# 9 Cluster Analysis

Learning Objectives

* Understand Cluster Analysis and its applications
* Learn the types of clusters and how they are represented
* Learn how the clustering technique works in practice
* Understand the K-means technique and its pseudocode
* Appreciate the many advantages and disadvantages of ANNs

### INTRODUCTION

Cluster analysis is used for automatic identification of natural grouping of things. It is also known as the segmentation technique. In this technique, data instances that are similar to (or near) each other are categorized into one cluster. Similarly, data instances that are very different (or far away) from each other are moved into different clusters.

Clustering is an unsupervised learning technique as there is no output or de- pendent variable for which a right or wrong answer can be computed. The correct number of clusters or the definition of those clusters is not known ahead of time. Clustering techniques can only suggest to the user how many clusters would make sense from the characteristics of the data. The user can specify a different, larger or smaller, number of desired clusters based on their making business sense. The cluster analysis technique will then define many distinct clusters from analysis of the data, with cluster definitions for each of those clusters. However, there are good cluster definitions, depending on how closely the cluster parameters fit the data.

#### Caselet: Cluster Analysis

*A national insurance company distributes its personal and small commercial insurance products through independent agents. They wanted to increase their sales by better understanding their customers. They were interested in increasing their market share by doing some direct marketing campaigns, however without creating a channel conflict with the independent agents. They were also interested in examining different customer segments based on their needs, and the profitability of each of those segments.*

*They gathered attitudinal, behavioral, and demographic data using a mail survey of 2000 U.S. households that own auto insurance. Additional geo-demographic and credit information was added to the survey data. Cluster analysis of the data revealed five roughly equal segments.*

* Non-Traditional *interested in using the Internet and/or buying insurance at work.*
* Direct Buyers *interested in buying via direct mail or telephone.*
* Budget Conscious *interested in minimal coverage and finding the best deal.*
* Agent Loyals *expressed strong loyalty to their agents and high levels of personal service.*
* Hassle-Free *similar to Agent Loyals but less interested in face-to-face service.*

*(Source: *greenbook.org*)*

1. *Which customer segments would you choose for direct marketing? Will these create a channel conflict?*
2. *Could this segmentation apply to other service businesses? Which ones?*

### APPLICATIONS OF CLUSTER ANALYSIS

Cluster analysis is used in almost every field where there is a large variety of transactions. It helps provide characterization, definition, and labels for populations. It can help identify natural grouping of customers, products, patients, and so on. It can also help identify outliers in a specific domain and thus decrease the size and complexity of problems. A prominent business application of cluster analysis is in market research. Customers are segmented into clusters based on their characteristics—wants and needs, geography, price sensitivity, and so on. Here are some examples of clustering

*Market Segmentation* Categorizing customers according to their similarities, for instance by their common wants and needs, and propensity to pay can help with targeted marketing.

*Product Portfolio* People of similar sizes can be grouped together to make small, medium and large sizes for clothing items.

*Text Mining* Clustering can help organize a given collection of text documents according to their content similarities into clusters of related topics.

### DEFINITION OF A CLUSTER

An operational definition of a cluster is that, given a representation of *n* objects, find *K* groups based on a measure of similarity, such that objects within the same group are alike but the objects in different groups are not alike.

However, the notion of similarity can be interpreted in many ways. Clusters can differ in terms of their shape, size, and density. Clusters are patterns and there can be many kinds of patterns. Some clusters are the traditional types, such as data points hanging together. However, there are other clusters, such as all points representing the circumference of a circle. There may be concentric circles with points of different circles representing different clusters. The presence of noise in the data makes the detection of the clusters even more difficult.

An ideal cluster can be defined as a set of points that is compact and isolated. In reality, a cluster is a subjective entity whose significance and interpretation requires domain knowledge. In the sample data below (Figure 9.1), how many clusters can one visualize?

x

x x

x

x x x x

x

x x x

x

FIGURE 9.1 Visual Cluster Example

It seems like there are two clusters of approximately equal sizes. However, they can be seen as three clusters, depending on how we draw the dividing lines. There is not a truly optimal way to calculate it. Heuristics are often used to define the number of clusters.

### REPRESENTING CLUSTERS

The clusters can be represented by a central or modal value. A cluster can be defined as the *centroid* of the collection of points belonging to it. A *centroid* is a measure of central tendency. It is the point from where the sum total of squared distance from all the points is the minimum. A real-life equivalent would be the citycenter as the point that is considered the most easy to use by all constituents of the city. Thus, all cities are defined by their centers or downtown areas.

A cluster can also be represented by the most frequently occurring value in the cluster, i.e., a cluster can be defined by its modal value. Thus, a particular cluster representing a social point of view could be called the ‘soccer moms’, even though not all members of that cluster need currently be a mom with soccer-playing children.

