2025

Data Mining & Techniques Cluster Analysis

International Journal of Science, Engineering and Technology
ISSN: 2395-4752

Mrs. Annam Rupa
Assistant Professor
annam.Rupa@Gmail.Com
VMTW

Mrs.Abbagouni Swetha
Assistant Professor
abbagouniswetha1990@Gmail.Com
VMTW

Mrs. Ratcha Jamuna
Assistant Professor
jamuna.kongari1@gmail.com
VMTW

INDEX

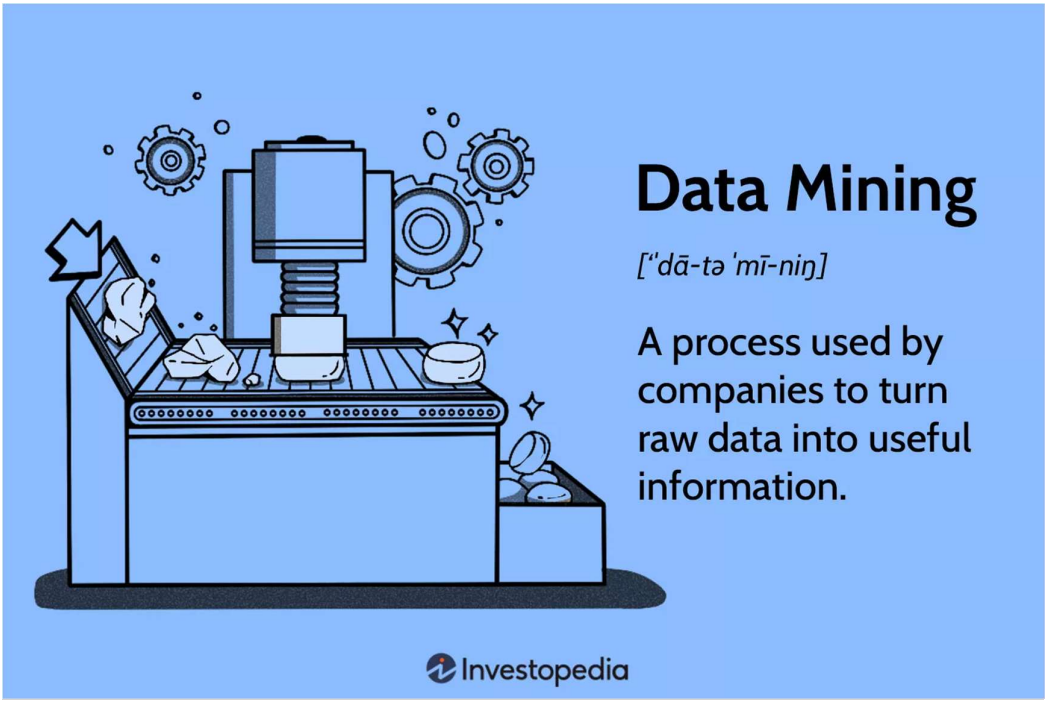
International Journal of Science, Engineering and Technology
ISSN (Print): 2395-4752

What is Data Mining?

Data mining is the process of searching and analyzing a large batch of raw data in order to identify patterns and extract useful information.

KEY TAKEAWAYS

- Data mining is the process of analyzing a large batch of information to discern trends and patterns.
- Data mining can be used by corporations for everything from learning about what customers are interested in or want to buy to fraud detection and spam filtering.
- Data mining programs break down patterns and connections in data based on what information users request or provide.
- Social media companies use data mining techniques to commodify their users in order to generate profit.
- This use of data mining has come under criticism as users are often unaware of the data mining happening with their personal information, especially when it is used to influence preferences.



How Data Mining Works

Data mining involves exploring and analyzing large blocks of information to glean meaningful patterns and trends. It is used in credit risk management, fraud detection, and spam filtering. It also is a market research tool that helps reveal the sentiment or

opinions of a given group of people. The data mining process breaks down into four steps:

. Data is collected and loaded into data warehouses on site or on a cloud service.
. Business analysts, management teams, and information technology professionals access the data and determine how they want to organize it.
. Custom application software sorts and organizes the data.
. The end user presents the data in an easy-to-share format, such as a graph or table.

Where is Data Mining (DM) Used?

Numerous sectors, including healthcare, retail, banking, government, and manufacturing, use Data Mining extensively.

For instance, if a business wants to recognize trends or patterns among the customers who purchase particular goods, it can use data-gathering techniques to examine past purchases and create models that anticipate which customers will want to purchase merchandise based on their features or behavior. Data mining, therefore, aids businesses in creating more effective sales techniques in the retail industry.

