

Tuo-July

Q:7 For the given dataset below compute average using moving window of size 10 & 11 & 12.

$$\begin{aligned} \text{Data} = & 10.5 + 11.6 + 12.3 + \\ & 13.8 + 14.9 + 15.6 + 16.1 + 17.8 + 18.6 + 19.3 + 20.5 \\ & 16.9 + 17.8 + 18.6 + 19.3 + 20.5 \end{aligned}$$

. 10

$$= \frac{183}{12} = 15.25$$

$$\begin{aligned} \text{Sum}^2 = & (10.5 - 15.25)^2 + (11.6 - 15.25)^2 + (12.3 - 15.25)^2 \\ & + (13.8 - 15.25)^2 + (14.9 - 15.25)^2 + (15.6 - 15.25)^2 + (16.1 - 15.25)^2 \\ & + (17.8 - 15.25)^2 + (18.6 - 15.25)^2 + (19.3 - 15.25)^2 + (20.5 - 15.25)^2 \end{aligned}$$

$$= 122.5$$

$$= \sqrt{122.5} = \sqrt{10.25} = 3.1937$$

2020/05/05 10:02

Suppose each of the three numbers is written as a sum of two squares  
of consecutive integers. Then we have  
that the square remains the same when it is multiplied by 1000.  
Hence - the original number must end in 000, 100, 400, 700 or 900.

$$= \frac{1}{10} = \frac{1}{10} \cdot \frac{2^k}{2^k} = 2^k$$

$$100 + 99 \cdot 9 + 2 \cdot 9 \cdot 9 + 99 \cdot 9 + 99 \cdot 9 + 99 \cdot 9 + 99 \cdot 9$$

10

$$= 2861$$

$$\frac{1}{10} = \frac{1}{10} \cdot \frac{2^k}{2^k} = 2^k$$

$$\begin{aligned} 2^k &= (29 \cdot 0 - 28 \cdot 1)^2 + (29 \cdot 0 - 28 \cdot 1)^2 + (29 \cdot 1 - 28 \cdot 0)^2 \\ &+ (29 \cdot 0 - 28 \cdot 1)^2 + (29 \cdot 1 - 28 \cdot 1)^2 + (29 \cdot 1 - 28 \cdot 0)^2 \\ &+ (29 \cdot 2 - 28 \cdot 1)^2 + (29 \cdot 1 - 28 \cdot 2)^2 \\ &+ (29 \cdot 4 - 28 \cdot 3)^2 \end{aligned}$$

10

$$= 2162 + 0.0801 + 0.0801 + 0.1521 + 0.2901 +$$
  
$$0.2901 + 0.3981 + 0.3981 + 0.6761 + 0.6761$$

10

$$= 23.8961$$

$$= \sqrt{23.3}$$

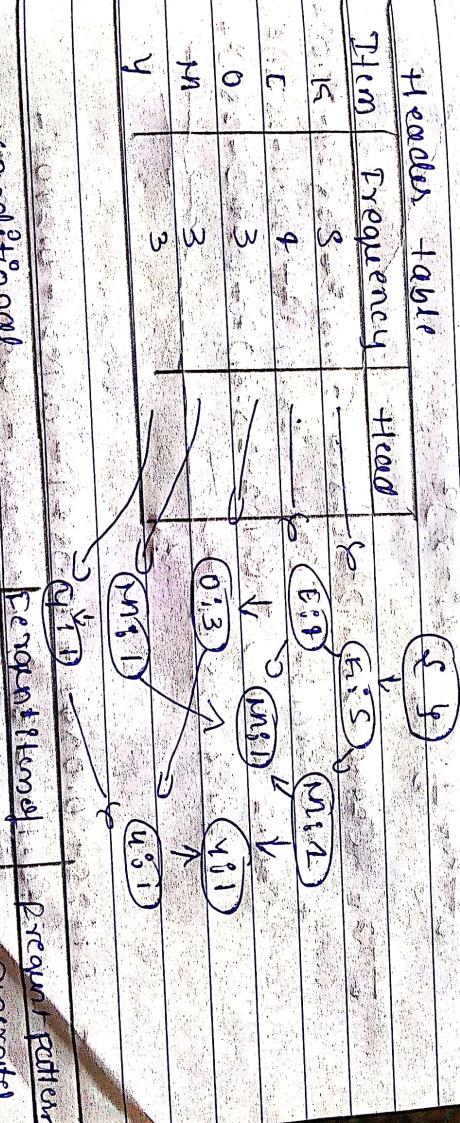
$$\delta = 4.81$$

2020/05/05 10:02

### P - P Growth

Items	Frequency	
E	4	
O	3	TED
M	3	T100 MONKEY
Y	3	T200 DO, MARY
N	2	T300 M, N, K
C	2	T400 M, V, CRY
D	1	T500 CO, OLEL
A	1	
V	1	
I	1	

