



# Introduction to Clustering

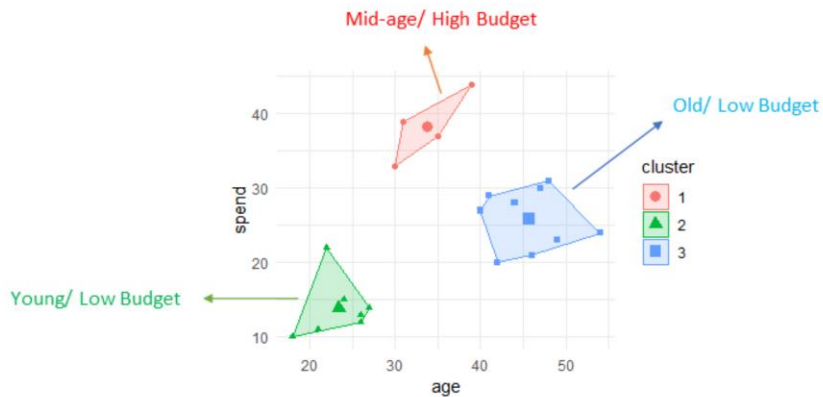
Here, we discuss a popular unsupervised learning method called clustering.

## Definitions

- Clustering: the process of grouping a set of objects into classes of similar objects. Ideally:
  - Objects within a cluster should be similar.
  - Objects from different clusters should be dissimilar.
- Clustering is the most common form of unsupervised learning where there is no predefined classes for a training data set
- Clustering can be used:
  - as a stand-alone tool to gain an insight into data distribution
  - as a preprocessing step of other algorithms in intelligent systems

The goal of clustering is to segment the data into a set of homogeneous clusters of records for the purpose of generating insight. Separating a dataset into clusters of homogeneous records is also useful for improving performance of supervised methods, by modeling each cluster separately rather than the entire, heterogeneous dataset. Clustering is used in a vast variety of business applications, from customized marketing to industry analysis.

## Example



This is an example of clustering. Note the separation of items based on Age and Spending. Any number of independent variables can be used, though one should ask “how do you form the clusters, and how many should there be?”

## Examples of Business Applications of Clustering

- **Customer Segmentation:** Clustering helps marketers work on target areas, and segment customers based on purchase history, interests, or activity monitoring.
- **Fraud Detection:** Utilizing past historical data on fraudulent claims, it is possible to isolate new claims based on its proximity to clusters that indicate fraudulent patterns.
- **Document Classification:** Cluster documents in multiple categories based on tags, topics, and the content of the document.

Cluster analysis is used to form groups or clusters of similar records based on several measurements made on these records. The key idea is to characterize the clusters in ways that would be useful for the aims of the analysis. This idea has been applied in many areas. Biologists, for example, have made extensive use of classes and subclasses to organize species. A spectacular success of the clustering idea in chemistry was Mendeleev's periodic table of the elements.

One popular use of cluster analysis in marketing is for market segmentation: customers are segmented based on demographic and transaction history information, and a marketing strategy is tailored for each segment. In countries such as India, where customer diversity is extremely location-sensitive, chain stores often perform market segmentation at the store level, rather than chain-wide (called "micro segmentation").

In finance, cluster analysis can be used for creating balanced portfolios: Given data on a variety of investment opportunities (e.g., stocks), one may find clusters based on financial performance variables such as return (daily, weekly, or monthly), volatility, beta, and other characteristics, such as industry and market capitalization.

## Clustering Considerations

- Similar to K-NN algorithm, we use different measures of distance to quantify the similarity/dis-similarity of different observations.
- Same considerations applies here:
  - The scale of variables are important. Using variables with different scales results in misleading calculations of distances
  - Choosing relevant variables for applying clustering algorithm is also important. Non-related variables should be excluded.
  - Finally, the choice of right distance metric is also important.

Cluster analysis can be thought of as an algorithm that measures the distance between records, and according to these distances, forms clusters. This is similar to the k-NN algorithms, which also use distance measures to identify neighbors.

As distance values, especially for numeric variables, are affected by scale, it is important that we *normalize* values before computing distances. As with other methods, only relevant variables should be included in your analysis.

## Number of Clusters

- Defining the right number of clusters is of great importance because :
  - Too few clusters does not allow for proper separation of observations.
  - Too many clusters, on the other hand, result in less meaningful clusters
- Some clustering algorithms require the number of clusters to be set in advance, while there are others which do not require the number of clusters as their input.
- In addition, there exists a number of methods that can help identifying an appropriate number of clusters.

In general, there are two types of clustering algorithms for a dataset of  $n$  records: hierarchical and non-hierarchical clustering.

Hierarchical methods can be either agglomerative or divisive. Agglomerative methods begin with  $n$  clusters and sequentially merge similar clusters until a single cluster is obtained. Divisive methods work in the opposite direction, starting with one cluster that includes all records. Hierarchical methods are especially useful when the goal is to arrange the clusters into a natural hierarchy.

Non-hierarchical methods, such as k-means, use a prespecified number of clusters. These methods are generally less computationally intensive and are therefore preferred with very large datasets. As we primarily study k-means in this module, one issue is in specifying the number of clusters beforehand. We will see approaches to address this issue.

“The idea of dividing a market up into homogeneous segments and targeting each with a distinct product and/or message, is now at the heart of marketing theory”

- Market Segmentation, Michael J Croft

WWW.KENT.EDU

And, those of you paying attention will realize that marketing is now at the individual level, or a cluster of size 1.