Data Preprocessing

Dr Satyaranjan Jena

Agenda

- Why data preprocessing?
- Data cleaning
- Data integration and transformation
- Data reduction
- Discretization and concept hierarchy generation
- Summary

Why Data Preprocessing?

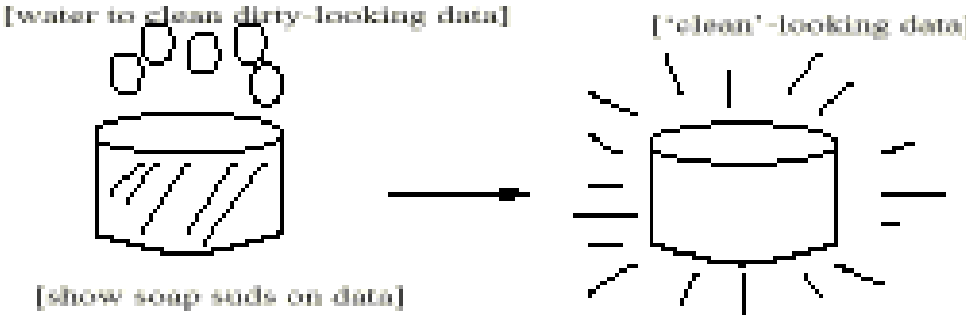
- Data in the real world is dirty
  - incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
  - noisy: containing errors or outliers
  - inconsistent: containing discrepancies in codes or names
- No quality data, no quality mining results!
  - Quality decisions must be based on quality data
  - Data warehouse needs consistent integration of quality data
- A multi-dimensional measure of data quality:
  - A well-accepted multi-dimensional view:
    - accuracy, completeness, consistency, timeliness, believability, value added, interpretability, accessibility
  - Broad categories:
    - intrinsic, contextual, representational, and accessibility.

Major Tasks in Data Preprocessing

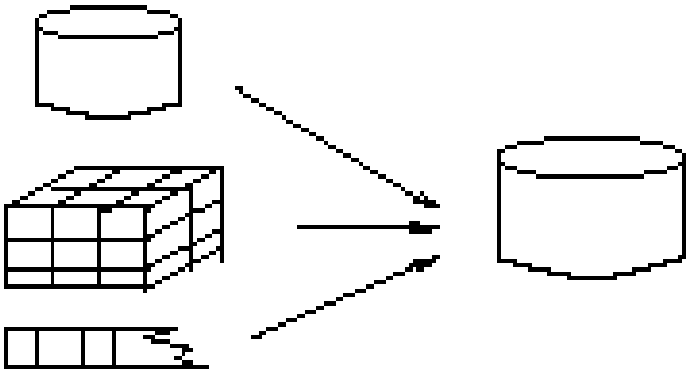
- Data cleaning
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- Data integration
  - Integration of multiple databases, data cubes, files, or notes
- Data transformation
  - Normalization (scaling to a specific range)
  - Aggregation
- Data reduction
  - Obtains reduced representation in volume but produces the same or similar analytical results
  - Data discretization: with particular importance, especially for numerical data
  - Data aggregation, dimensionality reduction, data compression,generalization

