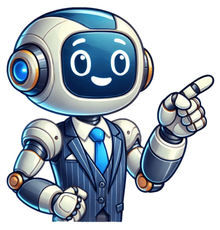
Click to verify



Cluster analysis is a statistical technique widely used in research and practical applications to group objects, data points, or cases into clusters based on their similarities. It is a cornerstone in data mining, pattern recognition, and machine learning, providing insights into the underlying structure of data. This article delves into the concept of cluster analysis, its types, methods, and practical examples. Cluster analysis is the process of organizing a set of objects into groups (clusters) such that objects within a cluster are more similar to each other than to those in other clusters. Similarity is typically measured using metrics like distance (e.g., Euclidean distance) or correlation. It is an unsupervised learning method, meaning it does not rely on labeled data and instead seeks to uncover hidden patterns. Cluster analysis is applied across various fields, including marketing (for customer segmentation), biology (for classifying species), and social sciences (for identifying behavioral patterns). Clustering techniques can be broadly categorized into the following types: In hard clustering, each data point belongs exclusively to one cluster. This approach is rigid and works best when clusters are well-defined and distinct. Example: K-means clustering, where each point is assigned to the nearest cluster centroid. In soft clustering (or fuzzy clustering), data points can belong to multiple clusters with varying degrees of membership. This method is useful when clusters overlap. Example: Fuzzy C-means clustering assigns a membership probability to each cluster for every data point. Hierarchical clustering builds a tree-like structure (dendrogram) that represents data groupings at different levels. It can be further divided into: Agglomerative Clustering: Starts with each object as a separate cluster and merges them iteratively. Divisive Clustering: Starts with a single cluster containing all objects and divides them iteratively. This type of clustering identifies dense regions of data points separated by sparser regions. It is effective for clusters with arbitrary shapes and varying densities. Example: DBSCAN (Density-Based Spatial Clustering of Applications with Noise) and OPTICS (Ordering Points to Identify the Clustering Structure). Model-based clustering assumes that data is generated from a mixture of probability distributions and uses statistical models to find clusters. Example: Gaussian Mixture Models (GMMs) use normal distributions to identify clusters. The methods used in cluster analysis depend on the type of clustering and the nature of the data. Below are the commonly used methods: How it works: Partitions data into a predefined number of clusters (k). Each cluster is represented by its centroid, and data points are assigned based on proximity to these centroids. Advantages: Fast and efficient for large datasets. Limitations: Assumes clusters are spherical and of equal size. How it works: Forms a hierarchy of clusters either by merging or splitting clusters iteratively. Advantages: Does not require specifying the number of clusters in advance. Limitations: Computationally expensive for large datasets. How it works: Groups points that are closely packed together and marks points in low-density regions as noise. Advantages: Effective for clusters of arbitrary shapes. Limitations: Struggles with clusters of varying density and high-dimensional data. How it works: Fits the data to a mixture of Gaussian distributions and assigns probabilities for cluster membership. Advantages: Handles overlapping clusters well. Limitations: Requires specifying the number of distributions. How it works: Allows data points to belong to multiple clusters with different degrees of membership, based on similarity. Advantages: Suitable for datasets with overlapping clusters. Limitations: Computationally intensive. Cluster analysis finds applications in numerous fields. Here are some notable examples: Objective: Customer segmentation for personalized marketing strategies. Example: Using K-means clustering to categorize customers into groups based on purchasing behavior, such as frequent buyers, occasional buyers, and potential customers. Objective: Identifying patient subgroups for targeted treatment plans. Example: Applying hierarchical clustering to group patients based on symptoms, medical history, or genetic data. Objective: Classifying species or genes based on similarities. Example: Using hierarchical clustering to construct phylogenetic trees that represent evolutionary relationships. Objective: Understanding group behaviors and societal trends. Example: Using DBSCAN to identify distinct social groups in a population based on demographic and survey data. Objective: Segmenting images into regions with similar characteristics. Example: Applying Gaussian Mixture Models to separate objects from the background in digital images. Provides insights into data structure without requiring labels. Works across various domains and data types. Enables data compression by summarizing large datasets into clusters. Results can be sensitive to the choice of parameters (e.g., the number of clusters). Performance can degrade with high-dimensional data. Some methods assume specific cluster shapes, limiting their applicability. Cluster analysis is a powerful and versatile tool for uncovering patterns and relationships in data. Its diverse types and methods make it suitable for a wide range of applications, from customer segmentation to scientific research. While it offers significant benefits, careful consideration of the method and parameters is crucial to ensure meaningful and accurate results. Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. ACM Computing Surveys, 31(3), 264-323. Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of KDD, 96(34), 226-231. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer. Kaufman, L., & Rousseeuw, P. J. (2005). Finding Groups in Data: An Introduction to Cluster Analysis. Wiley. Rokach, L., & Maimon, O. (2005). Clustering methods. In Data Mining and Knowledge Discovery Handbook (pp. 321-352). Springer. Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. IEEE Transactions on Neural Networks, 16(3), 645-678. The first thing to note about cluster analysis is that is is more useful for generating hypotheses than confirming them. Unlike the vast majority of statistical procedures, cluster analyses do not even provide p-values. In fact, while there is some unwillingness to say quite what cluster analysis does do, the general idea is to take observations and break them into groups. While there is a somewhat infinite number of methods to do this, there are three main bodies of methods, for two of which Stata has built-in commands. The first of these methods are partitioning methods, the second are agglomerative and the third are divisive. Partitioning methods assign each observation to the group with the nearest value (often mean or median). Agglomerative methods begin with each observation in its own group, then puts the two closest values together creating one group of two observations (all the rest of the groups remain single), then putting the next two closest values together so there are two groups of two (and all the other single groups) and continuing the process until the desired number of clusters is reached. The third method is something like a reverse of the agglomerative process, starting with one group containing all observations and working until each group contains a single observation. Divisive methods are very uncommon in the literature due to their time consuming nature and as a result Stata has no command for performing them. Once you have created a cluster, you can add notes to it using cluster note [cluster name] : [note content] and list the notes attached to your clusters via cluster notes without any arguments. You can also generate new grouping variables based on your clusters using the cluster generate [new variable name] command after a cluster command. For more on this ability see help cluster generate or Stata's Multivariate Statistics [MV] cluster generate entry. Note: Cluster notes are considered part of your data and will not be saved unless you save them. Back to Advanced Methods Data mining is the process of finding patterns, relationships and trends to gain useful insights from large datasets. It includes techniques like classification, regression, association rule mining and clustering. In this article, we will learn about clustering analysis in data mining.Understanding Cluster AnalysisCluster analysis is also known as clustering, which groups similar data points forming clusters. The goal is to ensure that data points within a cluster are more similar to each other than to those in other clusters. For example, in e-commerce retailers use clustering to group customers based on their purchasing habits. If one group frequently buys fitness gear while another prefers electronics. This helps companies to give personalized recommendations and improve customer experience. It is useful for:Scalability: It can efficiently handle large volumes of data.High Dimensionality: Can handle high-dimensional data.Adaptability to Different Data Types: It can work with numerical data like age, salary and categorical data like gender, occupation.Handling Noisy and Missing Data: Usually, datasets contain missing values or inconsistencies and clustering can manage them easily.Interpretability: Output of clustering is easy to understand and apply in real-world scenarios.Distance MetricsDistance metrics are simple mathematical formulas to figure out how similar or different two data points are. Type of distance metrics we choose plays a big role in deciding clustering results. Some of the common metrics are:Euclidean Distance: It is the most widely used distance metric and finds the straight-line distance between two points.Manhattan Distance: It measures the distance between two points based on grid-like path. It adds the absolute differences between the values.Cosine Similarity: This method checks the angle between two points instead of looking at the distance. It’s used in text data to see how similar two documents are.Jaccard Index: A statistical tool used for comparing the similarity of sample sets. It’s mostly used for yes/no type data or categories.Types of Clustering TechniquesClustering can be broadly classified into several methods. The choice of method depends on the type of data and the problem you're solving.1. Partitioning MethodsPartitioning Methods divide the data into k groups (clusters) where each data point belongs to only one group. These methods are used when you already know how many clusters you want to create. A common example is K-means clustering.In K-means the algorithm assigns each data point to the nearest center and then updates the center based on the average of all points in that group. This process repeats until the centres stop changing. It is used in real-life applications like streaming platforms like Spotify to group users based on their listening habits.2. Hierarchical MethodsHierarchical clustering builds a tree-like structure of clusters known as a dendrogram that represents the merging or splitting of clusters. It can be divided into:Agglomerative Approach (Bottom-up): Agglomerative Approach starts with individual points and merges similar ones. Like a family tree where relatives are grouped step by step.Divisive Approach (Top-down): It starts with one big cluster and splits it repeatedly into smaller clusters. For example, classifying animals into broad categories like mammals, reptiles, etc and further refining them.3. Density-Based MethodsDensity-based clustering group data points that are densely packed together and treat regions with fewer data points as noise or outliers. This method is particularly useful when clusters are irregular in shape.For example, it can be used in fraud detection as it identifies unusual patterns of activity by grouping similar behaviors together.4. Grid-Based MethodsGrid-Based Methods divide data space into grids making clustering efficient. This makes the clustering process faster because it reduces the complexity by limiting the number of calculations needed and is useful for large datasets.Climate researchers often use grid-based methods to analyze temperature variations across different geographical regions. By dividing the area into grids they can more easily identify temperature patterns