University of Reading

CSMDM21 - Data Analytics and Mining

An Overview of Data Preprocessing

Module convenor
Dr. Carmen Lam
carmen.lam@reading.ac.uk
Department of Computer Science

Lecture notes and videos powered by
Prof. Giuseppe Di Fatta

Data Preprocessing

Data Preprocessing:

- Aggregation
- Sampling
- Dimensionality Reduction
- Feature subset selection
- Feature engineering
- Discretization and Binarization
- Attribute Transformation

Aggregation

- Combining two or more attributes (or objects) into a single attribute (or object)

- Purpose
  - Data reduction
    - Reduce the number of attributes or objects
  - Change of scale
    - Cities aggregated into regions, states, countries, etc
  - More "stable" data
    - Aggregated data tends to have less variability

Sampling

- Sampling is the main technique employed for data selection.
  - It is often used for both the preliminary investigation of the data and the final data analysis.

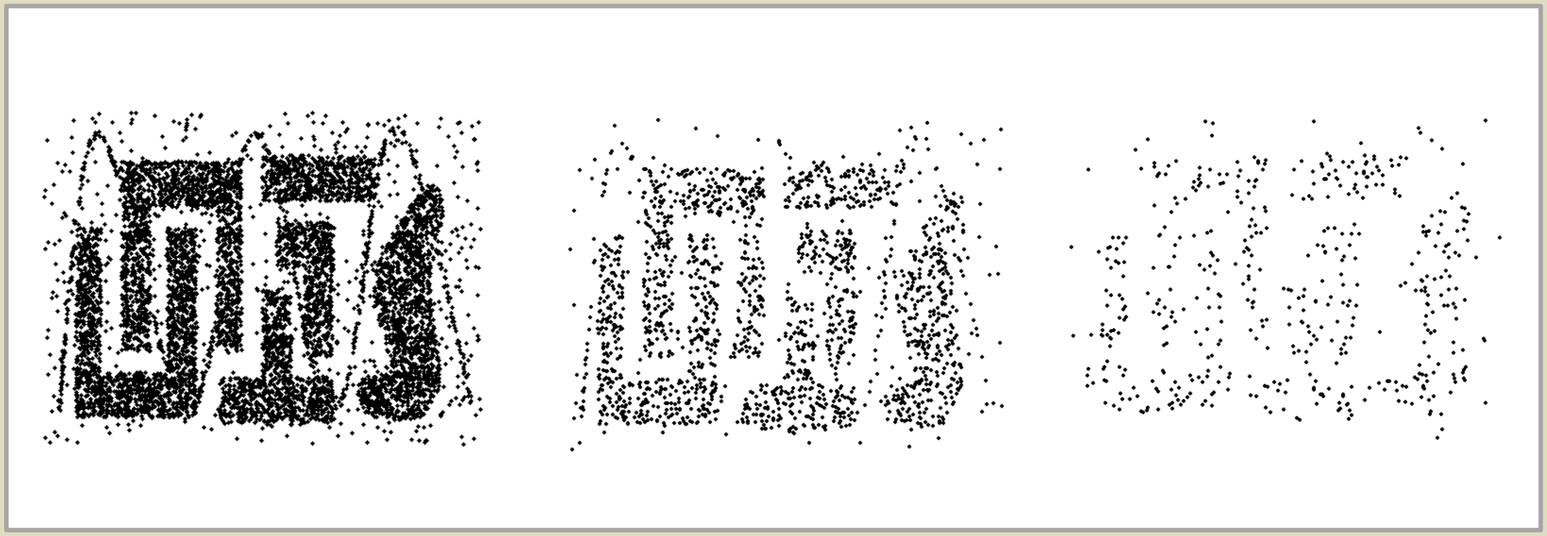
- In statistical analysis we typically use samples because obtaining the entire set of data of interest (entire population) is too expensive, time consuming or not feasible.
  - using a sample will work almost as well as using the entire data sets, if the sample is representative.
  - A sample is representative if it has approximately the same property (of interest) as the original set of data

- Sampling is used in Data Analytics and Mining because processing the entire set of data of interest is too expensive or time consuming.

