INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Introduction to Machine Learning — 2022/2023
Unsupervised Learning

> These exercises should be solved using Python notebooks (Jupyter) due to the ability to generate a report integrated with the code. It is assumed you are proficient with programming. All answers must be justified and the results discussed and compared to the appropriate baselines.
> Each exercise is scored 1 point. Max score of the assignment is 4 points. Optional exercises help achieving the max score by complementing the errors or mistakes in the mandatory exercises.
> **Deadline:** November 15th, 2021

This exercise will demonstrate how a learning algorithm can distinguish between two distributions of points generated with different parameters.

Exercise 1

Generate 2D points using a multivariate Gaussian distribution. Use the following code to generate two sets, each with 500 points, with different centers and save them to one variable, shuffled. The generated graph should be similar to the one presented in Fig. 1.

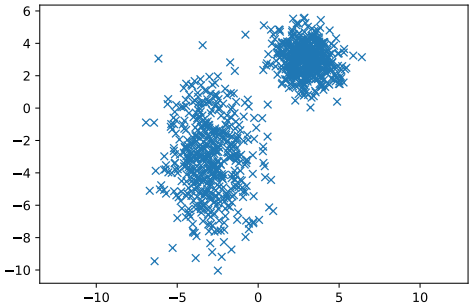

Figure 1: 2D points generated from two multivariate Gaussian distributions

```
import matplotlib.pyplot as plt
import numpy as np
import random

mean = [3, 3]
cov = [[1, 0], [0, 1]]
a = np.random.multivariate_normal(mean, cov, 500).T

mean = [-3, -3]
cov = [[2, 0], [0, 5]]
b = np.random.multivariate_normal(mean, cov, 500).T

c = np.concatenate((a, b), axis = 1)
c = c.T
np.random.shuffle(c)
c = c.T

x = c[0]
y = c[1]

plt.plot(x, y, 'x')
plt.axis('equal')
plt.show()
```

Generate a similar dataset, but with labels 1 or 2 depending if they were generated using the first or the second set of parameters.

Start by choosing two random points in the dataset r_1 and r_2 and apply the following adaptation rule:

for all $x \in$ the dataset **do**
  if x is closer to r_1 than to r_2 **then**
    $r_1 \leftarrow (1-\alpha) \times r_1 + \alpha \times x$
  else if x is closer to r_2 than to r_1 **then**
    $r_2 \leftarrow (1-\alpha) \times r_2 + \alpha \times x$
  end if
end for

Repeat for 10 times a passage through all the elements of the dataset with $\alpha = 10E - 5$ and save:

a) the consecutive values of r_1 and r_2 for the first passage;

b) the values of r_1 and r_2 at the end of each passage.

Plot a) and b) upon the dataset plot, but different colors. What do you conclude about the evolution of the two points in the different situations? Is there any relation between the (most common) final values of r_1 and r_2 and the parameters used to generate the dataset?

Exercise 2

Instead of changing the value for each example, accumulate the values of the difference $(x - r)$ and change the value only when all examples have been observed.

for all x **do**
  $d \leftarrow d + (x - r)$
end for
$r \leftarrow r + (\alpha / n\_examples) * d$

a) Plot the consecutive positions of r_1 and r_2 and compare with the plot in exercise 1. What do you observe?

b) Plot with one colour the points closest to r_1 and with another the points closest to r_2. What do you observe? Use the data set and plot with four colors:

- color 1 – points closer to r_1 labeled 1;
- color 2 – points closer to r_1 labeled 2;
- color 3 – points closer to r_2 labeled 1;
- color 4 – points closer to r_2 labeled 2.

What do you observe?

c) Repeat the experiment 30 times and plot the final values of r_1 and r_2 over the dataset.

Exercise 3

Implement a simplified version of agglomerative hierarchical clustering, as proposed in the following algorithm.

while there are more than two points **do**
  FIND the closest two points
  REPLACE both points by their average
end while

Exercise 4

Implement the DBScan algorithm as described in `https://www.youtube.com/watch?v=_A9Tq6mGtLI` and demonstrate it.