# **Assignment 1:** *Clustering and Maximum Likelihood*

Date: *Jan 14, 2025*
Due : *Jan 28, 2025*

**Acknowledgment** This assignment has been adapted in part from the materials of the courses ECE421 by N. Papernot and ECE1513 by S. Emara.

**Code of Honor** Assignments are designed to enhance your understanding and advance your skills, constituting a significant portion of your final assessment. They must be completed individually, as engaging in any form of academic dishonesty violates the principles of the Code of Honor. If you encounter any challenges while solving the assignments, please contact the instructional team for guidance.

**How to Submit** This assignment has 4 questions. For each question, you need to upload one file on Crowdmark. If the question asks for your code, you need to copy and paste your code in the file that you submit. The uploaded solution for Question 1 is expected to be typed (containing images required). For Questions 2, 3 and 4, the submitted solution can be typed or handwritten.

**Grading** The grades add up to 100 and comprise roughly 10% of the final mark.

**Deadline** The deadline for your submission is on **Jan 28, 2025 at 11:59 PM**. Please note that this deadline is strict and **no late submission will be accepted.**

## Questions

Question 1 [45 Points] **(Clustering with $K$-Means)** We consider the UCI ML Breast Cancer Wisconsin dataset. Features (data samples $\boldsymbol{x}_i$) are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.

You can download the dataset using:

```
sklearn.datasets.load_breast_cancer
```

Unless specified otherwise, we refer to the UCI ML Breast Cancer Wisconsin dataset in this question when we use the word "*dataset*".

1. Implement $K$-means clustering by yourself. Your function should take in an array containing a dataset and a value of $K$, and return the cluster centroids along with the cluster assignment for each data point. You may choose the centroid initialization heuristic of your choice. Hand-in the code for full credit. For this question, you should not rely on any library other than `numPy` in `Python`.

2. Run the $K$-means clustering algorithm for values of $K$ varying between 2 and 7, at increments of 1. Justify in your answer which data you passed as the input to the $K$-means algorithm.

3. Plot the distortion achieved by $K$-means for values of $K$ varying between 2 and 7, at increments of 1. Hand-in the code and figure output for full credit. For this question, you may rely on plotting libraries such as `matplotlib`.

4. If you had to pick one value of $K$, which value would you pick? Justify your answer.

QUESTION 2 [20 Points] **(Lack of Optimality in $K$-Means)** In this question, we try to construct a demonstration that $K$-means clustering algorithm can converge to a solution that is not *globally optimal*. For simplicity, we consider one-dimensional samples. Recall that the optimal clustering (with $N$ samples) finds the centroids $\mu_k$ for $k = 1, \ldots, K$ and assignments $r_{n,k}$ for $k = 1, \ldots, K$ and $n = 1, \ldots, N$ that minimize the risk

$$\mathcal{J} = \frac{1}{N} \sum_{k=1}^{K} \sum_{n=1}^{N} r_{n,k} \left(x_n - \mu_k\right)^2 .$$

Consider the case where $K = 2$ and the dataset is made up of 4 points as follows:

$$\mathbb{D} = \{x_1 = 1, x_2 = 2, x_3 = 3, x_4 = 4\} .$$

Initialize $K$-means with the centroids $\mu_1 = 2$ and $\mu_2 = 4$ and find its convergence. Compare the risk after convergence against the optimal risk, i.e., minimum of $\mathcal{J}$. Explain your conclusion.
**Note:** *You may assume that if a point $x_i$ is equally distant to multiple centroids, then it is assigned to the centroid whose index is smallest, e.g., if $x_i$ is in the same distance from $\mu_1$ and $\mu_2$; then, it is assigned to cluster 1.*
**Hint:** *To find the minimum risk, it is enough to do it by explanation.*

QUESTION 3 [20 Points] **(Maximum Likelihood Estimation)** A random number generator is known to sample from the Gaussian distribution $\mathcal{N}\left(\mu, \sigma^2\right)$. The mean $\mu$ and variance $\sigma^2$ are however unknown to us. We observe the following 5 samples drawn by this generator

$$\mathbb{D} = \{x_1 = 1.76, x_2 = 3.60, x_3 = -2.20, x_4 = 2.22, x_5 = 1.45\} .$$

Using the dataset $\mathbb{D}$, find the maximum likelihood estimate of the mean value $\mu$ and variance $\sigma^2$.

QUESTION 4 [15 Points] **(Linear Algebra Review)** Consider the following matrix:

$$\mathbf{A} = \begin{bmatrix} 2 & -1 & 4 \\ 1 & 1 & 1 \\ 4 & 1 & 2 \end{bmatrix}$$

Compute the characteristic polynomial of $\mathbf{A}$, i.e., $\det\left(\mathbf{A} - \lambda \mathbf{I}\right) = 0$, and determine its eigenvalues.
**Hint:** *One of the eigenvalues is $\lambda = 1$.*
**Note:** *This Question if only for review and is not considered a question you would have in the exam.*