BIOS 835: Dimensions Reduction

Alexander McLain

October 2, 2023

Outline

Introduction

Principal Components

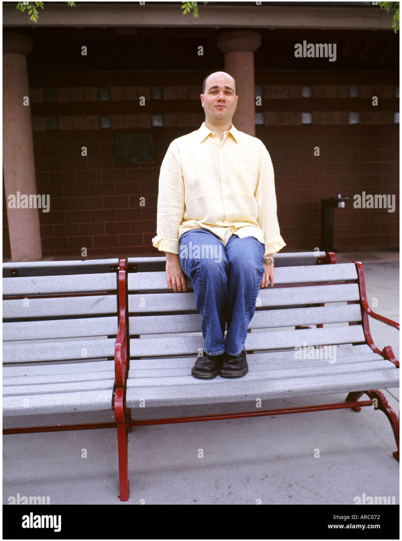
Multidimensional Scaling (MDS)

Isomap

t-SNE

Other methods

Conclusions


alamy
Image ID: ARC072
www.alamy.com

Why Dimension reduction?

- Dimension reduction refers to the process of reducing the number of random variables under consideration by obtaining a set of *principal variables*.
- It's a common step in data pre-processing to remove redundancy and irrelevant information, thereby helping in visualization, improving performance, and reducing computational costs.
    - **Feature Elimination** → simply eliminating irrelevant features.
    - **Feature Extraction** → transforming the data from a high-dimensional space into a space of fewer dimensions.

Why Dimension reduction?

- For our discussion here, we'll focus on feature extraction. Feature elimination, aka feature selection, will be discussed throughout the rest of the course.
- Also, we won't discuss the following supervised methods in this section
  - Linear Discriminant Analysis
  - Fisher's Linear Discriminant
  - Canonical Correlation Analysis

Local vs Global structure

- Dimension reduction techniques can be categorized based on how they preserve the *local* and *global* structures of the data.
- **Local structure** refers to the relationships between neighboring data points in the original high-dimensional space.
- Techniques that emphasize local structure aim to ensure that nearby points in the original space remain close to each other in the reduced-dimensional space.
- **Global structure** considers the broader relationships between data points that might be spread across the entire dataset.
- It aims to capture the data's overall trends and large-scale patterns.

Local structure

- Local structure focuses on capturing the fine-grained details and intricate patterns within clusters or groups of similar data points,
- preserving local structure is essential for retaining the inherent relationships between data points that are close to each other,
- techniques that emphasize local structure aim to ensure that nearby points in the original space remain close to each other in the reduced-dimensional space

Local structure example

- An example where local structure is desired:
  - Imagine you have a dataset of handwritten digits and want to reduce its dimensionality for visualization.
  - Each digit belongs to a particular class (0 to 9). In the original high-dimensional space, digits of the same class tend to cluster together, but there are subtle variations within each class due to different writing styles.
  - A dimensionality reduction technique that emphasizes local structure, would create a reduced-dimensional plot where digits of the same class are clustered closely together, highlighting the local relationships and similarities between digits.

Local structure examples

- **Gene Expression Data Analysis**: Biological data, such as gene expression profiles, may exhibit complex local patterns representing different cellular states or tissue types. Preserving local structure can be essential for identifying these patterns, leading to insights into biological processes or disease mechanisms.
- **Drug Discovery**: In chemoinformatics, preserving local structure might assist in clustering similar molecular structures, leading to the identification of new potential drug candidates.
- **Text Analysis and Document Clustering**: In natural language processing, preserving local structure can capture semantic similarities between closely related documents.
- **Customer Segmentation**: In marketing, customer data may contain complex local relationships that define micro-segments or niches. Preserving these local structures can help in creating more targeted marketing strategies.

Global structure

- Global structure considers the broader relationships between data points that might be spread across the entire dataset.
- It aims to capture the data's overall trends and large-scale patterns.
- Preserving global structure is important to understand the overall distribution and behavior of the data points, even when they are distant from each other in the high-dimensional space.
- Techniques that emphasize global structure ensure that the general layout and arrangement of data points are maintained in the reduced-dimensional space.