and trends.5. Model-Based MethodsModel-based clustering groups data by assuming it comes from a mix of distributions. Gaussian Mixture Models (GMM) are commonly used and assume the data is formed by several overlapping normal distributions.GMM is commonly used in voice recognition systems as it helps to distinguish different speakers by modeling each speaker’s voice as a Gaussian distribution.6. Constraint-Based MethodsIt uses User-defined constraints to guide the clustering process. These constraints may specify certain relationships between data points such as which points should or should not be in the same cluster. In healthcare, clustering patient data might take into account both genetic factors and lifestyle choices. Constraints specify that patients with similar genetic backgrounds should be grouped together while also considering their lifestyle choices to refine the clusters.Impact of Data on Clustering TechniquesClustering techniques must be adapted based on the type of data:1. Numerical DataNumerical data consists of measurable quantities like age, income or temperature. Algorithms like k-means and DBSCAN work well with numerical data because they depend on distance metrics. For example a fitness app cluster users based on their average daily step count and heart rate to identify different fitness levels.2. Categorical DataIt contain non-numerical values like gender, product categories or answers to survey questions. Algorithms like k-modes or hierarchical clustering are better for this. For example grouping customers based on preferred shopping categories like "electronics" "fashion" and "home appliances."3. Mixed DataSome datasets contain both numerical and categorical features that require hybrid approaches. For example, clustering a customer database based on income (numerical) and shopping preferences (categorical) can use k-prototype method.Applications of Cluster AnalysisMarket Segmentation: This is used to segment customers based on purchasing behavior and allow businesses send the right offers to the right people.Image Segmentation: In computer vision it can be used to group pixels in an image to detect objects like faces, cars or animals.Biological Classification: Scientists use clustering to group genes with similar behaviors to understand diseases and treatments.Document Classification: It is used by search engines to categorize web pages for better search results.Anomaly Detection: Cluster Analysis is used for outlier detection to identify rare data points that do not belong to any cluster.Challenges in Cluster AnalysisWhile clustering is very useful for analysis it faces several challenges:Choosing the Number of Clusters: Methods like K-means requires user to specify the number of clusters before starting which can be difficult to guess correctly.Scalability: Some algorithms like hierarchical clustering does not scale well with large datasets.Cluster Shape: Many algorithms assume clusters are round or evenly shaped which doesn’t always match real-world data.Handling Noise and Outliers: They are sensitive to noise and outliers which can affect the results.Cluster analysis is like organising a messy room—sorting items into meaningful groups making everything easier to understand. Choosing the right clustering method depends on the dataset and goal of analysis. Data mining is the process of finding patterns, relationships and trends to gain useful insights from large datasets. It includes techniques like classification, regression, association rule mining and clustering. In this article, we will learn about clustering analysis in data mining.Understanding Cluster AnalysisCluster analysis is also known as clustering, which groups similar data points forming clusters. The goal is to ensure that data points within a cluster are more similar to each other than to those in other clusters. For example, in e-commerce retailers use clustering to group customers based on their purchasing habits. If one group frequently buys fitness gear while another prefers electronics. This helps companies to give personalized recommendations and improve customer experience. It is useful for:Scalability: It can efficiently handle large volumes of data.High Dimensionality: Can handle high-dimensional data.Adaptability to Different Data Types: It can work with numerical data like age, salary and categorical data like gender, occupation.Handling Noisy and Missing Data: Usually, datasets contain missing values or inconsistencies and clustering can manage them easily.Interpretability: Output of clustering is easy to understand and apply in real-world scenarios.Distance MetricsDistance metrics are simple mathematical formulas to figure out how similar or different two data points are. Type of distance metrics we choose plays a big role in deciding clustering results. Some of the common metrics are:Euclidean Distance: It is the most widely used distance metric and finds the straight-line distance between two points.Manhattan Distance: It measures the distance between two points based on grid-like path. It adds the absolute differences between the values.Cosine Similarity: This method checks the angle between two points instead of looking at the distance. It’s used in text data to see how similar two documents are.Jaccard Index: A statistical tool used for comparing the similarity of sample sets. It’s mostly used for yes/no type data or categories.Types of Clustering TechniquesClustering can be broadly classified into several methods. The choice of method depends on the type of data and the problem you're solving.1. Partitioning MethodsPartitioning Methods divide the data into k groups (clusters) where each data point belongs to only one group. These methods are used when you already know how many clusters you want to create. A common example is K-means clustering.In K-means the algorithm assigns each data point to the nearest center and then updates the center based on the average of all points in that group. This process repeats until the centres stop changing. It is used in real-life applications like streaming platforms like Spotify to group users based on their listening habits.2. Hierarchical MethodsHierarchical clustering builds a tree-like structure of clusters known as a dendrogram