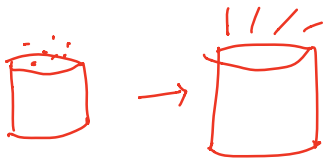
15/01/21 Lec-08 Chp-03 Data Preprocessing (from Han & Kamber, 3rd edn)

Major steps of data preprocessing :

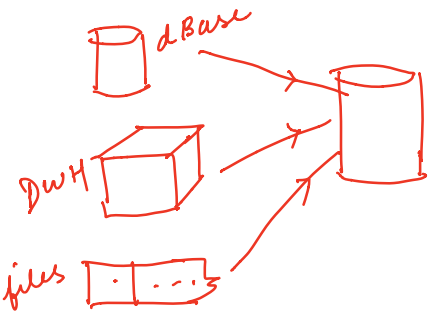
- major tasks (overview)
- data cleaning
- data integration
- data reduction
- data transformation and data discretization .

Data Quality : 1) accuracy 2) completeness 3) consistency 4) timeliness 5) believability 6) interpretability

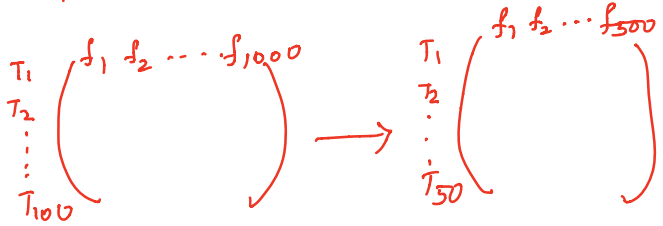
Major tasks (overview) :



→ data cleaning (filling missing values, smooth noisy data, identify and remove outliers, & resolve inconsistencies)



→ data integration



→ data reduction

<-2, 32, 100, 59, 48> → <-0.02, 0.32, 1, 0.59, 0.48>
0.06

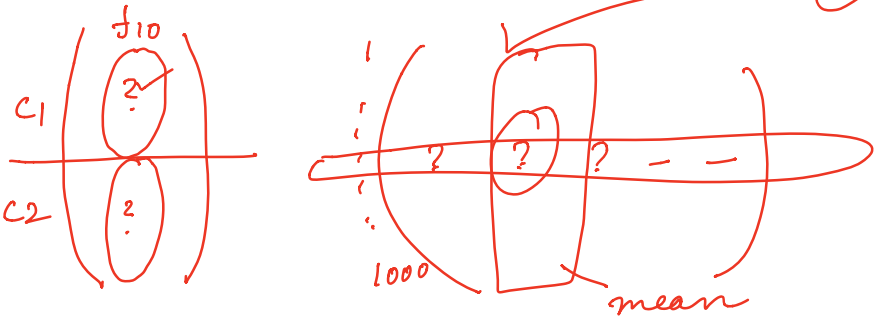
↓ data transformation

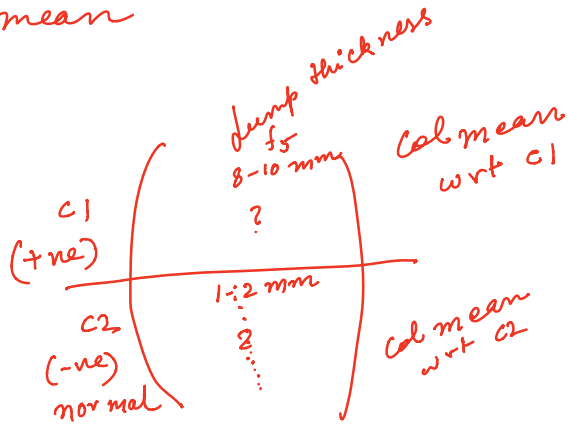
<A C TG A C C T G G> → ML → (f_1 f_2 ... f_n)

→ DATA CLEANING :

- missing values
  1. Ignore the tuple
  2. Fill in the missing values manually.
  3. Use a global constant to fill in missing values.
  4. Use a measure of central tendency of attribute (mean/median)
  5. Use attribute mean/median for all samples belonging to same class.
  6. Use the most probable value to fill in missing value.

- noisy data
  1. Binning (creating buckets)
     - Smooth bins by means
     - Smooth bins by median
     - Smooth by boundaries
  2. Regression
  3. Clustering





Bin 1 | Bin 2 | Bin 3

Binning: $\langle 4, 8, 15, | 21, 21, 24, | 25, 28, 34 \rangle$

→ Partition into equal frequency buckets.

Bin 1: $\langle 4 \ \ 8 \ \ 15 \rangle$ Bin 2 $\langle 21 \ \ 21 \ \ 24 \rangle$ Bin 3 $\langle 25 \ \ 28 \ \ 34 \rangle$

By boundaries: check min & max for a bin and identify the boundaries. Each bin value is then replaced by closest boundary value.

Bin 1 $\langle 4 \ \ 4 \ \ 15 \rangle$ Bin 2 $\langle 21 \ \ 21 \ \ 24 \rangle$ Bin 3 $\langle 25 \ \ 25 \ \ 34 \rangle$

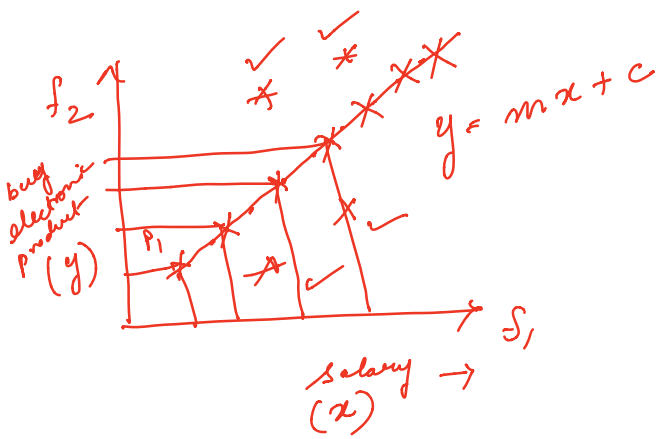
By mean: Each value is replaced by mean

Bin 1 $\langle 9 \ \ 9 \ \ 9 \rangle$ Bin 2 $\langle 22 \ \ 22 \ \ 22 \rangle$ Bin 3 $\langle 29 \ \ 29 \ \ 29 \rangle$

By median: Bin 1 $\langle 4 \ \ 4 \ \ 4 \rangle$ Bin 2 $\langle 21 \ \ 21 \ \ 21 \rangle$

Bin 3 $\langle 25 \ \ 25 \ \ 25 \rangle$

(2) **Regression:**



(3) **Clustering:**