Day 5
(10 Dec 2022)

Data cleansing

Noisy Data :—

- Binning :—

Smooth a sorted data value by consulting its "neighbourhood".

Steps:

~~Smoothing by bin means~~

-> Sort Data.

-> Partitioned into equal-frequency bins.

-> Apply.

- Smoothing by bin Means
- " Medians.
- " Boundaries.

eg:

After Partitions

4, 8, 15 -> Bin 1

21, 21, 24 -> Bin 2

25, 28, 34 -> Bin 3

=> Smoothing by Bin-Means:

Mean of Bin 1 => 9

So, (4, 8, 15) is => (9, 9, 9).

Bin 2 => 22

=> (22, 22, 22).

Bin 3 => 29

=> (29, 29, 29).

=> Smoothing by Bin - Median :-

Median of Bin 1 => 8

so, (4, 8, 15) is => (8, 8, 8).

. . . .

=> Smoothing by Bin - Boundaries :-

we have to change the values based on the closest value of Boundaries (first & last) value.

- first and last values remains constant.

(4, 8, 15) => (4, 4, 15)

(21, 21, 24) => (21, 21, 24)

(25, 28, 34) => (25, 25, 34)

diff (4, 8) = 4

diff (8, 15) = 7

4 < 7

so, replace values inside with small values.

Note: No. of Buckets = 2, No. of Data = 9 => 9/2 = 4.5 => 5 (Per bucket)

--> completed Data cleaning <--

Data Aggregation :-

=> combining Data:

-> Joining the data => [using Primary (or) candidate key].

-> Appending & stacking (mostly used in ML).

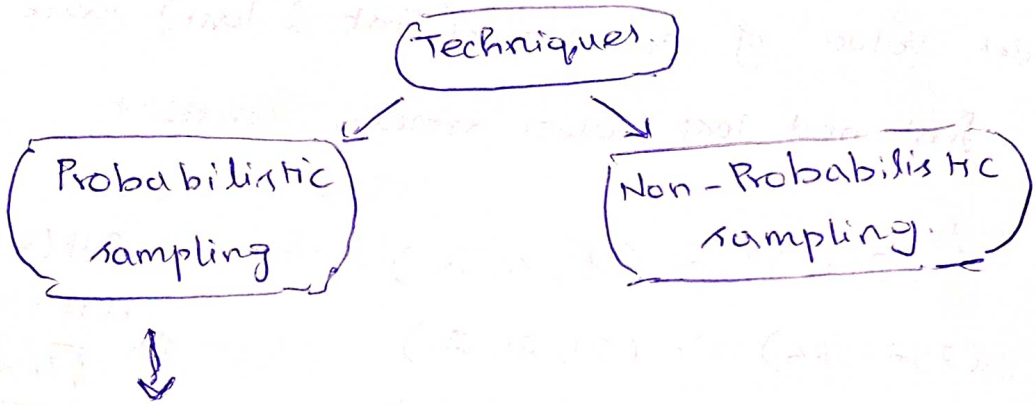
Data sampling :-

Use:-

=> Statisticians

=> Data Miners : (we will apply sampling over Training & Testing Dataset)

- To reduce the data size. To choose the best ML/DL algorithms..



=> Sampling without Replacement-

=> Sampling with Replacement..
(Boosting in ML)

=> Systematic sampling.
eg: Picking every fifth Mango from Mango Bucket-

=> Stratified sampling.
- eq no. of objects
- no. of objects drawn from each grp is proportional to the size

Data Similarity:-

1 => Similar 0 => Dissimilar.

Higher similarity value ⇒ Dissimilarity is More

$$\text{Similarity} = 1 - \text{dissimilarity}$$

Dissimilarity Matrix:-

$$d(i,i) = 0.$$

$$d(i,j) = d(j,i).$$

$$sim(i,j) = 1 - d(i,j).$$

Lower & upper Diagonal Matrix value will be same. Since, $d(i,j) = d(j,i)$

$$d(i,j) = \begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \ddots & 0 & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

Measures of Proximity:-

i) Numerical Data:-

Euclidean dist:

$$d(i,j) = \sqrt{(x_{i_1} - x_{j_1})^2 + (x_{i_2} - x_{j_2})^2 + \cdots + (x_{i_n} - x_{j_n})^2}$$

Manhattan or Hamming or city Block dist:-

$$d(i,j) = |x_{i_1} - x_{j_2}| + |x_{i_1} - x_{j_2}| + \cdots + |x_{i_n} - x_{j_n}|$$

(Sum of absolute Difference).