These tools can also be applied to:

- Predict Cancellations: Using past data, determine which clients are likely to cancel their orders.
- Product and Service Recommendation: Users should be given product and service recommendations based on their prior usage.
- Customer Segmentation: Customers should be divided into groups based on similar habits so that personalized marketing messages may be sent to each group.
- Fraud Detection: This is possible by using historical transaction data to spot and stop suspicious behavior.

.

KDD in Data Mining

KDD stands for Knowledge Discovery in Databases, which is the process of extracting useful knowledge from large amounts of data. It is an area of interest to researchers and professionals in various fields, such as artificial intelligence, machine learning, pattern recognition, databases, statistics, and data visualization. Data mining is a key component of the KDD process.

What is KDD in Data Mining

KDD (Knowledge Discovery in Databases) is a process of discovering useful knowledge and insights from large and complex datasets. The KDD process involves a range of techniques and methodologies, including data preprocessing, data transformation, data mining, pattern evaluation, and knowledge representation. KDD and data mining are closely related processes, with data mining being a key component and subset of the KDD process.

The KDD process aims to identify hidden patterns, relationships, and trends in data that can be used to make predictions, decisions, and recommendations. KDD is a broad and interdisciplinary field used in various industries, such as finance, healthcare, marketing, e-commerce, etc. KDD is very important for organizations and businesses as it enables them to derive new insights and knowledge from their data, which can be further used to improve decision-making, enhance the customer experience, improve business processes, support strategic planning, optimize operations, and drive business growth.

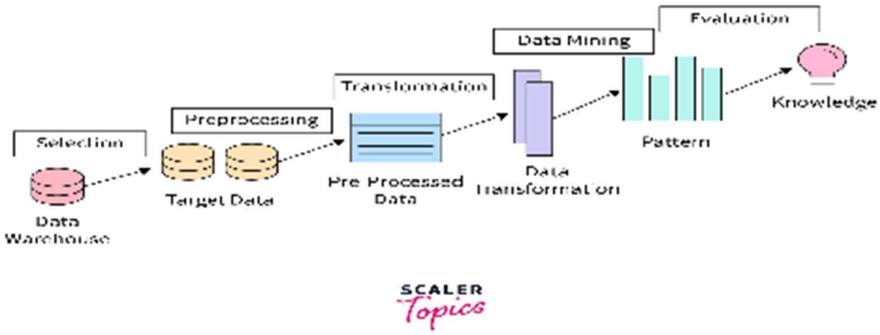
KDD Process in Data Mining

The KDD process in data mining is a multi-step process that involves various stages to extract useful knowledge from large datasets. The following are the main steps involved in the KDD process -

- **Data Selection** - The first step in the KDD process is identifying and selecting the relevant data for analysis. This involves choosing the relevant data sources, such as databases, data warehouses, and data streams, and determining which data is required for the analysis.
- **Data Preprocessing** - After selecting the data, the next step is data preprocessing. This step involves cleaning the data, removing outliers, and removing missing, inconsistent, or irrelevant data. This step is critical, as the data quality can significantly impact the accuracy and effectiveness of the analysis.
- **Data Transformation** - Once the data is preprocessed, the next step is to transform it into a format that data mining techniques can analyze. This step involves reducing the data dimensionality, aggregating the data, normalizing it, and discretizing it to prepare it for further analysis.
- **Data Mining** - This is the heart of the KDD process and involves applying various data mining techniques to the transformed data to discover hidden patterns, trends, relationships, and insights. A few of the most common data mining techniques include clustering, classification, association rule mining, and anomaly detection.
- **Pattern Evaluation** - After the data mining, the next step is to evaluate the discovered patterns to determine their usefulness and relevance. This involves assessing the quality of the patterns, evaluating their significance, and selecting the most promising patterns for further analysis.
- **Knowledge Representation** - This step involves representing the knowledge extracted from the data in a way humans can easily understand and use. This can be done through visualizations, reports, or other forms of communication that provide meaningful insights into the data.

- **Deployment** - The final step in the KDD process is to deploy the knowledge and insights gained from the data mining process to practical applications. This involves integrating the knowledge into decision-making processes or other applications to improve organizational efficiency and effectiveness.

In summary, the KDD process in data mining involves several steps to extract useful knowledge from large datasets. It is a comprehensive and iterative process that requires careful consideration of each step to ensure the accuracy and effectiveness of the analysis. Various steps involved in the KDD process in data mining are shown below diagram -



For a Hands-On Approach, Check out Scaler's Data Science Course that Offers Interactive Modules. Enroll and Get Certified by the Best!

