3iA Côte d'Azur
Interdisciplinary Institute for Artificial Intelligence

Data Analysis and Visualization

CentraleDigitalLab@Nice

Deborah Dore - PhD - ddore@i3s.unice.fr
MARIANNE, I3S, CNRS, INRIA, Université Côte d'Azur

Dimensionality Reduction

Overview

Dimensionality reduction is the process of reducing the number of random variables under consideration.

In other words, we use dimensionality reduction techniques to **transform an high-dimensional space into a low dimensional space** while retaining important information.

Dimensionality Reduction

Overview

Dimensionality reduction is the process of reducing the number of random variables under consideration.

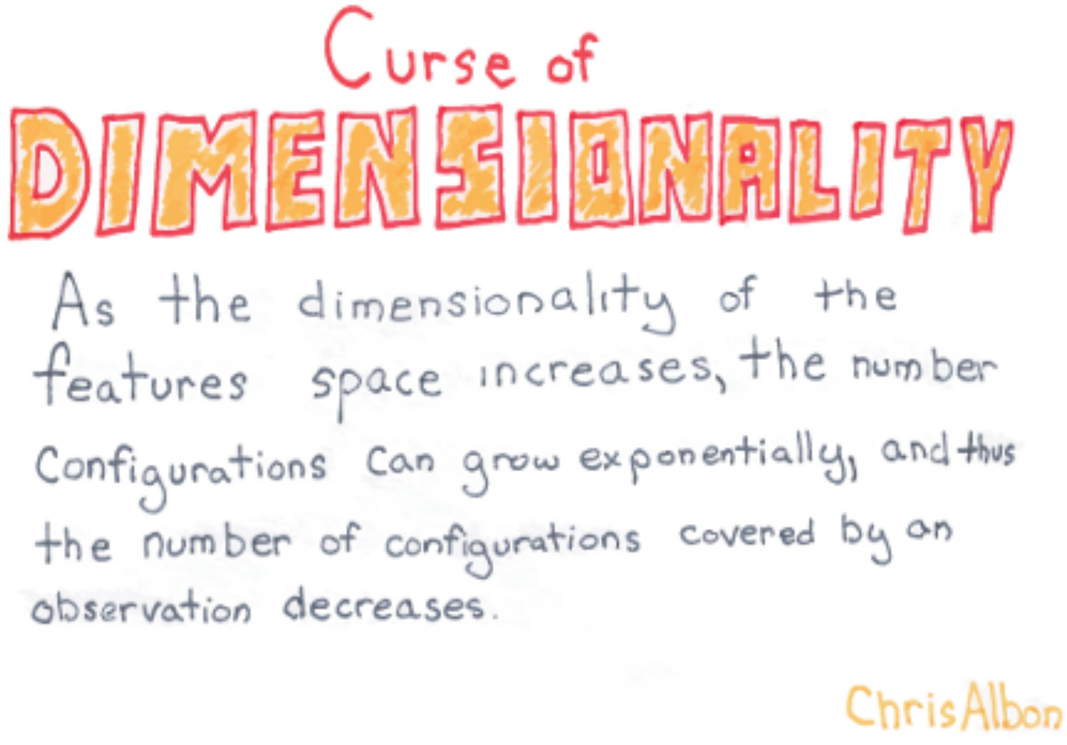
Why is it important?

- Simplifies data visualisation
- Reduces computational costs
- Mitigates the *curse of dimensionality*
- Improves model's performances by reducing noise

Dimensionality Reduction

Curse of Dimensionality

Coined by mathematician Richard E. Bellman, the **curse of dimensionality** references increasing data dimensions and its explosive tendencies. This phenomenon typically results in an increase in computational efforts required for its processing and analysis.



Dimensionality Reduction

Overview

Dimensionality reduction techniques:

- **Linear Methods:** PCA, MDS
- **Non Linear Methods:** LLE, Isomaps, t-SNE
- **Hybrid Approach**

Key Questions:

- *What is the underlying structure of the data?*
- *Is interpretability or performance the goal?*

Dimensionality Reduction

Principal Component Analysis (PCA)

PCA is a linear dimensionality reduction method that transforms data into a set of uncorrelated components (**principal components**) ordered by variance

The ***principal components*** are linear combinations of the original variables in the dataset and are ordered in decreasing order of importance

The total variance captured by all the principal components is equal to the total variance in the original dataset

The first principal component captures the **most variation** in the data, but the second principal component captures the maximum variance that is **orthogonal** to the first principal component, and so on

Dimensionality Reduction

Principal Component Analysis (PCA)

PCA is a linear dimensionality reduction method that transforms data into a set of uncorrelated components (principal components) ordered by variance

Possibile applications:

1. Image compressing
2. Noise reduction
3. EDA

Dimensionality Reduction

Principal Component Analysis (PCA)

PCA is a linear dimensionality reduction method that transforms data into a set of uncorrelated components (principal components) ordered by variance

