Assigned:		Due:
May 3, 2025	Homework 4.0	May 9, 2025

Please complete the assigned problems to the best of your abilities. Ensure that your work is entirely your own, external resources are only used as permitted by the instructor, and all allowed sources are given proper credit for non-original content.

1. Practicum Problems

These problems will primarily reference the lecture materials and the examples given in class using Python. It is suggested that a Jupyter/IPython notebook be used for programmatic components.

1.1 Problem 1

Load the auto-mpg sample dataset from the UCI Machine Learning Repository (auto-mpg.data) into Python using a Pandas dataframe. Using only the continuous fields as features, impute any missing values with the mean, and perform Hierarchical Clustering (Use sklearn.cluster.AgglomerativeClustering) with linkage set to average and the default affinity set to a euclidean. Set the remaining parameters to obtain a shallow tree with 3 clusters as the target. Obtain the mean and variance values for each cluster and compare these values to the values obtained for each class if we used origin as a class label. Is there
a Clear relationship between cluster assignment and class label?

Clustered Feature Means Demonstrate significant differences among the three groups in the dimensions of fuel consumption, displacement, horsepower, and weight, indicating that the clustering effectively captures the structure of the distribution of car performance.

Comparison with origin category:

Japanese cars typically have high fuel consumption and light weight;

American cars have low fuel consumption and high weight and displacement;

The clustering reflects the similarity of the cars' performance, but does not correspond one-to-one with the place of manufacture (ORIGIN). Explanation The structure of the clusters is not exactly equivalent to the class label, with some, but not sufficient, correspondence.

1.2 Problem 2

Load the Boston dataset (sklearn.datasets.load boston()) into Python using a Pandas dataframe. Perform a K-Means analysis on scaled data, with the number of clusters ranging from 2 to 6. Provide the Silhouette score to justify which value of k is optimal. Calculate the mean values for all features in each cluster for the optimal clustering - how do these values differ from the centroid coordinates?

After normalizing the Boston house price data, I used K-Means to cluster the data, setting the number of clusters between 2 and 6 to test each of them, and evaluating the clustering effect with the help of Silhouette coefficient. The final result shows that when the number of clusters is 2, the Silhouette Score is the highest, Silhouette Score = 0.3482, which indicates that the clustering scheme can better reflect the internal structure of the data. Analysis of the characteristic means within each cluster revealed that one category mainly represents areas with low crime, good air quality, and high housing prices, while the other category is concentrated in areas with high crime and low

housing prices. I also compared the coordinates of the cluster centers with the feature means of each category and found that they are very close to each other, indicating that the center of mass obtained from KMeans is well representative in representing the overall characteristics of the clusters.

1.3 Problem 3

Load the wine dataset (sklearn.datasets.load wine()) into Python using a Pandas dataframe. Perform a K-Means analysis on scaled data, with the number of clusters set to 3. Given the actual class labels, calculate the Homogeneity/Completeness for the optimal k - what information does each of these metrics provide?

After normalizing the Wine dataset, I used the KMeans clustering method and specified the number of clusters to be 3, based on the known number of true labels in the dataset. After obtaining the clustering results, two clustering evaluation metrics, Homogeneity and Completeness, were further computed. Homogeneity scored 0.8788, which indicates that most of the samples in each cluster are from the same real category, while Completeness scored 0.8730, which indicates that most of the samples from the same category are grouped in the same clusters. Both scores are high, reflecting the strong consistency between the clustering results and the real categories, so it can be assumed that KMeans is able to restore the original category structure of Wine data better.

END