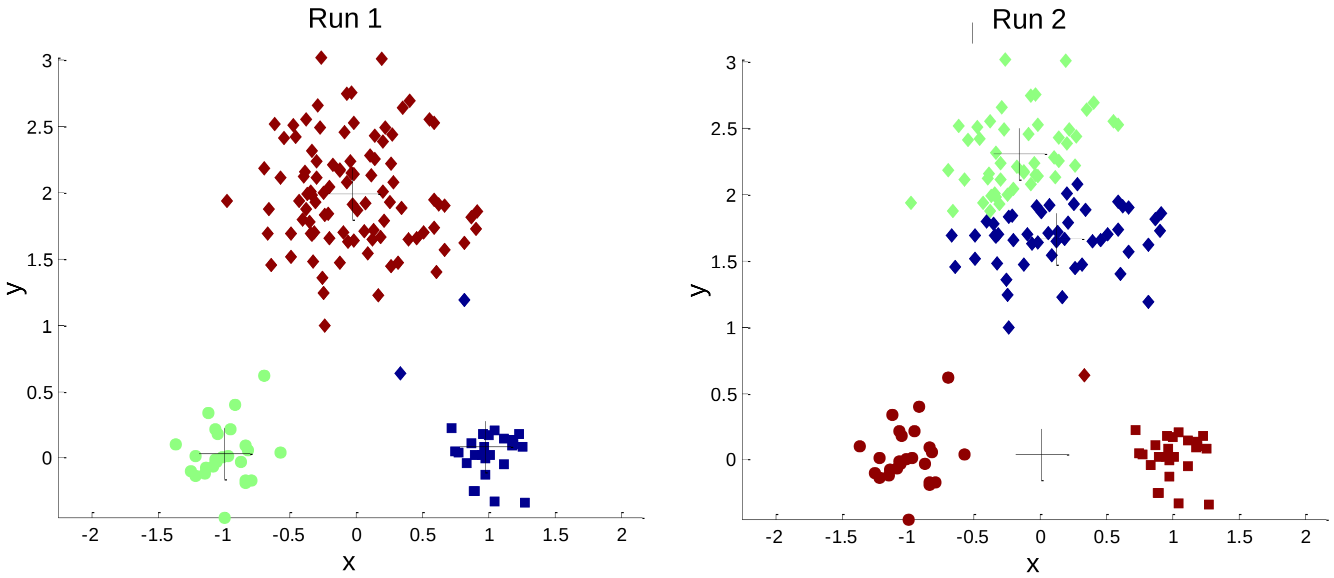
Foundations of Data Science & Analytics: Cluster Validation

Ezgi Siir Kibris

Introduction to Data Mining, 2nd Edition
by
Tan, Steinbach, Karpatne, Kumar

Clustering Comparison



We can qualitatively “see” that Run 1 produced a better result than Run 2, but how do we quantify this?

Evaluating Clusters

- Most common measure is Sum of Squared Error
  - For each data instance, its **error** is the distance to the nearest centroid
  - To get SSE, we square all errors and sum them

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist(m_i, x)^2$$

  - *x* is a data instance in cluster C_i and m_i is the centroid of cluster C_i
  - Given multiple runs of K-means, we typically choose the run with the smallest error

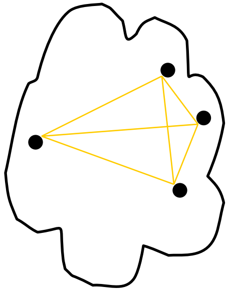
Cluster Validation

- For supervised clustering tasks, where class values are known, we can compute **cluster accuracy** by determining how many data instances ended up in the "correct" cluster for its class value.
- For unsupervised clustering tasks, we can **compare** two runs of k-means using their SSEs.
- But SSE is not the only measure to go by. It is also important to examine the **structure** of the clusters.