Forms of data preprocessing

Data Cleaning



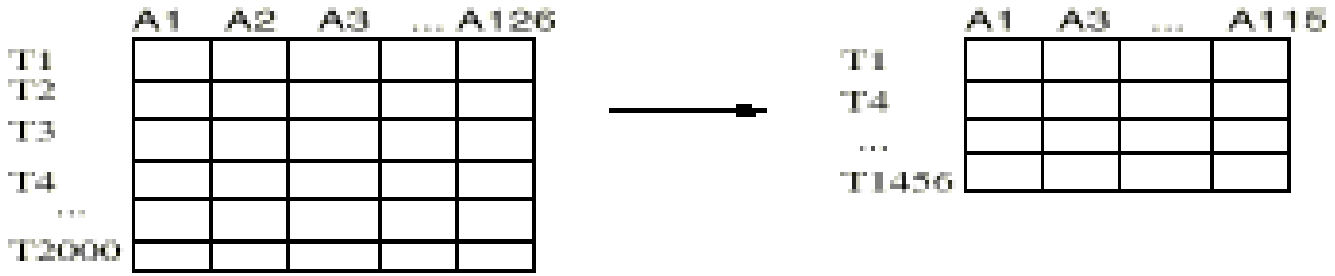
Data Integration



Data Transformation

-2, 32, 100, 59, 48 → -0.02, 0.32, 1.00, 0.59, 0.48

Data Reduction



Agenda

- Why preprocess the data?
- Data cleaning
- Data integration and transformation
- Data reduction
- Discretization and concept hierarchy generation
- Summary

Data Cleaning

- Data cleaning tasks
  - Fill in missing values
  - Identify outliers and smooth out noisy data
  - Correct inconsistent data

Missing Data

- Data is not always available
  - E.g., many tuples have no recorded value for several attributes, such as customer income in sales data
- Missing data may be due to
  - equipment malfunction
  - inconsistent with other recorded data and thus deleted
  - data not entered due to misunderstanding
  - certain data may not be considered important at the time of entry
  - not register history or changes of the data
- Missing data may need to be inferred

How to Handle Missing Data?

- Ignore the tuple: usually done when class label is missing (assuming the task is classification—not effective in certain cases)
- Fill in the missing value manually: tedious + infeasible?
- Use a global constant to fill in the missing value: e.g., “unknown”, a new class?!
- Use the attribute mean to fill in the missing value
- Use the attribute mean for all samples of the same class to fill in the missing value: smarter
- Use the most probable value to fill in the missing value: inference-based such as regression, Bayesian formula, decision tree

Noisy Data

- Q: What is noise?
- A: Random error in a measured variable.
- Incorrect attribute values may be due to
  - faulty data collection instruments
  - data entry problems
  - data transmission problems
  - technology limitation
  - inconsistency in naming convention
- Other data problems which requires data cleaning
  - duplicate records
  - incomplete data
  - inconsistent data

How to Handle Noisy Data?

- Binning method:
  - first sort data and partition into (equi-depth) bins
  - then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
  - used also for discretization (discussed later)
- Clustering
  - detect and remove outliers
- Semi-automated method: combined computer and human inspection
  - detect suspicious values and check manually
- Regression
  - smooth by fitting the data into regression functions

