Data Mining Fundamentals

Topics

- Data and Data Types
- Data Quality
- Data Preprocessing
- Similarity and Dissimilarity
- Data Exploration and Visualization

What is Data?

- Collection of **objects** defined by **attributes**
- An **attribute** is a property or characteristic of an object
  - Examples: eye color, temperature, daily revenue
  - Variable, field, characteristic, feature, predictor, etc.
- A collection of attributes describe an **object**
  - Record, point, case, sample, entity, entry, instance, etc.

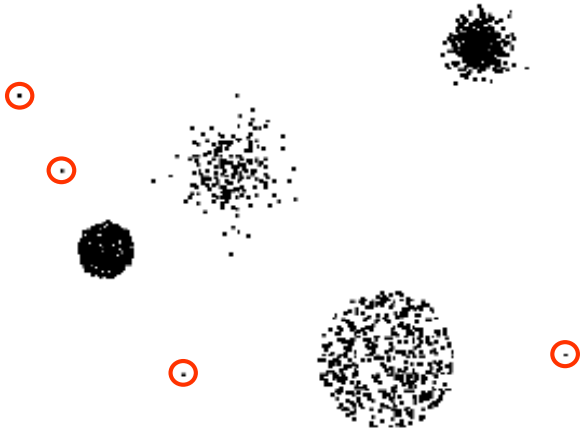
Attributes

Objects

Pid	Sex	Age	Pclass	Survived
2	Female	38	1	Yes
3	Female	26	3	Yes
5	Male	35	3	No
7	Male	54	1	No
13	Male	20	3	No
14	Male	39	3	No
21	Male	35	2	No
24	Male	28	1	Yes
34	Male	66	1	No
54	Female	29	2	Yes

Outliers

Data objects with characteristics that are considerably different than most of the other data objects in the data set



Types of Sampling

- Simple
  - There is an equal probability of selecting any particular item
- Stratified
  - Split the data into several partitions
  - Select fixed number of random samples from each partition
- Without replacement
  - As each item is selected, it is removed from the population
- With replacement
  - Objects are not removed from the population as they are selected for the sample
  - The same object can be selected more than once

End of Slide Sample

5 of 70 slides in presentation