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.

Biclustering for precise patient stratification from clinical features

Dr. JOSÉ-GARCÍA

Assignment 2:

16 October 2023

Reading

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

[4] To read about the Silhouette score sklearn and Wikipedia.

Part I: Experimentation with k-means

1. Considering your implementation of `k-means*` and the Iris dataset, run the algorithm 50 times. Measure the `accuracy` of the 50 clustering solutions generated by `k-means*` taking into account the `Iris true labels`.
2. Repete the experiment from **Step 1** using `sklearn.cluster.KMeans`.
3. Create a box plot comparing the accuracy performances of **Step 1** and **Step 2**.
4. Create a scatter plot to illustrate the best clustering solution (highest accuracy values) from the solutions generated in **Step 1**. Repeat the same procedure for creating a new scatter plot from the solution generated in **Step 2**.
5. To create a convergence plot of your `k-means*` implementation using the `Iris` dataset. The k-means minimizes the *within-cluster sum of squares* (WCSS). Thus plot at each iteration the WCSS value. The x-axis should indicate the number of iterations and the y-axis should indicate the WCSS value. Repeat this procedure 50 times, thus the plot will have 50 convergence lines.
6. Read about the digits dataset and then import it as:

```
from sklearn.datasets import load_digits
data, labels = load_digits(return_X_y=True)
```

7. Repeat Steps 1 to 5, using the `digits dataset` instead of the iris dataset.

Part II: Clustering on heterogeneous features

1. To create a GitHub repository for the Biclustering algorithm in Ref [2].
2. To create a flowchart of the algorithm described in Ref [2], Section 4: *k-Mean clustering for mixed data sets*.