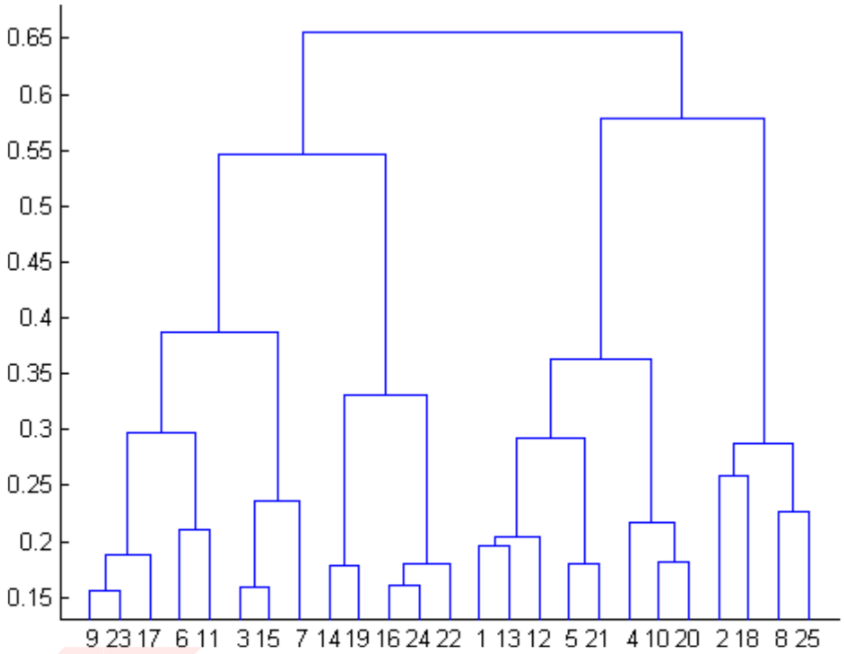
MACHINE LEARNING

Q1 to Q12 have only one correct answer. Choose the correct option to answer your question.

1. What is the most appropriate no. of clusters for the data points represented by the following dendrogram:



a) 2
b) 4
c) 6
d) 8

Answer : (b) 4

2. In which of the following cases will K-Means clustering fail to give good results?
   1. Data points with outliers
   2. Data points with different densities
   3. Data points with round shapes
   4. Data points with non-convex shapes

a) 1 and 2
b) 2 and 3
c) 2 and 4
d) 1, 2 and 4

Answer (d) 1,2 and 4

3. The most important part of is selecting the variables on which clustering is based.

a) interpreting and profiling clusters
b) selecting a clustering procedure
c) assessing the validity of clustering
d) formulating the clustering problem

Answer (d) formulating the clustering problem

4. The most commonly used measure of similarity is the____or its square.

a) Euclidean distance
b) city-block distance
c) Chebyshev's distance
d) Manhattan distance

Answer (a) Euclidean distance

MACHINE LEARNING

5. is a clustering procedure where all objects start out in one giant cluster. Clusters are formed by dividing this cluster into smaller and smaller clusters.
   a) Non-hierarchical clustering
   b) Divisive clustering
   c) Agglomerative clustering
   d) K-means clustering

Answer (b) Divisive clustering

6. Which of the following is required by K-means clustering?
   a) Defined distance metric
   b) Number of clusters
   c) Initial guess as to cluster centroids
   d) All answers are correct

Answer (d) All answers are correct

7. The goal of clustering is to-
   a) Divide the data points into groups
   b) Classify the data point into different classes
   c) Predict the output values of input data points
   d) All of the above

Answer (d) All of the above

8. Clustering is a-
   a) Supervised learning
   b) Unsupervised learning
   c) Reinforcement learning
   d) None

Answer (b) Unsupervised learning

9. Which of the following clustering algorithms suffers from the problem of convergence at local optima?
   a) K- Means clustering
   b) Hierarchical clustering
   c) Diverse clustering
   d) All of the above

Answer (d) All of the above

10. Which version of the clustering algorithm is most sensitive to outliers?
    a) K-means clustering algorithm
    b) K-modes clustering algorithm
    c) K-medians clustering algorithm
    d) None

Answer (a) K-means clustering algorithm

11. Which of the following is a bad characteristic of a dataset for clustering analysis-
    a) Data points with outliers
    b) Data points with different densities
    c) Data points with non-convex shapes
    d) All of the above

Answer (d) All of the above

12. For clustering, we do not require-
    a) Labeled data
    b) Unlabeled data
    c) Numerical data
    d) Categorical data

Answer (a) Labeled data

Q13 to Q15 are subjective answers type questions, Answers them in their own words briefly.

