Module-2

Clustering and Classification

- **Text Clustering**: Feature Selection and Transformation Methods, distance based Clustering Algorithms, Word and Phrase based Clustering, Probabilistic document clustering
- **Text Classification**: Feature Selection, Decision tree Classifiers, Rule-based Classifiers, Probabilistic based Classifiers, Proximity based Classifiers.

Clustering

- Text clustering, also known as document clustering or text categorization, is the process of grouping similar documents together based on their content. It falls under the umbrella of unsupervised learning, where the algorithm automatically discovers patterns and structures in the text data without relying on predefined categories or labels.
- Text clustering aims to organize a collection of text documents into clusters or groups, such that documents within the same cluster are more similar to each other than to those in other clusters.
- Unlike text classification, which assigns predefined labels to documents, text clustering does not require prior knowledge of document categories.

Feature Selection and Transformation method

Bag-of-Words (BoW):

- Bag-of-Words is one of the simplest and most commonly used techniques for feature extraction in text clustering.
- It represents each document as a vector where each dimension corresponds to a unique word in the vocabulary, and the value represents the frequency of that word in the document.
- BoW disregards the order of words in the document and only considers their frequencies, making it computationally efficient but losing some contextual information.

Term Frequency-Inverse Document Frequency (TF-IDF):

- TF-IDF is a statistical measure used to evaluate the importance of a word in a document relative to a corpus.
- It considers not only the frequency of a word in a document (TF) but also the rarity of the word across the corpus (IDF).
- TF-IDF helps in identifying words that are more discriminative and informative for clustering by penalizing common words and emphasizing rare ones.

Cont..

Word Embeddings:

- Word embeddings are dense vector representations of words in a continuous vector space, learned from large text corpora using techniques like Word2Vec, GloVe, or FastText.
- Word embeddings capture semantic similarities between words based on their contextual usage, enabling better representation of word meanings.
- In text clustering, word embeddings can be used to represent documents as vectors of word embeddings, capturing semantic relationships between words and improving clustering accuracy.

Topic Models:

- Topic models such as Latent Dirichlet Allocation (LDA) can be used to extract latent topics from a corpus and represent documents based on these topics.
- LDA assumes that each document is a mixture of topics, and each topic is a distribution over words.
- By representing documents as distributions over topics, topic models can capture the

Cont..

Word Frequency Filters:

- In some cases, it may be beneficial to filter out very frequent or very rare words before clustering to improve the quality of the features.
- Stop words (e.g., "the", "is", "and") are commonly removed as they often carry little semantic meaning.
- Rare words or words that appear in only a few documents may also be filtered out to reduce noise in the data.

Dimensionality Reduction Techniques:

- Text data often have high dimensionality due to the large vocabulary size, which can lead to computational challenges and overfitting.
- Dimensionality reduction techniques such as Principal Component Analysis (PCA) or Singular Value Decomposition (SVD) can be applied to reduce the dimensionality of the feature space while preserving the most important information.
- These techniques help in reducing computational complexity and improving the efficiency of clustering algorithms without sacrificing clustering performance.

Distance-Based Clustering Algorithms

- Distance-based clustering algorithms group data points based on their similarity or dissimilarity, often using distance metrics. Here are some commonly used distance-based clustering algorithms:
- K-Means Clustering
- Hierarchical Clustering
- DBSCAN (Density-Based Spatial Clustering of Applications with Noise)
- Agglomerative Clustering

K-Means Clustering:

