UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

Data and Web Science Group
B6, 26 – B1.15
68159 Mannheim

Data Mining – Python

Exercise 2: Cluster Analysis

2.1. Analyzing the Customer Data Set

1. Import the *customers* data set into Python with pandas. The customer data set is provided in ILIAS as an Excel file. Use the read_excel function.
2. Cluster the dataset using K-Means clustering. Experiment with different K values. Which values do make sense? What does the clustering tell you concerning your product portfolio? What does the clustering tell you concerning your marketing efforts in different regions?
3. Cluster the data set using Agglomerative Hierarchical Clustering (Using the dendrogram and linkage functions from scipy). What does the dendrogram tell you concerning your customer groups?
4. Flatten the hierarchical clustering so that you get 3 or 4 customer groups. Name these groups with appropriate labels.

2.2. Analyzing the Students Data Set

1. Aggregate the *students* data set (from Exercise 1) by student and calculate the average mark and the average number of attended classes.
2. Cluster the data set using the K-Means algorithm.
   Does one attribute dominate the clustering? What can you do about this? Assign suitable labels to your clusters.
3. Cluster the data set using Agglomerative Hierarchical Clustering. Experiment with different setting for calculating the cluster similarity. What is a good setting?
4. What does the dendrogram tell you about the distances between the different groups of students?

2.3. Clustering the Iris Data Set

1. Cluster the Iris data set (from Exercise 1) using different algorithms and parameter settings.
2. Does it make sense to normalize the data before applying the algorithms?
3. Try to choose an algorithm and parameter setting that reproduces the original division into the three different species.

2.4 Clustering the Geo Data Set

1. Within the geo data set (provided in ILIAS) the coordinates (x & y) of housings of inhabitants of an area are collected. Have a look at the data and visualize (scatter function).
2. Cluster the data using k-Means (k=3). Do the clusters represent the original areas?
3. Apply DBSCAN and play around with the epsilon. Can you reproduce the original areas using this cluster algorithm?

2.5. Clustering the Zoo Data Set

1. The *Zoo* data set describes 101 animals using 18 different attributes. The data set is provided in ILIAS as an ARFF file. Import the *Zoo* data set with the arff.loadarff function from scipy.
2. Cluster the data set using Agglomerative Hierarchical Clustering. Experiment with different parameter settings in order to generate a nice species tree.
3. Try to assign appropriate species names to the clusters at the upper levels.

ChatGPT Bonus Exercises

Reminder: Do not take the answers of ChatGPT at face value! Always cross-check with lecture slides, literature and/or the teaching staff!

C.1. Discuss Clustering Methods

- Use ChatGPT to discuss the problem of choosing the correct clustering algorithm for a use-case. When should you choose DBSCAN over K-Means and vice versa?
- Suppose you applied a DBSCAN algorithm but every point is marked as an outlier. Discuss with ChatGPT what might be the cause of this and how you can overcome the problem.

C.2. Learn about possibilities for visualizing higher dimensional clusterings in Python

- During the exercise we limit clusterings and visualizations to two dimensions for simplicity and understandability. Ask ChatGPT about possibilities and methods for visually evaluating higher dimensional clusterings.
- Cluster one of the exercise datasets by more than two dimensions and let ChatGPT create code for you to visualize this clustering using one or more of the proposed methods.

C.3. Self-Assessment

- Ask ChatGPT to create a pen and paper exercise for you that lets you practice the K-Means clustering algorithm using a small example dataset. After completing the exercise ask ChatGPT for the solution and critically evaluate it. You can also ask ChatGPT for code to run the exercise using scikit-learn.
- Ask ChatGPT to create an exam exercise for graduate students relating to the effect of outliers on the K-Means algorithm and solve the exercise. Get the answer from ChatGPT and critically evaluate it.
- Ask ChatGPT to create three multiple choice questions about clustering for graduate students, including about the topic of choosing the correct value for K, and solve them.