Exploring clinical heterogeneous data using unsupervised machine learning

Dr. JOSÉ-GARCÍA

Assignment 1:

22 September 2023

Reading

[2] Amir Ahmad and Lipika Dey. *A k-means clustering algorithm for mixed numeric and categorical data*

Implementation k-means

1. To create a Github or Gitlab account.
2. To create a new repository for the Master in Data Science project.
3. To implement the `k-means` algorithm.
4. To use the `Iris dataset` and the `k-means` algorithm, the number of clusters, k, should be set to 3. See below how to import the Iris dataset in Python.

```
from sklearn.datasets import load_iris
data = load_iris()
X = data.data
y = data.target # ground truth labels
```

5. To use a 2D scatter plot to visualize the clusters obtained by `k-means`.
6. To measure the accuracy obtained by `k-means`. Note. To relabel the k-means clustering if necessary to match the true labels (y).
7. To compare your `k-means` implementation with `sklearn.cluster.KMeans`. Compare the two solutions in terms of accuracy. Plot both solutions.