13. How is cluster analysis calculated?

Clustering is an unsupervised machine learning method that can identify groups of similar data points, known as clusters, from the data itself. For some clustering algorithms, such as K-means, one needs to know how many clusters there are beforehand. If the number of clusters is incorrectly specified, the results are not very informative (see Figure 1).

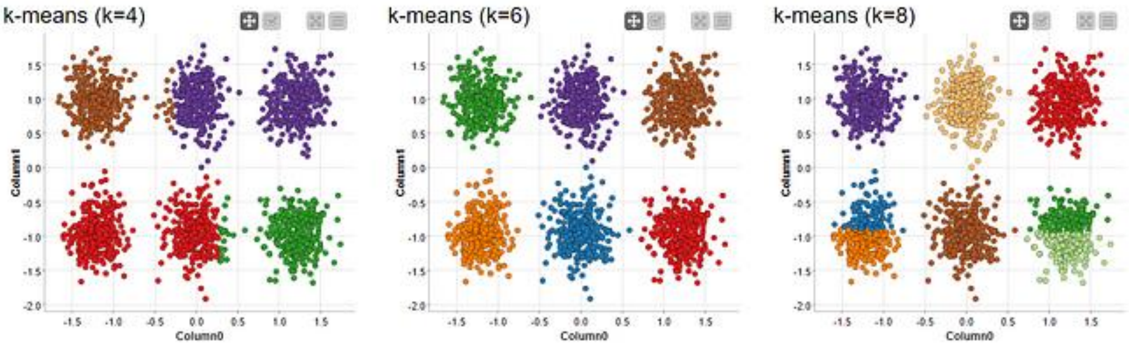


Figure 1: Clustering with different number of clusters, k=4, 6, & 8. Simulated data with 6 clusters.

Unfortunately in many instances we do not know how many clusters there are in our data.
One possible solution in determining the correct number of clusters is a brute-force approach. We try applying a clustering algorithm with different numbers of clusters. Then, we find the magic number that optimizes the quality of the clustering results. In this article, we first introduce two popular metrics to assess cluster quality.

We then cover three approaches to find the optimal number of clusters:

- The elbow method
- The optimization of the silhouette coefficient
- The gap statistic

- **Quality of Clustering Outcome**
  Before getting into different methods to determine the optimal number of clusters, we shall see how we can quantitatively assess the quality of clustering outcomes. Imagine the following scenarios. The same data set is clustered into three clusters (see Figure 2). As you can see, the clusters are defined well on the left, whereas the clusters are identified poorly on the right.

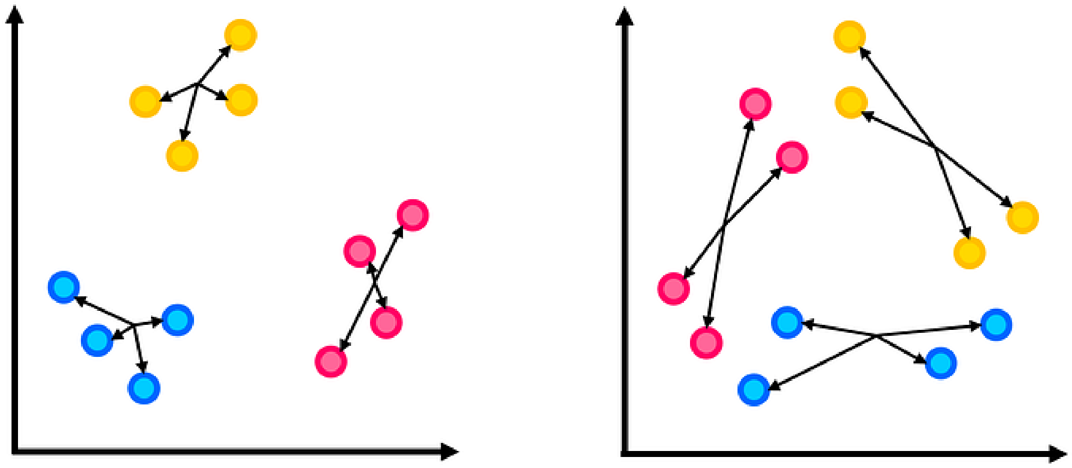


Figure 2: Examples of well-defined clusters (left) and poorly-defined clusters (right) based on the same data set. The arrows indicate the distance between the data points and their cluster centers.

