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

Handout #4: **Unsupervised Learning** (hierarchical clustering)

NOTE 1:

- All problems will require the use of `scikit-learn`[1] and `matplotlib`[2].
- Scikit web pages on **clustering methods**[3] and **clustering evaluation**[4] will be useful.
- In particular, the following objects/functions of `scikit-learn` will be necessary:

  `sklearn.cluster.AgglomerativeClustering`
  `sklearn.metrics.v_measure_score`
  `sklearn.metrics.davies_bouldin_score`
  `sklearn.metrics.cluster.contingency_matrix`
- You can make use of other functions from scikit-learn or any other Python library which may be useful.

P5. **Given the dataset** `dsxx5.txt`, cluster it according to the following combinations:

- the *complete linkage*, the *average linkage* and the *ward* algorithms,
- 2, 3, 4, and 5 clusters, and
- the Euclidean distance

and

(a) considering the **V-measure** (external measure):
   i) calculate the metric value for all cases
   ii) select the best number of clusters according to this metric for each algorithm
   iii) select the best algorithm and number of clusters according to this metric
   iv) plot separately the dataset using the true labelling and the labelling derived from the best algorithm and number of clusters
   v) find the contingency table for the best algorithm and number of clusters, and
   vi) determine the asignment of classes to clusters and the number of clustering errors that result from the best algorithm and number of clusters;

(b) considering the **Davies-Bouldin score** (internal measure):
   i) calculate the metric value for all cases
   ii) select the best number of clusters according to this metric for each algorithm
   iii) select the best algorithm and number of clusters according to this metric
   iv) plot separately the dataset using the true labelling and the labelling derived from the best algorithm and number of clusters
   v) find the contingency table for the best algorithm and number of clusters, and
   vi) determine the asignment of classes to clusters and the number of clustering errors that result from the best algorithm and number of clusters.

[1] `https://scikit-learn.org`, `https://scikit-learn.org/stable/modules/classes.html`
[2] `https://matplotlib.org/`
[3] `https://scikit-learn.org/stable/modules/clustering.html#clustering`
[4] `https://scikit-learn.org/stable/modules/clustering.html#clustering-performance-evaluation`

To finish, using one or two paragraphs, comment on whatever you consider adequate regarding the results obtained, e.g. a certain score is not indicating the real number of clusters and why, whether the mistakes are relevant or can be accepted, what should be discarded to get an acceptable result, etc.

NOTE 2: The previous problem requires loading dataset `dsgg5.txt` where `gg` is the group number:

```
import numpy as np
group = '01' # assuming group 1
ds = 5       # dataset 5
data = np.loadtxt('ds'+group+str(ds)+'.txt')
X = data[:, 0:2]
y = data[:, 2]
```

Class labels are 0, 1, 2, etc. All datasets include the true labelling for all samples.

P6. Repeat problem P5 (except points **a.iv** and **b.iv**) using the `digits` dataset directly available from **scikit**[5] considering each group the following combinations of classes:

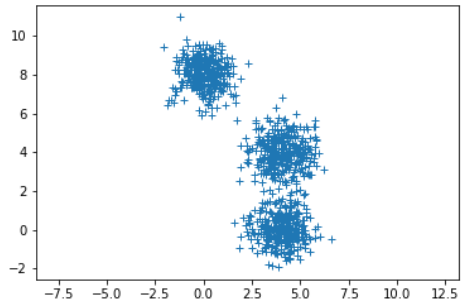
group	classes to consider
1	0, 1, 2
2	3, 4, 5
3	6, 7, 8
4	0, 2, 4
5	6, 8, 9
6	1, 3, 5
7	5, 7, 9

group	classes to consider
8	1, 7, 9
9	0, 4, 8
10	2, 3, 6
11	2, 5, 7
12	3, 6, 9
13	1, 5, 9

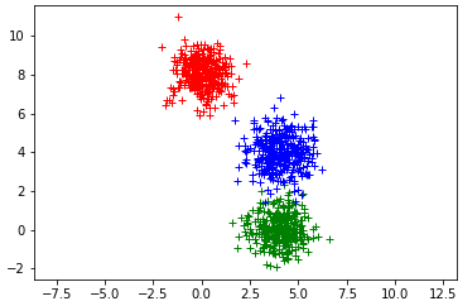
- A report of the work done has to be released by February 6, 2022 in electronic form as a notebook file (.ipynb).
- Provide the requested data and plots/figures at each point above. For figures, use appropriate titles, axis labels and legends to clarify the results reported.
- Suitable comments are expected in the source code.
- This work has to be done individually (see the number of group in *Aula Digital*).

Appendix: Example of dataset

`dsxx5` (unlabelled)



`dsxx5` (labelled)



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