Principles of Data Mining

Pham Tho Hoan
hoanpt@hnue.edu.vn
References


Principles of Data Mining

Table of Contents

Principles of Data Mining
Series Foreword
Preface
Chapter 1 - Introduction
Chapter 2 - Measurement and Data
Chapter 3 - Visualizing and Exploring Data
Chapter 4 - Data Analysis and Uncertainty
Chapter 5 - A Systematic Overview of Data Mining Algorithms
Chapter 6 - Models and Patterns
Chapter 7 - Score Functions for Data Mining Algorithms
Chapter 8 - Search and Optimization Methods
Chapter 9 - Descriptive Modeling
Chapter 10 - Predictive Modeling for Classification
Chapter 11 - Predictive Modeling for Regression
Chapter 12 - Data Organization and Databases
Chapter 13 - Finding Patterns and Rules
Chapter 14 - Retrieval by Content
Appendix - Random Variables
References
Index
List of Figures
List of Tables
List of Examples
What is data mining

- Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner.

- The relationships and summaries derived through a data mining exercise are often referred to as models or patterns. Examples include linear equations, rules, clusters, graphs, tree structures, and recurrent patterns in time series.

- Observational data ≠ experimental data, convenience (opportunity) samples ≠ random samples, huge data ≠ small data, data mining ≠ statistics

- Novelty >> triviality, novelty must be measured relative to the user's prior knowledge

- Simple relationships are more readily understood than complicated ones, and may well be preferred, but simple ones may not be useful.
Data mining and Knowledge Discovery in Data

1. Understand the domain and Define problems
2. Collect and Preprocess Data
3. Data Mining
   Extract Patterns/Models
4. Interpret and Evaluate discovered knowledge
5. Putting the results in practical use

KDD is inherently interactive and iterative

Data mining is a step in the KDD process consisting of methods that produce useful patterns or models from the data. Maybe 70-90% of effort and cost in KDD.
Types of data sets

- $n \times p$ data matrix \{real number, category, missing, noise\}
- Text, sequence, structure, pictures
- Transactions
- Etc.

Lost information
Model and pattern structures

- **A model structure**, as defined here, is a *global* summary of a data set; it makes statements about any point in the full measurement space. \( Y = aX + c \)

- **Pattern structures** make statements only about restricted regions of the space spanned by the variables. An example is a simple probabilistic statement of the form: if \( X > x_1 \) then \( \text{prob}(Y > y_1) = p_1 \); or \( p(Y > y_1 | X > x_1) = p_1 \). This structure consists of *constraints* on the values of the variables \( X \) and \( Y \), related in the form of a probabilistic rule.

- Once we have established the structural form we are interested in finding, the next step is to estimate its parameters from the available data. We refer to a particular model, such as \( y = 3:2x + 2:8 \), as a "fitted model," or just "model" for short (and similarly for patterns).
Data mining tasks

• **Exploratory Data Analysis (EDA)** the goal is simply to explore the data without any clear ideas of what we are looking for. Typically, EDA techniques are *interactive* and *visual*, and there are many effective graphical display methods for relatively small, low-dimensional data sets.

• **Descriptive Modeling** The goal of a descriptive model is to describe all of the data (or the process generating the data). Examples of such descriptions include models for the overall probability distribution of the data (*density estimation*), partitioning of the $p$-dimensional space into groups (*cluster analysis and segmentation*), and models describing the relationship between variables (*dependency modeling*).

• **Predictive Modeling: Classification and Regression** The aim here is to build a model that will permit the value of one variable to be predicted from the known values of other variables.

• **Discovering Patterns and Rules**

• **Retrieval by Content**
Components of data mining algorithms

1. **Model or Pattern Structure**: determining the underlying structure or functional forms that we seek from the data

2. **Score Function**: judging the quality of a fitted model

3. **Optimization and Search Method**: optimizing the score function and searching over different model and pattern structures.

4. **Data Management Strategy**: handling data access efficiently during the search/optimization
Score functions

• Without some form of score function, we cannot tell whether one model is better than another or, indeed, how to choose a good set of values for the parameters of the model.

• Several score functions are widely used for this purpose; these include likelihood, sum of squared errors, and misclassification rate (the latter is used in supervised classification problems).

• **Penalize** model complexity:

score(model) = error(model) + penaltyFunction(model),
Optimization and Search Methods

• The goal of optimization and search is to determine the structure and the parameter values that achieve a minimum (or maximum, depending on the context) value of the score function.

• Methods: Greedy Search Algorithm, Systematic Search and Search Heuristics, Branch-and-Bound, Gradient-Based Methods for Optimizing Smooth Functions, Univariate Parameter Optimization, Multivariate Parameter Optimization, Constrained Optimization, etc.
Data Management Strategy

• The ways in which the data are stored, indexed, and accessed.
An example

• Problem:
  – Input: a dataset of credit-card spending \{ (x_i, y_i), i=1, .., n \};
  – Output: a model which would allow us to predict a person's annual credit-card spending given their annual income.

• One solution: the model would not be perfect, but since spending typically increases with income, the model might well be adequate as a rough characterization.
  – Model structure: variable spending (f) is linearly related to the variable income (x): \ f(x) = ax + b
  – The score function: \ \sum [y_i - f(x_i)]^2
    The smaller this sum is, the better the model fits the data.
  – The optimization algorithm (to find a, b) is quite simple in the case of linear regression: \ a and \ b can be expressed as explicit functions of the observed values of spending and income.
Some questions

• Cùng bài toán trên (mô hình hóa quan hệ giữa x và y), xem xét 3 mô hình sau, anh chỉ thích mô hình nào?
  M1: $y = \frac{(y_1 + \ldots + y_n)}{n}$ với mọi $x$
  M2: $y = ax + b$ (với $a, b$ tìm được như trong slide trước)
  M3: if $(x=x_1)$ then $y=y_1$
    else if $(x=x_2)$ then $y=y_2$
    else ...
    if $(x=x_n)$ then $y=y_n$
    else $y=random-value$ (default)

• Mô hình nào phức tạp nhất? Mô hình nào phù hợp với dữ liệu huấn luyện nhất?
• Mô hình nào có khả năng dự đoán tốt nhất?