INDUSTRIAL AND MANAGEMENT ENGINEERING, IIT KANPUR

IME 672: Data Mining and Knowledge Discovery (3-0-0-0-9)

January 10, 2018

Course objectives:

The course introduces the fundamental approaches to knowledge discovery and data mining, and its main theoretical foundations. It will present important algorithms for data preprocessing, classification, clustering, association rule mining, and also show how to use these techniques on real problems. There is significant emphasize to interpret the result/knowledge obtained from these algorithms

Course contents:

- Introduction to machine learning, data mining, knowledge discovery, big data, data mining application examples, data mining tasks
- Data preparation for knowledge discovery - data understanding, data cleaning, data transformation, discretization, feature reduction, learning with unbalanced data
- Classification: decision trees, choosing the splitting attribute, information gain and gain ratio, handling numeric attributes (finding best split), dealing with missing values, pruning (pre-pruning, post-pruning, estimating error rates), from trees to rules
- Classification: naive Bayes classifier, neural networks, support vector machines
- Evaluation and Credibility: classification with train, test, and validation sets (handling unbalanced data), parameter tuning, predicting performance, evaluation on "small data": cross-validation, bootstrap, comparing data mining schemes
- Clustering: introduction, partitioned, hierarchical, density based
- Associations Rule Mining/ Market Basket Analysis: transactions, frequent itemsets, association rules, apriori algorithm, applications

Instructor: Dr. Faiz Hamid (fhamid@iitk.ac.in)

Class Room: C3, IME Building

Time: Tue, Wed (12:00 - 13:15)

Course Organization: All notices for the course will be sent by email to the course email list.

Home Assignments:

At the end of every chapter or week, home assignments will be given. The students are strongly advised to solve and master the material of the home assignment, submission is optional.

Exams and Quizzes:

- One mid-semester examination of two hours (weight: 35%)
- One end-semester examination of three hours (weight: 35%)
- Project work (weight: 20%)
- Assignment/quiz/presentation (weight: 10%)

Attendance:

It goes without saying that 100% attendance is compulsory. Any student who is granted leave by the Convener, DPGC/DUGC also must inform the instructor regarding his/her absence.

Recommended Books:

This being a PG course there is no prescribed text. However, the following books are recommended:

- Jiawei Han, Micheline Kamber & Jian Pei. Data Mining: Concepts and Techniques . Cengage Learning
- P. N. Tan, M. Steinbach & V. Kumar. Introduction to Data Mining. Pearson
- Jure Leskovec, Anand Rajaraman, Jeffrey David Ullman. Mining of Massive Datasets. Dreamtech Press
- Norman Matloff. The Art of R Programming
- Jared P. Lander. R for Everyone. Pearson

Faiz Hamid January 10, 2018