11752 Machine Learning
Master in Intelligent Systems
Universitat de les Illes Balears

Handout #4: **Unsupervised Learning**
(graded assignment)

This assignment deals with the **digits dataset** directly available from **scikit-learn**[1]. This dataset comprises 8×8-pixel images of hand-written digits 0-9 with approximately 180 samples per class. You are supposed to use the combination of **three** classes corresponding to your group, which is indicated at the **end of this handout**.

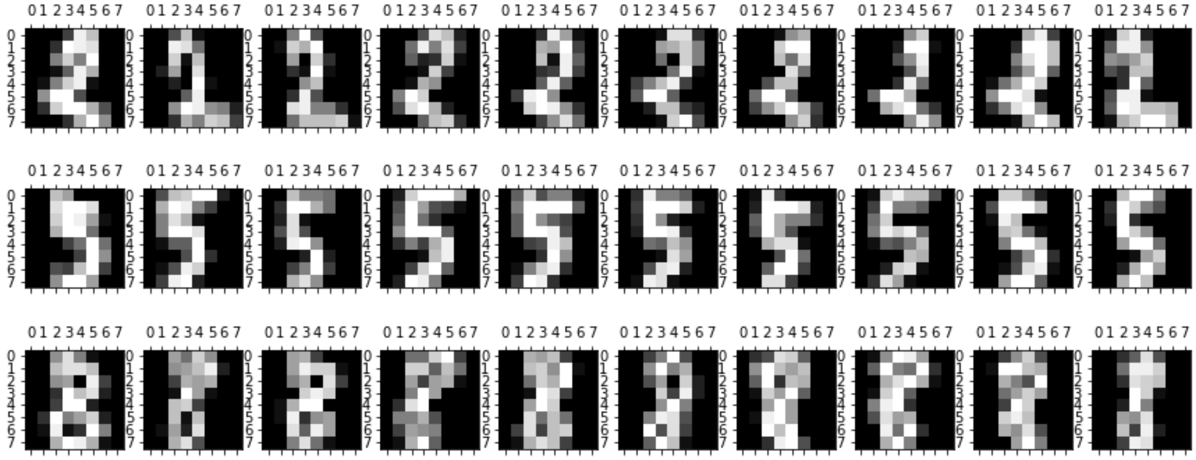

Figure 1: Samples corresponding to the 2-, 5- and 8-digit classes.

The following source code allows you to get access to the dataset samples and the corresponding labels 0-9:

```python
from sklearn.datasets import load_digits
digits = load_digits()
samples = digits.data
labels = digits.target
```

Listing 1: Loading of the **digits** dataset.

For the tasks which are described below you are supposed to:

(a) Consider the original dataset and a lower-dimensional version obtained through PCA retaining 95% of the variance.

(b) Cluster your dataset for m = 2, 3, 4 and 5 clusters and report on the performance attained in each case using the *v-measure*.

(c) For the best case among the 8 possible combinations resulting from (a) and (b):

   i. Compute the *contingency matrix*.

   ii. Determine the assignment of classes to clusters.

   iii. Identify the number of incorrectly clustered samples and calculate also the percentage of errors as *number of incorrectly clustered samples / total number of samples*.

   iv. Report also on the *homogeneity* and the *completeness* measures.

[1] `https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_digits.html`

v. In case there are mistakes, show one example of each case using the following source code (`X` is the matrix with the samples and `ndx` is the index of an incorrectly clustered sample, e.g. a sample from class 0 clustered as if it was from class 6):

```
import matplotlib.pyplot as plt
plt.figure()
plt.gray()
plt.matshow(X[ndx].reshape(8,8))
plt.title(’sample from class 0 clustered as class 6’)
plt.show()
```

T1. Consider the *Ward* algorithm and the Euclidean distance.

T2. Consider the *K-means* algorithm and the Euclidean distance.

T3. Consider the *Fuzzy K-means* algorithm and the Euclidean distance.

T4. Determine the best clustering methodology among the options above.

NOTE 1: Regarding T1, use the implementation of the *hierarchical agglomerative clustering* method available in *scikit-learn*.[2]

NOTE 2: Regarding T2 and T3, you have to use the implementation of the corresponding algorithm available in the adaptation of the *fuzzy_kmeans* library available in the course web page. Have a look at the implementation to understand how to make use of it.

NOTE 3: Scikit-learn web pages on **clustering methods**[3] and **clustering evaluation**[4] will be useful for this assignment. In particular, the following objects/functions of `scikit-learn` will be necessary:

`sklearn.metrics.cluster.contingency_matrix`

`sklearn.metrics.v_measure_score`

`sklearn.metrics.homogeneity_score`

`sklearn.metrics.completeness_score`

[2] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html

[3] https://scikit-learn.org/stable/modules/clustering.html#clustering

[4] https://scikit-learn.org/stable/modules/clustering.html#clustering-performance-evaluation

DELIVERY INSTRUCTIONS:

- To implement the solutions to tasks T1 - T4, you can either use a notebook file (`.ipynb`) or separate python files (`.py`). In the latter case, use a python file for each task and include inside all the source code that is needed to run the solution to the task.

  The name of the python files has to be `alltasks.ipynb`, or `task1.py`, `task2.py`, etc.

- Brief/suitable comments are expected in the source code.

- A report of the work done has to be delivered by/on February 11, 2024 in PDF form. The report can be generated by exporting the notebook file (after full execution) or using a separate text editor; you can find a template in `.docx` format in the course web page that you can adapt for the `.ipynb` case.

  Upload a Zip container to package the report (with name `report.pdf`) and the source code files (.ipynb or .py file(s)).

- This work can be done in groups of 2 students. Use the same group number that you employed for the previous assignment.

- IMPORTANT NOTICE: An excessive similarity between the reports/source code released can be considered a kind of plagiarism.

The classes to be used by each group can be found in the following table:

group	**classes**
1	1, 5, 6
2	3, 4, 5
3	6, 7, 8
4	0, 2, 4
5	6, 8, 9
6	4, 5, 6
7	5, 7, 9
18	0, 3, 7
19	0, 4, 8
20	2, 3, 6