that represents the merging or splitting of clusters. It can be divided into:Agglomerative Approach (Bottom-up): Agglomerative Approach starts with individual points and merges similar ones. Like a family tree where relatives are grouped step by step.Divisive Approach (Top-down): It starts with one big cluster and splits it repeatedly into smaller clusters. For example, classifying animals into broad categories like mammals, reptiles, etc and further refining them.3. Density-Based MethodsDensity-based clustering group data points that are densely packed together and treat regions with fewer data points as noise or outliers. This method is particularly useful when clusters are irregular in shape.For example, it can be used in fraud detection as it identifies unusual patterns of activity by grouping similar behaviors together.4. Grid-Based MethodsGrid-Based Methods divide data space into grids making clustering efficient. This makes the clustering process faster because it reduces the complexity by limiting the number of calculations needed and is useful for large datasets.Climate researchers often use grid-based methods to analyze temperature variations across different geographical regions. By dividing the area into grids they can more easily identify temperature patterns and trends.5. Model-Based MethodsModel-based clustering groups data by assuming it comes from a mix of distributions. Gaussian Mixture Models (GMM) are commonly used and assume the data is formed by several overlapping normal distributions.GMM is commonly used in voice recognition systems as it helps to distinguish different speakers by modeling each speaker’s voice as a Gaussian distribution.6. Constraint-Based MethodsIt uses User-defined constraints to guide the clustering process. These constraints may specify certain relationships between data points such as which points should or should not be in the same cluster. In healthcare, clustering patient data might take into account both genetic factors and lifestyle choices. Constraints specify that patients with similar genetic backgrounds should be grouped together while also considering their lifestyle choices to refine the clusters.Impact of Data on Clustering TechniquesClustering techniques must be adapted based on the type of data:1. Numerical DataNumerical data consists of measurable quantities like age, income or temperature. Algorithms like k-means and DBSCAN work well with numerical data because they depend on distance metrics. For example a fitness app cluster users based on their average daily step count and heart rate to identify different fitness levels.2. Categorical DataIt contain non-numerical values like gender, product categories or answers to survey questions. Algorithms like k-modes or hierarchical clustering are better for this. For example grouping customers based on preferred shopping categories like "electronics" "fashion" and "home appliances."3. Mixed DataSome datasets contain both numerical and categorical features that require hybrid approaches. For example, clustering a customer database based on income (numerical) and shopping preferences (categorical) can use k-prototype method.Applications of Cluster AnalysisMarket Segmentation: This is used to segment customers based on purchasing behavior and allow businesses send the right offers to the right people.Image Segmentation: In computer vision it can be used to group pixels in an image to detect objects like faces, cars or animals.Biological Classification: Scientists use clustering to group genes with similar behaviors to understand diseases and treatments.Document Classification: It is used by search engines to categorize web pages for better search results.Anomaly Detection: Cluster Analysis is used for outlier detection to identify rare data points that do not belong to any cluster.Challenges in Cluster AnalysisWhile clustering is very useful for analysis it faces several challenges:Choosing the Number of Clusters: Methods like K-means requires user to specify the number of clusters before starting which can be difficult to guess correctly.Scalability: Some algorithms like hierarchical clustering does not scale well with large datasets.Cluster Shape: Many algorithms assume clusters are round or evenly shaped which doesn’t always match real-world data.Handling Noise and Outliers: They are sensitive to noise and outliers which can affect the results.Cluster analysis is like organising a messy room—sorting items into meaningful groups making everything easier to understand. Choosing the right clustering method depends on the dataset and goal of analysis. Cluster analysis is a technique used in machine learning that attempts to find clusters of observations within a dataset. The goal of cluster analysis is to find clusters such that the observations within each cluster are quite similar to each other, while observations in different clusters are quite different from each other. The following examples show how cluster analysis is used in various real-life situations. Example 1: Retail Marketing Retail companies often use clustering to identify groups of households that are similar to each other. For example, a retail company may collect the following information on households: Household income Household size Head of household Occupation Distance from nearest urban area They can then feed these variables into a clustering algorithm to perhaps identify the following clusters: Cluster 1: Small family, high spenders Cluster 2: Larger family, high spenders Cluster 3: Small family, low spenders Cluster 4: Large family, low spenders The company can then send personalized advertisements or sales letters to each household based on how likely they are to respond to specific types of advertisements. Example 2: Streaming Services Streaming services often use clustering analysis to identify viewers who have similar behavior. For example, a streaming service may collect the following data about individuals: Minutes watched per day Total viewing sessions per week Number of unique shows viewed per month Using these metrics, a streaming service can perform cluster analysis to identify high usage and low usage users so that they can know who they should spend most of their advertising dollars on. Example 