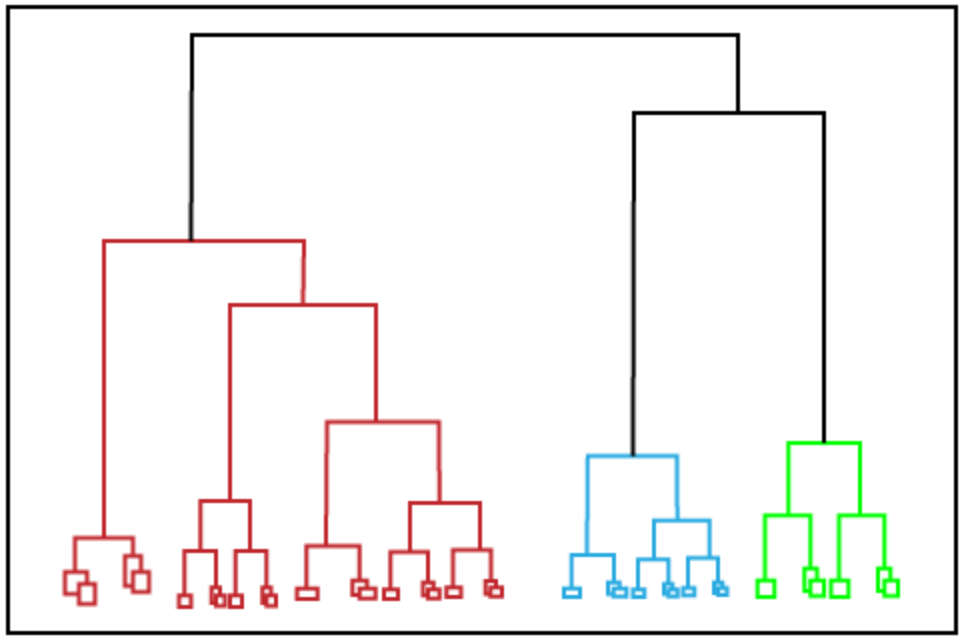
- K-Means aims to partition data into k clusters by minimizing the within-cluster sum of squares. It does so by iteratively assigning each data point to the nearest centroid and updating the centroids based on the mean of the data points assigned to each cluster.
- K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.
- It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabeled dataset on its own without the need for any training.
- It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters.
- The algorithm takes the unlabeled dataset as input, divides the dataset into k-number of clusters, and repeats the process until it does not find the best clusters. The value of k should be predetermined in this algorithm.

Cont..

- Operation:
  - Randomly initializes k centroids.
  - Assigns each data point to the nearest centroid based on a distance metric (typically Euclidean distance).
  - Updates centroids by computing the mean of the data points assigned to each cluster.
  - Iterates the assignment and update steps until convergence criteria are met (e.g., centroids stop moving significantly).
- **Advantages:**
  - Simple and easy to understand.
  - Computationally efficient, especially for large datasets.
  - Scales well to high-dimensional data.
- **Limitations:**
  - Requires specifying the number of clusters (k) beforehand.
  - Sensitive to the initial selection of centroids, which can lead to different solutions.
  - May converge to local optima, especially in the presence of outliers.

Hierarchical Clustering:

- Hierarchical clustering can be used as an alternative for the partitioned clustering as there is no requirement of pre-specifying the number of clusters to be created. In this technique, the dataset is divided into clusters to create a tree-like structure, which is also called a **dendrogram**. The observations or any number of clusters can be selected by cutting the tree at the correct level. The most common example of this method is the **Agglomerative Hierarchical algorithm**.



Cont..

The hierarchical clustering technique has two approaches:

1. Agglomerative: Agglomerative is a bottom-up approach, in which the algorithm starts with taking all data points as single clusters and merging them until one cluster is left.

- Operation**:** Starts with each data point as a single-cluster.
- Iteratively merges the two closest clusters until the desired number of clusters is reached