Advantages of KDD in Data Mining

KDD in data mining is a powerful approach for extracting useful knowledge and insights from large datasets. It is very important for organizations as it has a lot of advantages. Some of the advantages of KDD in data mining are -

- **Helps in Decision Making** - KDD can help make informed and data-driven decisions by discovering hidden patterns, trends, and relationships in data that might not be immediately apparent.
- **Improves Business Performance** - KDD can help organizations improve their business performance by identifying areas for improvement, optimizing processes, and reducing costs.
- **Saves Time and Resources** - KDD can help save time and resources by automating the data analysis process and identifying the most relevant and significant information or knowledge.
- **Increases Efficiency** - KDD can help organizations streamline their processes, optimize their resources, and increase their overall efficiency.
- **Enhances Customer Experience** - KDD can help organizations improve customer experience by understanding customer behavior, preferences, and requirements and giving personalized products and services.

- **Fraud Detection** - KDD can help detect fraud and identify fraudulent behavior by analyzing patterns in data and identifying anomalies or unusual behavior.
- **Enables Predictive Modeling** - KDD can enable organizations to develop predictive models that can forecast future trends and behaviors, providing a competitive advantage in the market.

Disadvantages of KDD in Data Mining

While KDD (Knowledge Discovery in Databases) is a powerful approach to extracting useful knowledge and insights from large datasets, there are also some potential disadvantages to consider -

- **Requires High-Quality Data** - KDD relies on high-quality data to generate accurate and meaningful insights. If the data is incomplete, inconsistent, or of poor quality, it can lead to inaccurate, misleading results and flawed conclusions.
- **Complexity** - KDD is a complex and time-consuming process that requires specialized skills and knowledge to perform effectively. The complexity can also make interpreting and communicating the results challenging to non-experts.
- **Privacy and Compliance Concerns** - KDD can raise ethical concerns related to privacy, compliance, bias, and discrimination. For example, data mining techniques can extract sensitive information about individuals without their consent or reinforce existing biases or stereotypes.
- **High Cost** - KDD can be expensive, and require specialized software, hardware, and skilled professionals to perform the analysis. The cost can be prohibitive for smaller organizations or those with limited resources.

Different Types of Data Mining Techniques

1. Classification

Data are categorized to separate them into predefined groups or classes. Based on the values of a number of attributes, this method of data mining identifies the class to which a document belongs. Sorting data into predetermined classes is the aim.

Predicting a variable that can have one of two or more different values (for example, spam/not spam; good or neutral/negative evaluation) given one or even more input factors called predictors is the most typical application of classification.

2. Clustering

The next data mining technique is clustering. Similar entries inside a database are grouped together using the clustering approach to form clusters. The clustering first identifies these groups inside the dataset and afterward classifies factors based on

their properties, in contrast to classification, which places variables into established categories.

For instance, you can group clients based on sales data, such as those who consistently purchase certain drinks or pet food and have consistent taste preferences. You may easily target these clusters with specialized adverts once you've established them.

Clustering has several uses, including the following:

- Web analytics
- Text mining
- Biological computation
- Medical Diagnosis

3. Association Rule Learning

Finding if-then patterns between two or more independent variables is done through association rule learning. The relationship between purchasing bread and butter is the most basic illustration. Butter is frequently purchased along with bread, and vice versa. Because of this, you can find these two products side by side at a grocery shop.

The connection might not be so direct, though. For instance, Walmart found in 2004 that Strawberry Pop-Tart sales peaked just before the hurricane. Along with stocking up on necessities like batteries, many also bought these well-liked treats.

In hindsight, the psychological motive is rather clear: having your favorite meal on hand during emergencies gives you a sense of security, and tarts with a long shelf life are the ideal choice. But data mining methods had to be used in order to identify this association.

4. Regression

The next data mining technique is Regression. A link between variables is established using regression. Its objective is to identify the appropriate function that best captures the relationship. Linear regression analysis is the term used when a linear function ($y = axe + b$) is applied.

Methods like multiple linear regression, quadratic regression, etc., can be used to account for additional kinds of relationships. Planning and modeling are the two most prevalent applications. One illustration is estimating a customer's age based on past purchases. We may also forecast costs based on factors like consumer demand; for

instance, if demand for vehicles in the US increases, prices on the secondary market would rise.

