ARTS1422 - Data - Visualization - Hw2

Problem

Hw2 mainly focuses on comparing Dimensionality Reduction Techniques, while we first need to apply the methods for the dataset and get the dimension-reduced data and then we need to create an interactive visualization so that we can find the difference among different methods performance.

Dataset

For the dataset, I choose the Iris recommended. Iris contains 150 samples and each row of data contains four features for each sample, containing calyx length, calyx width, petal length and petal width as well as the category information of the sample but since our clustering and dimensionality reduction methods are unsupervised so we truncate the category information.

Methodology

And for dimensionality reduce methods, I choose PCA, MDS and TSNE.

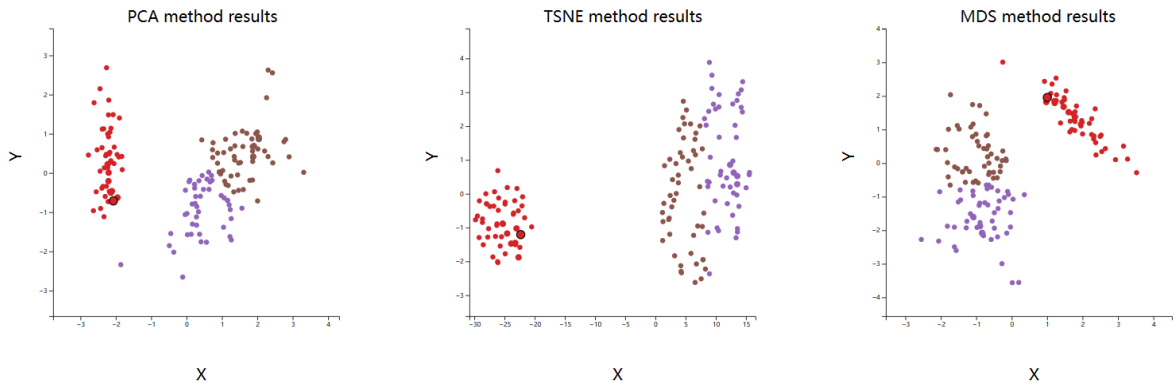
1. Principal component analysis aims to map high-dimensional data into a low-dimensional space through a linear transformation while retaining most of the variance in the data. This reduces the dimensionality of the data while maintaining the main features of the data. The main steps are to compute the covariance matrix of the data, then decompose the eigenvalues of the covariance matrix to obtain the principal components (eigenvectors), and then select the most important principal components to construct a new feature space.
2. t-SNE is a nonlinear dimensionality reduction technique designed to map high-dimensional data into a low-dimensional space while preserving the local structure between data points. This makes t-SNE particularly suitable for visualizing high-dimensional data. The core idea is to optimize the objective function so that the distribution of data points in the low-dimensional space reflects as much as possible the similarity between data points in the high-dimensional space.
3. MDS is designed to map high-dimensional data into a low-dimensional space while maintaining the distance relationship between data points as much as possible. This makes MDS particularly useful for visualizing the relative position and structure of data. The key idea is to achieve dimensionality reduction by minimizing the difference between the distances between data points in the high-dimensional space and the distances between data points in the low-dimensional space.

I also apply Kmeans algorighm for the dimension-reduced data, and set the category num = 3 which is equal to the original data, so that it can enhance the performance of the interactive visualization.

Evaluation

I try many times for calling dimensional reduction algorithms, and I find that the results of PCA and TSNE always stay the same, but MDS always changes. At the same time, all of three results show an clear visable V-structure in a degree. And you can see, MDS shows an V-structure with a certain angle or flip.. Also, the left part of TSNE is far away from the other two parts and is round like circle. Therefore, I think PCA has the best performance, that is because, it shows a correct V-structure, and MDS's and TSNE's visually show different structure and relationships.

Result



As shown above, when your mouse is over certain data points, it will also highlight all related points in three graphs.