2. Divisive: Divisive algorithm is the reverse of the agglomerative algorithm as it is a top-down approach.

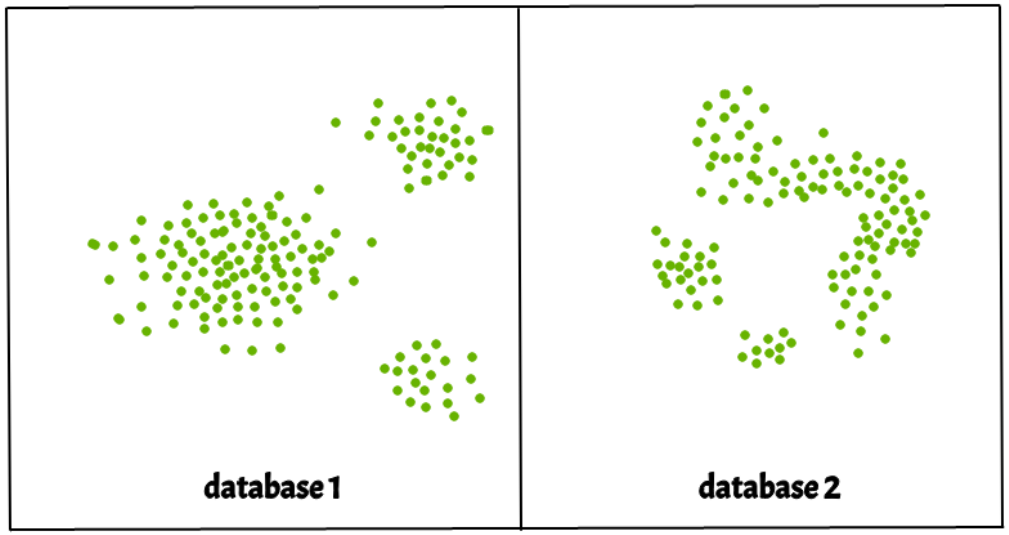
- Operation: Starts with all data points in a single cluster.
- Recursively splits the cluster until each data point is in its own cluster.

Cont..

- Advantages:
  - Does not require specifying the number of clusters beforehand.
  - Provides a hierarchical structure of clusters, allowing for different levels of granularity.
- Limitations:
  - Can be computationally expensive, especially for large datasets.
  - Dendrogram interpretation can be subjective, requiring manual inspection to determine the optimal number of clusters.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

- DBSCAN is a density-based clustering algorithm that groups together points based on density within neighborhoods.
- Clusters are dense regions in the data space, separated by regions of the lower density of points. The ***DBSCAN algorithm*** is based on this intuitive notion of “clusters” and “noise”. The key idea is that for each point of a cluster, the neighborhood of a given radius has to contain at least a minimum number of points.



Cont..

- Partitioning methods (K-means, PAM clustering) and hierarchical clustering work for finding spherical-shaped clusters or convex clusters. In other words, they are suitable only for compact and well-separated clusters. Moreover, they are also severely affected by the presence of noise and outliers in the data.
- Operation: It identifies core points (dense regions), expands clusters from core points, and labels points as noise if they are not in dense regions.
- Advantages: Can discover clusters of arbitrary shapes, robust to noise and outliers, does not require specifying the number of clusters beforehand.
- Limitations: Sensitive to the choice of distance metric and neighborhood size parameters, may struggle with clusters of varying densities

Word and phrase-based clustering techniques

- Word and phrase-based clustering techniques aim to group together similar words or phrases based on their semantic or syntactic similarities. These methods play a crucial role in various natural language processing (NLP) tasks, including document clustering, topic modeling, and semantic analysis.
- Word Embedding Clustering:
- Phrase Extraction
- N-Gram Clustering

Cont..

- Word embeddings : It represent words as dense vectors in a continuous vector space, where the position of each word vector reflects its semantic meaning.
  - Pre-trained word embeddings (e.g., Word2Vec, GloVe) are applied to transform each word in the text corpus into a numerical vector representation.
  - Clustering algorithms such as K-Means, hierarchical clustering, or DBSCAN can then be used to group similar word vectors together.
- Phrases extraction: are meaningful multi-word expressions that convey specific semantic or syntactic information.
  - Phrase extraction techniques aim to identify and extract meaningful phrases from text data using linguistic patterns, syntactic structures, or statistical measures.
  - Once phrases are extracted, clustering algorithms can be applied to group similar phrases together based on their semantic or syntactic similarities.

Cont..

- N-Grams: They are sequences of n consecutive words extracted from text data. They capture local syntactic and semantic relationships within text fragments.
  - N-Grams are extracted from the text corpus using a sliding window approach with a fixed length of n.
  - Clustering algorithms can then be applied to group similar N-Grams together based on their co-occurrence patterns, semantic similarities, or syntactic structures.

