Cluster evaluation

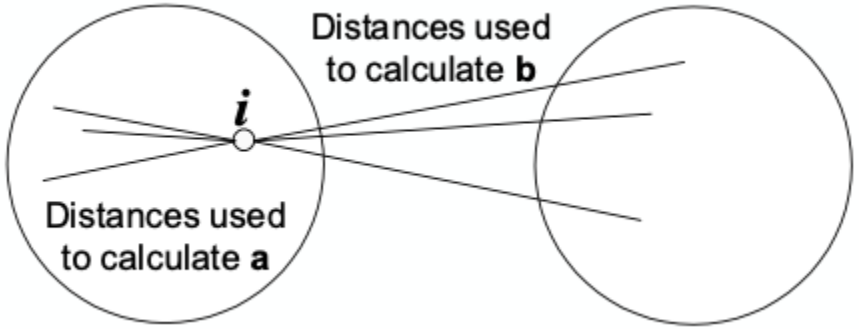
Dr. Huiping Cao

Internal Measures: Cohesion and Separation

- Cluster Cohesion: Measures how closely related are objects in a cluster
  - Example: SSE
- Cluster Separation: Measure how distinct or well-separated a cluster is from other clusters
  - Example: Squared Error

Internal Measures: Silhouette Coefficient

- **Silhouette coefficient** combines ideas of both cohesion and separation, but **for individual points**, as well as **clusters** and **clusterings**
- For an individual point, i
  - Calculate a = average distance of *i* to the points in its cluster
  - Calculate b= min (average distance of *i* to points in another cluster)
  - The silhouette coefficient for a point is then given by $s = \frac{b-a}{\max(a,b)}$