Minkowski :-

Supremum :-

$$d(i,j) = \max_f |x_{if} - x_{jf}|.$$

$$d(x_1, x_2) = \max\left(|x_{11} - x_{12}|, |x_{21} - x_{22}|\right)$$

ii) Categorical Attributes :- [0 - identical, 1 - dissimilar]

dissimilarity $\Leftarrow$ $d(i,j) = \frac{p-m}{p}$

$m \rightarrow$ no. of matches (same attributes).

$p \rightarrow$ total no. of attributes describing the objects.

$$\text{Tim}(i,j) = \frac{m}{p}$$

ONLY ONE ATTRIBUTE:

obj	color
1	R
2	B
3	G
4	R

$p \rightarrow$ no. of attribute $\Rightarrow$ only 'color' $\Rightarrow$ 1

	m	$\frac{p-m}{p}$
d(2, 1)	0	$\frac{1-0}{1} = 1$
d(3, 1)	0	$\frac{1-0}{1} = 1$
d(4, 1)	1	$\frac{1-1}{1} = 0$

after calc all the values

$d \Leftarrow$ Dissimilarity Matrix $=$

$$\begin{bmatrix} 0 & & \\ 1 & 0 & \\ 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

TWO ATTRIBUTES:-

obj	color	Position.

iii) Ordinal Attributes:-

Grades = { A, B, C } => m = 3

A > B > C

=> Sort in ascending order

Grades => { C, B, A }

1 2 3

Normalization:-

$$Z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

r => rank

m = 3

Grades => C B A

(Rank) r => 1 2 3

$$\frac{1-1}{3-1}, \frac{2-1}{3-1}, \frac{3-1}{3-1} \Rightarrow (0, 0.5, 1)$$

Sizes => S M L XL

(Rank) r => 1 2 3 4 m = 4

$$\frac{1-1}{4-1}, \frac{2-1}{4-1}, \frac{3-1}{4-1}, \frac{4-1}{4-1} \Rightarrow \left(0, \frac{1}{3}, \frac{2}{3}, 1\right)$$

=> So,
Ordinal Data was converted to Numerical Data

=> We can use any DISTANCE Measuring algorithm to calculate the dissimilarity.

Proximity measures of Binary Attributes :-

binary attr:

eg: Gender = {M, F}

Food = {V, NV}

Job = {Y, N}.

	G	F	J
Ahmed	1	0	1
Surekha	1	1	0

Contingency Matrix.

Surekha \ Ahmed	1	0	sum
1	1 (q)	1 (r)	2 (q+r)
0	1 (s)	0 (t)	1 (s+t)
sum	2	1	3

count

$(p) = q+r+s+t$

$$d(i,j) = \frac{r+s}{(q+s+r+t)} = \frac{1+1}{3} = \frac{2}{3}$$

(Dissimilarity)

Asymmetric Binary dissimilarity $= \frac{r+s}{q+s+r}$

(Jaccard co-efficient)

$$sim(x, y) = \frac{x \cdot y}{\|x\| \|y\|}$$

$x = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0)$

$y = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)$

$x^t \cdot y = (5 \times 3) + (0 \times 0) \ldots\ldots + (0 \times 1) = 25$

$\|x\| = \sqrt{\qquad} =$

$\|y\| = \sqrt{\qquad} =$

Proximity Measures of Mixed type of attributes:-

→ calculate similarity matrices for each of the attributes.

eg:

obj	color (categ.)	Quality (ordinal)	Quantity (numeric)
1	R	Excell	475
2	B	Fair	10
3	G	Good	1000
4	R	Excellent	500

$$d = \begin{matrix} \\ 1 \\ 2 \\ 3 \\ 4 \end{matrix} \begin{matrix} 1 & 2 & 3 & 4 \end{matrix} \begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ 0 & 1 & 1 & 0 \end{bmatrix} \qquad \begin{bmatrix} 0 & & & \\ 1.0 & 0 & & \\ 0.5 & 0.5 & 0 & \\ 0 & 1.0 & 0.5 & 0 \end{bmatrix}$$

(COLOR) (QUALITY)

↑ follow CATEGORICAL Method of dissimila.

