UNIT - IV

Decision Tree Induction :-

* Decision tree Induction is a common technique in data mining that is used to generate a predictive model from a dataset.
* This technique involves constructing a tree-like structure, where each internal node represents a test on an attribute, each branch represents the outcome of the test, and each leaf node represents a prediction.
* The goal of decision tree induction is to build a model that can accurately predict the outcome of a given event, based on the values of the attributes in the dataset.
* To build a decision tree, the algorithm first selects the attribute that best splits the data into distinct classes.
* This is typically done using a measure of impurity, such as entropy or the Gini index, which measures the degree of disorder in the data.
* The algorithm then repeats this process for each branch of the tree, splitting the data into smaller and smaller subsets until all of the data is classified.

* Decision tree induction is a popular technique in data mining because it is easy to understand and interpret, and it can handle both numerical and categorical data.

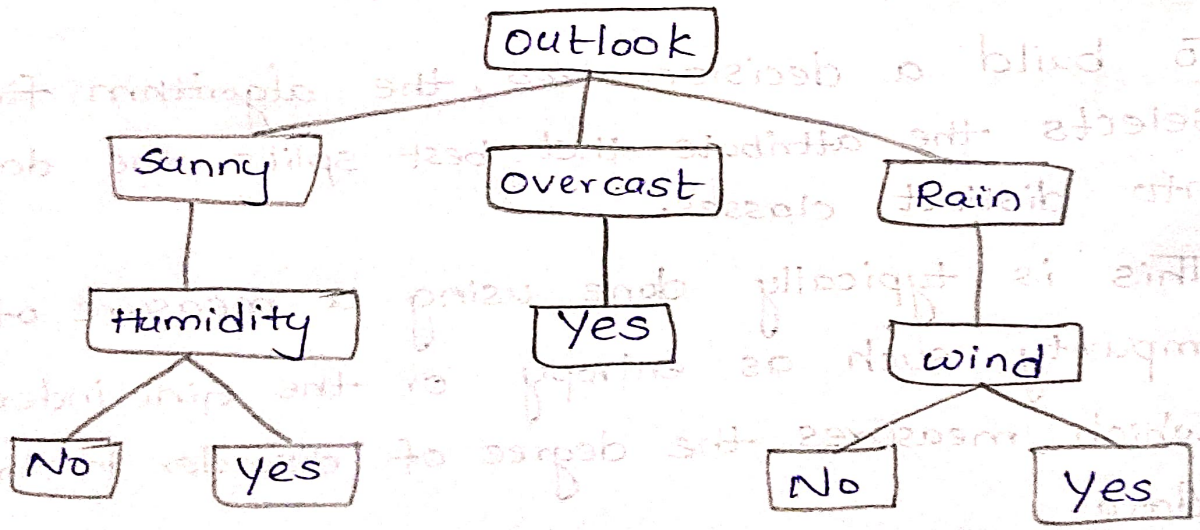
* Additionally, decision trees can handle large amounts of data, and they can be updated with new data as it becomes available.

* Decision trees can be prone to overfitting, where the model becomes too complex and does not generalize well to new data.

* Data scientists, often use techniques such as pruning to simplify the tree and improve its performance.

Example:-

Decision tree for play tennis:-



Advantages:-

1. Decision trees are a visual and intuitive model that can be easily understood by both experts and non-experts.

2. Decision trees can handle a mix of numerical and categorical data, which makes them suitable for many different types of datasets.

3. Decision trees can handle large amounts of data and can be updated with new data as it becomes available.

4. Can be used for both classification and regression tasks.

Disadvantages:-

1) Decision trees can become too complex and may not generalize well to new data. This can lead to poor performance on unseen data.

2) Decision trees can be sensitive to small changes in the data, and a small change in the data can result in a significantly different tree.

3) Decision trees can be biased towards attributes with many levels, and may not perform well on attributes with a small number of levels.

Bayesian classification:-

Bayesian classification is a statistical method used in data mining and machine learning that is based on Baye's theorem. It predicts the probability that a given data point belongs to a particular class.

* Bayesian classification works by calculating the probability that the data sample belongs to each possible class and then assigning it to the class with the highest probability.

