Topological Data Analysis(TDA)

Samarth Bansal(11630), Deepak Choudhary(11234)
Instructor : Dr.Amitabha Mukherjee

Introduction

Topology is the branch of mathematics that deals with the study of shape of data.

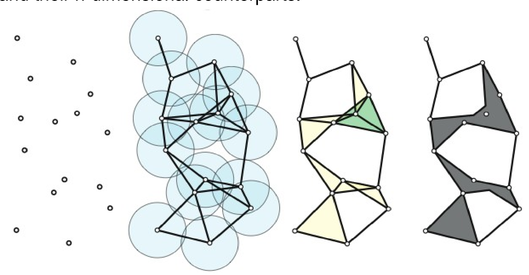
TDA:
Data Points ⟶ Geometric object ⟶ Topological summary

There are three key ideas of topology that make extraction of patterns via shape possible.[2]

- It studies shapes in a coordinate-free way.
- It studies the properties of shapes that are invariant under 'small deformations'.
- The third key idea is that of compressed representation of shapes. Betti numbers are used to distinguish topological spaces based on the connectivity of n-dimensional simplicial complexes.

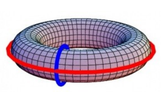
Persistence Homology

Simplicial complex is a topological space of a certain kind, constructed by "gluing together" points, line segments, triangles, and their n-dimensional counterparts.

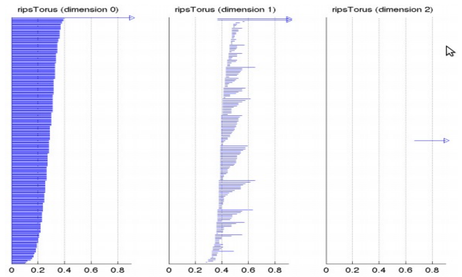


Persistent homology is a method for computing topological features of a space at different spatial resolutions.
Repeat the process throughout the poster as needed.

Betti numbers are used to distinguish topological spaces based on the connectivity of n-dimensional simplicial complexes.
Basically they record significant topological features of the shape.

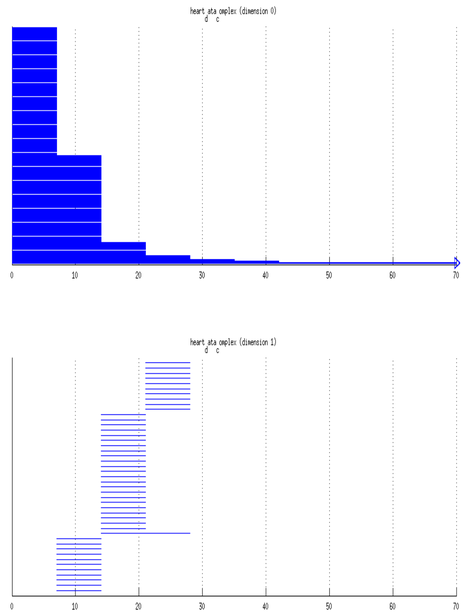


	$Betti_0$	$Betti_1$	$Betti_2$
Circle	1	1	0
Sphere	1	0	1
Torus	1	2	1



Barcode Analysis of Heart Disease Dataset

➔**Barcode**:
- ➔Each horizontal bar represents the birth-death of a separate homology class.
- ➔The i-th Betti number at any given parameter value is the number of bars.



Observations

The graph shows variation of the structure with different values of the threshhold parameter t.

By observing the barcode of dimension 0 we can infer that a single line persists in our data set.

This implies that b_0 is 1, meaning that there is a singly connected component.

From the graph of dimension 1, we can see that some structures come up and decay. There is no persitent homology in this structure in 1 dimesion.

