Chapter - 2

Data - Preprocessing

→ Raw data contained in databases is unpreprocened, incomplete and noisy.

→ Major reasons:

Data preprocessing is the procen of preparing raw data for analysis by cleaning and transforming it into a usable format.

→ Real datasets often having missing or incomplete entries.

→ Data may contain typos, duplicates or outliers.

→ Different features may have different scales not suitable for data mining models.

→ Some features are irrelevant, redundant or highly correlated.

→ values not consistency with policy or common sense.

→ Steps in Data Preprocessing :-

→ Data cleaning

→ Data Integration

→ Data Transformation

→ Data Reduction.

Data cleaning :-

→ It is the procen of identifying and correcting errors, or inconsistencies in the dataset.

→ It involves handling missing values, removing duplicates, and correcting incorrect or outlier data to ensure the dataset is accurate & reliable.

→ Clean data is essential for effective analysis, as it improves the quality of results & enhances the performance of data models.

A. Missing values : When data is absent from a dataset.

Sol?

→ Remove records with missing values.

→ Fill with mean / median / mode (imputation)

Noisy Data : It refers to irrelevant or incorrect data.

→ use smoothing techniques (Binning, moving avg.)

[Binning :- Data is sorted into equal segments and each segment is smoothed by replacing values with mean or boundary values]

→

1. Ignores the tuple :-

This is usually done when the class label is missing. This method is not very effective, unless the tuple contains several attributes with missing values.

2. Fill in the missing value manually :-

In general, this approach is time-consuming may not be feasible given a large dataset with many missing values.

2. a global constant to fill in the missing value.

Replace all missing attributes value by the same constant such as label like "Unknown" or "-∞".

4. use the attribute mean to fill in the missing value.

5. use the " " for all samples belonging to the same class as the given tuple.

6. use the most probable value to fill in the missing value.

Eg DT, regression & Bayesian formalism.

B. Noisy Data

→ Noise is a random error or variance in a measured variable.

→ Incorrect attribute values may due to
- → Faulty data collection instruments.
- → data entry problems
- → data transmission problems
- → technology limitations.
- → inconsistency in naming convention.

→ Other data problems.
- → Duplicate records
- → Incomplete data
- → Inconsistent "

→ (i) Binning method

→ First sort data and the sorted values are distributed into a number of "buckets", or "bins".

→ Then one can smooth by bin means, smooth by bin median, smooth by bin boundaries etc.

(i) By Bin means

Each value in a bin is replaced by the mean value of the bin

(ii) By Bin medians :

→ Each bin value is replaced by the bin median.

(iii) By bin boundaries :

→ minimum & maximum values in a given bin are identified as the bin boundaries.

→ Each bin value is then replaced by the closest boundary value.

→ In general, the larger the width, the greater effect of smoothing. So bins may be equal-width, or equal depth (frequency) partitioning : (distance)

Equal width (distance) partitioning :

→ It divides the range into N intervals of equal size

→ IF A & B are lowest & highest values of the attribute, the width of intervals will be $W = (B-A)/N$.

→ The most straight forward.

→ 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.

Q Sorted data for Price (in dollars):

4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

→ Partition into (equi-depth / equal-frequency) 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 medians:

Bin 1: 8.5, 8.5, 8.5, 8.5

Bin 2: 22.5, 22.5, 22.5, 22.5

Bin 3: 28.5, 28.5, 28.5, 28.5

→ Smoothing by bin boundaries:

Bin 1: 4, 4, 4, 15

Bin 2: 21, 21, 25, 25

Bin 3: 26, 26, 26, 34

Sol^n

Partition into (equi-width) bins:

lowest value A = 4, Highest value B = 34

S-1 Range = 34 - 4 = 30

S-2 width of interval = $\frac{30}{3} = 10$

↳ Here choosing 3 bin.

S-3 Bin intervals

Starting at min = 4

Bin 1: [4, 14)

Bin 2: [14, 24)

Bin 3: [24, 34)

S-4 Assign Data to Bins

Bin 1: 4, 8, 9

Bin 2: 15, 21, 21

Bin 3: 24, 25, 26, 28, 29, 34

Q want Equi-frequency

S-1 Data must be sorted.

S-2 take 3-bins.
Total value = 12
Each bin should have 12 ÷ 3 = 4 values.