3: Sports Science Data scientists for sports teams often use clustering to identify players that are similar to each other. For example, professional basketball teams may collect the following information about players: Points per game Rebounds per game Assists per game Steals per game They can then feed these variables into a clustering algorithm to identify players that are similar to each other so that they can have these players practice with each other and perform specific drills based on their strengths and weaknesses. Example 4: Email Marketing Many businesses use cluster analysis to identify consumers who are similar to each other so they can tailor their emails sent to consumers in such a way that maximizes their revenue. For example, a business may collect the following information about consumers: Percentage of emails opened Number of clicks per email Time spent viewing email Using these metrics, a business can perform cluster analysis to identify consumers who use email in similar ways and tailor the types of emails and frequency of emails they send to different clusters of customers. Example 5: Health Insurance Actuaries at health insurance companies often used cluster analysis to identify “clusters” of consumers that use their health insurance in specific ways. For example, an actuary may collect the following information about households: Total number of doctor visits per year Total household size Total number of chronic conditions per household Average age of household members An actuary can then feed these variables into a clustering algorithm to identify households that are similar. The health insurance company can then set monthly premiums based on how often they expect households in specific clusters to use their insurance. Additional Resources The following tutorials explain how to perform various types of cluster analysis using statistical programming languages: How to Perform K-Means Clustering in Python How to Perform K-Means Clustering in R How to Perform K-Medoids Clustering in R How to Perform Hierarchical Clustering in R Written by: Will Webster Reviewed by: Alex Mendoza Cluster analysis is a statistical method for processing data. It works by organizing items into groups – or clusters – based on how closely associated they are. Cluster analysis, like dimension reduction analysis (factor analysis), is concerned with data collection in which the variables have not been partitioned beforehand into criterion vs. predictor subsets. If we think of variables as individual data points or features that are being looked at, criterion subsets are the variables you’re trying to predict or explain, while predictor subsets are the variables you’re using to make those predictions. The objective of cluster analysis is to find similar groups of subjects, where the “similarity” between each pair of subjects represents a unique characteristic of the group vs. the larger population/sample. Strong differentiation between groups is indicated through separate clusters; a single cluster indicates extremely homogeneous data. Cluster analysis is an unsupervised learning algorithm, meaning that you don’t know how many clusters exist in the data before running the model. Unlike many other statistical methods, cluster analysis is typically used when there is no assumption made about the likely relationships within the data. It provides information about where associations and patterns in data exist, but not what those might be or what they mean. Product Tour: XM for Strategic Research When should cluster analysis be used? Cluster analysis is for when you’re looking to segment or categorize a dataset into groups based on similarities, but aren’t sure what those groups should be. While it’s tempting to use cluster analysis in many different research projects, it’s important to know when it’s genuinely the right fit. Here are three of the most common scenarios where cluster analysis proves its worth. Exploratory data analysis When you have a new dataset and are in the early stages of understanding it, cluster analysis can provide a much-needed guide. By forming clusters, you can get a read on potential patterns or trends that could warrant deeper investigation. Market segmentation This is a golden application for cluster analysis, especially in the business world. Because when you aim to target your products or services more effectively, understanding your customer base becomes paramount. Cluster analysis can carve out specific customer segments based on buying habits, preferences or demographics, allowing for tailored marketing strategies that resonate more deeply. Resource allocation Be it in healthcare, manufacturing, logistics or many other sectors, resource allocation is often one of the biggest challenges. Cluster analysis can be used to identify which groups or areas require the most attention or resources, enabling more efficient and targeted deployment. How is cluster analysis used? The most common use of cluster analysis is classification. Subjects are separated into groups so that each subject is more similar to other subjects in its group than to subjects outside the group. In a market research context, cluster analysis might be used to identify categories like age groups, earnings brackets, urban, rural or suburban location. In marketing, cluster analysis can be used for audience segmentation, so that different customer groups can be targeted with the most relevant messages. Healthcare researchers might use cluster analysis to find out whether different geographical areas are linked with high or low levels of certain illnesses, so they can investigate possible local factors contributing to health problems. Employers, on the other hand, could use cluster analysis to identify groups of employees who have similar feelings about workplace culture, job satisfaction or career development. With this data, HR departments can tailor their initiatives to better suit the needs of specific clusters, like offering targeted training programs or improving office amenities. Whatever the application, data cleaning is an essential preparatory step for successful cluster analysis. Clustering works at a data-set level where every point is assessed relative to the others, so the data must