**CSC-389 –Introduction To Data Mining and Data Warehousing**

**General Information**

| **Course Number** | **CSC-389 –Introduction To Data Mining and Data Warehousing** |
| --- | --- |
| **Credit Hours** | 3 (Theory Credit Hour = 3, Lab Credit Hour =0) |
| **Prerequisite** | CSC-252, Database Management Systems (DBMS) |
| **Course Coordinator** |  |

**Course Objectives**

| | Data mining refers to extracting or "mining" knowledge from large amount of data. Data mining has evolved from several areas including: databases, machine learning, algorithms, information retrieval, and statistics. Data warehousing involves data preprocessing, data integration, and providing on-line analytical processing (OLAP) tools for the interactive analysis of multidimensional data, which facilitates effective data mining. This course introduces data warehousing and data mining techniques and their tools.  Topics include: Introduction to Data Mining, Data Preprocessing, Data Warehouse and OLAP technology, Mining frequent patterns, Classification & Prediction and Cluster analysis. | | --- | |
| --- | --- |

**Catalog Description**

| **CSC-389** |
| --- |

**Course Content**

| **Session No.** | **Week No.** | **Topics** | **Suggested Readings**  **(Chapters)** |
| --- | --- | --- | --- |
| 01-04 | 1 | **Chapter 1.0 Introduction**  1.1 What Is Data Mining?  1.2 Motivating Challenges  1.3 The Origins of Data Mining  1.4 Data Mining Tasks | Jiawei: Chapter 01 |
| 05-12 | 2-3 | **Chapter 2 Data**  2.1 Types of Data  2.1.1 Attributes and Measurement  2.1.2 Types of Data Sets  2.2 Data Quality  2.2.1 Measurement and Data Collection Issues  2.2.2 Issues Related to Applications  2.3 Data Preprocessing  2.3.1 Aggregation  2.3.2 Sampling  2.3.3 Dimensionality Reduction  2.3.4 Feature Subset Selection  2.3.5 Feature Creation  2.3.6 Discretization and Binarization  2.3:7 Variable Transformation  2.4 Measures of Similarity and Dissimilarity  2.4.1 Basics  2.4.2 Similarity and Dissimilarity between Simple Attributes  2.4.3 Dissimilarities between Data Objects  2.4.4 Similarities between Data Objects  2.4.5 Examples of Proximity Measures  2.4.6 Issues in Proximity Calculation  2.4.7 Selecting the Right Proximity Measure | Jiawei: Chapter 02 |
| 13-20 | 4-5 | **Chapter 3 Exploring Data**  3.1 The Iris Data Set  3.2 Summary Statistics  3.2.1 Frequencies and the Mode  3.2.2 Percentiles  3.2.3 Measures of Location: Mean and Median  3.2.4 Measures of Spread: Range and Variance  3.2.5 Multivariate Summary Statistics  3.2.6 Other Ways to Summarize the Data  3.3 Visualization  3.3.1 Motivations for Visualization  3.3.2 General Concepts  3.3.3 Techniques  3.3.4 Visualizing Higher-Dimensional Data  3.3.5 Do's and Don'ts  3.4 OLAP and Multidimensional Data Analysis  3.4.1Representing Iris Data as a Multidimensional Array  3.4.2 Multidimensional Data: The General Case.  3.4.3 Analyzing Multidimensional Data  3.4.4 Final Comments on Multidimensional Data Analysis | Jiawei: Chapter 03 |
| **First Mid Exam** | | | |
| 21-32 | 6-8 | **Chapter 4 Data Warehousing and Online Analytical Processing**  4.1 Data Warehouse: Basic Concepts  4.1.1 What Is a Data Warehouse?  4.1.2 Differences between Operational Database Systems and Data Warehouses  4.1.3 But, Why Have a Separate Data Warehouse?  4.1.4 Data Warehousing: A Multitier Architecture  4.1.5 Data Warehouse Models: Enterprise Warehouse, Data Mart, and Virtual Warehouse  4.1.6 Extraction, Transformation, and Loading  4.1.7 Metadata Repository  4.2 Data Warehouse Modeling: Data Cube and OLAP  4.2.1 Data Cube: A Multidimensional Data Model  4.2.2 Stars, Snowflakes, and Fact Constellations: Schemas for Multidimensional Data Models  4.2.3 Dimensions: The Role of Concept Hierarchies  4.2.4 Measures: Their Categorization and Computation  4.2.5 Typical OLAP Operations  4.2.6 A Starnet Query Model for Querying Multidimensional Databases  4.3 Data Warehouse Design and Usage  4.3.1 A Business Analysis Framework for Data Warehouse Design  4.3.2 Data Warehouse Design Process  4.3.3 Data Warehouse Usage for Information Processing  4.3.4 From Online Analytical Processing to  Multidimensional Data Mining  4.4 Data Warehouse Implementation  4.4.1 Efficient Data Cube Computation: An Overview  4.4.2 Indexing OLAP Data: Bitmap Index and Join Index  4.4.3 Efficient Processing of OLAP Queries  4.4.4 OLAP Server Architectures: ROLAP versus MOLAP versus HOLAP  4.5 Data Generalization by Attribute-Oriented Induction  4.5.1 Attribute-Oriented Induction for Data Characterization  4.5.2 Efficient Implementation of Attribute-Oriented Induction  4.5.3 Attribute-Oriented Induction for Class Comparisons | Jiawei: Chapter 04 |
