Mean deviation $\frac{|\bar{x}-x_i|}{N}$

Std deviation — RMS

var — MS $\frac{\Sigma(\bar{x}-x_i)^2}{N}$

$\sqrt{\frac{\Sigma(\bar{x}-x_i)^2}{N}}$

PAGE No.

DATE / /

MODULE : 3

Q. 45 47 52 52 53 55 56 58 62 80

10 pts total

- Mean = 56

Make 2 sets

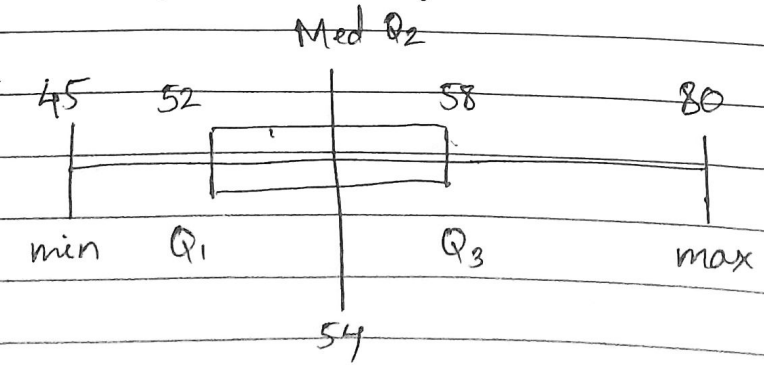
[45, 47, 52, 52, 53], [55, 56, 58, 62, 80]

keep separate

- Median : (53 + 55)/2 = 54
- Mode = 52 unimodal
- Max value : 80
- Min value : 45
- Mid range : (max + min)/2 = 62.5

- Quartile
  - Q1 = median of lower half part of data = 52
  - Q3 = median of upper half part of data = 58

- IQR = $Q_3 - Q_1$ = ht of box = 58 − 52 = 6
- Outlier = Bs 1.5 IQR = 9 (beyond the quartile)
- Semi inter quartile range = (Q3 − Q1)/2 = 6/2 = 3
- Mid quartile range = (Q3 + Q1)/2 = 55

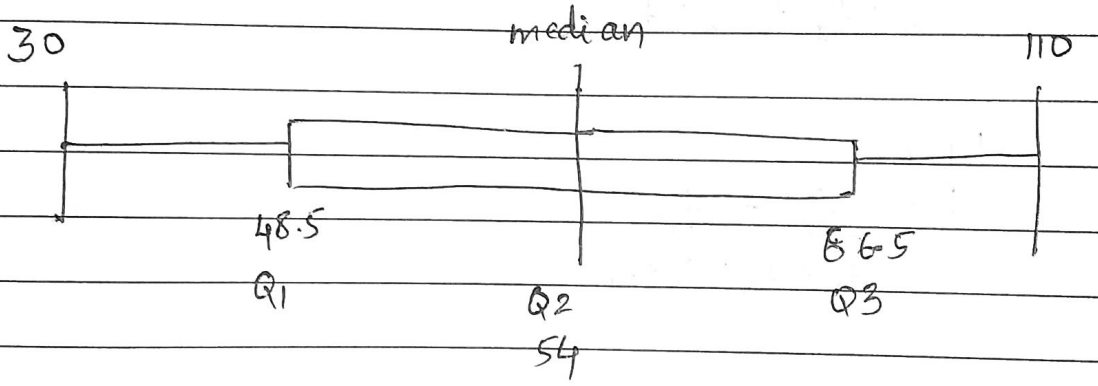


if odd no. of nos, make 2 sets excluding mid

PAGE No.
DATE / /

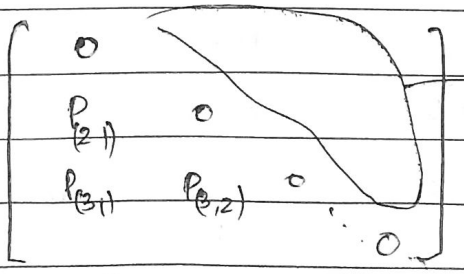
Q. 30 36 47 50 52 52 56 60 63 70 70

→
- Mean = 58
- Median = 54
- Mode = 52, 70 bimodal
- Max val = 110
- Min val = 30
- Mid range = (max + min) / 2 = 70
- Q1 = (47 + 50)/2 = 48.5
- Q3 = (63 + 70)/2 = 66.5



* (p) Similarity : max 1 [0, 1]
(q) Dissimilarity : often 0

Proximity: similarity / dissimilarity



→ These are considered to be same (symmetric matrix)

