COMP 7103B Data Mining Assignment 1

Jiao Yuqing 3035553721 Feb 25, 2019

Question 1 Frequent Itemset

a) Frequent itemsets: for pass 1 are **a,b,d**; for pass 2 are **ab, bd**. And no frequent items in pass 3.

ID	Baskets	Pass 1	Pass 2	Pass 3
1	a, b, c, e	a:1/b:1/c:1/e:1	ab:1	/
2	a, d, b	a:2/d:1/b:2	ad:1/db:1/ab:2	/
3	c, b	c:2/b:3	/	/
4	a, b, d, e	a:3/b:4/d:2/e:2	ab:3/bd:2/ad:2	/
5	b, d	b:5/d:3	bd:3	/
6	a, b	a:4/b:6	ab:4	/
7	a	a:5	/	/
Frequent Items(s>0.4)		a,b,d	ab,bd	/

b) support(b–>d) = 3/7=0.429; confidence(b->d)=support(b->d)/support(b)=3/6=0.5

c) Frequent itemsets: for pass 1 are **1,2,3,4,7**; for pass 2 is **(2,7)**.

ID	Baskets	Pass 1	Pass 2
1	1, 3, 4	1:1/3:1/4:1; (B1:2/B2:1)	(1,3):1/(3,4):1
2	4, 5	4:2/5:1; (B0:1)	/
3	2, 7	2:1/7:1; (B0:2)	(2,7):1
4	1, 6	1:2/6:1; (B1:3)	/
5	2, 7	2:2/7:2; (B0:3)	(2,7):2
6	3	3:2/	/
Frequent Items(s>0.33)		1,2,3,4,7; (B0,B1)	(2,7)

Question 2 K-means

1. P1,P2,P3,P4 belongs to Clustering 1; another 3 points, P5,P6,P7, belongs to Clustering 2.

	Centroid 1	Distance to centroid 1	Centroid 2	Distance to centroid 2	Result
Step 1 initial computing					
P1(0,0)	P1(0,0)	0	P4(1,1)	$\sqrt{2}$	1
P2(1,0.5)		$\sqrt{1.25}$		0.5	2
P3(1,0.5)		$\sqrt{1.25}$		0.5	2
P4(1,1)		$\sqrt{2}$		0	2
P5(4,0)		4		$\sqrt{10}$	2
P6(4,1)		$\sqrt{17}$		3	2
P7(5,1)		$\sqrt{26}$		4	2
Step 2 recomputing					

P1(0,0)	P1(0,0)	0	(8/3,2/3)	$2\sqrt{17}/3$	1
P2(1,0.5)		$\sqrt{1.25}$		$\sqrt{101}/6$	1
P3(1,0.5)		$\sqrt{1.25}$		$\sqrt{101}/6$	1
P4(1,1)		$\sqrt{2}$		$\sqrt{26}/3$	1
P5(4,0)		4		$2\sqrt{5}/3$	2
P6(4,1)		$\sqrt{17}$		$\sqrt{17}/3$	2
P7(5,1)		$\sqrt{26}$		$5\sqrt{2}/3$	2
Step 3 final computing- -**no change**					
P1(0,0)	(3/4,1/2)	$\sqrt{13}/4$	(13/3,2/3)	$\sqrt{173}/3$	1
P2(1,0.5)		1/4		$\sqrt{401}/6$	1
P3(1,0.5)		1/4		$\sqrt{401}/6$	1
P4(1,1)		$\sqrt{5}/4$		$\sqrt{101}/3$	1
P5(4,0)		$\sqrt{173}/4$		$\sqrt{5}/3$	2
P6(4,1)		$\sqrt{173}/4$		$\sqrt{2}/3$	2
P7(5,1)		$\sqrt{293}/4$		$\sqrt{5}/3$	2

2. P1=1; P2=2; P3=3.2; P4=7.4; P5=8.6; P6=9.6;
Set K=3, initial centroids are C1=1.01; C2=5.3; C3=9.59
In first clustering computing step:
P1,P2 ->C1; P3,P4->C2; P5,P6->C3;
Next clustering computing step: C1'=1.5; C2'=5.3; C3'=9.1
P1,P2,P3->C1'; P4,P5,P6->C3'. C2' is empty.
P1=0;P2=c;P3=2+c;P4=2+2c;P5=2+3c;P6=4+c,(initial P1,2,6)

Question 3 FI & AR Exercise

1) SSE=20.614168118035984, with default value: max_iter=300, n_init=10, tol=0.0001.

1619.???

