Unsupervised Learning
Association Rule & Problem Simplification

Zeham Management Technologies BootCamp

by SDAIA

August 7th, 2024

SDAIA
الهيئة السعودية للبيانات والذكاء الاصطناعي
Saudi Data & AI Authority

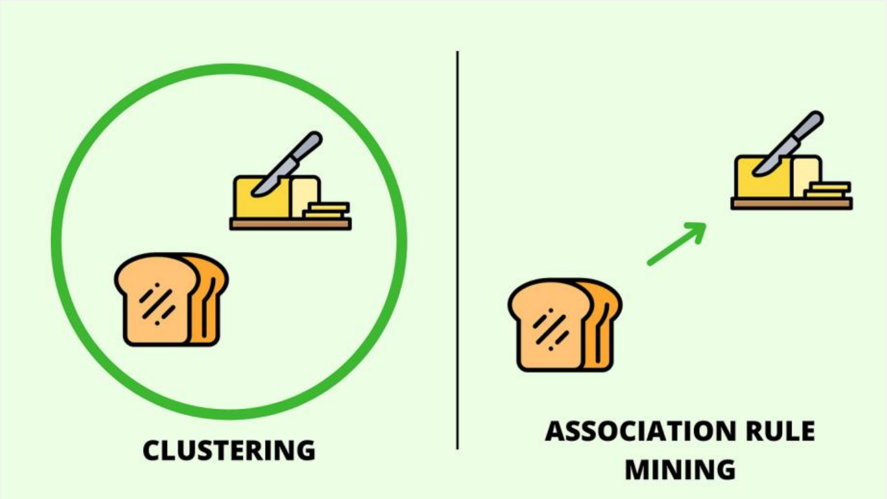
Agenda

- Association Rule.
- Problem Simplification.

Association Rule

What is Association rule?

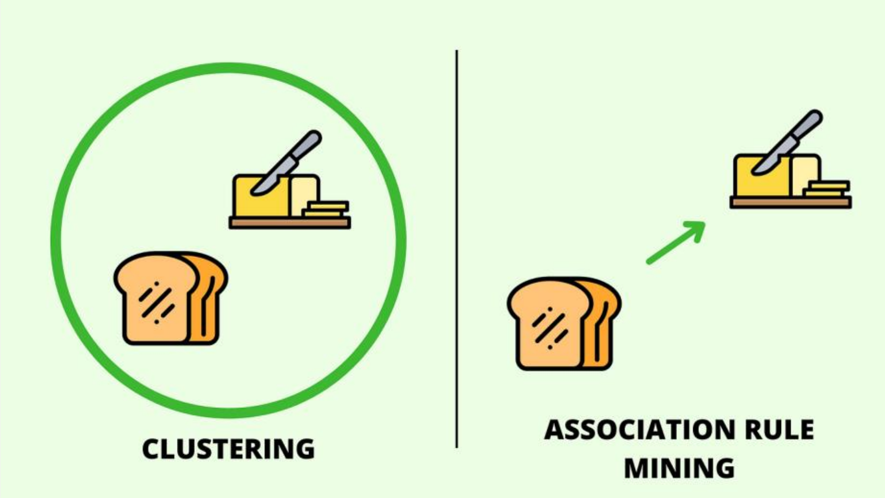
Association rule is a type of unsupervised learning where the algorithm works without a teacher since the data isn't labeled. This method is more about describing data rather than predicting outcomes. It's typically used to find interesting patterns or connections in large datasets. These patterns are often shown as rules or frequent itemsets.



Source

What is Association rule?

Association rule mining helps find new and interesting connections between different items in a dataset, like frequent patterns in transaction data or relational databases. It's often used for things like Market Basket Analysis (to see which items are bought together), Customer clustering in retail (to find which stores people visit together), Price Bundling, and more. It's a more advanced way of exploring "what if" scenarios, like "if this happens, then that might happen."



Source

How Association Rule Works

To understand how association rules work, it's important to be familiar with a few key terms.

