Density-Based Clustering

Applied Machine Learning in Engineering - Exercise 04

TU Berlin, Winter Term 2023/24

Prof. Dr.-Ing. Merten Stender – merten.stender@tu-berlin.de

Both problems will work on the data provided in the file `data_clustering.csv` and visualized in Figure 1.

Problem 1 - 30min

Implement an object-oriented method for z-scoring

$$\tilde{\mathbf{x}} = \underbrace{\frac{1}{\sigma\left(\mathbf{x}\right)}}_{\text{unit standard deviation}} \cdot \underbrace{\left(\mathbf{x} - \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\right)}_{\text{zero mean}}, \quad \mathbf{x} \in \mathbb{R}^{N,n} \tag{1}$$

a given data set (including the reverse transformation). Implement the following class methods

(a) `Zscorer.fit(X)` for estimating the mean and standard deviation for data `X`

(b) `Zscorer.transform(X)` for z-scoring the data

(c) `Zscorer.inverse_transform(X)` for reversing the z-scoring

(d) Make sure that you can z-score multi-dimensional data (see provided data file). Read the documentation of `numpy.mean` to see how to compute summary statistics along dimensions of arrays. Expect the data to have the shape $[N, n]$, where N is the number of samples, and n denotes the feature space dimension.

(e) Validate your result against the scikit-learn implementation in the `StandardScaler`.

Problem 2 - 60min

This exercise uses the DBSCAN algorithm from the `scikit-learn` package (official documentation).

(a) Use DBSCAN to find clusters in the given data set. How many clusters do you identify?

(b) Write a simple plotting function for visualizing the cluster label assignment. Use `plt.scatter` with the label assignment vector as additional argument.

(c) Vary DBSCAN hyperparameters and observe the results

(d) Use the silhouette coefficient to rate different clusterings.

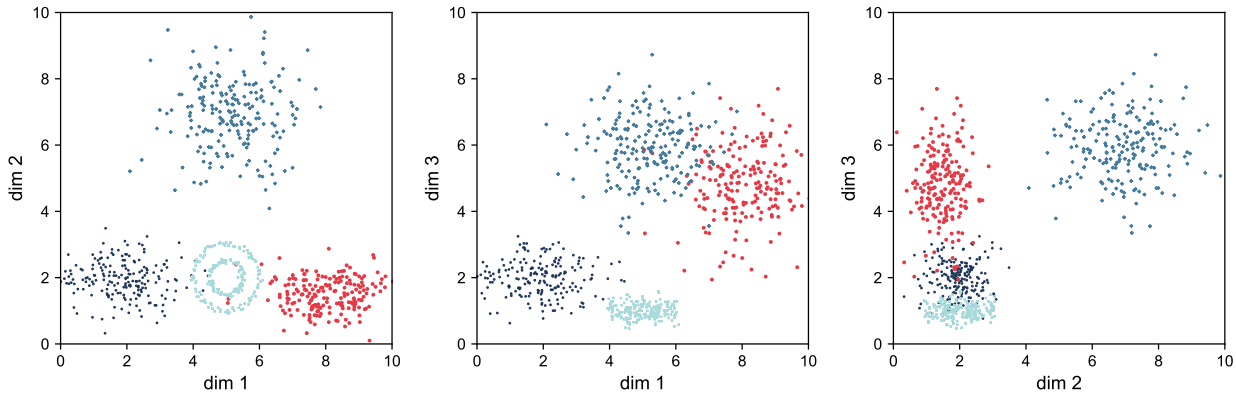

Figure 1: 3-dimensional data set with partially overlapping clusters