Why is that? Remember that the goal of clustering is to group data points in clusters so that (1) points within a cluster are as similar as possible, (2) points belonging to different clusters are as distinct as possible. This means that, in ideal clustering, the within-cluster variation is small whereas the between-cluster variation is large. Consequently a good clustering quality metric should be able to summarize (1) and/or (2) quantitatively.

One such quality metric is inertia. This is calculated as the sum of squared distances between data points and the centers of the clusters they belong to. Inertia quantifies the within-cluster variation.

Another popular metric is the silhouette coefficient, which attempts to summarize both within-cluster and between-cluster variation. At each data point, we calculate the distance to the cluster center in which the data point belongs (referred to as *a*), as well as the distance to the second best cluster center (referred to as *b*). Here, the second best cluster refers to the closest cluster that is not the current data point's cluster. Then based, on these two distances *a* and *b*, the silhouette *s* of that data point is calculated as s=(b-a)/max(a,b).

Under ideal clustering, the distance *a* is very small compared to the distance *b*, resulting in *s* being close to 1 (see Figure 3, left). If clustering is somewhat suboptimal, then the distances *a* and *b* may not differ dramatically (see Figure 3, center). In that case *s* is close to 0. If clustering is even worse, then the distance *a* may actually be larger than the distance *b* (see Figure 3, right). In such a scenario, *s* becomes negative, close to -1.

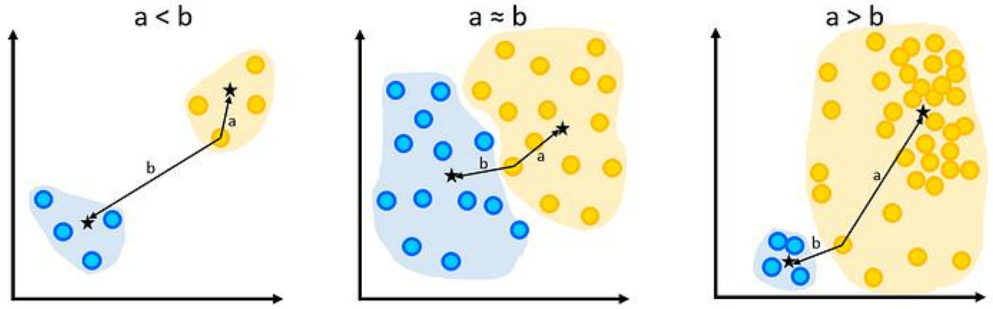


Figure 3: Scenarios where clustering is optimal (left), suboptimal (center), and even worse (right). The stars indicate cluster centers. Image by author.

Once *s* is calculated at all data points, the average of *s* determines a silhouette coefficient. A silhouette coefficient can be calculated for each cluster separately, or for all data points. A silhouette coefficient close to 1 indicates that a clustering algorithm is able to partition data into well-separated clusters.

- **Elbow Method**
- The inertia is a decreasing function of the number of clusters *k*. However, its rate of decrease is different above or below the optimal number of clusters *K*. For *k*<*K*, the inertia decreases rapidly, whereas the decrease is slow for *k*>*K*. Thus, by plotting the inertia over a range of *k*, one can determine where the curve bends, or elbows, at K. Figure 4 shows an inertia plot from our example in Figure 1. We can clearly see a bend, or the elbow, at *k*=6.

- This method, however, is somewhat subjective, as different people may identify the elbow at different locations. In our example in Figure 4, some may argue that *k*=4 is the elbow. Moreover, the elbow may not be always apparent, as we shall see later.