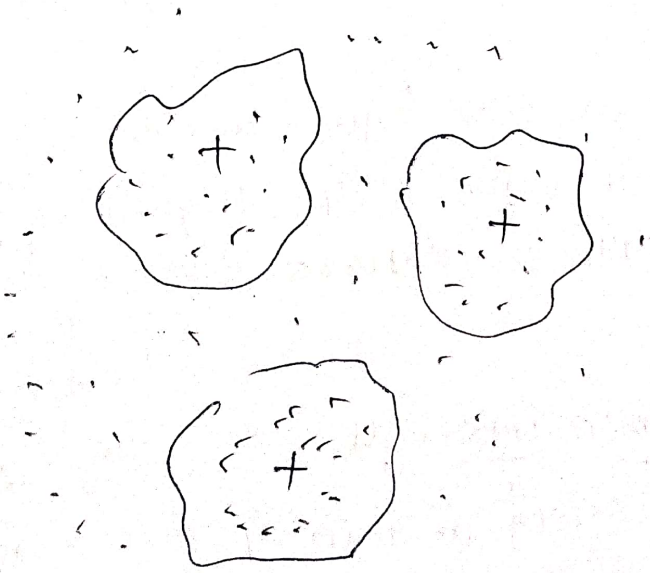
S-3 Bin 1 : {4, 8, 9, 15}
Bin 2 : {21, 21, 24, 25}
Bin 3 : {26, 28, 29, 34}

2. Regression : Data can be smoothed by fitting the data to a function such as regression.

(1) Linear regression : best line to fit two attributes (Dependent & independent variables)

(2) Multiple linear regression : It is an extension of linear regression. where more than two attributes are involved.

3. clustering : —



ntifying misclassifications

Misclassification means that a model predicts the wrong class label compared to the actual (true) class label in your dataset.

Brand	Frequency
USA	1
France	1
us	156
Europe	46
Japan	51

(Notice strange about this frequency distribution)

→ The frequency distribution shows 5 classes. However, two of the classes USA & France have count of only one each. what is clearly happening here is that two of the records have been inconsistently classified w.r.t. to the origin of manufacture.

→ To maintain consistency, the record have been labeled us & europe instead of USA & France.

Graphical methods for identifying outliers:

→ outliers are extreme values that go against the trend of remaining data.

→ Graphical methods for identifying outliers for numerical variables is to use "histogram" of the variable.

Q 45, 50, 52, 48, 47, 49, 51, 46, 500, 50, 48, 49, 5..

S-1 45, 46, 47, 47, 48, 48, 49, 49, 50, 50, 51, 51, 52, 52, ..

S-2 $Q_2 = 49$, $Q_1 = 47$, $Q_3 = 51$

S-3 $IQR = Q_3 - Q_1 = 51 - 47 = 4$

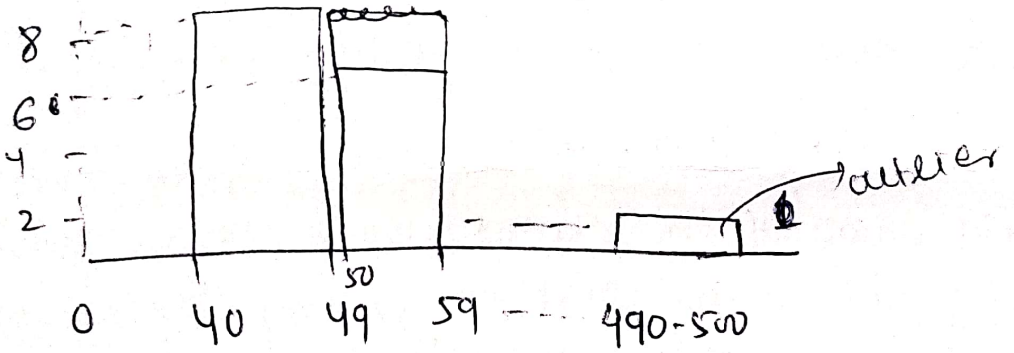
S-4 Outlier

lower Bound $= Q_1 - 1.5 \times IQR = 41$

upper " $= Q_3 + 1.5 \times IQR = 57$

S-5 anything $> 57 \rightarrow$ outlier

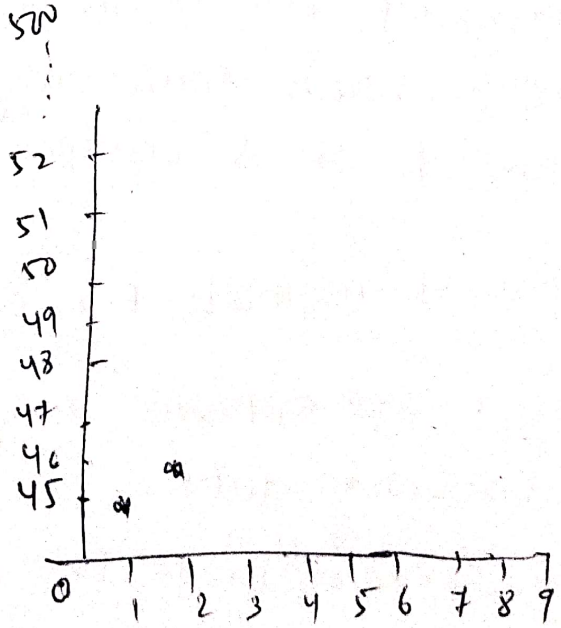
$\Rightarrow 500 > 57 \rightarrow$ "



outlier

Scatter plot

Student	score
1	45
2	50
3	52
4	48
5	47
.	
.	
.	
15	51



* Measures of center & spread

center (mean) $= \bar{x} = \frac{\Sigma x}{n}$ & Spread (SD) $= \sqrt{\frac{\Sigma (x-\bar{x})^2}{n-1}}$