be as complete as possible. Cluster analysis in action: A step-by-step example Here’s how an online bookstore used cluster analysis to transform its raw data into actionable insights. Step one: Creating the objective The bookstore’s aim is to provide more personalized book recommendations to its customers. The belief is that by curating book selections that will be more appealing to subgroups of its customers, the bookstore will see an increase in sales. Step two: Using the right data The bookstore has its own historical sales data, including two key variables: ‘favorite genre’, which includes categories like sci-fi, romance and mystery; and ‘average spend per visit’. The bookstore opts to hone in on these two factors as they are likely to provide the most actionable insights for personalized marketing strategies. Step three: Choosing the best approach After settling on the variables, the next decision is determining the right analytical approach. The bookstore opts for K-means clustering for the ‘average spend per visit’ variable because it’s numerical – and therefore scalar data. For ‘favorite genre’, which is categorical – and therefore non-scalar data – they choose K-medoids. Step four: Running the algorithm With everything set, it’s time to crunch the numbers. The bookstore runs the K-means and K-medoids clustering algorithms to identify clusters within their customer base. The aim is to create three distinct clusters, each encapsulating a specific customer profile based on their genre preferences and spending habits. Step five: Validating the clusters Once the algorithms have done their work, it’s important to check the quality of the clusters. For this, the bookstore looks at intracluster and intercluster distances. A low intracluster distance means customers within the same group are similar, while a high intercluster distance ensures the groups are distinct from each other. In other words, the customers within each group are similar to one another and the group of customers are distinct from one another. Step six: Interpreting the results Now that the clusters are validated, it’s time to dig into what they actually mean. Each cluster should represent a specific customer profile based solely on ‘favorite genre’ and ‘average spend per visit’. For example, one cluster might consist of customers who are keen on sci-fi and tend to spend less than $20, while another cluster could be those who prefer romance novels and are in the $20-40 spending range. Step seven: Applying the findings The final step is all about action. Armed with this new understanding of their customer base, the bookstore can now tailor its marketing strategies. Knowing what specific subgroups like to read and how much they’re willing to spend, the store can send out personalized book recommendations or offer special discounts to those specific clusters – aiming to increase sales and customer satisfaction. Cluster analysis algorithms Your choice of cluster analysis algorithm is important, particularly when you have mixed data. In major statistics packages you’ll find a range of preset algorithms ready to number-crunch your matrices. K-means and K-medoid are two of the most suitable clustering methods. In both cases (K) = the number of clusters. K-means The K-means algorithm establishes the presence of clusters by finding their centroid points. A centroid point is the average of all the data points in the cluster. By iteratively assessing the Euclidean distance between each point in the dataset, each one can be assigned to a cluster. The centroid points are random to begin with and will change each time as the process is carried out. K-means is commonly used in cluster analysis, but it has a limitation in being mainly useful for scalar data. K-medoids K-medoid works in a similar way to K-means, but rather than using mean centroid points which don’t equate to any real points from the dataset, it establishes medoids, which are real interpretable data-points. The K-medoids clustering algorithm offers an advantage for survey data analysis as it is suitable for both categorical and scalar data. This is because rather than measuring Euclidean distance between the medoid point and its neighbors, the algorithm can measure distance in multiple dimensions, representing a number of different categories or variables. K-medoids is less common than K-means in clustering analysis, but is often used when a more robust method that’s less sensitive to outliers is needed. Measuring clusters using intracluster and intercluster distances Evaluating the quality of clustering involves a two-pronged approach: assessing intracluster and intercluster distances. Intracluster distance is the distance between the data points inside the cluster. If there is a strong clustering effect present, this should be small (more homogenous). Intercluster distance is the distance between data points in different clusters. Where strong clustering exists, these should be large (more heterogenous). In an ideal clustering scenario, you’d use both measures to gauge how good your clusters are. Low intracluster distances – known as high intra-cluster similarity – mean items in the same cluster are similar, which is good; high intercluster distances – known as low inter-cluster similarity – mean different clusters are well-separated, which is also good. Using both measures gives you a fuller picture of how effective your clustering is. Key considerations in cluster analysis When getting started with cluster analysis, it makes sense to start with methods that assign each data point to a single, distinct cluster. It’s commonly accepted that within each cluster, the data points share similarities. The assumption here is that your data set is composed of different, unordered classes, and that none of these classes are inherently more important than the others. In some cases, however, we may also view these classes as hierarchical in nature, with sub-classes within them – here we could apply hierarchical clustering and hierarchical cluster analysis. Cluster analysis is often a “preliminary” step. That means before you even start, you’re not applying any previous judgments to split up your data; you’re working on the notion that natural clusters should exist within