Global structure example

- Consider a dataset representing population data from different countries.
- Each data point represents a country, and the dimensions represent various socio-economic indicators.
- In the original high-dimensional space, countries with similar socio-economic profiles might be scattered across the space, but they might follow certain global trends, such as a correlation between GDP and life expectancy.
- A dimensionality reduction technique that emphasizes global structure could.
  - create a reduced-dimensional representation where countries with similar overall socio-economic profiles are projected near each other
  - capture the major global trends, like countries with high GDP being located in one direction and countries with low GDP in another direction.

Global structure examples (cont)

- Other examples where the global structure is desired
  - **Financial Data Analysis**: Understanding the overall correlation structure between different stocks or assets is crucial in financial markets.
  - **Climate Studies**: Analyzing global weather patterns requires understanding large-scale relationships between different climate variables across regions.
  - **Social Network Analysis**: In studying social networks, understanding the overall community structure and how different groups are connected may be more important than focusing on individual connections.
  - **Medical Imaging**: In some medical imaging applications like MRI or CT scans, capturing the overall structure of the organ or tissue may be more relevant than focusing on local details.

Global and local structure

- One example where both global and local structure are desired is with single-cell RNA sequencing (scRNA-seq) data.
- **Local structure**
  - When analyzing scRNA-seq data, researchers are often interested in identifying subtle relationships between individual cells, such as small clusters representing specific cell subtypes in a developmental pathway.
  - In scRNA-seq, the local structure often captures biologically meaningful variations related to cell-type-specific gene expression patterns.
- **Global structure**
  - On the other hand, the global structure in scRNA-seq data might represent broader relationships between different cell types, tissues, or conditions.
  - Preserving global structure can be important when the focus is on understanding the overall relationships between clusters or large groups within the data.

Global and local structure

- The choice between local and global structure preservation depends on the characteristics of your data and the goals of your analysis.
- In some cases, preserving local structure might be more important, especially when dealing with clusters or non-linear relationships.
- In other cases, capturing global trends and distributions might take precedence, such as when visualizing the overall data distribution.
- A balanced approach might involve using a combination of techniques that capture both local and global structure.

Local structure methods

Examples of techniques that emphasize local structure include:

- **t-Distributed Stochastic Neighbor Embedding (t-SNE)**: t-SNE focuses on preserving pairwise similarities between data points, particularly emphasizing the local similarities in high-dimensional space. It is commonly used for data visualization and cluster analysis.
- **Locally Linear Embedding (LLE)**: LLE reconstructs each data point as a linear combination of its neighbors, emphasizing the local relationships. It aims to capture the underlying manifold structure.
- **Local PCA**: In cases where the data distribution is not globally linear, local PCA calculates principal components within small neighborhoods, capturing local linear relationships.

Global structure methods

Examples of techniques that emphasize global structure include:

- **Principal Component Analysis (PCA)**: PCA focuses on capturing the directions of maximum variance in the data. It emphasizes the global distribution of data points and projects them onto orthogonal axes that capture the largest variability.
- **Linear Discriminant Analysis (LDA)**: LDA aims to separate classes to maximize class separation. While it considers class relationships, it also captures global distribution differences.
- **Isomap**: Isomap constructs a lower-dimensional embedding that preserves pairwise geodesic distances, capturing the overall global structure of the data manifold.

Methods good for both

Examples of techniques that emphasize both local and global structures in the data (all since 2016):

- **Uniform Manifold Approximation and Projection (UMAP)**: is a nonlinear dimension reduction technique that is especially useful for visualizing high-dimensional data.
- **LargeVis**: provides a balance between preserving distances between neighboring points and maintaining the overall relationships between clusters or groups, making it suitable for tasks such as visualizing complex datasets like large-scale single-cell RNA-sequencing data.
- **PaCMAP (Parameter-free Conformal MAPping)**: is a dimensionality reduction technique designed to preserve both local and global structures. It was introduced as an alternative to existing methods that require careful parameter tuning.

PC Review

Recall the main idea of PC:

- Set $\boldsymbol{Z}_1 = \boldsymbol{X}\boldsymbol{v}_1$ such that $\boldsymbol{v}_1'\Sigma\boldsymbol{v}_1$ is maximized over all $\boldsymbol{v}$ such that $\boldsymbol{v}'\boldsymbol{v} = 1$.
- Set $\boldsymbol{Z}_2 = \boldsymbol{X}\boldsymbol{v}_2$ such that $\boldsymbol{v}_2'\Sigma\boldsymbol{v}_2$ is maximized over all $\boldsymbol{v}$ such that $\boldsymbol{v}'\boldsymbol{v} = 1$ and $Cov(\boldsymbol{Z}_1, \boldsymbol{Z}_2) = 0$.

