Machine Learning Paradigms for Complex Data

Assignment 1: Curse of dimensionality

Summer term 2024

Data generator

Implement a data generator to generate clustered data. The method shall generate a dataset of specified dimensionality d, where each dimension has the value range $[0, 100]$.
Furthermore, the data shall have k hidden clusters. The clusters shall either be in full-space or subspace. For each cluster, a certain number of dimensionality (i.e., the number of relevant dimensions for the cluster) can be specified. In the relevant dimensions, the objects of a cluster are uniformly distributed within a specific radius. In contrast, in the non-relevant dimensions, the objects are uniformly distributed in the data space (range $[0, 100]$).
Additionally, the generator shall have the option to generate outliers: objects that are uniformly distributed in the data space.
During the data generation, the following parameters are to be specified:

- number of dimensions d
- number of clusters k
- number of objects in the cluster (for each cluster)
- radius for the cluster in the relevant dimensions (for each cluster)
- dimensionality of the cluster (for each cluster)
- number of noise objects

The programming language is free to choose.

Analysis of high dimensional data

1. with the data generator create five datasets with increasing dimensionality d, vary d between 1 and 100. Create subspace clusters for each dataset. (The number of objects and dimensionality of subspaces are free to choose)
2. describe and visualize your datasets in a proper way
3. discover the curse of dimensionality problem in the generated datasets.
    - Determine for each object the ratio "farthest-neighbor-distance" / "nearest-neighbor-distance" by using the Euclidean distance and calculate the average ratio for all objects in the same dataset. Plot the average ratio for the sequence of datasets with increasing dimensionality. What conclusion can be drawn from this result with respect to the empty space problem/curse of dimensionality? Do you get the same results by using the Manhattan distance or the Maximum-Metric instead of Euclidean distance?
    - Assume the data space is partitioned into a regular grid (cf. Slide 24, Chapter 3) with four partitions per dimension. Generate for each dataset a histogram that counts the number of cells covering one object, two objects, three objects, etc. How do the histograms change by increasing the dimensionality of the data? What are your observations?
    - For every data set do a hypercube range query with length $s = 0.8$ after normalizing the value range to $[0, 1]$ in all dimensions placed arbitrarily in the data space. Plot a graph with increasing dimensions displaying the fraction of objects captured in each data set (cf. Slide 27, Chapter 3).
    - Apply one of the conventional full-space clustering models to the generated datasets. What can you observe from the results?

Submission

- summarize your solution as a 2-page report in pdf format[1] excluding figures
- submit the source code of data generator, and the report on Moodle **before** 11.06.2024 11:59 am
- share your datasets (.zip) to all lecture participants on Moodle **after** 11.06.2024 11:59 am. Your datasets will be used in Assignment 2.

[1] Example template: `https://www.acm.org/publications/proceedings-template`