Data Mining Theory

#1: Guidance

1. Definition of Data Mining
2. Glossary in machine learning
3. Syllabus

URL: https://ie.u-ryukyu.ac.jp/~tnal/2020/dm-theory/

Definition of Data Mining (1/2)

- Wikipedia: https://en.wikipedia.org/wiki/Data_mining
  - Data mining is an interdisciplinary subfield of computer science.[1][2][3] It is the **computational process of discovering patterns in large data sets** involving methods at the **intersection of artificial intelligence, machine learning, statistics, and database systems.**[1] The overall goal of the data mining process is **to extract information from a data set and transform it into an understandable structure for further use**.[1]
- DATA MINING CURRICULUM: http://www.kdd.org/curriculum/index.html
  - Recent tremendous technical advances in processing power, storage capacity, and inter-connectivity of computer technology is creating unprecedented quantities of digital data. Data mining, the science of **extracting useful knowledge** from such huge data repositories, has emerged as a young and **interdisciplinary field in computer science**. Data mining techniques have been widely applied to problems in industry, science, engineering and government, and it is widely believed that data mining will have profound impact on our society.

Definition of Data Mining (2/2)

- [book] Data Mining: Practical Machine Learning Tools And Techniques, 3rd edition
  - Preface
    - Data mining is the **extraction of implicit, previously unknown, and potentially useful information from data**. The idea is to build computer programs that sift through databases automatically, seeking regularities or patterns. Strong patterns, if found, will likely generalize to make accurate predictions on future data.
    - **Machine learning provides the technical basis of data mining**. It is used to extract information from the raw data in databases—information that is expressed in a comprehensible form and can be used for a variety of purposes. (...) This book is about the tools and techniques of machine learning that are used in practical data mining for finding, and describing, structural patterns in data.

Glossary in machine learning

- supervised, unsupervised learning
- classification, regression, clustering
- sample
- features, attributes
  - numerical value
  - categorical value
  - true or false
- supervisory signal, teacher, class, label, output data, target variable

- input, output
- training data / training set
- test data / test set
  - open test
  - close test
- model
- parameters
- learn, fit
- predict, estimate
- evaluation

After Part 1, you can explain those terms!

Syllabus

Course content and methods

- This class consists of mainly 2 parts. We learn about a general process to be ready for re-useable knowledges by data mining theories, or about machine learning models as a mining method.
- Each day has two ~~groups~~ presentations have to explain/introduce about assigned references.
  - basic theories (textbook) on part 1.
  - applications (journals) on part 2.
- ~~According to circumstances (belong to students), we will take consideration time as part 3 for practice about your-own work or interest applications.~~

Goals and objectives

- After Part 1,
  - You can explain about general process of data mining or a few models in machine learnings.
- After Part 2,
  - You can explain/introduce about newly conference papers.

- ~~If we select part 3,~~
  - ~~You can try to design/build the process to any your familiar topic.~~
  - ~~You can evaluate/consider about raw data and results of your application.~~

Flow for part 1

- Presenter
  - Before of class
    - Upload your presentation document(s) to google drive.
  - While class
    - 30 minutes for explanation .
    - 10 minutes for discussion.
  - After of class
    - If I (Toma) received questions from other students, I'll send them to you by the next day.
    - When you receive questions, make answers by the next lecture day.

- Other students
  - Before of class
    - Read textbook, execute example codes and list up about difficult points for you.
  - While class
    - Try to understand the target.
    - If you have any questions, please ask presenter.
  - After of class
    - If you have questions still, write them in google form by the day.

Flow for part 2

- Presenter
  - Before of class
    - Select a paper & write the title on the assignment sheet at least 7 days (1 week) in advance.
    - Upload your presentation.
  - While class
    - 20 minutes for explanation.
    - 10 minutes for discussion.
  - After of class
    - Same.

- Other students
  - Before of class
    - Read the abstract of paper in advance and list up about interesting points for you.
  - While class
    - Same as part 1.
  - After of class
    - Same.

Evaluation criteria and evaluation methods

- You must explain/introduce/discuss about assigned references.
- presentation (50%), presentation documents (20%), Q&A and discussions (30%)
  - Mini report about chapter 1 & 2 == 12 points
    - consists of 10 points for explanation, 2 points for Q&A
  - 1 presentation == 20 points & 10 points
    - 20 points for explanation & answering, 10 points for document
  - Discussion points
    - On time discussion == 4 points per day
    - After class discussion from google form == 1 point per day
    - When you joined 12 discussions with only google form, you got 1 x 12 times = 12 points for Q&A.

Course conditions

- Math (especially Linear Algebra and Statistics), experiences of programming.

Prior/Post learning

- read carefully references.
- Two candidates in Part 1
  - ~~– Data Mining: Practical Machine Learning Tools and Techniques, 3rd Edition~~
    - ~~• Targets: from chapter 1 to chapter 5, 13 assignments.~~
  - – Python Machine Learning
    - Targets: from chapter 1 to chapter 12, 12 assignments.

Message

- In part 2, we will read some best papers on below conferences in recent 5 years. But I hope that you suggest us any related papers in your interest.
- mainly application examples.
  - IEEE/WIC/ACM International Conference on Web Intelligence (IEEE/WIC/ACM WI)
  - ACM International Conference on Web Search and Data Mining (ACM WSDM)
  - ACM Special Interest Group on Information Retrieval (SIG-IR)
- mainly theoretical or technical papers.
  - IEEE International Conference on Data Mining (ICDM)
  - ACM SIGKDD Conference on Knowledge Discovery and Data Mining (ACM KDD)

Office hour and contact

- Office hour
  - 12:50-14:20 on Friday, eng. #1-705
- Contact
  - Naruaki TOMA
  - tnal@ie.u-ryukyu.ac.jp