$\vdots$

- Set $\boldsymbol{Z}_i = \boldsymbol{X}\boldsymbol{v}_i$ such that $\boldsymbol{v}_i'\Sigma\boldsymbol{v}_i$ is maximized over all $\boldsymbol{v}$ such that $\boldsymbol{v}'\boldsymbol{v} = 1$ and $Cov(\boldsymbol{Z}_k, \boldsymbol{Z}_i) = 0$ for all $k = 1, 2, \ldots, i = 1$.

$\vdots$

Advantages of PCA

1. Efficiency
2. Noise Reduction
3. Unsupervised
4. Global structure

Disadvantages of PCA

1. Linearity
2. Loss of Information
3. Sensitivity to Scaling
4. Not Robust to Outliers
5. Lack of Emphasis on Local Structure

Multidimensional Scaling Introduction

- Multidimensional Scaling (MDS) is a technique used for visualizing the similarity or dissimilarity of individual data points in a dataset.
- It aims to find a low-dimensional representation of the data where the distances between points in the low-dimensional space match the given similarities or dissimilarities as closely as possible.
- We'll discuss classical (metric) MDS. Other MDS variants, such as non-metric MDS, aim to preserve the rank order of the distances rather than the actual distances.

MDS Algorithm

1. Distance Matrix
   - Start with a matrix D of pairwise "dissimilarities" (e.g., distances) between n data points (i.e., subjects). The entry D_{ij} represents the dissimilarity between points i and j.
   - D is commonly the Euclidean distance between points.

MDS Algorithm

2. Double Centering: The goal is to convert the dissimilarity matrix into a form from which eigenvalues and eigenvectors can be extracted.
   - First, construct matrix H using the formula:

     $$H = I - \frac{1}{n}11^T$$

     where I is the identity matrix and 1 is a column vector of ones.
   - Then, create the matrix B using:

     $$B = -\frac{1}{2}HD^2H$$

     Here, D^2 is a matrix where each entry D^2_{ij} is the square of D_{ij}.

MDS Algorithm

3. Eigenvalue Decomposition:
   - Decompose B into its eigenvalues and eigenvectors.
   - Let Λ be the diagonal matrix of eigenvalues of B (sorted in descending order) and E be the matrix of corresponding eigenvectors.

MDS Algorithm

4. Low-Dimensional Representation:
   - For a k-dimensional representation (where k is much smaller than n):
   - Take the first k largest eigenvalues in Λ and form the $k \times k$ matrix Λ_k by keeping the top-left $k \times k$ block of Λ.
   - Let E_k be the matrix formed by the first k columns of E.
   - The coordinates of the data points in the k-dimensional space are given by:

     $$X = E_k \Lambda_k^{1/2}$$

     The resulting matrix X provides the coordinates of each data point in the k-dimensional space.
   - The distances between these points in this space should approximate the original pairwise dissimilarities D as closely as possible.

Key Features of MDS

1. Dimensionality Reduction
2. Visual Interpretation
3. Distance Preservation
4. Flexibility
5. Variants

Advantages of MDS

1. Intuitive Visualization: MDS can provide a clear visual representation of complex data, allowing for easier interpretation of relationships among data points.
2. Versatility: It can be applied to a wide range of disciplines and types of data, from market research to biology.
3. No Assumption on Linearity: Unlike methods like PCA, MDS does not assume linear relationships in the data.
4. Handles Non-Euclidean Distances: It's not restricted to Euclidean distances and can work with other types of dissimilarity measures.

Disadvantages of MDS

1. Computationally Intensive: Especially non-metric MDS can be computationally demanding, particularly for large datasets.
2. Sensitivity: The method can be sensitive to noise or outliers in the data, which might distort the resulting configuration.
3. Choice of Distance Metric: The quality and interpretability of MDS results can vary depending on the choice of distance or dissimilarity metric.
4. Lack of Unique Solution: Multiple configurations (mirrored or rotated) can produce almost the same goodness-of-fit for the same data and dissimilarity measure.
5. Interpretability: While MDS provides a visualization, interpreting the axes (dimensions) can sometimes be non-intuitive as they don't correspond to original features but are a construct to represent the data's distances best.