| 33-40 | 9-10 | **Chapter 6 Mining Frequent Patterns, Associations, and Correlations: Basic Concepts and Methods 243**  6.1 Basic Concepts  6.1.1 Market Basket Analysis: A Motivating Example  6.1.2 Frequent Itemsets, Closed Itemsets, and Association Rules  6.2 Frequent Itemset Mining Methods  6.2.1 Apriori Algorithm: Finding Frequent Itemsets by Confined Candidate Generation  6.2.2 Generating Association Rules from Frequent Itemsets  6.2.3 Improving the Efficiency of Apriori  6.2.4 A Pattern-Growth Approach for Mining Frequent Itemsets  6.2.5 Mining Frequent Itemsets Using Vertical Data Format  6.2.6 Mining Closed and Max Patterns  6.3 Which Patterns Are Interesting?—Pattern Evaluation Methods  6.3.1 Strong Rules Are Not Necessarily Interesting  6.3.2 From Association Analysis to Correlation Analysis  6.3.3 A Comparison of Pattern Evaluation Measures |  |
| **Second Mid Exams** | | | |
| 41-48 | 11-12 | **Chapter 8 Classification: Basic Concepts**  8.1 Basic Concepts  8.1.1 What Is Classification?  8.1.2 General Approach to Classification  8.2 Decision Tree Induction  8.2.1 Decision Tree Induction  8.2.2 Attribute Selection Measures  8.2.3 Tree Pruning  8.2.4 Scalability and Decision Tree Induction  8.2.5 Visual Mining for Decision Tree Induction  8.3 Bayes Classification Methods  8.3.1 Bayes’ Theorem  8.3.2 Naıve Bayesian Classification  8.4 Rule-Based Classification  8.4.1 Using IF-THEN Rules for Classification  8.4.2 Rule Extraction from a Decision Tree  8.4.3 Rule Induction Using a Sequential Covering Algorithm  8.5 Model Evaluation and Selection  8.5.1 Metrics for Evaluating Classifier Performance  8.5.2 Holdout Method and Random Subsampling  8.5.3 Cross-Validation  8.5.4 Bootstrap  8.5.5 Model Selection Using Statistical Tests of Significance  8.5.6 Comparing Classifiers Based on Cost–Benefit and ROC Curves | Jiawei: Chapter 08 |
| 49-56 | 13-14 | **Chapter 10 Cluster Analysis: Basic Concepts and Methods**  10.1 Cluster Analysis  10.1.1 What Is Cluster Analysis?  10.1.2 Requirements for Cluster Analysis  10.1.3 Overview of Basic Clustering Methods  10.2 Partitioning Methods  10.2.1 k-Means: A Centroid-Based Technique  10.4 Density-Based Methods  10.4.1 DBSCAN: Density-Based Clustering Based on Connected Regions with High Density | Jiawei: Chapter 09 |
| 57-64 | 15-16 | **Chapter 12 Outlier Detection**  12.1 Outliers and Outlier Analysis  12.1.1 What Are Outliers?  12.1.2 Types of Outliers  12.1.3 Challenges of Outlier Detection  12.2 Outlier Detection Methods  12.2.1 Supervised, Semi-Supervised, and Unsupervised Methods  12.2.2 Statistical Methods, Proximity-Based Methods, and Clustering-Based Methods  12.3 Statistical Approaches  12.3.1 Parametric Methods  12.3.2 Nonparametric Methods | Jiawei: Chapter 12 |
| **Final Exams** | | | |

**Text Book**

| 1. Jiawei Han and Micheline Kamber, “Data Mining: Concepts and Techniques”, Third Edition |
| --- |

**Reference Material**

| 1. Pang-Ning Tan, Michael Steinbach, and Vipin Kumar, **"Introduction to Data Mining**", Pearson Addison Wesley |
| --- |

**Course Learning Outcomes**

| **Course Learning Outcomes (CLO)** |
| --- |
| 1. Understanding the key processes of data mining principles and algorithms used in practical data mining. 2. Apply data mining techniques to solve problems in other disciplines in a mathematical way. 3. Design small projects with using data mining tools and techniques. (noncredit) |

**CLO-SO Map**

|  | **SO IDs** | | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| CLO ID | **GA1** | **GA2** | **GA3** | **GA4** | **GA5** | **GA6** | **GA7** | **GA8** | **GA9** | **GA10** | **GA11** | **GA12** |
| **CLO 1** | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **CLO 2** | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **CLO 3** | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

**Approvals**

| Prepared By | Dr Ghulam Murtaza Memon |
| --- | --- |
| Approved By | Not Specified |
| Last Update | 14 Feb 2021 |