	1	0
Sim	Max	Min
Dis	Min	Max

Symmetric binary : if the ~~value~~ attri can have 2 separate results ⇒ X sym binary attri ⇒ X sym bin

Pass → next class Fail → detain

Male Female ⇒ symmetric binary

Obj	Gendes	Similarity	Dissimilarity
Ram	M 1	p(Ram, sita) = 0	q = 1
Sita	F 0	p(Ram, Laxman) = 1	q = 0
Laxman	M 1		

$$q(i,j) = 1 - p(i,j)$$

* similarity measure with symmetric binary

↗ Sb coeff

$$S = \frac{\text{No. of matching attri values}}{\text{Total nos. of attris}}$$

OR

$$S = \frac{f_{00} + f_{11}}{f_{00} + f_{01} + f_{10} + f_{11}}$$

Db c

$$D = \frac{f_{01} + f_{10}}{f_{00} + f_{01} + f_{10} + f_{11}}$$

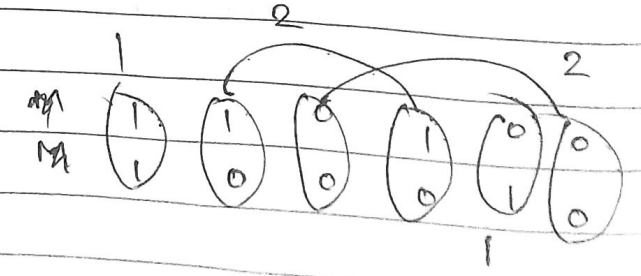
eg

Obj	Gender	Food	Caste	Edu	Hobby	Job
Hari	M	V	M	L	C	N
Ram	M	NV	M	I	T	N
Tomi	F	NV	H	L	C	Y

	M-1	V-1	M-0	L-1	C-0	N-0
	F-0	NV-0	H-1	I-0	T-1	Y-1

Contengency matrix / table

	R_0	R_1
H_0	2	1
H_1	2	1



$$S(\text{Hari, Ram}) = \frac{1+2}{|1-2|+2} = 0.5$$

* Proximity measure with asymmetric coeff

Jacard coeff

similarity J = No. of matching presence / No. of attri not involved in matching 00

$$= \frac{f_{11}}{f_{01}+f_{10}+f_{11}}$$

(d) dissim = $1 - J$

X	R_0	R_1
H_0	2	1
H_1	2	1

$$J = \frac{1}{1+2+1} = 0.25$$

* Proximity measure with categorical attri

$$s_{\text{cat}}(x,y) = \frac{\text{No. of matches}}{\text{Total no. of attri}} = \frac{m}{a}$$

$$d(x,y) = \frac{a-m}{a}$$

m = no. of matches

a =

*

w. ordinal attri

$$\hat{a}_i = \frac{i-1}{n-1}$$

$$s(x,y) = \sqrt{(\hat{a}_i - \hat{a}_j)^2}$$

8

Obj	Size	Quality
A	S(0.0)	A(0.66)
B	L(1.0)	Ex(1.0)
C	L(1.0)	C(0.0)
D	M(0.5)	B(0.33)

size = {S, M, L}

Qty Qual = {Ex, A, B, C}

1 2 3

S < M < L

~~Ex > A > B X~~

4 3 2 1

Ex > A > B > C

$$a_S = \frac{1-1}{3-1} = 0$$

$$a_M = \frac{2-1}{3-1} = \frac{1}{2}$$

$$a_L = \frac{3-1}{3-1} = 1$$

* with interval scale

1) Manhattan distance (L_2 Norm : r=1)

$$d = \sum_{i=1}^{n} |x_i - y_i|$$

eg $x = [7, 3, 5]$, $y = [3, 2, 6]$

$$|7-3| + |3-2| + |5-6| = 6$$

Hamming distance ~~with~~ when attri values $\in [0, 1]$

2) Euclidean Distance (L_2 Norm : $r = 2$)

3) Chebychev Distance (L_∞ Norm: $r \in R$)

$$d(x,y) = \max_{\forall i} \{ |x_i - y_i| \}$$

$$\text{Max} \{ |7-3|, |3-2|, |5-6| \} = 4$$

* for ratio - scale

$$X = A e^B$$

Apply the Naive Bayes classifies algorithm for buys computes classification and classify the type

X = (age = "young", income = "medium", student = "yes", credit-rating = "fair")

