kernel-Funktionen.
- Grundidee: Hohe Dichtebereiche bedeuten häufige Muster.

(5.2) Anomalie-Erkennung
Objekte mit niedriger Dichte oder Distanz zu Clustern gelten als Ausreißer.

(6) Evaluierung unüberwachter Modelle
- Es gibt keine klare "richtige" Label-Anordnung. Sinnvoll: interne Metriken.

(6.1) Silhouette-Koeffizient
Für jedes Objekt i:

$$a(i) = \text{durchschnittliche Abstand zu allen anderen Objekten im gleichen Cluster}$$

$$b(i) = \min_{\text{Clusters } C \neq C_i} \text{durchschnittlicher Abstand zu Objekten in } C$$

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

Der Durchschnitt über alle s(i) gibt eine Gesamteinschätzung.

(7) Vorverarbeitung und praktischer Umgang
- Standardisierung der Merkmale (z.B. z-Standardisierung).
- Wahl geeigneter Distanzmaße je nach Datenart.
- Skalierung der Features vor K-Means oder PCA.
- Visualisierung von Dimensionen (PCA-Scatterplots, Dendrogramme).

(8) Typische Anwendungsbeispiele
- Kundensegmente identifizieren (Clustering).
- Produkt-Eigenschaften reduzieren, um Visualisierung zu ermöglichen (Dimensionsreduktion).
- Muster und Ausreißer in Datenströmen erkennen (Mustererkennung).

Hinweis zu Lernzielen
Sie können die Konzepte von Clustering, Dimensionsreduktion und Mustererkennung im unüberwachten Lernen erklären, passende Algorithmen auswählen, grundlegende Formeln anwenden und typische Evaluierungsmetriken interpretieren.