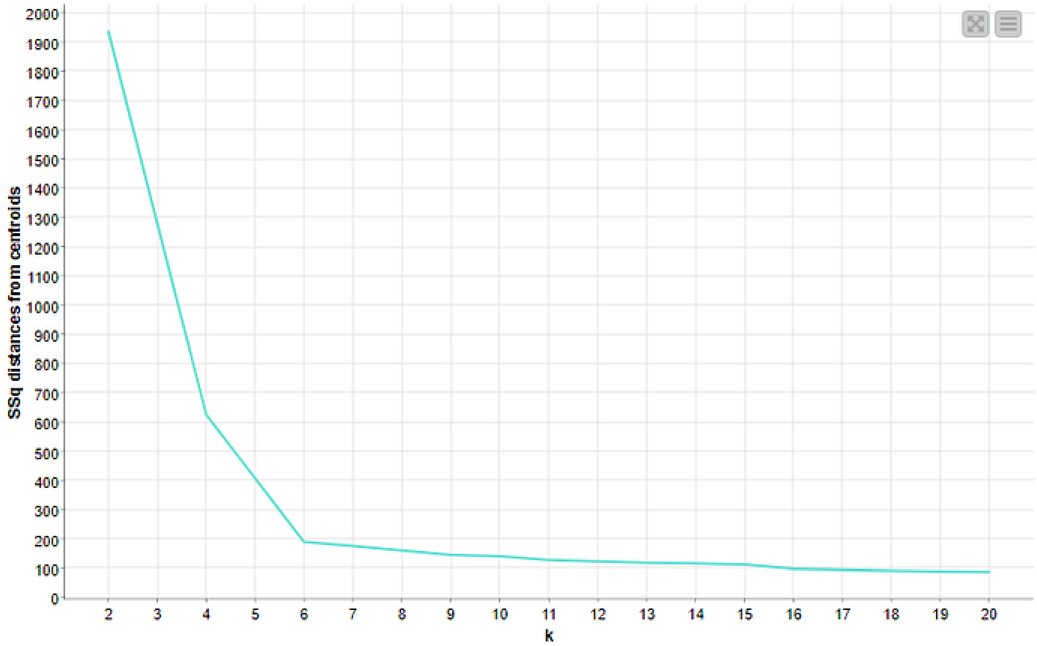


Figure 4: The plot of the inertia for different k, for the data set presented in Figure 1.

Silhouette Method

The silhouette coefficient may provide a more objective means to determine the optimal number of clusters. This is done by simply calculating the silhouette coefficient over a range of *k*, and identifying the peak as the optimum *K*. A KNIME component Optimized K-Means (Silhouette Coefficient) does exactly that. It performs K-Means clustering over a range of *k*, finds the optimal *K* that produces the largest silhouette coefficient, and assigns data points to clusters based on the optimized *K*. Figure 5 shows an example of a silhouette coefficient plot from our example data presented in Figure 1. As it can be seen, the silhouette coefficient peaks at *k*=6, and thus it is determined as the optimum K.

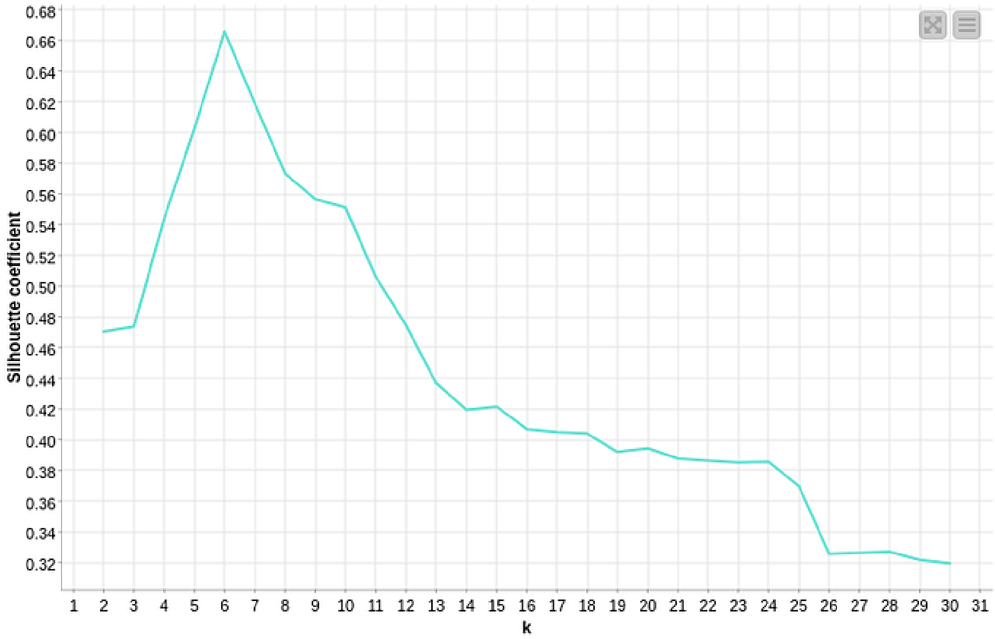


Figure 5: The plot of the silhouette coefficient for different k, for the data set presented in Figure 1.

