

MA722 - Foundations of Machine Learning Algorithms

Lab problem sheet 7

1. Implement the K-Means algorithm for the given dataset.

Point	X	Y
P1	1	2
P2	1.5	1.8
P3	5	8
P4	8	8
P5	1	0.6
P6	9	11
P7	8	2
P8	10	2

- (a) Store the data in a NumPy array.
 - (b) Randomly select $K = 2$ points as initial centroids.
 - (c) Compute the Euclidean distance of each point from the centroids.
 - (d) Assign each point to its nearest centroid.
 - (e) Update the centroids as the mean of the points assigned to each cluster.
 - (f) Repeat until the centroids no longer change significantly.
 - (g) Print the final centroids and the cluster members.
 - (h) Visualize the results using `matplotlib`.
2. Use the **Iris dataset** available in `sklearn.datasets` to perform K-Means clustering.
 - (a) Cluster the data into **3 clusters** (since there are 3 species of Iris).
 - (b) Compare the predicted cluster labels with the actual species labels.
 - (c) Plot the clusters using only the first two features: `sepal length` and `sepal width`.
 3. Given the dataset of students and whether they passed a fitness test based on their daily exercise time (hours) and junk food consumption (times per week):

Student	Exercise (hrs/day)	Junk Food (times/week)	Result
A	1.0	10	Fail
B	2.0	7	Fail
C	3.0	5	Pass
D	4.0	2	Pass

- Write a Python program to:

- (a) Compute the Euclidean distance between a new student's data (Exercise = 2.5, Junk Food = 5) and all other students.
 - (b) Use $K = 3$ nearest neighbors to classify whether the new student will **Pass** or **Fail**.
 - (c) Display distances, selected neighbors, and the final predicted label.
4. You are given the following dataset representing the size of a house (in square feet) and its corresponding price (in \$1000s):

House	Size (sqft)	Price (\$1000s)
A	1000	150
B	1500	200
C	2000	250
D	2500	300

Implement KNN (without scikit-learn) to:

- (a) Calculate the Euclidean distance between a new house (size = 1800) and all existing houses.
 - (b) Select the 2 nearest neighbors.
 - (c) Predict the house price by taking the **average** of those neighbors' prices.
5. Use the **California Housing dataset** from `sklearn.datasets.fetch_california_housing` to:
- (a) Split the dataset into training (80%) and testing (20%) subsets.
 - (b) Use a `KNeighborsRegressor` with $n_neighbors = 5$.
 - (c) Train the model and predict on the test data.
 - (d) Evaluate the performance using the **Mean Squared Error (MSE)** and the **R^2 score**.
