

ARTS1422—HW2

包佳诚 2021533098

Step1 Data Preparation & discription

We will be working with the famous “Iris” dataset that has been deposited on the UCI machine learning repository

(<https://archive.ics.uci.edu/ml/datasets/Iris>). (but I use `sklearn.datasets import load_iris` to get it)

The iris dataset contains measurements for **150** iris flowers from three different species.

The three classes in the Iris dataset are:

- Iris-setosa (n=50)
- Iris-versicolor (n=50)
- Iris-virginica (n=50)

And the four features of in Iris dataset are:

- sepal length in cm
- sepal width in cm
- petal length in cm
- petal width in cm

Step2 The methodology used for dimensionality reduction

- PCA
- t-SNE
- MDA

Step3 An evaluation of the different methods

PCA:

- Linear dimensionality reduction, preserving the direction in which the variance of the data is greatest
- Suitable for data compression, noise removal, visualization
- Simple and efficient, but unable to capture nonlinear relationships

t-SNE:

- Non-linear dimensionality reduction, preserving the local structure of high dimensional data
- Good at high-dimensional data visualization, display clustering effect
- The calculation cost is high, and the parameter selection needs to be careful

MDA:

- Oversee dimensionality reduction to maximize the separability of different categories
- Applies to categorizing tasks and displaying the category separability
- Data needs to be labeled and the calculation cost is high sensitive to Gaussian distributions and linear classification boundaries

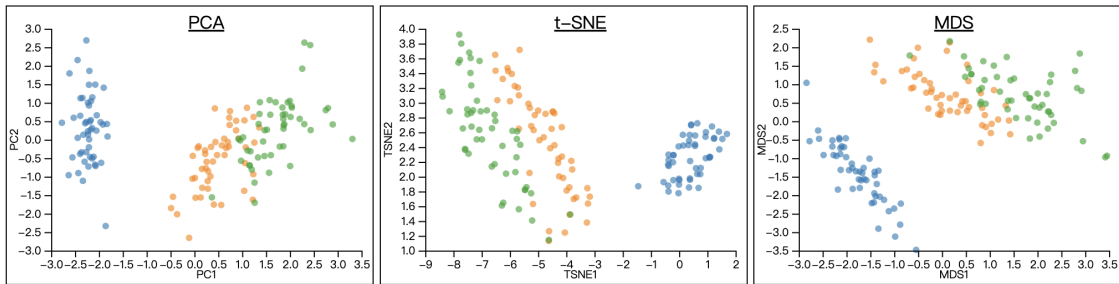
According to figures:

- PCA is good at separating Setosa classes, but not fully separating versicolor and virginica classes with nonlinear superposition.
- t-SNE: Retaining local structure, three classes form compact and clearly separated clusters, Versicolor and Virginica clusters overlap a little but are largely separate.
- MDA: Use supervision information to find differentiated linear subspaces. Setosa classes are completely separated, Versicolor and Virginica classes are also well separated but still overlap more than t-SNE method.

Step4 screenshots of visualization result

数据可视化HW2

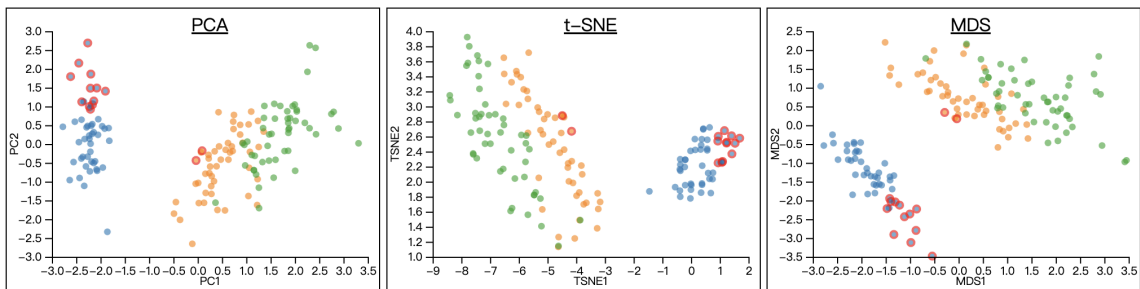
包佳诚 2021533098



Clear Highlights

数据可视化HW2

包佳诚 2021533098



Clear Highlights

- You can click the dots and corresponding dots will also be highlighted in other graphs. Click the highlighted dot will erase the highlighted corresponding dots.
- The button Clear Highlights will help you clear all the highlighted dots.