Gap Statistic

To talk about gap statistics, let’s consider clustering of a random data set with no cluster organization whatsoever. Say a random data set is clustered into *k* clusters, and the inertia is calculated based on the resulting clustering (see Figure 6). Despite the lack of underlying cluster organization, the clustered random data produces steadily decreasing inertiae (plural of inertia) as *k* increases. This is because the more cluster centers there are, the smaller the distance becomes between data points to the cluster centers, producing decaying inertiae. In contrast, as we have already seen in Figure 4, the rate of decrease in inertia varies whether *k* is below or above the optimum number of clusters *K* in a data set with cluster organization. When the inertia for the observed and random data are plotted together, the difference becomes apparent (see Figure 7). The gap statistic is calculated by comparing the inertiae from a (hopefully) clustered data set and a corresponding random data set covering the same ranges in the data space (Tibshirani et al., (2001)).

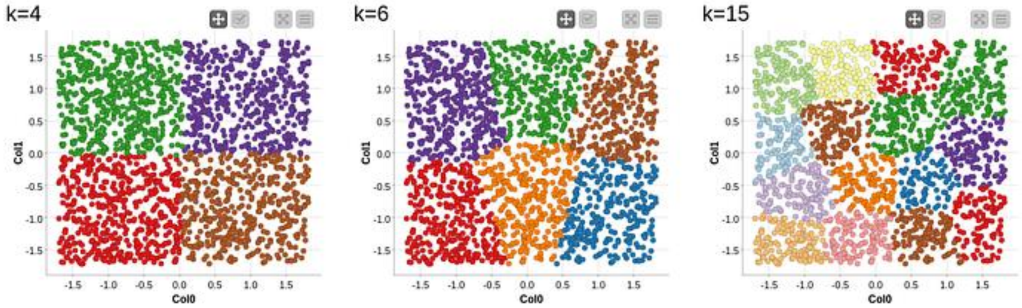


Figure 6: Uniformly distributed random data clustered into k=4 (left), 6 (center), and 15 (right) clusters.

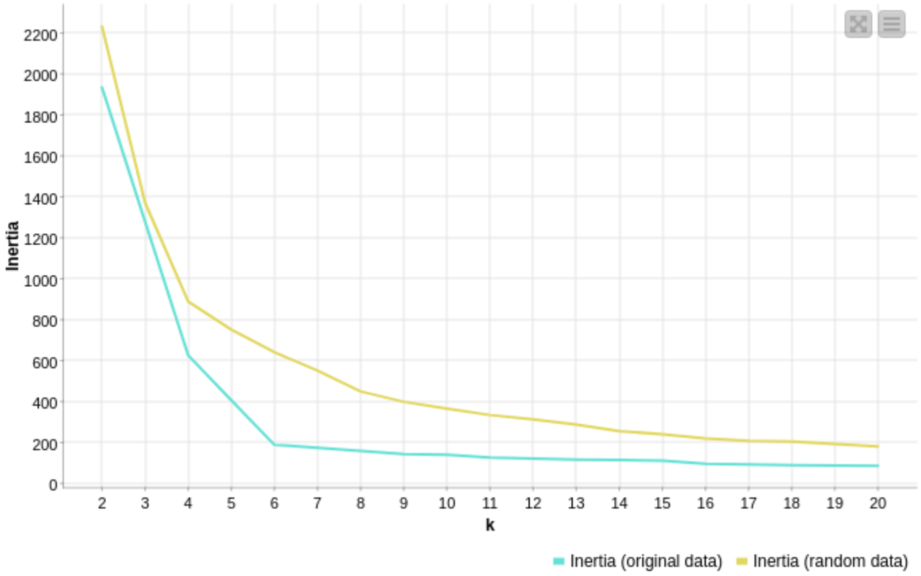


Figure 7: How the inertia decreases for the original data (from Figure 1) vs. the random data over a range of k.

