Fall 2016 INF2190H

Midterm Test, L0101

October 18, 2016, 6:30pm-8:00pm

This is a **closed** book and notes exam. You have **90 minutes** for a total of 30 points.

This booklet contains 8 pages, including the cover page and two pages as scratch paper at the back.

PLEASE WRITE YOUR NAME ON EACH PAGE !

Last name:____

First name:____

Student Number:____

Problem 1	__________	(out of 15)
Problem 2	__________	(out of 15)
TOTAL	__________	(out of 30)

Name: ____________________

1. **PROBLEM 1, (15 points)**
For each of the following statements, indicate whether they are true (T) of false (F). Each correct answer is worth 1 point.

(a) ____ Data Mining can only be a descriptive process.

(b) ____ Data Mining contains Clustering as one of its methods.

(c) ____ Association Rules are a method used to group similar objects together.

(d) ____ In Association Rule mining, if $X- > Y$ then $Y- > X$.

(e) ____ If the minimum support is set to 50%, the same holds for the minimum confidence.

(f) ____ The standard deviation of a set of numerical values is used to measure popularity of these values.

(g) ____ ”Mode” is the value that occurs more frequently in (categorical) data.

(h) ____ When discretizing a set of numerical values we *always* create bins with equal number of values.

(i) ____ Euclidean distance is derived from Minkowski distance.

(j) ____ Consider the following set of numbers $\{5, 5, 5, 20, 20, 20\}$. Their equi-width binning with $N = 3$ is the same as their equi-depth binning with 2 elements in each bin..

(k) ____ In k-means, k stands for the number of elements in each cluster.

(l) ____ k-means can only be performed on numerical data sets.

(m) ____ If we have three items, the total number of possible subsets is 7 (do not count the subset with no items in it)

(n) ____ Clustering is a supervised technique.

(o) ____ The value ”$Age = -20$”, indicates that our data is dirty.

2. **PROBLEM 2 (10 points)**

Circle the correct answer in the following questions.

Question 1 Which of the following is not a data mining algorithm?

(a) Ranking
(b) Clustering
(c) Association Rule Mining
(d) Classification

Question 2 Which of the following is *not* a Data Cleaning task ?

(a) Fill-in missing values
(b) Remove noisy data
(c) Remove outliers
(d) None of the above

Question 3 What is a frequent itemset ?

(a) A set of items with high confidence
(b) A set of items with high support
(c) A set of items bought in a supermarket
(d) A set of items that we store in a database

Question 4 Which of the following is a subset of the set {`nuts, bread,beer`}.

(a) {`milk, butter, nuts, beer`}
(b) {`beer, nuts`}
(c) {`diapers`}
(d) {`milk, butter, nuts, dipers`}

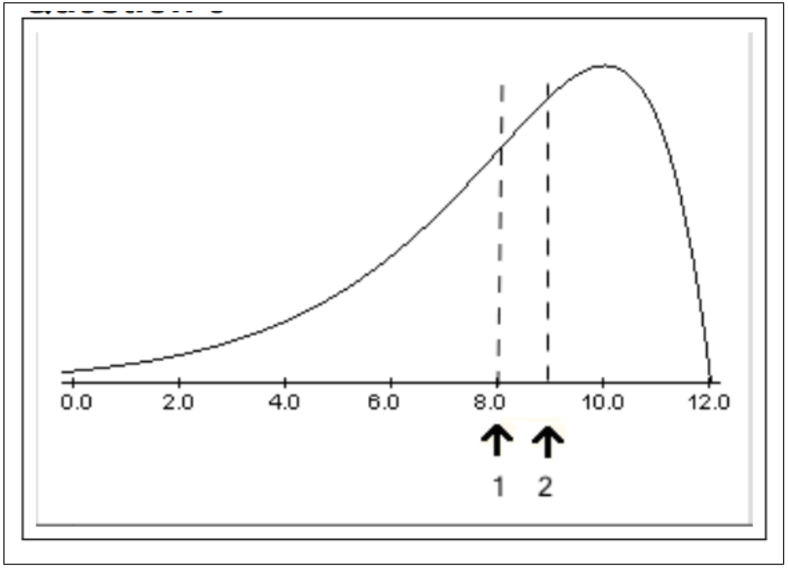
Question 5 Which of the following statements involves dirty data (inconsistencies)

(a) Age=30 and Birthday="12/12/1970
(b) Salary=-10
(c) P.Andritsos and Periklis Andritsos are two different people
(d) all of the above

Question 6 Given two items, X and Y such that $Y->X$, *Confidence* is the percentage of transactions where if Y is included then X is also included.