- **Itemset:** A collection of one or more items. For example, in a grocery store dataset, an itemset could be {milk, bread}.
- **Support:** This measures how frequently an itemset appears in the dataset. It's defined as the proportion of transactions that include the itemset.

$$\text{Support(X)}=\frac{Number\ of\ transactions\ containing\ X}{Total\ number\ of\ transactions}$$

How Association Rule Works Cont.

To understand how association rules work, it's important to be familiar with a few key terms.

- **Confidence:** This measures the likelihood that an item Y is bought when item X is bought. It is defined as the proportion of transactions containing both X and Y among transactions containing X.
- **Lift:** This evaluates the strength of a rule compared to the expected frequency of X and Y occurring together by chance. It is calculated as:

Formulas

$$\text{Confidence}(X \rightarrow Y) = \frac{Support(X \cup Y)}{Support(X)}$$

$$\text{Lift}(X \longrightarrow Y) = \frac{Confidence(X \Rightarrow Y)}{Support(Y)}$$

Or equevelantly:

$$\text{Lift}(X \longrightarrow Y) = \frac{Confidence(X \Rightarrow Y)}{Support(X)\ x\ Support(Y)}$$

How Association Rule Works Cont.

To understand how association rules work, it's important to be familiar with a few key terms.

- **Rule:** An association rule is expressed as $X \rightarrow Y$, meaning that the presence of itemset X implies the presence of itemset Y.

Association Rule Example

Let's consider a dataset of transactions from a grocery store. Suppose you have the following transactions:

1. {milk, bread}
2. {milk, diapers}
3. {bread, diapers}
4. {milk, bread, diapers}

Association Rule Example Calculations (Support)

- **Support:** To find the support for {milk, bread}, we count the number of transactions that include both milk and bread. In this case, there are 3 transactions out of 4 that include both milk and bread. So,

$$\text{Support}(\{\text{milk, bread}\}) = \frac{3}{4} = 0.75$$

Association Rule Example Calculations (Confidence)

- **Confidence:** To find the confidence of the rule {milk} → {bread}, we use the formula:

$$\text{Confidence}(\{\text{milk}\}\rightarrow\{\text{bread}\})=\frac{Support(\{milk,bread\})}{Support(\{milk\})}$$

The support of {milk} is 3/4 (since milk appears in 3 transactions). Thus:

$$\text{Confidence}(\{\text{milk}\}\rightarrow\{\text{bread}\}) =\frac{0.75}{0.75} = 1$$

Lift: To calculate the lift of the rule {milk} → {bread}:

$$\text{Lift}(\{\text{milk}\}\longrightarrow\{\text{bread}\})=\frac{Confidence(\{milk\}\Rightarrow\{bread\})}{Support(\{bread\})}$$

The support of {bread} is 3/4. Thus:

$$\text{Lift}(\{\text{milk}\}\longrightarrow\{\text{bread}\})=\frac{1}{0.75} = 1.33$$

Association Rule Example Results

The lift value of 1.33 indicates that the presence of milk increases the likelihood of buying bread by 33% compared to what would be expected by chance alone.

How Does Association Rule Works?

Let's now practice:

Tutorial:

Advanced Machine Learning/ 4- Unsupervised Learning Association Rule & Problem Simplification/LAB/PCD.ipynb

