Cluster Analysis in Data Mining: Course Overview

What Is Cluster Analysis?

❑ When flying over a city, one can easily identify fields, forests, commercial areas, and residential areas based on their features, without anyone's explicit "training" - this is the power of cluster analysis. This course will systematically study cluster analysis methods and help answer the following:

❑ What are the different proximity measures for effective clustering?

❑ Can we cluster a massive number of data points efficiently?

❑ Can we find clusters of arbitrary shape? At multiple levels of granularity?

❑ How can we judge the quality of the clusters discovered by our system?

The Value of Cluster Analysis

❑ What is the value of cluster analysis?

  ❑ Cluster analysis helps you partition massive data into groups based on its features.

  ❑ Cluster analysis will often help subsequent data mining processes such as pattern discovery, classification, and outlier analysis

❑ What roles does cluster analysis play in the Data Mining Specialization?

  ❑ You will learn various scalable methods to find clusters from massive data.

  ❑ You will learn how to mine different kinds of clusters effectively.

  ❑ You will also learn how to evaluate the quality of the clusters you find.

  ❑ Cluster analysis will help with classification, outlier analysis, and other data mining tasks.

Broad Applications of Cluster Analysis

- ❑ Data summarization, compression, and reduction

  - ❑ Examples: Image processing or vector quantization

- ❑ Collaborative filtering, recommendation systems, or customer segmentation

  - ❑ Finding like-minded users or similar products

- ❑ Dynamic trend detection

  - ❑ Clustering stream data and detecting trends and patterns

- ❑ Multimedia data analysis, biological data analysis, and social network analysis

  - ❑ Example: Clustering images or video/audio clips, gene/protein sequences, etc.

- ❑ A key intermediate step for other data mining tasks

  - ❑ Generating a compact summary of data for classification, pattern discovery, and hypothesis generation and testing

  - ❑ Outlier detection: Outliers - those "far away" from any cluster

# Major Reference Readings for the Course

❑ **Textbook**

❑ Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concepts and techniques (3rd ed.).* Morgan Kaufmann.

❑ **Chapters most related to the course**

❑ Chapter 2: Getting to Know Your Data (Section 2.4: Measuring Data Similarity and Dissimilarity)

❑ Chapter 10: Cluster Analysis: Basic Concepts and Methods

❑ Other references will be listed at the end of each lecture video.
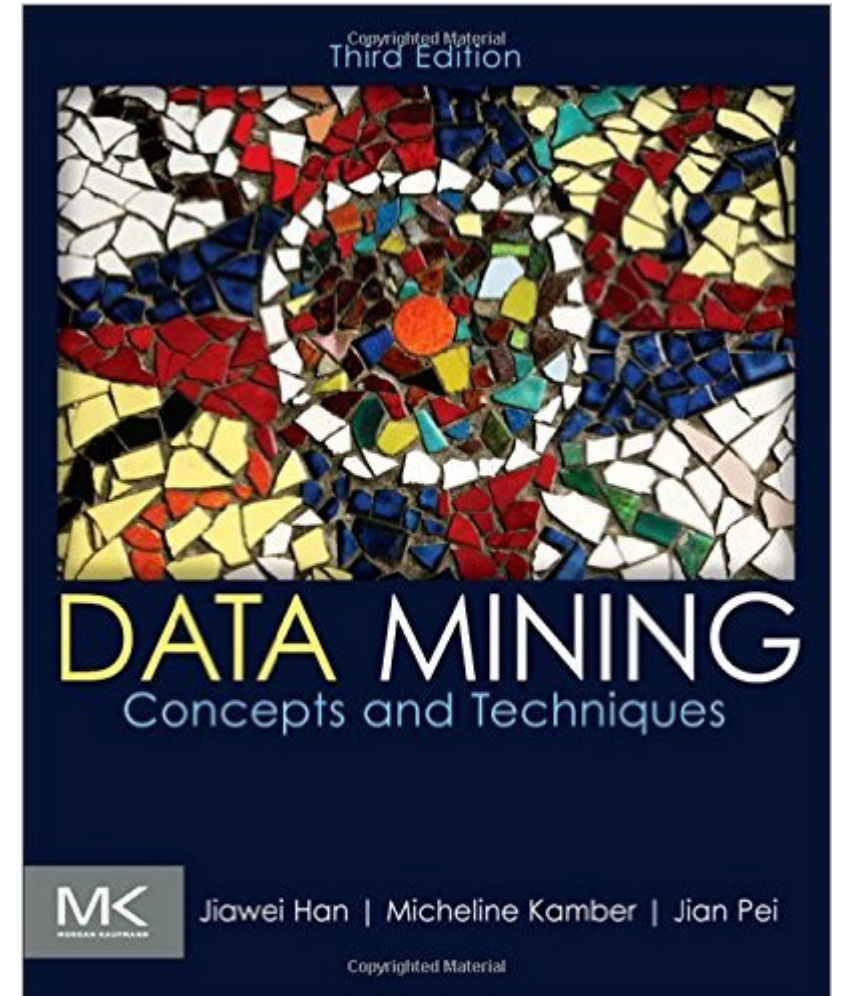
Course Structure

- ❑ Lesson 1: Cluster Analysis: An Introduction
- ❑ Lesson 2: Similarity Measures for Cluster Analysis

Module 1

- ❑ Lesson 3: Partitioning-Based Clustering Methods
- ❑ Lesson 4 (Part I): Hierarchical Clustering Methods (I)

Module 2

- ❑ Lesson 4 (Part II): Hierarchical Clustering Methods (II)
- ❑ Lesson 5: Density-Based and Grid-Based Clustering Methods

Module 3

- ❑ Lesson 6: Clustering Validation

Module 4

Course General Information

❑ Instructor:

Jiawei Han, Abel Bliss Professor

Department of Computer Science

University of Illinois at Urbana-Champaign

❑ Teaching assistants

❑ Course prerequisite:

Familiarity with basic data structures and algorithms

❑ Course assessments

❑ In-video questions

❑ Lesson quizzes

❑ Programming assignments