* The Baye's theorem probability is given as

$$P(c/x) = \frac{P(x/c) \times P(c)}{P(x)}$$

where:

$P(c/x)$ - probability of class c given data x

$P(x/c)$ - probability of data x given class c

$P(c)$ - probability of class c

$P(x)$ - probability of class x

Steps in Bayesian classification:-

1. Collect training data with known class labels.
2. Calculate prior probabilities of each class.
3. Find likelihood probabilities for each attribute given the class.
4. Apply the class baye's theorem to compute posterior probability.
5. Assign the class with the highest posterior probability to the new data.

Types of Bayesian classification:-

It is classified into two types

1) Naive Bayes Classifier
2) Bayesian Belief Networks

1) Naïve Bayes theorm :-

The Naïve Bayes classifier is a simple and powerful probabilistic classifier based on Baye's theorem, which assumes that all features in the data are independent of each other.

Example :-

whether	Temperature	play tennis
sunny	Hot	No
Overcast	Hot	Yes
Rainy	mild	Yes

Advantages :-

1) simple and fast
2) works well with large datasets.
3) Handles both continuous and categorical data.
4) performs well even with missing values.

Disadvantages :-

1) Assumes all features are independent.
2) Struggles when features are highly correlated.

Applications :-

1) Email spam detection
2) Sentiment analysis
3) Medical diagnosis
4) Document Categorization

2) Bayesian belief networks :-

Bayesian belief network is a graphical model used in artificial intelligence and Data mining to represent and reason about uncertainity using probabilities.

A Bayesian Belief Network is a directed acyclic graph (DAG) where:

1) Nodes represent random variables
2) Edges represent probabilistic dependencies between variables.
3) Each node has a conditional probability table that shows how it depends on its parent node.

Steps:-

1. Identify variables that affect the problem.
2. Draw a directed acyclic graph showing dependencies among variables.
3. Assign probabilities to each node.
4. Input evidence
5. Use Baye's theorem to update probabilities and find the likelihood of unknown variables.

Uses:-

1) Medical diagnosis
2) whether prediction
3) Fraud detection
4) decision support system

Advantages of Bayesian classification:-

1) Simple and easy to implement
2) works well with large datasets
3) performs with categorical data
4) can handle missing values

Disadvantages of Bayesian classification:-

1) Assumes independence between features.
2) can perform poorly if this assumption is violated.

Rule Based classification:-

Rule-based classification is a method in data mining that uses IF-THEN rules to classify data into different categories.

Each rule describes a relationship between attribute values and a class label.

It is a technique in which class decisions are taken based on various "if--then---else" rules.

The if-then rule can be written as

"IF condition THEN conclusion"

* The "IF" part of a rule is known as the rule antecedent or precondition.
* The "then" part is the rule consequent.
* In rule antecedent, the condition consists of one or more attribute tests.

Example:-

if age = young and income = high

then

buy-s-computer = yes

Steps in Rule-Based classification:-

1. Rule generation: Extract rules from the training data set.
2. Rule pruning: Remove unnecessary or redundant rules to simplify the model.
3. Rule ordering: Decide the order in which rules should be applied.
4. Rule matching: For a new data instance, find all rules whose conditions are satisfied.
5. Class prediction: Assign the class label based on the matched rule.

Properties of Rule-Based classification:-

There are two significant properties of rule-based classification. They are

1) Rules may not be mutually exclusive

2) Rules may not be exhaustive

1) Rules may not be mutually exclusive:-

* Many different rules are generated for the dataset, so it is possible and likely that many of them satisfy the same data record. This condition makes the rules not mutually

* Since the rules are not mutually exclusive, we cannot decide on classes that cover different parts of the data on different rules. To solve this problem, we have two-way:

1) The first way using an ordered set of rules. By ordering the rules, we set priority orders. Thus, this ordered rule is called a decision list. So the class with the highest

2) The second solution can be assigning votes for each class depending on their weights. So, in this, the set of rules remains ~~an~~ unordered.