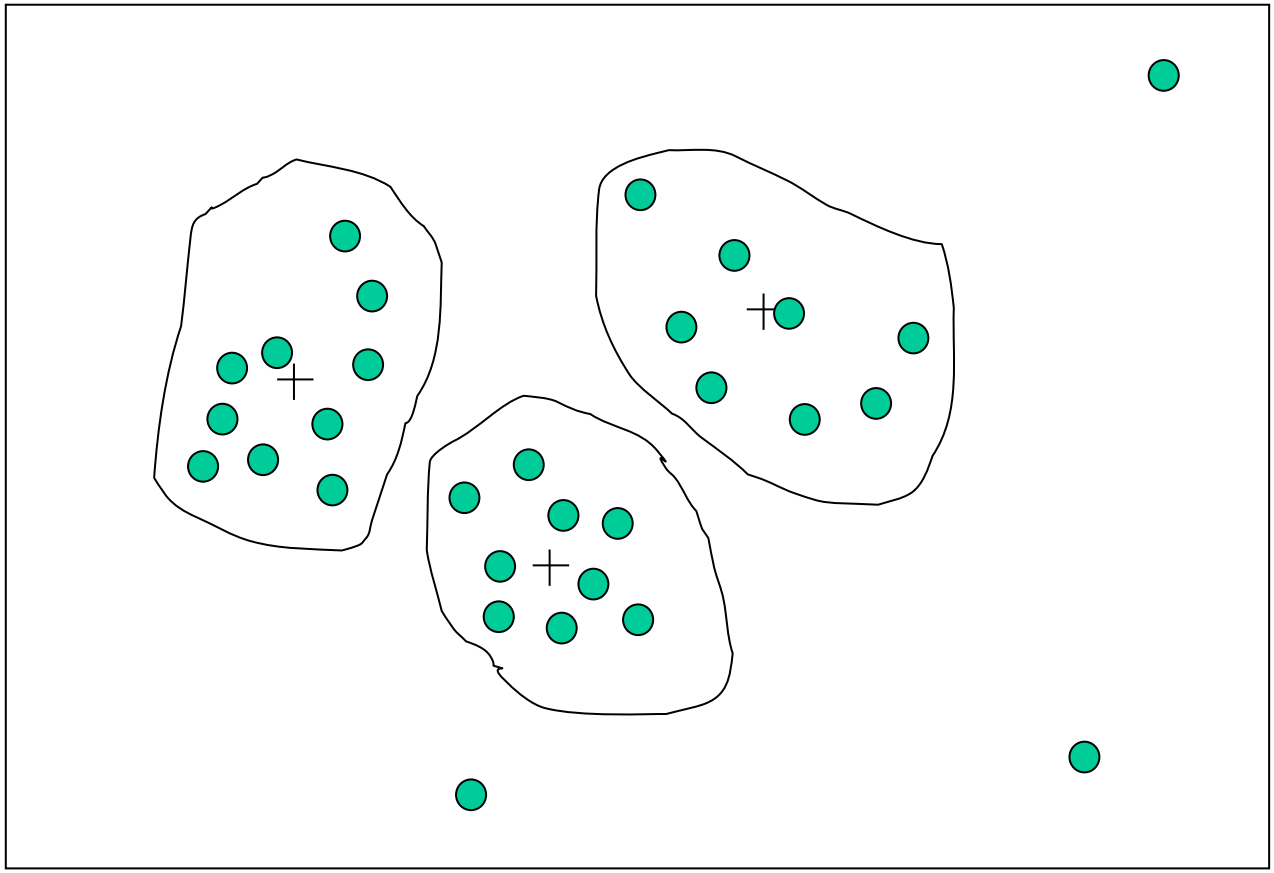
Simple Discretization Methods: Binning

- Equal-width (distance) partitioning:
  - It divides the range into *N* intervals of equal size: uniform grid
  - if *A* and *B* are the lowest and highest values of the attribute, the width of intervals will be: $W = (B\text{-}A)/N$.
  - The most straightforward
  - But outliers may dominate presentation
  - Skewed data is not handled well.
- Equal-depth (frequency) partitioning:
  - It divides the range into *N* intervals, each containing approximately same number of samples
  - Good data scaling
  - Managing categorical attributes can be tricky.

Binning Methods for Data Smoothing

* Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
* Partition into (equi-depth) bins:
  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25
  - Bin 3: 26, 28, 29, 34
* Smoothing by bin means:
  - Bin 1: 9, 9, 9, 9
  - Bin 2: 23, 23, 23, 23
  - Bin 3: 29, 29, 29, 29
* Smoothing by bin boundaries:
  - Bin 1: 4, 4, 4, 15
  - Bin 2: 21, 21, 25, 25
  - Bin 3: 26, 26, 26, 34

Cluster Analysis



Let us go through the above steps using the example below.

1.Consider 4 data points A,B,C,D as below

2.Choose two centroids AB and CD, calculated as

AB = Average of A, B

CD = Average of C,D

Cluster

	X1	X2
A	2	3
B	6	1
C	1	2
D	3	0

Observations

Cluster Analysis

	X1	X2
AB	4	2
CD	2	1

3. Calculate squared euclidean distance between all data points to the centroids AB, CD. For example distance between A(2,3) and AB (4,2) can be given by $s = (2–4)^2 + (3–2)^2$.

	A	B	C	D
AB	5	5	9	5
CD	4	16	2	2

A is very near to CD than AB

Cluster Analysis

4. If we observe in the fig, the highlighted *distance between (A, CD) is 4 and is less compared to (AB, A) which is 5. Since point A is close to the CD we can move A to CD cluster.*

5. There are two clusters formed so far, let recompute the centroids i.e, B, ACD similar to step 2.

$$ACD = Average\ of\ A, C, D$$

$$B = B$$

6. As we know K-Means is iterative procedure now we have to calculate the distance of all points (A, B, C, D) to new centroids (B, ACD) similar to step 3.