(a) True
(b) False

Question 7 Given the following graph of a distribution, what are the correct labels for 1. and 2. ?



(a) 1. is the ”mean” and 2. is the ”median”

(b) 1. is the ”median” and 2. is the ”mean”

(c) None of the above.

Question 8 In a normal distribution

(a) From $\mu - 2\sigma$ to $\mu + 2\sigma$, we find 99.7% of the data

(b) From $\mu - 3\sigma$ to $\mu + 3\sigma$, we find 68% of the data

(c) none of the above

Question 9 If A is a frequent itemset and B is a frequent itemset, then

(a) AB is a frequent itemset.

(b) AB is not a frequent itemset.

(c) we cannot tell if AB is also a frequent itemset

(d) non of the above

Question 10 If the minimum support is set to 50%, then

(a) the minimum confidence is set to 50% as well

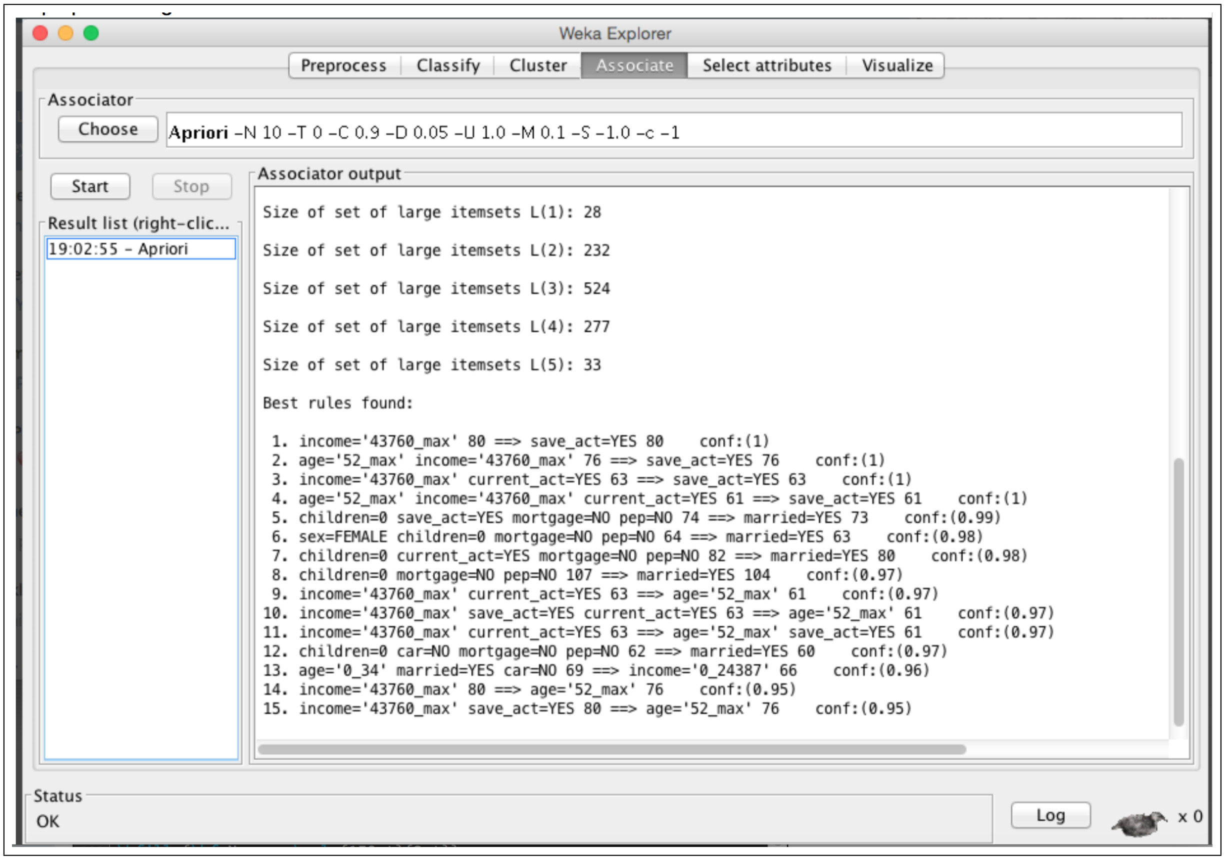
(b) the minimum confidence is set to at least 10% more

(c) we first set the minimum confidence and then the support

(d) we set confidence at a different number as we wish

(e) none of the above

Question 11 Given the following picture, explain rule 12.



Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1

19:02:55 - Apriori

```
Size of set of large itemsets L(1): 28

Size of set of large itemsets L(2): 232

Size of set of large itemsets L(3): 524

Size of set of large itemsets L(4): 277

Size of set of large itemsets L(5): 33

Best rules found:

 1. income='43760_max' 80 ==> save_act=YES 80    conf:(1)
 2. age='52_max' income='43760_max' 76 ==> save_act=YES 76    conf:(1)
 3. income='43760_max' current_act=YES 63 ==> save_act=YES 63    conf:(1)
 4. age='52_max' income='43760_max' current_act=YES 61 ==> save_act=YES 61    conf:(1)
 5. children=0 save_act=YES mortgage=NO pep=NO 74 ==> married=YES 73    conf:(0.99)
 6. sex=FEMALE children=0 mortgage=NO pep=NO 64 ==> married=YES 63    conf:(0.98)
 7. children=0 current_act=YES mortgage=NO pep=NO 82 ==> married=YES 80    conf:(0.98)
 8. children=0 mortgage=NO pep=NO 107 ==> married=YES 104    conf:(0.97)
 9. income='43760_max' current_act=YES 63 ==> age='52_max' 61    conf:(0.97)
10. income='43760_max' save_act=YES current_act=YES 63 ==> age='52_max' 61    conf:(0.97)
11. income='43760_max' current_act=YES 63 ==> age='52_max' save_act=YES 61    conf:(0.97)
12. children=0 car=NO mortgage=NO pep=NO 62 ==> married=YES 60    conf:(0.97)
13. age='0_34' married=YES car=NO 69 ==> income='0_24387' 66    conf:(0.96)
14. income='43760_max' 80 ==> age='52_max' 76    conf:(0.95)
15. income='43760_max' save_act=YES 80 ==> age='52_max' 76    conf:(0.95)
```

Status

OK

(a) There are 60 people in the data set that are not married.

(b) If there are no children and no car, and if mortgage and pep are set to NO, then there is a probability of 97% that married is set to YES.

(c) If there are no children and no car and pep is on NO, then the person is married.

Question 12 If the Manhattan distance between two objects is equal to 0.233, then

(a) the objects will be placed together in a k-means procedure

(b) their Euclidean distance is not necessarily the same

(c) the value of k will be small

(d) none of the above

Question 13 Euclidean distance

(a) is sensitive to outliers

(b) get computed between records of numerical data

(c) contains a square root

(d) all of the above

Question 14 Three frequent pairs (p, q), (r, s) and (t, u) have been found in association rule mining. What is the minumum number of rules that may be derived from these three pairs ?

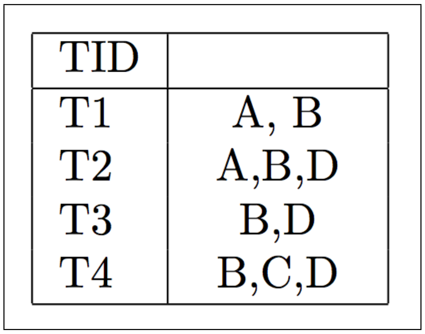
(a) 6

(b) 3

(c) 0

(d) 1

Question 15 Given the table of transactions



TID	
T1	A, B
T2	A,B,D
T3	B,D
T4	B,C,D

(a) The itemset $\{B, D\}$ has support 50%

(b) The itemset $\{C, D\}$ has support 75%

(c) Rule $B \rightarrow D$ has confidence 100%

(d) Rule $D \rightarrow B$ has confidence 100%

Name: ____________________

Scratch paper.

Name: ______________________

Scratch paper.