DATA MINING PRE-PROCESSING

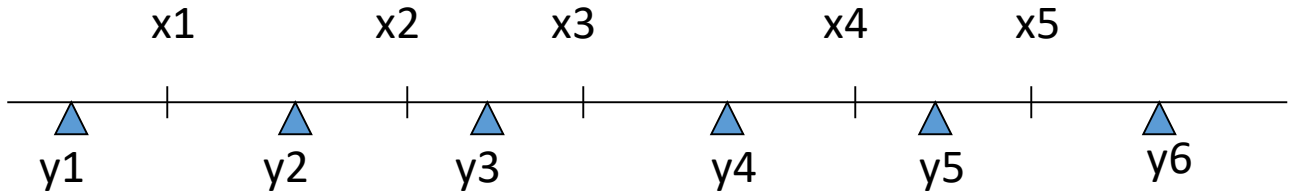
Outline

1. Why data preprocessing?
2. Data cleaning
3. Data integration and transformation
4. Data reduction
5. Discretization and concept hierarchy generation
6. Summary

5 Discretization and Concept Hierarchy Generation

Discretization/Quantization

- Three types of attributes:
  - Nominal — values from an unordered set
  - Ordinal — values from an ordered set
  - Continuous — real numbers
- Discretization/Quantization:
  - divide the range of a continuous attribute into intervals



  - Some classification algorithms only accept categorical attributes.
  - Reduce data size by discretization
  - Prepare for further analysis

Discretization and Concept Hierarchy

- Discretization
  - reduce the number of values for a given continuous attribute by dividing the range of the attribute into intervals. Interval labels can then be used to replace actual data values.
- Concept Hierarchies
  - reduce the data by collecting and replacing low level concepts (such as numeric values for the attribute age) by higher level concepts (such as young, middle-aged, or senior).

6 Summary

Summary

- Data preparation is a big issue for both warehousing and mining
- Data preparation includes
  - Data cleaning and data integration
  - Data reduction and feature selection
  - Discretization
- A lot a methods have been developed but still an active area of research

Terima Kasih