Experiment: Principal Component Analysis (PCA) vs Linear Discriminant Analysis (LDA) vs T-distributed Stochastic Neighbor Embedding (t-SNE)

Title:

Implementation of Linear Discriminant Analysis and Principal Component Analysis and T-distributed Stochastic Neighbor Embedding (t-SNE)

Aim:

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

Objective:

Students will learn:

- The implementation of the principal component analysis and Linear Discriminant analysis and T-distributed stochastic neighbor embedding on a dataset.
- Visualization and interpretation of results.

Problem Statement

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

Explanation / Stepwise Procedure / Algorithm

Principal Component Analysis (PCA)

PCA is a widely used dimensionality reduction technique that transforms high-dimensional data into lower-dimensional data while retaining most of the information. The goal of PCA is to identify the directions (principal components) in which the data varies the most and project the data onto those directions.

Here's how PCA works:

1. Standardize the data by subtracting the mean and dividing by the standard deviation.

2. Calculate the covariance matrix of the standardized data.

3. Compute the eigenvectors and eigenvalues of the covariance matrix.

4. Select the top k eigenvectors corresponding to the largest eigenvalues.

5. Project the original data onto the selected eigenvectors to obtain the lower-dimensional representation.

PCA is useful for:

- Reducing the dimensionality of high-dimensional data

- Identifying patterns and correlations in the data

- Improving the performance of machine learning algorithms

Linear Discriminant Analysis (LDA)

LDA is a supervised dimensionality reduction technique that aims to find a linear combination of features that separates classes of data. LDA is commonly used for classification problems and is particularly useful when the number of features is high and the number of samples is small.

Here's how LDA works:

1. Standardize the data by subtracting the mean and dividing by the standard deviation.

2. Compute the within-class scatter matrix and the between-class scatter matrix.

3. Compute the eigenvectors and eigenvalues of the scatter matrices.

4. Select the top k eigenvectors corresponding to the largest eigenvalues.

5. Project the original data onto the selected eigenvectors to obtain the lower-dimensional representation.

LDA is useful for:

- Reducing the dimensionality of high-dimensional data

- Improving the performance of classification algorithms

- Identifying the most discriminative features for classification

t-Distributed Stochastic Neighbor Embedding (t-SNE)

t-SNE is a non-linear dimensionality reduction technique that maps high-dimensional data to lower-dimensional data while preserving local structure. t-SNE is particularly useful for visualizing high-dimensional data and identifying patterns or clusters that may not be apparent in the original data.

Here's how t-SNE works:

1. Compute the similarity matrix of the high-dimensional data using a Gaussian kernel.

2. Convert the similarity matrix into a joint probability distribution.

3. Define a cost function that measures the difference between the joint probability distribution and the joint probability distribution of the lower-dimensional data.

4. Minimize the cost function using gradient descent to obtain the lower-dimensional representation.

t-SNE is useful for:

