Unsupervised Learning, Part 3

Manifold Learning and UMAP

UMAP

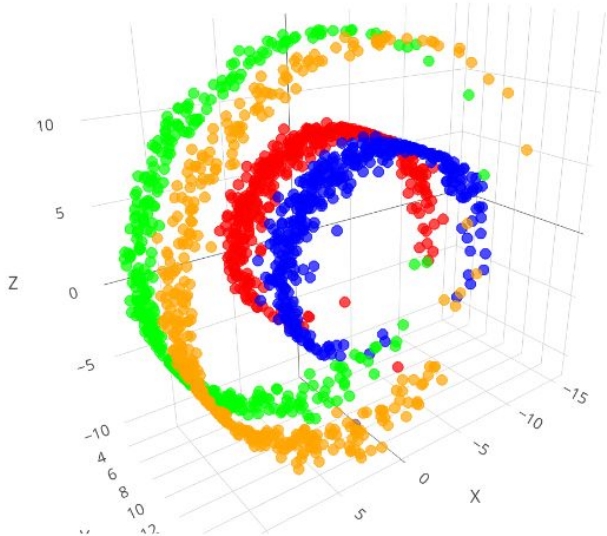
We have encountered PCA as a means of dimension-reduction.

Drawbacks of PCA:

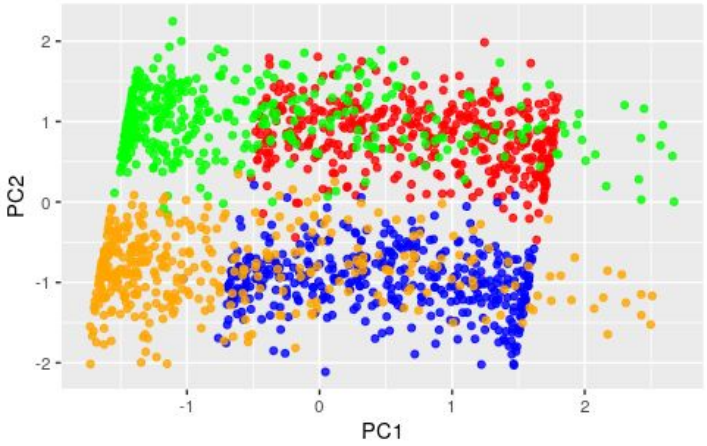
- Linear operation
- Relies on the first few principal components to capture most of the variation.

UMAP

If your dataset is highly nonlinear, this structure may be lost by PCA



PCA



UMAP

UMAP = **U**niform **M**anifold **A**pproximation and **P**rojection

General-purpose dimensionality-reduction

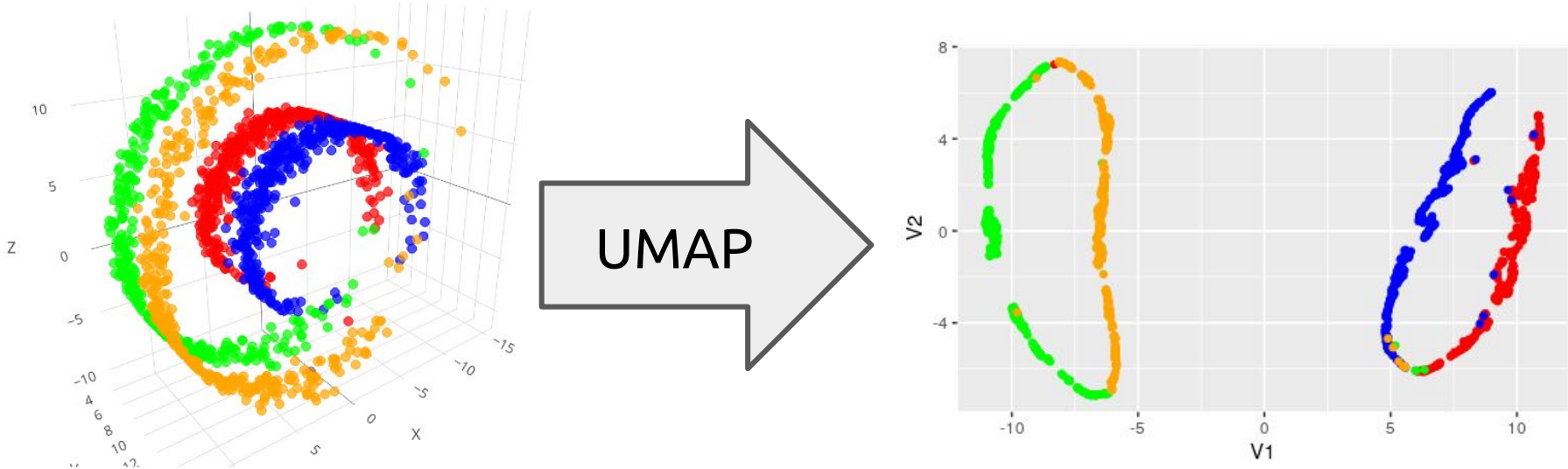
Uses:

- visualization
- preprocessing for machine-learning

Finds an embedding by searching for a low dimensional projection of the data that has the closest possible equivalent “fuzzy topological structure”

UMAP

UMAP does a better job at preserving more of the structure of the dataset.



Example Notebook

UMAP.Rmd