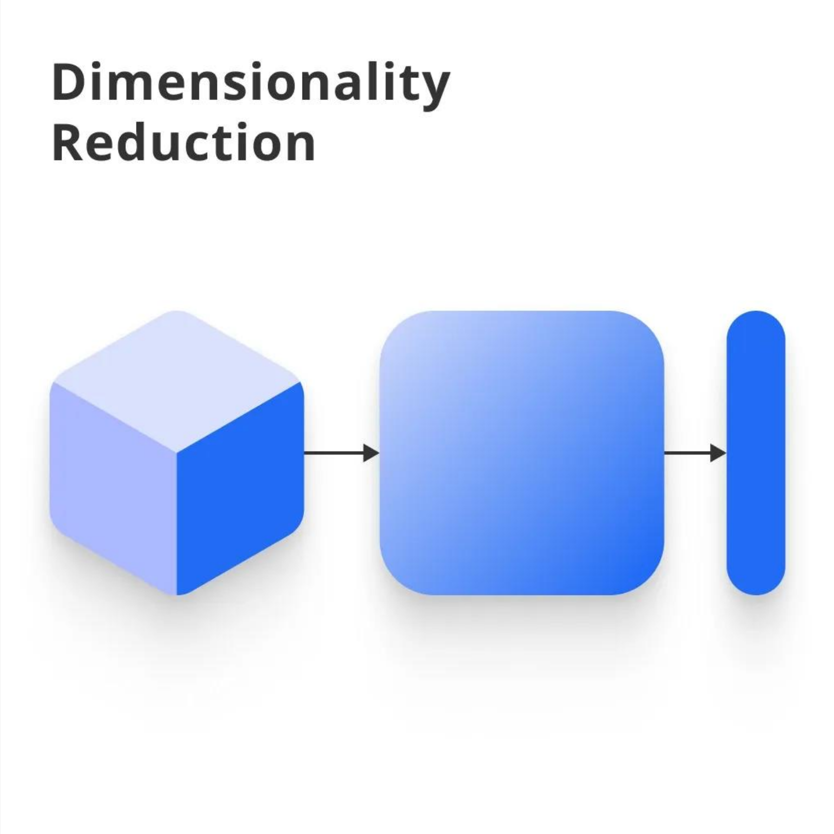
Exercise:

Advanced Machine Learning/ 4- Unsupervised Learning Association Rule & Problem Simplification/LAB/PCD_Exercise.ipynb

Problem Simplification with Dimensionality Reduction

Introduction to Dimensionality Reduction

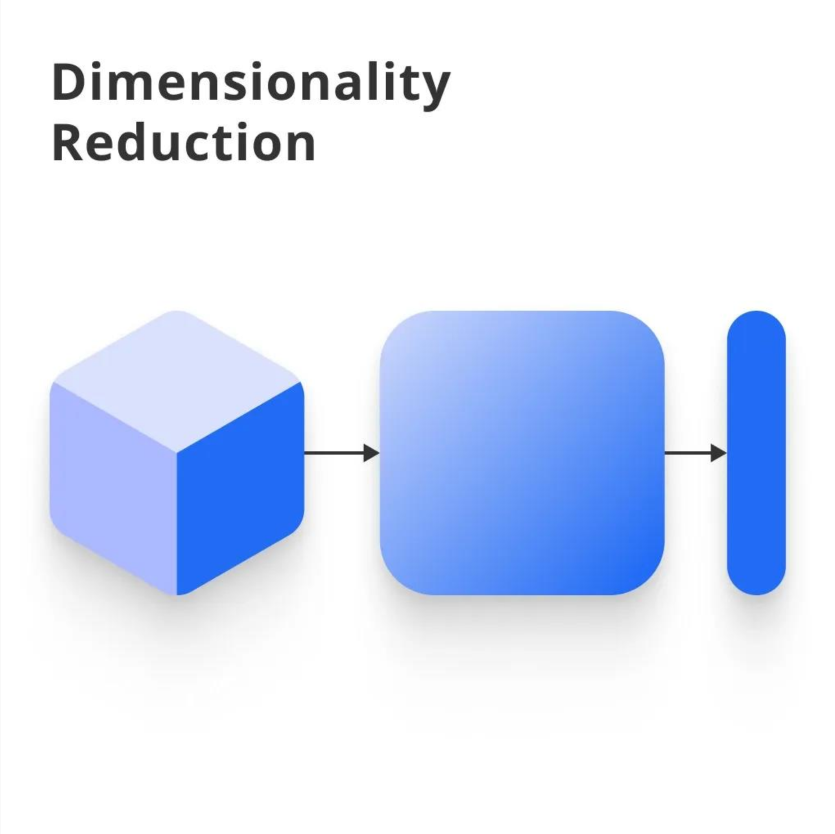
- Dimensionality reduction simplifies complex datasets by reducing the number of features.
- Helps in visualizing, analyzing, and modeling data more effectively.
- Common techniques: PCA (Principal Component Analysis) and SVD (Singular Value Decomposition).



Source

Why Use Dimensionality Reduction ?

