Data Mining Theory

#1: Guidance

1. Definition of Data Mining
2. Glossary in machine learning
3. Example: Iris flower data set
4. Syllabus

URL: https://ie.u-ryukyu.ac.jp/~tnal/2016/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