Id	Age	Income	Student	Credit rating	Buys Computes
1	young	high	no	fair	no
2	young	high	no	good	no
3	middle	high	no	fair	yes
4	old	medium	no	fair	yes
5	old	low	yes	fair	yes
6	old	low	yes	good	no
7	middle	low	yes	good	yes
8	young	medium	no	fair	no
9	young	low	yes	fair	yes
10	old	medium	yes	fair	yes
11	young	medium	yes	good	yes
12	middle	medium	no	good	yes
13	middle	hight	yes	fair	yes
14	old	medium	no	good	no

→ Total samples = 14

No. of Yes = 9

P(Yes) = 9/14

No. of No = 5

P(No) = 5/14

Age

$P(\text{young} \mid \text{Yes}) = \frac{2}{9}$ $\quad P(\text{young} \mid \text{No}) = \frac{3}{5}$

$P(\text{middle} \mid \text{Yes}) = \frac{4}{9}$ $\quad P(\text{middle} \mid \text{No}) = \frac{0}{5}$

$P(\text{old} \mid \text{Yes}) = \frac{3}{9}$ $\quad P(\text{old} \mid \text{No}) = \frac{2}{5}$

Income

$P(\text{high} \mid \text{Yes}) = \frac{2}{9}$ $\quad P(\text{high} \mid \text{No}) = \frac{2}{5}$

$P(\text{medium} \mid \text{Yes}) = \frac{4}{9}$ $\quad P(\text{medium} \mid \text{No}) = \frac{2}{5}$

$P(\text{low} \mid \text{Yes}) = \frac{3}{9}$ $\quad P(\text{low} \mid \text{No}) = \frac{1}{5}$

Student

$P(\text{Yes} \mid \text{yes}) = \frac{6}{9}$ $\quad P(\text{Yes} \mid \text{No}) = \frac{1}{5}$

$P(\text{No} \mid \text{Yes}) = \frac{3}{9}$ $\quad P(\text{No} \mid \text{No}) = \frac{4}{5}$

Credit rating

$P(\text{fair} \mid \text{Yes}) = \frac{6}{9}$ $\quad P(\text{fair} \mid \text{No}) = \frac{2}{5}$

$P(\text{good} \mid \text{Yes}) = \frac{3}{9}$ $\quad P(\text{good} \mid \text{No}) = \frac{3}{5}$

Data Sample: $X \equiv$ age = "young"
income = "medium"
student = "yes"
credit-rating = "fair"

$$P(Yes|X) = P(X|Yes) \cdot P(Yes)$$

$$= P(Young|yes) \cdot P(medium|yes) \cdot P(study|yes) \cdot P(Fair|Yes) \cdot P(Yes)$$

$$= \frac{2}{9} \times \frac{4}{9} \times \frac{6}{9} \times \frac{6}{9} \times \frac{9}{14} = 0.02821$$

$$P(No|X) = P(X|No) \cdot P(No)$$

$$= P(Young|No) \cdot P(medium|no) \cdot P(study)|no) \cdot P(Fair|No) \cdot P(No)$$

$$= \frac{3}{5} \times \frac{2}{5} \times \frac{1}{5} \times \frac{2}{5} \times \frac{5}{14} = 0.006857$$

As $P(Yes|X) > P(No|X)$

Thus, X belongs to the class buys computer

∴ The student buys a computer

$$E(S) = \underbrace{-p \log_2 p}_{\text{Yes}} - \underbrace{p \log_2 p}_{\text{No}}$$

$E(A_i)$ = Same (for all possi vals)

$IG(A_i) = E(S) - E(A_i)$

Q. ID3

Person	Gender	Height	Class
1	Female	1.6m	Short
2	Male	2m	Tall
3	Female	1.9m	Medium
4	Female	2.1m	Tall
5	Female	1.7m	Short
6	Male	1.85m	Medium
7	Female	1.6m	Short
8	Male	1.7m	Short
9	Male	2.2m	Tall

→ Ht <=1.7 H_1

>1.7 =1.85 H_2

>1.85 = 2 H_3

>2 H_4

		Ht	
1	F	H_1	S
2	M	H_3	T
3	F	H_3	M
4	F	H_4	T
5	F	H_1	S
6	M	H_2	M
7	F	H_1	S
8	M	H_1	S
9	M	H_4	T

$$E(S) = -\frac{4}{9} \log_2(4/9) - \frac{2}{9} \log_2(2/9) - 3/9 \log_2(3/9)$$

$$= 1.5305$$

Genders (9)

