BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE-PILANI, K.K.BIRLA GOA CAMPUS
SECOND SEMESTER 2013-14
COURSE HANDOUT

Course No. : **CS C415, CS F415,IS C415**
Course Title : **Data Mining**
Instructor-in-charge : **G.Aruna(garuna@goa.bits-pilani.ac.in) Chamber: A-406**

---

1. Objective and Scope

The course explores the concepts and techniques of data mining, a promising and flourishing frontier in database systems. Data Mining is automated extraction of patterns representing knowledge implicitly stored in large databases, data warehouses, and other massive information repositories. It is a decision support tool that addresses unique decision support problems that cannot be solved by other data analysis tools such as Online Analytical Processing (OLAP). The course covers data mining tasks like constructing decision trees, finding association rules, classification, and clustering. The course is designed to provide students with a broad understanding in the design and use of data mining algorithms. The course also aims at providing a holistic view of data mining. It will have database, statistical, algorithmic and application perspectives of data mining.

2. Text Book

a.i) Pang-Ning Tan, Micheal Steinbach, Vipin Kumar, "***Introduction to Data Mining",*** Pearson, 2009.

3. Reference Books

a.i) Han J & Kamber M, "***Data Mining: Concepts and Techniques",*** Morgan Kaufmann Publishers, 2001
a.ii) Hand D, Mannila H, & Smyth P, "***Principles of Data Mining***", MIT Press, 2001.
a.iii) Pujari A K, "***Data Mining Techniques***", University Press (India), 2001.
a.iv) Kimball R, ***"The Data Warehouse Toolkit",*** 2e, John Wiley, 2002.

4. Course Plan

Lecture No.	Learning Objective	Topic(s)	Chapter No.
1-2	To understand the definition and applications of Data Mining	**Introduction to Data Mining** • What is Data mining • Motivation & challenges • Origins of Data Mining • Data Mining Tasks	1
3-5	To understand the Data & Preprocessing of data in Data Mining	**Data** • Types of Data • Data quality • Data Preprocessing • Measures of Similarity & Dissimilarity	2
6-8	To understand the role of Data Exploration	**Exploring Data** • Some revision of basic statistical concepts • Visualization • OLAP and Multidimensional Data Analysis	3 (self study)
9-13	To understand Classification, Basic concepts, Decision trees & Model evaluation	**Classification** • Basics • General approach to solving a classification problem • Decision Tree Introduction • Model overfitting • Evaluating the performance of a classifier • Methods of comparing classifiers	4

14-16	To understand alternative techniques in classification	**Classification: Alternative Techniques** • Rule based classifiers • Nearest-neighbor classifiers • Bayesian Classifiers • Artificial Neural Network • Support vector machines • Ensemble methods • Class imbalance problem • Multiclass problem	5+Class Notes
17-22	To understand application and algorithms for association	**Association Analysis: Basic concepts and Algorithms** • Problem definition • Frequent itemset generation • Rule generation • Compact representation and frequent itemsets • Alternative methods for frequent itemsets • FP-Growth algorithm • Evaluation of Association Patterns • Effect of skewed Support Distribution	6+Class Notes
23-28	To understand advanced algorithms in Association analysis	**Association Analysis: Advance concepts** • Handling categorical attributes • Handling continuous attributes • Handling a concept hierarchy • Sequential Patterns • Subgraph Patterns • Infrequent Patterns	7
29-34	To introduce topics in clustering	**Cluster Analysis: Basic concepts and algorithms** • Overview • K-Means • Agglomerative hierarchical clustering • DBSCAN • Cluster evaluation	8
35-37	To introduce advanced topics in clustering	**Cluster Analysis: Additional Issues and Algorithms** • Characteristics of Data, Clusters and Clustering Algorithms • Prototype based clustering • Density based Clustering • Graph based Clustering • Scalable Clustering Algorithms • Choosing clustering technique	9 +Class Notes
38-40	To understand anomalies	**Anomaly Detection** • Preliminaries • Statistical Approaches • Proximity based outlier detection • Density based outlier detection • Clustering based Techniques	10

5. **Evaluation Schedule**

Component	Weigh tage(%)	Remarks	Date & Time
Test-I	20	Open Book / Closed Book	18.09.13, 8:30-9:30
Test-II	20	Open Book / Closed Book	28.10.13, 8:30-9:30
Project /Presentation	30	Open Book	Periodically
Comprehensive	30	Open Book / Closed Book	04-12-2013(FN)

6. **Project**
   The students are expected to work on HADOOP and implement any one of the Data Mining Techniques. The list of the techniques/algorithms will be displayed on moodle. The students has to come up with atleast 3 members in a team.

7. **Presentation**
   The students who opted for the presentation are expected to study two published papers from international journals in the specific topics , those will be mentioned on moodle.. And they have to give two seminars explaining those two papers and should come up with their own conclusions and observations.

8. **Make-up Policy:** Prior Permission is must and Make-up shall be granted only in genuine cases.

9. **Chamber Consulting Hours :** Every Tuesday 4pm-5pm .

Instructor-in-charge