Basic Data Mining Algorithms and their Scalability for Big Data
August 16-21, 2016

Schedule:

Day 1: August 16, 2016	
9:30am - 10:15am	Registration
10:15am -10:55am	Inaugural function followed by tea
11:00am -1:00pm	Lecture 1 (Introduction to Data Mining)
2:30pm - 4:30pm	Lab and Practice 1
4:45pm -6:45pm	Lecture 2 (Classification Algorithms: Decision Trees)
Day 2: August 17, 2016	
9:00am -10:00am	Lecture 2 (Classification Algorithm: Decision Trees)
10:00am - 11:00am	Lecture 3 (Association Analysis)
11:15am -1:15pm	Lecture 3 (Association Analysis)
2:30pm - 5:30pm	Lab and Practice 2
Day 3: August 18, 2016	
9:00am -12:00noon	Lab and Practice 3
12:15pm -1:15pm	Lecture 4 (Basic Clustering Algorithms)
2:30pm - 4:30pm	Lecture 4 (Basic Clustering Algorithms)
4:30pm - 5:00pm	Evaluation Session
Day 4: August 19, 2016	
9:00am -11:00am	Lecture 5 (Scalability of Algorithms for Big Data)
11:15am - 1:15pm	Lecture 5 (Scalability of Algorithms for Big Data)
2:30pm - 5:30pm	Lab and Practice 4
5:30pm - 7:00pm	Presentations by participants
Day 5: August 20, 2016	
9:00am - 1:00pm	Lab and Practice 5 (Designing algorithms using MapReduce Paradigm)
2:30pm - 6:30pm	R Programming
Day 6: August 21, 2016	
9:00am-1:00pm	R Programming
2:30pm -3:00pm	Valedictory Function

Detailed Course Contents:

1. Introduction to Data Mining (Total 4 hours)
    a. Lecture (2 hours)
        i. Applications and Need for data mining algorithms
        ii. Scalability issues for data mining tasks
        iii. Relationship to other fields such as statistics and machine learning
        iv. Different types of data: Relations, Graphs, Sequences, and Text
        v. Types and nature of patterns and knowledge to be discovered in data
    b. Lab and Practice (2 hours)
        i. Practice with representation and processing of data in MATLAB
2. Classification Algorithms: Decision Trees (6 hours)
    a. Lecture (3 hours)
        i. What is a Decision Tree: How does it work
        ii. Algorithms for inducing decision trees from data
        iii. Characteristics of decision tree induction algorithms
        iv. Overfitting and underfitting
        v. Evaluating the performance of a decision tree
        vi. Applications and Real life cases
        vii. Learning of tree ensembles
        viii. Induction of Decision Trees for Big Data: Issues of performance and algorithms
    b. Lab and Practice (3 hours)
        i. Build decision trees from test datasets using MATLAB functions
3. Association Analysis (6 hours)
    a. Lecture (3 hours)
        i. What are association rules
        ii. Apriori principle for frequent itemset generation
        iii. Association Rule Generation
        iv. Support, confidence, lift etc. metrics
    b. Lab and Practice (3 hours)
        i. MATLAB functions to generate association rules: test and practice
4. Basic Clustering Algorithms (6 hours)
    a. Lecture (3 hours)
        i. Why clustering?
        ii. Sequential Clustering Algorithms
        iii. Partitional Clustering Algorithms: K-means, bisecting k-means
        iv. Evaluating performance of clustering algorithms
    b. Exercise and Practices (3 hours)
        i. Exercises with clustering algorithms
5. Scalability of Algorithms for Big Data (8 hours)
    a. Lecture (4 hours)
        i. Types of Hardware for scaling: Scaling Up vs. Scaling Out
        ii. Hadoop Architecture and map-Reduce Algorithms
        iii. Foundational ideas of Map-Reduce Algorithms
        iv. Simple statistical Functions using MapReduce Formulations
    b. Lab and Practice (4 hours)
        i. Exercises in designing algorithms using MapReduce Paradigm
6. R Programming (8 hours)
    a. Introduction to R, Importing and cleaning data, Regression (4 hours)
    b. Decision tree and Neural network using R (4 hours)