CS 660 Data Mining for Decision Making

Credits: 3-0-1-4 **Approval:** Approved in 8th Senate

Prerequisites: IC 210: Probability, Statistics and Random Processes; IC 250: Data Structure and Algorithms

Intended for: UG/PG **Semester:** odd/even

Distribution: Discipline elective for CSE; CS elective for EE and ME

Course Preamble: In today's world, there is a rapid growth in data. Increasing amounts of data could be captured via the Internet, websites, point-of-sale devices, bar-code readers etc. Such data has tremendous relevance for managerial decisions. How could one find patterns in large amounts of collected data? This course titled, "Data Mining for Decision Making," involves learning a collection of techniques for extracting patterns and trends in large amounts of data. This course will provide a hands-on introduction to the data-mining area with an emphasis on aspects useful to business and management. Being built upon topics from artificial intelligence and statistical analyses, this course would form a good addition to the minor on Intelligent Systems at IIT Mandi.

Course Outline: The course will cover a number of algorithmic techniques like Naïve Bayes classifiers, decision trees, neural networks, clustering, logistic regression, multiple-linear regression, principal components analysis, discriminant analysis, and association rules (market basket analysis). Furthermore, this course will help students better understand the need and appropriate place for data mining, the major techniques used in data mining, and the important pitfalls to watch out for in this area. Each week, three 1-hour lectures will cover theoretical concepts and techniques. Furthermore, each week, a 1-hour tutorial will provide hands-on practice on the taught techniques.

Modules:

The course is divided into weekly modules, where a new topic is covered in each week. The details of the topics covered in each week are provided below:

Week 1: Introduction to Data mining [3 Lectures]

What is Data Mining? What is the Data Mining Process? Basic Data Mining Tasks, Problem Identification, Data Mining Metrics, Data Cleaning (pre-processing, feature selection, data reduction, feature encoding, noise and missing values, etc.), Key Issues, Opportunities for Data Mining.

Week 2: Naïve Bayes classifier [3 Lectures]

Two-class classifiers, Training and Test sets, Maximum-Likelihood estimation, Bayesian estimation, Classification of Test sets.

Week 3: Decision Trees [3 Lectures]

Classification and Regression Trees, Building and Selecting Decision Trees (concept of Information Gain), Obtaining Production Rules from Decision Trees, Handling missing values in Decision Trees.

Week 4: Neural Networks [3 Lectures]

Introduction to Artificial Neural Networks, Single-layer Networks, Multi-layer Networks, Backward Propagation Algorithm, Annealing the learning rate (Step decay, Adagrad, and RMSprop), Over-fitting and choice of Epochs.

Week 5: Instance-Based Learning [3 Lectures]

Instances, Activations, Recency, Frequency, Retrieval from Memory, Blending of instances.

Week 6: Clustering[3 Lectures]

Introduction to Cluster Analysis, Clustering Algorithms, Hierarchical Methods (Nearest neighbor, Farthest neighbor, Group average), Similarity Measures.

Week 7: Logistic Regression [3 Lectures]

Introduction to Logistic Regression, Logistic function, odds ratio, logit, Logistic Regression with more than two classes

Week 8: Multiple-Linear Regression [3 Lectures]

Introduction to Multiple-Linear Regression, Assumption made in a linear regression model, regression process, dropping irrelevant variables.

Week 9: Principal Components Analysis [3 Lectures]

Introduction to principal components analysis, dimensionality reduction, principal components and orthogonal least squares.

Week 10: Discriminant Analysis [3 Lectures]

Introduction to discriminant analysis, applications to two-classes, extension to more than 2-classes, canonical variate loadings, extension to unequal covariance structures.

Week 11: Association Rules [3 Lectures]

Introduction to association rules, support, confidence, Apriori Algorithm.

Week 13: Implementation Issues [3 Lectures]

Metrics for Model selection - MDL, BIC, AIC, Ethics, Legality, and Privacy; Staffing and Implementation

Week 14: The Future of Data Mining, Unstructured Data Mining, and conclusions [3 Lectures]

If time permits:

Topics in graph mining: Definition of Graphs, Subgraphs, Frequent Subgraphs and subgraphs, Detection Algorithms: Apriori-Based Approach, Pattern Growth Approach (gSpan), Graph Classification, and Graph Compression.

Textbooks:

1. Hand David, MannilaHeikki, and Smyth Padhraic. Principles of Data Mining. Boston, MA: MIT, 2004. ISBN: 8120324579

References:

1. Han, J., Kamber, M. & Pei, J. (2012). Data mining concepts and techniques, third edition Morgan Kaufmann Publishers
2. Berry and Linoff. Mastering Data Mining. New York, NY: Wiley, 2000. ISBN: 0471331236.
3. Delmater and Hancock. Data Mining Explained. New York, NY: Digital Press, 2001. ISBN: 1555582311.
4. T. Mitchell. Machine Learning. New York, NY: McGraw-Hill, 1997.
5. M. H. Dunham. Data Mining: Introductory and Advanced Topics. Pearson Education. 2001.
6. Samatova, N. F., Hendrix, W., Jenkins, J., Padmanabhan, K., & Chakraborty, A. (Eds.). (2013). *Practical Graph Mining with R*. CRC Press.
7. Wang, H. (2010). Managing and mining graph data (Vol. 40). C. C. Aggarwal (Ed.). New York, Springer.