(5) F — 3S, 1M, 1T

M (4) — 1S, 1M, 2T

$$E(G) = \frac{5}{9}\left[\frac{-3}{5}\log_2\frac{3}{5} - \frac{1}{5}\log_2\frac{1}{5} - \frac{1}{5}\log_2\frac{1}{5}\right]$$

$$\frac{4}{9}\left[-\frac{1}{4}\log_2\frac{1}{4} - \frac{1}{4}\log_2\frac{1}{4} - \frac{2}{4}\log_2\frac{2}{4}\right]$$

$E(G) = 1.4283$

$I_g(G) = 0.1022$

Height (9)

H1 — 4S

H2 — 1M

H3 — 1M, 1T

H4 — 2T

$$E(H) = \frac{4}{9}\left[\frac{-4}{4}\log_2\frac{4}{4}\right] + \frac{1}{9}\left[\frac{-1}{1}\log_2\left(\frac{1}{1}\right)\right]$$

$$+ \frac{2}{9}\left[\frac{-1}{2}\log(1/2) - \frac{1}{2}\log(1/2)\right]$$

$$+ \frac{2}{9}\left[\frac{-1}{1}\log_2\left(\frac{1}{1}\right)\right]$$

$E(H) = 0.2222$

$I_g(H) = 1.3083$

$I_g(H) > I_g(G)$

∴ H is the root element

DATE / / /

* Kmains

$D = \{1, 2, 6, 7, 8, 10, 15, 17, 20\}$

$K1 = \{1, 6, 8, 15, 20\}$

$K2 = \{2, 7, 10, 17\}$

→ K_1 mean = 10 K_2 mean = 9

closer to 10 closer to 9

$K_1(10) = \{10, 15, 17, 20\}$ $K_2(9) = \{1, 2, 6, 7, 8\}$

mean 15.5 mean 4.8

$K_1(15.5) = \{15, 17, 20\}$ $K_2(4.8) = \{1, 2, 6, 7, 8, 10\}$

17.333 5.667

$K_1(17.3) = \{15, 17, 20\}$ $K_2(5.6) = \{1, 2, 6, 7, 8, 10\}$

17.333 5.667

same means ⇒ stop

* Agglomerative

Q. Single link

I/P dist matrix

	BA	FI	MI	NA	RM	TO
BA	0	662	887	225	412	996
FI	662	0	295	468	268	400
MI	887	295	0	754	564	138
NA	225	468	754	0	219	869
RM	412	268	564	219	0	669
TO	996	400	138	869	669	0

• Single Link

Min dist

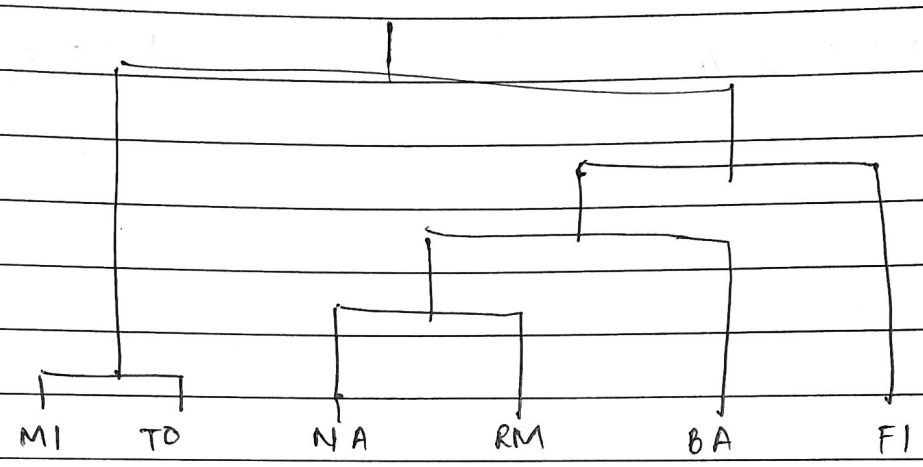
MI $\overset{138}{\rightleftharpoons}$ TO

	BA	FI	MI/TO	NA	RM
BA	0	662	877	255	412
FI	662	0	295	468	268
MI/TO	887	295	0	754	564
~~NA~~ ~~BA~~	255	468	754	0	219
RM	412	268	564	219	0

	BA	FI	MI/TO	NA/RM
BA	0	662	877	255
FI	662	0	295	268
MI/TO	877	295	0	564
NA/RM	255	268	564	0

MI/TO NA/RM

BA/NA/RM FI MI/



Q. Complete link

	1	2	3	4	5
1	1	0.90	0.10	0.65	~~0.20~~
2	0.90	1.00	0.70	0.60	0.50
3	0.10	~~0.70~~	1.00	~~0.40~~	0.30
4	0.65	0.60	0.40	1.00	0.80
5	0.20	~~0.50~~	0.30	0.80	1.00

