WEB MINING
MIDTERM Part#2
Due Date 08.05.2020

NOTE: Please answer the questions on a Word or similar program or answer it on a paper then either scan it or take a clear picture of it then email it to me at gurhangunduz@mu.edu.tr. DO NOT FORGET THAT CHEATERS WILL BE PUNISHED.

Also note that this exam will be 50% of your midterm grade. The other 50% will be the previous homework that was given to you. If you have not submitted your previous homework, I will give you another chance to submit it. But in order to provide justice to ones who submitted their homework on time, you will be able to get maximum 75 points(37.5 for midterm) from that homework. PLEASE KEEP IN MIND THAT CHEATERS WILL BE PUNISHED. For those who did not submit the previous homework, new due date is on April 30th.

Question 1 (60 pts): Given the following seven transactions and MIS(Milk) = 50%, MIS(Bread) = 70%, and 25% for all other items. And the support difference constraint is not used. **F1 = {{Beef}, {Cheese}, {Clothes}, {Bread}}** is given to you by running the first 3 lines of MS-Apriori algorithm. Find C2, F2, C3 and F3. Show how you find them!

Beef, Bread
Bread, Clothes
Bread, Clothes, Milk
Cheese, Boots
Beef, Bread, Cheese, Shoes
Beef, Bread, Cheese, Milk
Bread, Milk, Clothes

Algorithm MS-Apriori(T, MS, φ) // MS stores all MIS values
1 $M \leftarrow$ sort(I, MS); // according to MIS(i)'s stored in MS
2 $L \leftarrow$ init-pass(M, T); // make the first pass over T
3 $F_1 \leftarrow \{\{l\} \mid l \in L, l.count/n \geq \text{MIS}(l)\}$; // n is the size of T
4 **for** ($k = 2$; $F_{k-1} \neq \emptyset$; k++) **do**
5 **if** $k = 2$ **then**
6 $C_k \leftarrow$ level2-candidate-gen(L, φ) // $k = 2$
7 **else** $C_k \leftarrow$ MScandidate-gen(F_{k-1}, φ)
8 **endif**;
9 **for** each transaction $t \in T$ **do**
10 **for** each candidate $c \in C_k$ **do**
11 **if** c is contained in t **then** // c is a subset of t
12 $c.count$++
13 **if** $c - \{c[1]\}$ is contained in t **then** // c without the first item
14 $(c - \{c[1]\}).count$++
15 **endfor**
16 **endfor**
17 $F_k \leftarrow \{c \in C_k \mid c.count/n \geq MIS(c[1])\}$
18 **endfor**
19 **return** $F \leftarrow \cup_k F_k$;

Function **level2-candidate-gen(L, φ)**
1 $C_2 \leftarrow \emptyset$; // initialize the set of candidates
2 **for** each item l in L in the same order **do**
3 **if** $l.count/n \geq$ MIS(l) **then**
4 **for** each item h in L that is after l **do**
5 **if** $h.count/n \geq$ MIS(l) and $|sup(h) - sup(l)| \leq \varphi$ **then**
6 $C_2 \leftarrow C_2 \cup \{\{l, h\}\}$; // insert the candidate $\{l, h\}$ into C_2

Function **MScandidate-gen(F_{k-1}, φ)**

```
1   C_k ← ∅; // initialize the set of candidates
2   forall f_1, f_2 ∈ F_k // find all pairs of frequent itemsets
3     with f_1 = {i_1, … , i_{k–2}, i_{k–1}} // that differ only in the last item 4 and f_2 = {i_1, … , i_{k–2}, i'_{k–1}}
4       and i_{k-1} < i'_{k–1} and |sup(i_{k-1}) – sup(i'_{k–1})| ≤ φ do
5       c ← {i_1, …, i_{k–1}, i'_{k–1}}; // join the two itemsets f_1 and f_2
6       C_k ← C_k ∪ {c}; // insert the candidate itemset c into C_k
7       for each (k–1)-subset s of c do
8         if (c[1] ∈ s) or (MIS(c[2]) = MIS(c[1])) then
9           if (s ∉ F_{k–1}) then
10            delete c from C_k; // delete c from the set of candidates
11      endfor
12    Endfor
13    return C_k; // return the generated candidates
```

Question 2 (40pts): Suppose that we have the training data set in the Figure below, which has two attributes Attr1 and Attr2, and the Class. Compute all the probability values required to learn a naïve Bayesian classifier. Then predict the class of the following attributes

Attr1=x1, Attr2=x6 Class=?

Attr1=x2, Attr2=x4 Class=?

Attr1	Attr2	Class
x1	x4	T
x1	x5	T
x2	x6	T
x3	x5	T
x2	x6	T
x2	x6	F
x2	x5	F
x3	x4	F
x3	x6	F
x1	x4	F