5. Anomaly Detection

A data mining technique called anomaly detection is used to find outliers (values that deviate from the norm). For instance, it can identify unexpected sales at a store location during a specific week in e-commerce information. It can be used, among other things, to find credit or debit fraud and spot network attacks or disruptions.

6. Sequential Pattern Mining

A data mining technique known as sequential pattern mining finds significant connections between events. We can discuss a dependency between events when we can pinpoint a time-ordered sequence that occurs with a particular frequency.

Let's imagine we wish to look into how a drug or a specific therapeutic approach affects cancer patients' life expectancy. By including a temporal component in the study, sequential pattern mining makes it possible for you to do that.

This method can be used, among other things, in medicine to determine how to administer a patient's medicines and in security to foresee potential systemic attacks.

Sequential pattern mining has several uses, such as:

- DNA-sequencing studies
- Natural catastrophes
- Stock exchanges
- Shopping patterns
- Medical procedures

7. Artificial Neural Network Classifier

A process model supported by biological neurons could be an artificial neural network (ANN), also known as a "Neural Network" (NN). It is made up of a networked group of synthetic neurons. A neural network is a collection of connected input/output units with weights assigned to each connection.

In order to be able to anticipate the class label of the input samples correctly, the network accumulates information during the knowledge phase by modifying the

weights. Due to the links between units, neural network learning is also known as connectionist learning.

Neural networks require lengthy training periods, making them more suitable for applications where it is possible. They need a variety of parameters, like the network topology or "structure," which are often best determined empirically.

Since it is challenging for humans to understand the symbolic significance of the acquired weights, neural networks have come under fire for their poor interpretability. First, these characteristics reduced the appeal of neural networks for data mining.

However, neural networks' strengths include their high level of noise tolerance and their capacity to classify patterns for which they have not yet been taught. Additionally, a number of novel methods have been created to extract rules from trained neural networks. These problems affect how effective neural networks are at classifying data in data mining.

An artificial neural network is a machine that modifies its structure in response to information that passes through it during a learning phase. The learning-by-example principle underlies the ANN. Perceptron and multilayer perceptron are two of the most traditional neural network architectures.

8. Outlier Analysis

Data objects that do not adhere to the overall behavior or model of the data may be found in a database. These informational items are outliers. OUTLIER MINING is the process of looking into OUTLIER data.

When employing distance measurements, objects with a tiny percentage of "near" neighbors in space are regarded as outliers. Statistical tests that assume a distribution and probability model for the data can also be used to identify outliers.

Deviation-based strategies identify exceptions/outliers by examining variances in the primary features of items in a collection rather than using factual or distance metrics.

9. Prediction

The next data mining technique is Prediction. Data classification and data prediction both involve two steps. Despite the fact that we do not use the term "Class label attribute" for prediction because the attribute whose values are being forecasted is consistently valued (ordered) rather than category (discrete-esteemed and unordered).

Simply calling the attribute "the expected attribute" will do. Prediction can be thought of as the creation and use of a model to determine the class of an unlabeled item or the value or ranges of a particular attribute that an object is likely to possess.

10. Genetic Algorithms

The majority of evolutionary algorithms are genetic algorithms, which are adaptive heuristic algorithms. Natural selection and genetics are the foundations of genetic algorithms. These are clever uses of random search that are supported by historical data to focus the search on areas with superior performance in the solution space. They are frequently employed to produce excellent answers to optimization and search-related issues.

Natural selection is simulated by genetic algorithms, which means that only those species that can adapt to changes in their environment will be able to survive, procreate, and pass on to the next generation.

In order to solve an issue, they essentially replicate "survival of the fittest" among people of successive generations. Each generation consists of a population of people, and each person represents a potential solution or a point in the search space. A string of characters, integers, floats, and bits represents every person. This string resembles a chromosome.

Applications of Data Mining

Additionally, data mining methods are becoming more popular in practically every industry, including banking, logistics, finance, and science. Data mining is also used in intelligence and law enforcement:

- Based on past border crossings, customs officials can better identify the general profile of crossing violators and concentrate on particular groups of people.
- Because they are aware of when and where crimes are most likely to occur, police can pinpoint locations where they require to increase their manpower.

Data mining is employed in finance to:

- Locate investment opportunities
- Forecast share demand, allowing potential investors to make well-informed choices.