Possibile applications:

1. Image compressing
2. Noise reduction
3. EDA ← **Can you tell me why?**

Dimensionality Reduction

PCA's Steps

First Step. First we need to standardise the data:

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

Where: μ is the mean of independent features $\{\mu_1, \mu_2, \mu_3 \ldots \mu_n\}$

σ is the standard deviation of independent features $\{\sigma_1, \sigma_2, \sigma_3 \ldots \sigma_n\}$

Dimensionality Reduction

PCA's Steps

Second Step. We need to compute the covariance matrix:

The **covariance** measures the strength of joint variability between two or more variables, indicating how much they change in relation to each other

$$cov(x, y) = \frac{\sum_{i=1}^{n} (x_i - \mu_x)(y_i - \mu_y)}{n - 1}$$

The result can be:

- **positive:** *as x increases also y increases*
- **negative**: *as x increases, y decreases*
- **zero:** no direct relation

Dimensionality Reduction

PCA's Steps

Third Step. Calculate the eigenvalues and the eigenvectors

Let A be a **square** NxN **matrix** and X be a **non-zero vector** for which:

$$AX = \lambda X \text{ for some scalar values } \lambda.$$

λ **is known as the eigenvalue of matrix** A and X **is known as the eigenvector of matrix** A for the corresponding eigenvalue.

It can also be written as :

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

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

where I is the identity matrix of the same shape as matrix A

And the above conditions will be true only if $(A - \lambda I)$ will be non-invertible (i.e. singular matrix). That means: $|A - \lambda I| = 0$

From the above equation we can find the eigenvalues λ and therefore the corresponding eigenvectors can be found using the equation $AX = \lambda X$

Dimensionality Reduction

PCA's Steps

Fourth Step. Select top components

Dimensionality Reduction

PCA Advantages and Limitations

Advantages:

- Simple and computationally efficient
- Retains maximum variance in reduced dimensions

Limitations:

- Assumes linear relationships
- Sensitive to outliers
- Interpretation of principal components can be challenging

Dimensionality Reduction

Multidimensional Scaling (MDS)

Overview: **MDS** is a technique that maps high-dimensional data to a lower-dimensional space by preserving pairwise distances between points

MDS is based on the concept of distance and aims to find a projection of the data that minimizes the differences between the distances in the original space and the distances in the lower-dimensional space

Dimensionality Reduction

Multidimensional Scaling (MDS)

Overview: **MDS** is a technique that maps high-dimensional data to a lower-dimensional space by preserving pairwise distances between points

Applications:

- Visualizing dissimilarity matrices
- Clustering analysis

Dimensionality Reduction

MDS' Math

The mathematical foundation of **MDS** is the stress function, which **measures the difference between the distances in the original space and the distances in the lower-dimensional space:**

$$stress = \sqrt{\frac{1}{2n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} (d_{ij} - \hat{d}_{ij})^2}$$

Where: d_{ij} is the **distance** between data points i and j in the original space

$\hat{d}_{ij}$ is the **distance** between i and j in the lower dimensional space

n is the number of data points

The stress function is a measure of the deviation of the distances in the lower-dimensional space from the distances in the original space and is used to evaluate the quality of the projection

Dimensionality Reduction

Locally Linear Embedding (LLE)

LLE is an unsupervised approach designed to transform data from its original high-dimensional space into a lower-dimensional representation, all while striving to retain the essential geometric characteristics of the underlying non-linear feature structure

Applications:

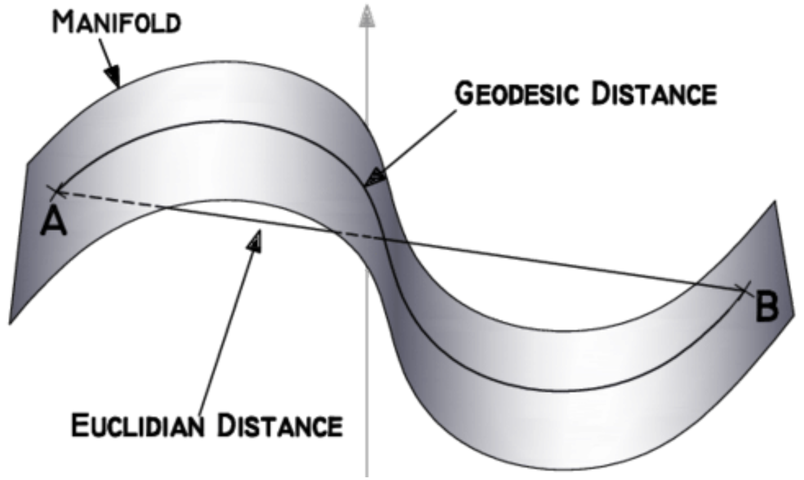
- Image processing
- Speech recognition

Dimensionality Reduction

Isomap

Isomap is a non-linear technique that extends MDS by preserving geodesic distances (distances along a manifold) instead of Euclidean distances.

