Clustering model evaluation

Prof. Junghyun Kim
Director, Engineering Systems Design Laboratory (ESDL)
Assistant Professor, School of Applied Artificial Intelligence
Handong Global University

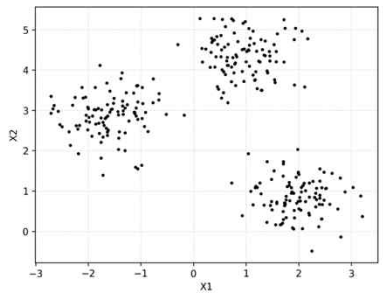
Course objectives

- By the end of this module, you will be able to answer the following questions:
  - How can we evaluate a clustering model?
    - Note that we primarily focus on partitioning-based clustering algorithms

Clustering model evaluation

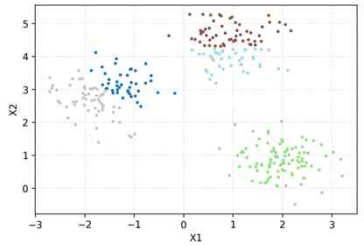
Clustering model evaluation

- We have set the number of clusters (K) to three for the datasets below
  - The decision stems from a clear observation within the datasets, making it evident to set K as 3

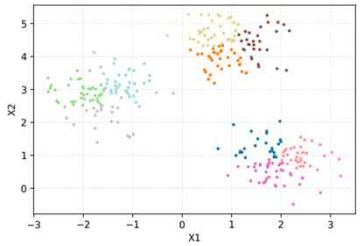


Clustering model evaluation

- In general, it is not easy to know how to set K in reality
  - The result might be quite bad if you would set it to the wrong value



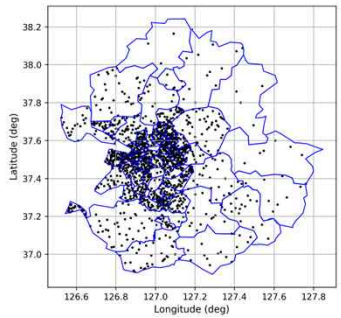
When $K = 5$



When $K = 9$

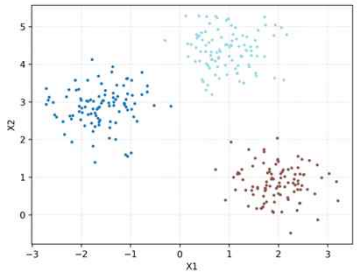
Clustering model evaluation

- Discussion
  - What if you have the following datasets? How would you find the optimal number of clusters?

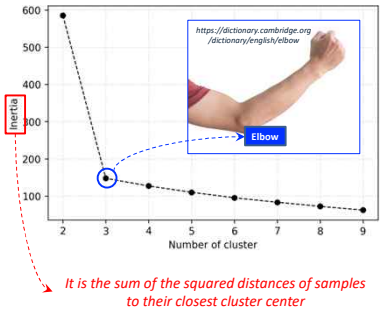


Clustering model evaluation

- Proposed approach 1. Elbow method
  - The elbow method is a technique used for determining the optimal number of clusters in a dataset



When $K = 3$



It is the sum of the squared distances of samples to their closest cluster center

Clustering model evaluation

- Proposed approach 1. Elbow method
  - Pros
    - Ease of use: it is simple to understand and implement
    - Quick evaluation: it provides a relatively quick way to estimate the optimal number of clusters
    - …
  - Cons
    - Subjectivity: determining the exact elbow point is somewhat subjective
    - Works best with well-separated clusters: it is most effective when clusters are well-separated. In other words, if data points are densely packed or have irregular shapes, it may be challenging to identify a clear elbow in the plot (i.e., inertia vs. number of cluster)
    - …

In practice, it is recommended to complement the elbow method with other techniques when determining the optimal number of clusters

Clustering model evaluation

- Proposed approach 2. Silhouette score
  - It is a metric used to evaluate the quality of clusters in a clustering analysis
  - It provides a measure of how similar a data point is to its own cluster (i.e., cohesion) compared to other clusters (i.e., separation)
    - A higher silhouette score indicates that the clusters may be well-separated and the data points within each cluster may be similar to each other
    - A lower silhouette score means that the clusters may be overlapping or poorly defined
  - It is calculated by the following equation:

$$\text{Silhouette score} = \frac{b - a}{\max(a, b)}$$

This represents how well the data point is clustered with its peers

This indicates the separation between the data point's cluster and neighboring clusters

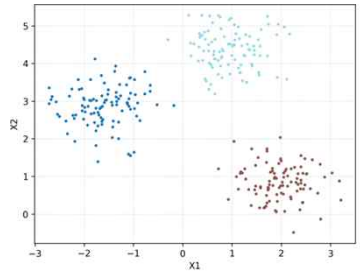
Where:

a = The mean distance between the data point and all other data points in the same cluster

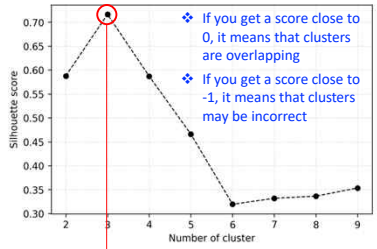
b = The smallest mean distance between the data point and all other data points in different clusters, excluding its own

Clustering model evaluation

- Proposed approach 2. Silhouette score
  - When using the silhouette score [-1, 1] to determine the optimal number of clusters, it typically computes the score for different values of K and choose the value that yields the highest score



When $K = 3$

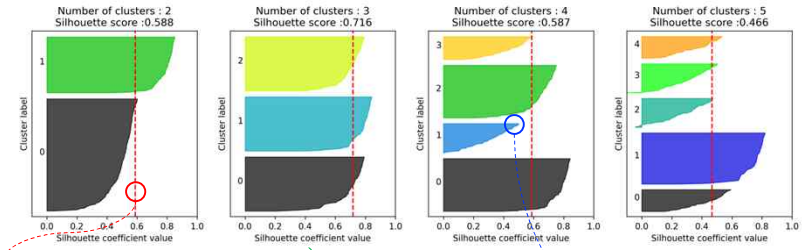


- If you get a score close to 0, it means that clusters are overlapping
- If you get a score close to -1, it means that clusters may be incorrect

A score close to +1 means that the instance may be well inside its own cluster and far from other clusters

Clustering model evaluation

- Proposed approach 3. Silhouette diagram
  - An even more informative visualization is obtained when we plot every data point's silhouette score, sorted by the cluster they are assigned to and by the value of the score, called "**Silhouette diagram**"



The vertical line represents the silhouette score

Most data points extend beyond the dashed line and closer to 1.0 (i.e., relatively good clusters)

The cluster may be bad as it stops short of the dash line

Course summary

- Throughout this module, you have learned:
  - How can we evaluate a clustering model?

THANK YOU

For more information, please reach out to Prof. Junghyun Kim at
junghyun.kim@handong.edu