Data Science: Course Overview

March 16, 2020

Sang-Wook Kim

Big Data Science Lab.
Department of Computer Science and Engineering
Hanyang University

Instructor

- Sang-Wook Kim (Big Data Science Lab.)
  - Areas of interest
    - Big data, machine learning, data mining, databases, recommender systems, and social network analysis
  - Contact information
    - Email: wook@hanyang.ac.kr
    - Phone: 02-2220-1736
    - URL: http://agape.hanyang.ac.kr (link to a course community page)
  - Teaching assistants
    - Dong-Hyuk Seo (email: hyuk125@agape.hanyang.ac.kr)
    - Taeri Kim (email: taerik@agape.hanyang.ac.kr)

Goal

- *To learn techniques and applications* of data mining in large databases
  - To understand the concepts of data mining
    - To find interesting patterns from a huge volume of data
  - To study a variety of data mining techniques
  - To understand the applications of data mining
  - To analyze real-world data by using data mining tools
  - To improve programming skills by developing data mining techniques and applications

Text Information

- Primary textbook
  - Jiawei Han, Micheline Kamber, and Jian Pei, *Data mining: concepts and techniques*, Morgan Kaufmann
- Secondary handouts
  - Related research papers
    - Will be provided when necessary
    - Available via Google scholar *in our university*

Issues to Be Touched

- Data Preprocessing
- Frequent Pattern Mining
- *Association Mining*
- Data Clustering
- Classification and Prediction
- Data Generalization
- Outlier Analysis
- Social Network Analysis
- Recommendation
- Other Big Data Issues

Pre-requisites

- Courses
  - Data structures (*mandatory*)
  - Databases (*highly recommended*)
- Programming Skill
  - Around 1,000~1,500 lines (*required*)
  - Debugging *with debuggers*

Grading Scheme

- Relative evaluation
  - A:B:CDF = **25%:35%:40%**
- Weights on graded parts
  - Midterm exam: **30%**
  - Final exam: **30%**
  - Term project: **30%**
    - **4 programing assignments (1,000~1,500 lines each)**
  - Attendance**+**: **10%**
    - **Bonus for participation (good questions or answers)**
- The students who took this class before (재수강) will be ***down-graded*** to a lower level (ex. A+ => B+; B0 => C0)

Notice

- Special grading policy
  - Grade ‘D‘ will be given if any two of programing projects are *not successfully fulfilled*
  - Grade ‘F‘ will be given if
    - S/he copies somebody else‘s program (i.e., from classmate or from the Internet) or ***allows others to copy her/his own program;***
    - S/he does not take either the midterm or final exam
- Attendance policy
  - No penalty up to 5 absence (after this point, penalized)
  - Two late attendance will be considered as one absence
  - Note: The attendance later than 10 minutes after the start time will be regarded as ABSENCE rather than late attendance

Notice

- No (audio/video) recording policy
  - Otherwise, this will be penalized significantly
- Do preview on our textbook for 5 minutes
- Visit our community site *at least once a week*
  - *For important announcement given*
- Online lecture video
  - It will be provided only in the attendance period (i.e., the week of its upload)
    - Watch the lecture *WITHIN this period* (i.e., not after nor before the period) for the attendance to be counted correctly
  - It will be no longer available after the period

Projects: General Information

- Four programming assignments
  - Three short-term projects
    - Frequent pattern mining: Apriori (will be announce today)
    - Classification: Decision tree
    - Clustering: DBSCAN
  - One long-term project
    - Recommender system (will be announce this week)
- Gitlab registration
  - For submission of programming assignments:
    - Make your account in gitlab by referring to the notice in our community site (due 3/20)

Exam schedule

- Time
  - Midterm exam: 5/25(mon) 19:00
  - Final exam: 6/17(wed) 19:00
    - If you have a problem with this schedule, please contact me via email by 3/18(wed)
    - Otherwise, this schedule will be finalized
- Place
  - will be announced later

Programming Assignment #1

- Title: *Apriori algorithm for association rule mining* in transactional databases
- Descriptions and requirements
  - Will be uploaded in our community site (***Today!***)
- Environment
  - OS: Windows, Mac OS, or Linux
  - Languages: C, C++, C#, Java, or Python (any version is ok)
- Goal
  - Find association rules using the ***Apriori*** algorithm

Programming Assignment #1

- Late submission policy (*for all short-term assignments*)
  - *20% penalty*: less than or equal to a week
  - *50% penalty*: less than or equal to two weeks
  - *Will not be accepted*, after two weeks
- Requirements unsatisfied
  - Significant penalty up to 30% will be given when the requirements are not fully-satisfied

Long-Term Project

- Title: *Recommender system* for movie data
- Descriptions and requirements
  - Will be uploaded in our community site (***this week!***)
- Environment
  - Same as assignment #1
- Goal
  - To predict the *ratings of movies* in test data by using the given training data containing movie ratings from users
  - You can choose any algorithm to predict
    - Ex. content-based and collaborative-filtering-based algorithms
    - For a content-based algorithm, you can refer to web page to get the content related to data (http://grouplens.org/datasets/movielens/)

Long-Term Project

- Note
  - This has a competition-based scoring system
  - As the accuracy of your model is higher, you will get a higher score
    - You will receive a minimum score at least 80 if you:
      - You submit your program before the deadline
      - Your program correctly works without any error
      - All requirements for this project are satisfied
- Late submission policy
  - *This assignment does not allow late submission!*
- Requirements unsatisfied
  - Significant penalty up to 30% will be given when the requirements are not fully-satisfied

Thank You!