In the actual calculation of a gap statistic, a number of random samples are generated, then clustered over a range of *k*, and the resulting inertia is recorded. This allows a number of inertiae for random cases. The original data set is also clustered over a range of *k*, resulting in a series of inertiae. The gap statistic, at *k* clusters, is calculated as

$$Gap(k) = \frac{1}{B}\sum_{i=1}^{B} log(W_k^{(i)}) - log(W_k)$$

Where Wk(i) is the inertia from the *i*-th random sample (i=1,2,…,B) with *k* clusters, and Wk is the inertia from the original data with *k* clusters. We also calculate its standard deviation as

$$s_k = \sqrt{1 + \frac{1}{B}}\sqrt{\frac{1}{B}\sum_{i=1}^{B}\left(log(W_k^{(i)}) - \overline{W}\right)^2}, \text{ where } \overline{W} = \frac{1}{B}\sum_{i=1}^{B} log(W_k^{(i)})$$

Then we find the optimal *K* as the smallest *k* that satisfies the condition

$$Gap(k) \geq Gap(k+1) - s_{k+1}$$

The calculation of a gap statistic involves a simulation. We call functions in R to calculate the gap statistic with some R scripting within a KNIME workflow. In particular, the clusGap() function is called to calculate the gap statistic at different *k*, and the maxSE() returns the optimal *K* satisfying the condition described above. Figure 8 shows the gap statistic plot of our example data set in Figure 1, based on *B*=100 iterations at each *k*. The red line represents the optimum *K* that satisfies the condition above.

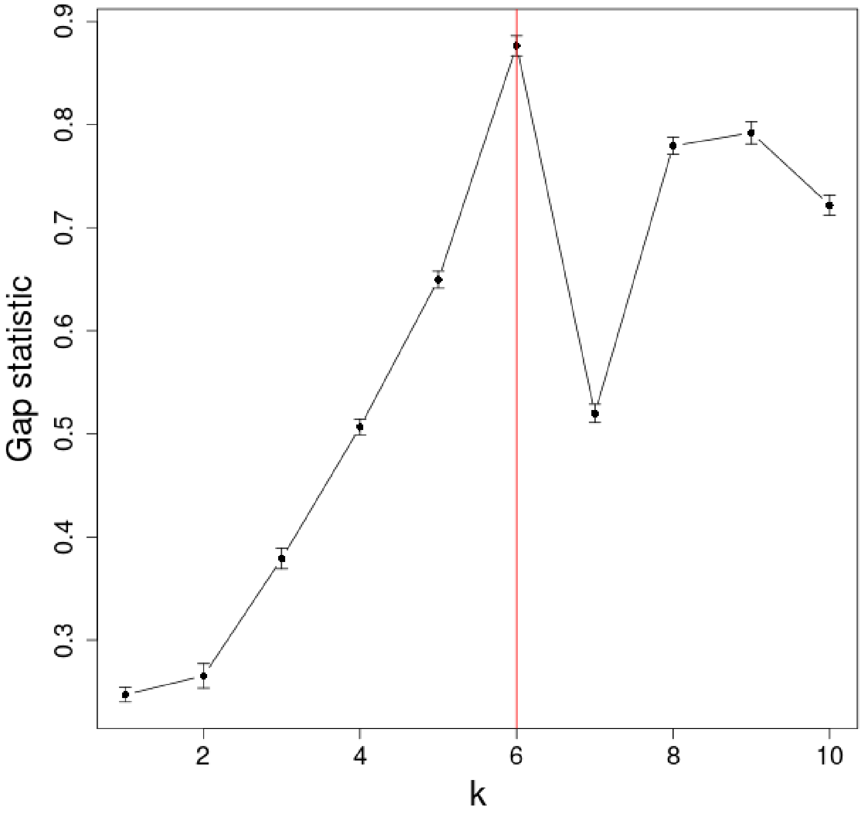


Figure 8: A plot of gap statistics as well as their standard deviations, based on B=100 iterations. The optimal k=6, satisfying the condition, is indicated by the red line. Image by author.

It should be noted that the optimal *K* determined by the gap statistic method may not be consistent. For example, when the gap statistic method is applied to our toy data many times, the resulting optimal *K* may be different (see Figure 9).