Rules may not be exhaustive in nature:-

It is not a guarantee that the rule will cover all the data entries. Any of the rules may leave some data entries. This case, on its occurance, will make the rules non-exhaustive. So, we can assign all the data entries not covered by any rules to the default class. Thus using the default class will be solve the problem of non-exhaustivity.

Advantages:-

1) Easy to understand and interpret.
2) simple to implement
3) can handle both categorical and numerical data.
4) Transparent decison making process.

Disadvantage:-

1) can become complex with large datasets.
2) Rules may conflict or overlap.
3) sensitive to noise in data.

Example:-

Training data:

Age	Income	Buys computer
Young	High	yes
Young	low	No
middle-aged	High	yes

Generated rule:-

if age = young and Income = High THEN
Buy_s_computer = yes

if age = young and Income = low THEN
Buys_computer = No

if age = middle-aged Then
Buys_computer = yes

Classification based on back propagation:-

Back propagation is a method used to train neural networks where the model learns from its mistakes. It works by measuring how wrong the output is and then adjust the weight step by step to make better prediction next time.

Working of Backpropagation:-

Neural network generate output vectors from input vectors on which neural network operates on it compares generated output with the desired output and generates an error report if the result does not match the generated output. Thein adjust weights accordingly to get desired output. It is based on gradient descent and updates weights by minimizing the error between predicted and actual output.

Training of backpropagation consist of three stages:

1) Forward propagation of input data.
2) Backward propagation of error.
3) updating weights to reduce the error.

Steps in Backpropagation:-

1. Data collection and preparation
   * Collect labeled training data.
     Ex:- Email data labeled as spam or not spam.
   * Normalize or scale the data for better performance.
2. Design the neural network
   * choose the input layer, hidden layers and output layer.

Example :

→ Inputs = number of words, links, sender info

→ output = 1 or 0

3. Initialize weights

Start with small random weights for all neuron connections.

4. ~~Fro~~ Forward Propagation

* pass the input data through the network layer by layer.
* Each neuron calculates its output using an activation function.
* The final layer gives the predicted class.

5. Calculate Error

* Compare the predicted output with the actual lebel.
* Error = (Actual output - predicted output)

6. Backward propagation

* The error is sent backward through the network.
* Each weight is adjusted based on how much it Contributed to the error.

7. Update weights

* Adjust weights using the learning rate (η):

$W_{new} = W_{old} + \eta \times \text{error} \times \text{input}$

8. Repeat until convergence

keep repeating the forward and backward steps until the total error becomes very small.

Applications:-

1. Handwritten digit recognition
2. Medical diagnosis
3. Credit risk analysis
4. Image or speech classification

Advantages:-

1) Can learn complex, non-linear relationships.
2) works well with large datasets.
3) Automatically adjusts to improve classification accuracy.

Disadvantages:-

1) Requires many training iterations
2) Can get stuck in local minima.
3) Needs proper selection of parameters.

Support vector machines:-

Support Vector Machine (SVM) is a supervised ML algorithm used for classification and regression tasks. It tries to find the best boundary known as hyperplane that seperates different classes in the data. It is useful when you want to do binary classification like spam vs. not spam or cat vs dog.

The main goal of svm is to maximize the margin between the two classes. The larger the margin the better the model performs on new and unseen data.

Key concepts of support Vector Machine:-

1) Hyperplane: A decision boundary separating different classes in feature space and is represented by the equation $wx+b=0$ in linear classification.

2) support Vectors: The closest data points to the hyperplane, crucial for determining the hyperplane and margin in SVM.

3) Margin: The distance between the hyperplane and the support vectors. SVM aims to maximize this margin for better classification performance.

4) kernel: A function that maps data to a higher-dimensional space enabling SVM to handle non-linearly seperable data.

5) Hard Margin: A maximum-margin hyperplane that perfectly seperates the data with misclassifications.

6) Soft Margin: Allows some misclassifications by introducing slack variables, balancing margin maximization and misclassification penalities when data is not perfectly separable.

7) C: A regularization term balancing margin maximization and misclassification penalties. A higher c value forces stricter penalty for misclassifications.

8) Hinge Loss: A loss function penalizing misclassified points or margin violations and is combined with regularization is SVM.

