Data Mining

Attribute : It is a data field representing a characteristic or feature of a data object.

- Nominal Attributes : The values are symbols or names of things
  e.g. hair-colour : black, brown, blonde, red, gray, white.
- Binary Attributes : It is a nominal attribute with only two categories, 0 or 1.

  0 means the attribute is ~~present~~ absent
  1 means " " " present

  e.g. smoker 1 or 0

- Ordinal Attributes : Attribute with possible values that have a meaningful order or ranking among them.

  e.g. drink_size : small, medium, large
  0 1 2

- Numeric Attributes : attributes that are measurable quantity represented in integer or real values.

  Interval scaled Attributes
  that do not have a zero point
  e.g. calendar dates

  Ratio scaled attributes
  that have a zero point
  e.g. Kelvin temp.
  years-of-experience.

- Discrete vs Continuous Attributes

Discrete attribute : finite or countably infinite set of values which may or may not be represented as integers.

e.g. customer-ID , zip codes

Continuous attribute : typically represented as floating point variables.

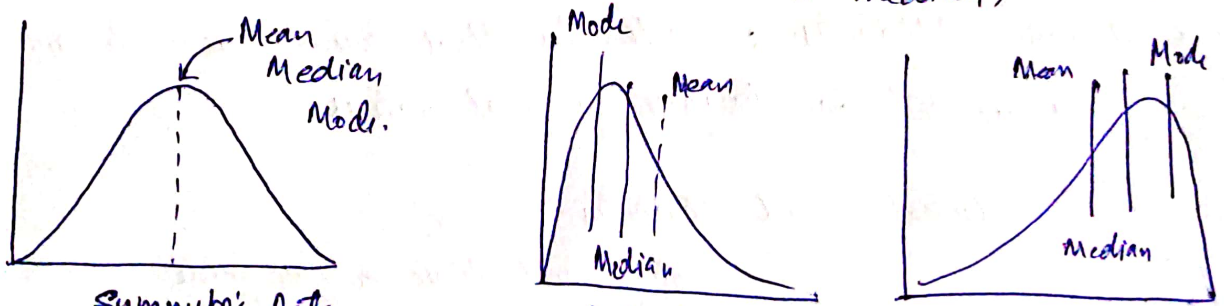
- Measures of Central Tendency.

Mean : $\bar{x} = \frac{\sum x_i}{N}$

Median : $L_1 + \left(\frac{N/2 - \sum freq}{freq_{median}} \right)$

Mode :

$$mean - mode \approx 3 \times (mean - median)$$



Symmetric Data — Mean ≈ Median ≈ Mode

Positively Skewed — Mean > median > mode

negatively Skewed — mean < median < mode ue

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^2 = \left(\frac{1}{N}\sum x_i^2\right)-\bar{x}^2$$

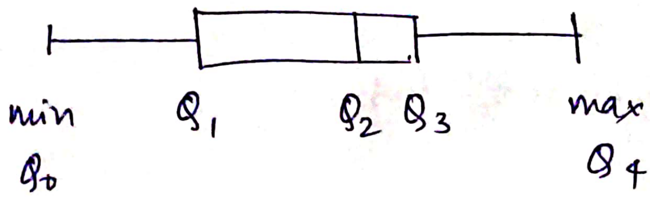
- Standard Deviation

$SD = \sqrt{\sigma^2}$

→ Pearson's product moment coeff.

$$r_{AB} = \frac{\sum (a_i-\bar{A})(b_i-\bar{B})}{n\,\sigma_A\sigma_B}$$

- Box Plot



IQR: $Q_3 - Q_1$

inter quartile range.

e.g. attribute values

6, 47, 49, 15, 42, 41, 7, 39,

43, 40, 36

sorted: 6, 7, 15, 36, 39, 40, 41, 42, 43, 47, 49

$$\begin{array}{ccccc} 6 & 15 & 40 & 43 & 49 \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ Q_0 & Q_1 & Q_2 & Q_3 & Q_4 \end{array}$$

Data Visualization

- Pixel Oriented Visualization

- Geometric Projection Visualization.

  e.g. 2D scatter plot

- Icon-Based Visualization Technique

  e.g. Chernoff Faces.

→ Data Matrix and Dissimilarity Matrix

Data Matrix   'n' objects described by 'p' attributes

$$\begin{bmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1p} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & x_{n3} & \cdots & x_{np} \end{bmatrix}_{n \times p}$$

object-by-attribute

Dissimilarity Matrix

(object-by-object structure)

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

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

...inity Measures for Nominal Attributes

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

p = total no. of attributes
m: matched attributes

- Proximity Measures for Binary Attributes

		Object j		
		1	0	
Object i	1	q	r	$q+r$
	0	s	t	$s+t$
		$q+s$	$r+t$	

→ ex: positive and negative outcomes of disease test

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

if two states are not equally important then $d(i,j) = \frac{r+s}{q+r+s}$

$sim(i,j) = 1 - d(i,j)$ = Jaccard coefficient

- Dissimilarity of Numeric Data
  - Minkowski distance

$$d(i,j) = \sqrt[h]{|x_{i1}-x_{j1}|^h + |x_{i2}-x_{j2}|^h + \dots |x_{ip}-x_{jp}|^h}$$

Euclidean is $h=2$, Manhattan is $h=1$

- supremum distance i.e. $h \to \infty$ $\quad d(i,j) = \lim_{h\to\infty} \left(\sum_{f=1}^{p} |x_{if}-x_{jf}|^h\right)^{\frac{1}{h}}$

or aka uniform norm

$$= \max_{f}^{p} (|x_{if}-x_{jf}|)$$

- Cosine Similarity

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

$$||x|| = \sqrt{x_1^2 + x_2^2 .. x_p^2}$$ or euclidean norm

- χ^2 correlation Test for Nominal Data

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

hyp: they are independent

if value is > then reject hyp

e.g.

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

e_{ij}:

90 $\frac{300 \times 450}{1500}$ 360 $\frac{1200 \times 450}{1500}$

210 $\frac{300 \times 1050}{1500}$ 840 $\frac{1200 \times 1050}{1500}$

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

$= 507.93$

Degree of freedom $= (2-1)(2-1) = 1$

Data Transformation Strategies

1. Smoothing : remove noisy data
2. Attribute construction : create new attributes from given ones
3. Aggregation : daily sales aggregated to form monthly sales.
3. Normalization : bringing values between -1.0 and 1.0 or 0.0 to 1.0
5. Discretization : raw values replaced by intervals.
6. Concept hierarchy generation for nominal data.

→ Normalization

- min-max

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

let's say you want to map to [0,1]

$$v_i' = \frac{v_i - min_A}{max_A - min_A}(1-0) + 1$$

- z-score normalization

$$v_i' = \frac{v_i - \bar{A}}{\sigma_A} \qquad \frac{v_i - mean}{std.\,dev.}$$

- Decimal Scaling

divide by 10^n

if A ranges from -986 to 917. max absolute value is 986, then divide by 10^3