Ch-2 Data Preprocessing

The data are incomplete → lacking attribute values, or containing only aggregate data
Ex: salary = " "
occupation = " "

→ noisy :- containing errors or outlier values
Ex: salary = -10

→ inconsistent - containing discrepancies
Age = '29', Birthday = 5/04/1985

Data cleaning → clean the data by filling in missing values,
smoothing noisy data
identifying or removing outliers
and resolving inconsistencies

Data integration → integrating multiple databases, data cubes or files

Data transformation → normalization and aggregation

Data reduction → a reduced representation of the data set that is much smaller in volume, yet produces the same analytical results.

(1) Data aggregation → building a data cube.

(2) attribute subset selection → removing irrelevant attributes through correlation analysis.

(3) dimensionality reduction.
Ex - DFT, PCA, DCT, DWT

(4) Numerosity reduction → replacing the data by smaller alternative forms as clusters or parametric models.

Data characteristics can be obtained from

(1) central tendency → mean, median, mode and midrange.

(2) Dispersion of data → quartiles, interquartile range (IQR) and variance.

There are 3 measures for central tendency.

→ Distributive measure → sum(), count(), max(), min()

Algebraic measure → mean() = sum()/count()
weighted mean.

Holistic measure.

Distributive measure → can be computed for a given data set by dividing the data into smaller subsets, and then computing the measure and then merging the results.

Algebraic measure → computed by applying an algebraic function to one or more distributive measures.

Ex → mean(),

weighted mean = $\bar{x} = \frac{w_1x_1 + w_2x_2 + \ldots + w_Nx_N}{w_1 + w_2 + \ldots + w_N}$.

Trimmed mean → mean obtained after chopping off values at the high and low extremes.

median → mid value if N is odd.
average of two middle values, if N is even.

Holistic measure → must be computed on the entire data set as a whole.

Ex:- approximate median $= L_1 + \left(\frac{N/2 - (\sum freq)_L}{freq_{median}} \right) width$

L_1 = lower boundary of the median interval.

N = total no. of values in the entire dataset.

$freq_{median}$ = frequency of the median interval

width = width of the median interval.

Mode → The value that occurs most frequently in the set.

Data sets with one, two or 3 modes are called unimodal, bimodal and trimodal.

A data set with 2 or more modes is called multimodal.

Midrange → Average of the largest and smallest values in the set.

Range, Quartiles, outliers and Boxplots

Let $x_1, x_2 \ldots x_N$ be a set of observations.

Kth percentile of a set of data is the value x_i having the property that K percent of the data entries lie at or below x_i.

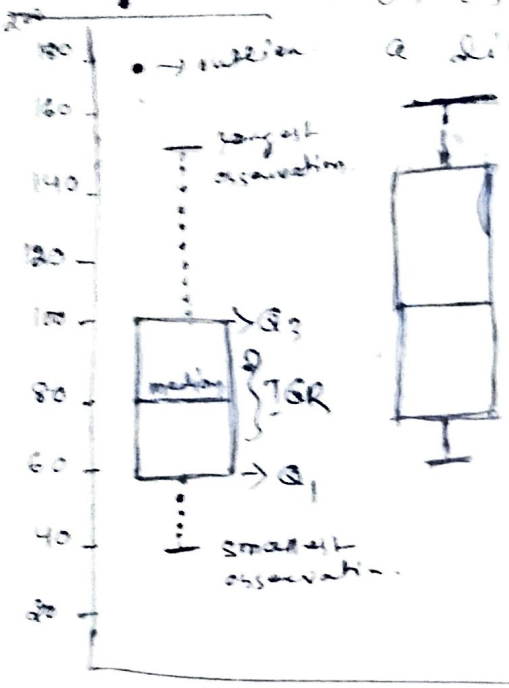
1st quartile (Q_1) → 25%

3rd " (Q_3) → 75%

IQR → interquartile range $= Q_3 - Q_1$

Outliers → are values falling at least 1.5 IQR times above the 3rd quartile or below the 1st quartile.

Boxplots → it is a popular way of visualizing a distribution.