* A support vector Machine is a supervised learning algorithm used mainly for classification and regression tasks in data mining and machine learning. It works by finding the best boundary that separates data points of different classes.

* SVM tries to find a line, plane or hyperplane that:

1) Maximizes the margin - the distance between the nearest data points of different classes.

2) The nearest data points that help form this boundary are called support vectors.

Steps used in SVM :-

1. Input data: Start with labeled training data (features + class labels)

2. Find hyperplane: The algorithm searches of a hyperplane that separates the classes with the widest margin.

3. Identify support vectors: The data points closest to the hyperplane are identified; these are most critical in defining the decision boundary.

4. Classify new data: when a new data point appears, SVM predicts its class based on which side of the hyperplane falls.

5. Use kernel functions: If the data is not linearly separable, SVM uses kernel tricks to project it into a higher dimensional space where it can be

separated.

Advantages:-

1. works well for high-dimensional data.
2. Effective when there is a clear margin of separation.
3. Memory efficient - uses only support vectors.

Disadvantages:-

1. Performance drops when data is noisy or overlapping.
2. Training time can be high for large datasets.
3. Choosing the right kernel is crucial and sometimes difficult.

Applications:-

1. Text classification
2. Image recognition
3. Handwriting detection
4. Bioinformatics
5. Sentiment analysis

Lazy learners:-

* Lazy learners are algorithms that differ the generalization process until a query or prediction request is made.
* They store all training data and use it directly to make predictions for new instances.

* The goal is to extract useful patterns or knowledge from large datasets. Lazy learners are used in this context as instance-based learning methods, where the system stores the data and makes decisions when needed, instead of building a global model in advance.

working of lazy learners:-

1. Store all training data in memory.
2. when a new ~~ist~~ instance is given, the system
   1. Measures similarity or distance between the new instance and all stored data.
   2. Selects the most similar data points.
   3. Makes a decision based on these neighbours.

Examples:-

The examples in data mining are:

1. K-Nearest Neighbour (K-NN)
2. Case-Based Reasoning (CBR)

1. K-Nearest Neighbour :-

The K-Nearest neighbour algorithm is a lazy learning method that classifies a new data point based on the classes of its nearest neighbours in the training dataset. It uses distant measures to find k most similar data points.

Steps:-

1. Store the training data
2. Choose the value of k

3. calculate the distance

4. Find the k nearest neighbours

5. Majority voting/averaging

6. output the result

Advantages:-

1. simple and easy to understand.

2. No training phase - fast to set up.

3. Can handle multi-class classification problems.

4. works well for non linear decision boundaries.

Disadvantages:-

1. slow for large datasets

2. High memory usage

3. sensitive to noisy or irrelevant features.

4. choice of k and distance metric affects accuracy.

2. Case-Based Reasoning (CBR) :-

Case-Based Reasoning is a lazy learning and problem solving technique in data mining and AI where new problems are solved using solutions of similar past problems.

Instead of building a model, the system store previous cases in a case library.

when a new problem appears, it:

1. Finds the most similar old case.

2. Reuses its solution.

3. Adapts the solution if needed.

4. Stores this new experience for future use.

* It is a 4 R's process
Retrieve, Reuse, Revise, Retain

Steps in CBR:-

1. Represent cases
2. Retrieve similar cases
3. Reuse the case
4. Revise the proposed solution
5. Retain the new case

Advantages:-

1. Learns continuously
2. Uses real experiences - not abstract rules
3. Easy to update with new cases.
4. Useful when exact rules or formulas are unavailable.

Disadvantages:-

1. Requires a large, well-organized case library.
2. Searching for similar cases can be computationally expensive.
3. Case adaption may be difficult for complex problems.

Applications:-

1. Technical support and fault diagnosis
2. Legal decision-making
3. Customer support systems

Model evaluation and selection:-

Model evaluation and selection involve systematically assessing multiple predictive or descriptive models to determine which best meets the goals of a data-driven task such as classification, regression, or clustering. Careful evaluation and selection are crucial for maximizing model generalizability and real-world performance.

