SNS COLLEGE OF TECHNOLOGY

Coimbatore-35
An Autonomous Institution

Accredited by NBA – AICTE and Accredited by NAAC – UGC with 'A+' Grade
Approved by AICTE, New Delhi & Affiliated to Anna University, Chennai

DEPARTMENT OF INFORMATION TECHNOLOGY

Data Mining and Warehousing

COURSE NAME: Data Mining and Warehousing

COURSE CODE: 19ITT301

SEMESTER: 5

CONTENTS:

- Data Mining Systems
- Knowledge Discovery Process
- Data Mining Techniques
- Issues
- Applications
- Data Objects and attribute types
- Statistical description of data
- Preprocessing

Classification in Data Mining:

Classification is a data mining technique used to predict the category or class of data objects based on predefined labels. It is widely used in various applications such as spam detection, credit scoring, medical diagnosis, and more.

1. Support Vector Machines (SVM)

Purpose: Classify data by finding the hyperplane that best separates the different classes.

Method: SVM creates a hyperplane or set of hyperplanes in a high-dimensional space, which can be used for classification, regression, or other tasks.

Example: Classifying emails as spam or not spam based on features like word frequency.

2. Decision Trees

Purpose: Make decisions by splitting the data into subsets based on the value of input features.

Method: A decision tree is a flowchart-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 class label.

Example: Classifying loan applicants as low, medium, or high risk based on attributes such as credit score, income, and employment history.

3. Generalized Linear Models (GLM)

Purpose: Extend linear regression models to allow for response variables that have error distribution models other than a normal distribution.

Method: GLM generalizes linear regression by allowing the linear model to be related to the response variable via a link function and by allowing the magnitude of the variance of each measurement to be a function of its predicted value.

Example: Predicting the probability of a customer defaulting on a loan.

4. K-Nearest Neighbors (KNN) Classifier

Purpose: Classify objects based on the closest training examples in the feature space.

Method: KNN is a non-parametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space, and the output is a class membership.

Example: Predicting the type of flower (e.g., Iris species) based on features like petal length and width.

5. Risk-Based Classification

Purpose: Classify data objects based on the risk associated with different outcomes.

Method: This method involves evaluating the potential risks and benefits associated with different classification decisions and assigning classes accordingly.

Example: Insurance companies classifying policyholders into different risk categories to determine premium rates.

6. Frequent Pattern-Based Classification

Purpose: Classify data based on frequently occurring patterns within the dataset.

Method: This method uses frequent itemset mining techniques to identify patterns that can be used for classification.

Example: Market basket analysis to classify customer buying behavior based on frequently purchased itemsets.

7. Rough Set Theory

Purpose: Handle vagueness and uncertainty in data classification.

Method: Rough set theory classifies objects into approximate sets, where objects that cannot be distinctly classified into a precise set are grouped based on their similarities.

Example: Medical diagnosis where symptoms may not lead to a clear-cut diagnosis but can be grouped into probable disease categories.

8. Fuzzy Logic

Purpose: Handle uncertainty and imprecision in classification problems.

Method: Fuzzy logic uses degrees of membership rather than crisp binary classification, allowing for more flexible and intuitive decision-making processes.

Example: Classifying weather conditions as "sunny," "partly sunny," or "cloudy" based on temperature, humidity, and other factors, allowing for overlapping class boundaries.

Classification in Data Mining