Advanced Data Mining

(was: 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/2023/adm/

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, 4th 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
- Data set
  - Training, validation and test set
- model
- parameters
- learn, fit
  - Tuning hyper-parameters
  - Optimize other parameters
- predict, estimate
- Evaluation
  - cross-validation

After this class, 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 and machine learning methods.
- **Part 1: reading & discussion about specified books**.
  - Each day has about 5-7 presentations to explain/introduce about assigned books.
    - **Book 1**: Data Mining, Practical Machine Learning Tools and Techniques, 4th edition.
      - You can download the ebook from the university's LAN at this link, or with VPN connection.
    - *(reference) Book 2: Python machine learning : machine learning and deep learning with Python, scikit-learn, and TensorFlow, 3rd edition.*
- **Part 2: writing & discussion about your research** from a data mining perspective.
  - On the last 3-4 weeks, we will discuss your own work.

Class plan (assignment)

- Please refer to "schedule.xlsx" in Teams for details.
  - Teams
    - You must log in to Teams with your university account (xx@cs.u-ryukyu.ac.jp).
    - Team code: **dsaks3m**
    - Team name: CLS_2023 ADM (Advanced Data Mining / データマイニング特論)

Goals and objectives

- Part 1.
  - After Book 1,
    - You can explain about general process of data mining & machine learning models.
- (after part 2)
  - You can learn about some specific application cases.

Flow for Part 1

- Presenter
  - Before of class
    - **Upload** your presentation document(s) to Teams.
  - While class
    - Please keep QA time. Basically, one presentation has about **15-30 mins for explanation and 10 mins for QA**.
  - After of class
    - When you receive questions on Teams, make answers by the next lecture day.

- Other students
  - Before of class
    - Read textbook and list up about difficult points for you.
    - Recommendation: run example codes.
  - 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 Teams by the day.

Evaluation criteria and evaluation methods

- You must explain/introduce/discuss about assigned books for Part 1. You must write/discuss about your research for Part 2.
- presentations (60%), discussions/questions (20%), final report (20%)
  - Presentations == 60 points
    - Scores for presenters
    - 50 points for explanation & answering, 10 points for presentation material.
  - Discussions/Questions == 20 points
    - Scores for participants. This refers to "questions to other presenters".
    - On time Question == 5 points
    - After class questions from Teams == 2 points per Question
  - Final report (Part 2)
    - Please explain about your research from a data mining perspective.
    - More than 500 words.
    - Follow the format of general academic conference manuscripts. For example, see ACM template: https://www.acm.org/publications/proceedings-template

Course conditions

- Math (especially Linear Algebra and Statistics)
- Experiences of programming. Highly recommended: Python or Java
- Teams & Zoom

Prior/Post learning

- Read references carefully.
- Please run example codes as possible.
  - Java (Weka) for book 1.
  - Python (scikit-learn) for book 2.

Message

- If you have interest, you can read some best papers on below conferences in recent years.
- 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
  - 10:20-11:50 on Wednesday, eng. #1-705
- Contact
  - Naruaki TOMA
  - tnal@ie.u-ryukyu.ac.jp
  - Teams