Model Evaluation in Data mining:-

Model evaluation is the process of quantifying how well a data mining model captures underlying data patterns, typically using statistical metrics and independent validation sets.

1. Classification: Accuracy, precision, recall, F1 score, ROC-AUC
2. Regression: Mean Absolute Error (MAE), Mean squard Error (MSE), Root mean squard Error (RMSE).
3. Clustering & Association: silhouette score, Davies-Bouldin index, support, confidence.

Validation strategies often involve splitting data into training and testing sets, performing cross-validation, and, in some cases, repeated random splits to average out model performance.

Model selection process:-

Model selection in data mining is the systematic process of comparing candidate models and their parameter settings.

Common steps include:

1. Understanding the task and data characteristics.
2. Identifying and select candidate models.
3. define evaluation metrics
4. Train and evaluate metrics
5. Hyperparameter tuning
6. Compare models
7. select final model

Popular techniques include grid search, random search, Bayesian optimization, and cross-validation-based comparision.

Evaluation metrics for classification Tasks:-

Classification models assign inputs to predefined labels. Their performance can be measured using accuracy, precision, recall and F1 score.

Accuracy:-

we will calculate the accuracy,

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

TP: True positivities

TN: True negatives

FP: False positives

FN: False negatives

accuracy_score() computes the proportion of correct predictions.

```
print ("Accuracy:"
        accuracy_score (y_test, y_pred))
```

output :-

Accuracy : 0.933333333333

Precision and recall :-

Precision

Precision measures how many predicted positives are actually positive.

$$\text{precision} = \frac{TP}{TP+FP}$$

* Focus on the correctness of positive predictions.
* High precision: few false positives.

Recall

Recall or sensitivity measures how many of the actual positive cases were correctly identified by the model. It is important when missing a positive case (false negative) is more costly than false positives.

$$\text{Recall} = \frac{TP}{TP+FN}$$

where:

* FN = False Negatives

In scenarios where catching all positive cases is important, recall is a key metric.

4. F1 score :-

The F1 score is the harmonic mean of precision and recall. It is useful when we need a balance between precision and recall as it combines both into a single number. A high F1 score means the model performs well on both metrics. Its range is [0,1].

Lower recall and higher precision gives us great accuracy but then it misses a large number of instances. More the F1 score better will be performance. It can be expressed mathematically in this way:

$$\text{F1 score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Techniques to improve classification accuracy:-

To improve classification accuracy, use data preprocessing such as cleaning, normalization, and feature engineering, and apply advanced techniques like ensemble methods and hyperparameter tuning. Additionally, use cross-validation for more robust model evaluation and ensure you are using a sufficiently large and diverse dataset.

Data preprocessing and feature engineering:-

Data cleaning: Handle missing values, remove duplicates, and corrects or remove outliers to improve data quality.

Feature engineering and selection: Create new, more informative features and select only the most relevant ones to reduce complexity and redundancy.

Data transformation: Apply techniques like normalization or generalization to make the data more suitable for the algorithm.

Data augmentation: For image data, techniques like rotation, scaling, and flipping can increase the size and diversity of the training set.

Data mining can be referred to as knowledge mining from data, knowledge extraction, data/pattern analysis, data archeology, and data dereding.

Techniques to evaluate the accuracy of classifiers:

1. Hold out
2. Random subsampling
3. Cross-validation
4. Bootstrapping

Holdout:-

In the holdout method, the largest dataset is randomly divided into three subsets:

* A training set is a subset of the dataset which are been used to build predictive models.
* The validation set is a subset of the dataset which is been used to assess the performance of the model built in the training phase. It provides a test platform for fine-tuning of the model's parameters and selecting the best-

performing model. It is not necessary for all modelling algorithms to need a validation set.

* Test sets or unseen examples are the subset of the dataset to assess the likely future performance of the model. If a model is fitting into the training set much better than it fit's into the test set, then overfitting is probably the cause that occured here.

Random subsampling :-

* Random subsampling is a variation of the holdout method. The holdout method is been repeated k times.

* The holdout subsampling involves randomly splitting the data into a training set and a test set.