### CLUSTERING TECHNIQUES

Cluster analysis is a machine-learning technique. The quality of a clustering result depends on the *algorithm*, the *distance* function, and the *application*. First, consider the distance function. Most cluster analysis methods use a distance measure to calculate the closeness between pairs of items. There are two major measures of distances – Euclidian distance (“as the crow flies” or straight line) is the most intuitive measure; the other popular measure is the Manhattan (rectilinear) distance, where one can go only in orthogonal directions. The Euclidian distance is the hypotenuse of a right triangle, while the Manhattan distance is the sum of the two legs of the right triangle. There are other measures of distance like Jacquard distance (to measure similarity of sets), or Edit distance (similarity of texts), and others.

In either case, the key objective of the clustering algorithm is the same, i.e., inter-cluster distance is maximized, and intra-clusters distance is minimized.

There are many algorithms to produce clusters. There are top-down, hierarchical methods that start with creating a given number of best-fitting clusters. There are also bottom-up methods that begin with identifying naturally occurring clusters.

The most popular clustering algorithm is the *K*-means algorithm. It is a top- down, statistical technique, based on the method of minimizing the least squared distance from the center points of the clusters. Other techniques, such as neural networks, are also used for clustering. Comparing cluster algorithms is a difficult task as there is no single right number of clusters. However, the speed of the algorithm and its versatility in terms of different dataset are important criteria.

Here is the generic pseudocode for clustering

1. Pick an arbitrary number of groups/segments to be created.
2. Start with some initial randomly chosen center values for groups.
3. Classify instances to closest groups.
4. Compute new values for the group centers.
5. Repeat steps 3 and 4 till groups converge.
6. If clusters are not satisfactory, go to step 1 and pick a different number of groups/segments.

The clustering exercise can be continued with a different number of clusters and different location of those points. Clusters are considered good if the cluster definitions stabilize, and the stabilized definitions prove useful for the purpose at hand. Else, repeat the clustering exercise with a different number of clusters and different starting points for group means.

### CLUSTERING EXERCISE

Here is a simple exercise to visually and intuitively identify clusters from the data as shown in Dataset 9.1. *X* and *Y* are the two dimensions of interest. The objective is to determine the number of clusters and the center points of those clusters.

|  |  |
| --- | --- |
| Dataset 9.1 |  |
| *X* | *Y* |
| 2 | 4 |
| 2 | 6 |
| 5 | 6 |
| 4 | 7 |
| 8 | 3 |
| 6 | 6 |
| 5 | 2 |
| 5 | 7 |
| 6 | 3 |
| 4 | 4 |

A scatter plot of 10 items in 2 dimensions shows them distributed fairly randomly. As a bottom-up technique, the number of clusters and their centroids can be intuited (Figure 9.2).



4,7

5,7

2,6

5,6

6,6

2,4

4,4

6,3

8,3

5,2

0 1 2 3 4 5 6 7 8 9

FIGURE 9.2 Initial Data Points and the Centroid (Shown as Thick Dot)

The points are distributed randomly enough such that it is considered as one cluster. The solid circle represents the central point (centroid) of these points.

However, there is a big distance between the points (2, 6) and (8, 3). So, this data can be broken into 2 clusters. The 3 points at the bottom right can form one cluster and the other 7 forms the other cluster. The two clusters look like as follows (Figure 9.3). The two circles are the new centroids.



4,7 5,7

2,6

5,6 6,6

2,4

4,4

6,3

8,3

5,2

0

1

2

3

4

5

6

7

8

9

FIGURE 9.3 Dividing into Two Clusters (Centroids Shown as Thick Dots)

The bigger cluster seems too far apart. So, it seems like the 4 points on the top form a separate cluster. The three clusters look like as follows (Figure 9.4).



4,7 5,7

2,6

5,6 6,6

2,4

4,4

6,3

8,3

5,2

0

1

2

3

4

5

6

7

8

9

FIGURE 9.4 Dividing into Three Clusters (Centroids Shown as Thick Dots)

This solution has 3 clusters. The cluster on the right is far from the other 2 clusters. However, its centroid is not too close to all the data points. The cluster at the top looks very tight-fitting, with a nice centroid. The third cluster, at the left, is spread out and may not be of much usefulness.

This was a bottom-up exercise in visually producing 3 best-fitting cluster definitions from the given data. The right number of clusters will depend on the data and the application for which the data would be used.

### *K*-MEANS ALGORITHM FOR CLUSTERING

*K*-means is the most popular clustering algorithm. It iteratively computes the clusters and their centroids. It is a top down approach to clustering. Starting with a given number of *K* clusters, say 3 clusters, that means 3 random centroids will be created as starting points of the centers. The circles are initial cluster centroids (Figure 9.5).