Types of Sampling

- **Simple Random Sampling**
  - There is an equal probability of selecting any particular item
- **Sampling without replacement**
  - As each item is selected, it is removed from the population
  - Each outcome depends on all previous outcomes
- **Sampling with replacement**
  - Objects are not removed from the population as they are selected for the sample.
    - In sampling with replacement, the same object can be picked up more than once
    - One outcome does not affect the other outcomes
- **Stratified sampling**
  - Split the data into several partitions; then draw random samples from each partition

Sample Size



8000 points **2000 Points** **500 Points**

Curse of Dimensionality

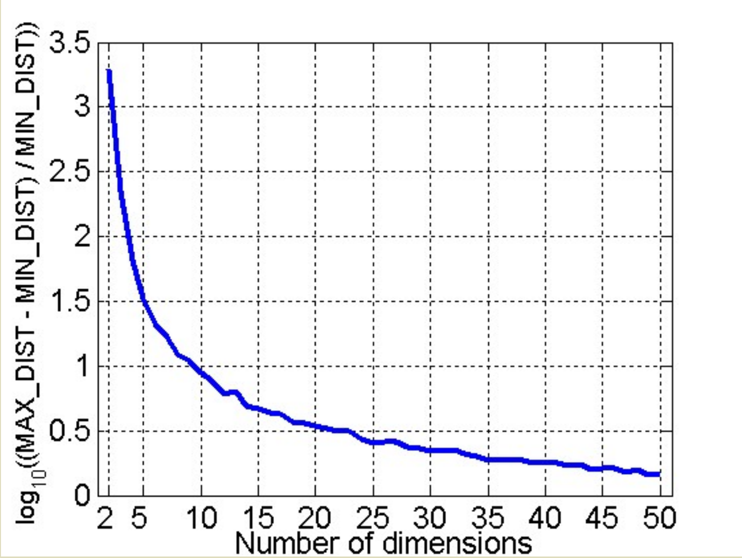
- When dimensionality increases, data becomes increasingly sparse in the space that it occupies.

- Definitions of *density* and *distance* between points, which is critical for clustering and outlier detection, become less meaningful.

Example

- Randomly generate 500 points in $\Re^n$
- Compute difference between max and min distance between any pair of points

$$f(n) = \log_{10}\left(\frac{\max\left(dist\left(v_i, v_j\right)\right) - \min\left(dist\left(v_i, v_j\right)\right)}{\min\left(dist\left(v_i, v_j\right)\right)} \right)$$



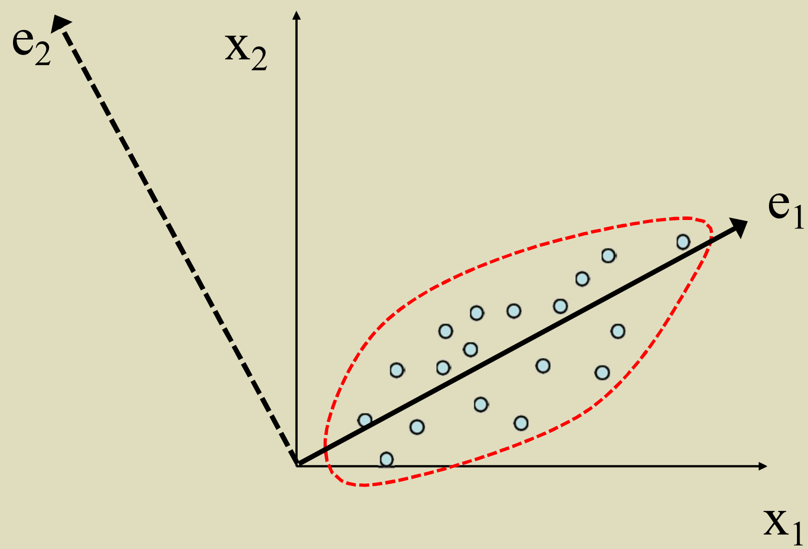
Dimensionality Reduction

- Purpose:
  - Avoid curse of dimensionality
  - Reduce amount of time and memory required by Data Analytics and Mining algorithms
  - Allow data to be more easily visualized
  - May help to eliminate irrelevant features or reduce noise

- Techniques
  - Principle Component Analysis
  - Singular Value Decomposition
  - Others: supervised and non-linear techniques

Dimensionality Reduction: PCA

- The goal is to find a projection that captures the largest amount of variation in data in the top dimensions (principal components).
  - e_1: 1st PC ➔ max variance
  - e_2: 2nd PC ➔ can be dropped, dimensionality reduction



Feature Subset Selection

- Another way to reduce dimensionality of data

- Redundant features
  - duplicate much or all of the information contained in one or more other attributes
  - Example: purchase price of a product and the amount of sales tax paid