the data. This initial approach differs from techniques like discriminant analysis, where you have a dependent variable guiding the classification. In cluster analysis, however, the focus is purely on inherent similarities within the data collection itself. So, the key questions for cluster analysis would be: What metrics will you use to measure the similarity between data points, and how will each variable be weighted when calculating this measure? Once you’ve determined the similarities, what methods will you use to form the clusters? After forming clusters, what descriptive metrics will help define the nature of each cluster? Assuming you’ve adequately described your clusters, what can you infer about their statistical significance? This should offer a clearer yet still approachable overview of the essential questions in cluster analysis. Non-scalar data in cluster analysis So far, we’ve mainly talked about scalar data – things that differ from each other by degrees along a scale, such as numerical quantity or degree. But what about items that are non-scalar and can only be sorted into categories? When you’re dealing with such categories like color, species and shape, you can’t easily measure the “distance” between data points like you can with scalar data. Various techniques, like using dummy variables or specialized distance measures, can be employed to include non-scalar data in your cluster analysis. Dummy variables are a way to convert categories into a format that can be provided to a mathematical model. For example, if you have a color category with options like red, blue and green, you could create separate “dummy” columns for each color, marking them as 1 if they apply and 0 if they don’t. Specialized distance measures, on the other hand, are custom calculations designed to figure out how “far apart” different categories are from each other. For example, if you’re clustering based on movie genres, a specialized measure might decide that “action” and “adventure” are closer to each other than “action” and “romance”. Ideally, the data for cluster analysis is categorical, interval or ordinal data. Using a mix of these types can complicate the analysis, as you’ll need to figure out how to meaningfully compare different kinds of data. It’s doable, but it adds an extra layer of complexity you’ll need to account for. Cluster analysis and factor analysis When you’re dealing with a large number of variables – for example a lengthy or complex survey – it can be useful to simplify your data before performing cluster analysis so that it’s easier to work with. Using factors reduces the number of dimensions that you’re clustering on, and can result in clusters that are more reflective of the true patterns in the data. Factor analysis is a technique for taking large numbers of variables and combining those that relate to the same underlying factor or concept, so that you end up with a smaller number of dimensions. For example, factor analysis might help you replace questions – like “Did you receive good service?”, “How confident were you in the agent you spoke to?” and “Did we resolve your query?” – with a single factor: customer satisfaction. This way you can reduce messiness and complexity in your data and arrive more quickly at a manageable number of clusters. Ready to dive into cluster analysis? Stats iQ™ makes its easy If you’re keen to perform cluster analysis but don’t have a stats degree, don’t worry – Stats iQ has you covered. The world’s most powerful statistical analysis tool does the heavy lifting, running the appropriate tests and translating complex results into straightforward language. Whether you’re looking to segment your market or explore new datasets, Stats iQ gives you the confidence to take the next step. How can financial brands set themselves apart through visual storytelling? Our experts explain how.Learn MoreThe Motorsport Images Collections captures events from 1895 to today’s most recent coverage.Discover The CollectionCurated, compelling, and worth your time. Explore our latest gallery of Editors’ Picks.Browse Editors' FavoritesHow can financial brands set themselves apart through visual storytelling? Our experts explain how.Learn MoreThe Motorsport Images Collections captures events from 1895 to today’s most recent coverage.Discover The CollectionCurated, compelling, and worth your time. Explore our latest gallery of Editors’ Picks.Browse Editors' FavoritesHow can financial brands set themselves apart through visual storytelling? Our experts explain how.Learn MoreThe Motorsport Images Collections captures events from 1895 to today’s most recent coverage.Discover The CollectionCurated, compelling, and worth your time. Explore our latest gallery of Editors’ Picks.Browse Editors' Favorites Cluster is a group of objects that belongs to the same class. In other words, similar objects are grouped in one cluster and dissimilar objects are grouped in another cluster. What is Clustering? Clustering is the process of making a group of abstract objects into classes of similar objects. Points to Remember A cluster of data objects can be treated as one group. While doing cluster analysis, we first partition the set of data into groups based on data similarity and then assign the labels to the groups. The main advantage of clustering over classification is that, it is adaptable to changes and helps single out useful features that distinguish different groups. Clustering analysis is broadly used in many applications such as market research, pattern recognition, data analysis, and image processing. Clustering can also help marketers discover distinct groups in their customer base. And they can characterize their customer groups based on the purchasing patterns. In the field of biology, it can be used to derive plant and animal taxonomies, categorize genes with similar functionalities and gain insight into structures inherent to populations. Clustering also helps in identification of areas of similar land use in an earth observation database. It also helps in the identification of groups of houses in a city according to house type, value, and geographic location. Clustering also helps in classifying documents on the web for information discovery. Clustering is also used in