Clustering

Machine Learning
uczeniemaszynowe@cs.uni.wroc.pl

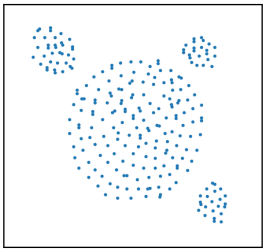
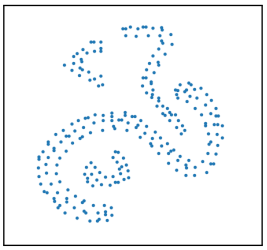
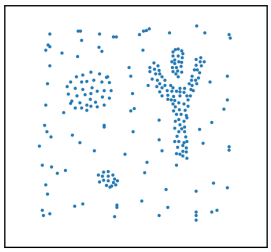
deadline: 25.11.2025 10:00
tutorial / lab

Task 1. [2 pts]

Construct an example of data – as simple as possible – on which the KMeans algorithm does not work directly (i.e., does not give the obvious expected clustering), but does work after standardization. Explain why.

Task 2. [4 pts]

The figures below present 3 datasets (the images are also available as PNG files on the course website). Recreate the datasets visible in the figures and cluster them using KMeans, DBScan, and hierarchical clustering. Choose algorithm parameters appropriately. Compare and discuss the results.

Hint: You do not need to manually rewrite the data from the figures. Load the images using the PIL package. Convert the image into a Numpy Array. Determine the indices of array elements corresponding to non-white pixels of the image. A single data point in the image may consist of several pixels, so use the KMeans algorithm to cluster these pixels into data points.

Task 3. [2 pts]

Load several arbitrary images (both photographs and drawings with a small number of colors). Use the KMeans algorithm to reduce the number of colors in the loaded images. Show and discuss the results.

Task 4. [2 pts]

Determine K points lying at the vertices of a regular K-gon in the plane. For each of these points, randomly generate M points from a two-dimensional Gaussian distribution centered at that vertex of the K-gon, with a diagonal covariance matrix $\sigma \cdot \mathbb{I} \in \mathbb{R}^{2\times 2}$, for a fixed parameter $\sigma \in \mathbb{R}$. For the dataset of $K \cdot M$ points thus created, use the KMeans algorithm to divide them into K groups (and recover the original groups). Choose the polygon size and parameter σ so that the groups are reasonably well separable.

1. Perform hierarchical clustering, draw a dendrogram, and discuss the results.
2. Run the KMeans algorithm several times for numbers of clusters from 2 to K, compare the obtained clusterings with those obtained in the previous step.
3. For a large amount of data (large K and large M — test K on the order of 10^2 or 10^3 and M on the order of 10^4 or 10^5 depending on available computational resources), compare the running time and accuracy of KMeans with the MiniBatchKMeans algorithm.