↑ follow ordinal

$$= \frac{x_{if} - x_{jf}}{max - min}$$

// Normalise the Numerical data when it is having larger values.

(QUANTITY)

$$\Rightarrow \begin{bmatrix} 0 & & & \\ 0.470 & 0 & & \\ 0.53 & 1.0 & 0 & \\ 0.025 & 0.4949 & 0.505 & 0 \end{bmatrix}$$

// after Normalisation

$$d(i,j) = \frac{\sum_{f=1}^{p} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{p} \delta_{ij}^{(f)}}$$

(Dissimilarity)

δ is 0. If any one of the value is Missing. (δ is 1) If both value are present

$$d(2,1) = \frac{(1 \times 1) + (1 \times 1.0) + (1 \times 0.470)}{1+1+1}$$

(1 × 1) from colour; (1 × 1.0) from Qual; (1 × 0.470) from Quantity

$$= \frac{2.470}{3} = 0.823.$$

$$d(4,1) = \frac{(1 \times 0) + (1 \times 0) + (1 \times 0.025)}{1+1+1}$$

$$= 0.0083$$

$\Rightarrow$ d(2,1)

.

.

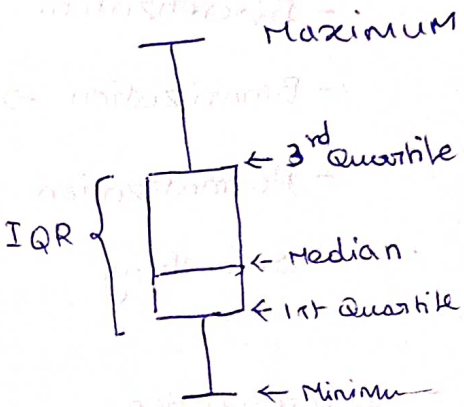
.

!

Likewise, we will find the values for all the values

Visualization Techniques for Data Exploratory Analysis:-

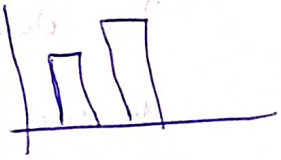
- Boxplot:



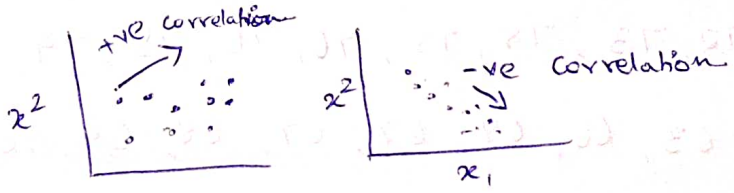
- Histogram :- (Univarient analysis)

→ x is Nominal. => Bar chart.

→ x is Numeric => Histogram.



- Scatter plot :- (Bivarient Analysis)



x_1, x_2 => 2 features. (x_1 = colour, x_2 =

=> Visualization.ipynb

Sampling..ipynb

Preprocessing.ipynb.

Handling Numeric Data:-

Techniques,

- Discretization → numeric data to discrete categories
- Binarization → " Binary
- Normalization → specific range.
- Smoothing

- Discretization :-

Supervised vs Unsupervised Discretization.
(not looking class labels at all).

Top down discretization (or) Splitting.

Bottom up discretization (or) Merging.

Bining Example:-

70, 70, 72, 73, 75, 75, 76, 76, 78, 79, 80, 81, 53,
56, 57, 63, 66, 67, 67, 67, 68, 69, 70, 70.

Sol: EQUAL WIDTH

1) Asc order: 53 56 57 63 66 67 67 67 68 69
70 70 70 70 72 73 75 75 76 76 78 79 80 81

2) width = $\frac{81-53}{3} = \frac{28}{3} = 9$

3) Bins (or) Buckets

Bin 1 [53, 62] (53 + 9) => 62 [Min + width]

Bin 2 [63, 72]

Bin 3 [73, 82]

EQUAL DEPTH

① same

② $\text{Depth} = \frac{\#\text{ count}}{\#\text{no. of bins}} = \frac{24}{3} = 8$

③

Bin 1 → 1st eight values.

Bin 2 → 2nd eight values.

Bin 3 → 3rd eight values.

Drawback of Equal Depth:-

If the values are occurred Multiple times, then the value can go to any of the Bins.

---

⑪ # Bining.ipynb