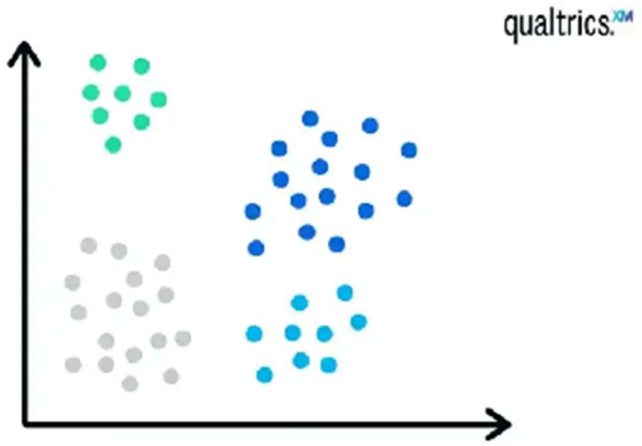
In the field of education, Data Mining aids in creating unique programs based on the following:

- The ways in which students study, such as whether they prefer to read, listen to or watch videos, or combine all three.
- Trends in the labor market make it possible to choose the educational concentration that is most pertinent.

We will now be looking at the various stages of the data mining process.

What is cluster analysis?

Cluster analysis is a statistical method for processing data. It works by organising 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.

When should cluster analysis be used?

Cluster analysis is for when you're looking to segment or categorise 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 'favourite 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

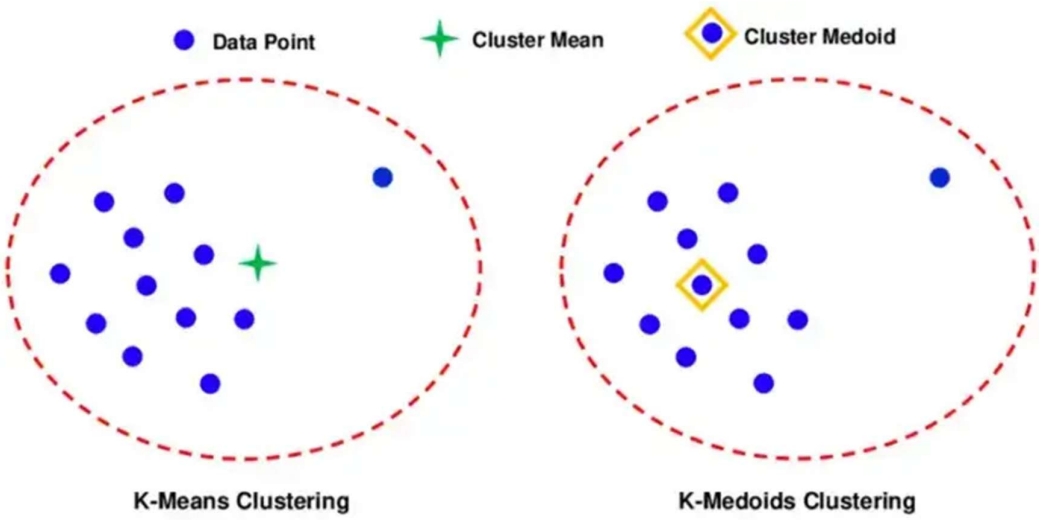
The final step is all about action. Armed with this new understanding of their customer base, the bookstore can now tailor its marketing strategies.

Knowing what specific subgroups like to read and how much they're willing to spend, the store can send out personalised 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 neighbours, 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

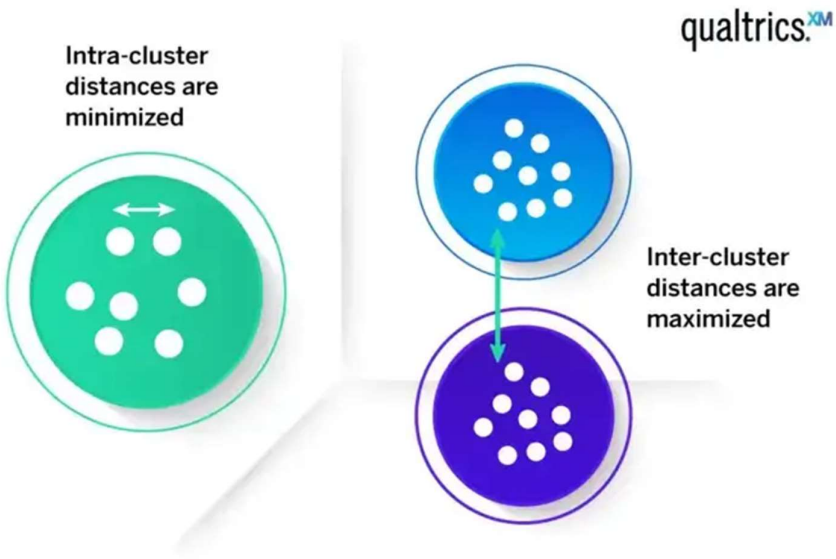
Evaluating the quality of clustering involves a two-pronged approach: assessing intracluster and intercluster distances.