Isomap Introduction

- Overall, Isomap combines graph theory, geodesic distance estimation, and multidimensional scaling to create a lower-dimensional representation that captures the underlying structure of the data.
- The accuracy of the representation depends on factors such as the choice of k (number of neighbors) and the characteristics of the "data manifold", i.e., the underlying or essential structure/shape of the data.
- Isomap aims to capture the underlying structure of high-dimensional data by preserving pairwise distances between data points.

Isomap Algorithm

1. Nearest Neighbor Graph
   - Given a high-dimensional dataset with N data points, Isomap starts by constructing a nearest neighbor graph.
   - For each data point i, the algorithm identifies its k nearest neighbors based on some distance metric (often Euclidean distance). The graph represents the pairwise connectivity between data points.

Isomap Algorithm

1. Nearest Neighbor Graph
   - Given a high-dimensional dataset with N data points, Isomap starts by constructing a nearest neighbor graph.
   - For each data point i, the algorithm identifies its k nearest neighbors based on some distance metric (often Euclidean distance). The graph represents the pairwise connectivity between data points.
2. Geodesic Distance Estimation
   - Capture the intrinsic manifold structure by approximating the geodesic distances between data points. Approximated via a graph of nearest neighbors.
     For each pair of data points i and j, the geodesic distance g_{ij} is approximated by the shortest path distance in the nearest neighbor graph:

$$g_{ij} \approx \text{shortest\_path}(i, j)$$

Isomap Algorithm

3. Distance Matrix
   - The estimated geodesic distances form a distance matrix G of size $N \times N$, where g_{ij} contains the distance between data points i and j.
4. Embedding in Lower-Dimensional Space
   - This is achieved through multidimensional scaling (MDS) on G.
5. Stress Function and Optimization
   - The stress function is used to quantify the discrepancy between the pairwise Euclidean distances g'_{ij} in the reduced space and the original geodesic distances g_{ij}.

Advantages of Isomap

1. Isomap can handle non-linear data distributions and capture complex underlying structures.
2. It is relatively robust to noise and outliers compared to linear methods like PCA.
3. Can be effective for data visualization when the intrinsic geometry of the data is important.

Disadvantages of Isomap

1. Isomap might struggle with datasets that have varying densities or disconnected regions on the manifold.
2. The choice of the number of neighbors (k) can impact the results and requires careful tuning.
3. Computationally more intensive than linear techniques like PCA.

t-SNE Introduction

- t-Distributed Stochastic Neighbor Embedding (t-SNE) is a dimensionality reduction technique used for visualizing high-dimensional data in a lower-dimensional space, often in two dimensions.
- Unlike linear techniques such as PCA, t-SNE focuses on preserving the local structure of the data, making it particularly effective for revealing clusters and patterns in complex data distributions.
- Below is an overview of the key aspects of t-SNE algorithm:

t-SNE Algorithm

1. Similarity Calculation
   - Given a high-dimensional dataset with N data points, t-SNE starts by calculating the pairwise similarities or affinities p_{ij} between all data points i and j.
   - The similarities can be computed using a Gaussian kernel based on the Euclidean distances between data points:

$$p_{ij} = \frac{\exp(-\|x_i - x_j\|^2/(2\sigma_i^2))}{\sum_{k \neq i} \exp(-\|x_i - x_k\|^2/(2\sigma_i^2))}$$

     Here, σ_i is the bandwidth of the Gaussian kernel for data point i, which is often set based on a perplexity parameter.
   - σ_i or the perplexity control the balance between focusing on local versus global structure. It influences the number of neighbors considered for each data point.

t-SNE Algorithm

2. Lower-Dimensional Affinities
    - t-SNE constructs a lower-dimensional similarity matrix q_{ij} for the lower-dimensional embedding.
    - It uses a t-distribution with one degree of freedom (Cauchy distribution) to model the distribution of pairwise distances in the lower-dimensional space:

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq l}(1 + \|y_k - y_l\|^2)^{-1}}$$

Here, y_i and y_j are the lower-dimensional representations of data points i and j.

t-SNE Algorithm

3. Optimization Objective
   - t-SNE aims to find a lower-dimensional representation $Y = \{y_1, y_2, \ldots, y_N\}$ that minimizes the divergence between P and Q:

     $$KL(P\|Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

     This divergence measures how well the lower-dimensional affinities q_{ij} approximate the original similarities p_{ij}.

t-SNE Algorithm

4. Gradient Descent
   - t-SNE minimizes the KL divergence using gradient descent. The details of this are beyond the scope of the course.
5. Initialization
   - t-SNE uses a random initialization for the lower-dimensional coordinates y_i (why "stochastic" in t-SNE's name) and iteratively refines them to minimize the KL divergence.
   - It employs a learning rate that decreases during the optimization to ensure convergence.

