19CS301 DATA MINING TECHNIQUES

source: *https://c1.sfdcstatic.com/content/*

Hours Per Week :

L	T	P	C
3	-	2	4

Total Hours :

L	T	P	CS	WA/RA	SSH	SA	S	BS
45	-	30	5	5	30	20	5	5

PREREQUISITE COURSES: Probability and Statistics, Database Management Systems.

COURSE DESCRIPTION AND OBJECTIVES:

This course introduces the basic concepts, principles, methods, implementation techniques, and applications of data mining, with a focus on three major data mining functions: (1) Association rule mining (2) Classification and (3) cluster Analysis. In the first part of the course student will learn why Association rule mining. In classification student will learn basic concepts of classification and methodologies used for classification. This includes KNN, Naive Bayes, Decision tree and Neural Network based methods. In clustering students will learn different clustering methods. It also focuses on issues relating to the feasibility, usefulness, effectiveness and scalability of techniques for the discovery of patterns hidden in large data sets.

COURSE OUTCOMES:

Upon completion of the course, the student will be able to achieve the following outcomes:

COs	Course Outcomes	POs
1	Investigate various patterns that can be extracted from different types of data.	4
2	Apply various pre-processing techniques and classification algorithms on different domains of data.	1
3	Build decision making systems using data mining algorithms for a given real time data set.	3
4	Construct models using modern tools such as WEKA, R and python etc.	5

SKILLS:

- ✓ *Pre-process the given data.*
- ✓ *Find the correlation among the attributes.*
- ✓ *Apply classification, association rule mining and clustering algorithms on data sets.*
- ✓ *Evaluate the performance of classification and clustering methods.*

UNIT– I **L- 9**

INTRODUCTION: What is data mining?; Why Data mining?; What kinds of data can be mined?; What kinds of patterns can be mined?; Which technologies are used?; What kinds of applications are targeted?; Major issues in data mining; Data objects and attribute types; Basic statistical descriptions of data, Data matrix versus dissimilarity matrix.

UNIT – II **L- 9**

DATA PREPROCESSING: Overview - data quality, major tasks in data preprocessing; Data cleaning - missing values, noisy data; Data Integration - entity identification problem, redundancy and correlation analysis tuple duplication; Data value conflict detection and resolution; Data reduction - PCA, attribute subset selection, regression and log linear models; Histogram; Data transformation - data transformation by normalization; Discretization by binning; Histogram Analysis.

UNIT – III **L- 9**

MINING FREQUENT PATTERNS, ASSOCIATIONS AND CORRELATIONS: Market basket analysis; Frequent Item sets; Closed item sets and association rules; Frequent Item set Mining Methods - apriori algorithm, generating association rules, improving apriori, FP growth method, vertical format method; Which patterns are interesting?; Pattern evaluation method; Pattern Mining in multilevel multidimensional space.

UNIT – IV **L- 9**

CLASSIFICATION BASIC CONCEPTS: What is classification?, General approach to classification, Decision tree induction - attribute selection measures; Tree pruning; Bayes Classification methods - Bayes theorem; Naïve Bayesian classification; Classification by back propagation - a multilayer feed forward neural network; Defining a network topology; Back propagation; K nearest neighbor classifier; Support vector machine, Linearly separable and inseparable cases, Model evaluation and selection; Techniques to improve classification accuracy; Other classification methods - KNN; generic algorithms; Fuzzy algorithm.

UNIT - V **L- 9**

CLUSTER ANALYSIS: Partition methods - K means and K medoid; Hierarchical methods; Agglomerative and divisive method; Density based methods - DBSCAN; Optics; Grid based methods- STING; Cluster evaluation methods; Clustering high dimensional data; Problems, Challenges and major methodologies.

LABORATORY EXPERIMENTS

LIST OF EXPERIMENTS **TOTAL HOURS: 30**

The Students pursue the following experiments by using the open source analytical tools such as R, Python, Weka, Rapid Miner etc.

Students experiment the following on UCI/ Kaggle/ NCBI data repository.

1. Apply the following data pre-processing techniques on a given dataset to illustrate the need of the pre-processing in data mining.
   a) Data Cleaning
   b) Data Normalization
   c) Data Discretization
   d) Computation of correlation coefficient to analyze the data behaviour
   e) Dimensionality reduction using PCA and Wavelets
2. Evaluate the need of feature selection on a given dataset using Information Gain as a metric.
3. Write a program to extract the interesting association rules from a given dataset using Apriori and Frequent Pattern growth algorithms.
4. Apply the following classifiers on a given dataset and analyze their performance.
   a) J48 and visualize the decision tree
   b) Naive Bayes
   c) Support Vector Machine
   d) Multi Layer Perceptron
5. Evaluate the performance of partitioning and hierarchical based clustering algorithms on a given dataset.

TEXT BOOK:

1. Jiawei Han, Micheline Kamber and Jian Pei, “Data mining Concepts and Techniques”, 3rd edition, Morgan Kaufmann. 2012.

REFERENCE BOOKS:

1. Pang-Ning Tan, Michael Steinbach and Vipin Kumar, “Introduction to Data Mining”,2nd edition, Pearson, 2018
2. Jure Leskovec, Anand R aja raman and Jeffrey D Ullman, “Mining of Massive Datasets”, 5th edition, Stanford University, 2014.