Mapper

Ananth Shreekumar

May 18, 2020

1 What is Mapper?

Mapper [SMC07] is an unsupervised algorithm that is used to construct a Simplicial Complex that represents the structure of data. It reveals topological features of the data so that the data can be explored better.

To construct the graph, we require:

1. Filter function(s) that map the data to a lower dimension.
2. Covers for the range of each filter function with overlapping between each pair of consecutive intervals.

The filter function(s) is(are) used to represent the data in a lower dimension space. Dimensionality-reduction techniques like PCA or t-SNE could be used here if required.

2 The Mapper Algorithm

The covers for each filter function divide the range into polytopes in space as shown in figure 1.

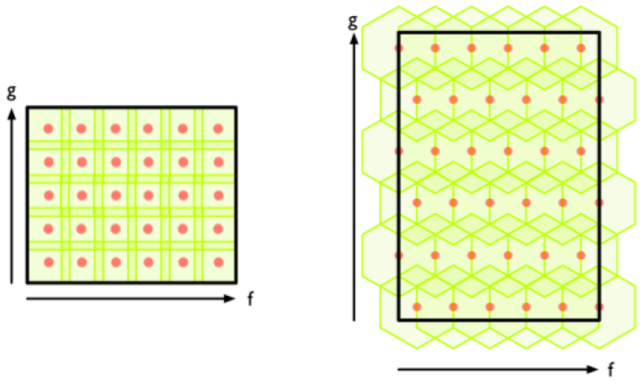

Figure 1: The functions **f** and **g** divide the space into polytopes in $\mathbb{R}^2$. Source: [SMC07]

The algorithm to construct the Simplicial Complex is as follows:

1. For each polytope, find a clustering of points that belong to that polytope and consider each cluster to represent a 0-Simplex (referred to as a vertex). Maintain a list of all clusters **L**.
2. For all vertices in **L**, if the intersection of any **n** vertices is a non-empty set of points, then add an $(\mathbf{n}-\mathbf{1})$-Simplex (referred to as an edge) corresponding to the associated vertices.

3 An Example

Consider data created using the following function on sklearn.

```
make_circles(n_samples = 10000, noise = 0.05, random_state = 44, factor = 0.5)
```

The data is a pair of concentric circles as shown in figure 2. The filter function is $\mathbf{f}(\mathbf{x}, \mathbf{y}) = \mathbf{y}$. The data as seen through it is shown in figure 3.

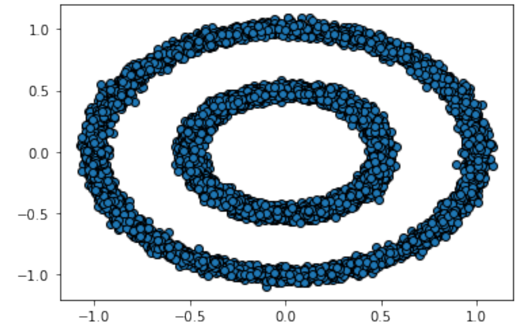

Figure 2: Data

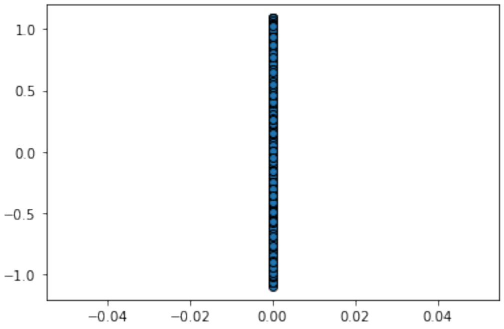

Figure 3: Data through the filter function

Let the cover of the range of the filter function be in the interval (-1.3, 1,3), with a uniform interval length of 0.1 and an overlap of 0.03. Explicitly, the cover is (-1.3, -1.2), (-1.23, -1.13), ... (1.22, 1.3). The data is split using the filter function $\mathbf{f}(\mathbf{x}, \mathbf{y}) = \mathbf{y}$ as shown in figure 4. The data points in the original data retrieved by the inverse of $\mathbf{f}$ after they are clustered are shown in figure 5. Note that overlaps are not visible because of the plotting software used.

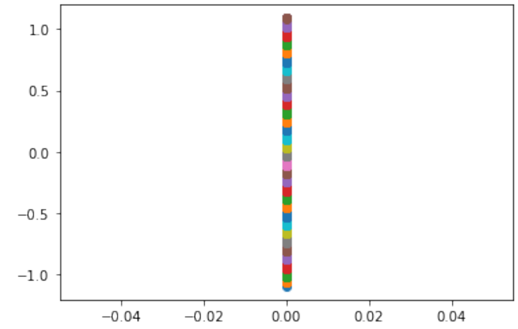

Figure 4: Data split through lens

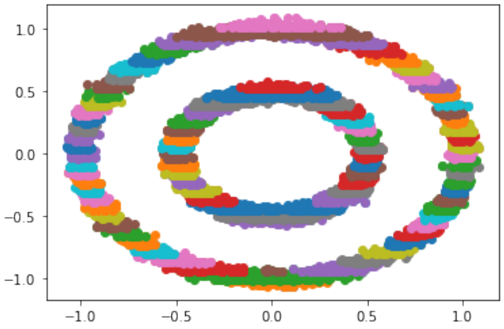

Figure 5: Data split in domain

Since the intervals in the cover are overlapping, consecutive intervals can contain common points. Each cluster in the data is made a node in a graph, and edges are added between two nodes only if the two nodes contain a common data point. The graph obtained after this operation is shown here.

References

[SMC07] Gurjeet Singh, Facundo Mémoli, and Gunnar E Carlsson. *Topological methods for the analysis of high dimensional data sets and 3d object recognition.*, pages 91–100. 2007.