- Reduces computational cost and storage requirements.
- Helps in removing noise and redundant features.
- Simplifies data for easier interpretation and visualization.



Source

Principal Component Analysis(PCA)

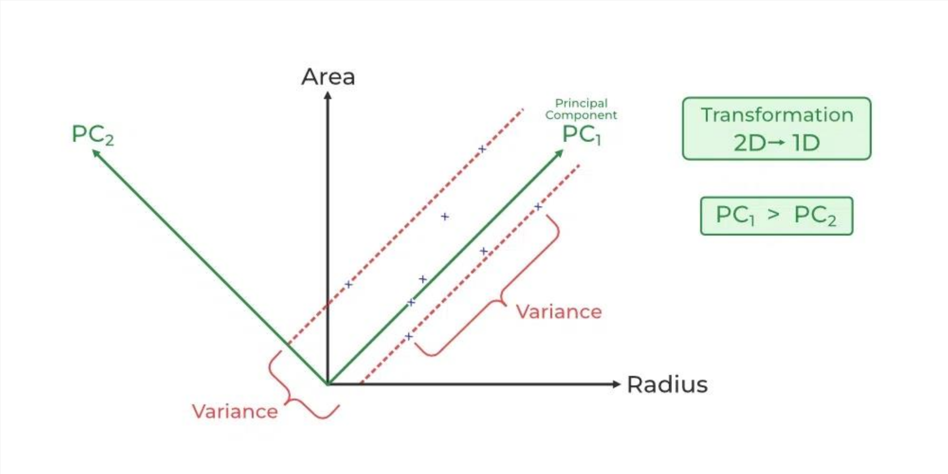
What is Principal Component Analysis(PCA)?

Principal Component Analysis (PCA), introduced by Karl Pearson in 1901, reduces high-dimensional data to lower dimensions while maximizing variance in the lower space.

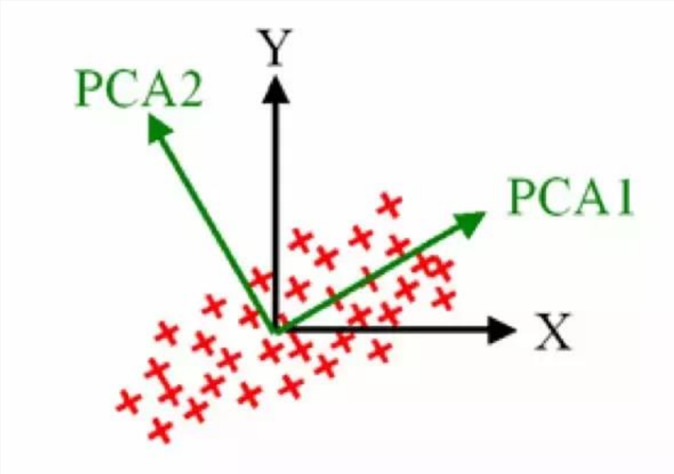
- Principal Component Analysis (PCA) is a statistical method that uses orthogonal transformation to turn correlated variables into uncorrelated ones. It is widely used in exploratory data analysis and machine learning.
- PCA is an unsupervised technique to explore relationships among variables, also known as general factor analysis, where regression finds a best-fit line.
- The main goal of PCA is to reduce a dataset's dimensions while preserving key patterns or relationships without prior knowledge of target variables.

What is Principal Component Analysis(PCA)?

Principal Component Analysis (PCA) reduces a dataset's dimensions by finding a smaller set of variables that retain most of the original information, useful for regression and classification.



Source



Source

What is Principal Component Analysis(PCA)?

1. Principal Component Analysis (PCA) reduces dimensions by identifying orthogonal axes, or principal components, that capture the most variance. These components are linear combinations of the original variables and are ranked by importance. The total variance of the principal components equals that of the original dataset.

2. The first principal component captures the most variance, while the second captures the maximum variance orthogonal to the first, and so on.

What is Principal Component Analysis(PCA)?

3. PCA is used for data visualization, feature selection, and data compression. It helps plot high-dimensional data in 2D or 3D for easier interpretation, identifies key variables in feature selection, and reduces dataset size in data compression while retaining important information.

