

**SECOND SEMESTER 2019-2020**

# Course Handout Part II

Date: Dec-2019

In addition to part-I (General Handout for all courses appended to the time table) this portion gives further specific details regarding the course.

*Course No.* : CS G520/SS G520

## Course Title : **Advanced Data Mining**

## Instructor-in-Charge : **Jabez Christopher**

**Course Description &Scope:**

Advanced Data Mining is a specialization course of Data Mining. Topics covered go beyond conventional record data mining to mining complex data structures and complex data: Tree/graph, biological data, web/text data, stream data, spatiotemporal data, time series data, high-dimensional data. A substantial portion of the course will focus on research projects, where students will study and work on a well-defined research problem. The course also deals with applications such as mining social networking sites, bioinformatics and medical informatics.

**Objectives of the Course:**

* To expose key research areas in data mining.
* Emphasize on the design and implementation of efficient and optimized algorithms for data mining.
* Emphasize the use of WEKA, MATLAB and R to implement Data preprocessing and Data Mining tasks.
* To improve research and presentation quality thereby enable students to comprehend and critically analyze data mining research.

**Text Book**

TB: Aggarwal, Charu C. **Data mining: the textbook**. Springer, 2015.

**Reference books**

1. R1: Hadzic F., Tan H. & Dillon T. S. “***Mining data with Complex Structures***” Springer, 2011
2. R2:Han J. &Kamber M., “***Data Mining: Concepts and Techniques”,*** Morgan Kaufmann Publishers, Second Edition, 2006
3. R3: Tan P. N., Steinbach M & Kumar V. “***Introduction to Data Mining”*** Pearson Education, 2006
4. R4: Mitsa, Theophano**. Temporal data mining**. Chapman and Hall/CRC, 2010.
5. R5:Chowriappa, Pradeep, and SumeetDua. **Data mining for bioinformatics**. CRC Press, 2012.
6. R6:Dunham, Margaret H**. Data mining: Introductory and advanced topics**. Pearson Education India, 2006.
7. R7: Ross, T. J. (2005). **Fuzzy logic with engineering applications**. John Wiley & Sons.

**Reading Material:** Research papers and other reading material will be provided on the course website.

[RM 01] Bradley, Paul S., Usama M. Fayyad, and Olvi L. Mangasarian. "Mathematical programming for data mining: Formulations and challenges." INFORMS Journal on Computing 11, no. 3 (1999): 217-238.

[RM 02] Mangasarian, Olvi L. "Mathematical programming in data mining." Data mining and knowledge discovery 1, no. 2 (1997): 183-201