- Irrelevant features
  - contain no information that is useful for the Data Mining task at hand
  - Example: students' ID should be irrelevant to the task of predicting students' Grade Point Average (GPA)

Feature Subset Selection Methods

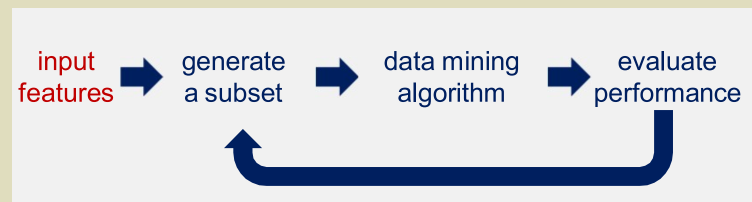
- <u>Exhaustive method</u>:
  - **Brute-force approach:**
    - Try all possible feature subsets as input to the Data Mining algorithm. Given n features, there are 2^n subsets.

input features: a, b, c

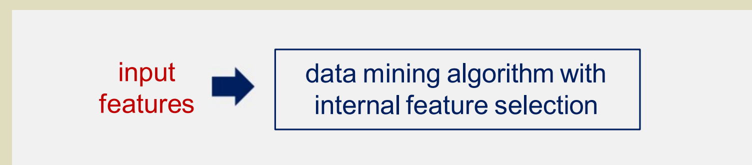
all feature subsets: {}
|power set|: $2^3 = 8$
{a}, {b}, {c}
{a, b}, {a, c}, {b, c}
{a, b, c}

- <u>Heuristic methods:</u>
  - **Filter approaches:**
    - Features are selected before Data Mining algorithm is run (e.g., correlation filter).

  - **Wrapper approaches:**
    - Use the Data Mining algorithm as a black box to find best subset of attributes (e.g., forward selection, backward elimination)



  - **Embedded approaches:**
    - Feature selection occurs as part of the Data Mining algorithm (e.g., LASSO)



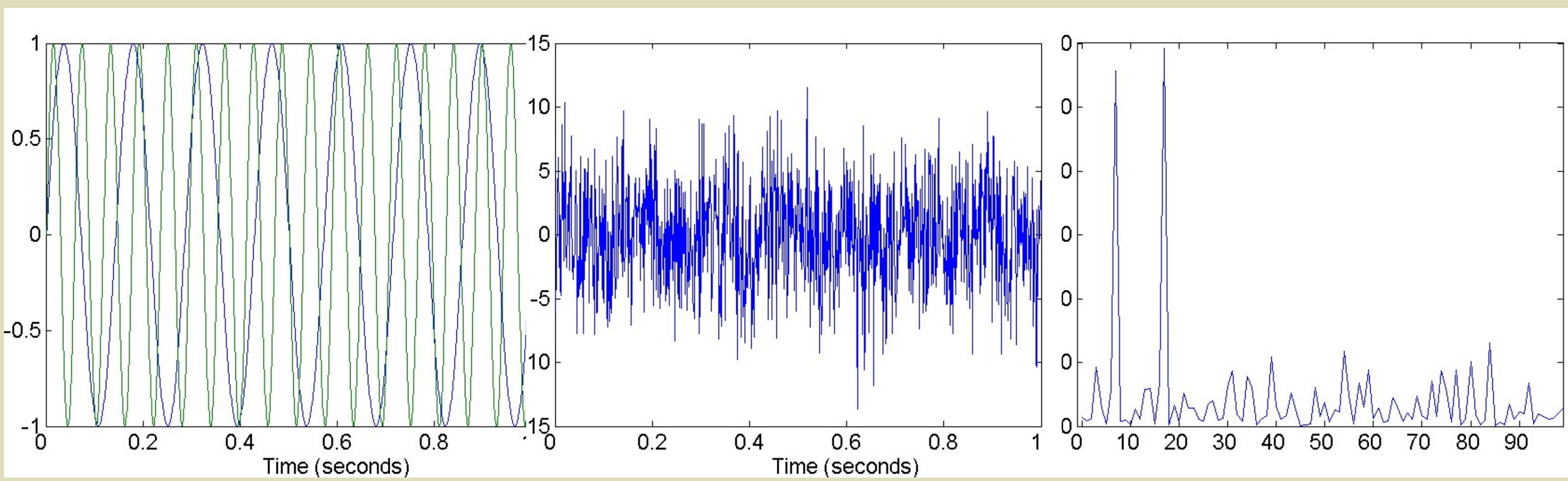
Feature Engineering

- Create new attributes that can capture the important information in a data set much more efficiently than the original attributes.

