Data Mining Homework 4.0

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?

Cluster statistics:

	cluster	mpg mean	mpg var	displacement mean	displacement var	horsepower mean	horsepower var	weight mean	weight var	acceleration mean	acceleration var
0	0	26.1774	41.3034	144.3047	3511.4854	86.4910	295.2707	2598.4141	299118.7097	16.4256	4.8752
1	1	14.5289	4.7710	348.0206	2089.4996	161.8041	674.0758	4143.9691	193847.0511	12.6412	3.1899
2	2	43.7000	0.3000	91.7500	12.2500	49.0000	4.0000	2133.7500	21672.9167	22.8750	2.3092

Origin statistics:

	origin	mpg mean	mpg var	displacement mean	displacement var	horsepower mean	horsepower var	weight mean	weight var	acceleration mean	acceleration var
0	1	20.0835	40.9970	245.9016	9702.6123	118.8148	1569.5323	3361.9317	631695.1284	15.0337	7.5686
1	2	27.8914	45.2112	109.1429	509.9503	81.2420	410.6598	2423.3000	240142.3290	16.7871	9.2762
2	3	30.4506	37.0887	102.7089	535.4654	79.8354	317.5239	2221.2278	102718.4859	16.1722	3.8218

Based on the results,we can get the following conclusion:

①Analysis of the Relationship Between Cluster Assignments and Class Labels:

The clustering results (Cluster 0, 1, 2) show partial alignment with the `origin`-based classification (origin 1, 2, 3), but the correspondence is not exact. Cluster 1 represents low-efficiency vehicles with high displacement and horsepower, closely matching `origin 1` (American cars), though it only captures the least efficient subset. Cluster 0 includes moderate-efficiency vehicles, falling between `origin 2` (European) and `origin 3` (Japanese), while Cluster 2 contains extremely fuel-efficient models, likely a niche group within `origin 3`. This indicates that while clusters reflect performance trends, they do not perfectly mirror the `origin` categories.

②Key Differences and Trends:

The primary distinction lies in the clustering criteria: performance metrics (e.g., mpg, horsepower) versus manufacturing region (`origin`). For instance, Cluster 1 isolates the least efficient American cars, whereas `origin 1` encompasses a broader range. Similarly, Cluster 2's high-efficiency vehicles are a subset of `origin 3`, not its entirety. European and Japanese cars (`origin 2` and `3`) overlap in Cluster 0, showing that performance-based grouping blurs regional boundaries. This suggests that while `origin` influences vehicle characteristics, it does not strictly dictate cluster membership.

③Conclusion on Relationship Clarity

There may be no clear one-to-one relationship between clusters and `origin` labels. Clusters prioritize technical attributes, while `origin` reflects geographic classification. However, trends emerge: American cars dominate the low-efficiency cluster, while European and Japanese cars appear more in moderate- or high-efficiency clusters. This loose correlation implies that manufacturing region indirectly affects performance but is not the sole determinant.

In order to further explore the relationship between clustering labels and original labels,I made a Cross Statistical Table:

