中国科学院研究生院

试题专用纸

课程编号：712009Z

课程名称：数据挖掘

任课教师：刘莹

姓名________________ 学号________________ 成绩________________

1. Suppose a hospital tested the *age* and *body fat* for 18 randomly selected adults with the following result:

age	39	57	41	58	47	58	49	60	50	61	23	52	23	54	27	54	27	56
%fat	31.4	30.2	25.9	34.1	27.4	32.9	27.2	41.2	31.2	35.7	9.5	34.6	26.5	42.5	7.8	28.8	17.8	33.4

(a) Use smoothing by bin means to smooth the *age* data, using a bin depth of 6. Illustrate your steps. (5 points)

23 23 27 27 39 41 | 47 49 50 52 54 54 | 56 57 58 58 60 61

(b) Partition the *age* into 3 bins by equal-width partitioning, and use bin boundary to smooth each bin. (5 points)

$(61-23)/3 = 38/3 \approx 13$ 23 − 36 36 − 49 49 − 62

(c) Use min-max normalization to transform the value 49 for *age* onto the range [0.0, 1.0]. (5 points)

$(49-23)/(61-23)+0 = 26/38 = 13/19$

(d) Use z-score normalization to transform the value 41.2 for *body fat*, where the standard deviation of *body fat* is 9.25. (5 points)

$\frac{41.2 - \text{body}}{9.25}$

2. Given a transaction database below, let min_support = 50% and min_confidence = 75%:

support = $P(A\cap B)$

Conference = $\frac{P(A\cap B)}{P(A)}$

TID	Items-bought
T1	{a,b,e}
T2	{a,b,c,e}
T3	{a,b,d,e}
T4	{a,c,d,e}
T5	{b,c,e}
T6	{b,d,e}
T7	{c,d}
T8	{a,b,c}
T9	{a,b,e}
T10	{a,b,e}

support = 5

C_1	L_1	C_2	L_2
a >=5	a	ab >=5	ab
b >=5	b	ac 3	ae
c >=5	c	ae >=5	be
d 4	e	bc 3	
e >=5		be >=5	
		ce 3	

$C_3 = L_3$

abe 5

(a) Find all frequent itemsets using Apriori method. Write up frequent itemsets and candidate set at each level. (10 points)

(b) Find all frequent itemsets using FP-growth method. Write up the conditional pattern base for each item, and the conditional FP-tree for each item. (10 points)

(c) Using the resulting frequent itemsets, find all strong associations in terms of the following rule format:

For any transaction *x*, buys(*x*, item1) ∧ buys(*x*, item2) ⇒ buys(*x*, item3) [s=?%, c=?%]. (5 points)

abe: support = $\frac{5}{10}$ = 50%

confidence (ab ⇒ e) = $\frac{P(abe)}{P(ab)} = \frac{50\%}{60\%} > 75\%$

confidence (ae ⇒ b) =

confidence (be ⇒ a) =

3. Given a data set below with three attributes {A, B, C} and two classes {C1, C2}. Build a decision tree, using information gain to select and split attribute. (15 points)

select B.

C1 = p C2 = n p = 6 n = 4

$I(p,n) = -\frac{6}{10}\log\frac{6}{10} - \frac{4}{10}\log\frac{4}{10}$

For A: A=0 p=3 n=1

A=1 p=3 n=3

$E(A) = \frac{4}{10} I(3,1) + \frac{6}{10} I(3,3)$

For B: B=0 p=5 n=0

B=1 p=1 n=4

$E(B) = \frac{5}{10} I(5,0) + \frac{5}{10} I(1,4)$

For C: C=0 p=4 n=3

C=1 p=2 n=1

$E(C) = \frac{7}{10} I(4,3) + \frac{3}{10} I(2,1)$

Instance	A	B	C	Class	
1	0	0	0	C1	×
2	0	0	1	C1	×
3	0	1	0	C1	×
4	0	1	1	C2	
5	1	0	0	C1	×
6	1	0	0	C1	×
7	1	1	0	C2	
8	1	0	1	C1	×
9	1	1	0	C2	
10	1	1	0	C2	

4. Consider the following data set. Use Naïve Bayesian Classifier to predict the class label for a test sample (A=0, B=1, C=0). (10 points)

令 X = {A=0, B=1, C=0}

$P(C?|X) = \frac{P(C?\,X)}{P(X)}$

$= \frac{P(X|C?)\cdot P(C?)}{P(X)}$

比较 P(X|C?) 即可

$P(X|C1) = P(A=0|C1)\cdot P(B=1|C1)\cdot P(C=0|C1)$

$= \frac{2}{5} \times \frac{1}{5} \times \frac{1}{5} = \frac{2}{125}$

$P(X|C2) = \frac{3}{5} \times \frac{2}{5} \times 0 = 0$

∴ label ⇒ C_1

Record	A	B	C	Class
1	0	0	0	C1
2	0	0	1	C2
3	0	1	1	C2
4	0	1	1	C2
5	0	0	1	C1
6	1	0	1	C1
7	1	0	1	C2
8	1	0	1	C2
9	1	1	1	C1
10	1	0	1	C1

5. Given a data set of 8 sample points. Perform K-means to generate 3 clusters. Suppose initially we assign point 1,2,3 as the center of each cluster. Note: list the clusters at each iteration. (15 points)

for (1,1) (2,1) $(\frac{11}{4}, 4)$

{1} {2,4,5,6} {3}

d(1,2) ✓

{1,2} {4,5,6} {3}

for $(1, \frac{3}{2})$ $(\frac{10}{3}, \frac{14}{3})$ (2,1)

$d(1,4) = \sqrt{4+9} = \sqrt{13}$

$d(2,4) = \sqrt{4+4} = \sqrt{8}$ ✓

$d(3,4) = \sqrt{1+9} = \sqrt{10}$

$d(1,5) = \sqrt{4+16}$

$d(2,5) = \sqrt{4+9} = \sqrt{13}$ ✓

$d(3,5) = \sqrt{1+16}$

$d(1,6) = \sqrt{9+16}$

$d(2,6) = \sqrt{9+9} =$ ✓

$d(3,6) = \sqrt{4+16}$

ID	Attribute 1	Attribute 2
1	1	1
2	1	2
3	2	1
4	3	4
5	3	5
6	4	5

{1,2}

{3}

{4,5,6}

6. Suppose that a large store has a transaction database that is distributed among four locations. Transactions in each component database have the same format, namely T_j: $\{i_1, \ldots, i_m\}$, where T_j is a transaction identifier, and i_k ($1 <= k <= m$) is the identifier of an item purchased in the transaction. Propose an efficient algorithm to mine global association rules (without considering multilevel associations). You may present your algorithm in the form of an outline. Your algorithm should not require shipping all of the data to one site and should not cause excessive network communication overhead. (15 points)