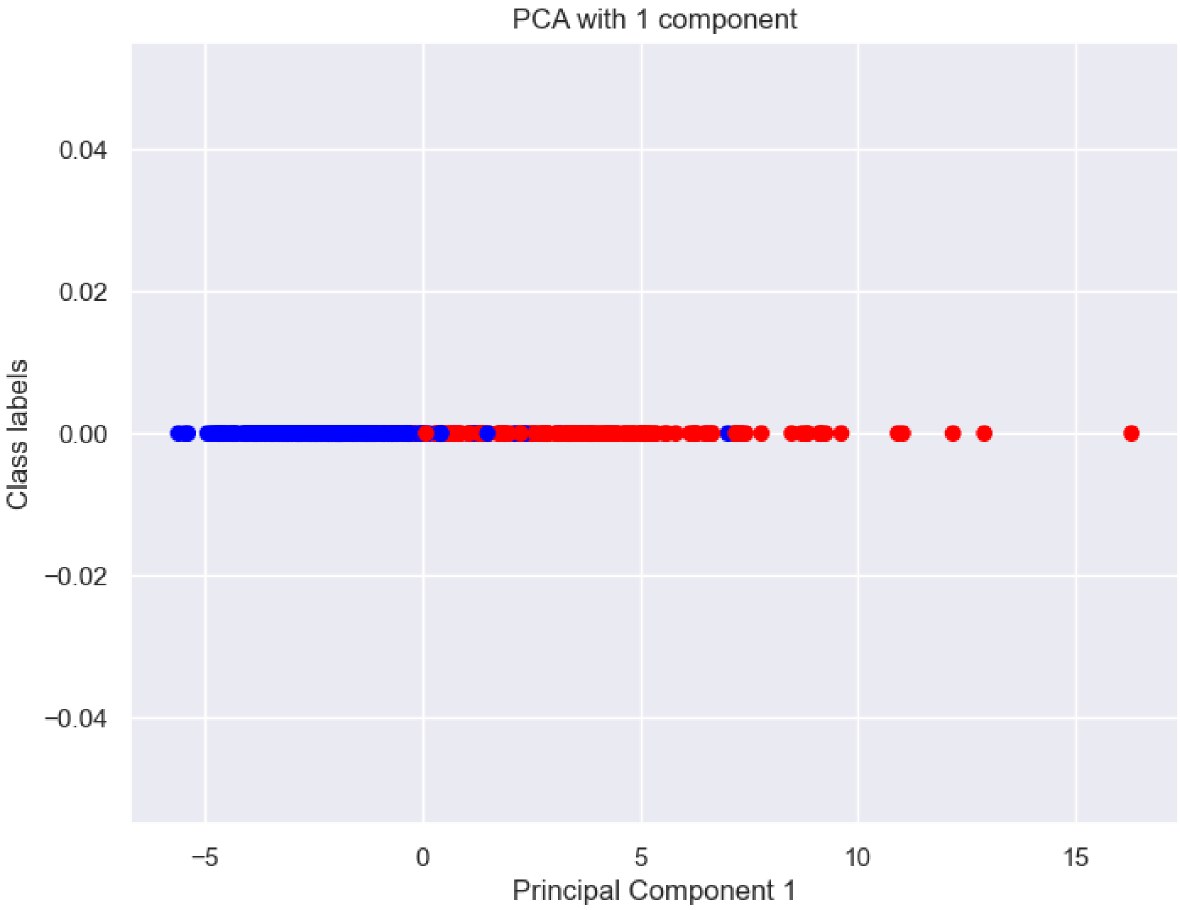
- Visualizing high-dimensional data in a lower-dimensional space