Data Integration

Data Integration - the merging of data from multiple data sources. It helps to reduce and avoid redundancies and inconsistencies in the resulting dataset which improve the accuracy and speed of the subsequent data mining procen.

→ Entity Identification Problem ⇒

→ It occurs when we try to determine whether two or more records from different data sources refer to the same real-world entity.

Solve
→ Rule based matching
→ Probabilistic matching. (Smillarity score)
→ Machine learning model.

Data Transformation

→ It is the process of converting data from one format, structure or value representation to another to ensure compatibility, consistency and better quality for analysis.

Types

1. Smoothing :-
   → Removes noise from data.
   → Ex Binning, regression & clustering.

2. Aggregation
   Summarizing data.
   Ex Daily sales data → Monthly sales totals.

3. Generalization :-

Replacing detailed data with higher-level concepts.

Eg: 23 years → "young Adult" or "youth" / middle-aged & senior — Age }

4. Normalization (scaling)

Rescaling values to a standard range (Eg: [0, 1] or z-score)

Eg: converting students marks from 0-100 to 0-1 scale.

5. Attribute construction (Feature Engineering)

Creating new attributes from existing ones.

Eg: from DOB create "Age"

MIN-MAX NORMALIZATION

Min-max normalization works by performing a linear transformation on the original data.

$$x'_{mm} = \frac{x - min(x)}{range(x)} = \frac{x - min(x)}{max(x) - min(x)}$$

where x = original value

$min(x)$ = minimum value of the attribute

$max(x)$ = maximum value of the attribute

x' = normalized value (b/w 0 & 1)

1st row 50, 60, 70, 80, 90, 100

min = 50, max = 100

For normalize 80:

$$x' = \frac{80-50}{100-50} = \frac{30}{50} = 0.6$$

So the Normalized value of 80 is 0.6.

Q suppose min & max values for the attribute income are $12,000 & $98,000 respectively. we would like to map income to the range [0.0, 1.0]. By min-max normalization, a value of $73,600 for income is transformed to ____

Sol?

$$\frac{73,600 - 12,000}{98,000 - 12,000} = 0.716$$

Q minimum weight of 1613 pounds & max = 4997 pounds

range = 4997 − 1613 = 3384 pounds.

(i) Find min-max normalization weight for a ultra weight vehicle of 1613 pounds.

$$X_{mm} = \frac{X - \min(X)}{\text{range}(X)} = \frac{1613 - 1613}{3384} = 0$$

(ii) For midrange $= \frac{\max(X) + \min(X)}{2} = \frac{4997 + 1613}{2} = 3305$ pounds

So min-max normalization for midrange is

$$X_{mm} = \frac{3305 - 1613}{3384} = 0.5$$

For heaviest vehicle.

$$X_{mm} = \frac{4497 - 1613}{3384} = 1$$

Z-Score Standardization

Already done in appendix

Decimal Scaling :-

Decimal scalling is a normalization technique where the normalized value lies between $-1 \leq 1$.

$$x^*_{decimal} = \frac{x}{10^d}$$

where x = original value
x^* = Normalized value
d = smallest integer or no. of digits in the data value such that $|x^*| < 1$ for all values.

Q $-987, -120, 56, 300, 850$

Sol^n maximum absolute value is 987
to make it less than 1, divide by 10^3 ($987 < 1000$)
So $d = 3$

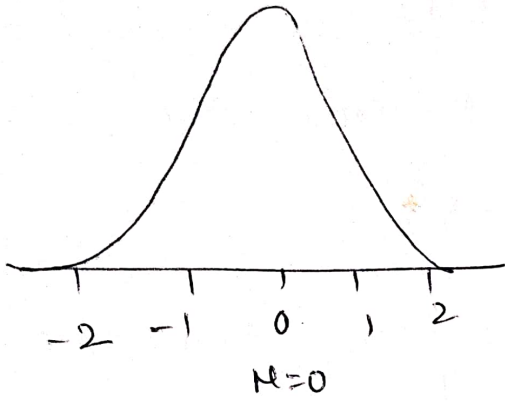
Now $$x^* = \frac{x}{1000}$$

original	Normalized
-987	-0.987
-120	-0.120
56	0.056
300	0.300
850	0.850

ie Normalized value lies between $-1 \leq 1$

Transformation to Achieve Normality :-

The normal distribution is a continuous probability distribution commonly known as bell curve, which is symmetric.