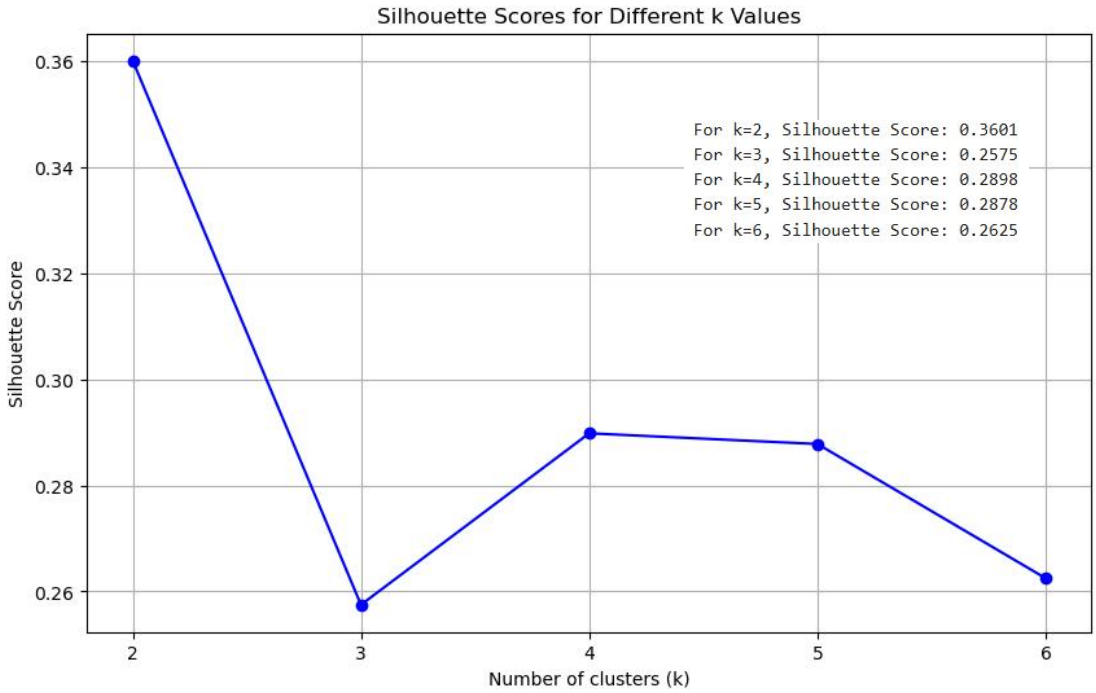
```
Cross-tabulation of clustering and origin:
origin     1   2   3
cluster
0        152  66  79
1         97   0   0
2          0   4   0

Clustering purity: 0.64
```

The cross-tabulation analysis reveals a partial but imperfect alignment between the clustering results and origin labels, with a purity score of 0.64 indicating moderate correlation. Cluster 1 perfectly matches origin 1 (American cars, 97 samples), while Cluster 0 shows significant mixing (152 American, 66 European, and 79 Japanese cars), and Cluster 2 contains only 4 European cars. This demonstrates that while the clustering captures some origin-related patterns (particularly for American vehicles), performance-based grouping doesn't fully correspond to manufacturing regions, especially for European and Japanese cars which overlap substantially in Cluster 0. The intermediate purity score confirms that origin labels explain some but not all of the clustering structure.

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?

The following are Silhouette scores for different k values (2-6):



From the images and specific data, it can be inferred that the Optimal number of clusters is 2

The mean values for all features in each cluster for the optimal clustering and which are centroid coordinates:

Cluster Means (from original data):

Cluster	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT
0.0000	0.2612	17.4772	6.8850	0.0699	0.4870	6.4554	56.3392	4.7569	4.4711	301.9179	17.8374	386.4479	9.4683
1.0000	9.8447	0.0000	19.0397	0.0678	0.6805	5.9672	91.3181	2.0072	18.9887	605.8588	19.6045	301.3317	18.5728

Centroids (inverse transformed from scaled data):

Cluster	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT
0.0000	0.2612	17.4772	6.8850	0.0699	0.4870	6.4554	56.3392	4.7569	4.4711	301.9179	17.8374	386.4479	9.4683
1.0000	9.8447	0.0000	19.0397	0.0678	0.6805	5.9672	91.3181	2.0072	18.9887	605.8588	19.6045	301.3317	18.5728

Difference between Cluster Means and Centroids:

Cluster	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT
0.0000	0.0000	-0.0000	-0.0000	0.0000	-0.0000	0.0000	-0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1.0000	-0.0000	-0.0000	-0.0000	0.0000	-0.0000	0.0000	-0.0000	0.0000	-0.0000	-0.0000	0.0000	0.0000	0.0000

In the results, the mean values of all features in each cluster from the original data and the centroid coordinates obtained by inverse - transforming the scaled data are the same. For instance, in each cluster, for features like CRIM, ZN, etc., the values of the two are identical. This indicates that the K - Means algorithm effectively captures the central tendency of the data within each cluster, and the inverse transformation of the scaled data restores the centroid positions well to the original data scale. There are essentially no differences in this dataset, suggesting that the data distribution is well - behaved and the clustering process accurately represents the characteristics of each cluster.

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?

