CHAPTER 4

Introduction to Cluster Analysis

OBJECTIVE

This chapter offers a comprehensive introduction to cluster analysis and its applications. You will explore the fundamentals of a specific clustering algorithm known as the k-means method. Additionally, you'll be introduced to an Excel workbook and template, which will be used in Chapter 5 to guide you through both the manual and automatic implementation of the k-means procedure.

Contents

4.1 Introduction

What Is Cluster Analysis? The dictionary defines a "cluster" as a collection or group of items with similar or different characteristics. In that sense, Clustering or Cluster analysis can be seen as a data analysis technique that creates clusters of objects that are closely related with respect to a given set of characteristics. Items that belong to the same cluster are located in "close proximity[1]" to each other with respect to a number of characteristics. Although cluster analysis is an old statistical that have been used for several decades, it has been rejuvenated in the field of computer science where clustering is defined as a process that organizes items into groups using unsupervised machine learning algorithms.

Cluster analysis is a useful, powerful and yet straightforward method for understanding data patterns. The main goal of clustering is to identify the clusters and categorize all items accordingly. Cluster analysis has been successfully used for identifying anomalies or outliers. These are cases that stand out from the rest of the data. For example, banks frequently use anomaly detection to fight fraud.

Let me review a few data analysis problems and how they can be addressed using cluster analysis. This will help clarify what your expectations should be when using this technique.

[1] The notion of close proximity will need to be defined more rigorously later in this chapter.

4.1.1 *When to Use Cluster Analysis?*

Cluster analysis is implemented using a specific clustering method that eventually translates to a formal computer algorithm. In this book I have confined myself to k-Means clustering, which is only one of many clustering methods that have been proposed in the literature. Although I will briefly review a few clustering methods, only k-Means will be treated in details. There is a wide variety of strange ways that nature can cluster data points, and neither k-Means nor any one other clustering method can successfully identify all these clusters. Consequently, cluster analysis should always be presented with the clustering method that was used.

Figure 4.1 is a good example of the way many data points are clustered in practice. That is, they are nicely distributed around a centered point (see the red diamond shape in the middle of cluster 1, and the blue diamond shape in the middle of cluster 2). It is how far a data point strays away from the cluster center that determines, which cluster it belongs.

In some weird situations that I will briefly discussed later in this chapter, nature can cluster your data points following a complicated geometric shape. In this case, an ordinary clustering technique such as the k-Means may produce unsatisfactory results. I recommend conducting a PCA before clustering. This allows you to create a scatterplot similar to that of Figure 4.1 using the first 2 composite score variables. Such a graph will give you a glimpse

into the types of clusters that you are dealing with. You will then be in a position where you can decide whether k-Means clustering will resolve your problem or not.

> *I recommend performing Principal Component Analysis before any other multivariate data analysis can be considered. The first 2 principal components will generally give a good two-dimensional snapshot of the dataset you want to analyze.*

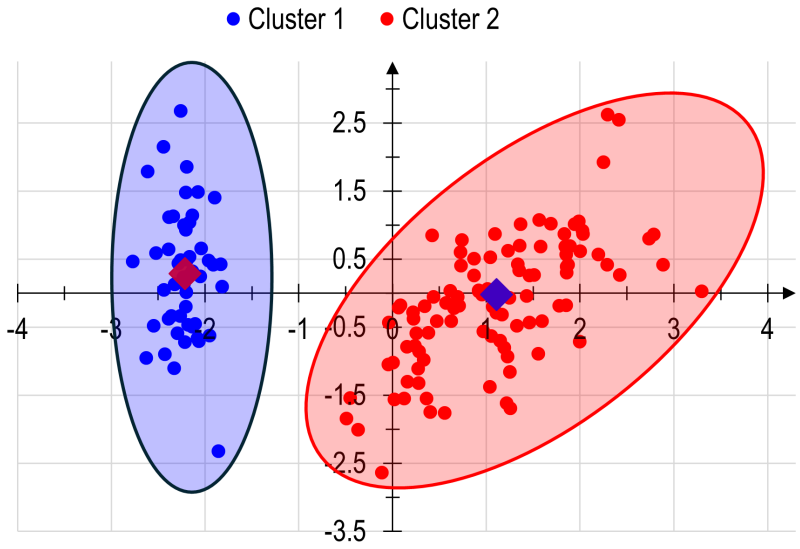

Figure 4.1: Scatterplot of the *Iris* dataset based on the first 2 composite score variables

Different Data Types and How to Handle Them

Cluster analysis aims at dividing your dataset into homogeneous groups. However, the methods used for quantifying the degree of homogeneity among items require all item characteristics

to be of quantitative type. Many clustering methods, which are implemented in the form of machine learning algorithms cannot handle categorical data, unless you find a way to recode them as numerical. The way to recode them varies considerably depending on whether the categorical variable is nominal[2] (e.g. *Gender="Male","Female"*) or ordinal[3] (e.g. *Age Group="<5", "5-17", "18+"*). But can you still use categorical in clustering? The answer is yes, you can.