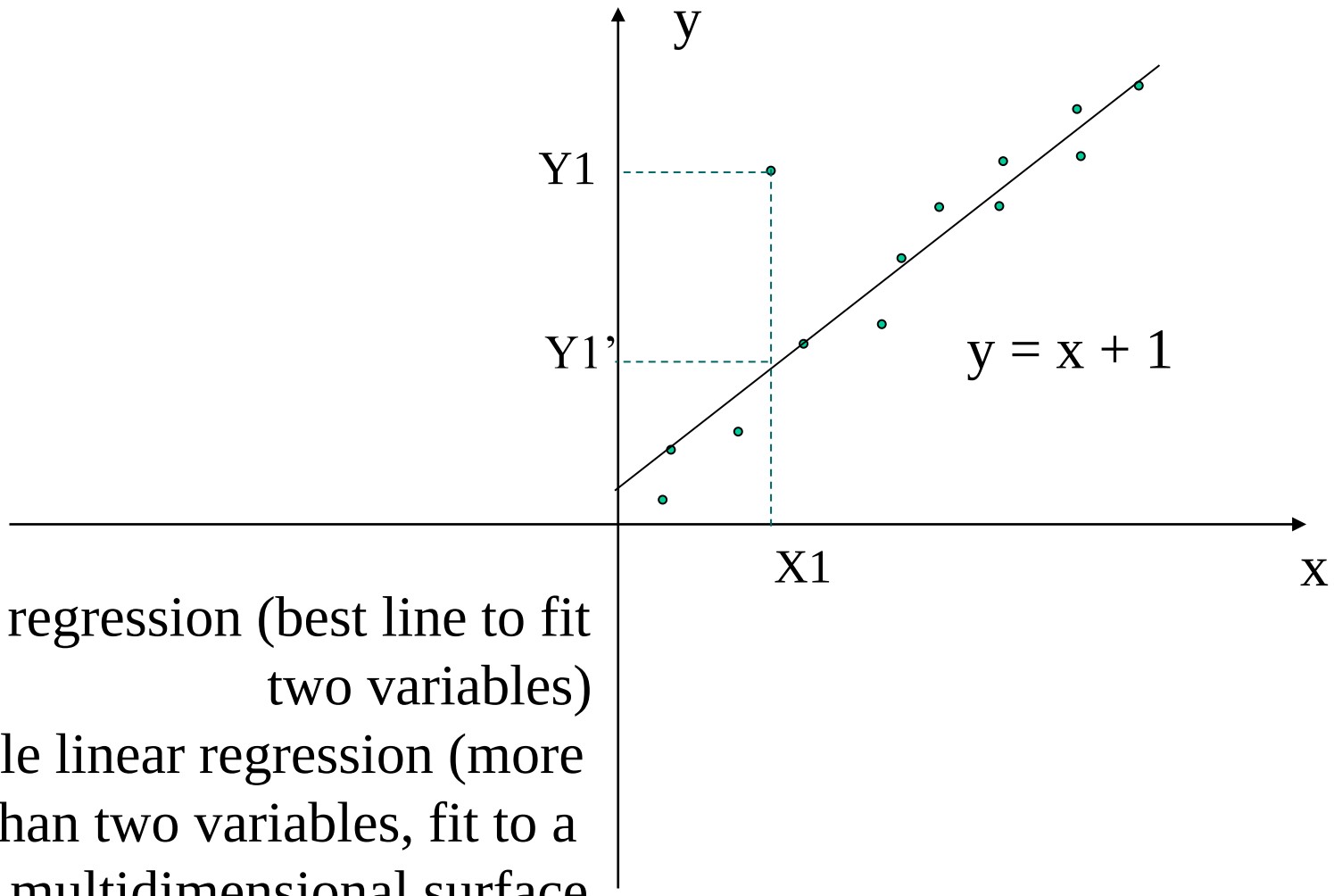
	X1	X2
B	6	1
ACD	2	1.67

New centroids B, ACD

7. In the above picture, we can see respective cluster values are minimum that A is too far from cluster B and near to cluster ACD. All data points are assigned to clusters (B, ACD) based on their minimum distance. The iterative procedure ends here.

