Machine Learning Algorithm: Dimension Reduction

PCA, UMAP, t-SNE

K. Kadri

Outline

Why Dimension Reduction?

- Real datasets often have many features
- Visualization becomes impossible in high dimension
- Models may suffer from:
  - noise
  - redundancy
  - curse of dimensionality

Goal

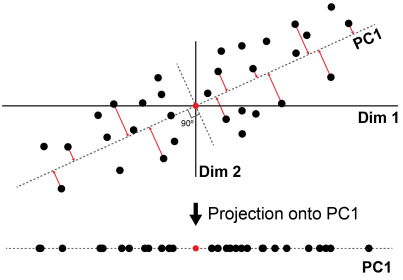
Project data into a lower-dimensional space while preserving structure.

PCA: Intuition

Key idea

Find directions of maximum variance in the data.

- Rotate the coordinate system
- New axes = principal components
- PC1 explains most variance
- PC2 orthogonal to PC1, explains next most



PCA: Mathematical Formulation

Projection

$$Z = XW$$

$$W = \text{eigenvectors of } X^\top X$$

Meaning

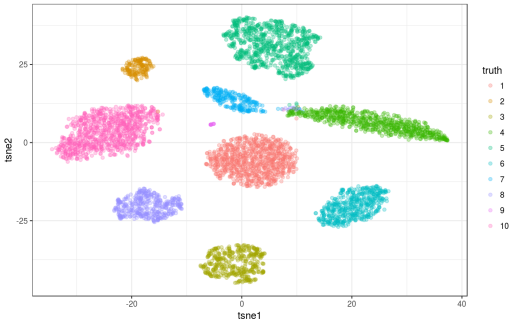
Eigenvectors define principal axes, eigenvalues measure explained variance.

t-SNE: Intuition

- Goal: preserve local neighborhoods
- Points close in high-dimension should remain close in 2D
- Converts distances into probabilities

Key idea

Clusters emerge naturally, but global structure may be distorted.



UMAP: Intuition

- Based on manifold theory and topological graphs
- Builds a neighborhood graph in high dimension
- Optimizes a graph layout in low dimension

Strengths

- preserves both local and some global structure
- faster than t-SNE
- stable embeddings

Comparison: PCA vs t-SNE vs UMAP

PCA

- Linear
- Fast
- Good for explainability

t-SNE

- Non-linear
- Excellent clustering visualization
- Poor global structure

UMAP

- Non-linear
- Keeps global + local structure
- Very fast