- Three general methodologies:
  - Feature Extraction
    - domain-specific
  - Mapping data to a new space
  - Feature construction
    - combining features

Example: Mapping Data to a New Space

- **Fourier transform**
- **Wavelet transform**



Two Sine Waves **Two Sine Waves + Noise** **Frequency**

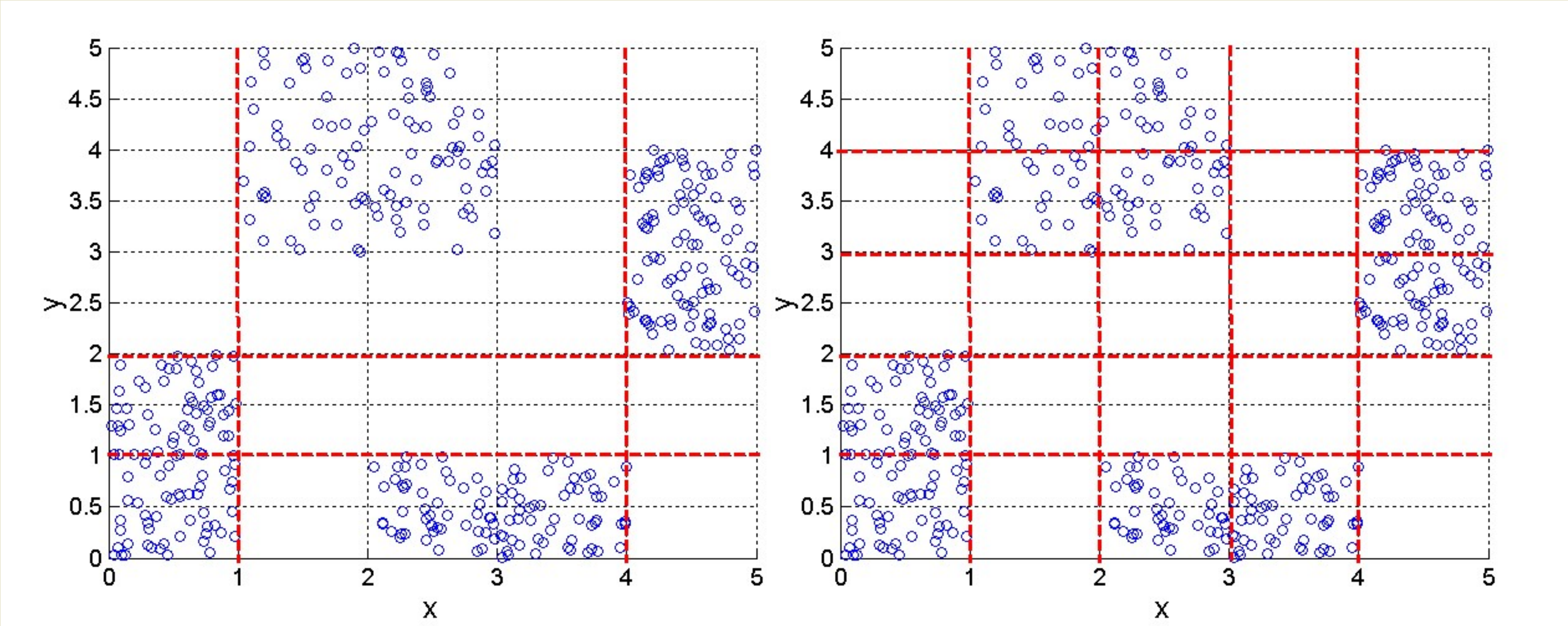
Discretization and Binarization

- Different Data Mining applications require specific data formats
  - Categorical only (discretization)
  - Binary only (binarization)
  - Interval/Ratio only (binarization)

- Discretization: transforming interval attribute into categorical

- Binarization: transforming non-binary attribute into a set of binary attributes

Discretization Using Class Labels



3 categories for both x and y **5 categories for both x and y**

Attribute Transformation

- A function that maps the entire set of values of a given attribute to a new set of replacement values such that each old value can be identified with one of the new values.
  - E.g., simple functions: x^k, $\log(x)$, e^x, $|x|$
  - Normalization and Standardization
    - ***Normalization*** is the transformation of the variable vectors into vectors of unit length.
    - ***Standardization*** transforms the variable vector into a vector of unit length, with a mean of zero and a standard deviation of one.

$$x' = \frac{(x - \mu_x)}{\sigma_x}$$

Next week:

- Introduction to KNIME