Data Mining Course: Introduction

Cam Tu Nguyen

阮锦绣

Software Institute, Nanjing University
nguyenct@lamda.nju.edu.cn
ncamtu@gmail.com

Prerequisites

- Basic Databases
- Algorithms:
  - Dynamic Programming, basic data structures
- Basic statistics
  - Fundamental distributions (Gaussian, Bernoulli distributions, etc)
  - Regression
- Programming
  - Your choice, but Java/Python will be very useful.

Course Outline

- Fundamental:
  - Data Exploration & Preparation
  - Classification
    - Basic Algorithms: Decision Trees, KNN, Linear Models
    - Evaluation: Cross-validation
  - Clustering
    - Basic Algorithms: K-means, Hierarchical Clustering; Example of Hierarchical Clustering; Incremental Clustering
    - Evaluation
  - Association rules, Frequent items

Course Outline (2)

- Special Topics
  - Recommendation Systems
  - Information Retrieval & PageRank
  - Text Mining & Topic Analysis
  - Introduction to Representation Learning
    - Neural network & Deep Learning
  - Large Scale Data mining on Map-Reduce and/or Spark

Possible Projects (10% credit bonus)

- **Document Management Application**
  - Browsing documents by predefined classes
  - Discover the most dominant cluster in a corpus of document
  - Data set: Wikipedia pages

- **Word association discoveries**
  - Given a word, find words that are frequently associated with it.
  - Make a word game, that given associated words as hints and we need to guess a hidden word.
  - Data set: Wikipedia pages

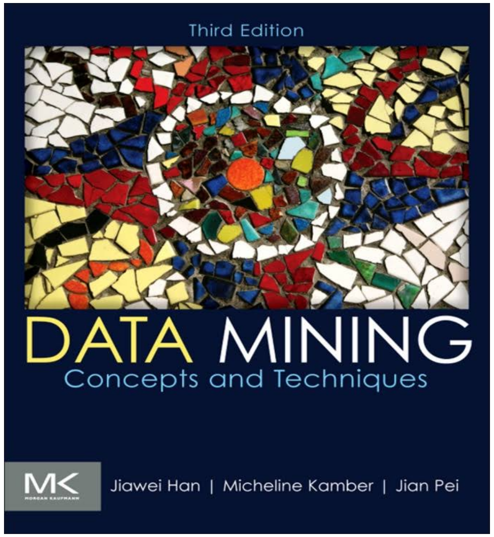
- Other topics are possible …

- Reports
  - A report describing data mining & evaluation methods
  - A demo application/website

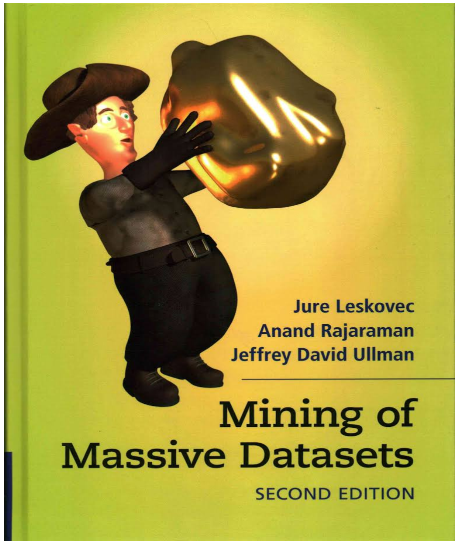
Possible Projects

- Team work:
  - It is possible to work in team of 3-4 students
  - Each student needs to have clear role in the project.
- Registering for teams and projects
  - Need to send out by October 1st .
- Project report Due
  - Possible dates: November 15th

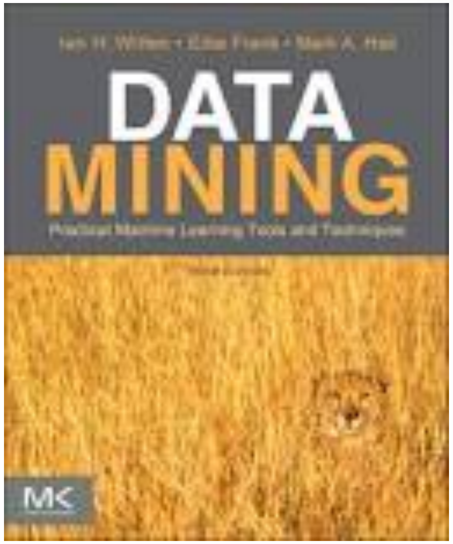
Text Books



Data Mining: Concepts and Techniques



Mining of Massive Datasets


Ian H. Witten • Eibe Frank • Mark A. Hall
DATA MINING
Practical Machine Learning Tools and Techniques
Third Edition
MK

Data Mining: Practical Machine Learning Tools and Techniques.