4. PCA assumes that more variance in a feature indicates more information, so features with higher variation carry more information.

PCA Explanation

Step 1: Standardization

Step2: Covariance Matrix Computation

Step 3: Compute Eigenvalues and Eigenvectors of Covariance Matrix to Identify Principal Components

PCA Explanation (Standardization)

First, we need to standardize the dataset so that each variable has a mean of 0 and a standard deviation of 1.

$$Z = \frac{X - \mu}{\sigma}$$

PCA Explanation (Covariance Matrix Computation)

Covariance measures how two or more variables change together, showing their joint variability. The formula for covariance is:

$$\text{conv}(x1, x2) = \frac{\sum_{i=1}^{n}(x1_i - \bar{x}1)(x2_i - \bar{x}2)}{n-1}$$

Covariance can be positive, negative, or zero:

- Positive: As x_1 increases, x_2 also increases.
- Negative: As x_1 increases, x_2 decreases.
- Zero: No direct relationship.

PCA Explanation (Compute Eigenvalues and Eigenvectors of Covariance Matrix to Identify Principal Components)

Let A be a square $n*n$ matrix and X be a non-zero vector for which

$$AX = \lambda X$$

For some scalar value λ then λ is the eigenvalue of matrix A, and X is the eigenvector of A for that eigenvalue.

This can also be written as:

$$AX - \lambda X = 0$$
$$(A - \lambda I)X = 0$$

where I is the identity matrix with the same dimensions as matrix A. The conditions hold true if $(A - \lambda\ I)$ is non-invertible (i.e., a singular matrix). This means:

$$|A - \lambda I| = 0$$

From the above equation, we can find the eigenvalues (λ), and the corresponding eigenvector can be found using the equation $AX = \lambda X$.

How Principal Component Analysis(PCA) Works?

Let's now practice:

Tutorial:

Advanced Machine Learning/ 4- Unsupervised Learning Association Rule & Problem Simplification/LAB/PCD.ipynb

Exercise:

Advanced Machine Learning/ 4- Unsupervised Learning Association Rule & Problem Simplification/LAB/PCD_Exercise.ipynb

Singular Value Decomposition (SVD)

Singular Value Decomposition (SVD)

Singular Value Decomposition (SVD) is a mathematical process that breaks down a matrix into three simpler parts. This breakdown reveals important information about the matrix, such as its properties and how it transforms data. SVD is widely used in various fields, including data analysis.

Mathematics behind SVD:

The SVD of mxn matrix A is given by the formula $A = U \sum V^T$

where:

U: mxm matrix of the orthonormal eigenvectors of AA^T

V^T: transpose of a nxn matrix containing the orthonormal eigenvectors of $A^T A$

$\sum$: diagonal matrix with r elements equal to the root of the positive eigenvalues of AA^T or $A^T A$ (both matrics have the same positive eigenvalues anyway).

Example of SVD:

Find the SVD for the matrix $A = \begin{bmatrix} 3 & 2 & 2 \\ 2 & 3 & -2 \end{bmatrix}$

Example of SVD:

First, we need to compute the singular values by finding eigenvalues of AA^{T}.

$$A \cdot A^T = \begin{bmatrix} 3 & 2 & 2 \\ 2 & 3 & -2 \end{bmatrix} \cdot \begin{bmatrix} 3 & 2 \\ 2 & 3 \\ 2 & -2 \end{bmatrix} = \begin{bmatrix} 17 & 8 \\ 8 & 17 \end{bmatrix}$$

The characteristic equation for the above matrix is:

W-λI=0

$AA^T - \lambda I = 0$

$\lambda^2 - 34\lambda + 225 = 0$

$= (\lambda - 25)(\lambda - 9)$

so our singular values are: $\sigma_1 = 5$; $\sigma_2 = 3$

Example of SVD:

Now we find the right singular vectors i.e orthonormal set of eigenvectors of $A^T A$.

The eigenvalues of $A^T A$ are 25, 9, and 0, and since $A^T A$ is symmetric we know that the eigenvectors will be orthogonal.