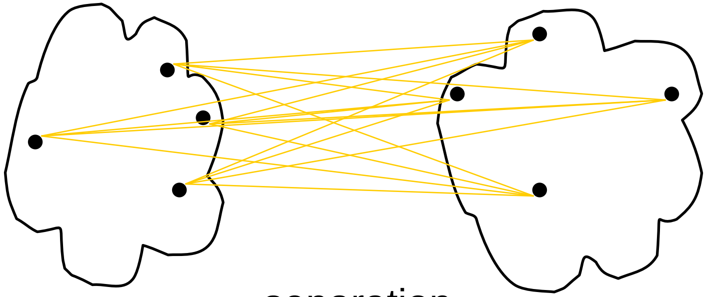
Cluster Cohesion and Separation

- **Cohesion** measures the closeness of data instances within a cluster

- **Separation** measure how well-separated a cluster is from other clusters



cohesion

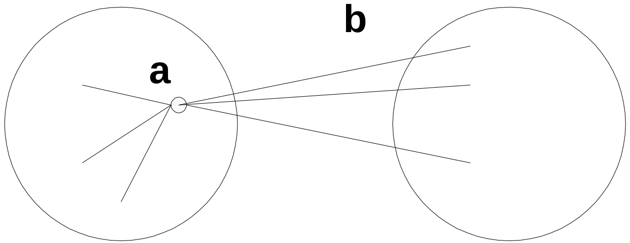


separation

Silhouette Coefficient

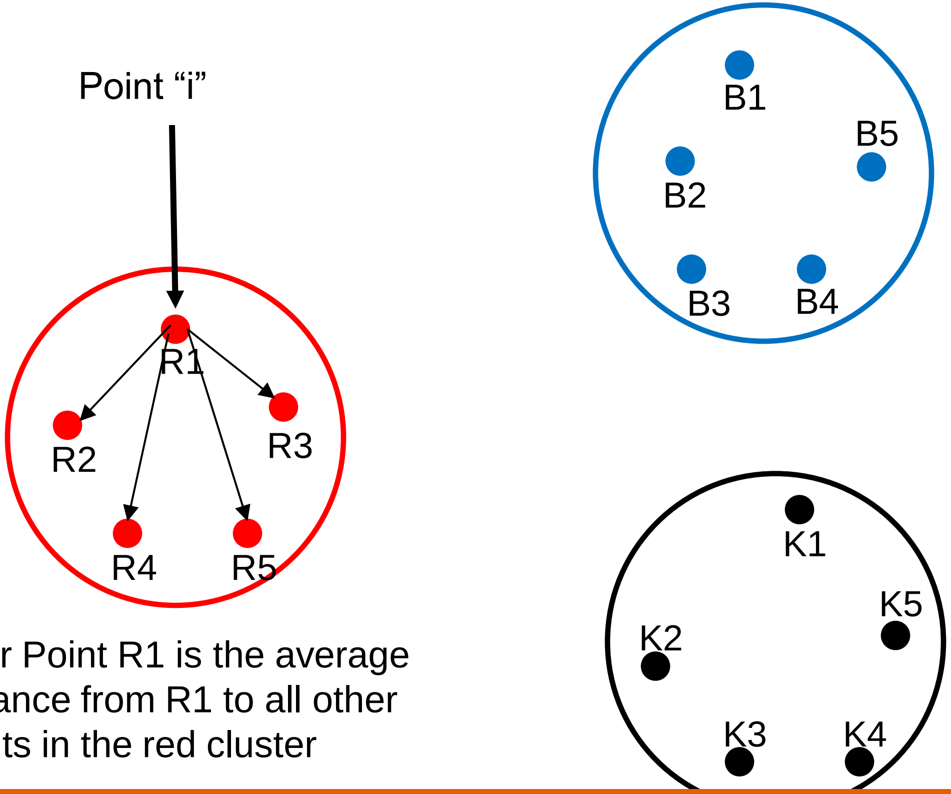
- Silhouette Coefficient combines the ideas of cohesion and separation into a single metric.
- For each data instance, *i*
  - Calculate $\boldsymbol{a_i}$ = average distance of *i* to all points in its cluster
  - Calculate $\boldsymbol{b_i}$ = minimum of the average distance of *i* to all points in each respective cluster
  - The silhouette coefficient, $\boldsymbol{s_i}$ for a point is

