Dimensionality Reduction and it's Types

Gauri Sharan

October 27, 2023

Outline

1. Introduction
2. Advantages and Disadvantages
3. Dimensionality Reduction Types
4. Feature Selection
5. Feature Extraction
6. Conclusion

Introduction

- Dimensionality reduction is a technique used in machine learning and data analysis to reduce the number of features (not samples) in a dataset.
- It simply refers to the process of reducing the number of attributes in a dataset while keeping as much of the variation in the original dataset as possible.
- It is a data preprocessing step - we perform dimensionality reduction before training the model.

Why "Dimension"?

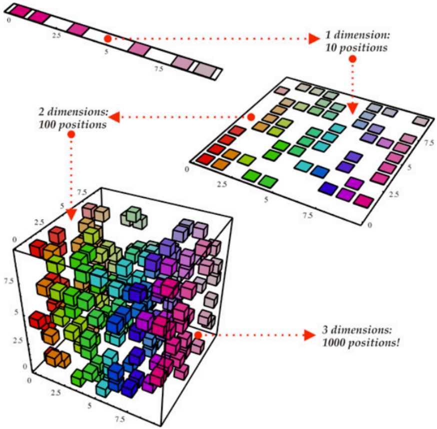


Figure: From 3D to 2D to 1D

Why "Dimension"?

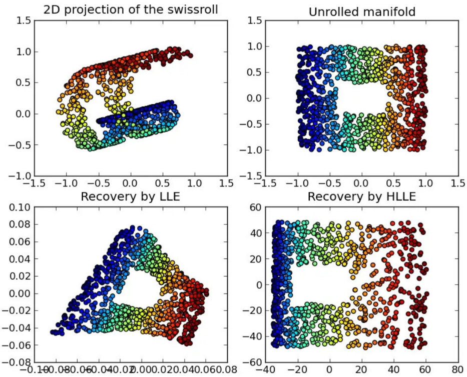


Figure: Unrolling swissroll dataset from 3D to 2D

Benefits of applying Dimensionality Reduction

- Reduces chances of overfitting by reducing complexity
- By reducing the dimensions of the features, the space required to store the dataset also gets reduced.
- Less Computation training time is required for reduced dimensions of features.
- Reduced dimensions of features of the dataset help in visualizing the data quickly.
- It removes the redundant features (if present) by taking care of multicollinearity.

Disadvantages of applying Dimensionality Reduction

- Some data may be lost due to dimensionality reduction.
- In the PCA dimensionality reduction technique, sometimes the principal components required to consider are unknown.
- Many dimensionality reduction methods make certain assumptions about the data distribution or the nature of the relationships between variables. If these assumptions are not met, the techniques may not perform well.

Types of Dimensionality Reduction

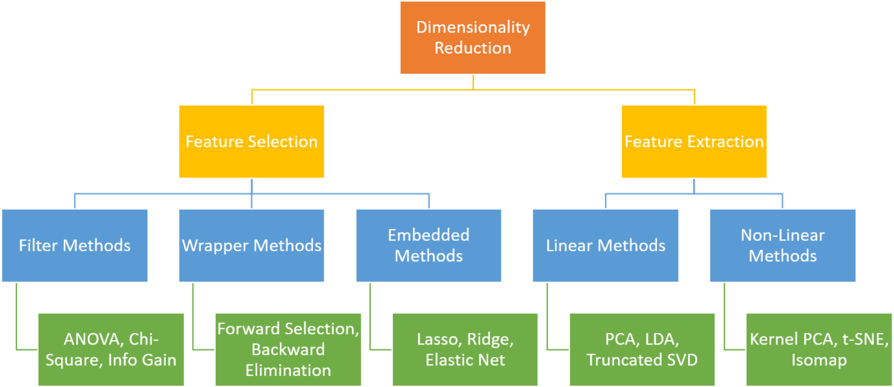


Figure: Dimensionality Reduction Types

Feature Selection

Feature selection is the process of selecting the subset of the relevant features and leaving out the irrelevant features present in a dataset to build a model of high accuracy. In other words, it is a way of selecting the optimal features from the input dataset.

Types of Selection Methods -

- Filter methods -
  The dataset is filtered, and a subset that contains only the relevant features is taken.
  Common techniques - Chi-Square Test, ANOVA, Information Gain, etc.

