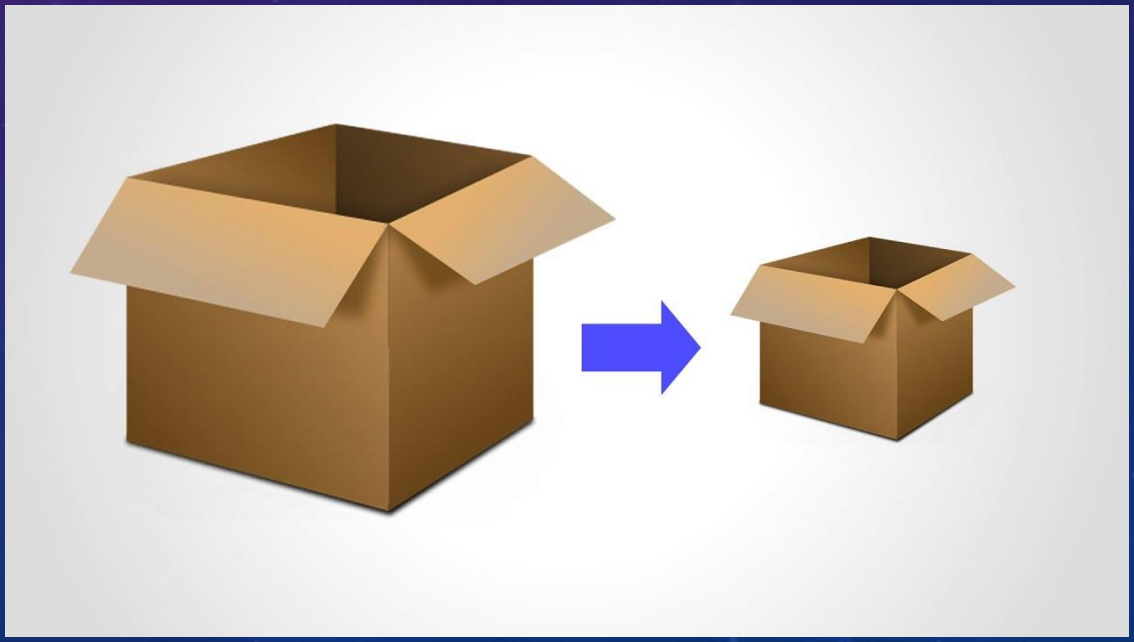
DIMENSIONALITY REDUCTION

Presented By: Rohit Prajapati
Date :2020-09-15

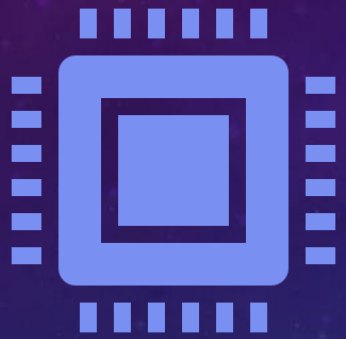
CONTENTS

WHAT IS DIMENSIONALITY REDUCTION?

- Dimensionality reduction is the technique of reducing the number of features in a data set.
- Dimensionality reduction is the process of reducing the number of variables under consideration, by obtaining a set of principal variables.



WHY DIMENSIONALITY REDUCTION?



Higher number of feature needs more space and computation.

