Invitation to TDA – Practical Exercises

Paweł Dłotko, Davide Gurnari, Niklas Hellmer

University of Warsaw, Summer semester 2022

Problems indicated with multiple * are project level.

Problem 1 Generate a collection of points sampled from (a) a unit circle, (b) a torus, (c) a Möbius band. Find appropriate parametrizations of those manifolds and present a code to sample the desired number of points from them. Visualize the obtained point clouds. Use the methodology of procedural programming to have one sampling procedure for all 2 and 3 dimensional point clouds.

Problem 2 (Concentration of measure); Sample k points from $[0,1]^n$ for $n = 2, 3, 10, 20, 100, 1000, 10000, 1000000$. Compute an average and the standard deviation of the distance between those points. What are the conclusions you may take out of it?

Problem 3 (Johnson-Lindenstrauss projection) Write down a procedure that projects an $n-$dimensional point cloud into randomly selected k dimensions. Define a metric distortion as in the Johnson-Lindenstrauss Lemma and write an appropriate *while* loop that search for a projection that minimizes the distortion. Repeat the experiment for 10 different random point clouds in dimension n for $n = 1000, 10000$. What is the obtained metric distortion? How many iterations were required to find the right projection? How is it related to the fact that Johnson-Lindenstrauss lemma speaks about the existence of an appropriate projection with probability 1?

Problem 4 Consider the possibility o using random projections of high dimensional datasets for the sake of speed up the search of k nearest neighbors. Check what are the conditions that need to be checked on the projected data, write down an implementation and test the running times.

Problem 5 Search for Anscombe Anscombe and Datasaurus dataset. Compute the summary statistics of the sets in those collections. What can you say about them, based on those summary statistics? Then visualize the datsets. What are your conclusions? Is there a way to detect instances similar to this one when the data are sampled from much higher dimensional space?

Problem 6 *** Consider the possibility of applying PCA to speed up the spatial search data structures. The proposed solution should take an advantage of the point cloud projected to a few principal components (where efficient spacial search tools, like k-d-trees may be utilized). How the amplitude of the remaining components can be used to control the error of such a method? Does it make sense, from the point of computational complexity, to use PCA in this context?

Problem 7 ***** A typical and very well studied textural corpora is the collection of all documents from English (or any other language) Wikipedia. Those articles can be downloaded, processed using term-frequency-inverse-document-frequency technique to provide vectors in high dimensional space. Those points can be subsequently analyzed using tools presented in this book, in particular mapper-type algorithms. Your task is to adapt the mapper algorithms for this datasets. Use appropriate metadata (indicating if we are dealing e.g. with a scientific, popular, political or other article) to colour the obtained model of the space. What are the conclusions you may get based on this analysis?

Problem 8 Check how much the cosine similarity suffers from the concentration-of- measure type phenomena.

Problem 9 Write a program using interval arithmetic to prove that a function $f(x) = x^2 - 2$ have a zero in an interval $[1, 2]$.

Problem 10 ** Implement a class of interval arithmetic (setting up appropriate rounding on the processor is an option). The class should implement the four basic arithmetic operations and a few of elementary functions.

Problem 11 Write a program to find a partially constant approximation $\hat{f}$ of a function $f(x) = e^{\frac{x^2}{2}}$ on an interval $[0, 10]$ that is not farther away that a predefined constant ϵ from $f(x)$.

Problem 12 Write a program which combines interval arithmetic and automatic differentiation to locate one of the minimum of the function $f(x) = sin(cos(tanh(cos(3 * x))))$ that is located close to $x = 2$.

Problem 13 Construct simplicial complexes in GUDHI's `SimplexTree` via

1. triangulations from theoretical problem 16,
2. mesh files provided in class,
3. Vietoris-Rips complexes for point cloud samples,

for each of the following spaces:

1. 2-sphere (at least three different triangulations),
2. Möbius band,
3. real projective plane,
4. torus,
5. Klein bottle.

Run the `compute_persistence()` method of the `SimplexTree`s with parameters `persistence_dim_max = True` and `homology_coeff_field = k` for various numbers of k. Print the output of the `betti_numbers` method. Compare the results.

Problem 14 Consider a rectangle $[-2, 2] \times [-2, 2]$ and build a 100 by 100 cubical complex therein. For each top dimensional cube, assign the value of a distance function from the unit circle $x^2 + y^2 = 1$ on that cube. Visualize the obtained cubical complex. Hint: you may use Gudhi library in the process.

Problem 15 Sample a random collection of N points from a uniform distribution in $[0, 1]^n$ for $n \in \{1, 2, 3, 5, 10, 15, 20\}$. Using Gudhi library, construct Vietoris-Rips and Čech complexes for some values of the parameter ϵ. Compare the time of creation and number of simplices. Draw a conclusions on the ranges of N and n for which each complex should be used. Moreover, investigate how the choice of ϵ affects the size on the complex.

Problem 16 Sample a random collection of N points from a uniform distribution in $[0, 1]^n$ for $n \in \{1, 2, 3, 5, 10, 15, 20\}$. Using Gudhi library, construct Alpha and Witness complexes for some values of the parameter ϵ. Compare the time of creation and number of simplices. Draw a conclusions on the ranges of N and n for which each complex should be used. Moreover, investigate how the choice of ϵ affects the size on the complex.

Problem 17 Sample a collection of N points from n–dimensional normal distribution for $n \in \{2, 3, 4\}$. Construct an Vieroris-Rips and Witness complex of this collection of points. Why Vietoris-Rips construction is having difficulties and how are the overcommed by the Witness construction?

Problem 18 In the two previous exercises, compare the persistence diagrams from Witness to the ones from Alpha and VR using Bottleneck and Wasserstein distance.

Problem 19 In the setting of problem 15, compare the persistence diagrams from Čech to those from VR using Bottleneck and Wasserstein distances.

Problem 20 Sample 100 points with noise from a sphere, a circle, a torus and the unit cube, respectively. Repeat 20 times for each shape. For each of the 80 point clouds, compute persistence diagrams in dimensions 0,1,2. Compare the computation time. Compute Bottleneck and Wasserstein distance matrices. Visualize the distance matrices. Use multi-dimensional scaling on the distance matrices and visualize the result in 2D. Train a k neighbors classifier to distinguish the four shapes by their persistence diagrams using the distance matrices. Compare the classification accuracy for Bottleneck vs. Wasserstein in dimensions 0,1,2.