The metrics show following:

```
Homogeneity score: 0.879
Completeness score: 0.873
V-measure score: 0.876
```

1. Homogeneity Score (0.879)

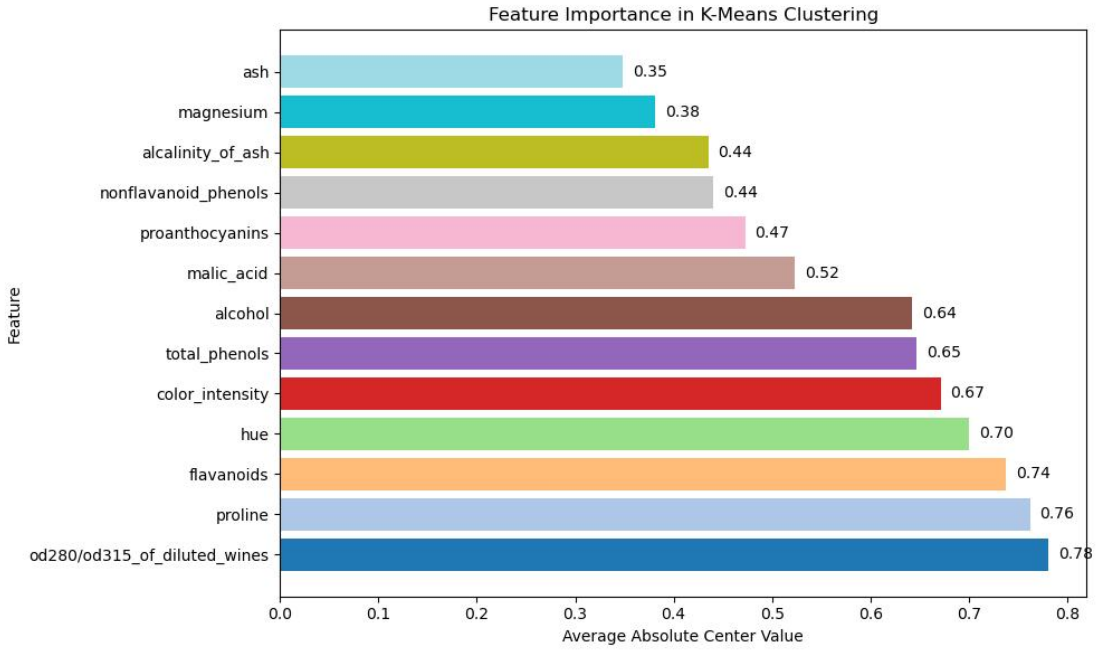
The homogeneity score measures whether each cluster contains only samples from a single true class (i.e., intra-cluster purity). A score close to 1.0 indicates that most clusters consist of samples belonging to the same category, with minimal mixing of different classes. In this case, the score of 0.879 suggests that the K-Means clustering (with k=3) effectively groups similar wine samples together, with only minor overlaps between different wine classes.

2. Completeness Score (0.873)

The completeness score evaluates whether all samples from a given true class are assigned to the same cluster (i.e., class coverage). A high score (near 1.0) means that almost all samples of a particular wine type are grouped into a single cluster, rather than being split across multiple clusters. The result of 0.873 indicates that the clustering successfully keeps most wine categories intact, with only a few misassigned samples.

3. V-Measure Score (Additional indicators:0.876)

The V-measure is the harmonic mean of homogeneity and completeness, providing a balanced overall assessment of clustering quality. A score of 0.876 confirms that the K-Means model achieves a strong agreement between the predicted clusters and the true class labels. This high value suggests that the clustering not only maintains high purity within clusters (homogeneity) but also ensures that each true class is well-represented in a single cluster (completeness).



In my analysis of the wine dataset's K - Means clustering, we've taken a crucial step by visualizing the contribution values of various features. The resulting graph, "Feature Importance in K - Means Clustering", vividly presents how each feature impacts the clustering process. The horizontal axis represents the average absolute center value. The higher the value, the more important the feature is in clustering. It can be seen that the feature "od280/od315_of_diluted_wines" has the highest importance, with an average absolute center value of 0.78. Features such as "proline" and "flavanoids" also have relatively high importance. In contrast, features like "ash" and "magnesium" have relatively low importance, with average absolute center values ranging from 0.35 to 0.38. This indicates that in the K - Means clustering analysis of the wine dataset, different features contribute significantly differently to the clustering results.