* On the training set the data is been trained and the mean square error (MSE) is been obtained from the predictions on the test set.

* As MSE is dependent on the split, this method is not recommended. So a new split can give you a new MSE.

* The overall accuracy is been calculated as $E = 1/k$

Cross - Validation :-

* k-fold cross-Validation is been used when there is only a limited amount of data available, to achieve an unbiased estimation of the performance of the model.

* Here, we divide the data into k subsets of equal sizes.

* we build models k times, each time leaving out one of these subsets from the training, and use it as the test set.

* If k equals the sample size, then this is called a "Leave-one-out".

Bootstrapping:-

* Bootstrapping is one of the techniques which is used to make the estimations from the data by taking an average of the estimates from smaller data samples.

* The bootstrapping method involves the iterative resampling of a dataset with replacement.

* On resampling instead of only estimating the statistics once on complete data, we can do it many times.

* Repeating this multiple times helps to obtain a vector of estimates.

* Bootstraping can compute variance, expected value, and other relevant statistics of these estimates.

Clustering techniques:-

Clustering techniques in data mining are methods used to group similar data points into clusters based on their attributes and patterns. The main types include:

Agglomerative Hierarchical clustering: A bottom-up approach that treats each data point as a single

cluster initially and merges pairs iteratively based on distances, forming a dendrogram without needed a predefined number of clusters. It is computationally expensive but flexible in cluster shape choice.

K-Means clustering:- A partitioning method that divides data into k clusters by iteratively assinging points to the nearest centroid and recalculating centroids based on cluster members. k must be predetermined, and it works best with spherical clusters.

EM (Expectation-Maximization) Clustering: An advanced method using Gaussian distributions to form clusters. It estimates parameters iteratively to maximize the likelihood of data fitting guassian mixture models, overcoming limitations of k-means for non-circular clusters.

Density-Based clustering: Groups together points that are closely packed while marking points in low-density regions as noise. This method can find arbitarily shaped clusters and is robust to outliers.

Grid based clustering: Divides the data space into a finite number of cells forming a grid structure for efficient clustering, suitable for large datasets.

Model-based clustering: Assumes data is generated by a mixture of underlying probability distributions, such as Gaussian mixture goals.

Constraint-based clustering:- uses user defined constraints, such as must link or cannot link rules, to guide cluster formation.

Clustering analysis:-

Cluster 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 experience. It is useful for:

1. Scalability: It can efficiently handle large volumes of data.
2. High dimensionality: can handle high-dimensional data.
3. Adaptability of different data types: It can work with numerical data like age, salary and categorical data like gender, occupation.
4. Handling Noisy and missing data: Usually, datasets contain missing values or inconsistencies and clustering can manage them easily.
5. Interpretability: Output of clustering is easy to understand and apply in real-world scenarios.

Types of clustering techniques:-

Clustering 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 Methods:-

* Partitioning methods divide the data into k groups where each data points belongs to only one group. These methods are used when many clusters you want to create. A common example is k-means clustering.

* In k-means 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 centers stop changing. It is used in real-life applications like streaming platforms like spotify to group users based on their listening habits.

2. Hierarchical methods:-

Hierarchical 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:

1) Agglomerative Approach (Bottom-up): It starts with indivisual points and merges similar ones. Like a family tree where relatives or grouped step by step.

2) Divisive approach (Top-down): It starts with one big cluster and splits it repeatedly into smaller clusters. For examples, classifying animals into broad categories like mammels, reptiles, etc and further refining them.

3. Density - Based methods:-

* Density-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 behaviours together.

4. Grid-based Methods:-

* Grid-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.

Evaluation of clustering:-

Evaluation of clustering in data mining involves measuring how effectively a clustering algorithm groups similar data points together and separates dissimilar ones, using objective metrics or subjective assessment methods. Typical evaluation techniques fall into two main categories: internal and external measures.