Header Table



Frequent itemset  
generate

conditional  
patternbase

laptop

(FIS)

K13 E13

(MIS)

K13 (K13)

(K13)

(K13; O13; N13; V13; I13)

(K13)

(K13; O13; N13; V13; I13)

(K13)

10:02

Taco Room Test

L1

Item	Count	Item	Count
A	1	L1	4
C	2	Item	Count
D	1	Set	50
E	4	EY	2
I	1	PM	3
K	5	FO	3
M	3	FY	3
N	8	MO	1
O	3	MY	2
V	1		
U	3		

2020/05/05 10:00

L2

Item	Count	Item	Count
8ut	-	8ut	-
BK	4	EKO	3
ED	3	KMO	1
GM	3	AMY	2
KO	3	KOY	2
GU	3		

L3

Item	Count	Item	Count
8ut	-	8ut	-
EKO	3	ED	3
KMO	1	8ut	-
AMY	2	Set	-
KOY	2	ED	3

## Similacly

$P(Cx)$  buys - computer  $\approx 0.600 \times 0.100 = 0.060$

$$= 0.039$$

10 find the class  $C^0$ , which maximized

$P(Cx|C^0) P(C^0)$  were computed;

3  $P(Cx)$  buys - computer = yes  $P(\text{buys} - \text{computer})$

= NO

$$= 2.0 \cdot 0.039 \times 0.603$$

$= P(Cx)$  buys computer  $\approx 0.028$

$$= 2.0 \cdot 0.019 \times 0.357$$

$$= 0.007$$

∴ The Naive bayesian classifier predicts

buys - computer = yes for tuples x

Ques Suppose a city's average temperature follows  
in July in the last 10 years in value  
descending order are  $29.0, 28.9, 29.0, 29.1$   
 $29.1, 29.2, 29.2, 29.3, 29.4$ . Assuming that the  
average temperature follows a normal distribution  
determined by  $\mu$  &  $\sigma$  illustrate outlier  
detection using maximum likelihood

$\forall x \in$  transaction

Hem - bought

$T_{100}$  Y M, N, E, Y

$T_{200}$  Y D, O, N, K, E, Y

$T_{300}$  Y M, A, K, E, Y

$T_{400}$  Y M V C R U Y

\* The ratio of information from each of  
parents (Dad) =  $1010 / 1010$  = 1.000  
Mother (Mum) =  $0.899 / 0.899$  = 1.000  
 $\rightarrow 0, 0.99, 1.00$

2020/05/05 10:02

Scanned with CamScanner

- b) Massif - the Apple  
 $\rightarrow \text{age} = \text{youth}$  in case of section 1 student  
Yes = creating = having = family  
# computers brought on the training tuples  
P(Churn) = computer = NO =  $\frac{2}{9} = 0.333$   
P(Churn) = computer = YES =  $\frac{7}{9} = 0.667$   
to compute  $P(\text{Churn})$  for  $i=1, 2$  we compute  
The following conditional probabilities  
 $P(\text{Churn} = \text{young} | \text{computer} = \text{YES}) = \frac{2}{9} = 0.222$   
 $P(\text{Churn} = \text{young} | \text{computer} = \text{NO}) = \frac{3}{5} = 0.600$   
 $P(\text{Churn} = \text{no student} | \text{computer} = \text{YES}) = \text{No} = \frac{1}{9} = 0.111$   
 $P(\text{Churn} = \text{no student} | \text{computer} = \text{NO}) = \frac{4}{5} = 0.800$   
 $P(\text{student} = \text{yes} | \text{young} - \text{computer} = \text{YES}) = \frac{6}{9} = 0.667$   
 $P(\text{student} = \text{no} | \text{young} - \text{computer} = \text{NO}) = \frac{2}{5} = 0.400$   
Getting these probabilities we obtain  
 $P(\text{Churn} = \text{young} - \text{computer} = \text{YES}) = \frac{2}{9} \times \frac{6}{9} = 0.133$   
 $P(\text{Churn} = \text{young} - \text{computer} = \text{NO}) = \frac{3}{5} \times \frac{2}{5} = 0.240$   
 $P(\text{Churn} = \text{no student} - \text{young} - \text{computer} = \text{YES}) = \frac{1}{9} \times \frac{4}{5} = 0.044$   
 $P(\text{Churn} = \text{no student} - \text{young} - \text{computer} = \text{NO}) = \frac{4}{5} \times \frac{1}{9} = 0.089$   
 $\therefore 0.240 \times 0.044 + 0.089 = 0.099$

	senior	low	"	high	info
5	mid-age	low	"	high	0.16
6	youth	medium	no	lowest	0.03
7	"	low	yes	low	0.03
8	senior	medium	"	"	0.03
9	mid-age	"	no	highest	0.03
10					