For branch 1,
median price of items sold is $80
Q_1 is $60
Q_3 is $100

minimum → 40
max → 150

outliers — 175 and 202 marked as the value dot as the value dot greater than (1.5 x IQR)

(Boxplot for unit price data for items sold at four branches)

Variance and standard deviation

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$$ → variance of N observations $x_1, x_2, \ldots x_N$

Standard deviation, $\sigma = \sqrt{\sigma^2}$

Graphical representation of Data.

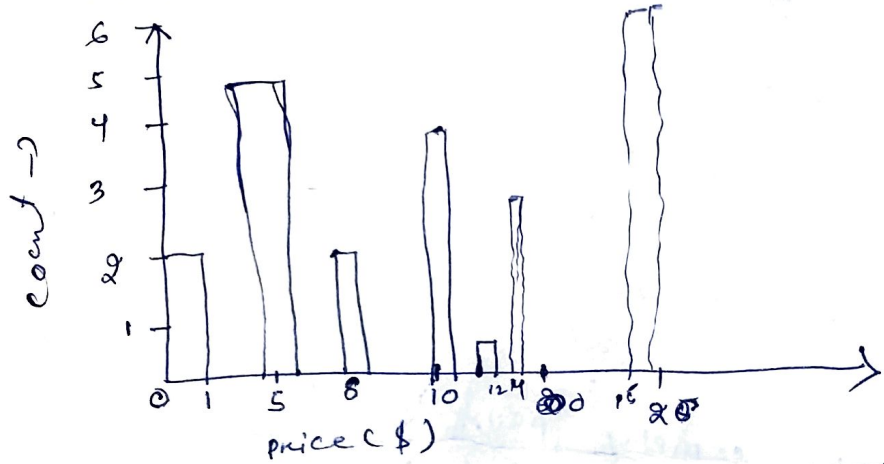
To display data distribution, we consider

(1) Histogram. or Bar chart.

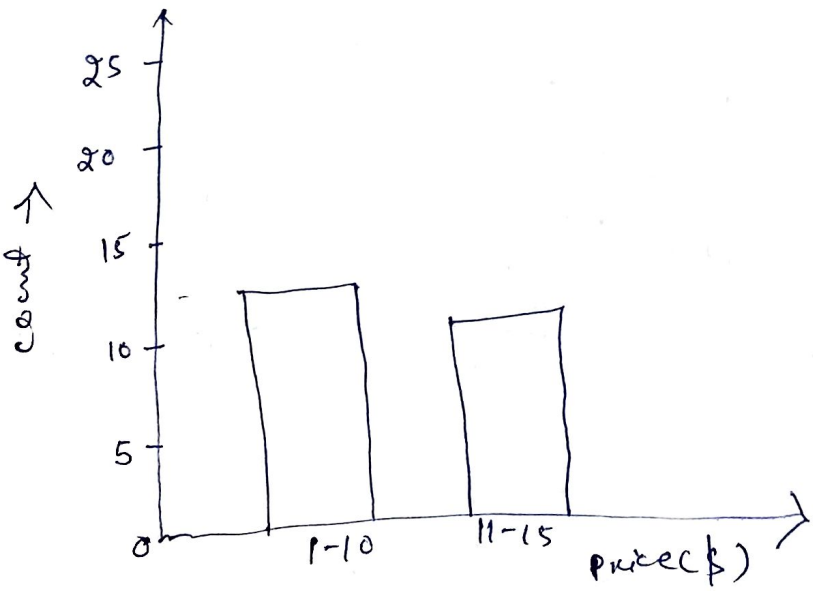
~~(2) Quantile plot~~

Ex The following data are a list of prices commonly sold items. The numbers have been sorted.

1, 1, 5, 5, 5, 5, 5, 8, 8, 10, 10, 10, 10, 12, 14, 14, 14,
15, 15, 15, 15, 15, 15, 18, 18, 18, 18, 18, 18, 18, 18, 20,
20, 20, 20, 20, 20, 20, 21, 21, 21, 24, 25, 25, 25, 25, 25, 28, 28, 30, 30, 30.



(Each bucket represents one price-value)



Equal-width histogram for price.