Internal vs External Evaluation:-

* Internal Evaluation: Assesses clustering quality using only data features and the resulting clusters, without reference to known labels. Metrics include:
  1) Sum of square distance (SSD)
  2) Average silhouette coefficient: Quantifies cohesion and separation with clusters.
  3) Calinski-Harabasz index: Assesses compactness and separation.
* External Evaluation: Compares clustering output to a gold standard when available measures include:
  1) Rand Index
  2) Purity
  3) silhouette coefficient

Clustering high dimensional data:-

Clustering of the High-dimensional data return the group of objects which are clusters. It is required to group similar types of objects together

to perform the cluster analysis of high dimensional data, but the high-dimensional data space is huge and it has complex data types and it has complex data types and attributes. A major challenge is that we need to find out the set of attributes that are present in each cluster. A cluster is defined and characterized based on the atributes present in the cluster. Clustering High-dimensional data we need to search for clusters and find out space for the existing clusters.

The high-dimensional data is reduced to low-dimension data to make the clustering and search for clusters simple. Some applications need the appropriate models of clusters, especially the high-dimensional data. Clusters in the high-dimensional data are significantly small. The conventional distance measures can be ineffective.

Subspace clustering methods:

There are 3 subspace clustering methods:

1) Subspace search methods
2) Correlation-based clustering methods
3) Biclustering methods

1. Subspace search methods:

A subspace search method searches the subspaces for clusters. Here, the cluster is a group of similar types of objects in a subspace. The similarity between the clusters is measured by using distance or density

features. CLIQUE algorithm is a subspace clustering method. subspace search methods search a series of subspaces

There are two approaches in subspace search methods:

1) Bottom-up approach
2) Top-down approach

2. Correlation-Based clustering:-

Correlation-based approaches discover the hidden clusters by developing advanced correlation models. Correlation-based models are preferred if is not possible to cluster the objects by using the subspace search methods. Correlation-Based clustering includes the advanced mining techniques for correlation cluster analysis. Biclustering methods are the correlation-based clustering methods in which both the objects and attributes are clustered.

3. Biclustering methods:-

Biclustering means clustering the data based on the two factors. we can cluster both objects and attributes at a time in some applications. The resultant clusters are biclusters. To perform the biclustering there are four requirements:

1) only a small set of objects participate in a cluster.
2) A cluster only involves a small no of attributes.
3) The data objects can take part in multiple clusters, or the objects may also include in any clusters.

Clustering with Constraints :-

Clustering with constraints is a semi-supervised learning approach that incorporates additional domain knowledge into the clustering process through constraints to guide cluster formation. The two common types of constraints are:

* Must - link constraints : specify that two data points must belong to the same cluster.
* cannot - link constraints : specify that two data points must not belong to the same cluster.

Constraint based clustering algorithms use these constraints to produce clusters that better reflect prior knowledge or desired properties, often optimizing an objective function that balances cluster homogeneity and constraint satisfaction. Some algorithms abort if constraints cannot be fully satisfied, while others minimize constraint violations.

There are various methods for handing constraints in clustering, including hard constraints, soft constraints, and approaches generating super- instances for must- link constraints through equivalence relations.

Constraints can also involve cluster properties such as minimum cluster size, expected number of clusters, and constraints on distance or similarity measures tailored to the application context.

outlier analysis- outlier detection methods:-

outliers:-

outliers are data point that is essentially a statistical anamoly, a data point that significantly deviates from other observations in a dataset. outliers can arise due to measurement errors, natural variation, or rare events, and they can a disproportionate impact on statistical analysis and machine learning models if not appropriately handled.

Types of outliers:-

outliers can be classified into various types based on their characteristics:

1. Global outliers: Also known as point anamolies, these data points significantly differ from the rest of the data set.

2. Contextual outliers: These are data points that are considered outliers in a specific context. For example, a high temperature may be common summer but an outlier in winter.

3. Collective outliers:- A collection of data points that deviate significantly from the rest of the dataset, even if indivisual points within the collections are not outliers.

outlier detection:-

Outlier detection is a process of identifying observations or data points that significantly deviate from the majority of data.

Methods for outlier detection:-

Outlier detection is a critical task in data analysis, crucial for ensuring the quality and reliability of conclusions drawn from data. Different techniques are tailored for varying data types and scenarios, ranging from statistical methods for general data sets to specialized algorithms for spatical and temporal data.

