CS280 Elements of Data Processing

Instructor Information	Quan Li Home Institution: ShanghaiTech University Email: liquan@shanghaitech.edu.cn Office Hours: Determined by Instructor		
Term	June 27, 2022 - July 22, 2022	**Credits**	4 units
Class Hours	Monday through Friday, 120 mins per teaching day		
Discussion Sessions	2.5 hours each week, conducted by teaching assistant(s)		
Total Contact Hours	66 contact hours (1 contact hour = 45 mins, 3000 mins in total)		
Required Texts (with ISBN)	Recommended texts: J. Han, M. Kamber and J. Pei, Data Mining: Concepts and Techniques, 3rd ed., Morgan Kaufmann, 2012. ISBN: 978-0-12-381479-1. Introduction to Data Mining. Pang-Ning Tan, Michael Steinbach, Vipin Kumar Boston : Pearson Addison		
Prerequisite	▪ Statistics and Probability would help, ▪ but not necessary ▪ Pattern Recognition would help, ▪ but not necessary ▪ Databases ▪ Knowledge of SQL and relational algebra ▪ But not necessary ▪ One programming language ▪ One of Java, C++, Perl, Matlab, etc. ▪ Will need to read Java Library		

Course Overview

This course covers both theoretical foundations and practical techniques and tools for data processing, data mining, knowledge discovery, and data visualization. Topics include data preprocessing, data warehouse and OLAP technology, advanced data cube technology, mining frequent patterns & association, classification, cluster analysis, outlier analysis, web data mining, and data visualization.

Learning Outcomes

The students will be able to:

1. Have a fundamental understanding on data, data representation and storage, processing, data mining, knowledge discovery, and data visualization.
2. Identify and use current data processing techniques, skills, and tools to perform effective data processing and analysis.
3. Have a basic knowledge of information retrieval, data mining, and knowledge discovery.

Program Outcomes

This course addresses the following program outcomes:

- An ability to apply knowledge of computing and mathematics appropriate to the discipline
- An ability to analyze a problem, and identify and define the computing requirements appropriate to its solution
- An ability to design, implement, and evaluate a computer-based system, process, component, or program to meet desired needs
- An ability to use current techniques, skills, and tools necessary for computing practice
- The capability for critical and independent thinking and skills for lifelong learning
- Respect for academic integrity and the ethics of scholarship

Grading Policy

Assignments	30%
Midterm	35%
Final Exam	35%

Grading Scale is as follows

Number grade	**Letter grade**	**GPA**
90-100	A	4
85-89	A-	3.7
80-84	B+	3.3
75-79	B	3
70-74	B-	2.7
67-69	C+	2.3
65-66	C	2
62-64	C-	1.7
60-61	D	1
≤59	F (Failure)	0

Class Schedule

Date	Lecture
Day 1	Course Introduction and Overview
Day 2	Getting to Know Your Data: Data Objects and Attribute Types, Basic Statistical Descriptions of Data, Measuring Data Similarity and Dissimilarity
Day 3	Data Warehouse: Basic Concepts, Data Warehouse Modeling: Data Cube and OLAP
Day 4	Data Cube Computation: Preliminary Concepts and Methods
Day 5	Association Rule Mining, Apriori Algorithm, FP-tree
Day 6	Association Rule Mining, Apriori Algorithm, FP-tree
Day 7	Mining Frequent Patterns & Association: Advanced Methods
Day 8	Classification: Basic Concepts, K Nearest Neighbor Classification Methods, Decision Tree Induction, Bayes Classification Methods, Model Evaluation and Selection, Techniques to Improve Classification Accuracy: Ensemble Methods
Day 9	Classification: Basic Concepts, K Nearest Neighbor Classification Methods, Decision Tree Induction, Bayes Classification Methods, Model Evaluation and Selection, Techniques to Improve Classification Accuracy: Ensemble Methods
Day 10	Midterm
Day 11	Classification: Advanced Methods: Classification by Backpropagation, Support Vector Machines, Additional Topics Regarding Classification

Day 12	Cluster Analysis: Basic Concepts, Partitioning Methods, Hierarchical Methods, Density-Based Methods, Grid-Based Methods, Evaluation of Clustering
Day 13	Cluster Analysis: Basic Concepts, Partitioning Methods, Hierarchical Methods, Density-Based Methods, Grid-Based Methods, Evaluation of Clustering
Day 14	Advanced Clustering Analysis: Probability Model-Based Clustering, Clustering High-Dimensional Data, Clustering Graphs and Network Data
Day 15	Outlier and Outlier Analysis, Outlier Detection Methods, Statistical Approaches, Proximity-Base Approaches, Clustering-Base Approaches, Classification Approaches, Outlier Detection in High Dimensional Data
Day 16	Web Data Mining: PageRank & HITS
Day 17	Data Visualization: Basic Concepts and Principles
Day 18	Data Visualization Techniques: Geographical Data Visualization, High-dimensional Data Visualization, Time-series Data Visualization, Hierarchical Data Visualization, Network Data Visualization, and Cross-Media Data Visualization
Day 19	Data Visualization Techniques: Geographical Data Visualization, High-dimensional Data Visualization, Time-series Data Visualization, Hierarchical Data Visualization, Network Data Visualization, and Cross-Media Data Visualization
Day 20	Final Exam