UMAP

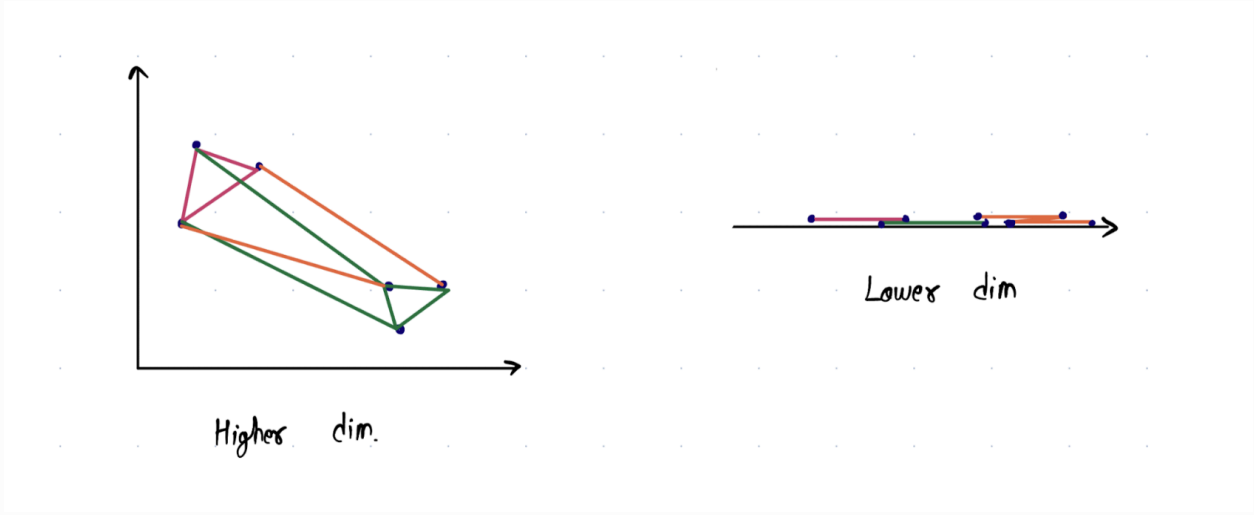
- Stands for Uniform Manifold Approximation and Projection
- Uses the underlying concepts of algebraic topology and topological data analysis.

UMAP will build a weighted graph using data points as nodes

- and have a sort of probabilistic distance among them.
- where weights represent the likelihood of a connection.

Such that

- Graph(Higher dimension) ~ Graph(lower dimension)



If UMAP is able to build a similar graph in lower dim

- This means it is able to proportionally preserve the distance between the nodes/ datapoints. (this is called fuzzy simplicial complex)

Hyperparam of UMAP

- n_neighbors
- Min_dist

What happens when we increase/decrease the n_neighbours?

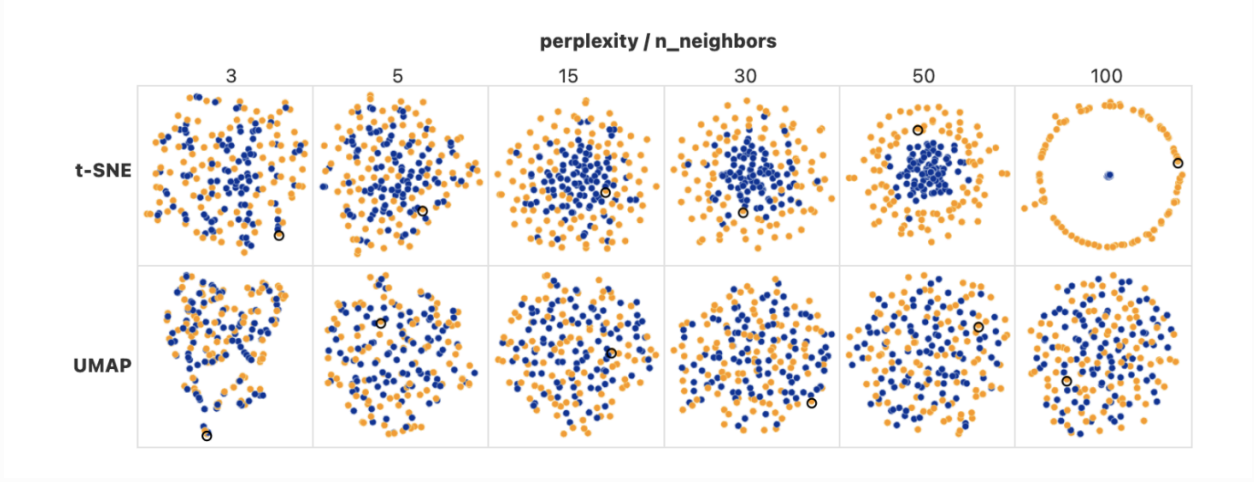
- As we increase n_neighbors, UMAP will preserve global structures very well
- As we decrease n_neighbors, 2D UMAP does not make any sense as it does not efficiently preserve global structures

min_dist is the **minimum distance allowed to be the separation between close points in the lower dimension embedding**

- If we **increase min_dist**, the **shape of the projection** of UMAP will **scatter** more

Failcase of UMAP

When there is a sparse cluster outside and a dense cluster inside



Important Points

- Hyperparameters in UMAP really matter
    - It is not easy to find optimal hyperparameters and one often has too many trials to find one.
- Cluster sizes are not important
    - After fitting UMAP, it is not guaranteed that the sizes of the clusters will be preserved

- Distances between clusters are not important
    - it is not guaranteed that these distances between the clusters will be preserved
- Random Noise doesn't always look random
    - After fitting UMAP on some random distribution of data, you may be able to find clusters(patterns), which says that the output from UMAP is not guaranteed to be random too
- May need to plot more than one plot
    - This means that every time one fits a UMAP model with the same hyperparameters, the same results are not guaranteed