If you want to use *Gender* for example in the clustering algorithm, one option is substitute it with two binary variables named *Male* and *Female*, which take the 2 values 0 and 1, indicating the absence or the presence of the characteristic. For ordinal data such as *Age Group*, each level of the variable will take a value such as $0, 1, 2, \cdots$ until the number of categories.

4.1.2 *Applications of Cluster Analysis*

Cluster analysis has applications in various industries and fields. Here is a short list of some disciplines that make use of this methodology.

1. *Marketing*:
   Cluster analysis is popular in marketing, especially in customer segmentation. This method of analysis helps to both target customer segments and perform sales analysis by groups.

[2]Nominal data cannot be ranked in any logical way.
[3]Ordinal data generally have a natural order of magnitude.

2. *Business Operations*:
   Businesses can optimize their processes and reduce costs by analyzing clusters and identifying similarities and differences between data points. For example, you can identify patterns in customer data and improve customer support processes for a particular group that may require special attention.

3. *Earth Observation*:
   Using a clustering algorithm, you can create a pixel mask for objects in an image. For example, you can use image segmentation to classify vegetation or built-up areas in a satellite image.

4. *Data Science*:
   You can use cluster analysis for predictive analytics. By applying machine learning techniques to clusters, you can create predictive models to make inferences about a particular data set.

5. *Outlier Detection*:
   When you need to detect outliers in your data, cluster analysis provides an effective method compared to traditional outlier detection methods based on the use of standard deviation.

   Cluster analysis can help you detect anomalies. While outliers are observations distant from the mean, they don't necessarily represent abnormalities. On the other hand, anomalies relate to identifying rare events or observations that deviate greatly from the mean.

6. *Text Clustering:*
   Text clustering is the process of grouping similar documents or pieces of text into clusters or categories. In the field of Natural Language Processing (NLP), text clustering is a fundamental and versatile technique that plays a pivotal role in various applications such as information retrieval, recommendation systems, content organization, and sentiment analysis.

4.1.3 *Types of Clustering Methods*

In this section, I want to provide a short description of a few clustering methods other than the k-Means. For a more comprehensive list of clustering methods, see Xu and Tian (2015). You may also obtain additional information from the link: https://developers.google.com/machine-learning/clustering/clustering-algorithms.

1. *Centroid-Based Clustering*
   This type of clustering forms groups around a central point, which may or may not be part of the dataset. In centroid-based clustering, a common algorithm is k-Means, which partitions the dataset into k clusters. Each data point is assigned to the cluster with the nearest centroid, calculated as the mean of the points within the cluster.

2. *Density-Based Clustering:*
   Density-based clustering focuses on the concentration of data points within a region. Clusters are formed based on a threshold, which defines the minimum number of points required

within a specified radius to form a dense region. This method is particularly effective at detecting noise and separating it from meaningful clusters. The most widely used algorithm for this approach is DBSCAN (Density-Based Spatial Clustering of Applications with Noise).

3. *Hierarchical Clustering:*
Hierarchical clustering organizes data into a tree-like structure of nested clusters, making it especially useful for hierarchical datasets, such as taxonomies. A notable example is the study "*Comparison of 61 Sequenced Escherichia coli Genomes*" by Lukjancenko et al. (2010). One key advantage of hierarchical clustering is the flexibility it provides, allowing users to select any number of clusters by cutting the tree at the desired level.

Centroid-based clustering and density-based clustering are two of the most commonly used clustering techniques. For a deeper understanding of these methods, you can watch an informative YouTube video from the following link: https://www.youtube.com/watch?v=Se28XHI2_xE.

4.2 Clustering and the Notion of Proximity

In this section, I present a high-level overview of the clustering procedure. A more detailed explanation will follow in Chapter 5.

Consider the small dataset shown in Figure 4.2, which consists of 15 observations extracted from a larger dataset. The full dataset, described and available for download at the following