44
-5

112-1 Introduction to Data Science

Midterm Exam

Student ID: 11204604 Name: Tse Wen Hong

1. **(10%)** Please briefly describe the major tasks in data preprocessing.

Ch. 3

Key word! Data Cleaning Data integration ...

Try to find out the relation in between the attributes, and the relation in between the data. So that we can make better prediction or judgement according the finding result.

-5

2. **(20%)** Given two objects represented by the tuples **X1 (20, 10, 5, 3)** and **X2 (30, 8, 6, 16)**:
   (a) Compute the Euclidean distance between the two objects.
   (b) Compute the Manhattan distance between the two objects.
   (c) Compute the Minkowski distance between the two objects, using h=3.
   (d) Compute the Pearson correlation coefficient between the two objects. Please describe the relationship between X1 and X2.

(a) i. $(20-30)^2+(10-8)^2+(5-6)^2+(3-16)^2 = 100+4+1+169 = 274$

ii. $\sqrt{274} = 16.5529$

Euclidean distance is

(b) $|20-30|+|10-8|+|5-6|+|3-16| = 10+2+1+13 = 26$

Manhattan distance is: 26

(c) i. $(20-30)^3+(10-8)^3+(5-6)^3+(3-16)^3 = 1000+8+1+2197 = 3206$

Minkowski distance is: $\sqrt[3]{3206}$

(sorry my calculator have no power function)

(d) i. $u_x = (20+10+5+3)/4 = 9.5$

$u_y = (30+8+6+16)/4 = 15$

$\frac{Cov(A,B)}{\delta_A \delta_B}$

ii. (turn to the back)

(ii)

x-μ	10.5	0.5	-4.5	-6.5
y-μ	15	22	24	14

(ii) $r = \dfrac{(10.5\times15)+(0.5\times22)+(-4.5\times24)+(-6.5\times14)}{\sqrt{10.5^2+0.5^2+(-4.5)^2+(-6.5)^2}\times\sqrt{15^2+22^2+24^2+14^2}}$

$= \dfrac{157.5+11-108-91}{\sqrt{110.25+0.25+20.25+42.25}\times\sqrt{225+484+576+196}}$

$= \dfrac{-30.5}{\sqrt{173}\times\sqrt{1481}}$

$= \dfrac{-30.5}{13.1529\times38.4838}$

$= \dfrac{-30.5}{506.1753}$

$r =$ ~~-0.60~~ -0.0602

~~x~~ 0.75

先写出公式 再套用。 $r = \dfrac{Var(A,B)}{\sigma_A\sigma_b}$

3. **(20%)** Suppose a group of 12 students with the test scores listed as follows:

19, 71, 48, 63, 35, 85, 69, 81, 72, 88, 99, 95.

Partition them into **four bins** by each of the following methods.

(a) equal-frequency (equi-depth) partitioning

(b) equal-width partitioning

19, 35, 48 | 63, 69, 71 | 72, 81, 85 | 88, 95, 99

(a) bin 1 = 19, 35, 48

bin 2 = 63, 69, 71

bin 3 = 72, 81, 85

bin 4 = 88, 95, 99

(b) (i) (99 − 19) / 4 = 20

(ii) bin 1 = 19, 35

bin 2 = 48

bin 3 = 63, 69, 71, 72

bin 4 = 81, 85, 88, 95, 99

4. **(30%)** For the following group of data with mean=600 and variance= 260000

100, 200, 400, 800, 1500

−26

(a) Normalize the above group of data by min-max normalization with min = 0 and max = 10.

(b) In z-score normalization, what value should the third number 400 be transformed to?

(c) Normalize the above values by decimal scaling.

(a) $100 \Rightarrow \frac{100-100}{1500-100} \times (10-0) = 0$

$200 \Rightarrow \frac{100}{1490} \times 10 = 0.6711$ ✗

$400 \Rightarrow 300/1490 \times 10 = 2.0134$ ✗

$800 \Rightarrow 700/1490 \times 10 = 4.6979$ ✗

$1500 \Rightarrow 10$

(b) (c) 本来已经会了

这两天都在搞 Apriori，现在脑子乱了。

需再回去複習

3. **(20%)** Suppose a group of 12 students with the test scores listed as follows:

19, 71, 48, 63, 35, 85, 69, 81, 72, 88, 99, 95.