2) A) **max_iter**(Maximum number of iterations of the k-means algorithm for a single run.) and **n_init**(Number of time the k-means algorithm will be run with different centroid seeds) might impact the value of SSE most. The main reason possibly is that if iterations/run and run time/centroid increase, which means that the clustering computing process more times and after massive times calculating, we might find minimum SSE. B) Both max_iter and n_init increase, while the value of SSE decrease, and result will be better.

max_iter	**n_init**	**SSE**
300	**10**	20.614168118035984
10000	**10**	19.07382224559073
30000	**10**	17.6364294376268
300	**5**	21.54088413034848
300	**10**	20.614168118035984
300	**15**	17.692323335713237

3) Green:techology//Orange:Financial//Pink:Medical//Red:Oil//Blue:Retail
[Retail]cluster 1 : ['American Express', 'Boeing', 'Kraft', 'Procter & Gamble', 'AT&T', 'Merck', 'Johnson & Johnson']
[Financial]cluster 2 : ['Bank of America']

[Retail]cluster 3 : ['The Home Depot', 'Wal-Mart', 'United Technologies', 'ExxonMobil', 'Travelers']

[Technology]cluster 4 : ['Walt Disney', 'Hewlett-Packard', '3M']

[Technology]cluster 5 : ['IBM', 'General Electric']

[Technology]cluster 6 : ['Cisco Systems', 'Alcoa']

[Technology]cluster 7 : ['DuPont', 'Caterpillar', 'Verizon', 'Intel']

[Retial]cluster 8 : ['Chevron', 'Microsoft', 'Pfizer', 'JPMorgan Chase', 'McDonalds', 'Coca-Cola']

4) Cluster：[5 4 6 1 7 5 7 2 7 2 0 0 6 4 1 7 6 3 0 5 5 5 3 0 3 4 4 2 6 4]
SSE=0.4804725547237445.

The value of SSE is smaller than the normal one. And the classifier is more efficient. Normalizing the data is important to ensure that the distance measure accords equal weight to each variable. Without normalization, the variable with the largest scale will dominate the measure.

[Technology]cluster 1 : ['IBM', 'The Home Depot', 'General Electric', 'ExxonMobil']

[Technology]cluster 2 : ['Cisco Systems', 'Alcoa']

[Financial]cluster 3 : ['Bank of America', 'Microsoft', 'Coca-Cola']

[Financial]cluster 4 : ['Wal-Mart', 'United Technologies', 'Travelers']

[Retail]cluster 5 : ['Boeing', 'Procter & Gamble', 'JPMorgan Chase', 'McDonalds', 'Johnson & Johnson']

[Medical]cluster 6 : ['American Express', 'Kraft', 'AT&T', 'Merck', 'Pfizer']

[Technology]cluster 7 : ['Chevron', 'Walt Disney', 'Hewlett-Packard', '3M']

[Technology]cluster 8 : ['DuPont', 'Caterpillar', 'Verizon', 'Intel']

Question 4 Clustering Exercise

1.
```
Supp Conf   Rule
0.37 0.90   Shape=4 -> Density=3
0.26 0.90   Margin=1 Density=3 -> BI-RADS=4
0.30 0.91   Margin=1 Severity=0 -> BI-RADS=4
0.30 0.91   BI-RADS=4 Margin=1 -> Severity=0
0.24 0.90   BI-RADS=5 Shape=4 -> Density=3
0.25 0.91   BI-RADS=5 Shape=4 -> Severity=1
0.30 0.90   Shape=4 Severity=1 -> Density=3
0.24 0.93   Margin=1 Density=3 Severity=0 -> BI-RADS=4
0.24 0.91   BI-RADS=4 Margin=1 Density=3 -> Severity=0
0.22 0.91   BI-RADS=5 Shape=4 Severity=1 -> Density=3
0.22 0.91   BI-RADS=5 Shape=4 Density=3 -> Severity=1
```

2. According to the data, if Margin=4,Shape=4, might represent malign. Otherwise, Margin=1 and Shape=1 or 2 might show benign.

```
Supp Conf   Rule
0.17 0.91   BI-RADS=4 Shape=1 -> Severity=0
0.16 0.90   BI-RADS=4 Shape=2 -> Severity=0
0.30 0.91   BI-RADS=4 Margin=1 -> Severity=0
0.25 0.91   BI-RADS=5 Shape=4 -> Severity=1
0.14 0.90   Shape=2 Margin=1 -> Severity=0
0.16 0.91   BI-RADS=4 Shape=1 Margin=1 -> Severity=0
0.15 0.92   BI-RADS=4 Shape=1 Density=3 -> Severity=0
0.12 0.94   BI-RADS=4 Shape=2 Margin=1 -> Severity=0
0.24 0.91   BI-RADS=4 Margin=1 Density=3 -> Severity=0
0.11 0.91   BI-RADS=5 Shape=4 Margin=4 -> Severity=1
0.22 0.91   BI-RADS=5 Shape=4 Density=3 -> Severity=1
0.12 0.91   BI-RADS=5 Margin=4 Density=3 -> Severity=1
0.14 0.90   Shape=1 Margin=1 Density=3 -> Severity=0
0.13 0.92   BI-RADS=4 Shape=1 Margin=1 Density=3 -> Severity=0
```

3. When we set support >=0.1, confidence >=0.9 and select data if BI-RADS<=2 and Severity=1,which means that the BI-RADS assessment is not always accurate. Finally find that “**BI-RADS=2 -> Severity=1**”(s=0.25 c=1.00); “**BI-**

RADS=0 Age=69 -> Severity=1”(s=0.25 c=1.00); “**BI-RADS=0 Margin=5 -> Severity=1**”(s=0.25 c=1.00).

