**Experiment: Principal Component Analysis (PCA) vs Linear Discriminant Analysis (LDA) vs T-distributed Stochastic Neighbour Embedding (t-SNE) vs Multi-Dimensional Scaling (MDS)**

**Title:**

**Implement a Multidimensionality Scaling Algorithm**[**MDS**](https://lms.sitpune.edu.in/mod/resource/view.php?id=18926)**on a specific dataset and compare its outcomes with other dimensionality reduction techniques such as PCA LDA and T-SNE**

**Aim:**

**Comparing the results of MDS with LDA, PCA and t-SNE for better suitability**

**Objective:**

Students will learn:

* The implementation of the Multi-Dimensional Scaling, principal component analysis and Linear Discriminant analysis and T-distributed stochastic neighbour embedding on a dataset.
* Visualization and interpretation of results.

**Problem Statement**

APPLY AND IMPLEMENT MDS ALGORITHM ON A SPECIFIC DATASET OF YOUR CHOICE AND COMPARE THE OUTCOMES WITH T-SNE, PCA AND LDA FOR THE SAME

**Explanation / Stepwise Procedure / Algorithm**

**Dimensionality Reduction Techniques**

**Principal Component Analysis (PCA)**

PCA reduces high-dimensional data while keeping most of its information. It identifies key directions (principal components) where data varies the most and projects it onto them.

**Steps:**

1. Standardize the data.
2. Compute the covariance matrix.
3. Find eigenvectors and eigenvalues.
4. Select top k eigenvectors.
5. Project data onto these eigenvectors.

**Uses:**

* Reducing dimensions
* Finding patterns in data
* Improving machine learning performance

**Linear Discriminant Analysis (LDA)**

LDA is a supervised technique that finds the best way to separate different classes. It is useful when features are many, but samples are few.

**Steps:**

1. Standardize the data.
2. Compute within-class and between-class scatter matrices.
3. Find eigenvectors and eigenvalues.
4. Select top k eigenvectors.
5. Project data onto these eigenvectors.

**Uses:**

* Reducing dimensions
* Improving classification performance
* Identifying key features for classification

**t-Distributed Stochastic Neighbor Embedding (t-SNE)**

t-SNE maps high-dimensional data to a lower-dimensional space while preserving local relationships. It is mainly used for visualization.

**Steps:**

1. Compute data similarity using a Gaussian kernel.
2. Convert it into a probability distribution.
3. Define a cost function for differences between high- and low-dimensional data.
4. Minimize the cost function.

**Uses:**

* Visualizing high-dimensional data
* Detecting clusters and patterns
* Preserving local structure in data

**Singular Value Decomposition (SVD)**

SVD breaks a matrix into three smaller matrices, capturing key patterns. It is widely used in image compression, recommendations, and noise reduction.

**Steps:**

1. Decompose matrix X into U, Σ, and VT:
   * U: Left singular vectors
   * Σ: Singular values (importance)
   * VT: Right singular vectors
2. Keep top k singular values and vectors.
3. Use these components to create a lower-dimensional representation.

**Uses:**

* Reducing dimensions
* Removing noise
* Feature extraction
* Applications in text mining & image processing

**Multidimensional Scaling (MDS)**

MDS represents high-dimensional data in lower dimensions while maintaining pairwise distances. It helps visualize similarities in data.

**Steps:**

1. Compute the dissimilarity matrix.
2. Convert it for eigenvalue decomposition or define a cost function.
3. Perform decomposition or use an optimization algorithm.
4. Select top k dimensions.
5. Assign new coordinates to data points.

**Uses:**

* Visualizing data in 2D/3D
* Understanding relationships between points
* Preserving distances in dimensionality reduction
* Market research and psychology analysis

**Figures/Diagrams**

* MDS and LDA, PCA and t-SNE plots plotted for the dataset.
* Comparison between MDS, LDA,PCA and t-SNE.

A group of colorful dots

AI-generated content may be incorrect.

**Challenges Encountered**

1. Different techniques work in different ways, making it hard to choose the best one.
2. Some methods, like t-SNE and MDS, take longer to process large datasets.
3. Understanding the reduced data can be tricky, as some details may be lost.

**Conclusion**

* Dimensionality reduction makes data easier to analyze and improves performance.
* Each method has its strengths, so the choice depends on the data and purpose.
* Comparing results helps in selecting the most suitable technique.