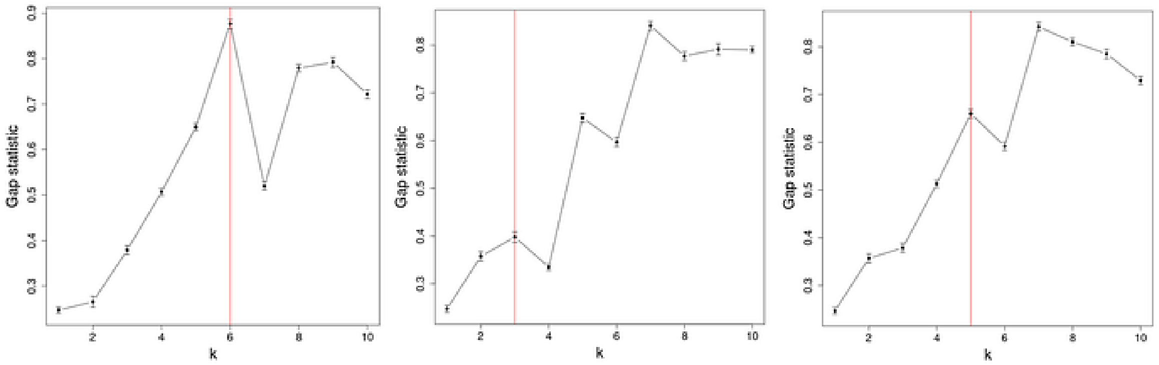


Figure 9: Examples of gap statistics plots. The optimum k may not be consistent, depending on the simulation outcome. Image by author.

14. How is cluster quality measured?

Measures for Quality of Clustering:

If all the data objects in the cluster are highly similar then the cluster has high quality. We can measure the quality of Clustering by using the Dissimilarity/Similarity metric in most situations. But there are some other methods to measure the Qualities of Good Clustering if the clusters are alike.

1. Dissimilarity/Similarity metric: The similarity between the clusters can be expressed in terms of a distance function, which is represented by d(i, j). Distance functions are different for various data types and data variables. Distance function measure is different for continuous-valued variables, categorical variables, and vector variables. Distance function can be expressed as Euclidean distance, Mahalanobis distance, and Cosine distance for different types of data.

2. Cluster completeness: Cluster completeness is the essential parameter for good clustering, if any two data objects are having similar characteristics then they are assigned to the same category of the cluster according to ground truth. Cluster completeness is high if the objects are of the same category.

Let us consider the clustering C1, which contains the sub-clusters s1 and s2, where the members of the s1 and s2 cluster belong to the same category according to ground truth. Let us consider another clustering C2 which is identical to C1 but now s1 and s2 are merged into one cluster. Then, we define the clustering quality measure, Q, and according to cluster completeness C2, will have more cluster quality compared to the C1 that is, Q(C2, Cg) > Q(C1, Cg).

3. Ragbag: In some situations, there can be a few categories in which the objects of those categories cannot be merged with other objects. Then the quality of those cluster categories is measured by the Rag Bag method. According to the rag bag method, we should put the heterogeneous object into a rag bag category.

Let us consider a clustering C1 and a cluster C ∊ C1 so that all objects in C belong to the same category of cluster C1 except the object o according to ground truth. Consider a clustering C2 which is identical to C1 except that o is assigned to a cluster D which holds the objects of different categories. According to the ground truth, this situation is noisy and the quality of clustering is measured using the rag bag criteria. we define the clustering quality measure, Q, and according to rag bag method criteria C2, will have more cluster quality compared to the C1 that is, Q(C2, Cg)>Q(C1, Cg).

4. Small cluster preservation: If a small category of clustering is further split into small pieces, then those small pieces of cluster become noise to the entire clustering and thus it becomes difficult to identify that small category from the clustering. The small cluster preservation criterion states that are splitting a small category into pieces is not advisable and it further decreases the quality of clusters as the pieces of clusters are distinctive. Suppose clustering C1 has split into three clusters, C11 = {d1, . . . , dn}, C12 = {dn+1}, and C13 = {dn+2}.

Let clustering C2 also split into three clusters, namely C1 = {d1, . . . , dn−1}, C2 = {dn}, and C3 = {dn+1,dn+2}. As C1 splits the small category of objects and C2 splits the big category which is preferred according to the rule mentioned above the clustering quality measure Q should give a higher score to C2, that is, Q(C2, Cg) > Q(C1, Cg).