4. “**Age=35 -> Severity=0**” support ~~= 1 and confidence =1~~.It’s valuable, because 0.01;0.92 people who is 35 years old will not have breast cancer, so they do not need to join the BI-RADS assessment. In addition, this rule also indicates that the severity value might related to the age.
5. Set minimum age = 60,“old = Orange.data.Table([d for d in age if d["Age"] >= "60"]) #Q5 change the value”. Then find rules:

```
Supp Conf   Rule
0.49 0.93 Age>=60 BI-RADS=5 -> Density=3
0.49 0.92 Age>=60 BI-RADS=5 -> Severity=1
0.53 0.95 Age>=60 Shape=4 -> Density=3
0.33 0.92 Age>=60 Margin=4 -> Density=3
0.18 0.95 Age>=60 Margin=5 -> Density=3
0.62 0.93 Age>=60 Severity=1 -> Density=3
0.13 0.93 Age>=60 BI-RADS=4 Shape=4 -> Density=3
0.14 0.90 Age>=60 Margin=1 Severity=0 -> BI-RADS=4
0.11 0.93 Age>=60 BI-RADS=4 Margin=4 -> Density=3
0.37 0.95 Age>=60 BI-RADS=5 Shape=4 -> Density=3
0.37 0.95 Age>=60 BI-RADS=5 Shape=4 -> Severity=1
0.22 0.91 Age>=60 BI-RADS=5 Margin=4 -> Density=3
0.22 0.93 Age>=60 BI-RADS=5 Margin=4 -> Severity=1
0.14 0.95 Age>=60 BI-RADS=5 Margin=5 -> Density=3
0.14 0.95 Age>=60 BI-RADS=5 Margin=5 -> Severity=1
0.46 0.94 Age>=60 BI-RADS=5 Severity=1 -> Density=3
0.46 0.93 Age>=60 BI-RADS=5 Density=3 -> Severity=1
0.11 0.93 Age>=60 Shape=4 Margin=3 -> Density=3
0.24 0.93 Age>=60 Shape=4 Margin=4 -> Density=3
0.16 0.98 Age>=60 Shape=4 Margin=5 -> Density=3
0.46 0.95 Age>=60 Shape=4 Severity=1 -> Density=3
0.12 0.94 Age>=60 Margin=3 Severity=1 -> Density=3
0.28 0.92 Age>=60 Margin=4 Severity=1 -> Density=3
0.16 0.94 Age>=60 Margin=5 Severity=1 -> Density=3
0.10 0.93 Age>=60 Margin=1 Density=3 Severity=0 -> BI-RADS=4
0.17 0.93 Age>=60 BI-RADS=5 Shape=4 Margin=4 -> Density=3
0.17 0.94 Age>=60 BI-RADS=5 Shape=4 Margin=4 -> Severity=1
0.12 1.00 Age>=60 BI-RADS=5 Shape=4 Margin=5 -> Density=3
0.11 0.94 Age>=60 BI-RADS=5 Shape=4 Margin=5 -> Severity=1
0.36 0.96 Age>=60 BI-RADS=5 Shape=4 Severity=1 -> Density=3
0.36 0.96 Age>=60 BI-RADS=5 Shape=4 Density=3 -> Severity=1
0.36 0.91 Age>=60 BI-RADS=5 Shape=4 -> Density=3 Severity=1
0.20 0.93 Age>=60 BI-RADS=5 Margin=4 Severity=1 -> Density=3
0.20 0.94 Age>=60 BI-RADS=5 Margin=4 Density=3 -> Severity=1
0.13 0.95 Age>=60 BI-RADS=5 Margin=5 Severity=1 -> Density=3
0.13 0.95 Age>=60 BI-RADS=5 Margin=5 Density=3 -> Severity=1
0.21 0.93 Age>=60 Shape=4 Margin=4 Severity=1 -> Density=3
0.14 0.98 Age>=60 Shape=4 Margin=5 Severity=1 -> Density=3
0.16 0.94 Age>=60 BI-RADS=5 Shape=4 Margin=4 Severity=1 -> Density=3
0.16 0.95 Age>=60 BI-RADS=5 Shape=4 Margin=4 Density=3 -> Severity=1
0.11 1.00 Age>=60 BI-RADS=5 Shape=4 Margin=5 Severity=1 -> Density=3
0.11 0.94 Age>=60 BI-RADS=5 Shape=4 Margin=5 Density=3 -> Severity=1
0.11 0.94 Age>=60 BI-RADS=5 Shape=4 Margin=5 -> Density=3 Severity=1
```