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.



The Benefits of Clustering

Clustering supports data analysis in several ways, including:

- Better understanding of data. By grouping similar data points together and identifying patterns that might not be immediately obvious, clustering helps analysts better understand complex datasets.

- Improved decision-making. Clustering can help businesses and organisations identify trends or other important patterns, which in turn can provide validation, and inform optimal business decisions in areas such as product development or marketing strategies.

- Time-saving. Clustering can save a lot of time. Rather than analysing each data point individually, clustering groups similar data points together, which means that large datasets can be analysed more quickly and efficiently.

APPLICATIONS FOR CLUSTERING

The applications for clustering are virtually limitless, and they are becoming more important as data use continues to grow. As the Institute of Electrical and Electronics Engineers (IEEE) pointed out at its International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) in 2017: "Data is the goldmine in today's ever competitive world."

MARKETING

Many businesses use clustering to support their marketing activities. For example, clustering can be used to group customers based on data such as their:

- Shopping preferences.
- Online engagement and behaviours.
- Personal demographics.

With this insight, marketers can create more targeted – and effective – campaigns.

HEALTHCARE

Clustering is used in healthcare services to group patients based on their medical history, symptoms, or treatments. This can help doctors and other medical professionals identify patterns, set benchmarks, and make better diagnoses.

FINANCE

Clustering is used in finance to group stocks based on their performance, risk level, and other key metrics. This information can then be used to make investment decisions.

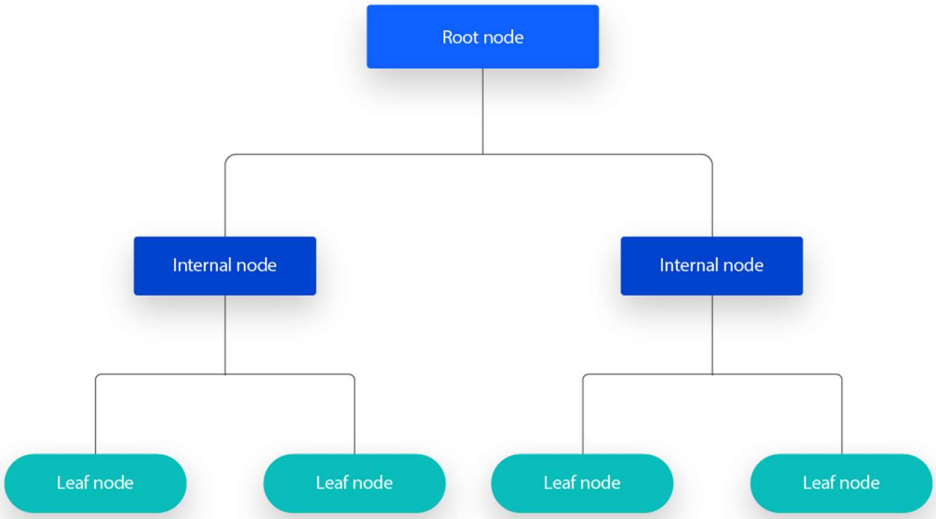
SOCIAL MEDIA

Clustering is also used by social media platforms to group users based on their behaviours or interests. This data can then be leveraged to create targeted advertising campaigns, or to recommend content to users.

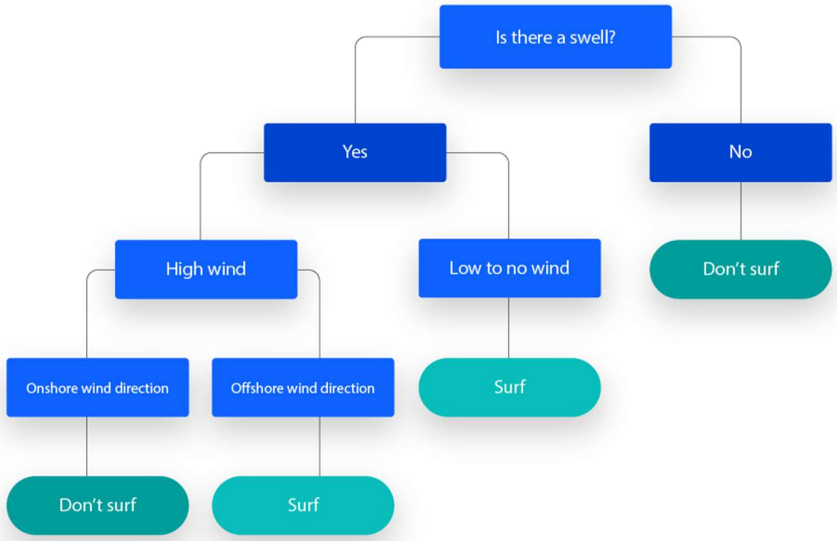
What is a decision tree?