(a) Induction a decision using Information, info + infor(D) =  $\sum_{j=1}^q p_j \log_2 (p_j)$  (expected information)

\* The class label attribute buys-computer has two distinct values (yes, no)

\* q-tuples of class yes

$\begin{matrix} S \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \end{matrix}, \begin{matrix} " \\ " \\ " \\ " \\ " \end{matrix}, \begin{matrix} " \\ " \\ " \\ " \\ " \end{matrix}, \begin{matrix} " \\ " \\ " \\ " \\ " \end{matrix}$

\* Info(D) =  $-q \cdot \log_2 \left( \frac{q}{14} \right) - S \cdot \log_2 \left( \frac{S}{14} \right) = 0.930 \text{ bits}$

If tuples are partitioned according to age,

\* lets start with the attribute age

\* age category youth

\* There are 2 yes & 3 no tuples

\* age category "middle aged"

\* There are 4 yes & 3 no tuples

\* Age category "senior"

There are 3 yes & 2 no tuples

\* Info(D) =  $\sum_{j=1}^q \frac{|D_j|}{|D|} \times \text{infor}(D_j)$

\* Info(D) =  $S \times \frac{2}{14} \log_2 \frac{2}{3} + 3 \times \frac{3}{14} \log_2 \frac{3}{2} + 4 \times \frac{4}{14} \log_2 \frac{4}{7} + 2 \times \frac{2}{14} \log_2 \frac{2}{3}$

= 0.694 bits

$$P(A \cap B) = P(A)P(B)$$

$$P(\text{game} \mid D) = \frac{1000}{6000} = 0.17$$

$$P(\text{video} \mid D) = \frac{1300}{16000} = 0.081$$

0.75

$$P(\text{game} \cap \text{video}) = \frac{4000}{10000} = 0.4$$

(a)

$$\text{left} = P(\text{game} \cap \text{video}) = 0.4 = 0.8 * 0.75$$

\*

0.88 is less than 1  
∴ there is a strong correlation between the occurrence

of game & video

Q4 (a) Briefly outline the process of attribute selection in decision tree induction using information gain measure

(b) classifying the tuple

$x = (\text{age} = \text{young}, \text{income} = \text{medium}, \text{student} = \text{yes}, \text{credit\_rating} = \text{fair})$

using naive bayesian classification on the following training data set

prob	age	income	student	credit rating	class
1	young	high	no	fair	no
2	"	low	no	excellent	no
3	middle	high	yes	fair	yes
4	senior	medium	yes	good	yes

We found a linear relationship between the number of children and the proportion of children with the following contingency table. We can see that there is no correlation between age and number of children with the game. So we can say that the game is popular among children.

Table 2: Contingency table summarizing the financial data with respect to game popularity per age group.

Age	Game	Not Game
≤ 10	3000	3000
11-15	4500	2500
16-20	2000	3000
21+	6000	4000
Total	15500	10500

We can see the association between age and game popularity by doing a Chi-square test. The null hypothesis is that there is no association between age and game popularity. The expected value for each cell is calculated as follows:

Expected values =  $\frac{\text{Row Total} \times \text{Column Total}}{\text{Grand Total}}$

Age	Game	Not Game
≤ 10	4000	3000
11-15	3000	2500
16-20	2000	3000
21+	6000	4000

$\chi^2$  (observed - expected)<sup>2</sup> / expected

$$\frac{(4000 - 3000)^2 + (3000 - 2500)^2 + (2000 - 3000)^2 + (6000 - 4000)^2}{4000}$$

$$12 + 25 + 100 + 400 = 567$$

2020/05/05 10:03

### 3) Normalization by decimal scaling

$$\begin{aligned} \text{3 digits } 200 &= 200 / 1000 = 0.2 \\ 400 &= 400 / 1000 = 0.4 \\ 600 &= 600 / 1,000 = 0.6 \\ 700 &= 700 / 1,000 = 0.7 \\ 800 &= 800 / 1,000 = 0.8 \\ 1000 &= 1000 / 1000 = 1.0 \end{aligned}$$

Q(2) Given the following transactional data  
minimum support count as 2 generate  
candidate items on frequent items using  
apriori algorithm

