Inclass 16: Model Evaluation and Selection

[SCS4049] Machine Learning and Data Science

Seongsik Park (s.park@dgu.edu)

Department of Artificial Intelligence, Dongguk University

K-means clustering: inertia

A performance metric, called *inertia*

intertia = the mean squared distance

between each instance and its closest centroid

Elbow rule: any lower value would be dramatic, while any higher value would not help much.

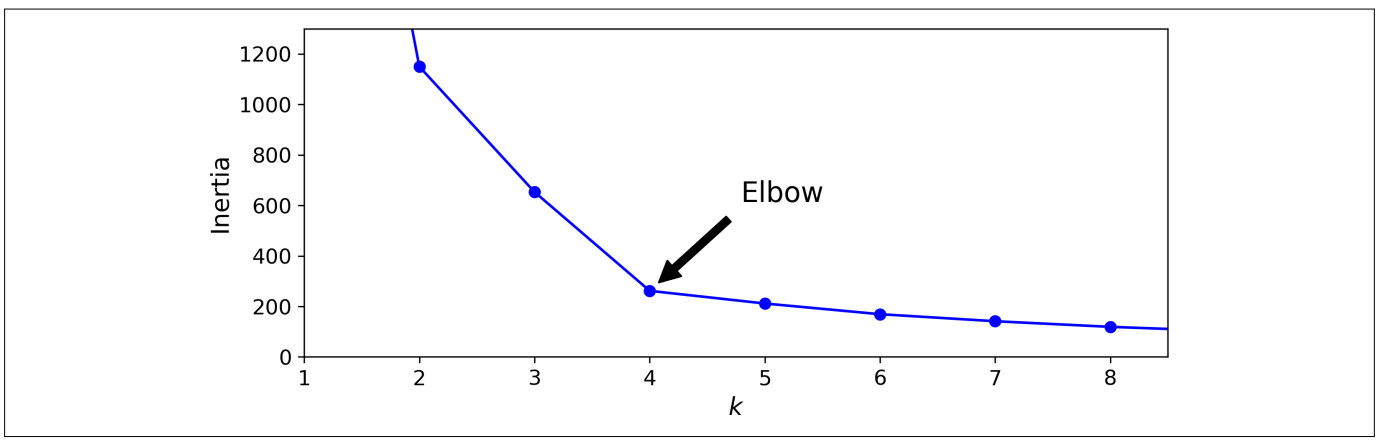

Figure 9-8. Selecting the number of clusters k using the "elbow rule"

K-means clustering: silhouette score

A more precise approach (but also more computationally expensive) is to use the *silhouette score*, which is the mean *silhouette coefficient* over all the instances.

- An instance's silhouette coefficient is equal to $(b - a)/\max(a, b)$
- where a is the mean distance to the other instances in the same cluster (i.e., the mean intra-cluster distance)
- and b is the mean nearest-cluster distance (i.e., the mean distance to the instances of the next closest cluster, defined as the one that minimizes b, excluding the instance's own cluster).

K-means clustering: silhouette score

The silhouette coefficient can vary between -1 and +1.

- A coefficient close to +1 means that the instance is well inside its own cluster and far from other clusters,
- while a coefficient close to 0 means that it is close to a cluster boundary,
- and finally a coefficient close to -1 means that the instance may have been assigned to the wrong cluster.