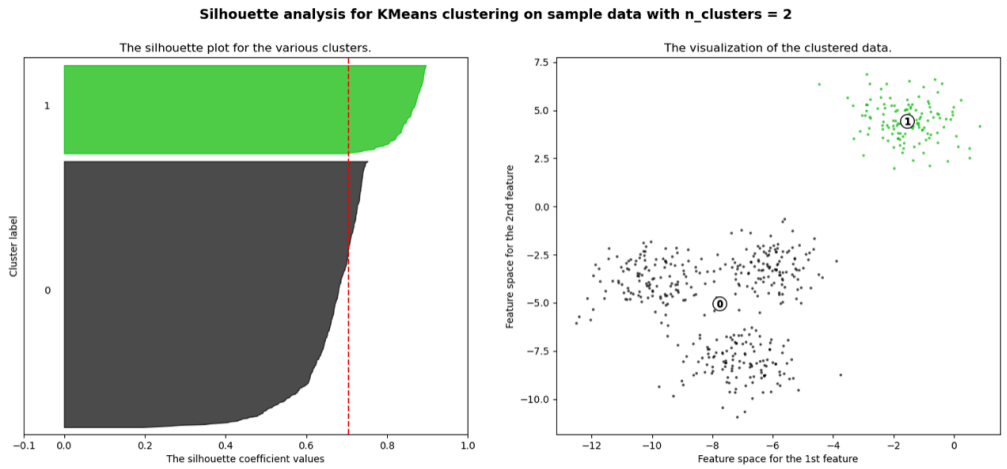


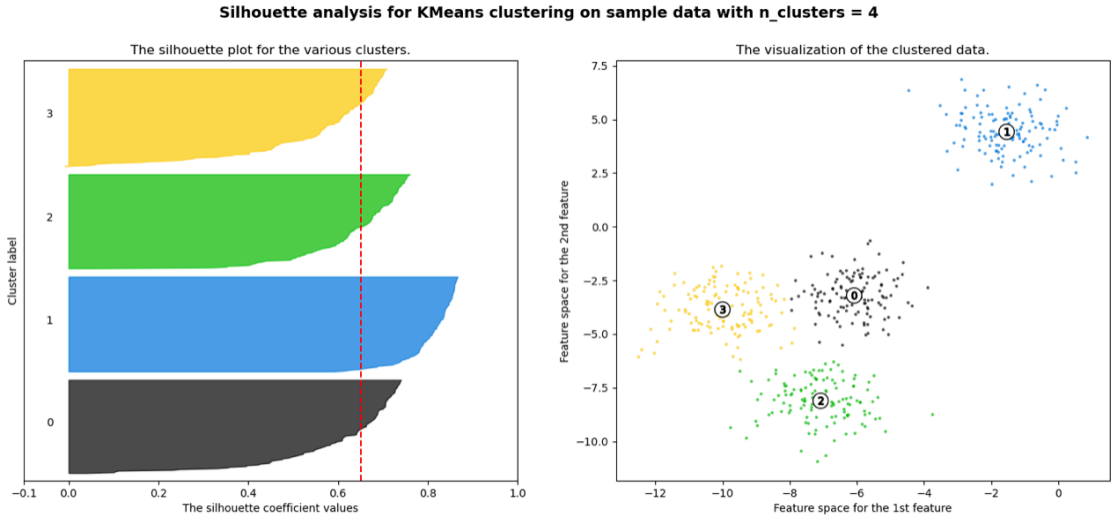
Internal Measures: Silhouette Coefficient (Cont.)

- Silhouette coefficient ranges between -1 and 1.
  - A negative value is undesirable because this corresponds to a case in which a is greater than b. Negative values indicate that those samples might have been assigned to the wrong cluster.
  - A positive value is desired. “+1” indicate that the sample is far away from the neighboring clusters.
  - A value of 0 indicates that the sample is on or very close to the decision boundary between two neighboring clusters.
- We can compute the average Silhouette coefficients of a cluster by simply taking the average of the silhouette coefficients of points belong to the cluster.
- An overall measure of the goodness of a clustering can be obtained by computing the average silhouette coefficient of all points.

Use Silhouette Coefficient to Determine the Number of Clusters

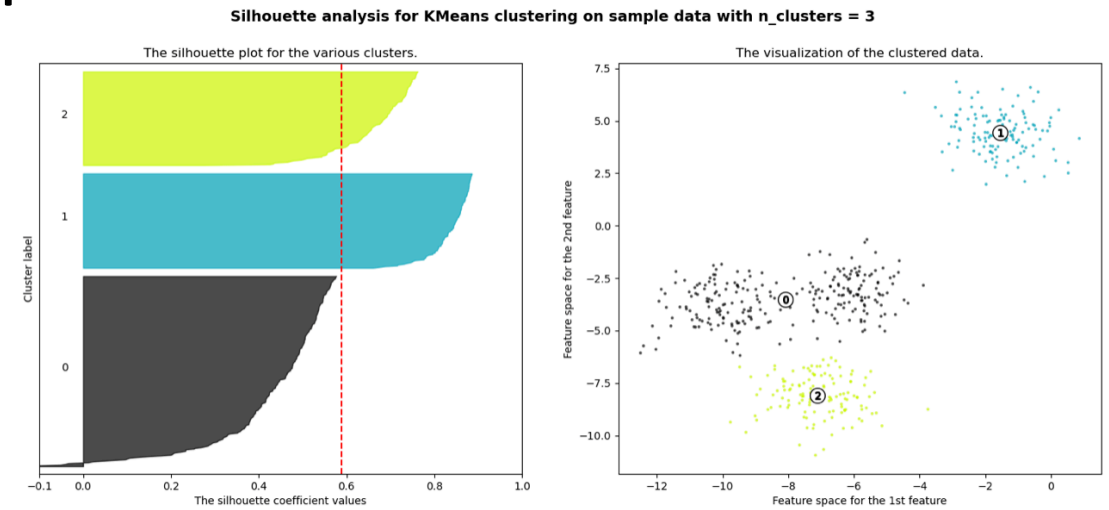
- A bad pick if there are clusters with below average silhouette scores and if there are wide fluctuations in the size of the silhouette plots.
- Silhouette analysis is more ambivalent in deciding between 2 and 4.

