Cluster Analysis - Cheatsheet

Onur Yilmaz

June 1, 2023

Clusteranalyse

Die Clusteranalyse ist ein Verfahren zur Gruppierung von Datenobjekten in Cluster, wobei Objekte in einem Cluster ähnlicher sind als Objekte in anderen Clustern. Die Clusteranalyse ist nützlich, um Muster und Strukturen in Daten zu identifizieren.

Cluster

Ein Cluster ist eine Gruppe von Datenobjekten, die ähnlich zueinander sind und sich von Objekten in anderen Clustern unterscheiden. Das Ziel der Clusteranalyse ist es, Datenobjekte in homogene Cluster zu gruppieren.

K-Means-Algorithmus

Der K-Means-Algorithmus ist ein populärer Algorithmus für die Clusteranalyse. Er teilt Datenobjekte in K Cluster ein, wobei K eine vorher festgelegte Anzahl ist.

Elbow-Methode

Die Elbow-Methode ist eine Technik zur Bestimmung der optimalen Anzahl von Clustern im K-Means-Algorithmus. Dabei wird die Summe der quadrierten Abweichungen (Sum of Squared Errors, SSE) für verschiedene Werte von K berechnet und graphisch dargestellt. Der "Ellenbogenpunkt" in der Grafik gibt den optimalen Wert für K an, an dem die SSE nicht mehr signifikant abnimmt.

Beispiele in Python

Hier sind Beispiele für die Implementierung der Clusteranalyse mit dem K-Means-Algorithmus und der Elbow-Methode in Python:

```
1 from sklearn.cluster import KMeans
2 import matplotlib.pyplot as plt
```

```
3
4  X = load_data()
5
6
7  sse = []
8  k_values = range(1, 10)
9  for k in k_values:
10 kmeans = KMeans(n_clusters=k)
11 kmeans.fit(X)
12 sse.append(kmeans.inertia_)
13
14
15 plt.plot(k_values, sse, 'bx-')
16 plt.xlabel('Anzahl der Cluster (K)')
17 plt.ylabel('SSE')
18 plt.title('Elbow-Methode')
19 plt.show()
20
21 kmeans = KMeans(n_clusters=3)
22 kmeans.fit(X)
23
24 labels = kmeans.predict(X)
```