

اسلام 25، ما وقتی داده‌ها را جمع کرده می‌بینیم، داده‌های تکراری می‌بینیم این Duplicate & Redundancy است، اینها کیفیت داده‌ها را پایین می‌آورند. مثلاً یک نفر ممکن است در یک جای چند بار نامش می‌آید در این سایت‌های اینترنتی. این باعث می‌شود داده‌ها را در اورب کند و این را دارد.

## Duplicate Data

- Data set may include data objects that are duplicates, or almost duplicates of one another
  - Major issue when merging data from heterogeneous sources
- Examples:
  - Same person with multiple email addresses
- Data cleaning
  - Process of dealing with duplicate data issues

اسلام 26 = Data Preprocessing چندین دارد که در زیر آمده است. 92, 8, 7, 5, 4, 3, 2, 1 Data mining

## Data Preprocessing

- 1- ● Aggregation
- 2- ● Sampling
- 3- ● Dimensionality Reduction
- 4- ● Feature subset selection
- 5- ● Feature creation
- 6- ● Discretization and Binarization
- 7- ● Attribute Transformation

Aggregation همان طور که از اسمش مشخص است، هدف از آن Data Reduction است یا تغییر اسکیل (Scale) است. یا پایداری یا ثبات بیشتر در داده‌ها است. در واقع Aggregation کاری که انجام می‌دهد این است که داده‌ها را به گونه‌ای که بتوانیم از آن‌ها استفاده کنیم، به گونه‌ای که بتوانیم از آن‌ها استفاده کنیم. مثلاً ما می‌خواهیم میزان برق مصرفی را در نردگام‌ها ببینیم، لیکن راه این است که اطلاعاتی که به ما می‌دهند چگونه است؟! مصرف برق خانگی، مصرف برق معابر، مصرف برق صنعتی، چیدمان این معارف برق را می‌دهند. آنرا اینها را چیدمان می‌کنند.









**توضیح ۱- SRS** ساده ترین نوع نمونه گیری است، یعنی ما می خواهیم از یک جامعه نمونه گیری کنیم، هر آنیم یک فهرستی درست می کنیم، به صورت یک برنامه کامپیوتری یا از اعداد تصادفی استفاده می کنیم، البته در داده های استفاده از این امکانی غیر ممکن است. و بعد از آن ها انتخاب می کنیم. SRS سراسری ترین نوع نمونه گیری است، و معمولاً حاشیای خیلی راحت است خیلی راحت می تواند از این فواید ها استفاده کرد اما ۲ ایراد دارد: همان ۱) آسانی که نمونه گیری داشته باشد: ۱) هزینه ای بالا است ۲) دسترسی کم است. فرض کنیم فرض داریم که تعداد کل تهران را می خواهیم بررسی کنیم، وزارت آموزش و پرورش جزو وزارت های است که

## Types of Sampling

- 1- ● Simple Random Sampling (SRS)
  - There is an equal probability of selecting any particular item

- 2- ● Sampling without replacement (SWR)
  - As each item is selected, it is removed from the population

- 3- ● 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

- 4- ● Stratified sampling
  - Split the data into several partitions; then draw random samples from each partition

© Tan, Steinbach, Kumar

Introduction to Data Mining

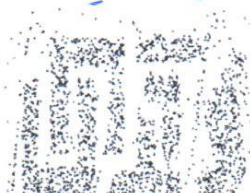
4/18/2004

31

## Sample Size



8000 points



2000 Points



500 Points

© Tan, Steinbach, Kumar

Introduction to Data Mining

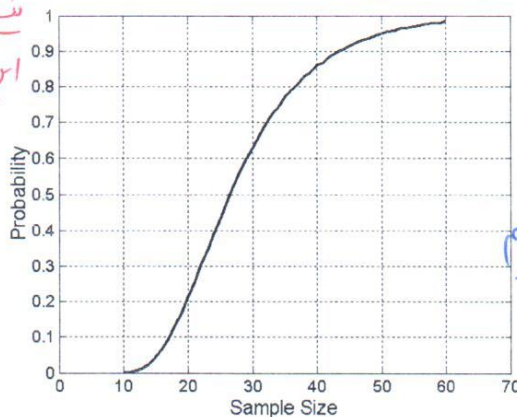
4/18/2004

32



## Sample Size

- احتمال افزائش پسر اقلند.



© Tan, Steinbach, Kumar

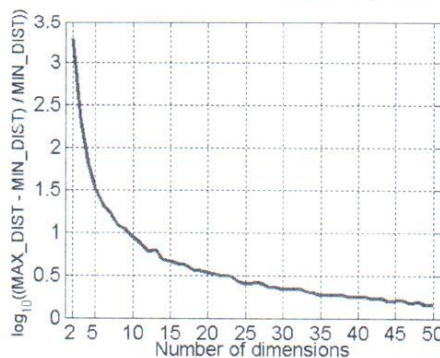
Introduction to Data Mining

4/18/2004

33

## 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



- Randomly generate 500 points
- Compute difference between max and min distance between any pair of points

© Tan, Steinbach, Kumar

Introduction to Data Mining

4/18/2004

34