4,7

5,7

2,6

5,6

6,6

2,4

4,4

6,3

8,3

5,2

0 1 2 3 4 5 6 7 8 9

FIGURE 9.5 Randomly Assigning Three Centroids for Three Data Clusters

*Step 1* For a data point, distance values will be from each of the three centroids. The data point will be assigned to the cluster with the shortest distance to the centroid. All data points will thus, be assigned to one data point or the other (Figure 9.6). The arrows from each data element show the centroid that the point is assigned to.

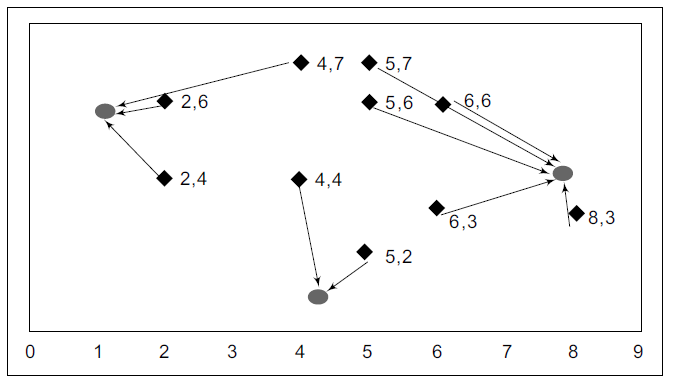


FIGURE 9.6 Assigning Data Points to Closest Centroid

*Step 2* The centroid for each cluster will now be recalculated such that it is closest to all the data points allocated to that cluster. The dashed arrows show the centroids being moved from their old (shaded) values to the revised new values (Figure 9.7).

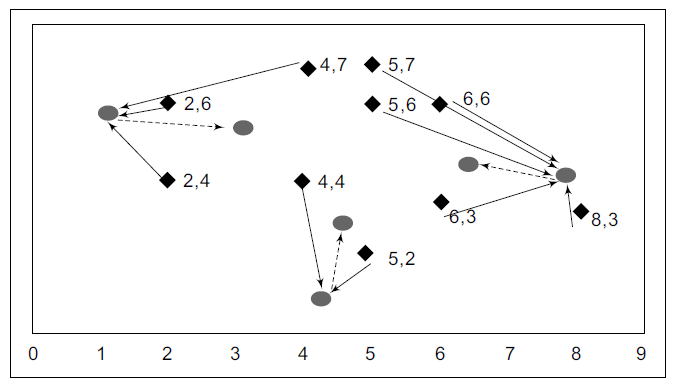


FIGURE 9.7 Recomputing Centroids for Each Cluster

*Step 3* Once again, data points are assigned to the three centroids closest to it (Figure 9.8).

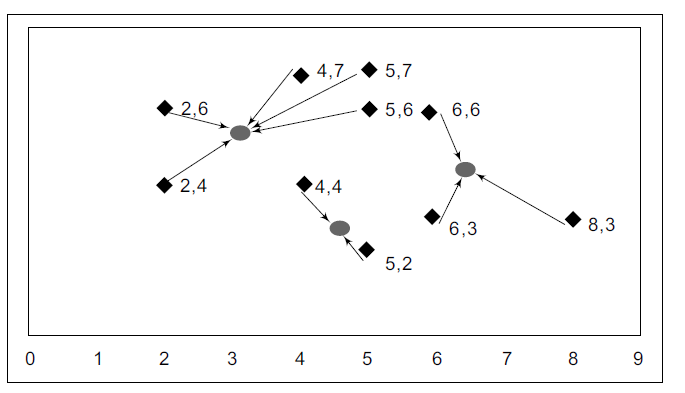


FIGURE 9.8 Assigning Data Points to Recomputed Centroids

The new centroids will be computed from the data points in the cluster until finally, the centroids stabilize in their locations. These are the three clusters computed by this algorithm.

The three clusters shown are – a 3-datapoints cluster with centroid (6.5, 4.5), a 2-datapoint cluster with centroid (4.5, 3) and a 5-datapoint cluster with centroid (3.5, 3) (Figure 9.9).