1. Standard deviation method:-

Standard deviation method is based on the assumption that the data follows a normal distribution. Data points outside of three standard deviations from the mean are considered outliers. It is commonly used for univariate data analysis where the distribution can be assumed to be approximately normal.

Step 1: Calculate the average and standard deviation of the data set, if applicable.

Step 2: Define the lower and upper bounds for outliers.

Step 3: Identify outliers as data points that fall outside these bounds.

Example:

Dataset: [1, 2, 2, 3, 1, 3, 10].

Find an outlier using the standard deviation method.

Mean, $\mu = \frac{(1+2+2+3+1+3+10)}{7} = \frac{22}{7} \approx 3.14$

standard deviation, $s \approx$

$$\sqrt{\frac{(1-3.14)^2+(2-3.14)^2+(2-3.14)^2+(3-3.14)^2+(1-3.14)^2+(3-3.14)^2+(10-3.14)^2}{7}}$$

≈ 2.90

Lower and upper bounds for outliers:

* Lower bound $= 3.142857 - 2 \times 2.917 = 3.142857 - 5.834 = -2.691$
* Upper bound $= 3.142857 + 2 \times 2.917 = 3.142857 + 5.834 = 8.977$

Any values outside $[-2.691, 8.977]$ is an outlier, thus 10 is an outlier.

2. IQR method:-

The Interquartile Range method focuses on the spread of the middle 50% of data. It calculates the IQR as the difference between the 75th and 25th percentiles of the data and identifies outliers as those points that fall below 1.5 times the IQR below the 25th percentile or above 1.5 times the IQR above the 75th percentile. This method is robust to outliers and does not assume a normal distribution.

Step 1: Find Q_1 (25th percentage) and Q_3 (75th percentage)

Step 2: $IQR = Q_3 - Q_1$

Step 3: Find lower bound : $Q_1 - 1.5 \times IQR$ and upper bound Lower bound : $Q_1 - 1.5 \times IQR$

Example :-

Dataset $X = \{3, 5, 7, 9, 11, 13, 30\}$, find outlier using IQR method.

* Q_1 (25th percentile): Median of first half = Median of $[3, 5, 7] = 5$

* Q_3 (75th percentile): Median of second half = Median of $[11, 13, 30] = 13$

$IQR = Q_3 - Q_1 = 13 - 5 = 8$

Lower bound : $Q_1 - 1.5 \times IQR = 5 - 1.5 \times 8 = 5 - 12 = -7$

Upper bound : $Q_3 + 1.5 \times IQR = 5 + 1.5 \times 8 = 13 + 12 = 25$

Therefore the interval is -7 to 25.

30 lies outside the interval therefore is an outlier.

Z-Score method :-

The Z-score method calculates the no. of standard deviations each data point is from the mean. A Z-score thershold is set, commonly 3, and any data point with a Z-score exceeding this thershold is considered an outlier. This method assumes a normal distribution and is sensitive to extreme values in small datasets.

Step 1: Calculate the mean

Step 2: Compute standard deviation

Step 3: Calculate z-scores

Step 4: Apply thershold rule: mild outlier : $|z|>2$ and extreme outlier: $|z|>3$.

Example:-

$X = \{4, 5, 5, 6, 7, 8, 20\}$ find outlier using the z-score method.

Mean, $\bar{x} = \frac{4+5+5+6+7+8+20}{7} = \frac{55}{7} \approx 7.86$

standard deviation, $s \approx \sqrt{\frac{(4-7.86)^2 + (5-7.86)^2 * 2 + (6-7.86)^2 + (7-7.86)^2 + (8-7.86)^2 + (20-7.86)^2}{6}}$

≈ 5.36

z-scores: $z_i = \frac{x_i - 7.86}{5.36}$

z-score for all data points : 4: -0.72, 5: -0.53, 5: -0.53, 6: -0.35, 7: -0.16, 8: 0.03, 20: 2.26

Thershold $z>2$: = 20 is an outlier since $|2.26|>2$.

Applications:-

1. Financial fraud detection
2. Cyber security
3. AI/ML modelling
4. Anamoly detection in big data & cloud systems
5. e