1, 3

	1, 3	2	4	5
1, 3	1.0	0.9	0.65	0.83
2	0.9	1.0	0.76	0.5
			1.00	
4	0.65	0.6	~~0.4~~	0.8
5	0.3	0.5	0.8	1.00

note

PAGE No.
DATE / /

1,3,5

	1,3,5	2	4
1,3,5	1.0	0.9	0.8
2	0.9	1.0	0.6
4	0.8	0.6	1.0

2,4

	1,3,5	2,4
1,3,5	1.0	0.9
2,4	0.9	1.0

1,3,5,2,4

1,3,5,2,4

DATE / / /

* Market Basket Analysis

if you buy x, you are more likely to buy y.

eg Bread ⟶ Milk

if you buy bread, you are more/less likely to buy milk

$x \longrightarrow y$, $x \cap y \neq \phi$ ie $x \neq y$ for singe item

* Support

min support threshold = 50%.

1: 1, 3, 5

2: 1, 8, 14, 17, 12

3: 4, 8, 6, 12, 9, 104

4: 2, 1, 8

support of $\{8, 12\} = 2$ (or 50%.)

$\{1, 5\} = 1$ (25%)

$\{1\} = 3$ (75%.)

50 = 50

75 > 50 ⟹ $\{8, 12\}$, $\{1\}$ are frequent. (high support)

* Confidence

min confi thr = 50%.

Or.

confi $\{\{8 \Rightarrow 12\}$ = $\frac{2}{3}$ 66%.

(⇒ : assoc rule)

confi $\{1 \Rightarrow 5\}$ = $\frac{1}{3}$ 33%.

$\{8, 12\}$ is strong (high confidence)

Sup → how many times bought x

confi → associativity

Q. Min Confi thre = 50%.

Sup = 30%.

1: 3, 5, 8

2 : 268

3 :

4

support {5,8} = $\frac{4}{10}$ (40%)

confi {5 => 8} = $\frac{4}{5}$ (80%)

(meaningful) Strong and Frequent

Q. min = 50, 50%.

Sup {9, 3} = $\frac{1}{10}$ (10%.)

confi {9 => 3} = $\frac{1}{1}$ (100%)

↑ confi ↓ sup → not meaningful

Q * Apriori Algorithm

Q. Min support = 30%, Min confidence = 75%.

Database D

TID	Items
1	A, B, D
2	B, C, D
3	A, B
4	B, D
5	A, B, C

→ C_1 → Scan D

itemset	Sup
A	3
B	5
C	2
D	3

L_1 ↓ Min sup 30%.

itemset
{A}
{B}
{C}
{D}

(K+1) → Scan D → C_2

itemset	Sup
{A, B}	3
{A, C}	0 x
{A, D}	1 x
{B, C}	2
{B, D}	3
{C, D}	1 x

Min sup 30% ← L_2

{A, B}
{B, C}
{B, D}

↓

		Sup	
Scan D → C_3	{A, B, C}	1	x
	{A, B, D}	1	x
	{B, C, D}	1	x

∵ All combin^{ns} are < 30% support, so no rules for associa^{n} can be generated

Q. min support = 30%, min confi = 80%.

→

Transaction	Items
T1	Bread, Jelly, Butter
T2	Bread, Butter
T3	Bread, Milk, Butter
T4	Coke, Bread
T5	Coke, Milk

C_1 → Scan D

	Sup	
Bread	4	
Jelly	1	x
~~Bread~~ Butter	3	
Milk	2	
Coke	2	

↓ Min sup 30%.

~~{Br, J}~~	1	{Br}
{Br, Bt}	3	~~{J}~~
{Br, M}	1	{Bt}
{Br, C}	1	{M}
~~{J, Bt}~~	~~1~~	{C}
~~{J, M}~~	~~1~~	
~~{J, C}~~	~~1~~	
{Bt, M}	1	
{Bt, C}	1	
{M, C}	1	

{Br, Bt} 3 ← {Br, Bt} 3 ← {Br}, {Bt}

Bread -> Butter = 3/4 = 75% x <80

Butter -> Bread = 3/3 = 100%.

If item comes twice, count only 1

PAGE No.
DATE / /

Q. Slide 29 Chpt 5

O, K, E - 3

Associa :

O —> K, E 3/3 ✓

K —> O, E 3/5
E —> K, O 3/4
O, K —> E 3/3 ✓

O, E —> K 3/3 ✓

K, E —> O 3/4