Feature Selection

- Wrapper methods -
  In this method, some features are fed to the ML model, and evaluate the performance. The performance decides whether to add those features or remove to increase the accuracy of the model. This method is more accurate than the filtering method but complex to work.
  Common Techniques - Forward Selection, Backward Selection, Bi-directional Elimination etc.

- Embedded methods -
  They check the different training iterations of the machine learning model and evaluate the importance of each feature.
  Common Techniques - LASSO, Elastic Net, Ridge Regression, etc.

Feature Extraction

Feature extraction techniques transform the original features into a lower-dimensional representation.

Common techniques:

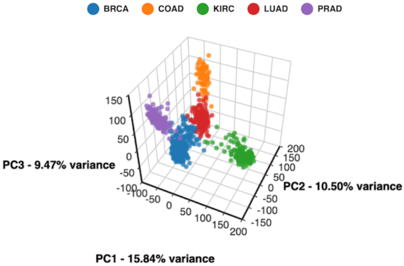
- Linear Methods -
  - Principal Component Analysis (PCA)
  - Linear Discriminant Analysis (LDA)
  - Truncated Singular Value Decomposition (SVD) etc.
- Non-Linear Methods -
  - Kernel PCA
  - t-distributed Stochastic Neighbor Embedding (t-SNE)
  - Isomap etc.

Principal Component Analysis

Principal Component Analysis is a statistical process that converts the observations of correlated features into a set of linearly uncorrelated features with the help of orthogonal transformation. These new transformed features are called the Principal Components. It is one of the popular tools that is used for exploratory data analysis and predictive modeling.

Principal Component Analysis

PCA works by considering the variance of each attribute because the high attribute shows the good split between the classes, and hence it reduces the dimensionality. Some real-world applications of PCA are image processing, movie recommendation system, optimizing the power allocation in various communication channels.



Graphical Examples

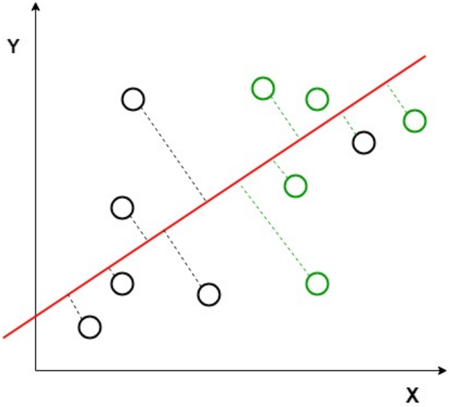


Figure: LDA Visual

Graphical Examples

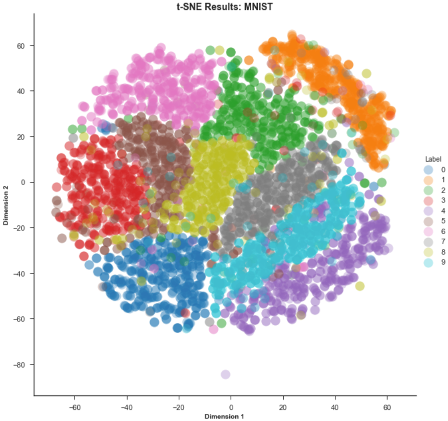
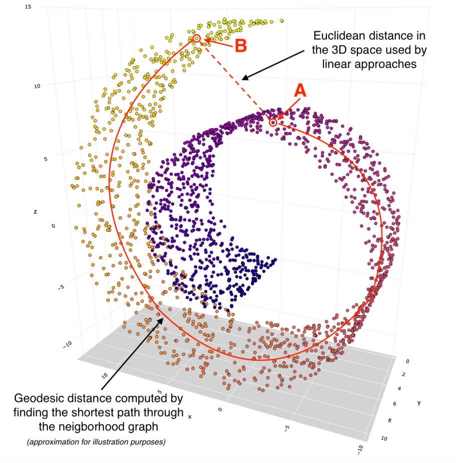


Figure: t-SNE Visual

Graphical Examples



Euclidean distance in the 3D space used by linear approaches

Geodesic distance computed by finding the shortest path through the neigborhood graph

(approximation for illustration purposes)

Figure: Isomap Visual

Conclusion

In conclusion, dimensionality reduction is a crucial tool in data analysis and machine learning for managing high-dimensional data and improving model efficiency and interpretability.