華東師範大學
EAST CHINA NORMAL UNIVERSITY

East China Normal University

DATA 22 Elements of Data Processing

Instructor: Haiyong Bao

Email: hybao@sei.ecnu.edu.cn

Home University: East China Normal University

Semester: December 19, 2022 to January 7, 2023

Course Hour: Monday through Friday, 160 mins per teaching day;

Total Contact Hours: 64 contact hours

Credits: 4

Designated Textbook with ISBN:

J. Han, M. Kamber and J. Pei, Data Mining: Concepts and Techniques, 3rd ed., Morgan Kaufmann, 2012. ISBN: 978-0-12-381479-1.

Bing Liu, Web Data Mining, Springer, 2011. ISBN: 978-3-642-26891-5.

Tilo Wendler, and Sören Gröttrup, Data Mining with SPSS Modeler: Theory, Exercises and Solutions, Springer, 1st ed., 2016. ISBN: 978-3-319-28707-2

Course Prerequisite:

Students are expected to have some knowledge of the concepts and terminology associated with statistics, database systems, and machine learning. Students are also expected to have some programming experience in any programming languages.

**Notes: The course might be moved to online delivery due to COVID-19 pandemic. Students will be notified once such decision is made.*

Course Overview

This course covers both theoretical foundations and practical techniques and tools for data processing and data mining. Topics include data representation, cleaning, transformation and analysis, visualization, general theory of data classification, Logistic Regression, Decision Trees, K-nearest neighbor, Naïve Bayes, Support Vector Machines, general theory of cluster analysis, K-Means Partitioning Clustering, model evaluation, social and ethical implications of data analytics.

Learning Outcomes

Upon completion of this course, students should be able to:

1. Have a fundamental understanding on data, data representation and storage, processing, visualization, and management.
2. Have a basic knowledge of key classification methods and cluster methods.
3. Identify and use current data processing techniques, skills, and tools to perform effective data processing and analysis.
4. Have respect for academic integrity and the ethics of scholarship

Grading Scale and Notes

The following definitions will be used as a guide for the assignment of grades:

Number Grade	Letter Grade	Definitions
94-100	A	Extraordinary distinction, indicating a full mastery of course content and excellent work.
90-93	A-	
87-89	B+	Strong performance demonstrating a high level of attainment, indicating a good comprehension of the course material and the student's full engagement with the course requirements and activities.
84-86	B	
80-83	B-	
77-79	C+	Acceptable performance, demonstrating an adequate and satisfactory comprehension of the course material and the student has met the basic requirements for completing assignments and participating in class activities.
70-76	C	
60-69	D	A marginal performance in the required exercises demonstrating a minimal passing level of attainment.
0-59	F	An unacceptable performance. The F grade indicates that the student's performance has revealed almost no understanding of the course content.

Assessment Policy

Assessment	Final Grade
Attendance	10%

Quiz	20%
Presentation	20%
Mid-Term Examination	20%
Final Examination	30%

Course Schedule

Week	Lecture	Reading/Assignments/ Examination
1	Definition of Data Processing, Definition of Data Mining, Introduction to Data Mining, and Targeted Applications	HKP: 3.1, 1.1-1.2, 1.5-1.6
	Data Representation, Type of Attributes Basic Statistical Description of Data, Introduction to SPSS and SPSS Modeler	HKP: 2.1-2.2 WG: 1
	Data Integration and Cleaning: Missing Values and Outlier Detection and Removal	HKP: 3.2, 12.1-12.2 WG: 2.7
	Data Transformation and Data Discretization, Data Visualization	HKP: 3.5, 2.3, 3.3 WG: 4.2
	Principal Components Analysis and Principal Factor Analysis	HKP: 3.4 WG: 6.3,6.4
2	Mining Frequent Patterns, Associations, and Correlations, Association Rules	HKP：6.2, 6.3
	Midterm	Data Analysis Report 1
	Entropy and Information Gain	HKP: 8.2.2
	Presentation	Data Analysis Report 1
	Classification Methods: Logistic Regression, Decision Trees, K-Nearest Neighbor	HKP: 8.2, 9.5.1 WG: 8.3, 8.7, 8.8
3	Classification Methods: Naïve Bayes, Combining Classifiers, Support Vector Machines	HKP: 8.3, 9.3 WG: 8.5
	Model Evaluation and Selection	HKP: 8.5.1-8.5.5
	Cluster Analysis: General Theory of Cluster Analysis, K-Means Partitioning Clustering	HKP: 10.1,10.2 WG: 7.2, 7.4
	Data Linkage, Privacy and Bloom Filters, Social and Ethical Implications of Big Data Analytics, Cloud Computing Project	HKP: 13.4
	Final Exam	Data Analysis Report 2

Reading List:

- **As given in the table of Course Schedule.**