15. What is cluster analysis and its types?

Cluster analysis, also known as clustering, is a method that groups similar data points together. The goal of cluster analysis is to divide a dataset into groups (or clusters) such that the data points within each group are more similar to each other than to data points in other groups. This process is often used for exploratory data analysis and can help identify patterns or relationships within the data that may not be immediately obvious. There are many different algorithms used for cluster analysis, such as k-means, hierarchical clustering, and density-based clustering. The choice of algorithm will depend on the specific requirements of the analysis and the nature of the data being analyzed.

Cluster Analysis is the process to find similar groups of objects in order to form clusters. It is an unsupervised machine learning-based algorithm that acts on unlabelled data. A group of data points would comprise together to form a cluster in which all the objects would belong to the same group.

The given data is divided into different groups by combining similar objects into a group. This group is nothing but a cluster. A cluster is nothing but a collection of similar data which is grouped together.

For example, consider a dataset of vehicles given in which it contains information about different vehicles like cars, buses, bicycles, etc. As it is unsupervised learning there are no class labels like Cars, Bikes, etc for all the vehicles, all the data is combined and is not in a structured manner.

Now our task is to convert the unlabelled data to labelled data and it can be done using clusters.

The main idea of cluster analysis is that it would arrange all the data points by forming clusters like cars cluster which contains all the cars, bikes clusters which contains all the bikes, etc.

Simply it is the partitioning of similar objects which are applied to unlabelled data.

Clustering Methods:
The clustering methods can be classified into the following categories:

- Partitioning Method
- Hierarchical Method
- Density-based Method
- Grid-Based Method
- Model-Based Method
- Constraint-based Method

Partitioning Method: It is used to make partitions on the data in order to form clusters. If "n" partitions are done on "p" objects of the database then each partition is represented by a cluster and n < p. The two conditions which need to be satisfied with this Partitioning Clustering Method are:

- One objective should only belong to only one group.
- There should be no group without even a single purpose.

In the partitioning method, there is one technique called iterative relocation, which means the object will be moved from one group to another to improve the partitioning

Hierarchical Method: In this method, a hierarchical decomposition of the given set of data objects is created. We can classify hierarchical methods and will be able to know the purpose of classification on the basis of how the hierarchical decomposition is formed. There are two types of approaches for the creation of hierarchical decomposition, they are:

- **Agglomerative Approach:** The agglomerative approach is also known as the bottom-up approach. Initially, the given data is divided into which objects form separate groups. Thereafter it keeps on merging the objects or the groups that are close to one another which means that they exhibit similar properties. This merging process continues until the termination condition holds.
- **Divisive Approach:** The divisive approach is also known as the top-down approach. In this approach, we would start with the data objects that are in the same cluster. The group of individual clusters is divided into small clusters by continuous iteration. The iteration continues until the condition of termination is met or until each cluster contains one object.

Once the group is split or merged then it can never be undone as it is a rigid method and is not so flexible. The two approaches which can be used to improve the Hierarchical Clustering Quality in Data Mining are: –

- One should carefully analyze the linkages of the object at every partitioning of hierarchical clustering.
- One can use a hierarchical agglomerative algorithm for the integration of hierarchical agglomeration. In this approach, first, the objects are grouped into micro-clusters. After grouping data objects into microclusters, macro clustering is performed on the microcluster.

Density-Based Method: The density-based method mainly focuses on density. In this method, the given cluster will keep on growing continuously as long as the density in the neighbourhood exceeds some threshold, i.e, for each data point within a given cluster. The radius of a given cluster has to contain at least a minimum number of points.

Grid-Based Method: In the Grid-Based method a grid is formed using the object together,i.e, the object space is quantized into a finite number of cells that form a grid structure. One of the major advantages of the grid-based method is fast processing time and it is dependent only on the number of cells in each dimension in the quantized space. The processing time for this method is much faster so it can save time.

Model-Based Method: In the model-based method, all the clusters are hypothesized in order to find the data which is best suited for the model. The clustering of the density function is used to locate the clusters for a given model. It reflects the spatial distribution of data points and also provides a way to automatically determine the number of clusters based on standard statistics, taking outlier or noise into account. Therefore it yields robust clustering methods.

Constraint-Based Method: The constraint-based clustering method is performed by the incorporation of application or user-oriented constraints. A constraint refers to the user expectation or the properties of the desired clustering results. Constraints provide us with an interactive way of communication with the clustering process. The user or the application requirement can specify constraints.