Text classification

- Text classification is a natural language processing (NLP) task that involves categorizing text documents into predefined categories or classes based on their content. It's a fundamental technique used in various applications such as sentiment analysis, spam detection, topic labeling, and document organization.
- Text classification aims to automatically assign predefined categories or labels to text documents based on their content.
- It involves training a machine learning model on a labeled dataset, where each document is associated with a known category or class.

Text Classification: Feature selection

- Feature selection in text classification is a crucial step that involves choosing the most relevant features (words, phrases, etc.) from the text data to train a classification model effectively. Feature selection helps improve classification accuracy, reduce computational complexity, and prevent overfitting. feature selection techniques in text classification:
- Term Frequency-Inverse Document Frequency (TF-IDF)
- Chi-Square Test
- Information Gain
- Mutual Information
- Word Embeddings
- Topic Modeling

Decision Tree Classifier

- A Decision Tree Classifier is a supervised machine learning algorithm used for both classification and regression tasks. In classification, it partitions the data into subsets based on the features, aiming to create a tree-like model of decisions.
- It operate based on a series of if-then-else decision rules. The algorithm recursively splits the dataset into subsets based on the values of input features, aiming to minimize impurity or maximize information gain at each step.
- The decision tree splits the feature space into regions or leaves, where each leaf corresponds to a class label in classification tasks.
- Feature Selection: At each node, the algorithm selects the best feature to split the data based on criteria like Gini impurity or information gain.
- Tree Construction: Starting from the root, data is recursively split into subsets based on selected features until stopping criteria are met.
- Tree Pruning: Optional technique to prevent overfitting by removing branches or nodes that don't improve accuracy on unseen data.
- Prediction: Traverses the tree from root to leaf, following decision rules based on feature values, assigning the leaf's class label to the input data point.

Cont..

Advantages:

- Simple to understand and interpret, as the decision rules mimic human decision-making processes. It Can handle both numerical and categorical data.
- Implicitly performs feature selection by selecting the most informative features for splitting.

Limitations:

- Prone to overfitting, especially when the tree is deep and complex.
- May not capture complex relationships in the data, as it makes axis-parallel splits.
- Can be sensitive to small variations in the data, leading to different trees for similar datasets.

Applications:

- Text classification, such as spam detection, sentiment analysis, or document categorization.
- Medical diagnosis, predicting diseases based on symptoms and patient data.
- Customer segmentation and churn prediction in marketing.
- Credit risk assessment and fraud detection in finance.

Rule-based classifiers

- Rule-based classifiers are a type of machine learning model that makes predictions by applying a set of if-then rules to the input data. These rules are typically derived from the training data or specified by domain experts.
- Rule-based classifiers operate on the principle of applying a set of rules sequentially to make predictions about the class label of a given instance.
- Each rule consists of conditions (if) that describe the feature values or patterns in the data and corresponding predictions (then) for the class label.

Cont..

- Rule Generation:
  - Rules can be generated manually by domain experts or automatically from the training data using techniques like association rule mining, decision tree induction, or rule learning algorithms.
- Rule Application:
  - To classify a new instance, the classifier evaluates the input features against each rule's conditions in a sequential manner.
  - When a rule's conditions are satisfied, the corresponding class label prediction is made, and the process stops

Cont..

Advantages:

- Interpretable and transparent, allowing easy understanding of the decision-making process.
- Can incorporate domain knowledge explicitly into the rule formulation.
- Robust to noise and outliers, as rules can be designed to handle specific scenarios.

Limitations:

- Limited expressiveness compared to other models like neural networks or ensemble methods.
- May require extensive domain expertise to define rules accurately.
- Prone to rule conflicts and redundancies, especially in complex datasets.

Applications:

- Rule-based systems are widely used in expert systems, where human expertise is encoded into a set of rules to make decisions or provide recommendations.
- They are also used in fields such as medicine, finance, and law for decision support, diagnosis, risk assessment, and compliance checking.
- Rule-based classifiers can be effective in scenarios where interpretability and transparency are critical, such as regulatory compliance or auditing.

