Exploring clinical heterogeneous data using unsupervised machine learning

2 February 2023

Assignment III

Reading

- Z. Huang. *Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values*. Data Mining and Knowledge Discovery, 1998. See Paper.

- To read about the Silhouette score sklearn and Wikipedia.

Experimentation with k-modes

1. To implement the **k-modes+** algorithm (see Section 4 of Huang's manuscript).
2. To download the Vote dataset from **HERE**. It contains 435 instances, 16 features, and 2 classes (k number). The last column in the dataset corresponds to the true labels.
3. To use your k-modes+ algorithm on the Vote dataset. Set k=2 and the number of iterations = 50. Repeat the experiment 50 times.
4. To report the average accuracy and Rand index from the experiment in Step 3.
5. To use the k-modes implementation from the library **kmodes 0.12.2** in python and repeat the experiment similar to Step 3..
6. To create a box plot comparing the accuracy performances obtained by kmodes+ and k-modes.
7. To share your GitHub repository indicating the source implementation

- NOTE. In case of any question regarding the assignment III, email me to adan.josegarcia@univ-lille.fr

Dr. JOSÉ-GARCÍA
Building ESPRIT S3.11
https://adanjoga.github.io