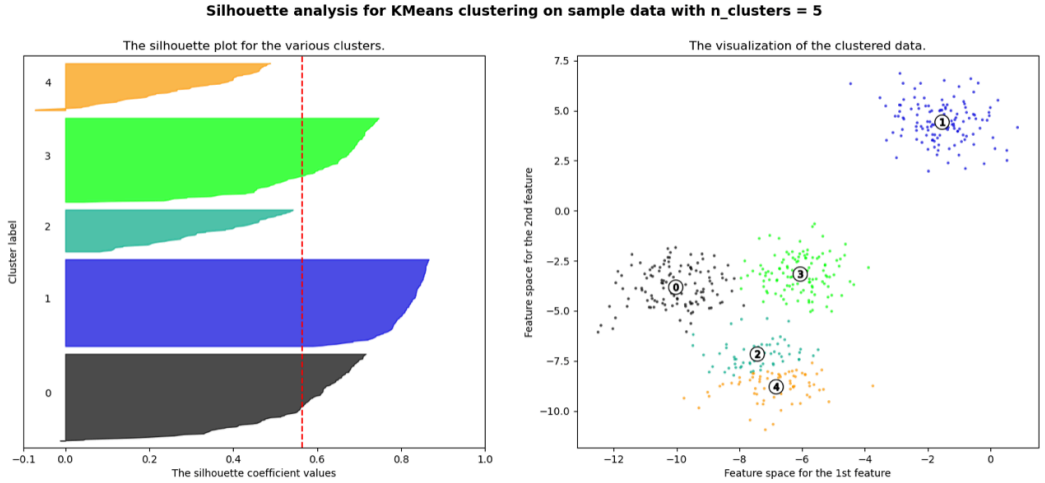


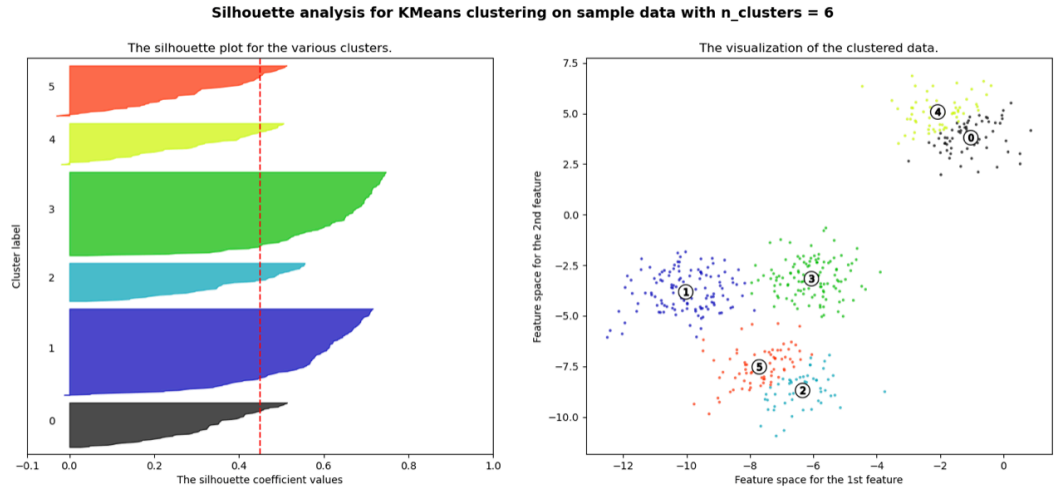


Use Silhouette Coefficient to Determine the Number of Clusters (cont.)

- It is bad to pick K=3, 5, 6







Use Silhouette Coefficient to Determine the Number of Clusters (cont.)

- The thickness of the silhouette plot can show the cluster size.
  - The silhouette plot of Cluster 0 for K = 2, is bigger in size owing to the grouping of the 3 sub clusters into one big cluster.
  - When K = 4, all the plots are more or less of similar thickness and hence are of similar sizes as can be also verified from the labelled scatter plot on the right.

Use datasets for classification

- Use the features to do clustering analysis
- Use the class labels to verify the clustering results

References

- https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html
- Selecting the number of clusters with silhouette analysis on KMeans clustering. https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html
- Chapter 7: Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, and Vipin Kumar: Introduction to Data Mining, 2nd Edition.