For λ=25,

$$A^T A - 25 \cdot I = \begin{bmatrix} -12 & 12 & 2 \\ 12 & -12 & -2 \\ 2 & -2 & -17 \end{bmatrix}$$ which can be row-reduces to : $$\begin{bmatrix} 1 & -1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

Example of SVD:

A unit vector in the direction of it is:

$$v1 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \\ 0 \end{bmatrix}$$

Similarly, for λ = 9, the eigenvector is:

$$v2 = \begin{bmatrix} \frac{1}{\sqrt{18}} \\ \frac{-1}{\sqrt{18}} \\ \frac{-4}{\sqrt{18}} \end{bmatrix}$$

Example of SVD:

For the 3rd eigenvector, we could use the property that it is perpendicular to v1 and v2 such that:

$$v1^T v3 = 0$$

$$v2^T v3 = 0$$

Solving the above equation to generate the third eigenvector:

$$v3 = \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} a \\ -a \\ -a/2 \end{bmatrix} = \begin{bmatrix} \frac{2}{3} \\ \frac{-2}{3} \\ \frac{-1}{3} \end{bmatrix}$$

Example of SVD:

Now, we calculate U using the formula: $U = \frac{1}{\sigma} * A \ * V$

and this gives $U = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{bmatrix}$

Hence, our final SVD equation becomes:

$$A = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} 5 & 0 & 0 \\ 0 & 3 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{18}} & \frac{-1}{\sqrt{18}} & \frac{4}{\sqrt{18}} \\ \frac{2}{3} & \frac{-2}{3} & \frac{1}{3} \end{bmatrix}$$

Applications of SVD:

Dimensionality reduction: SVD can be used to reduce the dimensionality of data while preserving as much information as possible.

Image compression: By reducing the rank of an image matrix, SVD can significantly compress image data without losing too much quality

Comparing PCA and SVD

PCA	SVD
Focuses on maximizing variance.	Decomposes matrices into singular values.
Works well with dense data.	Effective for sparse and large datasets.

Both techniques simplify data, but choose based on specific dataset needs.

How Does SVD works?

Let's now practice:

Tutorial:

Advanced Machine Learning/ 4- Unsupervised Learning Association Rule & Problem Simplification/LAB/SVD.ipynb

Exercise:

Advanced Machine Learning/ 4- Unsupervised Learning Association Rule & Problem Simplification/LAB/SVD_Exercise.ipynb

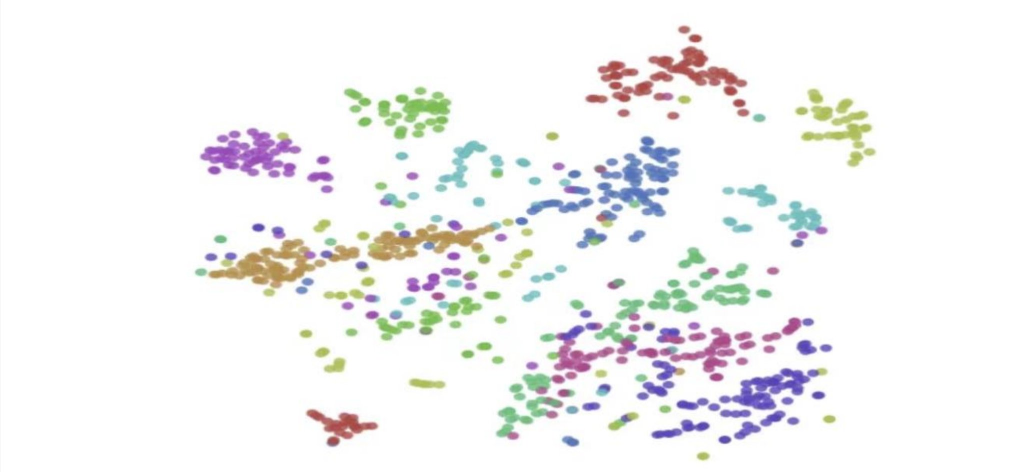
t-Distributed Stochastic Neighbor Embedding (t-SNE)

t-Distributed Stochastic Neighbor Embedding (t-SNE)

t-SNE is a technique that simplifies complex data by representing it in a lower-dimensional space, typically two or three dimensions.

