Two-July

Q8:7 For the given dataset below identify outliers using maximum likelihood 10.5 11.6, 11.7 12.3, 13.8, 14.9, 15.6 16.1 17.8, 18.6, 19.3 20.5

$$\hat{\mu} = \frac{10.5 + 11.6 + 11.7 + 12.3 + 13.8 + 14.9 + 15.6 + 16.1 + 17.8 + 18.6 + 19.3 + 20.5}{12}$$

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

$$\hat{\sigma}^2 = (10.5 - 15.25)^2 + (11.6 - 15.25)^2 + (11.7 - 15.25)^2 + (12.3 - 15.25)^2 + (13.8 - 15.25)^2 + (14.9 - 15.25)^2 + (15.6 - 15.25)^2 + (16.1 - 15.25)^2 + (17.8 - 15.25)^2 + (18.6 - 15.25)^2 + (19.3 - 15.25)^2 + (20.5 - 15.25)^2$$

$$\frac{2 \cdot 122.5}{12}$$

$$= \sqrt{10.20}$$

$$= \hat{\sigma} = 3.1937$$

Scanned with CamScanner

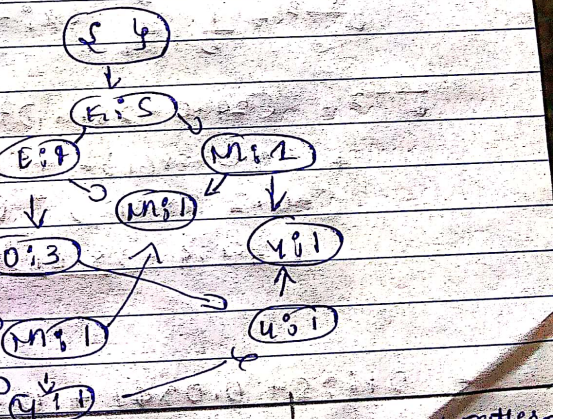
F-P Growth

Items	Frequency
K	5
E	4
O	3
M	3
Y	3
N	2
C	2
D	1
A	1
V	1
I	1

TID	item bought	order Frequent
T100	M O N E E Y	K E O M Y
T200	D O N K E Y	K E O Y
T300	M A K E	K E M
T400	M O V C A Y	K M Y
T500	C O O K I E	K E O

Headder table

Item	Frequency	Head
K	5	
E	4	
O	3	
M	3	
Y	3	



Item	Conditional Pattern base	Frequent item set	Frequent pattern generated
E	K:4	K:4	(K E:4)
O	K E:3	K:3 E:3	K:O:3)(E:O:3)
M	(K E O:1)(K E:1)(K:1)	K:3	(K M:3)
Y	(K E O M:1)(K E O:1)(K M:1)	K:3	(K Y:3)

Tree Traversal

C1

Item	count
Set	4
A	1
C	2
D	1
E	4
I	1
K	5
M	3
N	8
O	3
V	1
Y	3

L1

Item	count
Set	
E	4
K	5
M	3
O	3
Y	3

C2

Item	count
Set	4
EK	4
EM	2
EO	3
EY	2
KM	3
KO	3
KY	3
MO	1
MY	2
MY	2

L2

item	count
Set	
EK	4
EO	3
KM	3
KO	3
KY	3

C3

item	count
Set	
EKO	3
KMO	1
KMY	2
KOY	2

L3

item	count
Set	
Set	
EKO	3

Similarly

P(x, buys computer = NO) = 0.600 × 0.400 × 0.200 × 0.400

= 0.019

To find the class, c_i, that maximized $P(x|c_i) P(c_i)$ we compute:

$P(x|buys\ computer = yes) P(buys\ computer = NO)$

$= 0.044 \times 0.643$

$= 0.028$

$P(x|buys\ computer = NO) P(buys\ computer)$

$= 0.019 \times 0.357$

$= 0.007$

∴ The Naive bayesian classifier predicts buys computer = yes for tuple x

Q.5 Suppose a city's average temperature values in July in the last 10 years in value ascending order are 28.0, 28.9, 29.0, 29.1, 29.2, 29.2, 29.3, 29.4, 29.4. Assuming that the average temperature follows a normal distribution determined by μ, σ illustrate outlier detection using maximum likelihood

∀ x ∈ transaction

TID	Item - bought
T_{100}	{M, N, K, E, Y}
T_{200}	{D, O, N, K, E, Y}
T_{300}	{M, A, K, E}
T_{400}	{M, U, C, K, Y}

* The gain in information from each of partitioning would be
Gain (age) : Info (D) - Info$_{age}$ (D)
= 0.940 - 0.694
= 0.246 bits

b) classify the tuple
X = age = youth, income = medium, student = yes, credit-rating = fair)

* compute based on the training tuples
P(buys computer = yes) = 9/14 = 0.643
P(buy computer = NO) = 5/14 = 0.357

* To compute P(X|Ci) for i = 1, 2 we compute the following conditional probabilities
P(age = youth | buys computer = yes) 2/9 = 0.222
P(" " " NO) 3/5 = 0.600
P(income = medium | " " = yes) 4/9 = 0.444
P(" " " = NO) 2/5 = 0.400
P(student = yes | buys computer = yes) = 6/9 = 0.667
P(" " = NO) 1/5 = 0.200
P(credit rating fair | " = yes) = 6/9 = 0.667
P(" " NO) 2/5 = 0.400

using these probabilities we obtain
P(X | buys computer = yes)
P(age = youth | buys computer = yes) x
P(income = medium | " = yes) x
P(credit rating = fair | " yes) x
P(student yes | " yes)
= 0.222 x 0.444 x 0.667 x 0.667
= 0.044

