WORKSHOP SOLUTION - POINT 9

What are the underlying mathematical principles behind UMAP?

The theoretical foundations for UMAP are largely based in manifold theory and topological data analysis. Much of the theory is most easily explained in the language of topology and category theory.

At a high level, UMAP uses local manifold approximations and patches together their local fuzzy simplicial set representations to construct a topological representation of the high dimensional data. Given some low dimensional representation of the data, a similar process can be used to construct an equivalent topological representation. UMAP then optimizes the layout of the data representation in the low dimensional space, to minimize the cross-entropy between the two topological representations.

The construction of fuzzy topological representations can be broken down into two problems: approximating a manifold on which the data is assumed to lie; and constructing a fuzzy simplicial set representation of the approximated manifold.

MATHEMATICAL PRINCIPLES:

- Manifold assumption

- Fuzzy simplicial set

- Riemannian geometry

- Stochastic optimization

What are the underlying mathematical principles behind UMAP?

UMAP is useful for generating visualizations, but if you want to make use of UMAP more generally for machine learning tasks it is important to be able to train a model and later pass new data to the model and have it transform that data into the learned space.

While UMAP can be used for standard unsupervised dimension reduction the algorithm offers significant flexibility allowing it to be extended to perform other tasks, including making use of categorical label information to do supervised dimension reduction, and even metric learning.

UMAP can be used as an effective preprocessing step to boost the performance of density-based clustering.