Exercice 1 *Zero Inflated Poisson (ZIP) distribution.* A random variable X following the usual Poisson distribution with parameter λ, $\mathscr{P}(\lambda)$, with the probability mass function

$$\mathbb{P}(X=k) = \exp(-\lambda)\frac{\lambda^k}{k!}, \quad k = 0,1,\dots \tag{1}$$

is widely used to model many naturally occurring events where X represents the "number of events per unit of time or space". Note that X takes only nonnegative integer values. However, the $\mathscr{P}(\lambda)$ distribution may not be useful (or it gives a bad fit) when X takes the value 0 with a high probability. In such a case a modified version of a regular $\mathscr{P}(\lambda)$ distribution known as the zero-inflated Poisson (ZIP) distribution becomes useful. The ZIP distribution with parameters π and λ, denoted by ZIP($\pi, = \lambda$), has the following probability mass function :

$$\mathbb{P}(X=k) = \begin{cases} \pi + (1-\pi)\exp(-\lambda) & \text{if } k = 0 \\ (1-\pi)\exp(-\lambda)\frac{\lambda^k}{k!} & \text{if } k = 1,2,\dots \end{cases} \tag{2}$$

This model represents a **mixture** between a discrete distribution such that $\mathbb{P}(X=0) = 1$ with probability π and a classical Poisson distribution with parameter λ with probability $1-\pi$.

1. Write a function `zip(pi, l, size)` that simulates `size` independent realizations of a ZIP distribution with parameters π and λ. Verify your code by, for example, drawing a few bar diagrams.
2. EM algorithm
   (a) Find the likelihood of the parameters of a ZIP distribution for a sample $(x_i)_{1\le i\le N}$ assumed to come from such a distribution.
   (b) Describe the EM algorithm in the specific case of the ZIP distribution.
   (c) Write a function `emzip(x, itmax, tol)` which estimates the parameters π and λ of a ZIP distribution from a sample by maximizing the log-likelihood using the EM algorithm.
   (d) Using the function

   `scipy.optimize.minimize`

   construct another function `directzip` which estimates the parameters π and λ of a ZIP distribution from a sample by maximizing the log-likelihood directly (without using the EM algorithm)
3. The table 1 contains the number of tornado occurrences in Lafayette Parish, Louisiana per year from 1950 through 2009. Compute the estimators of π and λ and provide the **goodness**

of fit (GOF) using the χ^2 test statistic

$$\Delta_{\text{GOF}} = \sum_{i=0}^{k} (O_i - E_i)^2 / E_i$$

1950–1959	0	0	0	1	0	0	0	1	0	0
1960–1969	1	0	0	0	1	1	0	0	0	2
1970–1979	0	0	0	0	1	3	0	2	1	0
1980–1989	1	0	0	1	0	1	0	0	2	1
1990–1999	0	1	2	0	0	1	0	1	2	0
2000–2009	0	0	3	0	2	0	1	1	3	0

TABLE 1 – Number of tornado occurrences in Lafayette Parish, Louisiana per year from 1950 through 201

Exercice 2 *DBScan.* If ϵ is 2 and minpoint is 2, what are the clusters that DBScan would discover with the following 8 examples :

$$A_1 = (2,10), A_2 = (2,5), A_3 = (8,4), A_4 = (5,8),$$
$$A_5 = (7,5), A_6 = (6,4), A_7 = (1,2), A_8 = (4,9).$$

1. What is the Epsilon neighborhood of each point ?
2. Draw the 10 by 10 space and illustrate the discovered clusters.
3. What if ϵ is increased to $\sqrt{10}$?

You will find the distance between the points in the notebook `TD - 2023 - Clustering`

Exercice 3 *TP sur les données de textures.* Réaliser le TP qui est déposé sur le serveur Discord.