Curse of Dimensionality

CURSE OF DIMENSIONALITY

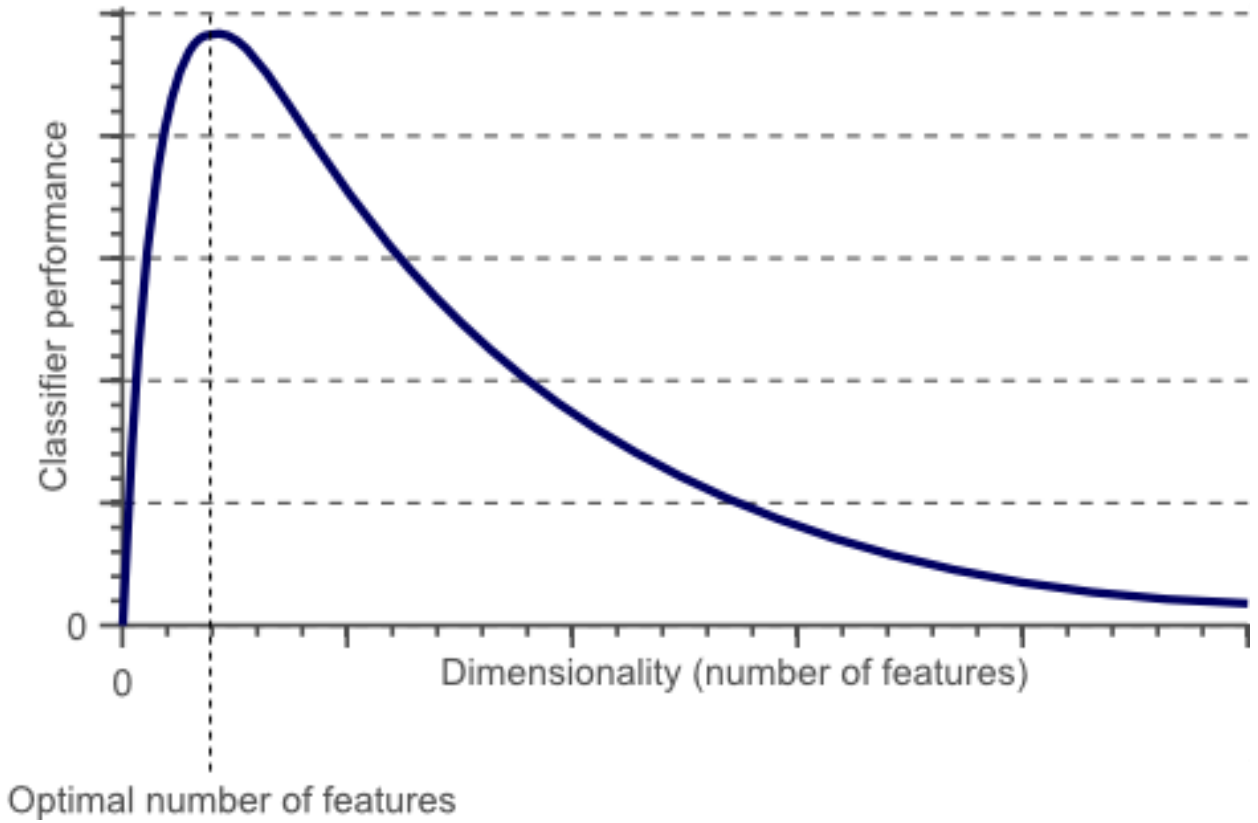


Figure 1: The curse of Dimensionality

CURSE OF DIMENSIONALITY

It refers to data having too many features meaning a vector of high dimension.

High features leads to overfitting of our model.

Visualization and observation of data becomes difficult.

The model becomes more complex.

TECHNIQUES OF DIMENSIONALITY REDUCTION

Feature Selection Method

Linear Dimensionality Reduction Method

Non-linear Dimensionality Reduction Method

Auto Encoders

FEATURE SELECTION METHOD

- Feature selection is the process of identifying and selecting relevant features for your sample.
- Feature engineering is manually generating new features from existing features, by applying some transformation or performing some operation on them.
- Manually or programmatically
  - Variance Threshold
  - Univariate Selection

LINEAR DIMENSIONALITY REDUCTION

- Most common and well-known method that apply linear transformation for dimensionality reduction.

PCA (Principle Component Analysis)

Factor Analysis

LDA (Linear Discriminant Analysis)

PCA (Principle Component Analysis)

PCA rotates and projects data along the direction of increasing variance. The features with the maximum variance are the principal components.

Factor Analysis

Reduce features into factors.
The observed data are expressed as function to know which is more important.

LDA (Linear Discriminant Analysis)

Projects data in a way that the class separability is maximised.