TID	list of item IDs
T <sub>1</sub>	I <sub>1</sub> I <sub>2</sub> I <sub>3</sub>
T <sub>2</sub>	I <sub>2</sub> I <sub>4</sub>
T <sub>3</sub>	I <sub>2</sub> I <sub>3</sub>
T <sub>4</sub>	I <sub>1</sub> I <sub>2</sub> I <sub>4</sub>
T <sub>5</sub>	I <sub>1</sub> I <sub>3</sub>
T <sub>6</sub>	I <sub>2</sub> I <sub>3</sub>
T <sub>7</sub>	I <sub>1</sub> I <sub>3</sub>
T <sub>8</sub>	I <sub>1</sub> I <sub>2</sub> I <sub>3</sub> I <sub>5</sub>
T <sub>9</sub>	I <sub>1</sub> I <sub>2</sub> I <sub>3</sub>

= There are 9 transactions in the db i.e. DFB  
minimum support counts =  $\frac{9}{2} = 4.5$  min. sup = 2

The corresponding relative support is  $\frac{2}{9} = 22\%$

Scan # for count of each candidate	Item Set	sup count	compare and update		Item Set	sup Count
			Support Count with minimum Support Count	6		
	I <sub>1</sub>	6			T <sub>1</sub>	6
	I <sub>2</sub>	7			I <sub>2</sub>	7
	I <sub>3</sub>	6			I <sub>3</sub>	6
	I <sub>4</sub>	3			I <sub>4</sub>	2
	I <sub>5</sub>	2			I <sub>5</sub>	2

2021/05/05 20:00

$$S.D = \sqrt{\frac{(200 - 616.66)^2 + (400 - 616.66)^2 + (600 - 616.66)^2 + (400 - 616.66)^2 + (800 - 616.66)^2 + (1000 - 616.66)^2}{6-1}}$$

$$= \sqrt{(-833.33)^2 + (-433.33)^2 + (-33.33)^2 + (694.44)^2 + (336.67)^2 + (1469.99)^2} + 5$$

$$\sqrt{186208.69} \\ 5$$

$$\sqrt{37241.738}$$

$$\sigma = 192.98$$

$$\mu = 616.66$$

$$SD = 192.98$$

$$Z - \text{SCORE} \quad z - \mu = \frac{200 - 616.66}{192.98} = -2.1591$$

$$= \frac{400 - 616.66}{192.98} = -1.1227$$

$$= \frac{600 - 616.66}{192.98} = -0.0863$$

$$= \frac{700 - 616.66}{192.98} = 0.4318$$

$$= \frac{800 - 616.66}{192.98} = 0.9500$$

$$= \frac{1000 - 616.66}{192.98} = 1.9864$$

Mean

2020/05/05 10:03

$$\text{For data : } 700$$

$$V'_1 = \frac{700 - 200}{1000 - 200} * (1 - 0) + 0$$

$$= \frac{500}{800} * 1 + 0$$

$$= 0.625$$

$$\text{For data : } 800$$

$$V'_2 = \frac{800 - 200}{1000 - 200} * (1 - 0) + 0$$

$$= \frac{600}{800} * 1 + 0$$

$$= 0.75$$

$$V'_3 = \frac{1000 - 200}{1000 - 200} * (1 - 0) + 0$$

$$= \frac{800}{800} * 1 + 0$$

$$= 1$$

data Data after normalization

900	0
400	0.25
600	0.5
700	0.625
800	0.75
1000	1

## 2) Z-Score Normalization

$$\text{Mean : } \frac{900 + 400 + 600 + 700 + 800 + 1000}{6} = 3400$$

$$= 566.66$$

$$\text{standard deviation } \sqrt{\frac{\sum (\text{every individual value} - \text{mean})^2}{6}}$$

## Data Mapping

Ques - 18  
why data should be normalized? Normaliz.

(Q1) the group of data 900, 900, 600, 700, 800, 1000 using the following methods  
1) min-max normalization by setting min = 0 & max = 1

2) score normalization

3) Normalization by decimal scaling.

→ Data normalization is to minimize or even exclude duplicate data. This is very essential and important issue because it is increasingly problematic to keep in data in relational data bases which store identical data in more than one place.

D) min-max normalization by setting

$$\text{min} = 0 \text{ & } \text{max} = 1$$

$$v'_i = \frac{v_i - \text{min}_A}{\text{max}_A - \text{min}_A} \rightarrow \text{New min}_A$$

$$v'_S = \frac{900 - 200}{1000 - 200} * (1-0) + 0$$

$$= \frac{0}{800} * (1) + 0 \\ = 0$$

For data = 400

$$v'_I = \frac{400 - 200}{1000 - 200} * (1-0) + 0$$

$$= \frac{200}{800} * 1 + 0$$

$$= 0.25$$

$$\text{For data } 600 : v'_J = \frac{600 - 200}{1000 - 200} * (1-0) + 0$$

$$= \frac{400}{800} * 1 + 0 \quad 2020/05/05 10:03$$