→ For standard normal distribution z has mean $\mu = 0$ & SD (σ) = 1 but distribution may still be skewed.

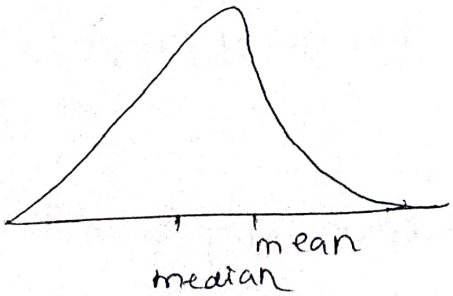
$$\text{Skewness} = \frac{3(\text{mean} - \text{median})}{\text{Standard deviation}}$$

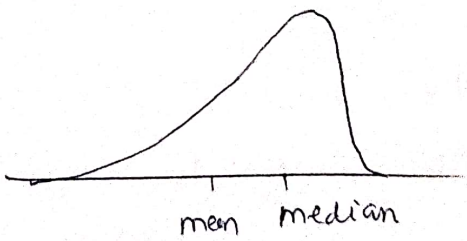
For weight

$$\text{Skewness} = \frac{3(3095.490 - 2835)}{852.646} = 0.6$$

For weight - z

$$\text{Skewness} = \frac{3(0 - (-0.2))}{1} = 0.6$$





mean > median → positive skewness

mean < median
⇒ Negative skewness

→ To eliminate skewness apply a transformation to the data.

→ common transformations are (i) natural log transformation

$$\bar{x} = \log(x+1)$$

(ii) Square root transformation.

$$\bar{x} = \sqrt{x}$$

(iii) inverse square root transformation.

$$\bar{x} = \frac{1}{\sqrt{x}}$$

Flag Variables

→ Flag variables (dummy or indicator variable) is a categorical variable taking only two values. 0 & 1.

Eg: categorical predictor sex. (female & male)

So flag variable as sex-flag.

IF sex = female then sex-flag = 0

IF sex = male then sex-flag = 1.

Eg: categorical predictor region has k=4 (north, east, south, west)

then

north-flag : IF region = north then north-flag = 1, otherwise north-flag = 0

east-flag : IF region = east then east-flag = 1, " east-flag = 0

south-flag : IF region = south " south-flag = 1 " south-flag = 0.

Classifying categorical Variables :-

→ Reclassifying categorical variables means combining, regrouping or transforming categories of variables into fewer, more meaningful groups to improve data analysis or model performance.

→ when a categorical variables has too many categories or irrelevant distinctions, we merge or rename them to make analysis easier and more accurate.

→ Why

→ Too many categories exist.
→ some categories have very few observations.
→ Several " have similar in meaning or behaviour.
→ To create more useful or predictive grouping for DM Algorithms.

Ex so states could reclassified as the variable region, like Northeast, Southeast, North central, southwest & west.
OR can be reclassified as economic-level like richer state, midrange state and poorer state.

Adding An index Field

→ An index field is a new variable added to the dataset to uniquely identify each record. It is especially useful when the dataset does not contain a unique ID or key.

Reason

1. uniquely identify each record.
2. Data tracking
3. Debugging
4. Joining datasets.

DMPA

Removing variables that are not useful

→ Some variables in the dataset may not contain useful information for DM task. These variables should be identified and removed to simplify the dataset and improve the performance of the model.

→ Reason

→ Reduce noise (Irrevant variables)
→ Improve computation
→ Simplify interpretation.
→ Avoid redundancy.

Ex ① Unary variables.

→ It is a variable that has only one value for all records in a dataset.

ans

ID	Gender	Country	Salary
1	Male	India	40K
2	Male	India	30K
3	Female	India	20K

② Variables that are very nearly unary

→ A nearly unary variables that has almost the same value for all records with only a few records differing.

Ex

ID	Membership	Age	
1	gold	38	
2	gold	28	
3	gold	45	
4	silver	30	→ only differ

Variables that should Probably not be Removed

Some variables that appear unimportant at first may later turn out to be quite useful.

Types of variables

→ categorical identifiers that group records meaningfully.
→ variables with potential interaction effects.
→ variables with missing or rare values.
→ " that are meaningful to interpretation.

Eg

ID	Name	department	salary	Age
1	John	HR	40k	25
2	Rani	sales	50k	28
3	Raj	HR	30k	40