8. To conclude, we have started with two centroids and end up with two clusters, K=2.

Regression



- Linear regression (best line to fit two variables)
- Multiple linear regression (more than two variables, fit to a multidimensional surface

Regression

Linear Regression Formula

As we know, linear regression shows the linear relationship between two variables. The equation of linear regression is similar to that of the slope formula. We have learned this formula before in earlier classes such as a linear equation in two variables. Linear Regression Formula is given by the equation

Y= a + bX

We will find the value of a and b by using the below formula

Y= a + bX

We will find the value of a and b by using the below formula

$$\mathbf{a=}\ \frac{(\sum_Y)(\sum_{X^2}) - (\sum_X)(\sum_{XY})}{n(\sum_{x^2}) - (\sum_x)^2}$$

$$\mathbf{b=}\ \frac{n(\sum_{XY}) - (\sum_X)(\sum_Y)}{n(\sum_{x^2}) - (\sum_x)^2}$$

Regression

- **Problem 3**

The values of y and their corresponding values of y are shown in the table below

x	0	1	2	3	4
y	2	3	5	4	6

a) Find the least square regression line $y = a x + b$.

b) Estimate the value of y when $x = 10$.

Regression

3. a) We use a table to calculate a and b.

x	y	x y	x^2
0	2	0	0
1	3	3	1
2	5	10	4
3	4	12	9
4	6	24	16
$\Sigma x = 10$	$\Sigma y = 20$	$\Sigma x\,y = 49$	$\Sigma x^2 = 30$

Regression

We now calculate a and b using the least square regression formulas for a and b.

$a = (n\Sigma x\, y - \Sigma x \Sigma y) / (n\Sigma x^2 - (\Sigma x)^2) = (5*49 - 10*20) / (5*30 - 10^2) = 0.9$

$b = (1/n)(\Sigma y - a\, \Sigma x) = (1/5)(20 - 0.9*10) = 2.2$

b) Now that we have the least square regression line $y = 0.9\, x + 2.2$, substitute x by 10 to find the value of the corresponding y.

$y = 0.9 * 10 + 2.2 = 11.2$

How to Handle Inconsistent Data?

