IT Security 2024/2025
Exercise Sheet 8
– Topological Data Analysis –

Christian Bungartz

Publication: 28.11.2024
Deadline: 04.12.2024 10:00

Lightning Survey ⚡

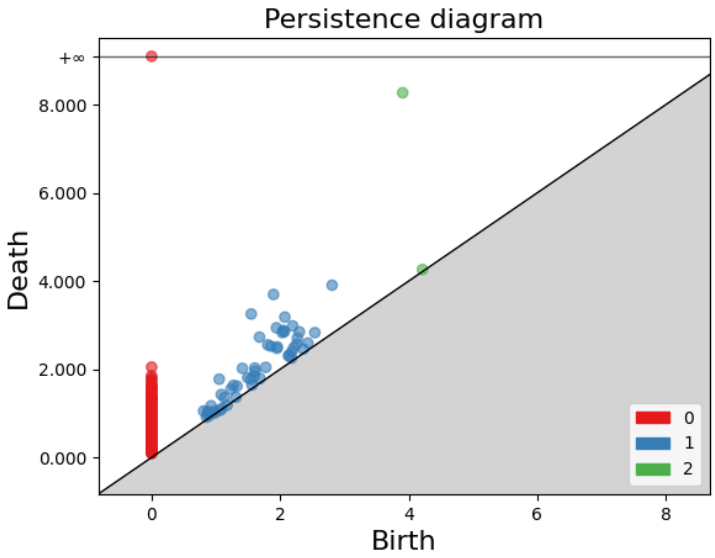

Figure 1: Persistence diagram of a point cloud sampled from an unknown shape.

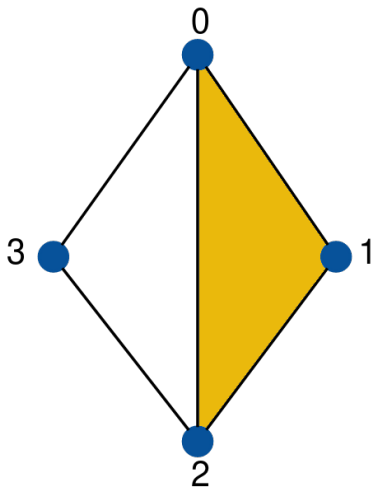

Figure 2: Simplical complex

Exercise 1 (Topological Data Analysis - Theory, 1+1 points)**.** Solve the following subtasks and submit your solution as `theory.txt` or `theory.pdf`:

a) What does the persistence diagram in Figure 1 tell you about the underlying point cloud? What shape could the point cloud be sampled from?

b) Consider Figure 2. What is the corresponding abstract simplex complex? How many zero- and one-dimensional holes does the complex have?

Exercise 2 (Persistent Homology - Practice, 3+3 points)**.** In this exercise you will compute the persistent homology of given point clouds. To this end, familiarize yourself with the `gudhi`[1] library and its documentation. Use the seed `42` where applicable.

a) Given the point cloud stored in `torus.csv`, compute its Vietoris-Rips complex and its persistent homology up to dimension three. Visualize the corresponding persistence diagram and barcodes. Submit your script as `torus.py`, and the persistence diagram and barcodes as `torus-diagram.png` and `torus-barcode.png`, respectively.

[1] `https://gudhi.inria.fr/python/latest/`

b) In the folder `iot_behavior` you find 174 point clouds, each represented by a `CSV` file. Each file contains 50 data points describing the behavior of an IoT device in comparison to four other homogeneous devices over a time period of t. Write a Python script `iot.py` that computes the persistent homology of the Vietoris-Rips complex of each point cloud and then generates a 2D embedding employing multidimensional scaling (MDS)[2]. Visualize the resulting embedding using the labels to color the points. In addition to your script, submit the embedding as `mds.csv` (without labels, header, and index), and the visualization as `iot.png`.

Exercise 3 (Mapper, 1+1 points)**.** Topological data analysis also finds applications in exploratory data analysis. In the following, you will get to know the Mapper algorithm, which is a method for visualizing high-dimensional data sets. To this end, familiarize yourself with the `kmapper`[3] library and its documentation. Use the seed `42` where applicable.

a) Describe in your own words (not more than 300 words) how the Mapper algorithm works. You may use diagrams to illustrate your explanation. Submit your answer as `mapper.txt` or `mapper.pdf`:

b) The file `phishing.csv` contains a data set describing characteristics of phishing websites. Write a Python script, that uses `kmapper` to visualize the dataset. Your script should utilize `Principal Component Analysis`[4] as a lens (projection) into 3D space. The clusterer should be `AgglomerativeClustering`[5] with `metric="euclidean"`, `linkage="complete"`, and `n_clusters=3`. Submit your scripts `phishing.py` and the resulting HTML file `phishing.html`.
Bonus: Experiment with different parametrization for both mapper as well as the projection and clustering algorithm.

[2] `https://scikit-learn.org/stable/modules/generated/sklearn.manifold.MDS.html`
[3] `https://kepler-mapper.scikit-tda.org/en/latest/`
[4] `https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html`
[5] `https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html`