$$s_i = \left|1 - \frac{a_i}{b_i}\right|$$



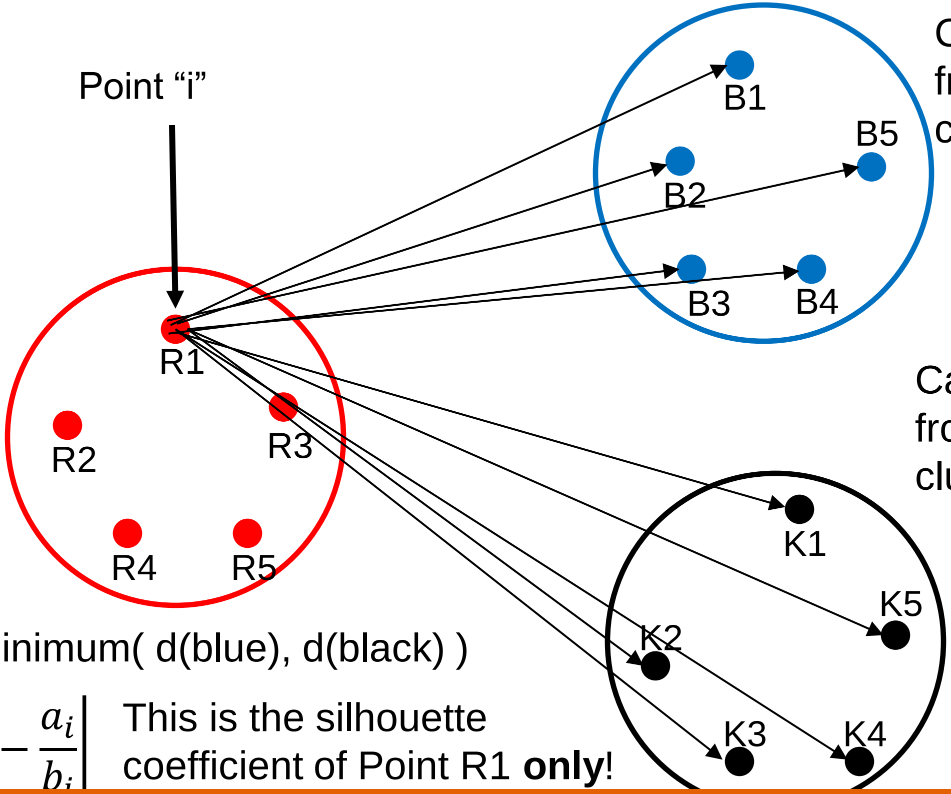
  - Ranges between 0 and 1
  - The closer to 1 the better
  - Average Silhouette Coefficient is calculated for a clustering and compared with another clustering

Silhouette Coefficient Example - a_i



a_i for Point R1 is the average distance from R1 to all other points in the red cluster

Silhouette Coefficient Example – b_i,s_i



Calculate the average distance from R1 to all points in blue cluster: d(blue)

Calculate the average distance from R1 to all points in black cluster: d(black)

b_i = minimum(d(blue), d(black))

$$s_i = \left|1 - \frac{a_i}{b_i}\right|$$

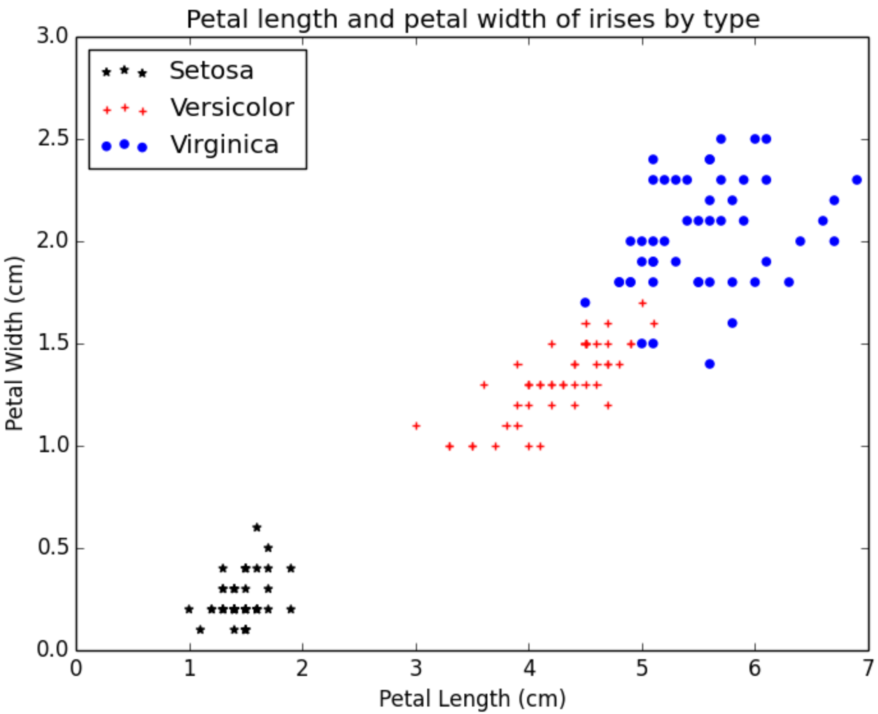
This is the silhouette coefficient of Point R1 **only**!

