Data Mining

The workshop is practical in nature. All contents are presented in the context of using WEKA and other tools, the most widely-used open source machine learning software in the world.

The workshop includes hands-on exercises using WEKA and other tools and data sets to do actual data mining. The course assumes no prior knowledge; but it is not a trivial treatment. It covers the basics and advances of machine learning, and surveys a majority of the most commonly used algorithms.
This is a practical course with design exercises. Participants are required to bring their laptops to the course.

This workshop:
o Explains how data mining algorithms work;
o Helps you select appropriate approaches to particular problems; how to compare and evaluate the results of different techniques;
o Shows you how to use the WEKA machine learning workbench.
o Covers performance improvement techniques, including input preprocessing and combining output from different methods.

Course Content:

1. Introduction to Data Mining

o What is data mining?
o Related technologies - Machine Learning, DBMS, OLAP, Statistics
o Data Mining Goals
o Stages of the Data Mining Process
o Data Mining Techniques
o Knowledge Representation Methods
o Applications
o Example: weather data

2. Data Warehouse and OLAP

o Data Warehouse and DBMS

o Multidimensional data model

o OLAP operations

o Example: loan data set

3. Data preprocessing

o Data cleaning

o Data transformation

o Data reduction

o Discretization and generating concept hierarchies

o Installing Weka 3 Data Mining System

o Experiments with Weka - filters, discretization

4. Data mining knowledge representation

o Task relevant data

o Background knowledge

o Interestingness measures

o Representing input data and output knowledge

o Visualization techniques

o Experiments with Weka - visualization

5. Attribute-oriented analysis

o Attribute generalization

o Attribute relevance

o Class comparison

o Statistical measures

o Experiments with Weka - using filters and statistics

6. Data mining algorithms: Association rules

o Motivation and terminology

o Example: mining weather data

o Basic idea: item sets

o Generating item sets and rules efficiently

o Correlation analysis

o Experiments with Weka - mining association rules

7. Data mining algorithms: Classification

o Basic learning/mining tasks

o Inferring rudimentary rules: 1R algorithm

o Decision trees

o Covering rules

o Experiments with Weka - decision trees, rules

8. Data mining algorithms: Prediction

o The prediction task

o Statistical (Bayesian) classification

o Bayesian networks

o Instance-based methods (nearest neighbor)

o Linear models

o Experiments with Weka - Prediction

9. Evaluating what\'s been learned

o Basic issues

o Training and testing

o Estimating classifier accuracy (holdout, cross-validation, leave-one-out)

o Combining multiple models (bagging, boosting, stacking)

o Minimum Description Length Principle (MLD)

o Experiments with Weka - training and testing

10. Mining real data

o Preprocessing data from a real medical domain (310 patients with Hepatitis C).

o Applying various data mining techniques to create a comprehensive and accurate model of the data.

11. Clustering

o Basic issues in clustering

o First conceptual clustering system: Cluster/2

o Partitioning methods: k-means, expectation maximization (EM)

o Hierarchical methods: distance-based agglomerative and divisible clustering

o Conceptual clustering: Cobweb

o Experiments with Weka - k-means, EM, Cobweb

12. Advanced techniques, Data Mining software and applications

o Text mining: extracting attributes (keywords), structural approaches (parsing, soft parsing).

o Bayesian approach to classifying text

o Web mining: classifying web pages, extracting knowledge from the web

o Data Mining software and applications.

Duration: The duration of this workshop will be two consecutive days, with eight hour session each day in a total of sixteen hours properly divided into theory and hands on sessions.

Fees: Rs. 1200/- per head inclusive of all taxes.