It was developed as an alternative to conventional methods like Principal Component Analysis (PCA) for preserving the intrinsic geometry of complex datasets.



Dimensionality Reduction

Isomap

Isomap is a non-linear technique that extends MDS by preserving geodesic distances (distances along a manifold) instead of Euclidean distances.

Key features:

1. **Geodesic distance preservation**: Isomap aims to maintain the geodesic distances between data points in the lower-dimensional representation

2. **Neighborhood graph construction**: The algorithm starts by creating a neighborhood network to approximate the manifold structure of the data

3. **Graph distance calculation**: It uses graph distances to estimate geodesic distances between all pairs of points

4. **Eigenvalue decomposition**: Isomap performs eigenvalue decomposition on the geodesic distance matrix to find the low-dimensional embedding

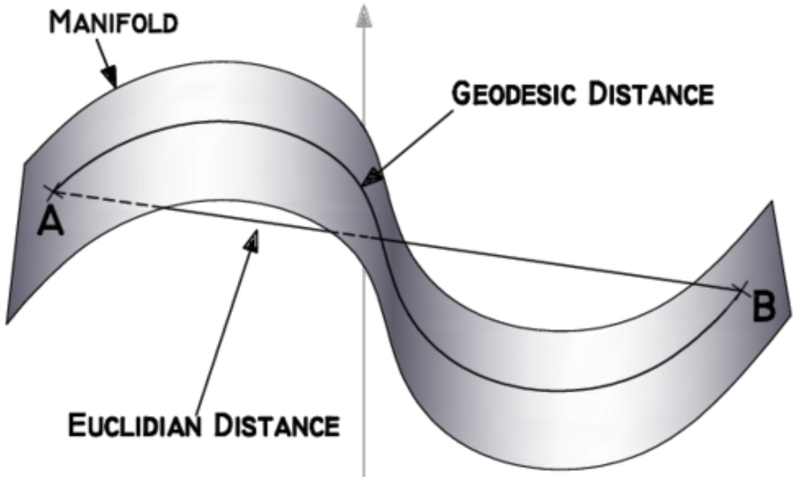
Dimensionality Reduction

Isomap

Isomap is a non-linear technique that extends MDS by preserving geodesic distances (distances along a manifold) instead of Euclidean distances.

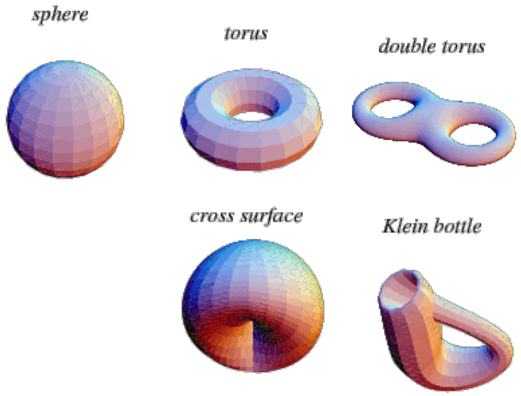
Applications:

- Analyzing intrinsic structures in data



Dimensionality Reduction

What is a manifold?



A **Manifold** is a topological space that locally resembles Euclidean space near each point.

Key characteristics:

- **Local flatness**: When zoomed in sufficiently, they appear flat like Euclidean space
- **Global complexity**: While locally simple, manifolds can have intricate global properties
- **Dimensionality**: Manifolds can have various dimensions, from one-dimensional curves to higher-dimensional spaces

Dimensionality Reduction

T-Distributed Stochastic Neighbor Embedding (t-SNE)

t-SNE is a nonlinear dimensionality reduction technique used for visualizing high-dimensional data in two or three dimensions

The t-SNE algorithm works in two main stages:

1. It constructs a probability distribution over pairs of high-dimensional objects, assigning higher probabilities to similar objects and lower probabilities to dissimilar ones

2. It defines a similar probability distribution over the points in the low-dimensional map and minimizes the Kullback-Leibler divergence between the two distributions

Dimensionality Reduction

T-Distributed Stochastic Neighbor Embedding (t-SNE)

t-SNE is a nonlinear dimensionality reduction technique used for visualizing high-dimensional data in two or three dimensions

Key features:

- It preserves local structures in the data, allowing for better visualization of clusters and patterns
- It uses gradient descent to optimize the lower-dimensional embedding
- The algorithm is stochastic, meaning multiple runs can produce different results
- It has a time complexity of $O(n^2)$ and space complexity of $O(n^2)$ for n data points

Dimensionality Reduction

Choosing the Right Technique

To Consider:

- Linear vs. non-linear relationships.
- Size of the dataset.
- Objective: visualization vs. feature reduction.

Rule of thumb :

- PCA: Linear data and interpretability.
- t-SNE: Visualizing clusters.
- Isomap/LLE: Manifold learning

Demo with Notebook_Dimensionality_reduction.ipynb