Homework #5 (due 12/05/2024, Thursday)

1. The K Means Clustering Method

(1) Create your own Python code (no use of sklearn package) to implement the K-means clustering algorithm.
(2) Run your code on the IRIS data set with $K = 3$ and visualize the clustering results on 2D with any two features. Show different clusters with different colors. Compare the results with the target values of the data set to evaluate the accuracy of your code.
(3) Compare the accuracies of the algorithm on the IRIS data set when different distance definitions are used in the algorithm. Please consider Euclidean distance, *Manhantan* distance and *Chebyshev* distance.

Let $\boldsymbol{x} = [x_1 \quad \cdots \quad x_m]^T \; and \; \boldsymbol{y} = [y_1 \quad \cdots \quad y_m]^T$ be two vectors, then the distances between $\boldsymbol{x} \; and \; \boldsymbol{y}$ are:

- *Euclidean* distance

$$d(\boldsymbol{x}, \boldsymbol{y}) = \left(\sum_{i=1}^{m} (x_i - y_i)^2 \right)^{\frac{1}{2}}$$

- *Manhantan* distance

$$d(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{m} |x_i - y_i|$$

- *Chebyshev* distance

$$d(\boldsymbol{x}, \boldsymbol{y}) = \max_i |x_i - y_i|$$

2. The PCA Method

Use the PCA method in the sklearn package to visualize the IRIS data set in 2D plane. Show different classes with different colors.