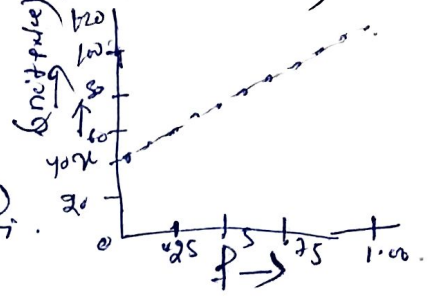
Quantile plot

→ Let x_i, for $i = 1$ to N, be the data sorted in increasing order so that x_1 is the smallest observation and x_N is the largest.

→ Each observation, x_i is paired with a percentage, f_i which indicates that approximately $100 f_i$% of the data are below or equal to the value x_i.

→ $f_i = \frac{i - 0.5}{N}$.

→ x_i is plotted against f_i.



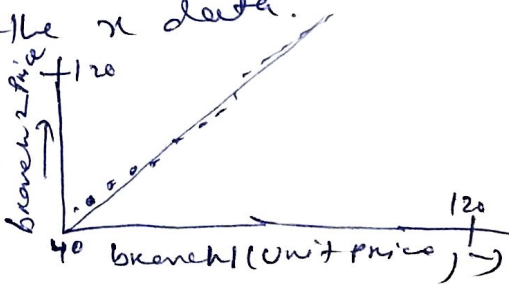
Quantile - quantile plot (q-q)

Let $x_1 \dots x_N$ be the data from the 1st branch and $y_1, \dots y_M$ " " " " 2nd branch where each data set is sorted in increasing order.

If $M = N$, simply plot y_i Vs. x_i, where y_i and x_i are both $(i - 0.5)/N$ quantiles of their respective data sets.

If $M < N$, here y_i is the $(i - 0.5)/M$ quantile of the y data, which is plotted against the $(i - 0.5)/M$ quantile of the x data.

Each point corresponds to the same quantile for each dataset, shows the price of items sold at Branch 1 vs. branch 2 at that quantile.



Scatter plot

Here, each pair of values is treated as a pair of coordinates in an algebraic sense and plotted as points on the plane.

Data cleaning

How to handle missing values?

(1) Ignore the tuple where there is a missing
But not an efficient way.

(2) Fill it manually. (tedious, infeasible)

(3) Fill it with global constant (e.g. unknown or ∞)
Not an efficient method.

(4) By the attribute mean.

(5) By most probable value (this is determined by regression or Bayesian formula or decision tree)

How to handle noisy data?

Noise is a random error or variance of a measured variable.

Happened due to → fault data collection
data entry problems
technology limitation

It can be handled by the following techniques:—

(1) Binning → 1st sort the data, then partition into bins or buckets of equal length and then one can smooth by 3 ways.

(2) Regression → Data can be smoothed by fitting the data to a function, such as regression.

(3) clustering → one can detect the outliers by clustering and remove it.

Binning →

(1) Smoothing by bin means → each value in the bin is replaced by the mean value of the bin.

(2) Smoothing by bin medians → replaced by bin median.

(3) Smoothing by bin boundaries → the minimum and max. values are identified as the bin boundaries. Each bin value is then replaced by the closest boundary value

Ex 4, 8, 15, 21, 21, 24, 25, 28, 34 (price data in $)

partition into (equal-frequency) bins:

Bin 1: 4, 8, 15

Bin 2: 21, 21, 24

Bin 3: 25, 28, 34

Smoothing by bin means

Bin 1: 9, 9, 9

Bin 2: 22, 22, 22.

Bin 3: 29, 29, 29.

Smoothing by bin boundaries

Bin 1: 4, 4, 15

Bin 2: 21, 21, 24

Bin 3: 25, 25, 34

Ex 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

use smoothing by means, using a bin depth of 3.

Merging of data from multiple data sources.

Data Integration –

Various Issues on Data Integration –

- Schema integration
  - ⇓
  - Cust_ID on one database / Cust-NO on another database. (when integrating problem arises)
- Object matching
  - ⇓
  - Brian Lara = Lara Dutta or Bill Clinton = William Clinton
- Redundancy
  - ⇓
  - An attribute is redundant if it can be derived from another attribute or a set of attributes. Ex: Annual revenue.