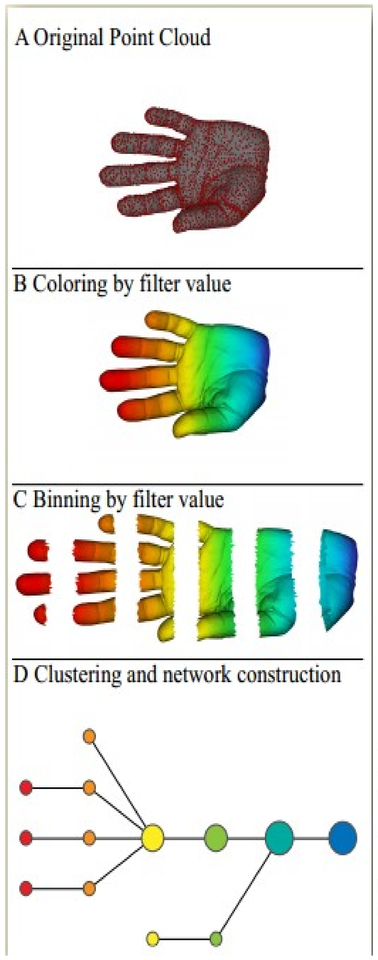
Thus b_1 for our data is 0.

These number imply that there is no persistent circle in our data.

MAPPER

Point Cloud ⟶ Filter Function ⟶ Binning ⟶ Clusters

Mapper method is based on the idea of partial clustering of the data along with a set of other functions defined for the data.



Point Cloud Data
N-dimesional
(Can be pre-processed as well)

Filter
f(x,y,z) ⟶ x

Binning
$f^{-1}(a_i, b_i)$
Put data into overlapping bins

Cluster each bin and create a network

Vetex = cluster of bin

Edge = non-empty intersection between clusters

Choices Made for Mapper Function

1) Data Model
- •How to model the data?
- •Distance metric : L2, Explicit etc

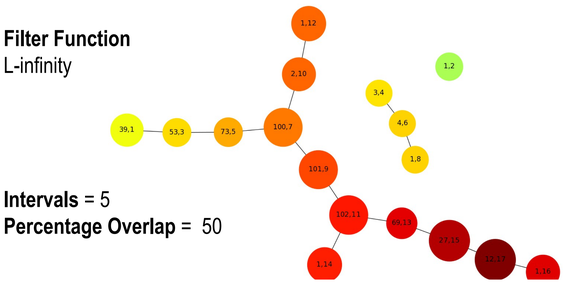
2) Filter Function
Possible Filter Functions
- •PCA
- •L-infinity centrality
- •Norm

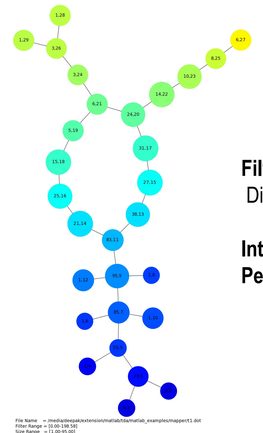
3) Resolution Parameters
- •Length of Interval
- •Percentage Overlap

MAPPER Algorithm on Heart Disease Dataset

Filter Function
L-infinity

Intervals = 5
Percentage Overlap = 50





Filter Function
Distance from age dimension

Intervals = 5
Percentage Overlap = 50

Subgroup Identification

We studied the heart dataset by extracting its shape and then obtaining insights from that. This method, employing TDA, helps us to identify subgroups in data that are difficult to obtain via traditional methods. The graphical network representation makes it easier to make sense of data. Different Subgroups might emerge as we use different filter functions, distance metric and resolution parameters.

References

[1] G. Carlsson, Topology and data. Bull. Amer. Math. Soc. 46, 255 (2009)
[2] Carlsson, G. et al. (2013) 'Extracting insights from the shape of complex data using topology',
[3] Gurjeet Singh, Facundo Memoli, and Gunnar Carlsson.(2007) Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition.

Dataset : Heart Disease Data Set, Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.

Softwares and programs

1) JavaPlex (Maintained by Comptop.stanford.edu)
2) Mapper (Maintained by Comptop.stanford.edu)
3) Matlab
4) Graphviz