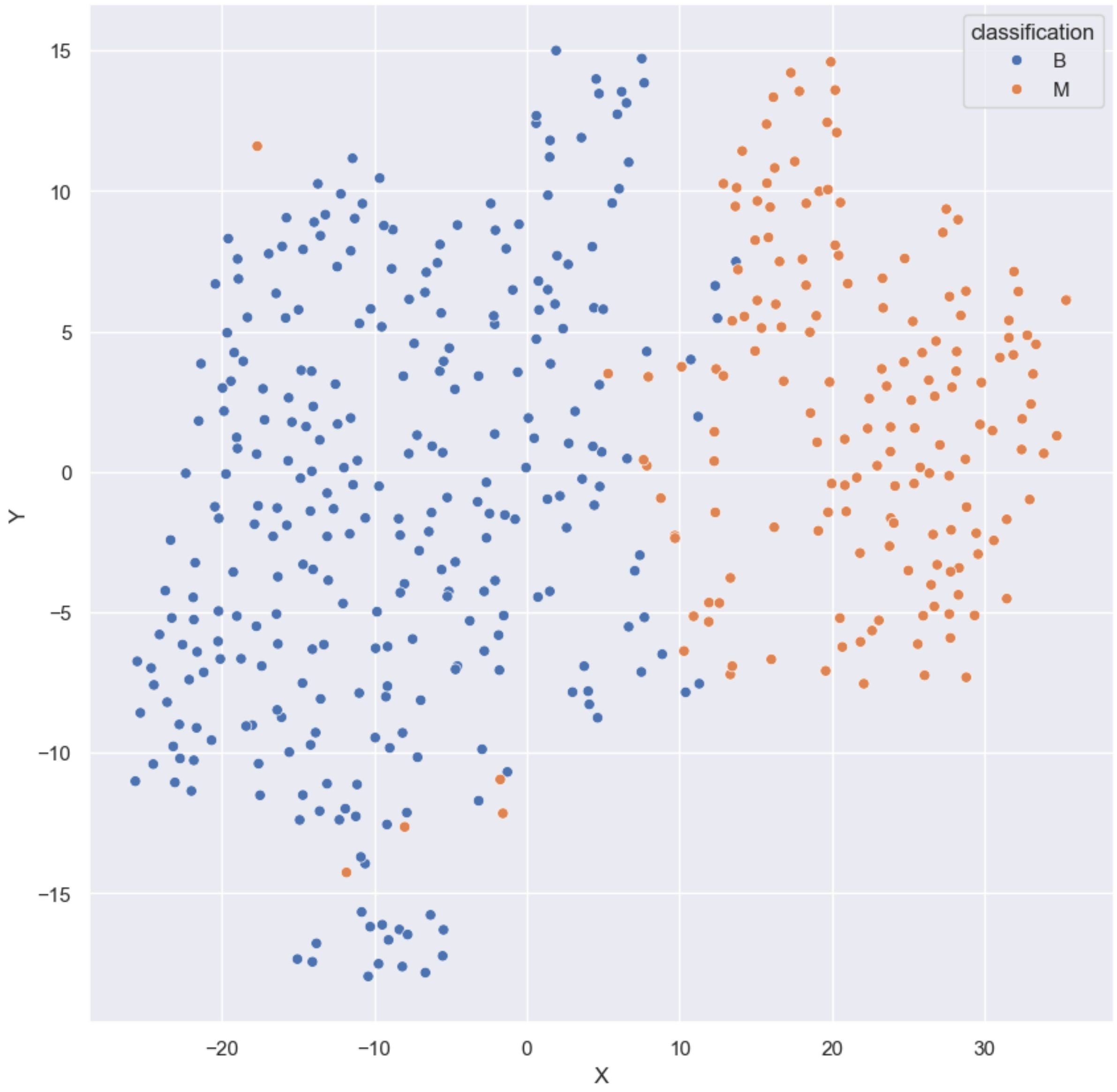
- Identifying patterns and clusters in the data

