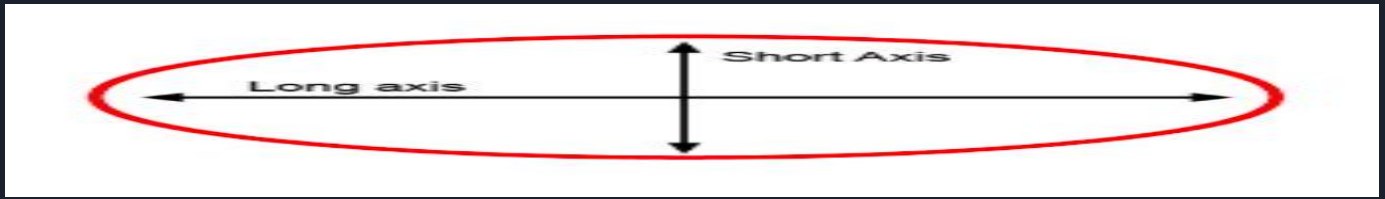
Topological Data Analysis

Different from the usual ML techniques

ML or TDA?

ML is a technique which gives you information about something, some statistical values in general. TDA is a technique which tells you about the shape of the data. Sometimes, it's important to understand the shape instead of a statistic.

For example, if you have data that forms an ellipse, a PCA based approach will give you the long axis as the most important axis and discard the other one.



ML or TDA?

In TDA, you will get a shape because TDA tries to make a simplicial complex from the data. And there are a lot of techniques by which you can identify the data as an ellipse and not a straight line, in a very short period of time.

Photo Credit: Kevin Brown/State Fair of Texas.

ML or TDA?

TDA is widely used for clustering or unsupervised task. But it is quicker than most of the other ML models.
The most important aspect is knowing about the shape of the data, because it can express a lot of information.

So, is TDA better?
No!
TDA alone might not be the best approach because of the hyperparameters. However, it can be very well combined with ML techniques to get more accurate and faster results.

Some examples are coming shortly!

Techniques in TDA

There are two major techniques:

1. Persistent Homology of the data: Homology is calculation of holes in some dimensions. Persistent means we are looking for the time when this hole appears and till the time it disappears.

   https://miro.medium.com/max/1400/1*9dqnnfbqz0WfjlhONkD2iQ.gif

Techniques in TDA

There are two major techniques:

2. Mapper: This is not too mathematical, however, it uses projections of different kind. There are different metrics which can be used according to the data we have.

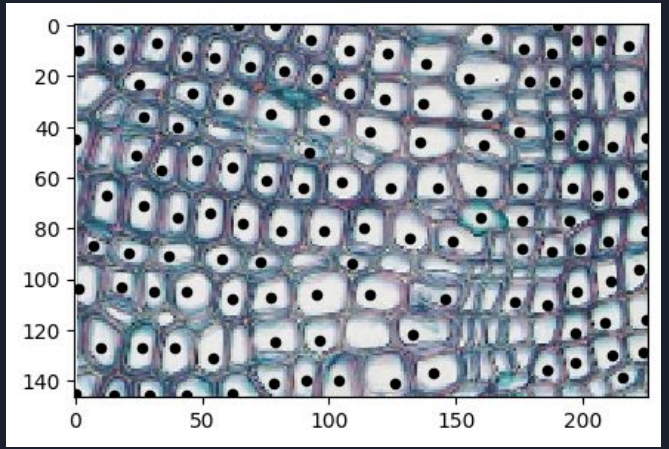
   Also, we create a simplicial complex using filters and then apply clustering techniques, which is the reason why it’s faster than the usual algorithms.

Some Examples

Counting example:

You can count the number of cells within a few

seconds using ripser/homology.



H_0 tells us about the number of connected components in the graph/surface. In this case, it tells us about the number of cells.

Soure: https://github.com/pandey-tushar/intro_to_TDA/blob/master/Ripser.ipynb

A lot of examples for mapper.

Mapper provides us with an interactive html file where you can see the data and how it's related to the other data points. You can also see which clustering techniques and projections are used to conclude the result.

The results on heart data https://pandey-tushar.github.io/intro_to_TDA/heart_data3.html

A result on classification of handwritten digits.
https://pandey-tushar.github.io/intro_to_TDA//digits_ylabel_tooltips.html

Trying to look into shape of the hand using simple knn projection.

https://pandey-tushar.github.io/intro_to_TDA/smoothed.html

Conclusion

Overall, Mapper and Ripser packages can help fasten a lot of work, specially with large datasets.

The codes I have written, I believe can be further optimized which means it can have even lesser running time.

The interpretation by Mapper is wonderful since it tells you not only the clusters but how are different nodes inside the cluster related to the other nodes and how many edges are there in any misclassified node etc.