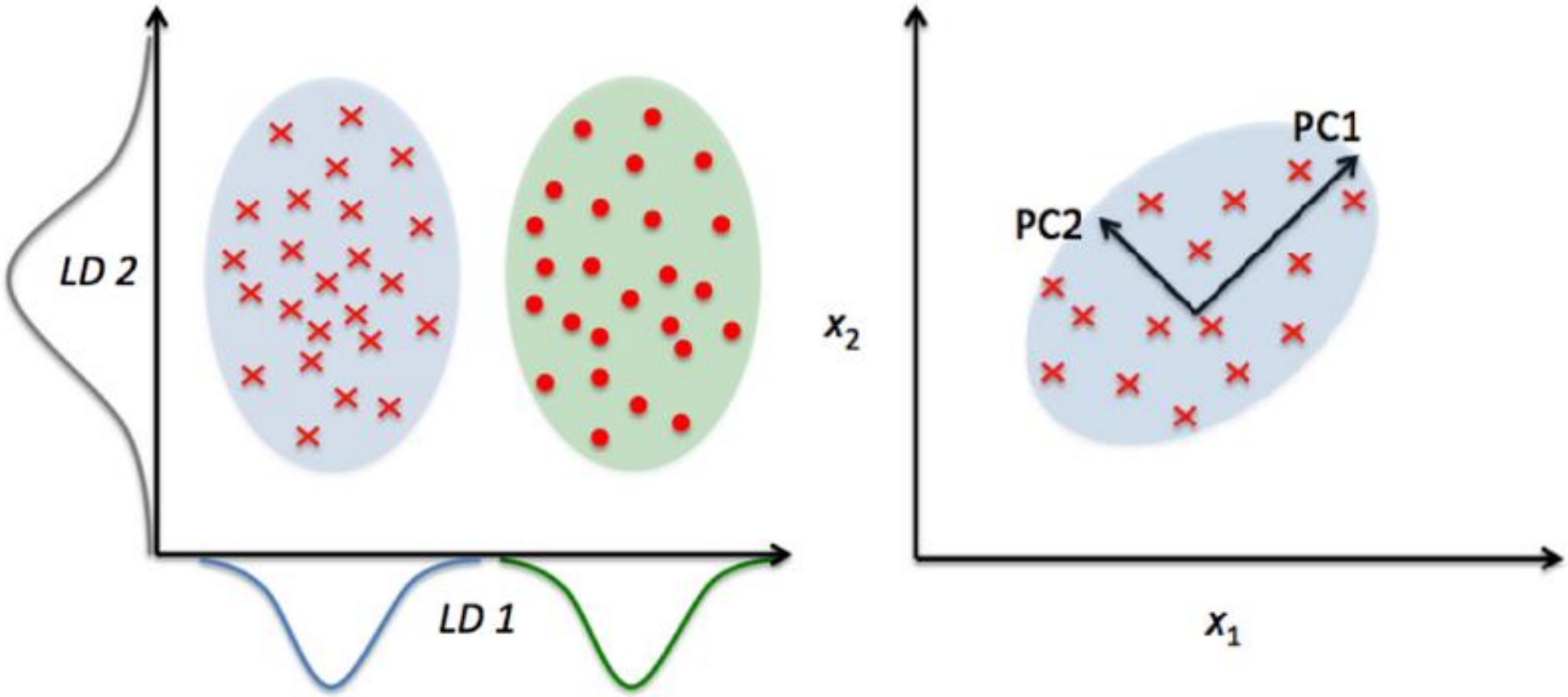


Figure 2: PCA and LDA

PCA orients data along the direction of the component with maximum variance whereas LDA projects the data to signify the class separability

NON-LINEAR DIMENSIONALITY REDUCTION METHOD

- It is based on the manifold hypothesis which says that in a high dimensional structure, most relevant information is concentrated in small number of low dimensional manifolds.

MULTI-DIMENSIONAL SCALING (MDS)

ISOMETRIC FEATURE MAPPING (ISOMAP)

LOCALLY LINEAR EMBEDDING (LLE)

HESSIAN EIGENMAPPING (HLLE)

SPECTRAL EMBEDDING (LAPLACIAN EIGENMAPS)

T-DISTRIBUTED STOCHASTIC NEIGHBOR EMBEDDING (T-SNE)