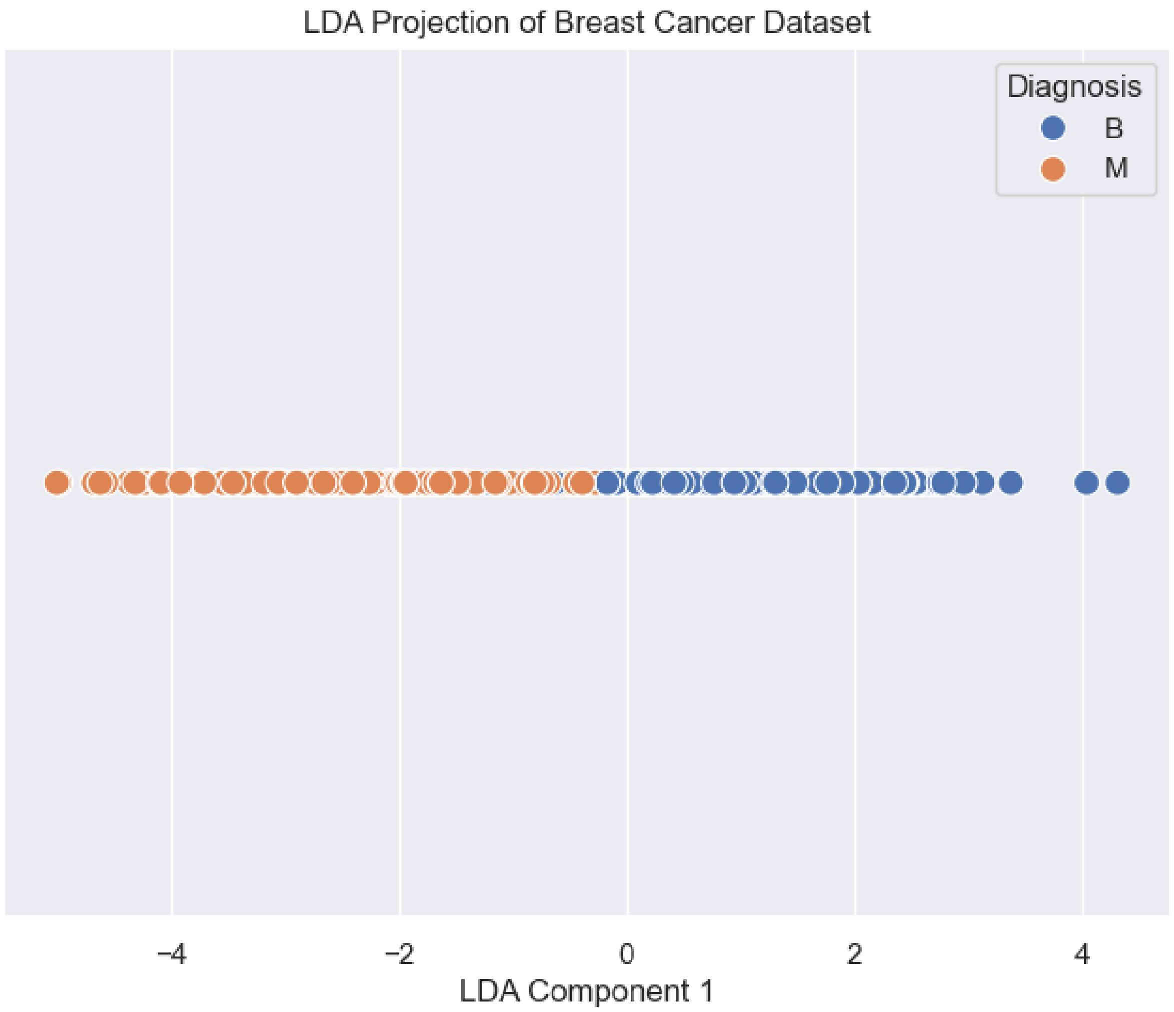
- Reducing the dimensionality of high-dimensional data while preserving local structure

Figures/Diagrams

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


PCA with 1 component
0.04
0.02
0.00
−0.02
−0.04
Class labels
−5
0
5
10
15
Principal Component 1


classification
B
M
15
10
5
0
−5
−10
−15
Y
−20
−10
0
10
20
30
X



Challenges Encountered

1. 1. T-SNE needs careful selection of parameters like perplexity and learning rate for good results, or else the outcome may not be accurate.
2. T-SNE can be slow and expensive for large datasets, making it challenging to apply to big data and get quick results.
3. Unlike PCA, which preserves global structure, T-SNE focuses on local structure, making it harder to understand the results, especially for complex high-dimensional data.
4. Comparing T-SNE with LDA can be challenging because LDA is a supervised method that relies on class labels, whereas T-SNE is unsupervised, making it difficult to evaluate their performance on the same dataset.

Conclusion

- In conclusion, T-SNE is a powerful tool for dimensionality reduction, but it requires careful parameter selection and can be slow for large datasets.

- Unlike PCA, which preserves global structure, T-SNE focuses on local structure, providing a unique perspective on the data.
- While LDA is a supervised method that excels in certain tasks, T-SNE's unsupervised nature makes it a valuable addition to any data analyst's toolkit.
- Ultimately, the choice between T-SNE, PCA, and LDA depends on the specific needs of the project and the characteristics of the data.