Partition them into **four bins** by each of the following methods.

(a) equal-frequency (equi-depth) partitioning

(b) equal-width partitioning

19, 35, 48, | 63, 69, 71 | 72, 81, 85 | 88, 95, 99

(a) bin 1 = 19, 35, 48
bin 2 = 63, 69, 71
bin 3 = 72, 81, 85
bin 4 = 88, 95, 99

(b) (i) (99−19)/4 = 20
(ii) bin 1 = 19, 35
bin 2 = 48,
bin 3 = 63, 69, 71, 72,
bin 4 = 81, 85, 88, 95, 99

這就是 Decimal Scaling
$\frac{18}{10^4} = 0.01$
$\frac{200}{10^4} = 0.02$

4. **(30%)** For the following group of data with mean=600 and variance= 260000

100, 200, 400, 800, 1500 ← J=4 (4位數)

−26

(a) Normalize the above group of data by min-max normalization with min = 0 and max = 10.

(b) In z-score normalization, what value should the third number 400 be transformed to?

(c) Normalize the above values by decimal scaling.

$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \times (Y_{max} - Y_{min}) + Y_{min}$

(a) $100 \Rightarrow \frac{100-100}{1500-100} \times (10-0) = 0$

$200 \Rightarrow \frac{100}{1490} \times 10 = 0.6711$ × 減錯了 $\Rightarrow \frac{200-100}{1400} = \frac{100}{1400} = 0.0714 \times 10 = 0.714$

$400 \Rightarrow 300/1490 \times 10 = 2.0134$ ×

$800 \Rightarrow 700/1490 \times 10 = 4.6979$ ×

$1500 \Rightarrow 10$

(b) (c) 本來已經忘了

New value $= (X-\mu)/\delta$

這兩天都在想 Apriori，現在腦子亂了。需再回去複習

$\mu = 600$

$\delta = \sqrt{(100-600)^2 + (200-600)^2 + (400-600)^2 + (1500-600)^2} = 250000 + 160000 + 40000 + 810000 =$

-20

5. **(20%)** A database has 5 transactions. Let minimum support be 60% and minimum confidence be 80%.

Customer	Date	Items_boughts
100	10/15	{I, P, A, D, B, C}
200	10/15	{D, A, E, F}
300	10/16	{C, D, B, E}
400	10/18	{B, A, C, K, D}
500	10/19	{A, G, T, C}

(1) Support count = 60% × 5 = 3

(a) List the frequent k-itemset for the largest k. (10%) ⇒ {B, C, D}

(b) List all the strong association rules (with support and confidence) for the following shape of rules.

$\forall x \in transaction, buys(x, item_1) \wedge buys(x, item_2) \Rightarrow buys(x, item_3)$ [sup., conf.] (10%)

$m = 5$

C_1

item	count	support
A	4	80%
B	3	60%
C	4	80%
D	4	80%
~~E~~	~~2~~	~~40%~~
F	1	
G	1	
I	1	
K	1	
T	1	

$P(D|A) = \frac{60\%}{80\%} = 75\%$

$P(A|D) = \frac{60\%}{80\%} = 75\%$

{ B, C ⇒ D conf = 3/3 = 100%
B, D ⇒ C 100%
C, D ⇒ B 100%

C_2

		sup.	
{A, B}	1	20%	
{A, C}	2	40%	
{A, D}	3	60%	
{B, C}	3	60%	60%
{B, D}	2	40%	60%
{C, D}	2	40%	60%

C_3

ABC	2	
ACD	2	
ABD	2	
BCD	3	3/5 = 60%

confidence:

{A, D} ⇒ B $P(B|\{A,D\}) = \frac{60\%}{60\%} = 100\%$

{A, D} ⇒ C $P(C|\{A,D\}) = \frac{80\%}{60\%} = 133\%$

~~{A, D} ⇒ B~~

frequent k-itemset is {A, D}

strong association rule is

{A, D} ⇒ B confidence 100%

{A, D} ⇒ C confidence 133%