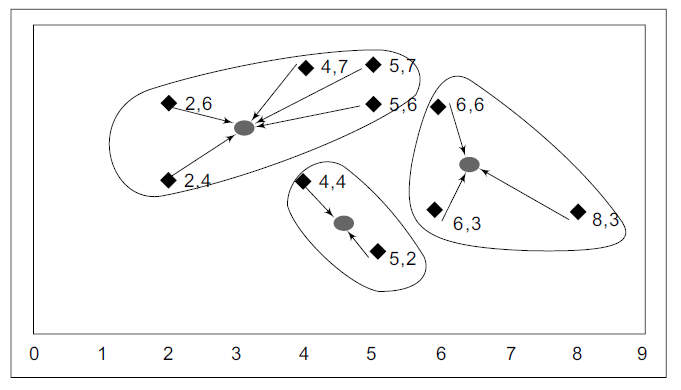


FIGURE 9.9 Recomputing Centroids for Each Cluster till Clusters Stabilize

These cluster definitions are different from the ones derived visually. This is a function of the random starting centroid values. The centroid points used earlier in the visual exercise were different from that chosen with the *K*-means clustering algorithm. The *K*-means clustering exercise should, therefore, be run again with this data, but with new random centroid starting values. With many runs, the cluster definitions are likely to stabilize. If the cluster definitions do not stabilize, that may be a sign that the number of clusters chosen is too high or too low. The algorithm should also be run with different values of *K*.

Here is the pseudocode for implementing a *K*-means algorithm.

Algorithm K-Means (K number of clusters, D list of data points)

1. Choose *K* number of random data points as initial centroids (cluster-centers)
2. Repeat till cluster-centers stabilize
   1. {Allocate each point in *D* to the nearest of *K* centroids;
   2. Compute centroid for the cluster using all points in the cluster}

### SELECTING THE NUMBER OF CLUSTERS

The correct choice of the value of *K* is often ambiguous. It depends on the shape and scale of the distribution points in a dataset and the desired clustering resolution of the user. Heuristics are needed to pick the right number. One can graph the percentage of variance explained by the clusters against the number of clusters (Fig. 9.10). The first cluster will add more information (explain a lot of variance), but at some point the marginal gain in variance will fall, giving a sharp angle to the graph, looking like an elbow. Beyond that elbow point, adding more clusters will not add much incremental value. That would be the desired *K*.

To engage with the data and to understand the clusters better, it is often better to start with a small number of clusters such as 2 or 3, depending upon the dataset and the application domain. The number can be increased subsequently, as needed from an application point of view. This helps understand the data and the clusters progressively better.

2.0



|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  | \* |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |

1.8

Average within-cluster of sum of squares

1.6

1.4

1.2

1.0

0.8

0.6

0.4

Elbow for *K*Means clustering

0.2

1 2 3 4 5 6 7 8 9

Number of clusters

FIGURE 9.10 Elbow Method for Determining Number of Clusters in a Dataset

### ADVANTAGES AND DISADVANTAGES OF *K*-MEANS ALGORITHM

There are many advantages of the *K*-means algorithm.

* *K*-means algorithm is simple, easy to understand and easy to implement.
* It is also efficient, in that, the time taken to cluster *K*-means rises linearly with the number of data points.
* No other clustering algorithm performs better than *K*-means, in general.

There are a few disadvantages too

* The user needs to specify an initial value of *K*.
* The process of finding the clusters may not converge.
* It is not suitable for discovering cluster shapes that are not hyperellipsoids (or hyperspheres).

Neural networks can also be deployed for clustering, using the appropriate objective function. The neural network will produce the appropriate cluster centroids and cluster population for each cluster.

## Conclusion

Cluster analysis is a useful, unsupervised learning technique that is used in many business situations to segment the data into meaningful small groups. *K*-means algorithm is an easy statistical technique to iteratively segment the data. How- ever, there is only a heuristic technique to select the right number of clusters.

## Questions

1. What is unsupervised learning? When is it used?
2. Describe three business applications in your industry where cluster analysis will be useful.
3. Data about height and weight for a few volunteers is available. Create a set of clusters for the following data shown in Dataset 9.2, to decide how many sizes of T-shirts should be ordered.

|  |  |
| --- | --- |
| Dataset 9.2 |  |
| Height | Weight |
| 71 | 165 |
| 68 | 165 |
| 72 | 180 |
| 67 | 113 |
| 72 | 178 |
| 62 | 101 |
| 70 | 150 |
| 69 | 172 |
| 72 | 185 |
| 63 | 149 |
| 69 | 132 |
| 61 | 115 |

## True/False

1. Cluster analysis is used for market segmentation.
2. A good set of clusters should be of significantly different sizes.
3. Usually the clusters can each be represented by a central point, or the modal point, also called the ‘centroid’.
4. A decision tree can help in determining the likely number of clusters in a given dataset.
5. The key objective of all clustering algorithms is the same, i.e., inter-cluster distance is maximized and intra-cluster distance is minimized.
6. There are multiple ways of measuring the distance between data points.
7. An unlimited number of variables can be used for cluster analysis.
8. *K*-means is the most popular clustering algorithm.
9. *K*-means automatically produces the right number of clusters in the data.
10. Amazon uses clustering technique to recommend new products for its customers to buy.