

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?

```

The mean and variance of each cluster:
      mpg      ... acceleration
      mean      var      ...      mean      var
cluster
0      27.365414  41.976309  ...    16.298120    5.718298
1      13.889062   3.359085  ...    13.025000    3.591429
2      17.510294   8.829892  ...    15.105882   10.556980

[3 rows x 10 columns]

The mean and variance for each category when using origin as the class label:
      mpg      displacement  ...      weight acceleration
      mean      var      mean  ...      var      mean      var
origin
1      20.083534  40.997026  245.901606  ...  631695.128385   15.033735   7.568615
2      27.891429  45.211230  109.142857  ...  240142.328986   16.787143   9.276209
3      30.450633  37.088685  102.708861  ...  102718.485881   16.172152   3.821779

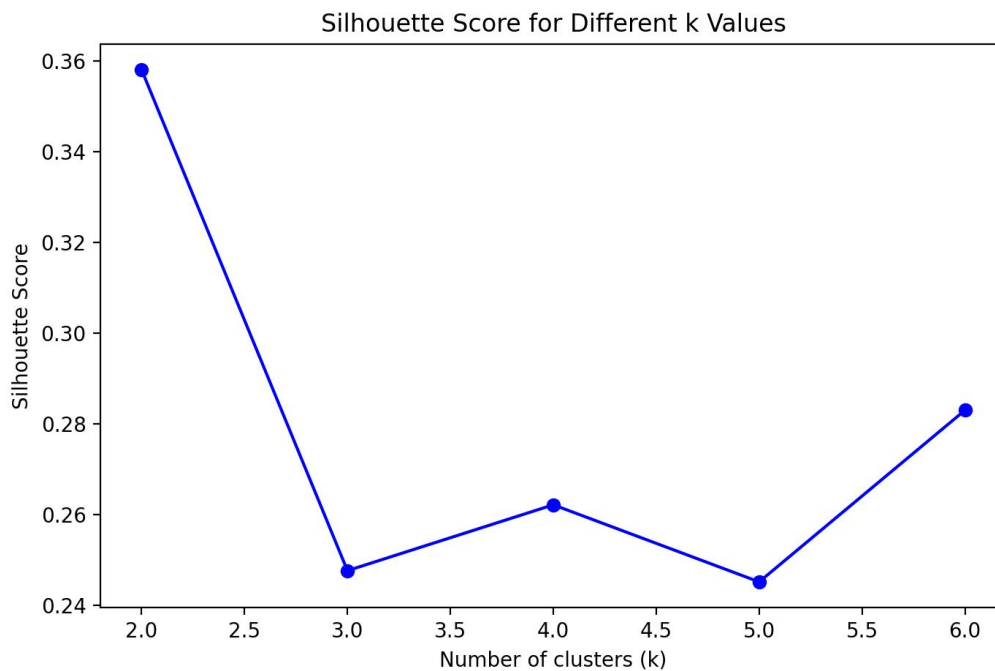
[3 rows x 10 columns]

```

The mean value of mpg for cluster 0 is 27.365414, while the mean value of mpg for the category labeled "origin" as 2 is 27.891429, which is relatively close; however, the mean value of acceleration for cluster 1, 13.889062, differs significantly from the mean values of mpg for each category label. The mean value of acceleration for cluster 0, 16.298120, does not show a clear and consistent correspondence compared to the values of 15.033735 for "origin" as 1, 16.787143 for "origin" as 2, and 16.172152 for "origin" as 3. There is no regular matching pattern in the mean values of this feature across different clusters and different category labels. Through the observation of these mean values and variances of the features, no very clear and stable correspondence between the cluster allocation and the category labels based on "origin" was found. There is no obvious relationship between the two.

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?



```
For k=2, the silhouette score is 0.3580
For k=3, the silhouette score is 0.2477
For k=4, the silhouette score is 0.2622
For k=5, the silhouette score is 0.2452
For k=6, the silhouette score is 0.2830
```

```
Optimal number of clusters: 2
```

```
Cluster means (original scale):
```

	Unnamed: 0	crim	zn	...	ptratio	b	lstat
0	193.620896	0.287682	1.716418e+01	...	17.794030	384.797612	9.519254
1	370.807018	10.129061	1.243450e-14	...	19.751462	301.578129	18.792398

```
[2 rows x 14 columns]
```

```
Centroid coordinates (scaled):
```

	Unnamed: 0	crim	zn	...	ptratio	b	lstat
0	-0.409936	-0.387039	0.248957	...	-0.305855	0.308357	-0.439278
1	0.803091	0.758234	-0.487722	...	0.599189	-0.604091	0.860574

```
[2 rows x 14 columns]
```

Cluster feature mean: It is the average value of all samples in each cluster across all features in the original data scale. It reflects the average level of each feature in the original data space for this cluster. For example, in "Cluster means", the mean value of the crim feature in the 0th cluster is 0.287682, indicating that the average level of per capita crime rate of the samples in the 0th cluster is 0.287682.

Centroid coordinates: They are the coordinates of the center of each cluster in the standardized feature space after data standardization. Standardization involves scaling the data proportionally so that the mean is 0 and the variance is 1. The centroid

coordinates represent the position of the cluster center in this standardized space. For example, in "Centroid coordinates (scaled)", the centroid coordinate of the crim feature in the 0th cluster is -0.387039, reflecting the relative position of the cluster center in the standardized crim feature dimension.

Different numerical ranges:

The numerical range of cluster feature mean depends on the range of values of the original data. The range of values of the original data for different features may vary greatly. For instance, the value of the zn feature might be in the range of tens, while the value of the tax feature might be in the range of hundreds.

Centroid coordinates, due to the data being standardized, usually fluctuate around the mean of 0 and generally do not have very large absolute values. Most data will be distributed within a range of the mean plus or minus several times the standard deviation.

In summary, although both cluster feature mean and centroid coordinates are used to describe the central characteristics of the cluster, one is based on the original data scale and the other is based on the standardized data scale. There are significant differences in meaning and numerical range.

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?

```
Homogeneity score: 0.8788432003662366  
Completeness score: 0.8729636016078731
```

The homogeneity score is 0.8788432003662366, which measures whether each cluster contains only samples of a single category. The maximum score is 1. In the K-means clustering results of this wine dataset, most clusters have relatively pure sample categories within them, and the mixing degree of samples from different categories is relatively low.

The completeness score is 0.8729636016078731, which measures whether all samples of the same category are classified into the same cluster. The full score is 1. This result indicates that most of the wine samples of the same category have been successfully classified into the same cluster, but there are still a small number of samples of the same category that are scattered into different clusters.

Overall, these two indicators indicate that the clustering effect of the wine dataset based on the K-means algorithm (with $k = 3$) is relatively good, but there is still room for further optimization.

END