outlier detection applications such as detection of credit card fraud. As a data mining function, cluster analysis serves as a tool to gain insight into the distribution of data to observe characteristics of each cluster. The following points throw light on why clustering is required in data mining − Scalability − We need highly scalable clustering algorithms to deal with large databases. Ability to deal with different kinds of attributes − Algorithms should be capable to be applied on any kind of data such as interval-based (numerical) data, categorical, and binary data. Discovery of clusters with attribute shape − The clustering algorithm should be capable of detecting clusters of arbitrary shape. They should not be bounded to only distance measures that tend to find spherical cluster of small sizes. High dimensionality − The clustering algorithm should not only be able to handle low-dimensional data but also the high dimensional space. Ability to deal with noisy data − Databases contain noisy, missing or erroneous data. Some algorithms are sensitive to such data and may lead to poor quality clusters. Interpretability − The clustering results should be interpretable, comprehensible, and usable. Clustering Methods 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 Suppose we are given a database of n objects and the partitioning method constructs k partition of data. Each partition will represent a cluster and k ≤ n. It means that it will classify the data into k groups, which satisfy the following requirements − Each group contains at least one object. Each object must belong to exactly one group. Points to remember − For a given number of partitions (say k), the partitioning method will create an initial partitioning. Then it uses the iterative relocation technique to improve the partitioning by moving objects from one group to other. Hierarchical Methods This method creates a hierarchical decomposition of the given set of data objects. We can classify hierarchical methods on the basis of how the hierarchical decomposition is formed. There are two approaches here − Agglomerative Approach Divisive Approach Agglomerative Approach This approach is also known as the bottom-up approach. In this, we start with each object forming a separate group. It keeps on merging the objects or groups that are close to one another. It keep on doing so until all of the groups are merged into one or until the termination condition holds. Divisive Approach This approach is also known as the top-down approach. In this, we start with all of the objects in the same cluster. In the continuous iteration, a cluster is split up into smaller clusters. It is down until each object in one cluster or the termination condition holds. This method is rigid, i.e., once a merging or splitting is done, it can never be undone. Approaches to Improve Quality of Hierarchical Clustering Here are the two approaches that are used to improve the quality of hierarchical clustering − Perform careful analysis of object linkages at each hierarchical partitioning. Integrate hierarchical agglomeration by first using a hierarchical agglomerative algorithm to group objects into micro-clusters, and then performing macro-clustering on the micro-clusters. Density-based Method This method is based on the notion of density. The basic idea is to continue growing the given cluster as long as the density in the neighborhood 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 this, the objects together form a grid. The object space is quantized into finite number of cells that form a grid structure. Advantages The major advantage of this method is fast processing time. It is dependent only on the number of cells in each dimension in the quantized space. Model-based methods In this method, a model is hypothesized for each cluster to find the best fit of data for a given model. This method locates the clusters by clustering the density function. It reflects spatial distribution of the data points. This method also provides a way to automatically determine the number of clusters based on standard statistics, taking outlier or noise into account. It therefore yields robust clustering methods. Constraint-based Method In this method, the clustering is performed by the incorporation of user or application-oriented constraints. A constraint refers to the user expectation or the properties of desired clustering results. Constraints provide us with an interactive way of communication with the clustering process. Constraints can be specified by the user or the application requirement. Take ownership of your education at Reed College. Pursue knowledge, conduct hands-on research, consider diverse perspectives, and embark on an exhilarating journey of intellectual exploration. At Reed, we support students' progress through thoughtfully written feedback—de-emphasizing letter grades—which leads to more discussion and collaboration. Go Beyond the ABCs From assisting professors in their labs to volunteering in our nuclear reactor, Reedies actively contribute to scholarship. Research from Day One The culmination of your academic journey at Reed, your senior thesis contains original research and showcases your intellectual growth. Contribute to the Thesis Tower In this signature Reed program, explore how people living in diverse historical contexts have engaged fundamental questions about existence. Oh, the Humanities! From the shared experience of Humanities 110 to your senior thesis, study broadly across disciplines while diving deep into your major. Throughout, you have the space to ask complex questions, embrace paradoxes, and develop new perspectives. Discover Reed Academics At Reed, we look beyond the numbers. That’s why we waived application fees and embraced a test-blind policy. We find students through a holistic examination of grades, accomplishments, and perspectives. If you’re passionate about learning, we want to know more about you. Learn How to Apply to Reed Hike the 28-acre Reed canyon, relax on our rolling lawns, and study in our beautiful buildings. Take a short bike ride to downtown Portland, Oregon, or plan a trip to our ski cabin on Mount Hood. At Reed, avenues for discovery and growth are everywhere. Visit Reed Explore Reed Virtually