K-means clustering: silhouette score

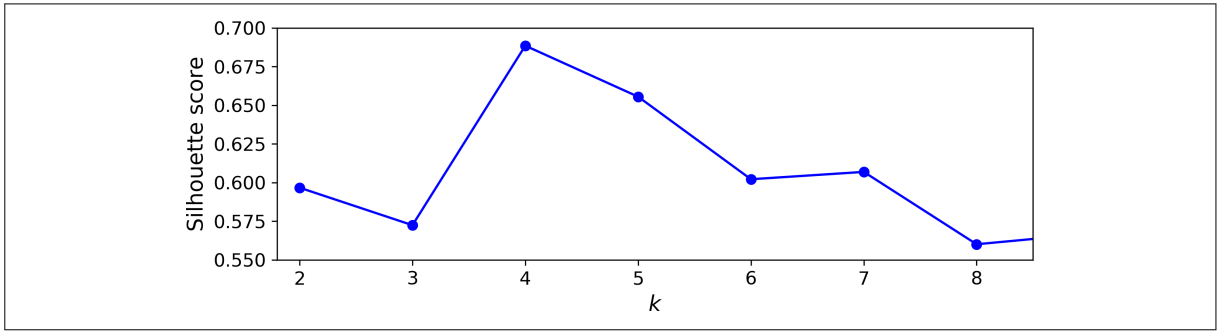

Figure 9-9. Selecting the number of clusters k using the silhouette score coefficient. This is called a *silhouette diagram* (see Figure 9-10):

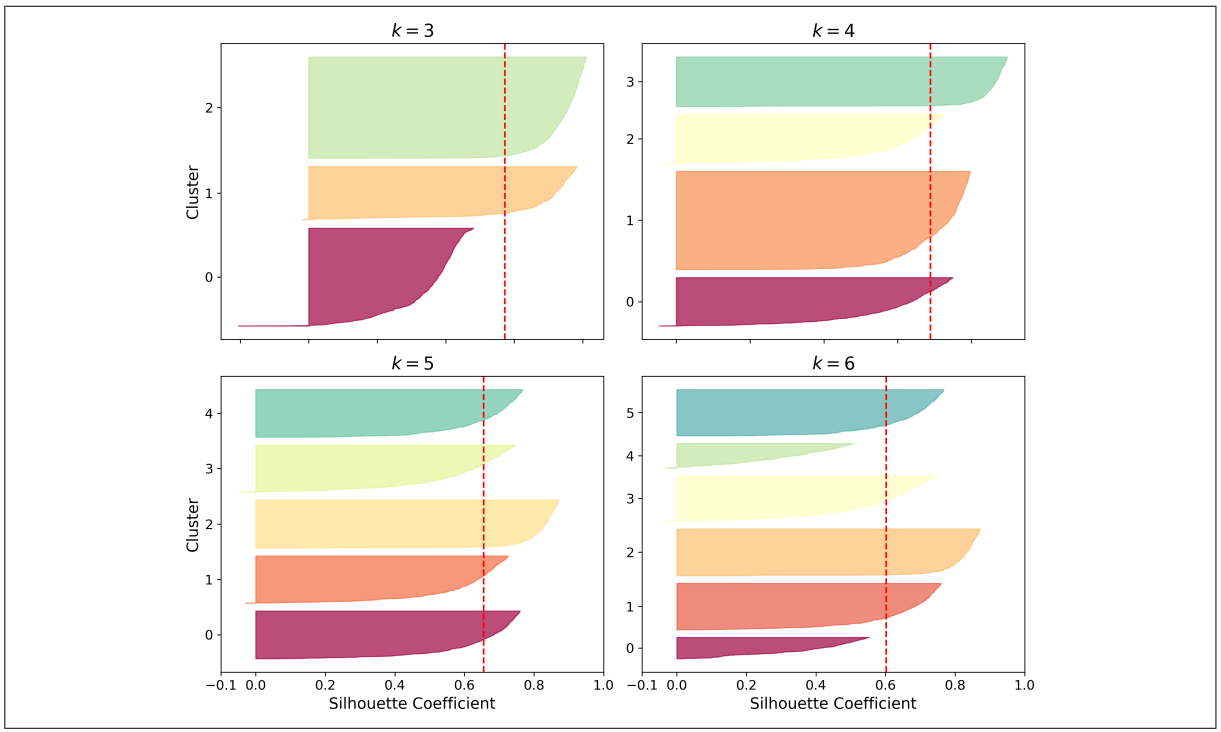

Figure 9-10. Silouhette analysis: comparing the silhouette diagrams for various values of k