Probabilistic-based Classifiers

- Probabilistic-based classifiers are machine learning models that make predictions by estimating the probability of each class given the input features. These classifiers explicitly model the probability distributions of the classes and use Bayes' theorem to calculate the posterior probabilities.
- Probabilistic classifiers, such as Naive Bayes, Logistic Regression, or Gaussian Naive Bayes, compute the probability of each class given the input features using Bayes' theorem.
- They assume that the features are conditionally independent given the class label, allowing for simplified probability calculations.

1. Probability Estimation:
    1. Probabilistic classifiers estimate the likelihood of observing the input features given each class (likelihood) and the prior probability of each class in the dataset.
    2. They combine these probabilities using Bayes' theorem to calculate the posterior probability of each class given the input features.
2. Decision Making:
    1. To classify a new instance, the classifier selects the class with the highest posterior probability as the predicted class label.
    2. In the case of binary classification, a decision threshold can be applied to convert posterior probabilities into class predictions.

Advantages:

- Provide uncertainty estimates through class probabilities, allowing for more informed decision-making.
- Handle missing data gracefully and robust to irrelevant features.
- Efficient for large-scale datasets and computationally inexpensive compared to some other models.

Limitations:

- Naive Bayes assumes feature independence, which may not hold true in practice and can lead to suboptimal predictions.
- Logistic Regression may struggle with nonlinear relationships between features and classes.
- Sensitive to imbalanced class distributions and may require additional techniques like class weighting or oversampling.

Applications:

- Text classification tasks such as sentiment analysis, spam detection, or document categorization.
- Medical diagnosis and disease prediction based on patient symptoms and diagnostic tests.
- Customer churn prediction and recommendation systems in marketing and e-commerce.
- Fraud detection and credit risk assessment in finance and banking.

Proximity-based Classifiers

- Proximity-based classifiers, such as k-Nearest Neighbors (k-NN) and Support Vector Machines (SVM), classify instances by measuring their proximity or similarity to labeled instances in the training data. These classifiers assign a class label to a new instance based on the majority class labels of its nearest neighbors.
- Proximity-based classifiers make predictions based on the assumption that similar instances tend to belong to the same class.
- They measure the proximity or distance between instances in the feature space and use this information to assign class labels.

operation

1. Nearest Neighbor Search:
    1. In k-NN classification, the algorithm finds the k nearest neighbors of the new instance in the training data based on a distance metric (e.g., Euclidean distance, cosine similarity).
    2. In SVM classification, the algorithm constructs a decision boundary (hyperplane) that maximizes the margin between different classes in the feature space.
2. Majority Voting:
    1. Once the nearest neighbors are identified, the classifier assigns the class label that appears most frequently among the neighbors (for k-NN) or determines the side of the decision boundary on which the new instance lies (for SVM).

CONT..

Advantages:

- Simple and intuitive approach to classification, requiring minimal assumptions about the underlying data distribution.
- Naturally handles multi-class classification problems and does not require explicit probabilistic assumptions.
- Robust to noisy data and outliers, as it focuses on local patterns rather than global distributions.

Limitations:

- Computationally expensive for large datasets, especially for high-dimensional feature spaces.
- Sensitive to the choice of distance metric or kernel function, which may require careful tuning.
- Requires careful selection of hyperparameters (e.g., k in k-NN, kernel parameters in SVM) to achieve optimal performance.

Applications:

- Pattern recognition tasks such as image classification, handwritten digit recognition, and facial recognition.
- Recommender systems for personalized product recommendations based on user preferences and behavior.
- Anomaly detection in network security, identifying unusual patterns or behaviors in network traffic.
- Text clustering and document similarity analysis, grouping similar documents or articles based on their content.

Sample Question

- What is the significance of feature selection in text clustering?
- Explain the concept of rule-based classifiers.
- Compare various distance-based clustering algorithm
- Can you explain the role of distance metrics in distance-based clustering algorithms, and how they impact the clustering results
- word and phrase-based clustering techniques compared to traditional clustering methods?
- Explain Distance-Based Clustering Algorithms in details.
- Elaborate Proximity-based Classifiers.
- Explain Probabilistic-based Classifiers
- Explain Decision tree classifier.