Mateo W. Racca

UNIVERSITAT ROVIRA i VIRGILI
Escola Tècnica
Superior d'Enginyeria

NEURAL AND EVOLUTIONARY COMPUTING

(MESIIA): Assignment #3: Unsupervised learning with PCA, t-SNE, k-means, AHC and SOM

Part 1

The chosen dataset is commonly known as the HCV dataset (Hepatitis C Virus) from the UCI Machine Learning Repository. Below is the detailed description of the dataset and a link to the source webpage:

HCV Dataset Description:

Domain: Medical / Hepatology.

Features: The dataset includes various medical measurements such as liver enzymes, bilirubin, albumin, and other blood chemistry measurements. It also contains demographic information like age and gender.

Target: The dataset categorizes patients into different categories based on the stage of Hepatitis or other clinical diagnoses related to HCV. The possible values are '0=Blood Donor', '0s=suspect Blood Donor', '1=Hepatitis', '2=Fibrosis', '3=Cirrhosis'

Data Points: It initially has 615 instances, each representing different patients or clinical cases.

The dataset can be retrieved from: https://archive.ics.uci.edu/dataset/571/hcv+data

Preprocessing

Firstly, the dataset is loaded, and we separate the features we want to analyze from the target labels. The target labels, which represent the classes in the dataset, are set aside because unsupervised learning algorithms work without labeled output data.

We then categorize the features into numerical and categorical types since they require different treatments. Numerical features are standardized to have a mean of zero and a standard deviation of one. This normalization is crucial because unsupervised algorithms like PCA and k-means can be skewed by features that operate on larger scales.

Categorical features, in this case binary, are encoded to ensure they are represented numerically, making them suitable for mathematical operations performed by the algorithms.

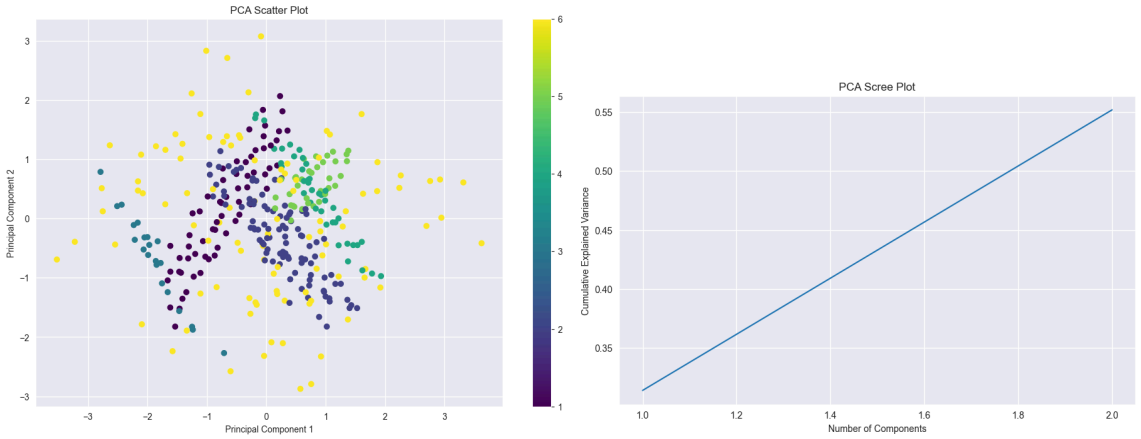
After applying these transformations, the cleaned and transformed data is saved back into a CSV file.

Part 2:
Let's walk through the process for both the NEC and HCV datasets:

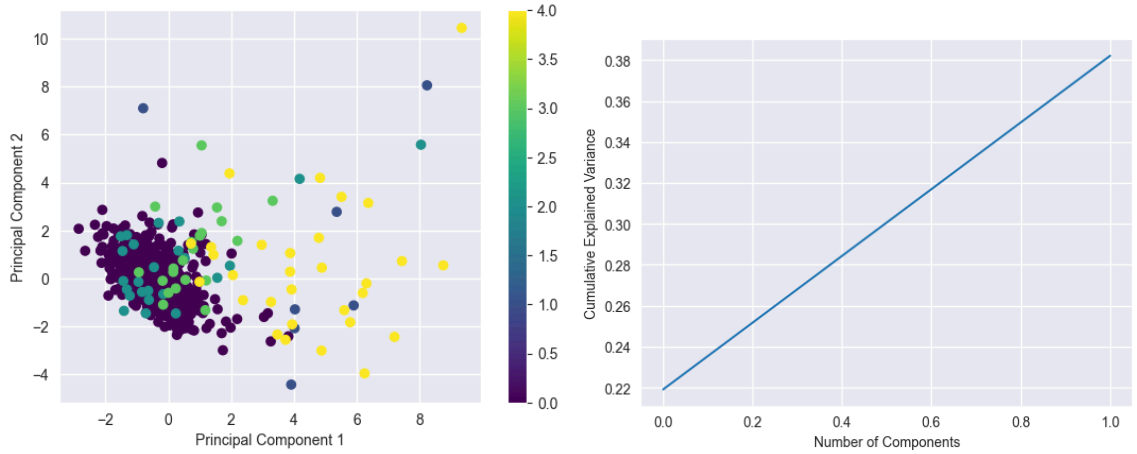
For both the NEC and HCV datasets, the application of various unsupervised learning techniques offers a deep and multifaceted understanding of the data:

PCA (Principal Component Analysis): This technique reduces the dimensionality of the data while retaining most of the variance. The scatter plot visualizes the distribution of data points along the principal components, revealing inherent structures or clusters. The scree plot shows the explained variance ratio of each principal component, helping to decide how many components to retain for optimal data representation.

A3 data:

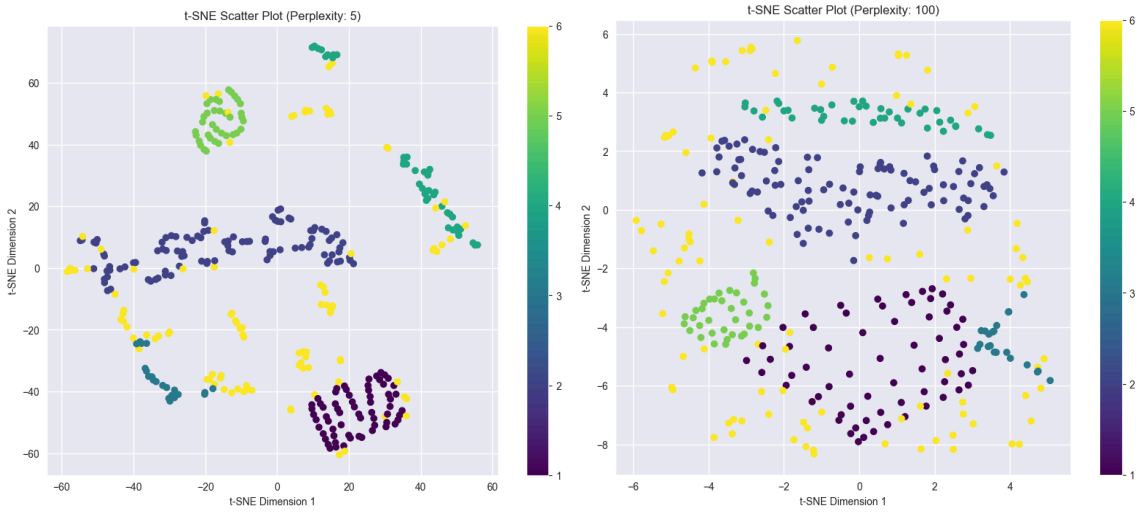


HCV data:

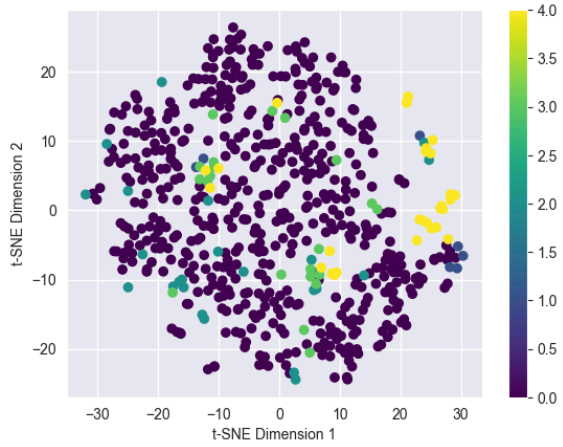


t-SNE (t-Distributed Stochastic Neighbor Embedding): A non-linear technique particularly effective for visualizing high-dimensional data in two or three dimensions. t-SNE plots can reveal clusters or groupings in the data that are not apparent in PCA due to t-SNE's ability to capture local structures. Different perplexity settings can significantly affect the results, offering insights into the data's underlying structure.

A3 data:



HCV data:

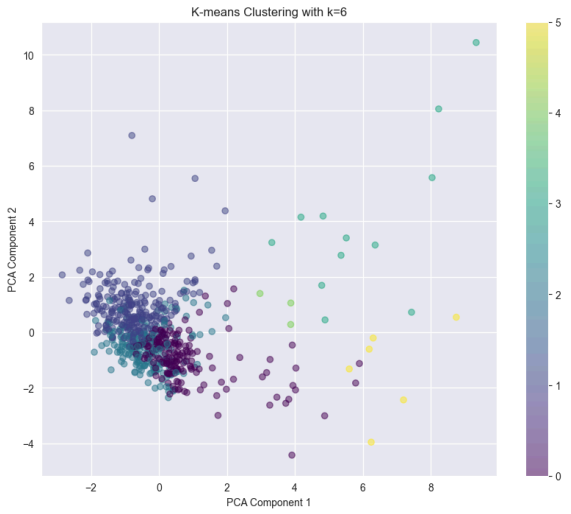


k-means Clustering: This algorithm partitions the data into k clusters, where each data point belongs to the cluster with the nearest mean. By varying the number of clusters (k) and visualizing the results using scatter plots (often plotted against the first two PCA components for easy visualization), you can assess how well k-means can segment the

data. When k equals the actual number of classes, a confusion matrix provides a clear view of the clustering performance against true labels.

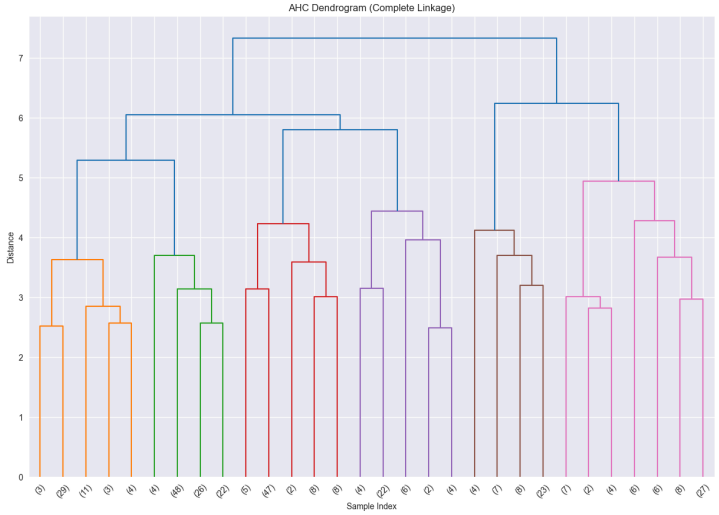
A3 data:

HCV data:

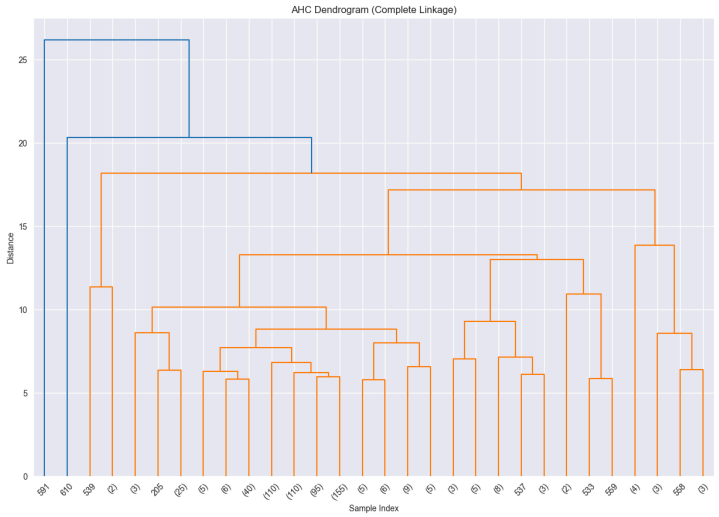


AHC (Agglomerative Hierarchical Clustering): This method builds a hierarchy of clusters and is visualized using a dendrogram. It provides a tree-like structure of the data, which is useful for understanding the hierarchical grouping of data points and determining the number of clusters by 'cutting' the dendrogram at a suitable level.

A3 data:

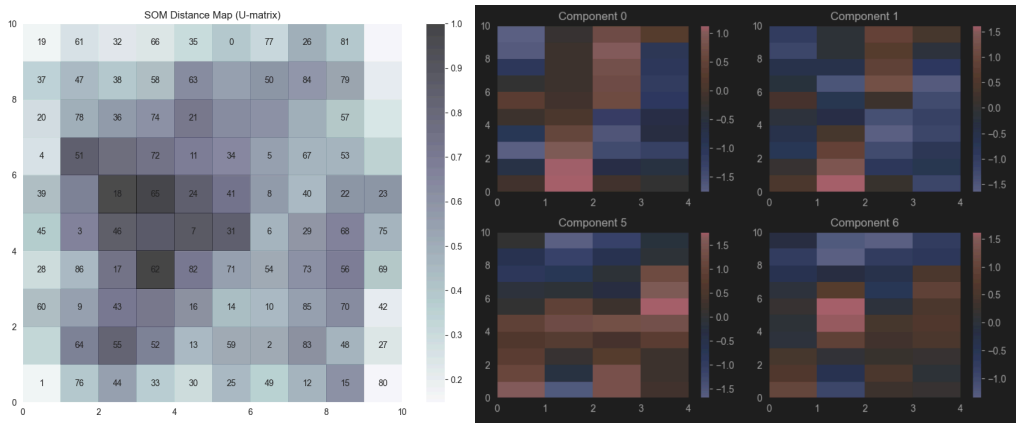


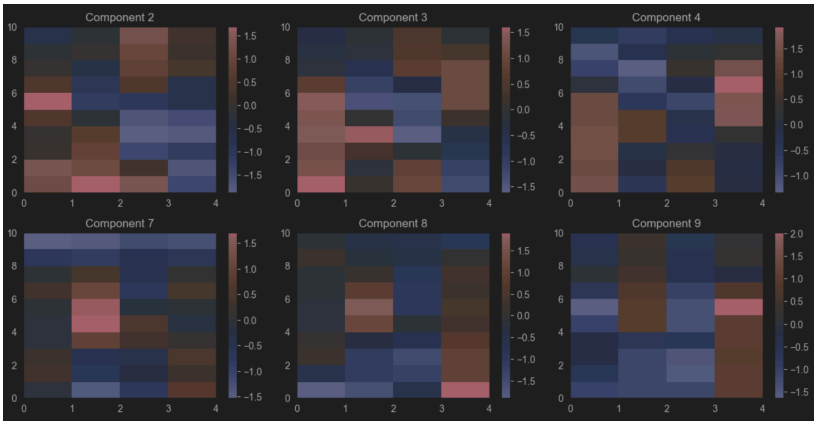
HCV data:



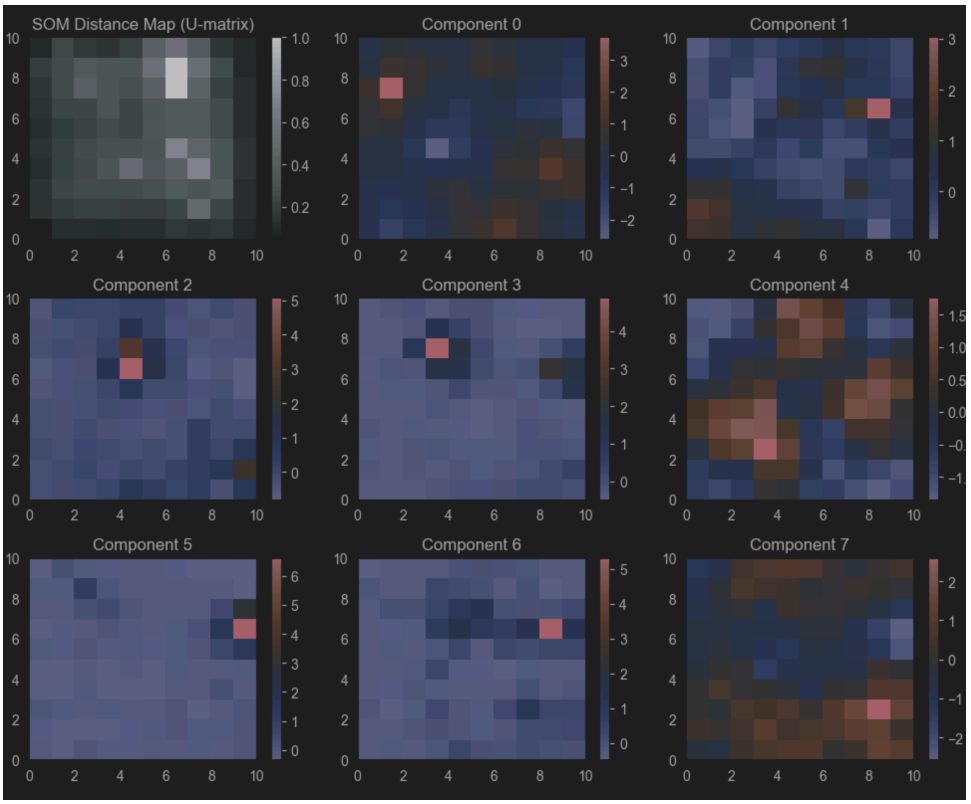
SOM (Self-Organizing Maps): SOMs reduce the dimensionality of the data while preserving topological and metric relationships. The output is a two-dimensional map of neurons, where similar data points are mapped close to each other. The U-matrix visualizes the distances between neurons, helping to identify potential clusters. Component planes show the weights of each input feature, offering insights into feature contributions and relationships.

A3 data:





HCV data:



All the process can be found on the notebooks of each dataset in the github repository: https://github.com/raccamateo/NEC_A3

The application of unsupervised learning techniques to the NEC and HCV datasets has provided a comprehensive insight into the underlying structure and distribution of the data. Here are some final thoughts on the findings from each method applied:

Results:

The PCA scatter plots for both datasets show the spread of data points across the first two principal components. For the A3 data, points are more evenly distributed, while the HCV data shows a dense cluster, suggesting more variance.

The scree plots indicate how much variance each principal component accounts for. In both cases, the first few components capture a substantial part of the variance,

with a sharp drop-off, indicating that these components are significant for data representation.

The t-SNE plots for the A3 dataset show clear clustering at different perplexity levels, suggesting distinct groupings in the data. For the HCV data, the plot shows a dense cluster with some outliers, which may point to a core structure within the data, surrounded by less common or more varied data points.

The k-means clustering scatter plots with PCA reduced dimensions indicate that for the A3 dataset, there's a visible separation between clusters, while for the HCV dataset, the clustering is less distinct. This could imply that the A3 dataset has clearer group separations or that the chosen value of k is more appropriate for its structure.

The AHC dendrograms for both datasets illustrate the hierarchical clustering structure. In the A3 data, the dendrogram shows a more balanced cluster split, while the HCV dendrogram displays some large clusters and a few small ones, which could represent outlier groups or sub-clusters within the data.

The SOM's U-matrix and component planes for the A3 dataset reveal different intensities and potential cluster regions, which are less distinct in the HCV data.

It is key for the activity to remark that these visualizations are essential for understanding the complex, high-dimensional structures. They provide insight into the inherent clustering and can help guide further analysis, such as identifying features that are important for class separation or informing the development of predictive models.