5	senior	low	yes	fair	yes
6	mid-age	low	yes	excellent	yes
7	youth	medium	no	fair	no
8	"	low	yes	"	yes
9	senior	medium	yes	"	yes
10	midage	"	no	excellent	yes

(a) Induction a decision using information gain

* $\text{Info}(D) = -\sum_{i=1} p_i \log_2(p_i)$ (Expected information)

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

* $\frac{9}{5}$ tuple of class yes &

* " " " " No

* $\text{Info}(D) = -\frac{9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right) = 0.940$ 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

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

* Info to age $(D) = \frac{5}{14} \times -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}$

$\frac{4}{14} \times \left(-\frac{4}{4}\log_2\frac{4}{4}\right) + \frac{5}{14} \times \left(-\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}\right)$

$= 0.694$ bits

* $lift(A,B) = \frac{P(A \cup B)}{P(A)\,P(B)}$

$P(\text{game } \&)$ $\frac{6000}{10000} = 0.6$

$P(\text{video } \&)$ $\frac{7500}{10000} = 0.75$

0.75

$P(\text{game wideo})$ $\frac{4000}{10000}$ 40%

lift is $\frac{P(\text{game video})}{P(\text{game} * P(\text{video})} = \frac{0.4}{0.6 * 0.75} = 0.88$

0.88 is less than 1

∴ is a ve correlation between the occurence of game & video

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

(b) classify the tuple

x = (age = youth, income = medium, student = yes, credits_rating = fair)

using naive bayesion classification on the following training data set

Rid	age	Income	student	credit rating	class buys_comp
1	youth	high	NO	fairs	no
2	"	"	"	excellent	no
3	middle	"	"	fair	yes
4	senior	medium	"	"	yes

3) Normalization by decimal scaling

3 digits 200 = 200/1000 = 0.2
400 = 400/1000 = 0.4
600 = 600/1000 = 0.6
700 = 700/1000 = 0.7
800 = 800/1000 = 0.8
1000 = 100/1000 = 0.1

Q:2) Given the following transactional data sseming minimum support count as 2 generate candidate itemsets on frequent itemsets using apriori Algorithm

T_{ID}	list of item IDS
T_1	I_1 I_2 I_5
T_2	I_2 I_4
T_3	I_2 I_3
T_4	I_1 I_2 I_4
T_5	I_1 I_3
T_6	I_2 I_3
T_7	I_1 I_3
T_8	I_1 I_2 I_3 I_5
T_9	I_1 I_2 I_3

=> There are 9 transactions in the db i.e D:2 P

minimum support counts = 2 i.e min. sup = 2

The corresponding Relative support is 2/9 = 22%.

Scan 1 for count of each candidate

Item set	sup count
I_1	6
I_2	7
I_3	6
I_4	2
I_5	2

compare candidate support count with minimum support count

6

Item set	sup Count
I_1	6
I_2	7
I_3	6
I_4	2
I_5	2

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

$$= \sqrt{\frac{(-833.32) + (-433.32) + (-33.32) + (6948.55) + (3613.55) + (146994.55)}{5}}$$

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

$$\sqrt{37241.738}$$

$$\sigma = 192.98$$

μ - Mean = 616.66

SD = 192.98

$$Z\text{-score} \quad \frac{x-\mu}{\sigma} = \frac{200-616.66}{192.98} = -2.1591$$

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

$$= \frac{600-616.66}{19298} = -0.0863$$

$$= \frac{700-616.66}{19298} = 04318$$

$$\frac{800-616.66}{192.98} = 0,9500$$

$$= \frac{1000-616.66}{192.98} \quad 19864$$

Mean

2020/05/05 10:03

For data : 700

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

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

$= 0.625$

For data : 800

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

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

$= 0.75$

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

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

$= 1$

data	Data after normalization
200	0
400	0.25
600	0.5
700	0.625
800	0.75
1000	1

2) z-Score Normalization:

Mean : $\frac{200 + 400 + 600 + 400 + 800 + 1000}{6} = \frac{3700}{6}$

$= 616.66$

Standard deviation $\sqrt{\frac{\sum (\text{Every individual value} - \text{mean})^2}{N-1}}$

2020/05/05 10:03

Data Mining

Jun - 18

Q1) why data should be normalized? Normalize the group of data 900, 400, 600, 700, 800, 1000 using the following methods

1) min - max normalization by setting min = 0 & max = 1

2) Z - Score normalization

3) Normalization by decimal scaling.

→) Data normalization is to minimize or even exclude duplicated 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

1) min-max Normalization by setting

min = 0 & max = 1

$$v_i' = \frac{v_i - \min A}{\max A - \min A} (\text{New max}_A - \text{New min}_A) + \text{New min}_A$$

$$v_1' = \frac{900 - 200}{1000 - 200} * (1-0) + 0$$

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

$$= 0$$

For data = 400

$$v_2' = \frac{400 - 200}{1000 - 200} * (1-0) + 0$$

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

$$= 0.25$$

For data 600 : $v_3' = \frac{600 - 200}{1000 - 200} * (1-0) + 0$

$$= \frac{400}{800} * 1 + 0$$

2020/05/05 10:03