A decision tree is a non-parametric supervised learning algorithm, which is utilized for both classification and regression tasks. It has a hierarchical, tree structure, which consists of a root node, branches, internal nodes and leaf nodes.

As you can see from the diagram below, a decision tree starts with a root node, which does not have any incoming branches. The outgoing branches from the root node then feed into the internal nodes, also known as decision nodes. Based on the available features, both node types conduct evaluations to form homogenous subsets, which are denoted by leaf nodes, or terminal nodes. The leaf nodes represent all the possible outcomes within the dataset.



As an example, let's imagine that you were trying to assess whether or not you should go surf, you may use the following decision rules to make a choice:



This type of flowchart structure also creates an easy to digest representation of decision-making, allowing different groups across an organization to better understand why a decision was made.

Decision tree learning employs a divide and conquer strategy by conducting a greedy search to identify the optimal split points within a tree. This process of splitting is then repeated in a top-down, recursive manner until all, or the majority of records have been classified under specific class labels.

Whether or not all data points are classified as homogenous sets is largely dependent on the complexity of the decision tree. Smaller trees are more easily able to attain pure leaf nodes—i.e. data points in a single class. However, as a tree grows in size, it becomes increasingly difficult to maintain this purity, and it usually results in too little data falling within a given subtree. When this occurs, it is known as data fragmentation, and it can often lead to overfitting.

As a result, decision trees have preference for small trees, which is consistent with the principle of parsimony in Occam's Razor; that is, "entities should not be multiplied beyond necessity." Said differently, decision trees should add complexity only if necessary, as the simplest explanation is often the best. To reduce complexity and prevent overfitting, pruning is usually employed; this is a process, which removes branches that split on features with low importance. The model's fit can then be evaluated through the process of cross-validation.

Another way that decision trees can maintain their accuracy is by forming an ensemble via a random forest algorithm; this classifier predicts more accurate results, particularly when the individual trees are uncorrelated with each other.

Types of decision trees

Hunt's algorithm, which was developed in the 1960s to model human learning in Psychology, forms the foundation of many popular decision tree algorithms, such as the following:

- **ID3:** Ross Quinlan is credited within the development of ID3, which is shorthand for "Iterative Dichotomiser 3." This algorithm leverages entropy and information gain as metrics to evaluate candidate splits. Some of Quinlan's research on this algorithm from 1986 can be found here.
- **C4.5:** This algorithm is considered a later iteration of ID3, which was also developed by Quinlan. It can use information gain or gain ratios to evaluate split points within the decision trees.
- **CART:** The term, CART, is an abbreviation for "classification and regression trees" and was introduced by Leo Breiman. This algorithm typically utilizes Gini impurity to identify the ideal attribute to split on. Gini impurity measures how often a randomly chosen attribute is misclassified. When evaluating using Gini impurity, a lower value is more ideal.

The latest AI News + Insights

Discover expertly curated insights and news on AI, cloud and more in the weekly Think Newsletter.

How to choose the best attribute at each node

While there are multiple ways to select the best attribute at each node, two methods, information gain and Gini impurity, act as popular splitting criterion for decision tree models. They help to evaluate the quality of each test condition and how well it will be able to classify samples into a class.

Entropy and information gain

It's difficult to explain information gain without first discussing entropy. Entropy is a concept that stems from information theory, which measures the impurity of the sample values. It is defined with by the following formula, where:

$$\text{Entropy}(S) = -\sum_{c \in C} p(c)\log_2 p(c)$$

- S represents the data set that entropy is calculated
- c represents the classes in set, S
- p(c) represents the proportion of data points that belong to class c to the number of total data points in set, S

Entropy values can fall between 0 and 1. If all samples in data set, S, belong to one class, then entropy will equal zero. If half of the samples are classified as one class and the other half are in another class, entropy will be at its highest at 1. In order to select the best feature to split on and find the optimal decision tree, the attribute with the smallest amount of entropy should be used.

