

Assigned:
May 3, 2025

Homework 4.0

Due:
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?

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?

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?

END