NON-LINEAR DIMENSIONALITY REDUCTION

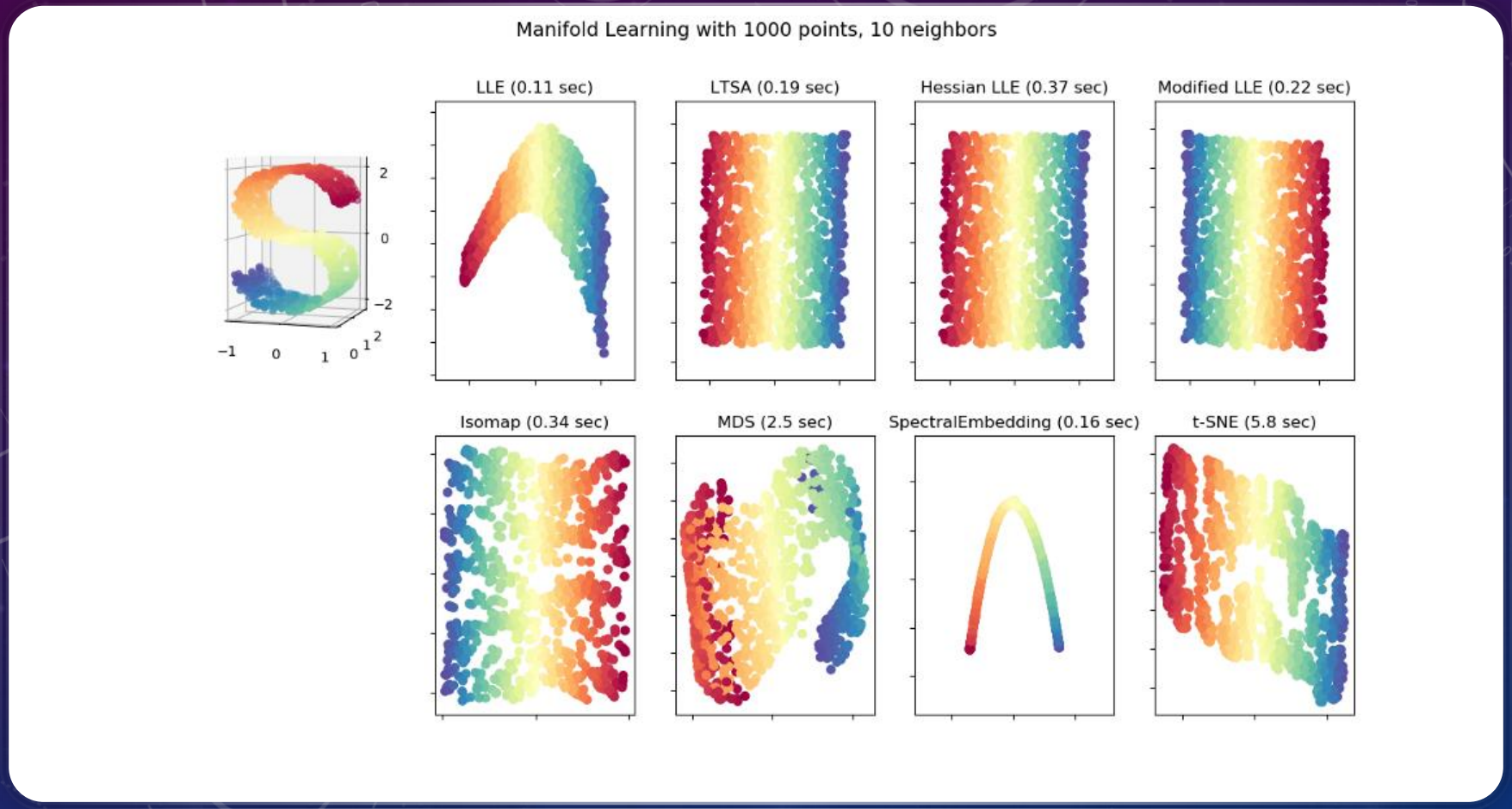
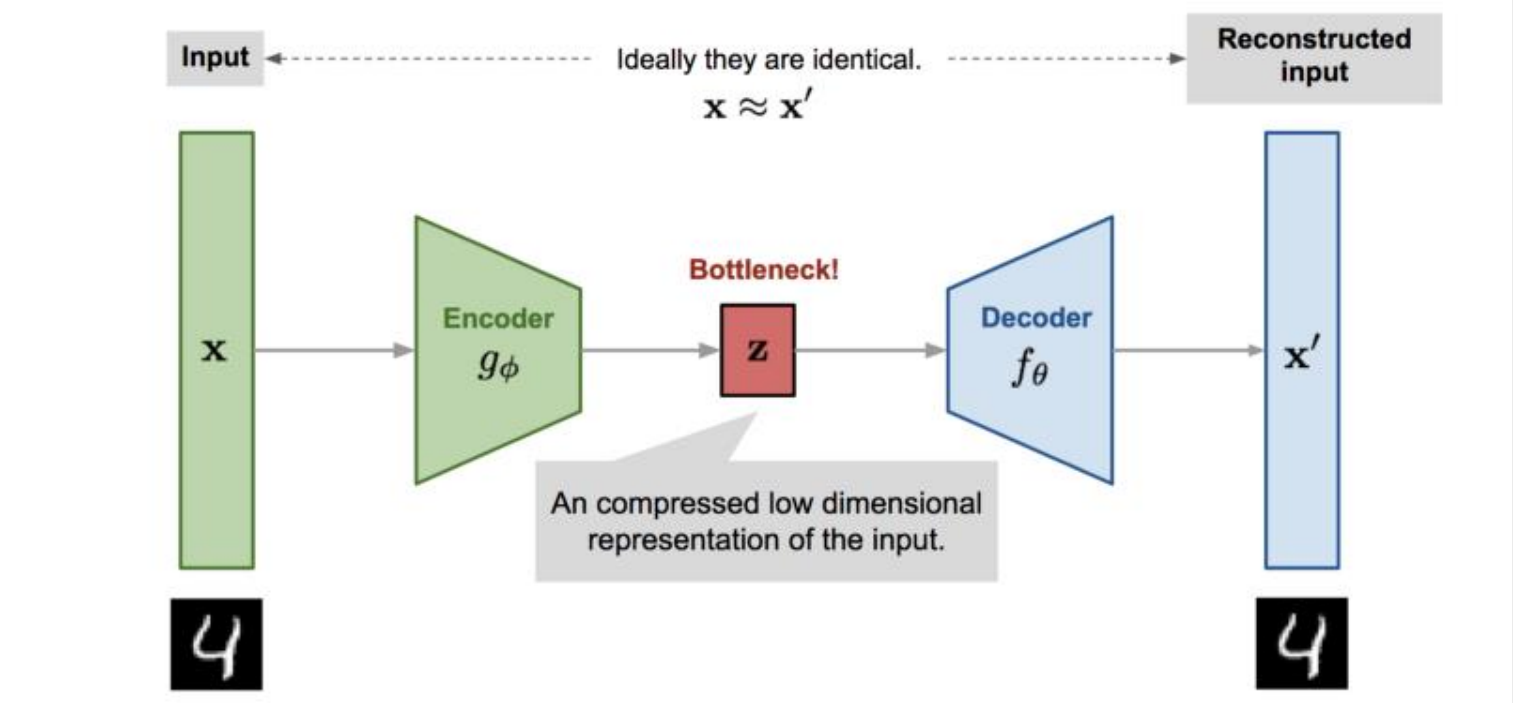


Figure 3: Shows the resulting projection from applying different manifold learning methods on a 3D S-Curve

AUTOENCODERS

- A type of artificial neural network that aims to copy their inputs to their outputs.
- They compress the input into a **latent-space representation**



ADVANTAGES

- LESS MISLEADING DATA MEANS MODEL ACCURACY IMPROVES.
- LESS DIMENSIONS MEAN LESS COMPUTING. LESS DATA MEANS THAT ALGORITHMS TRAIN FASTER.
- LESS DATA MEANS LESS STORAGE SPACE REQUIRED.
- LESS DIMENSIONS ALLOW USAGE OF ALGORITHMS UNFIT FOR A LARGE NUMBER OF DIMENSIONS
- REMOVES REDUNDANT FEATURES AND NOISE.

THANK YOU!

REFERENCES

https://towardsdatascience.com/dimensionality-reduction-for-machine-learning-80a46c2ebb7e
https://machinelearningmastery.com/dimensionality-reduction-for-machine-learning/
https://towardsdatascience.com/the-curse-of-dimensionality-50dc6e49aa1e
https://www.geeksforgeeks.org/dimensionality-reduction/