Redundancies can be detected by correlation analysis

⇓

(i.e by analyzing how strongly one attribute implies the other. i.e causally effect)

For Numeric attributes, the correlation between two attributes can be evaluated by computing the correlation co-efficient

⇓

also called Pearson's product-moment coefficient (product)

⇒ The correlation doesn't imply causality. i.e if A & B are correlated, this doesn't necessarily imply that A causes B or B causes A

Ex:- We may find that attributes representing the no. of hospitals and the no. of car thefts in a region are correlated. This does not mean that one causes the other.

Chi-square test (χ^2) -

→ The correlation relationship between two attributes (A & B) can be evaluated by χ^2 test.
(of categorical data) i.e discrete
Ex: Discrete attribute: ZIPcode, profession
" continous " : temp, height, weight

1: Assume a null hypotheses.
2: Compute value of $\alpha(P)$ → Degree of freedom
3: Calculate value of χ^2 & DOF
4: write the rule to proof the hypotheses.
5: else write the summary based on decession.

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(O_{ij} - e_{ij})^2}{e_{ij}}$$

(A has c distinct values)
B has r " "

or

$$\chi^2 = \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

O → observed value (frequency)
E → expected value (frequency)

$$E = \frac{\text{Col total} * \text{Row total}}{\text{Grand total}}$$

$$DOF = (\text{NO. of col} - 1) * (\text{NO. of rows} - 1)$$

Eg Suppose that a group of 1500 people was surveyed. Each person was polled as to whether their preferred type of reading was fiction or nonfiction. some have two attribute, gender, prefered_reading

	male	female	~~total~~
fiction	250	200	~~450~~
non-fiction	50	1000	~~1050~~

Nonfiction → real face based story
Fiction → imaginary story

Are gender & prefered_reading are correlated?

$\alpha(P) = 0.001$

Ans

	male	female	Total
fiction	250	200	= 450
non-fiction	50	1000	= 1050
Total =	300	1200	= 1500

1500

(2)

for fiction

	observed value	Expected value	χ^2
Male	250	$\frac{300 \times 450}{1500} = 90$	$\frac{(250-90)^2}{90}$
Female	200	$\frac{1200 \times 450}{1500} = 360$	$\frac{(200-360)^2}{360}$

for non-fiction

Male	50	$\frac{300 \times 1050}{1500} = 210$	$\frac{(50-210)^2}{210}$
Female	1000	$\frac{1200 \times 1050}{1500} = 840$	$\frac{(1000-840)^2}{840}$

$$\Rightarrow \chi^2 = \frac{(250-90)^2}{90} + \frac{(200-360)^2}{360} + \frac{(50-210)^2}{210} + \frac{(1000-840)^2}{840}$$

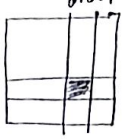
$$= |507.93| = 508$$

DOF = $(2-1) * (2-1) = 1$ ($\because$ 2×2 matrix)

$\alpha(p) = 0.001$

At (DOF, $\alpha(p)$)

i.e at (1, 0.001) from χ^2 distribution table



its value is 10.828

Since χ^2 value which is computed as 508 is above this value,

→ the hypotheses can be rejected

i.e Gender & Preferred Reading } are not independent

⇒ These two attributes are strongly co-related

If χ^2 > value at (DOF, $\alpha(p)$),

strongly co-related

For Numeric, it is correlation co-efficient

i.e $r_{A,B} = \dfrac{\sum_{i=1}^{N}(a_i b_i) - N\bar{A}\bar{B}}{N\sigma_A\sigma_B}$

N → no. of tuples
A, B → Attributes (Numeric)
a_i → values of A
b_i → " of B

$$\boxed{-1 \leq r_{A,B} \leq 1}$$

If $r_{A,B} > 0$, A & B are positively correlated
(i.e if A ↑, B also increases)

If $r_{A,B} = 0$, A & B are independent
(No correlation between them)

If $r_{A,B} < 0$, A & B are negatively correlated
i.e if A ↑, B ↓