Key Features of t-SNE

1. Local Structure Preservation
2. Probability Distributions
3. Mapping Process
4. Perplexity
5. Stochastic Nature

Advantages of t-SNE

1. Excellent for visualizing clusters and patterns in high-dimensional data. Revealing clusters, groups, and patterns in high-dimensional datasets.
2. Can reveal intricate relationships and structures that might be challenging for linear techniques.
3. Exploratory data analysis: Gaining insights into data distribution and relationships.
4. Identifying potential outliers or anomalies.

Disadvantages of t-SNE

1. Computational Complexity: t-SNE can be computationally expensive, particularly with large datasets.
2. Sensitivity to Hyperparameters: t-SNE has hyperparameters, such as the perplexity and learning rate, which can significantly affect the results.
3. Lack of Global Structure Preservation
4. Difficulty in Reproducibility: Due to the stochastic nature of the algorithm and the sensitivity to initial conditions and hyperparameters, t-SNE may provide different visualizations each time it's run.

UMAP

- Uniform Manifold Approximation and Projection (UMAP) is a manifold learning technique used for dimension reduction.
- The steps for UMAP are a mix of what we've seen with t-SNE and Isomap
    1. Nearest Neighbor Graph via a distance metric, k-NN and "fuzzy" membership.
    2. Lower-Dimensional Representation: optimization objective is based on minimizing the cross-entropy between the high-dimensional graph and the low-dimensional representation. Similar to t-SNE
    3. Stochastic Gradient Descent: similar to t-SNE

Advantages of UMAP

1. UMAP preserves both local and global structures, providing a balance between details and the overall view of the data.
2. It can be used with various distance metrics, making it adaptable to different types of data.
3. UMAP is generally faster than similar manifold learning techniques like t-SNE, especially on large datasets.
4. It can be used in a wide range of applications, including clustering, visualization, and general-purpose dimensionality reduction.

Disadvantages of UMAP

1. Hyperparameter Sensitivity: The results can be sensitive to the choice of hyperparameters, although this is often less pronounced than in methods like t-SNE.
2. Stochastic Nature: The algorithm has a stochastic element, which means that repeated runs can lead to different results.
3. Interpretability: While it is designed to preserve local and global structures, interpreting these structures and understanding what they mean in the original data context can sometimes be challenging.

PaCMAP

- PaCMAP (Parameter-free Conformal Mapping) is a dimensionality reduction technique designed to create low-dimensional representations of high-dimensional data while preserving both local and global structures.

PaCMAP

- PaCMAP (Parameter-free Conformal Mapping) is a dimensionality reduction technique designed to create low-dimensional representations of high-dimensional data while preserving both local and global structures.
- Key features:
  1. Conformal Prediction: Conformal prediction is a framework in machine learning that provides confidence measures for predictions. PaCMAP leverages conformal prediction to assess the trustworthiness of the mapping of data points from high-dimensional space to low-dimensional space.
  2. PaCMAP automates the process of parameter selection by utilizing conformal prediction.
  3. Local and Global Preservation: Local relationships between neighboring data points are captured, and global trends across the entire dataset are maintained in the lower-dimensional representation.

Regression with other dimension reduction methods

- We can apply the same idea as PCR with other dimension reduction techniques.
- However, the dataset with reduced dimension, say $\boldsymbol{z}$, is not linear in $\boldsymbol{X}$ for other methods.
  - commonly not possible to get β values on the original scale
  - standard errors of original β values are even harder!
- Some methods (e.g., t-SNE), it's not directly possible to find $\boldsymbol{z}$ for a new subject or get the original β values.

Summary

- Dimension reduction is a versatile tool in machine learning and data analysis.
- It helps simplify datasets, reduce computational costs, and provide insights through visualization.
- The choice of method and its implementation should align with the specific goals and nature of the data being analyzed (i.e., local vs. global).
- There are many other methods that we have not discussed.