Compute a_i, b_i, s_i for all other points in the clustering

How to Choose a k Value

- One “reliable” method for choosing a **k** value for an unsupervised clustering problem is to use the average silhouette coefficient
- General Method
  - Choose several **k** values (2, 3, 5, 7, for example)
  - Run k-means several times for each **k** value and compute the average silhouette coefficient for the best run
  - The **k** value with the highest average silhouette coefficient is what the number of clusters, **k**, “should” be

Iris Scatter Plot



Since we already know the class labels, we know that there should be three clusters, but what happens if we calculate the average silhouette coefficient for 2, 3, 5, and 7 clusters?

SSE/Average Silhouette Coefficient Results

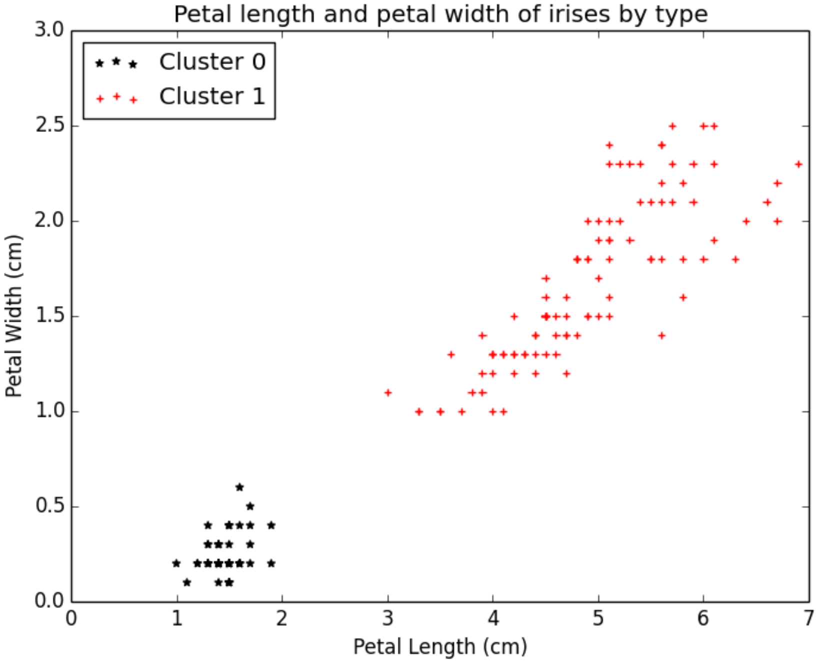
K	SSE
2	12.14
3	6.998
5	5.131
7	3.758

As more clusters are generated, the SSE decreases because there are more clusters spread across the data, causing the instances in the clusters to be closer together.

K	Avg Silhouette Coefficient
2	0.689
3	0.551
5	0.450
7	0.306

The silhouette coefficient considers both **within** cluster separation and **between** cluster separation. This is an important distinction between it and SSE. This result indicates that the iris data set should only have **2** clusters!

Cluster Visualization (k=2)



Domain experts have stated that there are 3 types of irises, but unsupervised learning says there are only 2!

Assignment 10