Information gain represents the difference in entropy before and after a split on a given attribute. The attribute with the highest information gain will produce the best split as it's doing the best job at classifying the training data according to its target classification. Information gain is usually represented with the following formula,

$$\text{Information Gain}(S,a) = \text{Entropy}(S) - \sum_{v \in values(a)} \frac{|S_v|}{|S|} \text{Entropy}(S_v)$$

where

- *a* represents a specific attribute or class label
- *Entropy(S)* is the entropy of dataset, S
- $|Sv|/|S|$ represents the proportion of the values in S_v to the number of values in dataset, *S*.

Let's walk through an example to solidify these concepts. Imagine that we have the following arbitrary dataset:

Day	Outlook	Temp	Humidity	Wind	Tennis
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Weak	Yes

For this dataset, the entropy is 0.94. This can be calculated by finding the proportion of days where "Play Tennis" is "Yes", which is 9/14, and the proportion of days where "Play Tennis" is "No", which is 5/14. Then, these values can be plugged into the entropy formula above.

Entropy (Tennis) = -(9/14) log2(9/14) – (5/14) log2 (5/14) = 0.94

We can then compute the information gain for each of the attributes individually. For example, the information gain for the attribute, "Humidity" would be the following:

Gain (Tennis, Humidity) = (0.94)-(7/14)*(0.985) – (7/14)*(0.592) = 0.151

As a recap,

- 7/14 represents the proportion of values where humidity equals "high" to the total number of humidity values. In this case, the number of values where humidity equals "high" is the same as the number of values where humidity equals "normal".
- 0.985 is the entropy when Humidity = "high"
- 0.59 is the entropy when Humidity = "normal"

Then, repeat the calculation for information gain for each attribute in the table above, and select the attribute with the highest information gain to be the first split point in the decision tree. In this case, outlook produces the highest information gain. From there, the process is repeated for each subtree.

Gini Impurity

Gini impurity is the probability of incorrectly classifying random data point in the dataset if it were labeled based on the class distribution of the dataset. Similar to entropy, if set, S, is pure—i.e. belonging to one class) then, its impurity is zero. This is denoted by the following formula:

$$\text{Gini Impurity} = 1 - \sum_i (p_i)^2$$

Advantages and disadvantages of decision trees

While decision trees can be used in a variety of use cases, other algorithms typically outperform decision tree algorithms. That said, decision trees are particularly useful for data mining and knowledge discovery tasks. Let's explore the key benefits and challenges of utilizing decision trees more below:

Advantages

- **Easy to interpret:** The Boolean logic and visual representations of decision trees make them easier to understand and consume. The hierarchical nature of a decision tree also makes it easy to see which attributes are most important, which isn't always clear with other algorithms, like neural networks.

- **Little to no data preparation required:** Decision trees have a number of characteristics, which make it more flexible than other classifiers. It can handle various data types—i.e. discrete or continuous values, and continuous values can be converted into categorical values through the use of thresholds. Additionally, it can also handle values with missing values, which can be problematic for other classifiers, like Naïve Bayes.

- **More flexible:** Decision trees can be leveraged for both classification and regression tasks, making it more flexible than some other algorithms. It's also insensitive to underlying relationships between attributes; this means that if two variables are highly correlated, the algorithm will only choose one of the features to split on.

Disadvantages

- **Prone to overfitting:** Complex decision trees tend to overfit and do not generalize well to new data. This scenario can be avoided through the processes of pre-pruning or post-pruning. Pre-pruning halts tree growth when there is insufficient data while post-pruning removes subtrees with inadequate data after tree construction.

- **High variance estimators:** Small variations within data can produce a very different decision tree. Bagging, or the averaging of estimates, can be a method of reducing variance of decision trees. However, this approach is limited as it can lead to highly correlated predictors.

- **More costly:** Given that decision trees take a greedy search approach during construction, they can be more expensive to train compared to other algorithms.

References

- https://www.simplilearn.com/types-of-data-mining-techniques-article
- https://www.techtarget.com/searchbusinessanalytics/definition/data-mining
- file:///C:/Users/roopa/Downloads/Cleaning_Big_Data_Streams_A_Systematic_Literature_.pdf
- file:///C:/Users/roopa/Downloads/FinalID-8592-WiDS2023.pdf
- https://www.ibm.com/think/topics/decision-trees
- https://www.qualtrics.com/en-au/experience-management/research/cluster-analysis/
- https://online.keele.ac.uk/clustering-and-cluster-analysis-an-explainer/