This allows for easier visualization and understanding of data, especially when dealing with non-linear relationships between data points.

It's particularly useful for exploring and analyzing complex datasets in fields like machine learning and data science.



Core idea of t-SNE

- Maps high-dimensional data to a lower-dimensional space (usually 2D or 3D).
- Preserves local relationships between data points.
- Calculates similarity between data points in the high-dimensional space.
- Represents similarities as probabilities.
- Creates a similar probability distribution in the lower-dimensional space.
- Minimizes the difference between the two probability distributions.
- Effectively captures local structure of the data.
- Useful for visualizing and understanding complex datasets.

Why learn t-SNE, since we have already PCA technique for dimensionality reduction ?

Preserving local relationships in t-SNE means maintaining the proximity of similar data points when transforming high-dimensional data into a lower-dimensional space.

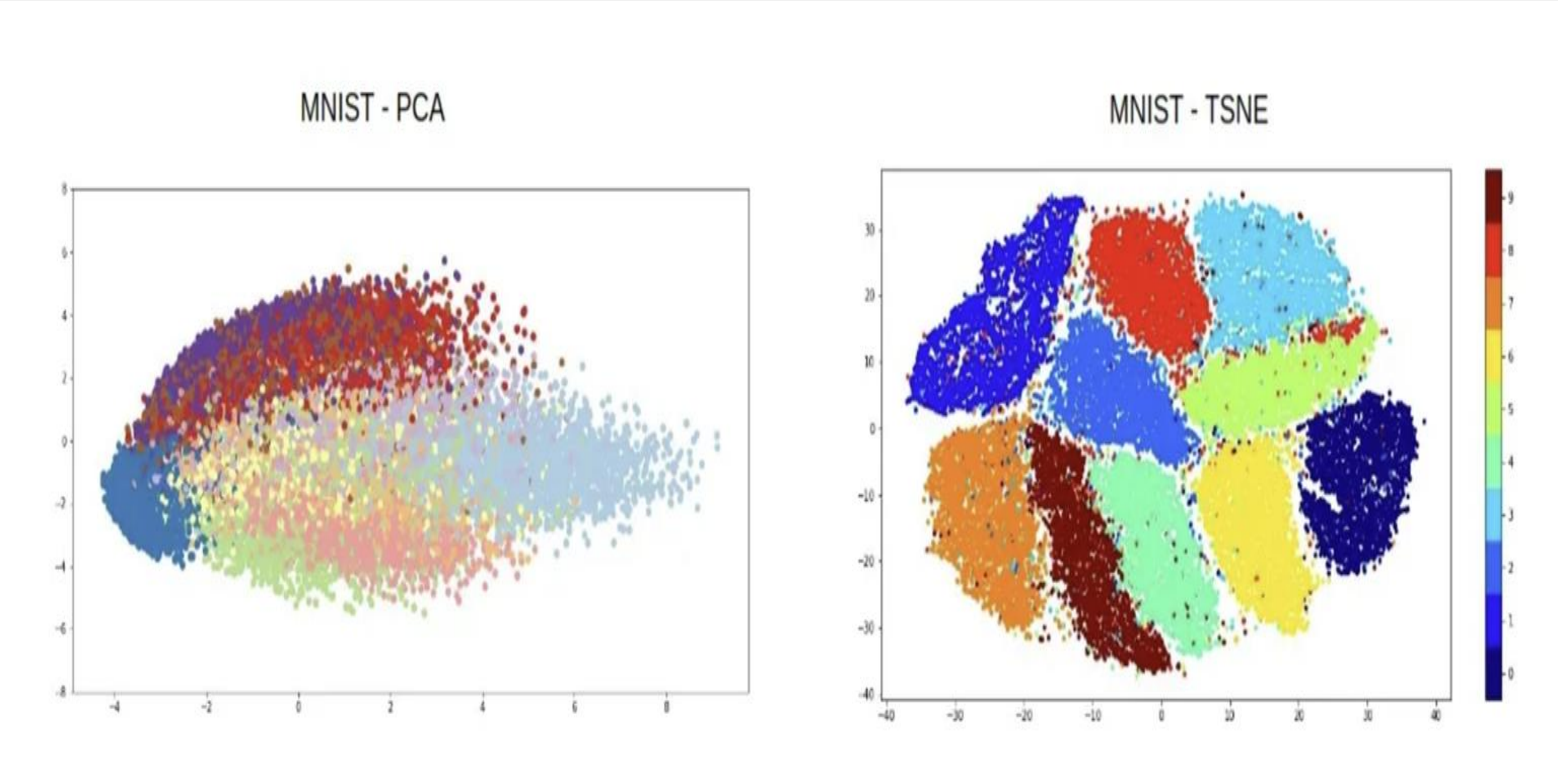
Essentially, points that are close together in the original dataset should remain close after dimensionality reduction, while distant points should stay apart.