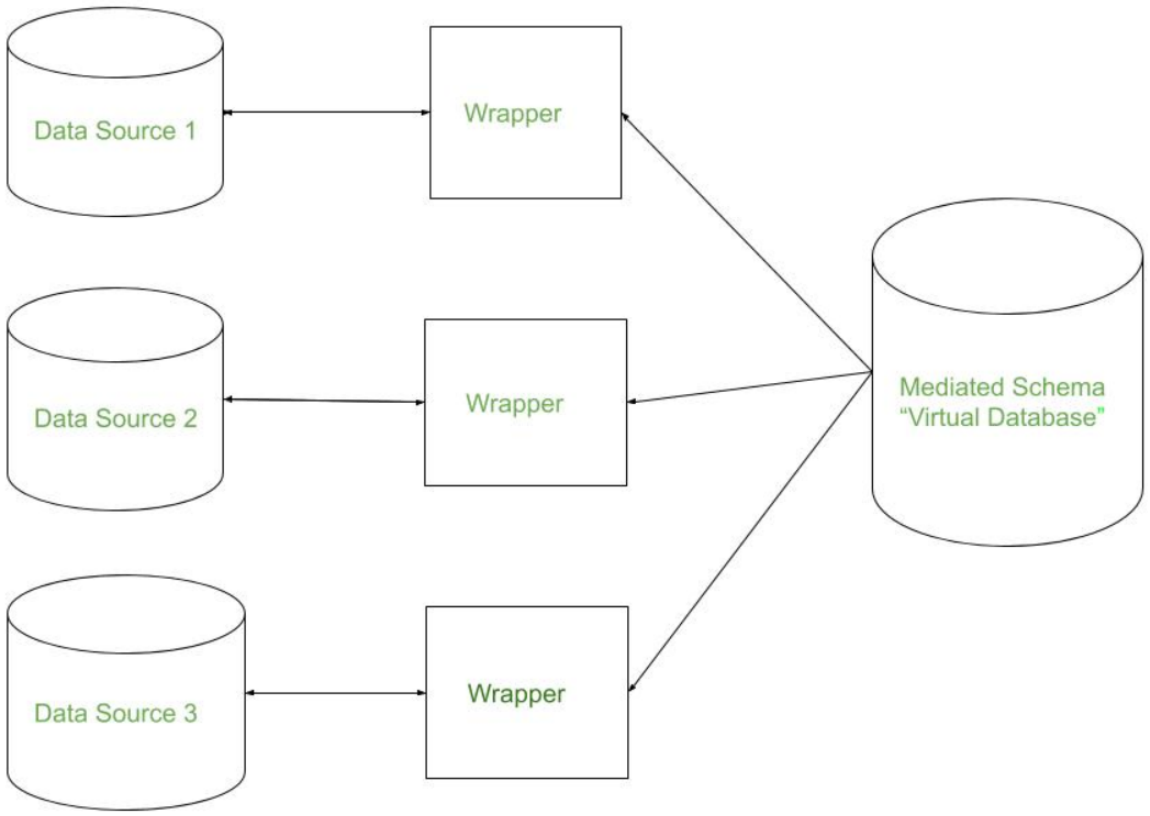
- Manual correction using external references
- Semi-automatic using various tools
  - To detect violation of known functional dependencies and data constraints
  - To correct redundant data

Agenda

- Why preprocess the data?
- Data cleaning
- Data integration and transformation
- Data reduction
- Discretization and concept hierarchy generation
- Summary

Data Integration

Data integration is a process where data from many sources goes to a single centralized location, which is often a data warehouse. The end location needs to be flexible enough to handle lots of different kinds of data at potentially large volumes. Data integration is deal for powering analytical use cases.



Data Integration

Issues in Data Integration:

There are three issues to consider during data integration: Schema Integration, Redundancy Detection, and resolution of data value conflicts. These are explained in brief below.

1. **Schema Integration:**
   - Integrate metadata from different sources.
   - The real-world entities from multiple sources are referred to as the entity identification problem.

2. Redundancy:

- An attribute may be redundant if it can be derived or obtained from another attribute or set of attributes.
- Inconsistencies in attributes can also cause redundancies in the resulting data set.
- Some redundancies can be detected by correlation analysis.

Data Integration

3. Detection and resolution of data value **conflicts:**

- This is the third critical issue in data integration.
- Attribute values from different sources may differ for the same real-world entity.
- An attribute in one system may be recorded at a lower level of abstraction than the “same” attribute in another.

Handling Redundant Data in Data Integration

- Redundant data occur often when integrating multiple DBs
  - The same attribute may have different names in different databases
  - One attribute may be a "derived" attribute in another table, e.g., annual revenue
- Redundant data may be able to be detected by correlational analysis

$$r_{A,B} = \frac{\Sigma(A-\overline{A})(B-\overline{B})}{(n-1)\sigma_A\sigma_B}$$

- Careful integration can help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

Data Transformation

- Smoothing: remove noise from data (binning, clustering, regression)
- Aggregation: summarization, data cube construction
- Generalization: concept hierarchy climbing
- Normalization: scaled to fall within a small, specified range
  - min-max normalization
  - z-score normalization
  - normalization by decimal scaling
- Attribute/feature construction
  - New attributes constructed from the given ones

Data Transformation: Normalization

Particularly useful for classification (NNs, distance measurements, nn classification, etc)

- min-max normalization

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

- z-score normalization

$$v' = \frac{v - mean_A}{stand\_dev_A}$$

- normalization by decimal scaling

$$v' = \frac{v}{10^j}$$ Where j is the smallest integer such that Max($|v'|$)<1