**Course Plan:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Lecture No.** | **Learning objectives** | **Topics to be covered** | **Reference** |
| 1 | To understand the objectives of the  course | Introduction to the course & Lab Topics | --- |
| 2,3 | Refresh the basics of KDD and Data Mining tasks. | Overview of Knowledge Discovery   * Data Integration * Data Preprocessing * Data Mining | R2 Ch 1, 2, 3;  TB Ch 1 [RM 01] |
| 4-8  9,10 | Overview of Data Mining   * Association Analysis * Classification * Clustering   Statistics for Data Mining | R2 Ch 6, 8, 10;  [RM 02]  Class Notes |
|  | **WEB DATA MINING** | | |
| 11-15 | Concentrate on use of Data Mining techniques to Web and Social Networks, and understand their their use in Web analytics, user profiling and personalization. | Social Network Analysis: Preliminaries and Properties   * Homophily * Triadic Closure * Dynamics of Network Formation * Power-Law Degree Distributions * Measures of Centrality and Prestige   Community Discovery   * Kernighan–Lin Algorithm * Girvan–Newman Algorithm * Multilevel Graph Partitioning: METIS * Spectral Clustering | TB Ch. 19 |
|  | **BIOLOGICAL DATA MINING** | | |
| 16-20 | Enable students to find patterns and statistical dependencies in large biological databases and to gain an understanding of the underlying system from which the data were obtained.  Use computational biology & data mining tasks to analyze vast experimental data generated by high-throughput technologies, and thereby enables the generation of new hypotheses. | **Data Transformations**   * Normalization Techniques for Gene Expression Analysis. * Data Preprocessing of Mass Spectrometry Data * Data Preprocessing for Genomic Sequence Data * Ontologies in Bioinformatics   **Applications of Classification & Clustering in Bioinformatics**.   * Distance Metric in Gene Expressions for Coexpressed Genes * Gene Expression Clustering Using Mutual Information Distance Measure * Gene Expression Data Clustering Using aLocal Shape-Based Clustering * Fuzzy k-Means Clustering on Gene Expression   **Temporal Data Mining in Medicine and Bioinformatics** | **R5 Ch. 5**  **R5 Ch. 6** |
|  | **SPATIAL &TEMPORAL DATA MINING** | | |
| 21-25 | To study how to investigate temporal data and understand models and methods for representation. | **Mining with Contextual Spatial Attributes.**  **Trajectory Mining**  **Temporal Data Similarity Computation,**  **Representation, and Summarization**   * Temporal Data Types and Preprocessing * Temporal Data Preprocessing * Time Series Similarity Measures * Time Series Summarization Methods * Temporal Event Representation * Temporal Knowledge Representation   in Case-Based Reasoning Systems | TB Ch.16  R4 Ch. 2 |
| 26-28 | **Forecasting Model and Error Measures**  **Event Prediction**  **Time Series Forecasting**   * Moving Averages * Exponential Smoothing * Time Series Forecasting via Regression * Forecasting Seasonal Data via Regression * Random Walk * Autocorrelation * Autoregression * ARIMA Models | R4 Ch. 4  TB Ch. 14 |
|  | **RECENT TRENDS** | | |
| 29-36 | Understand the use of Optimization techniques to catalyze the performance of Algorithms and parameter-tuning. | **Optimization Algorithms for Data Mining**  Lagrangian Methods  Gradient Descent Algorithm  Simulated Annealing  Evolutionary Algorithms (GA)  Swarm Intelligence Algorithms   * Particle Swarm Optimization | Research articles (to be decided) |
| 37-40 | Understand how to handle uncertainty in decision-making systems by incorporating fuzzy logic and FIS | **Soft Computing Approaches for Data Mining**  Handling Uncertainty in Data  Resolving Uncertainty using Fuzzy Logic  Fuzzy Inference Systems  Evolving Fuzzy Systems | R7 Ch 1, 4, Class notes&  Research articles (to be decided) |
| 41, 42 | --- | **RESEARCH GAPS & FUTURE DIRECTIONS** | ­­­--- |

**Evaluation Scheme:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Component** | **Duration** | **Weightage (%)** | **Date & Time** | **Nature of Component** |
| Mid-Semester Test | 1½ hours | 20 | 6/3 - 9.00 - 10.30AM | Closed |
| Quizzes (2) | ½ hour | 5 |  | Closed |
| Labs Assignments& Projects\* | -- | 30 |  | Open |
| Term Paper & Seminar | -- | 10 |  | Open |
| Comprehensive Exam | 3 Hours | 35 | 12/05 - FN | Closed |

\*Lab-projects will emphasize on the use of WEKA/MATLAB/R to implement preprocessing and data mining algorithms on datasets from open repositories. In addition to this the students are expected to survey, review & compare research works in similar areas and prepare articles and reports. Evaluation would be a continuous assessment model during the practical sessions.

**Chamber Consultation Hours:**To be announced in the class.

**Make-up Policy:**

**Notices:** All the notices concerning this course will be displayed on the CMS.

**Academic Honesty and Integrity Policy**: Academic honesty and integrity are to be maintained by all the students throughout the semester and no type of academic dishonesty is acceptable.

.

**INSTRUCTOR-IN-CHARGE**