This ensures that the underlying structure of the data is preserved in the visualization.

Why learn t-SNE, since we have already PCA technique for dimensionality reduction ?

t-SNE	PCA
Non-linear dimensionality reduction technique.	Linear dimensionality reduction technique.
Focuses on preserving small pairwise distances (local similarities).	Best for data with a linear structure.
Better for capturing complex relationships in data.	Identifies principal components by projecting data onto lower dimensions.
Ideal for visualizing high-dimensional data.	Preserves large pairwise distances (variance).

Why learn t-SNE, since we have already PCA technique for dimensionality reduction ?



How Does t-SNE Works?

Let's now practice:

Tutorial:

Advanced Machine Learning/ 4- Unsupervised Learning Association Rule & Problem Simplification/LAB/t-SNE.ipynb

Exercise:

Advanced Machine Learning/ 4- Unsupervised Learning Association Rule & Problem Simplification/LAB/t-SNE_Exercise.ipynb

Independent Component Analysis (ICA)

Independent Component Analysis (ICA)

Purpose of ICA: Separates a complex signal into its individual, independent components.

Method: ICA uses statistical and computational techniques to find a linear transformation of the data.

Goal: To maximize the statistical independence of the transformed components.

Advantages Independent Component Analysis (ICA)

Signal Separation: ICA is a technique that separates mixed signals into their individual components, which is useful in a variety of applications such as signal processing.

Non-Parametric Approach: ICA doesn't make assumptions about the data's distribution, making it flexible and reliable method.

Advantages Independent Component Analysis (ICA)

Unsupervised Learning: ICA doesn't require pre-labeled data, making it adaptable to various datasets.

Feature Extraction: ICA can identify crucial patterns within data, which can be helpful for tasks like sorting and categorizing information.

Disadvantages Independent Component Analysis (ICA)

Non-Gaussian Assumption: ICA works best when the original signals have non-normal distributions. If the signals follow a normal (Gaussian) pattern, ICA might not be effective.

Linear Mixing Assumption: ICA assumes that the signals are combined in a simple, linear way. If the signals are mixed in a more complex, nonlinear manner, ICA might not produce accurate results.

Disadvantages Independent Component Analysis (ICA)

High computational cost: ICA requires a lot of computing power, especially when dealing with large amounts of data. This can make it difficult to use ICA in real-world situations

Difficulty finding a solution: ICA may have trouble finding a solution, especially when the data is complex and has many sources. This can be a problem when using ICA in practical applications.

Benefits of Dimensionality Reduction

- Simplifies data for easier analysis and visualization.
- Reduces computational cost and storage.
- Removes noise and irrelevant features.
- Enhances model performance by reducing overfitting.

Conclusion

- Dimensionality reduction techniques like PCA and SVD help simplify complex datasets.
- They capture the essential features while reducing the number of dimensions.
- Experiment with both techniques to determine the best fit for your data.

How Does Independent Component Analysis (ICA) Works?

Let's now practice:

Tutorial:

Advanced Machine Learning/ 4- Unsupervised Learning Association Rule & Problem Simplification/LAB/ICA.ipynb

Exercise:

Advanced Machine Learning/ 4- Unsupervised Learning Association Rule & Problem Simplification/LAB/ICA_Exercise.ipynb