Invitation to TDA – Practical Exercises

Paweł Dłotko, Davide Gurnari, Niklas Hellmer

University of Warsaw, Summer 2022

Problem 1 Generate a collection of points sampled from (a) a unit circle, (b) a torus, (c) A Mobious 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.